Progress of Warpdrive

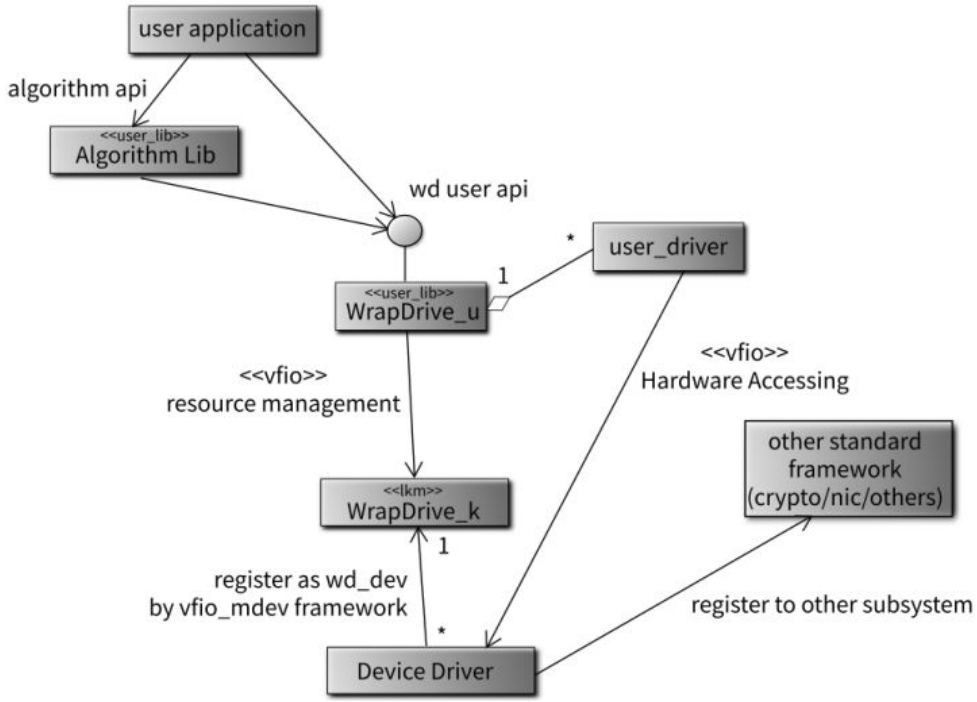
(A Common Accelerator Framework for User Space)

Zaibo Xu from Huawei

Contents

- Looking back
- Progress and Example
- Current Challenges
- Plans

Looking back on Warpdrive
---on SFO 2017 & HK 2018



What is Warpdrive (WD)?

1. An accelerator framework for user space, leveraging hardware accelerators with native performance.
2. Based on VFIO and MDEV, allowing for direct access hardware with improved security and efficiency.
3. Hardware accelerator is accessed via a ‘queue’, the minimal working unit for user. The queue’s DMA priority is controlled by the user.
4. Support SVA, being compatible with Jean-Philippe Brucker (JPB)’s SVA patch set.
5. Dynamic ‘queue’ allocation mechanism with DMA access supported with multiple processes.
6. Compatibility (native driver, normal Mdev, NO-IOMMU, Other subsystem)

Notes: ‘Wrapdrive’ now is renamed as ‘Warpdrive’, and kernel part of ‘Wrapdrive’ is called as ‘Share Domain Mediated Device’, SDMDEV for short.

Looking back on Warpdrive
--- on SFO 2017 & HK 2018

Why Warpdrive?

- Try to get the native performance of accelerator.
- Break the limit of one device serving only one process with ‘queue’.
- User space DMA should be controlled under security.

Status back then

- Shared virtual Address (SVA) from JPB was in the phase of V1.
- Substream-ID was still not supported by SMMU driver .
- Warpdrive supported multiple processes with SVA enabling, SVA without I/O page fault running OK on Hisilicon D06 board.

Warpdrive Progress

- WD RFC v2 patch set has been released.
- Updating on WD.
- Enhancements of JPB's SVA proposal.
- Example: WD SVA on Hisilicon D06 board.

Warpdrive Progress

WD RFC patch set has been released.

The following discussions are ongoing:

1. Whether WD should go around VFIO, and create CDEV and DMA buffer pool for each accelerator while driver probing.(Example: KFD solution from AMD for GPU)
2. Whether IOMMU should be aware of Mdev or need auxiliary domain [1]. (Analogous RFC solution on VFIO operating Mdev, related to the intel 'scalable mode' virtualization)
3. Whether resources should be pre-assigned at Mdev being created, analogizing to original VM scenario of Mdev.
4. Whether an accelerator device should be shared between user and kernel land.

[1]https://patchwork.kernel.org/cover/10539191/; https://lkml.org/lkml/2018/8/30/118

Updating on WD

1. MDEV:
   a) Created by 'root' before Warpdrive processes start.
   b) Destroyed by the 'root' after Warpdriver processes end.

2. 'Queues' held by a process are released automatically if the process exists unexpectedly.

3. PASID setting logic is moved from user space to the VFIO core.

4. Resources are pre-assigned at MDEV creation.

Warpdrive Progress

Enhancements of JPB's SVA proposal

1. PASID is not exposed to user space.

2. VFIO ATTACH and DETACTH commands operate on the IOMMU group of the MDEV's parent device.

3. Enable user space DMA mapping from multiple processes by creating private I/O page tables.

4. Break the limitation of SVA's using scenario of primary and secondary processes.

5. Keep all the UAPIs of VFIO unchanged while enabling SVA.

Warpdrive Progress

---A relative topic

Why JPB's SVA using scenario is not general?

Overall, as to be compatible with VFIO-PCI[1]:

1. Need expose PASID to user space.
2. Layout user processes as primary and secondary processes as DPDK.
3. BIND/UNBIND IOCTL CMDs are not general APIs of VFIO

[1] https://patchwork.kernel.org/patch/10394927/

Warpdrive Progress
---example

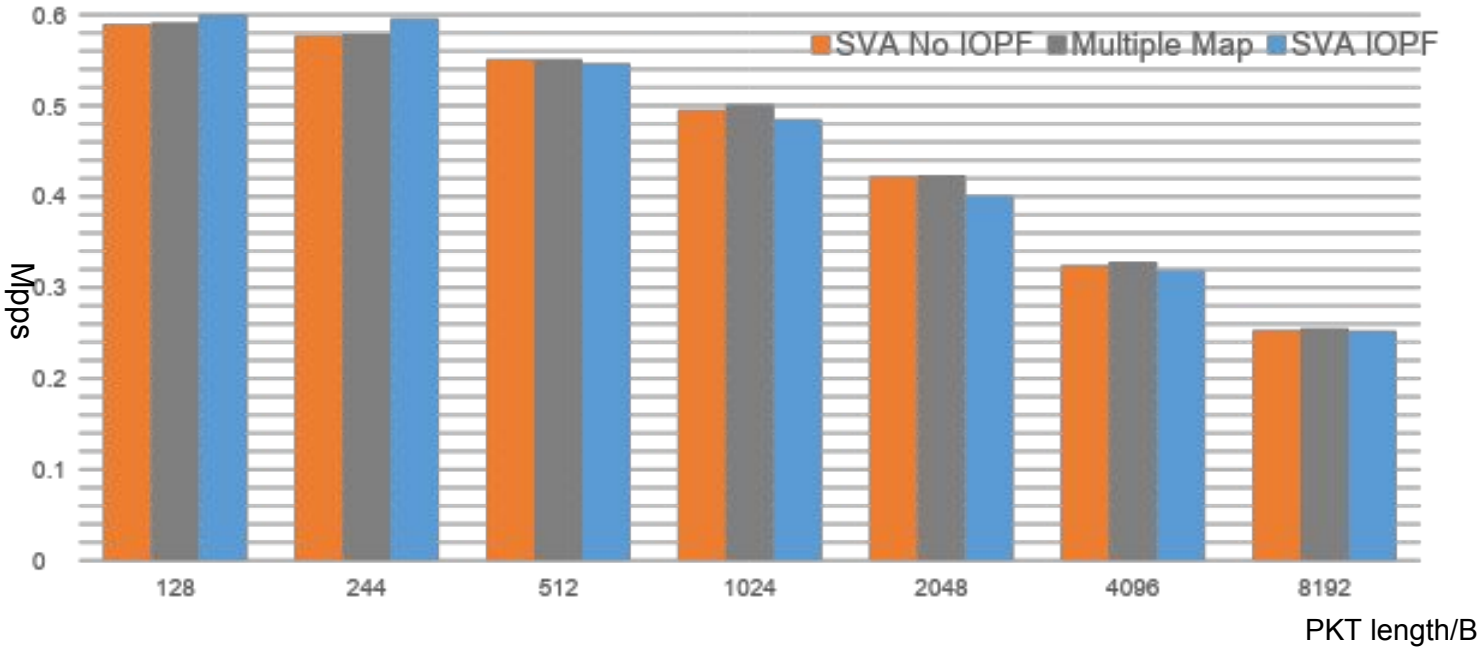
WD SVA stall mode on Hisilicon D06 board.

1. Based on JPB's SVA V2 patch set.
2. Using pseudo PCIe device: ZIP accelerator.
3. Platform device stall mode is running on ZIP.
4. Multiple processes of SVA scenario is OK.
5. ZIP PCIe device should be distorted as supporting 'STALL' in IOMMU config.

Question: Should the 'STALL' mode be supported as a common mode for pseudo PCIe and platform devices?

Warpdrive Progress

---example

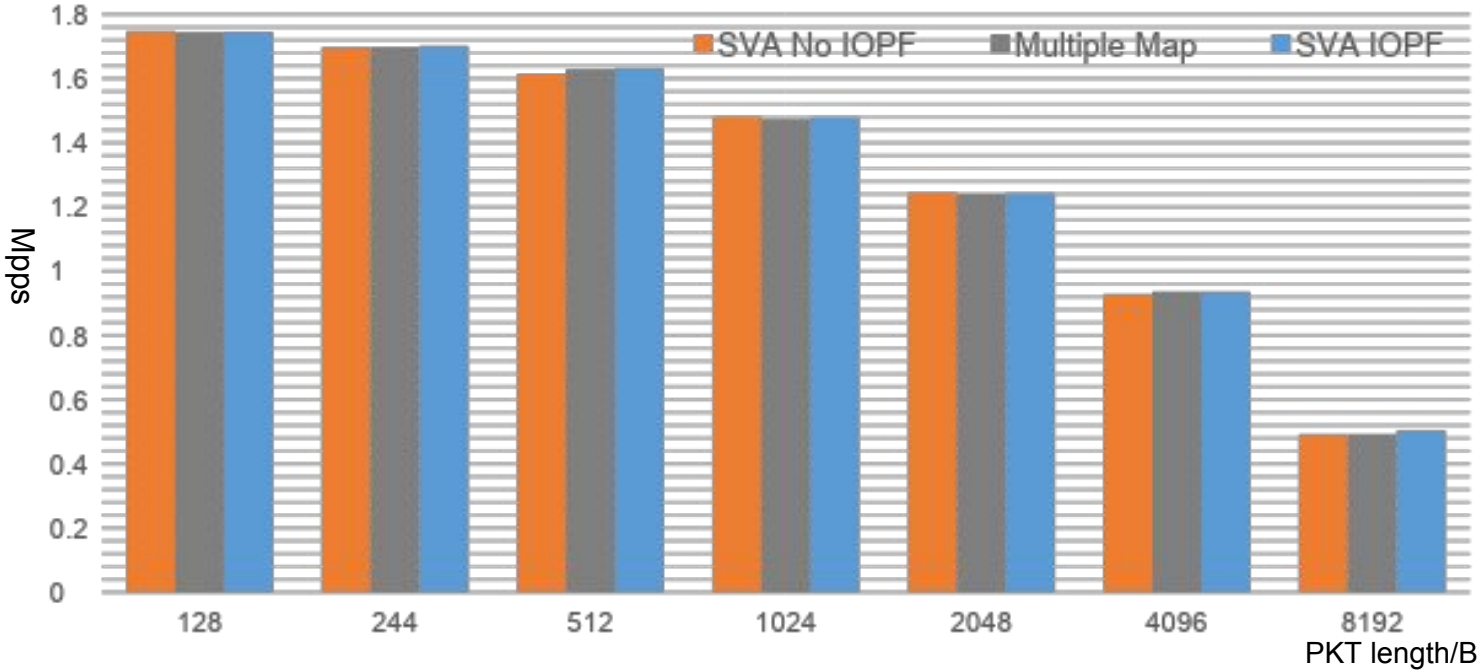


Throughput of ZIP accelerator on Hisilicon D06 board (1 processes & 1 queues)

Notes: 3 scenarios are tested, each with 1 process being run synchronous mode at the same time.

Warpdrive Progress

---example



Throughput of ZIP accelerator on Hisilicon D06 board (3 processes & 3 queues)

Notes: 3 scenarios are tested, each with 3 processes being run synchronous mode at the same time.

Current Challenges

1. Still existing security risk since several users work on one device.

- ‘Queue’ is not isolated as much as a task in the OS.

2. Coexist with Linux kernel Crypto / AF-ALG (controversial topic)

- Ecosystem is ready because devices support PASID/PRI/ATS/ATC and corresponding software such as VFIO and SVA is also in place.
- User space should be able to leverage accelerators directly with high performance now that devices can DMA from user space in a fast and secure way.

3. IOMMU/VFIO software status

- No fixed solution on PASID & IOMMU domain .

Plans

1. RFC v3 will be released shortly taking into account Intel's work on IOMMU aware MDevs.
2. A general solution for using SVA based on VFIO/Warpdrive will be brought up.
3. More work will be done on user space Warpdrive.

Thank You!

And question?

xuzaibo@huawei.com