Hardware-Assisted PMU Virtualization

Manali Shukla, Sandipan Das

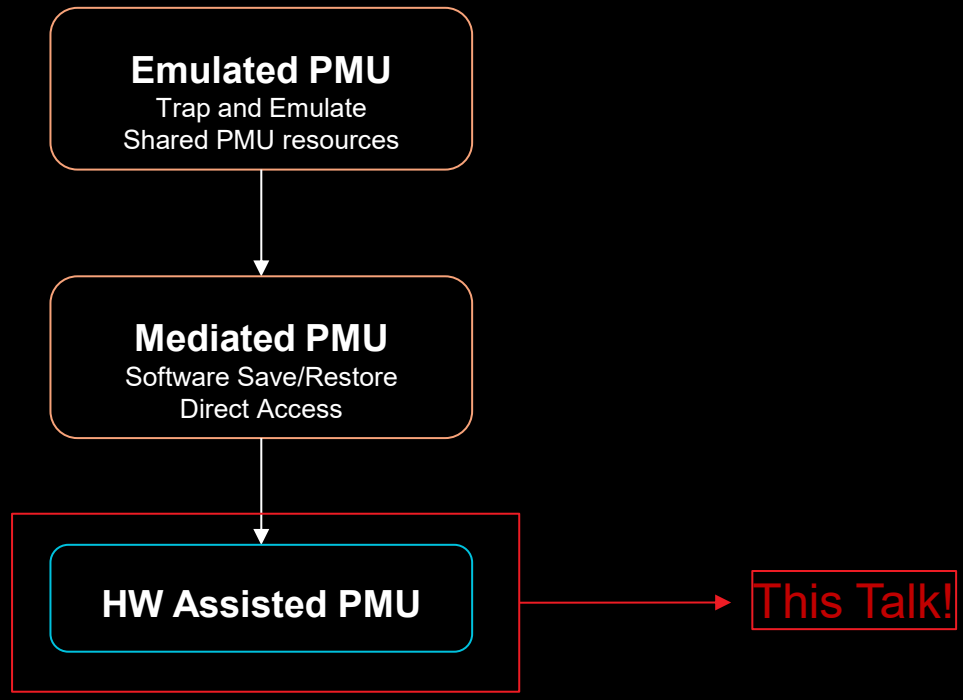
AMD
together we advance_

Agenda

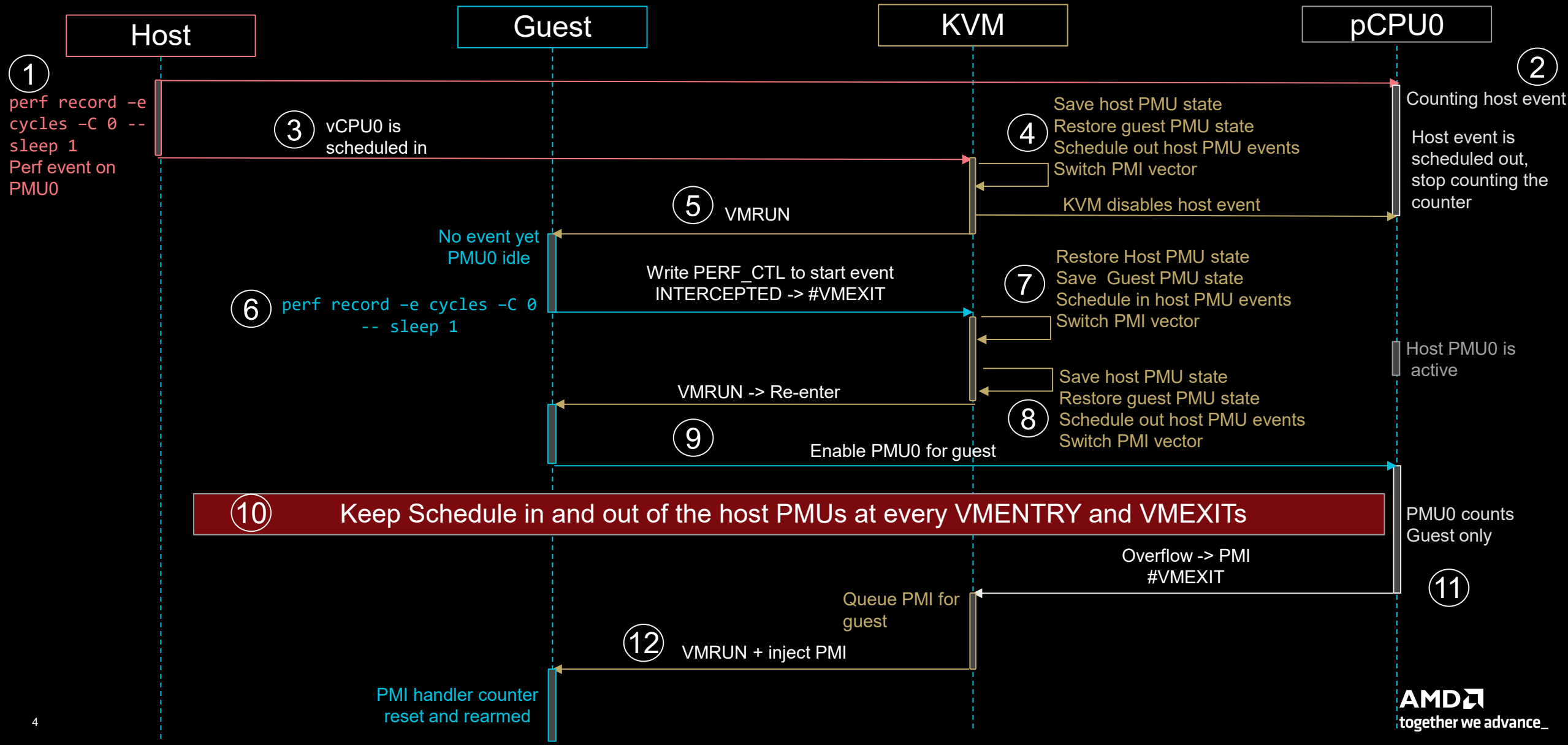
- Understand Hardware-Assisted PMU Virtualization
- Explore the PMU features it supports
- Cover the CPU support for these features

Where we left off: Emulated PMU → Mediated PMU

- **Emulated PMU** - *trap & emulate*: PMU resources are shared between host and guest and blurs the host–guest boundary.
- **Mediated PMU** - the host yields the PMU to the guest around the world switch; the guest gets **direct access to PMU resources**.
- **This talk -** what **hardware** adds on top



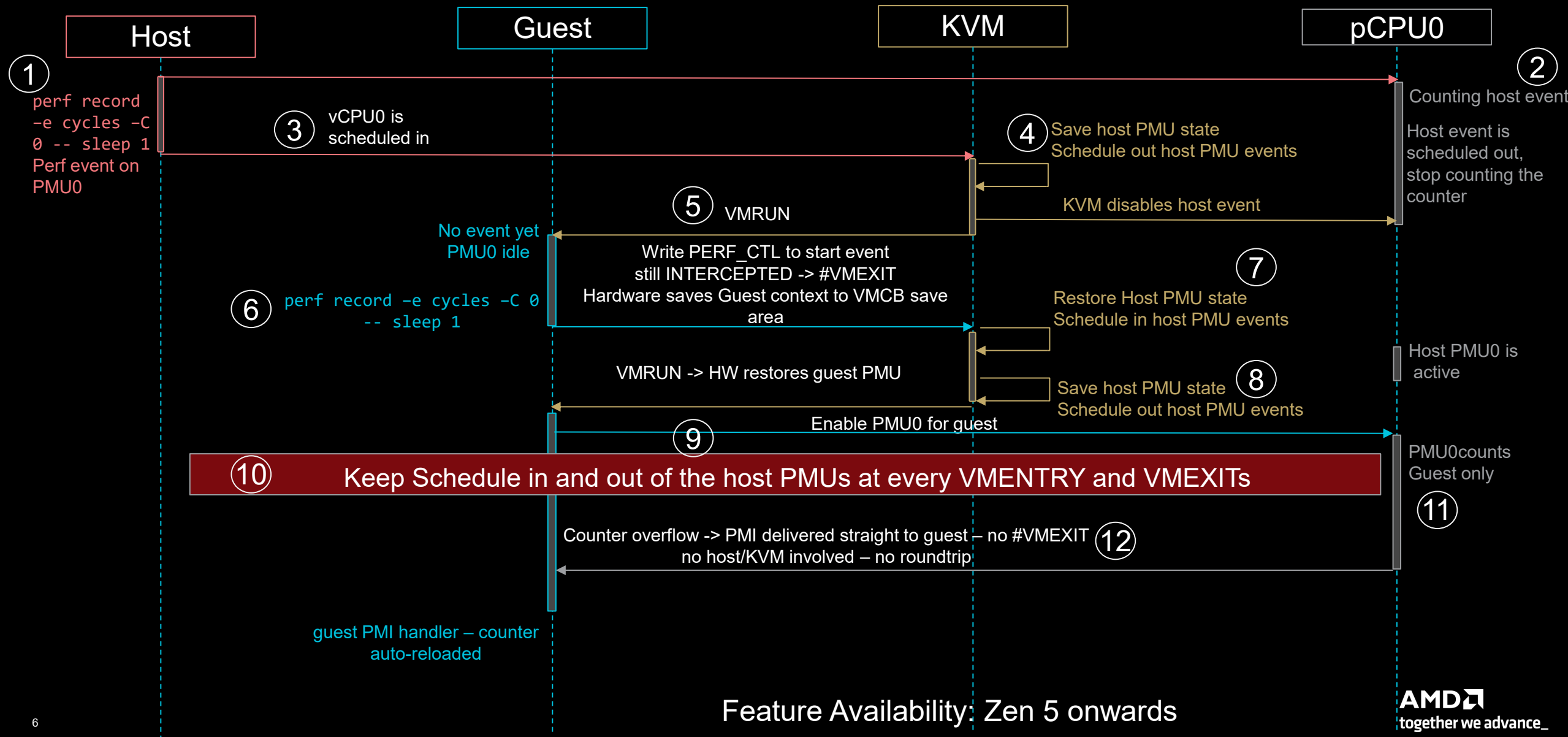
Recap: Mediated PMU



Virtualization Building Blocks (AMD Specific)

- **VMCB** → the vCPU's control block. Its **State Save Area** holds the guest's CPU/MSR state across switches.
- **VMRUN / #VMEXIT** → the **world switch**: enter the guest, or snap back to the hypervisor. Every exit costs time.
- **Swap type C** → hardware swaps the **guest** copy via the VMCB, but the **host** copy is the hypervisor's job. The **mediated PMU** does that host save/restore
- MSR interception → cleared for all MSRs

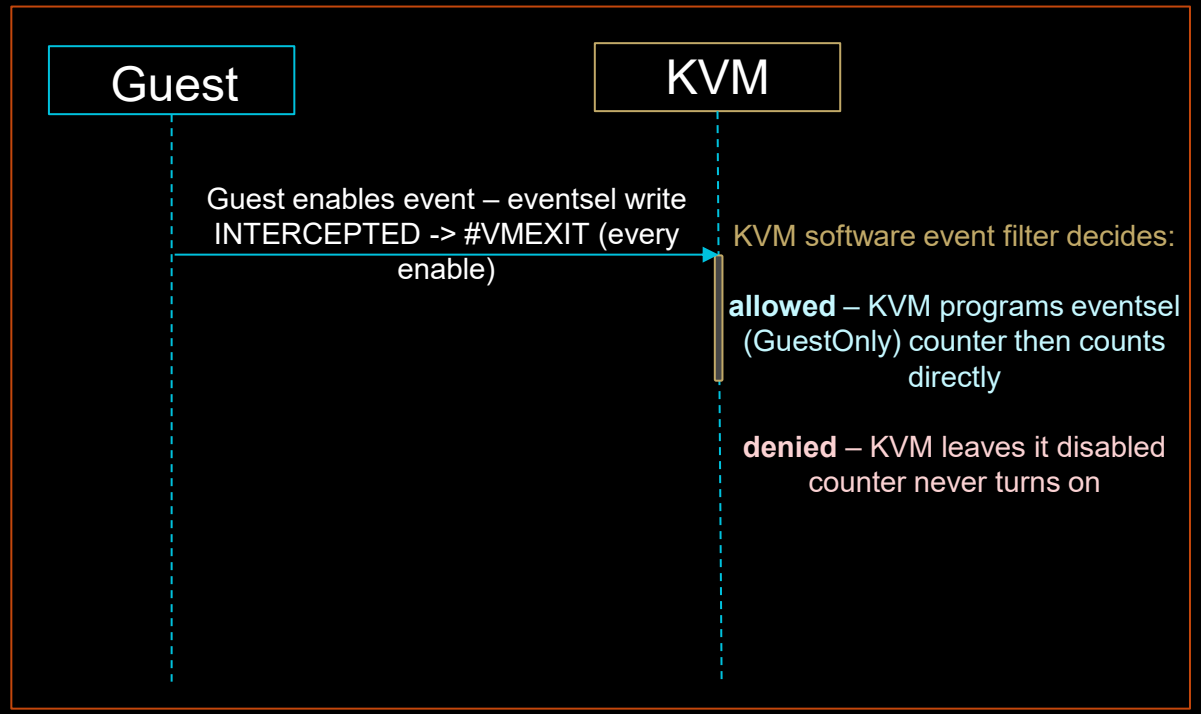
VPMC: Hardware Assisted PMU virtualization



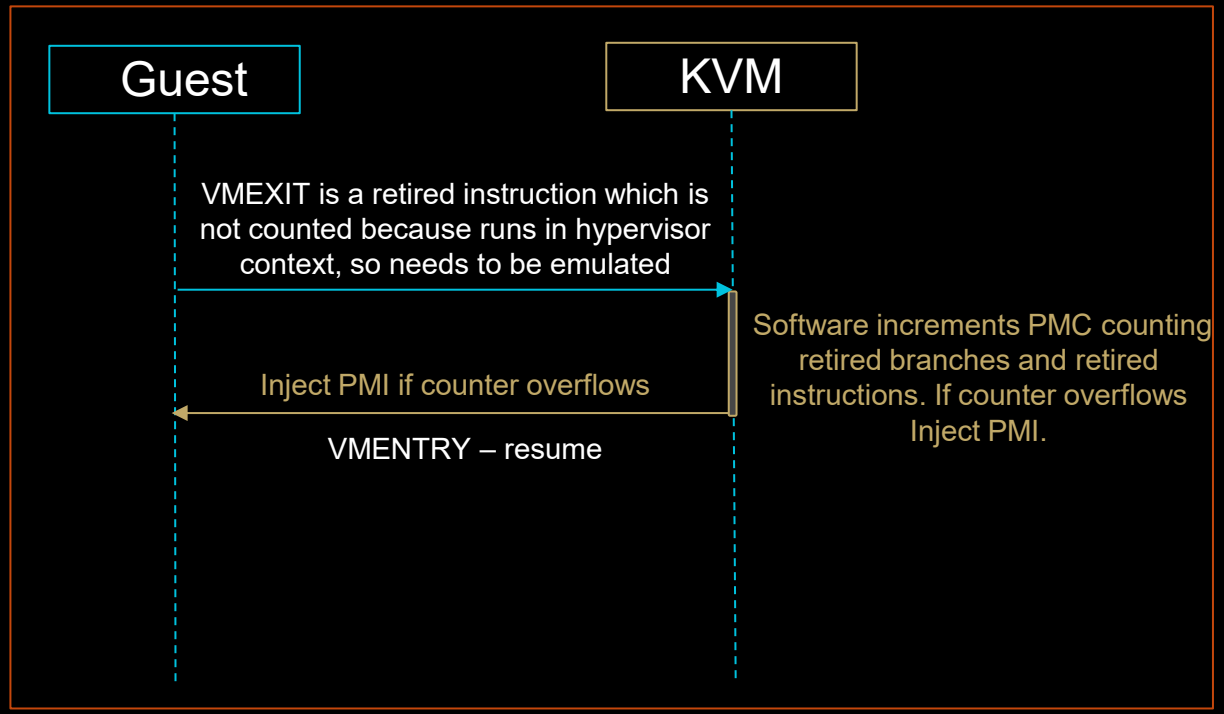
Feature Availability: Zen 5 onwards

VPMC: Selective interception

Software Event filtering



Instruction Emulation



Delivering Interrupts

- When only AVIC is enabled:
  - ✓ Guest's LVTPC is honored
  - ✓ Guest can pick VNMI or INTR mode

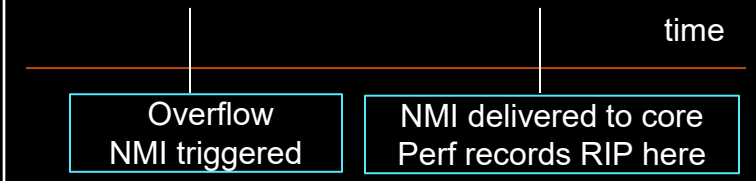
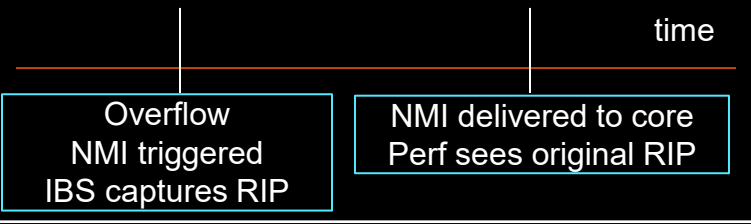
- When only VNMI is enabled:
  - ✓ Overflows always delivered as NMI

When AVIC and VNMI are enabled, hardware automatically prefers **AVIC**.

Software-synthesized overflows (emulated instructions) still use KVM's interrupt injection path.

Instruction Based Sampling

An AMD feature for *precise* profiling. Instead of just counting events, IBS tags a random µop and records rich detail about its execution - exact RIP, data address, cache behavior, latency, branch resolution.

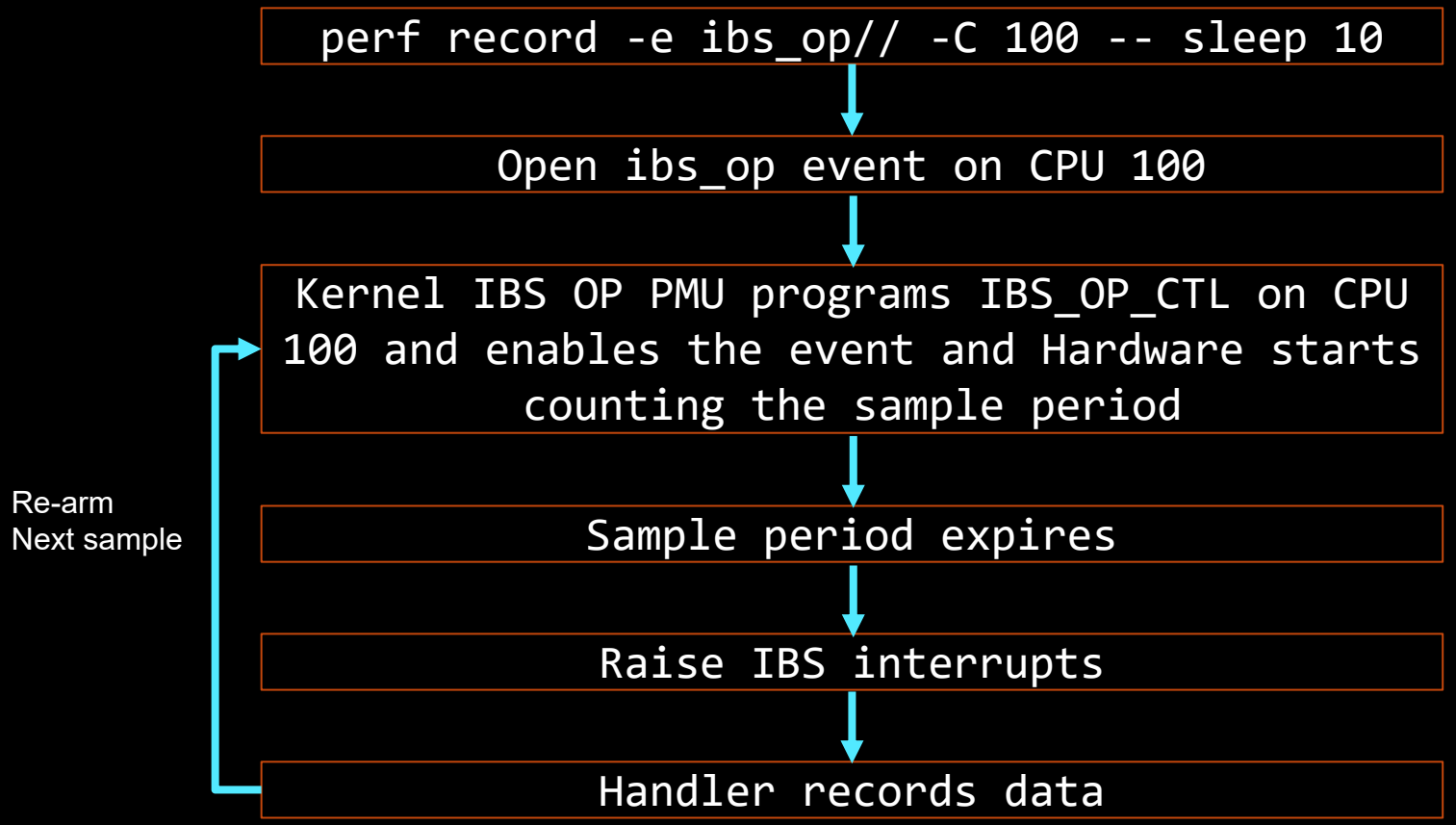
Core PMU	IBS
Imprecise 	Precise 
No Additional data from the hardware	Hardware provides interesting information along with the samples
Counting (perf stat) & Sampling (Perf record) modes	Only Sampling (Perf record) mode
overflow → **skid**	**exact** op, no skid
6 Counters (PMCs)	1 counter for each PMU (op and fetch)

IBS: Real Examples

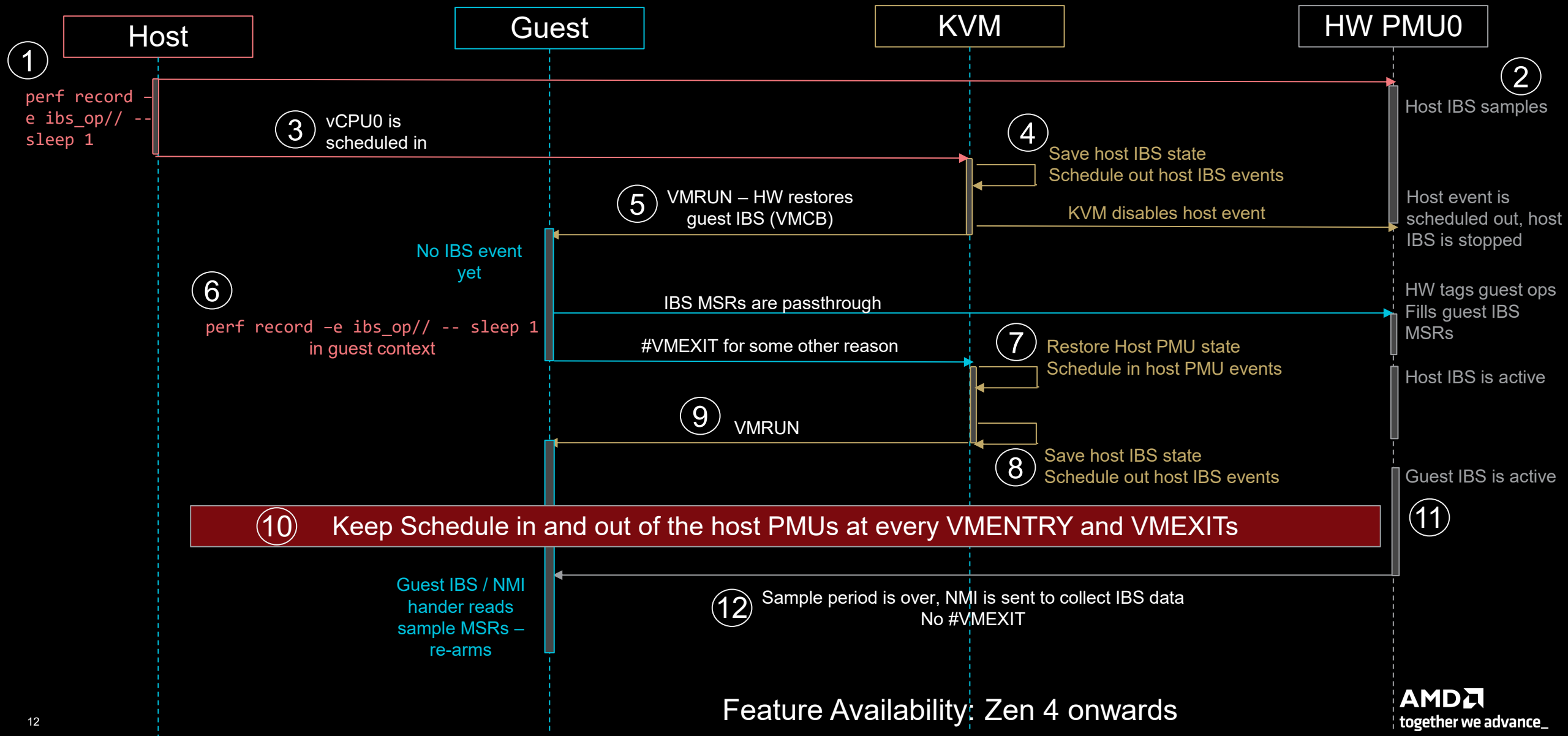
https://arstechnica.com/gadgets/2022/09/20-year-old-linux-workaround-is-still-slowing-down-amd-systems/

How did IBS help with finding the root cause of this issue?

The IBS sampling flow



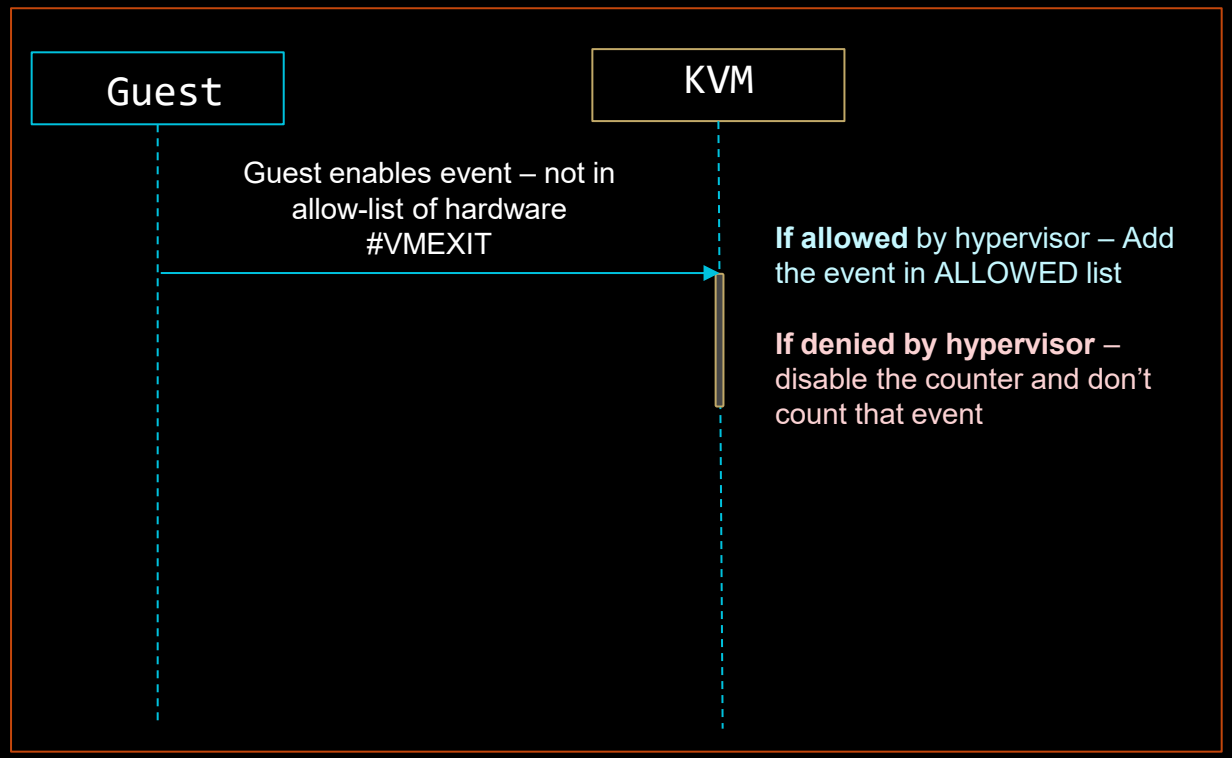
IBS virtualization



Feature Availability: Zen 4 onwards

Hardware-Assisted PMC filtering

- **Software based PMC filtering**
  - KVM intercepts every event-enable and decides whether to allow it. Flexible, but pays a VM-exit each time.
- **Hardware-Assisted PMC Filtering**
  - KVM hands the allow-list to the CPU; the guest then counts at native speed with the policy enforced underneath. Allow-listed events run **passthrough** (no exit), while non-allow-listed events are **intercepted** on enable - cutting the number of VM-exits.
  - Support the KVM being able to filter which PMC events a guest is allowed to enable to help prevent PMC-based side channel attacks.



Mediated PMU vs Hardware-Assisted PMU

Guest instance (single vCPU) – Perform a top-down analysis

```
$ sudo perf stat –M PipelineL1  -- sysbench --cpu-max-prime=1000 --threads=1 cpu run
```

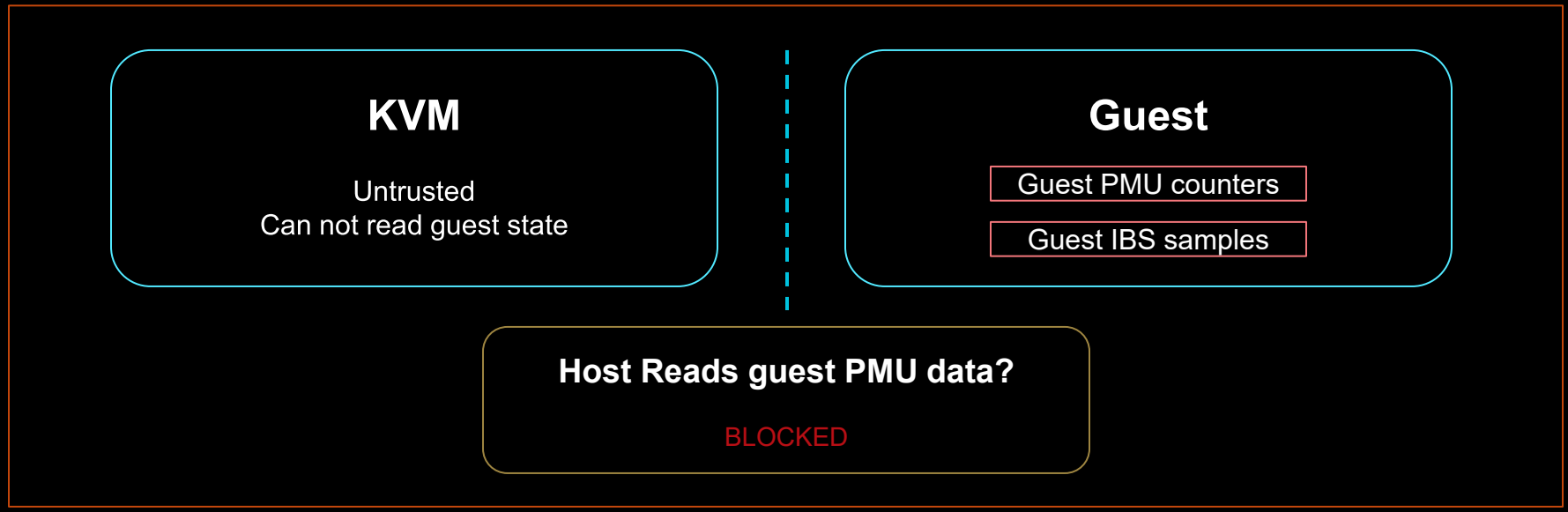
Host instance – Count VM-Exits for hardware counter accesses

```
$ sudo perf stat –e kvm:kvm_msr --filter="((ecx >= 0xc0010000 && ecx <= 0xc0010007) ||
(ecx >= 0xc0010200 && ecx <= 0xc001020b) || (ecx >= 0xc0000300 && ecx <= 0xc0000303))"
```

101,528 kvm:kvm_msr	101,264 kvm:kvm_msr	7,511 kvm:kvm_msr
Mediated PMU No of Exits: 101,528	VPMC No of Exits: 101,264	Hardware-Assited PMC filtering No of Exits: 7511

“VPMC and Mediated PMU will always intercept PERF_CTL MSRs irrespective of the event code and whether we are enabling or disabling the counters”

Confidential Guests



“VPMC virtualizes the counters, PMC filtering makes that safe, and VIBS adds precise, in-guest instruction sampling — all built on the Mediated-PMU model so guests can profile at near-native speed without frequent VM-Exits.”

Current Status

- PMC virtualization (RFC has been posted)
  https://lore.kernel.org/kvm/cover.1762960531.git.sandipan.das@amd.com/

- IBS virtualization (V3 has been posted)
  https://lore.kernel.org/kvm/202603100600022.15120-1-manali.shukla@amd.com/

References

- Mediated PMU (Patches are already available in upstream)
  https://lore.kernel.org/all/20251206001720.468579-1-seanjc@google.com/

- AMD64 Guest PMC Event Filtering
  https://docs.amd.com/api/khub/documents/2eieIP7KVNk7sKFJ9Ew~Ng/content

- AMD64 Architecture Programmer's Manual
  https://docs.amd.com/v/u/en-US/40332_4.09_APM_PUB

Copyright and disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. The information contained herein is subject to change and may be rendered inaccurate releases, for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

THIS INFORMATION IS PROVIDED 'AS IS." AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS, OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION. AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, DIRECT, INDIRECT, SPECIAL, OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION


AMD