Нижегородский государственный университет им. Н.И. Лобачевского
Факультет вычислительной математики и кибернетики

Образовательный комплекс «Введение в принципы функционирования и применения современных мультиядерных архитектур (на примере Intel Xeon Phi)»

Лекция №2
Архитектура Intel Xeon Phi

Линев А.В.

При поддержке компании Intel

Нижний Новгород
2013

Содержание

Введение

В данном разделе курса описывается аппаратная архитектура и программная модель сопроцессора Intel Xeon Phi. Рассматриваются основные архитектурные блоки и особенности сопроцессора: ядро, блок векторной обработки данных, встроенная высокопроизводительная двунаправленная кольцевая шина, полностью когерентные кэши L2 и принципы взаимодействия компонент. Основное внимание уделяется элементам, наиболее существенно влияющим на производительность вычислений и понимание способов оптимизации программ для архитектуры Intel Xeon Phi.

1. Используемая терминология

Хост, **хост-система**, **базовая система** – вычислительная система на базе Intel Xeon и совместимых с ним процессоров, в которой установлен сопроцессор Intel Xeon Phi. На хосте может работать операционная система семейств Red Hat Enterprise Linux 6.x или SUSE Linux Enterprise Server SLES 11.

Операционная система хоста, **ОС хоста** – операционная система, установленная на хост-системе.

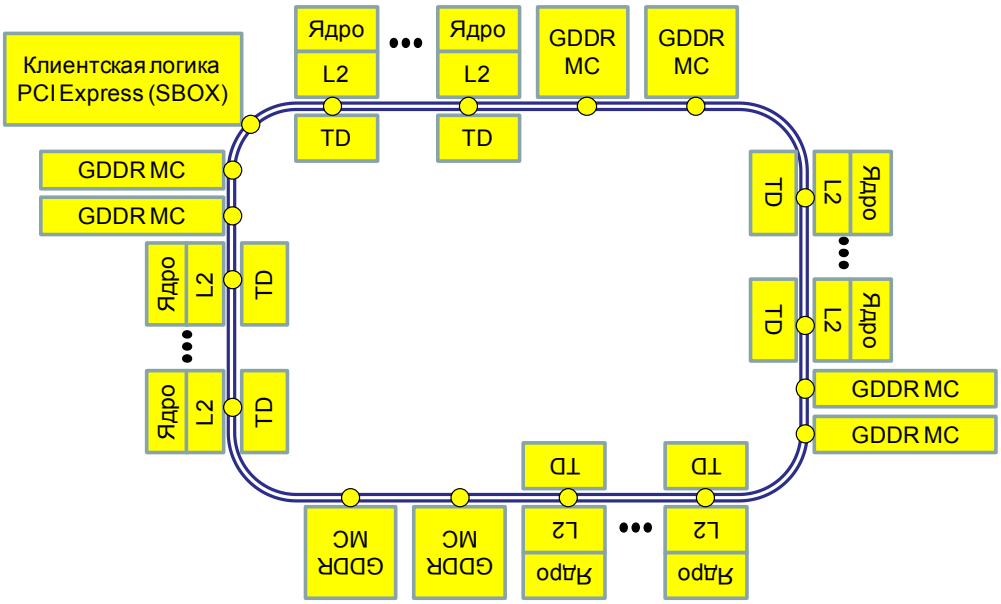
Сопроцессор, **целевая система** – сопроцессор Intel Xeon Phi и установленное на нем программное обеспечение.

uOS, **Micro Operating System**, **операционная система сопроцессора** – операционная система, установленная на сопроцессоре, базируется на ядре Linux.

MPSS, **Intel Manycore Platform Software Stack** – совокупность программного обеспечения системного и пользовательского уровней, обеспечивающая запуск и выполнение программ на сопроцессоре Intel Xeon Phi.

2. Архитектура сопроцессора Intel Xeon Phi

Сопроцессор Intel Xeon Phi включает до 61 процессорных ядер, соединенных высокопроизводительной встроенной кольцевой шиной. 8 контроллеров памяти обслуживают 16 каналов GDDR5, обеспечивая суммарную производительность 5,5 GT/s (миллиардов пересылок в секунду, при ширине шины 64 байта это дает пропускную способность 352 GB/s). Отдельный компонент реализует клиентскую логику PCI Express (см. рис. 1).



TD – Каталог тегов (Tag Directory), GDDR MC – Контроллер памяти

Рис. 1. Основные компоненты сопроцессора Intel Xeon Phi [1]

Каждое ядро является полнофункциональным и поддерживает выборку и декодирование инструкций из 4 потоков команд. Для повышения эффективности работы с памятью в сопроцессоре реализован распределенный каталог тегов кэша, позволяющий использовать более эффективный протокол для поддержания когерентности кэшей всех ядер. Контроллеры памяти обеспечивают теоретическую пропускную способность 352 гигабайта в секунду. Приведем основные характеристики компонент сопроцессора.

- Исполнительное ядро (Core) выполняет выборку и декодирование инструкций 4 аппаратных потоков. Поддерживается выполнение 32- и 64-битного кода, совместимого с архитектурой Intel64. Ядро содержит 2 конвейера (U-конвейер и V-конвейер) и может выполнять 2 инструкции за такт. V-конвейер способен выполнять не все типы инструкций, возможность параллельного выполнения команд на U- и V-конвейерах задается набором правил. Внеочередное выполнение инструкций не поддерживается, также не реализованы команды Intel Streaming SIMD Extensions (SSE), MMX и Advanced Vector Extensions (AVX). Ядро включает по 32 Кб 8-канальных множественно-ассоциативных кэшей инструкций и данных (L1 I-Cache и L1 D-Cache).

Ядро сопроцессора Intel Xeon Phi содержит следующие компоненты (см. рис. 2).

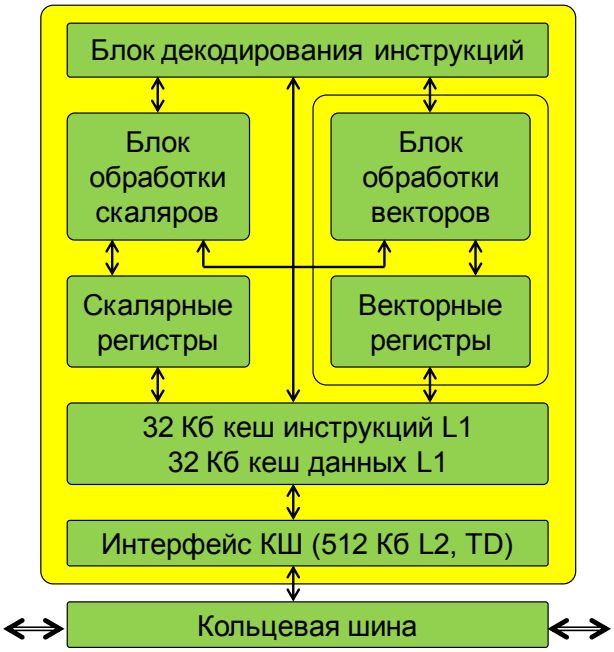


Рис. 2. Основные компоненты ядра сопроцессора Intel Xeon Phi [2]

- 512-битный блок векторных вычислений (vector processor unit, VPU) включает расширенный блок математических вычислений (extended math unit, EMU) и способен посылать на выполнение по одной векторной операции на каждом такте (то есть обработать 16 чисел с плавающей точкой одинарной точности или 16 32-битных целых чисел или 8 чисел с плавающей точкой двойной точности). Для операций «умножение и сложение» (multiply-add, FMA) это дает выполнение 32 операций над числами с плавающей точкой за такт. Блок векторных вычислений содержит 32 512-битных регистра (zmm0-zmm31) и дополнительно обеспечивает выполнение операций заполнения и перестановки содержимого векторного регистра, вычисление для вещественных чисел одинарной точности степеней 2 (2x) и двоичного логарифма (log2x), обратного значения (1/x) и обратного квадратного корня (1/sqrt(x)). При выполнении операций один из аргументов может считываться из оперативной памяти с выполнением при необходимости преобразования типа.
- Интерфейс кольцевой шины (Core-Ring Interface, CRI/L2) обеспечивает подключение ядра к высокопроизводительному встроенному интерконнекту сопроцессора – кольцевой шине, а также включает 512 Кб 8-канального множественно-ассоциативного кэша L2, усовершенствованный программируемый контроллер

прерываний (advanced programmable interrupt controller, APIC) и каталог тегов (Tag Directory, TD).

Каталог тегов (Tag Directory, TD) является частью распределенного каталога, обеспечивающего отслеживание всех адресов памяти, по которым происходило изменение данных всеми ядрами сопроцессора, и когерентность кэшей L2 всех ядер. Каждый тег содержит адрес, состояние и идентификатор владельца (кэша L2 какого-либо ядра) строки данных кэша. Пространство адресов оперативной памяти поровну разделено между каталогами тегов различных ядер, и ядро, на котором отсутствуют нужные данные, посылает запрос к соответствующему каталогу тегов через кольцевую шину.

- Контроллер памяти (GBOX, GDDR MC на Рис. 1) включает три основных компонента: интерфейс кольцевой шины (FBOX), планировщик запросов (MBOX) и интерфейс к устройствам GDDR. Каждый контроллер памяти включает два независимых канала доступа к памяти. Все контроллеры памяти сопроцессора действуют независимо друг от друга.
- Компонент SBOX реализует клиентскую логику PCI Express, включая механизм прямого доступа к памяти (Direct Memory Access, DMA) и ограниченные возможности по управлению питанием.
- Двунаправленная кольцевая шина обеспечивает передачу данных между компонентами сопроцессора.

Сопроцессор Intel Xeon Phi содержит 61 ядро, но он исполняет собственную операционную систему, и одно ядро выделено для исполнения кода ОС, обслуживания прерываний и т.п. Поэтому в расчетах производительности предполагается, что для вычислений используется 60 ядер из имеющихся 61.

Теоретическая производительность сопроцессора Intel Xeon Phi с 60 ядрами и частотой 1,1 ГГц может быть вычислена следующим образом [1]:

- 16 (длина вектора) * 2 flops(FMA) * 1.1 (GHZ) * 60 (число ядер) = 2112 GFLOPS – для вещественных чисел одинарной точности и
- 8 (длина вектора) * 2 flops (FMA) * 1.1 (GHZ) * 60 (число ядер) = 1056 GFLOPS – для вещественных чисел двойной точности.

2 flops за такт удается получить благодаря использованию инструкции «умножение и сложение» (multiply-add, FMA).

3. Конвейер ядра Intel Xeon Phi

Ядра Intel Xeon Phi обеспечивают выполнение 32- и 64-битного кода, совместимого с архитектурой Intel64 без поддержки расширений MMX, AVX и SSE (всех версий). Блок векторных вычислений, содержащийся в каждом ядре, дополнительно реализует набор операций над 512-битными векторами.

Конвейер ядра Intel Xeon Phi содержит 7 этапов, блок векторных вычислений также имеет конвейерную структуру и состоит из 6 этапов (см. рис. 3). Все этапы основного конвейера кроме последнего (WB), поддерживают спекулятивное выполнение. Каждое ядро может выполнять инструкции 4 потоков, что позволяет уменьшить потери из-за латентности доступа к памяти, выполнения векторных инструкций и т.д.

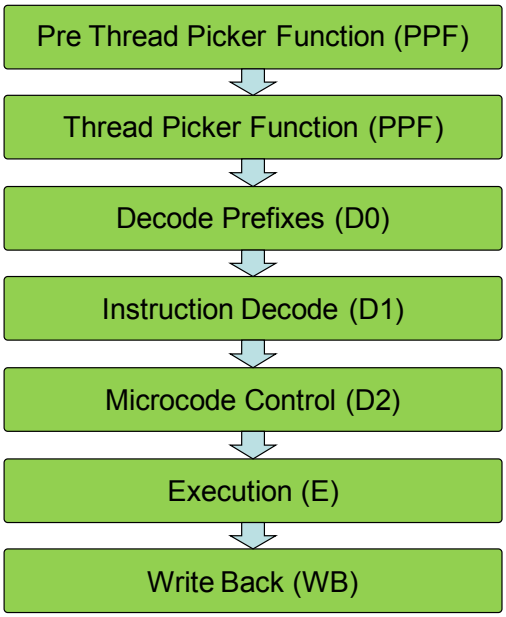


Рис. 3. Конвейер ядра Intel Xeon Phi

Выборка команд разбита на 2 этапа – PPF (pre thread picker) и PF (thread picker). На этапе PPF выполняется чтение инструкций потока исполнения в буфер предвыборки. На этапе PF производится выбор потока, инструкции которого будут выполняться, и передача пары инструкций для декодирования. Для каждого из четырех исполняющихся на ядре потоков имеется буфер предварительной выборки, который может содержать 2 инструкции для выполнения на U- и V-конвейерах ядра. Выбор инструкции для исполнения производится из заполненных буферов предвыборки согласно простому циклическому алгоритму (round robin).

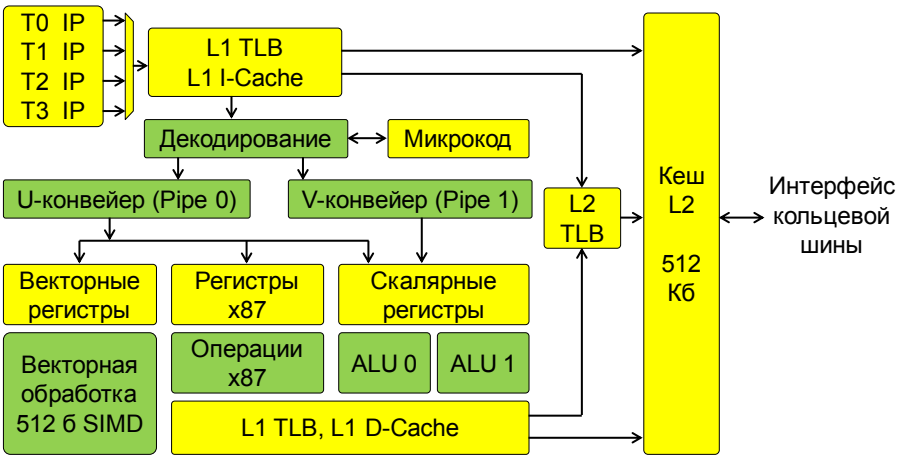


Рис. 4. Архитектура ядра Intel Xeon Phi [1]

Реализация выборки команд накладывает ограничение на выполнение потоков – на двух последовательных тактах не могут выбираться инструкции одного и того же потока. Таким образом, для полной загрузки ядра необходимо выполнять на нем по крайней мере два потока одновременно, а при работе только одного потока выборка инструкций будет выполняться через такт, что приведет к потере половины производительности. Для полной загрузки ядра достаточно выполнения на нем 2 потоков, однако, с учетом того, что заполнение буфера предварительной выборки требует 4-5 тактов при попадании в кэш инструкций и значительно больше при промахе, для обеспечения полной загрузки может потребоваться 3-4 потока.

После выбора пары инструкций для исполнения они отправляются на декодирование, состоящее из двух этапов – декодирование префиксов (Decode prefixes, D0) и декодирование инструкции (Instruction decode, D1), – которые выполняют декодирование двух инструкций за такт. На этапе D0 выполняется декодирование префиксов со штрафом от 0 до 2 тактов (для префиксов, унаследованных от старых архитектур). На этапе D1 выполняется декодирование инструкций с учетом результата декодирования префиксов. Далее следует этап управляемого выполнения микрокоманд (Microcode control, D2), на котором производятся операции чтения данных из регистров общего назначения, вычисления адреса и поиска и чтения данных из кэша.

Декодированные инструкции отправляются на этап исполнения (Execution, E), реализованный в виде двух конвейеров – U и V. Первая инструкция всегда отправляется на U-конвейер, для второй инструкции проверяется возможность одновременного выполнения с первой согласно набору правил парного выполнения команд, и в случае положительного решения она отправляется на V-конвейер. Скалярные целочисленные операции выпол-

няются арифметико-логическими устройствами (ALU), для скалярных и векторных операций с вещественными числами используется дополнительный 6-стадийный конвейер. Векторные инструкции выполняются в основном на U-конвейере.

Большинство инструкций с целыми числами и масками имеют латентность 1, большинство векторных инструкций – 4 или более при использовании операций чтения/записи с заполнением или перестановкой.

4. Иерархия памяти

Каждое ядро сопроцессора Intel Xeon Phi имеет собственные кэши L1 и L2, все ядра совместно используют оперативную память сопроцессора. Кэши L1 и L2 являются инклюзивными, то есть все данные, хранящиеся в кэше L1, хранятся также в кэше L2. В обоих кэшах при замещении используется псевдо-LRU алгоритм [3].

4.1. Кэши L1 и L2

Кэши первого уровня (кэш инструкций L1 I-Cache и кэш данных L1 D-Cache) имеют размер по 32 Кб, размер строки 64 байта, степень ассоциативности 8. Кэш L1 имеет среднюю латентность доступа 3 такта, поскольку использование регистров общего назначения в качестве базовых или индексных требует 3 или более тактов для формирования адреса («чистая» латентность кэша L1 составляет 1 такт). Средняя load-to-use латентность составляет 1 такт – целочисленное значение, загруженное на текущем такте из кэша, может быть использовано на следующем такте целочисленной инструкцией; однако для векторных инструкций load-to-use латентность может быть больше.

Кэш второго уровня L2 имеет размер 512 Кб, размер строки 64 байта, степень ассоциативности 8, 32 Гб кэшируемых адресов (размер адреса 35 бит), «чистую» латентность доступа 11 тактов и среднюю 14-15 тактов. Кэш L2 имеет аппаратное потоковое устройство предвыборки, способное выполнять избирательную предвыборку инструкций для исполнения и данных для операций чтения и записи. При выявлении потокового обращения к памяти устройство может инициировать до 4 составных запросов предвыборки. Поддерживается работа с 16 потоками, что позволяет инициировать параллельную предвыборку до 4 Кб данных.

В ядре Intel Xeon Phi кэш-промах в L1 или L2 не приводит к блокировке работы всего ядра. При возникновении кэш-промаха в ходе выполнении операции чтения поток, инструкция которого вызвала промах, будет приостановлен до поступления данных. Все остальные потоки ядра при этом будут продолжать свое исполнение. Каждый из кэшей L1 и L2 могут ини-

циировать до 38 одновременных запросов к внешним данным (на чтение или запись). С учетом возможностей блока клиентской логики PCI Express, включающего контроллер прямого доступа к памяти и способного обслуживать одновременно до 128 запросов, общее возможное число запросов данных в сопроцессоре вычисляется как *38 * (число ядер) + 128*. Это позволяет активно использовать программную предвыборку данных для уменьшения потерь из-за кэш-промахов.

Кэш L2 является частью блока интерфейса кольцевой шины (CRI/L2), который также включает каталог тегов (TD). Распределенный между ядрами каталог тегов обеспечивает доставку запросов данных к клиентам кольцевой шины – другим ядрам или встроенным контроллерам памяти. Для контроля когерентности кэшей используется комбинация протоколов: MESI (Modified, Exclusive, Shared, Invalid) на каждом ядре и GOLS3 (Globally Owned Locally Shared) для распределенного каталога тегов. Распределенный каталог сопроцессора (TD) разделен на 64 части, каждая из которых отвечает за контроль глобального состояния когерентности части строк кэша.

Дополнительно необходимо отметить следующие особенности реализации кэшей L1 и L2:

- обращения к кэшу L1 можно выполнять в последовательные такты, между обращениями к кэшу L2 должен выдерживаться интервал в 1 такт (то есть к кэшу L2 можно обращаться в лучшем случае через такт);
- на каждом конкретном такте для L1 и L2 допускается выполнение либо чтения из кэша, либо записи в кэш, но не чтения и записи одновременно.

В сопроцессоре Intel Xeon Phi реализованы только две схемы взаимодействия между кэшем и основной памятью – отсутствие кэширования (uncacheable, UC) и отложенная запись (write-back, WB).

4.2. Поддержка виртуального адресного пространства

Описание работы оперативной памяти включает множество различных схем и параметров ее использования. Однако возможность использования той или иной аппаратной возможности зависит не только от наличия ее реализации, но и от свойств используемой операционной системы. На сопроцессоре Intel Xeon Phi исполняется специальная ОС, основанная на ядре Linux (kernel.org), в которую внесены минимальные дополнения для обеспечения совместимости. Далее мы будем в основном описывать только те возможности и параметры, которые поддерживаются операционной системой сопроцессора на момент написания раздела.

Сопроцессор Intel Xeon Phi поддерживает 32-битные физические адреса при работе в 32-битном режиме, 36-битные адреса при использовании технологии PAE (Physical Address Extension) в 32-битном режиме, 40-битные физические адреса при работе в 64-битном режиме. Операционная система сопроцессора поддерживает работу только в 64-битном режиме.

Процессам предоставляется линейное виртуальное адресное пространство (ВАП) и возможность использовать 64-битные адреса. Для поддержки ВАП используется стандартная схема архитектуры x86_64 – страничная адресация с 4 уровнями таблиц страниц.

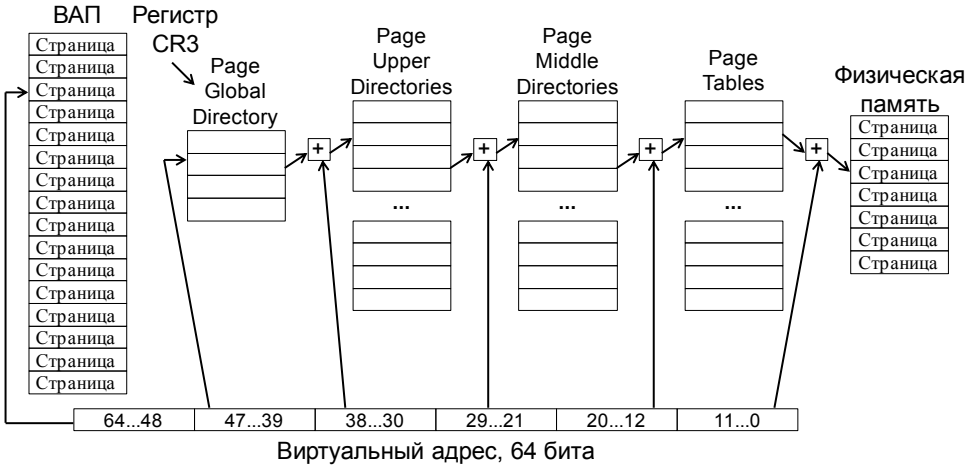


Рис. 5. Схема трансляции виртуальных адресов в физические

Поддерживаемые размеры страниц – 4 Кб и 2 Мб. Реализована поддержка запрета исполнения страницы (Execute Disable, NX; может использоваться для запрета исполнения инструкций, хранящихся вне региона кода, например, в стеке), но отсутствует признак глобальности страницы (Global Page bit; используется для страниц, отображенных в ВАП разных процессов на одни и те же физические страницы, например, части адресного пространства ядра).

Кэш дескрипторов страниц (translation look-aside buffer, TLB) имеет двухуровневую архитектуру. Каждое ядро содержит:

- L1 TLB для данных, содержащий 64 записи для страниц размером 4 Кб и 8 записей для страниц размером 2 Мб;
- L1 TLB для инструкций, содержащий 32 записи для страниц размером 4 Кб;
- универсальный L2 TLB, содержащий 64 записи для хранения дескрипторов страниц размером 2 Мб или записей таблицы страниц второго уровня (Page Middle Directory) для страниц размером 4 Кб.

Таблица 1 Характеристики кэшей TLB сопроцессора Intel Xeon Phi

Кэш	Размер страницы	Число записей	Степень ассоциативности	Соответствующий размер памяти
L1 TLB данных	4 Кб	64	4	256 Кб
	2 Мб	8	4	16 Мб
L1 TLB инструкций	4 Кб	32	4	128 Кб
L2 TLB	4 Кб, 2 Мб	64	4	128 Мб

Все TLB реализованы как 4-канальные множественно-ассоциативных кэши. Несколько выполняемых потоков могут использовать одни и те же записи TLB при условии совпадения у них значений регистров CR3, CR0.PG, CR4.PAE, CR4.PSE, EFER.LMA (фактически это означает, что данные потоки принадлежат одному процессу).

4.3. Доступ к оперативной памяти

Сопроцессор Intel Xeon Phi содержит 8 встроенных контроллеров памяти, каждый из которых обслуживают по два 32-битных канала GDDR5, обеспечивая суммарную производительность 5,5 GT/s (миллиардов пересылок в секунду) или 352 GB/s. Латентность доступа к памяти составляет более 300 тактов. Контроллеры памяти непосредственно подключены к кольцевой шине сопроцессора и осуществляют преобразование запросов на чтение/запись памяти в команды GDDR5 и планирование их исполнения с учетом физической организации и характеристик памяти для максимизации итоговой пропускной способности.

Компонент сопроцессора, реализующий клиентскую логику PCI Express (SBOX), также обеспечивает работу механизма прямого доступа к памяти (DMA). 8 независимых каналов DMA, работающих на той же частоте, что и ядра сопроцессора, могут выполнять следующие типы передачи данных:

- из GDDR5-памяти сопроцессора в оперативную память хоста;
- из оперативной памяти хоста в GDDR5-память сопроцессора;
- из GDDR5-памяти в GDDR5-память в пределах сопроцессора.

Выполнение операции передачи данных может быть запрошено как со стороны центрального процессора хоста, так и со стороны сопроцессора, при этом буфер передачи должен быть выделен на той стороне, которая инициирует передачу. При выполнении одной транзакции PCI Express может быть передано от 64 байт (1 строка кэша) до 256 байт данных.

5. Набор инструкций сопроцессора Intel® Xeon Phi

Сопроцессор Intel Xeon Phi совместим с архитектурой Intel64 за исключением расширений MMX, AVX и SSE, и дополнительно поддерживает собственный набор инструкций для работы с векторными данными. Основные свойства нового набора команд.

- Новый набор команд ориентирован в первую очередь на повышение производительности HPC-приложений (High Performance Computing). Поддерживаются операции с 32-битными целочисленными и вещественными числами, множество операций преобразования типов данных, часто используемых в высокопроизводительных приложениях, также поддерживаются арифметические операции с 64-битными вещественными числами и логические операции с 64-битными целыми.
- Наличие 32 512-битных векторных регистров помогает компенсировать латентность обращения к данным. Каждый вектор может обрабатываться как набор из 16 32-битных или 8 64-битных целочисленных или вещественных значений.
- Используются тернарные инструкции, в которых явно указываются два источника входных данных и получатель результата. У операции «умножение и сложение» (multiply-add, FMA) получатель также является третьим источником.
- Реализованы специальные регистры масок, использование которых позволяет осуществлять условное выполнение операций над элементами векторов, а также осуществлять векторизацию циклов, содержащих операции ветвления.
- Специальные инструкции (scatter/gather) позволяют упаковывать в векторы для обработки данные, произвольно расположенные в памяти, что позволяет выполнять векторизацию алгоритмов, использующих сложные структуры данных.

6. Литература

1. Reinders J. An Overview of Programming for Intel Xeon processors and Intel Xeon Phi coprocessors. [http://software.intel.com/en-us/blogs/2012/11/14/an-overview-of-programming-for-intel-xeon-processors-and-intel-xeon-phi].
2. Loc Q Nguyen et al. Intel Xeon Phi Coprocessor Developer's Quick Start Guide. [http://software.intel.com/en-us/articles/intel-xeon-phi-coprocessor-developers-quick-start-guide].
3. Pseudo-LRU. [http://en.wikipedia.org/wiki/Pseudo-LRU].

4. Intel Xeon Phi Coprocessor System Software Developers Guide. [http://software.intel.com/en-us/articles/intel-xeon-phi-coprocessor-system-software-developers-guide].
5. Rahman R. Intel Xeon Phi Core Micro-architecture [http://software.intel.com/en-us/articles/intel-xeon-phi-core-micro-architecture].