Поддержка механизма Peer-to-Peer в микропроцессоре Эльбрус-32С

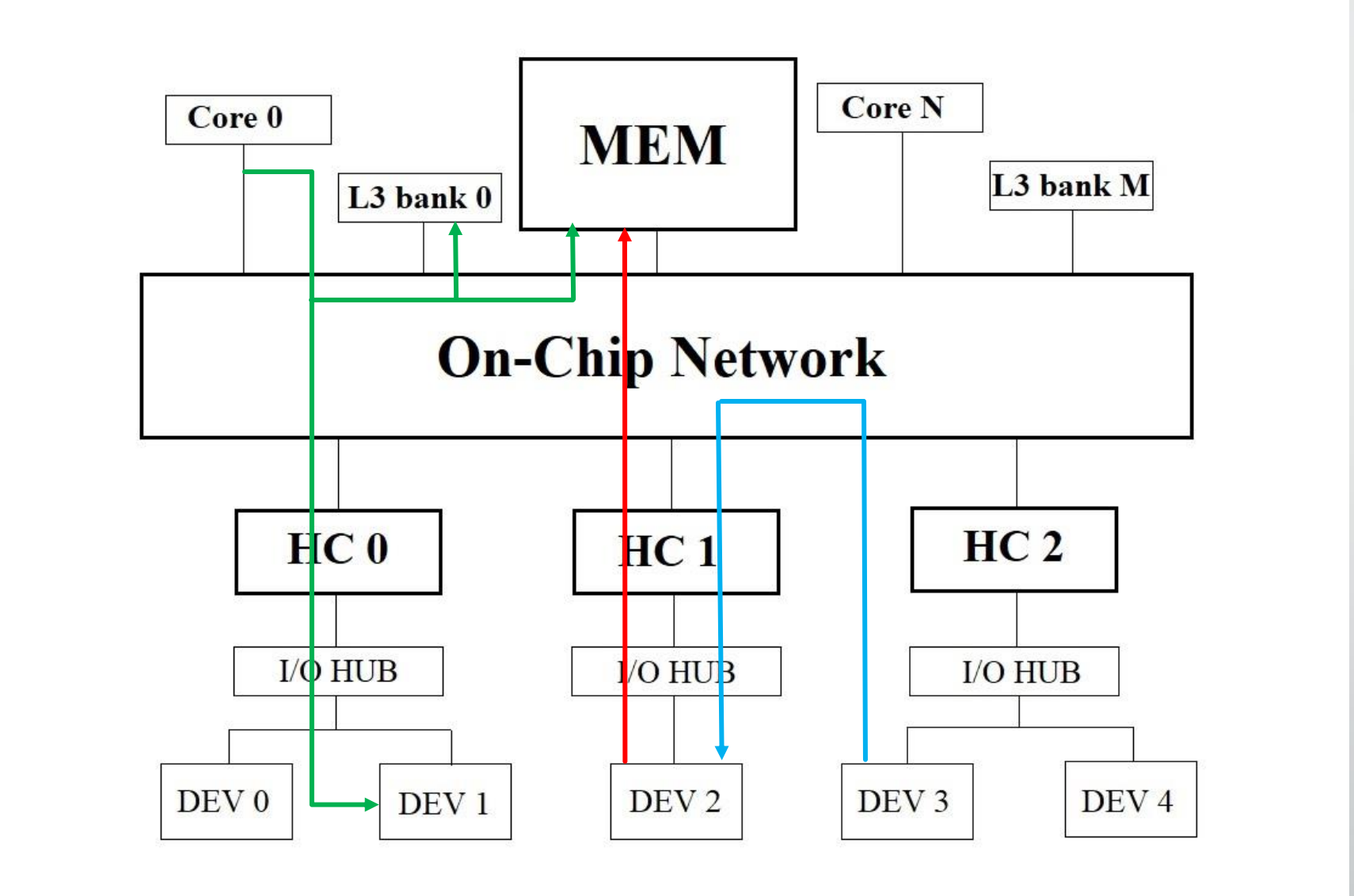
Студент: Валюков Николай

Научный руководитель: д.т.н. Фельдман В. М.

Режимы работы внешних устройств

- PIO (Programmable Input/Output) - передача данных между устройством и памятью выполняется за счёт ядра
- DMA (Direct Memory Access) - устройство напрямую делает чтение/запись в память
- P2P (Peer-to-Peer) - устройство напрямую делает чтение/запись в другое устройство

- PIO
- DMA
- P2P


Core 0
MEM
Core N
L3 bank 0
L3 bank M
On-Chip Network
HC 0
HC 1
HC 2
I/O HUB
I/O HUB
I/O HUB
DEV 0
DEV 1
DEV 2
DEV 3
DEV 4

Цель работы: поддержка механизма P2P в МП Эльбрус-32С

Задачи:

- Модифицировать протокол ESP (Elbrus System Protocol) для поддержки механизма P2P
- Разработать RTL-описание НС с поддержкой механизма P2P
- Обеспечить масштабируемость решения в системе-на-кристалле с несколькими НС

ESP (Elbrus System Protocol)
Протокольный уровень

- **IRQ** - класс первичных запросов
- **SRQ** - класс снуп-запросов
- **DAT** - класс ответов с данными
- **RSP** - класс ответов без данных

ESP (Elbrus System Protocol)

Протокольный уровень

Первичные запросы

- IRQ PIO - первичные запросы в IO
- IRQ A2B - первичные запросы в L3
- IRQ A2M - первичные запросы в оперативную память
- IRQ PRM - первичные запросы в оперативную память с высоким приоритетом

Поддержка P2P:

+ HC отсылает IRQ PIO для P2P-запросов

ESP (Elbrus System Protocol)

Протокольный уровень

Ответы без данных

- ACK B2H - Snoop Acknowledgment, ответ-подтверждение от L3 кэша
- ACK C2U - Snoop Acknowledgment, ответ-подтверждение от ядра
- HAK - Home Acknowledgment, подтверждение от Home-устройства (L3, HMU, HC)
- RLS – Release, сообщение о завершении транзакции

ESP (Elbrus System Protocol)
Транспортный уровень

- **LC0** - канал первичных (IRQ) запросов, уровень приоритета 0
- **LC1** - канал снуп-запросов (SRQ), уровень приоритета 1
- **LC2** - канал ответов с данными (DAT), уровень приоритета 2
- **LC3** - канал ответов без данных (RSP), уровень приоритета 3

Порядок пакетов с одинаковым адресом назначения сохраняется в одном логическом канале

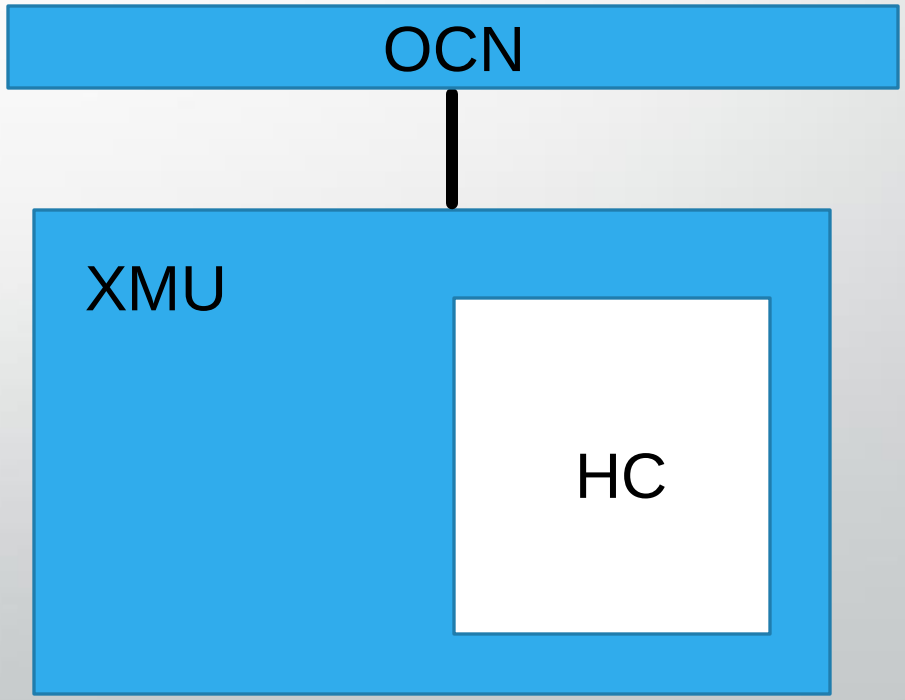
ESP (Elbrus System Protocol)

Сетевой уровень

- C2X – Core to XMU
- B2X – L3 Bank to XMU
- H2X – HMU to XMU
- X2B – XMU to L3 Bank
- X2H – XMU to HMU

Поддержка P2P:

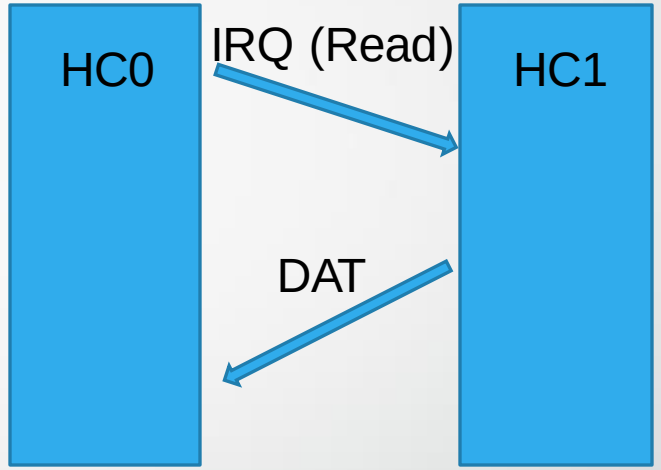
+ X2X – XMU to XMU



ESP для P2P транзакций

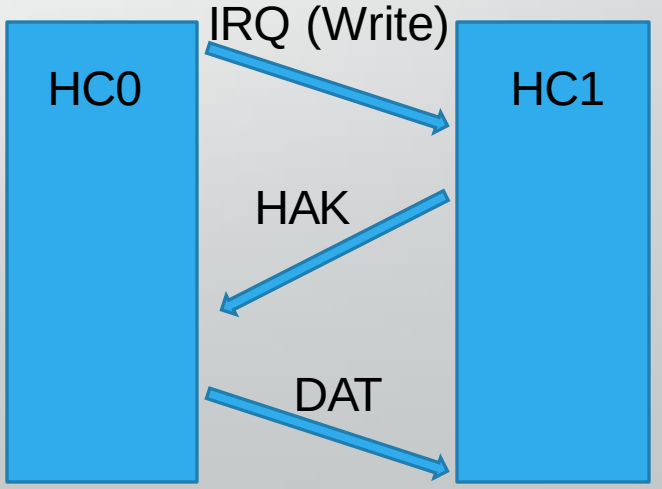
- **Чтение**

1) HC #0 отправляет в HC #1 первичный запрос IRQ (Read)
2) HC #1, получив пакет IRQ, отправляет данные чтения в HC#0 пакетом DAT



- **Запись**

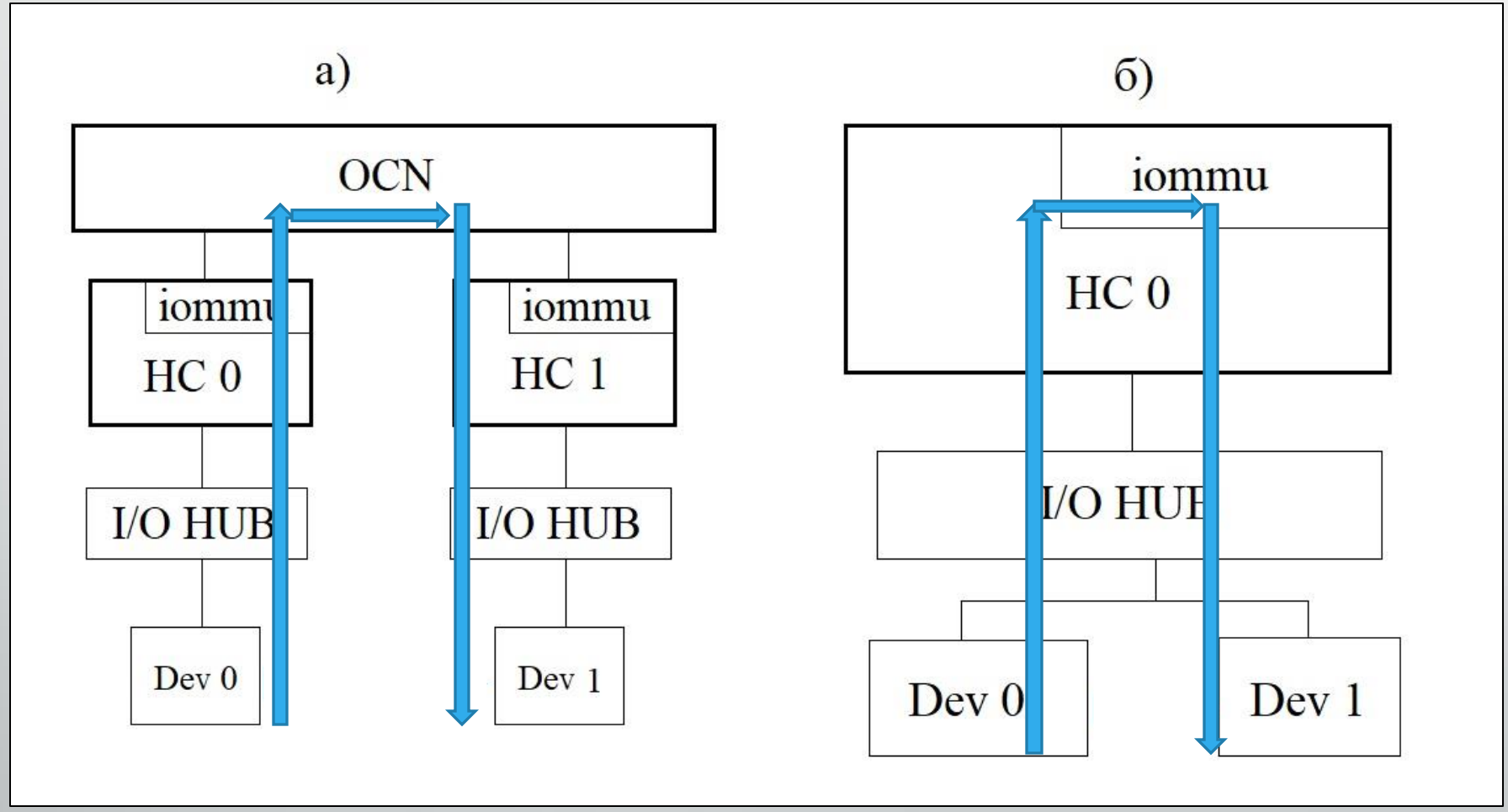
1) HC #0 отправляет в HC #1 первичный запрос IRQ (Write)
2) HC #1, получив пакет IRQ, отправляет пакет HAK "выдай данные записи" в HC #0
3) HC #0, получив пакет HAK, отправляет данные записи пакетом DAT



Этапы обработки P2P-транзакций

1) Запрос на чтение/запись поступает из I/O HUB в HC
2) Трансляция адреса в IOMMU
3) Определение адресного пространства запроса (оперативная память или другое устройство)
4) Выдача запроса в сеть или обратно в свой I/O HUB

Два случая обработки P2P-транзакций



Масштабируемость решения в системе с несколькими НС

Проблема DMA-кэша

	E16C	E32C
Кол-во строк в DMA-кэше	64	64
Кол-во НС в одном МП	1	3
Кол-во МП в кластере	4	4
Кол-во строк в глобальном справочнике	256	768

Решение: отказ от DMA-кэша, замена на буфер записей

Проблемы отказа от DMA-кэша:

- Упорядоченность PCI-E-транзакций
- Возможная блокировка в устройствах L3, HMU

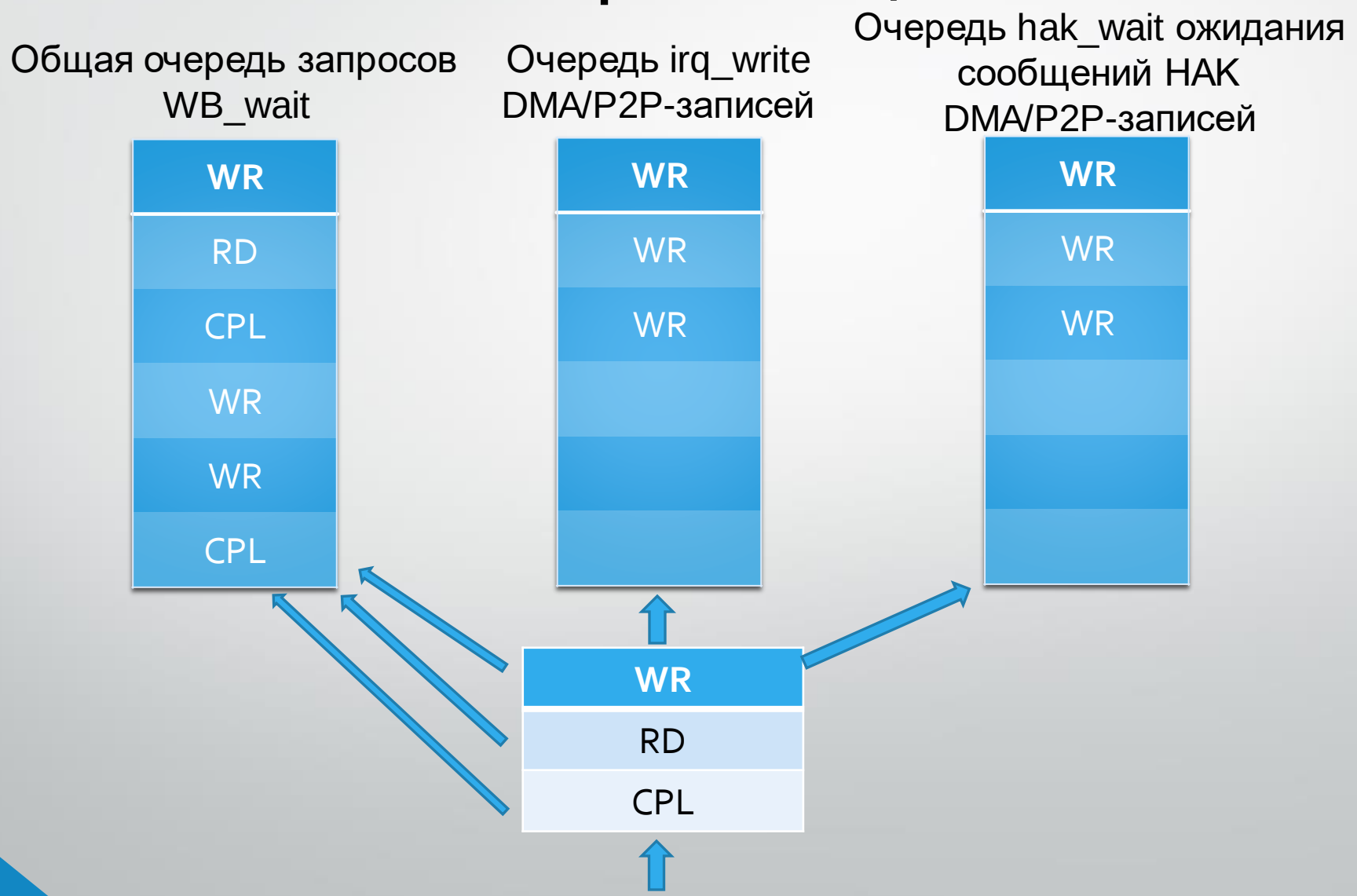
Упорядоченность PCI-E-транзакций

Возможность обгона строка -> столбец	WRITE	READ	CPL
WRITE	нет/да**	да	да
READ	нет/да*	да	да
CPL	нет/да**	да	да

* - разрешен обгон, если у второго запроса выставлен признак IDO и у обоих запросов разный SID

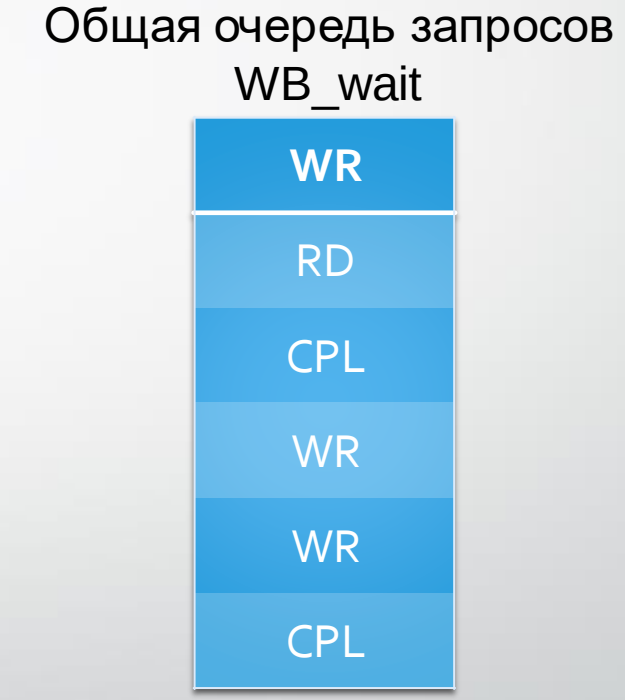
** - разрешен обгон, если у второго запроса выставлен признак RO. Также разрешен обгон, если у второго запроса выставлен признак IDO и у обоих запросов разный SID

Реализация упорядоченности PCI-E-транзакций



Реализация упорядоченности PCI-E-транзакций

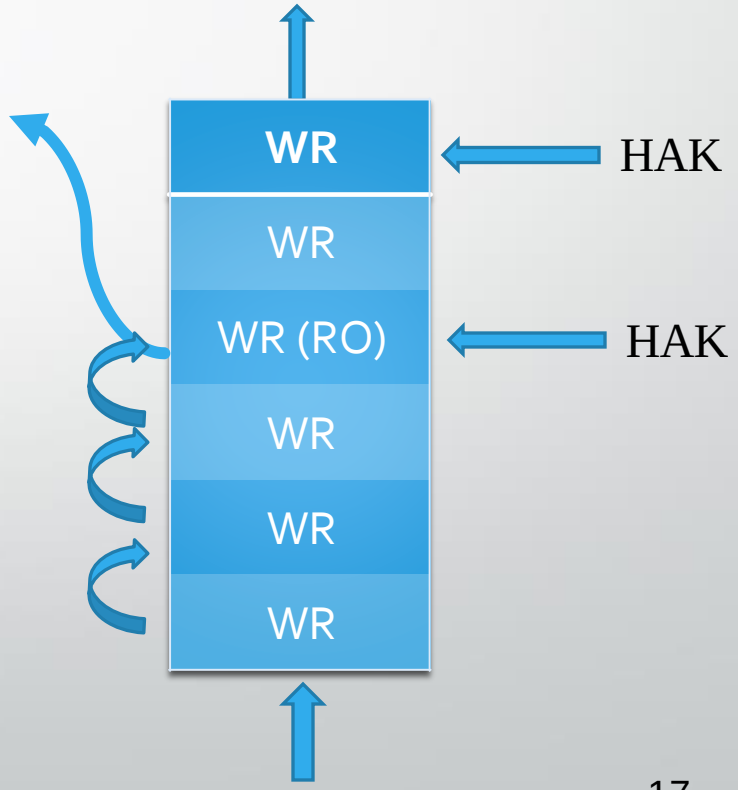
- DMA/P2P-записи изымаются из "головы" очереди WB_wait только после отправки данных и получения сообщения RLS
- DMA/P2P-чтения отправляют IRQ по достижению "головы" очереди, после чего сразу изымаются из очереди
- PIO CPL отправляют RLS (для PIO-записей) или DAT (для PIO-чтений) по достижению "головы" очереди, после чего сразу изымаются из очереди



Реализация упорядоченности PCI-E-транзакций

- Сообщения HAK могут приходить в произвольном порядке
- Выдача данных записи происходит после получения сообщения "дай данные" (HAK) запросом, находящимся в "голове" очереди hak_wait, для соблюдения упорядоченности запросов

Очередь ожидания HAK для записей hak_wait

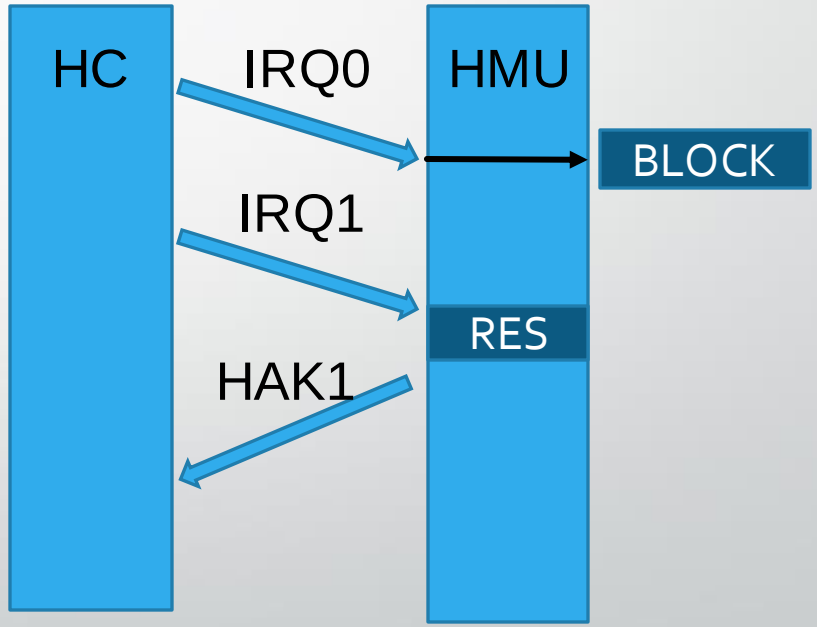


Оптимизация: если у запроса по записи выставлен признак RO (Relaxed Ordering), то ему разрешается выдавать данные вне очереди

Механизм выхода из блокировки

Таймер начинает отсчитывать после достижения запросом по DMA-записи "головы" очереди hak_wait. Если таймер отсчитает определённое количество тактов, запускается механизм выхода из блокировки:

1) Отправка сообщения RLS (WRCNL) для оповещения об отмене операции по DMA-записи всем абонентам системы, которые прислали сообщение HAK ("дай данные")
2) Ожидание сообщения HAK для запроса, находящегося в "голове" очереди ожидания HAK_wait
3) Повторная отправка сообщений IRQ (W) для всех отменённых операций



Результаты:

- Разработана модификация протокола ESP для поддержки P2P
- Разработано RTL-описание HC с поддержкой P2P
- Обеспечена масштабируемость системы с несколькими HC за счет отказа от DMA-кэша
- Реализована оптимизация обработки DMA/P2P-записей с признаком Relaxed Ordering

Реализация упорядоченности PCI-E

- Из очереди irq_write отправляются запросы IRQ (W) независимо от положения запроса в общей очереди запросов WB_wait