openEuler

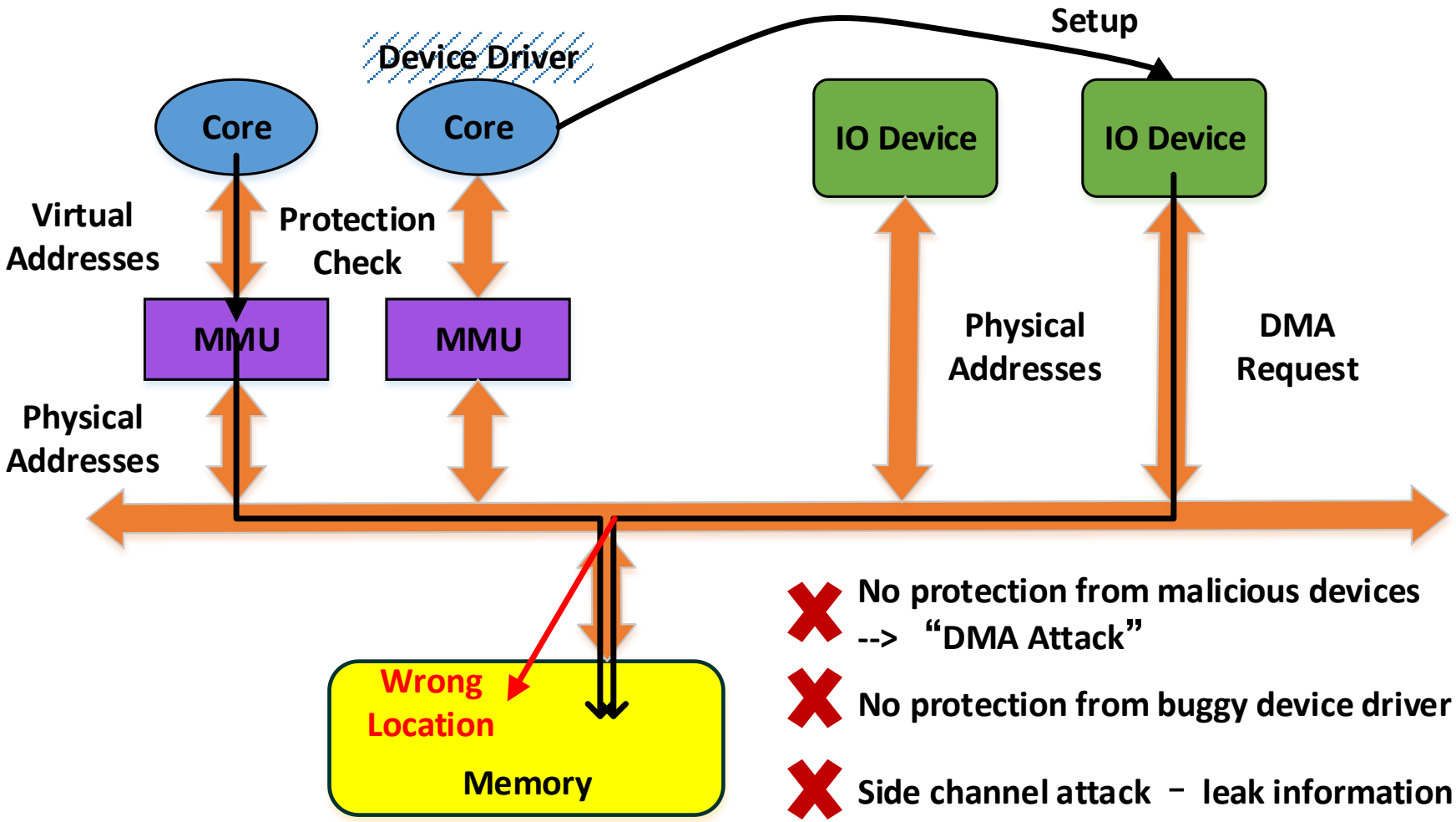
SVA(Shared Virtual Memory)机制介绍

目录

CONTENT

openEuler

IOMMU简介-TRADITIONAL DMA



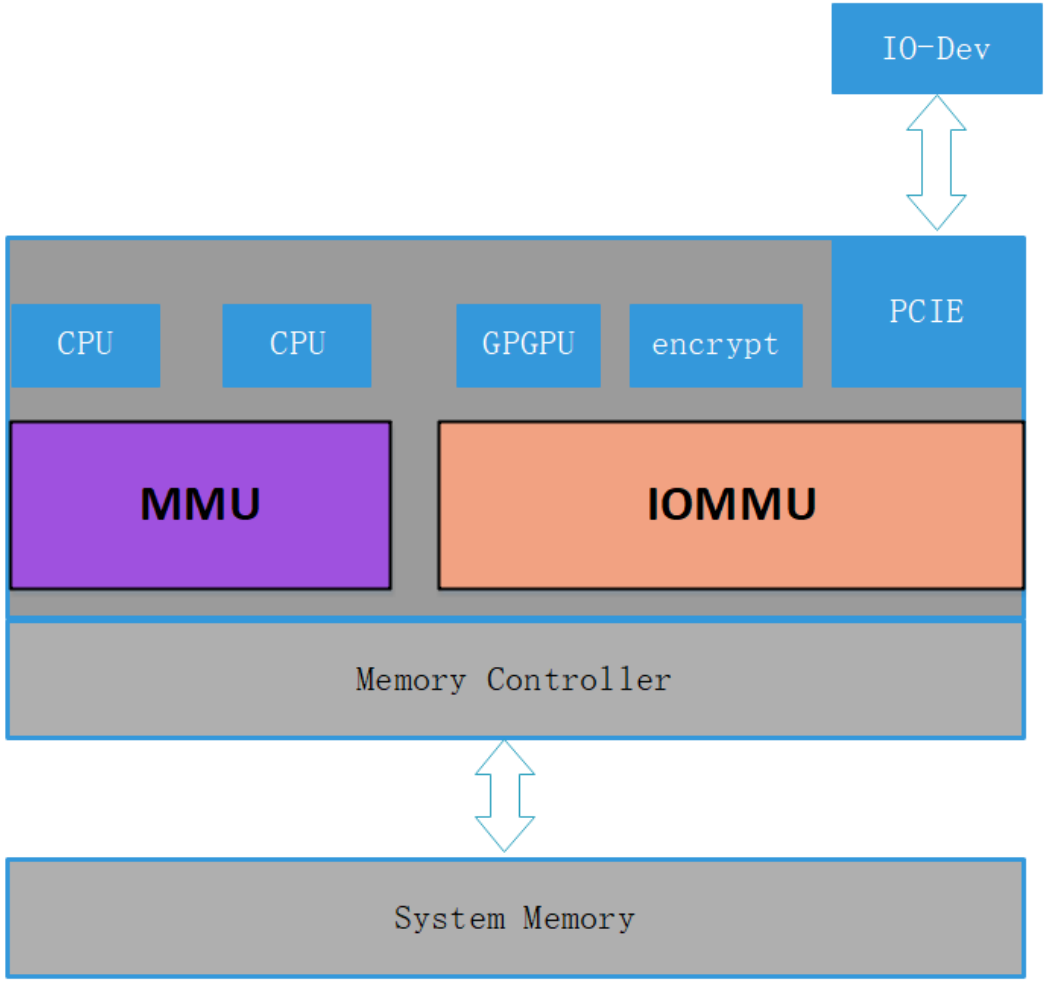
图片来自： IOMMU_TUTORIAL_ASPLOS_2016.pdf

openEuler

IOMMU简介-新型的IO Device

随着计算量的暴增，使用专用/通用的加速设备辅助计算。

general-purpose computation on a GPU (GPGPU), encryption accelerators, digital signal processors, etc.



openEuler

IOMMU介绍-ARM的SMMU

A System Memory Management Unit (SMMU) performs a task that is analogous to that of an MMU in a PE, translating addresses for DMA requests from system I/O devices before the requests are passed into the system interconnect. It is active for DMA only. Traffic in the other direction, from the system or PE to the device, is managed by other means – for example, the PE MMUs.

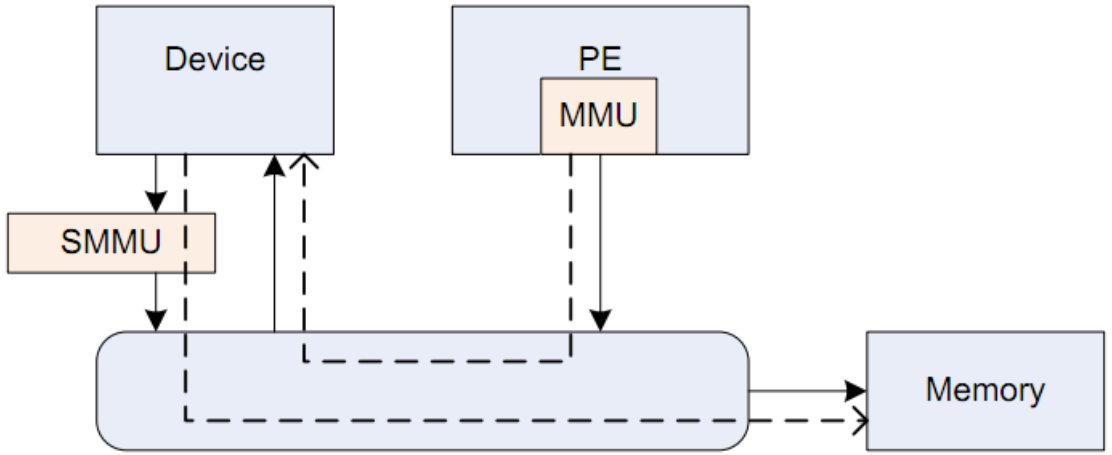
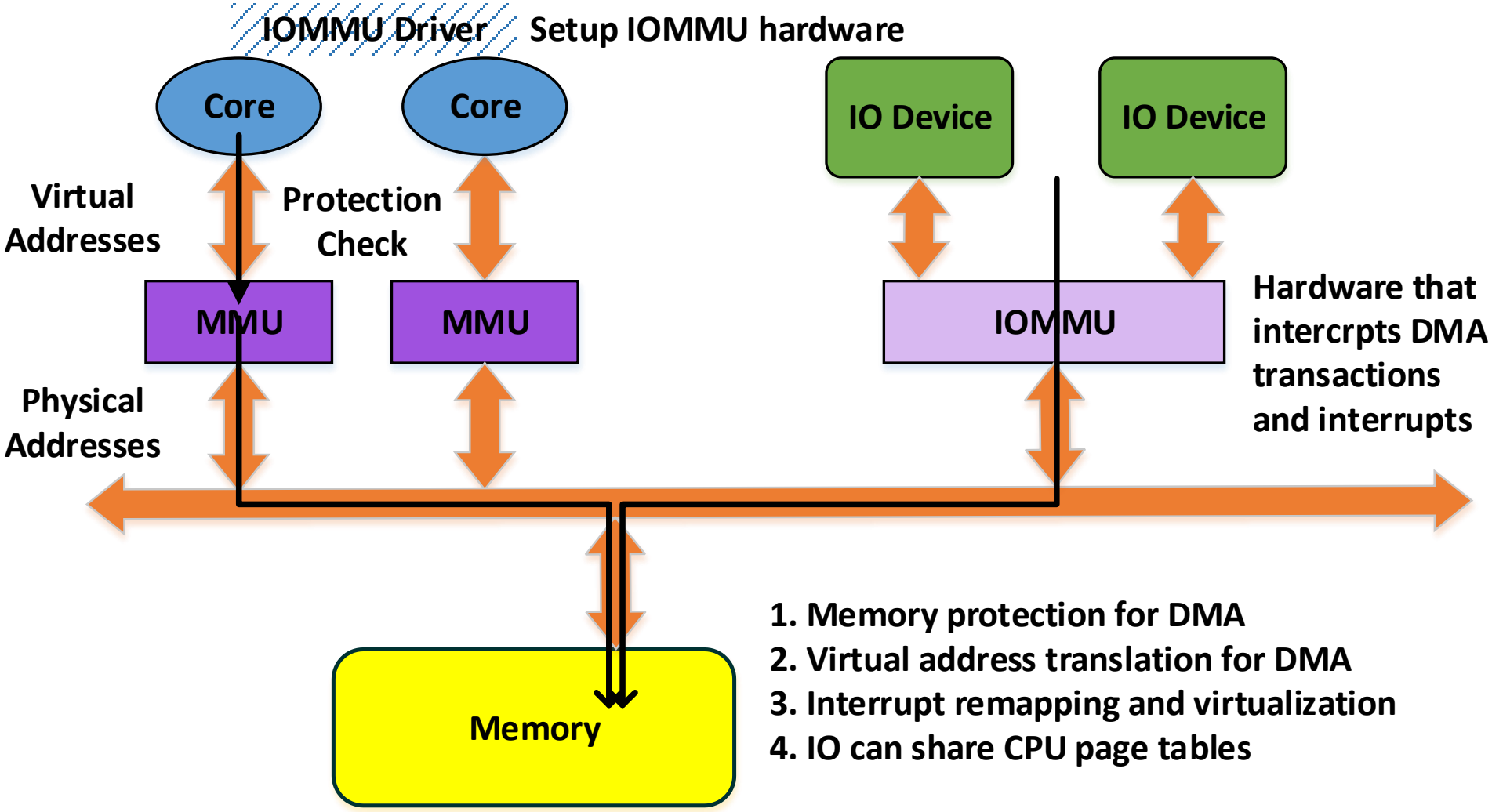


Figure 1: System MMU in DMA traffic

图片来自： ARM SMMUv3 Spec

openEuler

IOMMU简介-使能IOMMU



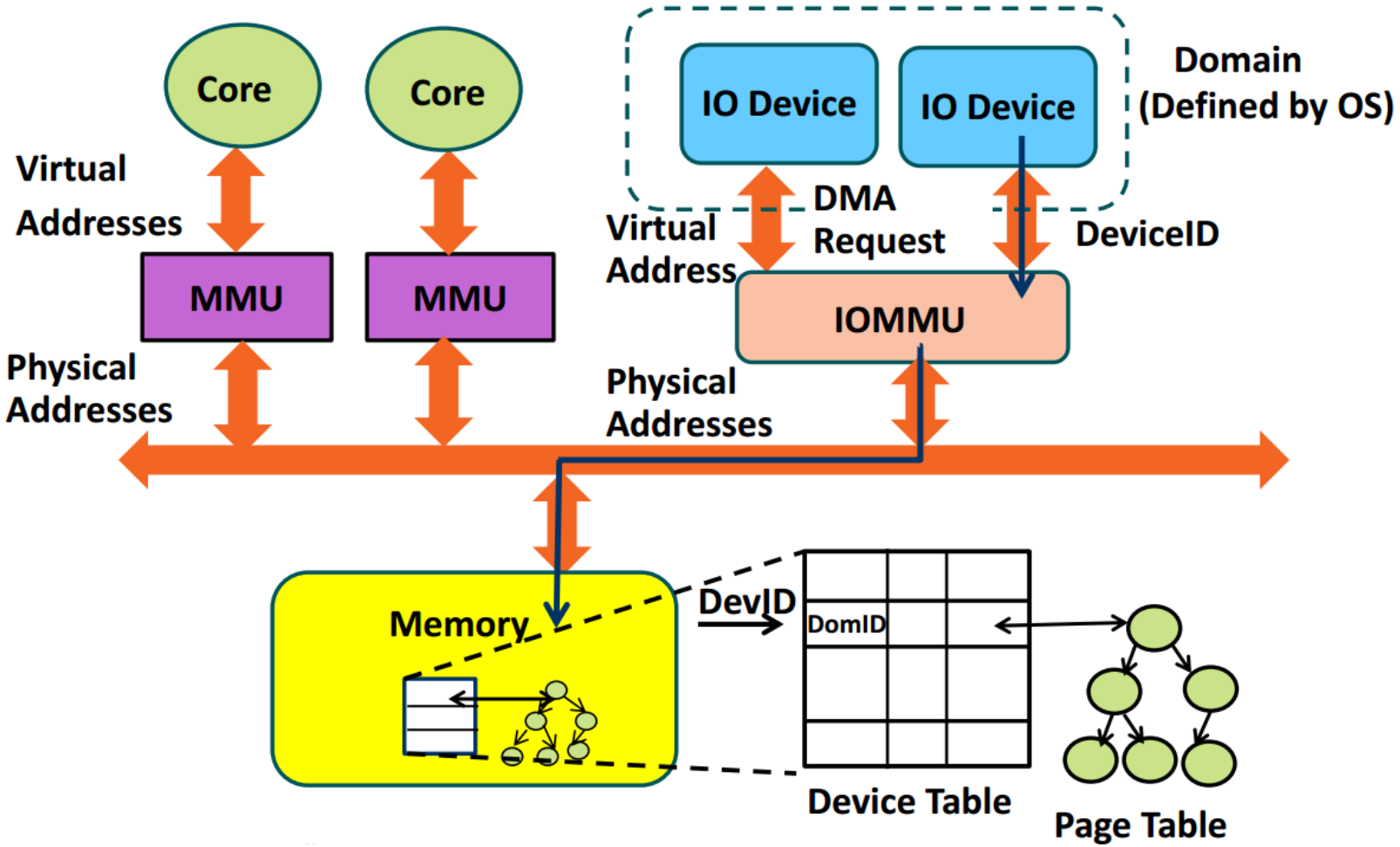
图片来自： IOMMU_TUTORIAL_ASPLOS_2016.pdf

openEuler

IOMMU简介-DMA Mapping

1、IOMMU使用独立的页表，为设备提供地址虚拟化能力；

2、不同的页表可以为不同的设备提供隔离的地址空间



图片来自： IOMMU_TUTORIAL_ASPLOS_2016.pdf

openEuler

IOMMU简介-DMA Mapping编程示例

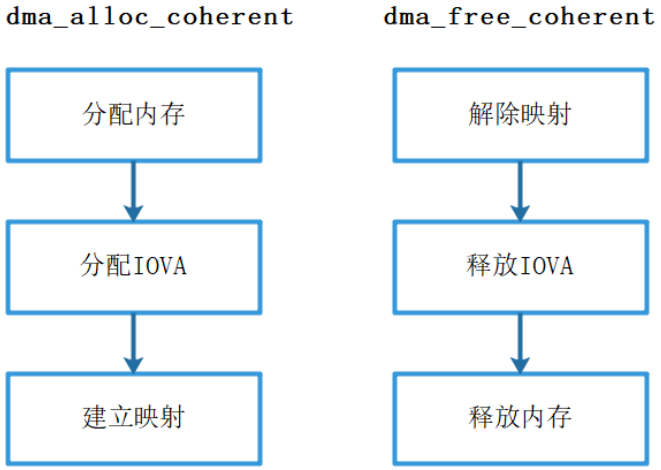
1、dma_alloc_coherent(dev, size)

- 申请内存、申请IO虚拟地址空间，建立映射；
- 也有独立的接口可以分开调用，更灵活

2、设备使用IOVA发起DMA

3、dma_free_coherent(dev, iova, size)

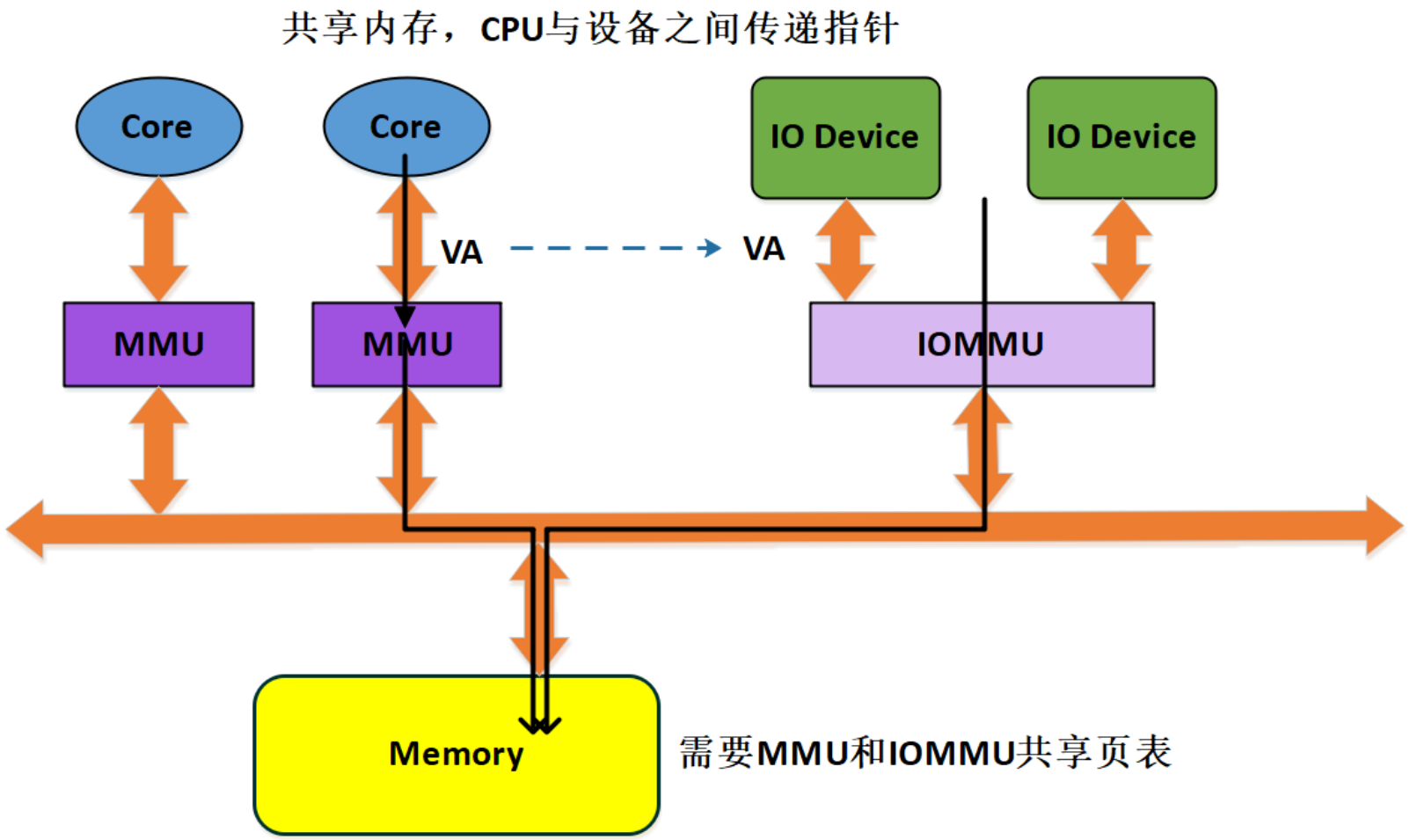
- 接触映射，释放内存以及虚拟地址范围；




```
static const struct dma_map_ops iommu_dma_ops = {
        .alloc = __iommu_alloc_attrs,
        .free = __iommu_free_attrs,
        .mmap = __iommu_mmap_attrs,
        .get_sgtable = __iommu_get_sgtable,
        .map_page = __iommu_map_page,
        .unmap_page = __iommu_unmap_page,
        .map_sg = __iommu_map_sg_attrs,
        .unmap_sg = __iommu_unmap_sg_attrs,
        .sync_single_for_cpu = __iommu_sync_single_for_cpu,
        .sync_single_for_device = __iommu_sync_single_for_device,
        .sync_sg_for_cpu = __iommu_sync_sg_for_cpu,
        .sync_sg_for_device = __iommu_sync_sg_for_device,
        .map_resource = iommu_dma_map_resource,
        .unmap_resource = iommu_dma_unmap_resource,
        .mapping_error = iommu_dma_mapping_error,
};
```

openEuler

SVA简介-简化编程模型



图片来自： IOMMU_TUTORIAL_ASPLOS_2016.pdf

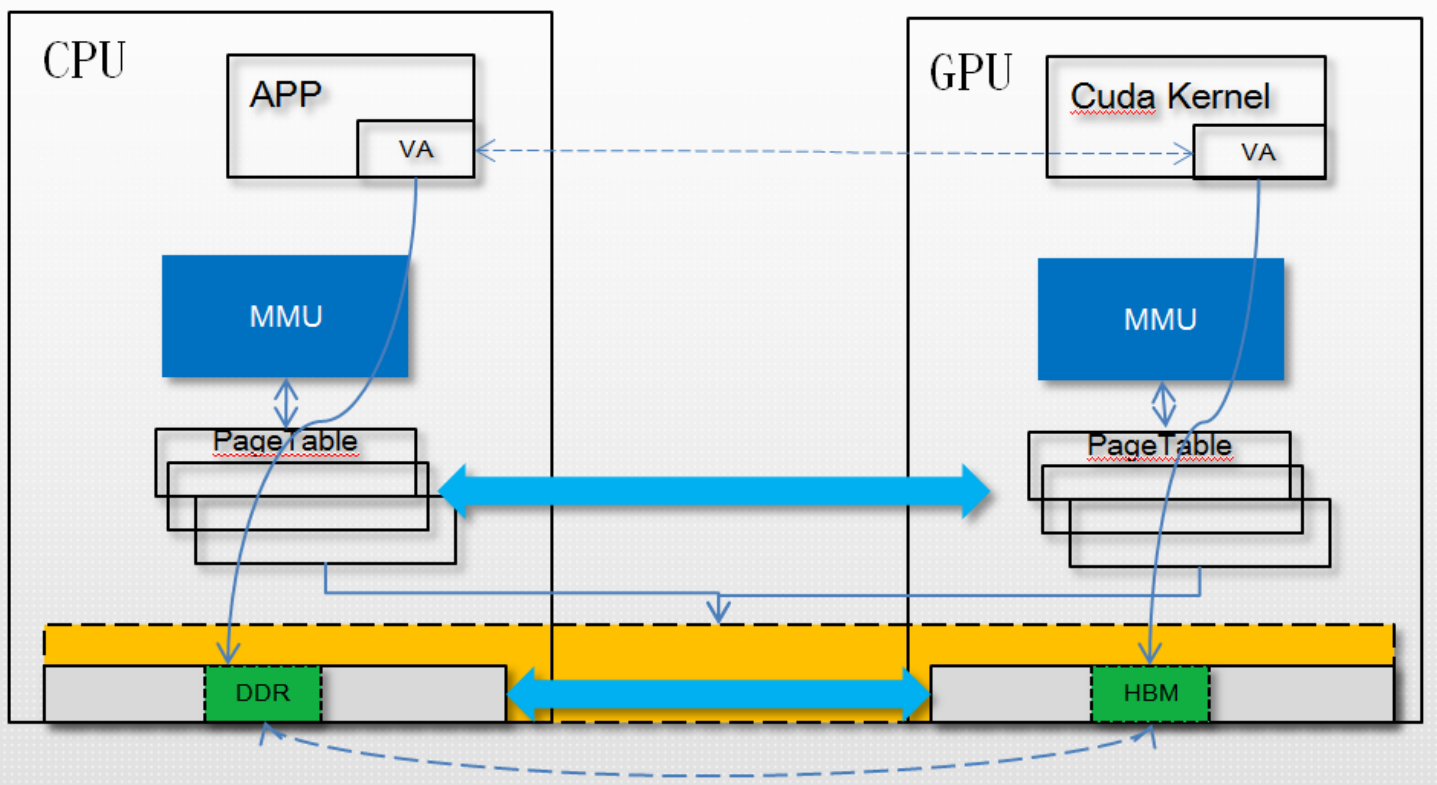
openEuler

SVA简介-GPU加速卡

1、GPU独立的内存资源；

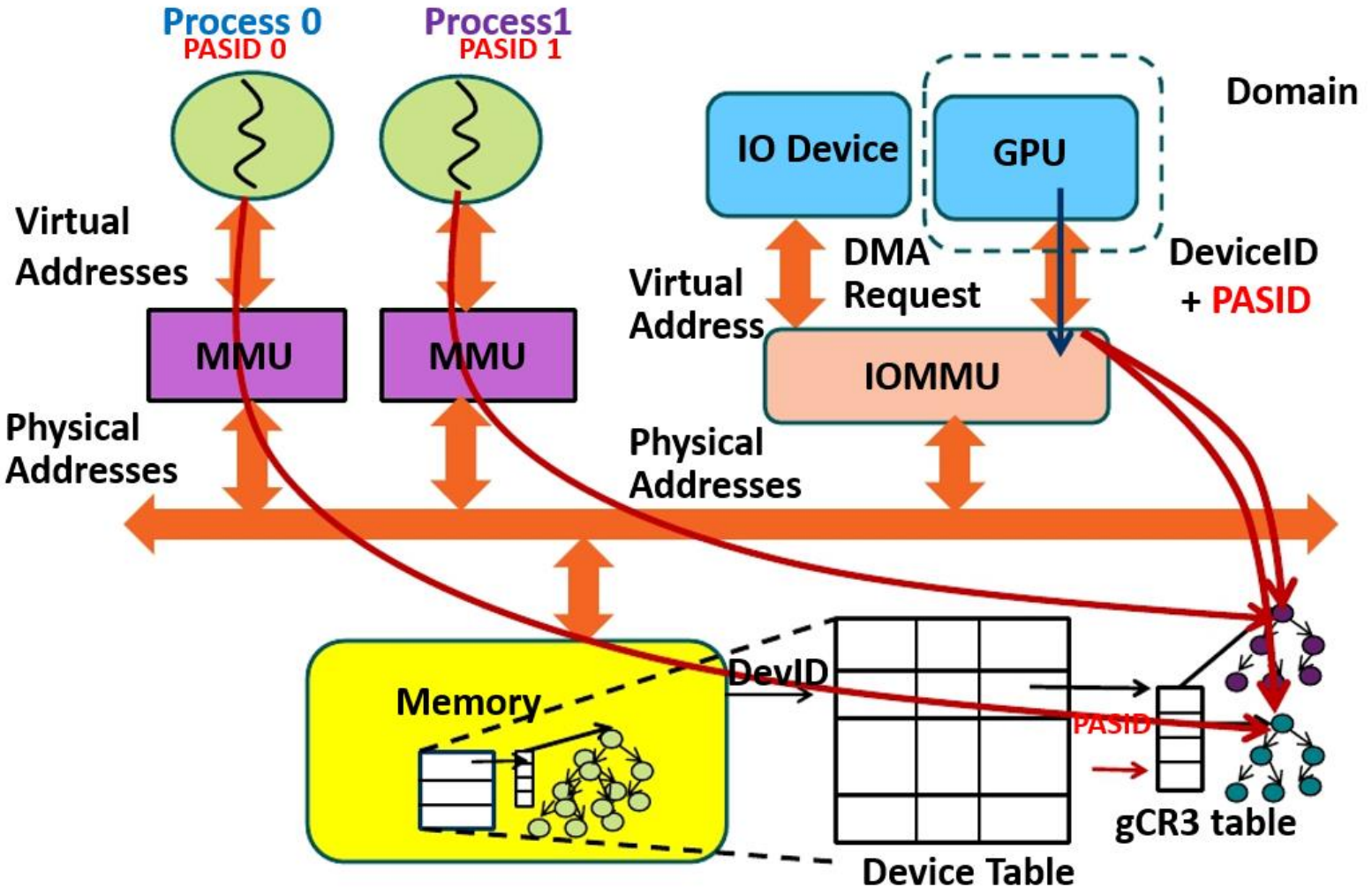
2、CPU与GPU需要完成页表同步以及数据拷贝。

需要大量的数据拷贝以及独立的资源可能资源利用率并不高



openEuler

SVA介绍-SVM/SVA-Shared Vitual Memory/Address



图片来自： IOMMU_TUTORIAL_ASPLOS_2016.pdf

openEuler

SVA简介-简化编程模型

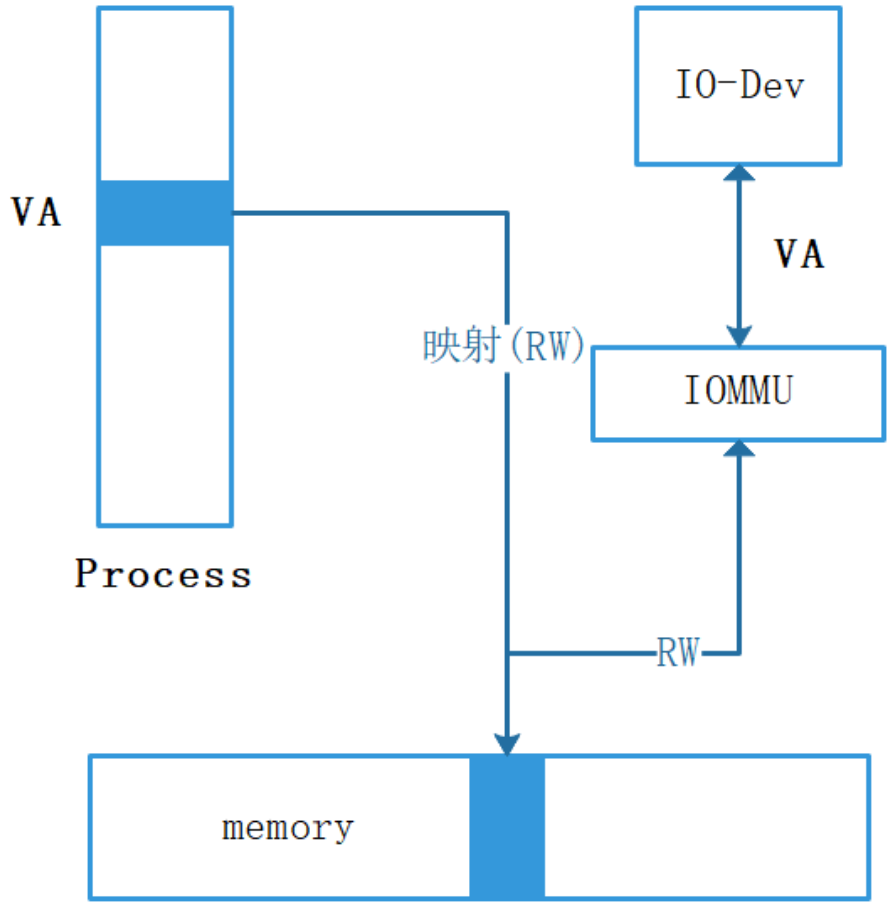
1、malloc/mmap(size)

➢ 申请内存，建立映射

2、设备使用VA发起DMA

3、free/munmap(va)

➢ 解除映射，释放内存

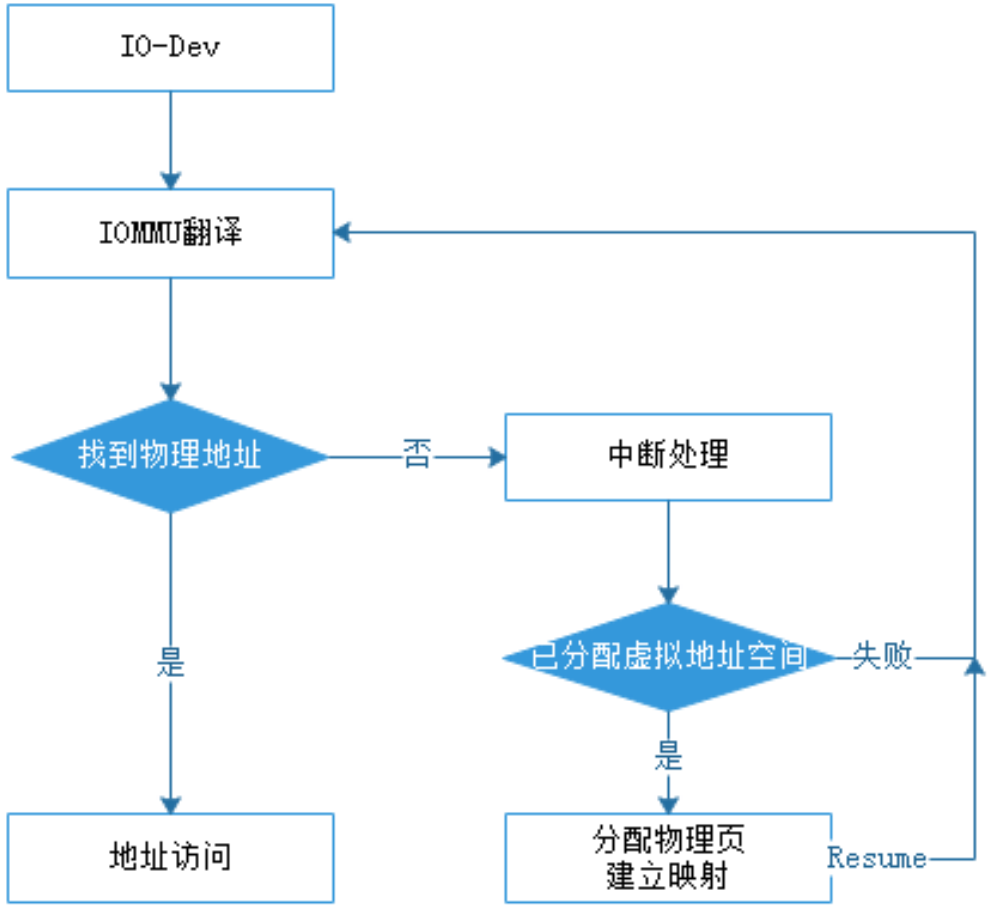


openEuler

SVA简介-缺页处理

1、IOMMU在Pagetable walk时，找不到VA对应的页表项，上报中断；

2、中断处理中，判断VA地址是否为已分配，未分配，直接返回失败；已分配但没有对应物理页，则分配相应的物理页，建立映射；并通知IOMMU retry。



openEuler

SVA简介-pin内存

包含两种意思：

1、避免被swap出去；

2、避免页面被迁移；

mlock可以防止页面被swap；

get_user_page防止页面被swap以及迁移；

1、malloc/mmap(size)

- 申请内存，建立映射，pin住内存

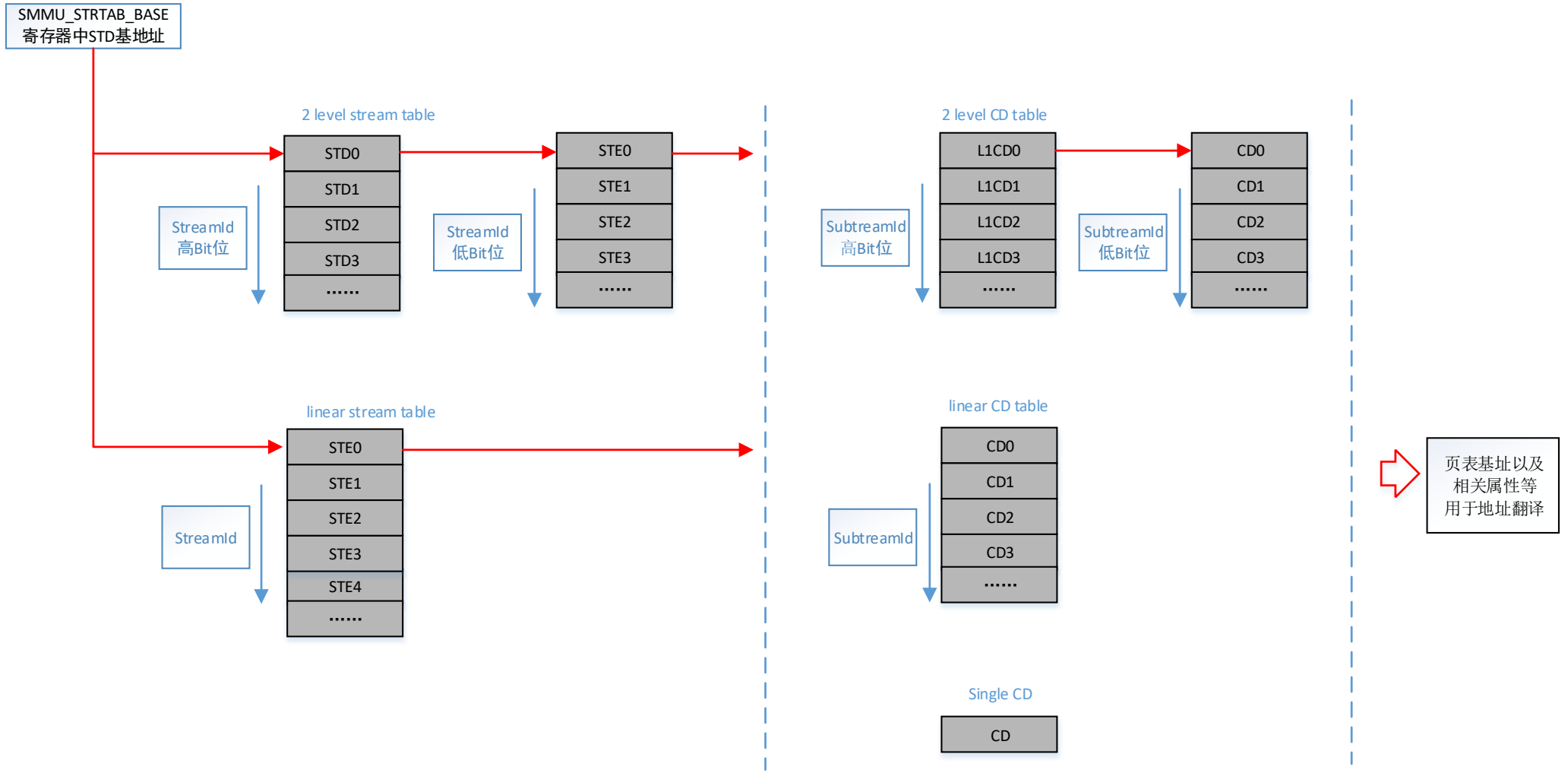
2、设备使用VA发起DMA

3、free/munmap(va)

- Unpin内存，解除映射，释放内存

openEuler

SMMUv3示例-实现



StreamId = DevID, SubStreamId = PASID

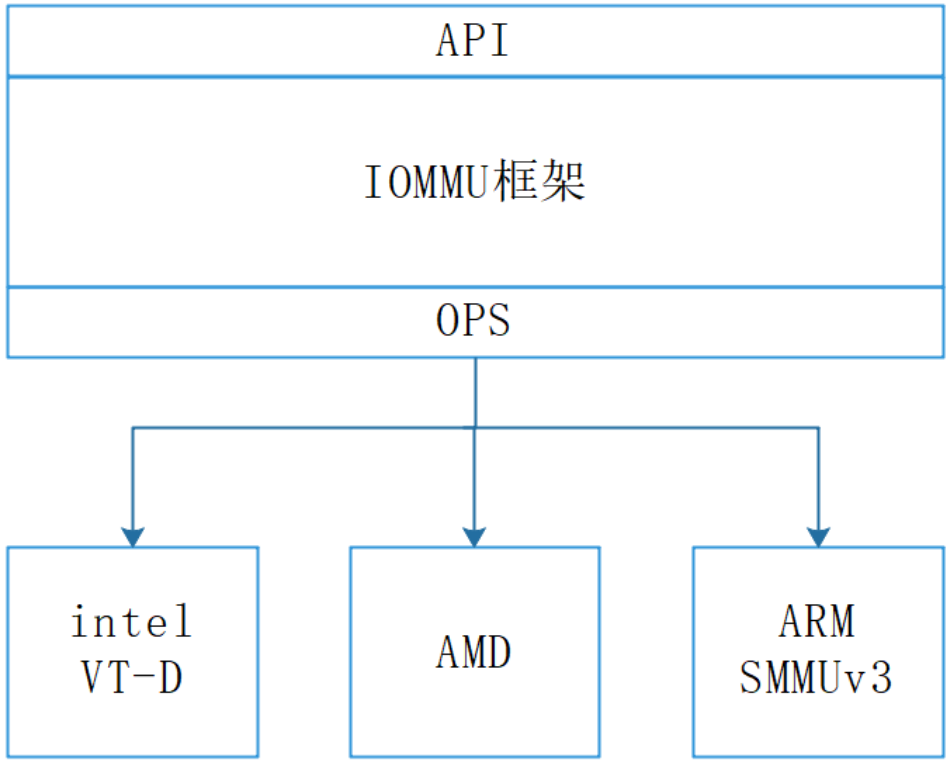
openEuler

SMMUv3示例-IOMMU框架

1、IOMMU框架提供了一套API接口，比如：

2、IOMM框架屏蔽了底层具体的IOMMU设备的差异，底层只需要完成具体的OPS实现即可。

3、同时也实现了DMA设备与IOMMU设备probe时的自动建立关联等。



openEuler

SMMUv3示例-共享页表

将进程的地址空间mm bind到dma设备dev，返回一个分配的全局唯一的pasid

```
2372: /**
2373:  * iommu_sva_bind_device() - Bind a process address space to a device
2374:  * @dev: the device
2375:  * @mm: the mm to bind, caller must hold a reference to it
2376:  * @pasid: valid address where the PASID will be stored
2377:  * @flags: bond properties
2378:  * @drvdata: private data passed to the mm exit handler
2379:  *
2380:  * Create a bond between device and task, allowing the device to access the mm
2381:  * using the returned PASID. If unbind() isn't called first, a subsequent bind()
2382:  * for the same device and mm fails with -EEXIST.
2383:  *
2384:  * iommu_sva_device_init() must be called first, to initialize the required SVA
2385:  * features. @flags is a subset of these features.
2386:  *
2387:  * If IOMMU_SVA_FEAT_IOPF isn't requested, the caller must pin down using
2388:  * get_user_pages*() all mappings shared with the device. mlock() isn't
2389:  * sufficient, as it doesn't prevent minor page faults (e.g. copy-on-write).
2390:  *
2391:  * On success, 0 is returned and @pasid contains a valid ID. Otherwise, an error
2392:  * is returned.
2393:  */
2394: int iommu_sva_bind_device(struct device *dev, struct mm_struct *mm, int *pasid,
2395: »       »       »       unsigned long flags, void *drvdata)
2396: {
```

openEuler

SMMUv3示例-设备节点

SMMU节点

```
238 >-------smmu8:smmu_fte@102000000 {
239 >-------->-------compatible = "arm,smmu-v3";
240
241 >-------->-------#address-cells = <2>;
242 >-------->-------#size-cells = <2>;
243 >-------->-------index = <8>;
244 >-------->-------reg = <0x1 0x02000000 0x0 0x20000>;
245 >-------->-------alignment = <0x10000>;
246
247 >-------->-------interrupt-parent = <&gic>;
248
249 >-------->-------interrupts = <67 0x02000002 0x00000001>,
250 >-------->------->-------     <68 0x02000002 0x00000001>,
251 >-------->------->-------     <69 0x02000002 0x00000001>;
252 >-------->-------#iommu-cells = <0x1>;
253 >-------->-------dma-coherent;
254 >-------->-------hisilicon,broken-prefetch-cmd;
255 >-------->-------hisilicon,broken-spi;
256 >-------->-------#iommu-spi-base = <0x1 0x9000040>;
257 >-------};
```

Device节点

```
363 >-------->-------peri_fte: peri_fte@102040000 {
364 >-------->------->-------compatible = "hisilicon, fte_smmu_drv";
365 >-------->------->-------iommus = <&smmu8 0x7f8a>;
366 >-------->------->-------reg = <0x01 0x02040000 0x0 0x10000>,
367 >-------->------->-------      <0x01 0x0c000000 0x0 0x10000>;
368 >-------->------->-------alignment = <0x1000>;
369 >-------->------->-------smmu_id = <8>;
370 >-------->-------};
```

1、SMMU节点中配置了当前设备的配置空间、中断等等信息；

2、后端的DMA设备节点中配置了其使用的哪个SMMU设备以及对应的sid(该sid是与master设备驱动约定的，大家都使用这个sid才能匹配上)

openEuler

参考资料

1、IOMMU_TUTORIAL_ASPLOS_2016.pdf

2、IHI_0070_D_a_System_Memory_Management_Unit_Arm_Architecture_Specification.pdf

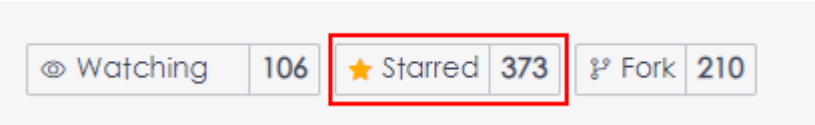
openEuler

✓ openEuler kernel gitee 仓库

源代码仓库

https://gitee.com/openeuler/kernel

欢迎大家多多 Star，多多参与社区开发，多多贡献补丁。



✓ maillist、issue、bugzilla

可以通过邮件列表、issue、bugzilla 参与社区讨论

欢迎大家多多讨论问题，发现问题多提 issue、bugzilla

https://gitee.com/openeuler/kernel/issues

https://bugzilla.openeuler.org

kernel@openeuler.org

✓ openEuler kernel SIG 微信技术交流群

请扫描右方二维码添加小助手微信

或者直接添加小助手微信（微信号：openeuler-kernel）

备注“交流群”或“技术交流”

加入 openEuler kernel SIG 技术交流群

技术交流

openEuler

Thank you

openEuler