CPU. MMX, hypher threading

[1]Кафедра информационных технологий и систем
Национальная металлургическая академия Украины

14 сентября 2011 г.

MMX (Multimedia Extensions — мультимедийные расширения) — коммерческое название дополнительного набора инструкций, выполняющих характерные для процессов кодирования/декодирования потоковых аудио/видео данных действия за одну машинную инструкцию. Впервые появился в процессорах Pentium MMX.

SISD (англ. Single Instruction, Single Data) или ОКОД (Одиночный поток Команд, Одиночный поток Данных) — архитектура компьютера, в которой один процессор выполняет один поток команд, оперируя одним потоком данных. Относится к фон-Неймановской архитектуре.

SISD компьютеры это обычные, ”традиционные” последовательные компьютеры, в которых в каждый момент времени выполняется лишь одна операция над одним элементом данных (числовым или каким-либо другим значением). Большинство персональных ЭВМ до последнего времени, например, попадает именно в эту категорию. Иногда сюда относят и некоторые типы векторных компьютеров, это зависит от того, что понимать под потоком данных.

SIMD (англ. single instruction, multiple data — одиночный поток команд, множественный поток данных, ОКМД) — принцип компьютерных вычислений, позволяющий обеспечить параллелизм на уровне данных.

SIMD-компьютеры состоят из одного командного процессора (управляющего модуля), называемого контроллером, и нескольких модулей обработки данных, называемых процессорными элементами. Управляющий модуль принимает, анализирует и выполняет команды. Если в команде встречаются данные, контроллер рассылает на все процессорные элементы команду, и эта команда выполняется на нескольких или на всех процессорных элементах. Каждый процессорный элемент имеет свою собственную память для хранения данных.

Одним из преимуществ данной архитектуры считается то, что в этом случае более эффективно реализована логика вычислений. До половины логических инструкций обычного процессора связано с управлением выполнением машинных команд, а остальная их часть относится к работе с внутренней памятью процессора и выполнению арифметических операций. В SIMD компьютере управление выполняется контроллером, а "арифметика" отдана процессорным элементам.

Фактически вся история развития компьютеров представляет собой непрерывную гонку между быстродействием центрального процессора и прочих систем - памяти и внешних устройств. Особенно это заметно в системах мультимедиа, где идет обработка звука и изображения, цифровое представление которых занимает большие объемы памяти. Для эффективной обработки звука и видео при относительно низкой пропускной способности системной магистрали (шины) все большее количество функций переносится в аппаратуру - модемы, видео- и звуковые адаптеры. Это вызывает их заметное удорожание в сравнении с общей стоимостью компьютера, что особенно неприятно в обстановке быстрого морального старения всей компьютерной аппаратуры.

Особенно данная проблема стала актуальна в начале 1990-х годов, когда ПК стал доступен широким массам пользователей и все активнее стал превращаться в средство развлечений. Первым процессором ощутившим нехватку ресурсов для мультимедийных приложений по тому времени стал Pentium.

На самом деле, неспособность ПК с процессором Pentium эффективно обрабатывать в реальном времени звук и видео без специальных карт происходит уже не столько от общего быстродействия процессора или шины, которые в большинстве случаев вполне достаточны, а от характера его набора команд обработки данных, известного под названием CISC. Этот набор, состоящий из относительно сложных арифметико-логических команд, ориентирован на типовые задачи обработки данных, без специальной "заточки"под особые приложения. Эта выгодная для большинства приложений, архитектура оказывается совершенно неэффективной при скоростной и специфической обработке больших массивов данных, поскольку сложная система команд используется на считанные проценты, а накладные

Технология MMX представляет собой компромиссное решение, объединяющее пути, используемые в компьютерах SPARC и Silicon Graphics (технология RISC - Reduced Instruction Set Computer, компьютер с упрощенным набором команд), а также в компьютерах с параллельной архитектурой (технология SIMD: Single Instruction, Multiple Data - одна команда, много данных): классический процессор Pentium (CISC) с добавлением ряда простых (RISC) команд параллельной обработки данных (SIMD).

Аббревиатура MMX происходит от выражения MultiMedia eXtension - расширение для мультимедиа, которое реализовано фирмой Intel в своей новой серии процессоров MMX с тактовой частотой 166 и более МГц. Исторически сложилось так, что почти любое новое решение в области персональных компьютеров широко рекламируется и преподносится как эпохальное, сулящее невиданный доселе расцвет компьютерным технологиям, однако все мы помним, сколько раз подобная шумиха оборачивалась весьма скромным реальным эффектом.

Процессор Pentium MMX отличается от "обычного"Pentium по шести основным пунктам:

1. добавлено 57 новых команд обработки данных;
2. увеличен в два раза объем внутреннего кэш (16 кб для команд и столько же - для данных);
3. увеличен объем буфера адресов перехода (Branch Target Buffer - BTB), используемого в системе предсказания переходов (Branch Prediction);
4. оптимизирована работа конвейера (Pipeline);
5. увеличено количество буферов записи (Write Buffers);
6. введено так называемое двойное электропитание процессора.

Набор из 57 новых команд и является основным отличием; остальные пять - не более, чем сопутствующие изменения. Хотя увеличенный объем кэш и внутренних буферов и оптимизированный конвейер несколько ускоряют работу любых приложений, однако основное увеличение производительности - до 60% - возможно только при использовании программ, правильно применяющих технологию MMX в обработке данных.

Как уже говорилось, в Pentium MMX добавлено 57 новых команд обработки данных и, соответственно - четыре новых типа данных. За одну операцию команда MMX обрабатывает 64-разрядное двоичное слово (так называемое квадраслово, или QWord). Новые типы данных образуются от упаковки в квадраслово обычных типов - байтов (по 8), слов (по 4) или двойных слов (по 2). Четвертый тип представляет собой само квадраслово.

Таким образом, одна элементарная MMX-операция имеет дело либо с одним квадрасловом, что похоже на обычную операцию большой разрядности, либо с двумя двойными словами, четырьмя словами или восемью байтами, причем выполнение происходит одновременно и каждый элемент данных обрабатывается независимо от других. Подобные групповые операции преобладают во время обработки изображения (группы точек) и звука (группы значений амплитуды).

Набор MMX-команд состоит из команд пересылки данных, упаковки/распаковки, сложения/вычитания, умножения, сдвига, сравнения и поразрядных логических. Команды упаковки и сложения/вычитания могут работать в двух режимах: обычном, когда переполнение разрядной сетки вызывает "заворачивание"(wraparound) значения результата, и специальном, когда оно приводит к ограничению (clipping) результата до минимально или максимально допустимого значения. Режим ограничения в терминологии Intel называется Saturation (насыщение) - в нем особенно удобно выполнять смешивание цветов изображение или амплитуд звуковых сигналов, поскольку при обычном переполнении результат не имеет никакого смысла.

Команда умножения представлена тремя видами: первые два выполняют попарное умножение четырех слов с выбором либо старшей, либо младшей части результата, а третий выполняет операцию вида ab + cd для каждой пары из четырех слов операндов, что очень удобно при вычислении математических рядов.

Команды сдвига реализуют логический и арифметический сдвиги своих операндов (арифметический сдвиг отличается от логического тем, что при сдвиге вправо освободившиеся разряды заполняются копией знакового разряда, а не нулями, отчего он пригоден для умножения/деления знаковых операндов на степени двойки). Логические поразрядные команды выполняют операции И (AND), ИЛИ (OR), Исключающее ИЛИ (XOR), а также комбинированную команду И с инверсией одного из операндов (AND NOT), удобную для реализации "обратного выбора"по битовой маске.

Команды сравнения работают несколько необычно по сравнению с общепринятой логикой: вместо установки признаков для последующих команд перехода они генерируют единичные битовые маски для тех операндов, которые удовлетворяют условию, и нулевые - для остальных операндов. Последующие логические поразрядные операции могут выделить, погасить или как-то иначе обработать отмеченные таким образом операнды, которые в этом случае могут представлять собой точки изображения или отсчеты звукового сигнала.

Для обработки данных и хранения промежуточных результатов в Pentium MMX используются восемь 64-разрядных регистров MM0..MM7, которые физически совмещены со стеком регистров математического сопроцессора. При выполнении любой из MMX-команд происходит установка "режима MMX"с отметкой этого в слове состояния сопроцессора (FPU Tag Word). С этого момента стек регистров сопроцессора рассматривается как набор MMX-регистров; завершает работу в режиме MMX команда EMMS (End MultiMedia State). С одной стороны, такая реализация позволила обеспечить нормальную работу приложений, использующих MMX, в многозадачных системах, не поддерживающих эту технологию, поскольку все подобные системы создают собственную копию

Так как MMX - достаточно узкоспециализированное расширение системы команд процессора, нельзя ожидать кардинального ускорения работы только от самого факта перехода на процессор MMX. Как уже было сказано, на приложениях общего характера, незнакомых с MMX, реальная производительность возрастает лишь на единицы процентов, хотя тесты могут показывать ее возрастание на 20-30% - это происходит из-за цикличности большинства тестов, когда большая часть цикла попадает в увеличенный внутренний кэш. При использовании "чистого"MMX-кода, удачно подходящего к специфике решаемой задачи, быстродействие переписанного участка может возрасти в 5-6 раз, однако это ускорение будет локальным и неизбежно компенсируется "типовыми"участками программы, поэтому

Расширение MMX включает в себя восемь 64-битных регистров общего пользования MM0—MM7. Для совместимости со способами сохранения состояния процессора в существующих ОС Intel была вынуждена объединить в программной модели процессора восемь регистров MMX с мантиссами восьми регистров FPU (Математический сопроцессор). Аппаратно это могут быть разные устройства, но с точки зрения программиста - это одни и те же регистры. Таким образом, нельзя одновременно пользоваться командами Математического сопроцессора и MMX.

Команды технологии MMX работают с 64-разрядными целочисленными данными, а также с данными, упакованными в группы (векторы) общей длиной 64 бита. Такие данные могут находиться в памяти или в восьми MMX-регистрах.

Команды технологии MMX работают со следующими типами данных:

- упакованные байты (восемь байтов в одном 64-разрядном регистре) (англ. packed byte);
- упакованные слова (четыре 16-разрядных слова в 64-разрядном регистре) (packed word);
- упакованные двойные слова (два 32-разрядных слова в 64-разрядном регистре) (packed doubleword);
- 64-разрядные слова (quadword).

Hyper-Threading Technology (HTT)) — это торговая марка компании Intel для реализации технологии "одновременной мультипоточности" (англ. Simultaneous multithreading) на микроархитектуре Pentium 4. Расширенная форма суперпоточности (англ. Super-threading), впервые появившаяся в процессорах Intel Xeon и позднее добавленная в процессоры Pentium 4.

Эта технология увеличивает производительность процессора при определённых рабочих нагрузках путём предоставления "полезной работы" (англ. useful work) исполнительным устройствам (англ. execution units), которые иначе будут бездействовать; к примеру, в случаях кэш-промаха. Процессоры Pentium 4 с включённым Hyper-threading операционная система определяет как два разных процессора вместо одного. В процессорах Core 2 Duo поддержка технологии Hyper-threading не была реализована. В процессорах Core i7 снова используется Hyper-threading, при этом каждое физическое ядро процессора определяется операционной системой как два логических. Так же эта технология присутствует в мобильных процессорах Core i3, Core i5 и в некоторых процессорах серии Atom.

Многопоточность — свойство платформы (например, операционной системы, VM и т. д.) или приложения, состоящее в том, что процесс, порождённый в операционной системе, может состоять из нескольких потоков, выполняющихся ”параллельно”, то есть без предписанного порядка во времени. При выполнении некоторых задач такое разделение может достичь более эффективного использования ресурсов вычислительной машины.

Такие потоки называют также потоками выполнения (от англ. thread of execution); иногда называют ”нитями” (буквальный перевод англ. thread) или неформально ”тредами”.
Сутью многопоточности является квазимногозадачность на уровне одного исполняемого процесса, то есть все потоки выполняются в адресном пространстве процесса. Кроме этого, все потоки процесса имеют не только общее адресное пространство, но и общие дескрипторы файлов.
Выполняющийся процесс имеет как минимум один (главный) поток.

Многопоточность (как доктрину программирования) не следует путать ни с многозадачностью, ни с многопроцессорностью, несмотря на то, что операционные системы, реализующие многозадачность, как правило реализуют и многопоточность.

К достоинствам многопоточности в программировании можно отнести следующее:

- Упрощение программы в некоторых случаях, за счет использования общего адресного пространства.
- Меньшие относительно процесса временные затраты на создание потока.
- Повышение производительности процесса за счет распараллеливания процессорных вычислений и операций ввода/вывода.

В процессорах с использованием этой технологии каждый физический процессор может хранить состояние сразу двух потоков, что для операционной системы выглядит как наличие двух логических процессоров (англ. Logical processor). Физически у каждого из логических процессоров есть свой набор регистров и контроллер прерываний (APIC), а остальные элементы процессора являются общими. Когда при исполнении потока одним из логических процессоров возникает пауза (в результате кэш-промаха, ошибки предсказания ветвлений, ожидания результата предыдущей инструкции), то управление передаётся потоку в другом логическом процессоре.

Таким образом пока один процесс ждёт, например данные из памяти, вычислительные ресурсы физического процессора используются для обработки другого процесса.

Были представлены следующие преимущества Hyper-threading: улучшенная поддержка многопоточного кода, позволяющая запускать потоки одновременно; улучшенная реакция и время отклика; увеличенное количество пользователей, которое может поддерживать сервер.
По словам Intel, первая реализация потребовала всего 5-процентного увеличения площади кристалла, но позволяла увеличить производительность на 15 — 30%.

Intel утверждает, что прибавка к скорости составляет 30% по сравнению с идентичным процессорами Pentium 4 без технологии ”Simultaneous multithreading”. Однако прибавка к производительности изменяется от приложения к приложению: некоторые программы вообще несколько замедляются при включённой технологии Hyper-threading. Это, в первую очередь, связано с ”системой повторения” (англ. replay) процессоров Pentium 4, занимающей необходимые вычислительные ресурсы, отчего и начинают ”голодать” другие потоки