百度安全 有AI更安全 Baidu Security Advanced AI, Stronger Security

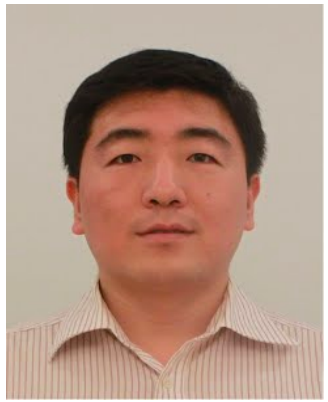
Mission Impossible
Steal Kernel Data from User Space

Yueqiang Cheng, Zhaofeng Chen, Yulong Zhang, Yu Ding, Tao Wei

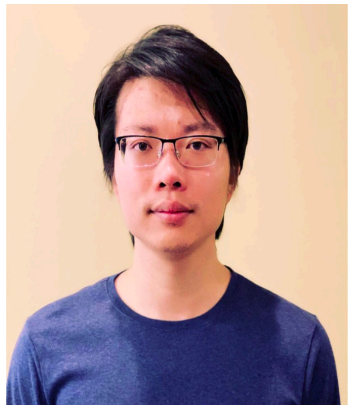
Baidu Security

About Speakers

Baidu X-Lab



Dr. Yueqiang Cheng



Mr. Zhaofeng Chen



Mr. Yulong Zhang



Dr. Yu Ding



Dr. Tao Wei

Our Security Projects:

MesaLock MesaLink MesaPy MesaTEE MesaArmor

How to Read Unauthorized Kernel Data From User Space?

Why is Hard?

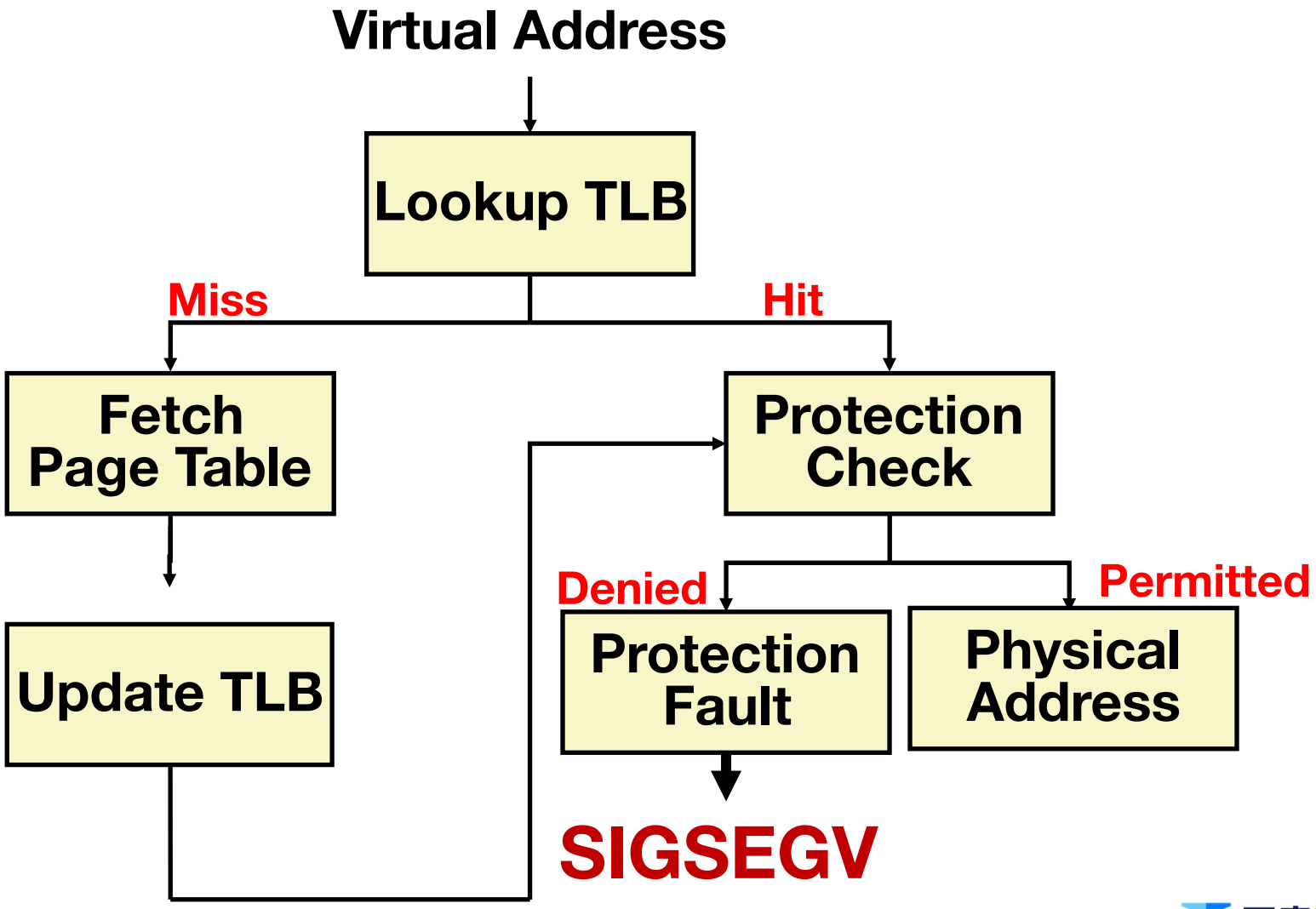
Strong Kernel-User Isolation (KUI)

Enforced by MMU via Page Table

Assume Kernel has NO implementation bug:

No kernel vulnerability to arbitrarily read kernel data

Memory Access in KUI



Permission Checkings

1: Page Table Permissions

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	
Address of page directory[1]																				Ignored							PCD	PWT	Ignored			CR3
Bits 31:22 of address of 4MB page frame										Reserved (must be 0)						Bits 39:32 of address[2]			PAT	Ignored			G	1	D	A	PCD	PWT	U/S	R/W	1	PDE: 4MB page
Address of page table																				Ignored				0	Ign	A	PCD	PWT	U/S	R/W	1	PDE: page table
Ignored																															0	PDE: not present
Address of 4KB page frame																				Ignored			G	PAT	D	A	PCD	PWT	U/S	R/W	1	PTE: 4KB page
Ignored																															0	PTE: not present

2: Control Registers, e.g., SMAP in CR4

Image from Intel sdm

百度安全 有AI更安全 Baidu Security Advanced AI, Stronger Security

No Way to Go?



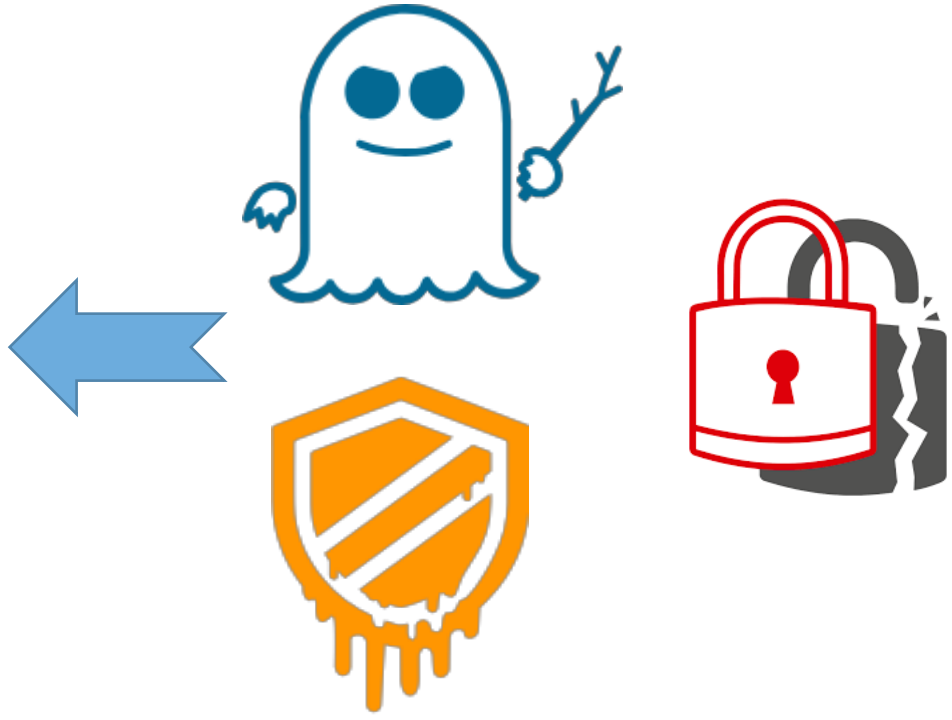
1.Unprivileged App +
2.KUI Permission Checking +
3.Bug-free Kernel

However, in order to gain high performance, CPU …

1.Unprivileged App +

2.Permission Checking +

3.Bug-free Kernel





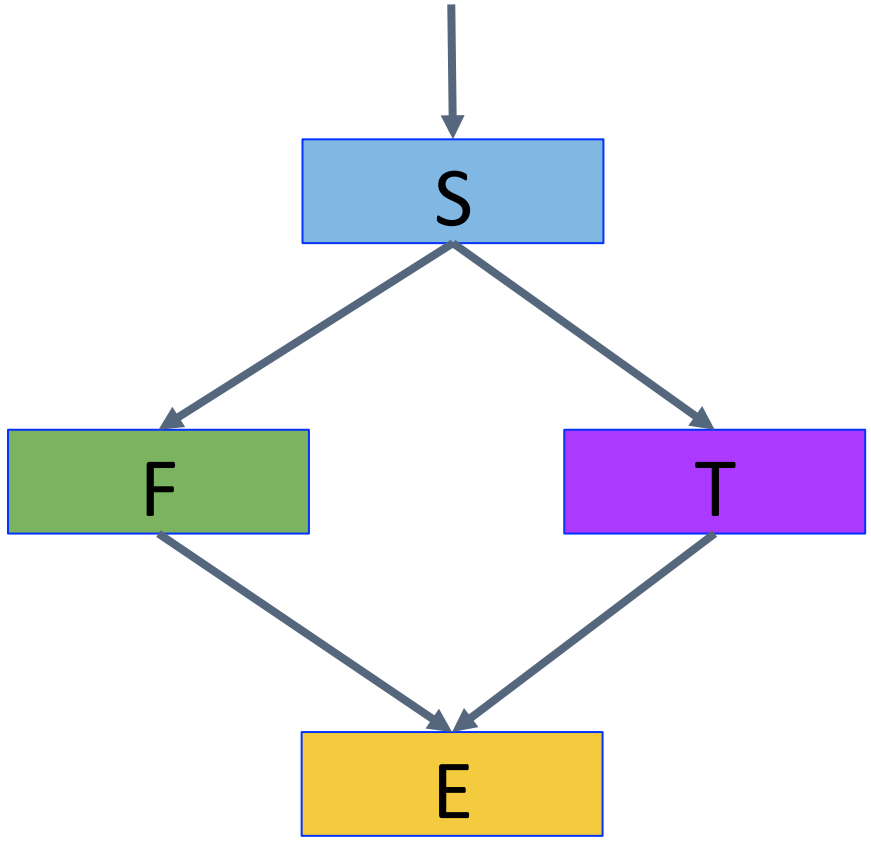




Microarchitecture

Speculative Execution + Out-of-order Execution

Speculative Execution



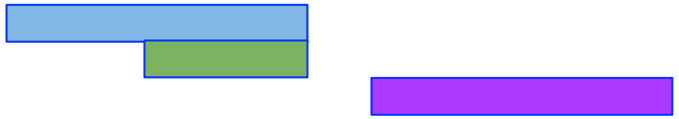
No Speculative Execution



Correct Prediction



Misprediction

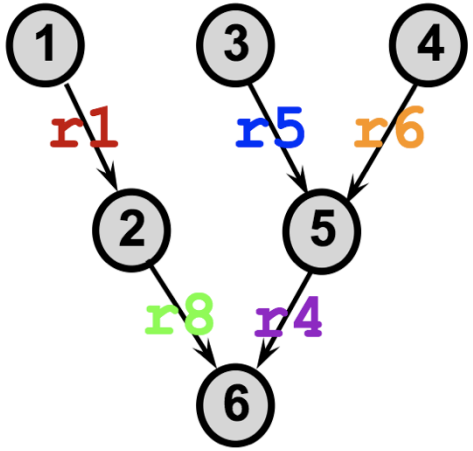


Out-of-order Execution

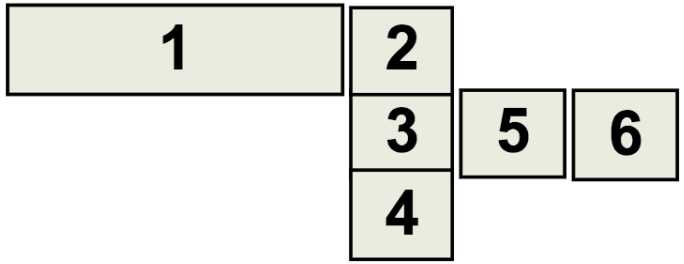
Example:

```
(1) r1 ← r4 / r7
(2) r8 ← r1 + r2
(3) r5 ← r5 + 1
(4) r6 ← r6 - r3
(5) r4 ← r5 + r6
(6) r7 ← r8 * r4
```

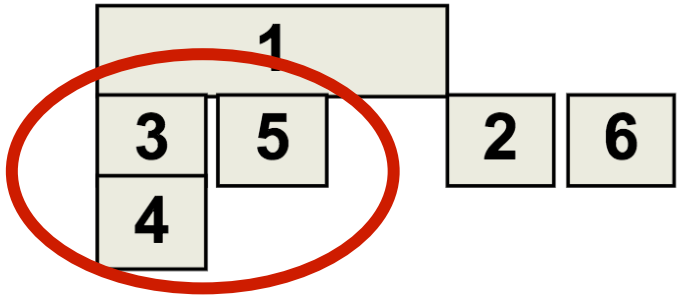
Data Flow Graph



In-order execution



Out-of-order execution



Images are from Dr. Lihu Rappoport

百度安全 有AI更安全 Baidu Security Advanced AI, Stronger Security

Speculative Execution + Out-of-order Execution

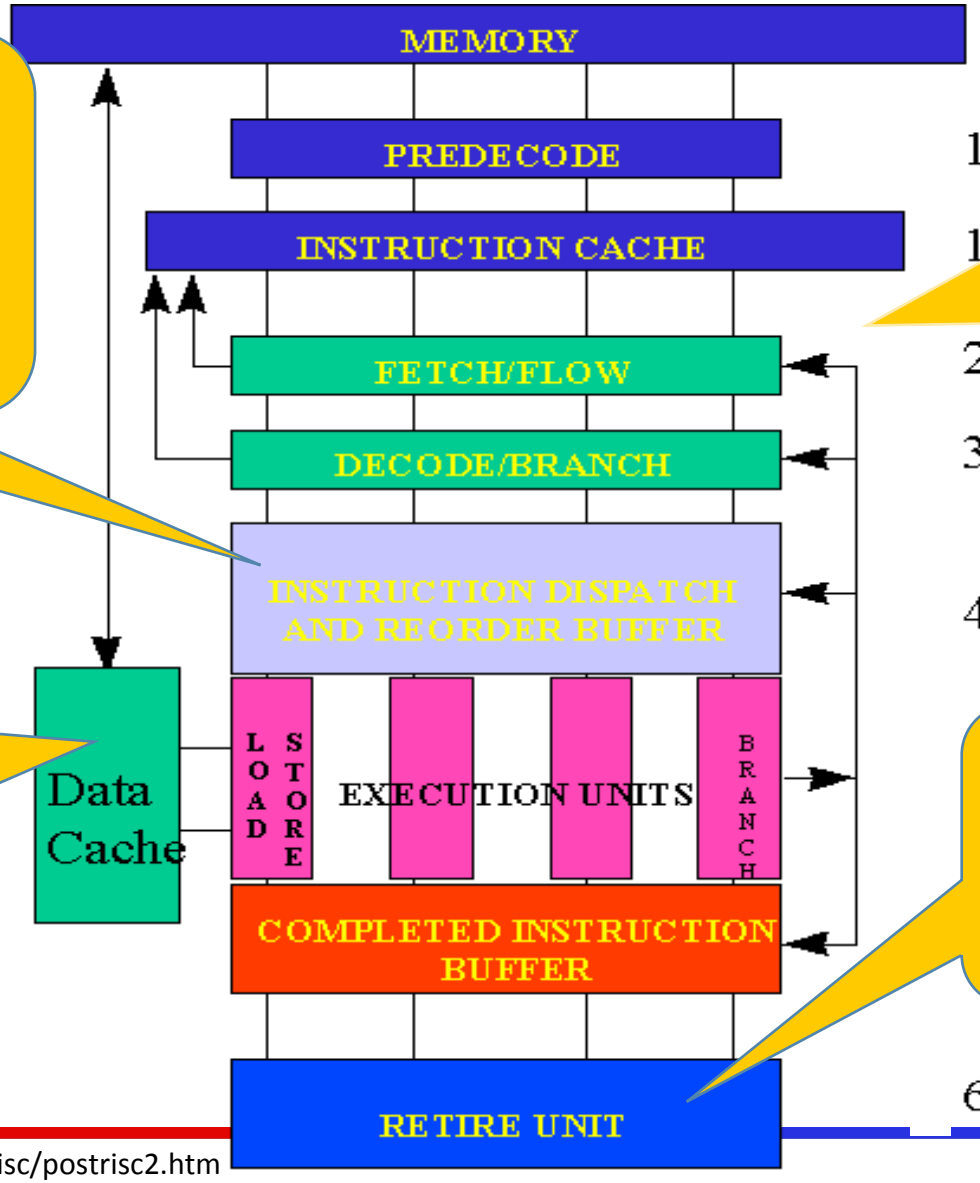
Enough?

Not Enough !!!

Delayed Permission Checking + Cache Side Effects

Execution Engine executes in a out-of-order way

Branch Predictor in Front End Serving Speculative Execution



Side effects in cache are still there!!!

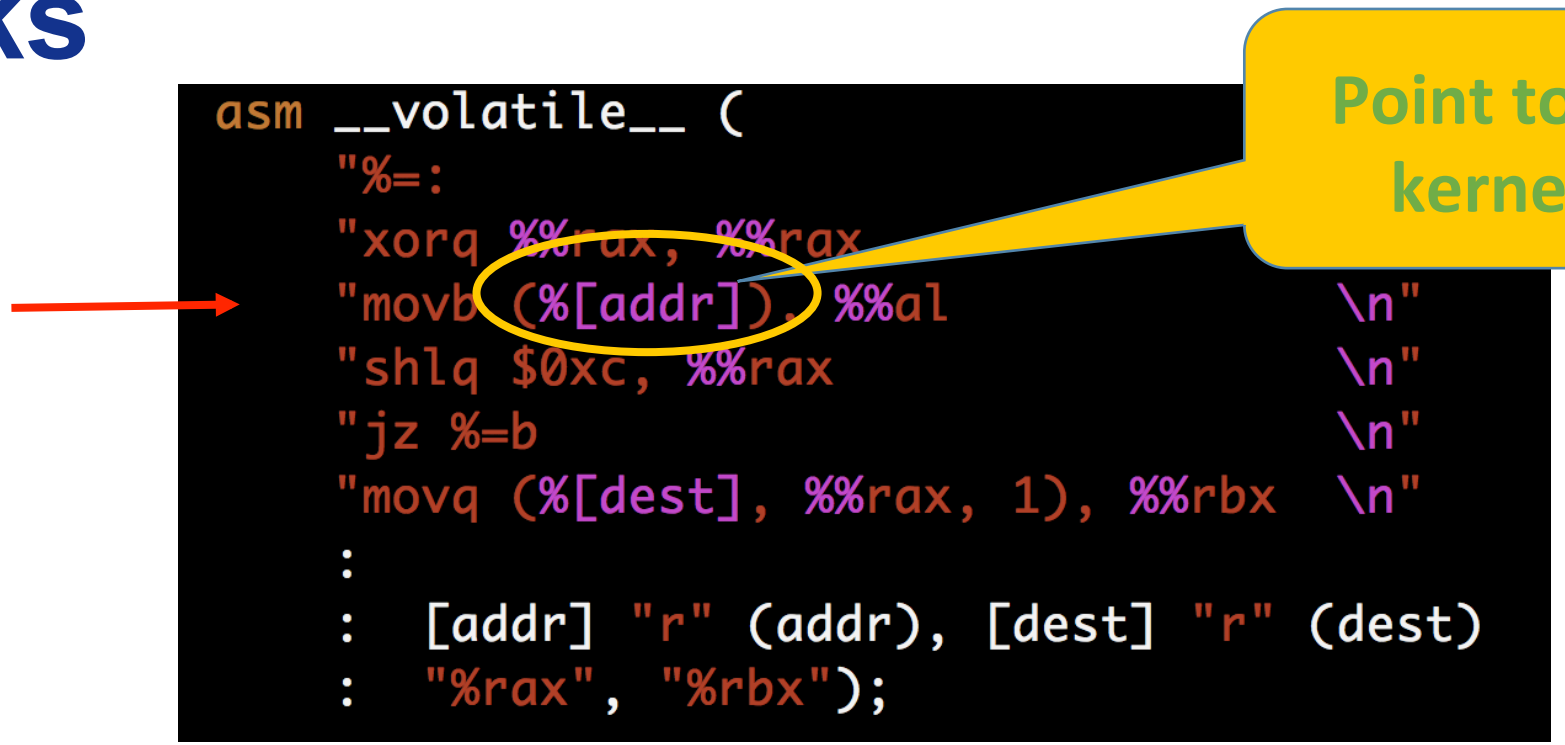
Permission checking is delayed to Retire Unit

Image from https://www.cse.msu.edu/~enbody/postrisc/postrisc2.htm

百度安全 有AI更安全 Baidu Security Advanced AI, Stronger Security

How Meltdown (v3) Works

1. The content of an attacker-chosen memory location, which is inaccessible to the attacker, is loaded into a register.



How Meltdown (v3) Works

2. A transient instruction accesses a cache line based on the secret content of the register.

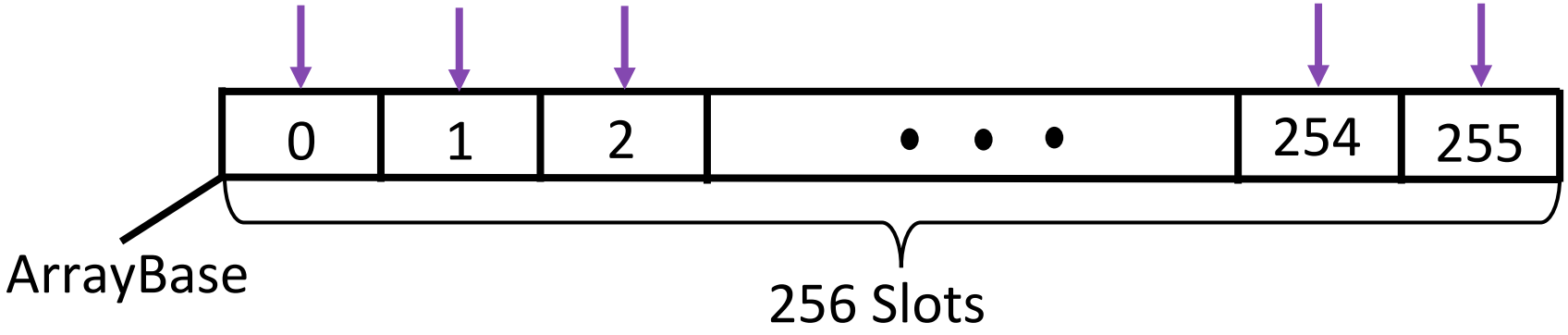
Bring data into cache

This number should >= 0x6

```
asm __volatile__ (
    "%=:                                    \n"
    "xorq %%rax, %%rax                      \n"
    "movb (%[addr]), %%al                   \n"
    "shlq $0xc, %%rax                       \n"
    "jz %=b                                 \n"
    "movq (%[dest], %%rax, 1), %%rbx        \n"
    :
    : [addr] "r" (addr), [dest] "r" (dest)
    : "%rax", "%rbx");
```

How Meltdown (v3) Works

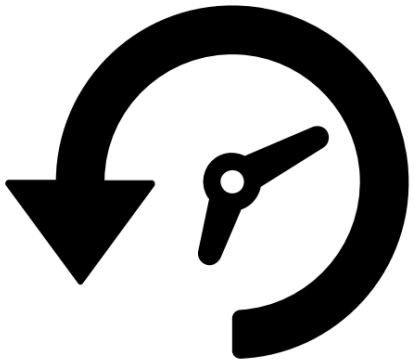
3. The attacker uses Flush+Reload to determine the accessed cache line and hence the secret stored at the chosen memory location.



The selected index is the value of the target byte

e.g., if the selected index is **0x65**, the value is **'A'**

ForeShadow Attack



Put secrets in L1



Unmap Page Table Entry



Meltdown

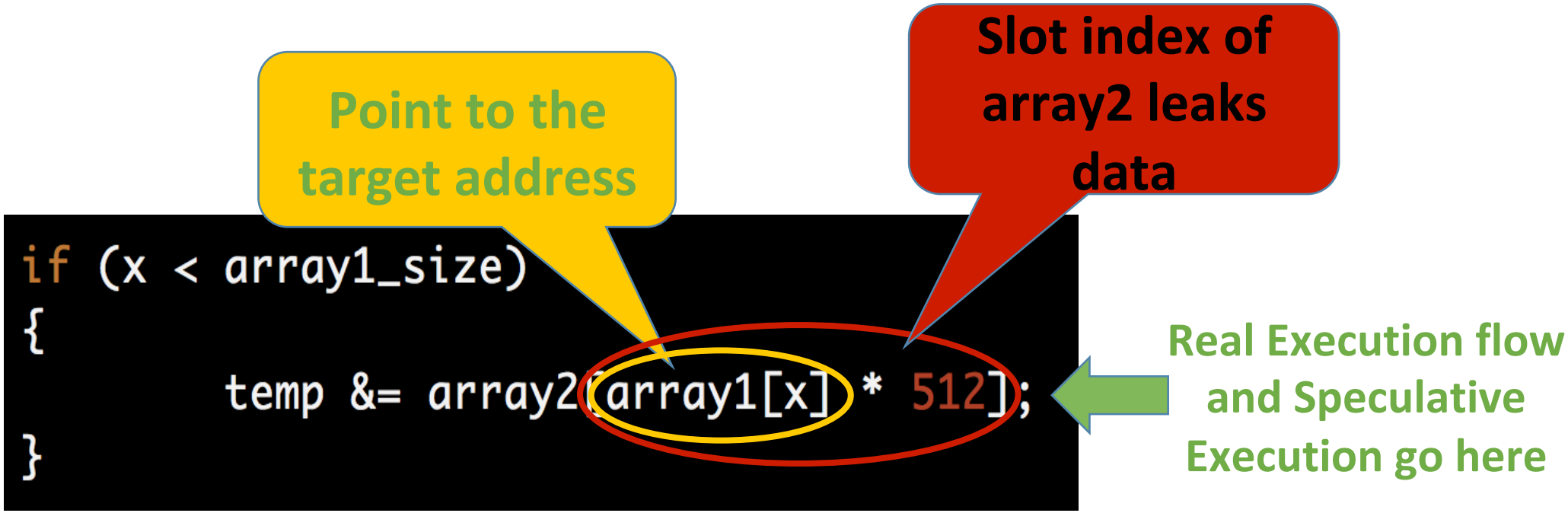


How about Spectre (v1/v2)?

How Spectre (v1) Works

1. The setup phase, in which the processor is mistrained to make "an exploitable erroneous speculative prediction."

e.g., x < array1_size



How Spectre (v1) Works

2. The processor speculatively executes instructions from the target context into a microarchitectural covert channel.

e.g., x > array1_size

Execution flow should go here

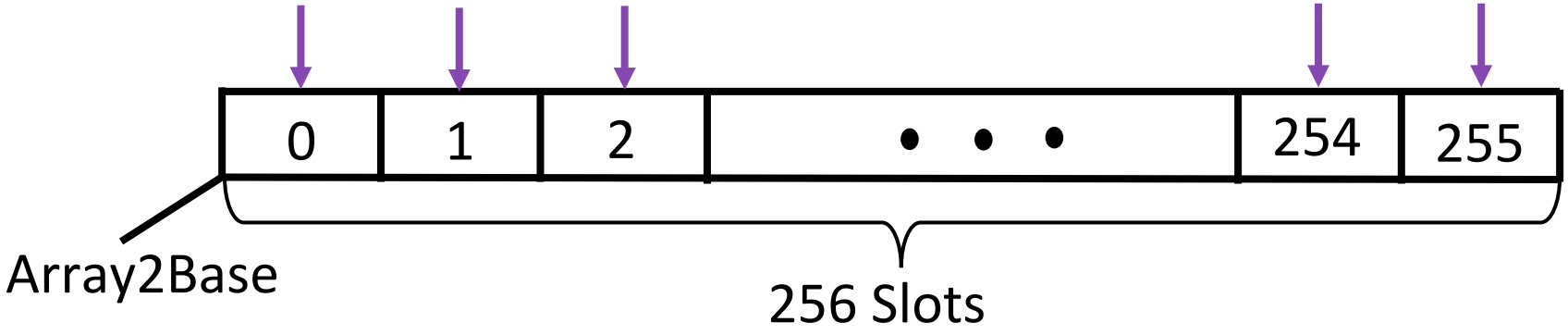
```
if (x < array1_size)
{
        temp &= array2[array1[x] * 512];
}
```

Speculative Execution goes here!

A slot of array2 is loaded into cache

How Spectre Works

3: The sensitive data is recovered. This can be done by timing access to memory addresses in the CPU cache.



The selected index is the value of the target byte

e.g., if the selected index is **0x66**, the value is **'B'**

How Spectre Read Kernel Data

- ✓ **array1** and **array2** are in user-space
- ✓ **x** is controlled by the adversary

array1+x points to secret

Slot index of array2 leaks kernel data

```
if (x < array1_size)
{
        temp &= array2[array1[x] * 512];
}
```

Happy! We Get Kernel Data Now



1.Unprivile[illegible]pp +

2.Permissi[illegible]hecking +

3.Bug-free Kernel

However...



KPTI

Meltdown

Spectre (Gadget in User Space)

FAILED

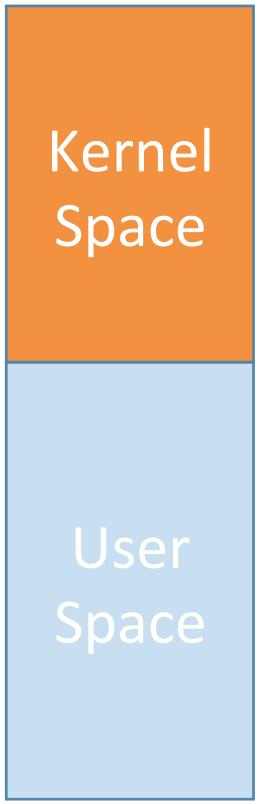


SMAP

Spectre (Gadget in Kernel Space)

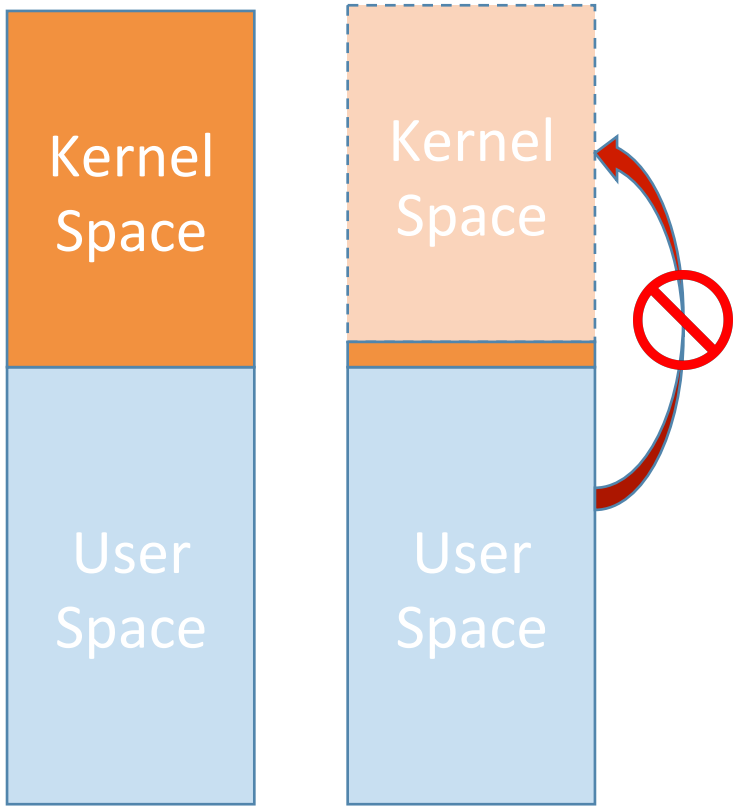
FAILED

KPTI



User/kernel mode

Before KPTI



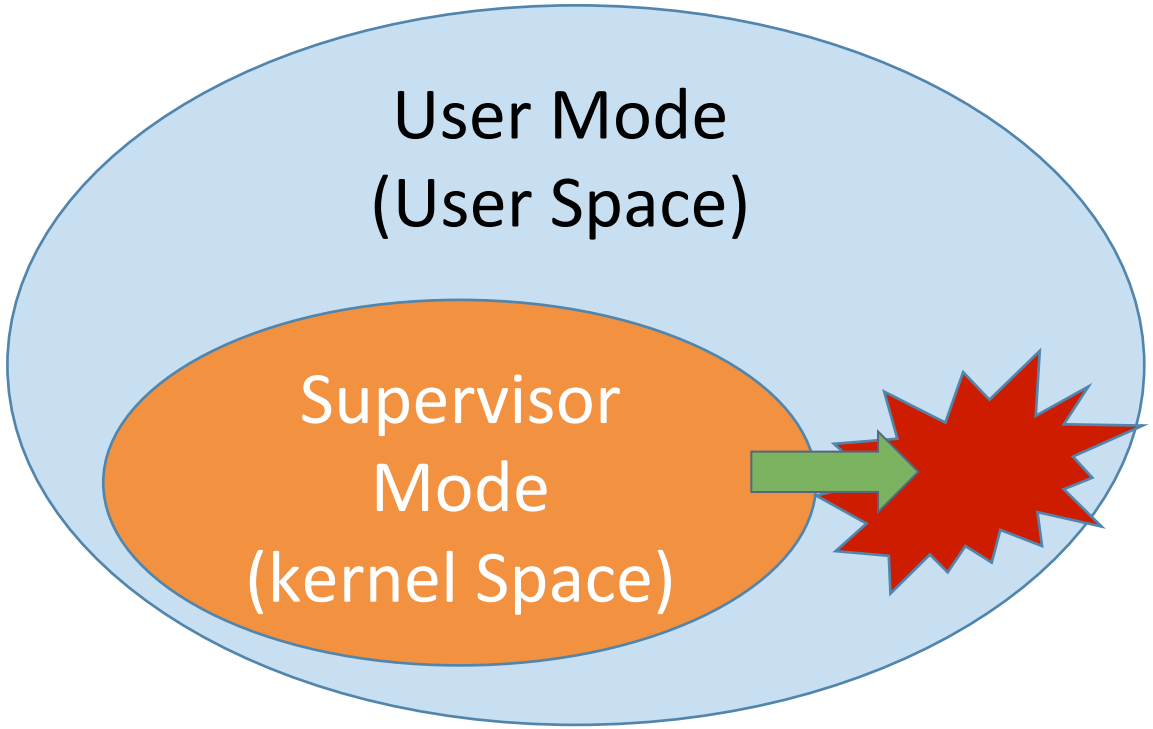
kernel mode

User mode

After KPTI

PCID helps performance

SMAP



- ✓ SMAP is enabled when the SMAP bit in the CR4 is set
- ✓ SMAP can be temporarily disabled by setting the EFLAGS.AC flag
- ✓ **SMAP checking is done long before retirement or even execution**

Even we put the Spectre gadget into the kernel space, SMAP will stop it

百度安全 有AI更安全
Baidu Security Advanced AI, Stronger Security

Attack and Mitigation Summary

Techniques	Steal Kernel Data?	Mitigations	After Mitigation, kern. Data Leakage?
Spectre	Yes	KPTI + SMAP	NO
Meltdown	Yes	KPTI	NO
ForeShadow	Yes	KPTI	NO

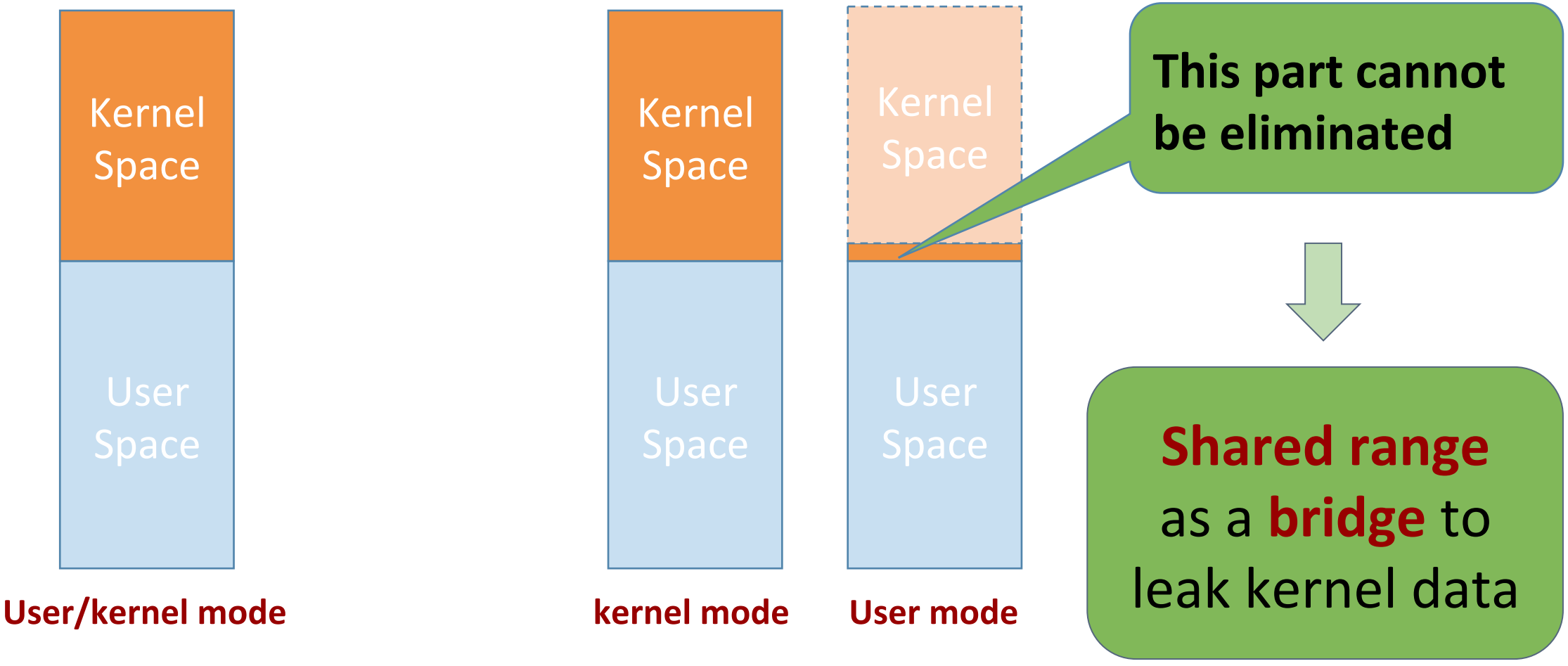
Only for Kernel Data Leakage. For other aspects, the summary is not included here.

Despair...



Image from http://nohopefor.us/credits

Hope in Despair



New Variant Meltdown v3z

Breaking SMAP + KPTI + user-kernel Isolation

1: Use **new gadget** to build data-dependence between target kernel data and the bridge (bypass SMAP)

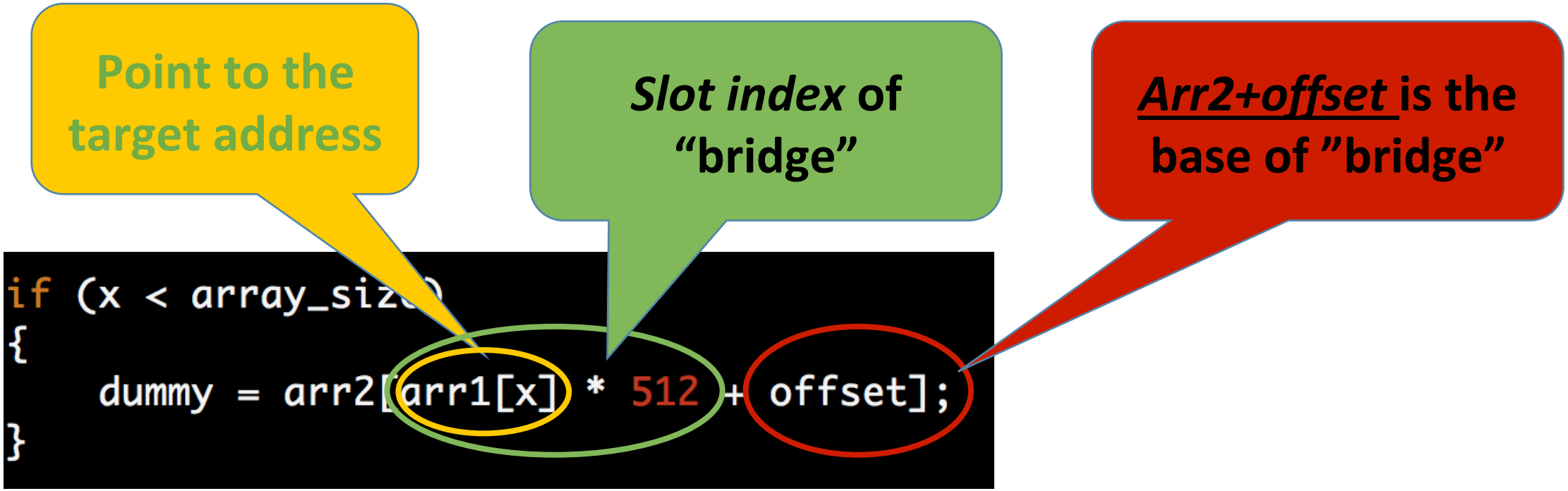
2: Use **Reliable Meltdown** to probe bridge to leak kernel data

(bypass KPTI and KUI)



1st Step: Trigger New Gadget

Similar to Spectre gadget, but not exact the same



x **and** ***offset*** **should be controlled by the adversary!!**

How to Trigger the New Gadget

There are many sources to trigger the new gadget

1: Syscalls
2: /proc and /sys etc. interfaces
3: Interrupt and exception handlers
4: eBPF
5: …

How to Find the New Gadget

Source Code Scanning

We use smatch for Linux Kernel 4.17.3,

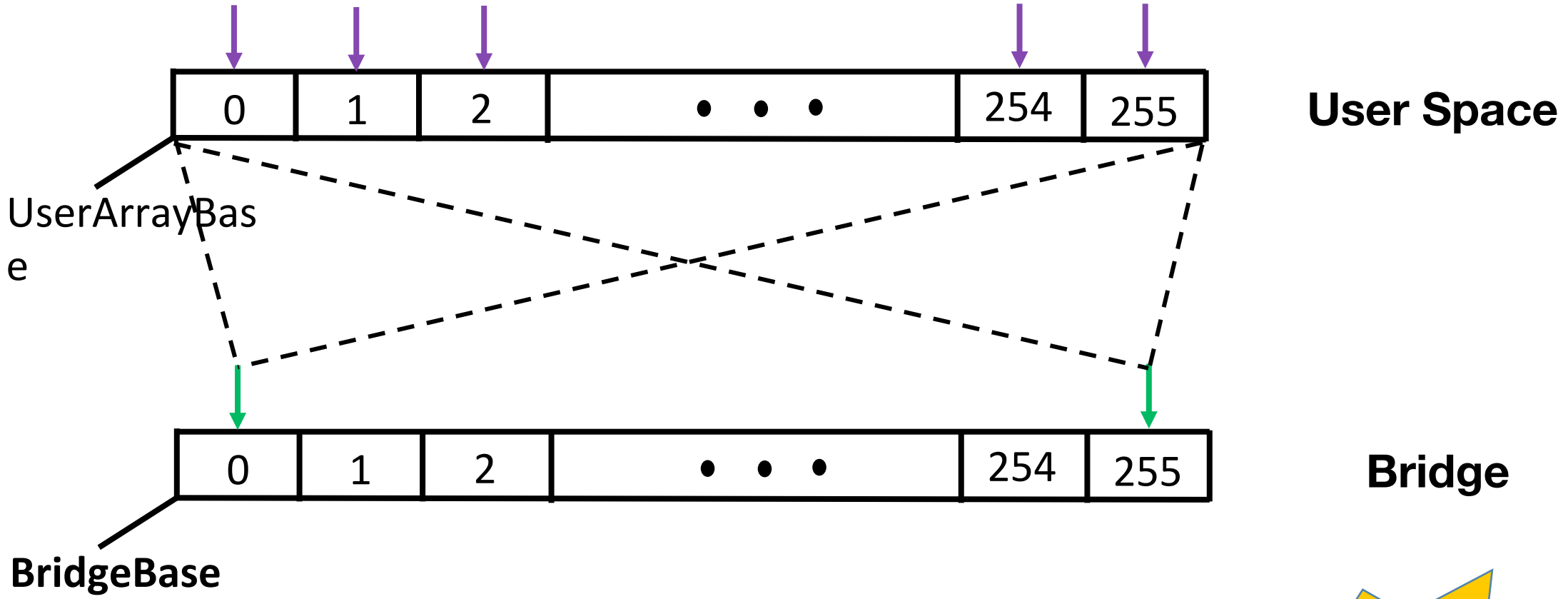
- ➢ Default config: 36 gadget candidates
- ➢ Allyes config: 166 gadget candidates

However, there are many restrictions to the gadget in real exploits

- ✓ Offset range
- ✓ Controllable invocation
- ✓ Cache noise
- ✓ …

Binary Code Scanning??

2nd Step: Probe Bridge

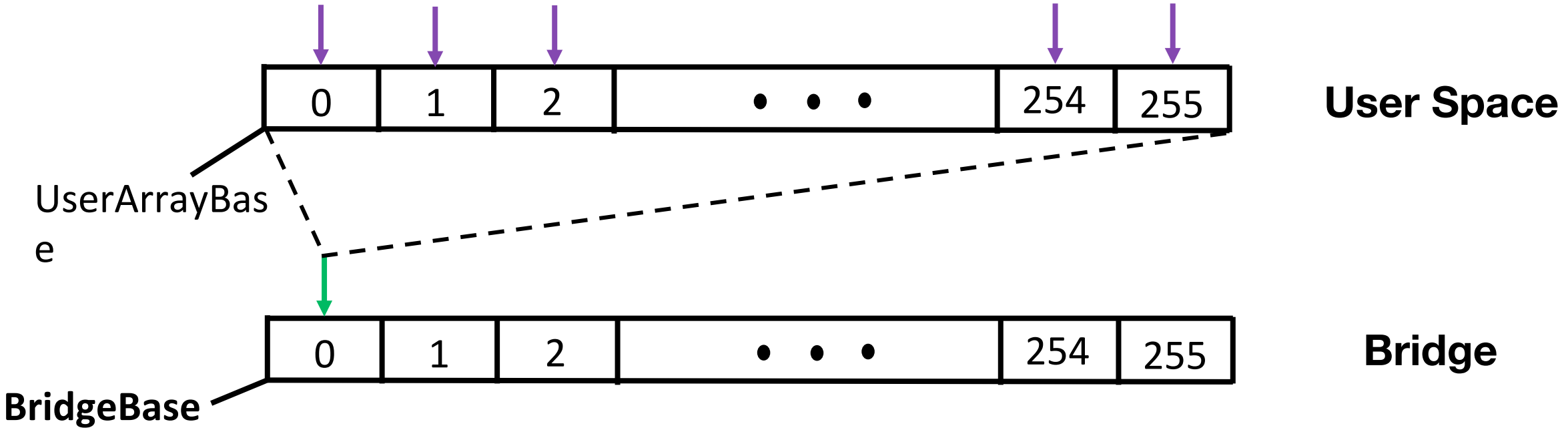


Obviously, in each round there are **(256*256) probes**
To make the result reliable, usually we need **multiple rounds**

Inefficient

Baidu Security
Advanced AI , Stronger Security

Make it Practical/Efficient



Why do we need to probe 256 times in Meltdown?

If we know the value of the slot 0 of the BridgeBase, we probe it only **once**.

Can we know the values in advance?

No for Meltdown (v3)

Meltdown is able to read kernel data.
But, it requires that the target **data** is in the **CPU L1d cache.**

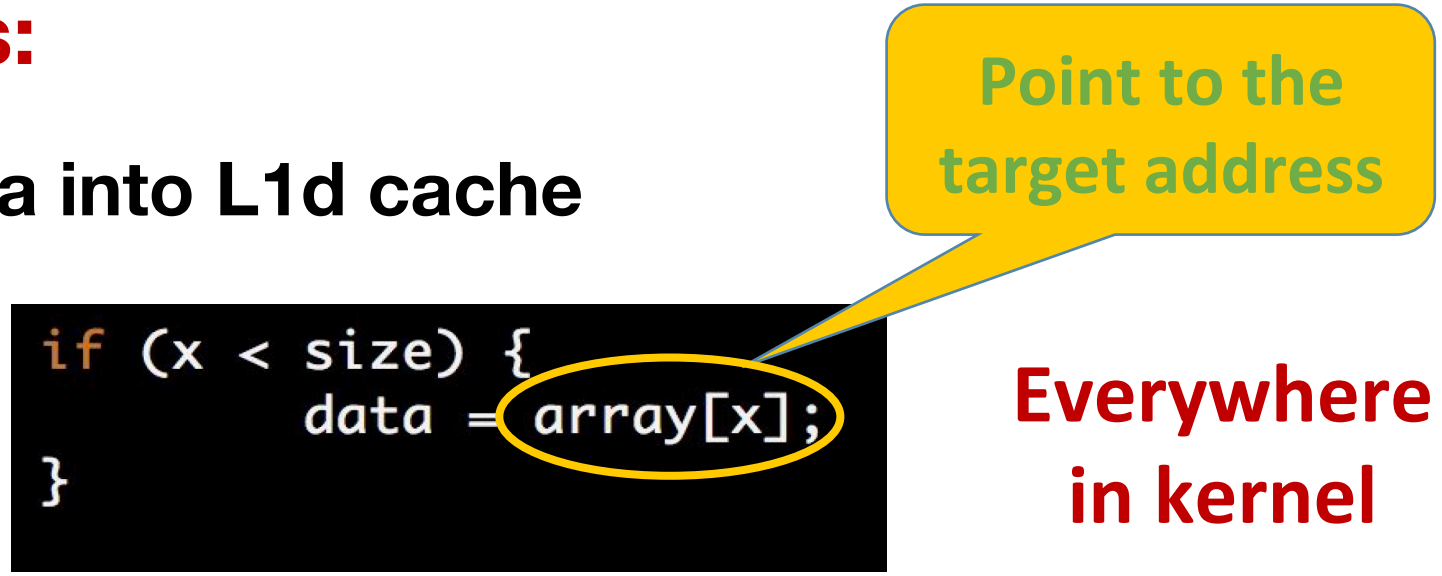
If the target data is **NOT** in L1d cache, **0x00** returns.

We need reliably reading kernel data!

Reliable Meltdown (V3r)

V3r has two steps:

1st step: bring data into L1d cache



Everywhere in kernel

2nd step: use v3 getting data

We test it on Linux 4.4.0 with Intel CPU E3-1280 v6, and MacOS 10.12.6 (16G1036) with Intel CPU i7-4870HQ

Put Everything Together

Offline phase:

- Use v3r dumping bridge data, and save them into a table

Online phase:

- 1st step: Build data dependence between target data and bridge slot
- 2nd step: Probe each slot of the bridge

Efficiency:

- from **several minutes** (even around 1 hour in certain cases) to only **several seconds** to leak **one byte**.

Demo Settings

Kernel: Linux 4.4.0 with SMAP + KPTI

CPU: Intel CPU E3-1280 v6

In kernel space, we have a
secret msg, e.g., **xlabsecretxlabsecret**,
location is at, e.g., **0xffffffffc0e7e0a0**

Countermeasure Discussions

Software Mitigations

- ✓ Patch kernel to eliminate all expected gadgets
- ✓ Minimize the shared “bridge” region
- ✓ Randomize the shared “bridge” region
- ✓ Monitor cache-based side channel activities

Countermeasure Discussions

Hardware Mitigations

- ✓ Do permission checking during or even execution stage
- ✓ Revise speculative execution and out-of-order execution
- ✓ Use side channel resistant cache, e.g., exclusive/random cache
- ✓ Add hardware-level side channel detection mechanism

Take Away

- **Trinational Spectre and Meltdown can NOT steal kernel data with KPTI + SMAP + KUI enabled.**

- **Our new Meltdown variants is able to break the strongest protection (KPTI + SMAP + KUI).**

- **All existing kernels need to be patched to mitigate our new attack**



Mission Impossible

Steal Kernel Data from User Space

Yueqiang Cheng

Baidu Security

Q&A image is from https://i.redd.it/wbiwgnokgig11.jpg