Девять секретов производительности процессоров Intel

12 Фев 2016 Илья Манусов

Новейшее функциональное расширение *AVX512 (Advanced VectorExtension 512-bit),* также известное под именем AVX3, состоит из девяти технологий, каждая из которых может опционально поддерживаться или не поддерживаться конкретным процессором. Ряд внедрений рассматриваемого семейства, связанных с операциями повышенной разрядности, уже применяется в сопроцессорах Intel Xeon Phi и опционально доступны в процессорах Xeon на основе микроархитектуры Skylake. Напомним, что в отличие от процессоров Xeon, сопроцессоры Xeon Phi предназначены для установки PCI Express слот.

I. Базовая функциональность AVX512

Принятое сокращение: *AVX512F (Foundation)*. Это базовый или минимальный набор выполняемых команд и программно-доступных ресурсов, необходимых для обработки 512-битных векторов. Поддержка AVX512F подразумевает расширение разрядности векторных регистров до 512 бит и увеличение количества этих регистров до 32. Для сравнения, функциональное расширение предыдущего поколения (AVX2), реализованное в процессорах Haswell, подразумевает использование 16 регистров разрядностью 256 бит. Традиционно, «старые» регистры являются частью «новых». В данном случае это означает, что 16 256-битных регистров YMM0–YMM15 отображаются на младшие 256-битные «половинки» 512-битных регистров ZMM0–ZMM15.

Как следует из несложных подсчетов, новое 512-битное операционное устройство способно обрабатывать 8 64-битных чисел двойной точности либо 16 32-битных чисел одинарной точности за одну векторную команду.

В базовый набор AVX512F также входит *предикатное выполнение* векторных операций. Это означает, что при обработке чисел, упакованных в 512-битном регистре, операция может быть выполнена или отменена, индивидуально для каждого числа. Например, при обработке 16 32-битных чисел одинарной точности, 16-битный предикат, содержащий все «единицы» обеспечит выполнение операции для всех чисел. Если все биты предиката нулевые, операция не выполняется. А установив, например два младших бита предиката, можно выполнить операцию для двух первых чисел, оставив остальные числа незатронутыми. Для хранения предикатов вводится 8 дополнительных 64-битных регистров K0–K7.

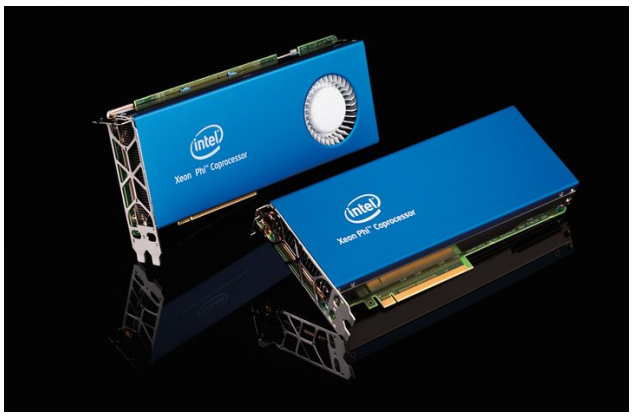

Рис 1. *Сопроцессор Xeon Phi, предназначенный для установки PCI Express слот*

II. Аппаратное выявление конфликтов

Принятое сокращение: *AVX512CD (Conflict Detection)*. Используется для эффективного обнаружения ситуаций, при которых заданные программные процедуры не могут быть выполнены параллельно в силу зависимости по данным. Вспомним, что при оптимизации программных циклов применяется метод, подразумевающий переход от последовательного к параллельному выполнению итераций цикла. Цикл «разворачивается» в линейную последовательность операций, затем эти операции выполняются параллельно с использованием векторных регистров. Такой прием допускается только в том случае, если между итерациями цикла нет «конфликтов» по данным. Иначе, если некоторая итерация использует данные, которые должна подготовить предыдущая итерация, их параллельное выполнение приведет к ошибке.

III. Опережающая загрузка данных

Принятое сокращение: *AVX512PF (Prefetch)*. Механизм опережающей загрузки усовершенствован с целью загрузки произвольно фрагментированных данных. Опережающая загрузка в простейшем виде, применяется в процессорах Intel еще со времен Pentium III. Она состоит в заблаговременном чтении операндов из оперативной памяти в кэш-память с помощью специальных команд (prefetch hints). Это минимизирует непроизводительные паузы в работе процессора, поскольку в момент затребования данных они уже загружены из памяти. Теперь эту операцию можно выполнить не только для одной ячейки памяти, но и для списка ячеек, адреса которых находятся в векторном регистре.

Рис 2. *Процессор Intel Xeon на основе микроархитектуры Skylake*

IV. Вычисление экспоненты и обратных величин

Принятое сокращение: *AVX512ER (Exponential and Reciprocal)*. Эта группа команд формирует аппроксимированные (приближенные) результаты для экспоненты, обратной величины и обратной величины квадратного корня. Относительная погрешность, в зависимости от типа команды составляет до 2 в степени минус 23 либо 2 в степени минус 28. Команды этой группы используются для эффективной поддержки ситуаций, в которых допустимо пожертвовать точностью ради производительности. Сразу оговоримся, под экспонентой здесь понимается возведение двойки (а не числа e) в заданную степень. Эта спорная терминологическая особенность на совести инженеров Intel.

V. Операции переменной разрядности

Принятое сокращение: *AVX512VL (Vector Length)*. Эта функциональность обеспечивает использование возможностей AVX512, в частности, описанных выше предикатов и 32 векторных регистров) для операндов, размер которых 128 и 256 бит.

VI. Обработка байтов и 16-битных операндов

Принятое сокращение: *AVX512BW (Byte and Word)*. Обеспечивает использование возможностей AVX512 для целочисленных операций разрядностью 8 и 16 бит.

VII. Обработка 32 и 64-битных операндов

Принятое сокращение: *AVX512DQ (Double word and Quad Word)*. Обеспечивает использование возможностей AVX512 для операций разрядностью 32 и 64 бита.

VIII. Совмещенное умножение-сложение для 52-битных операндов

Принятое сокращение: *AVX512IFMA (Integer Fused Multiply and Add)*. Перемножение целых чисел без знака, разрядностью 52 бита. В открытых документах не удалось найти объяснение использования столь нетипичного формата чисел. Вспомнив, что разрядность мантиссы для числа двойной точности — 52 бита, позволим себе предположить, что целью была возможность целочисленной (а потому быстрой) обработки мантиссы как самостоятельной величины.

IX. Операции с байтами в составе векторных регистров

Принятое сокращение: *AVX512VBM (Vector Byte Manipulation)*. Сюда входят инструкции для перестановки и избирательной пересылки байтовых операндов, расположенных в векторных регистрах.

Резюме

Как видно из вышеизложенного, расширение AVX512 подразумевает введение новых процессорных команд и архитектурных ресурсов. Поэтому, прибавку производительности получит только то программное обеспечение, которое написано и оптимизировано с учетом новой технологии.

Очевидно, для работы такой технологии требуется поддержка и со стороны операционной системы, поскольку увеличение количества и разрядности регистров приводит к увеличению количества информации, которую требуется сохранять и восстанавливать при переключении контекста процессора в многозадачной среде. Отметим, что компания Intel максимально унифицировала процесс сохранения-восстановления контекста.