**Progetto SOA**

**0** – **Introduzione:**

LKM 🡪 loadable linux kernel != base kernel

(hanno 3 maxi categorie)

- device driver

- filesystem driver

- system calls

**1** – **Specifiche progetto:**

• Scambio dati basato su TAG

• Montare un modulo

* + - Ricerca della tabella delle system call
    - Scrivere 4 chiamate di sistema
      * Get
      * Send
      * Receive
      * Control
    - Il software deve gestire 256 tag
    - Messaggi almeno 4 KB
    - 32 livelli
    - Device driver

**2 – Moduli Linux:**

• Comando per vedere tutti i moduli caricati

lsmod

• Comando per caricare i moduli

insmod + modulo

• Comando per vedere tutti i moduli caricati

rmmod + modulo

* **Hello world moduli**

<https://www.geeksforgeeks.org/linux-kernel-module-programming-hello-world-program/>

Modulo = pezzo di codice che può essere caricato o scaricato a domanda.

Un modulo può essere aggiunto in due modi:

1) Aggiungere un modulo alla “kernel source tree” e ricompilare il kernel.

2) Aggiungere un modulo in corsa.

* **Codice di Quaglia sui moduli**

**Kprobe-interceptor** 🡪 prende il contenuto dello standard input e lo passa al modulo.

**Hook\_init** 🡪 atomic counter che non ci da problemi di concorrenza.

large target pid 🡪 il processo per cui vogliamo osservare lo standard input.

current 🡪 all’interno della the\_hook prende l’indirizzo di memoria corrente del thread.

copy\_from\_user e inserisco il terminatore di stringa.

**User**: oggetto che scrive su standard input e legge da standard output.

**Kprobe usage** 🡪 aggiungiamo altri moduli al nostro hook.

**Finish task with** 🡪 determina che si è concluso con un cambio di thread.

**Tail\_hook** 🡪 parla di hook\_function che è tipicamente pari a 0 ma se lo relaziono a qualcos’altro lui lo va a chiamare. Lo colleghiamo a kernel function pointer exposition che espone un function pointer (indirizzo di memoria associato a una funzione).

**3 – Makefile:**

La seconda riga del makefile ci dice che usiamo due moduli per creare un unico kernel object.

**4 – Multithread:**

Argomento utile per il progetto.

**5 – Riconciliazione temporale:**

Argomento utile per il progetto, top/bottom half.

**6 – System call:**

Trap e interrupt 🡪 migliorano il livello di privilegio.

Entry della tabella degli interrupts sono GATE.

Gestire le trap 🡪 meccanismo dei gate 🡪 inseriti in una tabella di livello kernel, interrupts table, come detto ogni entry identifica un gate che è a sua volta identificato da un id segmento ea quale offset punta.

Il gate ci indica anche qual è il valore massimo che può passare attraverso quel gate (passaggio cross segmento) e va confrontato con il CPL di ogni thread (di quello che fa l’accesso). Quando eseguiamo in user mode abbiamo un certo CPL e vogliamo iniziare a operare in modalità kernel dobbiamo usare un gate 🡪 quando entriamo il CPL viene portato a 0.

Passare da privilegio 0 a 3 è comunque un passaggio di cross segmento ma non ha bisogno dei gate.

Se viene fatto un passaggio cross segmento illegale viene lanciata un’eccezione e quindi un gestore di eccezione.

Passare a un segmento con un livello di privilegio migliore si può fare attraverso l’istruzione INT 🡪 che prende come unico parametro il displacement all’interno della tabella delle trap che descrive il gate che vogliamo utilizzare (poi il gate indica anche il segmento dove vogliamo spiazzarci).

**7 – Gate**

ITD 🡪 interrupt descriptor table 🡪 ciascuna entry è un gate.

IDTR ci permette di accedere a IDT 🡪 il parametro che passiamo a INT deve essere compatibile con la size della IDT che è indicato in IDTR (riprogrammabile).

CPL contenuto nel registro CS.

Nel gate ci sono altri metadati che ci indicano come deve essere settato lo stato del processore all’ingresso del gate.

Il gate che ci fa accedere al kernel on demand è unico, in particolare negli x86 a 32 bit è all’interno della entry a spiazzamento 0x80 (esadecimale) (INT).

Gli altri gate vengono usati per altri motivi, come ad esempio accedere agli interrupt.

Il dispatcher, che è in grado di dispatchare ogni servizio all’interno del kernel, è il blocco di codice a cui ci rimanda quest’unico gate.

Per far capire al (unico) dispatcher quale servizio vogliamo attivare utilizziamo un’ulteriore tabella puramente software 🡪 system call table 🡪 ciascuna entry è un puntatore a una funzione di livello kernel. Quindi il dispatcher deve ricevere l’indice numerico del servizio che deve richiamare.

Il dispatcher eesegue già a CPL 0, attraverso CALL<offstet> andiamo a chiamare la syscall. Il suo output andrà a finire nei registri di sistema. Poi il dispatcher ci deve riportare in modalità user (DTI).

**8 – System call** **2:**

1. Servizio non dispatchabile dall’attuale dispatcher, quindi dobbiamo crearne uno nuovo nostro
2. Oppure usando lo stesso seguendo le regole che segue per altre syscall

Dobbiamo sicuramente andare a modificare la system call table 🡪 trovare una entry libera e agganciare il pointer alla nostra funzione oppure espanderla (ma è molto difficile).

La syscall che implementiamo deve essere compliant a livello di formato e di regole rispetto al dispatcher.

Indicizzare le syscall 🡪 associare il codice numerico k a una funzione (o meglio ad un pointer di una funzione) 🡪 con unistd\_63 cambiano i codici numerici rispetto a unistd\_32 (che è ancora supportato).

A livello user le info riguardanti l’indicizzazione le troviamo all’interno di alcuni header come, ad esempio, unistd\_64.h

Un user per accedere al gate deve passare il codice numerico e anche i parametri.

Il meccanismo dei gate necessita di linguaggio assembly.

**9 – Stub di system call:**

Entra in modo kernel e chiama la reale attività, gli stub vengono costruiti attraverso macro livello user (che hanno anche istruzioni machine dependent) a cui passiamo i 7 parametri.

Syscall0(type, nome)

0 🡪 perchè prende 0 parametri

Type 🡪 del valore di ritorno

Nome 🡪 della funzione C deve essere generata

**10 – Utili:**

<https://www.generacodice.com/it/articolo/36965/>

asm line prende i parametri dai registri di processore e non dallo stack

**11 – Syscall:**

Immagine che contiene testo, ricevuta

Descrizione generata automaticamente

I registri MSR ci aiutano perché avito di accedere alla memoria ogni volta. Sta tutto nel processore.

Abbiamo 2 dispatcher diversi proprio grazie a MSR.

Per utilizzarli in passato si usava sysenter (32bit) ma ora usiamo syscall (64bit) 🡪 con questa usiamo i registri MSR e non serve lo stack pointer.

Non abbiamo quindi uno stack switch fino a che non prende il controllo il kernel; per tornare in modo user a 64bit dobbiamo usare sysret (trova in rcx il punto di ritorno).

Troviamo syscall() in stdlib.h 🡪 API che diventa una sorta di dispatcher livello user.

**12 – System call table:**

Difficilmente possiamo operare resizing della tabella, eseguiamo quindi l’hacking della system call table per cercare entry libere. Tutte le entry ora sono occupate ma alcune puntano allo stesso servizio x, ovvero un servizio che implementa tutti i servizi che non hanno una reale implementazione 🡪 ni\_syscall.

**13 – Sys ni\_syscall:**

La tabella possiamo trovarla in:

• arch/i386/kernel/entrys.c

• arch/x86/entry/syscall64.c

sys\_qualcosa ci permette di accedere al frontend di ‘qualcosa’ all’interno del kernel.

Quaglia a lezione ci ha fatto vedere l’uso delle macro per utilizzarle come trampolino per i frontend delle syscall che servono ad offuscare la stack area.

**14 – RCU:**

Read-Copy-Update è un meccanismo di [sincronizzazione](https://it.wikipedia.org/wiki/Sincronizzazione) presente in alcuni sistemi operativi moderni, spesso usato in alternativa ai [lock](https://it.wikipedia.org/wiki/Lock" \o "Lock) di scrittura-lettura. Sono caratterizzati da un ritardo di lettura nella pratica nullo, compensato da una fase di scrittura più dispendiosa. Per questo motivo si adattano particolarmente alle strutture dati frequentemente lette e raramente scritte. Il funzionamento ricorda il pattern publish/subscribe: la lettura avviene all'interno di sezioni critiche in cui il thread lettore vede una versione coerente del dato, mentre chi scrivere deve "pubblicare" il nuovo valore. In questo modo si crea il concetto di versioni multiple dell'oggetto condiviso.

Il concetto chiave del meccanismo RCU è che il thread lettore non può entrare in attesa. Questo causa i ben noti problemi di sincronizzazione: come garantire che nessuno legga se l'aggiornamento del dato non è ancora terminato e come garantire che una sequenza di letture garantisca il ritorno dello stesso valore.

Nello schema RCU il lettore indica che è presente, legge e indica che non è più presente. Lo scrittore invece prende il lock, scrive e aspetta che tutti abbiano letto per rilasciare il lock.

**15 – Utili:**

Ricordare di fare alla fine il cleen up dei moduli rimettendo le ni\_syscall.

**16 – Codice:**

1. Lato user 🡪 scrivere il codice che lato user richiama la nostra system call tramite syscall()
   1. Il codice user Quaglia lo spiega a lezione 16 al minuto 97.
2. Lato kernel 🡪 all’interno del modulo da mantenere inserire la funzione che hackera la tabella delle system call e il frontend delle chiamate di sistema (ovvero la loro effettiva implementazione).

**17 – Implementazione:**

* ­Primo step: abbiamo eseguito l’helloworld dei moduli kernel in cui abbiamo scritto e montato un modulo che quando montato eseguiva delle stampe sul ring buffer del kernel e stessa cosa faceva quando veniva smontato.
* Secondo step: abbiamo aggiunto un makefile con delle rules che automatizzasse il nostro processo.
* Terzo step: abbiamo preso ed eseguito il codice del prof. che andava a trovare le entry della syscall table in cui avevamo sys\_ni\_syscall ovvero entry libere che puntavano a ni\_syscall. In questo modo vengono segnate le entry che possono essere hackerate dal programmatore ovvero in cui è possibile aggiungere delle nuove syscall.

Per fare ciò il codice cerca la pagina in cui la syscall table inizia e quindi in cui cercare le n entry che puntano a ni\_syscall da utilizzare per le nostre syscall hackerate.

Quando abbiamo trovato le n aree di memoria che puntano a ni\_syscall possiamo far cessare la ricerca e stampiamo nel log del kernel il risultato della nostra ricerca ovvero le entry della syscall tabel in cui abbiamo trovato le sys\_ni\_syscall.

Ripuliamo tutto e smontiamo il modulo.

* Quarto step: abbiamo modificato il codice del prof. per adattarlo al nostro scopo, abbiamo aggiunto un array ed inserito una funzione che ci restituisce le 4 entry in cui possiamo inserire le hacked syscall come da specifica in modo da poterle utilizzare in un secondo momento. Una volta che la nostra ricerca ha avuto esito positivo, ovvero abbiamo trovato le 4 entry e abbiamo salvato i relativi dati nell’apposita struttura, abbiamo creato 2 dummy hacked syscall per testarle. La prima eseguiva la stampa del parametro mentre la seconda eseguiva la stampa della somma di due valori passati come parametri.

Ripuliamo tutto e smontiamo il modulo.

* Quinto step: abbiamo affinato la scrittura delle hacked syscall in un secondo file che viene incluso nel main file in cui abbiamo il frontend mockato delle syscall e che viene richiamato a livello user dal file user che esegue la demo di quello che è stato implementato fin’ora.

**18 – Scelte implementative:**

* ­32 livelli, 256 tag, messaggio 4KB.
* Il tag è una struct che ha dentro di sé un puntatore al livello che è anch’esso una struct che ha come campi, tra gli altri, il buffer (messaggio) e l’intero che definisce il livello corrispondente.
* Le syscall vengono implementate come ‘servizi’.
  + Tag\_get(int key, int command, int permission) 🡪 la chiamata di sistema istanzia o apre il servizio Tag associato alla key dipendente dal valore command. Il valore di ritorno dovrebbe indicare il tag service descriptor.
  + Tag\_send(int tag, int level, char\* buffer, size\_t sizeBuff) 🡪 prende da buffer e manda su un livello dentro un tag, tutti i thread bloccati si svegliano e leggono. La chiamata di sistema, quindi, consegnerà al Tag con descrittore il messaggio in buffer, tutti i thread in attesa sul valore level dovrebbero essere ripresi e ricevere; se nessuno è in attesa il messaggio viene scartato.
  + Tag\_receive(int tag, int level, char\* buffer, size t\_sizeBuff) 🡪 il thread si blocca e aspetta che arrivi un messaggio. Il servizio permette quindi a un thread di ricevere il messaggio e può fallire con un segnale mentre un utente è in attesa.
  + Tag\_ctl(int tag, int command) 🡪 con il commando possiamo differenziare i comportamenti: AWAKE\_ALL sveglia tutti i thread in attesa di messaggi e REMOVE per rimuore il tag (un tag non può essere rimosso se sono presenti thread in attesa su di esso).

**19 – Flusso del programma:**

* Init module 🡪 monta il modulo
* Syscall\_table\_finder() 🡪 guarda nella tabella, riempie l’hacked\_syscall\_table (indirizzo della tabella) e hacked\_ni\_syscall (indirizzo della 1° sys\_ni\_syscall).
  + Syscall\_number\_finder chiama syscall\_table\_finder e fill\_ni\_syscall filla l’array con le entry delle ni\_syscall.