Notes Software Engineering

Brendon Mendicino

April 28, 2023

Contents

1 Memoria

La memoria si compone si suddivide in:

- registri;
- cache
- memoria primaria;
- memoria secondaria;

Uno principali problemi di accesso alla memoria è quello di assicurarne la sua protezione, ovvero evitare che un programma in memoria riesca ad accedere alle zone di memoria di altri programmi. Una volta fatto il bootstrap, sia il SO che i programmi verranno caricati in memoria, per evitare che ogni processo abbia una vista al di fuori del suo scope, si possono usare dei controlli a livello della CPU, esistono dei registri chiamati **base** e **limit**, che attraverso dei meccanismi riescono a effettuare la protezione della memoria. Esiste anche un altro problema che è quello della **relocation**, che consiste nel come mappare gli indirizzi delle varie istruzioni di salto e dei vari puntatori una volta che il programma viene caricato in memoria. Se si utilizzassero degli indirizzi statici (creati solo in fase di compilazione prendendo come riferimento d'inizio del programma l'indirizzo 0), una volta che il programma viene caricato in memoria le istruzioni punterebbero sempre allo stesso indirizzo, ma i programmi non partono tutti dall'indirizzo 0, sorge dunque il probelema di come gestire le istruzioni che fanno riferimento ad altre parti del programma. Immagginiamo che gli indirizzi vengano creati una mentre il programma viene caricato in memoria, cosa succede se quel programma viene mosso in un altra parte della memoria? I rifermenti degli indirizzi sarabbere tutti sbagliati o addirittura potrebbero puntare a pezzi di altri programmi in esecuzione. Per questo motivo quando si utilizzano valori di indirizzi, essi vengono sommati al **base register**, mentre le boundry del programma in memoria vengono salvate nel **limit register**, grazie a questa tecnica **hardware** è possibile scrivere i programmi come se gli indirizzi del programma partissero da 0.

1.1 Indirizzamento

Gli indirizzi vengono rappresentati in modo diverso a differenza della fase di vita di un programma: durante la compilazione vengono considerati come simbolici, durante la compilazione agli indirizzi è fatto il **bind** ad un indirizzo relativo (base) in modo da poter essere rilocato (questa somma è fatto dall'hardware), durante il linking o durante il loading. Il **binding** si può fare:

- **in compilazione**: fatto quando molto semplice come nei sistemi embedded, ad esempio quando esistono solo due programmi;

- **in fase di load**: viene fatta la rilocazione durante il caricamento in memoria;
- **in esecuzione**: viene fatto il binding degli indirizzi in modo dinamico;

Per risolver questo problema ci si affida all'hardware, che è incaricato di fare la traduzione: l'indirizzo rimane lo stesso (logico) all'interno del processore e del programma, prima di arrivare all'address bus viene tradotto in indirizzo fisico, esiste dunque una dicotomia tra indirizzo logico e fisico, in questo modo quando si scrive un programma, l'indirizzo parte sempre da 0. Esistono due tipi d'indirizzamento che sono di tipo logico, dove l'intervallo degli indirizzi utilizzabili è logico, e un indirizzamento fisico, dove il range è limitato dalla memoria del sistema. Per effettuare questa traduzione da indirizzo fisico a indirizzo logico e viceversa si utilizza una **MMU** (Memory Management Unit). Il modo più facile per realizzare una MMU, è quello di usare un **relocation register**, ovvero un registro che contiene il valore da aggiungere un indirizzo logico per fare un indirizzo fisico, il modello più semplice di MMU è fatto da un sommatore e un comparatore. Il **relocation register** va a sostituire il **base register**.

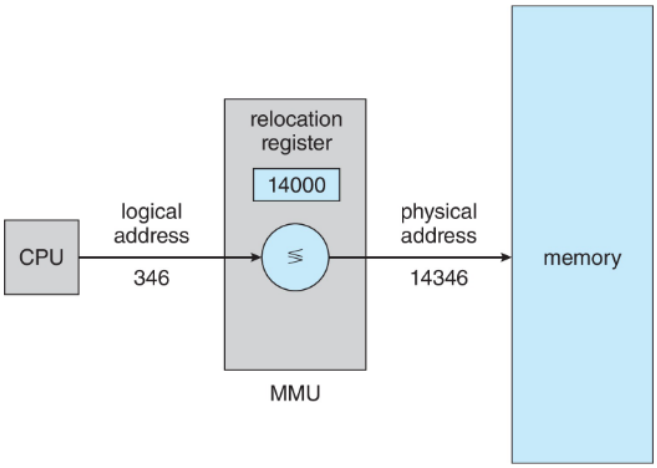


Figure 1: Basic MMU

Per aumentare le prestazioni ed usare la memoria in modo più efficente si possono usare delle tecniche dinamiche.

- Si parla di **dynamic loading** quando, un programma viene caricato in memoria principale in modo frammentato, utilizzando solo i componenti che effettivamento vengono chiamati;
- Si parla di **dynamic linking** quando i file che contengono le funzioni che devono essere linkate (come le librerie standard) non vengono inserite all'interno dell'eseguibile, ma gli indirizzi vengono risolti in modo dinamico durante l'esecuzione;

- Il **link statico** è quando si crea un eseguibile con tutte le funzioni dentro, di fatto il loader carica tutto quando in memoria.

Il **dynamic loading** vuol dire che una routine non è caricata finché non è necessaria, questo può essere fatto quando il programmatore ne è consapevole, infatti il processo di load non è trasparente:

```
void myPrinf(**args) {
  static int loaded = 0;
  if (!loaded) {
    load("printf");
    loaded = 1;
  }
  printf(args);
}
```

Il **linker-assisted DL** usa una chiamata fasulla che prima chiama la load della funzione linkata dinamicamento e poi la invoca, questi piccoli pezzi di codice vengono detti **stub**.

Le **shared libraries** sono in grado di condividere le risorse, infatti se più processi utilizzano la stessa funzione essa viene messa a disposizione a livello globale e per ogni nuova chiamata non ci sarà bisogno di chiamare una load.

1.2 Allocazione in Memoria

Come si alloca memoria per un programma (immagine)? La più semplice è l'allocazione contigua, dove si vede la RAM separata in due partizioni, una per il SO e una per i processi (indirizzi più bassi), per caricare un processo si parte da un indirizzo d'inizio e un indirizzo di fine, la MMU vista prima funziona solo con i casi di allocazione contigua (basilare). Una tecnica più efficente è l'**allocazione contigua con partizione variabile**, che consiste in: quando ci sono più buchi ci sono delle politiche differenti per inserire nuovi programmi:

- **first-fit**: il primo che si trova;
- **best-fit**: il buco con la dimensione più piccola;
- **worst-fit**: il buco con la dimensione più grande;

La **frammentazione** è definita come la *sparsità dei buchi all'interno della memoria*. Si dice **esterna** perché è al di fuori dei processi, si dice **interna** quando è interna al processo, ovvero che ha più memoria di quello che serve. La frammentazione ha bisogno di **compattazione**, il SO sposta i pezzi e poi si riparte, partiziona la memoria in zona libera e zona occupata, per fare una **deframmentazione** (o compattazione) vuol dire creare solo due partizioni (parte processi e zona libera), per effettuare la compattazione bisogna che i processi si possano spostare, inoltre un processo non

può essere spostato se in quel momento sta effettuando delle operazioni di IO, una soluzione per il problema dell'IO è il **latch job** in cui solo la parte che sta facendo IO non può essere spostata, oppure si utilizza un **buffer IO del kernel**, si fa IO solo in buffer del sistema operativo.

Definition 1.1 – Backing Store

Con **backing store** si definisce uno spazio della memoria secondaria in cui non si salvano programmi ma vengono immagazzinati dei dati che altrimenti dovrebbero andare in RAM.

La **Paginazione** risolve i problemi della allocazione contigua, al suo posto si utilizzano della **pagine** che sono l'unità minima di allocazione e di trasferimento, questo risolve il problema della frammentazione.

Definition 1.2 – Pagine

Le partizioni della memoria logica.

Definition 1.3 – Frame

Le partizioni della memoria fisica.

Definition 1.4 – Blocco

...

Tipicamente la loro dimensione è un multiplo di due, sia pagine che frame hanno la stessa dimensione, da qualche parte andranno salvate le informazioni di mapping tra pagine e frame, si utilizza una **frame table**, dove ogni riga corrisponde a una mappatura. Anche utilizzando la paginazione si ha frammentazione interna.

Un indirizzo generato dall CPU si divide in:

- **numero di pagine (p)**:
- **numero di offset (d)**:
- **numero di frame (f)**:

Un indirizzo è composto da: $(\underbrace{p-d}_{\text{parte logica da sostituire}} , \underbrace{d}_{\text{accesso interno al frame}})$

Example 1.1

Se si hanno frame piccoli diminuisce la frammentazione ma aumentano anche le righe della tabella.

Nelle righe della tabella si aggiungono dei bit in più per rappresentare delle informazioni aggiuntive rispetto al frame:

- **parte di protezione**: specifica che la parte di codice non può essere scritta;
- **modify bit**: pagina modifica;
- **page present/absent**: pagina presente in memoria;
- **page referenced**;
- **caching disabled**;

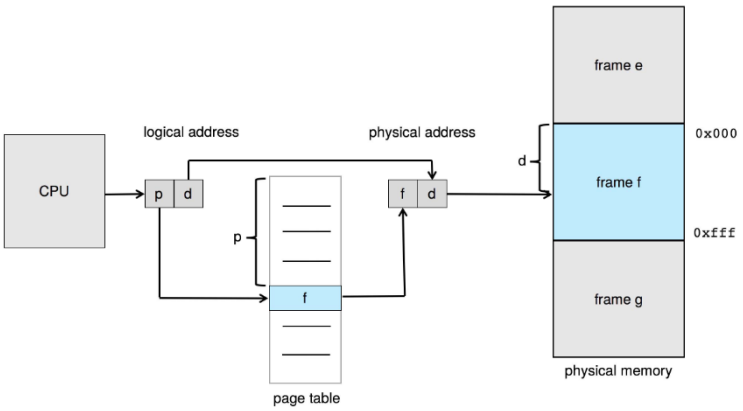


Figure 2: Page Table

Come si implementa una **page table**? La page table si trova in memoria RAM, per sapere dove si trova si utilizzano due rigistri: **page table base register** e **base table lenght register**, infatti la base table si trova in memoria contigua. Per velocizzare questa operazione si può spostare questa tabella all'interno della CPU (le operazioni di accesso alla memoria sono costose), si usa la **Translation Look-aside Buffer** (TLB), un tipo di memoria in cui si accede per contenuto (memoria associativa). Si aggiunge anche un altra informazione **ASID** in cui viene salvata l'infromazione del processo a cui la pagina appartiene, se non è presente questa informazione i processi si contendono la TLB. Quando avviene un **TLB miss** se reinserisce la pagina nella TLB e poi si ritenta, utilizzando una politica scelta. Anche usando una TLB si mantiene comunque la page table in RAM in caso di miss, in modo da recuperare il frame ed inserirlo nella TLB.

Lo schema finale è la combinazione di TLB e page table.

Il **Tempo di accesso effettivo** (EAT) in memoria è il tempo che mi costa accedere alla RAM:

$$EAT = h \cdot M + (1 - h) \cdot 2M$$

- h = TLB hit ratio;

- M = Memory access time;

L'accesso alla memoria è raddoppiato perché si deve prima recupare il frame dalla page table e poi si può accedere alla memoria.

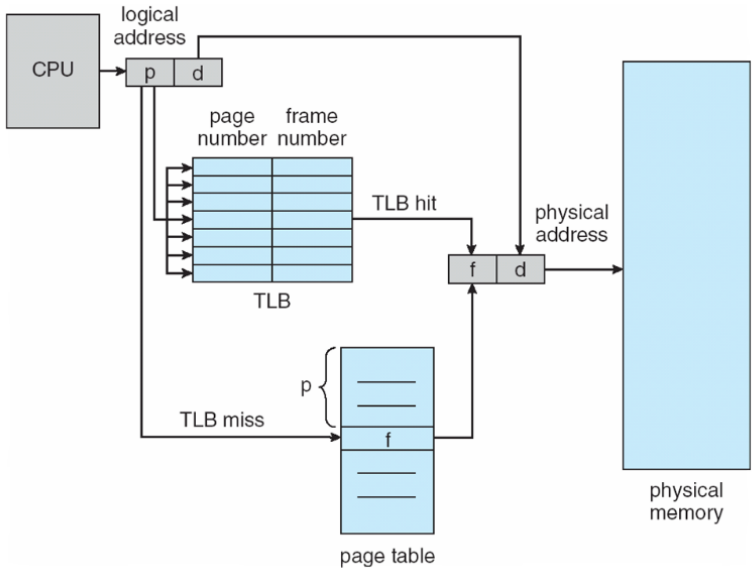


Figure 3: Paginazione Con TLB

La protezione è implementata associando un bit di protezione per un frame specifico, e quindi per la rispettiva pagina, grazie a questo bit si possono definire parti di codice che indica che il frame è in solo scrittura. Anche un altro bit associato è il **validity** che indica quando un frame è valido, ovvero quando c'è ed è associata a un indirizzo in RAM, o quando non è valida che vuol dire che il frame non esiste o che non è associato a un indirizzo, se si cerca di accedere un frame non valido viene lanciata una **trap**.

La tabella della pagine permette di **condividere le pagine** (anche se in realtà sono i frame), se più processi hanno delle pagine che sono comuni vengono condivise da entrambi i processi.

La page table è una struttura dati del kernel:

Example 1.2

HP: spazio logico di indirizzamento di 32 bit, una pagina di 4KB (2^{12}), la page table avrà un milione di entry ($2^{32}/2^{12}$), ogni entry della page table sono 4 byte, quindi una page table è grande 4MB. Adesso è possibile avare:

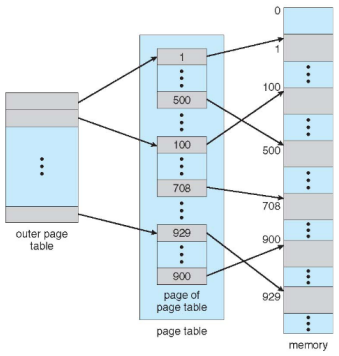
- **page table contigua**:

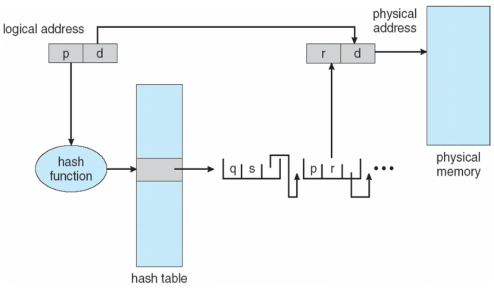
Se la dimensione è critica (troppo grande):

- page table gerarchica;

- hash page table;
- inverted page table;

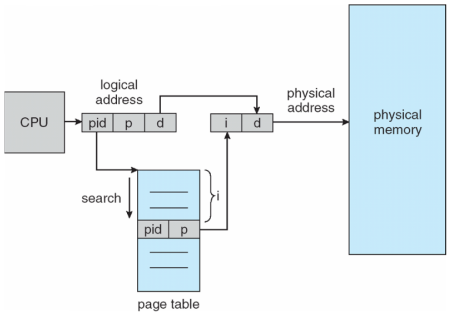
La **la page table gerarchica**, la page table viene divisa in blocchi più piccoli non contigui, si usa una page table di livello superiore che porta alle page table di livello inferiore, questi blocchi devono rimanere contigui, anche se diventano molto piccoli. Il page number viene diviso in due parti, una per la tabella outer ed una per la tabella inner (può avere anche più di due livelli), quando si va su 64 bit e la outer dienta molto grande si può usare solo una parte dell'indirizzamento se l'eseguibile è molto piccolo.





La **hashed page table** permette di creare una tabella di hash, con una funzione di hash direttamente implementata in hardware che dato p in input ritorna l'hash a cui è associato f. Vanno implementate delle liste di collisione, e quindi vieni immagazzinato che p che è la chiave di accesso e come suo valore f. Usando una tabella di hash si potrebbe pensare anche di usare una page table condivisa tra tutti i processi, ma a quel punto andrà inserito l'ID del processo all'interno della chiave primaria.

La **inverted page table** è una tabella condivisa tra tutti i processi, la sua grandezza viene dimensionata ripetto alla grandezza della RAM e non sulla grandezza dei singoli processi, all'interno ogni frame fisico è associato a una pagina. In questo caso le entry che corrispondono a un frame hanno come valore di ritorno la pagina a cui è associato, il problema è che si vuole la traduzione da pagine ad entry e non il contrario, la prima strategia che si può utilizzare è la scansione lineare. Per rendere l'accesso più veloce viene messa una tabella di hash prima di arrivare alla tabella invertita.



1.3 Swapping

Si suppunga che a un certo punto si voglia far partire un processo e che la memoria sia piena, cosa si può fare? Una soluzione è lo **swapping**, ovvero la memoria associata a quel processo viene messa nel backing store (in modo temporaneo) che in linux è la partizione di swap, e viene caricato il nuovo processo, grazie allo swapping si possono avere processi che hanno più memoria occupata di quanto la RAM sia grande.

Example 1.3

Quanto costa il **context switch**? Memoria piena e viene fatto partire un programma, dunque deve essere portato un processo nel backing store. 100MB da fare la swap con velocità di trasferimento di 50MB/sec.

- swap out = 2000ms
- più swap in dello stessa grandezza
- il tempo totole è di 4s

Se il processo sta facendo IO non si può fare lo swap out, una soluzione è non spostare un processo che sta facendo IO ma spostarne un altro, oppure usare un buffer del kernel.

Sui telefoni non esiste lo swaping, vengono in auito le **lifecycle delle Activity e dei Fragment**. Il vantaggio di fare swapping è l'utilizzo del **paging swapping**, dove lo swap in/out diventa page in/out, che consiste nel trasferire solo le pagine e non l'intero processo.

1.4 Esercizi

Esericizi:

Example 1.4

- memoria fisica da 512MB
- allocazione minima 64B
- 128MB al SO
- tabella dei processi contiene: contiene indirizzo iniziale e size
- allocazione worst fit
- partizioni libere sono contenute in lista ordinata con dimensione descrescente
-

Si supponga ...

Example 1.5

Si descrivano brevemente

- tlb: 0.9 hit ratio
- page table a due livelli
- indirizzo in tre parti, 10 11 11 bit
-

Si fa una outer con cui si utillizzano 10bit, ed una inner con 11bit. Per calcolare il numero di inner ci sono due strade, il numero di byte preciso che contengono 100MB oppure si prende la prima potenza di 2 che contiene 100MB. Il resto delle righe delle outer overanno l'invalid bit settato ad 1.

2 Memoria Virtuale

Con **memoria virtuale** si indica che delle pagine esistono solo virtualmente ma non fisicamente, in questo modo si vuole supportare uno spazio d'indirizzamento più grande dello spazio d'indirizzamento fisico. Si parte dalla premessa che *un programma non bisogna di essere tutto in memoria per essere eseguito*, si pensi a un programma parzialmente caricato in memoria che contiene solo i componenti necessari a lui in quel momento, in questo caso la grandezza di un programma non è più vincolato dalla grandezza della RAM, si parla allora di **memoria virtuale**. Avere questa memoria virtuale porta una serie di vantaggi che velocizzano tutte le operazioni di IO tra la RAM e la memoria secondaria. La conseguenza è che viene introdotta la **demand paging**, ovvero che un pagina viene caricata solo quando è richiesta. Il **demand paging** è molto simile allo swapping, infatti il caso limite è il **lazy swapper** dove inizialmente non si ha nessun frame caricato, e ad ogni richiesta viene preso un frame.

Quasi tutto il meccansico si basa sul *validity bit*, infatti l'invalid bit può rappresetnare una pagina che non ha un frame associato, i casi sono due: il frame non esiste proprio, il frame c'è ma deve essere recuperato, infatti il programma viene comunque caricato in memoria (backing store), perché lo spazio è tanto.

Il **page fault**, è una *trap* che viene scatenata quando si cerca di accedere a una pagina con invalid bit, il processo viene bloccato e si passa in modalità kernel, il sistema operativo agisce in due modi:

- **invalid refernce** $\implies$ abort;
- **non in memoria** $\implies$ viene recuperato il frame;

Quando il frame viene portato in memoria dal disco, la page table viene aggiornata e viene fatto ripartire il processo. Un caso estremo di esecuzione è quello di avere nessuna pagina disponibile e poi caricarle in modo incrementale.

Quanto costa una page fault?

1. Trap al sistema operativo, vengono salvati tutti i registri e lo stato del processo;
2. gestisci l'interrupt;
3. l'OS deve verificare che quel frame esiste;
4. recupera il frame facendo IO;
5. quando si fa il trasferimento, l'OS viene fatto uscire dal processore e un altro processo inizia a girarare;
6. viene tirata un eccezione al termine dell'IO;

L'IO è la parte più onerosa, per calcolare l'EAT:

- p = probabilità page fault
- pfo = page fault overhead;
- spo = swap page out;
- spi = swap page in;
- M = memory access
- $EAT = (1 - p) \cdot M + p(pfo + spo + spi)$

Example 2.1

- memory access time = $20ns$;
- avg page-fault service time = $8ms$;

Il page fault è molto onerso, anche con p molto piccoli i tempi sono comunque molto lenti. L'approccio migliore è ottimizzare il tempo del service ed ottimizzare p.

Copy-on-write si genera una copia di un frame solo quando uno dei processi decidono di scrivere su un frame, questo approccio viene utilizzato quando viene chiamata una `fork()`, infatti la pagine dei due processi non venogno copiate ma sono condivise, solo quando una dei due decide di scrivere la page veien copiata ed ogni processo ha i suoi dati modificati.

Che succede se non ci sono frame disponibili? Un processo che richiede una nuova pagina, **rimpiazza** (**page replacement**) una pagina non in uso che viene salvata sul disco, è possibile rimpiazzare una pagina dello stesso processo o anche di altri processi. Questa caratteristica va aggiunta nella funzione di gestione del page fault all'interno del sistema operativo, in questo caso deve essere aggiunto il **modify/dirty bit**, per gestire se delle pagine vengono modificate, se una pagina scelta ha il modify bit settato allore il frame ha un valore diverso rispetto rispetto al disco, allora primad i rimpiazzarla deve essere copiata, altrimenti è inutile salvare perché tanto si trova già sul disco.

Page replacement, se non esiste un frame libero, si deve scegliere un **victim frame** da rimpiazzare, quindi oltread i tempi per portare una pagina in memoria si deve anche portare una pagina su disco. **Algoritmi di Page Replacement**:

- (il processo per il momento ha un numero fisso di frame), il primo algoritmo si basa su una **string reference** che indica il numero di pagina che vengono richieste in sequenza da un processo. Viene definita la **page fault frequency**:

Definition 2.1 – Page Fault Frequency

$$f(A, m) = \sum_{\forall w} p(w) \frac{F(A, m, w)}{len(w)}$$

- A = algoritmo;
- w = string reference;
- p = probability;
- m = number of availably page frames;
- F = number of page fault generated;

- **FIFO**: con 3 frames,
- **Least Recently Used (LRU)**: viene ripiazzata la pagina usata per ultima. In questo caso per ogni pagina esiste un tempo associato alla pagina, questo è molto oneroso dal punto di vista hardware, infatti non viene adottata.
- **LRU algorithm**: ogni pagina ha un contatore, per ogni page fault si deve fare una ricerca del minimo. Esiste un altra implementazione a stack di numeri di pagina.
- **LRU Approximation Algorithm**: ogni riga ha un **reference bit**, inizialmente è zero e quendo si fa una reference alla pagina viene messo a 1, esiste un momento nel passato in cui i reference bit sono stati azzerati, la victim viene presa nel gruppo dei frame con reference bit a 0. **Second-chance algorithm**: si usa una fifo con reference bit, quando si incontra una pagina con 1 viene settato a 0, se si incontra uno 0 allora viene preso il victim.
- **Enhanced Second-Chance Algorithm**: si utilizza una combinazione di reference e modify per scegliere il frame:
  - (0, 0): ...
  - ...
- **Counting Algorithm**:
  - **Least Frequently Used (LFU):**
  - **Most Frequently Used (MFU):**

Esercizi:

Example 2.2
...

- **Page-Buffering Algorithm**: La vittima viene messa nel *free pool*, ..., il SO potrebbe dare il controllo all'applicazione per la gestione della memoria.

Allocazione fissa:

- si da a tutti la stessa misura (equal);
- si da di più a chi ha più bisogno (proporzionale): ad ogni processo si danno i frame in maniera proporzionanta.

Global vs Local allocation: un processo può scegliere una frame vittima solo tra i suoi anche tra quelli degli altri? **Locale** visto nell'esempio di sopra. **Globale**, esista una lista di free frame, si fa in modo che questa lista non arrivi mai a 0, l'idea di massima è che superata una soglia si iniziano a scegliere delle vittime, fino ad arrivare di nuovo ad un altra solgia, che superata permette il reinserimento dei frame.

... NUMA ...

Le prestazioni migliori si ottengono quando un processo gira su un unico processore.

Quando il sistema inizia a saturare la memoria in modo aggressivo, il SO ha due possibilità: bloccare il sistema ed aspettare che un processo termini, oppure inziare a rimuovere frame, il **thrashing** è tenere il processore sempre attivo con dei processi.

Come si possono realizzare pochi page fault in modo statistico? Per minimizzare il numero di page fault si sfrutta il principio di località, basandosi su questo concetto si 'e sviluppato il modello di *working set* (l'insieme di pagine su cui si lavora), si definisce una finestra temporale che in cui si definisce Δ (finestra temporale e il range di pagine a cui si fanno accesso), si suppone che ci sono 10 mila accessi in memoria, WSS_i (dimensione del set di pagine a cui si è fatto accesso, del processo P_i) = numero totale di pagine in cui si è fatto riferimento nei Δ precedenti. Ci aspettiamo:

- se Δ troppo piccolo: non si riesce a contenere l'intero programma;
- Δ troppo grande si tengono troppe pagine in memoria;

Il **working set model**, dato un intervallo Δ si contano gli id delle pagine, se una pagina non è presente nell'intervallo viene rimosso dal *working set*, il difetto è che si deve generare un'istruzione per buttare via una pagina quando esce dal working set e soprattutto si deve tenere traccia di chi deve essere buttatto fuori. Questo non si può fare in modo esatto, si usa questo principio: si tiene nel *recidence set* un po più del working set, si suppongo che un timer interrompa ogni $\Delta/2$, ogni ogni pagina si tengono 2 bit di riferimento (ref bit 1, ref bit 2), uno dei due viene settato automaticamente dall'HW, ad ogni intervallo per ogni pagina a cui si fa accesso viene

settato il ref bit ad 1, il ref bit che viene selezionato viene scambiato ad ogni intervallo, quando si cambia intervallo si può decidere di rimuovere le pagine che hanno $(0, 0)$, questo approccio si può estendere con più ref bit.

Il problema del working set è che non guarda quante volte si fa accesso ad una pagina, si stabilisce una **frequenza di page-fault**, se il numero di page fault è alto allora si allarga la window, se il numero di page fault è basso allora la window si restringe. Funzionamento dell'algoritmo:

- si attiva l'algoritmo solo quando c'è un page-fault;
- quando si attiva si misura il tempo intercorso rispetto all'ultimo page fault τ, per vedere se si sta andando bene o male;
- se τ è minore di una costante c si prende un nuovo frame e lo si aggiunge al *resident set*;
- se τ è maggiore di c si rimuove dal *resident set* tutte le pagine con reference bit a zero e si settano i reference bit a zero della pagine rimanenti;

Normalmente il working set dovrebbe coincidere con i monti nel queale il SO ha delle spike di utilizzo dei frame, dovuti a vari motivi.

Il kernel è allocato in modo contiguo, il motivo è che la maggior parte delle strutture dati presenti al suo interno sono allocate in modo contiguo, ad esempio la **page table**.

Esempio di due tipi di allocazione:

- **Buddy System**: L'idea è che va bene l'allocazione contigua va bene ma l'allocazione può essere solo una potenza di 2, questo causa un sacco di frammentazione interna.
- **Slab Allocator**: L'idea è allocare per pagina ma si prendono in modo contiguo. In linux c'è una struttura di slab allocation.

Per diminuire il numero di page fault, si può allocare un po di pagine iniziali ed essere bravi ad indovinarle, questa tecnica viene detta **prepaging**.

Quanto deve essere grande una pagina? Per diminuire la frammentazione si vorrebbere pagine piccole, per diminuire la dimensione della page table si volgiono pagine grandi. Un altro parametro è dato dalle operazioni di IO, infatti una volta che si legge dal disco (molto lento), tanto vale prendere tanta roba. La grandezza prende in considerazione anche la *località* dei dati. Se si guarda l'efficienza della TLB avere poche entry è molto meglio (pagine grandi). Solitamente le pagine hanno una grandezza compresa tra 2^{12} e 2^{22} (la grandezza è sempre un multiplo di 2).

Il **TLB reach**, è la percentuale di spazio che la TLB vede sulla memoria (quantità di memoria vista dalla TLB)

$$TLBreach = TLBsize \cdot PageSize$$

La struttura dei programmi ha anche un effetto sui page fault.

Example 2.3

Si supponga di voler azzerare una matrice in un doppio loop annidato.

```
for (int i = 0; i < 128; i++)
  for (int j = 0; j < 128; j++)
    matrx[i][j] = 0;

```

```
for (int j = 0; j < 128; i++)
  for (int i = 0; i < 128; i++)
    matrx[i][j] = 0;

```

Supponendo che una riga stia all'interno di una singola pagina, il primo programma causa 128 page faults, mentre il secondo causa 128×128 page fault.

2.1 Esercizi

Gestione della memoria 3.

3 OS161

3.1 Configurazioni del Kernel

Ogni versione del kernel corrisponde a un file di configurazione tutto in maiuscolo, la prima versione è `DUMBVM` e tutti i file di configurazione si trovano in `$HOME/kern/conf`.

Ogni volta che si aggiunge un nuovo modulo al kernel, questo va abilitato, per abilitare nuovi moduli si va in `conf.kern` e si aggiunge

```
defoption hello
optfile hello main/hello.c
```

dove `hello.c` è il nome del nuovo modulo. `defoptions` *`name`* crea un file `opt-`*`name`*`.h` che è un file `.h` di configurazione, esso contiene una define per specificare se un opzione è abilitata o meno.

```
/* Automatically generated; do not edit */
#ifndef _OPT_HELLO_H_
#define _OPT_HELLO_H_
#define OPT_HELLO 1
#endif /* _OPT_HELLO_H_ */
```

Questo serve nei file del kernel per abilitare l'aggancio di altri moduli che vengono aggiunti.

Per compilare i nuovi moduli si può creare un nuovo file di configurazione nella cartella `conf`, solitamente è meglio partire da uno già esistente come `DUMBVM`, farne una copia e poi aggiungere i nuovi file opzionali. Una volta creato il nuovo file di configurazione, ad esempio `HELLO`, si esegue il comando `./config HELLO` per creare la configurazione compilata della versione del kernel, si dovrà poi andare in `kern/compile/HELLO` e poi eseguire

```
$ bmake depend
$ bmake
$ bmake install
```

3.2 Memoria

Nella versione base (`DUMNBVM`) di OS1616 l'allocazione è allocazione contigua sia per i processi che per il kernel, inoltre questo allocatore non rilascia la memoria. Il file `dumbvm.c` è l'allocatore del mips, `vm.c` è un allocatore opzionale ma è vuoto.

`ram_stealmem` nel file `ram.c`, perdende della pagine dalla ram e le alloca.

L'allocatore è diviso in allocatore User e allocatore Kernel.

Il kernel è basato su mips (32 bit) quindi ha 4GB di memeria logica, è diviso in:

- `kuseg`; (0x00000000, 0x80000000) [user]
- `kseg0`: (0x80000000, 0xa0000000) [kernel]

- `kseg1`: (0xa0000000, 0xc0000000) [kernel]
- `kseg2`: (0xc0000000, 0xffffffff) [kernel]

Dall'indirizzo si capisce se il programma è kernel o user, i processi user non possono vedere lo spazio del kernel mentre il kernel può vedere lo spazio user.

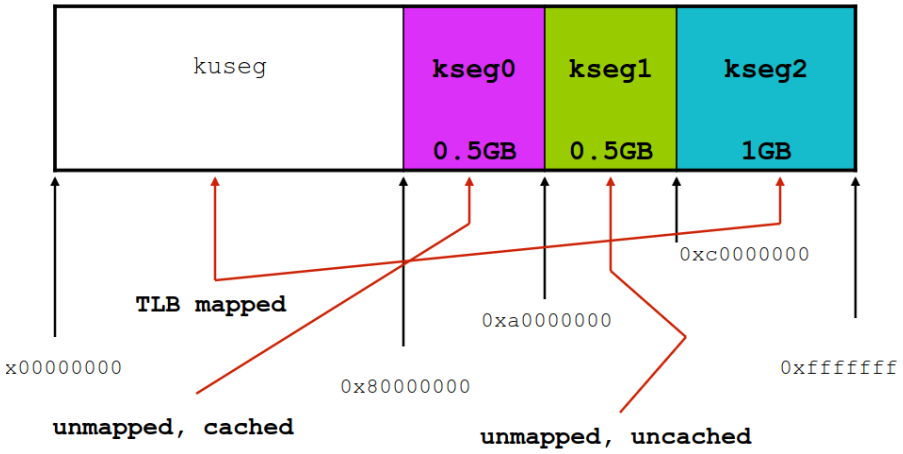


Figure 4: Mips Virtual Address Space

Il kernel utilizza `kseg0` e `kseg1` per TLB e cache. Se si fa accesso alla memoria `kuseg` e `kseg2` i loro indirizzi sono mappati sulla TLB, negli altri spazi non si intende usare la TLB e quindi sono *invisibili* alla TLB. Il `kseg1` è una memoria **uncached** e potrà essere usata per parlare con i dispositivi di IO, il motivo è che questi dispositivi non funzionano bene con le cache. Quando il kernel viene caricato la memoria la prima funzione chiamata è `ram_bootstrap`, partizionando la memoria in:

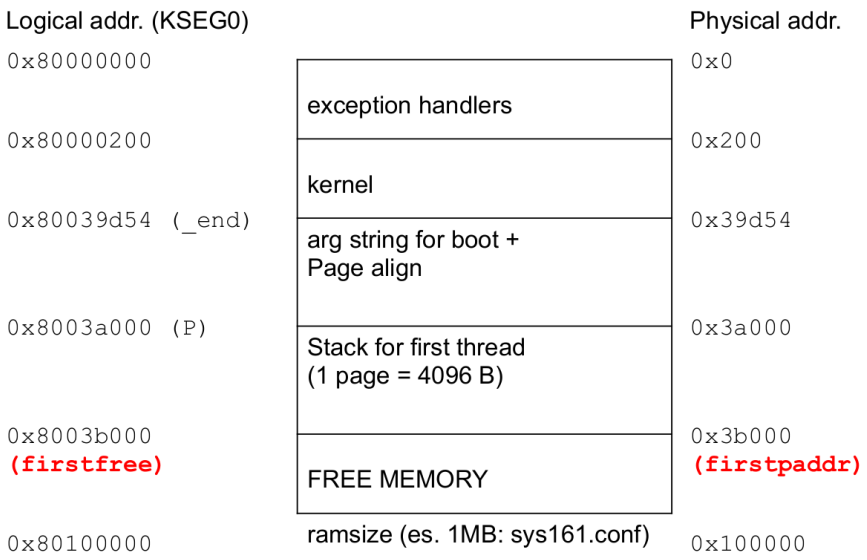


Figure 5: Kernel Loader Initial

Una volta che viene fatto il bootstrap, si hanno a dipsosizione il `firstfree` (primo indirizzo di ram libera, logico), `firstpfree` (primo indirizzo fisico) e `lastpaddr` (ultimo indirizzo fisico).

Quando si vuole allocare della memoria, si usa `ram_stealmem` a cui gli vanno passete il numero di pagine.

```
// kern/arch/mips/vm/ram.c
paddr_t ram_stealmem(unsigned long npages) {
  paddr_t paddr;
  size_t size = npages * PAGE_SIZE;
  if (firstpaddr + size > lastpaddr) {
    return 0;
  }
  paddr = firstpaddr;
  firstpaddr += size;
  return paddr;
}
```

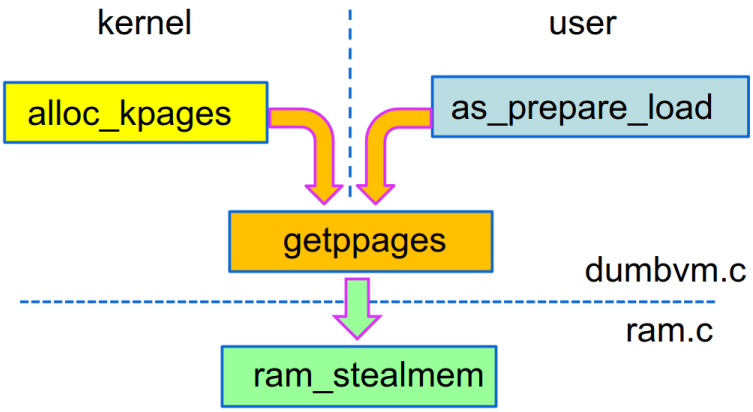


Figure 6: Interfaccia Dumbvm

Le funzioni di free devono essere implementate.

Se si volesse gestire la paginazione, si dovrebbe scindere i l'allocatore per kernel e user, infatti l'user dovrebbe vedere delle pagine mentre il kernel della memoria contigua. Per ora la soluzione proposta è:

- allocazione contigua (per pagine) comune;
- l'allocatore in bumbvm: ...
- bitmap implementata come char array, dove ogni posizione rappresenta se la pagina è libera o presa;
-

3.3 Overview

In os161 si parla di thread di kernel e poi di processi utenti (in futuro un thread di kernel si sgancia dal kernel e diventa processo user), ogni thread ha us suo **context** di esecuzione, in cui sono contenuti dei dati e il suo stack. Gli user thread sono generati da un altro processo thread, ad esempio i *POSIX pthread* permettono di creare dei nuovi thread.

La differenza tra un processo ed un thread è: un processo ha uno spazio di indirizzamento in cui sono contenute le istruzioni ed i dati, poi esistono due aree che sono lo stack e lo heap, il processo arriva da un file eseguibile (tutto o in parte). Un processo può essere *single-threaded* o *multi-threaded*, tutti i dati globali sono condivisi e visibili dai thread, ogni thread ha privato il suo **context** (registri, stack, program counter).

In os161 il contesto di kernel thread, è: stack, registri, dati. I due modi di implementare la creazione dei nuovi thread è diviso in:

- i thread sono gestiti dalla libreria;
- i thread sono gestiti dal kernel;

Questo è dovuto al fatto che: il kernl vede solo dei processi (al posto dei thread) e da come è fatto lo scheduler dei thread. In os161 un thread è fatto da:

```
/* see kern/include/thread.h */
struct thread {
  char *t_name; /* Name of this thread */
  const char *t_wchan_name; /* Name of wait channel, if sleeping */
  threadstate_t t_state; /* State this thread is in */
  /* Thread subsystem internal fields. */
  struct thread_machdep t_machdep;
  struct threadlistnode t_listnode;
  void *t_stack; /* Kernel-level stack */
  struct switchframe *t_context; /* Saved register context (on stack) */
  struct cpu *t_cpu; /* CPU thread runs on */
  struct proc *t_proc; /* Process thread belongs to */
  ...
};
```

Esistono 3 funzioni per gestire dei thread:

- `thread_fork`: crea un nuovo thread;
- `thread_exit`: termina il thread;
- `thread_yield`: sospende l'esecuzione del thread;

Quando viene creato un nuovo thread viene fatto un context switch. Come viene fatto il context switch? Si utilizza `switchframe_switch`, il codice scritto è in assembler MIPS perché vanno salvati anche i registri e questo non è possibile farlo in C.

Quando viene creato un nuovo user thread, viene generato un nuovo stack per il processo user, lo stack kernel non viene rimosso ma rimane in memoria, quando il processo farà una chiamata ad una systema call e passarà in *modalità kernel* e userà lo stack kernel che ha lasciato indietro.

I processi sono rappresentati dalla `struct proc`, questa struct ha tutte le informazioni relative al processo, un campo in esso contenuto (`struct addrspace`) contiene al suo interno che definisce lo spazio d'indirizzamento del processo, se ci fosse paginazione dovrebbe contenere la tabella della pagine (non è il caso in `DUMBVM`). Nella versione base di OS161 non ha un puntatore a ogni thread che ha (ha solo il numero di thread), mentre i thread hanno un puntatore al processo padre. Un processo user non esegue istruzioni solo in modalità user ma richiede del lavoro in modalità kernel, usando le system call.

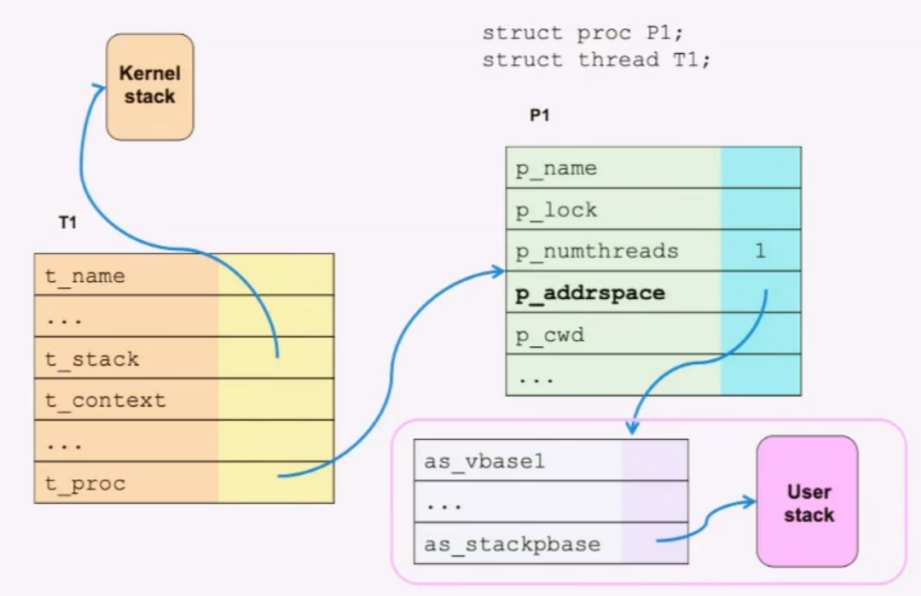


Figure 7: Process Stack

Per creare un programma user da os161 si fa

```
p <elf_file> {<args>}
```

quello che avviene è:

1. si chiama

   ```
   proc_create_runprogram
   ```

   che crea lo spazio d'indirizzamento del processo

2. legge il file ELF per caricarlo in memoria

3. si passa da kernel thread user thread

... La `runprogram` al suo interno fa partire la `enter_new_process`, che si appoggia ad una strutturea siimila allo `switchframe` che si chiama `trapframe`, questa struttura server per fare un *cambio di contesto* (context switch), la `enter` crea un nuovo processo come se questo processe stesse ritornando da una trap, infatti il `trapframe` è proprio la struttura in cui vengono salvati i dati alla chiamata di una trap.

3.4 Systemcall

Le **systemcall** sono un modo per i processi di utenti di far eseguire al kernel delle operazioni.

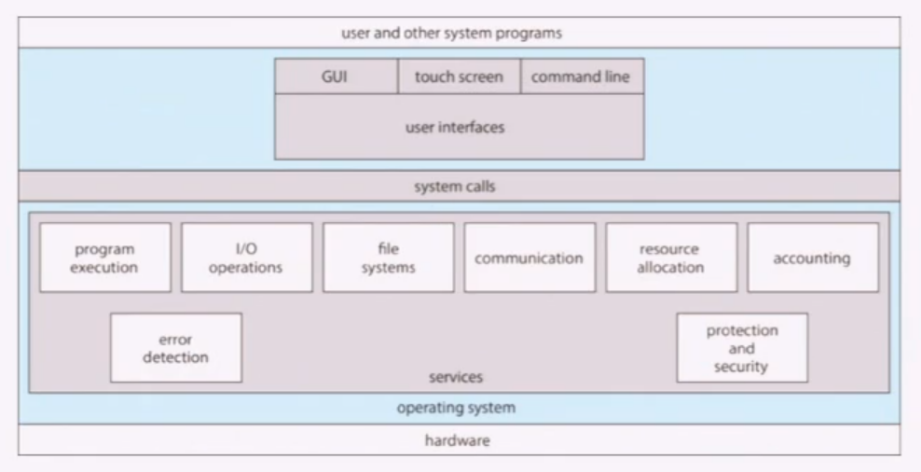


Figure 8: Os Overview

In os161 quando viene invocato una system call i dati vengono passati nel `trapframe` (come in linux), uno user program chiamerà la system call con il suo numero associato e inserisce i dati necessari.

Di system call ce ne sono di vario genere, tra le più importanti ci sono: sincronizzazione, file system, tempo, ... In MIPS esiste un solo handler per gestire gli interrupt, le eccezioni e le system calls, questo è fatto perchè la differenza sarà gestita in hardware, il syscall handler prende un numero ad attraverso uno switch decide che tipo di trap è stata scatenata (hardware o software). Nel processo utente il `trapframe` ed il `contextframe` si trovano entrami nello stack del processo.

Esistono due tipi di programmi che interagiscono che sono i processi utenti ed il kernel, come già visto i processi utenti usano indirizzi virtuali, ma il kernel ha bisogno di utilizzare indirizzi virtuali? La possibilità sono:

- Il kernel è in memoria fisica: la CPU è a consoscenza della modalità di operazione, quindi quando è il kernel ad operare viene spenta la traduzione da indirizzo logico a fisico.
- Il kernel si trova in un spazio di indirizzamento virtuale separato: in questi casi non ci sarà paginazione del suo spazio di indirizzamento
-

In linux tutti gli indirizzi sono separati, infatti si potrebbero avere due indirizzi identici che in due contesti diversi puntano a memoria diversa, in linux questo non piace, perché il kernel può metter mano all'interno dei processi, per questo motivo esistono indirizzi utente ed indirizzi kernel.

Nel MIPS la traduzione logico/fisica avviene nella TLB, che contiene: numero di pagina virtuale, numero di frame fisico, falgs, ID indirizzo (non usato in OS161).

In `dumbvm` l'allocazione delle pagine è contigua, vengono definiti dei campi per l'`addrspace` di un processo e sono:

```
\struct addrspace {
#if OPT_DUMBVM
  vaddr_t as_vbase1; /* base virtual address of code segment */
  paddr_t as_pbase1; /* base physical address of code segment */
  size_t as_npages1; /* size (in pages) of code segment */
  vaddr_t as_vbase2; /* base virtual address of data segment */
  paddr_t as_pbase2; /* base physical address of data segment */
  size_t as_npages2; /* size (in pages) of data segment */
  paddr_t as_stackpbase; /* base physical address of stack */
#else
/* Put stuff here for your VM system */
#endif
};
```

In questo modo basta conoscere il `base` ed il numero di pagine che il processo possiede. In questo modo si evita di avere una page table per tenere traccia delle pagine.

Se non è presente la entry nella TLB viene generato un `page fault`.

3.5 Sincronizzazione

I problemi delle programmazione concorrente sono: deadlock, starvation, race conditions. Si chiama **sezione critica** un pezzo di codice che accede a della memoria condivisa che può essere eseguito da un solo processo. Le sezioni critiche possono essere garantite da: mutua esclusione, progresso, bounded wait. Si possono usare due approcci che garantiscono la sincronizzazione, questo è discriminato dal fatto che il sistema sia **preemptive** o **non-preemptive**, ovvero se le primitive di sincronizzazione possono essere bloccate o se non possono essere bloccate (in questo caso in kernel mode non esistono problemi di race conditions).

Example 3.1 – Algoritmo di Peterson

Supponiamo che `load` e `sotre` sia atomiche, e che siano due variabili condivise tra due processi `turn` (dice a chi tocca) e `flag[2]` (indica se il processo 1 o 2 è pronto ad entrare nella sezione critica),

```
/* Processo Pi, l'altro e' Pj */
while (true)
{
  flag[i] = true;
  turn = j;
  while (flag[j] && turn == j);

  /* Critical section */

  flag[i] = false;
}
```

si può dimostrare che se lettura e scrittura sono atomiche questo algoritmo funziona e la mutua esclusione è garantita. Il problema è che nelle architetture moderne questo non è vero, perché non è detto che le letture e scritture funzionino in modo atomico, soprattutto nei processori multicore e con pipeline.

A causa di questi problemi si sono inventati dei metodi appositi per effettuare la sincronizzazione, i **lock**. I modi di acquisire un lock sono:

- spin o busy-waiting
- lasciare il processore

Per realizzare queste primitive l'hardware mette a disposizione **test and set** e **compare and swap** in modo atomico. La **test and set** fa

```
bool test_and set(bool *lock)
{
  bool prev = *lock;
  *lock = true;
  return prev;
}
```

mentre la **compare and swap**

```
itn compare_and_swap(int *value, int expected, int new_value)
{
  int temp = *value;

  if (*value == expected)
    *value = new_value;

  return temp;
}
```

entrambe sarebbero le realizzazioni in C, ma queste operazioni sono realizzate in hardware. Insieme a queste due istruzioni esistono anche le **variabili atomiche**, che permettono operazioni atomiche su di loro.

In OS161 esistono gli **spinlock** ovvero dei lock con busy-waiting, vengono usati quando la critical section non è molto complicata, vengono implementati con `test-and-test-and-set`, che è un leggera variante del `test-and-set` dove lo spinlock viene prima testata in modo non atomico e poi in modo atomico

```
void spinlock_acquire(struct spinlock *splk)
{
  // ...
  while (1)
  {
    if (spinlock_data_get(&splk->splk_lock) != 0)
      continue;
    if (spinlock_data_testandset(&splk->splk_lock) != 0)
```

```
      continue;

    break;
  }
  // ...
}
```

questo viene fatto per motivi di performance, infatti fare il `testandset` inficia sulla performance dell'intero sistema.

Esistono poi anche i **semafori**, che sono più potenti dei lock, ne esistono di counting e di binary. Le primitive sono dette `wait` (`P`) e `signal` (`V`), dove `wait` decrementa il semaforo se maggiore di 0 e `signal` incrementa il semaforo. I semafori solitamente non vengono implementati con del busy-waiting, per questo motivo i semfori al loro interno contengono una lista di attesa, quando la `wait` fallisce si inserisce un nuovo processo nella lista. Quando sarà fatta la `signal` si dovrà decidere quale processo svegliare. Un modo di rendere un operazione atomica è quello di disabilitare l'handler di un interrupt, questo va bene in un single core, ma in un multicore questo non si può fare.

Un semaforo contiene al suo interno una lista di processi, ovvero i processi in attesa. In os161 esistono `thread_sleep` e `thread_wakeup`, la differenza tra `sleep` e `yield` è che dopo `sleep` il thread è stato di blocco mentre con `yield` il processo è ready.

I `lock` sono degli oggetti simili a dei semafori binari, i `lock` però sono adatti solo alla mutua esclusione, il vincolo è che che fa il lock-acquire deve fare anche il lock-realease.

Esistono dei casi (**wait on codition**) in cui data una condizione basata su più variabili, si manda una `signal` per fare un'operazione su queste variabili

```
/* shared state vars with some initial value */
int x,y,z;
/* mutual exclusion for shared vars */
struct lock *mylock = lock_create("Mutex");
/* semaphore to wait if necessary */
struct semaphore *no_go = sem_create("MySem", 0);
compute_a_thing
{
  lock_aquire(mylock); /* lock out others */
  /* compute new x, y, z */
  x = f1(x); y = f2(y); z = f3(z);
  if (x != 0 || (y <= 0 && z <= 0)) V(no_go);
    lock_release(mylock); /* enable others */
}

//...
use_a_thing
{
  lock_aquire(mylock); /* lock out others */
  if(x == 0 && (y > 0 || z > 0))
```

```
    P(no_go);
  /* Now either x is non-zero or y and z are
  non-positive. In this state, it is safe to run
  "work" on x,y,z, which may also change them */
  work(x,y,z);
  lock_release(mylock); /* enable others */
}
```

In questo caso se `use_a_thing` arriva per primo al lock ci sarà un deadlock. Una soluzione è

```
use_a_thing
{
  lock_aquire(mylock); /* lock out others */
  while (x == 0 && (y > 0 || z > 0))
  {
    lock_release(mylock); /* no deadlock */
    P(no_go);
    lock_aquire(mylock); /* lock for next test */
  }
  /* Now either x is non-zero or y and z are
  non-positive. In this state, it is safe to run
  "work" on x,y,z, which may also change them */
  work(x,y,z);
  lock_release(mylock); /* enable others */
}
```

Per questo motivo esistono le **condition variable**. I **Monitor** sono dei costrutti di alti livello che contengono delle code di attesa, ..., che mettono tutto insieme.

La differenza tra signal e condition variable è che quando la signal viene segnalata, se un semaforo arriva dopo può comunque prenderla, la condition varaible quando segnala se nessuno prende il segnale allora viene perso, se un semaforo arriva dopo non vede la signal.

Esistono anche i **wait channel**, in OS161 quando un

3.6 Esame

4 Rust

Rust è un linguaggio di programmazione moderno, che non permette di avere undefined behaveiur e soprottuto è un linguaggio **statically strongly typed**, inlotre è capace di inferire i tipi di dati in modo molto potente. Rust fornisce delle *astrazioni a costo nullo*, ed è anche *interoperabile* con C. La memoria in rust è gestita dal programmatore solo in parte, il motivo è che il complilatore forza il programmatore a scrivere del codice senza errori, ed il compilatore si occupa di effettuare la gestione della memoria. Esempio di **dangling pointer** in C.

```c
typedef struct Dummy { int a; int b; } Dummy;

void foo(void) {
  Dummy *ptr = (Dummy *)malloc(sizeof(Dummy));
  Dumyy *alias = ptr;
  ptr->a = 2048;
  free(ptr);
  alias->b = 25;  // WARNING!!
  free(alias);
}
```

Grazie alle feature di rust si possono prevenire:

- **dangling pointers**
- **doppi rilasci**
- **corse critiche**
- **buffer overflow**
- **iteratori invalidi**
- **overflow aritmetici**

Inoltre rust propone delle convenzioni per aiutare il programmatore ad avere del codice ideomatico.

Rust parte dall'assunzoine che un valore può essere **posseduto** da una singola variabile, se la variabile esce al di fuori del suo scope il valore viene deallocato. Quando si fanno delle assegnazioni l'**ownership** del valore viene trasferito alla variabile alla quale è stata fatto l'assegnazione, inoltre è anche possibile **dare in presetito** (**borrow**) un valore; l'insieme di queste idee sono alla base della sicurezza che fornisce Rust. I puntatori sono tutti controllati in fase di compilazione, oltre all'**ownership** ogni puntatore ha un **tempo di vita** (**lifetime**), tutti gli accessi al di fuori della lifetime sono negati dal compilatore, in alcuni casi non può essere fatto l'infering della lifetime e questo andrà specificato esplicitamente usando la notazione `<'...>`. La **sicurezza dei thread** è incorporata nel sistema dei tipi e anch'essi hanno una lifetime. Rust inoltre non ha stati nascosti, come in java con le eccezioni.

I due strumenti per gestire l'ambiente di sviluppo in rust sono:

- `$ rustup`: installer and updater di rust;
- `$ rustc`: compilatore;
- `$ cargo`: gestione dei progetti e delle dipendenze;

I comandi di base di `cargo` sono:

- `$ cargo new project-name`
- `$ cargo new --lib library-name`
- `$ cargo build`
- `$ cargo run`

Terminologia in rust:

- **crate**: unità di compilazione, crea un eseguibile o una libreria;
- **create root**: radice del progetto e solitamente contiene il `main.rs`, oppure `lib.rs` in caso di librerie;
- **module**: il progetto è organizzato in *moduli*, se si crea un modulo come cartella il file .rs andrà nominato `mod.rs`, all'interno possono essere presenti moduli interni;
- **package**:

Come in kotlin i blocchi condizionali hanno dei valori di ritorno. In rust non esiste l'eriditarità, però esistono i **trait**, che sono simili alle interfacce, offrendo la possibilità di poter implementare dei metodi, ad esempio alcuni di questi sono `Display` e `Debug`, ogni tipo può implementare più tratti.

I tipi elementari di rust sono:

- numerici: `i8, i16, i32, i64, i128, isize` (valore nativo del processore).
- numerici senza segno: `u8, ...`
- naturali: `f32, f64`
- logici: `bool`
- caratteri: `char` (32 bit, rappresentazione **Unicode**)
- unit: `()` rappresenta una tupla di 0 elementi, corrisponde a `void` in C/C++

Per rappresentare le stringhe invece si use la codifica `utf-8`, che codifica i caratteri unicode in blocchi di 8 bit, la codifica utilizza la maggior parte dei primi 8 bit come caratteri standard, se si vogliono usare simboli più strani si combina un byte con il successivo, se il carattere è ancora più raro si utilizzano più byte, fino ad arrivare a 4.

Una **tupla** è una struttura che permette di contenere più tipi di valori, per accedere alla posizione corrispondente si utilizza la notazoine `<variabile>.<numero>`.

Rust ha un meccanismo al suo interno per rappresentare i puntatori, infatti quando viene creato un valore in una funzione si può decidere di passare l'ownership alla funzione chiamante, che poi ha la responsabilità di liberla oppure di passare ancora l'ownership, in rust esistono tre tipi di puntatori:

- **reference**;
- **box**;
- **native pointer**;

L'utilizzo dei puntatori nativi (che sono gli stessi del C), è possibile solo in un blocco `unsafe { ... }`.

Le **reference** sono puntatori senza possesso e possono essere assegnati con `&`:

```
let r1 = &v;
let r1 = &mut v;
```

In questo caso `r1` borrows il valore `v`, per acceder al valore si utilizza la notazione per **derefenziare**: `*r1`. Quando si crea un puntatore che `&` si può un creare un puntatore in sola lettura, si vuole anche scrivere all'indirizzo del puntatore si deve usere la keyword mut: `&mut v`, l'accesso in scrittura è *esclusivo*, ovvero può essere assegnato ad una sola variabile e quando viene asseganto in modo mutabile nussun altro può utilizzarlo, infatti è *mutuamente esclusivo*.

```
fn main() {
  let mut i = 32;

  let r = &i;     // r is of type "int ref";
  println!("{}", *r);

  i = i+1;        // ERROR: i was borrwed
  println!("{}", *r);
}
```

```
fn main() {
  let mut i = 32;

  let r = &mut i;
  println!("{}", i);  // ERROR: i was mutably borrowed

```

```
  *r = *r+1;
  println!("{}", *r);
}
```

In questo caso la variabile `i` non è accessibile in alcun modo per tutta l'esistenza di `r`.

In altre situazioni c'è l'esigenza di allocare un dato, per prolungare la sua vita all'infuori della funzione in cui viene creato, l'equivalente della `malloc` in rust è `Box<T>` che alloca un parte di memoria presa dall'*heap*, per creare un valore che che si vuole venga conservato, si passa a `Box` il valore che voglio si venga conservato:

```
let b = Box::new(v);
```

Rust è in grado di fare l'infer del tipo di `v` e riesce ad allocare lo spazio necessario, quando `b` non sarà più visibile il `Box` verrà liberato dalla memoria.

```
fn useBox() {
  let i = 4;                      // i si trova nello stack
  let mut b = Box::new((5, 2));   // *b si trova nell'heap, b si trova nello stack

  (*b).1 = 7;

  println!("{:?}", *b);   // (5, 7)
  println!("{:?}", b);    // (5, 7)
}
```

Quando `println!` si trova un puntatore lo dereferenzia automaticamente. Quando si arriva alla fine della funzione `a` e `b` vengono rimossi dallo stack, dato che il valore a cui puntava `b` non possiede un owner in vista quel valore viene rimosso dell'heap.

```
fn makeBox(a: i32) -> Box<(i32,i32)> {
  let r = Box::new((a, 1));
  return r;
}

fn main() {
  let b = makeBox(5);
  let c = b.0 + b.1;
}
```

In questo caso il valore l'ownership del valore puntato da `r` passa la sua ownership a `b` che si trova nel `main`, quando il `main` termina e la vita di `b` finise il valore puntato viene anch'esso liberato.

I **Puntatori nativi** sono `*const T` e `*mut T`, il `*` indica un puntatore nativo e possono essere utilizzati solo nei blocchi `unsafe ...` .

Rust supporta nativamente anche gli **array**, in rust gli array sono composti da dati omogenei allocati contiguamente nello stack. Si possono inizializzare nel seguente modo:

```
let a: [i32; 5] = [1, 2, 3, 4, 5];
let b = [0; 5];     // array lungo 5, inizializzato a 0
```

In rust gli array hanno la conoscenza di quanto sono lunghi. Se si cerca di accedere ad indice fouri dal range di utilizzo rust fa `panic`, se `panic` viene lanciato in nel thread principale il programma termina, se viene lanciato in un altro thread viene terminato il thread. Per accedere all'array si utilizzano gli **Slice** (invece di utilizzare i puntatori), che vengono creati come riferimento ad una porzione di un array:

```
let a = [1, 2, 3, 4];
let s1: &[i32]  = &a; // s1 contiene 1, 2, 3, 4
let s2 = &a[0..2]     // s2 contiene 1, 2
let s3 = &a[2..]      // s3 contitne 3, 4
```

Per questa sua natura viene detto **fat pointer**, perchè oltre a contenere il puntatore contiene anche la lunghezza. Come nel caso degli array se si va al di fuori del range viene generato un `panic`.

Il problema degli array è che quando vengono creati con una dimensione, questa rimane fino alla del loro lifetime, un array variabile può essere creato con `Vec<T>` che rappresenta una sequenza di oggetti di tipo `T`, al contrario degli array gli oggetti presenti in `Vec` sono allocati nell'heap.

```
fn useVec() {
  let mut v: Vec<i32> = Vec::new();

  v.push(2);
  v.push(4);

  let s = &mut v;
  s[1] = 8;
}
```

Quando viene inizializzato `Vec` nello stack vengono inseriti 3 parametri: pointer, size, capacity. Appena viene fatto push viene allocata della memoria nell'heap, e i parametri size e capacity vengono aggiornati. Se vuole creare uno slice `s`, nello stack viene messo un fat pointer, ovvere il puntatore all'heap ed il size corrente. Quando `Vec<T>` viene inizialmente inizializzato, nel campo puntatore viene inserito l'allinemento che il tipo `T` dovrà avere in memoria.

Lo stesso discorso fatto per i `Vec` vale per gli oggetti di tipo `String`. Quando si crea una stringa tra le virgolette viene creato un `&str` e non una `String`, questo dato andrà inserito nella sezione del programma dei *dati inizializzati*, al contrario delle `String` queste non modificabili, se si vuole utilizzare una stringa si utilizza `"...".as_str()`, che ritorna uan stringa allocata nell'heap.

```
fn main() {
  let hello: &str = "hello";

  let mut s = String::new();    // il valore (futuro) puntato da s viene allocato
    nell'heap

  s.push_str(hello);            // l'heap viene inizializzato una capacita' ed un
    size uguale al size della stringa puntata da "hello"
```

```
  s.push_str(" world!");
}
```

In rust le keyword `let` e `let mut` sono delle istruzioni, mentre un blocco racchiuso tra `...` è un espressione e quindi ha un valore di ritorno, anche il costrutto `if` è un espressione, **il valore viene ritornato a patto che l'ultima istruzione del blocco non abbia un punto e virgola**. Si può asseganre un etichetta con `'<nome-etichetta>:` ad un espressione per poter specificare il ritorno:

```
fn main() {
  'outer: loop {
    'inner: loop {
      //...
      continue 'outer;
    }
  }
}
```

Per leggere gli argomenti passati da linea di comando si utilizza:

```
fn main() {
  let args: Vec<String> = args().skip(1).collect();

}
```

Si può utilizzare la libreria `clap` per fare il parsing degli argomenti. per farlo si può andare nel file `.toml` ed inserire la dipendenza.

Per fare IO da console si utilizza `stdin`:

```
fn main() {
  let mut s = String::new();

  if io::stdin().read_line(&mut s).is_ok() {
    println!(..., s.trim());
  } else {
    println!("Failed!");
  }

  // Oppure...
  io::stdin().read_line(&mut s).unwrap();
  println!(..., s.trim());
}
```

4.1 Ownership

Il concetto dell'ownership in rust permette al **borrow checker** di controllare l'utilizzo corretto dei puntatori, in rust *un valore può essere posseduto da una e una sola variabile*, il fallimento di questa condizione comporta ad un fallimento della compilazione. Possere un valore vuol dire *occuparsi del suo rilascio*. Se una variabile mutabile viene

riasseganta, ed il valore precedente implementava il tratto `Drop` esso viene prima liberato e poi avviene l'assegnazione. Quando si fa un assegnazione da una variabile all'altra: `a = b`, il valore posseduto da `b` viene spostato in `a`, questa operazione viene detta **movimento**, ed il valore non è più accessibile da `b`, dal punto di vista fisico il movimento è una *copia*.

```
fn main() {
  let mut s1 = "hello".to_string();
  println!("s1: {}", s1);

  let s2 = s1;
  println!("s2: {}", s2);

  // s1 non e' piu' accessibile
}
```

Il movimento è il comportamento standard in rust, se però un tipo implementa il tratto `Copy` quando avviene un'essegnazione entrabe le variabili possono continuare ad accedere al valore, però i tratti `Copy` e `Drop` sono mutuamente esclusivi. In generale i tipi copiabili sono i: numeri, booleani, le reference `&`, ..., mentre i `&mut` non sono copiabili. Il tratto `Copy` può essere implementato solo se il tipo implementa il tratto `Clone`, che permette di fare una clonazione dell'oggetto **in profodità**, inoltre la clonazione è implementata dal compilatore, mentre copia e movimento sono gestite dal compilatore e veine implementata con `memcpy()`.

```
fn main() {
  let mut s1 = "hi".to_string();
  let s2 = s1.clone();  // viene creato nello stack una nuova variabile e nell'
    heap un nuovo valore
  s1.push('!');

  println!("s1: {}", s1); // hi!
  println!("s2: {}", s2); // hi
}
```

In rust il comportamento di base è il movimento, mentre in C l'unico comportamento è quello della copia, mentre in C++ si può copiare e si può muovere ma questo va scritto esplicitamente.

```
struct Test(i32);

impl Drop for Test {
  fn drop(&mut self) {
    println!("Dropping Test({}) @ {:p}", self.0, self);
  }
}

fn alfa(t: Test) {
  println!("Invoking alfa() with Test({}) @ {:p}", t.0, &t);
}

```

```
fn main() {
  let t = Test(12);

  alfa(t);

  println!("Ending...");
}
```

Quello che succede in questo caso è che il metodo di `drop()` viene chiamato scritto prima di `Ending...`, il motivo è che `t` viene mossa all'interno della funzione `alfa()`. Quando non si vuole passare la proprietà del valore ad una funzione si può decidere di **prestarglielo**, se si vuole fare ciò l'argomento della funzione deve essere una reference.

```
fn alfa(t: &Test) {
  println!("Invoking alfa() with Test({}) @ {:p}", t.0, t);
}

fn main() {
  let t = Test(12);

  alfa(&t);

  println!("Ending...");
}
```

Se si vuole modificare il valore all'interno della funzione si cambiare la firma in: `fn alfa(t: &mut Test)`.

In rust i riferimenti permettono il prestito dei valori di una variabile in solo scrittura, mentri i riferimenti mutabili permettano lettura e scrittura, ma se viene prestato la variabile originale non può leggere. A differenza del tipo di dato che si maneggia, si possono avere 3 tipi di puntatori: **reference**, **fat pointer**, **double pointer**. Quando il compilare ha tutte le informazioni sul tipo di dato ho bisogno di una reference semplice, se tuttavia non è in grado di inferire tutti i dati ricorre ad un *fat pointer* (come nelle slice), che contiene il puntatore all'oggetto e la dimensione dell'oggetto. Nel caso di un tipo che contiene delle implementazioni, viene usato un double pointer, dove il primo punta all'oggetto, ed il secondo punta alla **vtable**, questa è una tabella che contiene i puntatori alle funzioni che vengono implementati per un tratto, ed esiste una vtable per ogni tratto implementato.

```
let f: File = File::create("text.txt");
let r3: &dyn Write = &f;   // double pointer
```

Il borrow checker garantisce che tutti gli accessi ad un riferimento avvengano solo all'interno di una **lifetime**, in alcuni casi il complilatore non è in grado di fare l'infering della lifetime, per assegnare una lifetime ad un riferimento si utilizza la sintassi:

```
&'a Type
```

L'unico nome riservato è `&'static Type`, questo vuol dire che la lifetime di quella reference sopravvive per l'intera durata del programma.

4.2 Clap

`clap` è una libreria di rust che permette di fare il parsing degli argomenti da linea di comando, una implementazione possibile è la seguente:

```rust
use clap::Parser;

/// Simple program to greet a person
#[derive(Parser, Debug)]
#[command(version, long_about = None)]
struct Args {
  /// Name of the person to greet
  #[arg(short, long)]
  name: String,
  /// Number of times to greet
  #[arg(short, long, default_value_t = 1)]
  count: u8,
}

fn main() {
  let args = Args::parse();
  for _ in 0..args.count {
    println!("Hello {}!", args.name)
  }
}
```

4.3 Slice

Uno slice è una vista su una sequenza contigua di elementi, gli slice in queanto riferimento non possiedono il valore, oltre ad essere dei fat pointer.

5 Tipi Composti

In rust si possono creare delle `struct` che è un tipo *prodotto*, ovvero che la sua grandezza è il prodotto cartesiano dei domini dei parametri contenuti al suo interno. Per convenzione si usa il *CamelCase* per i nomi delle `struct`. Le `struct` possono essere di due tipi: `named struct {}` dove ogni campo ha un nome, `tuple struct ()` dove i campi non hanno nome e vegono invocati attraverso un numero come posizione. In rust quando viene dichiarata una `struct` i campi possono essere riordinati per permettere una maggiore efficienza nell'utilizzo della memoria, a differenza di C/C++ che dispongono i parametri in memoria secondo l'ordine di scrittura, per questo motivo rust permette di utilizzare l'attributo `#[repr(...)]` che modifica la rappresentazione in memoria dei parametri.

```
#[repr(C)]
struct Test {
  a: i8,
  b: i32,
  c: i8,
};
```

In questo caso la struttura verrà disposta in memoria come in C.

5.1 Visibilità

In rust si possono definire dei moduli con la keyword `mod`, questo si può fare sia all'interno del file `main` che in altri file.

```
mod example {
  struct S2 {
    a: i32,
    b: bool,
  }
}

fn main() {
  let s2 = S2 { a: 0, b: true };
  // ERRORE: tutto cio' che sta in un modulo e' privato
}
```

Se si vuole utilizzare un modulo esso deve essere importato, se si vuole utilizzare la `struct S2` bisogna renderla pubblica, infatti qualsiasi cosa all'interno di un modulo una visibilità privata di default (ovvero che è visibile solo all'interno del modulo).

```
mod example {
  pub struct S2 {
    pub a: i32,
    pub b: bool,
  }
}

```

```
use example::S2;

fn main() {
  let s2 = S2 { a: 0, b: true };
  // Adesso compila
}
```

Si possono implementare anche delle funzioni per `S2`:

```
mod example {
  pub struct S2 {
    a: i32,
    b: bool,
  }

  impl S2 {
    pub fn new(a: i32, b: bool) -> Self {
      Self { a, b }
    }
  }
}

use example::S2;

fn main() {
  let s2 = S2::new(0, true);
}
```

Nei blocchi `impl` se si vuole fare riferimento all'istanza della `struct` si utilizza la keyword `self`. Le funzioni che hanno come primo parametro `self` diventano dei **metodi** e vengono dette **funzioni associate**.

In rust non è possible fare l'*overloading* dei metodi, da convenzione si usa `fn with_attributes(...)` (e.g. `String::with_capacity(10)`).

Il paradigma **RAII** (Resource Acquisition Is Initialization) voul dire: le risorse vengono incapsulate in un oggetto che possiede solo un *costruttore* ed un *distruttore*, in rust un concetto in cui viene implementato RAII è la struct `Box`. Ad esempio su C++ se può creare una classe che viene allocato nello stack che incapsula un oggetto e lo libera quando viene rimossa dallo stack.

```
class Car {};

void a_function() {
  Car *ptr = new Car;
  function_that_can_throw();
  if (!check()) return;
  delete ptr;
}
```

In questo caso non esiste garanzia che `ptr` venga liberato dalla memoria, grazie al paradigma RAII si può creare una classe wrapper che libererà l'oggetto in modo

automatico.

```
class CarManager {
  Car *car;
public:
  CarManager() {
    this->car = new Car;
  }
  ~CarManager() {
    delete this->car;
  }
}

class Car {};

void a_function() {
  CarManager car;
  function_that_can_throw();
  if (!check()) return;
}
```

C++ ha proprio un modo per gestire queste casistiche già implementato nella libreria standard, si chiama `unique_pointer`

```
void a_function() {
  std::unique_ptr<Car> car = std::make_unique<Car>();
  function_that_can_throw();
  if (!check()) return;
}
```

In Rust tutto questa gestione delle memoria è fatta in automatico grazie al Borrowing System e soprattutto **senza overhead**.

5.2 Enum

All'interno di valori `enum` si possono associare dei valori oppure anche dei campi.

```
enum HttpResponse {
  Ok = 200,
  NotFound = 404,
  InternalError = 500,
}

enum HttpResponse {
  Ok,
  NotFound(String),
  InternalError {
    desc: String,
    data: usize,
  }
}
```

In rust i tipi `enum` sono dei tipi somma, ovvero che il loro dominio è la somma di tutti i suoi campi, gli `enum` sono i tipi più simili alle `union` di C, in rust però i tipi `enum` contengono anche un `tag` per poterli distinguere. Grazie agli `enum` è possibile fare del parrtern matching con `match` e con `if let`:

```
enum Shape {
  Square { s: f64 },
  Circle { r: f64 },
  Rectangle { w: f64, h: f64 },
}

fn compute_area(shape: Shape) -> f64 {
  match shape {
    // Destructory assignment
    Square { s } => s*s,
    Circle { r } => r*r*3.14,
    Rectangle { w, h } => w*h,
  }
}

fn process(shape: Shape) {
  // Destructory assignment
  if let Shape::Square { s } = shape {
    println!("Square side: {}", s);
  }
}
```

La **destrutturazione** può essere utilizzata anche nelle operazioni semplici.

```
let Point { x, y } = point;
println!("{}{}", x, y);
```

Quando si fa destructuring è possibile prendere i valori interni agli oggetti come riferimento, anteponendo la keyword `ref`, il motivo è che usando questi construtti, se la `struct` non implementa `Copy`, il valore viene consumato dal blocco.

5.3 Monad

In rust non esiste `null`, l'assenza di valori o gli errori vengono gestiti con: `Option<T>` che contiene `Some(t)` o `None` e `Result<T,E>` che contiene `Ok(T)` o `Err(E)`. Sia `Option` che `Result` sono implementati come `enum`.

5.4 Polimorfismo

Il polimorfismo è la capacità di associare comportamenti comuni a un insieme di tipi differenti. In C++ esiste l'*ereditarietà multipla*, che permette di fare il polimorfismo.

```
#include <iostream>

class Alfa {
```

```
  int i;
public:
  virtual int m() { return 1; }
  virtual unsigned size() { return sizeof(*this); }
};

class Beta: pulbic Alfa {
  double d;
public:
  int m() { return 2; }
  unsigned size() { return sizeof(*this); }
};

int main() {
  Alfa* ptr1 = new Alfa;
  Beta* ptr2 = new Beta;

  std::cout << ptr1->m() << std::endl;  // 1
  std::cout << ptr2->m() << std::endl;  // 2

  std::cout << sizeof(*ptr1) << std::endl;  // 4
  std::cout << sizeof(*ptr2) << std::endl;  // 16

  delete ptr1;
  delete ptr2;
}

// Senza virtual
int main2() {
  Alfa* ptr1 = new Alfa;
  Alfa* ptr2 = new Beta;

  std::cout << ptr1->m() << std::endl;  // 1
  std::cout << ptr2->m() << std::endl;  // 1

  std::cout << sizeof(*ptr1) << std::endl;  // 4
  std::cout << sizeof(*ptr2) << std::endl;  // 4

  delete ptr1;
  delete ptr2;
}

// Con virtual
int main2() {
  Alfa* ptr1 = new Alfa;
  Alfa* ptr2 = new Beta;

  std::cout << ptr1->m() << std::endl;  // 1
  std::cout << ptr2->m() << std::endl;  // 2

  std::cout << sizeof(*ptr1) << std::endl;  // 16
```

```
  std::cout << sizeof(*ptr2) << std::endl;  // 16
  std::cout << ptr1->size() << std::endl;  // 16: "8vtable + 4int + 4padd"
  std::cout << ptr2->size() << std::endl;  // 24: "8vtable + 4int + 4padd + 8
    double"

  delete ptr1;
  delete ptr2;
}
```

Quando in una classe C++ appare `virtual`, gli oggetti della classe acquisiscono un campo in più detto `vtable`, che risolve il metodo chiamato dall'oggetto, che possono puntare alle funzioni di `Alfa` o di `Beta`.

In rust **non esiste l'ereditarietà**, però in rust i **Tratti** sono per rust come le **Interfaccie** per Java. *Un tratto esprime la capacità di eseguire una certa funzione*, as esempio: `std::io::Write`, `std::iter::Iterator`, ... In rust dunque non esiste l'overhead come in C++ causato dal polimorfismo, il motivo è che il compilatore risolve tutti i riferimenti in modo statico, l'unico momento in cui quel costo si presenta è quando si creano dei riferimenti dinamici (`&dyn SomeTrait`). Per implementare un tratto in rust si scrive:

```
trait T1 {
  fn num() -> i32;
  fn return_self() -> Self;
}

struct OtherType {};
impl T1 for OtherType {
  fn num() -> i32 { 2 }
  fn return_self() -> Self { OtherType }
}
```

Quando si crea un tratto si può definire un **Tipo Associato**.

```
trait T2 {
  type AssociatedType;
  fn f(arg: Self::AssociatedType);
}

struct OtherType {};
impl T2 for OtherType {
  type AssociatedType = &str;
  fn f(arg: Self::AssociatedType) {}
}
```

Rust permette un forma di dipendenza tra tratti e vengono detti **sub-trait** e **super-trait**:

```
trait Super {
  fn f(&self) { println!("Super"); }
  fn g(&self) {}
```

```
}

trait Sub: Super {
  fn f(&self) { println!("Sub"); }
  fn h(&self) {}
}

struct SomeType;
impl Super for SomeType;
impl Sub for SomeType;

fn main() {
  let s = SomeType;
  s.f();  // ERRORE: chiamata ambigua
  <SomeType as Super>::f(&s);
  <SomeType as Sub>::f(&s);
}
```

Esistono due tratti per poter fare dei confronti tra tipi su rust, che sono `Eq` e `PartialEq`. Un caso particolare di confronto sono i floating point. Quando si vogliono gestire dei confronti di ordine, ovvero quando un tipo è maggiore/minore, si utilizza il tratto `PartialOrd`. Quando si usa `PartialX` l'uguaglianza non è riflessiva mentre mentre senza `Partial` la comparazione è riflessiva.

Grazie ai tratti si possono ottere dei risultati simili a qualli che in C++ era l'operation overloading, infatti è possibile definire le operazioni di addizione, molitiplicazione, accesso con indice (dove l'indice è un tipo generico), ecc... Alcuni tratti molto caratteristici sono `Derfef` e `DerefMut`, da questo tratto deriva una serie di classi di Rust che vengono dette **Smart Pointers** che ci permettono di trattarli come dei puntatori senza che siano ne un puntatore e ne un riferimento, ma che restituisce un riferimento, si fa accesso alla variabile usando `*t`, infatti il tratto `Box` è proprio uno smart pointer, il tipo al loro interno è solo movibile e non copiabile.

Per effettuare le conversioni tra tipi esistono anche i tratti `From` e `Into`. Quando non si ha `Into` è possibile inferirlo da `From`. Esistono anche `TryFrom` e `TryInto` che restituiscono un `Result`.

5.5 Funzione Generiche

È possibile definire una funzione che implementa un *tipo generico* che al momento non implementa un tipo specifico.

```
fn max<T>(t1: T, t2: T) -> T where T: Ord {
  if t1 < t2 { t2 } else { t1 }
}
```

Quando si utilizza una funzione generica per ogni tipo nuovo che viene passato alla funzione, viene generata dal compilatore una nuova versione della funzione specializzata per il tipo passato, questo processo viene detto **monomorfizzazione**, perché

si passa da *polimorfismo* a *monomorfismo*. I tipi generici si possono usare anche all'interno delle struct.

```
struct MyS<T> where T: SomeTrait {
  a: T,
}

impl<T> MyS<T> where T: SomeTrait {
  fn a(b: T) -> Self {
  }
}
```

Con i tipi generici si possono aggiungere dei vincoli sui tipi generici, per farlo esistono due sintassi: si specificano all'interno di `<>` oppure specificandoli con la keyword `where`.

Example 5.1 – Dynamic vs Generic

```
fn dynamic(w: &mut dyn Write) { ... }
fn generic<T>(w: &mut T) where T: Write { ... }

```

La differenza tra `dynamic` e `generic` è che: `dynamic` viene compilata una sola volta, ma viene passato un fat pointer e per accedere ai metodi di `w` si dovrà passare prima dalla v-table, mentre in `generic` viene passato solo il puntatore semplice però il codice viene duplicato, causato dalla monomorfizzazione.

5.6 Lifetimes

In Rust ogni riferimento ha tempo di vita, il compilatore è in grado d'inferire il tempo di vita dei riferimenti nella maggior parte dei casi, quando questo non è possibile deve essere scritti esplicitamente, la sintassi per farlo è: `fn test<'a>(number: &'a i32)`, dove le lifetimes sono precedute da un apice e vengono dalle lettere dell'alfabeto in modo ordinato.

Example 5.2 – Uso delle Lifetimes

Si preda come esempio il caso in cui un funzione che prende in input due riferimenti e ne ritorna un'altro:

```
fn max<'a>(num1: &'a i32, num2: &'a i32) -> &'a i32 {
  if num1 > num2 {
    return num1;
  } else {
    return num2;
  }
}

```

In questo caso le lifetimes coincidono, proviamo ad usarla:

```
fn main() {                 // lifetimes:
  let n = 10;               // n+
  let res;                  //  |
  {                         //  |
    let m = 19;             //  | m+ 'a+
    res = max(&n, *m);      //  |  |   |
  }                         //  |  |   |
  println!("max: {}", res); //  |
  // ERRORE-----------^     //  |
}                           //  |
```

In questo caso il codice non compilerà perché quando si utilizzano delle lifetimes si assume come intervallo **l'interseizione delle lifetimes** e quindi viene presa la più piccola, in questo casa `'a` concide con il tempo di vita di `b`.

5.7 Closures

Una **funzione di ordine superiore** è detta tale se prende in input a da in output una funzione. In rust è possibile assegnare ad una variabile una funzione is suo tipo sarà `fn(T1, T2, ..., Tn) -> U`. Quando una variabile ha un tipo `fn` questa sarà legata ad una funzione classica, se invece si usano le *lambda* esistono altri tre tipi:

- `FnOnce`: struttura che può essere invocata una sola volta.
- `FnMut`: la struttura può essere chiamata più volte e a ogni chiamata può alterare il suo stato.
- `Fn`: la struttura può essere chiamata più volte ma non può mutare il suo stato.

Il corpo della funzione lambda può fare riferimento alla variabili locali nello scope in cui sono state create, in questo caso la lambda acquisisce uno stato e le variabili vengono dette **variabili libere**, in questo caso si dice che le viabili vengono **catturate**. La cattura può avvenire per **riferimento**, e quindi la variabile sarà accessibile solo nel suo tempo di vita, oppure la cattura può avvenire per **movimento**, dove la variabile viene mossa all'interno della funzione lambda e quindi ne prende il possesso.

```
let number = 10;
let with_reference = |...| { number ... };
let with_movement = move |...| { number ... };
```

Il tipo di una lambda è inferito in compilazione a differenza degli utilizzi, oppure si può specificare:

```
fn generator<F>(base: &str) -> F
where F: impl FnMut() -> String {
  let mut v = 0;
  return move || {
```

```
    v += 1;
    format!("{}{}", base, v)
  };
}
```

6 Errori ed Eccezioni

In Rust non esistono le eccezioni ma usa invece `Result<T, E>` e `Option<T>`. In Java quando si lancia un eccezione si ritorna al più basso contesto di gestione di eccezioni, dove si va alle differenti `catch` fino a trovare un match, in C++ le eccezioni funzionano quasi in modo identico

```
int f2() {
  if (...)
    throw std::logic_error("err");
  return 1;
}
int f1() {
  try {
    return f2();
  }
  catch (std::locic_error e) {
    return -1;
  }
}
```

in C++, alla radice di programma c'è un `try catch` in grado di catchare tutte le eccezioni non gestite. Quando viene scritto un blocco `try`, nello stack viene inserito un nuovo **exception context**, che avrà il riferimento a quello precedente, il corrente invece si trova in un registro dalla CPU, in questo modo quando viene lanciata un eccezione in modo ricorsivo si controllano tutti gli exception context fino a trovare un match. A differenza di Java in C++ si può lanciare qualsiasi valore invece di classi `throwable`, quando viene lanciata un eccezione il pattern RAII permette di pulire lo stack prima di lasciare la funzione.

In Rust non esistono le eccezioni e adotta un approccio funzionale con `Result` e `Option`, il tipo `Result` è una **Monade** che incapsula il valore o l'errore. In Rust si può terminare l'interno processo con `exit` oppure si può terminare il thread corrente con `panic!`. Gestire gli errori è tedioso, allora Rust mette a disposione l'operatore `?` che in caso di successo ritorna `Ok` altrimenti esce dalla funzione e ritorna `Err`. Per gestire più tipi di errori come valori di ritorno si può astrarre l'errore a uno generico usando come valore di ritorno

```
Result<(), Box<dyn error::Error>>
```

poi si può anche generare un nostro errore, per fare questo si implementa il tratto `Error`, il problema è che si devono implementare una serie di tratti, che sono richi-

esti da `Error`: `Display, Debug, From`. Un crete chiamato `thiserror`, permette di implentare direttamente il tratto `Error` per ogni tipo usando la macro `#[derive]`

```
#[derive(Error, Debug)]
enum SumFileError {
  #[error("IO error {0}")]
  Io(#[from] io::Error)

  #[error("Parse error {0}")]
  Io(#[from] ParseIntError)
}
```

Un altro create è `anyhow`, che gestisce gli errori in modo ideomatico.

7 Iteratori

Un iteratore è una struttura dati che ha un suo stato, che premette di esplorare i dati contenuti all'interno di un altra struttura. Rust usa il tratto `Iterator`, che ha un solo metodo da implementare

```
trait Iterator {
  type Item;
  fn next(&mut self) -> Option(Self::Item);
}
```

poi esistono 3 metodi che servono a trasformare un oggetto in un iteratore: `iter(), into_iter(), iter_mut()`.

Un **adattatore** è un metodo che ogni iteratore offre che permette di trasformare l'iteratore originale in uno diverso, fino ad arrivare ad un consumatore finale

```
let sum = list
  .iter()
  .filter(|x| x.len() < 4)
  .map(|x| x.into::<i32>())
  .sum();
```

alcuni adattori sono: `map`, `take_while`, `skip`, `skip_while`, `peekable`, `fuse`, `rev`, ... I **consumatori** invece consumano un adattatore.

8 Collezioni di Dati

Le collezioni mettono a disposizione: `new(), len(), clear(), is_empy(), iter()`.

`VecDeque` è una coda in cui si può inserire/rimuovere in modo costante in testa e in coda, in Rust viene implementato come buffer circolare.

`Mappe`, in Rust sono presnti due tipi di mappe che sono `HashMap` e `BTreeMap`, il vantaggio sul tratto delle chiavi è che possibile fare delle operazioni di ricerca e di sostituzione una sola volta, questo grazie al tratto `Entry<'a,K,V>`

```rust
let mut animals = HashMap::<&str, u32>::new();

animals.entry("dog")
  .and_modify(|e| *e += 1)
  .or_insert(1);
```

`BinaryHeap<T>` è una collezione di elementi con una priorità associata, infatti `T` deve implentare il tratto `Ord`, `peek()` mi permette di vedere l'elemento più grande in tempo unitario.

9 IO

Nei sistemi operativi i file sono stati implementati in modo gerarchico per ragioni storiche, anche se sarebbe molto più effincente rappresentare tutto con una mappa e la rappresentazione gerarchica solo come un tag. Esiste anche il problema della compatibilità tra i diversi SO e le diverse system call, Rust cerca di unificare le diverse rappresentazioni nella libreria standard, per ovviare a questo esistono `Path` e `PathBuf`, la rappresentazione interna dei nomi dei file è fatta con `OsString` che permette di avere compatibilità tra i diversi SO. Per aprire un file si utilizza `std::fs::OpenOption` o `File`. La libreria mette a disposizione quattro tratti fondamentali che sono `Read`, `BufRead`, `Write`, `Seek`.

10 Smart pointer

In C e C++ è lecito creare dei puntatori e derefernziarli con `*`, in Rust cosí come avere un tratto per usare l'uguaglianza tra due struct, o l'ordine, è lecito implementare il tratto per la dereferenziazione pur non essendo un puntatore, anche questo è possibile farlo in C++ con l'operator overloading. In Rust questo quando si vuole che una struttura viva fin tanto che esiste qualcuno che la referenzia si vuole che questa struttura venga mantenuta in vita, grazie al conteggio delle referenze. Il bisogno di gestire queste casistiche in C++ si sono introdotti gli **smart pointer**, questo stesso concetto è stato portato in Rust. In Rust ad esempio è impossibile creare una struttura ciclica, perché sorge la domanda, *chi possiede chi?*

In C++ si accede a un pezzo di memoria dinamica ci si deve assicurare che quel pezzo di memoria esiste e che il possessore sia uno solo, per questo motivo esiste il tipo `std::unique_ptr<T>` (che non può essere duplicato) e `std::shared_ptr<T>` (che può essere copiato).

In Rust vengono ripresi gli stessi concetti di C++ con `Box`, `Rc` (Reference Count), `Arc` (Atomic Reference Count), ... Una struct che implementa il tratto `Deref/DerefMut`, permette di utilizzare delle struct come se fossero delle referenze

```rust
trait Deref {
  type Target: ?Sized;
```

```
  fn deref(&self) -> &Self::Target;
}
trait DerefMut: Deref {
  type Target: ?Sized;
  fn deref_mut(&mut self) -> &mut Self::Target;
}
```

as esempio `Box<T>` si comporta come se possedesse una reference al tipo `T`, in più quando `Box` verra droppato rilascerà i dati contenuti al suo interno allocati sull'heap. `Rc` permette invece di contare il numero di volte che il dato è referenziato, solo quando il numero di conteggi `strong` e `weak` saranno zero la struttura `T` non verrà rilasciata. Per risolvere il problema dei cilci si utilizza `Weak`, per distinguere i puntatori che tengono in vita e quelli che non lo fanno, per ovviare al fatto di puntare ad una refernza nulla, per fare fare questo `Weak` offre `upgrade()`, che permette di ritornare ad `Rc` e se il dato è ancora esistente ci permette di accedere altrimenti no.

```
let five = Rc::new(5);
let weak_five = Rc::downgrade(&five);

let strong_five = weak_five.upgrade();
assert!(string_five.is_some());

drop(strong_five);
drop(five);

assert!(weak_five.upgrade().is_none());
```

Esistono delle situazioni in cui si potrebbe volere un possesso condiviso ad un dato sullo stack, questo è un problema perchè un solo una variabile può possedere un dato come mutabile, Rust permette però di cambiare un dato anche possedendo un refernce non mutabile, il borrow checker dice che: esiste un unico riferimento mutabile, ci sono più riferimenti, in alcuni casi questa condizione è troppo restrittiva, per questo motivo esiste `Cell<T>` che permette di scambiare un riferimento con un riferimento mutabile.