Nested Virtualization on ARM

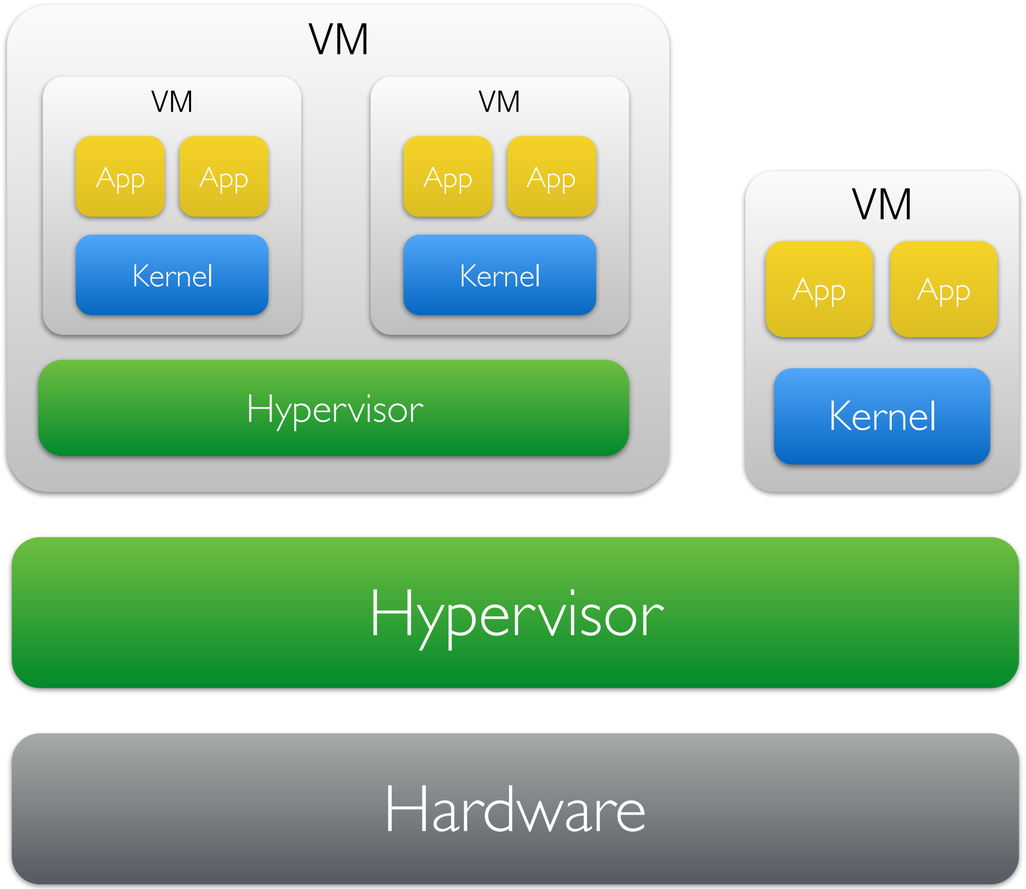
NEVE: Nested Virtualization Extensions

Jin Tack Lim Christoffer Dall Shih-Wei Li
Jason Nieh Marc Zyngier

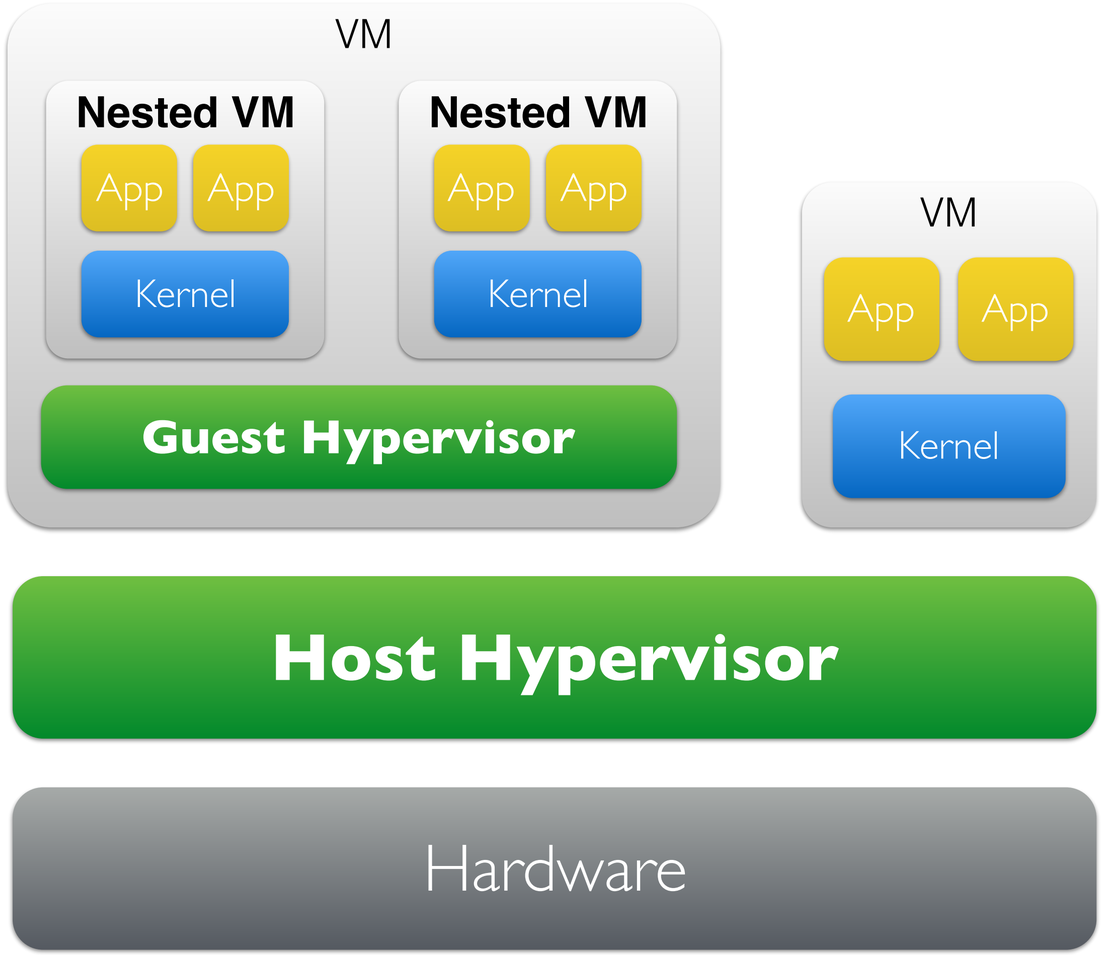
jitack@cs.columbia.edu christoffer.dall@linaro.org shih-wei@cs.columbia.edu,
nieh@cs.columbia.edu marc.zyngier@arm.com

LEADING COLLABORATION IN THE ARM ECOSYSTEM

Nested Virtualization



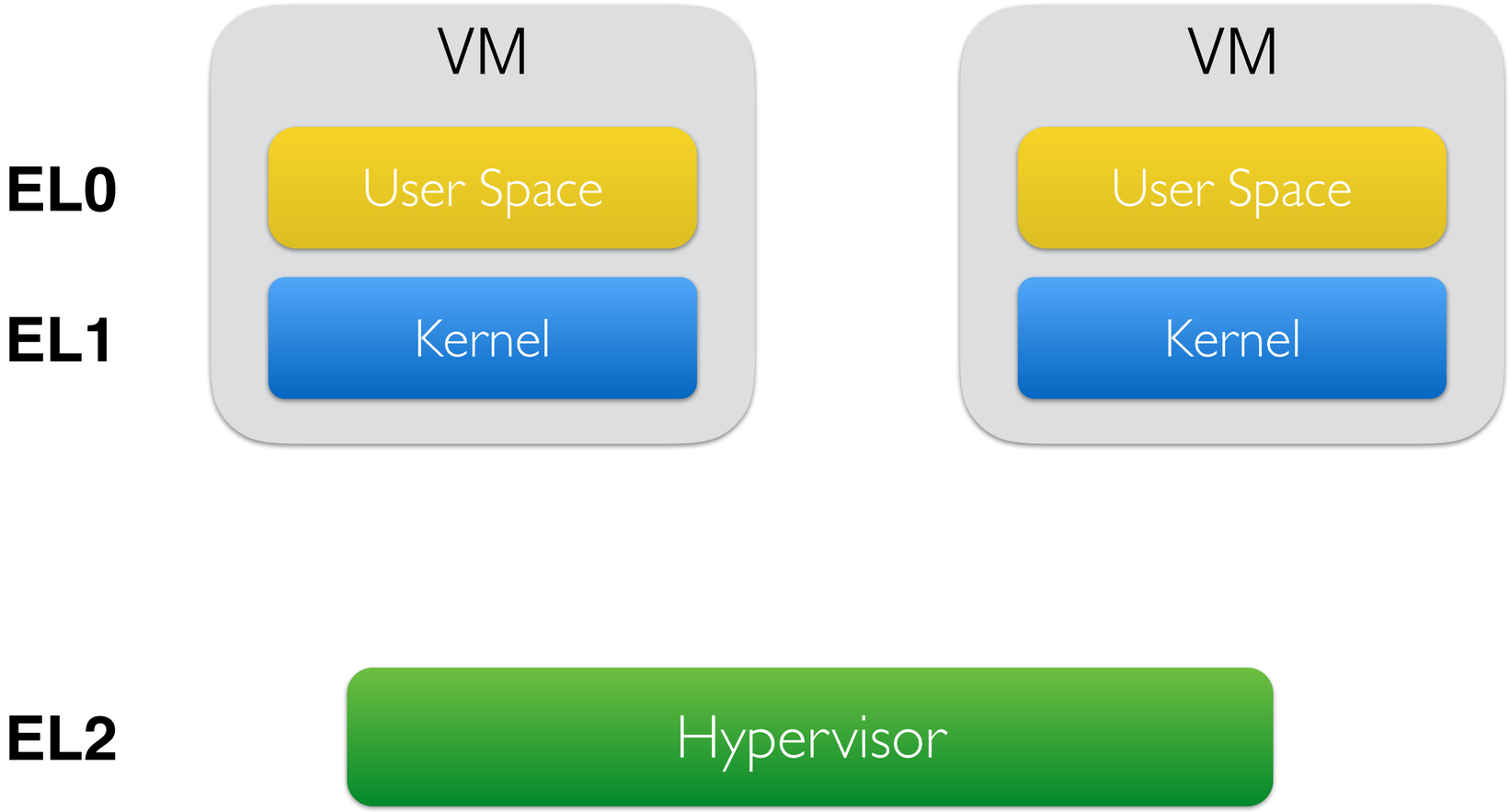
Terminology



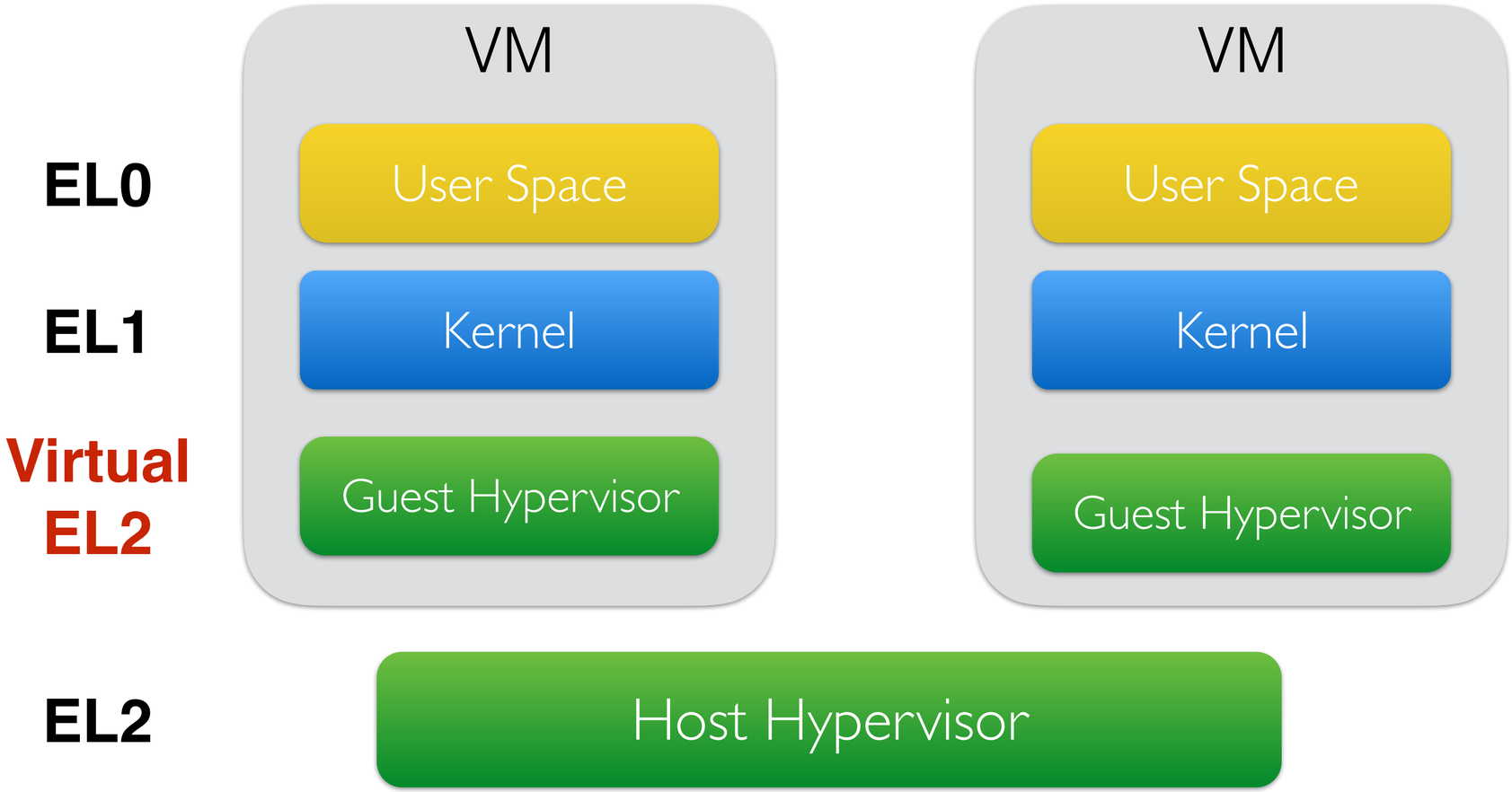
Use Cases

1. Run guest operating systems with built-in virtualization.
2. IaaS hosting private clouds
3. Test your hypervisor in a VM
4. Debug your hypervisor in a VM
5. Develop hypervisors using a cloud

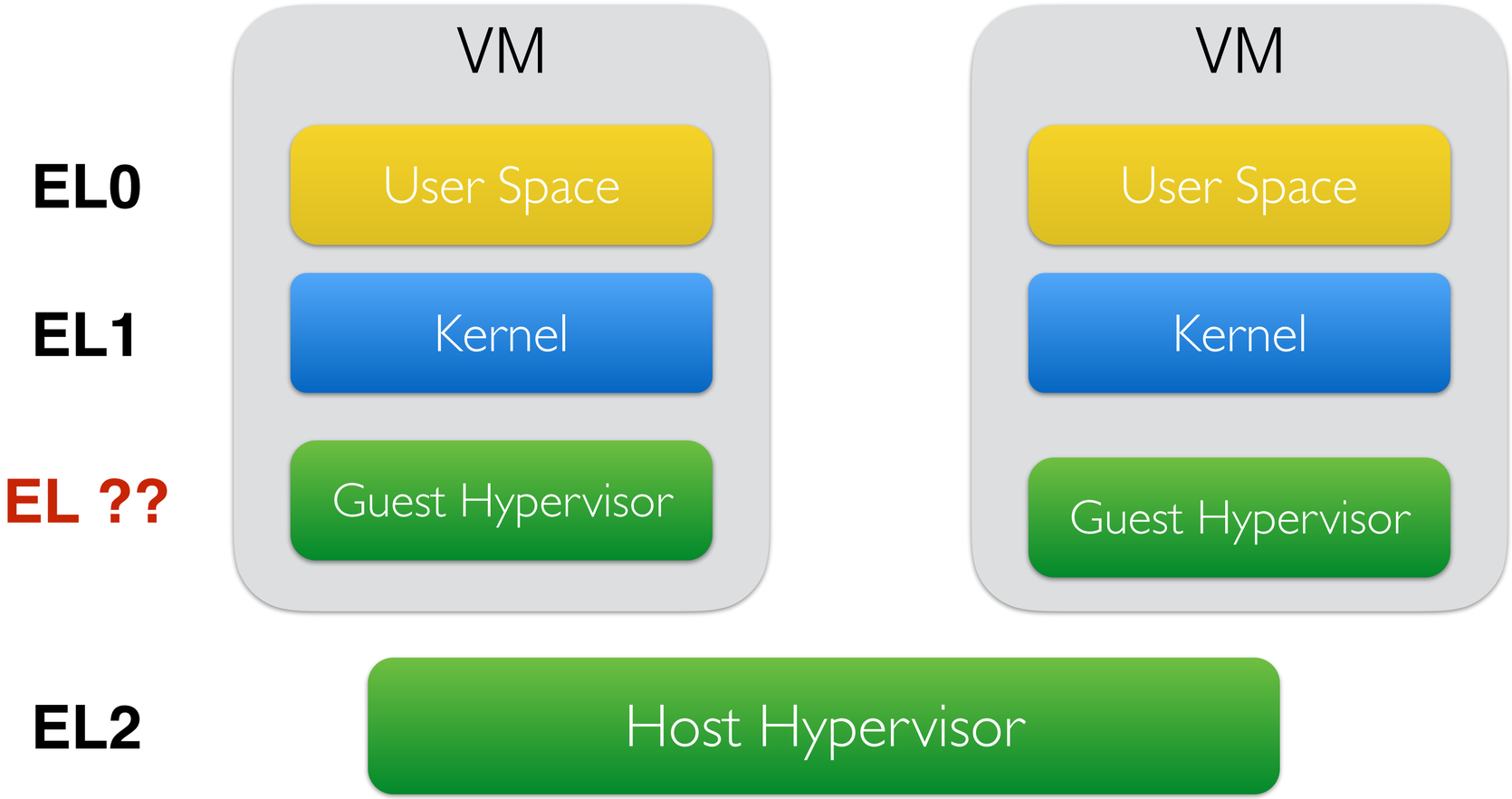
ARM Virtualization Extensions



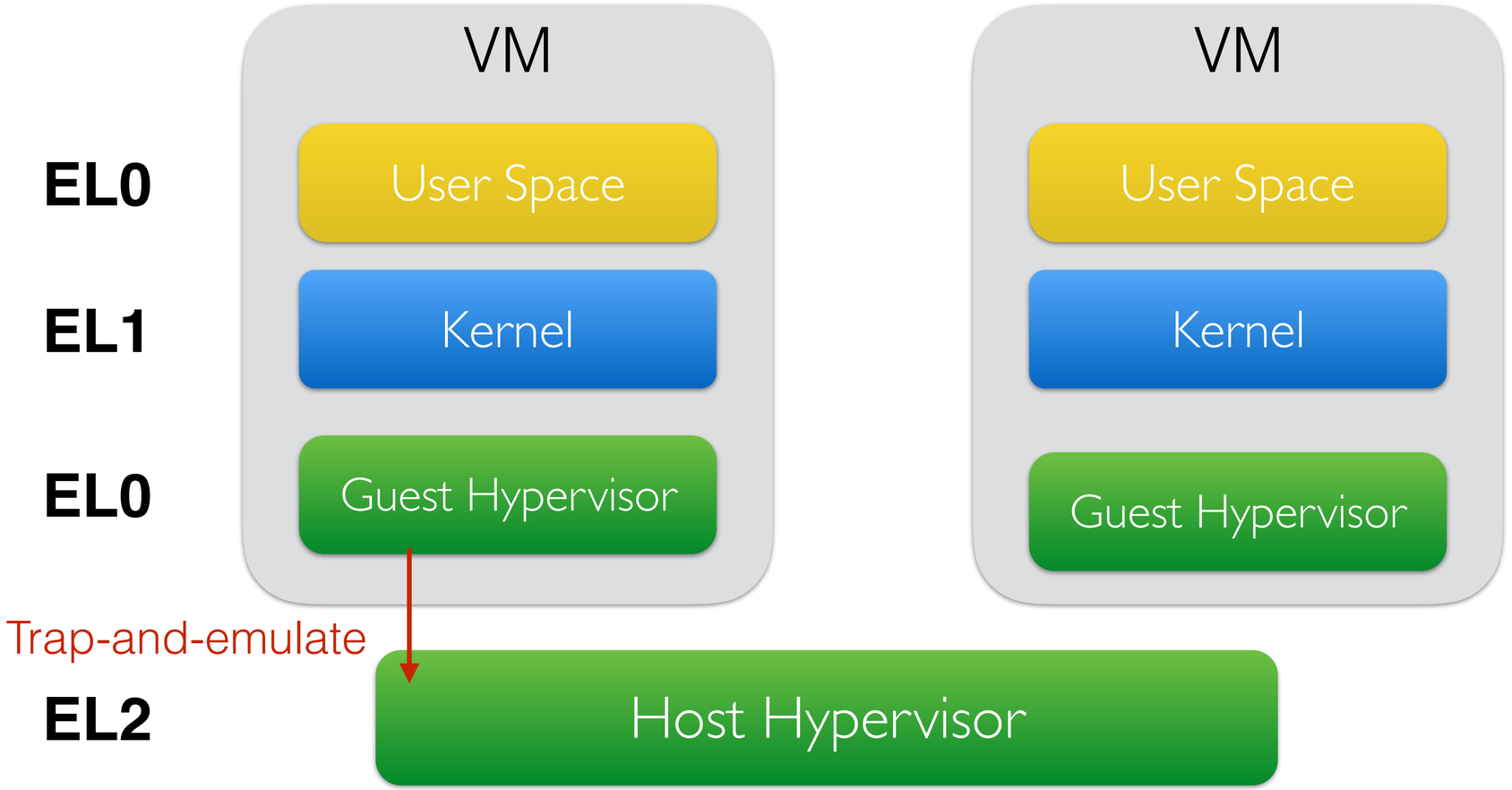
ARM Nested Virtualization



ARM Nested Virtualization

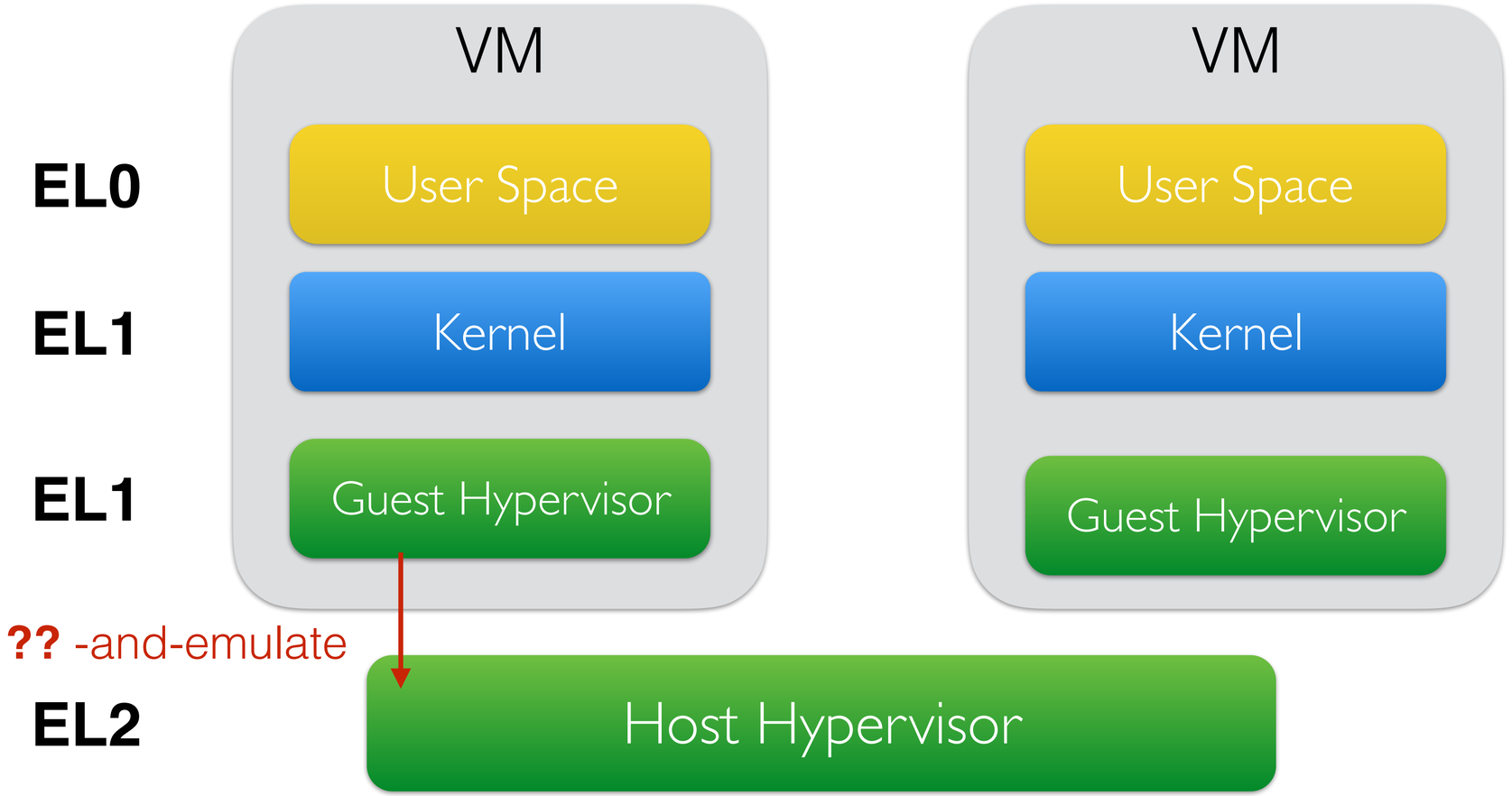


ARMv8.0 Nested Virtualization



Linaro

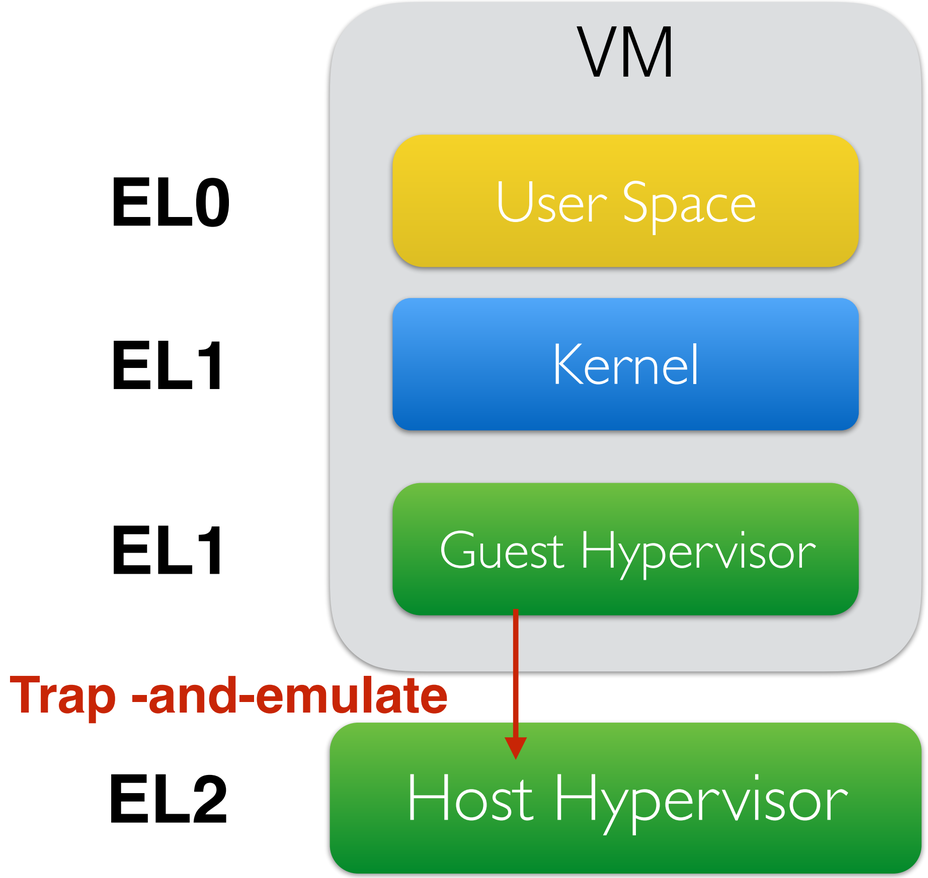
ARMv8.0 Nested Virtualization



Linaro

ARMv8.3 Nested Virtualization

- Gives you software emulation of vEL2 in EL1
- HCR_EL2.NV:
  - Traps EL2 operations executed in EL1 to EL2
  - Traps **`eret`** to EL2
  - CurrentEL reports EL2 even in EL1



Linaro

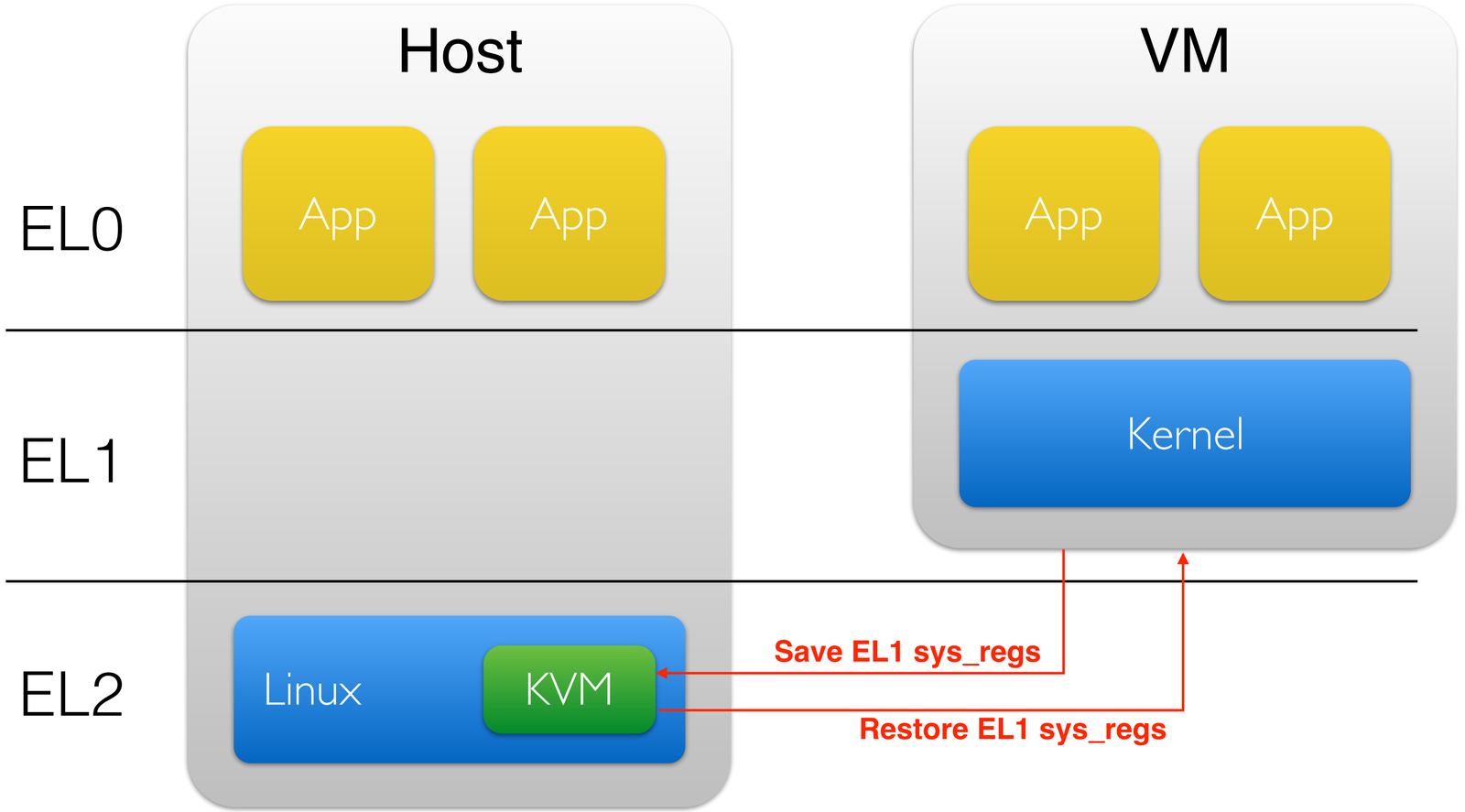
KVM/ARM Nested Virtualization Implementation

- **EL2 Emulation**
- Stage 2 MMU Virtualization
- Hyp Timer Virtualization
- Nested Virtual Interrupts

Nested CPU Virtualization

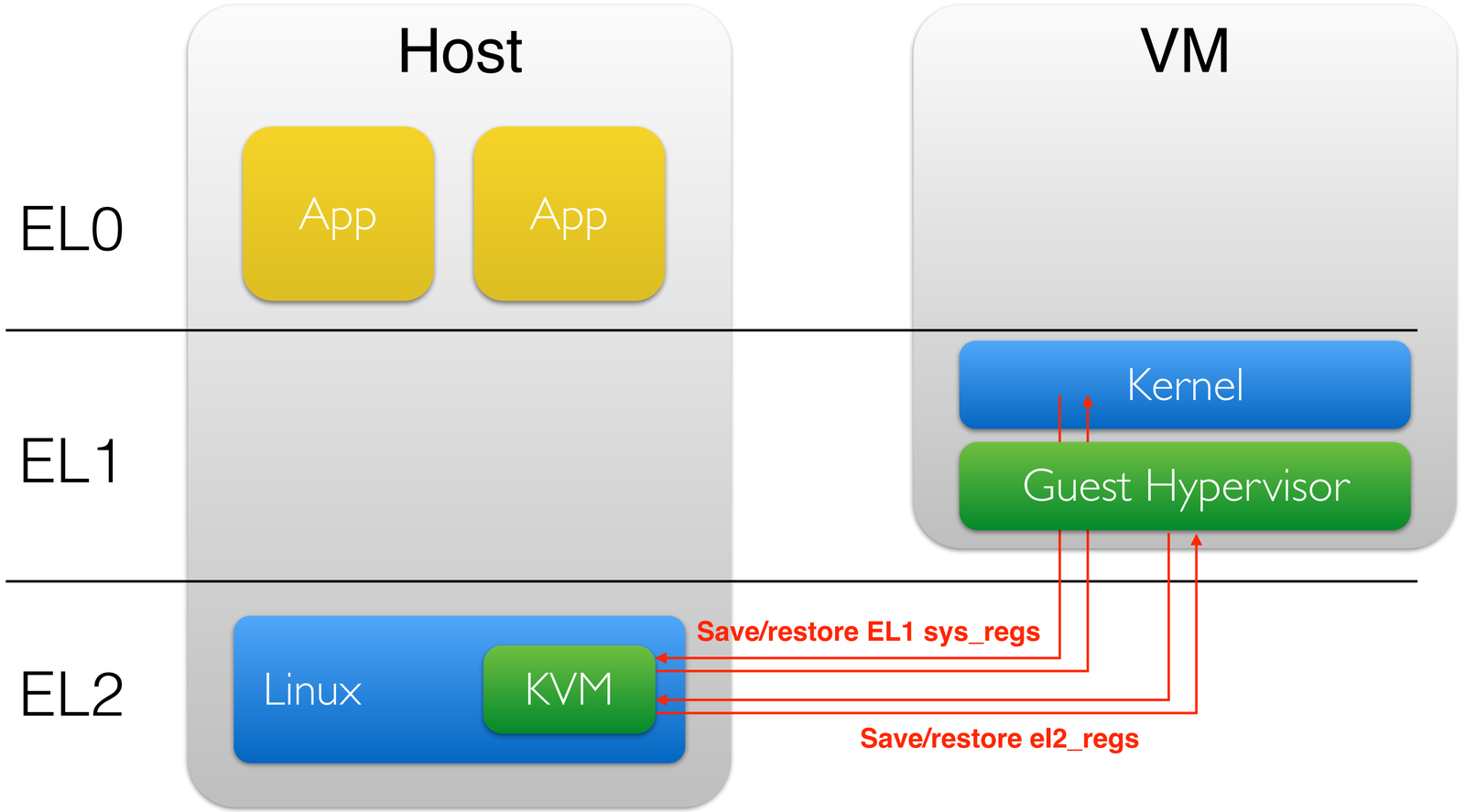
```
 struct kvm_cpu_context {
     u64 sys_regs[NR_SYS_REGS];
+    u64 el2_regs[NR_EL2_REGS];
 }

 struct kvm_vcpu_arch {
     …
     struct kvm_cpu_context ctxt;
 }
```

Hypervisor-VM Switch



Linaro

Hypervisor-Hypervisor Switch



Emulating EL2 in EL1

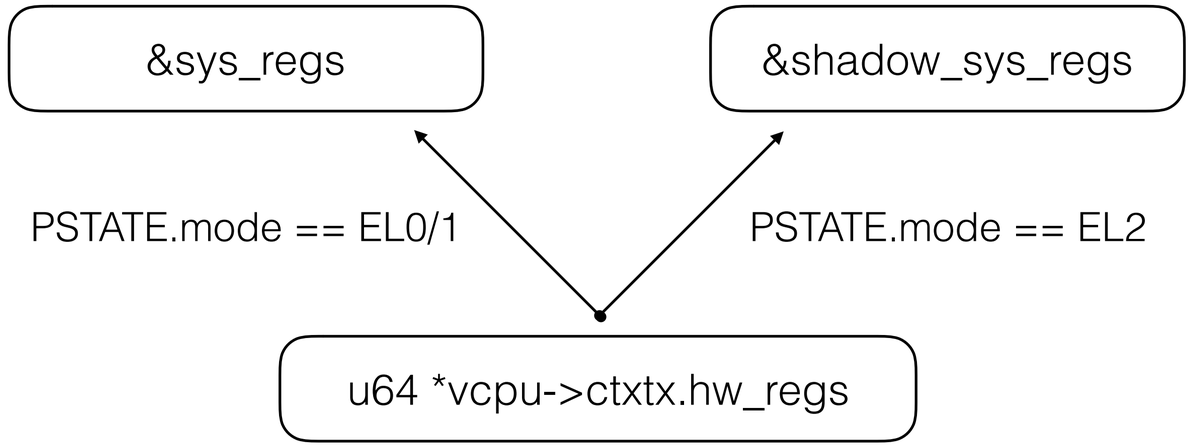
- Define mapping of EL2 registers to EL1 registers
- Example: TTBR0_EL2 to TTBR0_EL1
- Example: SCTLR_EL2 adapted to SCTLR_EL1
- Shadow EL1 registers

Nested CPU Virtualization

```
 struct kvm_cpu_context {
     u64 sys_regs[NR_SYS_REGS];
+    u64 el2_regs[NR_EL2_REGS];
+    u64 shaow_sys_regs[NR_SYS_REGS];
 }

 struct kvm_vcpu_arch {
     …
     struct kvm_cpu_context ctxt;
 }
```

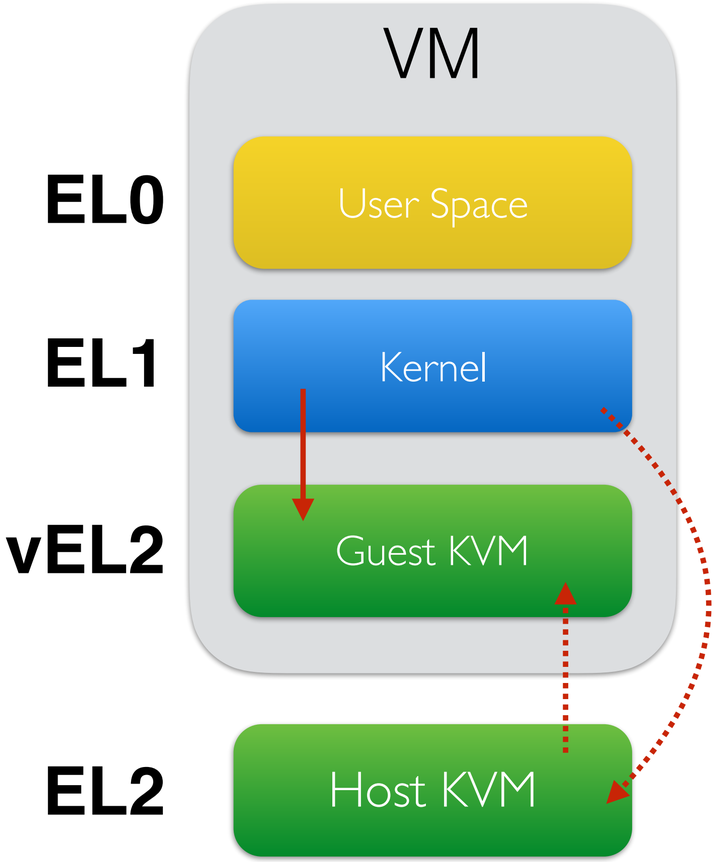
Linaro

Shadow Registers



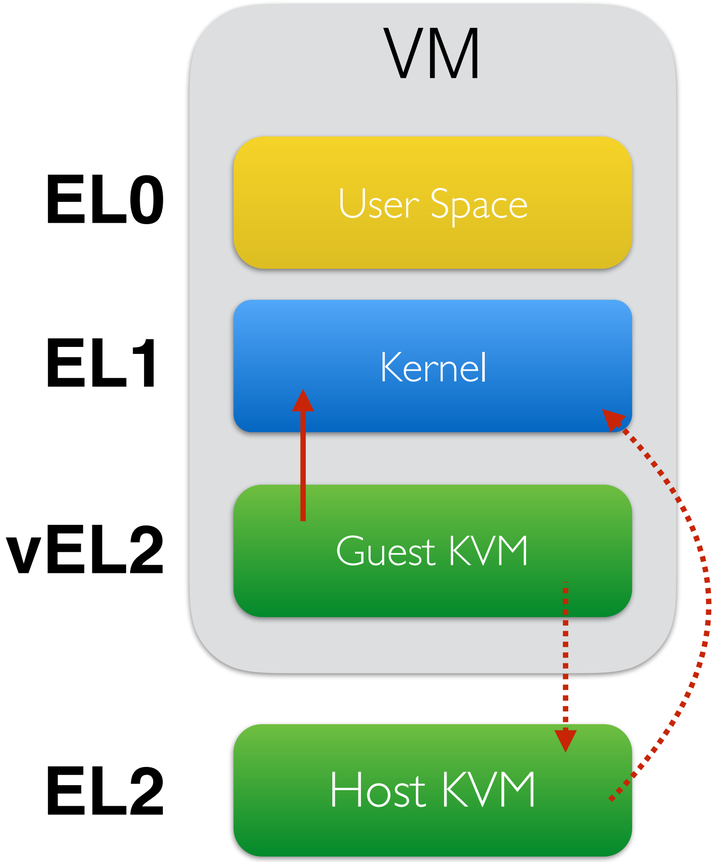
Virtual Exceptions

- Trap to virtual EL2
- “Forward” exceptions
- Emulate virtual exceptions



Virtual Exceptions

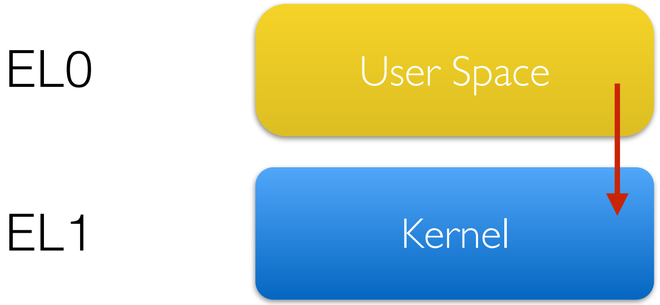
- Returning from virtual EL2
- Trap **`eret`** to EL2 (ARMv8.3)
- Emulate virtual exception return



KVM/ARM Nested Virtualization Implementation

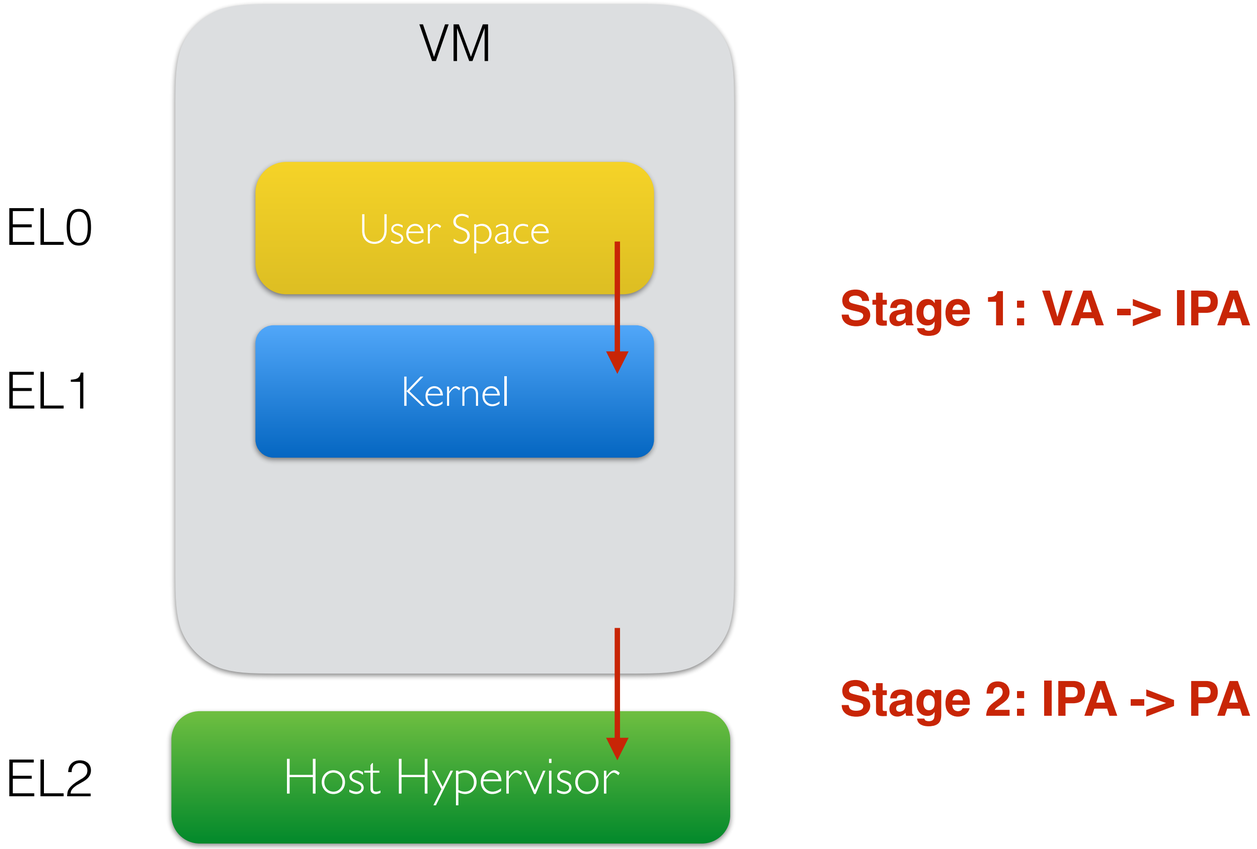
- EL2 Emulation
- **Stage 2 MMU Virtualization**
- Hyp Timer Virtualization
- Nested Virtual Interrupts

Memory Virtualization

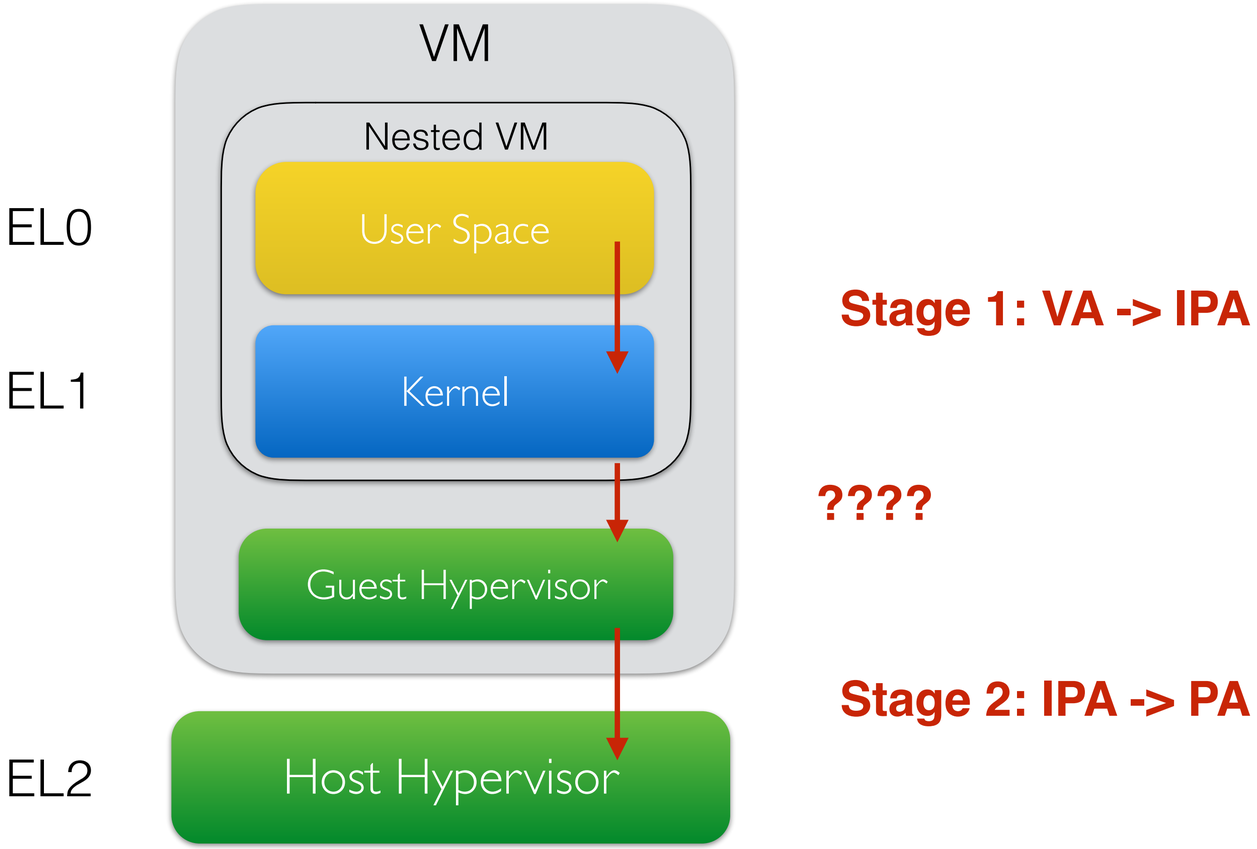


Stage 1: VA -> IPA

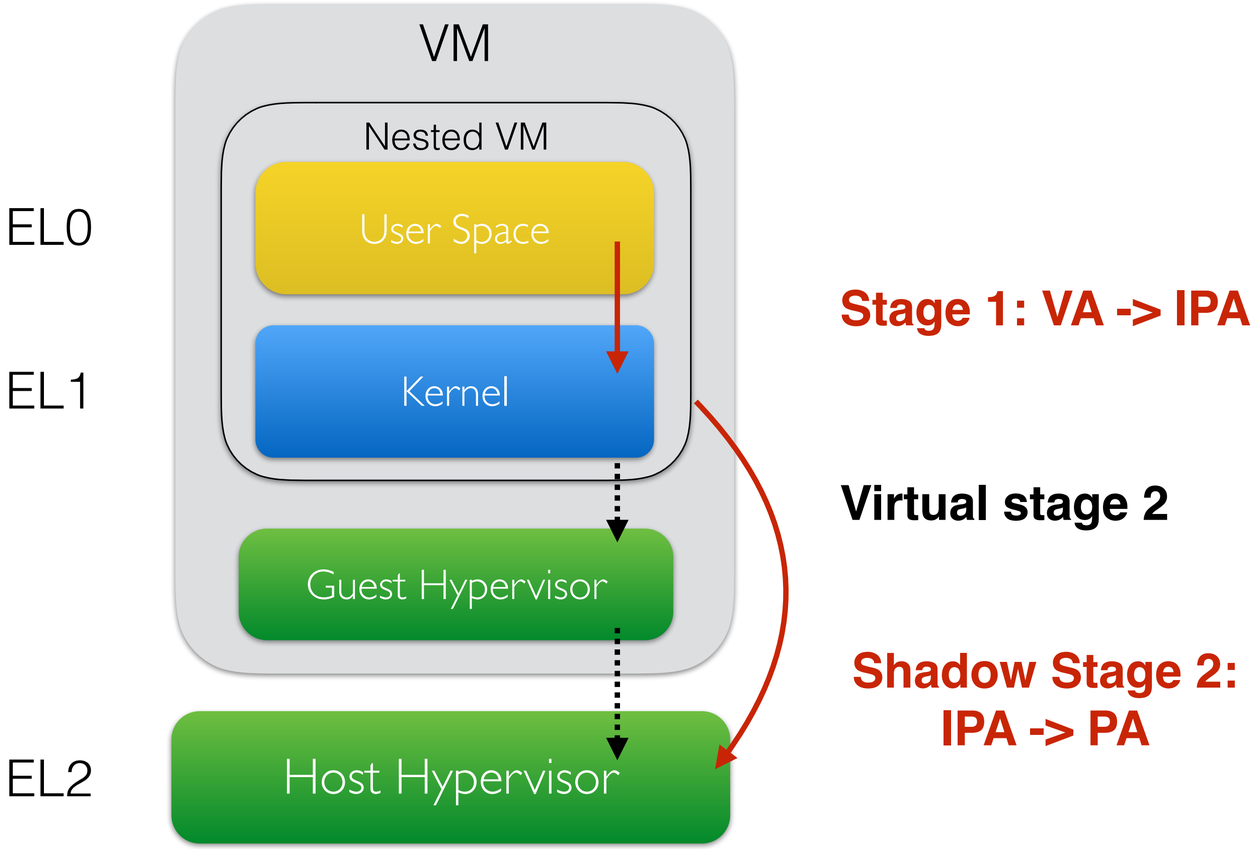
Memory Virtualization



Memory Virtualization



Memory Virtualization



KVM/ARM Nested Virtualization Implementation

- EL2 Emulation
- Stage 2 MMU Virtualization
- **Hyp Timer Virtualization**
- Nested Virtual Interrupts

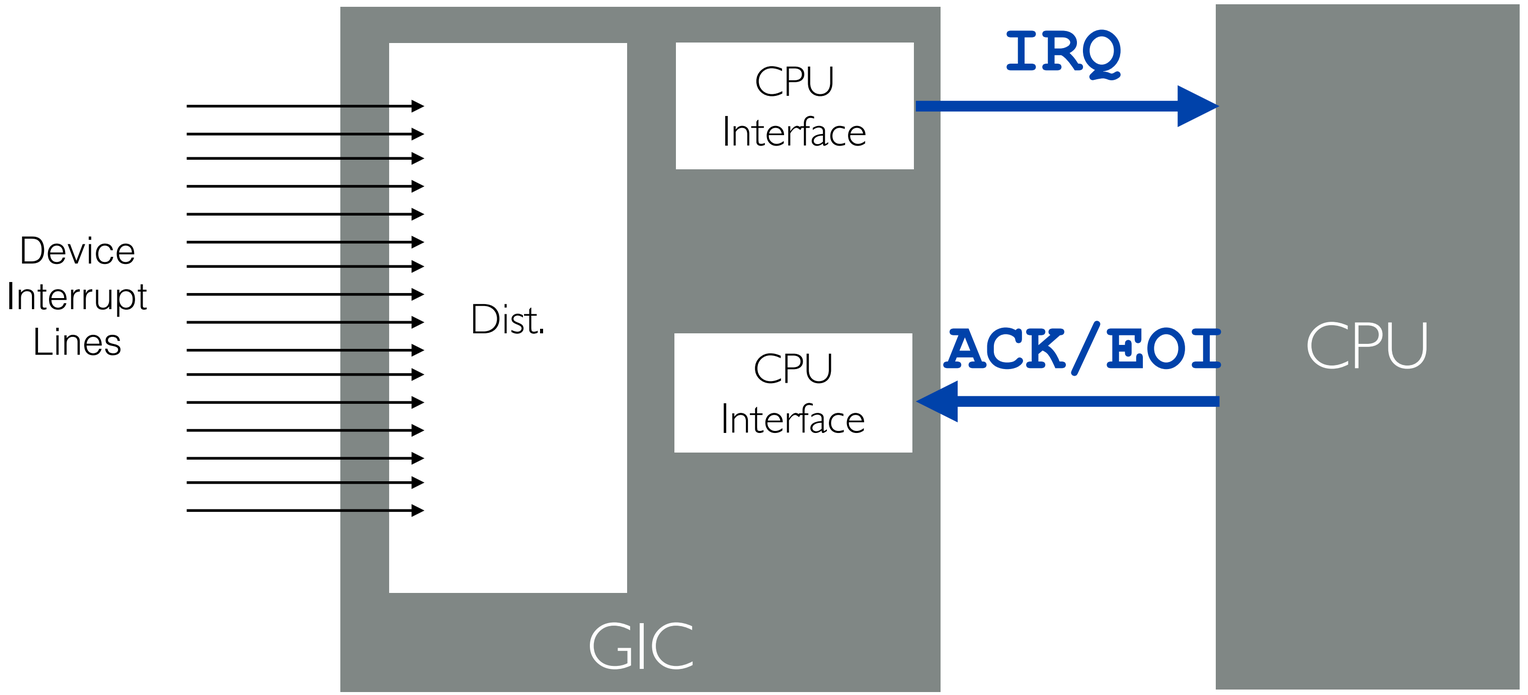
Nested Timer Virtualization

- ARM provides a virtual and physical timer in EL1
- EL2 provides a separate EL2 “hyp” timer
- Nested KVM/ARM supports a virtual CPU with EL2 and the hyp timer

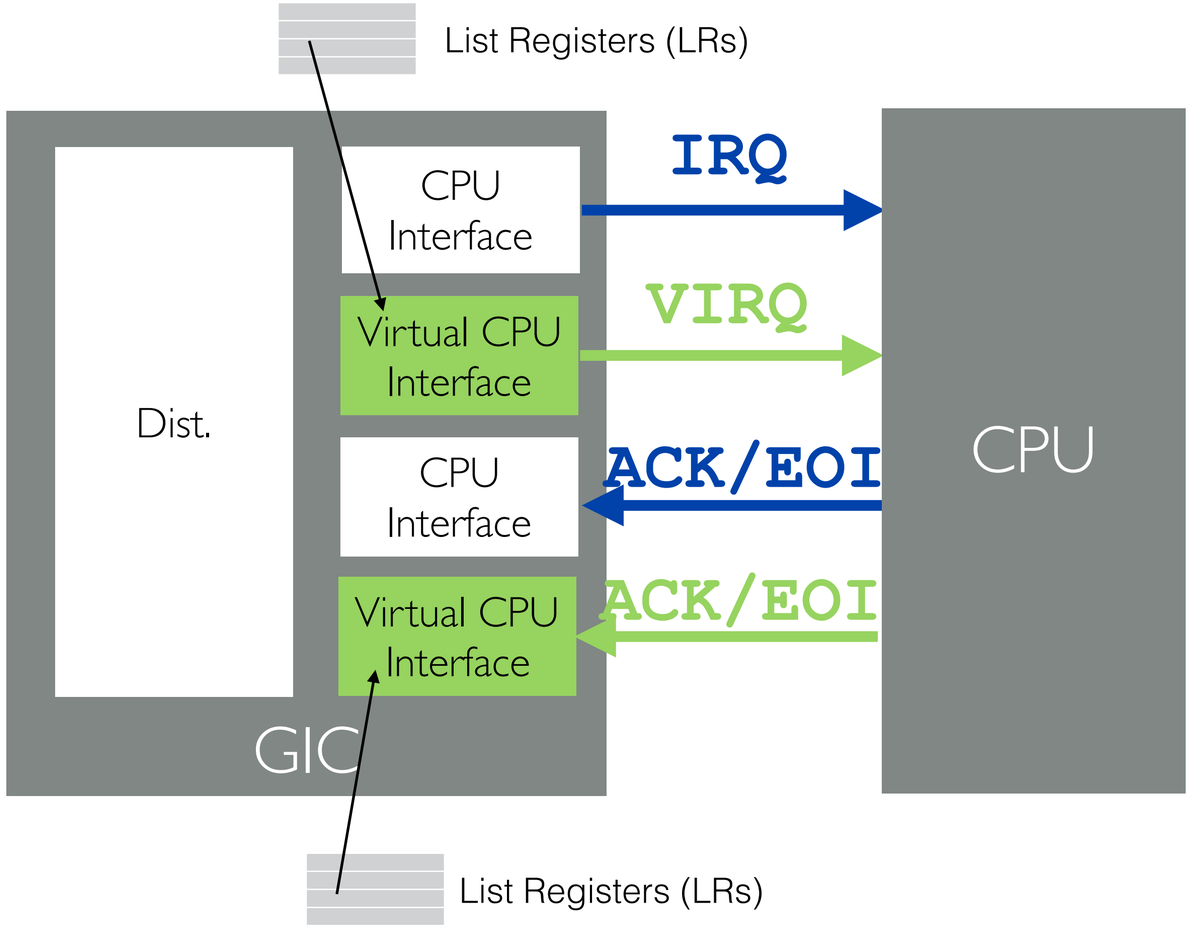
KVM/ARM Nested Virtualization Implementation

- EL2 Emulation
- Stage 2 MMU Virtualization
- Hyp Timer Virtualization
- **Nested Virtual Interrupts**

ARM Generic Interrupt Controller (GIC)

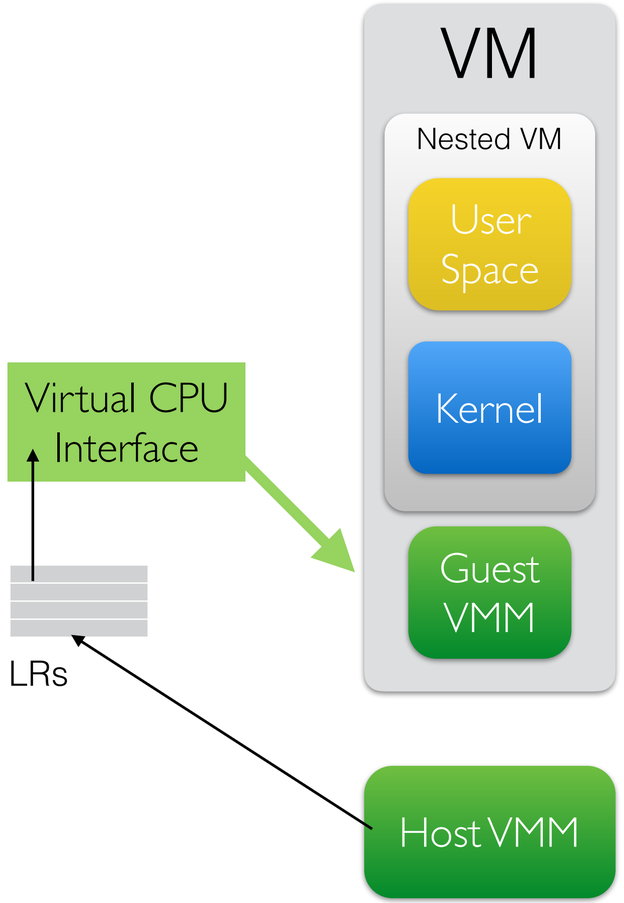


ARM Generic Interrupt Controller (GIC)



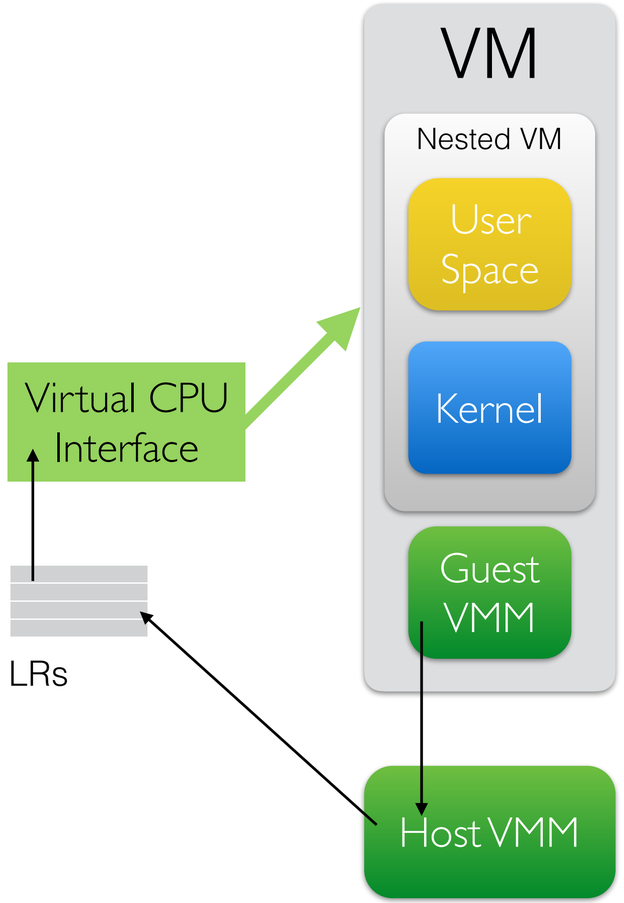
Nested Interrupt Virtualization

- Deliver virtual interrupts from the host to the VM



Nested Interrupt Virtualization

- Deliver virtual interrupts from the guest hypervisor to the nested VM
- Shadow list registers
- The nested VM can ACK and EOI virtual interrupts without trapping

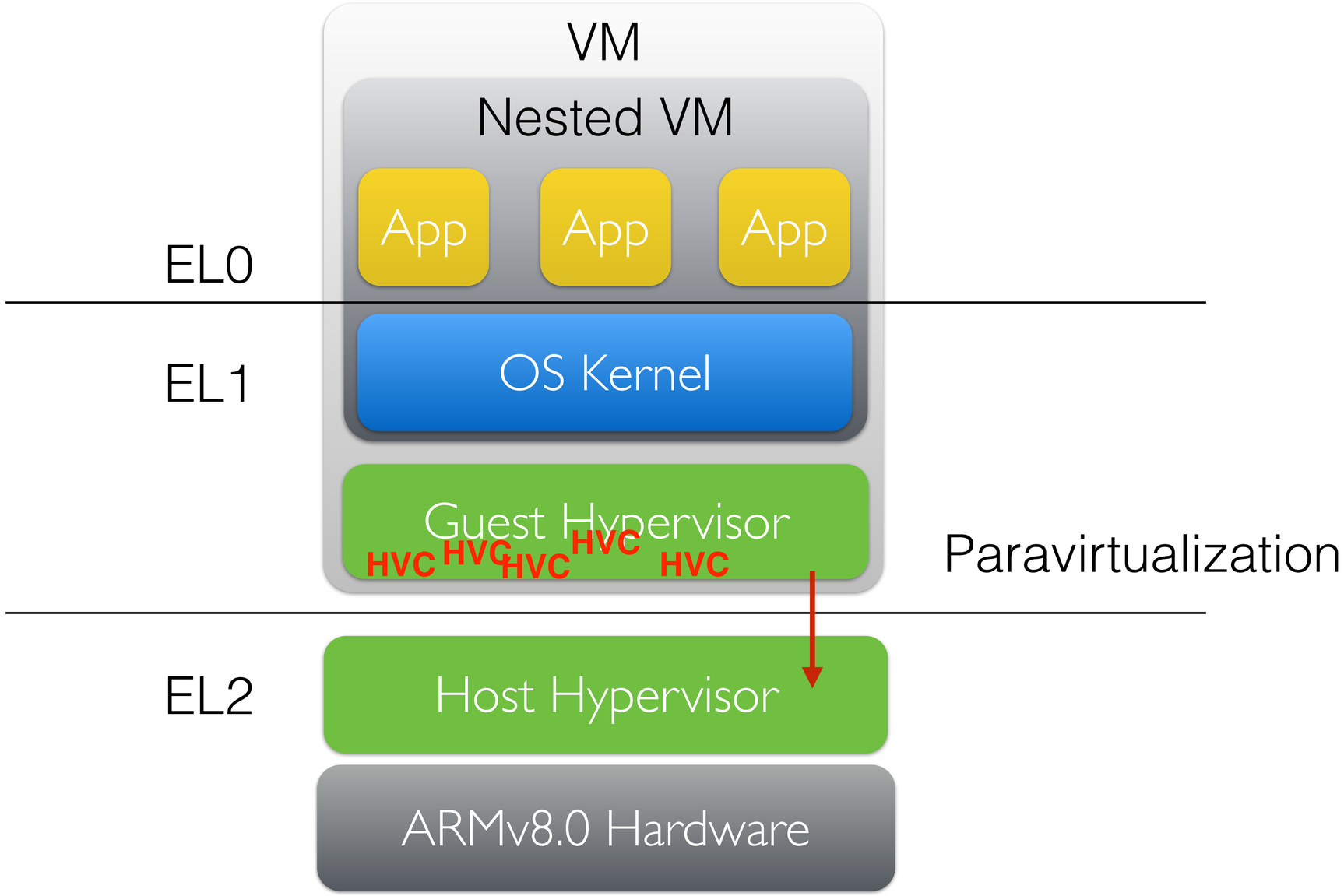


Linaro

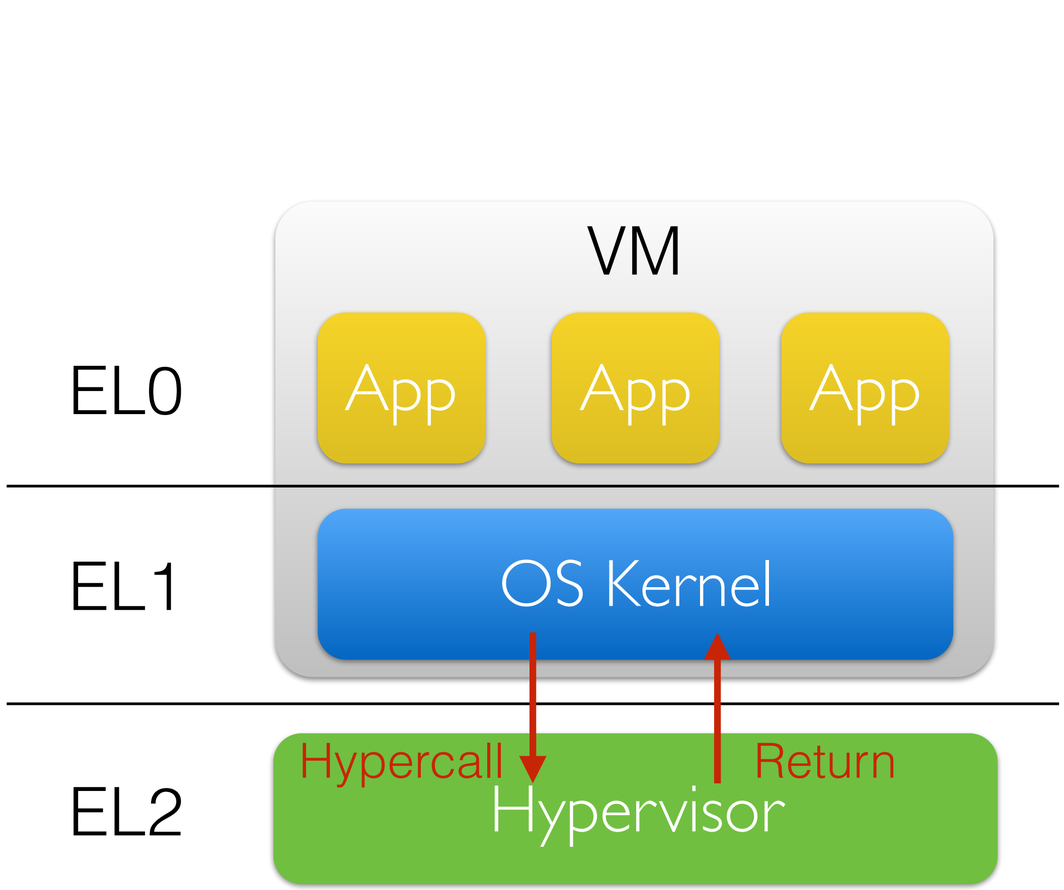
Performance Evaluation

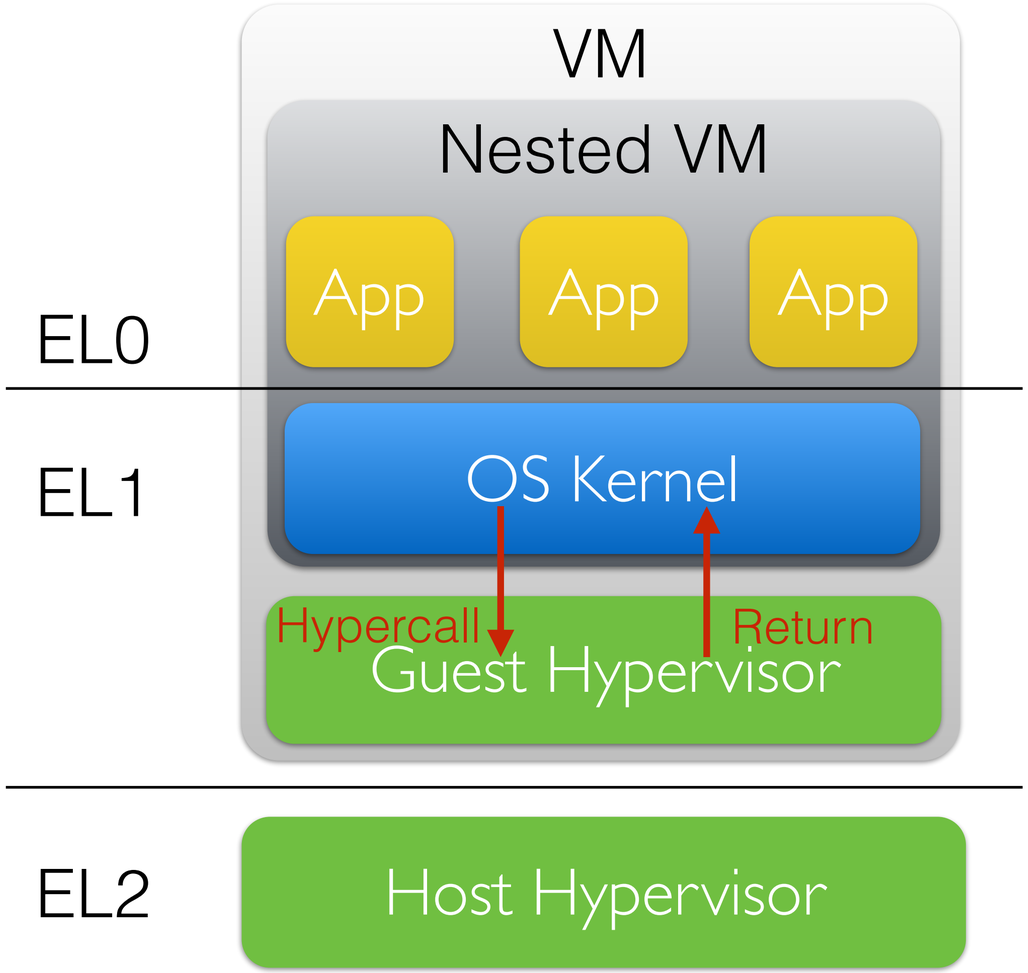
- Problem: No ARMv8.3 hardware available.
- Solution: Use ARMv8.0 hardware with the software modification

Emulating v8.3 on v8.0



Hypercall MicroBenchmark





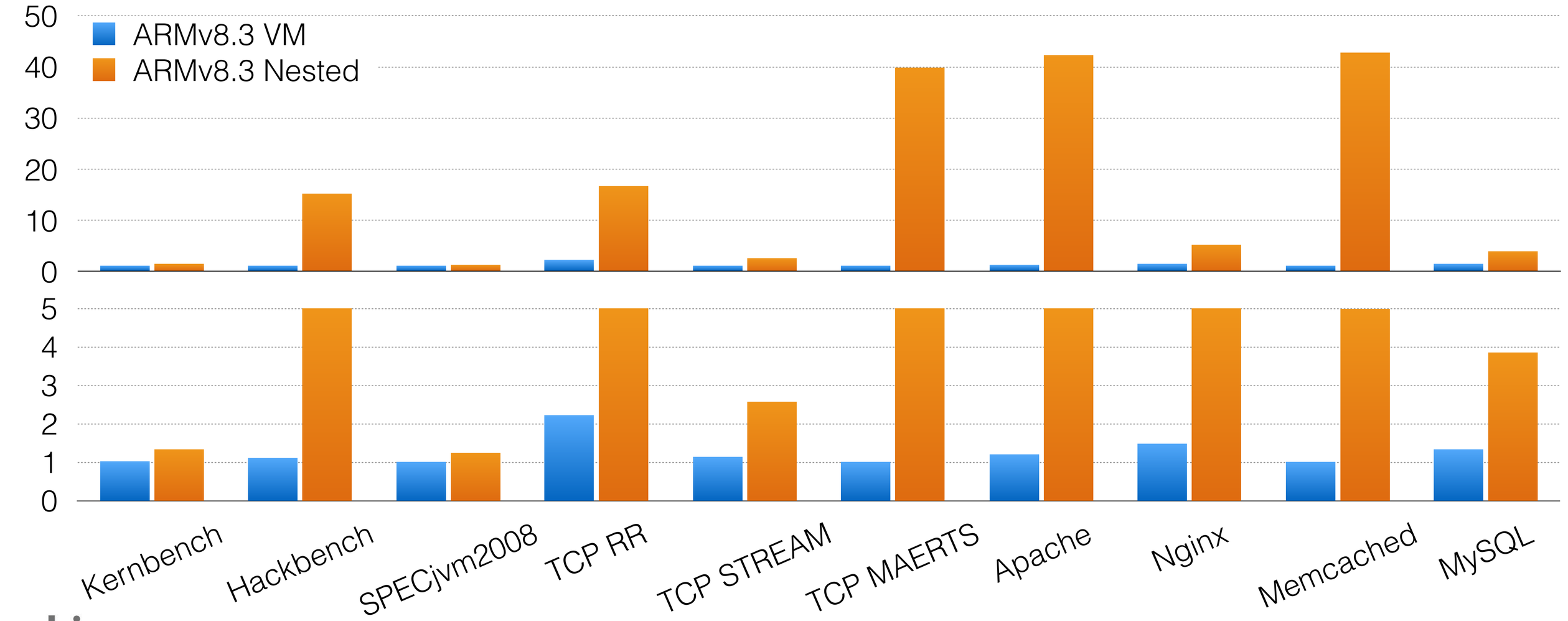
Linaro

Hypercall MicroBenchmark

	ARMv8.3	
	VM	Nested VM
Cycle counts	2,729	422,720
Ratio to VM	1	155x

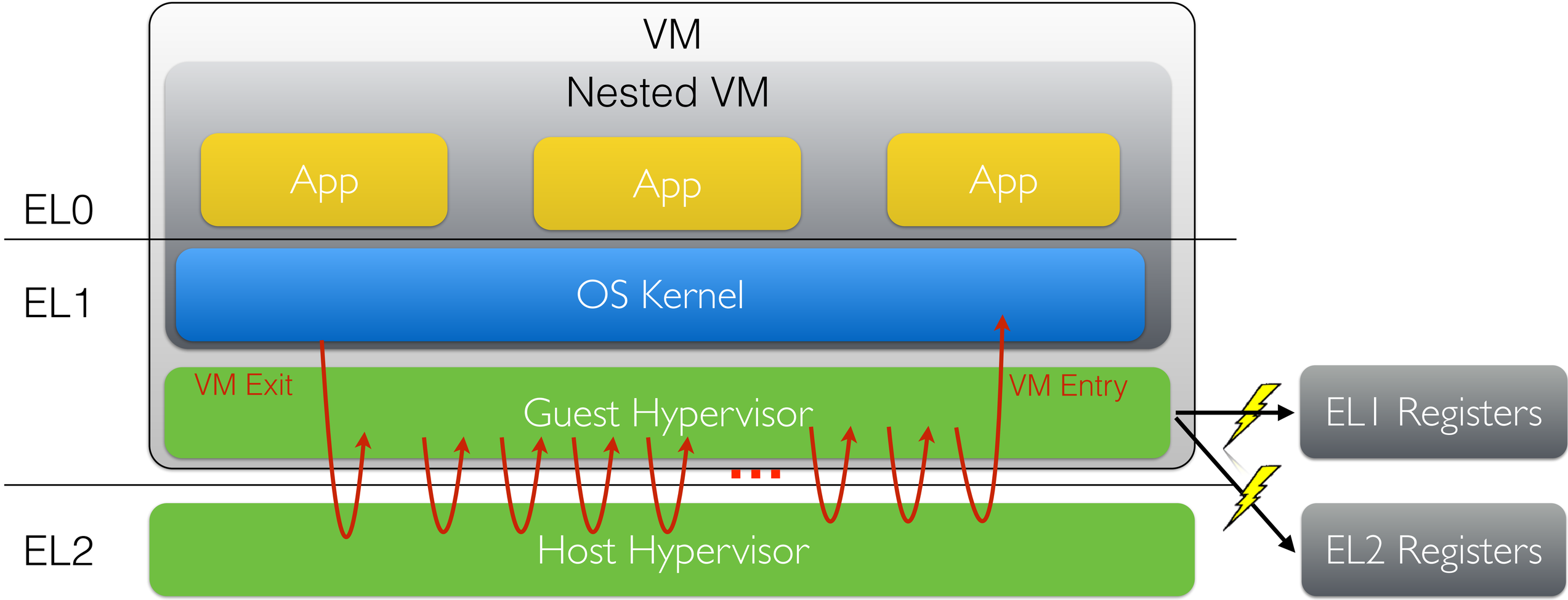
Linaro

Application Benchmarks



Linaro

Nested VM Exit/Entry on ARM



> 120 traps

Linaro

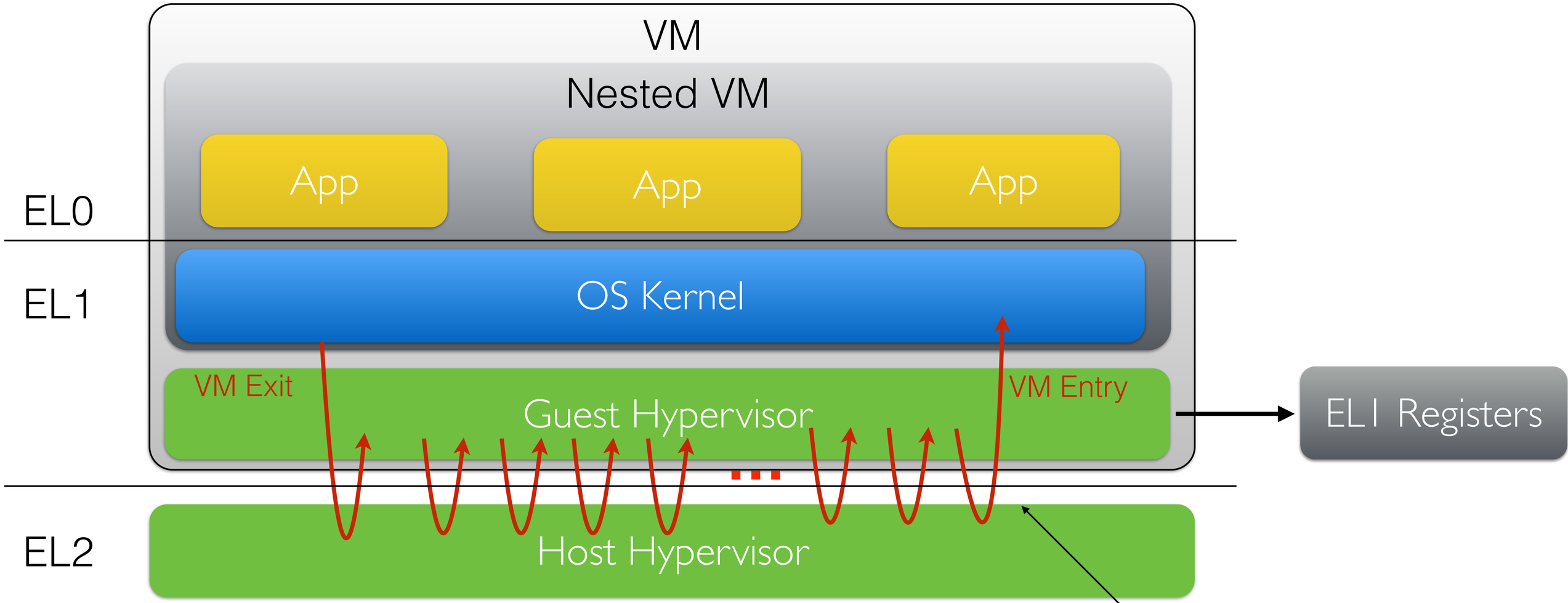
NEVE: NEsted Virtualization Extensions for ARM

- Supports unmodified guest hypervisors and OSes
- Improves performance by providing register redirection

Register Classification

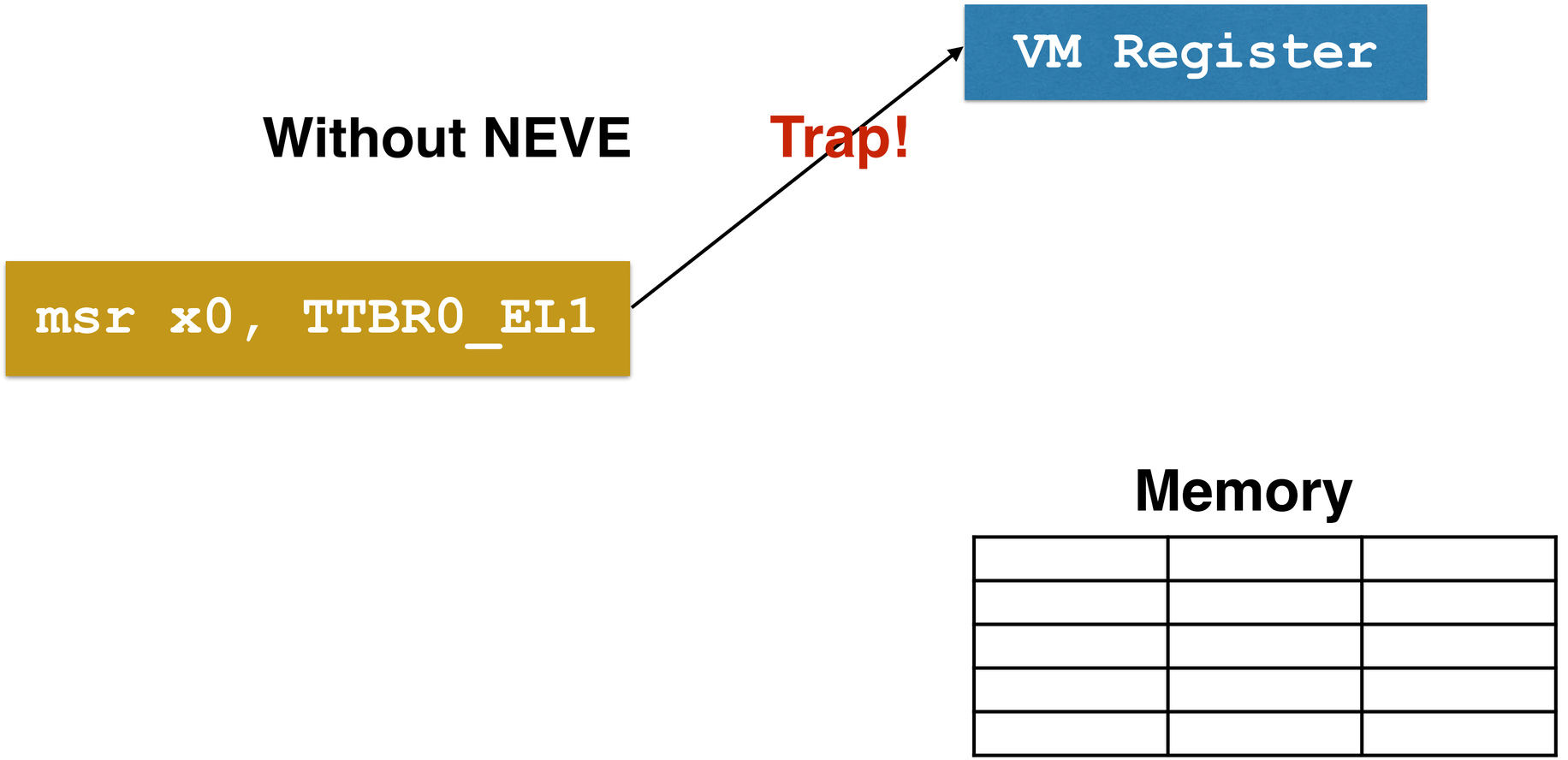
- VM registers: EL1 registers only affecting the nested VM’s execution
- Hypervisor registers: EL2 registers affecting the hypervisor’s execution

VM Registers

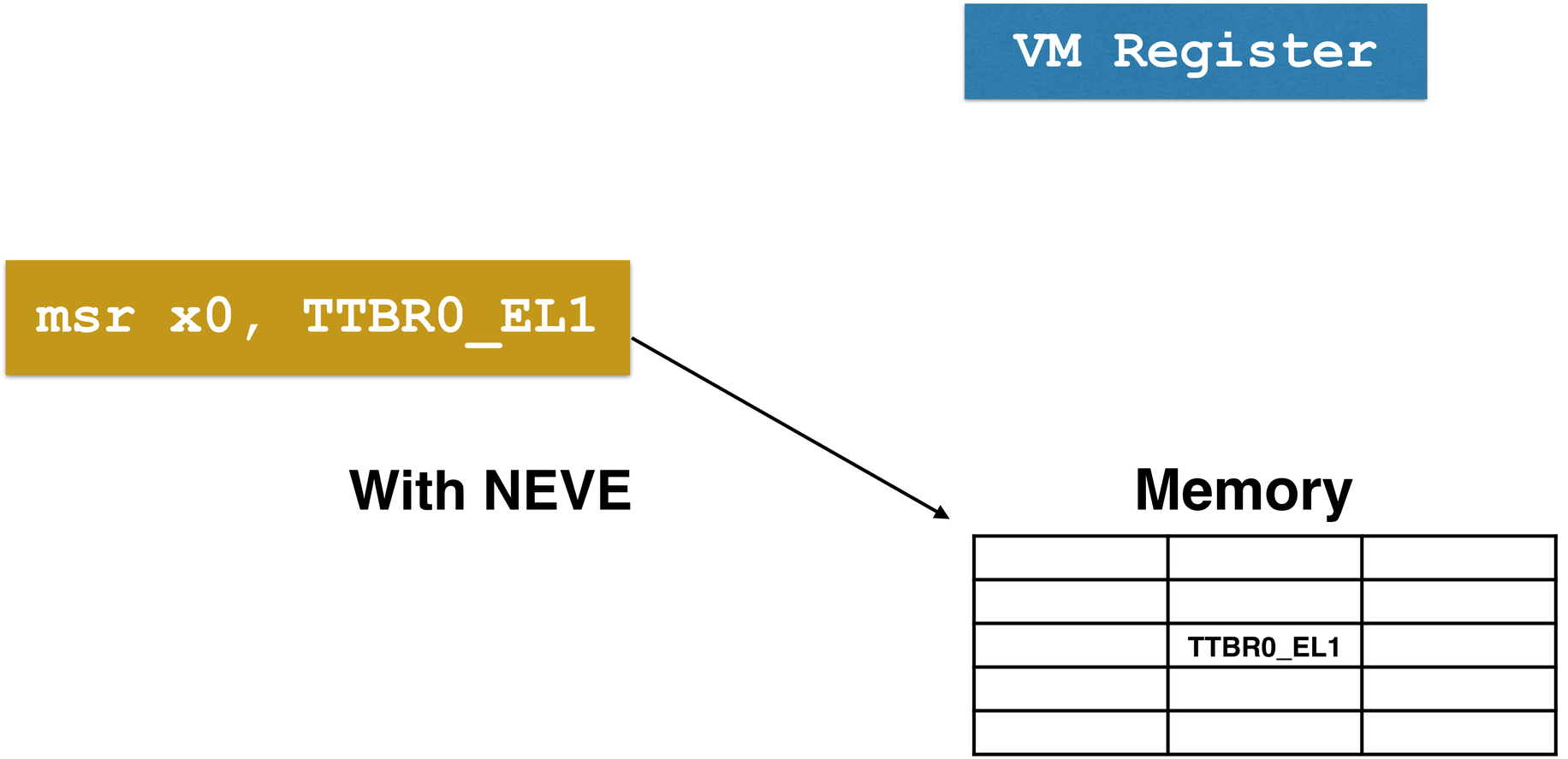


This is when VM register states are used

VM Registers: Logging to Memory

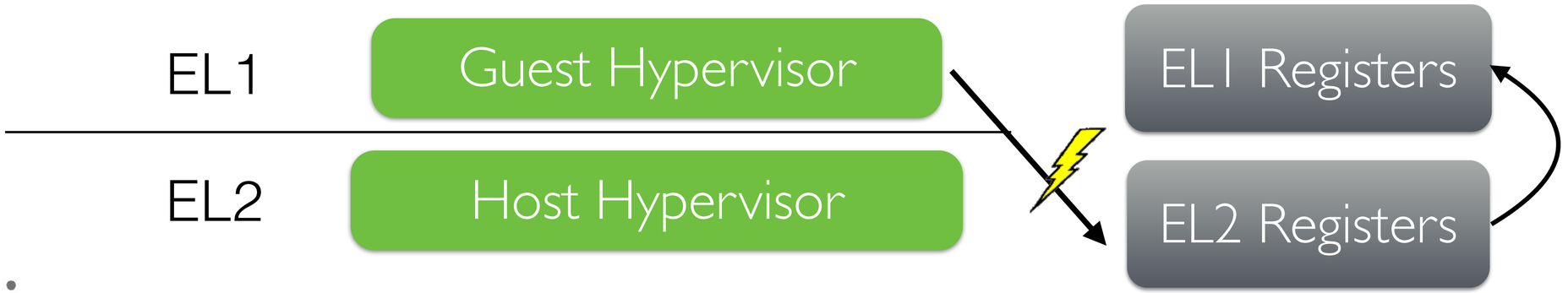


VM Registers: Logging to Memory



Hypervisor control registers

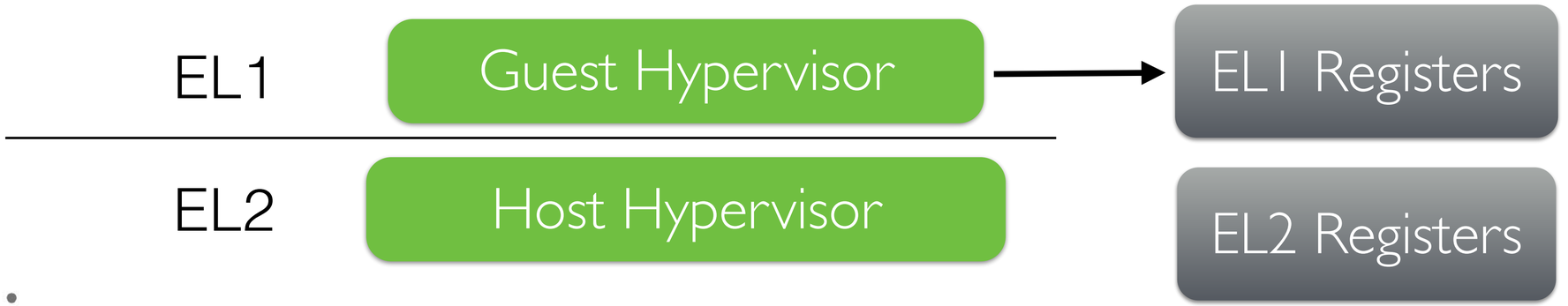
- Can’t apply the technique for VM registers
  - They have an immediate impact (EL2 system registers)
- Traps are handled by redirecting to EL1 registers in software



Linaro

Hypervisor control registers

- Can’t apply the technique for VM registers
  - They have an immediate impact (EL2 system registers)
- Traps are handled by redirecting to EL1 registers in software
- Redirect in hardware instead!



Hypercall MicroBenchmark

	ARMv8.3		NEVE
	VM	Nested VM	Nested VM
Cycle counts	2,729	422,720	**92,385**
Ratio to VM		155x	**34x**
Trap counts	1	126	**15**

Application Workloads

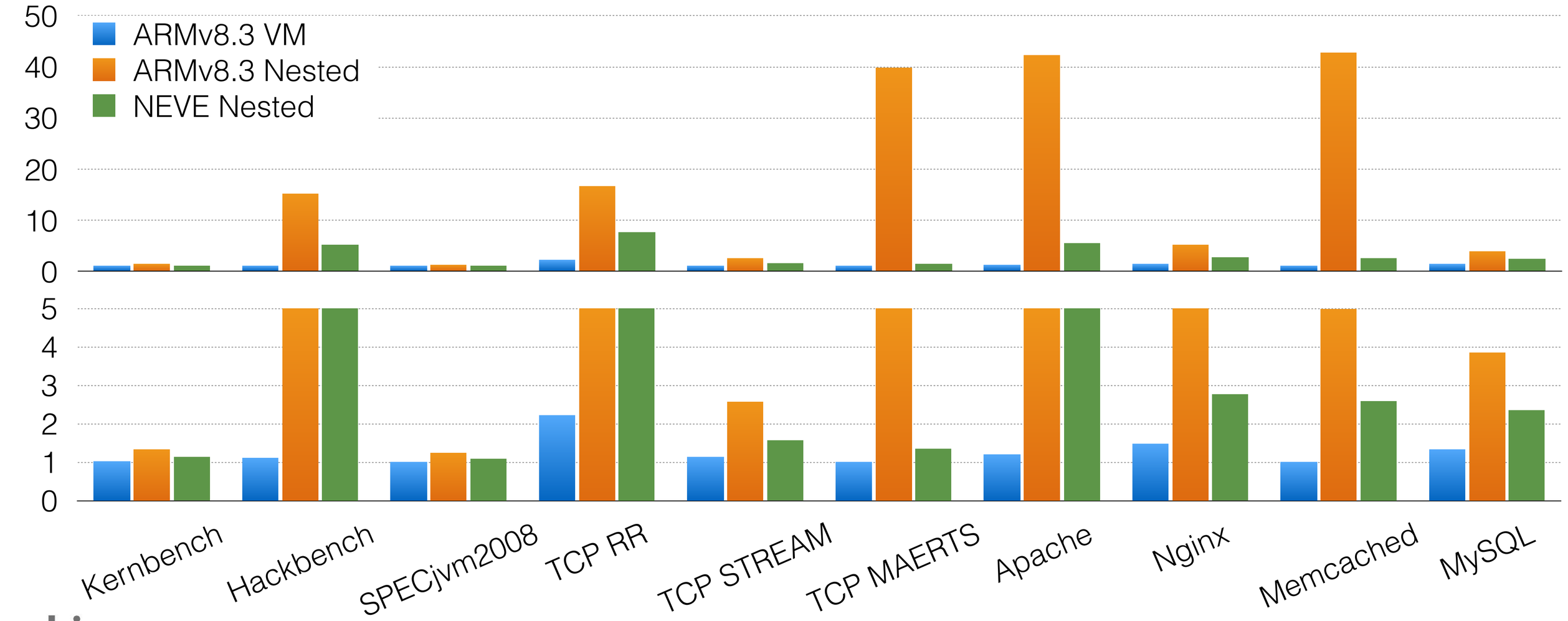
Application	Description	Application	Description
Kernbench	Kernel compile	Netperf TCP_RR	Network performance
Hackbench	Scheduler stress	Netperf TCP STREAM	Network performance
SPECjvm2008	Java Runtime	Netperf TCP MAERTS	Network performance
MySQL	Database management	Apache	Web server stress
Memcached	Key-Value store	Nginx	Web server stress

Experimental Setup

- ARM Hardware
  - APM X-Gene (ARMv8.0)
  - 8-way SMP
  - 64 GB RAM
- x86 Hardware
  - Intel E5-2630 v3
  - VMCS Shadowing
  - 8-way SMP
  - 128 GB RAM
- Native/VM/Nested VM
  - 4-way SMP
  - 12 GB RAM
  - Virt I/O (VM/nested VM)
  - 10 Gb Ethernet
- Software
  - KVM on KVM
  - v4.10

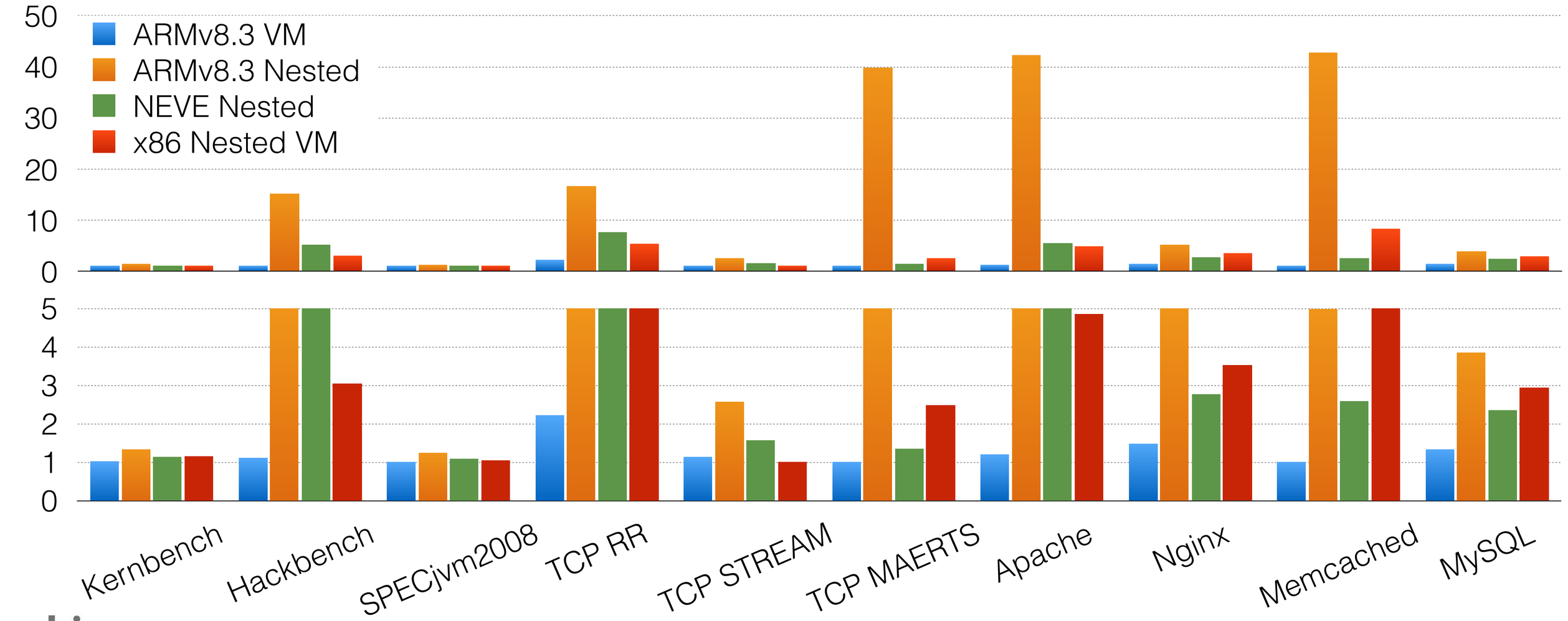
Linaro

Application Benchmarks



Linaro

Application Benchmarks



Linaro

Conclusion

- We have an implementation of KVM/ARM for v8.3
- Evaluated nested virtualization performance by emulating ARMv8.3
- Nested virtualization on ARMv8.3 incurs high overhead
  - Due to the exit multiplication problem
- NEVE enhances performance significantly by reducing number of traps
- NEVE is used as basis for extended nested virtualization support in ARMv8.4
- NEVE to appear at SOSP later month - read the paper for more details

Code

- Nested CPU Virtualization patches for ARMv8.3 [RFC v2]: https://lists.cs.columbia.edu/pipermail/kvmarm/2017-July/026388.html
- Nested Memory Virtualization patches for ARMv8.3 [RFC]: https://lists.cs.columbia.edu/pipermail/kvmarm/2017-October/027286.html
- v8.3 and NEVE Paravirtualization on Linux v4.12-rc1: https://github.com/columbia/nesting-pub
- QEMU Patches: https://github.com/columbia/qemu-pub nested-v2.3.0-model