The Devastation of Meltdown

J.Nider

2019-11

Background

Virtual Memory

OoO Execution

Linux Memory Management

The Exploit

The Fix

The Damage

Background

- Meltdown allows an unprivileged app to read ALL memory of a victim machine
- Official name: CVE-2017-5754 “Rogue Data Cache Load” (RDCL)
- Caused by a race condition in out-of-order CPU’s
- NSA potentially knew about this since 1995

Scope

- Affects almost all Intel processors [1990-2018]
- IBM POWER7,8 Z
- ARMv8 A-series
- AMD is not vulnerable!!

- All operating systems are affected (Linux, Windows, Android, etc)
- Containers are affected (LXC, Docker, OpenVZ, etc)
- Hardware-supported virtual machines are not (KVM, VMWare ESX, etc)

Background

Virtual Memory

OoO Execution

Linux Memory Management

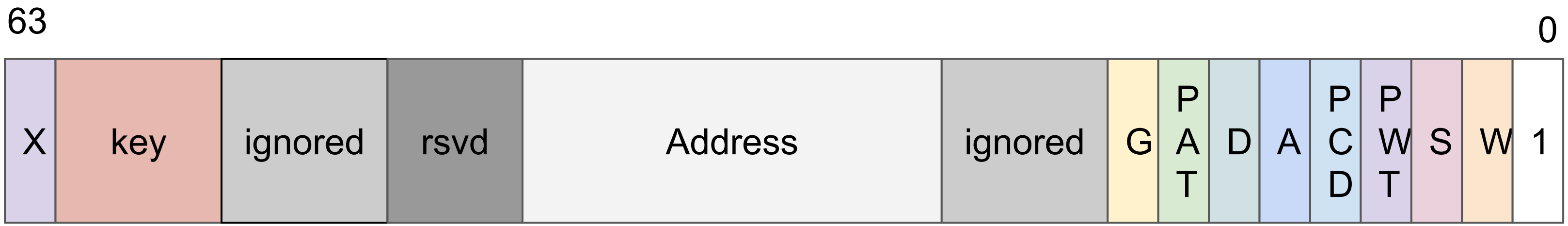
The Exploit

The Fix

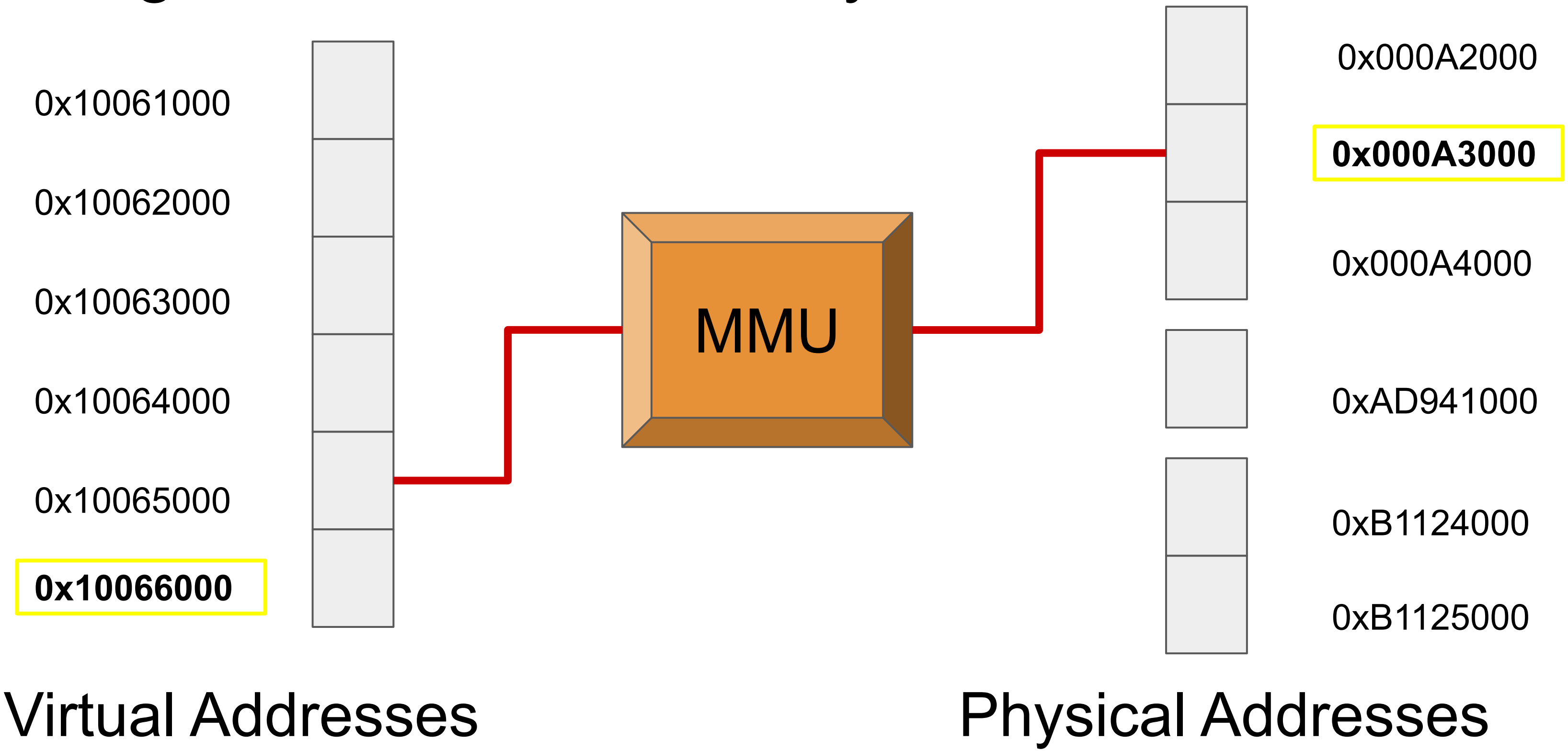
The Damage

Virtual Memory

- Memory is organized into pages
  - Page sizes range from 4KB to up 1GB
- MMU maps virtual address to physical addresses
  - Usually through page tables, but there are other mechanisms (such as hashing)
- Each page has attributes
  - Describes permissions (RWX), (S) and caching (C)
- Translations are cached in TLB (translation lookaside buffer)



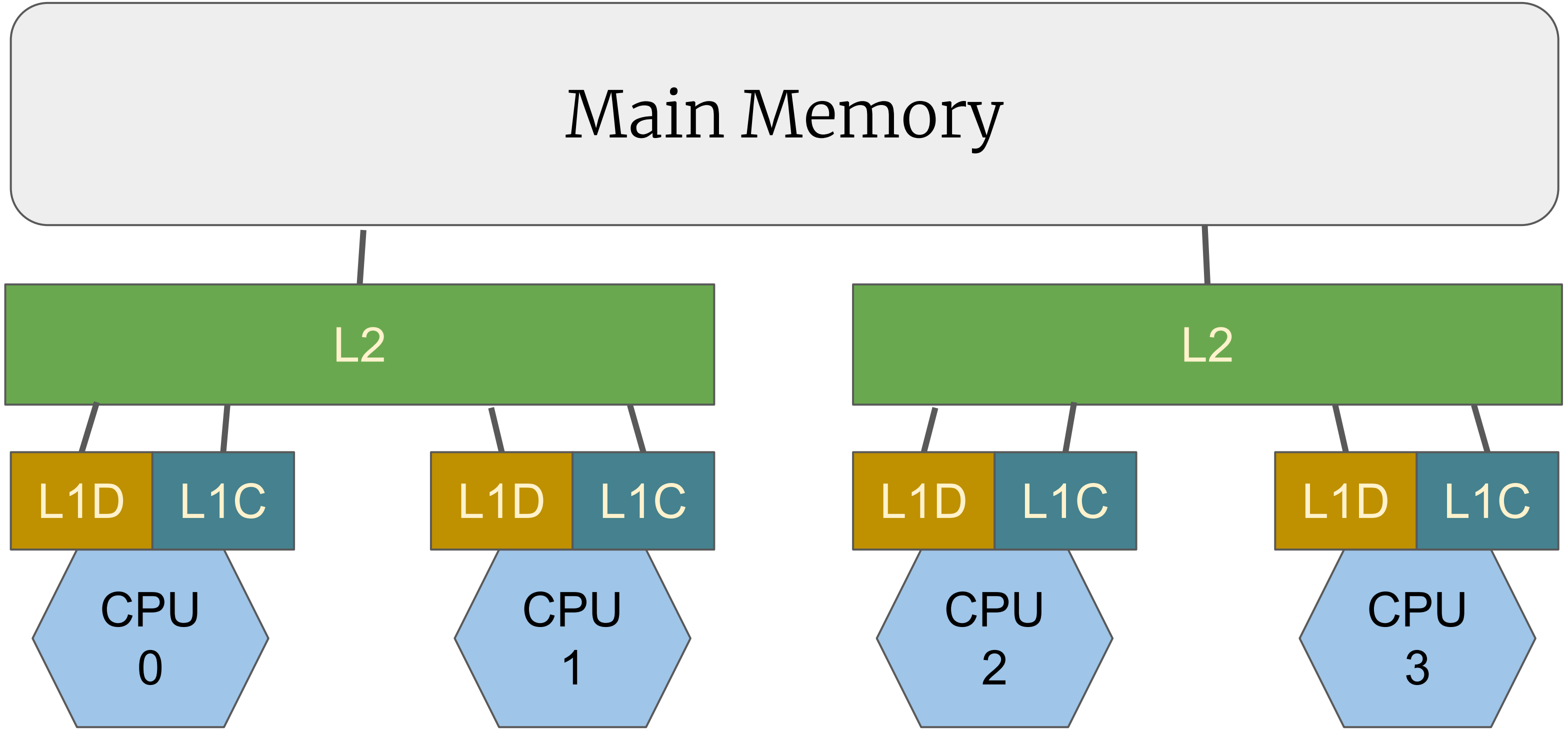
Background: Virtual Memory



Cache Organization

- Reading from main memory is slow!
  - In the range of 400-800ns
  - Therefore we want to avoid main memory as much as possible
- So we cache (make a copy) of any data in a smaller, faster memory
  - Much faster - in the range of 10-100ns
  - Faster memories are more expensive
- We can make a hierarchy with different attributes
  - Capacity
  - Access time
  - Mapping (Direct, Associativity)
  - Multiple ways
- **Cache is not part of the Instruction Set Architecture!**
  - **It is part of the microarchitecture**

Cache Organization



Background

Virtual Memory

OoO Execution

Linux Memory Management

The Exploit

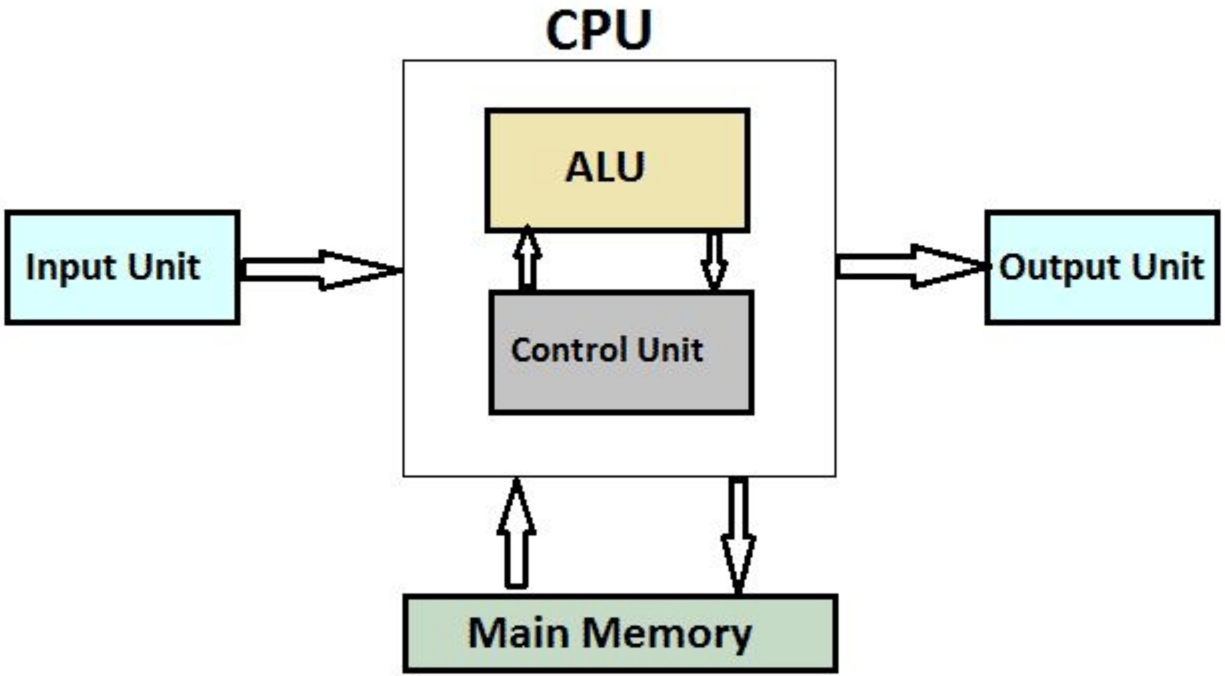
The Fix

The Damage

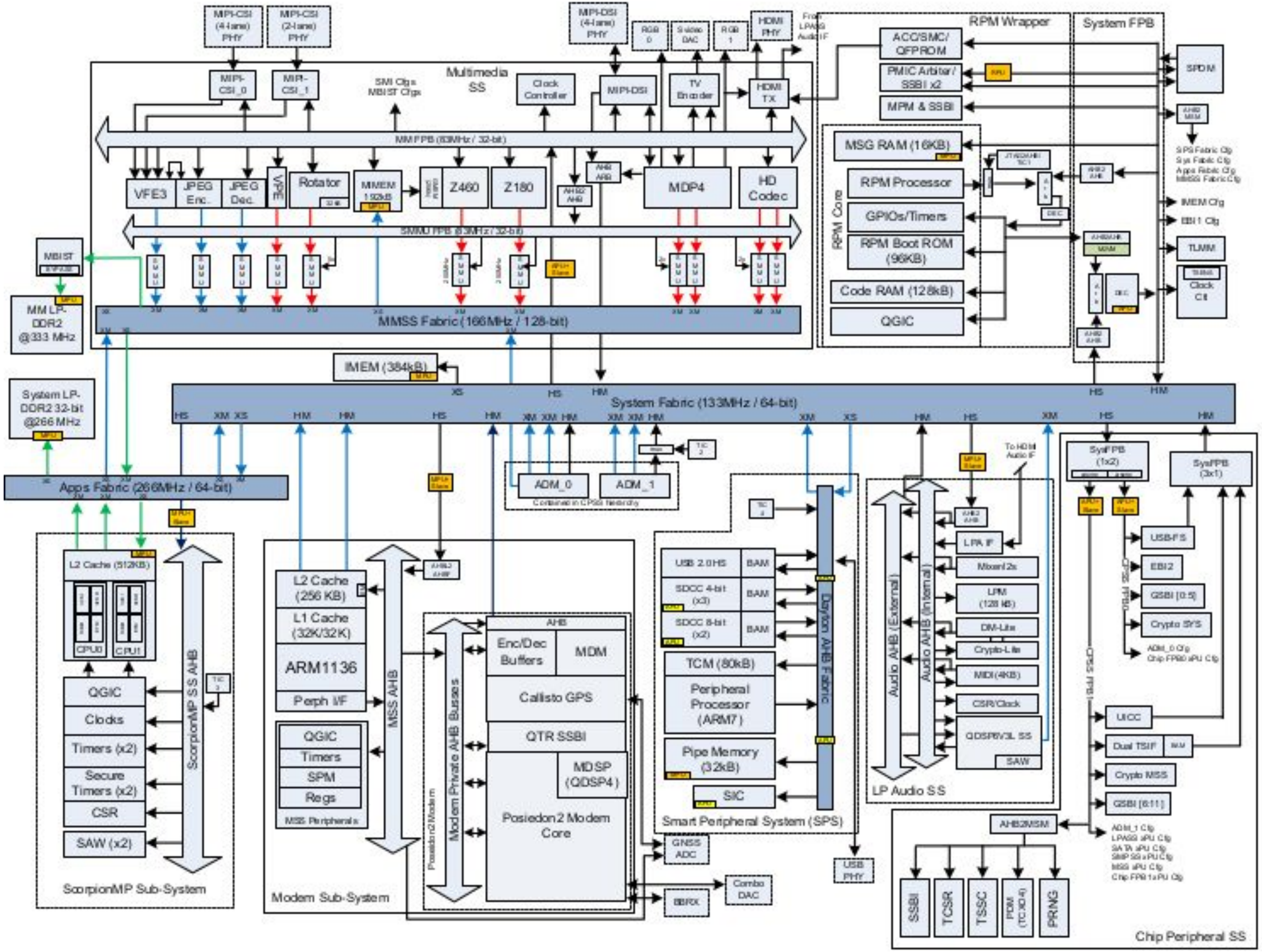
Out-of-order Execution

- CPUs are made up of many hardware blocks
    - Integer units
    - Floating point units
    - Internal registers
    - Many more
- Not all hardware is used for each instruction
- Some instructions wait even though there are not dependencies
- We want to work as fast as possible

CPU Architecture - Simplified



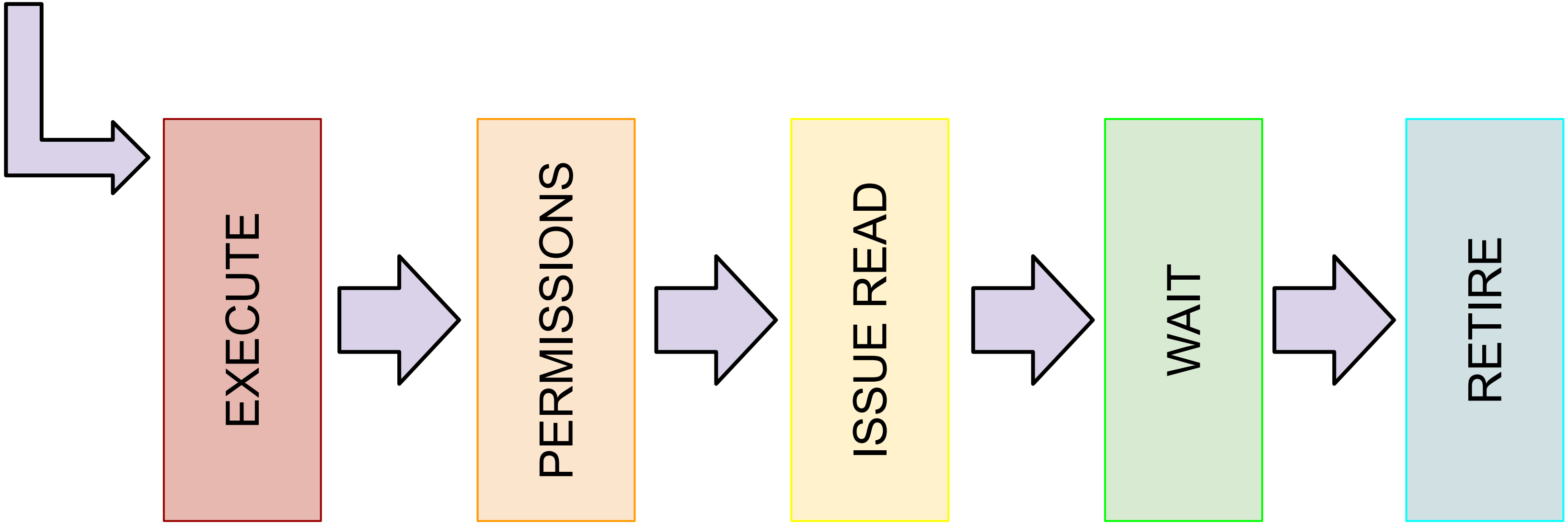
Real Life



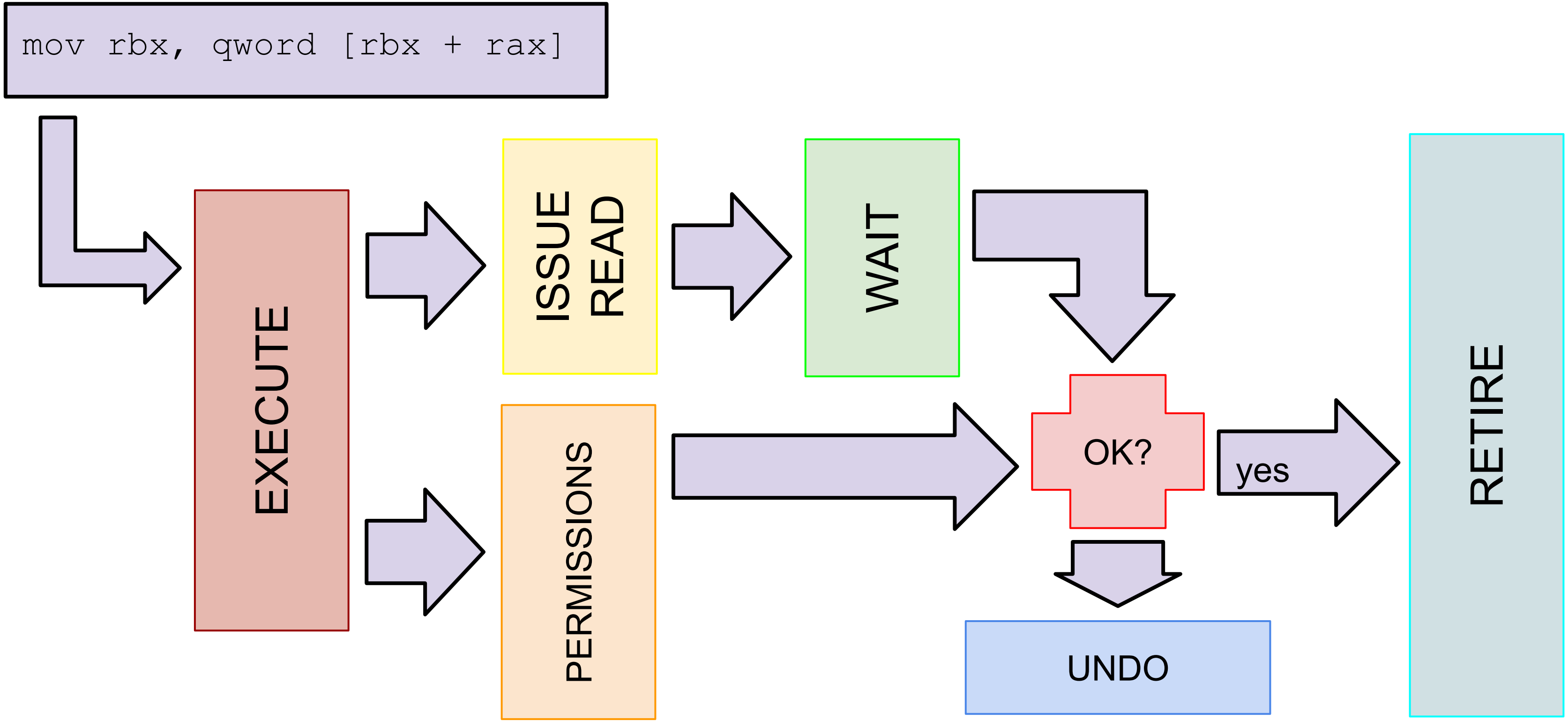
Multimedia SS
MM FPB (83MHz / 32-bit)
SMMU FPB (83MHz / 32-bit)
MMSS Fabric (166MHz / 128-bit)
VFE3
JPEG Enc.
JPEG Dec.
VPE
Rotator
MMEM 192kB
Z460
Z180
MDP4
HD Codec
MIPI-CSI_0
MIPI-CSI_1
MIPI-DSI
TV Encoder
HDMI TX
Clock Controller
RPM Wrapper
System FPB
ACC/SMC/ QFPROM
PMIC Arbiter/ SSBI x2
MPM & SSBI
MSG RAM (16KB)
RPM Processor
GPIOs/Timers
RPM Boot ROM (96KB)
Code RAM (128kB)
QGIC
RPM Core
SPDM
TLMM
IMEM (384kB)
System Fabric (133MHz / 64-bit)
System LP-DDR2 32-bit @266 MHz
MM LP-DDR2 @333 MHz
MBIST
Apps Fabric (266MHz / 64-bit)
L2 Cache (512KB)
CPU0
CPU1
QGIC
Clocks
Timers (x2)
Secure Timers (x2)
CSR
SAW (x2)
ScorpionMP SS AHB
ScorpionMP Sub-System
L2 Cache (256 KB)
L1 Cache (32K/32K)
ARM1136
Perph I/F
QGIC
Timers
SPM
Regs
MSS Peripherals
MSS AHB
Modem Private AHB Busses
Enc/Dec Buffers
MDM
Callisto GPS
QTR SSBI
MDSP (QDSP4)
Posiedon2 Modem Core
Modem Sub-System
ADM_0
ADM_1
USB 2.0HS
SDCC 4-bit (x3)
SDCC 8-bit (x2)
BAM
TCM (80kB)
Peripheral Processor (ARM7)
Pipe Memory (32kB)
SIC
Dayton AHB Fabric
Smart Peripheral System (SPS)
GNSS ADC
BBRX
Combo DAC
USB PHY
Audio AHB (External)
Audio AHB (Internal)
LPA IF
Mixer/I2s
LPM (128 kB)
DM-Lite
Crypto-Lite
MIDI(4KB)
CSR/Clock
QDSP6V3L SS
SAW
LP Audio SS
SysFPB (1x2)
SysFPB (3x1)
USB-FS
EBI2
GSBI [0:5]
Crypto SYS
UICC
Dual TSIF
Crypto MSS
GSBI [6:11]
AHB2MSM
SSBI
TCSR
TSSC
PDM (TCXO4)
PRNG
Chip Peripheral SS

Memory Read (sequential)

```
mov rbx, qword [rbx + rax]
```



Memory Read (parallel)



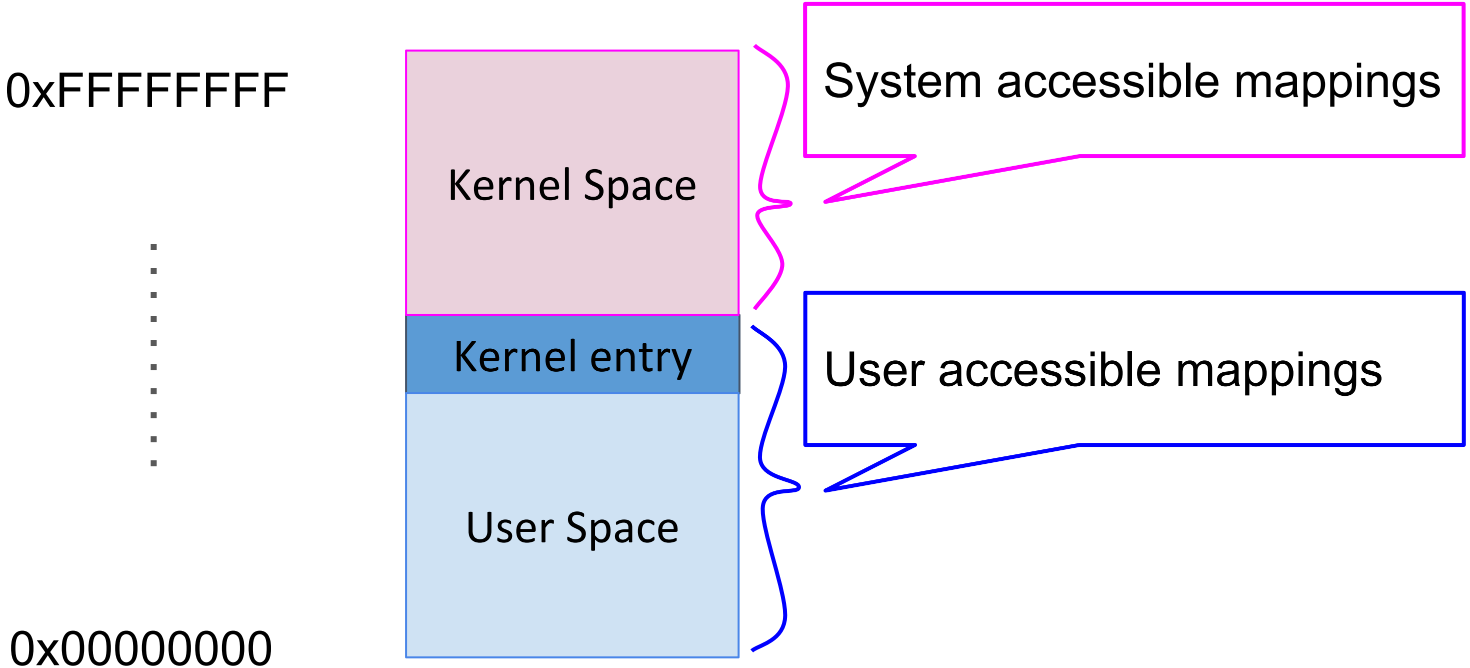
Background
Virtual Memory
CPU Pipelining
Linux Memory Management
The Exploit
The Fix
The Damage

Linux Memory Management

- The operating system uses the hardware to give us an abstraction
  - We can make assumptions without understanding the details of the hardware
- Makes sure we can get physical memory when we need it
- Use the hardware to protect our memory from other programs

Start	**End**	**Size**	**Description**
`0000000000000000`	`00007fffffffffff`	128 TB	user-space virtual memory
`0000800000000000`	`ffff7fffffffffff`	~16 EB	empty
`ffff800000000000`	`ffffffffffffffff`	128 TB	Kernel-space virtual memory

Linux Virtual Address Space Layout (w/o KPTI)



Linux Direct Map

- All physical memory is directly mapped in kernel virtual memory space
- Basis for `phys2virt` and `virt2phys` macros
- Used primarily for drivers and ‘mm’ functions
- This makes memory manipulation code small, fast and efficient
- This is also a big security risk!

Start	End	Size	Description
`ffff888000000000`	`ffffc87fffffffff`	64 TB	Direct Map

Background

Virtual Memory

CPU Pipelining

Linux Memory Management

The Exploit

The Fix

The Damage

The Exploit

- We want to read kernel memory - how?

Two conditions must hold

1. Mapping of physical page in our virtual address space
2. Permission bit to allow unprivileged access to page

The Exploit

- We want to read kernel memory - how?

Two conditions must hold

1. Mapping of physical page in our virtual address space ✓
2. Permission bit to allow unprivileged access to page

The Exploit

- We want to read kernel memory - how?

Two conditions must hold

1. Mapping of physical page in our virtual address space ✓
2. Permission bit to allow unprivileged access to page ✗

The Exploit in C

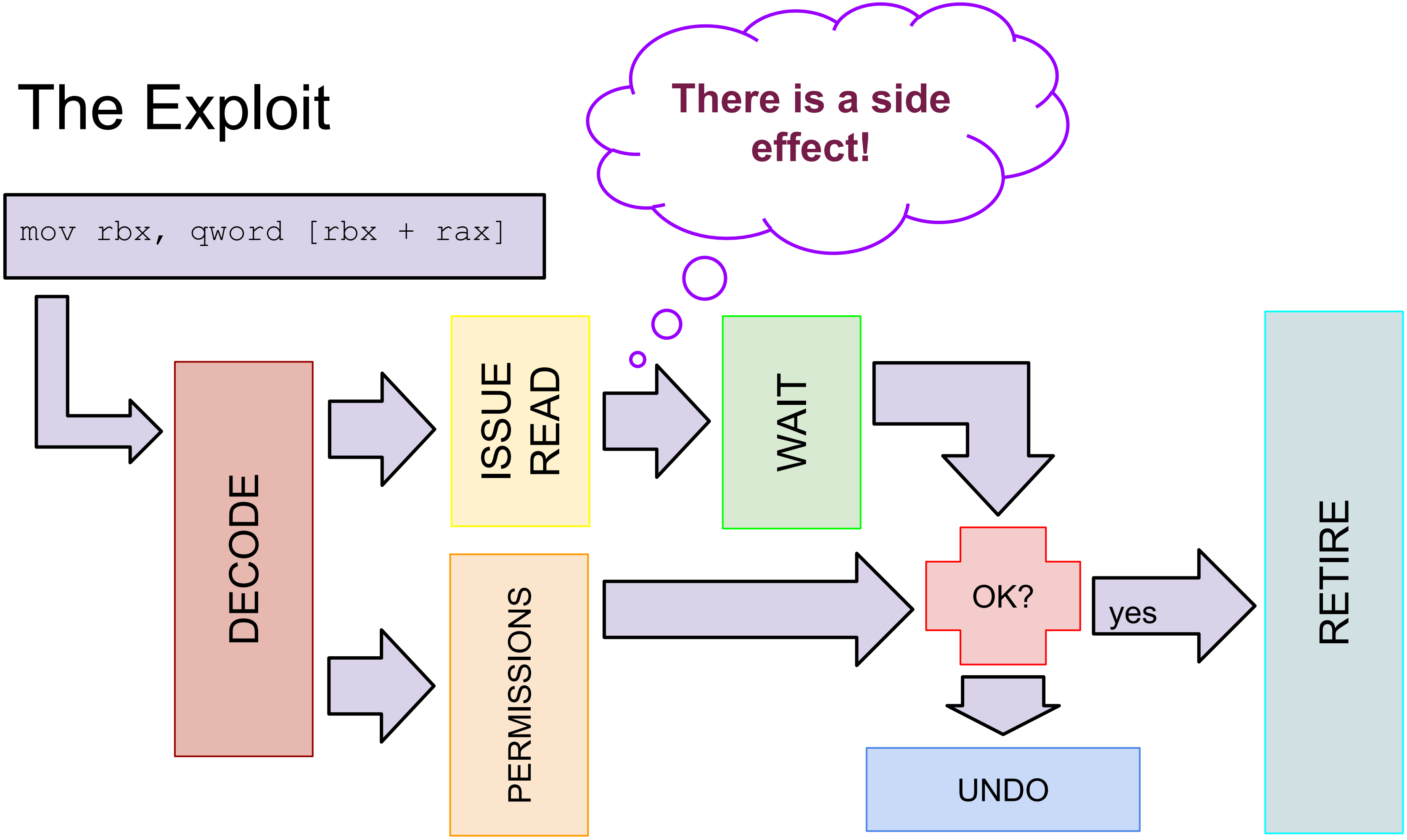
We could do this in C, but it is clearer in assembly language

```
unsigned long rax = 0;
char probe[4096 * 256];          the probe array
unsigned long rcx = 0xffff800000000000;  pointer to a kernel address
char val;
while (!rax) {
    rax = *(byte*)rcx;           (no permission!)
    rax <<= 12;                  shift the secret value by the page size
    val = probe[rax];            secret value becomes index into probe array
}
```

The Exploit

`xor rax, rax`	
`retry:`	
`mov al, byte [rcx]`	rcx is a pointer to a kernel address (no permission!)
`shl rax, 0xc`	shift the secret value by the page size secret value becomes index into probe array
`jz retry`	
`mov rbx, qword [rbx + rax]`	rbx is the base address of the probe array

The Exploit


There is a side effect!
mov rbx, qword [rbx + rax]
DECODE
ISSUE READ
WAIT
PERMISSIONS
OK?
yes
UNDO
RETIRE

The Exploit

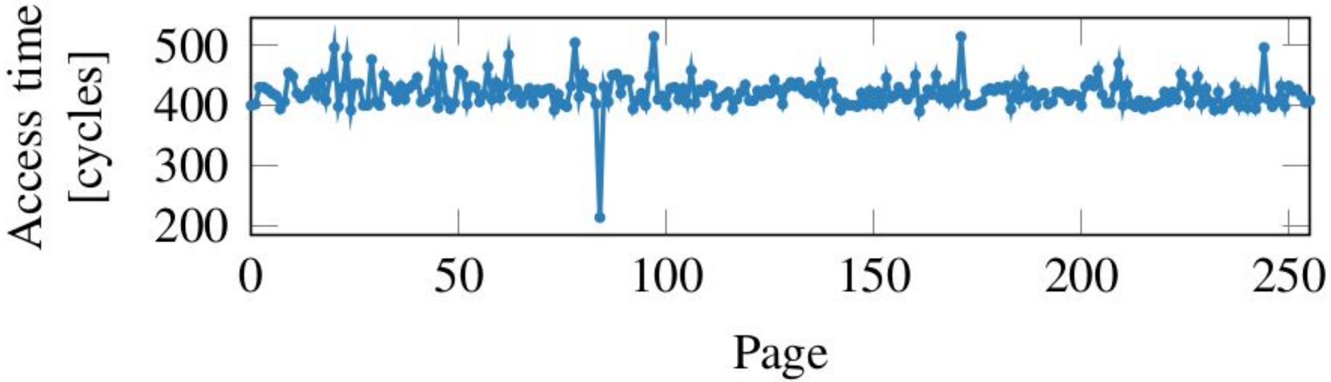
```
xor rax, rax

retry:

mov al, byte [rcx]
```
Exception!

```
shl rax, 0xc

jz retry

mov rbx, qword [rbx + rax]
```

Already scheduled and perhaps executed

Flush + Reload

- Make sure the cache is empty (clflush)
- Perform attack
- Read all entries in the probe array, and measure access time
- One measurement might stand out!
- Index of cached page is the value of the secret byte



Accessing All Memory

- Now we know how to access kernel memory!
    - Not very fast, but it works
- But how to access memory of another process?
    - Linux manages all processes (including their hierarchy) in a linked list
    - The head of this task list is stored in the init_task structure
- Use the direct memory map
    - Must find the page tables belonging to another process
    - Perform a page walk to find the physical page for a particular virtual address
    - Access that physical page through the direct map

Performance

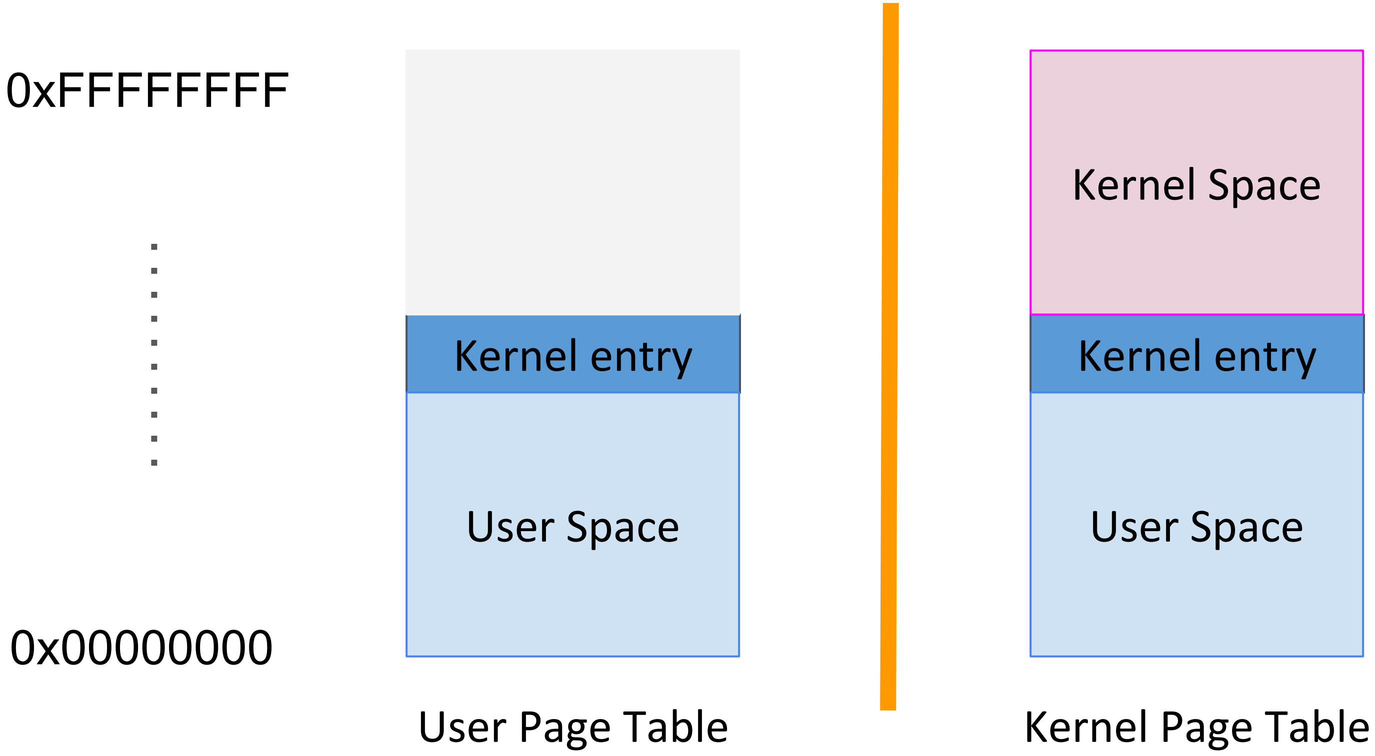
- Flush-Reload is the bottleneck of the attack
- Instead of 8 bits (=256 entries), send 1 bit (=2 entries) of information
  - Much faster
  - Less reliable (noise bias to ‘0’)
- Can read memory at rates between 4KB/s - 500KB/s

Background
Virtual Memory
CPU Pipelining
Linux Memory Management
The Exploit
The Fix
The Damage

The Fix

- KPTI - Kernel Page Table Isolation
- Based on KAISER patches
- Removes kernel mappings from user process virtual memory
- Requires a pair of page tables for each process
    - One for user space
    - One for kernel space
- Drastically increases overhead during context switch

Linux Virtual Address Space Layout (with KPTI)



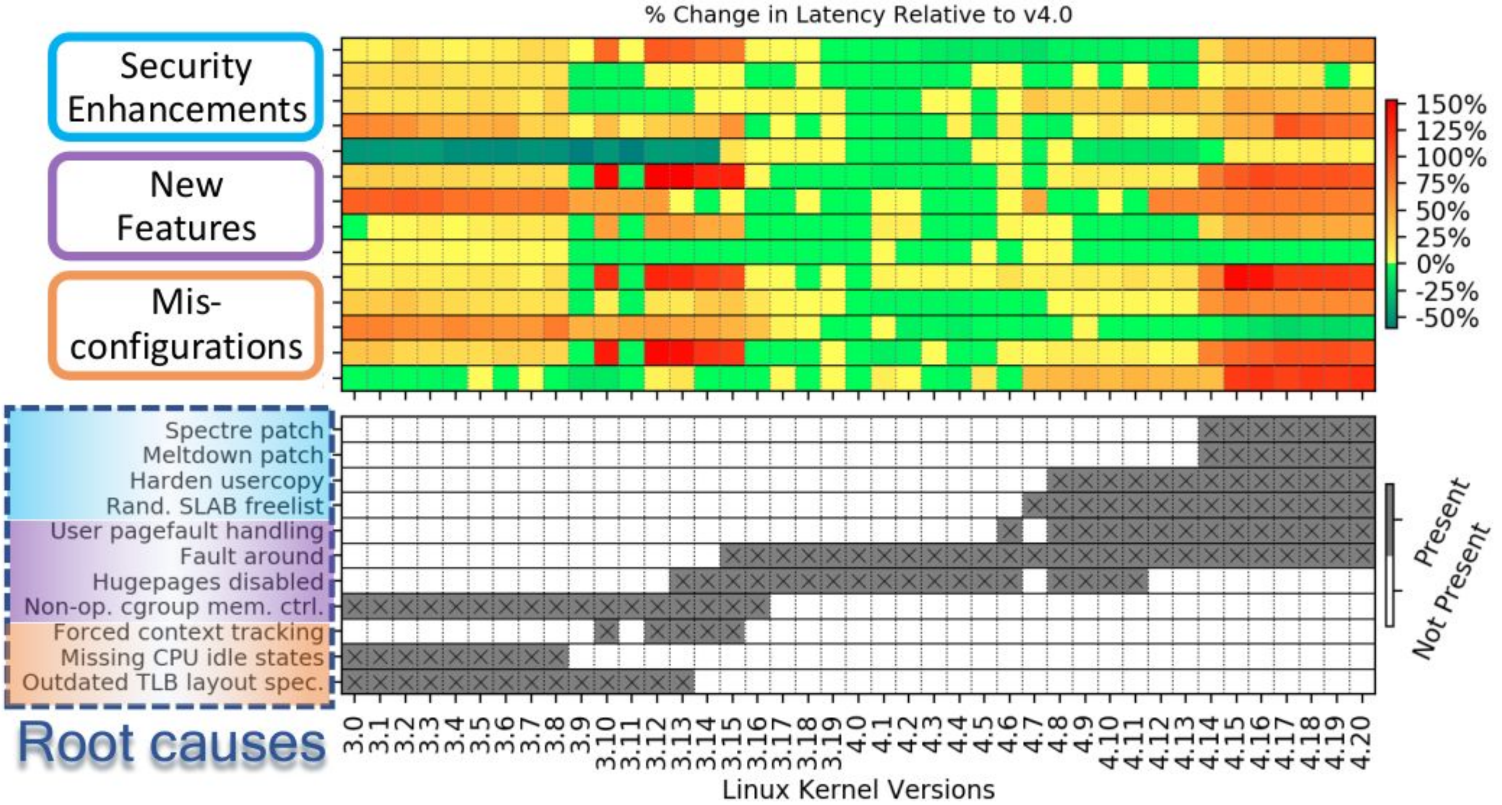
Background
Virtual Memory
CPU Pipelining
Linux Memory Management
The Exploit
The Fix
The Damage

The Damage

- Measurements are very dependent on the number of syscalls
- The overhead was measured to be 0.28% according to KAISER's original authors
- a Linux developer measured it to be roughly 5% for most workloads and up to 30% in some cases
- for database engine PostgreSQL the impact on read-only tests on an Intel Skylake processor was 16–23% (without PCID)
- Redis slowed by 6–7%

- Linux kernel compilation slowed down by 5% on Haswell

The Damage


% Change in Latency Relative to v4.0
Security Enhancements
New Features
Mis-configurations
150%
125%
100%
75%
50%
25%
0%
-25%
-50%
Spectre patch
Meltdown patch
Harden usercopy
Rand. SLAB freelist
User pagefault handling
Fault around
Hugepages disabled
Non-op. cgroup mem. ctrl.
Forced context tracking
Missing CPU idle states
Outdated TLB layout spec.
Root causes
Present
Not Present
3.0 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 3.10 3.11 3.12 3.13 3.14 3.15 3.16 3.17 3.18 3.19 4.0 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9 4.10 4.11 4.12 4.13 4.14 4.15 4.16 4.17 4.18 4.19 4.20
Linux Kernel Versions

Making It Hurt Slightly Less

- PCIDs allow a logical processor to cache information for multiple linear-address spaces
- Allows us to bypass the TLB flush on syscall entry/exit

- PostgreSQL read-only tests on an Intel Skylake processor was 7–17% (or 16–23% without PCID)

Conclusions

- Even the most commonly used, professionally made chips have bugs
- Operating systems can be used to mask these bugs
- Even so, the bugs are costly!

Meltdown

Spectre

L1TF

RIDL

Fallout

More??

References

https://sosp19.rcs.uwaterloo.ca/slides/ren.pdf

https://meltdownattack.com/meltdown.pdf

https://www.kernel.org/doc/Documentation/x86/x86_64/mm.txt

https://en.wikipedia.org/wiki/Kernel_page-table_isolation