Виртуализация подсистемы ввода-вывода микропроцессоров архитектуры Эльбрус

Никита Поляков

12 марта 2020

Аппаратная поддержка виртуализации в Эльбрус v6

эльбрус

- Вычислительные ресурсы
- Оперативная память
- **Периферийные устройства (устройства ввода-вывода)**
- ***Прерывания (внешние и*** межпроцессорные***)***

Способы виртуализация периферийных устройств

эльбрус

- **эмуляция** – гипервизор предоставляет гостевой ОС *виртуальное* устройство
- **проброс устройства (*assignment*)** – гостевой ОС предоставляется *физическое* периферийное устройство, которое управляется напрямую драйвером гостевой ОС с минимальным вмешательством гипервизора
- ***I/O device sharing*** – расширение проброса устройства для устройств, поддерживающих несколько функциональных интерфейсов (PCI function)

Аппаратная поддержка проброса устройства

- **доступ к регистрам** периферийного устройства
- **DMA remapping –** трансляция адресов DMA-обращений
- **Interrupt remapping** - изоляция и маршрутизация внешних прерываний от устройств гостя
- **Interrupt posting** - прямая доставки прерываний виртуальному процессору
- **Reliability –** журналирование и оповещение гипервизора об ошибках трансляции DMA или прерываний

- ➢ перехват гипервизором
- ➢ **аппаратная** трансляция в **IOMMU**
- ➢ **EPIC + IOMMU**
- ➢ **EPIC**
- ➢ **IOMMU** (только оповещение)

Использование DMA remapping в ОС

- защита ОС
- поддержка legacy устройств (32-битные адреса)
- изоляция устройств (домены)
- использование общего виртуального пространства несколькими устройствами

IOMMU

I/O Memory Management Unit – устройство трансляции адресов DMA

- Режимы работы
  - нативный режим: VA -> PA
  - режим виртуализации – двухуровневая трансляция: GVA -> GPA -> PA
- встроенные кэши трансляций
- размеры страниц: 4КБ, 2МБ и 1Гб
- структуры трансляции древовидные 4-уровневые
- конвейер из 3-х стадий, максимальные темп – 1 адрес за такт

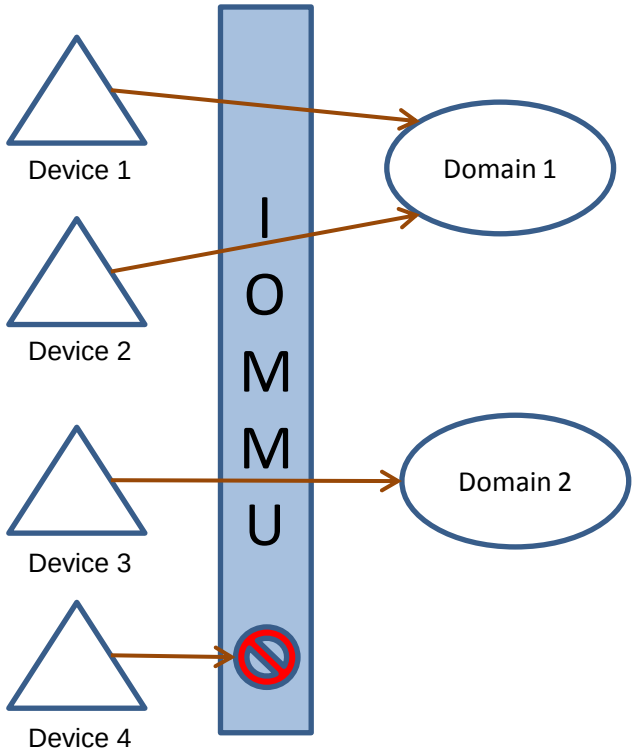
IOMMU

Нативный режим

В данном режиме выполняется одноуровневая трансляция VA -> PA с возможностью разделения устройств на домены

Варианты работы:

1) общие структуры трансляции всех устройств
2) разделение устройств на домены:
   - работа по виртуальному адресу
   - работа по физическому адресу
   - устройству запрещен доступ по DMA



Домены трансляции адреса

- ***Домен*** – изолированная среда, к которой привязана некоторая область физической памяти.
- Каждое внешнее устройство, работающее по DMA, привязывается к одному из доменов.
- Изоляция домена достигается за счет того, что все устройства домена имеют доступ только к той области физической памяти, которая привязана к домену.
- Информация о трансляции адреса для каждого устройства хранится в **Таблице устройств** (Device Table, DT)
- В нативном режиме поддерживается до 4096 доменов

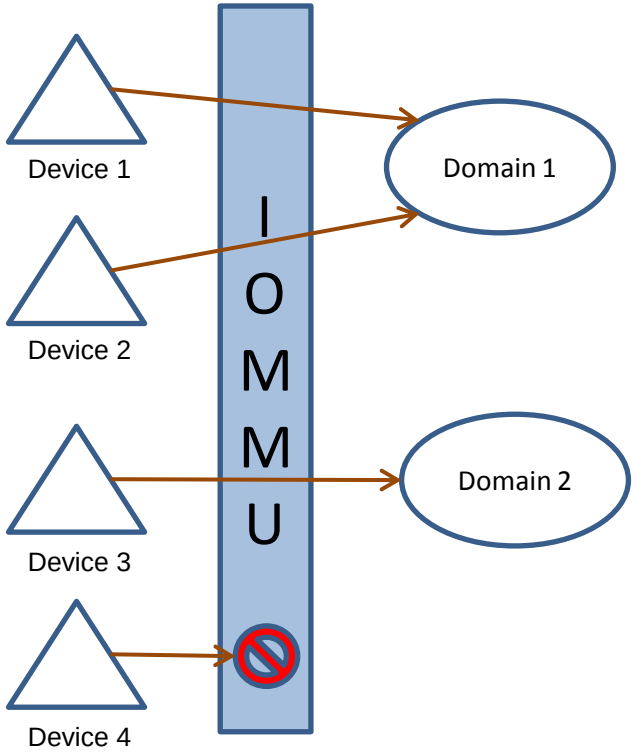
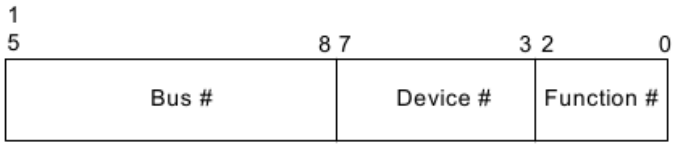


Таблица устройств (Device Table)

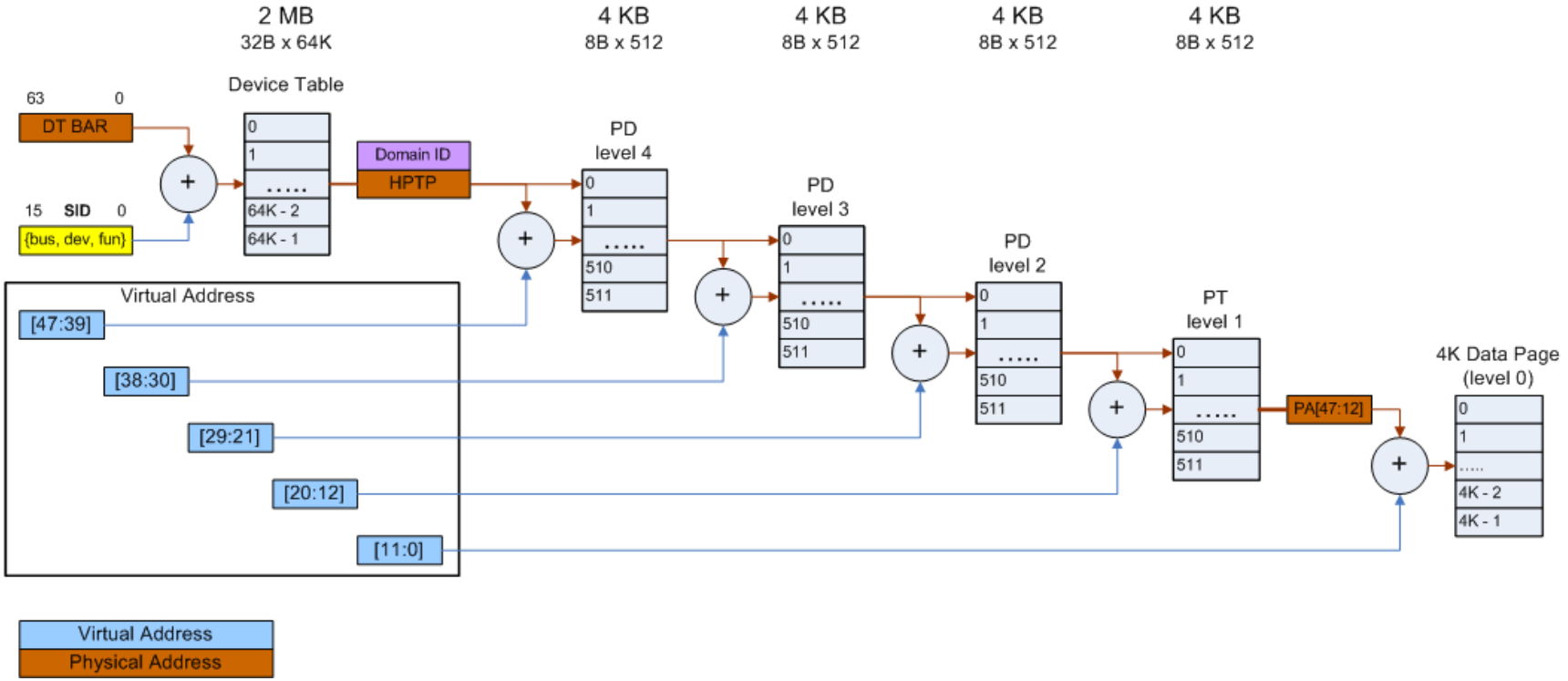
- Таблица устройств состоит из 32-байтовых элементов - Device Table Entry (DTE)
- Таблица индексируется Source ID (SID)



- DTE содержит в числе прочего
  - Domain ID – идентификатор домена
  - адрес (указатель) корня структур трансляции
  - размер виртуального пространства
- Размер таблицы – 2 МБ
- Указатель на таблицу хранится в регистре DT BAR

IOMMU

Нативный режим. Структуры трансляции



IOMMU

Режим виртуализации

В данном режиме выполняется

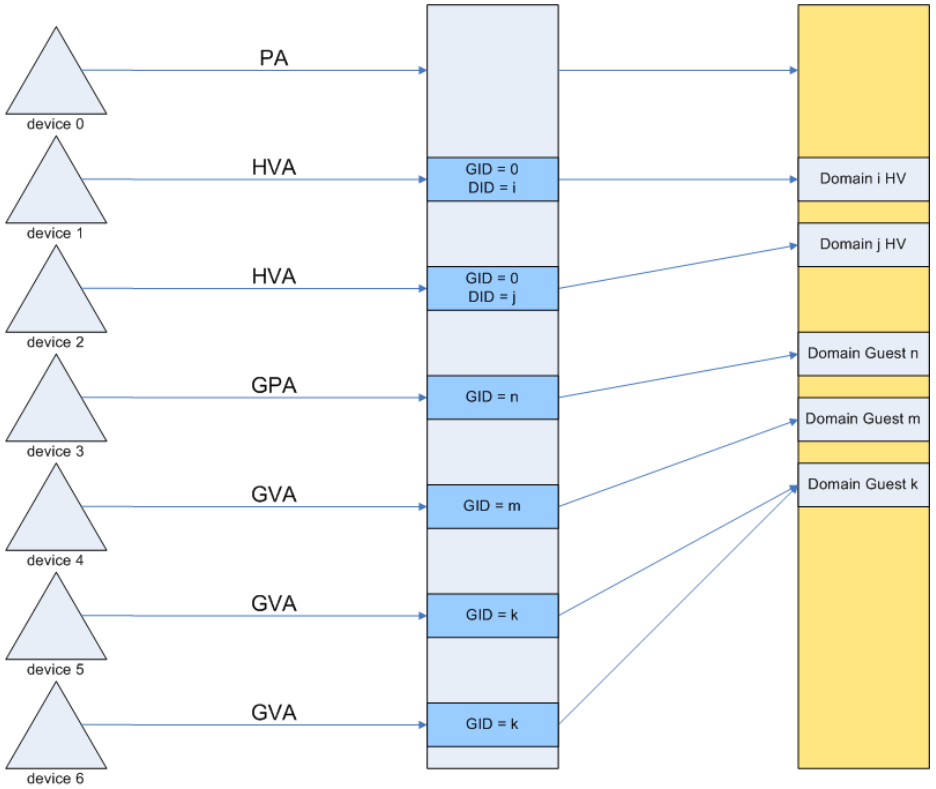
двухуровневая трансляция

GVA -> GPA -> PA

Варианты работы устройства

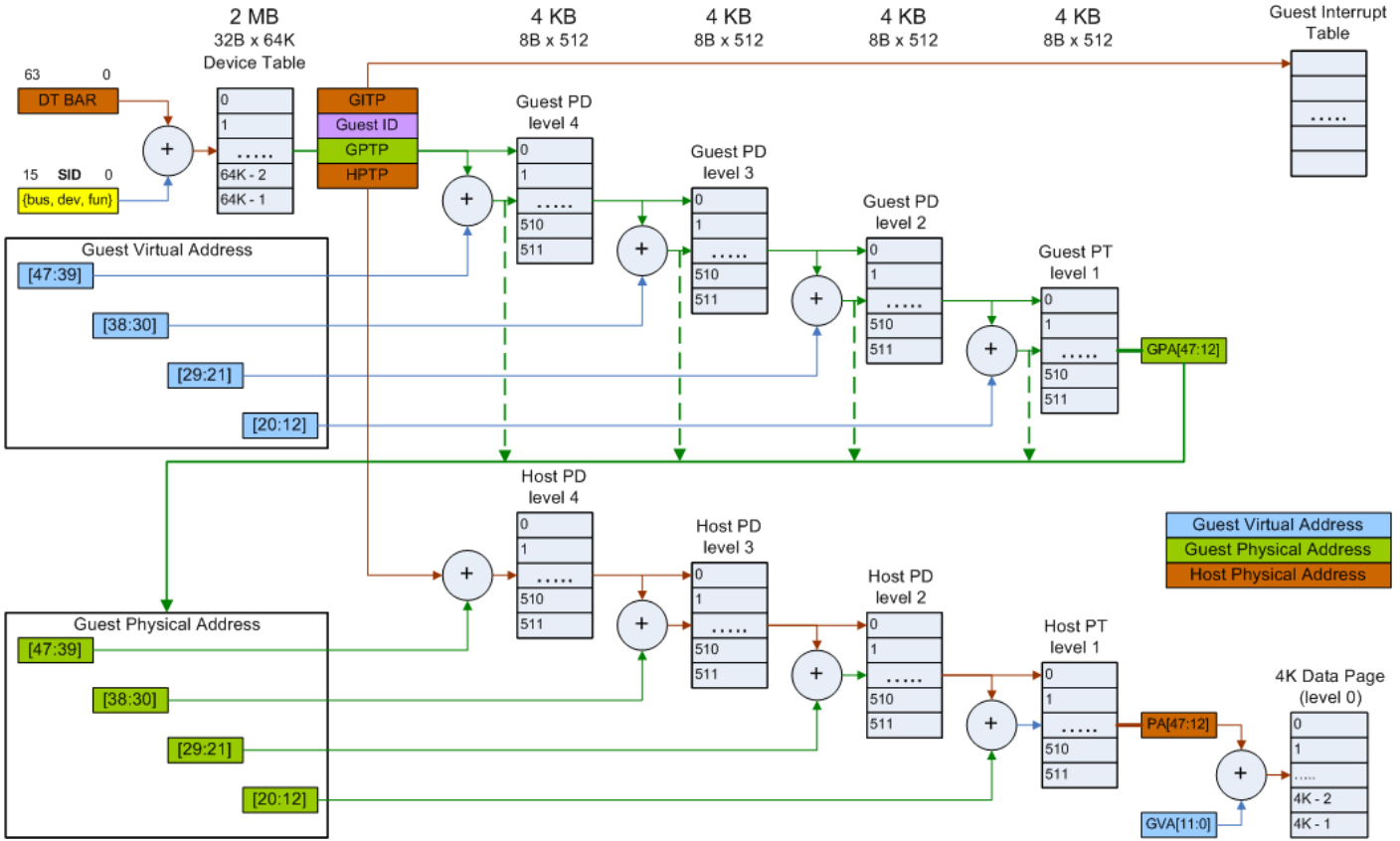
1) работать под управлением ГВ
   a) по PA
   b) по HVA в с разделением на домены ГВ
2) работать под управлением Гостя
   1) по GPA
   2) по GVA в общем домене

SID (Source ID) => Guest ID / ГВ Domain ID

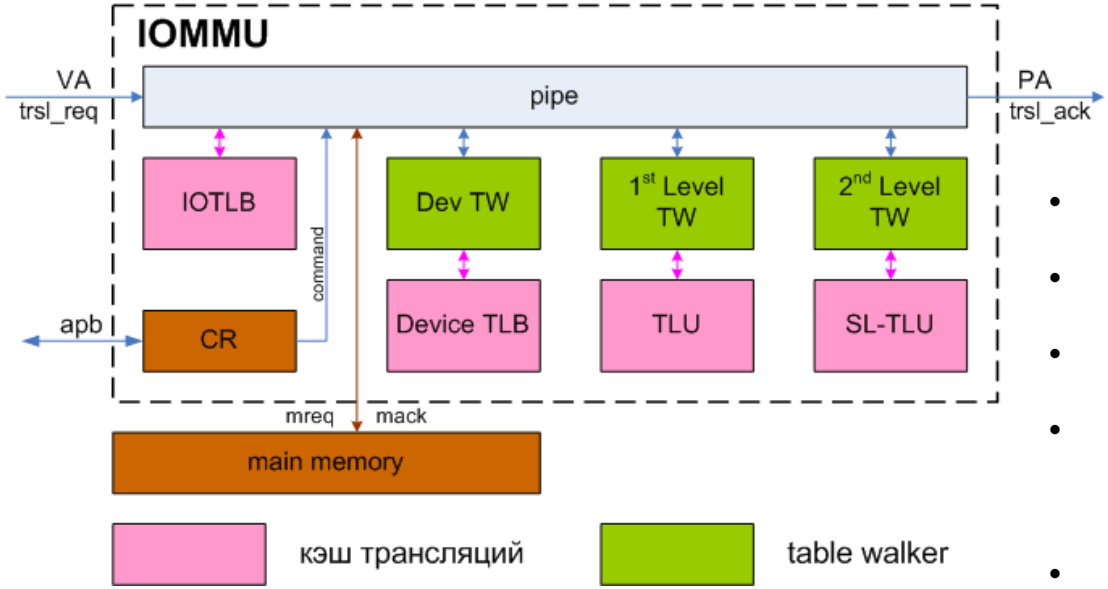


IOMMU

Режим виртуализации. Структуры трансляции



Структура IOMMU



- pipe – конвейер, 3 стадии
- IOTLB – кэш конечных трансляций
- Device TLB – кэш DTE
- TLU – кэш трансляций первого уровня: **HVA -> PA** и **GVA -> GPA**
- SL-TLU - кэш трансляций второго уровня: **GPA -> PA**