La Bibbia di Sistemi operativi

Mario Petruccelli
Università degli studi di Milano

A.A. 2018/2019

Sommario

1 Virtualization

1.1 Introduzione

Processi Un processo, informalmente, è un programma in esecuzione. Un programma a sua volta, è una sequenza finita di istruzioni scritte in un linguaggio comprensibile all'esecutore (CPU). L'esecuzione di un programma da parte del processore è:

- **Fetch** Prelievo istruzione dalla memoria.
- **Decode** Decodifica dell'istruzione.
- **Execute** Esecuzione dell'istruzione.

1.1.1 Virtualizzazione

La virtualizzazione consiste nel prendere una risorsa fisica e trasformarla in una più generale, potente e facile da adoperare forma virtuale di se stessa.

Virtualizzazione della CPU L'illusione consiste nel far credere che il sistema abbia un elevato numero di cpu virtuali. Avere più CPU permeterrebbe a più programmi di essere eseguiti in **parallelo** nonostante il processore fisico effettivo sia uno solo. Se due processi vogliono essere eseguiti entrambi ad un certo tempo, oppure vogliono accedere alla stessa periferica, quale dei due ha la priorità? La risposta viene data con l'introduzione delle politiche di priorità (**politiche di scheduling**).

Virtualizzazione della memoria Consiste nel fabbricare l'illusione che ogni processo abbia il proprio spazio di indirizzi virtuali privato (**address space**) al quale accede e sarà il sistema operativo ad occuparsi di mappare nella memoria fisica.

Con la virtualizzazione è fondamentale riuscire a distinguere i processi in esecuzione. Per fare ciò viene associato un **PID** (process id) ad ogni job. il PID è un numero univoco.

1.1.2 Concorrenza

Si riferisce a tutta quelle serie di problemi che sorgono, e che vanno risolti, quando all'interno dello stesso programma più entità lavorano in parallelo. Le entità in questione si chiamano **threads**.

1.1.3 Persistenza

La persistenza è legata alla memorizzazione dei dati all'interno della memoria. La non volatilità delle memorie ha introdotto la possibilità di memorizzare dati in modo persistente. Il software nel sistema operativo che generalmente gestisce i dischi è chiamato **file system**.

1.1.4 Protezione ad anelli

Un modello di potezione implementato dal sistema operativo è quello ad anelli. Ci sono 5 livelli e 3 anelli differenti. A ciascun anello corrisponde un relativo livello di sicurezza.

- **Level 1** *Hardware level* qui vengono eseguiti, ad esempio, i device drivers visto che essi richiedono accesso diretto all'hardware dei dispositivi (microcontroller).
- **Level 2** *Firmware level* Il firmware sta in cima al livello elettronico. Contiene in software necessario dal dispositivo hardware e dal microcontroller.
- **Level 3: ring 0** *Kernel level* Questo è il livello dove opera il kernel, dopo la fase di bootload siamo qui.
- **Level 4: ring 1 e 2** *Device drivers* I device drivers passano atraverso il kernel per accedere all'hardware.
- **Level 5: ring 3** *Application level* Qui è dove viene eseguito normalmente il codice utente.

1.2 Processi

Sistema multiprogrammato Sistema nel quale è possibile eseguire più programmi contemporaneamente, idea alla base della virtualizzazione.

1.2.1 Multiprogrammazione

Time sharing prevede che il tempo di CPU sia equamente diviso fra i programmi in memoria.

Real time sharing La politica di scheduling è differente. Alcuni processi vanno serviti prima di altri.

1.2.2 Virtualizzazione della CPU

L'illusione consiste nel rendere indipendenti il numero di processi dal numero di processori. Si vuole disaccoppiare le entità logiche (*processi*), dalle entità fisiche (*processori*), in modo tale che ad ogni processo venga assegnato un processore logico mappato su processore fisico.

I concetti fondamentali alla base della virtualizzazione sono:

- **Time sharing** Meccanismo mediante il quale il tempo di CPU viene diviso equamente fra i processi.
- **Context switch** Meccanismo che consente di interrompere l'esecuzione di un processo in corso sulla CPU fisica e assegnare quest'ultima ad un nuovo processo.

1.2.3 Processi

Un processo è un programma in esecuzione, il sistema operativo deve fornire alcune interfacce (**APIs**) per la gestione dei processi che permettano di fare:

- **Create** Creazione di un nuovo processo.
- **Destroy** Eliminazione forzata di un processo. Molti processi termineranno per conto loro, ma l'utente potrebbe voler eliminare processi non ancora terminati.
- **Wait** Mette in attesa un processo.
- **Miscellaneous control** Sospensione di un processo per farlo ripartire dopo un certo tempo.
- **Status** Interfacce che restituiscono lo stato e altre informazioni di un processo.

Creazione di un processo La prima cosa che deve fare il sistema operativo per eseguire un programma è caricare il suo codice ed eventuali dati statici da disco a memoria, nell'address space del processo.

- **Allocazione dello stack** Un po' di memoria deve essere creata per lo stack del programma (*variabili locali, parametri delle funzioni e indirizzi di ritorno*).
- **Allocazione dello heap** Un po' di memoria deve essere creata per lo heap del programma (*dati allocati dinamicamente*).
- **Inizializzazione I/O** Standard input, output ed error.
- **Salto ed esecuzione** Salto all'entry point ed esecuzione. (*main*)

Stato di un processo

- **Running** È in esecuzione sul processore.
- **Ready** In attesa di essere eseguito dal processore.
- **Blocked** In stato di block, il processo sta esegundo qualche operazione (*es: I/O*).

Strutture dati Il sistema operativo deve tenere traccia delle informazioni fondamentali di un processo per poter ripristinare l'esecuzione di un processo interrotto. Esse sono:

- Porzioni di memoria coinvolte.
- Valori dei registri di CPU usati dal processo.
- Stato dei dispositivi di I/O usati dal processo.

Questi dati sono organizzati in strutture chiamate **Process Control Block (PCB)**, salvate in un per-process **kernel stack**, il quale risiede nel kernel space.

1.2.4 Process API

La creazione di un processo avviene tramite la `fork()`, la quale genera un processo identico a quello in esecuzione. Tale processo prende il nome di padre, quello generato viene chiamato figlio. L'esecuzione del processo figlio parte dall'istruzione successiva alla `fork()`. La `fork()` ritorna al figlio 0, al padre il **PID** del figlio e **-1** in caso di errore. Il processo figlio avrà il **proprio** address space, registri, PC, ecc...

La `wait()` è una funzione che forza il padre ad aspettare che il processo figlio termini la propria esecuzione. Senza, l'output potrebbe essere **non-deterministico** e potrebbero crearsi processi orfani o zombie. Esiste anche la `waitpid` che viene usata se si ha a che fare più di un figlio.

Per eliminare un processo esiste la funzione `kill()`. Solo il padre può distruggere il figlio. Ciò può portare alla creazione di processi **zombie** (processi terminati la cui **PCB** è ancora in memoria).

La `exec()` serve per generare un processo che fa qualcosa di diverso da quello padre. `exec(nome_programma, arg)` prende il nome di un eseguibile e alcuni argomenti, carica il codice e i dati statici di quell'eseguibile, sovrascrivendo il code segment corrente all'interno del PCB del figlio. Heap, stack e altre parti di memoria vengono re-inizializzate. Rimane la relazione padre-figlio.

1.3 Context Switch

1.3.1 Shell

Come mai `fork()` ed `exec()` sono due system call separate? Per rispondere introduciamo la **shell**.

La shell è un programma del sistema operativo `Unix` il cui compito è riconoscere ed eseguire altri programmi; si può dire che essa sia il genitore di tutti i processi che vengono mandati in esecuzione. Nello specifico, essa esegue una `fork()`, cambia il file descriptor se richiesto, ed infine invoca la `exec()`. Poi si mette in attesa che il programma abbia terminato prima di tornare in attesa di istruzioni. Esistono due tipi di shell, grafica (*terminale*) e interattiva (*aprire programmi col mouse*).

La separazione di `fork()` ed `exec()` è dovuta alla presenza della shell, con la quale possiamo andare ad effettuare alcune modifiche dopo la `fork()` e prima dell'`exec()`, come ad esempio la sostituzione del file descriptor.

```
$> wc file.c > n.txt
```

La shell esegue la `fork()` per poter mandare in esecuzione il programma `wc`. Prima di sostituire il codice del padre all'interno del PCB del figlio, sostituisce il file descriptor relativo allo standard output con n.txt. Successivamente esegue l'`exec()` producendo

l'output desiderato all'interno di n.txt. Queste manipolazioni non sarebbero possibili se `fork()` ed `exec()` fossero un'unica system call perchè non si avrebbe accesso al PCB del figlio prima dell'exec().

1.3.2 Direct execution

Il concetto di direct execution è semplice: il programma viene eseguito direttamente sulla CPU fisica.

Quando il sistema operativo desidera iniziare l'esecuzione di un programma, viene fatto quanto segue:

- Crea una entry nella lista dei processi.
- Alloca la memoria per il programma.
- Carica il programma in memoria.
- Imposta lo stack con `argc/argv`.
- Pulisce i registri.
- Esegue la chiamata a `main()`.
  Si ha un salto dalla zona kernel al `main`. Il processo a questo punto deve:
- Eseguire il codice del `main()`.
- Ritornare dal main a fine esecuzione.
  Dal processo si torna alla zona kernel. Il sistema operativo infine:
- Rimuove la entry dalla lista dei processi.

Tuttavia la direct execution solleva alcune problematiche:

- Il sistema operativo non può assicurarsi che un programma in esecuzione non faccia qualcosa che non dovrebbe fare.
- Il sistema operativo non può fermare un processo in esecuzione.

Il primo problema si risolve con l'introduzione dello **user mode**. Il codice che viene eseguito in questa modalità di elaborazione è limitato in termini di istruzioni eseguibili. Nasce quindi anche la **kernel mode**, modalità in cui opera il sistema operativo e che consente di eseguire tutte le istruzioni privilegiate.

Per permettere ad un processo di eseguire istruzioni privilegiate vengono introdotte delle **system call**. Per eseguirle, un programa deve eseguire un'istruzione **trap** (*interrupt via software*). Questa istruzione salta nel kernel, aumenta i privilegi a kernel mode, esegue le operazioni privilegiate e ritorna al processo scalando i privilegi tramite un'istruzione **return-from-trap**. Durante questo procedimento bisogna assicurarsi di

salvare i registri del chiamante. Per sapere dove la trap deve saltare, il kernel imposta una **trap table** al boot time. Non è il processo utente a specificare l'indirizzo dei **trap handlers** perchè potrebbe saltare ovunque nel sistema.
Per specificare la system call, generalmente viene assegnato un **system-call-number** che solitamente viene inserito in un registro appropriato.

1.3.3 Switch tra processi

Cooperative approach Soluzione via software che consiste nel programmare il processo in modo che, dopo un certo numero di secondi di utilizzo della CPU, il comando torni al sistema operativo. Il problema è che se vengono creati loop infiniti nel programma, la CPU non verrebbe mai condivisa.

Time interrupt Soluzione via hardware che consiste nel creare una nuova componente che genera un segnale elettrico (**time interrupt**) dopo un certo lasso di tempo. Ci sarà quindi un orologio interno che invierà un segnale al piedino del microprocessore. L'hardware deve inoltre fermare l'esecuzione del processo corrente, salvarne lo stato per dare il controllo allo **scheduler**, che nel caso decidesse di cambiare processo, farà eseguire al sistema operativo codice a basso livello che prende il nome di **context switch**.

Context switch Ciò che deve fare il sistema operativo è salvare alcuni valori dei registri per il processo in corso di esecuzione (*nel kernel stack*) e ripristinarne altri per il processo scelto. Viene eseguita una return-from-trap per mandare in esecuzione il processo scelto.
Interrupt, system call ed eccezioni sono eventi che inducono il mode switch.

1.4 Scheduling policy

Dati n processi, a quale assegno il processore? La scelta è fatta dallo **scheduler**, un modulo del sistema operativo che implementa una politica decisionale.

CPU burst è l'intervallo di tempo in cui viene usata intensamente la CPU.

I/O burst è l'intervallo di tempo in cui viene usato intensamente I/O.

CPU bound processi con CPU burst lunghi, ad esempio compilatori, simulatori, calcolo del tempo, ecc...

I/O Bound processi con I/O burst lunghi, ciò comporta maggiore interattività con l'utente.

Stato di IDLE è lo stato in cui è una risorsa accesa e funzionante ma non utilizzata.

Un processo in esecuzione si trova o in CPU burst o in I/O burst. Lo scheduler, per essere efficiente, deve ottimizzare l'uso delle risorse in modo tale che, se la CPU è occupata con l'esecuzione di un processo, i dispositivi di I/O lo sono con un altro e viceversa. L'ottimizzazione della CPU viene dunque portata mediante lo scheduler. Per valutare la bontà di un algoritmo di scheduling si devono introdurre delle metriche di valutazione.

$$T_{turnaround} = T_{termine} - T_{arrivo}$$

$$T_{response} = T_{first-exec} - T_{arrivo}$$

$$T_{wait} = T_{turnaround} - T_{job}$$

1.4.1 Algoritmo FIFO

L'algoritmo FIFO (*First In First Out*) mette in esecuzione il primo processo arrivato. Il problema a cui può portare questo algoritmo è l'**effetto convoglio**, ovvero quando un certo numero di piccoli consumatori di una risorsa vengono messi in coda dietro un enorme consumatore.

1.4.2 Algoritmo SJF

L'algoritmo SJF (*Shortest Job First*) mette in esecuzione il processo con CPU burst minore. In questo modo si evita l'effetto convoglio, ma solo se i processi arrivano allo stesso istante.

1.4.3 Algoritmo STCF

L'algoritmo STCF (*Shortest Time to Completion First*) ogni volta che arriva un processo, lo compara il processo in esecuzione e lascia il processore a quello che ha CPU burst minore.

SJF e STCF funzionano molto male per quanto riguarda il tempo di risposta (spesso possono indurre anche al verificarsi della starvation). Inoltre non conoscono a priori il CPU burst di un processo, perciò sono solo algoritmi teorici.

1.4.4 Round Robin

L'algoritmo **Round Robin** assegna un **quanto di tempo** ad ogni processo. Viene inizializzato un timer che, una volta arrivato a zero, forza un context switch. Il quanto di tempo va scelto bene, altrimenti si hanno troppi context switch se è troppo piccolo, o degenera in FIFO se è troppo grande.

1.5 Multilevel feedback scheduler

Il problema che **MLFQ** (*MultiLevel Feedback Queue*) cerca di risolvere è:

- Ottimizzare il $T_{turnaround}$.
- Aumentare l'interattività utente/sistema, minimizzando il $T_{response}$.

L'approccio che si usa consiste nell'avere un certo numero di **code** distinte, ognuna assegnata ad un diverso **livello di priorità**. MLFQ sfrutta i diversi livelli di priorità per decidere quale processo eseguire: viene scelto quello all'interno della coda di priorità maggiore. Se ci sono più processi all'interno di una certa coda, viene usato **RR**.

- 1. If priority(A) > priority(B), A runs (B doesn't).
- 2. If priority(A) = priority(B), A & B run in RR.
- 3. Quando un processo entra nel sistema, viene posizionato nella coda di priorità massima.
- 4a. Se un processo utilizza tutto il lasso di tempo a disposizione durante l'esecuzione, la sua priorità viene ridotta.
- 4b. Se un processo libera la CPU prima di terminare il lasso di tempo a disposizione, il livello di priorità rimane invariato.
- 5. **Priority boost** Dopo un certo periodo di tempo, tutti i processi vengono spostati nella coda di priorità più alta. (*Evita la starvation dei long running jobs e il monopolio della CPU se qualche processo la rilascia poco prima del lasso di tempo.*)

1.5.1 Better accounting

La scelta del tempo è cruciale, se settato troppo grande, i long running jobs potrebbero ancora andare in starvation, se impostato troppo piccolo, i processi interattivi potrebbero non avere una porzione adeguata dela CPU. Per evitare che possa essere raggirato l'algoritmo di scheduling, lo scheduler tiene traccia di quanto tempo ha consumato un processo in un certo livello di **MLFQ**. Le regole 4a e 4b diventano:

- 4. Una volta che un processo ha usato il tempo a disposizione in un certo livello (indipendentemente da quante volte ha rilasciato la CPU), la sua priorità viene ridotta.

1.6 Address space

Un programma per essere eseguito deve risiedere in memoria. Essa può essere usata implicitamente (**stack:** `int x`) o esplicitamente (**heap:** `int *x = malloc(sizeof(int))`). Lo stack è gestito autonomamente, lo heap è gestito dal programmatore attraverso opportune funzioni (`malloc, realloc, free, ...`), per cui non si conosce a priori la dimensione.

1.6.1 Memory API

- `malloc()` Riceve in input un argomento di tipo `size_t` (numero di bytes), se ha successo restituisce un puntatore all'inizio della zona allocata nello **heap**, se fallisce restituisce NULL.
- `free()` Riceve in input un puntatore, la grandezza della regione da liberare viene tenuta nella libreria `memoryallocation`.

La system call per la gestione diretta della memoria è `int brk(void *addr)`

1.6.2 Memory errors

- **Dimenticarsi di allocare la memoria.** È da patchare il fatto che si possa indurre un `segfault` in modo tale da poter accedere al core dump della memoria e vedere dati sensibili (**attacchi core-dump**). È necessario eliminare questi dati dopo il loro utilizzo.
- **Non allocare abbastanza memoria.** Può portare a vulnerabilità come il **buffer overflow**.
- **Dimenticarsi di inizializzare memoria allocata.** Potrebbero esserci valori come 0 o valori random.
- **Dimenticarsi di liberare la memoria.** Il **memory leak** può portare ad un esaurimento della memoria disponibile.
- **Liberare la memoria prima di aver finito di usarla.** Questo errore è chiamato **dangling pointer**, può causare un crash o la sovrascrittura di memoria valida.
- **Liberare la memoria più di una volta.** Problema noto come **double free**, il risultato è indefinito, la libreria `memory-allocation` potrebbe confondersi e fare cose strane. I crash sono la cosa più comune.
- **Chiamata di `free()` incorretta.** La funzione si aspetta un puntatore prodotto in precedenza da una `malloc()`. Quando viene passato alla `free` un valore diverso, possono succedere cose brutte e pericolose.

1.6.3 Virtualizzazione della memoria

Con l'avvento della **multiprogrammazione** la memoria diviene una risorsa condivisa, bisogna iniziare a far fronte a tutte le problematiche che ciò comporta.

- **Protezione** un processo non può invadere lo spazio di un altro.
- **Interattività** Ci devono essere molti processi in esecuzione.

Il meccanismo di astrazione che si vuole implementare prende il nome di **address space**, esso è il punto di vista di un processo sulla memoria del sistema, ovvero l'astrazione che il sistema operativo gli fornisce.
Gli obiettivi della virtualizzazione della memoria sono riassunti come segue:

- **Trasparenza.** Il programmatore scrive il codice indipendentemente dalla grandezza della memoria.
- **Efficienza.** Il meccanismo di virtualizzazione non deve avere overhead troppo elevato.
- **Protezione.** Bisogna proteggere i processi da altri processi, dal sistema operativo, e viceversa.

1.6.4 Mapping

Il **mapping** consiste nel trovare una corrispondenza fra indirizzo logico e indirizzo fisico. Nei sistemi **monoprogrammati** ciò era facile poichè ogni programma veniva mappato a partire dall'indirizzo `64KB` fino alla fine. Il compilatore assegnava ai programmi indirizzi costanti. Nel caso della **multiprogrammazione** invece, il compilatore assegna indirizzi preliminarli al programma, i quali vengono successivamente rilocati.

1.6.5 Base e Bound

Questa tecnica di mapping utilizza due registri, **base** e **bound**. Assunzioni:

- Il programma viene caricato in locazioni contigue di memoria. (Un programma da `32KB` verrà caricato in `32KB` locazioni adiacenti)
- L'indirizzo logico è sempre minore dell'indirizzo fisico.

Mediante la rilocazione siamo in grado di calcolare l'indirizzo fisico come segue:

```
indirizzo fisico = indirizzo logico + Base
```

Base è un registro contenente il punto di partenza (indirizzo fisico) del programma. **Bound** è il registro limite. Se un processo prova a saltare in zone di un altro processo viene generato un errore di segmentazione.

1.6.6 MMU

MMU sta per **Memory Managment Unit** ed è una componente hardware per la rilocazione degli indirizzi. L'input è un indirizzo logico prodotto dalla CPU, l'output è l'indirizzo fisico. Generalmente questa traduzione viene fatta a runtime. Prima di eseguire l'istruzione a cui sto puntando, l'indirizzo logico viene tradotto in indirizzo fisico (**rilocazione dinamica**).
Ore che abbiamo la rilocazione dinamica, il sistema operativo deve fare le seguenti cose per implementare la memoria virtuale:

- Quando un nuovo processo viene creato, il sistema operativo dovrà cercare in una struttura dati (spesso chiamata **free list**) spazio libero per il nuovo address space e marcarlo come in uso.
- Quando un processo termina, deve riabilitare tutta la memoria allocata per il processo all'interno della free list e pulire ogni struttura dati associata ad esso.
- Quando avviene un context switch deve salvare nel PCB i registri base e bound e ripristinare quelli del nuovo processo.
- Quando un processo viene fermato è possibile muovere un address space da una locazione di memoria a un'altra. Basta deschedularlo, copiare l'address space dalla locazione corrente a quella nuova e infine aggiornare il registro **base**.

Il sistema operativo deve fornire degli **exception handler**. Per esempio, se un processo prova ad accedere a memoria al di fuori del suo **bound**, la CPU deve sollevare un'eccezione.

1.7 Segmentazione

1.7.1 Binding

Durante il processo di rilocazione vengono cambiati tutti gli indirizzi del programma per evitare che vadano fuori dallo spazio di indirizzamento previsto. Il **binding** è l'operazione che viene fatta per modificare gli indirizzi. Può essere:

- **Early binding.** Rilocazione degli indirizzi fatta a **compile time**. Il compilatore deve conoscere la posizione di partenza del programma in memoria, ma funziona solo quando il compilatore genera direttamente il codice assoluto (*sistemi embedded, monoprogrammati, ...*).
- **Delayed binding.** La rilocazione degli indirizzi viene fatta durante il trasferimento del programma da disco a memoria (*operazione svolta dal sistema operativo prima dell'introduzione dell'MMU*).
- **Late binding.** La rilocazione degli indirizzi viene fatta immediatamente prima di eseguire l'istruzione corrente, quindi a **runtime**. Per implementare questa tecnica serve l'MMU.

1.7.2 Segmentazione

Con la tecnica base e bound, c'è dello spazio potenzialmente non utilizzato tra lo stack e lo heap. L'idea alla base della **segmentazione** è quella di dividere il programma in **segmenti** che possono essere caricati in porzioni di memoria differenti siccome ad ognuno di essi è associata una coppia base-bound. I segmenti sono inseriti in modo indipendente all'interno della memoria fisica, in questo modo siamo in grado di evitare gli sprechi. Questo risparmio di memoria, tuttavia, complica notevolmente l'MMU, la quale deve gestire più segmenti presenti all'interno della memoria (ogni processo ha tre segmenti).
Il meccanismo funziona come segue:

- **Input:** indirizzo logico B (prodotto dal compilatore).
- Individua il segmento s di appartenenza dell'indirizzo B.
- Calcola l'offset k sottraendo all'indirizzo virtuale l'indirizzo di partenza (logico) del segmento ($k = B$ - indirizzo iniziale di s)
- Viene calcolato l'indirizzo fisico sommando k e il base register (Indirizzo fisico = $Base(s) + k$)

Se un processo cerca di produrre un indirizzo illegale, l'hardware rileverà che l'indirizzo è out of bounds, trap nel sistema operativo, il quale terminerà il processo (**segmentation fault**).
L'hardware per conoscere il segmento e l'offset taglia l'address space in segmenti basati sui primi bit dell'indirizzo virtuale (**approccio esplicito**). Nell'**approccio implicito** invece l'hardware determina il segmento in base a come è formato l'indirizzo. Se, ad esempio, l'indirizzo è stato generato dal program counter, appartiene al code segment; se è dello stack o del base pointer, deve appartenere al segmento stack. Ogni altro indirizzo viene interpretato come parte del segmento heap.

1.7.3 Stack

Siccome lo stack cresce al contrario, invece dei soli valori base e bound, l'hardware ha bisogno di sapere in quale direzione cresce il segmento (un bit settato a 1 se il segmento cresce positivamento, 0 negativamente). Il controllo del bound register viene fatto in valore assoluto.

1.7.4 Permessi

Code sharing. Per risparmiare memoria, a volte è utile condividere certi segmenti tra gli address spaces. Per supportare la condivisione abbiamo bisogno di **protection bits** da parte dell'hardware. Vengono aggiunti solamente pochi bit per segmento, a indicare quando un programma può leggerne, scriverne o eseguirne il codice contenuto.

1.7.5 Coarse grained and fine grained

Gli esempi visti fin'ora utilizzavano la tecnica **coarse grained** (poche fette relativamente grandi). Alcuni dei primi sistemi erano più flessibili e permettevano che gli address spaces consistessero in un gran numero di piccoli segmenti, questo concetto era espresso come segmentazione **fine grained**. Ciò richiede un ulteriore supporto hardware, una **segment table** all'interno della memoria.

1.7.6 Frammentazione

La segmentazione solleva un numero di nuove problematiche:

- Cosa dovrebbe fare il sistema operativo a fronte di un context switch? I segment registers devono essere salvati e ripristinati.
- Come viene gestito lo spazio libero in memoria fisica? Quando un nuovo address space viene creato, il sistema operativo deve essere in grado di trovare lo spazio in memoria fisica per i suoi segmenti.

Il problema generale è che la memoria fisica consuma velocemente piccoli spazi liberi, rendendo difficile l'allocazione di nuovi segmenti o la crescita di quelli già esistenti. Questo problema è noto come **frammentazione esterna**. Si può risolvere con la **deframmentazione**, compattando la memoria fisica e riarrangiando i segmenti esistenti, copiando i dati dei segmenti in una regione contigua di memoria e cambiando il valore dei loro segment registers. Questa operazione è piuttosto complessa e dispendiosa oltre che bloccante (non si possono eseguire processi in quanto il loro indirizzo sta cambiando). Un approccio più semplice è quello di usare un algoritmo per la gestione della **free-list** che tenta di mantenere un elevato spazio disponibile contiguo in memoria. Purtroppo però la frammentazione esisterà sempre a prescindere da quanto buono sia l'algoritmo per minimizzarla.

1.8 Paginazione

La **paginazione** nasce per gestire in modo ottimale lo spazio libero in memoria e l'address space di un programma. Consiste nel tagliare gli spazi in fette di una certa dimensione. Anzichè dividere l'address space di un processo in segmenti, esso viene diviso in unità di dimensione fissata, ognuna delle quali è chiamata pagina.
Vediamo la memoria fisica come un array di slots di dimensione fissata, chiamati **page frames**. Ogni frame può contenere una singola pagina di memoria virtuale. Ciò porta ad alcuni vantaggi:

- **Flessibilità.** Il sistema sarà in grado di supportare l'astrazione dell'address space efficacemente, a prescindere da come un processo ne fa uso. Non vogliamo, ad esempio, dover fare assunzioni riguardo la direzione di crescita dello heap e dello stack e come vengono usati.

- **Semplicità** della gestione dello spazio libero. Per esempio, supponiamo che il sistema operativo desideri inserire il nostro addess space da `64B` in memoria fisica. Siccome i programmi sono divisi in pagine di dimensione fissata, il problema della segmentazione viene ridotto di molto visto che, siccome il sistema operativo tiene traccia della free list, gli basta semplicemente prendere il primo frame disponibile e assegnarlo a una pagina.

Per memorizzare dove ogni pagina virtuale dell'address space è posizionata in memoria fisica, il sistema operativo tiene una struttura dati per ciascuno processo nota come **page table**. Il ruolo principale della page table è di memorizzare, per ogni pagina virtuale dell'address space, il corrispondente frame fisico.

1.8.1 Address translation

Per tradurre l'indirizzo virtuale generato da un processo, dobbiamo per prima cosa dividerlo in **Virtual Page Number (VPN)** e **offset**. Siccome si conosce la dimensione di ciascuna pagina, si può dividere l'indirizzo virtuale in:

- **VPN:** Bit più significativi che fanno da indice per accedere alla page table del processo per trovare il frame fisico corrispondente (**PFN**).
- **Offset:** Bit che servono per indirizzare la grandezza di una pagina.

A questo punto si traduce l'indirizzo virtuale in fisico sostituendo il **Physical Frame Number (PFN)** al VPN.

1.8.2 Page tables

Le page tables possono essere terribilmente grandi. Per esempio, immaginiamo un address space da `32 bit` con pagine da `4KB`. L'indirizzo virtuale sarà diviso in `20 bit` di VPN e `12 bit` di offset. `20 bit` di VPN implicano 2^{20} possibili traduzioni per ogni processo. Assumendo di aver bisogno di `4B` per **page table entry (PTE)** per mantenere la traduzione fisica più ogni altra informazione utile otteniamo `4MB` di memoria necessari per ogni page table. Con 100 processi in esecuzione, questo significa che il sistema operativo avrà bisogno di **400MB** di memoria.

Cosa contiene una page table? La page table è sempicemente una struttura dati usata per mappare gli indirizzi virtuali in indirizzi fisici. La forma più semplice è chiamata **page table lineare** che è semplicemente un array. Il sistema operativo indicizza l'array con il VPN e consulta la PTE a quell'indice per trovare il PFN desiderato.
Ogni PTE contiene diversi bit:

- **Valid bit.** Indica quando una particolare traduzione è valida. Per esempio, quando un programma inizia l'esecuzione, avrà code e heap a un'estremità del suo spazio di indirizzamento e lo stack dall'altra. Tutto lo spazio non utilizzato

in mezzo sarà marcato come invalido e se il processo tenterà di accedervi, verrà generata una trap al sistema operativo che lo terminerà. È cruciale per supportare un address space sparso.

- **Protection bits.** Indicano quando una pagina può essere letta, scritta o eseguita. Accedere a una pagina in modo non consentito da questi bit genererà una trap nel sistema operativo, il quale terminerà il processo.
- **Present bit.** Indica se la pagina in questione è in memoria fisica o su disco. Consente al sistema operativo di swappare le pagine liberando la memoria fisica.
- **Dirty bit.** Indica se la pagina è stata modificata da quando risiede in memoria.
- **Reference bit.** Viene usato per tenere traccia se una pagina è stata acceduta da quando risiede in memoria.

1.8.3 Quanto è lenta la paginazione?

Per ogni riferimento a memoria (sia per prelevare un'istruzione che per un load o store esplicito), la paginazione ne necessita uno aggiuntivo per prelevare la traduzione dalla page table. I riferimenti a memoria aggiuntivi sono costosi e in questo caso rallenteranno il processo di un fattore pari a due o più.

1.9 Translation Lookaside Buffer

Siccome le informazioni di mappatura risiedono generalmente in memoria fisica, la paginazione richiede un accesso aggiuntivo per ogni indirizzo virtuale generato dal programma. L'obbiettivo è snellire la tecnica introdotta, cercando di **diminuire il numero di accessi a memoria fisica** (alla page table). Viene aggiunta alla MMU una cache hardware delle traduzioni virtual-to-physical più popolari chiamata **translation lookaside buffer o TLB**. Per ogni indirizzo virtuale, l'hardware controlla per prima cosa il TLB per vedere se la traduzione desiderata è presente al suo interno.

```
1  VPN = (VirtualAddress & VPN_MASK) >> SHIFT
2  (Success, TlbEntry) = TLB_Lookup(VPN);
3  if (Success == True){ //TLB HIT
4    if (CanAccess(TlbEntry.ProtectBits == True){
5      Offset = VirtualAddress & OFFSET_MASK;
6      PhysAddr = (TlbEntry.PFN << SHIFT) | Offset;
7      Register = AccessMemory(PhysAddr);
8    }
9    else
10     RaiseException(PROTECTION_FAULT);
11 }
12 else{              //TLB MISS
13   PTEAddr = PTBR + (VPN * sizeof(PTE));
14   PTE = AccessMemory(PTEAddr);
15   if(PTE.Valid == False)
```

```
16      RaiseException(SEGMENTATION_FAULT);
17    else if (CanAccess(PTE.ProtectBits) == False)
18      RaiseException(PROTECTION_FAULT);
19      else{
20        TLB_Insert(VPN, PTE.PFN, PTE.ProtectBits);
21        RetryInstruction();
22      }
23  }
```

L'algoritmo che l'hardware segue funziona in questo modo:

- Estrae il VPN dall'indirizzo virtuale.
- Controlla se il TLB contiene la traduzione per il VPN. Se così fosse, abbiamo un **TLB hit**, la traduzione è cioè contenuta in cache.
- Se la CPU non trova la traduzione nella TLB abbiamo un **TLB miss**. L'hardware accede alla page table per trovare la traduzione e, assumendo che l'indirizzo virtuale generato dal processo sia valido e accessibile, aggiorna il contenuto del TLB con la nuova entry. Queste operazioni sono parecchio costose.
- Una volta che il TLB è aggiornato, l'hardware riprova l'istruzione, ottenendo un TLB hit.

1.9.1 Performance e località

Il TLB migliora le performance grazie al **principio di località**. Esso si divide in:

- **Spaziale.** Se la CPU sta eseguendo un'istruzione presente in memoria, vuol dire che con molta probabilità le prossime istruzioni da eseguire si troveranno fisicamente nelle vicinanze di quella in corso.
- **Temporale.** Se accedo all'istruzione 100 al tempo t_0, con molta probabilità accederò nuovamente ad essa negli istanti di tempo successivi.

1.9.2 TLB miss

Chi gestisce un TLB miss? Ci sono due possibili risposte:

- **Hardware.** L'HW deve sapere la posizione delle page tables in memoria (attraverso il page table register), oltre al loro formato esatto. In presenza di un miss, l'HW deve accedere alla page table, trovare la PTE corretta, estrarre la traduzione desiderata, aggiornare il TLB con la pagina contenente l'indirizzo fisico ricercato e riprovare l'istruzione.
- **Software (S.O).** Al verificarsi di un TLB miss, l'hardware solleva un eccezione per mettere in pausa il flusso corrente di istruzioni, aumenta i privilegi a livello kernel e salta a un trap handler. Questo trap handler è codice scritto all'interno

del sistema operativo, il cui scopo è la gestione esplicita dei TLB misses. Il codice cercherà la traduzione nella page table, userà "speciali" istruzioni privilegiate per aggiornare il TLB e, infine, eseguirà la **return-from-trap**. A questo punto, l'hardware riproverà l'istruzione (TLB hit).

TLB return from trap In questo caso, quando si torna da una TLB miss-handling trap, l'hardware deve ripristinare l'esecuzione dall'istruzione che aveva causato la trap nel sistema operativo.

Quando il TLB miss-handler è in esecuzione, il sistema operativo deve essere molto attento a non causare una catena infinita di TLB misses. Se ho un miss, viene generata un'eccezione. Bisogna fare un context switch per permettere al S.O. di gestire l'evento. Per mandarlo in esecuzione bisogna mettere l'indirizzo del TLB miss-handler nel PC. Questo indirizzo tuttavia, come tutti gli altri, viene passato all'MMU. Quest'ultima lo cerca nel TLB, ottenenedo un miss. Parte quindi un loop. La soluzione che viene adottata per risolvere questo problema consiste nel tenere il miss handler all'interno del TLB.

1.9.3 TLB - contenuto

Una address-translation cache tipica potrebbe avere 32, 64 o 128 entries ed essere ciò che viene chiamato **fully associative**. Ciò significa che una traduzione potrebbe essere ovunque nel TLB e l'hardware dovrà cercare in parallelo fino a trovare la traduzione desiderata. Una entry del TLB ha il seguento aspetto: `VPN | PFN | other bits`.
Tra gli other bits generalmente ci sono il **valid bit**, i **protection bits**, ...

1.9.4 TLB - Context Switch

Il TLB contiene traduzioni virtual to physical che sono valide per il processo in esecuzione ma prive di significato per gli altri. Bisogna assicurarsi che quando cambiamo processo, il processo che sta per essere eseguito non usi le traduzioni di quello precedente. Un approccio semplice ma inefficace è fare un **flush** (impostando tutti i valid bit a 0) del TLB a fronte di un context switch. Ogni volta che un processo verrà eseguito, incapperà in TLB misses.
Per ridurre questo overhead, alcuni sistemi aggiungono un supporto hardware per abilitare la condivisione del TLB attraverso context switches. In particolare alcuni sistemi hardware forniscono un campo **Address Space Identifier** (ASID) nel TLB.

1.10 Multi Level Page Tables

Le page tables sono grandi e consumano troppa memoria.

1.10.1 Bigger pages

Una possibile soluzione è quella di fare pagine più grandi. Il problema è che questo comporta a sprechi di spazio all'interno delle pagine stesse (**frammentazione interna**). La memoria si riempie subito di pagine contenenti parecchio spazio vuoto.

1.10.2 Paginazione e segmentazione

Assumiamo di avere un address space nel quale la porzione usata da stack e heap è piccola. Per esempio, usiamo uno spazio di indirizzamento da `16KB` con pagine da `1KB`. La page table relativa a questo address space sarà quindi:

PFN	valid	prot	present	dirty
10	1	r-x	1	0
-	0	_	-	-
-	0	_	-	-
-	0	_	-	-
23	1	rw-	1	1
-	0	_	-	-
-	0	_	-	-
-	0	_	-	-
-	0	_	-	-
-	0	_	-	-
-	0	_	-	-
-	0	_	-	-
-	0	_	-	-
-	0	_	-	-
28	1	rw-	1	1
4	1	rw-	1	1

Come possiamo osservare dalla figura, la maggior parte della page table non è utilizzata. Invece di avere una singola page table per l'intero address space del processo, perchè non averne una per segmento logico? Con la segmentazione avevamo un registro *base* che ci diceva dove ogni segmento risiedeva in memoria fisica e un registro *bound* che ne esprimeva la grandezza. Nel nostro approccio ibrido, abbiamo queste strutture nell'MMU; qui, non usiamo il *base* per puntare al segmento stesso, ma teniamo l'indirizzo fisico della page table di quel segmento. Il registro *bound* è usato per indicare la fine della page table relativa a un segmento.
Nell'hardware, assumiamo che ci siano tre paia di *base/bound*: una per code, heap e stack. In un context switch, questi registri devono essere cambiati per riflettere la locazione delle page tables del nuovo processo in esecuzione. In un TLB miss, l'hardware usa i bits del segmento per determinare quale coppia *base/bound* usare. L'hardware quindi prende il *base* corretto e lo combina col VPN come segue, per formare l'indirizzo della PTE:

```
SN = (VirtualAddress & SEG_MASK) >> SN_SHIFT
VPN = (VirtualAddress & VPN_MASK) >> VPN_SHIFT
AddressOfPTE = Base[SN] + (VPN * sizeof(PTE))
```

La differenza critica, sta nella presenza di un registro *bound* per segmento. Ogni *bound* contiene il valore della massima pagina valida nel segmento. Se il code segment sta usando le sue prime tre pagine (0, 1 e 2), la page table del segmento avrà solamente tre entries allocate e il bound register sarà impostato a 3. In questa maniera, il nostro approccio ibrido realizza un risparmio di memoria significativo rispetto alla classica page table lineare.

Purtroppo questo approccio si porta dietro con sè tutti i problemi della segmentazione, e se avessimo ad esempio heap molto grandi e sparsi in memoria, finiremmo con l'avere ancora una volta sprechi per via dele page table. Seconda cosa, il manifestarsi ancora una volta della frammentazione esterna.

1.10.3 Multi Level Page Tables

Un altro approccio potrebbe essere trasformare una page table lineare in qualcosa di simile a un albero.

- Per prima cosa, viene tagliata la page table in unità page-sized.
- Se una pagina di una PTE è invalida, non viene allocata.
- Per tenere traccia se una pagina delle page table è valida (e se valida, dove risiede in memoria), usiamo una nuova struttura chiamata **page directory**.

Ciò che fa la **multi-level table** è far scomparire parti della page table lineare e tenere traccia di quali pagine sono allocate.

La page directory, consiste in una serie di **page directory entries (PDE)**, le quali hanno un valid bit e un numero di page frame (**PFN** - in questo caso rappresenta l'indirizzo di memoria dove è situata una page table). Il bit di valità è un po' diverso, se la PDE è valida significa che almeno una delle pagine della page table a cui la entry punta (via PFN) è valida.

Vantaggi:

- La multi-level table alloca spazio **solamente per la page table in proporzione all'ammontare di address space in uso** (*se c'è tanto spazio in mezzo tra due pagine utilizzate, vengono comunque allocate solo due pagine*).
- Se implementata correttamente, **ogni porzione della page table entra ordinatamente in una pagina**, rendendo più facile la gestione della memoria; il sistema operativo prende semplicemente la prossima pagina libera quando ha bisogno di allocare o far crescere una page table.

Abbiamo aggiunto l'**indirezione**, che ci permette di posizionare pagine della page table ovunque vogliamo in memoria fisica. Questa tecnica ha un costo: in un TLB miss saranno necessari due caricamenti da memoria per prelevare la traduzione corretta dalla page table (una per la page directory e una per la PTE stessa, *trade off time-space*). Nel caso medio (TLB hit), le performance sono **identiche** alla page table lineare. Un altro aspetto negativo è la **complessità**, che sia l'hardware o il sistema operativo a gestire la consultazione delle page tables.

1.10.4 Più di due pagine

Bisogna evitare che la page directory diventi troppo grande, altrimenti l'obiettivo di fare in modo che ogni pezzo della multi-level page table entri in una pagina svanisce. Quando si ha a che fare con pagine piuttosto piccole che lasciano parecchi bit di VPN, è preferibile splittare la page directory stessa in più pagine, aggiungendo un'altra page directory sopra ad essa.

1.11 Page Fault e Swap

Per supportare address spaces di grandi dimensioni (per permettere ai processi di non preoccuparsi se c'è abbastanza spazio in memoria), il sistema operativo avrà bisogno di posizionare altrove le pagine che non sono largamente richieste.

1.11.1 Swap space

La prima cosa da fare è riservare un po' di spazio su disco per muovere le pagine avanti e indietro. Nei sistemi operativi, ci riferiamo a questa locazione come **swap space**. La sua dimensione è importante, in quanto determina il numero massimo di pagine di memoria che possono essere usate da un sistema ad un dato istante di tempo.

1.11.2 Present bit

Quando l'hardware guarda nella PTE, potrebbe scoprire che la pagina non è presente in memoria fisica. Il modo in cui l'hardware (o il sistema operativo in caso di software-managed TLB) determina ciò è attraverso un nuovo **present bit** in ogni PTE. Se il present bit è a 0, la pagina è da qualche parte su disco. A fronte di un **page fault**, viene invocato il sistema operativo, il quale manda in esecuzione un **page-fault handler**.

1.11.3 Page Fault

Se una pagina non è presente, il sistema operativo viene messo al comando per gestire il page fault sia nei sistemi hardware-managed TLB che nei software-managed TLB. Il sistema operativo può usare i bits della PTE relativi al PFN come indirizzo su disco. Quando l'I/O del disco è completato, il sistema operativo aggiorna la page table per marchiare la pagina come presente e aggiorna il campo PFN della PTE e riprova

l'istruzione. Questo nuovo tentativo potrebbe generare un TLB miss (è possibile aggiornare anche il TLB a seguito di un page fault per evitare questo scenario). Mentre viene fatto I/O il sistema operativo sarà libero di eseguire altri processi in ready.

1.11.4 Memoria piena

Il sistema operativo potrebbe voler prima swappare una o più pagine su disco per fare spazio a quelle nuove in procinto di caricare. Il processo di scegliere una pagina da sostituire è noto come **page replacement policy**. Spostare la pagina sbagliata può avere dei costi elevati in termini di performance (*si può causare una velocità disk-like, 10.000 o 100.000 volte più lento*).

A fronte di un page fault, il sistema operativo deve trovare il frame fisico per far risiedere la pagina, e se tale frame non c'è, bisogna aspettare che l'algoritmo di replacement venga eseguito e liberi delle pagine dalla memoria rendendole disponibili per l'utilizzo.

1.11.5 Replacements

Piuttosto che aspettare che si riempa la memoria, il sistema operativo tiene un piccolo ammontare di memoria libera. In molti sistemi, vengono utilizzati un **high watermark** (HW) e un **low watermark** (LW) per facilitare la decisione di quando iniziare a sfrattare le pagine. Quando il sistema operativo nota che ci sono meno di LW pagine disponibili, un thread (**swap deamon**) in background rensponsabile della liberazione della memoria viene eseguito. Il thread sfratta le pagine fino a quando non ce ne sono HW disponibili.

1.12 Replacement policies

Decidere quale pagina (o pagine) sfrattare è incapsulato all'interno della **politica di replacement** del sistema operativo.

1.12.1 Cache management

È possibile vedere il nostro obiettivo come la massimizzazione del numero di cache hits. Conoscere il numero di cache hits e misses ci permette di calcolare l'**average memory access time** (AMAT).

$$AMAT = T_M + (P_{MISS} * T_D)$$

Dove T_M rappresenta il costo di accesso a memoria, T_D il costo di accesso a disco, e P_{MISS} la percentuale di miss (da 0.0 a 1.0).

1.12.2 Optimal replacement policy

La politica ottimale di replacement conduce al minor numero di misses in generale. Se dobbiamo sfrattare delle pagine, perchè non selezionare quelle che verranno usate più avanti nel tempo? È un approccio semplice ma difficile da implementare.

1.12.3 FIFO policy

FIFO (*first in first out*) ha un buon punto di forza: è semplice da implementare. Purtroppo non è in grado di determinare l'importanza dei blocchi, se una pagina viene acceduta parecchie volte, FIFO deciderà comunque di sfrattarla.

1.12.4 Random policy

L'algoritmo random, che sceglie una pagina casuale da sostituire, ha proprietà simili al FIFO, è semplice da implementare ma non sceglie intelligentemente i blocchi da sfrattare.

1.12.5 LFU and LRU

Per evitare di sfrattare pagine importanti sfruttiamo il **principio di località**. Se un processo accede una pagina di recente, è molto probabile che quest'ultima verrà acceduta nuovamente nel futuro prossimo. Un tipo di informazione "storica" che potrebbe essere usata in una politica di page replacement è la **frequenza**. La politica **Least Frequently Used** (LFU) sostituisce le pagine usate meno di frequente. Simile è LRU **Least Recently Used** che sostituisce la pagina usata meno di recente.
Esistono anche una classe di algoritmi opposti, Most Frequently Used MFU e Most Recently used MRU.

1.12.6 LRU approssimato

Dato che scansionare tutti i tempi per trovare la pagina least recently used è molto costoso, possiamo usare un'approssimazione. L'idea richiede supporto hardware, nella forma di **use bit**. Questo bit è contenuto in ogni pagina del sistema e ogni volta che una di esse viene riferita (letta o scritta), lo use bit è settato dall'hardware a 1. L'hardware non pulisce mai il bit, è il sistema operativo che ha il compito di settarlo a 0. Ci sono molti modi ma il **clock algorithm** è un approccio molto semplice e funzionale. Immaginiamo tutte le pagine del sistema arrangiate in una lista circolare. Una **clock hand** punta a una pagina (non importa quale). Quando deve essere fatta una sostituzione, il sistema operativo controlla se la pagina puntata P ha lo use bit a 1 o a 0. Se a 1 non è un buon candidato per la sostituzione, il bit viene settato a 0 e la clock hand passa alla prossima pagina. L'algoritmo continua fino a quando non trova una pagina con use bit a 0. Se la pagina è **dirty**, deve essere riscritta su disco prima di essere sfrattata, quindi molti sistemi preferisco pulire pagine **clean**.

1.12.7 Trashing

Cosa dovrebbe fare il sistema operativo quando la memoria è semplicemente sovraccaricata e la richiesta di memoria dell'insieme dei processi in esecuzione eccede la memoria fisica disponibile? In questi casi il sistema è in costante paginazione (**trashing**).

- **Admission control.** Dato un insieme di processi, un sistema può decidere di non eseguirne un sottoinsieme.
- **Out of memory killer.** Quando la memoria è sovraccarica, questo demone sceglie un processo che sta usando intensamente la memoria e lo termina, riducendo piano piano l'utilizzo della risorsa. (*Linux*)

2 Concurrency

2.1 Threads e locks

Un **thread** è un sottoinsieme delle istruzioni di un processo, che può essere eseguito in maniera concorrente con altre parti di esso. L'obiettivo dei threads è quello di rendere più veloce l'esecuzione di un processo. Un programma multi-thread ha più punti di esecuzione (molteplici PCs, da ognuno dei quali vengono prelevate ed eseguite istruzioni). Possono essere visti come processi separati che **condividono lo stesso address space**. Ogni thread ha il proprio insieme privato di registri. Se ci sono due threads in esecuzione su un singolo processore, per switchare da T1 a T2 deve avvenire un context switch. Invece che il PCB avremo bisogno di un **Thread Control Blocks** (TCBs) per memorizzare lo stato di ogni thread di un processo. La differenza principale tra thread switch e context switch è che nel primo caso l'address space rimane lo stesso. Un altra grande differenza tra threads e processi riguarda lo stack. In un processo multi-thread, ogni thread è indipendente e potrebbe chiamare varie routines. Invece di un singolo stack nell'address space ce ne sarà uno per thread (**thread-local storage**).

Utilizzare i thread abilita la sovrapposizione dell'I/O con altre attività all'interno di un singolo programma.

2.1.1 Thread creation

Vogliamo creare un programma che generi due threads. Ogni threads eseguirà la funzione `mythread()` con argomenti diversi (stringa A o B). Una volta che un thread viene creato, potrebbe venire eseguito subito (dipende dallo scheduler) o essere messo in stato di ready.

```c
#include <stdio.h>
#include <assert.h>
#include <pthread.h>

void *mythread(void *arg) {
  printf("%s\n", (char *) arg);
  return NULL;
}


int main(int argc, char *argv[]) {
  pthread_t p1, p2;
  int rc;
  printf("main: begin\n");
  rc = pthread_create(&p1, NULL, mythread, "A"); assert(rc==0);
  rc = pthread_create(&p2, NULL, mythread, "B"); assert(rc==0);
  // join waits for the threads to finish
  rc = pthread_join(p1, NULL); assert(rc==0);
  rc = pthread_join(p2, NULL); assert(rc==0);
  printf("main: end\n");
```

```
21   return 0;
22 }
```

Il thread principale chiama `pthread_join()` che aspetta il completamento di un particolare thread. L'ordine in cui viene eseguito il programma, dipende solo dallo scheduler.

2.1.2 Dati condivisi

La **race condition** consiste nell'avere più threads che concorrono all'uso della stessa risorsa. Il risultato della computazione **non è deterministico** in quanto dipende esclusivamente dalle decisioni dello scheduler e da quanto siamo fortunati con i timer interrupts. Una **sezione critica** è un pezzo di codice che accede alle variabili condivise e non deve essere eseguita simultaneamente da più thread. Abbiamo bisogno della **mutua esclusione**, la quale garantisce che, se un thread è in esecuzione in sezione critica, agli altri verrà proibilito l'accesso.

2.1.3 Atomicità

Un modo per risolvere il problema potrebbe essere quello di avere istruzioni più potenti che, in un singolo passo, facciano esattamente ciò di cui abbiamo bisogno. In questo caso è l'hardware a garantire l'atomicità, tramite delle **synchronization primitives**.

2.1.4 Thread creation

```
1 #include <pthread.h>
2 int pthread_create( pthread_t *thread, const pthread_attr_t *attr, void *
      (*start_routine) (void*), void * arg );
```

- `pthread_t *thread` è un puntatore a una struttura di tipo `pthread_t` che useremo per interagire con i thread.
- `attr` viene usato per specificare ogni tipo di attributo che il thread potrebbe avere. (*Grandezza stack, informazioni riguardanti priorità di scheduling, ...*)
- Il terzo argomento è un **puntatore a funzione** e consiste nel nome della funzione che vogliamo far eseguire al thread creato.
- `arg` è l'argomento che deve essere passato alla funzione che il thread deve eseguire.

I puntatori sono `void` perchè permettono di passare ogni tipo di argomento facendo semplicemente un cast.

2.1.5 Thread completion

Se vogliamo aspettare il completamento di un thread, dobbiamo chiamare la routine `pthread_join()`

```c
#include <pthread.h>
int pthread_join( pthread_t thread, void **value_ptr );
```

- `thread` è usato per specificare quale thread stiamo aspettando.
- `**value_ptr` è un puntatore al valore di ritorno.

2.2 Locks

I lock vengono usati per introdurre la **mutua esclusione**, permettendo quindi di eseguire atomicamente la sezione critica. Per usare un lock, basta aggiungere il codice necessario attorno alla sezione critica come segue:

```c
lock_t mutex; //lock allocato globalmente
...
lock(&mutex);
x = x + 1;    //sezione critica
unlock(&mutex);
```

Un lock è una variabile, e come tale va **dichiarata e inizializzata**. Un lock, ad un certo istante di tempo, può trovarsi in due stati: **disponibile** o **acquisito**. Il funzionamento è questo:

- viene chiamata la routime `lock()` per acquisire il lock.
- Se disponibile, il thread chiamante riceve il lock e può entrare in sezione critica. Se acquisito, il thread chiamante rimarrà bloccato nella routine `lock()` fino a quando il thread in sezione critica non termina e invoca la `unlock()`.
- Una volta acquisito il lock, un thread può operare in sezione critica.

Il nome della libreria `POSIX` per un lock è **mutex**. È possibile proteggere la sezione critica con un unico grande lock (**coarse-grained**) ma è possibile usare svariati lock (**fine-grained**). Mediante il lock, il programmatore guadagna un po' di **controllo sullo scheduler**. I locks però devono avere le seguenti proprietà:

- **Correctnes.** Garantire la mutua esclusione.
- **Fairness.** Evitare che i thread vadano in starvation. Tuttu devono poter accedere alla sezione critica prima o poi.
- **Performance.** L'overhead di time dovuto all'introduzione dei lock non deve minare le performance.

Per progettare un lock funzionante, abbiamo bisogno di aiuto da parte dell'hardware e dal sistema operativo.

2.2.1 Controlling interrupts

Se gli interrupt vengono disabilitati prima di entrare in sezione critica e riabilitati dopo, abbiamo implementato un rudimentale lock.

```
void lock() { DisableInterrupts(); }
void unlock() { EnableInterrupts(); }
```

Vantaggi

- In intel `CLI` e `STI` sono già presenti nell'ISA del processore, non dobbiamo quindi apportare modifiche dal punto di vista hardware.
- Soluzione molto semplice.

Svantaggi

- È necessario avere fiducia nei threads (`CLI` e `STI` sono istruzioni privilegiate). Un loop infinito in sezione critica potrebbe causare una catastrofe.
- Non funziona bene in presenza di più processori, visto che non siamo in grado di disabilitare gli interrupts su tutte le CPU.
- Disabilitare gli interrupts per periodi di tempo troppo estesi può portare alla perdita di alcuni di essi. Ad esempio, la CPU può perdersi il fatto che il disco ha comunicato di aver terminato la read request.
- Inefficienza.

2.2.2 Load e Store

```
typedef struct __lock_t { int flag; } lock_t;

void init(lock_t *mutex) {
  // 0 -> lock is available, 1 -> held
  mutex->flag = 0;
}

void lock(lock_t *mutex) {
  while (mutex->flag == 1) // TEST the flag
    ; // spin-wait (do nothing)
  mutex->flag = 1; // now SET it!
}

void unlock(lock_t *mutex) {
  mutex->flag = 0;
}
```

Viene usata una flag per indicare se un thread è in possesso del lock. Il primo thread che entrerà in sezione critica chiamerà la routine `lock()`, la quale controllerà il valore di flag e la setterà a 1 nel caso in cui essa sia uguale a 0 (a indicare che il thread è ora in possesso del lock) o farà spin-lock in caso contrario. Una volta finito la sezione critica, il thread chiama la `unlock()` e pulisce flag (rilasciando il lock). Se un altro thread chiamasse la `lock()` mentre il primo è in sezione critica, esso farà **spin-wait** nel ciclo while fino a quando non verrà chiamata la `unlock()`. Ci sono però due problemi:

- **Correctness.** È possibile che entrambi i thread settino flag a 1 ed entrino in sezione critica. Se `T1` vede il lock a 0 e prima di settarlo si ha un interrput, sia `T2` che `T1` setteranno il flag a 1 ed entreranno in sezione critica.

- **Performance.** Questo ciclio è chiamato ciclio di busy waiting o **spinlock**. Il thread è in uno stato di attesa che mantiene occupato il processore, vengono quindi sprecati cicli di CPU

2.2.3 Test and Set

Viene implementato il supporto hardware con l'istruzione `TestAndSet`:

```
int TestAndSet( int *old_ptr, int new) {
  int old = *old_ptr; //prelevo vecchio valore di ptr
  olt_ptr = new;      //inserisco "new" in old_ptr
  return old;         //restituisco il vecchio valore
}
```

Restituisce il vecchio valore puntato da `ptr` e lo aggiorna a `new`. La chiave di questo meccanismo è che questa sequenza di operazioni viene eseguita **atomicamente**.

```
typedef struct __lock_t {
  int flag;
} lock_t;

void init(lock_t *lock) {
  // 0 indicates that lock is available, 1 that it is held
  lock->flag = 0;
}

void lock(lock_t *lock) {
  while (TestAndSet(&lock->flag, 1) == 1)
    ; // spin-wait (do nothing)
}

void unlock(lock_t *lock) {
  lock->flag = 0;
}
```

Con questo tipo di lock, siamo sicuri che un singolo thread potrà acquisire il lock ed entrare in sezione critica. Per funzionare correttamente su un singolo processore, ha bisogno di un **preemptive scheduler** (situazione per cui un processo viene temporaneamente interrotto e portato fuori dalla CPU), ad esempio che interromperà un

thread attraverso un timer. Senza, non ha senso siccome un thread in spinning non potrà mai rinunciare al lock.

2.2.4 Algoritmo di Peterson

L'idea è di garantire che due thread non entrino mai in sezione critica allo stesso tempo.

```
int flag[2];
int turn;

void init() {
  flag[0] = flag[1] = 0;  // 1->thread wants to grab lock
  turn = 0;               // whose turn? (thread 0 or 1?)
}

void lock() {
  flag[self] = 1;    // self: thread ID of caller
  turn = 1 - self;   // make it other threads turn
  while ((flag[1-self] == 1) && (turn == 1 - self))
    ;                // spin-wait
}

void unlock() {
  flag[self] = 0;    // simply undo your intent
}
```

Se una cella di `flag` viene settata a 1 indica che il corrispondente thread desidera entrare in sezione critica. `turn` indica il turno del thread per entrare in sezione critica.

2.2.5 Spin locks

Vantaggi

- Semplicità (poche righe di codice).
- **Correctness.** Fornisce la mutua esclusione correttamente.

Svantaggi

- **Fairness** Non siamo in grado di garantire che ogni thread entrerà in sezione critica.
- **Performance** In una CPU monoprocessore lo spin lock è molto costoso. Quando abbiamo N threads a contendersi il lock, nel caso peggiore verranno sprecato $N-1$ fette di tempo. In altri casi funziona ragionevolmente (*se il numero di threads è più o meno uguale al numero dei processori*).

2.2.6 Soluzioni

Dare la precedenza ad altri thread invece che fare spinlock.

```
void init() {
  flag = 0;
}

void lock() {
  while (TestAndSet(&flag, 1) == 1)
    yield();  //rilascia la cpu
}

void unlock() {
  flag = 0;
}
```

La primitiva `yield()` è una semplice **system call** che muove il chiamante dallo stato di running a quello di ready (*deschedula il thread*). Con due thread funziona bene ma con tanti che si contendono la sezione critica no. Se un thread acquisisce la CPU e viene interrotto prima di chiamare la `unlock()`, tutti gli altri chiameranno la `lock()`, troveranno il lock occupato e chiameranno la `yield()`. Oltre al problema della **starvation** abbiamo anche problemi di **performance**.

Sleeping instead of spinning Lo scheduler viene lasciato troppo al caso. `Solaris` mette a disposizione due routines:

- `park()` per mettere un thread chiamante in stato di sleep.
- `unpark(threadID)` per svegliare un particolare thread.

Queste due routines possono essere usate per costruire un lock che mette il chiamante a dormire se il lock è già acquisito e lo sveglia quando è disponibile. Per evitare la starvation si usa una **coda** per controllare chi è il prossimo a prendere il lock. Inoltre, viene usata una variabile `guard` per fare spin-lock attorno a `flag` e manipolare la coda in uso dal lock. Un thread potrebbe comunque essere interrotto durante l'acquisizione o il rilascio del lock, causando gli altri thread a fare spin-wait ancora una volta. Tuttavia, il tempo speso a fare spinning è limitato (solo poche istruzioni all'interno del codice di `lock` e `unlock`.

```
typedef struct __lock_t {
  int flag;
  int guard;
  queue_t *q;
} lock_t;

void lock_init(lock_t *m) {
  m->flag = 0;
  m->guard = 0;
  queue_init(m->q);
```

```
}

void lock(lock_t *m) {
  while (TestAndSet(&m->guard, 1) == 1)
    ; //acquire guard lock by spinning
  if (m->flag == 0) {
    m->flag = 1; // lock is acquired
    m->guard = 0;
  } else {
    queue_add(m->q, gettid());
    m->guard = 0;
    park();
  }
}

void unlock(lock_t *m) {
  while (TestAndSet(&m->guard, 1) == 1)
    ; //acquire guard lock by spinning
  if (queue_empty(m->q))
    m->flag = 0; // let go of lock; no one wants it
  else
    unpark(queue_remove(m->q)); // hold lock (for next thread!)
  m->guard = 0;
}
```

Se `guard` fosse dopo `park()` tutti i thread successivi avrebbero trovato `guard` a 1, andando in spin-lock. conseguentemente, lo scheduler sceglie i threads in modo da garantire che non avvenga nessun deadlock. `flag` non viene settata nuovamente a 0 quando un altro thread viene svegliato perchè passiamo il lock direttamente dal thread che lo rilascia al prossimo che lo acquisisce.

È Possibile che un thread sul punto di fare `park()` venga switchato al thread in possesso del lock e che quest'ultimo lo rilasci. Ciò potrebbe portare allo sleep permanente del primo thread (**waiting race**). Si può risolvere con una terza chiamata a `setpark()` (introdotto da `Solaris`) che serve per indicare che un thread è in procito di fare `park()`. Se quindi dovesse avvenire un thread switch e un altro thread chiamasse `unpark()` prima che `park()` sia effettivamente chiamata, la successiva `park()` ritorna immediatamente invece di dormire.

```
queue_add(m->q, gettid());
setpark();
m->guard = 0;
```

2.3 Condition Variables

Ci sono molti casi in cui un thread desidera controllare quando una condizione è vera prima di continuare la propria esecuzione. Per esempio un thread genitore potrebbe voler attendere il completamento del figlio prima di continuare (`join()`).

```
volatile int done = 0;
```

```
2
3  void *child(void *arg) {
4    printf("child\n");
5    done = 1;
6    return NULL;
7  }
8
9  int main(int argc, char *argv[]) {
10   printf("parent: begin\n");
11   pthread_t c;
12   Pthread_create(&c, NULL, child, NULL); // create child
13   while (done == 0)
14     ; // spin
15   printf("parent: end\n");
16   return 0;
17 }
```

Usare variabili condivise funziona ma è molto inefficiente, siccome il genitore spreca tempo di CPU a fare spin. Una **condition variable** è una coda esplicita in cui i threads possono mettersi quando la condizione di esecuzione non è quella desiderata. Quando lo stato cambia, il thread (uno o più) viene svegliato e può quindi riprendere la propria esecuzione. Una condition variable ha associate due operazioni:

- `wait()` Mette in sleep un thread.
- `signal()` Viene usata quando un thread ha cambiato qualcosa nel programma e vuole quindi svegliarne uno in sleep che aspettava il verificarsi di quella condizione.

La `pthread_cond_wait( pthread_cond_t *c, pthread_mutex_t *m)` prende anche un mutex come parametro. Si assume che questo mutex sia **locked** quando la `wait()` viene invocata. La responsabilità della wait è di liberare il lock e mettere il thread chiamante in stato di sleep (atomicamente). Quando un thread viene svegliato, deve acquisire nuovamente il lock prima di ritornare dalla `wait()` (per prevenire la race condition). Una soluzione al problema del `join()`:

```
1  int done = 0;
2  pthread_mutex_t m = PTHREAD_MUTEX_INITIALIZER;
3  pthread_cond_t c = PTHREAD_COND_INITIALIZER;
4
5  void thr_exit() {
6    Pthread_mutex_lock(&m);
7    done = 1;
8    Pthread_cond_signal(&c);
9    Pthread_mutex_unlock(&m);
10 }
11
12 void *child(void *arg) {
13   printf("child\n");
14   thr_exit();
15   return NULL;
16 }
```

```
17
18 void thr_join() {
19   Pthread_mutex_lock(&m);
20   while (done == 0)
21     Pthread_cond_wait(&c, &m);
22   Pthread_mutex_unlock(&m);
23 }
24
25 int main(int argc, char *argv[]) {
26   printf("parent: begin\n");
27   pthread_t p;
28   Pthread_create(&p, NULL, child, NULL);
29   thr_join();
30   printf("parent: end\n");
31   return 0;
32 }
```

2.3.1 Approcci sbagliati

Senza la variabile done se il figlio viene eseguito immediatamente e chiama la `thr_exit()`, essa chiamerà la `signal()`, ma non ci sono thread in sleep sulla condizione. Quando il genitore verrà eseguito chiamerà la `wait()` e rimarrà bloccato, nessun thread lo sveglierà mai.

Senza il lock se il genitore chiama `thr_join()` e controlla il valore di done, vedrà che è a zero e si metterà in sleep. Prima di chiamare la wait viene interrotto e viene eseguito il figlio. Quest'ultimo cambia `done` a 1 e chiama `signal()` ma non c'è nessun thread in attesa del verificarsi della condizione. Quando il genitore viene nuovamente eseguito, andrà a dormire per sempre.

Vanno usati sempre i **cicli while** per i controlli.

2.3.2 Produttore e consumatore

I thread produttori generano i dati e li inseriscono in un buffer, i consumatori prendono questi dati dal buffer e li consumano in un certo modo. Il buffer è una risorsa condivisa ed è necessario sincronizzare l'acesso ad essa per evitare la race condition.

```
1  int buffer;
2  int count = 0; // initially, empty
3
4  void put(int value) {
5    assert(count == 0);
6    count = 1;
7    buffer = value;
8  }
9
10 int get() {
11   assert(count == 1);
```

```
12    count = 0;
13    return buffer;
14  }
```

- `put()`, assumendo che il buffer sia vuoto, inserisce semplicemente un valore in esso e lo marca come "pieno" settando la variabile `count`.
- `get()` fa l'opposto, settando il buffer a vuoto (`count = 0`) e restituisce il valore prelevato.

```
1  cond_t cond;
2  mutex_t mutex;
3
4  void *producer(void *arg) {
5    int i;
6    for (i = 0; i < loops; i++) {
7      Pthread_mutex_lock(&mutex);
8      if (count == 1)      //questo verra' cambiato con un while
9        Pthread_cond_wait(&cond, &mutex);
10     put(i);
11     Pthread_cond_signal(&cond);
12     Pthread_mutex_unlock(&mutex);
13   }
14 }
15
16 void *consumer(void *arg) {
17   int i;
18   for (i = 0; i < loops; i++) {
19     Pthread_mutex_lock(&mutex);
20     if (count == 0)      //questo verra' cambiato con un while
21       Pthread_cond_wait(&cond, &mutex);
22     int tmp = get();
23     Pthread_cond_signal(&cond);
24     Pthread_mutex_unlock(&mutex);
25     printf("%d\n", tmp);
26   }
27 }
```

Con un singolo produttore e un singolo consumatore, il codice sopra funziona, ma se abbiamo ad esempio due consumatori, la soluzione due problemi critici:

- Supponiamo che ci siano due consumatori e un produttore:
  - C_1 acquisisce il lock e controlla che ogni buffer sia pronto per la consumazione ma non ce ne sono, chiama la `wait`.
  - P_1 viene eseguito, acquisisce il lock e controlla che tutti buffer siano pieni, ma non lo sono. Va avanti riempendo il buffer. Invoca la `signal()` per informare che il buffer è stato riempito.

  - C_1 si muove nella coda di ready dallo stato di sleeping sulla condition variable (*non viene ancora eseguito*).
  - P_1 continua fino a quando non realizza che il buffer è pieno, e poi va in sleep.
  - C_2 viene eseguito e consuma il valore nel buffer (salta la `wait` perchè il buffer è pieno).
  - C_1 viene ora eseguito, prima di ritornare dalla `wait()`, acquisisce nuovamente il lock, chiama la `get()` ma **non ci sono buffer da consumare**.

    Il problema è questo: dopo che il produttore sveglia C_1, ma prima che esso venga eseguito, lo stato del buffer è cambiato (per colpa di C_2). Segnalare un thread lo sveglia solamente dicendogli che lo stato è cambiato (in questo caso che un valore è stato messo nel buffer), ma non ci sono garanzie che quando il thread svegliato venga eseguito lo stato sia lo stesso (**Mesa semantics**).

    Soluzione: cambiare l'**if** in **while** in modo che se C_1 viene svegliato ricontrolla immediatamente lo stato della variabile condivisa. Se il buffer è vuoto, tornerà semplicemente a dormire.

- C'è una sola condition variable.
  - Vengono eseguiti prima C_1 e C_2 e vanno in sleep.
  - Il produttore mette un valore nel buffer e sveglia uno dei consumatori, diciamo C_1.
  - Il produttore torna indietro e prova a inserire più dati nel buffer, siccome è pieno il produttore chiamerà `wait()` e andrà a dormire.
  - C_1 è pronto per essere eseguito e **due threads stanno dormendo sulla condizione** (C_2 e P_1).
  - C_1 Si sveglia ritornando dalla `wait()`, controlla nuovamente la condizione e trova che il buffer è pieno. Consuma il valore e segnala la condizione, svegliando un thread in sleeping. **Quale dei due sveglia?** Se svegliasse il consumatore, troverà il buffer vuoto e si metterà in sleep. Il produttore viene lasciato a dormire. Tutti e tre i threads sono in stato di sleeping. Un **consumatore non dovrebbe poter svegliare altri consumatori**, ma solo produttori (e viceversa).

  La soluzione è usare due condition variables per segnalare correttamente quale tipo di thread andrebbe svegliato.

```
1 int buffer[MAX];
2 int fill = 0;
3 int use = 0;
4 int count = 0;
5
6 void put(int value) {
```

```
7   buffer[fill] = value;
8   fill = (fill + 1) % MAX;
9   count++;
10 }
11
12 int get() {
13   int tmp = buffer[use];
14   use = (use + 1) % MAX;
15   count--;
16   return tmp;
17 }
18
19 cond_t empty, fill;
20 mutex_t mutex;
21
22 void *producer(void *arg) {
23   int i;
24   for (i = 0; i < loops; i++) {
25     Pthread_mutex_lock(&mutex);
26     while (count == MAX)
27       Pthread_cond_wait(&empty, &mutex);
28     put(i);
29     Pthread_cond_signal(&fill);
30     Pthread_mutex_unlock(&mutex);
31   }
32 }
33
34 void *consumer(void *arg) {
35   int i;
36   for (i = 0; i < loops; i++) {
37     Pthread_mutex_lock(&mutex);
38     while (count == 0)
39       Pthread_cond_wait(&fill , &mutex);
40     int tmp = get();
41     Pthread_cond_signal(&empty);
42     Pthread_mutex_unlock(&mutex);
43     printf("%d\n", tmp);
44   }
45 }
```

I thread produttori aspettano sulla condizione `empty` e segnalano `fill`, per i consumatori l'inverso. Inoltre per abilitare più concorrenza ed efficienza si aggiungono più slot al buffer, in modo tale che più valori possano essere prodotti o consumati prima di andare in sleep. Un produttore dorme solo se tutti i buffer sono pieni, un consumatore dorme solo se tutti i buffer sono vuoti.

2.4 Semafori

Un semaforo è un oggetto con un valore di tipo integer utilizzabili sia come locks che come condition variables.

Il valore iniziale di un semaforo ne determina il comportamento.

```
#include <semaphore.h>
sem_t s;
sem_init(&s, 0, 1);
```

Il secondo valore di `sem_init()` viene settato a 0 e indica che il semaforo è condiviso tra thread dello stesso processo. Il terzo argomento è il valore con cui lo si inizializza. Per interagire con esso vengono introdotte due routines:

- `sem_wait(sem_t *s)` decrementa il valore del semaforo `s` di uno e aspetta se il valore del semaforo è negativo.
- `sem_post(sem_t *s)` incrementa il valore del semaforo `s` di uno, se ci sono uno o più threads in attesa ne sveglia uno.

Il valore del semaforo, quando negativo, è uguale al numero di thread in attesa.

2.4.1 Semafori binari: locks

Per implementare i lock, il codice sarà:

```
sem_t m;
sem_init(&m, 0, 1); //inizializza il semaforo a 1
sem_wait(&m);
//sezione critica
sem_post(&m);
```

Il valore di partenza del semaforo `m` a 1 è critico per la realizzazione del lock. I semafori **binari** possono trovarsi solamente in due stati: acquisito o disponibile.

2.4.2 Semafori per ordinare

I semafori possono essere usati anche come condition variables, esempio:

```
sem_t s;

void *child(void *arg) {
  printf("child\n");
  sem_post(&s); // signal here: child is done
  return NULL;
}

int main(int argc, char *argv[]) {
  sem_init(&s, 0, 0); //Inizializzo a 0 per forza
  printf("parent: begin\n");
  pthread_t c;
  Pthread_create(c, NULL, child, NULL);
  sem_wait(&s); // wait here for child
  printf("parent: end\n");
  return 0;
}
```

In questo modo ci aspettiamo di avere come risultato: `parent: begin, child, parent: end`

2.4.3 Semafori come produttore e consumatore

Si usano due semafori, `empty` e `full`, che indicano quando una entry del buffer è stata svuotata o riempita rispettivamente.

```
int buffer[MAX];
int fill = 0;
int use = 0;

void put(int value) {
  buffer[fill] = value;
  fill = (fill + 1) % MAX;
}

int get() {
  int tmp = buffer[use];
  use = (use + 1) % MAX;
  return tmp;
}
```

Produttori e consumatori:

```
sem_t empty;
sem_t full;

void *producer(void *arg) {
  int i;
  for (i = 0; i < loops; i++) {
    sem_wait(&empty);
    put(i);
    sem_post(&full);
  }
}

void *consumer(void *arg) {
  int i, tmp = 0;
  while (tmp != -1) {
    sem_wait(&full);
    tmp = get();
    sem_post(&empty);
    printf("%d\n", tmp);
  }
}

int main(int argc, char *argv[]) {
// ...
sem_init(&empty, 0, MAX); //MAX buffers are empty to begin with
sem_init(&full, 0, 0); // ... and 0 are full
// ...
```

```
28 }
```

Il produttore aspetta che il buffer diventi vuoto per poterlo riempire di dati e il consumatore attende che il buffer sia pieno prima di consumare i dati. Con un singolo consumatore e un produttore funziona bene, con `MAX = 10` abbiamo un problema di **race condition**:

- Abbiamo due produttori in procinto di chiamare la `put()`
- P_1 viene eseguito prima e inizia a riempire il buffer. Prima che esso incrementi `fill` a 1, viene interrotto.
- Il produttore P_2 inizia ad essere eseguito e anche lui, inserisce i suoi dati nello stesso punto del buffer sovrascrivendo quelli vecchi.

Manca la **mutua esclusione**, riempire il buffer e incrementare il contatore è una porzione di codice critica. Basta aggiungere i semafori binari come locks. Inizialmente può sembrare una buona idea metterli intorno a `wait` e `post`, ma questo potrebbe portare a un **deadlock**.

- Il consumatore viene eseguito e acquisisce mutex.
- Chiama la `sem_wait(&full)`. Possiede ancora il lock.
- Viene eseguito un produttore che, se fosse possibile, riempirebbe il buffer di dati e sveglierebbe il consumatore. Chiama `sem_wait(&mutex)`. Il lock è già in possesso del consumatore e il produttore è bloccato ad aspettare. Il consumatore ha il lock ed è in waiting, produttore non ha il lock, ed è in waiting.

Si sposta quindi il mutex intorno solo alla `get()` e alla `put()` e si ottiene il seguente codice:

```
1  sem_t empty;
2  sem_t full;
3  sem_t mutex;
4
5  void *producer(void *arg) {
6    int i;
7    for (i = 0; i < loops; i++) {
8      sem_wait(&empty);
9      sem_wait(&mutex); //(MOVED MUTEX HERE...)
10     put(i);
11     sem_post(&mutex); //(... AND HERE)
12     sem_post(&full);
13   }
14 }
15
16 void *consumer(void *arg) {
17   int i;
18   for (i = 0; i < loops; i++) {
```

```
19      sem_wait(&full);
20      sem_wait(&mutex); //(MOVED MUTEX HERE...)
21      int tmp = get();
22      sem_post(&mutex); //(... AND HERE)
23      sem_post(&empty);
24      printf("%d\n", tmp);
25    }
26  }
27
28  int main(int argc, char *argv[]){
29    // ...
30    sem_init(&empty, 0, MAX);//MAX buffers are empty to begin with
31    sem_init(&full, 0, 0);   // ... and 0 are full
32    sem_init(&mutex, 0, 1);  // mutex=1 because it is a lock
33    // ...
34  }
```

2.4.4 Implementazione dei semafori

```
1   typedef struct __Zem_t {
2     int value;
3     pthread_cond_t cond;
4     pthread_mutex_t lock;
5   } Zem_t;
6
7   // only one thread can call this
8   void Zem_init(Zem_t *s, int value) {
9     s->value = value;
10    Cond_init(&s->cond);
11    Mutex_init(&s->lock);
12  }
13
14  void Zem_wait(Zem_t *s) {
15    Mutex_lock(&s->lock);
16    while (s->value <= 0)
17      Cond_wait(&s->cond, &s->lock);
18    s->value--;
19    Mutex_unlock(&s->lock);
20  }
21
22  void Zem_post(Zem_t *s) {
23    Mutex_lock(&s->lock);
24    s->value++;
25    Cond_signal(&s->cond);
26    Mutex_unlock(&s->lock);
27  }
```

2.5 Problemi relativi alla concorrenza

2.5.1 Non-deadlock bugs

Atomicity-violation bug consiste nel non mettere dei **locks** attorno ad una variabile condivisa.

Order-violation bug consiste nel non assegnare un ordine a due thread che accedono alla stessa variabile. Ad esempio, un thread potrebbe dare per scontato che una variabile sia già inizializzata ed utilizzarla, ma se è ancora a `NULL` questo causerà un crash. Per risolvere questo bug, basta usare le **condition variables**.

2.5.2 Deadlock bugs

I deadlock avvengono quando un thread T_1 è in possesso del lock L_1 e in attesa di un altro L_2; il thread T_2 che possiede il lock L_2 è in attesa che L_1 venga rilasciato.

```
1 Thread 1:                    Thread 2:
2 pthread_mutex_lock(L1);      pthread_mutex_lock(L2);
3 pthread_mutex_lock(L2);      pthread_mutex_lock(L1);
```

Le ragioni per cui i deadlock avvengono possono essere:

- Troppo codice
- Dipendenze complesse tra i componenti.
- L'incapsulamento del codice.

Affinchè avvenga un deadlock devono essere valide quattro condizioni:

- **Mutua esclusione** I threads richiedono il controllo escluso delle risorse che acquisiscono.
- **Hold-and-wait** I thread tengono le risorse allocate da essi (ad esempio i locks che hanno già acquisito) finchè aspettano risorse addizionali (ad esempio i lock che vogliono acquisire).
- **No preemption** Le risorse (ad esempio i locks) non possono essere rimosse forzatamente dai threads che le stanno tenendo.
- **Circular wait** Esistono catene circolari di threads.

Se una qualunque di queste condizioni non è soddisfatta, un deadlock non può avvenire.

2.5.3 Prevenzione dei deadlocks

- **Circular wait** La miglior tecnica di prevenzione è fornire un acquisizione del lock ordinata. Per esempio, possiamo prevenire il deadlock acquisendo sempre L_1 prima di L_2. Questo ordine consente di evitare wait cicliche e quindi deadlock.

- **Hold-and-wait** Per evitare il deadlock dovuto a questa condizione basta acquisire tutti i locks in una volta atomicamente, in modo da evitare che non ci siano thread switch prematuri nel mezzo dell'acquisizione del lock. La soluzione però è problematica: richiede di sapere esattamente quali locks devono essere posseduti e di acquisirli tutti in una volta. Tutti i locks devono essere acquisiti in una volta anche se non è necessario possederli tutti.

- **No preemption** Per non dare tutti i locks come acquisiti fino a quando la `unlock()` non viene invocata, si può usare una routine come `pthread_mutex_trylock()` che prende il lock (se disponibile) e ritorna **success** o un codice di errore se il lock è già posseduto. Questo per evitare l'acquisizione multipla di locks, perchè potremmo aspettarne uno mentre siamo in possesso di un altro. Sorge un nuovo problema, la **livelock**.

```
top:
pthread_mutex_lock(L1);
if(pthread_mutex_trylock(L2) != 0) {
  pthread_mutex_unlock(L1);
  goto top;
}
```

  È possibile che due threads tentino ripetutamente questa sequenza e falliscano nell'acquisire il lock. Si può risolvere con un delay random prima di fare il loop back.

- **Mutua esclusione** Si possono utilizzare itruzioni hardware per evitare di usare i locks, in questo modo non sarebbe possibile che si verifichino deadlocks (livelock può comunque accadere).

2.5.4 Algoritmo del banchiere

Si potrebbe non eseguire concorrentemente threads che potrebbero causare deadlock. Un approccio che utilizza questa tattica è l'algoritmo banchiere, che permette di gestire istanze multiple di una risorsa.

Sua n il numero di processi e m il numero di tipi di risorse, abbiamo che:

- `disponibili[j] = k` se ci sono k istanze disponibile del tipo di risorsa R_j. Vettore lungo m.

- `massimo[i,j] = k` se il processo P_i può richiedere al più k istanze del tipo di risorsa R_j. Matrice $n \times m$.

- `assegnate[i,j] = k` se a P_i sono attualmente allocate k istanze di R_j. Matrice $n \times m$.
- `necessità[i,j] = k` se P_i può richiedere k ulteriori istanze di R_j per completare il proprio task.

`necessità[i,j] = massimo[i,j] - assegnate[i,j]`

```
lavoro[m]
fine[n]
lavoro = disponibili

for i = 0 to n:
  fine[i] = false

back:
if si trova i tale che fine[i] = false & richiesta[i] <= lavoro:
  lavoro = lavoro + assegnate[i]
  fine[i] = true
  goto back

for i = 0 to n:
  if fine[i] = false:
    il processo i e' in deadlock

il sistema e' in stato sicuro.
```

Se i deadlock sono rari, viene fatto un reboot e basta. Periodicamente un deadlock detector viene eseguito e costruisce un grafico delle risorse e lo controlla per individuare i cicli. Se accade un deadlock, il sistema ha bisogno di essere fatto ripartire.

3 Persistence

3.1 I/O devices

Più sono corti i bus, più veloci sono. La memoria è collegata alla CPU attraverso il memory bus, altri dispositivi I/O più generici attraverso I/O bus generici (*PCI*). Le periferiche, attraverso peripheral I/O bus (*SATA, USB, ...*). La bontà di bus si misura in **banda** (quantità di dati che riesce a far circolare).

3.1.1 Dispositivo generico

Un dispositivo ha due componenti e tre registri:

- **Interfaccia** Componente software o hardware con cui comunica con l'esterno.
- **Struttura interna** Parte del dispositivo responsabile dell'implementazione dell'astrazione che il dispositivo presenta al sistema (**drive**). Possono essere semplici chip o semplici CPU.
- **Status register** Può essere letto per vedere lo stato corrente del dispositivo.
- **Command register** Usato per ordinare al dispositivo di eseguire un certo compito.
- **Data register** Usato per trasferire dati da e verso il dispositivo.

Il **protocollo** che il sistema operativo usa per interagire con un dispositivo è:

- Polling sullo status register del dispositivo per aspettare che sia pronto a ricevere un comando.
- Manda qualche dato nel data register. Potrebbero essere necessarie diverse writes. Se la CPU principale è coinvolta con il trasferimento dati è **programmed I/O**.
- Scrive un comando nel command register. Fare ciò permette implicitamente al dispositivo di sapere che i dati sono presenti e che deve iniziare ad eseguire il comando richiesto.
- Il sistema operativo aspetta che il dispositivo termini, facendo polling sullo status register.

Questo protocollo è semplice, funziona ma è inefficiente.

3.1.2 Riduzione dell'overhead della CPU con interrupts

Per evitare l'overhead dovuto al polling, il sistema operativo avanza una richiesta, mette il processo chiamante in sleep e fa un context switch a un altro processo. Quando il dispositivo ha terminato l'operazione richiesta, solleverà un interrupt hardware (**interrupt I/O**) che costringerà la CPU a saltare nel S.O. a un **interrupt handler**. L'handler è un pezzo del codice del sistema operativo che terminerà la richiesta e sveglierà il processo in attesa dell'I/O.

Non è sempre la soluzione migliore perchè in presenza di un dispositivo rapido, si avrà un rallentamento del sistema dovuto ai molti context switch. In quel caso la soluzione migliore potrebbe essere il polling. Se non è conosciuta la velocità del dispositivo, la scelta migliore potrebbe essere un approccio **ibrido** (*polling per un po', se il dispositivo non ha finito vengono usati gli interrupts*).

3.1.3 DMA - Direct Memory Access

Un **DMA** è un dispositivo specifico che orchestra il trasferimento tra dispositivi e memoria principale senza che la CPU debba intervenire. Viene programmato dal S.O. il quale gli dirà dove risiedono i dati in memoria, quanti dati copiare e a quale dispositivo mandarli. Il sistema operativo ha terminato con il trasferimento e può procedere intanto con altri compiti. Quando il DMA ha finito, il **DMA controller** genera un interrput per avvertire il S.O. del completamento del trasferimento.

Vantaggi I tempi di CPU sono notevolmente ottimizzati.

Svantaggi

- Richiede hardware aggiuntivo.
- Il DMA controller ruba cicli di CPU.
- Per evitare collisioni (*memoria che serve sia al S.O. che al DMA*) si usa una **cache**.
- Bisogna sincronizzare RAM e DMA per evitare di caricare o scrivere dati non aggiornati. Se il DMA viene interrotto dalla CPU a causa di un page fault e la pagina dove stavano per essere caricati i dati viene messa su disco, dobbiamo evitare che il DMA scriva sulla pagina di un altro processo. Si usa un bit speciale che, se settato a 1, impedisce all'algoritmo di page replacement di toccare la pagina (**pinning**).

3.1.4 Interazione con i dispositivi

Come fa l'hardware a comunicare con i dispositivo? Esistono due metodo primari per farlo:

- **Istruzioni eplicite I/O** (*es x86: in, out*) per mandare dati a specifici registri del dispositivo. Il chiamante specifica un registro con i dati e il nome della porta del dispositivo. Sono istruzioni **privilegate**.
- **Memory mapped I/O** l'hardware rende i registri dei dispositivi disponibili come se fossero locazioni di memoria vere e proprie. Si usano le solite load e store.

3.1.5 Device driver

Per mantenere la maggior parte del sistema operativo *device-neutral*, nascondendo i dettagli delle interazioni coi dispositivi, al livello più basso, un pezzo di software nel sistema operativo deve conoscere i dettagli di come funziona un dispositivo. Questo pezzo di software è chiamato **device driver**. L'incapsulamento ha anche aspetti negativi, in presenza di un dispositivo con molte capacità speciali, quest'ultimo dovrà comunque presentare un'interfaccia generica al resto del kernel. Siccome i device drivers sono necessari per ogni dispositivo inserito nel sistema, il kernel è composto circa al 70% di drivers.

Il sistema operativo deve adattarsi ai dispositivi, introducendo i drivers appropriati. Se fosse il contrario, i ssitemi operativi sarebbero molto meno complessi e più compatibili.

3.2 Hard disks

Gli hard disk sono la principale forma di archiviazione persistente dei dati nei sistemi di computer e nello sviluppo della tecnologia dei file systems.

3.2.1 Geometria base

Un disco ha diversi piatti, ogni piatto ha due superfici, ognuna delle quali è divisa in anelli concentrici detti tracce, le quali sono a loro volta suddivise in settori. La velocità di rotazione è misurata in rotazioni per minuto (**RPM**). Si leggono e scrivono dati attraverso una testina. Ogni settore è una fetta da `512B`. È possibile fare operazioni multi-settore, una singola write da `512B` viene fatta atomicamente. Se si dovesse verificare una perdita di corrente, solo una write viene scritta male.

3.2.2 Hard disk semplice

Il disco deve aspettare che il settore venga fatto ruotare sotto la testtina (**rotational delay**). Il processo impegato per muovere il braccio del disco alla traccia corretta è chiamato **seek**. La seek consiste in: accelerazione, spostamento a piena velocità, decelerazione, e **settling**. La fase finale dell'I/O (scrittura o lettura) è conosciuta come **transfer**.

$$T_{I/O} = T_{seek} + T_{rotation} + T_{transfer}$$

$$Rate_{I/O} = Size_{transfer}/T_{I/O}$$

3.2.3 Disk scheduling

Nello scheduler dei processi non eravamo a conoscenza della lunghezza di ogni job, con il disk scheduling possiamo fare buone assunzioni di quanto impiegherà un job. Stimando il seek time e la rotational delay di una richiesta, il **disk scheduler** può sapere quanto quella richiesta impiegherà per essere soddisfatta. Cercherà di eseguire il principio di **SJF** *Shortest Job First*.

3.2.4 SSTF - Shortest Seek Time First

SSTF ordina la coda delle richieste di I/O per track, scegliendo le richieste sulla traccia più vicina da eseguire prima. Il problema è che al sistema operativo non è disponibile la geometria del drive, il quale vede il dispositivo come un array di blocchi. Può però implementare un algoritmo **NBF** *nearest block first*. Inoltre, se ci fosse un flusso costante di richieste alla traccia dove la testina è correntemente posizionata, causerebbe **starvation**.

3.2.5 Elevator - SCAN o C-SCAN

SCAN si muove avanti e indietro attraverso il disco, servendo le richieste in ordine di tracks. Un passo da una traccia all'altra viene detto **sweep**. Se si incontra una richiesta per una track che è già stata servita in questo sweep del disco, non viene gestita subito ma viene accodata fino al prossimo sweep. Non è ottimale in quanto non si avvicina al principio di **SJF** a cui si voleva puntare.

3.2.6 SPTF - Shortest Positioning Time First

Se il tempo di seek è molto più alto di quello di rotational delay, allora SSTF funziona correttamente. Tuttavia, se il tempo di seek fosse più veloce di quello di rotation, allora avrebbe più senso usare un algoritmo che valuti quale settore è meglio servire prima. **SPTF** (o SATF *shortest access time first*) è utile per migliorare le performance. È molto difficile implementarlo **in un sistema operativo**, in quanto non ha generalmente un'idea di dove siano collocate le tracks o di dove si trovi la testina del disco. SPTF viene quindi eseguito all'**interno di un drive**.

3.2.7 Problemi relativi ai dischi

- **Dove fare il disk scheduling?** Nei vecchi sistemi veniva fatto dal sistema operativo, in quelli moderni i dischi possono avere scheduler interni molto sofisticati in grado di implementare SPTF. All'interno del **disk controller** sono contenuti una serie di dettagli rilevanti, inclusa l'esatta posizione della testina.

- **I/O merging** Se due richieste di lettura sono in due settori adiacenti, l'operazione di I/O merging permette di fonderle in un'unica richiesta. È importante a livello di sistema operativo in quanto riduce il numero di richieste inviate al disco.

- **Quanto tempo si dovrebbe aspettare prima di eseguire un I/O su disco?** Si può servire il disco non appena si presenta una richiesta in modo che il dispositivo non sia mai in stato di IDLE (**work-conserving**) o si può attendere per una richiesta migliore e più efficiente (**non-work-conserving**) e in questo modo le performance aumentano notevolmente.

3.3 RAID

Il sistema **RAID** (*redundant array of inexpensive disks o redundant array of indipendent disks*) è una tecnica che, utilizzando dischi multipli in accordo, costruisce un disk system più veloce, grande e affidabile. Esternamente un RAID appare come un normale disco, un gruppo di blocchi che possono essere letti e scritti. Internamente invece è molto complesso e consiste in dischi multipli, memoria (volatile e non) e uno o più processori adoperati per la gestione del sistema. I RAIDs offrono molti **vantaggi**:

- **Performance** Usando dischi multipli in parallelo è possibile velocizzare notevolmente i tempi di I/O.
- **Capacità** Essendoci molti dischi siamo in grado di memorizzare dati di grandi dimensioni.
- **Affidabilità** I RAIDs tollerano la perdita di un disco e continuano a operare normalmente.

I sistemi che adoperano RAID non hanno bisogno di trattare problemi legati alla compatibilità.

3.3.1 Interfaccia e struttura interna

Quando un file system esegue una richiesta **I/O logica** al RAID, quest'ultimo calcola internamente quale disco (o dischi) devono essere acceduti ed eseguirà uno o più **I/O fisici** per svolgere il compito richiesto. Un RAID è generalmente costituito da un insieme di dischi **SCSI (o SATA)** e da un **SCSI (o SATA) controller** che esegue il firmware per dirigere le operazioni, una memoria volatile **DRAM** usata come buffer per blocchi di dati quando essi vengono letti o scritti e, in alcuni casi, una **memoria non volatile** usata come buffer per scrivere in modo sicuro.

Le idee base di un sistema RAID sono:

- Distribuire l'informazione su più dischi in modo da parallelizzare una parte delle operazioni di accesso ai dati e guadagnare in prestazioni.
- Duplicare su più dischi l'informazione memorizzata per maggiore affidabilità.

Quando il S.O. vuole accedere a un blocco logico, presenta il numero del blocco al controller del disco. Quest'ultimo utilizza il numero per risalire al settore che contiene il blocco desiderato, ne legge il contenuto e lo trasferisce in RAM (tramite DMA). Esistono diversi livelli di sistemi RAID:

3.3.2 RAID level 0: Striping

Non sono veri e propri RAID, in quanto non viene impiegata alcuna duplicazione dei dati. Il disco virtuale (ciò che viene visto dal S.O.) viene mappato dalla logica del RAID sui vari settori dei dischi, suddividendo i blocchi logici del disco virtuale in **strips** (una striscia può consistere in più blocchi). Per mappare l'array virtuale nelle locazioni fisiche del disco, dato un indirizzo logico A di un blocco si ha:

$$Disk = indirizzo\%n\_dischi$$

$$Offset = A/n\_dischi$$

La parte con la virgola viene troncata. Un RAID di livello 0 è tanto più efficace quanto più le richieste coinvolgono l'accesso a molti blocchi consecutivi e quanto più è alto il numero di dischi su cui sono suddivisi i blocchi. Se le operazioni su disco richiedono l'accesso a dati contenuti nello stesso disco, si hanno le stesse prestazioni di un disco normale. L'affidabilità è inferiore a quella di un semplice disco, perchè è formato da più dischi e il **Mean time to failure** (MTTF) si abbassa. Viene usato quando si ha bisogno di alte prestazioni senza particolari problemi di affidabilità (*streaming video o audio*).

Performance

- **Capacità** dati N dischi, ognuno composto da B blocchi, fornisce $N * B$ blocchi di capacità utile.
- **Affidabilità** bassa, ogni disk failure porterà a una perdita di dati.
- **Performance** eccellente. Tutti i dischi sono utilizzati, spesso in parallelo, per servire le richieste di I/O.
- **Single-block lantecy** uguale a quella di un singolo disco.
- **Steady-state throghput** dati N dischi e R un workload random. Per un elevato numero di I/Os random, possiamo ancora una volta usare tutti i dischi, ottenenendo $N * R$ `MB/s`

3.3.3 RAID level 1: Mirroring

Usa contemporaneamente striping e mirroring: tutti i dati sono suddivisi in strip (come nel livello 0) e duplucati su due dischi. È la soluzione RAID più costosa a parità di capacità di memorizzazione, ma anche la più affidabile e la più efficiente in lettura. I dischi di mirroring possono essere usati per fare letture in parallelo.

Performance

- **Capacità** N dischi e B blocchi con il mirroring level = 2 ha capacità $(N * B)/2$.
- **Affidabilità** eccellente.
- **Performance** elevate. Le richieste possono essere servite in parallelo come per il RAID 0.
- **Single-block latency** In lettura ha la stessa latenza di un disco singolo, in scrittura la write logica deve attendere il completamento di due writes fisiche. L'operazione di scrittura sarà più lenta rispetto a quella in un singolo disco.
- **Steady-state throughput** Quando scriviamo o leggiamo **sequenzialmente**, ad esempio nel blocco logico 0, il RAID internamente scriverà sui blocchi 0 e 1. La banda massima ottenuta durante una write o una read sequenziale è di $(N/2) * S$. Il workload random è il caso migliore in quanto con le read è in grado di fornire una banda pari a $N * R$ `MB/s`. Per quanto riguarda le writes, $(N/2) * R$ `MB/s` ognuna.

3.3.4 RAID level 4: saving space with parity

Usa la tecnica di striping a livello di blocchi e introduce anche un disco di parità per eventuali operazioni di **recovery**. Il blocco n del disco di recovery, conterrà la parità della strip n. Per computare la parità si usa la `XOR bitwise` (*bit a bit*). L'informazione di parità può essere usata per guarire da un fallimento. Per sapere quale contenuto aveva un blocco, semplicemente si leggono tutti i valori nelle righe (incluso il bit di parità).

Il RAID 4 garantisce lo stesso livello di mantenimento dei dati del RAID 1 in caso di guasto di un disco, ma al costo una maggiore inefficienza, in quanto ogni qualvolta che una strip di un disco viene modificata, occorre leggere e ricalcolarne la parità. Inoltre, il disco di recovery è coinvolto in ogni operazione di scrittura sul RAID e può facilmente diventare un **collo di bottiglia**.

Performance

- **Capacità** $(N - 1) * B$.
- **Affidabilità** tollera il fallimento su un solo disco, non di più.
- **Performance** basse.
- **Single-block latency** La latenza di una signola write richiede 2 read e 2 write (2 per il dato e 2 per la parità). Le reads possono essere fatte in parallelo, così come le writes. La latenza totale è quindi doppia rispetto al sinolo disco (con qualche differenza visto che dobbiamo aspettare che entrambe le reads vengano portate a termine).

- **Steady-state throughput** Nel caso di scrittura sequenziale in tutta una striscia, la banda equivale a $(N-1)*S$ `MB/s`. Anche nel caso di una random read, in quanto non abbiamo bisogno di leggere il disco di parità. Per quanto riguarda le random writes invece, dobbiamo aggiornare il valore di parità. Ci sono due metodi:
  - **Additive parity** per sapere il valore di parità si legge in parallelo il valore dei blocchi nella striscia e si esegue una `XOR` con il valore del nuovo blocco. Maggiore sono i dischi e più costosa diventa.
  - **Subtractive parity** Vengono letti il bit che bisogna cambiare e il bit di parità. Se il bit vecchio è uguale al nuovo, si lascia uguale il bit di parità, altrimenti diventa il suo opposto.

Quando vogliamo scrivere due dati in due strips diverse e in due dischi diversi, possiamo eseguire la scrittura in parallelo, ma il disco di parità dovrà essere acceduto sequenzialmente. Questo problema è noto come **small-write problem**. Le performance (*pessime*) di piccole random writes è pari $(R/2)$ `MB/s`.

3.3.5 RAID level 5: rotation parity

Funziona come il 4, ma per ridurre il carico sul disco di parità nelle operazioni di scrittura, distribuisce i blocchi di parità fra i vari dischi. Il difetto di questo approccio è che, in caso di guasto di un disco, è più complessa la ricostruzione. Questo livello fornisce comunque la migliore combinazione in termini di prestazioni, affidabilità e capacità di memorizzazione. La parità viene fatta ruotare in tutti i dischi in modo da eliminare l'effetto **collo di bottiglia** del livello 4.

Performance

- **Capacità** $(N-1)*B$.
- **Affidabilità** Uguale al RAID 4.
- **Performance** Rispetto al RAID 4 ha performance notevolmente migliorate grazie alla disposizione dei blocchi di parità all'interno dei dischi.
- **Single-write latency** Uguale al RAID 4.
- **Steady-state throughput** Nel workload random, il RAID 5 funziona leggermente meglio rispetto al RAID 4 perchè ora possiamo utilizzare tutti i dischi. Le performance di una write random è notevolmente migliorata visto che ora siamo in grado di parallelizzare le richieste e non dobbiamo più accedere sequenzialmente al disco di parità. Dato un elevato numero di richieste random, saremo in grado di tenere occupati praticamente tutti i dischi. Se questo è il caso, allora la larghezza di banda per piccole writes sarà di $(N/4)*R$ `MB/s`. Il fattore di perdità

4 è dovuto al fatto che ogni write genera 4 operazioni di I/O, che è semplicemente il costo dovuto all'impiego della parità (*read data, read parity, write data, write parity*).

3.3.6 Riassunto prestazioni

	RAID 0	**RAID 1**	**RAID 4**	**RAID 5**
Capacity	$N * B$	$(N * B)/2$	$(N-1) * B$	$(N-1) * B$
Reliability	0	$1_{sure\ \frac{N}{2}\ if lucky}$	1	1
Throughput				
- Sequential read	$N * S$	$(N/2) * S$	$(N-1) * S$	$(N-1) * S$
- Sequential write	$N * S$	$(N/2) * S$	$(N-1) * S$	$(N-1) * S$
- Random read	$N * R$	$N * R$	$(N-1)R$	$N * R$
- Random write	$N * R$	$(N/2) * R$	$\frac{1}{2}R$	$\frac{N}{4}R$
Latency				
- Read	T	T	T	T
- Write	T	T	2T	2T

3.4 File e directories

Il file è l'elemento alla base della memorizzazione. È un array lineare di bytes che può essere letto o scritto. I file sono contraddistinti da un low-level name noto come **inode number**.

Una **directory**, così come un file, ha un inode number associato. Il suo contenuto è una lista di coppie "user-readable-name"/"low-level-name". La directory principale è chiamata **root** (/), le altre sono chiamate sub-directories. Files e directories possono avere lo stesso nome a patto che siano in nodi diversi dell'albero delle directory.

3.4.1 Creazione di un file

È possibile creare file in svariati modi, la pratica più diffusa è l'invocazione della system call **open** utilizzando la flag `O_CREAT` che prende come argomenti il nome del file, e una serie di parametri per specificare i permessi e il tipo di operazioni eseguibile sul file.

```
int fd = open("foo",O_CREAT|O_WRONLY|O_TRUNC,S_IRUSR|S_IWUSR)
```

La open ritorna un intero definito **file descriptor**. Il file descriptor è un indice utilizzato dal file system per individuare un file aperto nella **file table**.

3.4.2 Lettura e scrittura (sequenziale) di un file

Ogni processo ha 3 file aperti, **standard input, output ed error**. Essi sono rappresentati dai file descriptors 0, 1 e 2. Per leggere un file viene invocata la `open()` e poi la `read()` per leggere ripetutamente i bytes dal file. Il primo argomento di `read()` è il

file descriptor, il secondo è un puntatore a un buffer dove verrà posizionato il risultato della chiamata, il terzo argomento è la dimensione del buffer. Una volta terminata la `read()`, restituisce il numero di bytes letti e viene invocata la `write()` con argomento il file descriptor 1 (std output). Il secondo argomento è ciò che deve stampare, cioè i bytes appena letti dalla `read()`, il terzo è la grandezza del buffer da stampare. Infine viene chiamata nuovamente la `read()` e se quest'ultima restituisce 0, specifica che non ci sono altri bytes da leggere e viene quindi invocata la `close()`. Il procedimento di scrittura di un file è pressochè identico.

3.4.3 Lettura e scrittura (non sequenziale) di un file

Per poter accedere direttamente alla posizione designata viene adoperato un **random offset** utilizzato per spostarsi all'interno del file. I possibili valori di whence (*da dove*) possono essere:

- **SEEK-CURE** ci si sposta di offset posizioni dal valore attuale. Ad esempio se sono al byte 50 e offset è pari a 100, la posizione di arrivo sarà il byte 150.
- **SEEK-END** Ci si sposta di offset posizioni dalla fine del file. Ad esempio `lseek(fd, 10, SEEK_END)` la posizione di arrivo sarà ottenuta spostandoci di 10 posizioni dalla fine del file `fd`.
- **SEEK-SET** viene spostato l'indice all'ofset specificato. Ad esempio, con offset pari a 50, la `lseek` si sposta al byte 50.

3.4.4 Scrittura immediata su disco

Quando viene invocata la write(), il chiamante sta generalmente chiedendo di salvare in modo persistente certi dati. Il file system, per questioni di performance, bufferizzerà queste writes in memoria per un certo istante di tempo (dai 5 ai 30 secondi). Dal punto di vista del chiamante, le richieste sembrano essere servite rapidamente e solamente in rari casi (ad esempio a seguito di un crash dopo l'invocazione della syscall `write()` ma subito prima che i dati vengano scritti su disco) vi sarà una perdita d'informazione. Tuttavia, alcune applicazioni richiedono più di questa "eventuale garanzia". I file systems forniscono, per le applicazioni che necessitano evitare perdita di dati, APIs di controllo addizionali. Nel mondo Unix, l'interfaccia usata per servire a questo scopo è conosciuta come `fsync(int fd)`. Quando un processo invoca la `fsync()` per un certo file descriptor, il file system risponde forzando tutti i dati non ancora scritti (*dirty*) su disco. La `fsync()` ritorna una volta che tutte le writes sono state completate.

3.4.5 Rinominare un file

Per rimominare un file, basta usare la move. `mv foo bar` invoca la `open()` per creare un file temporaneo, viene scritta la nuova versione del file attraverso la `write()` e in fine viene forzata la scrittura su disco con `fsync()`. Quando si è certi di aver salvato il

nuovo nome, viene rinominato il file temporaneo col nome effettivo (questo passaggio viene fatto atomicamente grazie alla "rename").

3.4.6 Informazioni dei file

Il file system mantiene una serie di informazioni note come **metadati** relative ad ogni file memorizzato. Per visionare i metadati di un file, si possono utilizzare le system call `stat()` o `fstat()`. I file system tengono i metadati in strutture chiamate **inode**.

3.4.7 Rimozione dei file

Per rimuovere un file viene usato il comando `rm`. Questo chiama la system call `unlink()`, che riceve come argomento il nome del file da rimuovere e restituisce 0 in caso di successo.

3.4.8 Creazione directories

Non è possibile scrivere direttamente in un directory, questo perchè il suo formato è considerato un metadato del file system. Per creare una directory viene adoperata la system call `mkdir()`. Quando viene creata è considerata vuota. Nello specifico contiene due entries: (. , se stessa), (.. , direcotry genitore).

3.4.9 Leggere le directories

Per leggere il contenuto di una direcotry si utilizza il comando `ls` che invoca le syscall `opendir(), readdir()` e `closedir()`.

3.4.10 Eliminare directories

Per eliminare una directory viene invocata la system call `rmdir()`. Prende come argomento il nome (user-readable-name) della directory che si vuole eliminare e, se essa è vuota serve la richiesta, altrimenti esce stampando un messaggio di errore. Questo per evitare perdita di dati importanti.

3.4.11 Hard links

La syscall `link()` riceve due argomenti: un vecchio e un nuovo pathname. Quando linchiamo un nuovo file-name a uno vecchio, essenzialmente stiamo creando un nuovo modo per riferirsi allo stesso file. Viene usato il prgramma `ln`.

Un **hard link** è l'associazione del nome di un file ad un inode del sistema. Quando creiamo un file, viene creata una struttura inode che terrà traccia di una serie di informazioni relative a tale file (grandezza, locazione dei suoi blocchi su disco, ...), e successivamente viene collegato (link) uno human-readable name a quel file. Infine il link viene posizionato nella directory in cui il file è stato creato. Il file system mantiene

un contatore noto come **link count** associato a ciascun inode del sistema. Se è a 0, il file system libera l'inode e i blocchi di dati relativi, eliminando realmente il file.

3.4.12 Symbolic links

Gli hard link sono limitati in quanto non è possibile creare un hard link a una directory (potrebbe generare un ciclio all'interno dell'albero delle directory) e non è possibile creare un hard link a un file in altri partizioni del disco (gli inode number sono unici all'interno di un particolare file system). Un symbolic link (`ln -s`) è un file vero e proprio e sono il terzo tipo di dati che il file system conosce (*file, directory e soft link*). Il contenuto del collegamento è il pathname al file a cui fa riferimento e non i dati effettivi. Se si elimina il filename originale, il collegamento punta a un pathname inesistente (**dangling reference**).

3.4.13 Making and mounting the file system

Come assembliamo un unico directory tree a partire da molti file systems? Un **file system** è un meccanismo mediante il quale i file sono collocati e organizzati su un dispositivo di archiviazione. Per creare un file system viene fornito un tool `mkfs`. A seguito di questo comando siamo in grado di creare un file system vuoto sul device passatogli come argomento. Per rendere l'albero accessibile si usa il comando `mount` (copia il file system tree e lo incolla nel punto specificato in mount).

3.5 File system implementation

Un file system è una componente **puramente software**.

3.5.1 VSFS - very simple file system

La prima cosa da fare è dividere il disco in N blocchi (*es* `4KB`) indirizzabili da 0 a $N-1$. Lo spazio complessivo sarà pari a $N * $ `4KB`. Questi blocchi sono divisi in:

- **data region** dove si contengono i dati utente (*gran parte della memoria*).
- **inodes table** dove si contengono gli inodes (*il numero di inodes rappresenta il numero massimo di files che possono essere mantenuti da VSFS*).
- **allocation structure** per tenere traccia di quando gli inodes o i data-blocks sono liberi. Vengono divise in
  - **data bitmap**
  - **inodes bitmap**
- **superblock** è un singolo blocco che contiene informazioni relative al file system stesso (numero inodes, data blocks, ...). Questo blocco include anche un magic number per identificare il tipo di file system.

La bitmap è una struttura piuttosto semplice, ogni bit viene usato per indicare quando il corrispondete oggetto/blocco è libero (0) o occupato (1).

3.5.2 Organizzazione dei file, l'inode

Gli **inodes** (*index node*) sono strutture adoperate da tutti file system. Ogni inode viene implicitamente riferito da un (**inumber**) di un file. In VSFS, dato un inumber `i`, per calcolare dove l'inode corrispondente è collocato su disco si fa:

```
i * (dimensione inode) + (indirizzo inizio inode region)
```

I dischi non sono indirizzabili attraverso i byte. Consistono in un vasto numero di settori indirizzabili, generalmente da `512B`.

```
blk = inumber * sizeof(inode_t) / blockSize

sector = ((blk * blockSize) + inodeStartAddr) / sectorSize
```

All'interno di un inode troviamo informazioni quali:

- **Tipo** di file (*directory, normale o link*).
- **Dimensione**
- **Numero di data-blocks**
- **Informazioni relative alla protezione**.
- **Informazioni relative al tempo** (*quando è stato creato, modificato, ultimo accesso, ...*).

Gli inodes possono riferirsi ai data blocks in due modi:

- **direct pointers** (disk addresses) all'interno dell'inode; ogni puntatore fa riferimento a un blocco su disco appartenente al file. Questo approccio è limitato se si vuole creare un file più grande della dimensione di un blocco moltiplicato per il numero di direct pointers.
- **indirect pointers** Invece di puntare a un blocco che contiene dati utente, punta a un blocco che contiene più puntatori, i quali puntano ai dati utente. In questo modo, un inode potrebbe avere un numero fissato di direct pointers (*esempio 12*) e un singolo indirect pointer. Se il file cresce abbastanza, viene allocato un indirect block (dalla data-block region del disco) e lo slot dell'inode relativo all'indirect pointer è impostato in modo tale che punti a esso. Questa organizzazione dei puntatori è un sistema di indicizzazione che prende il nome di **multi-level index**.

Per supportare file di dimensione ancora maggiore, basta aggiungere un altro puntatore all'inode: **double indirect pointer**. Questo puntatore si riferisce a un blocco che contiene puntatori a blocchi indiretti, ognuno dei quali contiene puntatori a data blocks. È possibile usare anche un **triple indirect pointer** se si vuole. Attraverso questo sistema di indicizzazione, siamo in grado di supportare file di dimensioni maggiori. I tipi di multi-level index introdotto sono:

- **Direct indexing.** Il file può essere grande massimo 12 blocchi, ciascuno di dimensione `4K`.
- **Single indirect indexing.** Il puntatore punta a un blocco che contiene altri puntatori. Ciascuno di questi punta a un blocco che contiene dati utente. Avendo puntatori da `4B` si ha che un blocco di puntatori suddiviso in fette da `4B` (1024 indirizzi da `4B` ciascuno, un file può essere grande `(12 + 1024) * 4KB = 4144KB`).
- **Double indirect indexing** il puntatore punta a un blocco di puntatori, ciascuno dei quali punta a un altro blocco di puntatori, ciascuno dei quali, infine, punta a un blocco che contiene i dati utente. Il file può essere grande `4GB`.
- **Triple indirect indexing** in questo caso riesco a indicizzare un numero enorme di blocchi, riuscendo a supportare file di dimensioni notevoli. Nello specifico si ha `(12 + 1024 +` 1024^2 `+` 1024^3 `)*4 KB = 4 TB.`

Il motivo per cui si utilizza un albero così sbilanciato è perchè la maggior parte delle volte i file sono di dimensione piccola. Viene ottimizzato il caso più frequente.

3.5.3 Organizzazione delle directories

In VSFS il contenuto di una directory è una coppia **entry name, inode number**. Le directories vengono spesso trattate dai file system come tipo di files speciali. In questo modo, una directory avrà un inode da qualche parte nell inode table (con il campo `type` dell'inode marcato come directory.

3.5.4 Free space management

La gestione dello spazio libero è molto importante, in VSFS ci sono due bitmap che fanno ciò. Se creiamo un file, dobbiamo allocare un inode per esso. Il file system cercherà nella bitmap un inode libero, lo setterà a 1 e lo allocherà al file. Un procedimento simile viene adoperato per quanto riguarda i blocchi di dati.

3.5.5 Reading to disk

A seguito della syscall `open()` il file system dovrà trovare l'inode del file avendo a disposizione solamente il pathname completo, per ottenere informazioni base riguardo a esso (*dimensione, permessi, ...*). Tutti gli attraversamenti partono dalla directory **root**.

3.5.6 Writing to disk

Scrivere su un file è un processo simile a quello descritto in precedenza. Per prima cosa è necessario aprire il file invocando la syscall `open()`. Quindi, l'applicazione può invocare la `write()` per aggiornarne il contenuto. Infine, il file viene chiuso tramite la syscall `close()`.

3.5.7 Caching e buffering

Per evitare che le prestazioni calino eccessivamente, i file systems usano spesso la memoria centrale (DRAM) in modo aggressivo, per "nascondere/far sostare" blocchi "importanti". I primi file systems introdussero una fixed-size cache con lo scopo di mantenere accessibili in modo rapido blocchi "popolari" (di uso frequente). Algoritmi quali LRU, decidevano quale blocco mantenere nel buffer. La file-cache veniva generalmente allocata a boot-time per essere approssimativamente il 10% della memoria totale. Nei moderni sistemi operativi viene utilizzato un metodo differente, noto come partizionamento dinamico. Vengono integrate pagine di memoria virtuale e pagine di file system in una unified-page-cache. Con questo approccio la memoria può essere allocata in modo più flessibile attraverso la memoria virtuale e il file system, in base alle necessità in un certo istante di tempo.

3.6 Crash consistency

Se durante l'aggiornamento del file system si verifica una perdita di energia o un crash del sistema, questo è noto come **crash-consistency problem**. Supponiamo di scrivere in un file. Per farlo, dobbiamo eseguire tre write: aggiornare l'inode, aggiungere un nuovo bloco nella regione dati e aggiornare la data bitmap. Se una di quete write non viene completata per via di un crash, il file system viene lasciato in uno stato "inconsistente".

3.6.1 Crash scenarios

- **Solamente il data-block è stato scritto su disco** Non rappresenta un problema per la crash consistency perchè è come se il blocco non fosse mai esistito. L'utente potrebbe perdere però dei dati importanti.

- **Solamente l'inode viene scritto su disco** e punterà a uno o più data blocks su disco dentro ai quali non è stata inserita alcuna informazione; essi conterranno **garbage data** del disco. C'è **file system inconsistency** in quanto l'inode punta al blocco dati mentre la bitmap lo ha segnato come libero.

- **Solamente la bitmap viene scritta su disco** La bitmap indica che il blocco è occupato ma non si ha nessun inode che punta ad esso (**inconsistency**). Ciò porta a dello **space leak**.

- **Inode e bitmap sono scritti su disco** Il blocco a cui punta l'inode contiene ancora una volta **garbage data**. Ciò non scaturisce file system incosistency.

- **Inode e blocco vengono scritti su disco** Il blocco non potrà essere acceduto e rischia di essere sovrascritto perchè marcato come libero nella data bitmap.

- **Bitmap e blocco vengono scritti su disco** Il blocco non sarà accessibile poichè nessun inode punterà ad esso. Non sappiamo quale sia il file associato a tale blocco.

3.6.2 FSCK - File System Checker

FSCK serve per trovare inconsistenze del file system e ripararle. Questo tool viene eseguito prima che il file system venga montato e reso disponibile. Una volta finito il lavoro, questo programma dovrebbe rendere il file system consistente e accessibile all'utente. FSCK esegue sequenzialmente le seguenti fasi:

- **Sanity check** Controlla che il superblock sia corretto controllando ad esempio che la grandezza del file system sia maggiore del numero di blocchi allocati e così via. L'obiettivo è trovare un superblock corrotto e sostituirlo con una copia funzionante.

- **Free blocks** Scansiona gli inodes, indirect blocks, double indirect blocks e così via. In questo modo produce una versione corretta della data bitmap (fidandosi delle informazioni contenute negli inodes). La stessa scansione viene fatta per gli inodes e la loro relativa bitmap.

- **Inode state** Ogni inode viene controllato per verificare se corrotto o se ha altri problemi. Se l'inode è ritenuto sospetto e ha problemi che non possono essere risolti facilmente, viene pulito e, successivamente, viene aggiornata la bitmap.

- **Inode links** Controlla il link count di ogni inode allocato e costruisce il proprio link count per ogni directory presente. Se il link count di FSCK è differente da quello del file system per una data directory si procede ad aggiornare quest'ultimo col valore corretto.

- **Duplicates** Controlla che non ci siano puntatori duplicati. Se un inode è corrotto, potrebbe venir pulito. Alternativamente, viene copiato il blocco, in questo modo ogni inode ha la propria copia.

- **Bad blocks** Viene fatto un controllo anche per i puntatori a blocchi possibilmente corrotti. Un puntatore è considerato corrotto se punta al di fuori di un range valido, e venogno rimossi i puntatori dall'inode o dagli indirect blocks.

- **Directory checks** Viene eseguito un controllo di integrità sulle directories. FSCK controlla contenuto di ogni directory, assicurandosi che "." e ".." siano le prime

entries e che ogni inode riferito a una entry di una directory sia allocato correttamente. Si assicura che non ci siano cicli.

Il problema di questo approccio è che è **estremamente lento**.

3.6.3 Journaling

L'idea base è la seguente: prima di aggiornare le strutture dati su disco, viene generata una "nota" (da qualche parte nel disco in una struttura conosciuta come log) che descrive l'operazione che il file system è in procinto di svolgere (**write-ahead**). In questo modo, se si dovesse verificare un crash durante l'aggiornamento delle strutture dati, è sempre possibile andare a consultare la nota e sapere esattamente cosa sistemare (e come). Dopo un crash siamo in grado di andare direttamente alla fonte del problema.

Nel file system `linux ext3` la nuova struttura chiave è il blocco **journal**, il quale occupa un piccolo spazio all'interno della partizione. Supponiamo di voler eseguire l'aggiornamento: `inode (I[v2]), bitmap (B[v2]), data block (Db)`. Prima di scrivere su disco dobbiamo scrivere nel log (journal) le informazioni necessarie. Oltre ai tre soliti blocchi, si scrivono i blocchi `TxB` (*transaction begin*) all'inizio e `TxE` (*transaction end*) alla fine. `TxB` contiene informazioni relative all'aggiornamento in attesa, tra cui, ad esempio, l'indirizzo finale dei blocchi oltre a un qualche tipo di **transaction identifier (TID)**. Il blocco finale, `TxE` è un marcatore della fine della transazione e conterrà anch'esso il TID. Una volta che il blocco journal risiede su disco, è possibile andare ad eseguire l'aggiornamento delle vecchie strutture dati. Questo processo è chiamato **checkpointing**. La sequenza delle operazioni svolte è:

- **Journal write** vengono inserite le informazioni necessarie nel blocco journal.
- **Checkpoint** vengono apportate le opportune modifiche su disco, alle locazioni corrette.

Se dovesse verificarsi un crash mentre stiamo scrivendo i dati nel blocco journal, per evitare di scrivere `garbage data`, il file system esegue la scritture in due steps. Per prima cosa scrive tutti i blocchi tranne `TxE` nel journal. Quando la scrittura di questi blocchi verrà portata a termine, il file system può procedere a scrivere anche `TxE`. Il punto chiave di questo approccio è l'**atomicità**. Il disco garantisce che le scritture da `512B` vengano eseguite atomicamente. Il blocco `TxE` dovrebbe quindi essere grande `512B` per fare in modo che funzioni correttamente. Il protoccolo di aggiornamento del file system sarà quindi:

- **Journal write** viene scritto il contenuto della transazione nel log (*TxB,metadati e dati*).
- **Journal commit** Viene scritto il *commit block* della transazione `TxE` nel log e si attende il completamento di questa write.
- **Checkpoint** Viene applicato l'aggiornamento delle opportune locazioni su disco.

3.6.4 Recovery

Se il crash avviene prima che la scrittura della transazione sia completa, l'update in attesa viene semplicemente saltato. Se il crash avviene dopo che la transazione ha scritto nel *commit block* ma prima che il checkpoint sia completato, il file system può riprendere l'aggiornamento: quando il sistema viene avviato, il recovery process del file system scansiona il log e cerca le transazioni che hanno provato ad accedere a disco (committed). Queste transazioni vengono eseguite nuovamente in ordine, nel tentativo di aggiornare le opportune strutture dati. Questo processo di recovery viene chiamato **redo-logging**. Se il crash avviene durante il checkpoint possiamo sempre leggere il contenuto del journal durante il recovery ed eseguire nuovamente l'aggiornamento (le writes vengono rieseguite tutte).

Il protocollo introdotto potrebbe generare traffico I/O addizionale. Per evitare questo problema, viene introdotto un **buffer globale** nel quale vengono inserite le transazioni. Tuttavia, se si continuano ad aggiungere transazioni, il buffer finirà col riempirsi molto facilmente. Si vengono a creare due problemi:

- Più grande è il log, più tempo ci vorra per eseguire il recovery.
- quando il log è pieno o quasi, non è possibile avanzare altre transazioni; questo rende il file system inutile.

Questi due problemi vengono risolti con la **circular log**, i file system trattano i logs come strutture dati circolari. Per fare ciò, quando una transazione viene completata, il file system deve liberare lo spazio relativa a essa. Un modo molto semplice per fare ciò consiste nel marcare le transazioni vecchie e non-checkpointed in un **superblock del journal**. Nel journal superblock ci sono abbastanza informazioni per capire quale blocco non ha ancora raggiunto la fase di checkpoint. Questo riduce notevolmente anche il recovery time, visto che non bisogna ripetere tutte le transazioni ma solo alcune. Lo schema del protocollo diventa quindi:

- **Journal write** Vengono scritti `TxB` e i tre blocchi mediani.
- **Journal commit** Viene scritto il blocco `TxE`.
- **Checkpoint** Vengono apportate le opportune modifiche alle strutture dati del file system.
- **Free** Se la transazione è completata, viene marcata come libera nel superblock.

Per evitare di scrivere due volte su disco le stesse cose (**data journaling**), vengono introdotte altre soluzioni come il **metadata journaling**, in cui vengono scritti nel journal solamente i metadati. In questo caso si devono per forza scrivere prima i dati utente, e poi fare tutto il processo di journaling.

3.6.5 Block reuse

Il metadata journaling porta ad un problema: abbiamo una directory `foo` nella quale viene aggiunta una entry; il contenuto della directory viene scritto nel log. Assumiamo che la locazione del datablock di `foo` sia `1000`. A questo punto l'utente elimina la directory `foo` e ogni file in essa, liberando il blocco `1000`. L'utente crea un nuovo file `foobar` che finisce col riutilizzare lo stesso blocco `1000`. L'inode di `foobar` viene quindi scritto su disco, così come i suoi dati. Supponiamo ora che si verifichi un crash e che tutte queste informazioni siano ancora nel log. Viene sovrascritto il blocco `1000` contenente i dati di `foobar` dal vecchio contenuto della directory `foo`. Esistono soluzioni a questo problema come la **revoke record**: eliminare una directory causa una scrittura del revoke record nel journal. Quando il journal verrà esaminato in fase di recovery, il sistema scansionerà prima il revoker record. I dati che sono stati revocati non vengono mai riscritti.