CHINA LINUX KERNEL
中国Linux内核开发者大会

華中科技大學
网络空间安全学院
School of Cyber Science and Engineering, HUST

第19届中国Linux内核开发者大会

赞助单位

華中科技大學 | HUAWEI | 阿里云 | FUJITSU | 腾讯云
oppo | 火山引擎 | vivo | 龙芯中科 LOONGSON TECHNOLOGY | 蚂蚁集团 ANT GROUP

支持单位

清华大学 Tsinghua University | 迪捷软件 DIGIPROTO | OpenEuler | OpenAnolis 龙蜥社区 | OpenCloudOS

支持社区&媒体

CSDN | 云巅论剑 | sf 思否 | 阅码场 yomocode | InfoQ 极客传媒
51CTO | 开源江湖

2024年10月　湖北 • 武汉

華中科技大學

RISC-V架构下的PMU虚拟化直通方案

路旭
字节跳动虚拟化工程师

01 PMU虚拟化背景介绍

02 RISC-V PMU演进&现状

03 RISC-V PMU直通方案

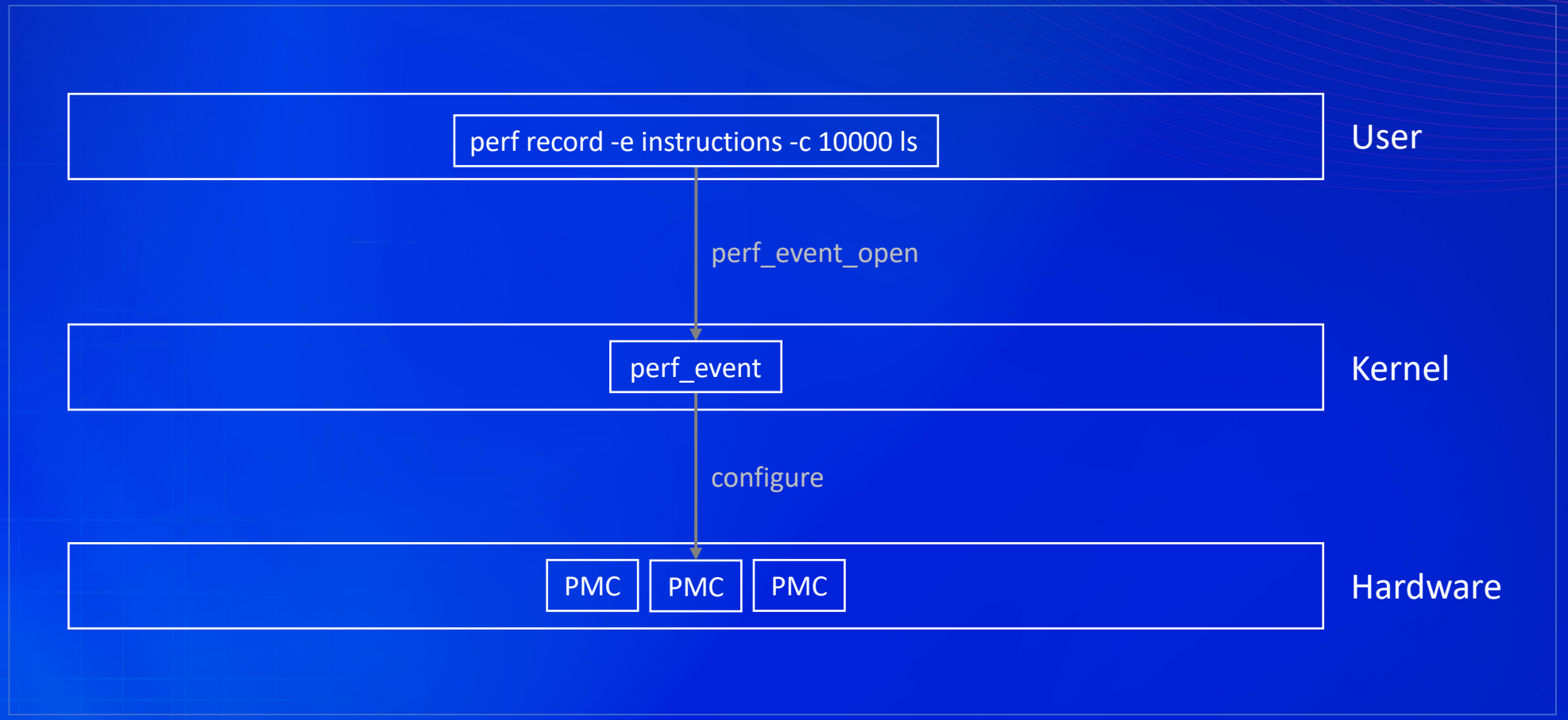
01 PMU虚拟化背景介绍

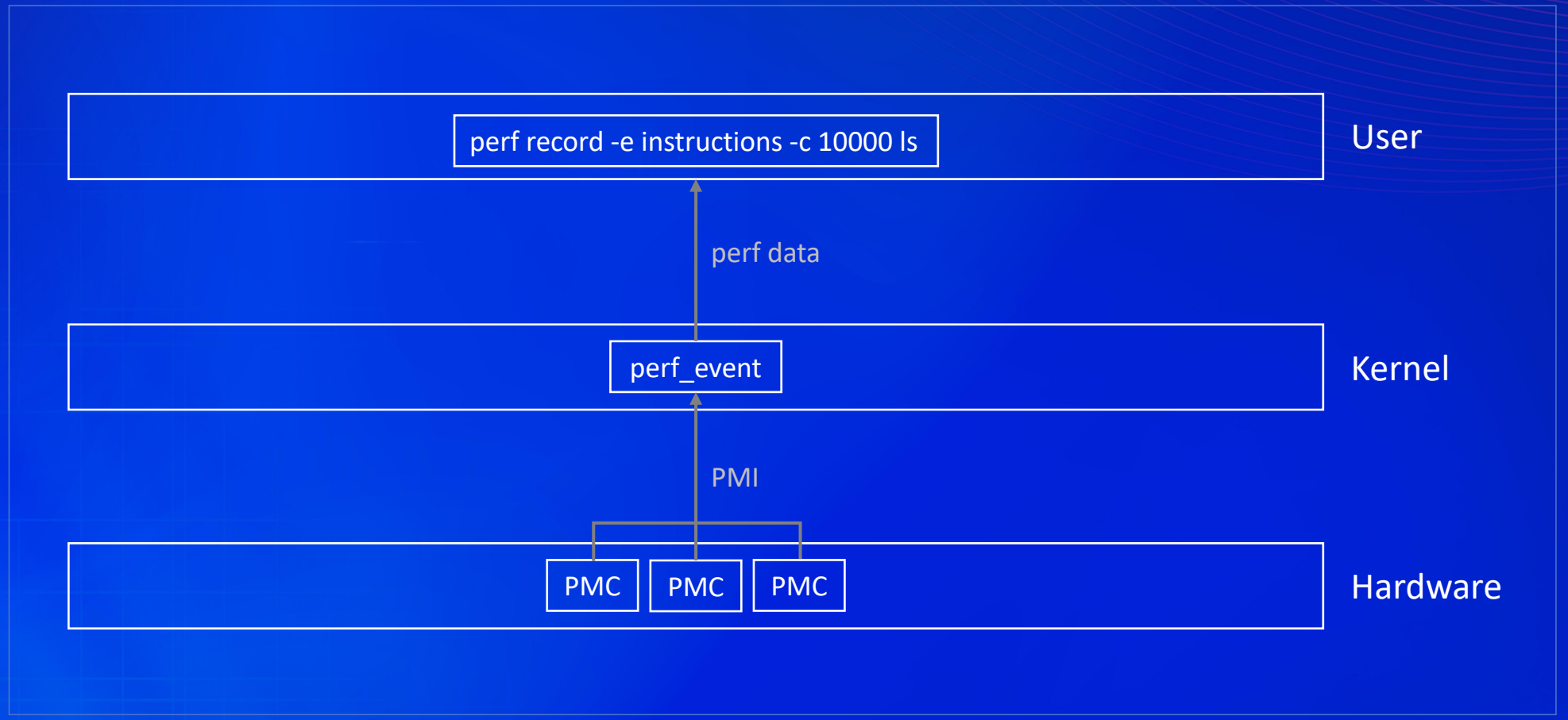
02 RISC-V PMU演进&现状

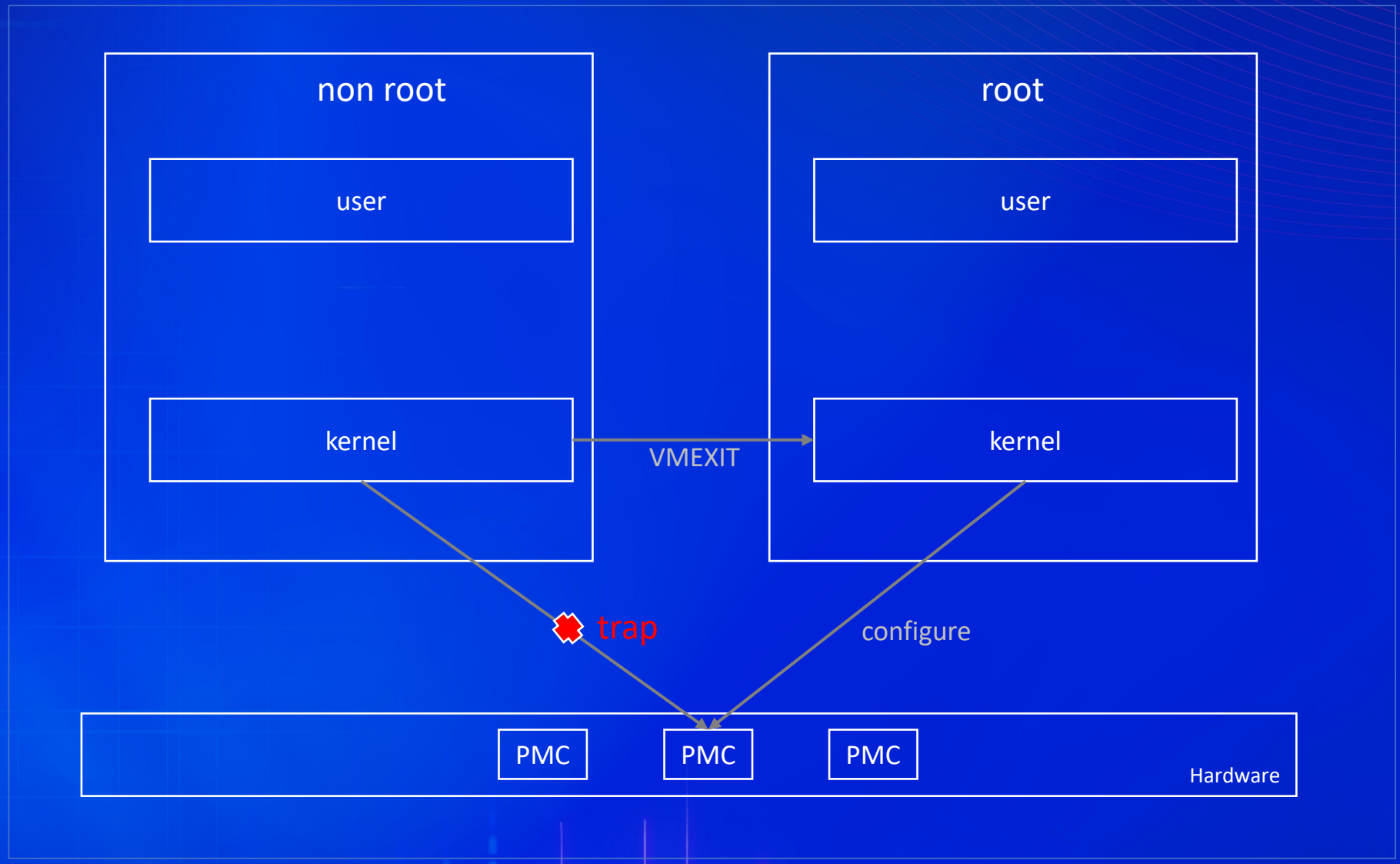
03 RISC-V PMU直通方案

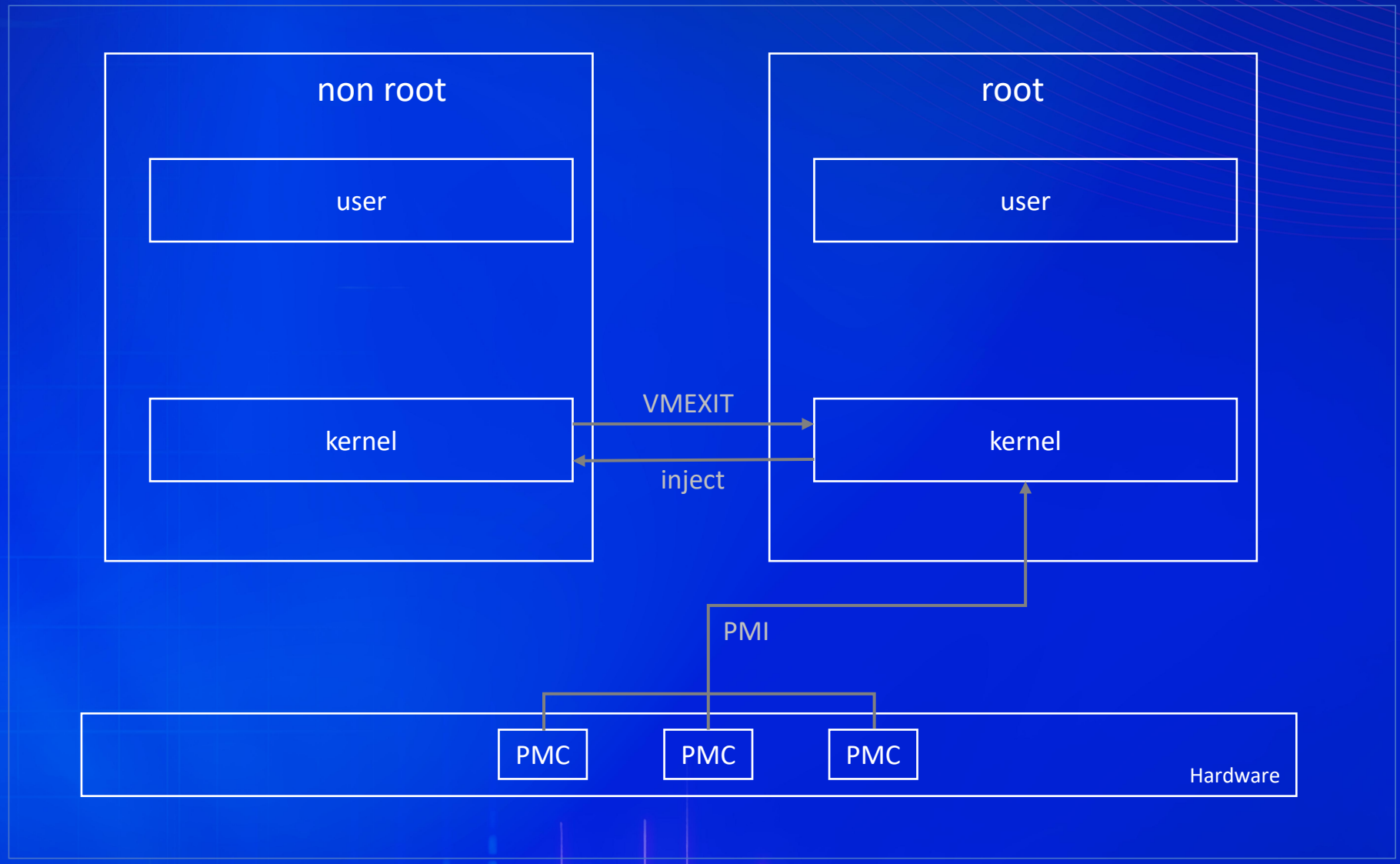
perf list

```
cpu-cycles OR cycles                                [Hardware event]
Instructions                                        [Hardware event]
branch-instructions OR branches                     [Hardware event]
…

alignment-faults                                    [Software event]
page-faults OR faults                               [Software event]
cpu-migrations OR migrations                        [Software event]
…

L1-dcache-load-misses                               [Hardware cache event]
L1-dcache-loads                                     [Hardware cache event]
L1-dcache-stores                                    [Hardware cache event]
…

alarmtimer:alarmtimer_cancel                        [Tracepoint event]
alarmtimer:alarmtimer_fired                         [Tracepoint event]
alarmtimer:alarmtimer_start                         [Tracepoint event]
…
```


perf record -e instructions -c 10000 ls
User
perf_event_open
perf_event
Kernel
configure
PMC
PMC
PMC
Hardware


perf record -e instructions -c 10000 ls
User
perf data
perf_event
Kernel
PMI
PMC
PMC
PMC
Hardware

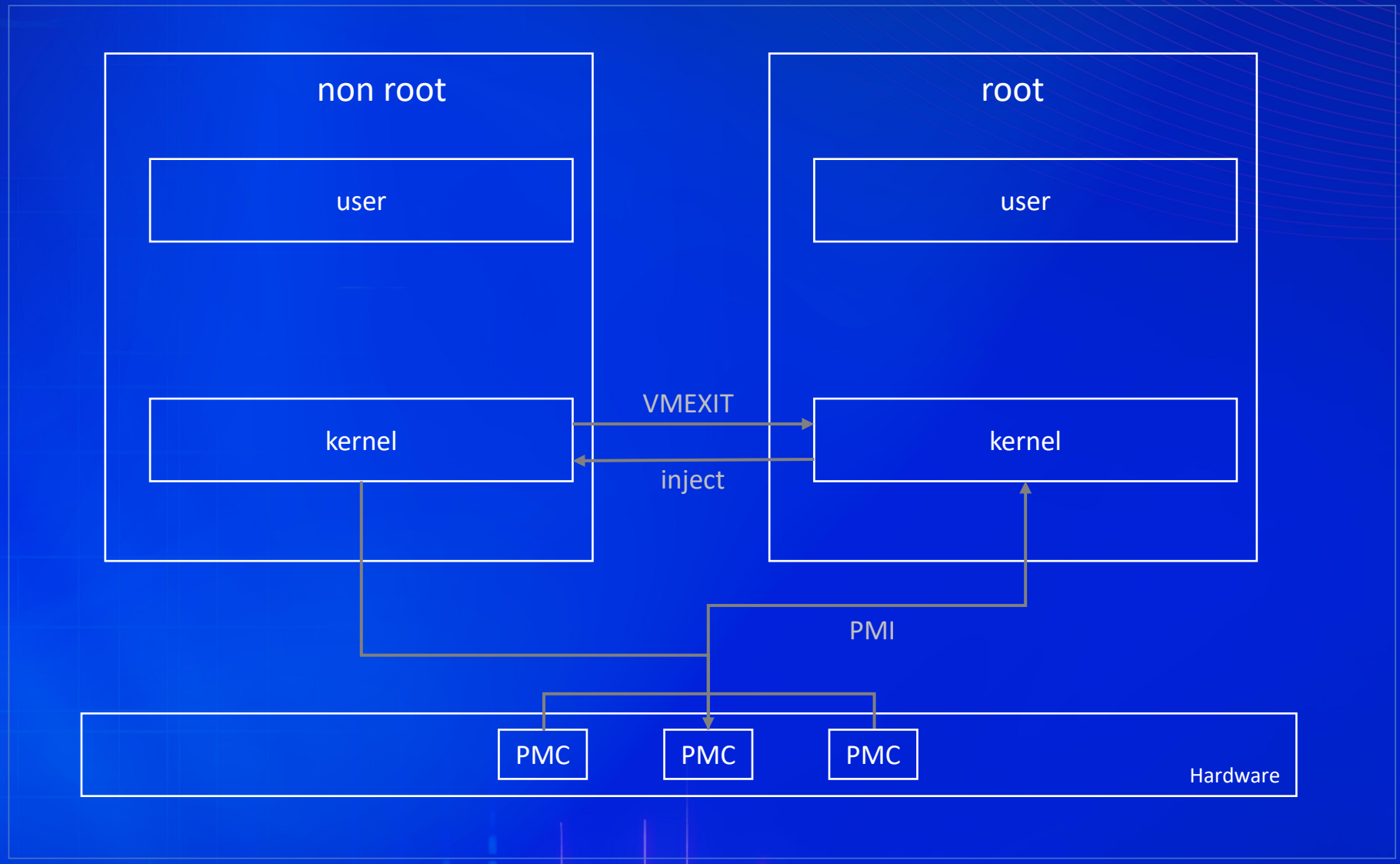

non root
user
kernel
root
user
kernel
VMEXIT
trap
configure
PMC
PMC
PMC
Hardware



sample per-cgroup instructions&cycles in VM: 1.03 millions rdpmc VMEXITs / 1.68 million total VMEXITs

Problem1: PMU not fully passed through

Problem2: Host loses profile capability



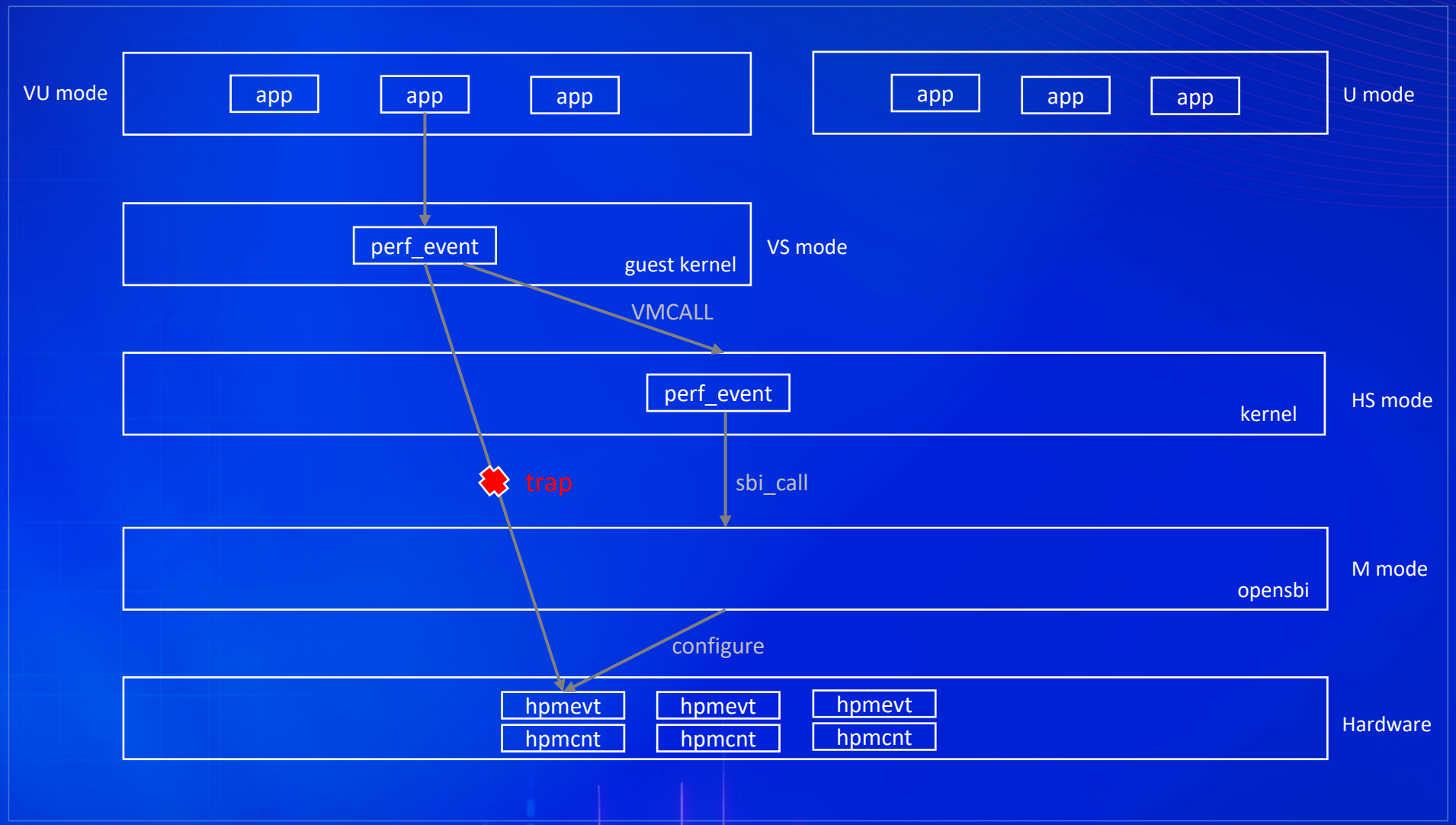
PMC Passthrough: https://lwn.net/Articles/959653/

01 PMU虚拟化背景介绍

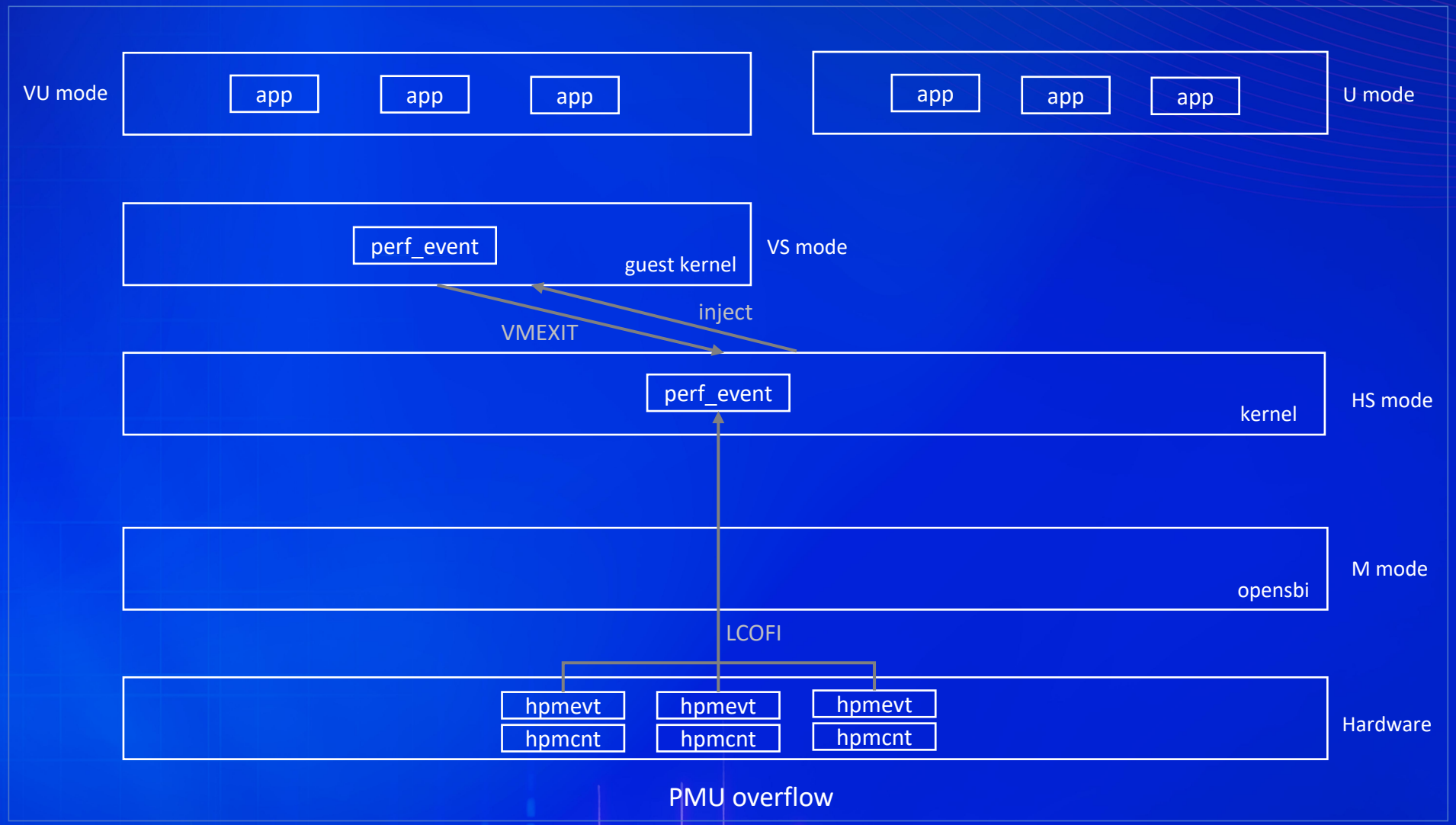
02 RISC-V PMU演进&现状

03 RISC-V PMU直通方案

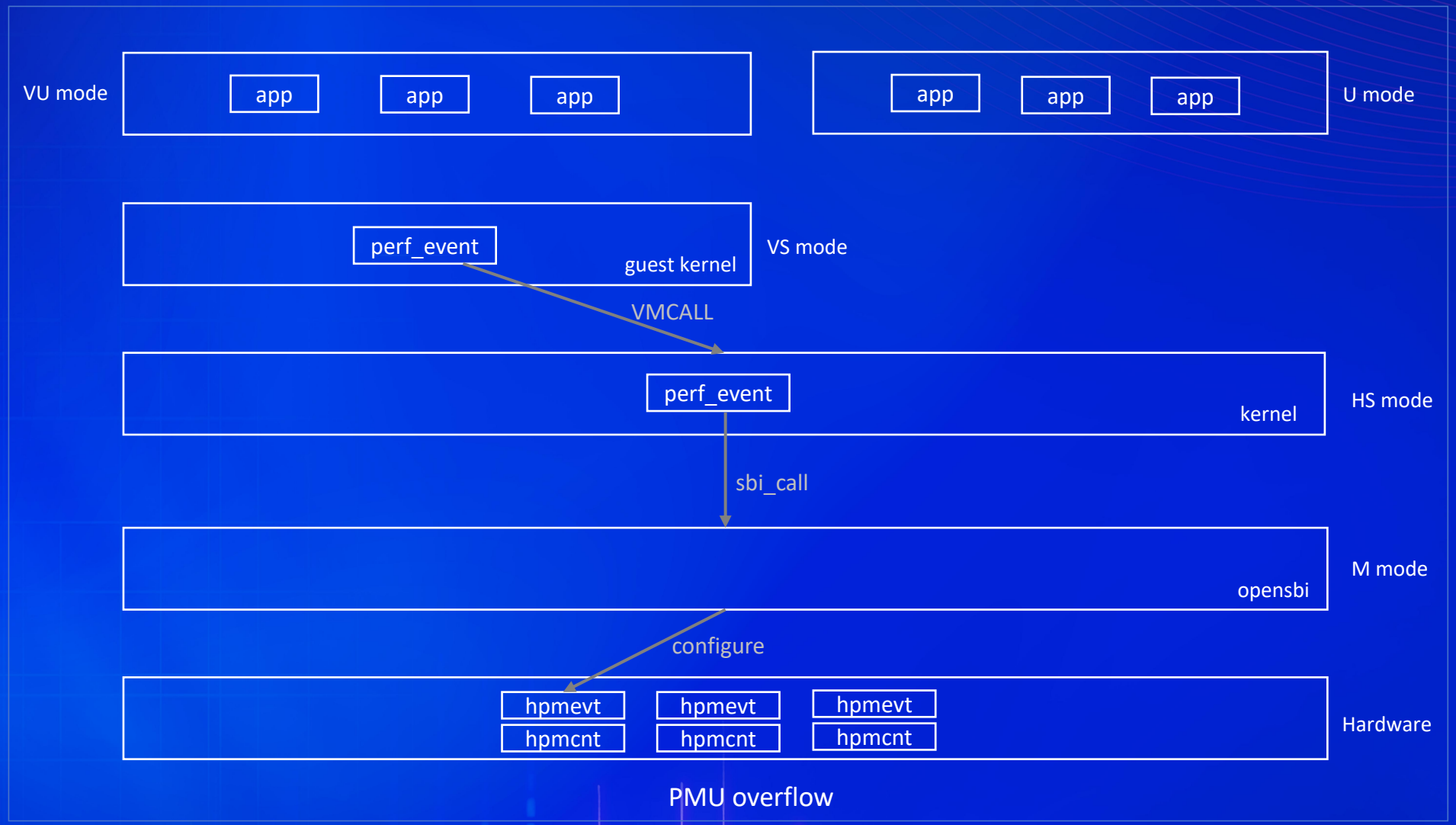
Original Core PMU Architecture



Original Core PMU Architecture

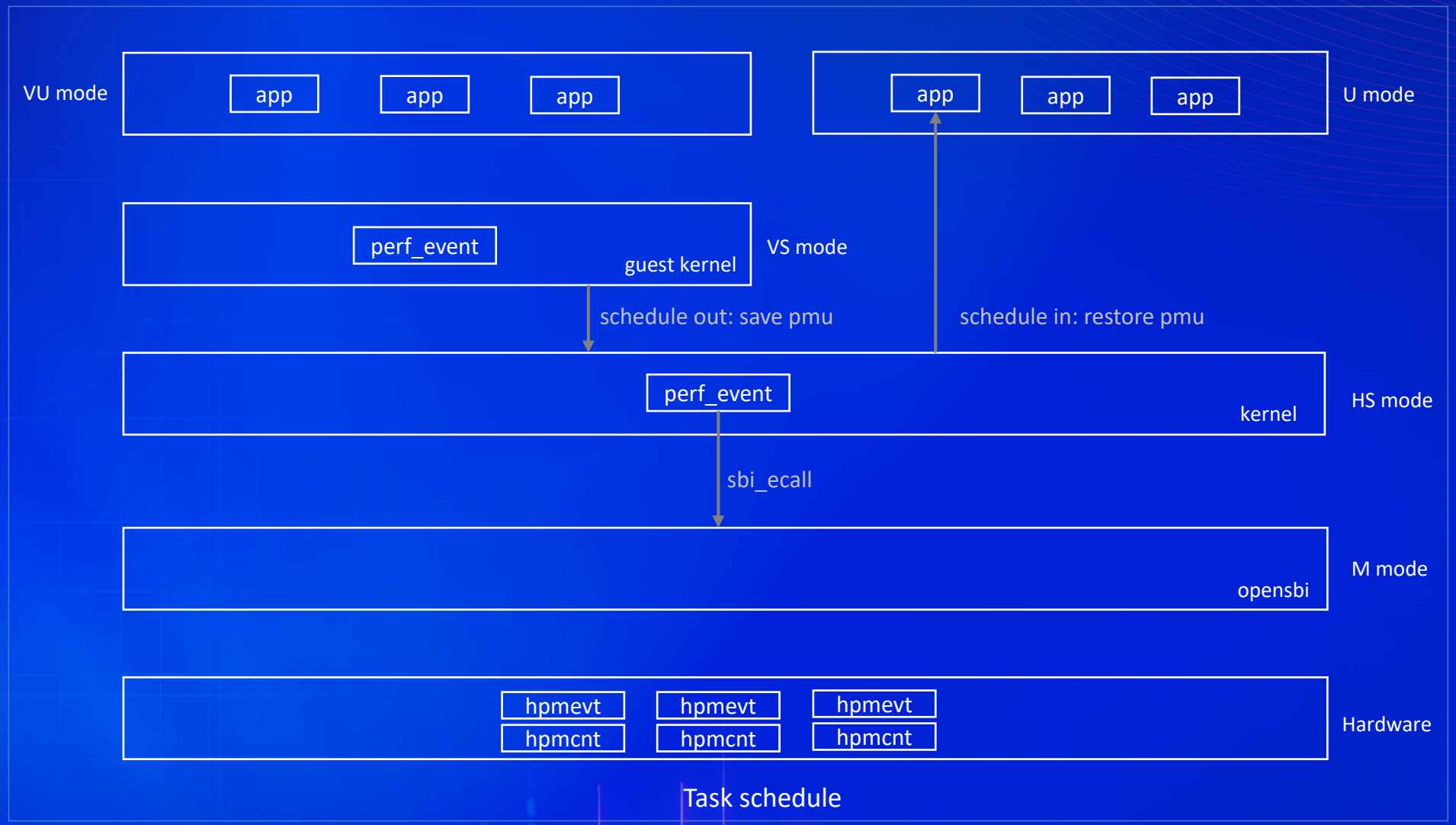


Original Core PMU Architecture



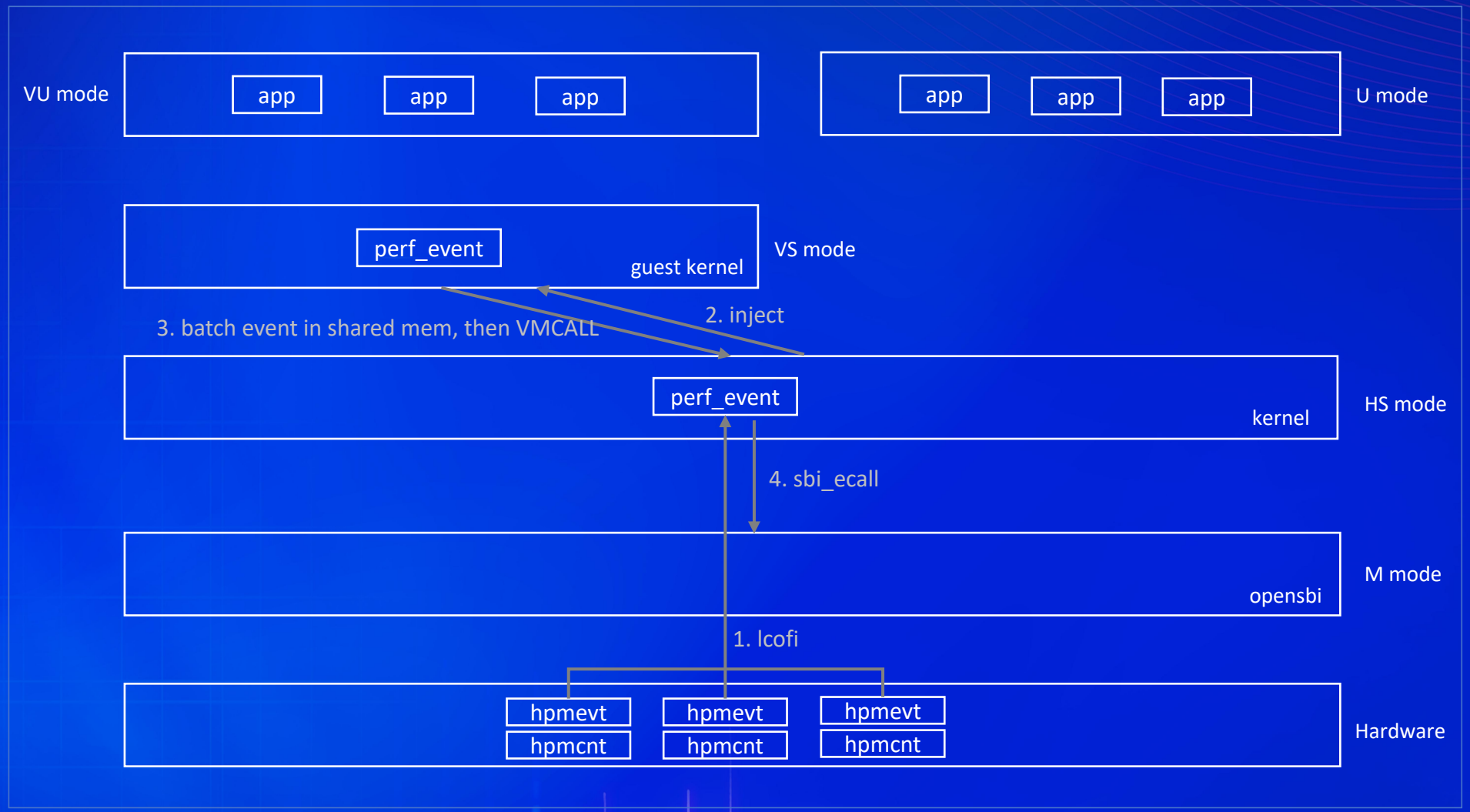
Original Core PMU Architecture

Original Core PMU Architecture



perf record -e instructions -c 1000000 ls: 5 ovf irqs, 57 VMEXITs, 202 sbi calls

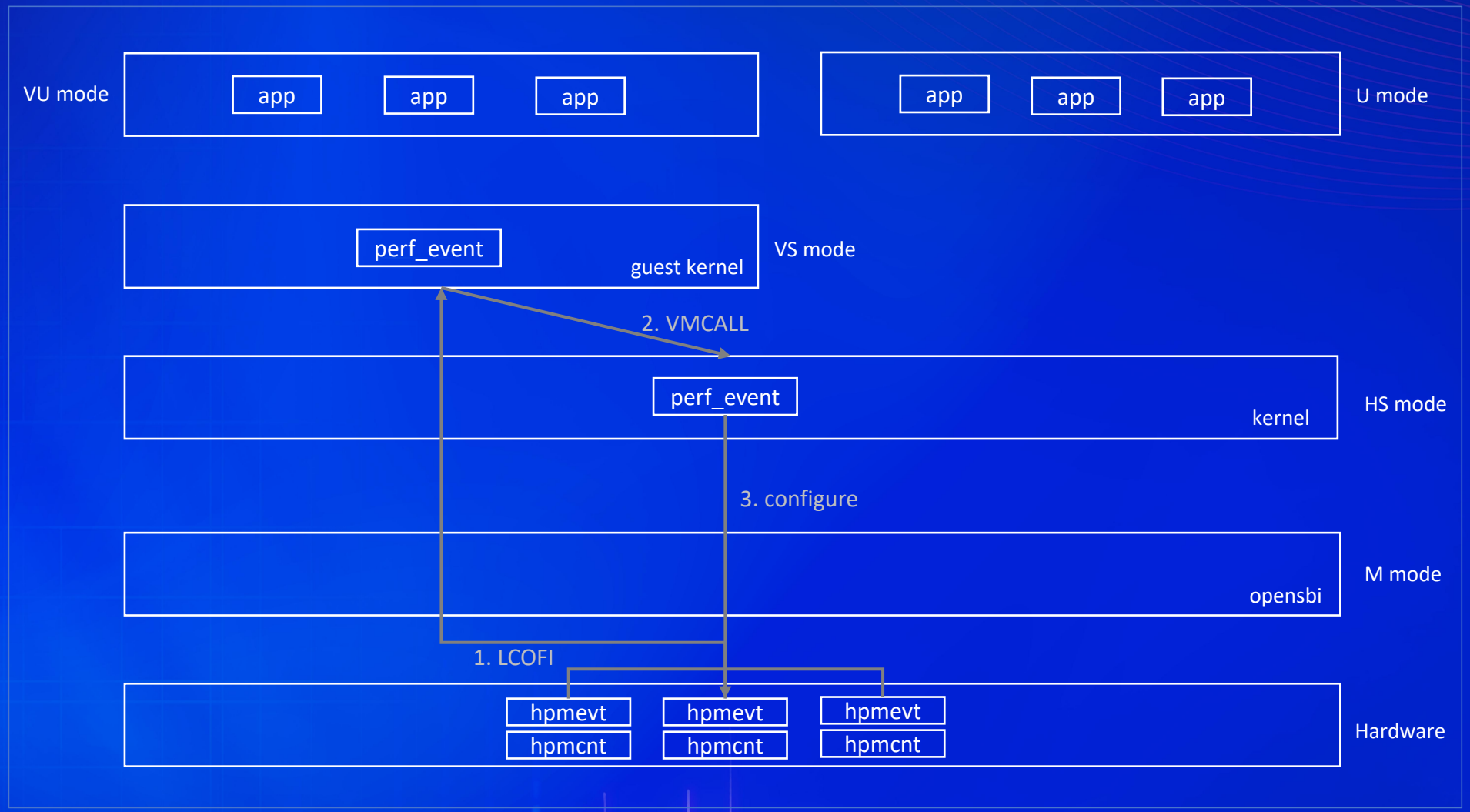
Original Core PMU Architecture + Snapshot



snapshot: https://lore.kernel.org/all/20240420151741.962500-10-atishp@rivosinc.com/

perf record -e instructions -c 1000000 ls: 5 ovf irqs, 58 VMEXITs, 204 sbi calls

Original Core PMU Architecture + Snapshot + Smcdeleg/Shlcofideleg



smcdeleg: https://lore.kernel.org/all/20240723-counter_delegation-v2-0-c4170a5348ca@rivosinc.com/

perf record -e instructions -c 1000000 ls: 5 ovf irqs, 47 VMEXITs, 0 sbi calls

Problem1: PMU not fully passed through

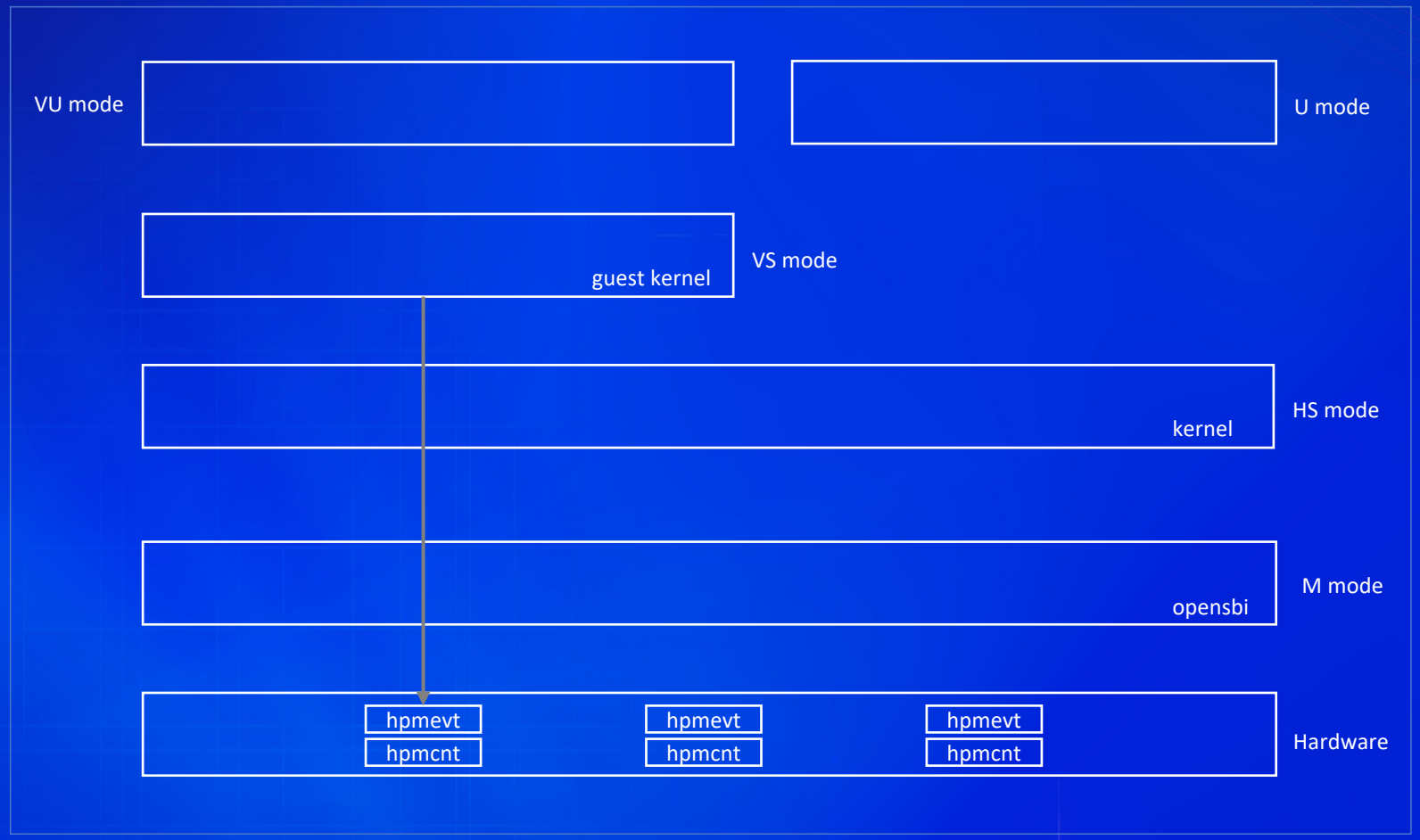
Problem2: Host loses profile capability

01 PMU虚拟化背景介绍

02 RISC-V PMU演进&现状

03 RISC-V PMU直通方案

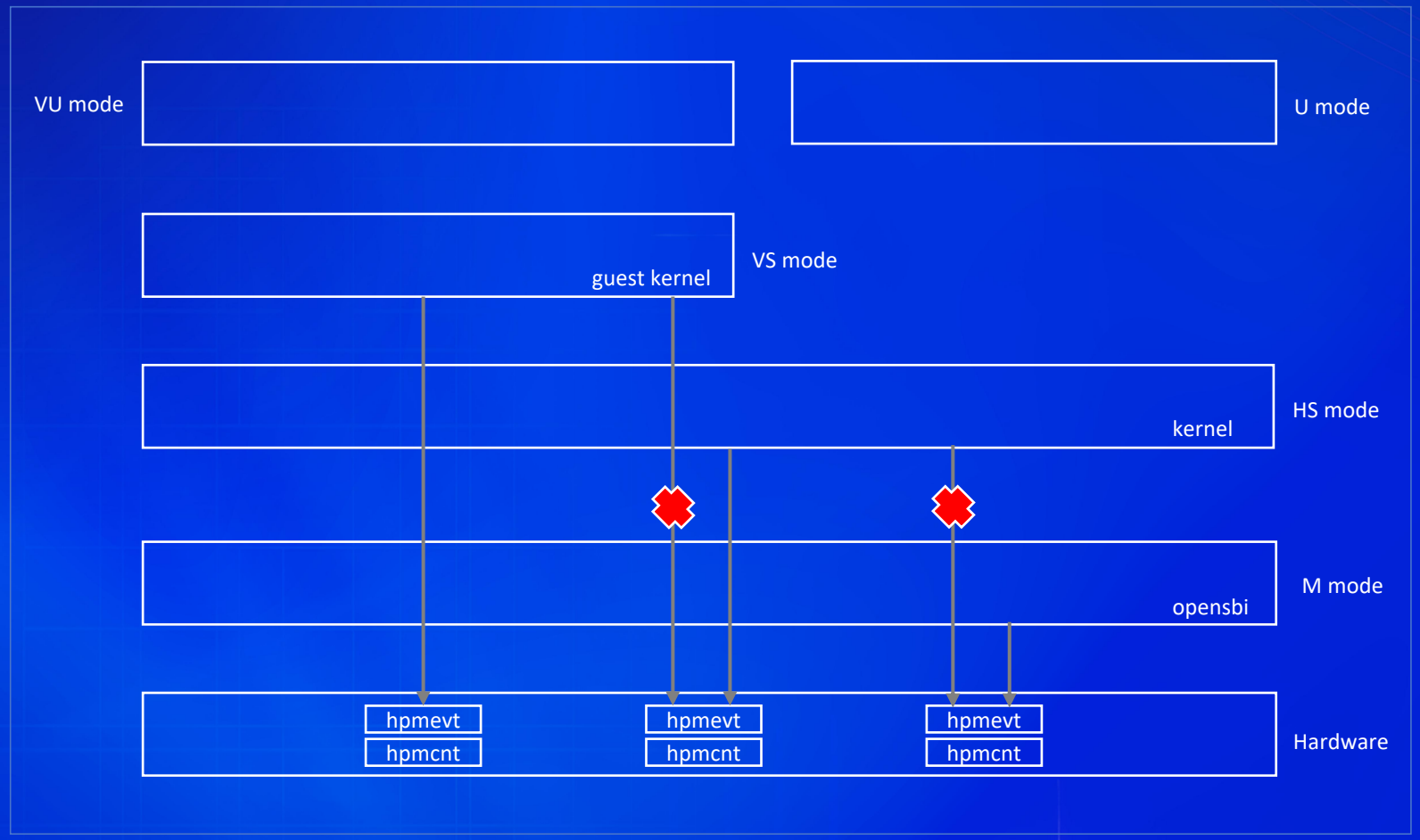
Solution for Problem1: PMU not fully passed through



- PMU CTRL & COUNT registers totally passed through
- For PMU registers passed through, hardware events limited to VS/VU mode only

Hardware requirement for thorough PMU passthrough

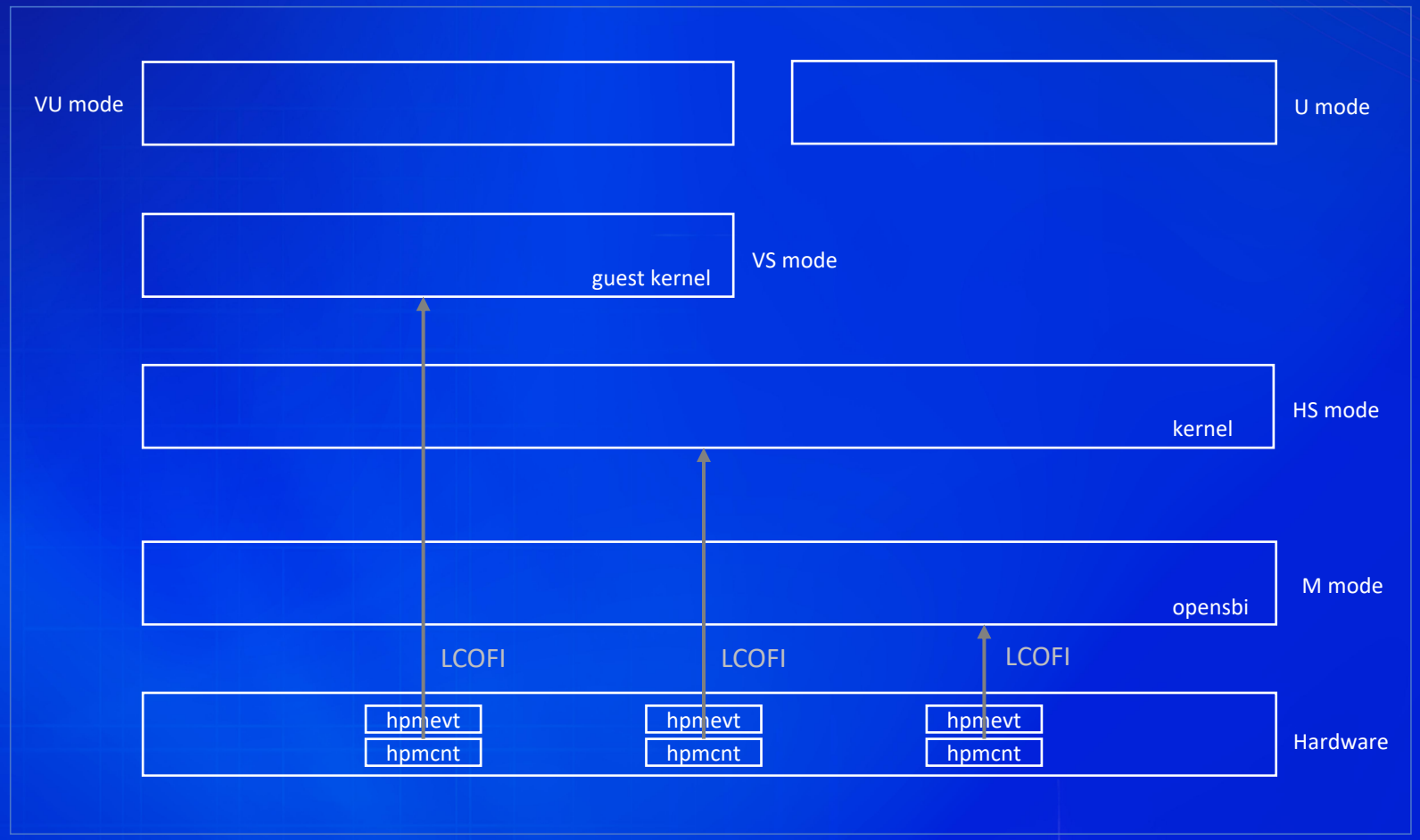
Solution for Problem2: Host loses profile capability



- PMU registers should support finer-grained delegation
- PMU overflow interrupt should support finer-grained delegation

Hardware requirement for host‘s profiling capabilities

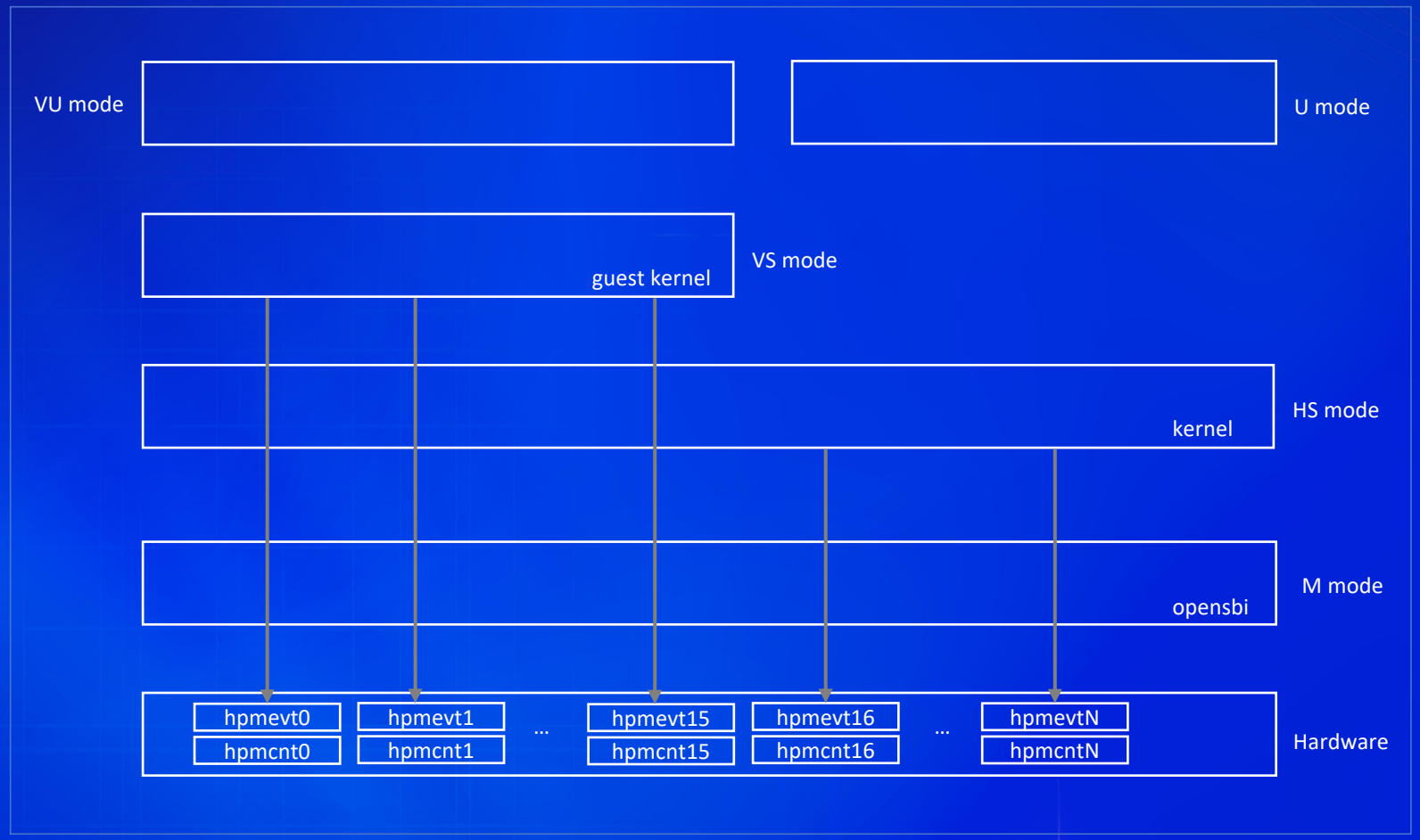
Solution for Problem2: Host loses profile capability



- PMU registers should support finer-grained delegation
- PMU overflow interrupt should support finer-grained delegation

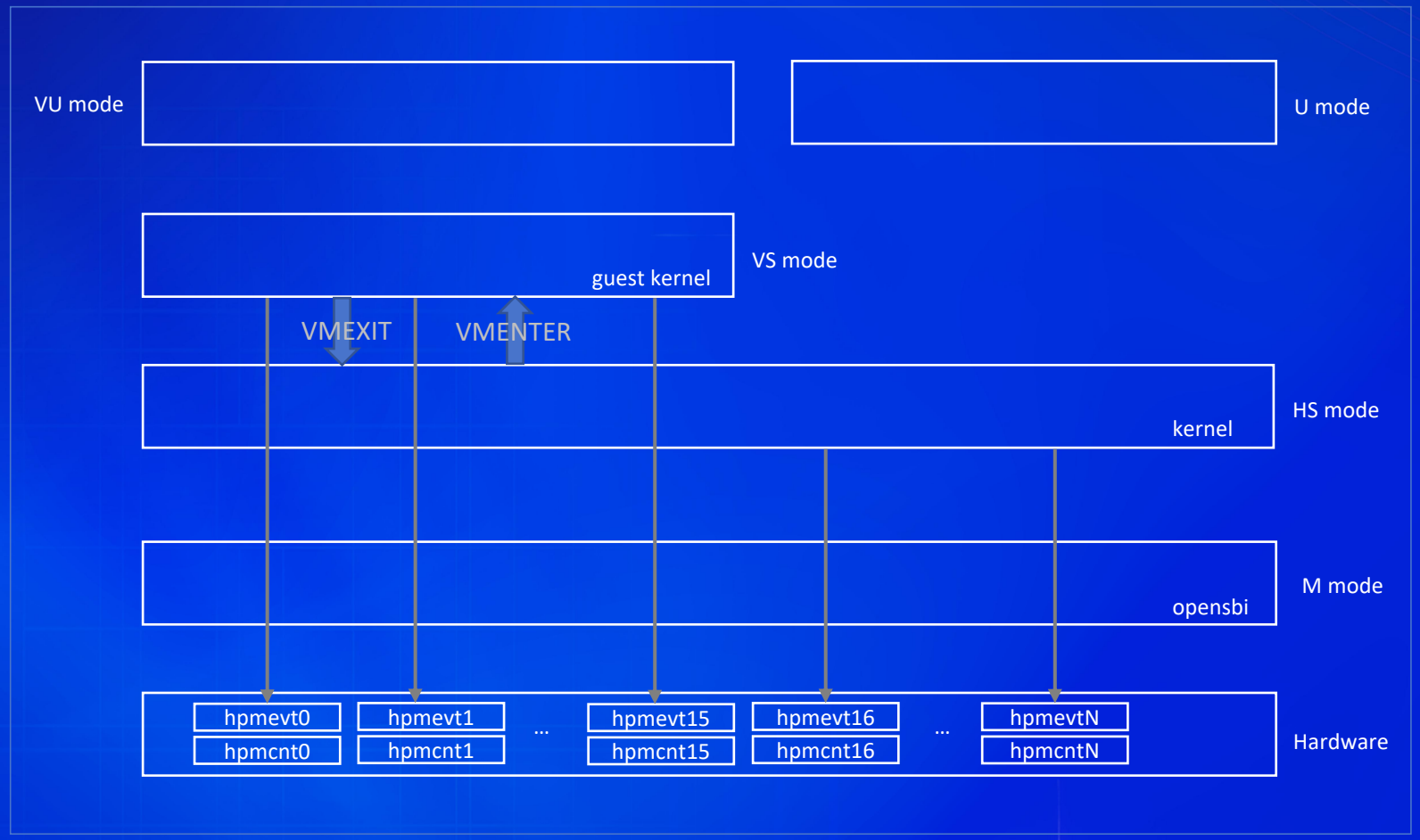
Hardware requirement for host‘s profiling capabilities

Software design for PMU passthrough



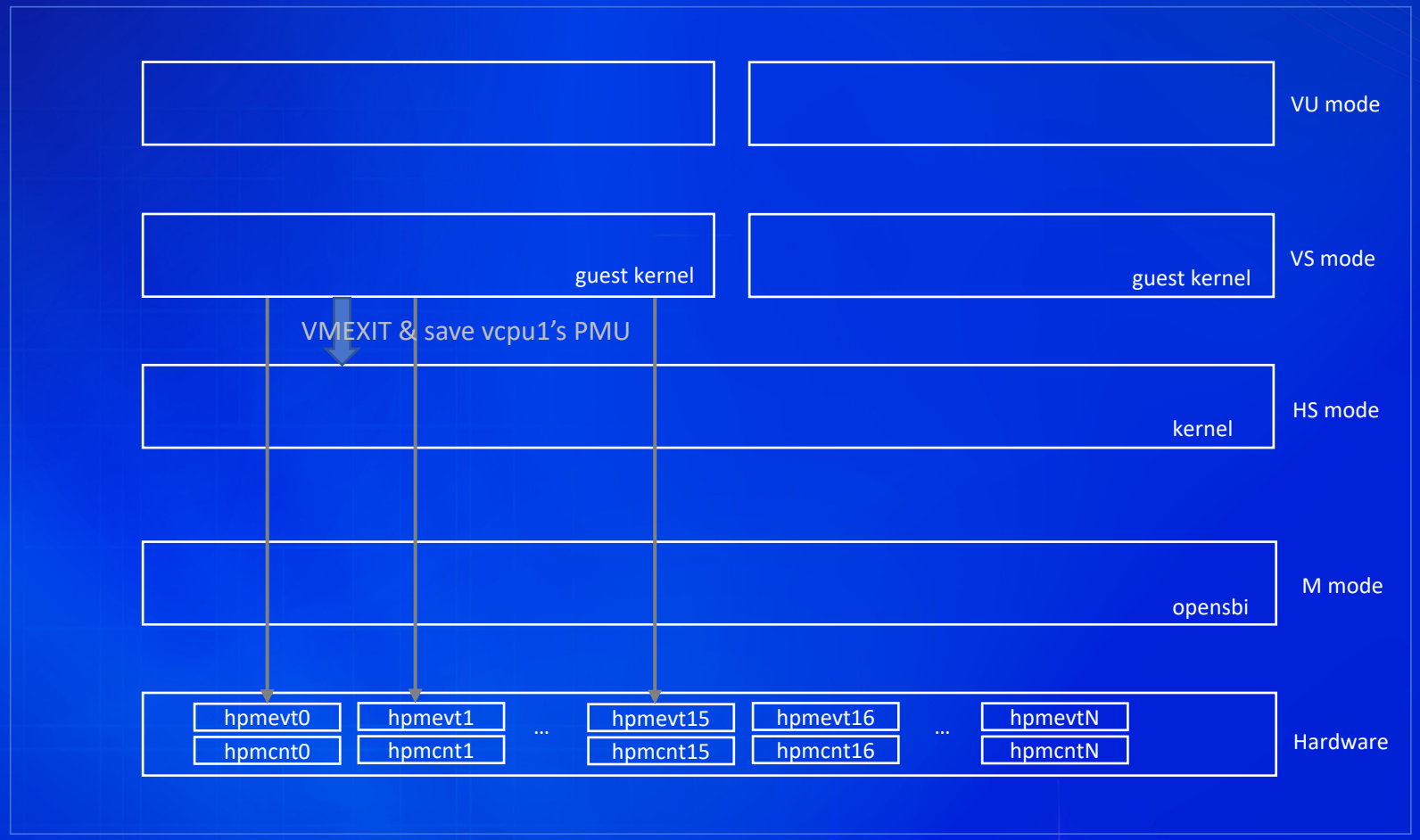
- For guest
  - The first several PMU registers/interrupt passed through (16, e.g.)
  - Guest sees and directly access all 16 PMU register pairs
- For host
  - Reserve the last PMU registers for host sampling (not delegated to VS)
  - PMU registers allocation method updated (choose the last indexes first)

Software design for PMU passthrough



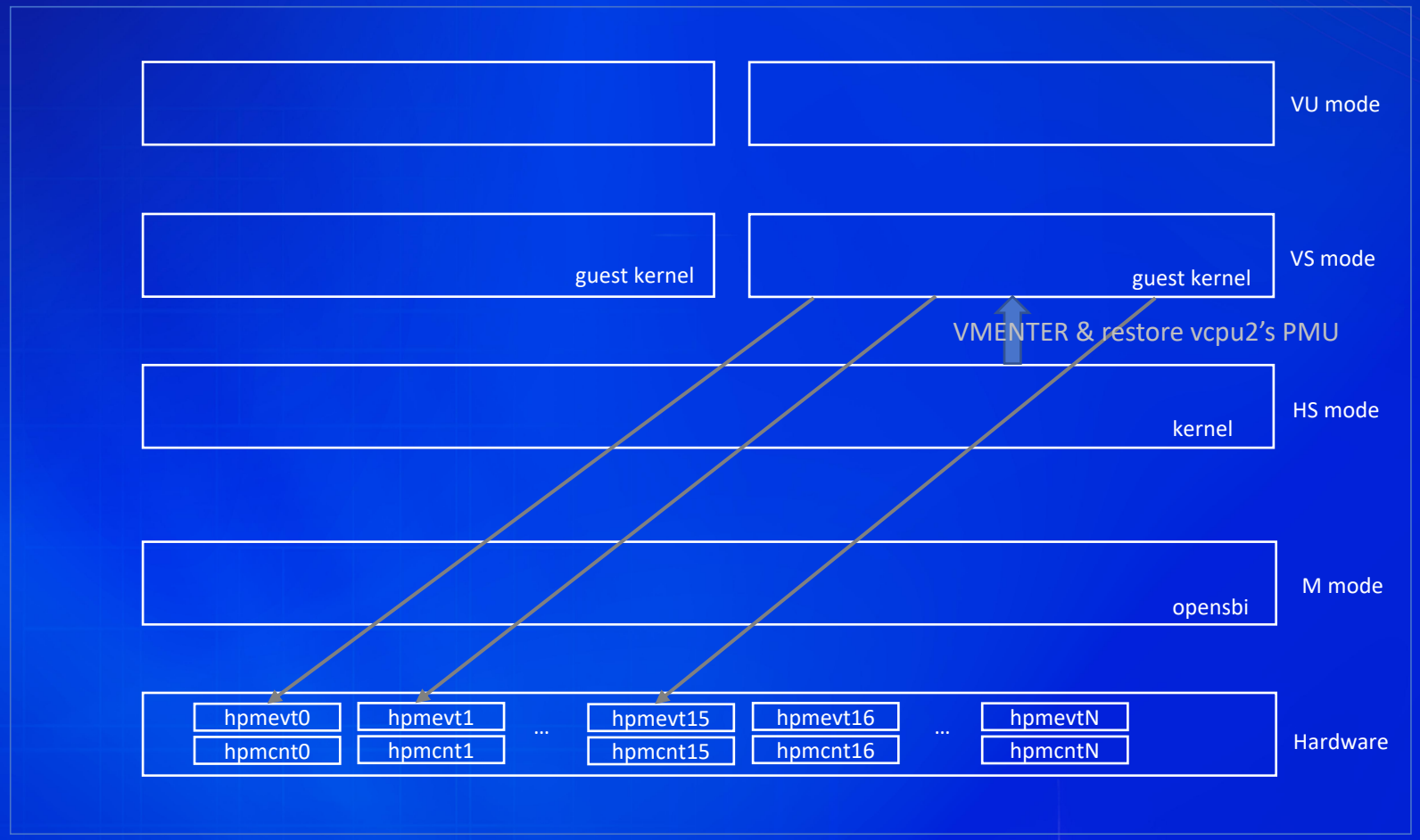
- For single VCPU
  - No need to save/restore PMU registers when VMEXIT/VMENTER as PMU passed-through do not sample non VS/VU mode

Software design for PMU passthrough

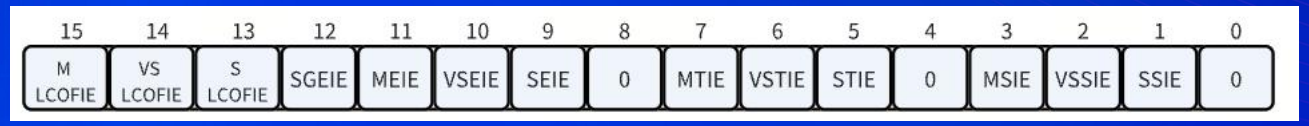


- For CPU oversubscription
  - Save VCPU's PMU registers only when
    - A different VCPU runs on the same CPU
    - VCPU is to be migrated
  - Restore VCPU's PMU registers when
    - VCPU runs on a new VCPU
    - VCPU is not the last one on the same CPU

Software design for PMU passthrough



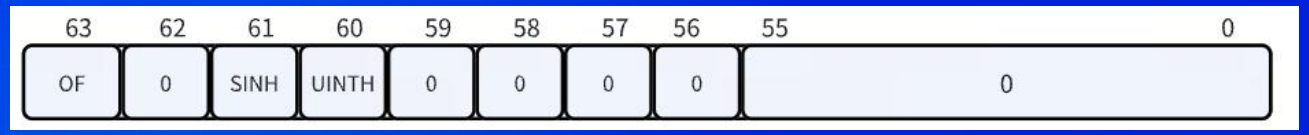
- For CPU oversubscription
  - Save VCPU's PMU registers only when
    - A different VCPU runs on the same CPU
    - VCPU is to be migrated
  - Restore VCPU's PMU registers when
    - VCPU runs on a new VCPU
    - VCPU is not the last one on the same CPU



MIE Register

M mode PMU

HS mode PMU



VS mode PMU

https://lists.riscv.org/g/sig-perf-analysis/topic/tech_performance_event_sampling/107881217

Thanks!