Scalable Work Submission in Device Virtualization

Hao Wu hao.wu@intel.com

Oct. 30th 2020

Disclaimers

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request. No product or component can be absolutely secure.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.

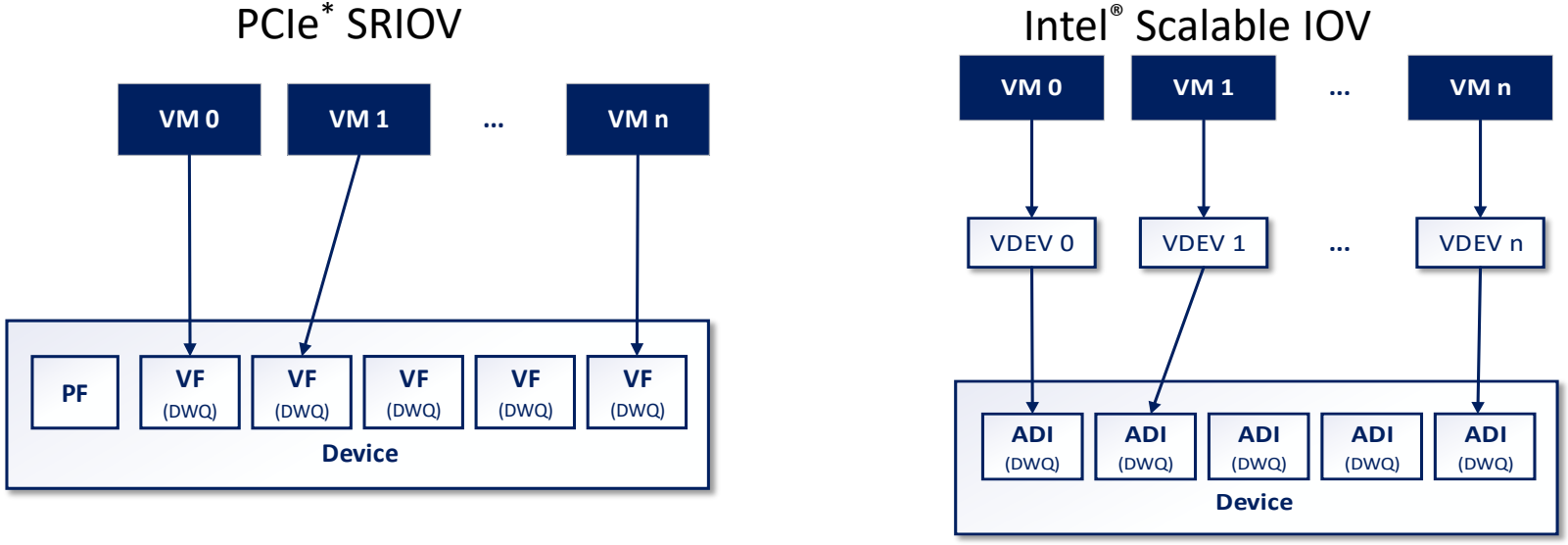
Intel and the Intel logo are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

Agenda

- Scalability in Device Virtualization
- ENQCMD Instruction
- ENQCMD Virtualization
- Example - SVA work submission

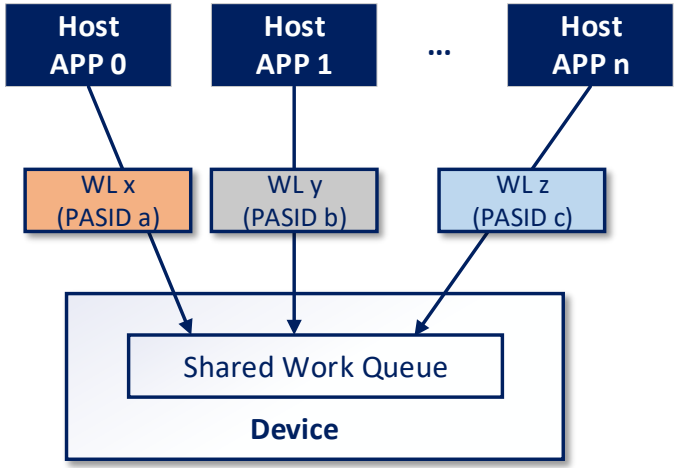
Scalability In Device Virtualization



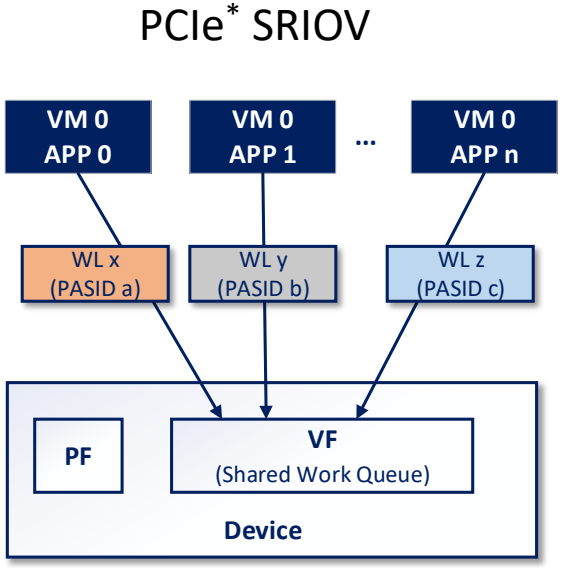
- **Dedicated work queue (DWQ)** implemented in VFs and ADIs
- Provide Scalability by **hard partitioning the hardware resources**
- Difficult to increase VFs / ADIs due to limited resources on some devices

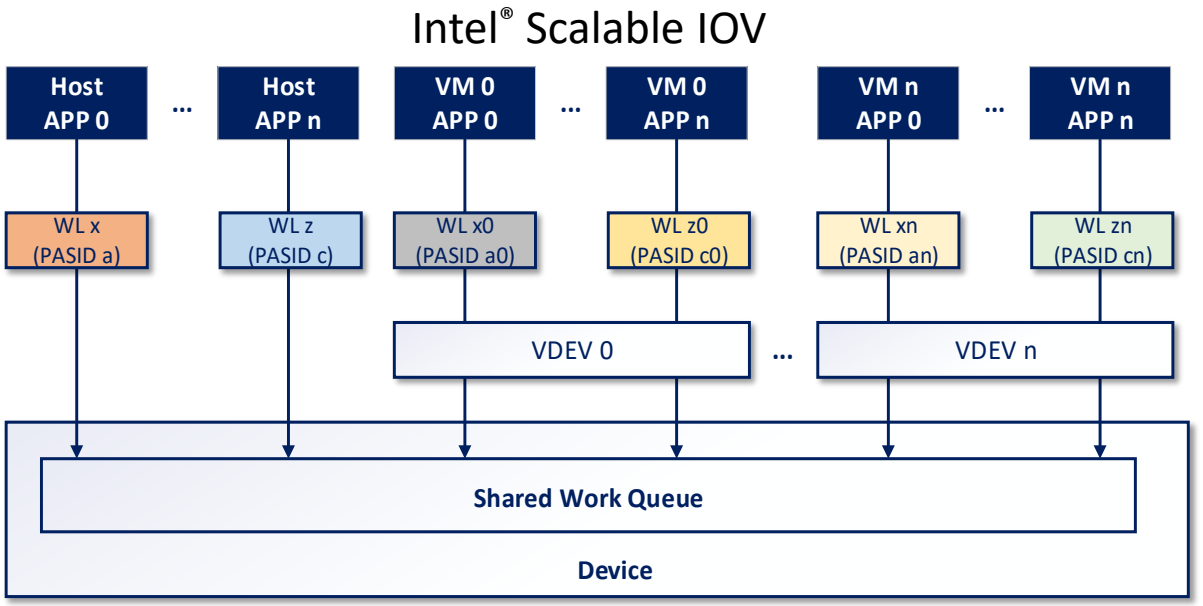
Shared Work Queue

- Typical usage: Shared Virtual Addressing (SVA)
  - Device uses the CPU virtual address for DMA
- Distinguish the context of different workloads by Process Address Space ID (PASID)
- DMA address translation at Requestor ID (RID) + PASID granularity per IOMMU



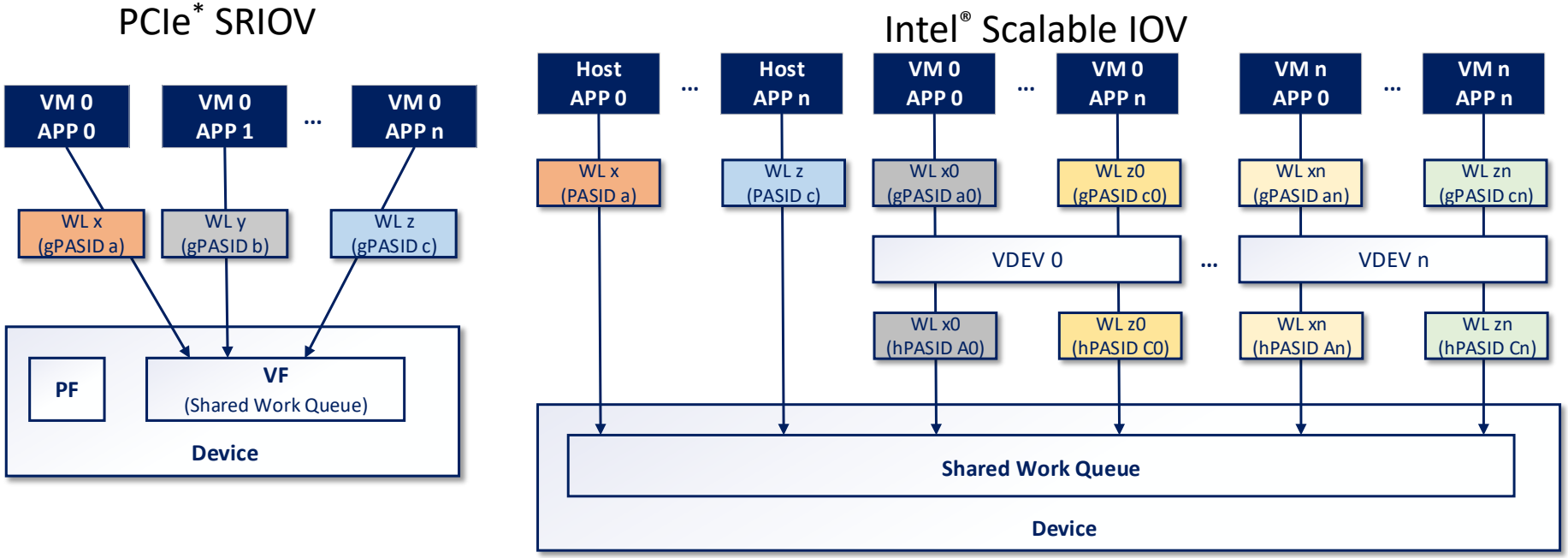
Shared Work Queue In Device Virtualization





- Allow sharing the same device interface by users in VMs and Host
- No hard limitation on user number
- Device can implement DWQ and SWQ together

Challenge In Device Virtualization



- Challenge: How to convert guest PASID to host PASID

ENQCMD Instruction - Overview

- New instruction on Intel® Platforms
- Atomically submit a work with PASID
  - Obtains PASID from IA32_PASID MSR
  - Enqueue store 64B command to enqueue register in device MMIO

- IA32_PASID is managed by XSAVE as PASID supervisor state
- Non-Posted instruction which carries back a status
  - ZF flag indicates if the command was accepted by device
  - Allows user to retry

* Figure is from Intel ® architecture instruction set extensions spec, see link in reference page page

ENQCMD Instruction - ENQCMDS

- ENQCMDS (Enqueue Command Supervisor)
  - Similar to ENQCMD
  - Used in kernel space only
  - Obtain PASID value from command data.

ENQCMD Instruction – Device Requirement

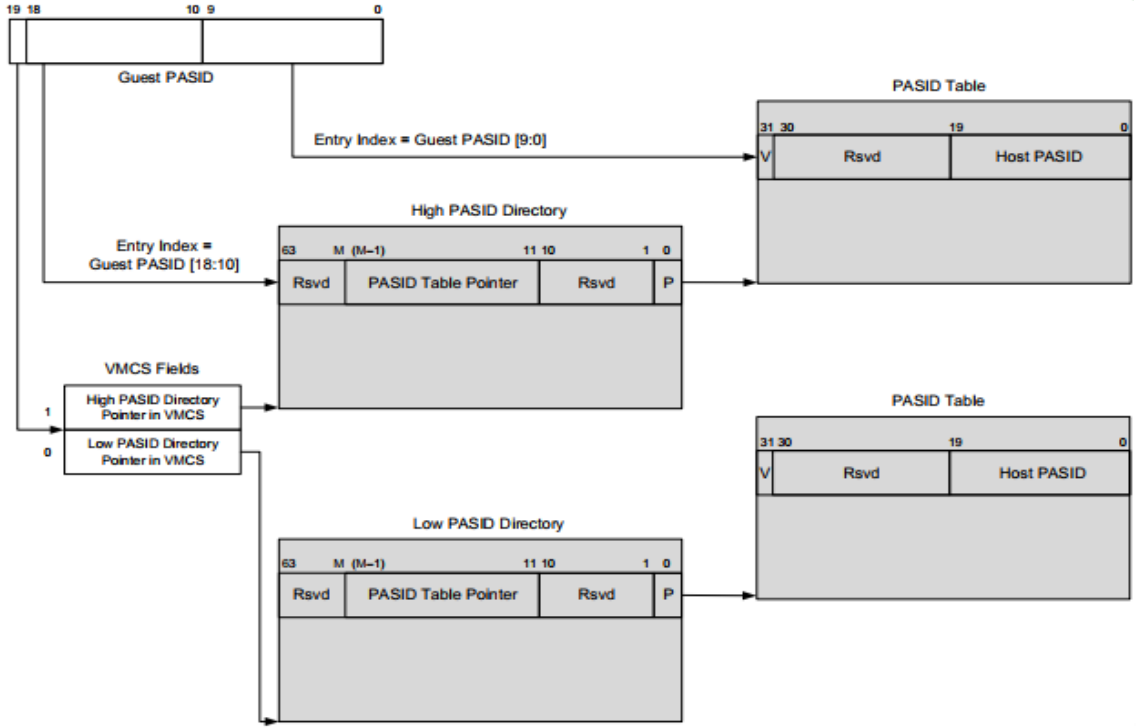
- Support DMWr (Deferrable Memory Write) completer capability.
- All switch ports and root ports have DMWr routing enabled.
- Intel® Data Streaming Accelerator is the first device which supports ENQCMD
  - https://lkml.org/lkml/2020/9/24/1056

ENQCMD Virtualization - Non-Root Mode Operation

- ENQCMD/ENQCMDS obtain guest PASID
- Perform guest PASID to host PASID Translation
- Enqueue store command data with host PASID to device

ENQCMD Virtualization - PASID Translation Support

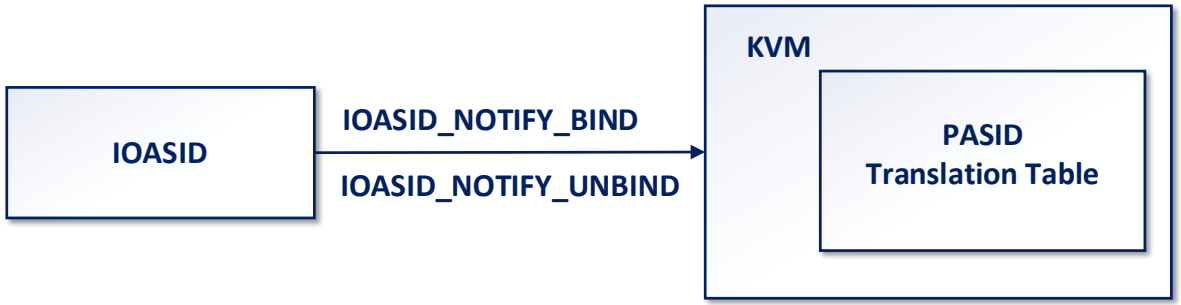
- New feature in VMX on Intel® Platforms
- Use PASID Translation Table for guest PASID to host PASID translation
- Trigger VM-Exit if fails to translate guest PASID



* Figure is from Intel ® architecture instruction set extensions spec, see link in reference page page

Manage PASID Translation Table in KVM

- Update translation per IO Address Space ID (IOASID) events
  - IOASID manages host PASID and its association to guest PASID
  - Monitor IOASID BIND/UNBIND events for translation update



Refer to “PASID Management in KVM” KVM Forum Session

- https://kvmforum2020.sched.com/event/eE4v/pasid-management-in-kvm-yi-liu-jacob-pan-intel

Manage PASID Translation Table in KVM

- PASID Translation Table is a per VM table shared by all VMCS

- Modification must be a rendezvous operation
  - Kick all VCPUs into root mode and block VM entry until modification is done
  - Required by SDM 24.11.4, when modify data structure which is referenced by pointers in VMCS and controls non root mode operation.

PASID Translation Failure (VM-Exit) Handling in KVM

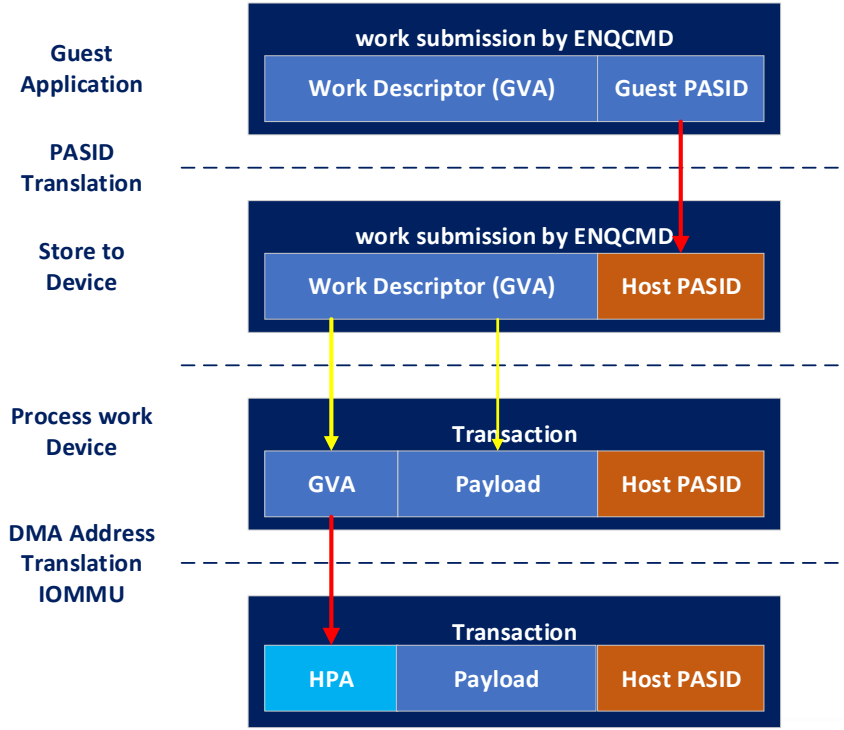
- Translation failure (VM-Exit) only happens with invalid guest PASID
  - Must be associated with a host PASID for DMA operation
- Set the ZF = 1 to indicate the failure and skip the instruction

IA32_PASID MSR virtualization in KVM

- Passthrough IA32_PASID MSR
- Enable virtualization support for XSAVE PASID supervisor state component

Example – SVA Work Submission In Guest

- Prepare a Work Descriptor
- Submit by ENQCMD
  - PASID is translated automatically
  - Store Work Descriptor (GVA) + Host PASID to Device
  - Check ENQCMD instruction status
- Device performs DMA operation
  - GVA + Payload + Host PASID
- DMA address translation per RDI + PASID by IOMMU
  - GVA -> HPA



References

- Kernel Doc “Shared Virtual Addressing (SVA) with ENQCMD”- Documentation/x86/sva.rst by Ashok Raj <ashok.raj@intel.com>
- ENQCMD in Intel® Architecture Instruction Set Extensions - https://software.intel.com/sites/default/files/managed/c5/15/architecture-instruction-set-extensions-programming-reference.pdf
- Intel® Scalable IOV: https://01.org/blogs/2019/assignable-interfaces-intel-scalable-i/o-virtualization-linux
- Intel® Data Streaming Accelerator Spec: https://software.intel.com/sites/default/files/341204-intel-data-streaming-accelerator-spec.pdf

Status

- ENQCMD native support: merged.
- IOASID extensions for notification: v3 submitted.
  - https://lkml.org/lkml/2020/9/28/1186
- ENQCMD virtualization support: will submit soon (internal review now)
  - TODO: Live migration support

Summary

- Dedicated Work Queue (DWQ) based on hard partitioning of resources, has scalability limitation in virtualization.
- Shared Work Queue (SWQ) with ENQCMD support allows more scalable usage in device virtualization, as same device interface can be shared by multiple users in host and VMs.
- Additional hardware support is required to support ENQCMD virtualization, e.g. PASID translation, XSAVE extension for PASID state and etc.


KVM
FORUM