Основные тенденции в архитектуре высокопроизводительных многоядерных процессоров.

Mikhail Isaev.

Введение.

Продолжая выполнять закон Мура и удваивать количество транзисторов каждые 24 месяца, полупроводниковая индустрия развивается, предлагая всё более и более производительные решения вычислительных проблем. Но если ещё совсем недавно развитие происходило в основном в направлении усложнения процессорного ядра и увеличении его производительности, то в последнее время стала намечаться тенденция к увеличению числа процессорных ядер на кристалле и ко всевозможному распараллеливанию, на уровне тредов, ядер или количества многоядерных процессоров в системе.

В наше время уже никого нельзя удивить четырёъядерными процессорами, которые уверенно обосновались в настольных компьютерах на работе и дома. Но уже сейчас некоторые кампании увеличили число ядер в одном микропроцессоре до 8 и более. Зачастую, такое увеличение требует неординарных технических и архитектурных решений и представляют собой достаточно большой скачёк вперёд в сравнении с предыдущими разработками.

Фирм, которые имеют рабочие или инженерные образцы процессоров с количеством ядер 8 и более не так много, поэтому все эти процессоры можно просто перечислить. Это AMD Opteron серии Magny-Cours, Intel Nehalem-EX, IBM Power7, IBM Cell, Sun UltraSPARC T3, а также ещё не вышедший AMD Buldozer, Intel Larrabee и не прошедший далее стадии разработки Sun Rock. Рассмотрим основные технические и технологические решения, предлагаемые данными процессорами, и попытаемся подвести черту над тенденциями, которые наблюдаются в разработке процессоров для современных высокопроизводительных решений.

AMD Opteron.

Начнём с решения кампании AMD. На данный момент топовый серверный процессор Opteron серии Magny-Cours состоит из 12 ядер, которые расположены на двух кристаллах в одной упаковке. Для создания этого процессора AMD использовало процессор серии Istanbul, состоящий из 6 ядер, объединённых на базе общего L3 кэша.

Технологически 12-ядерный процессор Opteron представляет собой упаковку LGA, имеющую два шестиядерных кристалла частотой до 2,3ГГц на одной упаковке с 1974 выводами. Каждый кристалл имеет 6 ядер, суммарный объём кэша кодов и данных L1 составляет 128Кб (64Кб данных и 64Кб инструкций), общий кэш инструкций и данных L2 составляет 512Кб на ядро, общий кэш L3, объединяющий 6 ядер, составляет 6Мб на 6 ядер, или 1Мб в пересчёте на ядро. Каждый кристалл имеет двухканальный контроллер DDR3 памяти, работающий на частоте 1,8ГГц, а также межпроцессорные линки, работающие со скоростью 6,4ГТ/с и связывающие два кристалла внутри упаковки и внешние кристаллы в системе до 4 сокетов. Кристаллы выпускаются по технологии 45nm SOI (кремний на изоляторе).

Особенностью данного процессора является появившийся в шестиядерном процессоре Istanbul кэш-справочник по памяти, который занимает по 1Мб кэша третьего уровня на каждом кристалле. Справочник является неполным, то есть, в нём есть состояния только части кэш-строк. В том случае, если в справочнике нет состояния какой-либо кэш-строки, рассылается широковещательный снуп-запрос (работает старая схема когерентности, применяющаяся в процессорах Opteron до использования справочника). Если же состояние необходимой кэш-строки найдено в кэше, то снуп-запрос рассылается непосредственному владельцу кэш-строки, в зависимости от состояния строки. Таким образом, данное нововведение заметно сокращает снуп-трафик.

Ещё одна особенность заключается в том, что иерархия кэшей в процессоре эксклюзивная, то есть, кэши всех трёх уровней хранят абсолютно разные данные. Это увеличивает суммарный объём кэшей, но несёт ряд неудобств, связанных с тем, что при подкачке данных в кэш более высокого уровня требуется сначала протравить данные из этого кэша, а также с тем, что при снуп-опросе требуется опросить кэши всех трёх уровней для поиска данных. Кэш третьего уровня является общим для 6 ядер, поэтому ядра объединены по данным кэшем L3.

Последняя особенность связана с использованием протокола когерентности MOESI. В данной версии он был расширен состоянием S1, которое говорит о том, что данные имеются в состояние Shared (то есть прочитанном из памяти и не модифицированном) только у одного ядра, у которого они могут быть взяты [1]. Это позволяет уменьшить количество обращений в память за данными, если таковые имеются у одного из ядер, так как доступ к чужим данным, находящимся в другом процессоре, быстрее и менее ресурсозатратен доступа в память.

AMD Buldozer.

Эта архитектура является новой разработкой AMD, поэтому точных данных по ней нет. Технологически, это будет процессор. Выпускаемый по технологии 45nm SOI в упаковке с 1974 выводами (та же технология и упаковка, что и у современных серверных процессоров AMD Opteron).

Исходя из предварительной схемы кристалла, формально это будет 8-ядерный процессор с неоднородной структурой L1-кэша данных (только для целочисленных данных), общими модулями предвыборки инструкций, декодировщиком, FPU и L2 кэшем на 2 ядра, а также общим L3 кэшем на 4 двухъядерных модуля. Хотя AMD и называет такой дизайн восьмиядерным, всё же, на мой взгляд, уместнее его называть четырёхъядерным процессором с двумя кластерами целочисленной арифметики и одним кластером арифметики с плавающей запятой на ядро. В такой формулировке данный процессор не имеет интерес с точки зрения организации микропроцессора с 8 и более ядрами.

Intel Nehalem-EX (Becton)

Совсем недавно в серверном сегменте вышло принципиально новое для Intel решение — процессор на архитектуре Nehalem-EX (Becton). Топовая модель представляет собой кристалл, работающий с частотой до 2,26ГГц с 8 ядрами, поддерживающими до двух ветвей в рамках технологии HyperThreading, размещённых на упаковке с 1567 выводами LGA, суммарный объём кэша кодов и данных L1 составляет 64Кб (32Кб данных и 32Кб инструкций), общий кэш инструкций и данных L2 составляет 256Кб на ядро, общий кэш L3, объединяющий 8 ядер, составляет 24Мб на 8 ядер, или 3Мб в пересчёте на ядро, контроллер DDR3 памяти имеет 4 канала и работает на частоте 1333МГц, 4 межпроцессорных линка работают со скоростью 6,4ГТ/с и связывают внешние кристаллы в системе до 32 сокетов. Кристаллы выускаются по технологии 45nm.

Особенностью иерархии кэшей процессора Nehalem-EX является то, что все они инклюзивные. Это несколько уменьшает их суммарный объём, так как в L2 кэше поддерживается копия состояний L1 кэша, а в L3 кэше в свою очередь поддерживается копия состояний L2 кэша, но это делает поиск кэша нов более удобным, так как они собраны в одном месте, более того, это место является общим для всех восьми ядер на кристалле.

Новый процессор имеет достаточно широкие и быстрые интерфейсы, так суммарно пропускная способность памяти составляет 50ГБ/с, а суммарная пропускная способность межпроцессорных линков составляет порядка 100ГБ/с, что является залогом расширяемости системы до 32 сокетов. Второй вклад в расширяемость системы вносит

новый протокол когерентности — MESIF. Появившееся F-состояние (Forward) несёт примерно то же смысл, что и S1-состояние в расширенном MOESI протоколе когерентности у AMD, то есть позволяет передавать кэш-строку из одного ядра в другое а не брать её из памяти, что сокращает трафик с памятью [2].

Для поддержания аппаратной когерентности на кристалле расположен чип-коммутатор, который является программируемым и объединяет все внутренние и внешние запросчики в системе. Поддержка когерентности, видимо, тоже осуществляется посредством этого коммутатора.

Разрабатывая большой общий L3 кэш, Intel столкнулась с рядом проблем при его проектировании. Основной сложностью заключалось сделать такой большой массив памяти (24Мб) достаточно быстрым и наделить его 8 интерфейсами с ядрами. Поэтому кэш был разбит на 8 частей по 3Мб, которые территориально принадлежат к одному из ядер, и эти участки имеют неравномерный доступ. Так, к своим участкам кэш-памяти ядра имеют достаточно быстрый доступ, а участки связаны между собой кольцевой шиной. Кольцевая шина представляет собой 4 кольца по 32 байта, то есть 1024 бит, и работает в оба направление, то есть по 512 бит на направление и обеспечивает пропускную способность порядка 250ГБ/с.

IBM Power7.

Фирма IBM всегда славилась сложными и бескомпромиссными решениями с абсолютным лидерством в производительности среди своих конкурентов. Новейший процессор Power7 не стал исключением. Он представляет собой кристалл, работающий на частоте до 4,25ГГц, с 8 ядрами, поддерживающими до четырёх ветвей в рамках технологии multi-threading, суммарный объём кэша кодов и данных L1 составляет 64Кб (32Кб данных и 32Кб инструкций), общий кэш инструкций и данных L2 составляет 256Кб на ядро, общий кэш L3, объединяющий 8 ядер, составляет 32Мб на 8 ядер, или 4Мб в пересчёте на ядро, два контроллера DDR3 памяти по 4 канала, 4 кристалла размещаются на одном мульти-чиповом модуле, которые, в свою очередь, объединяются в систему из восьми модулей, то есть, 32 процессоров по 8 ядер каждый. Кристаллы выпускаются по технологии 45nm SOI.

При разработке Power7 фирмой IBM было использовано достаточно много новейших технологий и инженерных решений. Так, к примеру, IBM отказалась от параллельного интерфейса с памятью и использует проприетарные буферы-памяти собственной разрабоки, контроллеры памяти связываются с буферами памяти посредством высокоскоростных LVDS-интерфейсов частотой 6,4ГГц, а уже буферы памяти связываются

непосредственно с планками оперативной памяти DDR3. Колоссальный объём общей L3 памяти организован не с помощью стандартных элементов — статической памяти SRAM, а с помощью динамической памяти eDRAM, новейшей технологии, также проприетарной и доступной в настоящее время лишь IBM, что позволило повысить плотность памяти при хорошем времени доступа к ней.

Основой расширяемости процессора стали очень высокоскоростные интерфейсы. Так, интерфейс с памятью работает с общей производительностью 100ГБ/с, а межпроцессорные линки суммарно выдают производительность 360ГБ/с.

С архитектурной точки зрения наиболее любопытны структура L3 кэша и протокол когерентности. L3 кэш разбит на 8 участков, обеспечивая каждому ядру быстрый доступ к своему участку. При этом L3 кэш выполняет много разнообразных задач, являясь и victim-cache для кэшей высшего уровня, и более объёмным кэшем нижнего уровня для своего ядра, и средством связи ядер между собой, и victim-cache для остальных участков L3 кэша. Всё дело в механизме передачи данных между ядрами, который основывается на создании копий одной и той же кэш-строки и копирования её в разные участки L3 кэша [3]. Частично из-за этого, а частично из-за организации системы (8 ядер на кристалле, 4 кристалла в упаковке, до 8 упаковок в системе), протокол когерентности усложнился, в сравнении с предыдущим процессором в этой серии Power6. Кэш-строка поддерживает 13 состояний, при этом нет статической точки сериализации в системе, то есть каждая строка в кэше может являться динамической точкой сериализации для соответствующей кэш-строки. Так же вся когерентность разбита на регионы: первый регион — это кристалл из 8 ядер, второй регион — модуль из 4 кристаллов, третий регион — 8 модулей по 4 кристалла. Широковещательные запросы рассылаются только в пределах региона, то есть поддерживается ступенчатый протокол когерентности. Это позволяет сократить общий поток снуп-запросов и ответов и оптимизировать передачу данных между ядрами и из памяти.

IBM Cell.

Выпускаясь уже давно, является, пожалуй, самым неоднозначным из всех многоядерных процессоров. Дело в том, что его ядра, по сути, не являются равнозначными. Он имеет одно или два ядра общего назначения архитектуры Power с двумя тредами, на котором может раскручиваться операционная система, и 8 вспомогательных вычислительных ядер с достаточно простой архитектурой. Первоначально процессор выпускался на технологии 90nm, однако для последней ревизии IBM использовала 45nm SOI. Частота процессора

достигает 5,3ГГц. Основным новшеством являлось то, что 8 вспомогательных ядер не были охвачены общим механизмом когерентности и не имели L2 кэшей. Вместо них использовалась локальная память, 256Кб в последней ревизии, которая доступна по DMA, и кольцевая шина, связывающая все ядра в системе. Кольцевая шина состоит из 4 колец по 16 байт и является двунаправленной. В качестве памяти используется двухканальный контроллер XDR RAM с пропускной способностью 25ГБ/с, используется межпроцессорная шина Flex I/O 32 ГБ/с на ввод, 44,8 ГБ/с на вывод.

Процессоры фирмы Sun.

Являясь самыми простыми среди всех представленных, процессоры фирмы Sun наименее интересны с точки зрения объединения процессорных ядер. 8 ядер на кристалле появились ещё в процессоре Sun UltraSPARC T1, они были объединены 3Мб общего L2 кэша. Лишь в последней версии Sun UltraSPARC T3 появились межпроцессорные линки и возможность объединять процессоры в системы из нескольких сокетов. Число ядер увеличилось до 16, а кэш вырос до 6Мб, но в целом организация системы осталась неизменной и достаточно простой. Ядра объединяются на базе общего L2 кэша с помощью коммутатора, 2 когерентных контроллера и 6 когерентных линков соединяют до 6 процессоров в систему. Каждый кристалл имеет 4 канала памяти DDR3. Дизайн ядра разработан для выпуска на технологии 45nm.

Sun UltraSPARC Rock является примером неудачного процессора. Предполагалось, что это будет 16-ядерный процессор, в котором будет 4 сегмента по 4 ядра, 16 ядер должен был объединять коммутатор, обеспечивая доступ к общему L2 кэшу, но процессор получался довольно-таки медлительным и имел слишком большое тепловыделение, поэтому проект был закрыт в связи с финансовыми сложностями компании. При его разработке использовалась новая технология памяти транзакций.

Intel Larrabee.

Новейший и ещё не законченный проект интела нельзя рассматривать, как многоядерный центральный процессор. По сути, это нечто среднее между процессором и видеокартой, спецвычислитель, который хотя и поддерживает инструкции x86 и выглядит как многоядерный процессор, но на самом деле имеет доступ по шине PCI-E и не обладает самостоятельностью, подобно видеокарте, то есть, должен управляться центральным процессором. Пока готовы только инженерные образцы, предварительно заявлено о выпуске по нормам 45nm и 32 nm, до 48 ядер на кристалл, до 2,5ГГц частотой.

Основной особенностью нового чипа является аппаратная поддержка когерентности, которой нет ни в видеокартах, ни в процессоре IBM Cell, который структурно отдалённо похож на Larrabee. В качестве механизма когерентности используется оптимизированный протокол MESI. Дополнительные состояния позволяют брать Shared-копию данных из соседнего ядра, не забирая их из памяти. Ядра, как и в процессоре Cell, объединены кольцевой шиной. Каждое ядро имеет 256Кб кэша L2 с быстрым доступом в составе общего L2, части L2 связываются кольцевой шиной. Шина шириной 1024 бит является двунаправленной и состоит из 4 колец, по 2 в каждую сторону. L2 кэш поддерживает механизмы прямого управления, в том числе прямую подкачку данных.

Общие выводы.

Наращивание ядер становится повсеместным трендом, а соединение процессоров в систему из 4-32 кристаллов с помощью межпроцессорных LVDS-интерфейсов практически становится стандартом де-факто. Причём скорость этих интерфейсов достаточно большая, в основном используется порядка 6,4ГТ/с. При этом очень остро встают вопросы когерентности в системе. На данный момент почти все решения предлагают аппаратную поддержку когерентности. Но чтобы обеспечивать заявленную производительность в реальных приложениях, приходится достаточно сильно расширять интерфейсы, обеспечивая один контроллер памяти на два-три ядра, то есть, для создания восьмиядерного процессора гармоничным кажется использование 4 каналов памяти, и точно не менее трёх, иначе интерфейс с памятью становится узким местом. При выстраивании механизма когерентности происходит усложнение протоколов. Когерентность делиться на регионы и запросы рассылаются ступенчато, чтобы сократить служебный когерентный трафик между процессорами. В купе с иерархией регионов используются директории для адресации когерентных запросов, вместо рассылки широковещательных. Протоколы когерентности дорабатываются и дополняются состояниями, которые позволяют меньше обращаться за данными в память и забирать Shared-копию данных из соседнего процессора при её наличии. Ядра обычно имеют какую-то общую структуру данных, L3 кэш, хотя доступ к разным его частям обычно несимметричен. Разделяется локальная часть, доступ к которой достаточно быстр, и остальные части, с более медленным доступом. Так как появляется кэш нового уровня, то L2 кэш обычно стараются уменьшить, часто пользуются лишь 256Кб кэша, общих для данных и кодов команд, при том, L1 составляет обычно по 32Кб для данных и для команд, суммарно 64Кб. Объём L3 обычно

составляется из размеров 1-2Мб на ядро. Больший удельный объём используется лишь фирмами, которые способны позволить себе уникальную технологию для изготовления такого большого объёма памяти.

Кольцевые шины на данный момент используются в процессоре IBM Cell, но там не поддерживается аппаратная когерентность, и шина используется только для обмена данными, Intel Larrabee, причём на ней поддерживается аппаратная когерентность, по шине передаются адресная часть, данные и служебная информация, и в Intel Nehalem-EX, но там кольцевая шина лишь связывает части L3 кэша по данным между собой, остальным. Должно быть, заведует коммутатор или коммутаторы, точной информации не раскрывается. Также кольцевые шины применялись в прошлом в графических чипах AMD с серии R600, но впоследствии от них было решено отказаться. Они также использовались только для передачи данных, когерентность между ядрами не поддерживалась.

Литература.

1. Pat Conway, Nathan Kalyanasundharam, Gregg Donley, Kevin Lepak, Bill Hughes, “Cache Hierarchy and Memory Subsystem of the AMD Opteron Processor”, IEEE Micro, vol. 30, no. 2, 2010, pp. 16-29.
2. Herbert H. J. Hum et al, “Forward state for use in cache coherency in a multiprocessor system”, US Patent 6922756, 2005.
3. Kalla, R., Sinharoy, B., Starke, W.J., Floyd, M., Power7: IBM's Next-Generation Server Processor, IEEE Micro, vol. 30, no. 2, 2010, pp. 7-15.