DuVisor

ARTIFACT EVALUATED usenix ASSOCIATION AVAILABLE

ARTIFACT EVALUATED usenix ASSOCIATION FUNCTIONAL

ARTIFACT EVALUATED usenix ASSOCIATION REPRODUCED

Security and Performance in the Delegated User-level Virtualization

Jiahao Chen*, Dingji Li*, Zeyu Mi, Yuxuan Liu,
Binyu Zang, Haibing Guan, and Haibo Chen

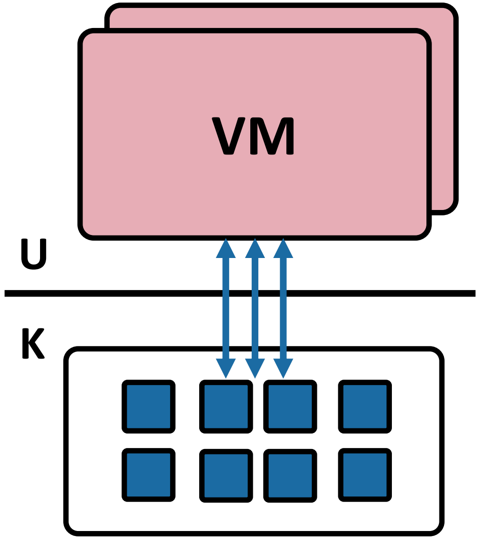
Institute of Parallel and Distributed Systems, SEIEE, Shanghai Jiao Tong University
Engineering Research Center for Domain-specific Operating Systems, Ministry of Education, China
MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
Shanghai Key Laboratory of Scalable Computing and Systems, Shanghai Jiao Tong University
**Co-first authors*

IPADS
INSTITUTE OF PARALLEL
AND DISTRIBUTED SYSTEMS

History of Virtualization

🟦 Virtualization Function

⬛ General Function

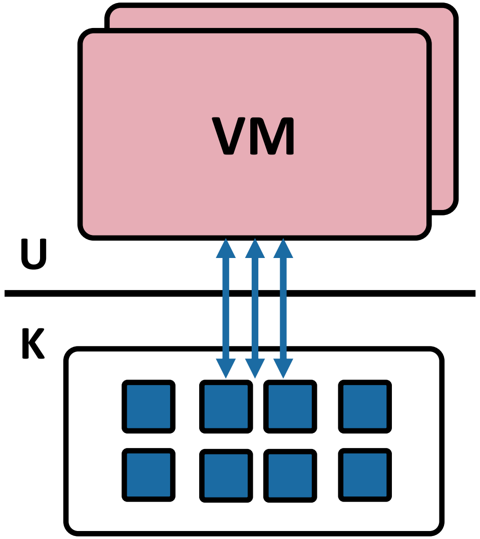


Stage 1 - Monolithic Hypervisor
E.g., IBM VM/370

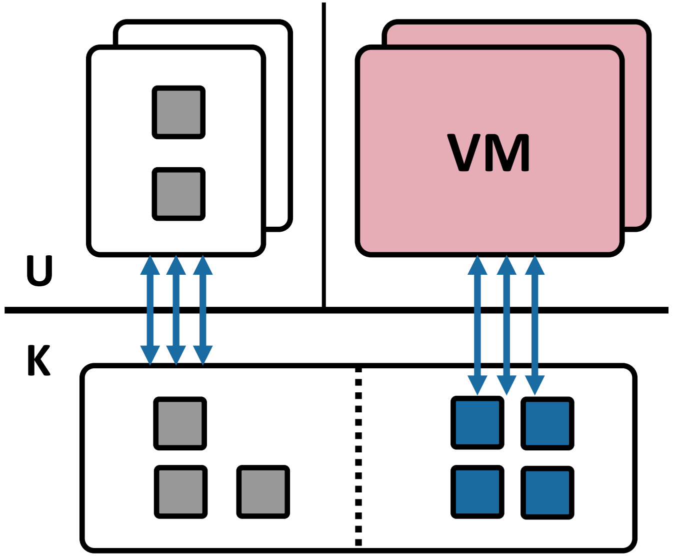
History of Virtualization

🟦 Virtualization Function

⬛ General Function



Stage 1 - Monolithic Hypervisor
E.g., IBM VM/370

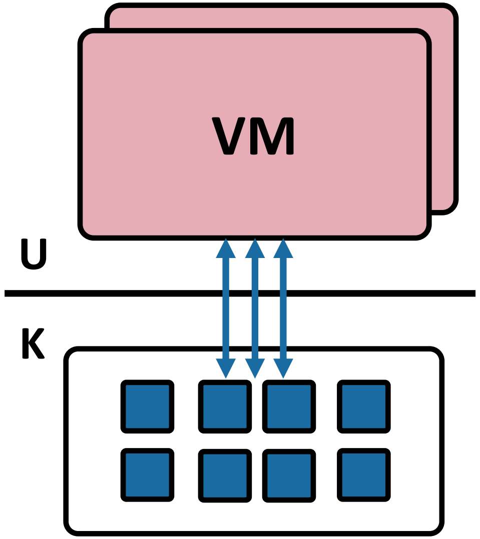


Stage 2 – Reusing Host OS
or Management VM
E.g., Xen (SOSP 2003)

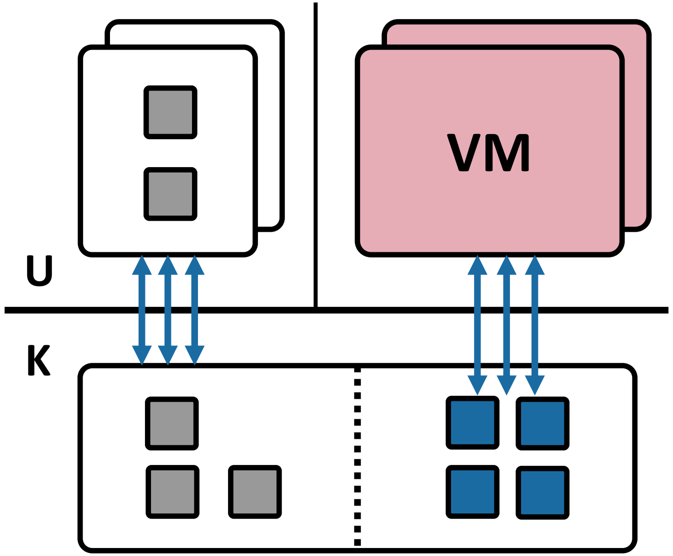
History of Virtualization

🟦 Virtualization Function

■ General Function



Stage 1 - Monolithic Hypervisor
E.g., IBM VM/370

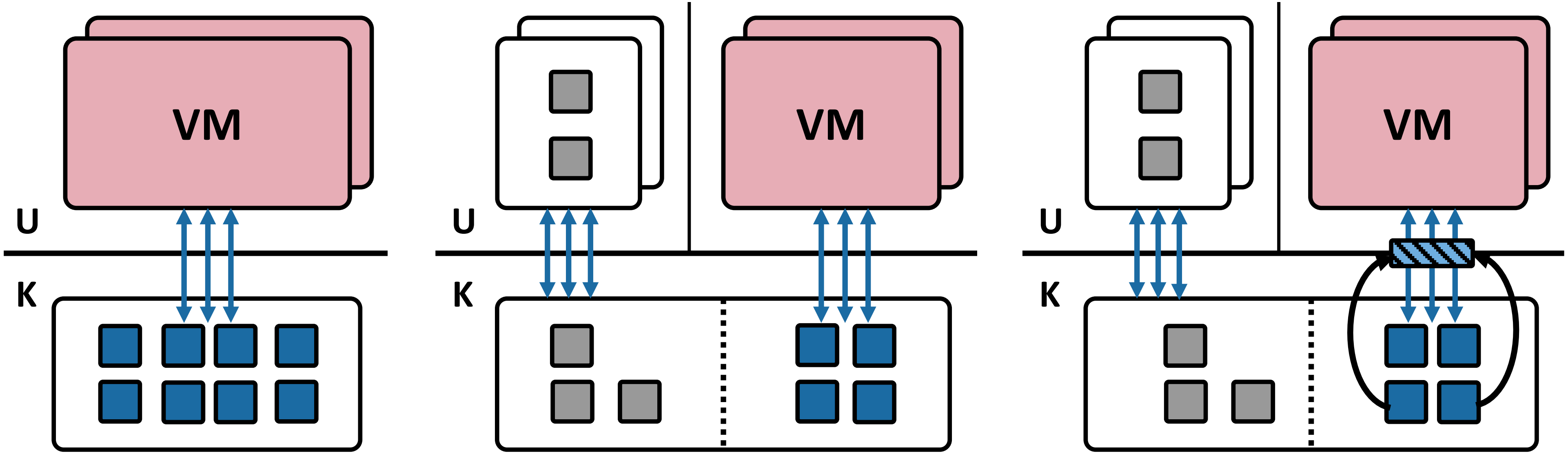


Stage 2 – Reusing Host OS
or Management VM
E.g., Xen (SOSP 2003)

History of Virtualization

🟦 Virtualization Function

◼ General Function



Stage 1 - Monolithic Hypervisor
E.g., IBM VM/370

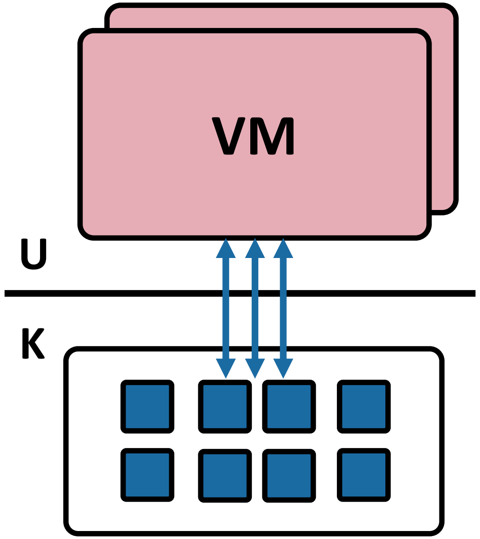
Stage 2 – Reusing Host OS
or Management VM
E.g., Xen (SOSP 2003)

Stage 3 – Hardware Virtualization

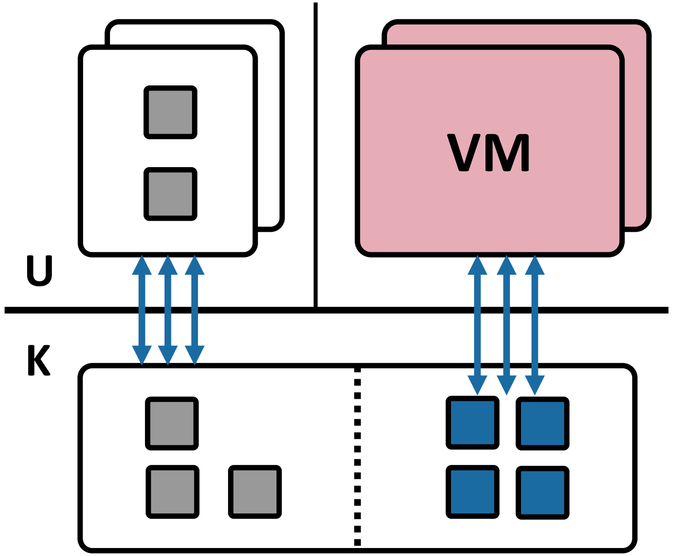
History of Virtualization

🟦 Virtualization Function

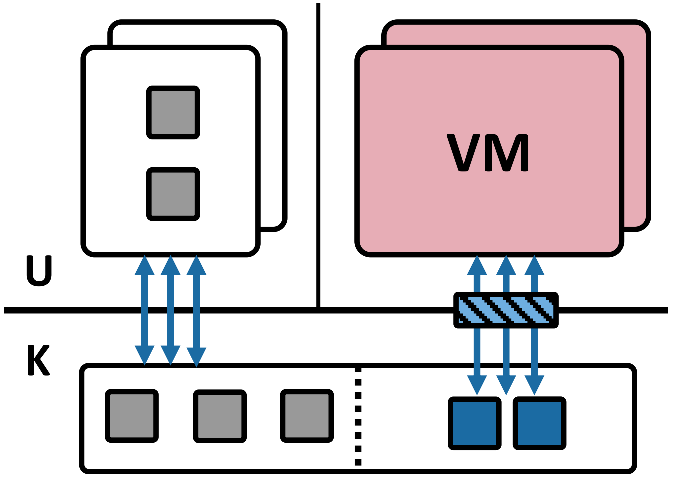
⬛ General Function



Stage 1 - Monolithic Hypervisor
E.g., IBM VM/370



Stage 2 – Reusing Host OS
or Management VM
E.g., Xen (SOSP 2003)

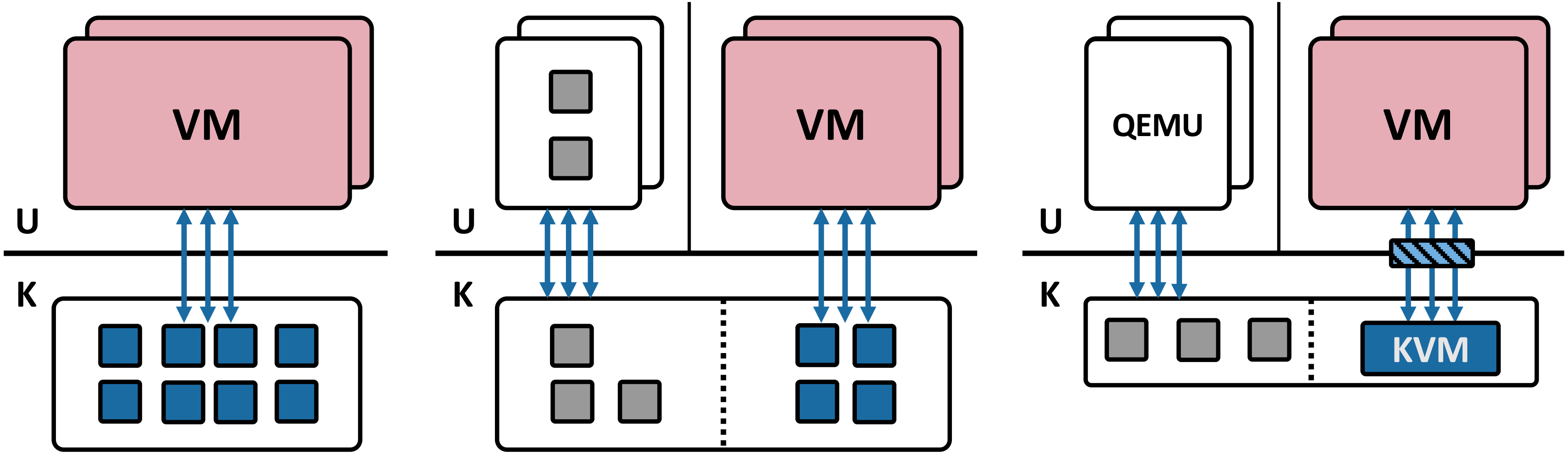


Stage 3 – Hardware Virtualization

History of Virtualization

🟦 Virtualization Function

⬛ General Function



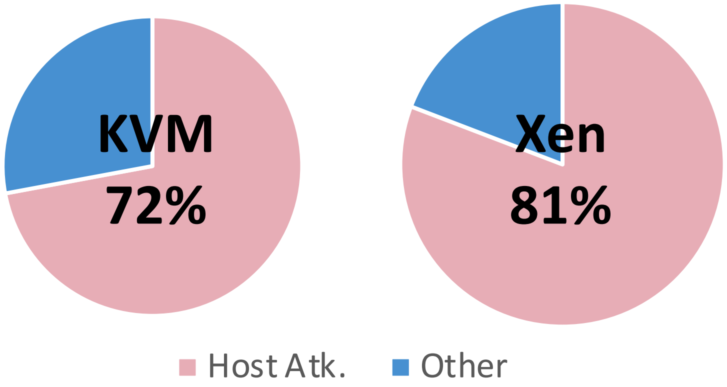
Stage 1 - Monolithic Hypervisor
E.g., IBM VM/370

Stage 2 – Reusing Host OS
or Management VM
E.g., Xen (SOSP 2003)

Stage 3 – Hardware Virtualization
E.g., QEMU/KVM

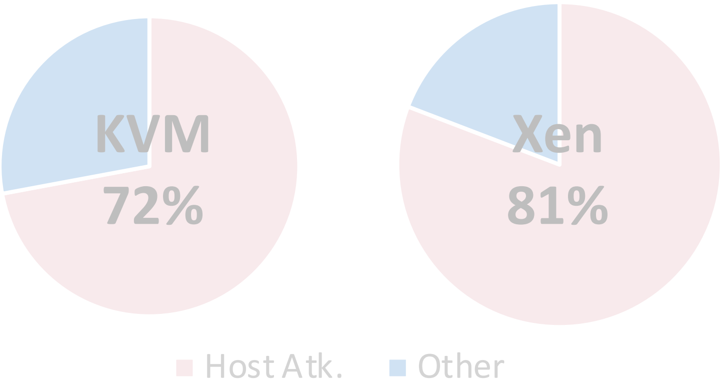
Vulnerabilities of Hypervisors

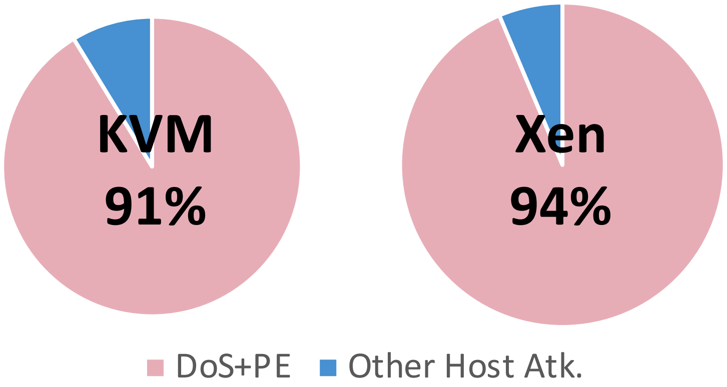
- Large vulnerability quantity
  - **About 500 CVEs** for KVM and Xen
  - Most of them are host-attacking



Vulnerabilities of Hypervisors

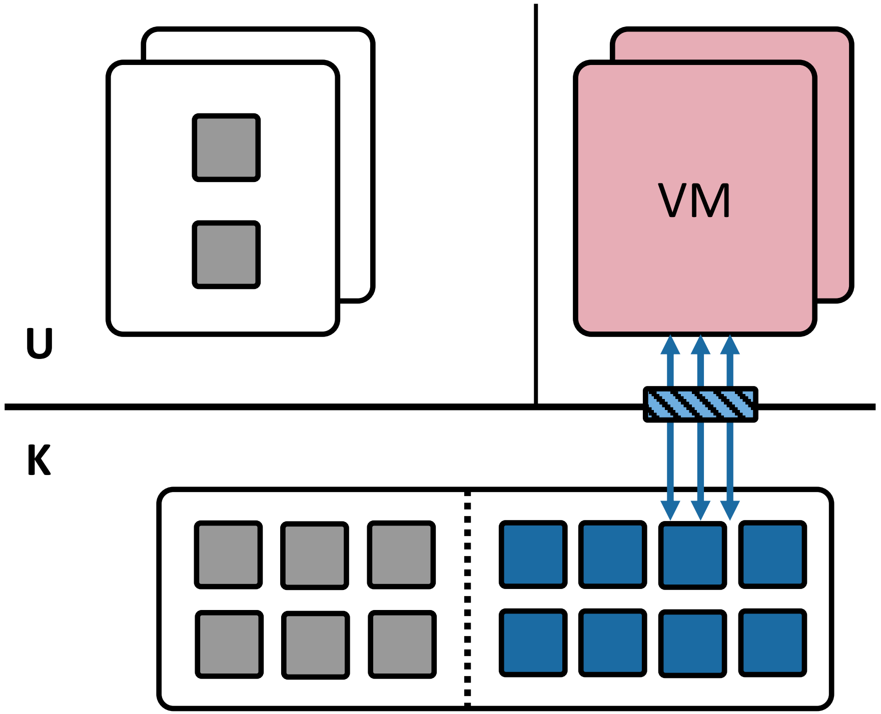
- Large vulnerability quantity
  - **About 500 CVEs** for KVM and Xen
  - Most of them are host-attacking
- Severe security threats
  - **Over 90%** of the Host-attacking CVEs cause DoS attacks
  - 26% and 34% cause privilege escalation
- Low exploit cost





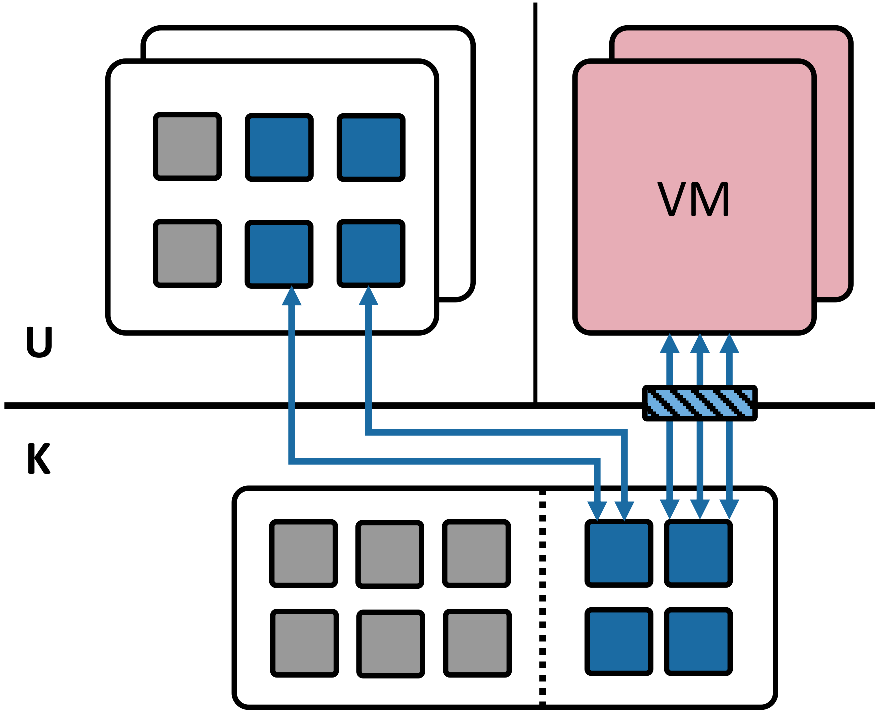
Prior Works

- Deprivilege large number of hypervisor components to the user mode
  - NOVA (EuroSys-2010)
  - DeHype (NDSS-2013)
- Part of the vulnerabilities are deprivileged to the user state along with their components



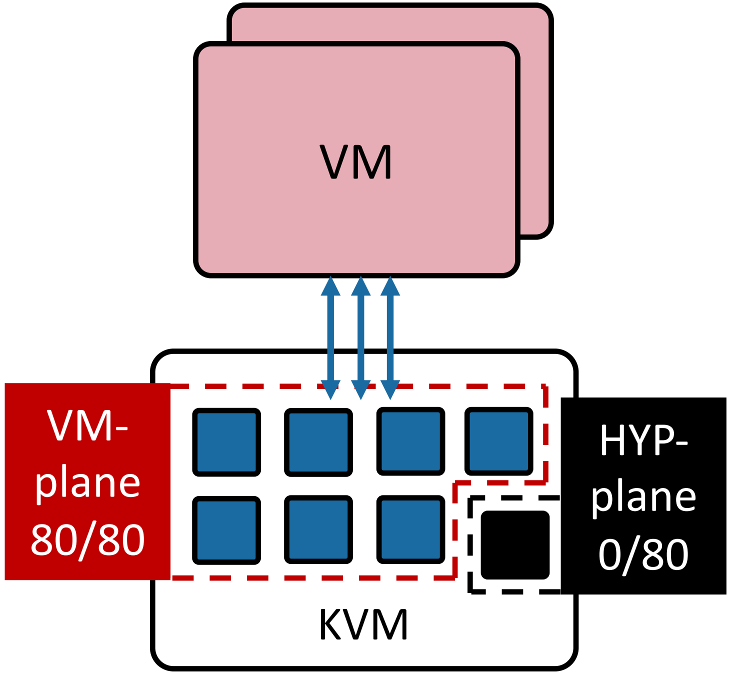
Prior Works

- Deprivilege large number of hypervisor components to the user mode
  - NOVA (EuroSys-2010)
  - DeHype (NDSS-2013)
- Part of the vulnerabilities are deprivileged to the user state along with their components



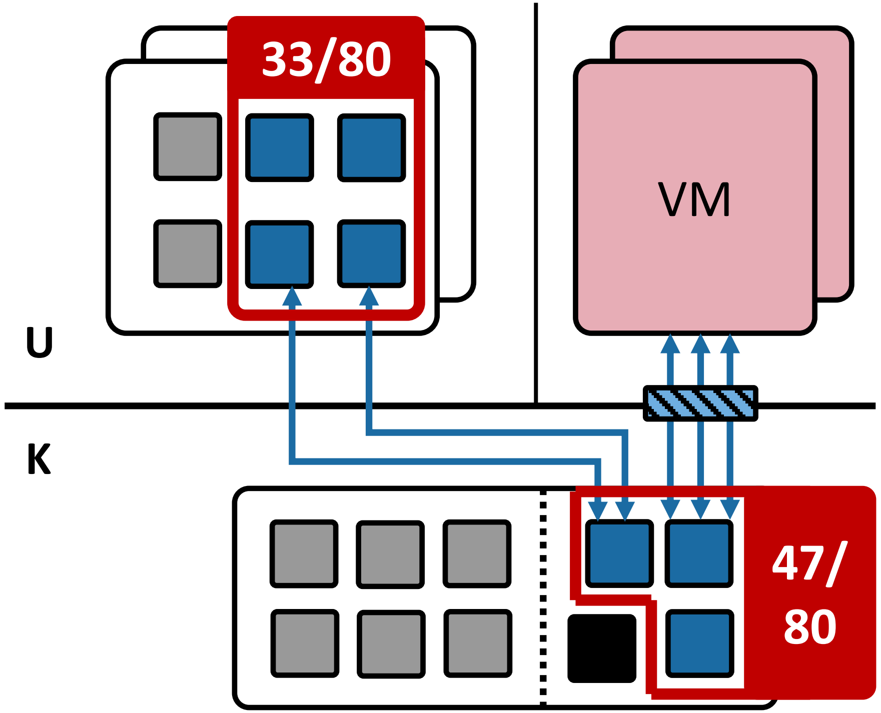
Limitations of Deprivileged Execution

- All 80 host-attacking CVEs reside in the VM-plane subsystems
  - VM-plane: Serve VMs directly
    - E.g., memory virtualization, ISA emulation
  - Hypervisor-plane: Serve VM-plane subsystems for hypervisor control
    - E.g., resource control, hypervisor initialization



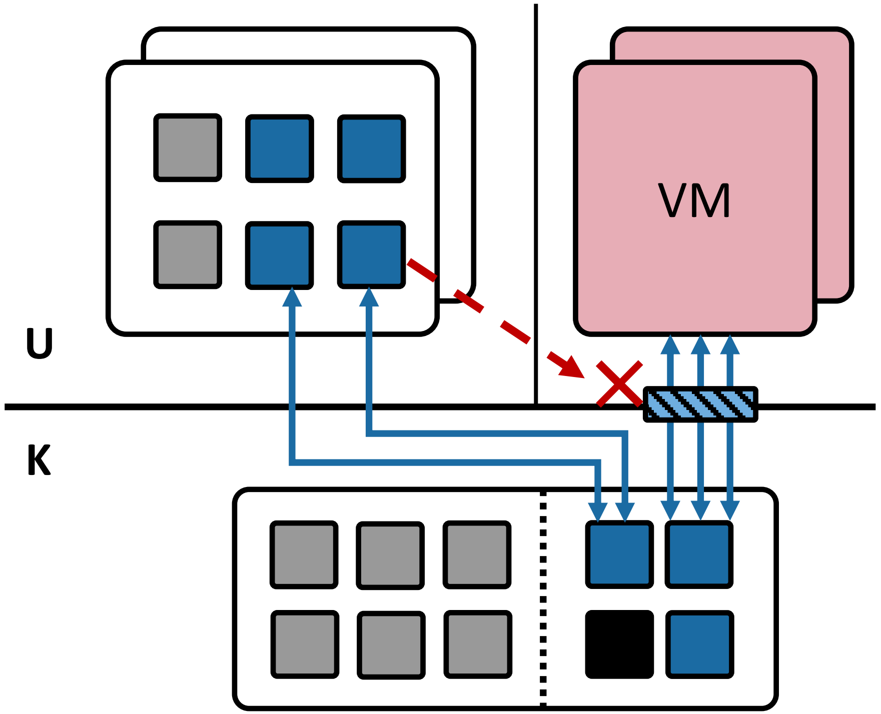
Limitations of Deprivileged Execution

- Non-eliminable in-kernel vulnerabilities
  - Only 33 of the 80 host-attacking CVEs on KVM are deprivileged



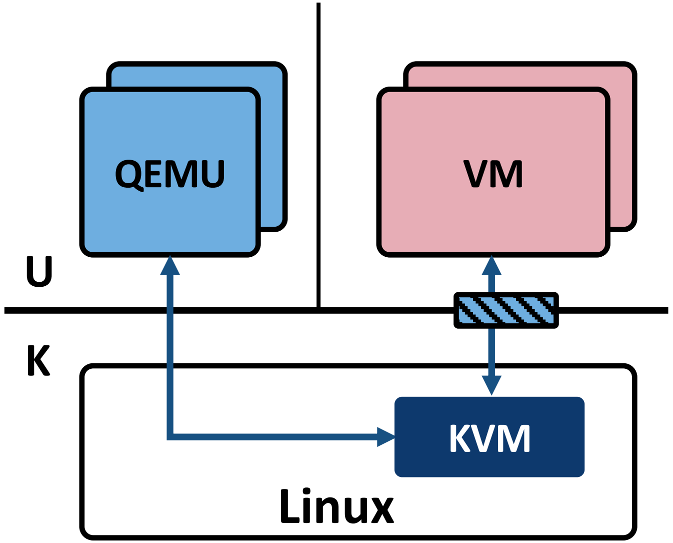
Limitations of Deprivileged Execution

- Non-eliminable in-kernel vulnerabilities
  - Only 33 of the 80 host-attacking CVEs on KVM are deprivileged
  - Several vulnerable components are constrained in the kernel to perform privileged operations



Limitations of Deprivileged Execution

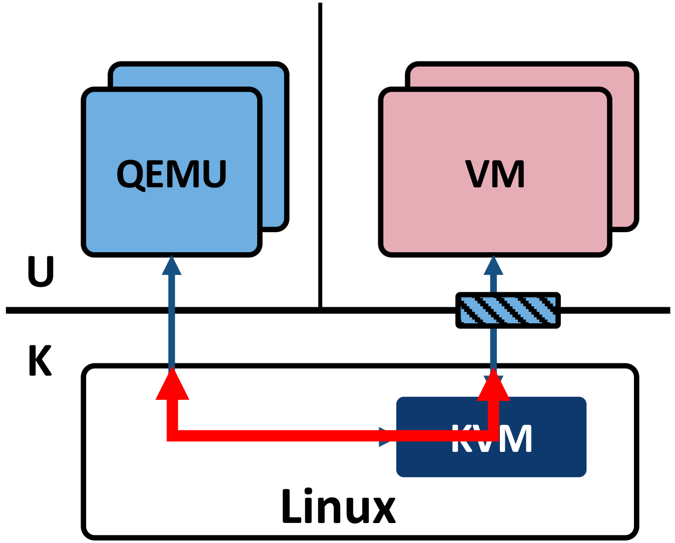
- Non-eliminable in-kernel vulnerabilities
- Redundant and costly mode switching



Platform	Total (Cycle)
ARM	5,919
RISC-V	7,202
x86	4,119

Limitations of Deprivileged Execution

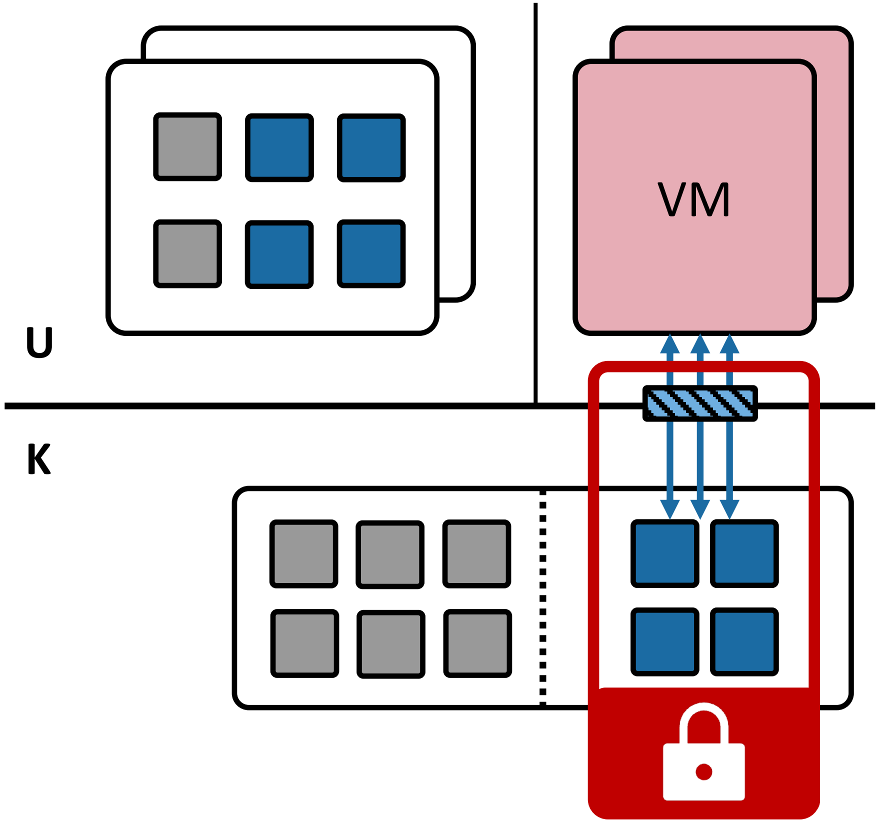
- Non-eliminable in-kernel vulnerabilities
- Redundant and costly mode switching



Platform	Kernel	User	Total (Cycle)
ARM	73.0%	1,596	5,919
RISC-V	43.5%	4,067	7,202
x86	58.6%	1,704	4,119

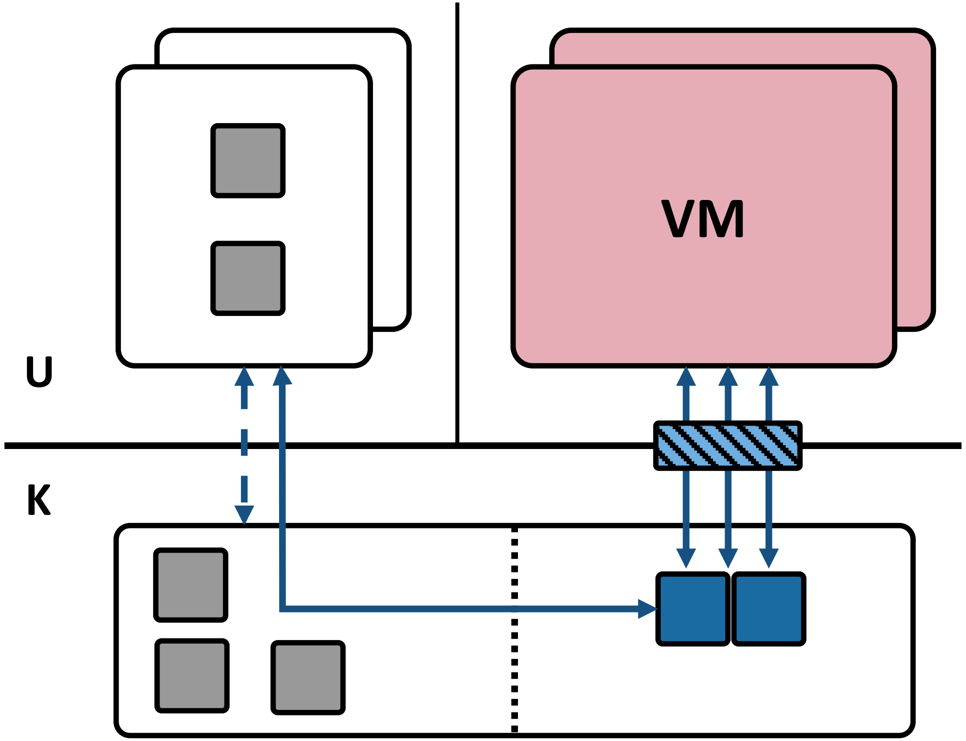
Root Cause

- The **unnecessary tight coupling** between the **hardware virtualization extensions** and **kernel mode**



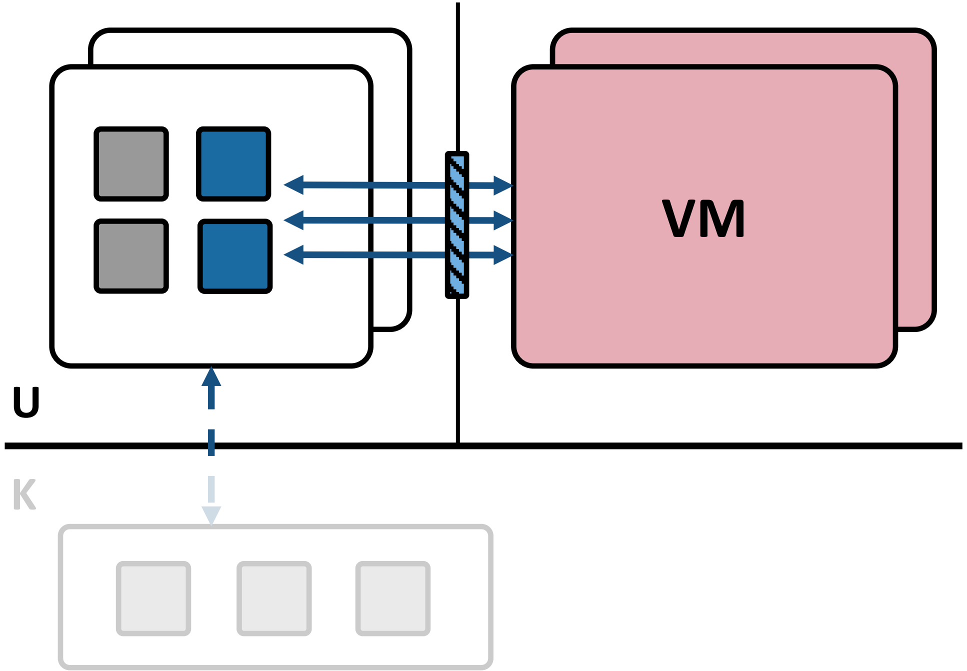
DuVisor

Delegated Virtualization



Stage 3 – Hardware Virtualization

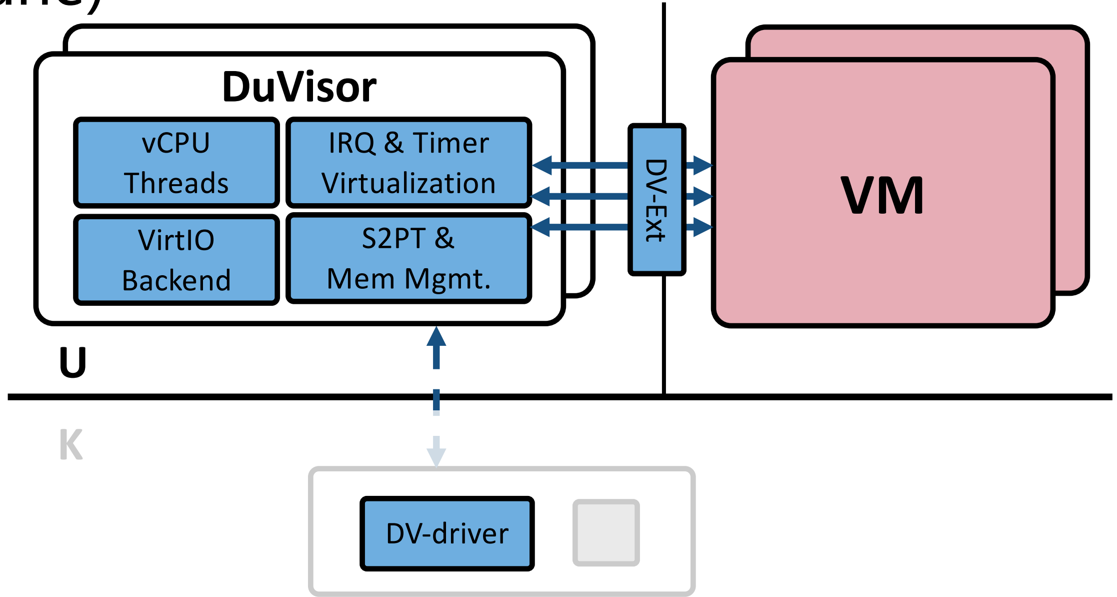
Delegated Virtualization



Delegated Virtualization

Delegated Virtualization

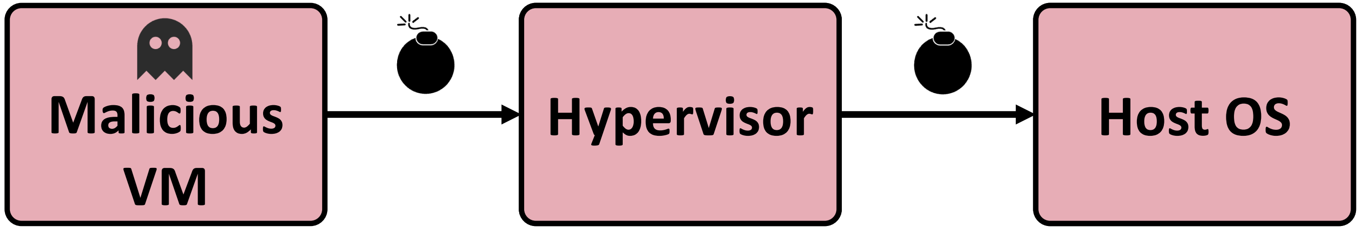
- The Delegated Virtualization Extension (DV-Ext)
- DuVisor hypervisor processes (VM-plane serving VMs directly)
- DV-driver (Hypervisor-plane)



Threat Model & Assumptions

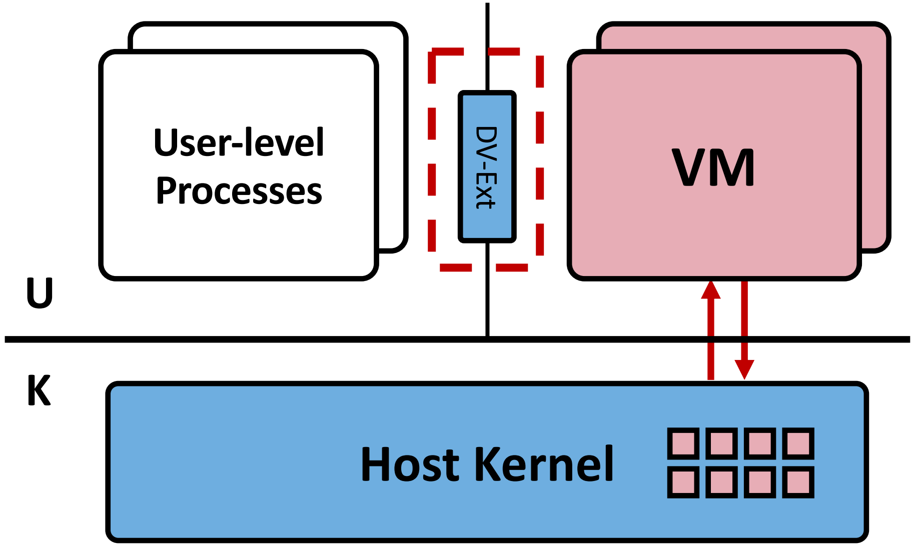
A hostile tenant compromises the hypervisor to further attack **the host kernel** and **other VMs**

- DuVisor **CAN** be compromised
- Hardware is correctly implemented
- The host kernel with DV-driver is trusted



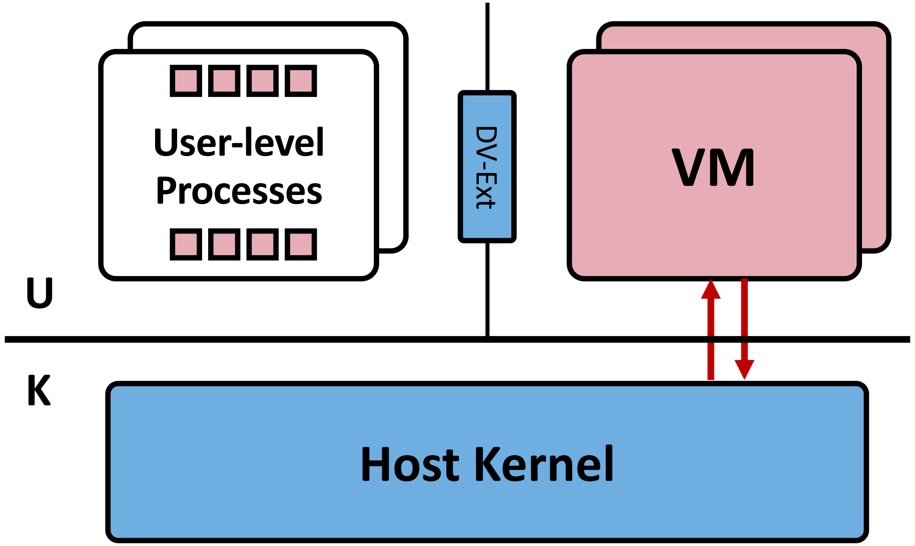
Delegated Virtualization Extension

- Virtualization registers and instructions for user mode



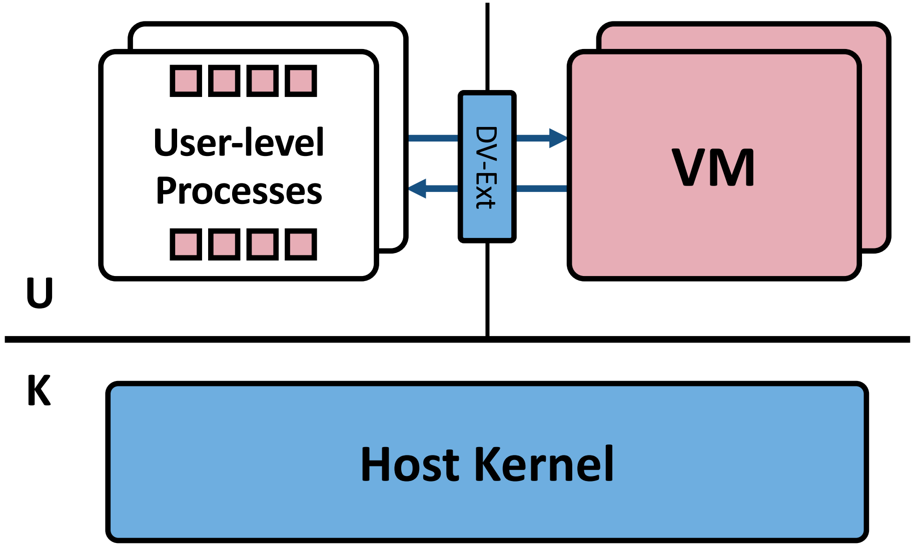
Delegated Virtualization Extension

- Virtualization registers and instructions for user mode

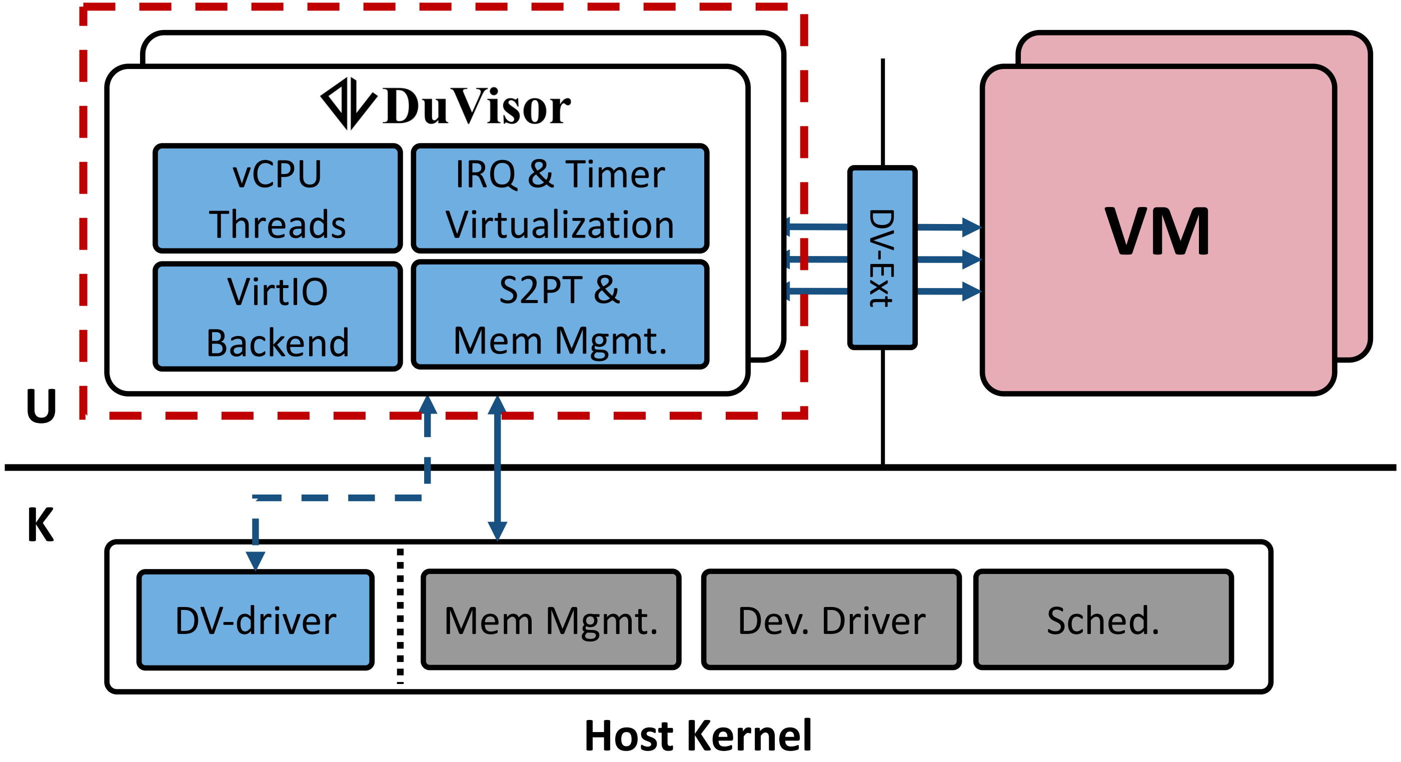


Delegated Virtualization Extension

- Virtualization registers and instructions for user mode
- Delegatable VM Exits (DVE)

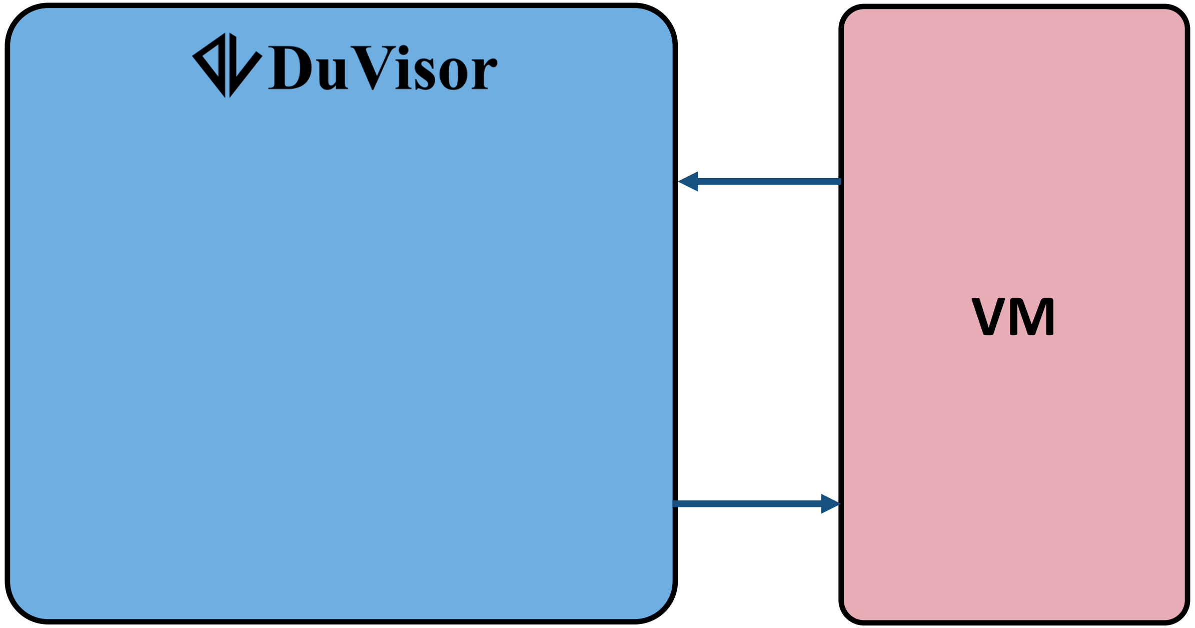


DuVisor


DuVisor
vCPU Threads
IRQ & Timer Virtualization
VirtIO Backend
S2PT & Mem Mgmt.
DV-Ext
VM
U
K
DV-driver
Mem Mgmt.
Dev. Driver
Sched.
Host Kernel

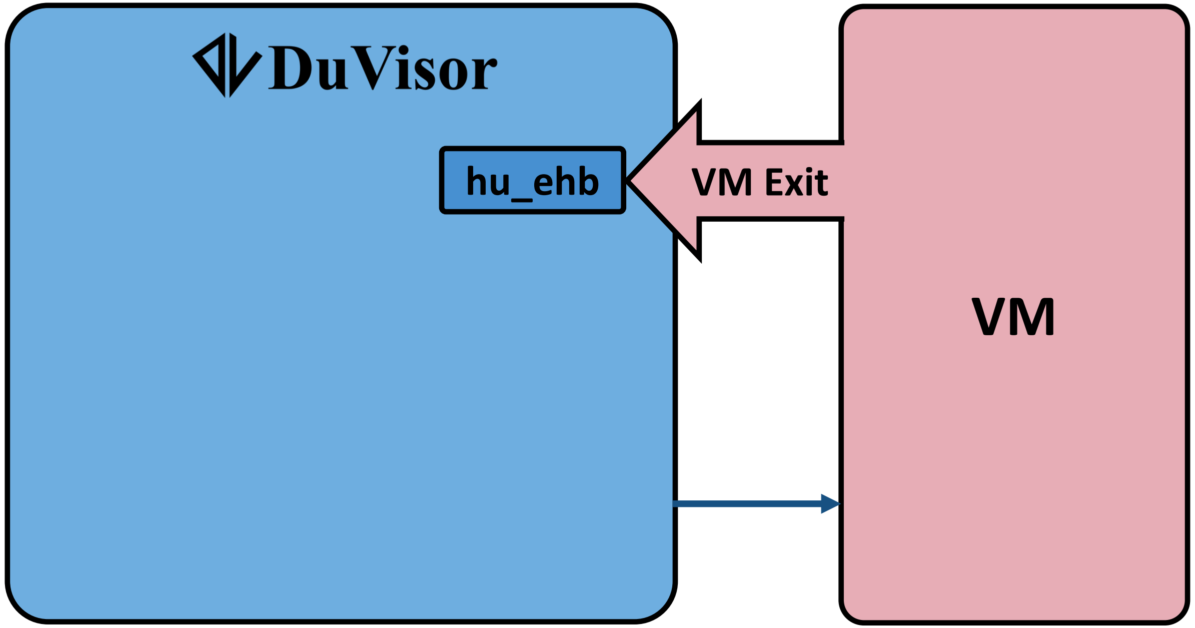
Handling VM Exits

- All exceptions that result in exits are sent to the user-level DuVisor



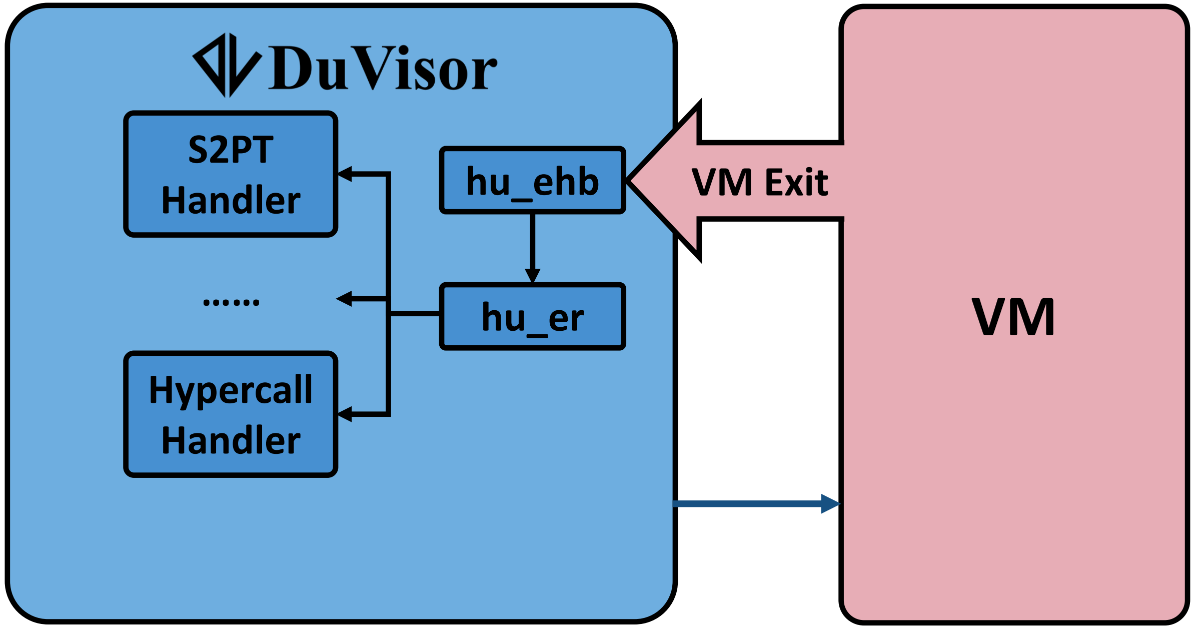
Handling VM Exits

- All exceptions that result in exits are sent to the user-level DuVisor



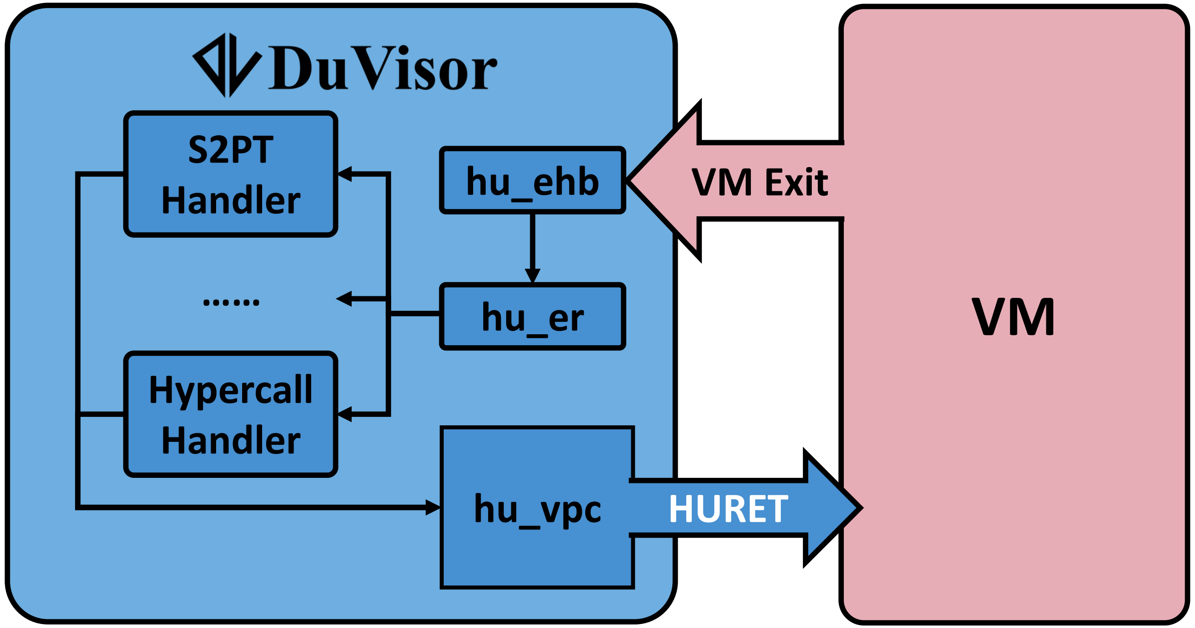
Handling VM Exits

- All exceptions that result in exits are sent to the user-level DuVisor



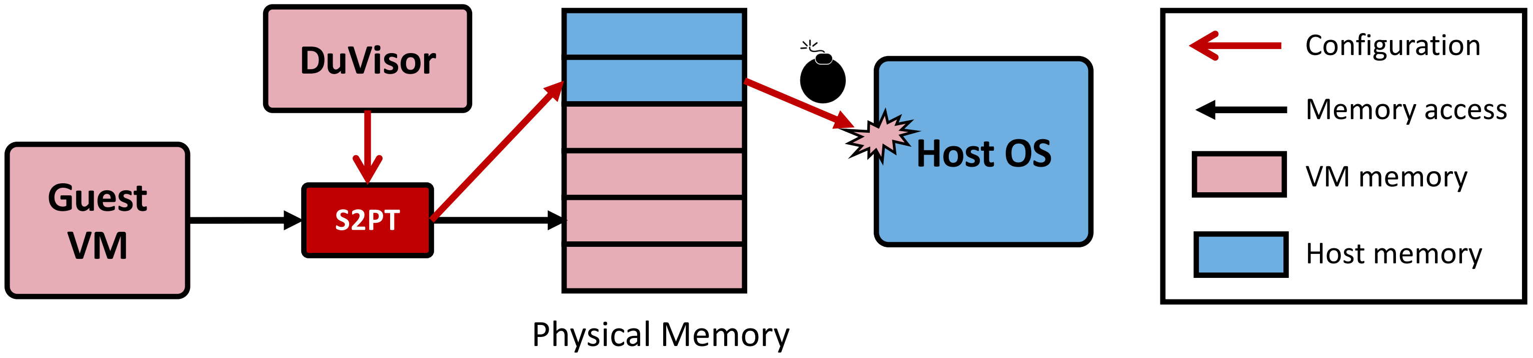
Handling VM Exits

- All exceptions that result in exits are sent to the user-level DuVisor



Restricted Memory Virtualization

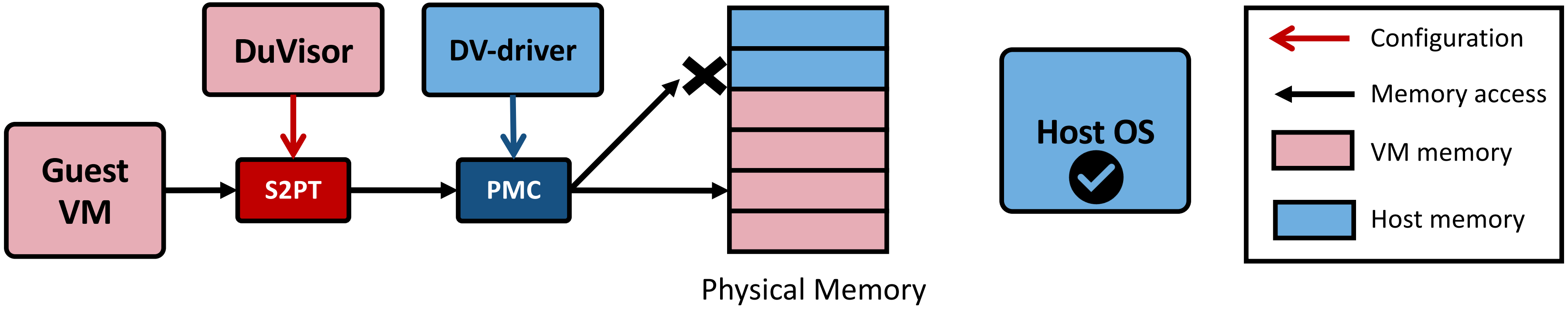
- Handle stage-2 page faults and provide memory virtualization in user mode without involving the kernel
- A malicious DuVisor may misconfigure the stage-2 page table



**S2PT stands for stage-2 page table*

Restricted Memory Virtualization

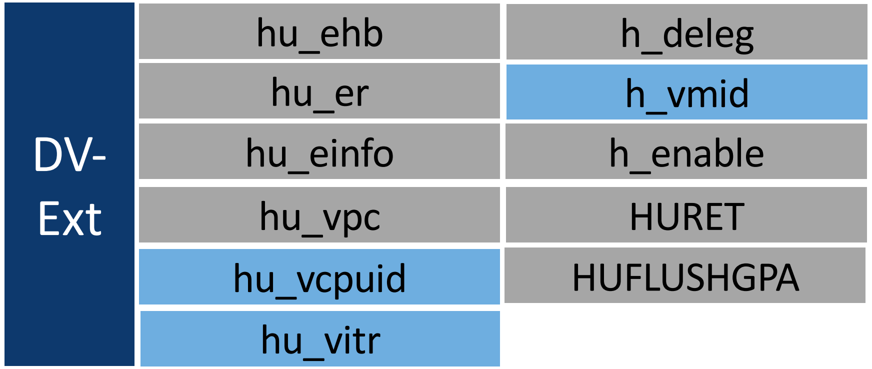
- Handle stage-2 page faults and provide memory virtualization in user mode without involving the kernel
- A malicious DuVisor may misconfigure the stage-2 page table
- Physical Memory Checking (PMC) limits the HPA that the VMs can access



*S2PT stands for stage-2 page table

DV-Ext Implementation

- Platform
  - RISC-V Rocket Core
  - 16KB L1 ICache, 16KB L1 DCache, 512KB shared L2 cache
  - 16GB DRAM
- Non-intrusive modifications
  - Reuse registers and instructions from **RISC-V N-Ext** and **H-Ext**
  - **481 lines of Chisel** to support DVE based on **RISC-V H-Ext**
  - **14 lines of Chisel** to support PMC based on **RISC-V PMP**
  - Only 3 registers implemented from scratch



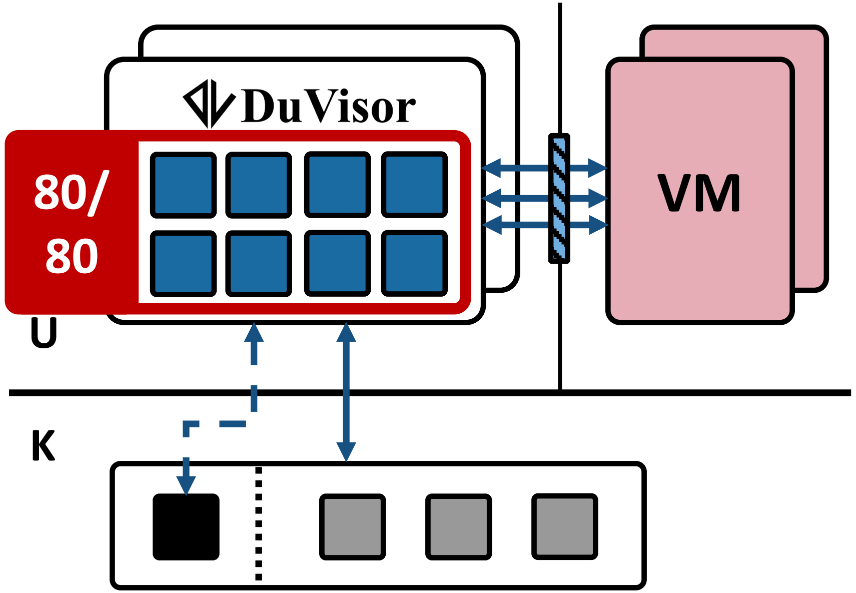
DV-Ext	hu_ehb	h_deleg
	hu_er	h_vmid
	hu_einfo	h_enable
	hu_vpc	HURET
	hu_vcpuid	HUFLUSHGPA
	hu_vitr	

Software Implementation

- DuVisor hypervisor
  - 7,128 LoC
    - **5,052 lines of Rust**, 166 lines of assembly, and 1,910 lines of C
  - Virtualization of CPU, memory, and interrupt
    - 4,984 lines of Rust and 166 lines of assembly
- DV-driver
  - A tiny Linux kernel module with **only 337 LoC**
  - **IOCTLs** for DuVisor

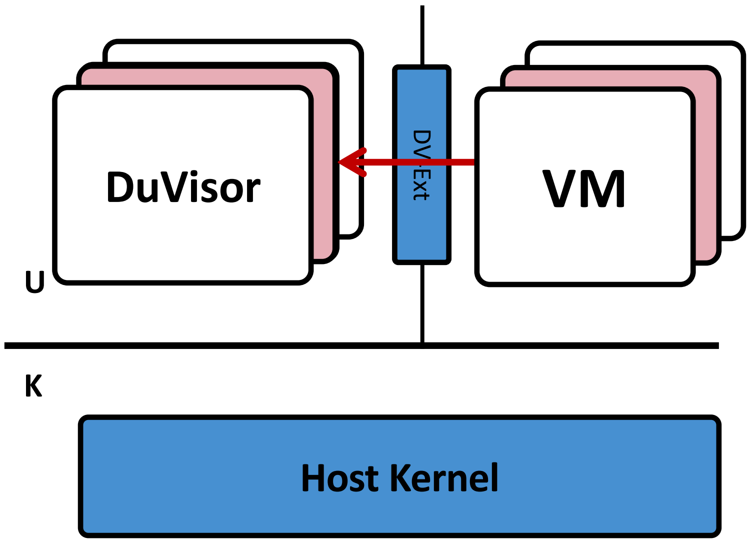
Security Evaluation

- Attack from guest to host kernel
  - All host-attacking CVEs are moved to user mode
    - Prior works: 47 left



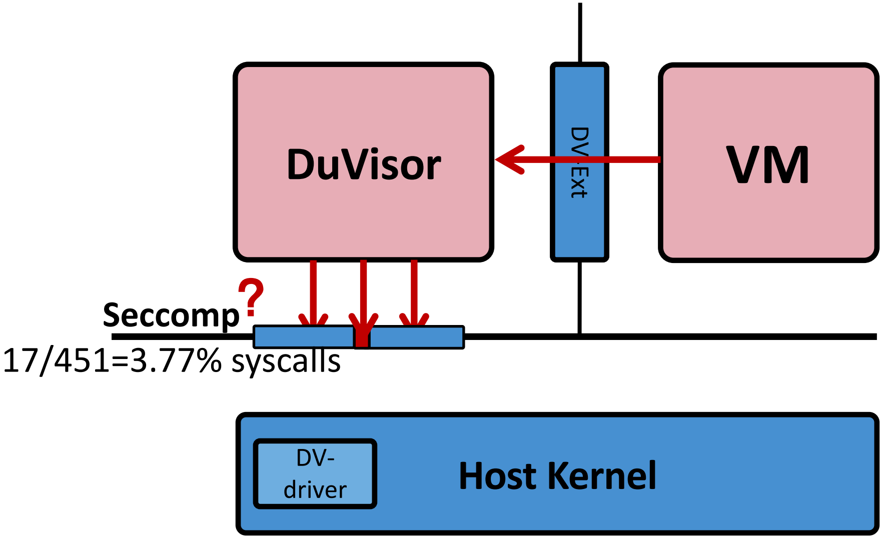
Security Evaluation

- Attack from guest to host kernel
- Attack from guest to DuVisor



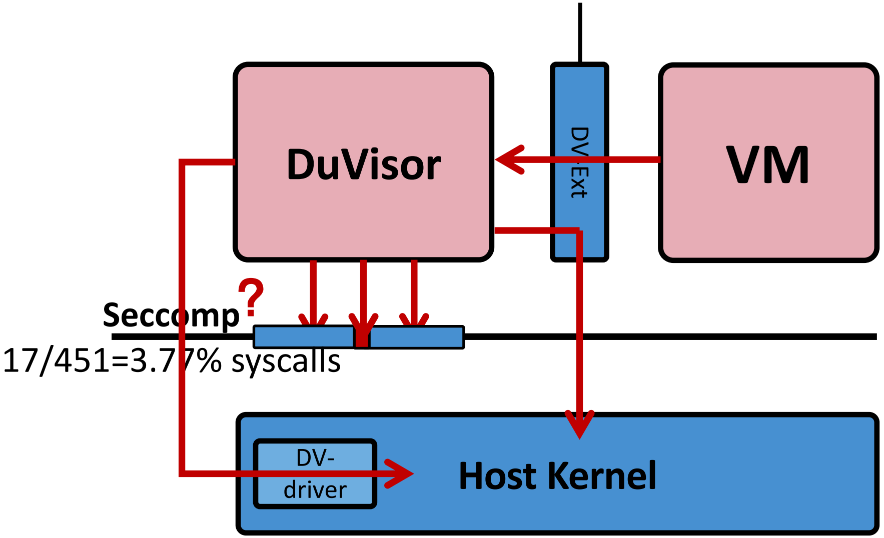
Security Evaluation

- Attack from guest to host kernel
- Attack from guest to DuVisor
- Attack from DuVisor to host kernel



Security Evaluation

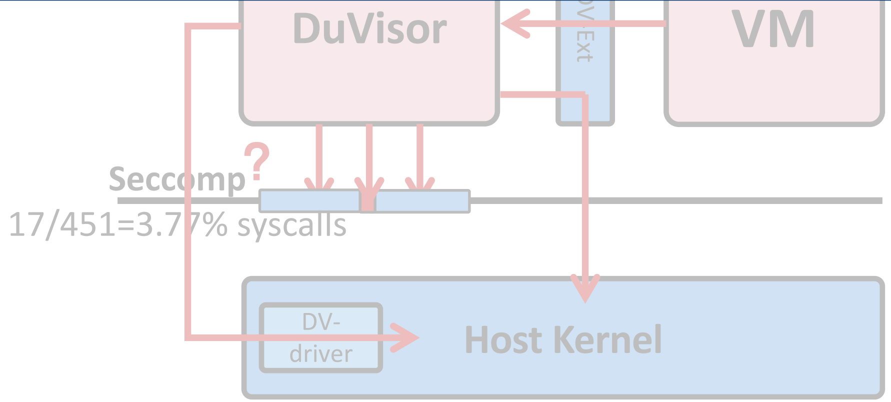
- Attack from guest to host kernel
- Attack from guest to DuVisor
- Attack from DuVisor to host kernel



Security Evaluation

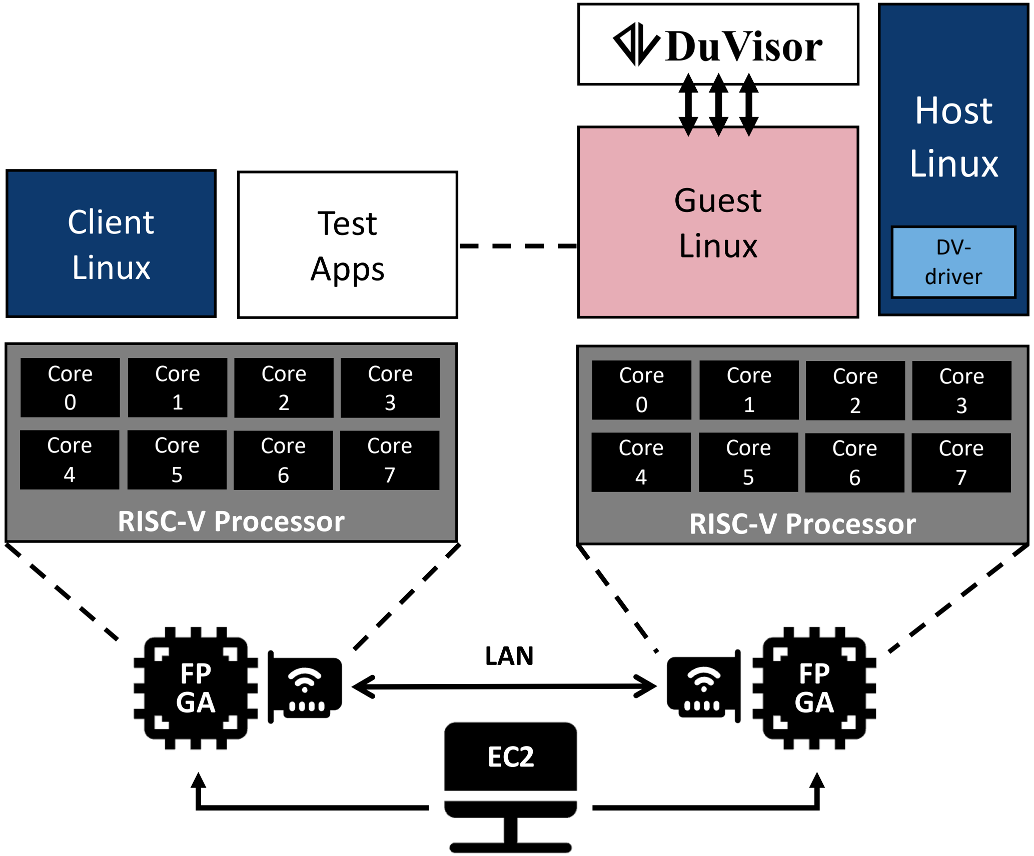
- Attack from guest to host kernel
- Attack from guest to DuVisor
- Attack from DuVisor to host kernel

Protect the host kernel from malicious guests

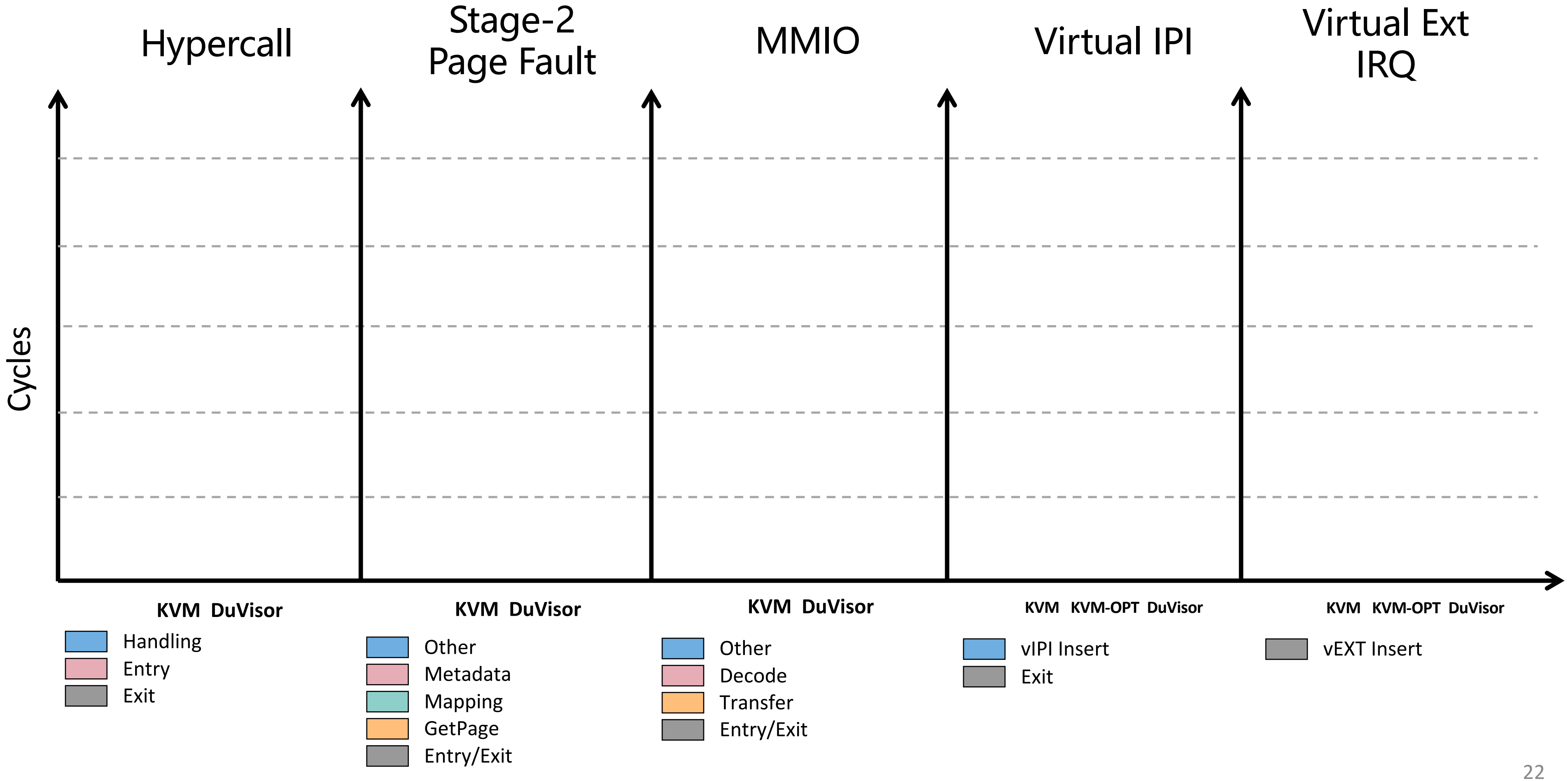


Performance Evaluation

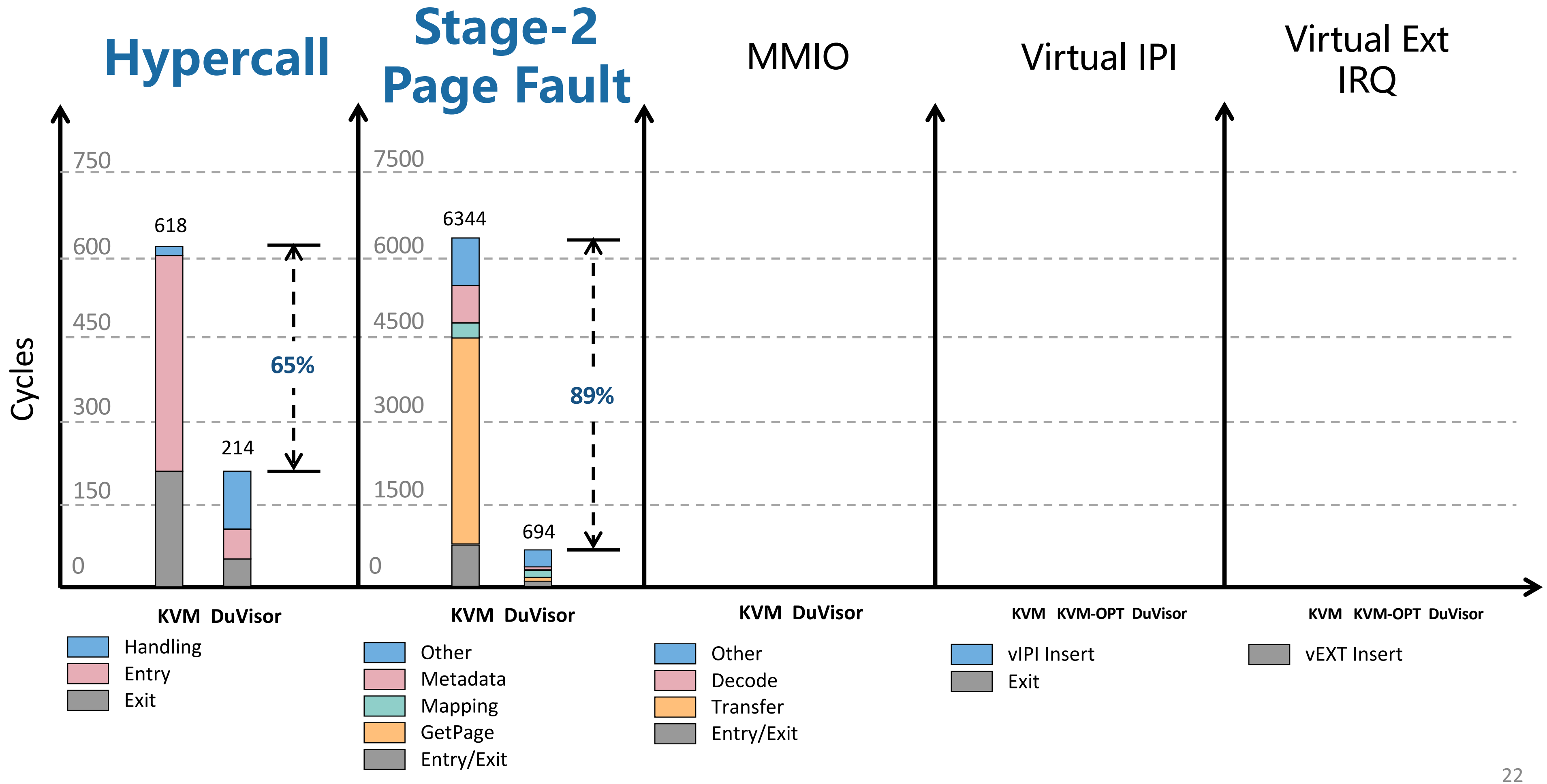
- Hardware
  - Firesim platform
    - **Two FPGA boards**
    - **8 RISC-V processors** on each FPGA board
    - 16GB RAM & 115GB storage
    - Local area network with IceNICs
    - Support for **H-Ext** and **DV-Ext**
- Software
  - Firmware
    - OpenSBI v0.8
  - DuVisor
    - Run in user mode
    - Linux equipped with the **DV-driver**
  - Baseline
    - **QEMU/KVM with H-Ext support**



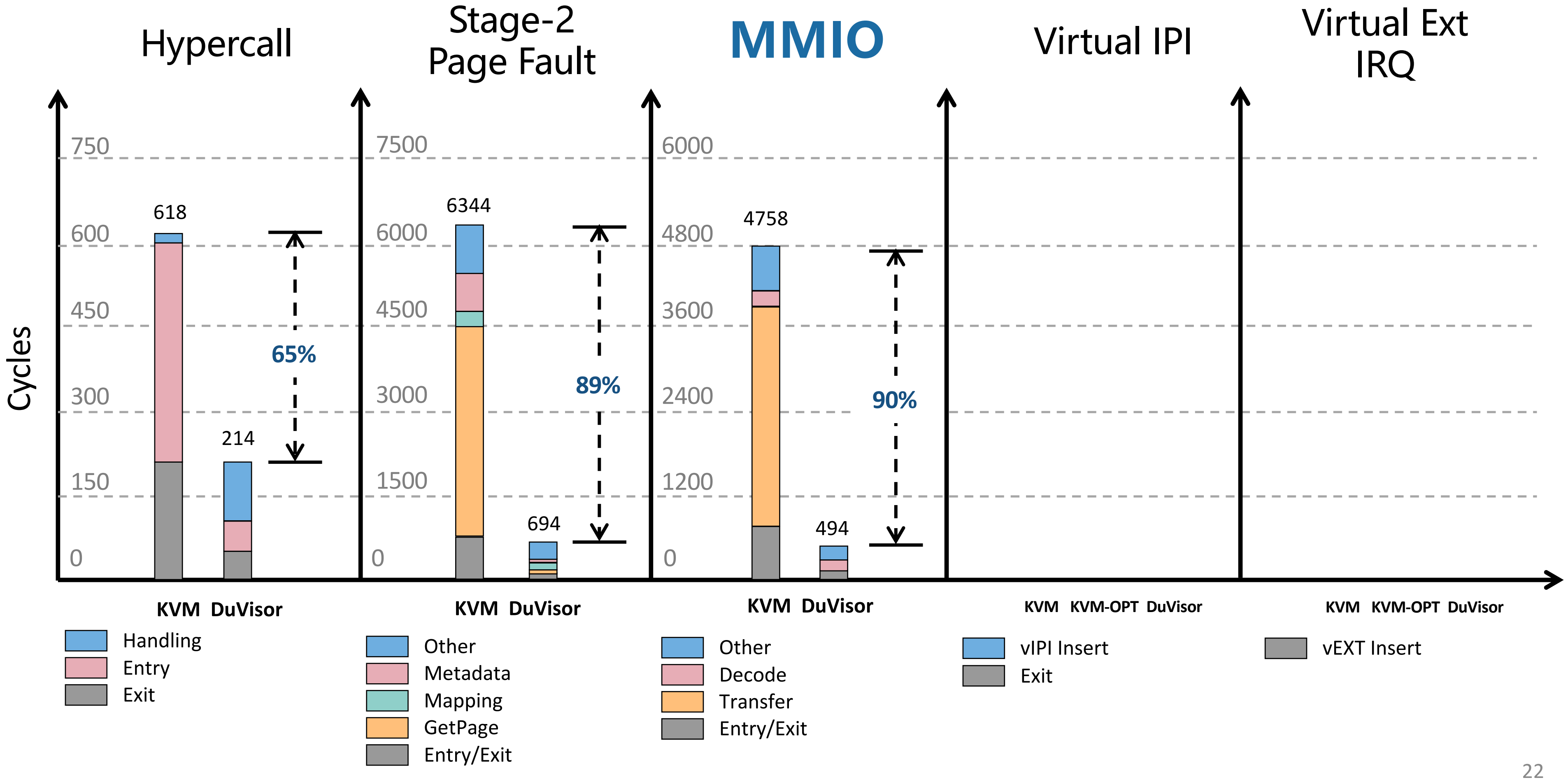
Microbenchmarks



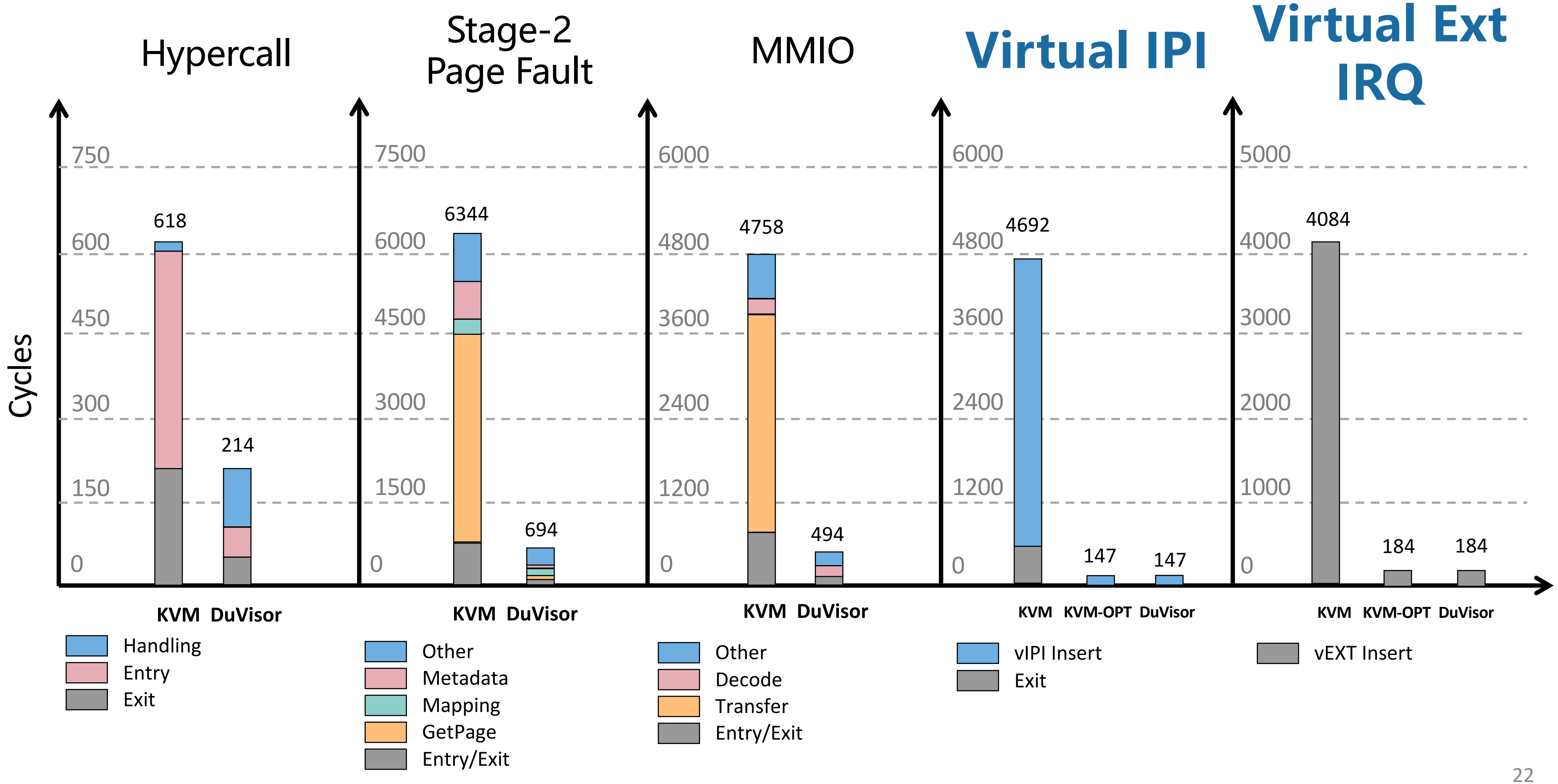
Microbenchmarks



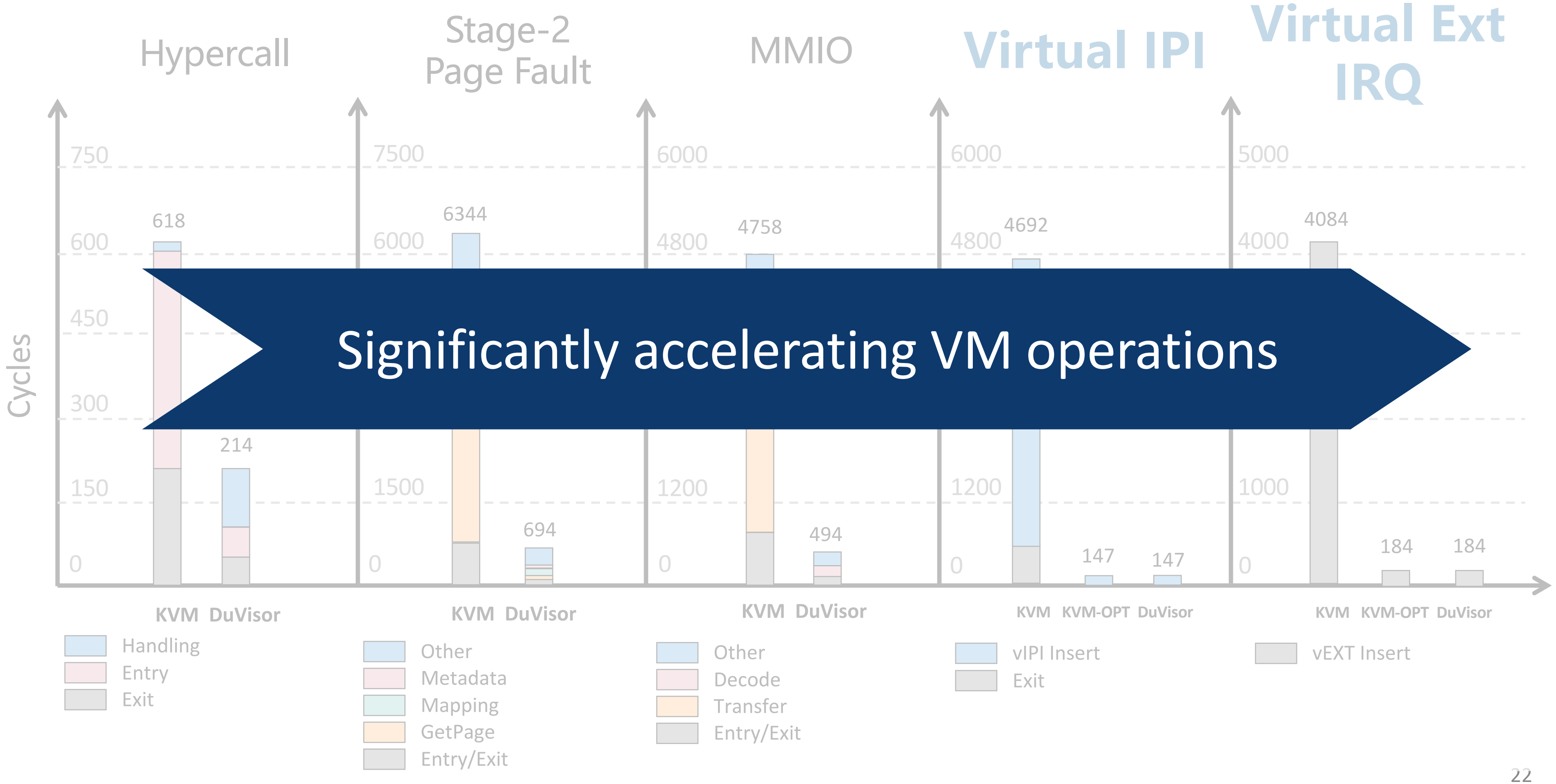
Microbenchmarks



Microbenchmarks

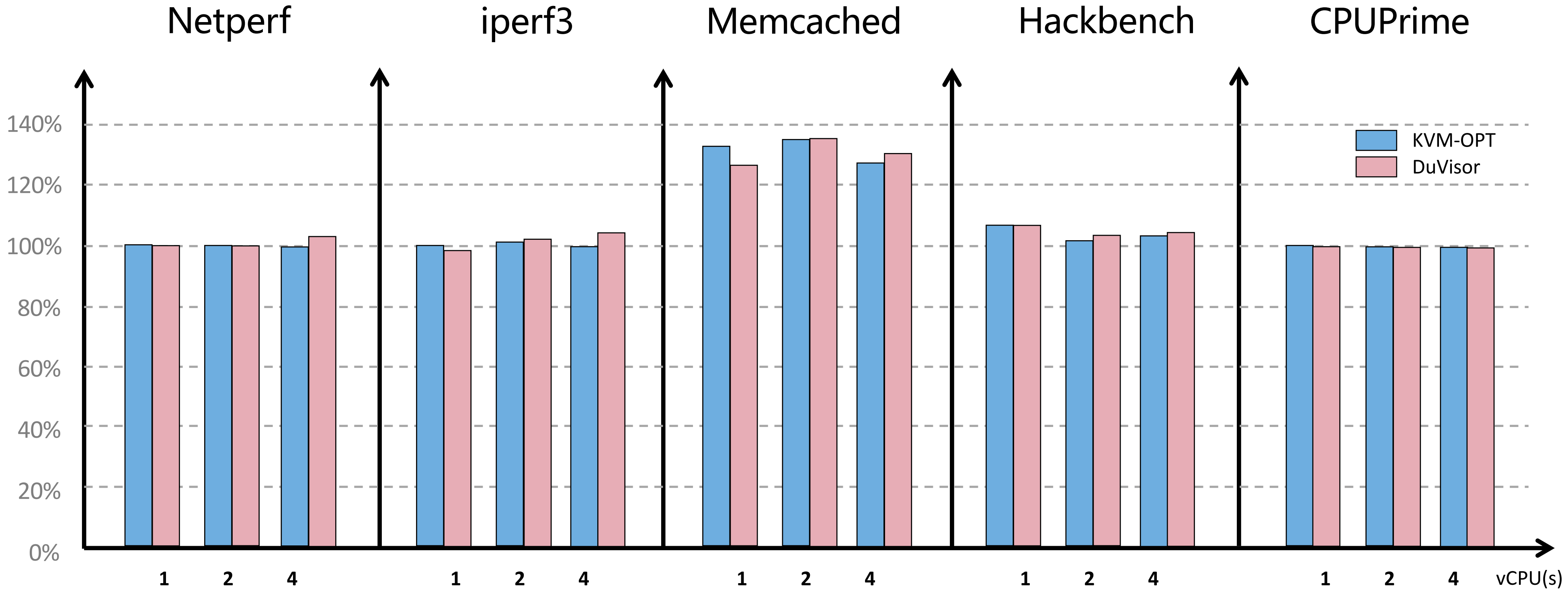


Microbenchmarks

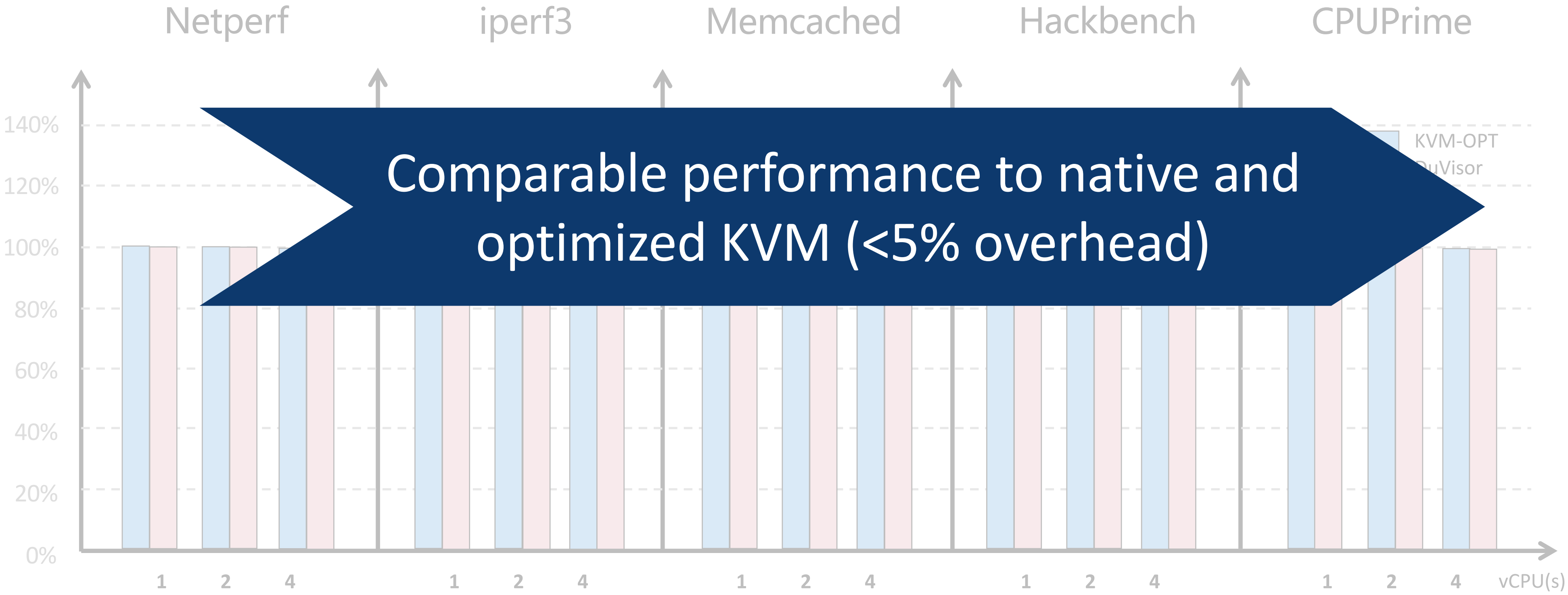


Significantly accelerating VM operations

Application Benchmarks



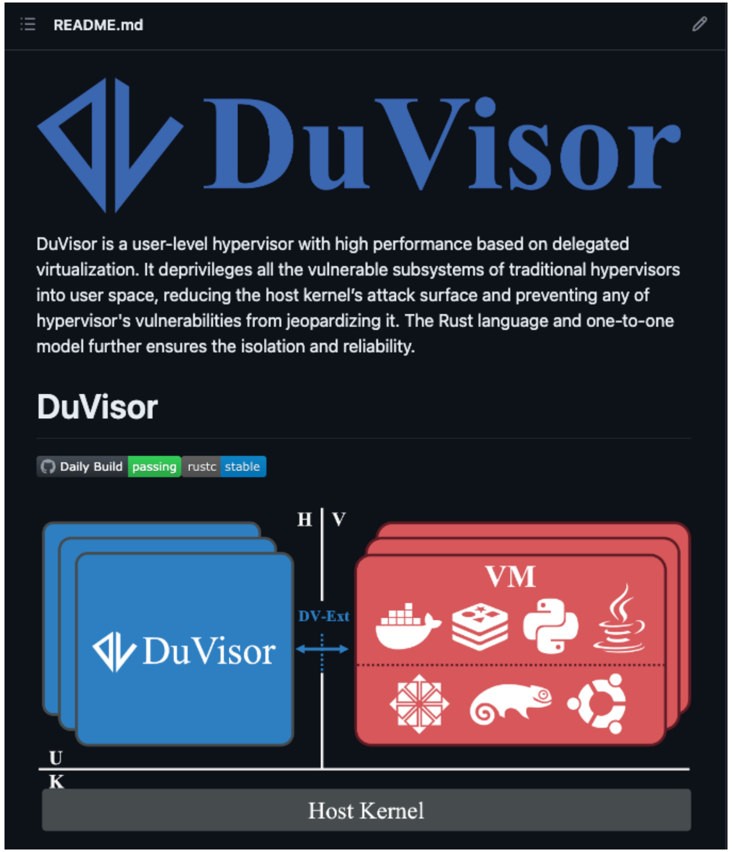
Application Benchmarks



Comparable performance to native and optimized KVM (<5% overhead)

Conclusion

- Delegated **User-level** Virtualization
  - A new direction for secure virtualization research and development
  - DV-Ext securely exposes hardware interfaces to user space
  - **DuVisor, a user-level hypervisor**, directly serves VM-hypervisor interactions in user space
  - Protection for the host kernel (& VMs) without performance degradation
- Open Source
  - https://github.com/IPADS-DuVisor/DuVisor
  - Firesim & QEMU



README.md

DuVisor

DuVisor is a user-level hypervisor with high performance based on delegated virtualization. It deprivileges all the vulnerable subsystems of traditional hypervisors into user space, reducing the host kernel's attack surface and preventing any of hypervisor's vulnerabilities from jeopardizing it. The Rust language and one-to-one model further ensures the isolation and reliability.

DuVisor

Daily Build passing rustc stable

Thanks

Q&A

SHANGHAI JIAO TONG UNIVERSITY 1896

IPADS
INSTITUTE OF PARALLEL
AND DISTRIBUTED SYSTEMS