Московский физико-технический институт
(государственный университет)
Факультет радиотехники и кибернетики
Кафедра информатики и вычислительной техники

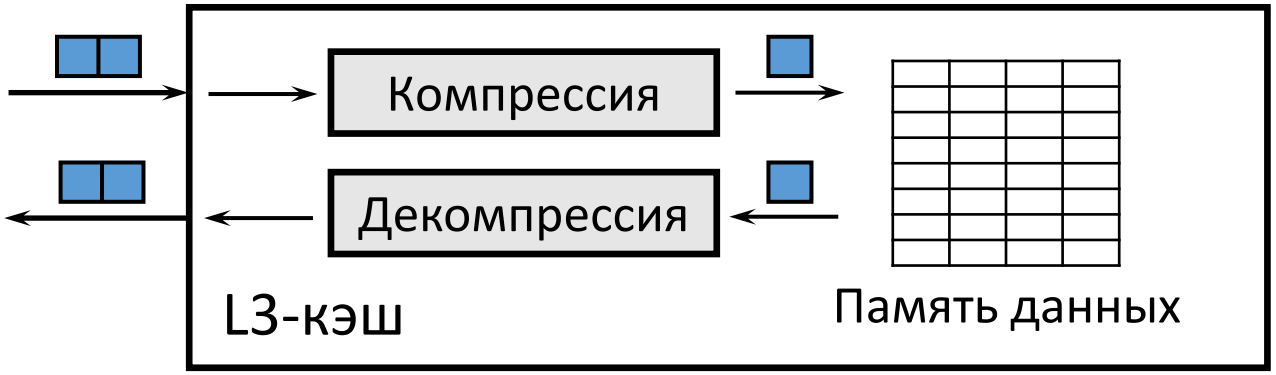
РАЗРАБОТКА МЕХАНИЗМА КОМПРЕССИИ ДАННЫХ ДЛЯ КЭШ-ПАМЯТИ ТРЕТЬЕГО УРОВНЯ ПРОЦЕССОРОВ С АРХИТЕКТУРОЙ «ЭЛЬБРУС»

Выпускная квалификационная работа
(магистерская диссертация)

Студент: Сурченко А.В.
Научный руководитель: Фельдман В.М.
Консультант: Кожин А.С.

Москва, 2020

Компрессия данных в кэш-памяти



- Применение:
  Увеличение эффективного объема кэш-памяти при незначительном изменении площади и времени доступа

- Алгоритм BΔI*-HL:
  - **Скорость** (задержка декомпрессии – 1 такт)
  - **Простота** (занимает не более 0,1% от площади банка L3)
  - **Эффективность** (доля сжатых строк – 24,2%, степень сжатия – 1.246)

  Характеристики алгоритма исследованы в ходе бакалаврской работы

Цель работы

Разработка механизма компрессии на базе алгоритма BΔI*-HL в кэш-памяти третьего уровня процессора «Эльбрус-16С»

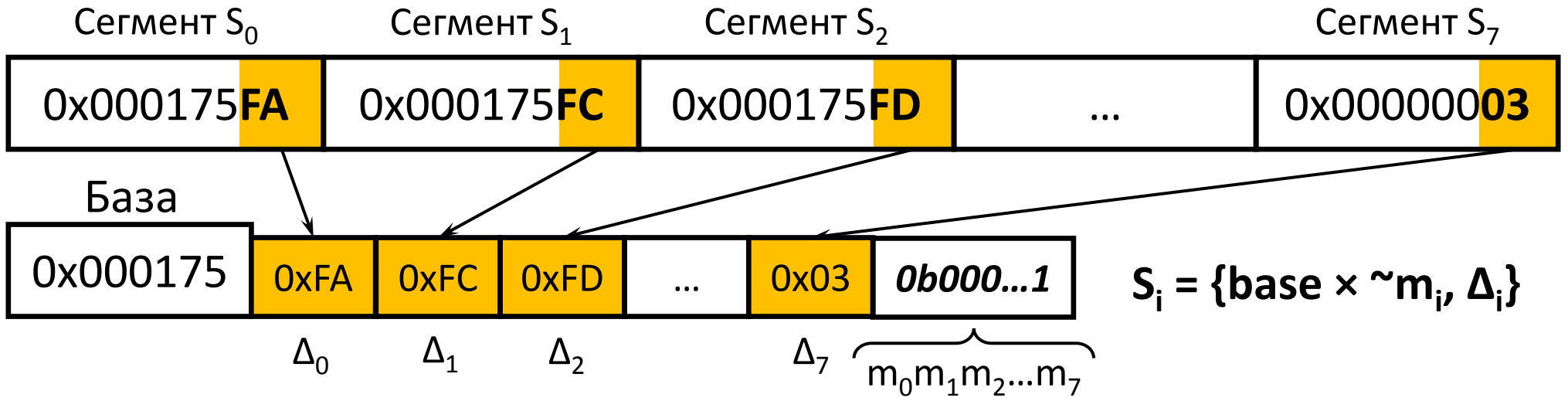
Задачи:

- разработка и RTL-описание алгоритма совместного размещения сжатых кэш-строк для увеличения эффективного объема L3-кэша
- доработка RTL-описания L3-кэша для интеграции механизмов компрессии и декомпрессии
- отладка разработанного решения на Verilog-модели и FPGA-прототипе процессора «Эльбрус-16С»

Требования:

- Внедрение механизма не должно вносить существенные изменения в структуру L3-кэша, влиять на алгоритмы вытеснения и протокол когерентности
- Внедрение механизма не должно уменьшать частоту работы L3-кэша и существенно увеличивать время доступа к нему

Принцип сжатия BΔI*-HL



- Кэш-строка параллельно разбивается на 4, 8, 16 и 32 сегмента равных размеров, проверяется равенство их старших разрядов нулю или между собой
- Сжатая кэш-строка содержит общие старшие разряды (база), набор младших разрядов – смещений (Δ), и маску, показывающую, относительно нуля или базы получено смещение (m)
- Размеры смещений выбираются так, чтобы сжатая кэш-строка была не больше половины размера полной
- Это повышает пропускную способность внутри кэша (кэш-строки передаются по половинам) и позволяет хранить две кэш-строки в одной ячейке памяти

Организация L3-кэша процессора «Эльбрус-16С»

L3-кэш состоит из памяти тэгов и состояний
(TA, Tag Array) и памяти данных (DA, Data Array)

В TA находится локальный справочник LD с
информацией о когерентном состоянии строк
в L3 и кэшах ядер

Совместное размещение сжатых кэш-строк

Проблема: как хранить больше данных в кэш-памяти, увеличив тем самым ее эффективный объем?

- Записи в памяти тэгов дополняются информацией о компрессии данных (позволяет не хранить нулевые данные)
- Тип компрессии тэге указывает на размер данных в кэш-строке (полный, половинный или нулевой)

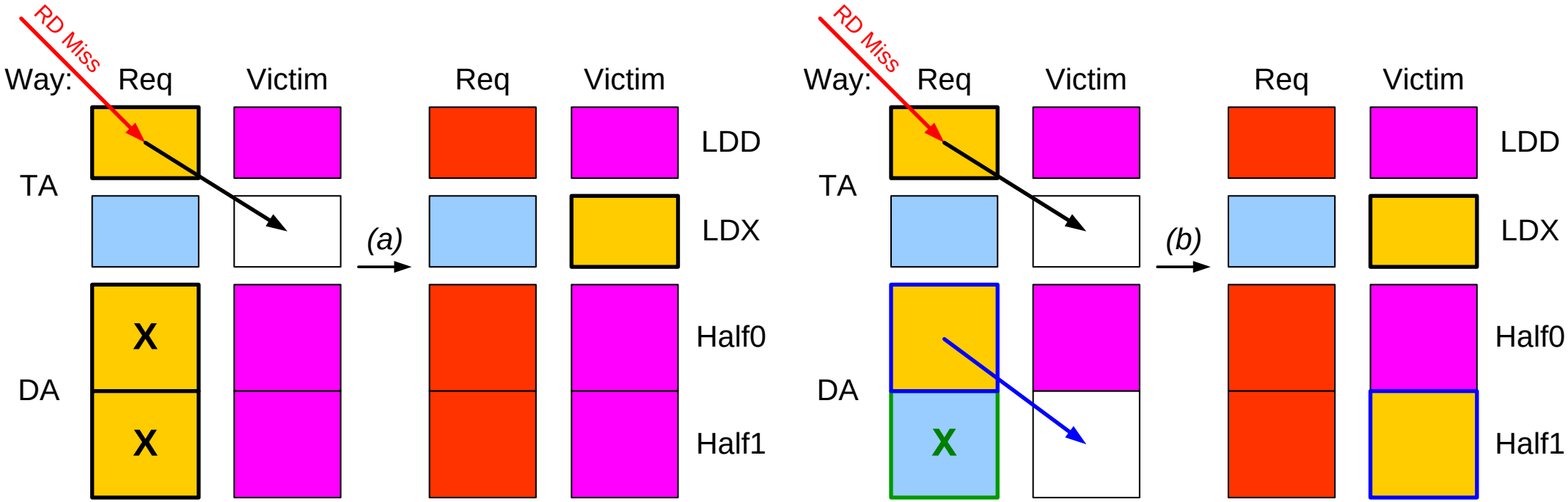
Действия при обработке запросов в L3-кэше без компрессии

В зависимости от того, какого типа запрос по протоколу когерентности приходит в кэш-память и его попадания/промаха выполняемые действия различаются:

Дополнительные действия при обработке запросов в L3-кэше с компрессией

Дополнительные действия при обработке запросов в L3-кэше с компрессией на примере RD Miss

Одинаковые цвета блоков на рисунке показывают зависимость между тэгом и данными. Цвет линий показывает зависимость между переносами и переносимыми тэгами/данными



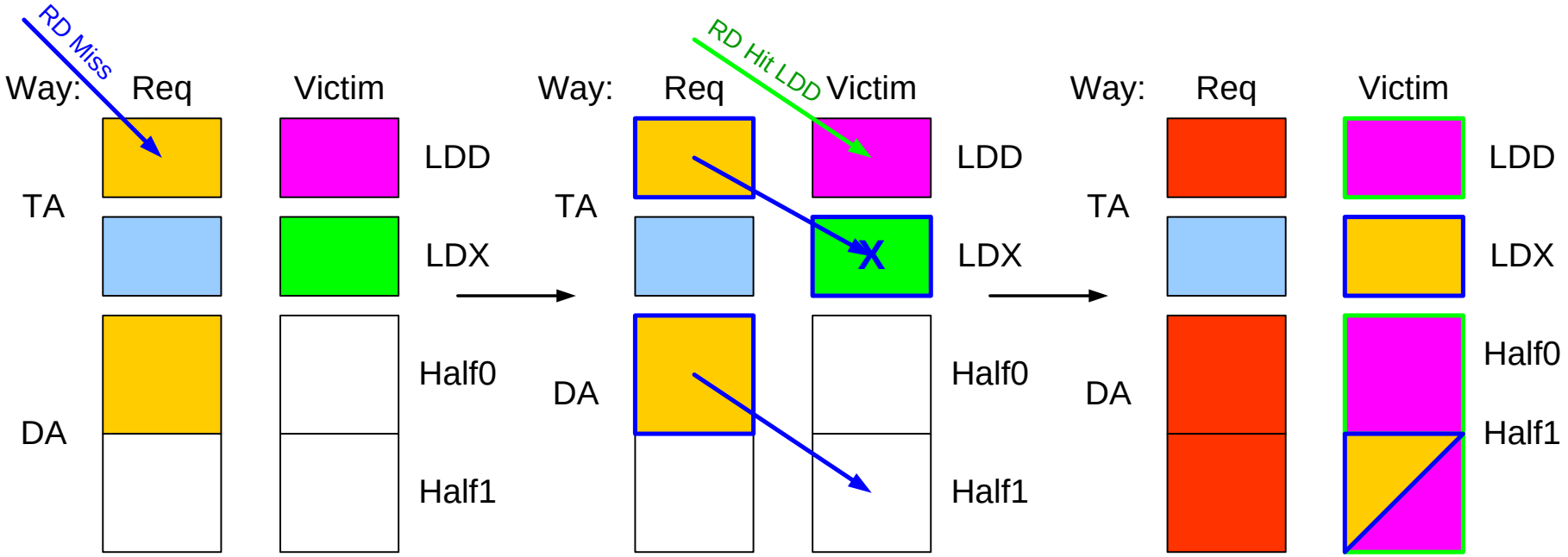
(a) – действия при L3-кэше без компрессии

(b) – действия при L3-кэше с компрессией. Занесение данных в кэш требует дополнительного вытеснения данных при тэге в LDX

Проблема потери актуальности информации о данных

Необходимо разработать ***ограничения для поддержки атомарности транзакции*** в пределах одного ассоциативного пути

Их отсутствие может привести к ***ошибке*** из-за потери актуальности информации о размерах данных, участвующих в транзакциях



RD Miss: соседствующих данных по месту записи нет, данные при вытеснении переносятся

RD Hit LDD: данных при тэге LDX нет, вытеснения не требуется

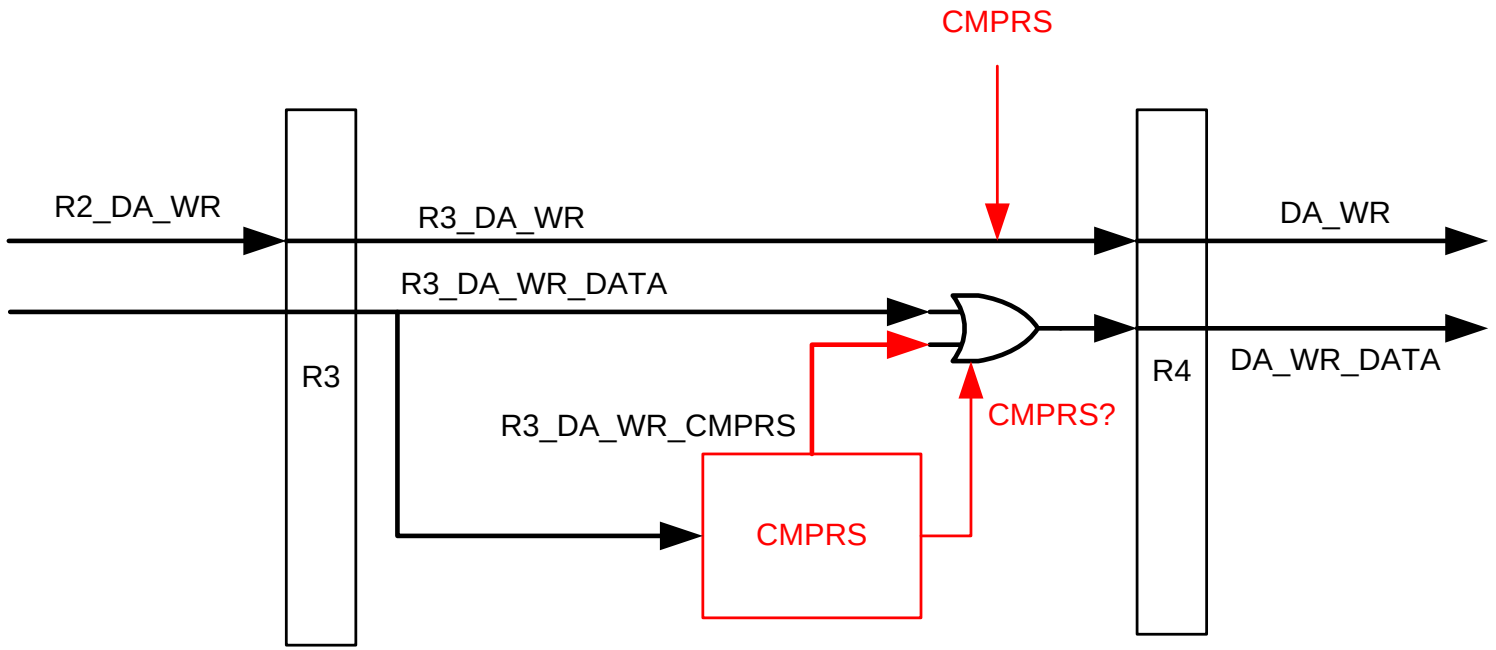
Возникает потеря данных одной из транзакций

Поддержка атомарности транзакций в пределах одного ассоциативного пути

Введенные ограничения для поддержки атомарности:

Поиск описанных ситуаций и решения о размещении данных принимаются при начале работы транзакции, предотвращая конфликты

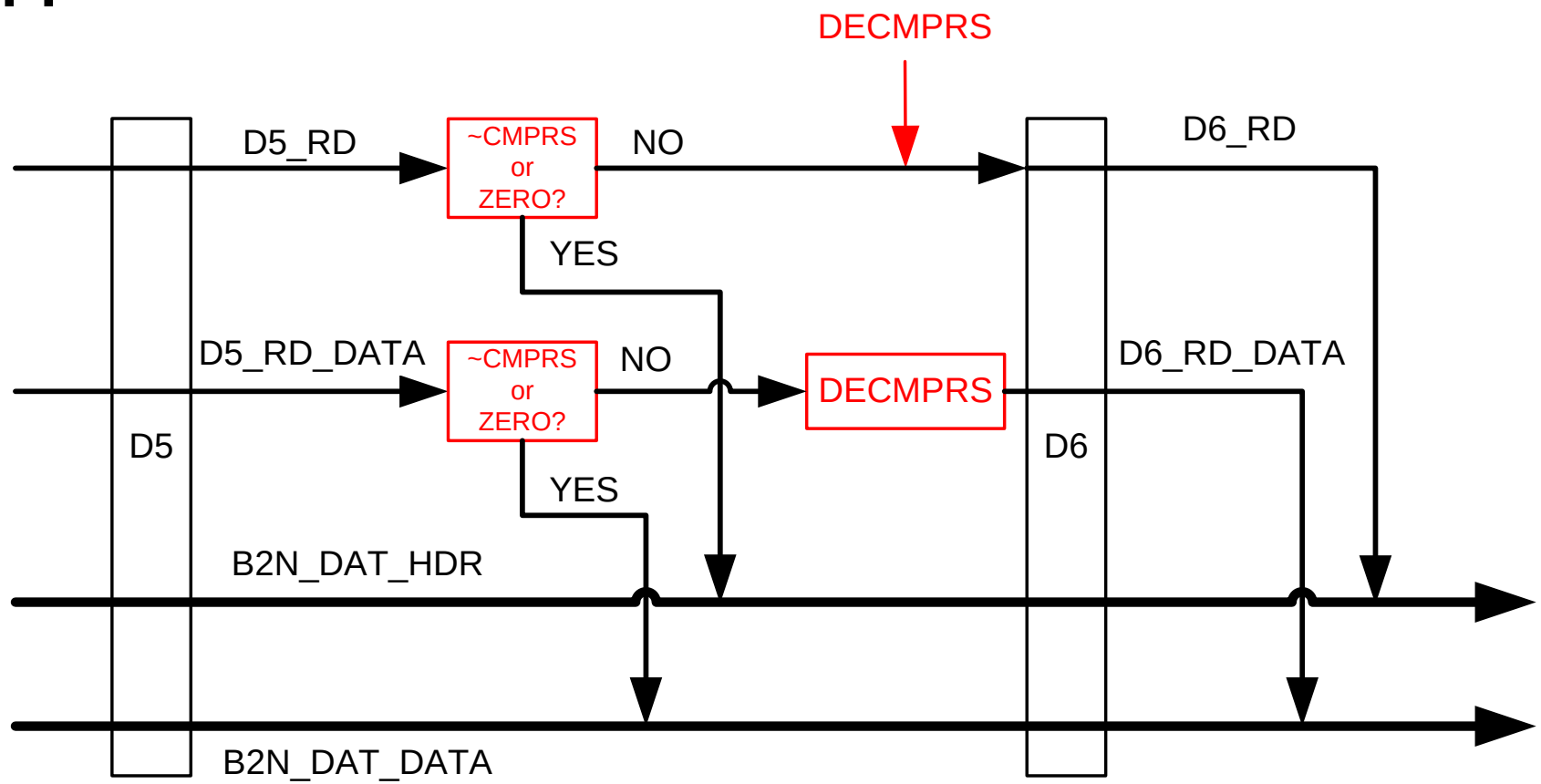
Интеграция механизма компрессии в конвейер записи данных



В случае, когда данные записи R3_DA_WR_DATA удалось сжать, производится модификация записи в память данных

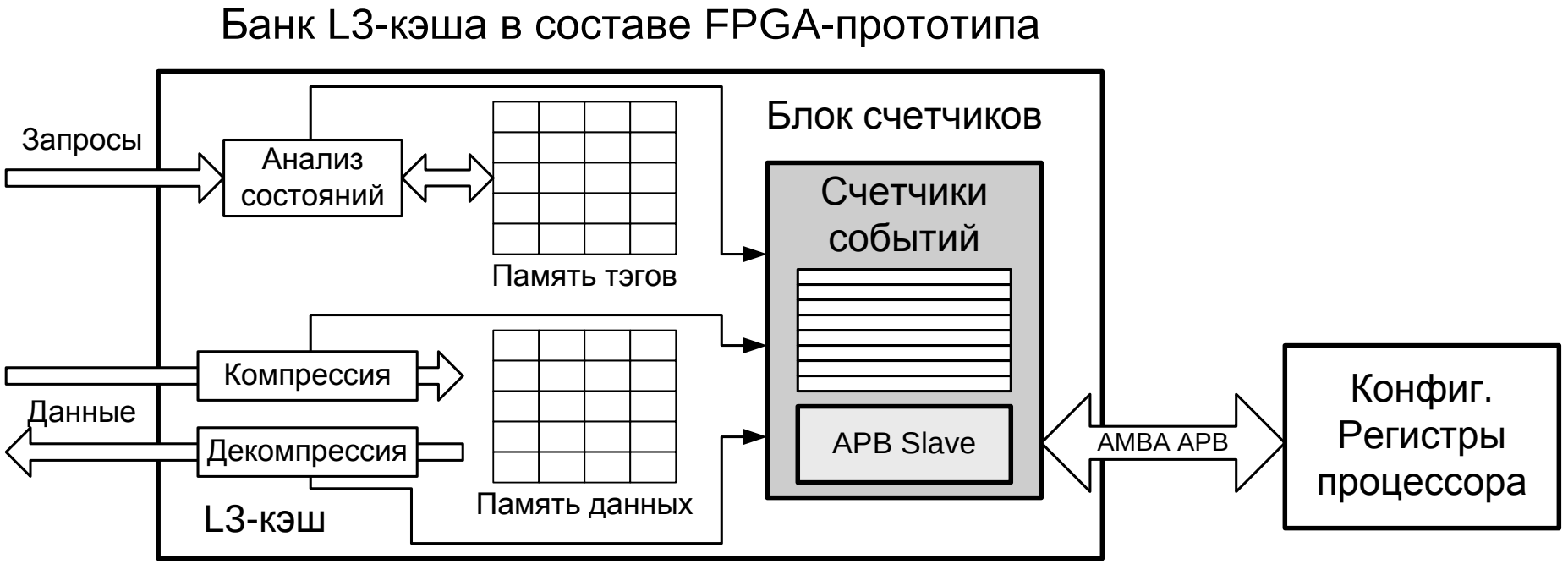
- Запись дополняется информацией о сжатии данных, которая затем записывается в соответствующий кэш-строке тэг
- В случае сжатия до половины размера запись передается в один такт вместо двух
- В случае нулевой строки запись отменяется, меняется только информация в тэге кэш-строки

Интеграция механизма декомпрессии в конвейер чтения данных



- При чтении из памяти данные D5_RD_DATA подвергаются декомпрессии, если они были сжаты и не являются нулевыми, т.к. в последнем случае значение данных заранее известно
- Декомпрессия задерживает отправку пакета с данными “DAT” на такт

Тестовый стенд



- Был разработан тестовый стенд, позволяющий собирать данные о работе алгоритмов компрессии и декомпрессии, совместном размещении данных и кэш-памяти в целом через программно-доступные счетчики событий
- Тестирование производилось на Verilog-модели, системных и автономных тестах, а также на FPGA-прототипе процессора «Эльбрус-16С»
- Прошли 184 теста на базе системных тестов на когерентность
- Удалось добиться прогона на автономной конфигурации из 2854 тестов

Характеристики разработанного механизма

Разработанный механизм компрессии соответствуют заявленным требованиям к проекту

- В кэш-памяти осталась неизменной работа алгоритмов вытеснения и протокола когерентности
- Разработанный механизм компрессии не потребовал существенных изменений в структуре кэш-памяти
- Не изменилось время доступа по чтению для несжатых кэш-строк, время доступа к сжатым строкам, кроме доступа к нулевым данным, увеличилось на один такт
- Результаты синтеза в САПР Synopsys Design Compiler и Intel Quartus:
  - частота работы не изменилась и составила 2ГГц
  - прирост аппаратуры – 4,8% и 10% от общего числа регистров и логических элементов банка L3-кэша соответственно
  - механизм занимает 1% от общей площади всего банка, размер L3-кэша при этом не меняется

Результаты работы

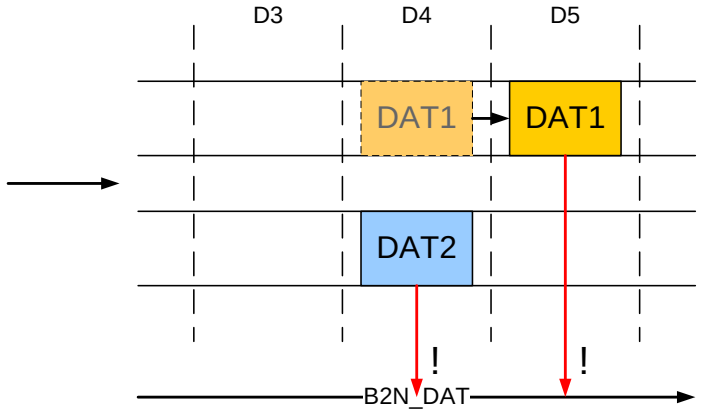
- Доработано RTL-описание кэш-памяти для поддержки работы со сжатыми кэш-строками
- Разработан алгоритм совместного размещения сжатых кэш-строк для увеличения эффективного объема кэш-памяти, разработано его RTL-описание
- Разработаны ограничения для поддержки атомарности транзакций, связанных с обработкой в L3-кэше запросов по протоколу когерентности
- Характеристики разработанного механизма соответствуют заявленным требованиям к проекту

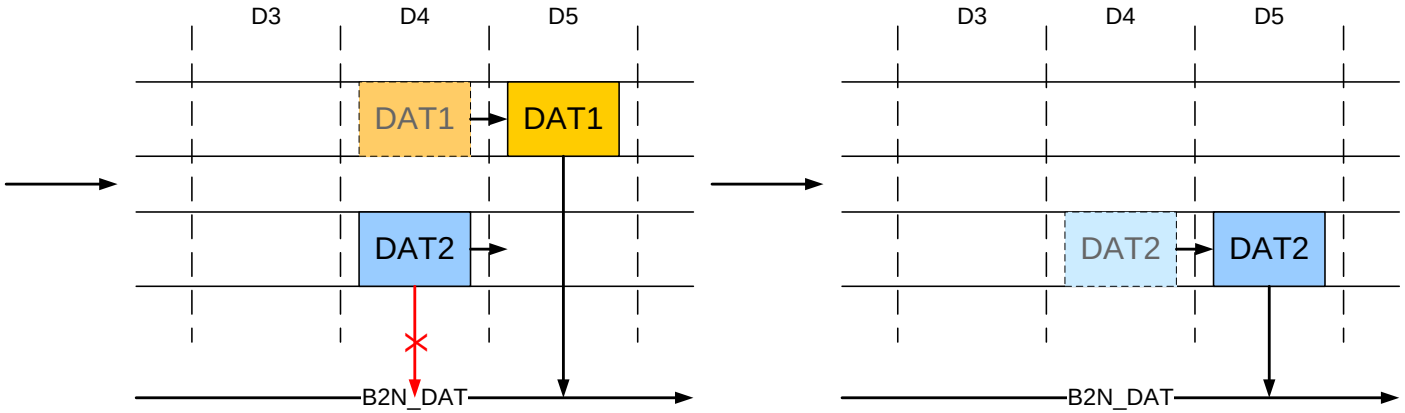
- В настоящее время механизм отлажен на системных и автономных тестах, начата отладка на FPGA-прототипе

Дополнительные материалы

Конфликт выдачи пакетов DAT

Поскольку пакеты могут отправляться с разных стадий конвейера, возможен конфликт при одновременной отправке двух пакетов





Для предотвращения конфликта передача пакета DAT2 на интерфейс задерживается на такт