VENTANA MICRO

# RISC-V Nested Virtualization

Anup Patel <apatel@ventanamicro.com>

# Outline

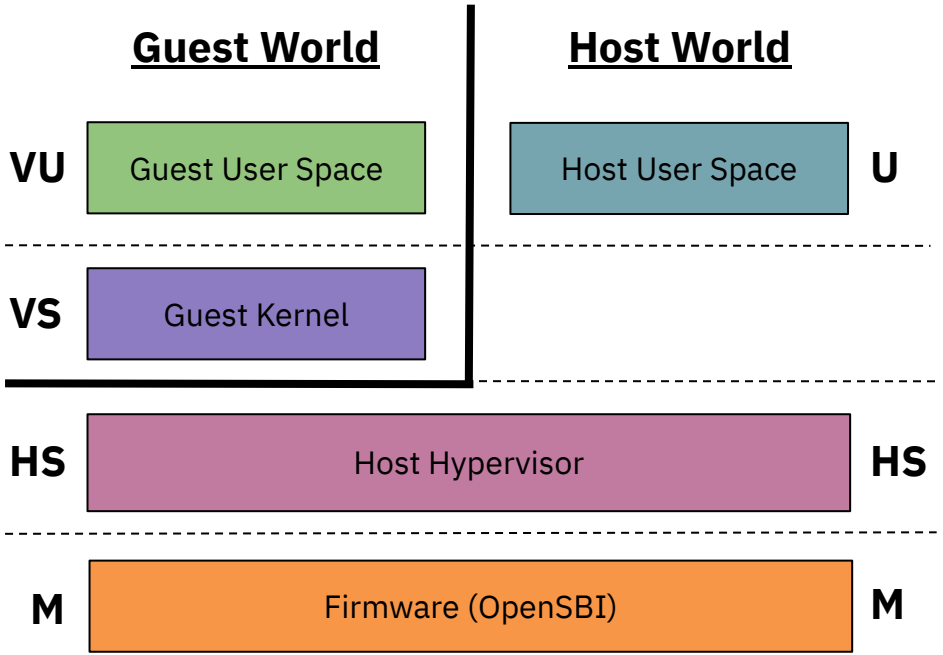
- RISC-V H-extension Overview
- RISC-V Nested Virtualization
- RISC-V Nested Virtualization Status And Demo

# RISC-V H-extension Overview

# RISC-V H-extension

VENTANA MICRO
RISC-V Performance Leader

- Suitable for both Type-1 and Type-2 Hypervisors
- HS-mode = S-mode + Hypervisor Support
  - HFENCE.[GVMA|VVMA] instructions
  - HLV/HSV instructions
  - “h<xyz>” CSRs for hypervisor capabilities
  - “vs<xyz>” CSRs contains VS-mode state
- Two additional modes for Guest
  - VS-mode = Virtualized S-mode
  - VU-mode = Virtualized U-mode
- In HS-mode (V=0)
  - “s<xyz>” CSRs are host S-mode CSRs
- In VS-mode (V=1)
  - “s<xyz>” CSRs are alias to “vs<xyz>” CSRs
  - sfence.vma is alias to hfence.vvma



# H-extension CSRs

VENTANA MICRO
RISC-V Performance Leader

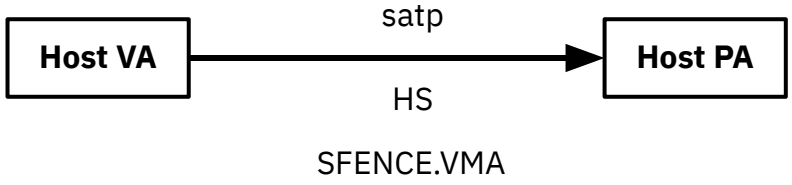
- **h<xyz> CSRs** (Additional CSRs for HS-mode)
  - Hypervisor Trap Setup
    - hstatus, hedeleg, hideleg, hie, hcounteren, hgeie
  - Hypervisor Configuration
    - henvcfg, henvcfgh (RV32)
  - Hypervisor Trap Handling
    - htval, hip, hvip, htinst, hgeip
  - Hypervisor Protection and Translation
    - hgatp
  - Hypervisor Counter/Timer Virtualization Registers
    - htimedelta, htimedeltah (RV32)

- **vs<xyz> CSRs** (Access to VS-mode state from HS-mode)
  - vsstatus, vsie, vstvec, vsscratch, vsepc, vscause, vstval, vsip, vsatp

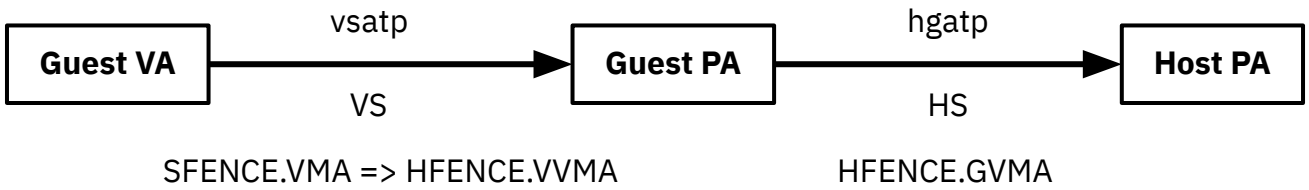
# H-extension Two-Stage MMU

VENTANA MICRO
RISC-V Performance Leader

**Host World**

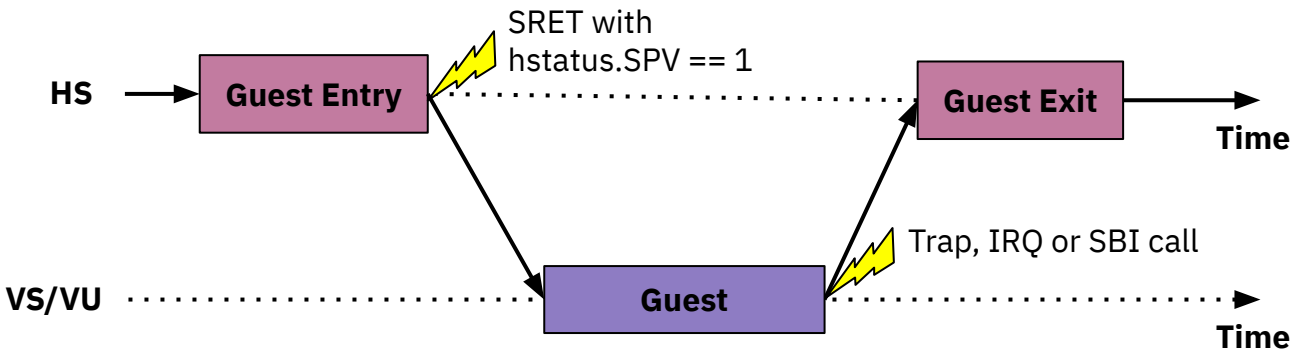



**Guest World**



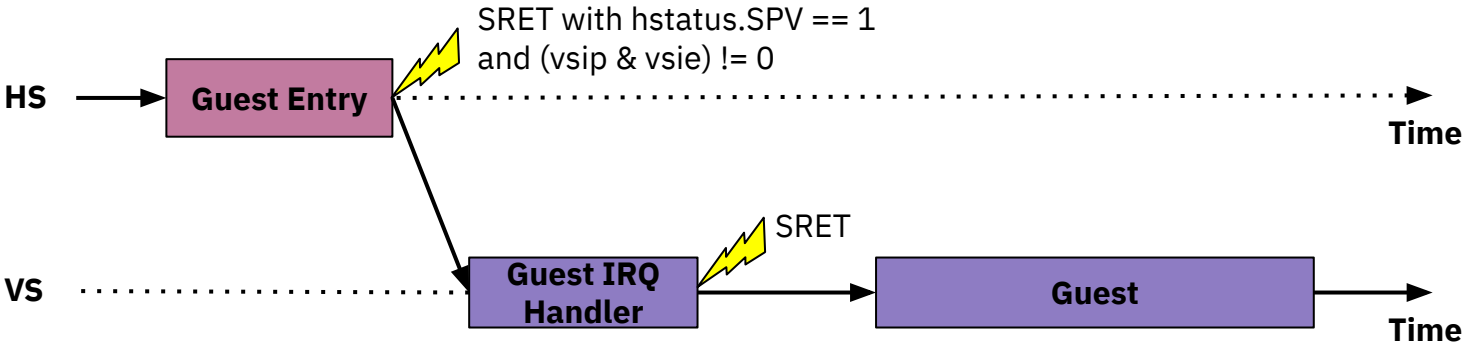
# H-extension Guest Entry/Exit

VENTANA MICRO
RISC-V Performance Leader



# H-extension Virtual Interrupts





# RISC-V Nested Virtualization

# What is Nested Virtualization ?

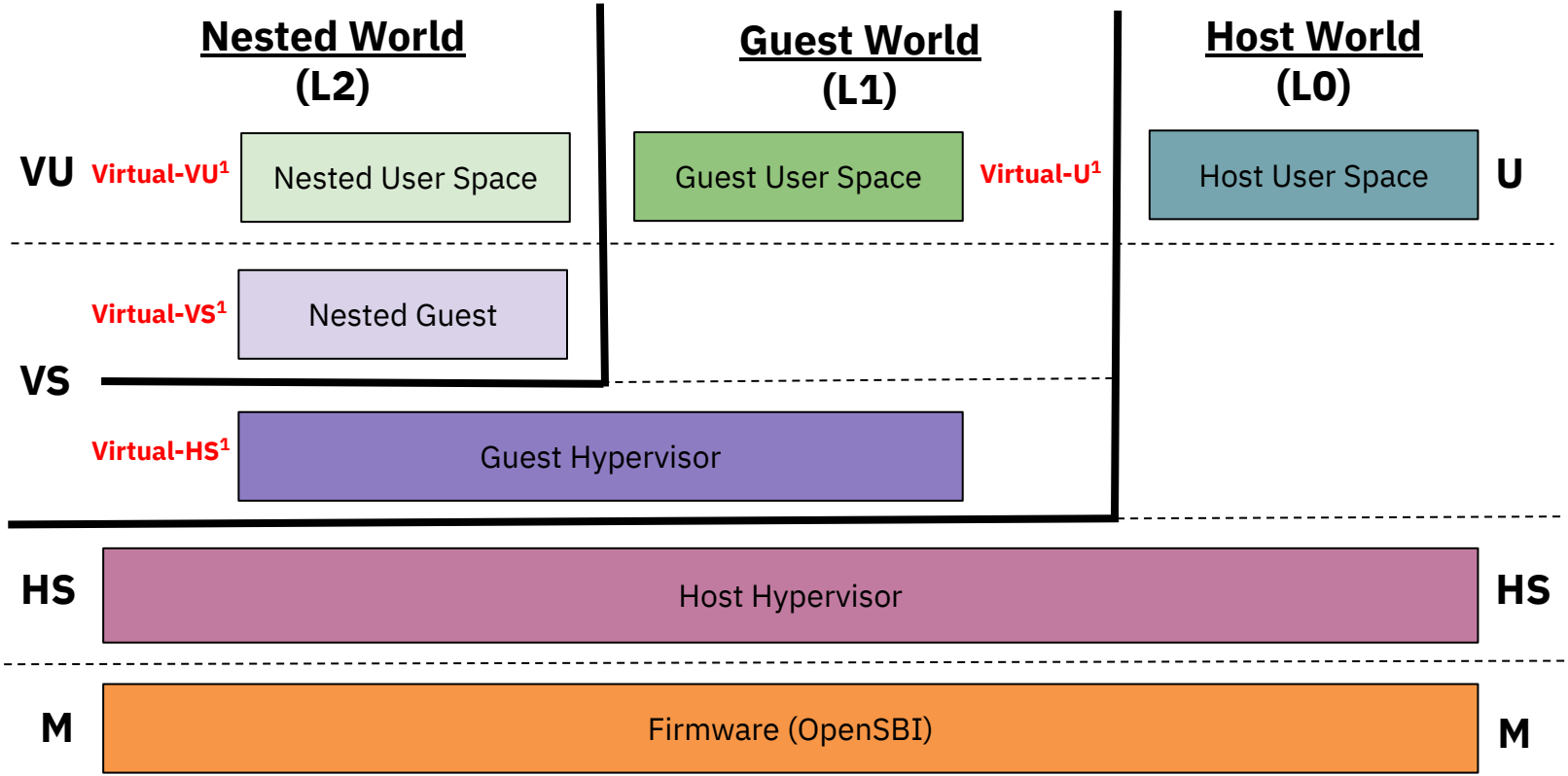
VENTANA MICRO
RISC-V Performance Leader

- Nested virtualization is the ability to run virtual machines (VMs) inside other VM
  - RISC-V nested virtualization means hypervisor emulates H-extension for Guest/VM

- **Use cases:**
  - Software Development
    - Mobile application development teams needing native mobile platform can use Android running as VM under a VM instance on public/private cloud
  - Software Quality Assurance
    - Testing/validating  Android where Android itself runs as VM under a VM instance on public/private cloud
    - Testing cloud infrastructure software managing a virtual cloud running on existing public/private cloud
  - Sales and Education
    - Customers/Students can understand, learn and experiment with cloud infrastructure software on existing public/private cloud
  - … and more to come in future …

# Nested Virtualization (Logical View)

VENTANA MICRO
RISC-V Performance Leader

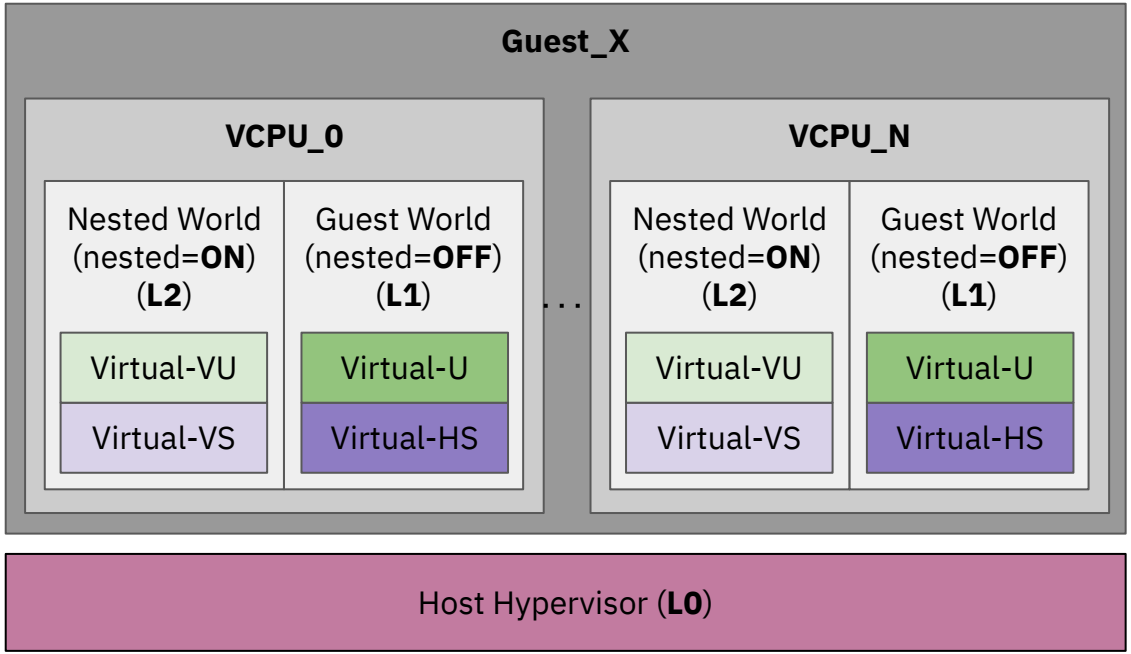



[1] Synthetic modes emulated by Host Hypervisor

# Nested Virtualization (Temporal View)

VENTANA MICRO
RISC-V Performance Leader

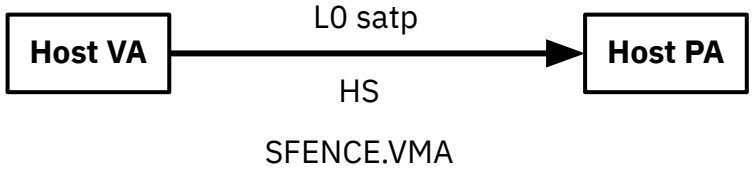
- At any point in time, a VCPU's "**nested**" virtualization state is:
  - **OFF:** Guest World (L1 = Virtual-VH or Virtual-U)
  - **ON:** Nested World (L2 = Virtual-VS or Virtual-VU)
- Host hypervisor (L0) will implement a special "**nested switch**" for changing "nested" virtualization state of a VCPU
- Host hypervisor (L0) emulates H-extension only for Virtual-HS/U
  - h<xyz> and vs<abc> CSRs
  - hfence, hvinval, hlv, and hsv instructions



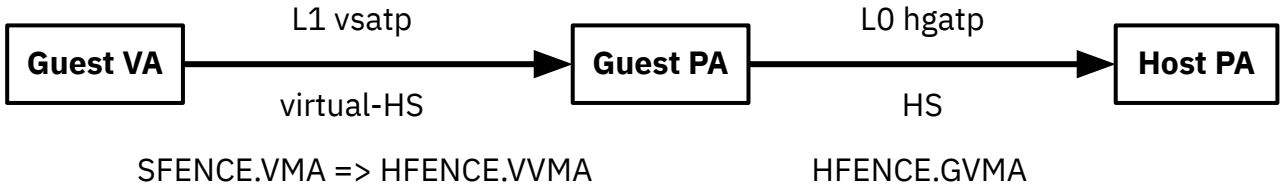
# Nested Three-Stage MMU

VENTANA MICRO
RISC-V Performance Leader

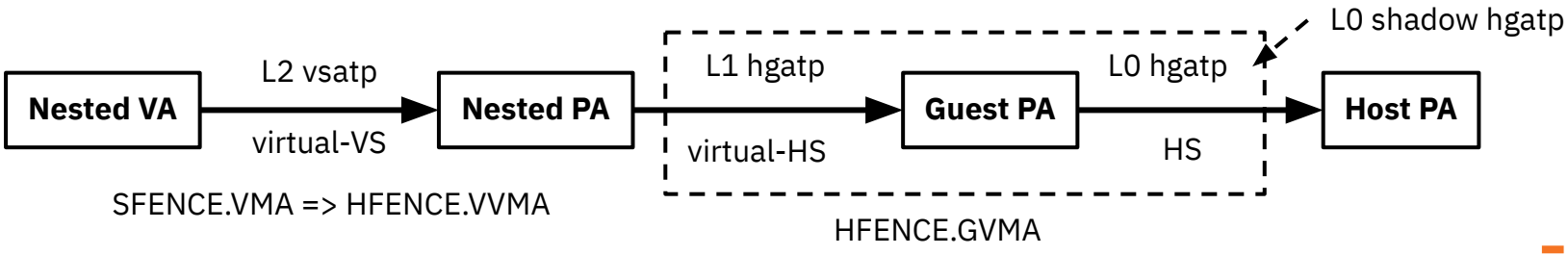
**Host World (L0)**




**Guest World (L1)**




**Nested World (L2)**

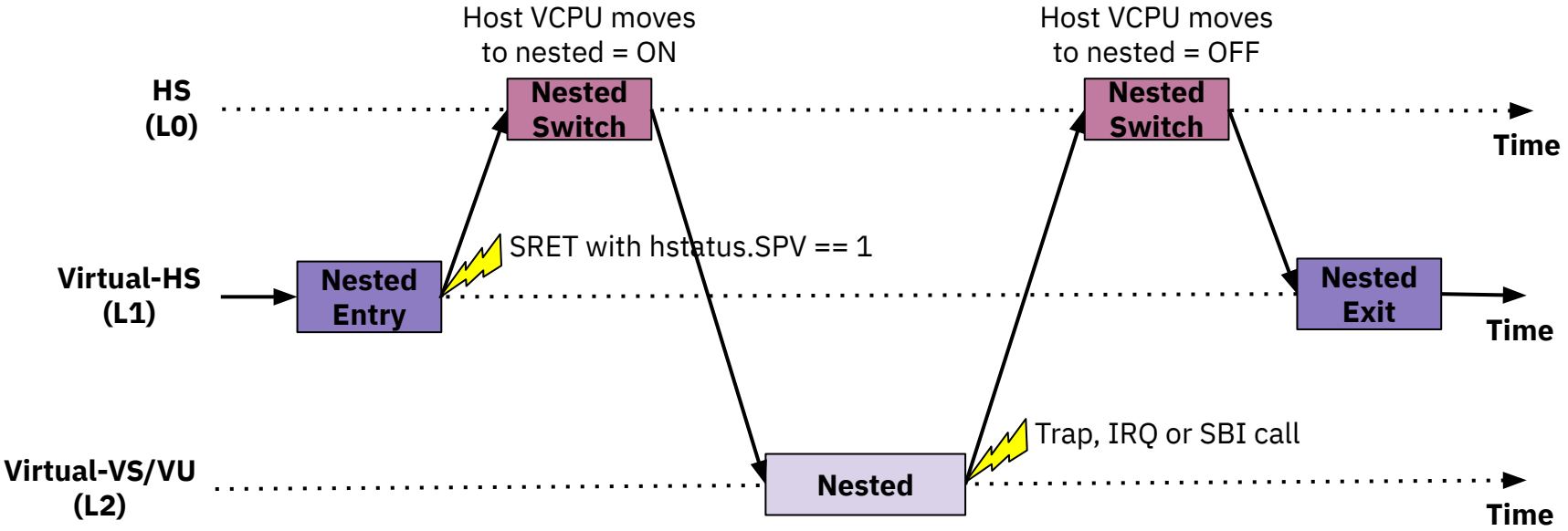


# Nested Entry/Exit

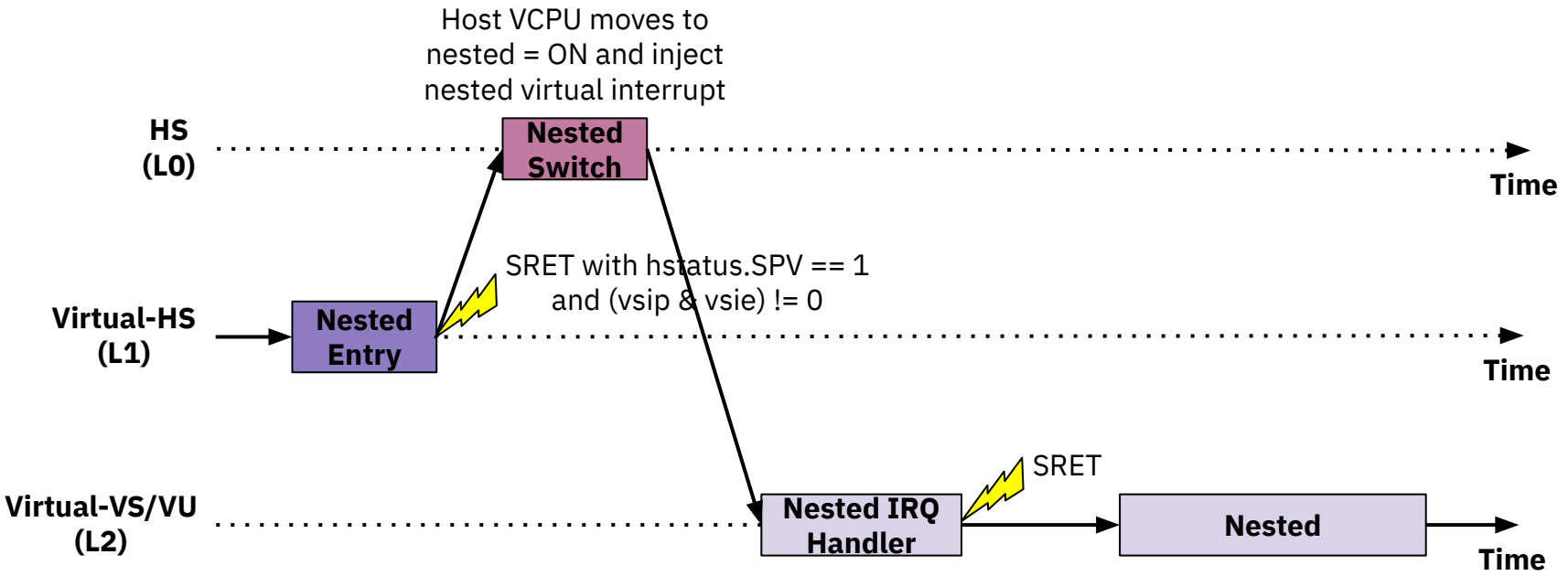
VENTANA MICRO
RISC-V Performance Leader

**Host hypervisor (L0) does lazy trapping of SRET instructions executed by Guest hypervisor (L1)**

- Enable SRET trapping when Guest hypervisor sets hstatus.SPV
- Disable SRET trapping when Guest hypervisor clears hstatus.SPV



# Nested Virtual Interrupts

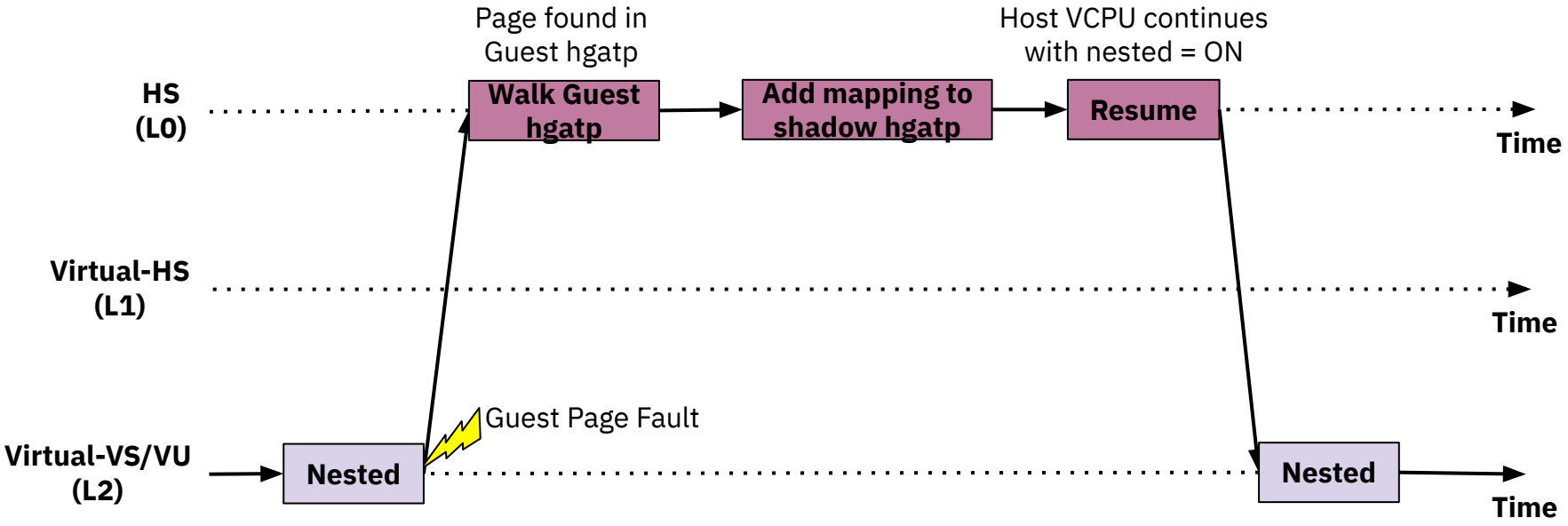


# Nested Guest Page Fault Handling

VENTANA MICRO
RISC-V Performance Leader

**To speed-up host hypervisor (L0) page table walks on Guest hgatp:**

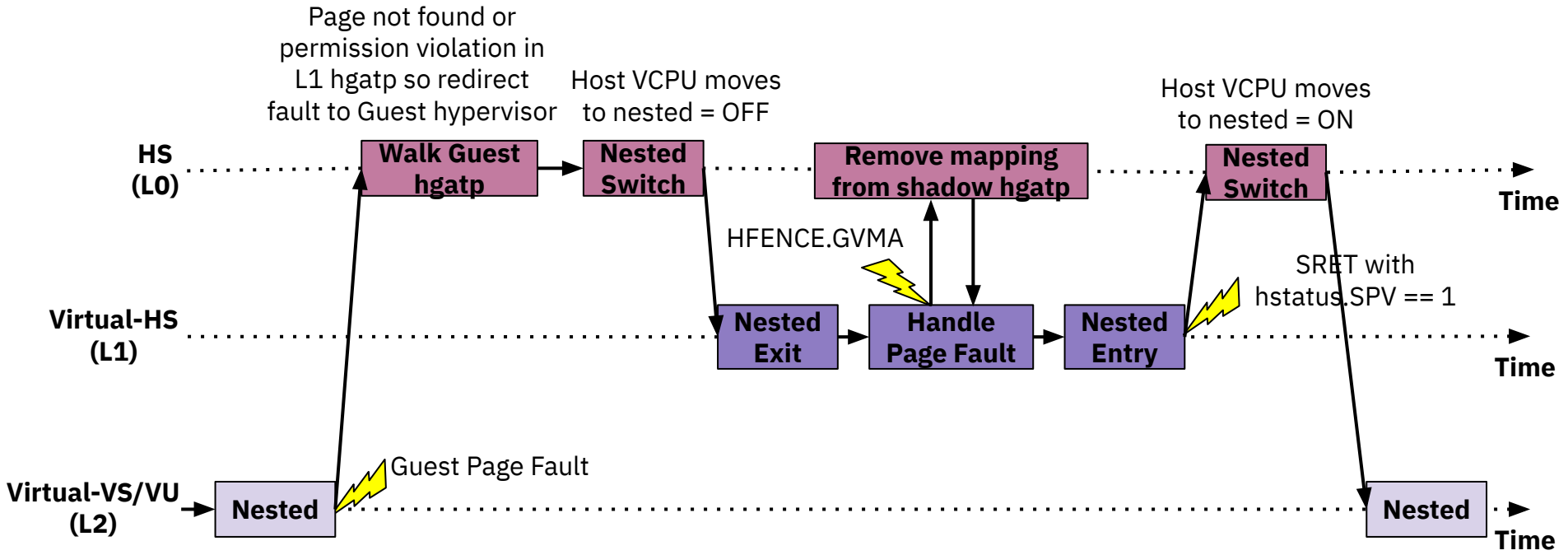
- Host hypervisor (L0) can restrict Guest hgatp.MODE to Sv39x4
- Guest hypervisor (L1) can use hugepages



# Nested Guest Page Fault Redirection

VENTANA MICRO
RISC-V Performance Leader

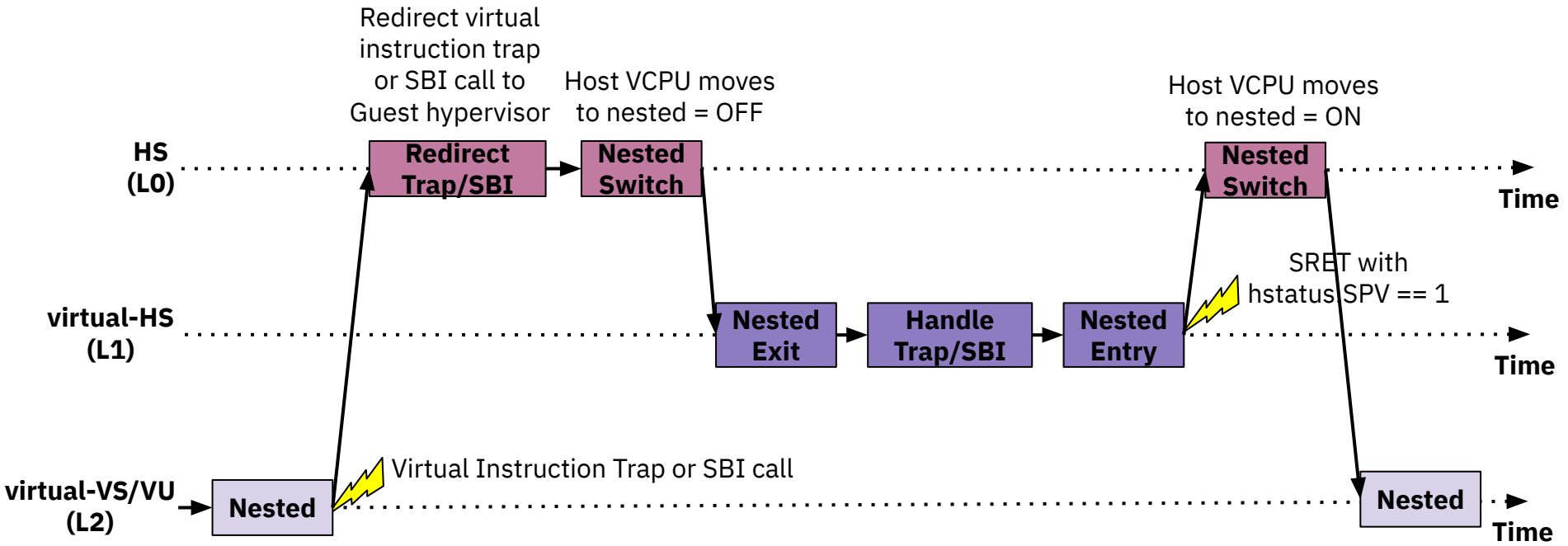
**Guest hypervisor (L1) can use hugepages to speed-up page table walks for Host hypervisor (L0)**



# Nested Recursion

VENTANA MICRO
RISC-V Performance Leader

**Nested Recursion = Guest hypervisor (L1) emulating H-extension for Nested world (L2)**



# RISC-V Nested Virtualization Status And Demo

# Nested Virtualization Status

VENTANA MICRO
RISC-V Performance Leader

- Complete proof-of-concept done on QEMU:
  - **Host World (L0):** Xvisor RISC-V
  - **Guest World (L1):** KVM RISC-V
  - **Nested World (L2):** Linux RISC-V
- QEMU and Xvisor patches already upstreamed
- Work in progress
  - KVM RISC-V nested virtualization support
    - Running KVM/Xvisor inside KVM
  - SBI nested acceleration (NACL) specification proposal
    - Host hypervisor and Guest hypervisor use shared memory to minimize traps
  - Xvisor RISC-V SBI NACL support
  - KVM RISC-V SBI NACL support
- **Live Demo !!!**

# Thank You !!!