Virtualization and Security

Daniel Sanchez
Computer Science & Artificial Intelligence Lab
M.I.T.
Evolution in Number of Users

IBM 1620
1959

Single User

Runtime loaded with program
Evolution in Number of Users

IBM 1620 1959  IBM 360 1960s

Single User  Multiple Users

Runtime loaded with program  OS for sharing resources
Evolution in Number of Users

- **IBM 1620**
  - **1959**
  - Single User
  - Runtime loaded with program

- **IBM 360**
  - **1960s**
  - Multiple Users
  - OS for sharing resources

- **IBM PC**
  - **1980s**
  - Single User
  - OS for sharing resources
Evolution in Number of Users

<table>
<thead>
<tr>
<th>Year</th>
<th>Device</th>
<th>User Type</th>
<th>OS for Sharing</th>
<th>OS for Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>1959</td>
<td>IBM 1620</td>
<td>Single User</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Runtime loaded</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>with program</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1960s</td>
<td>IBM 360</td>
<td>Multiple Users</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>OS for</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>sharing</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>resources</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1980s</td>
<td>IBM PC</td>
<td>Single User</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>OS for</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>sharing</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>resources</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990s</td>
<td>Cloud Servers</td>
<td>Multiple Users</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multiple OSs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Single-Program Machine

- Hardware executes a single program
- This program has direct and complete access to all hardware resources in the machine
Single-Program Machine

- Hardware executes a single program
- This program has direct and complete access to all hardware resources in the machine
Single-Program Machine

- Hardware executes a single program
- This program has direct and complete access to all hardware resources in the machine
- The instruction set architecture (ISA) is the interface between software and hardware
Operating System (OS) goals:
- Protection and privacy: Processes cannot access each other’s data
• Operating System (OS) goals:
  – **Protection and privacy**: Processes cannot access each other’s data
  – **Abstraction**: OS hides details of underlying hardware
    • e.g., processes open and access files instead of issuing raw commands to the disk
Operating System (OS) goals:

- **Protection and privacy**: Processes cannot access each other’s data
- **Abstraction**: OS hides details of underlying hardware
  - e.g., processes open and access files instead of issuing raw commands to the disk
- **Resource management**: OS controls how processes share hardware (CPU, memory, disk, etc.)
Operating System Mechanisms

• The OS kernel provides a **private address space** to each process
  – Each process is allocated space in physical memory by the OS
  – A process is not allowed to access the memory of other processes
The OS kernel provides a private address space to each process
- Each process is allocated space in physical memory by the OS
- A process is not allowed to access the memory of other processes
Operating System Mechanisms

- The OS kernel provides a **private address space** to each process
  - Each process is allocated space in physical memory by the OS
  - A process is not allowed to access the memory of other processes
- The OS kernel **schedules processes** into cores
  - Each process is given a fraction of CPU time
  - A process cannot use more CPU time than allowed
Operating System Mechanisms

• The OS kernel provides a **private address space** to each process
  – Each process is allocated space in physical memory by the OS
  – A process is not allowed to access the memory of other processes

• The OS kernel **schedules processes** into cores
  – Each process is given a fraction of CPU time
  – A process cannot use more CPU time than allowed

[Diagram of physical memory with sections for OS Kernel, Process 1, Process 2, and free memory.]
Operating System Mechanisms

- The OS kernel provides a **private address space** to each process
  - Each process is allocated space in physical memory by the OS
  - A process is not allowed to access the memory of other processes
- The OS kernel **schedules processes** into cores
  - Each process is given a fraction of CPU time
  - A process cannot use more CPU time than allowed
- The OS kernel lets processes invoke system services (e.g., access files or network sockets) via **system calls**
Virtual Machines

- The OS gives a Virtual Machine (VM) to each process
  - Each process believes it runs on its own machine...
  - ...but this machine does not exist in physical hardware
Virtual Machines

- The OS gives a **Virtual Machine (VM)** to each process
  - Each process believes it runs on its own machine...
  - ...but this machine does not exist in physical hardware
Virtual Machines

- The OS gives a **Virtual Machine (VM)** to each process
  - Each process believes it runs on its own machine...
  - ...but this machine does not exist in physical hardware
Virtual Machines

• The OS gives a **Virtual Machine (VM)** to each process
  – Each process believes it runs on its own machine...
  – ...but this machine does not exist in physical hardware

![Diagram showing OS Kernel, Virtual Machines, and Physical Hardware]

- ABI
- Virtual CPUs
- Virtual Memory
- Events
- Files
- Sockets
- Syscalls

OS Kernel (specially privileged process)

Physical Hardware

- Processor Memory
- Disk
- Network card
- Display
- Keyboard
Virtual Machines

- The OS gives a **Virtual Machine (VM)** to each process
  - Each process believes it runs on its own machine...
  - ...but this machine does not exist in physical hardware

![Diagram](image.png)
Virtual Machines

• A Virtual Machine (VM) is an *emulation* of a computer system
  – Very general concept, used beyond operating systems

OS Kernel (specially privileged process)

Physical Hardware

Processor Memory  Disk  Network card  Display  Keyboard...
Virtual Machines Are Everywhere

• Example: Consider a Python program running on a Linux Virtual Machine
Virtual Machines Are Everywhere

• Example: Consider a Python program running on a Linux Virtual Machine
Virtual Machines Are Everywhere

• Example: Consider a Python program running on a Linux Virtual Machine

Python program

Python Language
Example: Consider a Python program running on a Linux Virtual Machine

- Python program
- Python interpreter (CPython)
- Python Language
Virtual Machines Are Everywhere

- Example: Consider a Python program running on a Linux Virtual Machine

```
Python program
Python interpreter (CPython)
```

Python Language
Implements a Python VM
Virtual Machines Are Everywhere

• Example: Consider a Python program running on a Linux Virtual Machine

  Python program
  Python interpreter (CPython)
  Python Language
  Implements a Python VM
  Linux ABI
Virtual Machines Are Everywhere

- Example: Consider a Python program running on a Linux Virtual Machine

- Python program
- Python interpreter (CPython)
- Linux OS kernel
- Python Language
  Implements a Python VM
- Linux ABI
Virtual Machines Are Everywhere

- Example: Consider a Python program running on a Linux Virtual Machine

<table>
<thead>
<tr>
<th>Python program</th>
<th>Python Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python interpreter (CPython)</td>
<td>Implements a Python VM</td>
</tr>
<tr>
<td>Linux OS kernel</td>
<td>Linux ABI</td>
</tr>
</tbody>
</table>
Virtual Machines Are Everywhere

- Example: Consider a Python program running on a Linux Virtual Machine

- Python program
- Python interpreter (CPython)
- Linux OS kernel
- Python Language
  - Implements a Python VM
- Linux ABI
  - Implements a Linux-x86 VM
- x86 ISA
Virtual Machines Are Everywhere

• Example: Consider a Python program running on a Linux Virtual Machine

Python program

Python interpreter (CPython)

Linux OS kernel

VirtualBox

Python Language

Implements a Python VM

Linux ABI

Implements a Linux-x86 VM

x86 ISA
Virtual Machines Are Everywhere

• Example: Consider a Python program running on a Linux Virtual Machine

Python program
Python interpreter (CPython)
Linux OS kernel
VirtualBox

Python Language
Implements a Python VM

Linux ABI
Implements a Linux-x86 VM

x86 ISA
Implements an x86 system VM
Virtual Machines Are Everywhere

- Example: Consider a Python program running on a Linux Virtual Machine

```
Python program
Python interpreter (CPython)
Linux OS kernel
VirtualBox
```

- Python Language
  - Implements a Python VM
  - Linux ABI

- Linux-x86 VM
  - x86 ISA

- x86 ISA
  - Win/Linux/MacOS/... ABI
Virtual Machines Are Everywhere

- Example: Consider a Python program running on a Linux Virtual Machine

- Python program
- Python interpreter (CPython)
- Linux OS kernel
- VirtualBox
- OS kernel (Win/Linux/MacOS/...)

  - Python Language
    - Implements a Python VM
    - Linux ABI
  - Linux OS kernel
    - Implements a Linux-x86 VM
    - x86 ISA
  - VirtualBox
    - Implements an x86 system VM
    - Win/Linux/MacOS/... ABI
Virtual Machines Are Everywhere

- Example: Consider a Python program running on a Linux Virtual Machine

<table>
<thead>
<tr>
<th>Python program</th>
<th>Python Language</th>
<th>Implements a Python VM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python interpreter (CPython)</td>
<td>Linux ABI</td>
<td>Implements a Linux-x86 VM</td>
</tr>
<tr>
<td>Linux OS kernel</td>
<td>x86 ISA</td>
<td>Implements an x86 system VM</td>
</tr>
<tr>
<td>VirtualBox</td>
<td>Win/Linux/MacOS/... ABI</td>
<td>Implements an OS-x86 VM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OS kernel (Win/Linux/MacOS/...)</th>
<th>Linux ABI</th>
<th>Implements a Linux-x86 VM</th>
</tr>
</thead>
<tbody>
<tr>
<td>VirtualBox</td>
<td>x86 ISA</td>
<td>Implements an x86 system VM</td>
</tr>
<tr>
<td>OS kernel (Win/Linux/MacOS/...)</td>
<td>Win/Linux/MacOS/... ABI</td>
<td>Implements an OS-x86 VM</td>
</tr>
</tbody>
</table>
Virtual Machines Are Everywhere

• Example: Consider a Python program running on a Linux Virtual Machine

- Python program
- Python Language
  - Implements a Python VM
- Python interpreter (CPython)
- Linux ABI
  - Implements a Linux-x86 VM
- Linux OS kernel
- x86 ISA
  - Implements an x86 system VM
- VirtualBox
- Win/Linux/MacOS/... ABI
  - Implements an OS-x86 VM
- OS kernel (Win/Linux/MacOS/...)
- x86 ISA
Virtual Machines Are Everywhere

- Example: Consider a Python program running on a Linux Virtual Machine

<table>
<thead>
<tr>
<th>Component</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python program</td>
<td>Implements a Python VM</td>
</tr>
<tr>
<td>Python interpreter (CPython)</td>
<td>Linux ABI</td>
</tr>
<tr>
<td>Linux OS kernel</td>
<td>Implements a Linux-x86 VM</td>
</tr>
<tr>
<td>VirtualBox</td>
<td>x86 ISA</td>
</tr>
<tr>
<td>OS kernel (Win/Linux/MacOS/...)</td>
<td>Implements an x86 system VM</td>
</tr>
<tr>
<td>Hardware (e.g., your laptop)</td>
<td>Win/Linux/MacOS/... ABI</td>
</tr>
<tr>
<td></td>
<td>Implements an OS-x86 VM</td>
</tr>
<tr>
<td></td>
<td>x86 ISA</td>
</tr>
</tbody>
</table>
Virtual Machines Are Everywhere

- Example: Consider a Python program running on a Linux Virtual Machine

<table>
<thead>
<tr>
<th>Python program</th>
<th>Python Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python interpreter (CPython)</td>
<td>Implements a Python VM</td>
</tr>
<tr>
<td>Linux OS kernel</td>
<td>Linux ABI</td>
</tr>
<tr>
<td>VirtualBox</td>
<td>Implements a Linux-x86 VM</td>
</tr>
<tr>
<td>OS kernel (Win/Linux/MacOS/...)</td>
<td>x86 ISA</td>
</tr>
<tr>
<td>Hardware (e.g., your laptop)</td>
<td>Implements an x86 system VM</td>
</tr>
<tr>
<td></td>
<td>Win/Linux/MacOS/... ABI</td>
</tr>
<tr>
<td></td>
<td>Implements an OS-x86 VM</td>
</tr>
<tr>
<td></td>
<td>x86 ISA</td>
</tr>
<tr>
<td></td>
<td>Implements an x86 physical machine</td>
</tr>
</tbody>
</table>
Implementing Virtual Machines

• Virtual machines can be implemented entirely in software, but at a performance cost
  – e.g., Python programs are 10-100x slower than native Linux programs due to Python interpreter overheads
Implementing Virtual Machines

• Virtual machines can be implemented entirely in software, but at a performance cost
  – e.g., Python programs are 10-100x slower than native Linux programs due to Python interpreter overheads

• We want to support virtual machines with minimal overheads → need hardware support!
ISA Extensions to Support OS
ISA Extensions to Support OS

- Two modes of execution: user and supervisor
ISA Extensions to Support OS

- Two modes of execution: user and supervisor
  - OS kernel runs in supervisor mode
  - All other processes run in user mode
ISA Extensions to Support OS

- Two modes of execution: **user** and **supervisor**
  - OS kernel runs in supervisor mode
  - All other processes run in user mode
- **Privileged instructions and registers** that are only available in supervisor mode
ISA Extensions to Support OS

- **Two modes of execution:** user and supervisor
  - OS kernel runs in supervisor mode
  - All other processes run in user mode
- **Privileged instructions and registers** that are only available in supervisor mode
- **Traps (exceptions)** to safely transition from user to supervisor mode
ISA Extensions to Support OS

- Two modes of execution: user and supervisor
  - OS kernel runs in supervisor mode
  - All other processes run in user mode
- Privileged instructions and registers that are only available in supervisor mode
- Traps (exceptions) to safely transition from user to supervisor mode
- Virtual memory to provide private address spaces and abstract the storage resources of the machine
Process Mode Switching

Trap, e.g., i/o read() or exception

user mode

Switch to kernel mode; Pass arguments; Save app state

kernel mode

Trap handler

Check arguments Find trap handler addr

Kernel routine

Restore app state, Return to user
Protection – Single OS

- OS Kernel
  - User Process
  - User Process

Traps from User Processes to OS Kernel.
Supporting Multiple OSs

process₁ … processₙ process₁ … processₘ

OS Kernel₁ … OS Kernelₖ

Virtual Machine Monitor (VMM)

Hardware
A VMM (aka Hypervisor) provides a system virtual machine to each OS
Supporting Multiple OSs

- A VMM (aka Hypervisor) provides a system virtual machine to each OS
- VMM can run directly on hardware (as above) or on another OS
  - Precisely, VMM can be implemented against an ISA (as above) or a process-level ABI. Who knows what lays below the interface...
Motivation for Multiple OSs

Some motivations for using multiple operating systems on a single computer:

- Allows use of capabilities of multiple distinct operating systems.
Motivation for Multiple OSs

Some motivations for using multiple operating systems on a single computer:

- Allows use of capabilities of multiple distinct operating systems.

- Allows different users to share a system while using completely independent software stacks.
Motivation for Multiple OSs

Some motivations for using multiple operating systems on a single computer:

- Allows use of capabilities of multiple distinct operating systems.
- Allows different users to share a system while using completely independent software stacks.
Motivation for Multiple OSs

Some motivations for using multiple operating systems on a single computer:

- Allows use of capabilities of multiple distinct operating systems.
- Allows different users to share a system while using completely independent software stacks.
- Allows for load balancing and migration across multiple machines.
Motivation for Multiple OSs

Some motivations for using multiple operating systems on a single computer:

- Allows use of capabilities of multiple distinct operating systems.
- Allows different users to share a system while using completely independent software stacks.
- Allows for load balancing and migration across multiple machines.
- Allows operating system development without making entire machine unstable or unusable.
Virtualization Nomenclature

From (Machine we are attempting to execute)
- Guest
- Client
- Foreign ISA

To (Machine that is doing the real execution)
- Host
- Target
- Native ISA
Virtual Machine Requirements
[Popek and Goldberg, 1974]

- Equivalence/Fidelity: A program running on the VMM should exhibit a behavior essentially identical to that demonstrated when running on an equivalent machine directly.

- Resource control/Safety: The VMM must be in complete control of the virtualized resources.

- Efficiency/Performance: A statistically dominant fraction of machine instructions must be executed without VMM intervention.
Virtual Machine Requirements
[Popek and Goldberg, 1974]

Classification of instructions into 3 groups:

- **Privileged instructions**: Instructions that *trap* if the processor is in *user mode* and do not trap if it is in a more privileged mode.

- **Control-sensitive instructions**: Instructions that attempt to change the configuration of resources in the system.

- **Behavior-sensitive instructions**: Those whose behavior depends on the configuration of resources, e.g., mode.

Building an *effective* VMM for an architecture is possible if the set of sensitive instructions is a subset of the set of privileged instructions.
Sensitive instruction handling

Sensitive instruction

Non-VMM mode

VMM mode

Switch to VMM mode; Pass arguments; Save app state

VMM handler

Find handler addr

VMM routine

Restore app state, Return to guest
Protection – Multiple OS

- VMM
- OS Kernel
- User Process
- Trap

Sensitive
Deny
Virtual Memory Operations

TLB can be designed to translate guest virtual addresses (gVA) to a host physical address (hPA), but...
Virtual Memory Operations

TLB can be designed to translate guest virtual addresses (gVA) to a host physical address (hPA), but...

- TLB misses are a ‘sensitive’ operation
Virtual Memory Operations

TLB can be designed to translate guest virtual addresses (gVA) to a host physical address (hPA), but...

- TLB misses are a ‘sensitive’ operation
- TLB misses happen very very very frequently
Virtual Memory Operations

TLB can be designed to translate guest virtual addresses (gVA) to a host physical address (hPA), but...

- TLB misses are a ‘sensitive’ operation
- TLB misses happen very very very frequently

- So how expensive are TLB fills?
Nested Page Tables

- Guest VA
- Guest Page Table Base
- Index 1
- Index 2
- Offset
- PTP
- PTE
- L1 Table
- L2 Table
- PPN
- Offset
- Guest PA
Nested Page Tables

Guest VA

Index 1  Index 2  Offset

Guest Page Table Base

L1 Table

Guest PA == Host VA

Index 1  Index 2  Offset

Host Page Table Base

L1 Table

Host PA

L2 Table

PPN  Offset

Guest PA

PPN  Offset

Host PA
Shadow Page Tables

Diagram:
- **Guest VA**
  - Guest Page Table Base
    - L1 Table
      - PTP
    - L2 Table
      - PTE
  - Index 1
  - Index 2
  - Offset
- PPN
- Offset
- Guest PA
Shadow Page Tables

Guest VA

Index 1  Index 2  Offset

Guest Page Table Base

PTP

L1 Table

PTP

L2 Table

PPN  Offset

L1 Table

Index 1  Index 2  Offset

Shadow Page Table Base

PTP

L1 Table

PTP

L2 Table

PPN  Offset

Host PA
## Nested vs Shadow Paging

<table>
<thead>
<tr>
<th></th>
<th>Native</th>
<th>Nested Paging</th>
<th>Shadow Paging</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TLB Hit</strong></td>
<td>VA-&gt;PA</td>
<td>gVA-&gt;hPA</td>
<td>gVA-&gt;hPA</td>
</tr>
<tr>
<td><strong>TLB Miss (max)</strong></td>
<td>4</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td><strong>PTE Updates</strong></td>
<td>Fast</td>
<td>Fast</td>
<td>Uses VMM</td>
</tr>
</tbody>
</table>

On x86-64
Security and Side Channels

- ISA and ABI are **timing-independent** interfaces
  - Specify *what* should happen, not *when*
Security and Side Channels

- ISA and ABI are timing-independent interfaces
  - Specify *what* should happen, not *when*

- Hardware isolation mechanisms like virtual memory guarantee that architectural state will not be directly exposed to other processes...
Security and Side Channels

- ISA and ABI are **timing-independent** interfaces
  - Specify *what* should happen, not *when*

- Hardware isolation mechanisms like virtual memory guarantee that architectural state will not be directly exposed to other processes...

- ...but timing and other implementation details (e.g., microarchitectural state, power, etc.) may be used as **side channels** to leak information!
Cache-Based Side Channels

- Attacker can infer shared cache behavior of victim
  - e.g., prime+probe attack: Attacker fills cache with own data, then times accesses to data to see which hit and miss, inferring which lines the victim is using
  - Leaks address-dependent information, e.g., RSA [Percival 2005] and AES keys [Osvik et al. 2005]
Cache-Based Side Channels

- Attacker can infer shared cache behavior of victim
  - e.g., prime+probe attack: Attacker fills cache with own data, then times accesses to data to see which hit and miss, inferring which lines the victim is using
  -Leaks address-dependent information, e.g., RSA [Percival 2005] and AES keys [Osvik et al. 2005]
- Microarch side channels among threads running on same SMT core?
Cache-Based Side Channels

- Attacker can infer shared cache behavior of victim
  - e.g., prime+probe attack: Attacker fills cache with own data, then times accesses to data to see which hit and miss, inferring which lines the victim is using
  - Leaks address-dependent information, e.g., RSA [Percival 2005] and AES keys [Osvik et al. 2005]

- Microarch side channels among threads running on same SMT core?
Cache-Based Side Channels

- Attacker can infer shared cache behavior of victim
  - e.g., prime+probe attack: Attacker fills cache with own data, then times accesses to data to see which hit and miss, inferring which lines the victim is using
  - Leaks address-dependent information, e.g., RSA [Percival 2005] and AES keys [Osvik et al. 2005]
- Microarch side channels among threads running on same SMT core?

L1/L2/L3 caches
Branch & other predictors
Cache-Based Side Channels

- Attacker can infer shared cache behavior of victim
  - e.g., prime+probe attack: Attacker fills cache with own data, then times accesses to data to see which hit and miss, inferring which lines the victim is using
  - Leaks address-dependent information, e.g., RSA [Percival 2005] and AES keys [Osvik et al. 2005]

- **Microarch side channels among threads running on same SMT core?**

```
<table>
<thead>
<tr>
<th>Core 0</th>
<th>Core 1</th>
<th>Core 2</th>
<th>Core 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priv Caches</td>
<td>Priv Caches</td>
<td>Priv Caches</td>
<td>Priv Caches</td>
</tr>
<tr>
<td>L1/L2/L3 caches</td>
<td>Branch &amp; other predictors</td>
<td>ROB/Issue/FU contention</td>
<td></td>
</tr>
</tbody>
</table>
```
Exploiting Speculative Execution in Side-Channel Attacks

- OoO cores run instructions speculatively and out of order
- Problem: Speculative instructions can change microarchitectural state → can leak data via side channel
Exploiting Speculative Execution in Side-Channel Attacks

- OoO cores run instructions speculatively and out of order
- Problem: Speculative instructions can change microarchitectural state $\rightarrow$ can leak data via side channel
- Example: In x86, process page table can have kernel pages, but kernel pages only accessible in kernel mode
  - Avoids switching page tables on context switches
Exploiting Speculative Execution in Side-Channel Attacks

- OoO cores run instructions speculatively and out of order
- Problem: Speculative instructions can change microarchitectural state → can leak data via side channel
- Example: In x86, process page table can have kernel pages, but kernel pages only accessible in kernel mode
  - Avoids switching page tables on context switches
  - *What does the following code do when run in user mode?*

\[
\text{val} = *\text{kernel\_address};
\]
Exploiting Speculative Execution in Side-Channel Attacks

- OoO cores run instructions speculatively and out of order
- Problem: Speculative instructions can change microarchitectural state \(\rightarrow\) can leak data via side channel
- Example: In x86, process page table can have kernel pages, but kernel pages only accessible in kernel mode
  - Avoids switching page tables on context switches
  - *What does the following code do when run in user mode?*

```c
val = *kernel_address;
```

Causes a protection fault
Exploiting Speculative Execution in Side-Channel Attacks

• OoO cores run instructions speculatively and out of order
• Problem: Speculative instructions can change microarchitectural state → can leak data via side channel
• Example: In x86, process page table can have kernel pages, but kernel pages only accessible in kernel mode

  - Avoids switching page tables on context switches
  - *What does the following code do when run in user mode?*

    ```
    val = *kernel_address;
    ```

    Causes a protection fault

    In Intel processors, protection check happens late
    → Kernel data speculatively loaded into val register!
Meltdown
[Lipp et al. 2018]

1. Setup: Attacker allocates 256-line \texttt{probe\_array}, flushes all its cache lines
1. Setup: Attacker allocates 256-line `probe_array`, flushes all its cache lines

2. Transmit: Attacker executes

```c
uint8_t byte = *kernel_address;
probe_array[byte] = 1;
```
**Meltdown**
[Lipp et al. 2018]

1. **Setup**: Attacker allocates 256-line `probe_array`, flushes all its cache lines
2. **Transmit**: Attacker executes
   
   ```
   uint8_t byte = *kernel_address;
   probe_array[byte] = 1;
   ```

3. **Receive**: After handling protection fault, attacker times accesses to all cache lines of `probe_array`, finds which one hits → recovers `byte`
Meltdown
[Lipp et al. 2018]

1. Setup: Attacker allocates 256-line `probe_array`, flushes all its cache lines
2. Transmit: Attacker executes

   ```c
   uint8_t byte = *kernel_address;
   probe_array[byte] = 1;
   ```

3. Receive: After handling protection fault, attacker times accesses to all cache lines of `probe_array`, finds which one hits → recovers `byte`

- Result: Attacker can read arbitrary kernel data!
  - For higher performance, use transactional memory (protection fault aborts transaction on exception instead of invoking kernel)
  - Mitigation: Do not map kernel data in user page tables
Domain of victim

Transmitter

Access

Secret

Side channel

Attacker

Receiver

Secret

General Attack Schema
[Belay, Devadas, Emer]
General Attack Schema
[Belay, Devadas, Emer]

- Types of transmitter:
  1. Pre-existing (the victim itself leaks secret, e.g., RSA/AES keys)
  2. Programmed by attacker (e.g., Meltdown)
  3. Synthesized from existing victim code by attacker (e.g., Spectre)
• Consider the following kernel code, e.g., in a system call
  if (x < array1_size)
      y = array2[array1[x] * 4096];
Spectre variant 1 — Exploiting Conditional Branches [Kocher et al. 2018]

- Consider the following kernel code, e.g., in a system call
  
  ```c
  if (x < array1_size)
      y = array2[array1[x] * 4096];
  ```

1. Setup: Attacker invokes this kernel code with small values of \( x \) to train the branch predictor to taken
Spectre variant 1 — Exploiting Conditional Branches [Kocher et al. 2018]

- Consider the following kernel code, e.g., in a system call

```c
if (x < array1_size)
    y = array2[array1[x] * 4096];
```

1. Setup: Attacker invokes this kernel code with small values of $x$ to train the branch predictor to taken

2. Transmit: Attacker invokes this code with an out-of-bounds $x$, so that &array1[x] maps to some desired kernel address. Core mispredicts branch, fetches array2[array1[x] * 4096]’s line into the cache.
Spectre variant 1 — Exploiting Conditional Branches [Kocher et al. 2018]

• Consider the following kernel code, e.g., in a system call

```c
if (x < array1_size)
    y = array2[array1[x] * 4096];
```

1. Setup: Attacker invokes this kernel code with small values of x to train the branch predictor to taken.

2. Transmit: Attacker invokes this code with an out-of-bounds x, so that &array1[x] maps to some desired kernel address. Core mispredicts branch, fetches array2[array1[x] * 4096]’s line into the cache.

3. Receive: Attacker probes cache to infer which line of array2 was fetched, learns data at kernel address
   - array2 may or may not be accessible to attacker (can use prime+probe)
• Assume the BTB stores partial tags but full target PCs. How can this be exploited?
Spectre variant 2—Branch Target Injection [Kocher et al. 2018]

• Assume the BTB stores partial tags but full target PCs. How can this be exploited?
  1. Setup: Attacker chooses any jump in kernel code, mistrains BTB so that it predicts a target PC under the control of the attacker that leaks information, e.g.,
Spectre variant 2—Branch Target Injection [Kocher et al. 2018]

• Assume the BTB stores partial tags but full target PCs. How can this be exploited?
  1. Setup: Attacker chooses any jump in kernel code, mistrains BTB so that it predicts a target PC under the control of the attacker that leaks information, e.g.,

    ```c
    uint8_t byte = *kernel_address;
    probe_array[byte] = 1;
    ```
Spectre variant 2—Branch Target Injection [Kocher et al. 2018]

• Assume the BTB stores partial tags but full target PCs. How can this be exploited?
  1. Setup: Attacker chooses any jump in kernel code, mistrains BTB so that it predicts a target PC under the control of the attacker that leaks information, e.g.,

    ```c
    uint8_t byte = *kernel_address;
    probe_array[byte] = 1;
    ```

  2. Transmit & receive: Like in Spectre v1
Spectre variant 2—Branch Target Injection [Kocher et al. 2018]

• Assume the BTB stores partial tags but full target PCs. How can this be exploited?
  1. Setup: Attacker chooses any jump in kernel code, mistrains BTB so that it predicts a target PC under the control of the attacker that leaks information, e.g.,

        ```c
        uint8_t byte = *kernel_address;
        probe_array[byte] = 1;
        ```

  2. Transmit & receive: Like in Spectre v1

• Most BTBs store partial tags and targets...
  - Hard to get BTB to jump from a kernel address to a far-away user address
Spectre variant 2—Branch Target Injection [Kocher et al. 2018]

- Assume the BTB stores partial tags but full target PCs. How can this be exploited?
  1. Setup: Attacker chooses any jump in kernel code, mistrains BTB so that it predicts a target PC under the control of the attacker that leaks information, e.g.,

\[
\text{uint8_t byte} = \ast\text{kernel_address};
\]

\[
\text{probe_array[byte]} = 1;
\]

  2. Transmit & receive: Like in Spectre v1

- Most BTBs store partial tags and targets...
  - Hard to get BTB to jump from a kernel address to a far-away user address

- But most cores add an indirect branch predictor that stores full targets (e.g., to predict virtual function calls)
  - Spectre v2 exploits this predictor instead
Spectre variants and mitigations

- Spectre relies on speculative execution, not late exception checks → Much harder to fix than Meltdown
Spectre variants and mitigations

- Spectre relies on speculative execution, not late exception checks → Much harder to fix than Meltdown
- Several other Spectre variants reported
  - Leveraging the speculative store buffer, return address stack, leaking privileged registers, etc.
Spectre variants and mitigations

• Spectre relies on speculative execution, not late exception checks → Much harder to fix than Meltdown
• Several other Spectre variants reported
  – Leveraging the speculative store buffer, return address stack, leaking privileged registers, etc.
• Can attack any type of VM, including OSs, VMMs, JavaScript engines in browsers, and the OS network stack (NetSpectre)
Spectre variants and mitigations

• Spectre relies on speculative execution, not late exception checks → Much harder to fix than Meltdown

• Several other Spectre variants reported
  – Leveraging the speculative store buffer, return address stack, leaking privileged registers, etc.

• Can attack any type of VM, including OSs, VMMs, JavaScript engines in browsers, and the OS network stack (NetSpectre)

• Short-term mitigations:
  – Microcode updates (disable sharing of speculative state when possible)
  – OS and compiler patches to selectively avoid speculation
Spectre variants and mitigations

- Spectre relies on speculative execution, not late exception checks → Much harder to fix than Meltdown
- Several other Spectre variants reported
  - Leveraging the speculative store buffer, return address stack, leaking privileged registers, etc.
- Can attack any type of VM, including OSs, VMMs, JavaScript engines in browsers, and the OS network stack (NetSpectre)

Short-term mitigations:
- Microcode updates (disable sharing of speculative state when possible)
- OS and compiler patches to selectively avoid speculation

Long-term mitigations:
- Disabling speculation?
- Closing side channels?
Thank you!