Исследование топологий сетей на кристалле многоядерных процессоров с архитектурой «Эльбрус»

А. Кожин[1], Е. Кожин[2], Д. Шпагилев[3]

УДК 004.318 | ВАК 05.13.05

Сети на кристалле, играющие центральную роль при проектировании современных многоядерных процессоров, связывают между собой все основные компоненты системы. Их топологии определяют масштабируемость пропускной способности и времени доступа в подсистему памяти в зависимости от числа процессорных ядер. В статье приведены результаты исследований сетей на кристалле, имеющих топологии 2d mesh и 2d torus-mesh, для 8-, 12- и 16-ядерных процессоров с архитектурой «Эльбрус» с учетом особенностей их физического проектирования.

Современное развитие универсальных высокопроизводительных процессоров во многом основано на наращивании числа процессорных ядер. Большинство разработчиков, за исключением фирмы AMD, перешедшей к объединению нескольких кристаллов небольшой площади на общей подложке [1], продолжают проектировать многоядерные монолитные кристаллы, содержащие 28 [2] и более [3] процессорных ядер. На кристалле также размещается все большее количество банков кэш-памяти последнего уровня, контроллеров оперативной памяти (6–8 контроллеров на процессор), устройств ввода-вывода. Таким образом, современный высокопроизводительный процессор – это сложная распределенная система из десятков-сотен устройств, которые должны быть соединены в общую систему с учетом требований по задержкам и пропускной способности соединений между устройствами.

Многоядерные процессоры с архитектурой «Эльбрус» четвертого и пятого поколений объединяли восемь процессорных ядер и восемь банков общей кэш-памяти третьего уровня посредством распределенной двунаправленной кольцевой шины [4, 5]. При этом контроллеры оперативной памяти и устройств ввода-вывода подключались через центральный коммутатор [6].

При разработке новых поколений многоядерных процессоров «Эльбрус» требовалось нарастить количество процессорных ядер в системе и отказаться от использования центрального коммутатора для подключения периферийных устройств и контроллеров памяти. Соответственно, система связей устройств внутри процессора подлежала переработке. Основой решения стало построение распределенной сети на кристалле, в которой устройства через сетевые адаптеры подключаются к общей распределенной коммутационной среде произвольной топологии [7]. Это наиболее модульный и масштабируемый по числу устройств подход. Он позволяет, изменяя только параметры сети и сетевые адаптеры, использовать ранее разработанные устройства, такие как ядра, банки общей кэш-памяти последнего уровня, контроллеры доступа к памяти и вводу-выводу, для построения серии процессоров различных конфигураций. При этом основными характеристиками, задающими производительность коммутационной среды, являются топология сети, адресация в ней и технология маршрутизации пакетов.

Данная работа посвящена анализу топологий сети на кристалле для построения 8-, 12- и 16-ядерных процессоров с архитектурой «Эльбрус» шестого поколения и исследованию влияния рассмотренных вариантов на время доставки пакетов и эффективную пропускную способность сети.

СТРУКТУРА СЕТИ НА КРИСТАЛЛЕ

Все обмены между устройствами сетей на кристалле процессоров с архитектурой «Эльбрус» шестого поколения выполняются с помощью сообщений в соответствии с проприетарным системным протоколом. Сообщения имеют несколько типов: первичные запросы, когерентные снуп-запросы, ответы с данными, ответы без данных, подтверждения. Каждое сообщение передается в виде пакета определенного размера, для передачи между узлами сети пакетов разных типов выделяются отдельные физические и виртуальные каналы.

Исследуемые топологии сети на кристалле имеют «плиточную» структуру [7]: процессорное ядро вместе

---

[1] АО «МЦСТ», начальник сектора, Alexey.S.Kozhin@mcst.ru.

[2] АО «МЦСТ», ведущий инженер, Evgeny.S.Kozhin@mcst.ru.

[3] АО «МЦСТ», ведущий инженер, Danil.I.Shpagilev@mcst.ru.

с ближайшим банком кэш-памяти третьего уровня (L3-кэша) объединяются в «тайл» (tile) посредством централизованного коммутатора (рис. 1). Он имеет два порта подключения абонентов (ядра и банка L3-кэша, но в общем случае возможно подключение и других устройств) и четыре сетевых порта: North, South, Upper, Down. Коммутатор основан на разработке, выполненной для процессоров с архитектурой «Эльбрус» четвертого и пятого поколений, в которых процессорные ядра и банки общей кэш-памяти третьего уровня объединялись в двунаправленное кольцо [8]. Он имеет буферизующую структуру, поддерживает виртуальные каналы и Quality-of-Service (QoS). Пакеты из нескольких посылок передаются по принципу wormhole. Коммутатор является узлом сети на кристалле и соединяется с соседними узлами посредством двунаправленных каналов связи (линков, links), подключаемых к сетевым портам.

Коммутатор выполняет функции сетевого, канального и физического уровней стека системного протокола. Каждый пакет при помещении в сеть расширяется адресом назначения, который позволяет однозначно определить устройство-получателя этого пакета. Адрес назначения вычисляется исходя из карты маршрутизации и топологии сети. Сформированный пакет передается между узлами сети до тех пор, пока не достигнет получателя. В отличие от обычных компьютерных сетей, которые могут включать множество узлов и используют динамическую маршрутизацию, сети на кристалле имеют гораздо меньший размер и обычно основываются на статической маршрутизации пакетов. При статической маршрутизации прохождение пакета по сети зависит только от взаимного расположения устройства-отправителя и устройства-получателя и полностью определяется выбором топологии сети, которая не может перестраиваться во время работы.

ИССЛЕДОВАННЫЕ ТОПОЛОГИИ

Архитектура разработанного коммутатора позволяет реализовать достаточно большое разнообразие топологий сети на кристалле. Общепринятой топологией многоядерных процессоров (10 и больше ядер) считается 2d mesh со статической маршрутизацией X-Y или Y-X, когда каждому узлу присваивается декартова координата (X, Y). Пакет из узла-отправителя сначала перемещается вдоль одной оси, пока она не совпадет с координатой узла-получателя, затем – вдоль другой. Доказано, что такая маршрутизация позволяет не допустить возникновения взаимных блокировок в сети (в отличие от кольцевой топологии, в которой требуется введение дополнительных виртуальных каналов) [7]. Топология 2d mesh хорошо масштабируется вплоть до 64–256 ядер, имеет высокую пропускную способность и меньшую по сравнению с кольцом задержку доставки, но требует сбалансированности по обеим координатам – число узлов по оси X и по оси Y должно примерно совпадать. В этом случае будет максимальным параметр сети bisectional width – число линков, пересекающих «разрез» сети на две равные по числу узлов части в ее самом узком месте. Именно эта характеристика определяет достижимую пропускную способность сети [7].

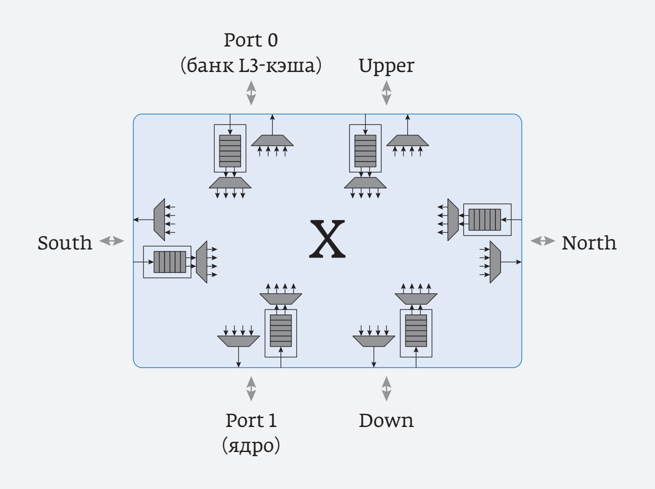

Рис. 1. Схема коммутатора сети на кристалле

На практике использование топологии 2d mesh при построении сети на кристалле может быть неэффективным решением, если ее конфигурация оказывается несбалансированной из-за ряда физических ограничений. При проектировании процессоров шестого поколения с архитектурой «Эльбрус» физические ограничения на топологию всего кристалла и особенности построения системы синхронизации привели к необходимости размещать процессорные ядра только в два ряда. С учетом этого требования сеть на кристалле с топологией 2d mesh для 8-, 12- и 16-ядерного процессоров имеет размеры 2 × 4, 2 × 6 и 2 × 8 соответственно, ее структура представлена на рис. 2а. Bisectional width для всех трех конфигураций будет равна 2 и не увеличивается с ростом числа узлов. В то же время сбалансированная 2d mesh сеть должна иметь размеры 2 × 4, 3 × 4, 4 × 4, а ее параметр bisectional width масштабируется вместе с ростом числа ядер благодаря большей связности сети. Также стоит отметить, что диаметр (максимальное расстояние между двумя узлами) несбалансированной 2d mesh сети больше, чем в сбалансированной.

Повысить bisectional width сети можно за счет увеличения ее связности, однако большинство более сложных топологий требует слишком больших накладных расходов (fat tree, полносвязная сеть) или использует слишком длинные линки при 2d-представлении (2d torus, hypercube) [7]. С учетом этих критериев, а также других аппаратных особенностей реализации, была выбрана

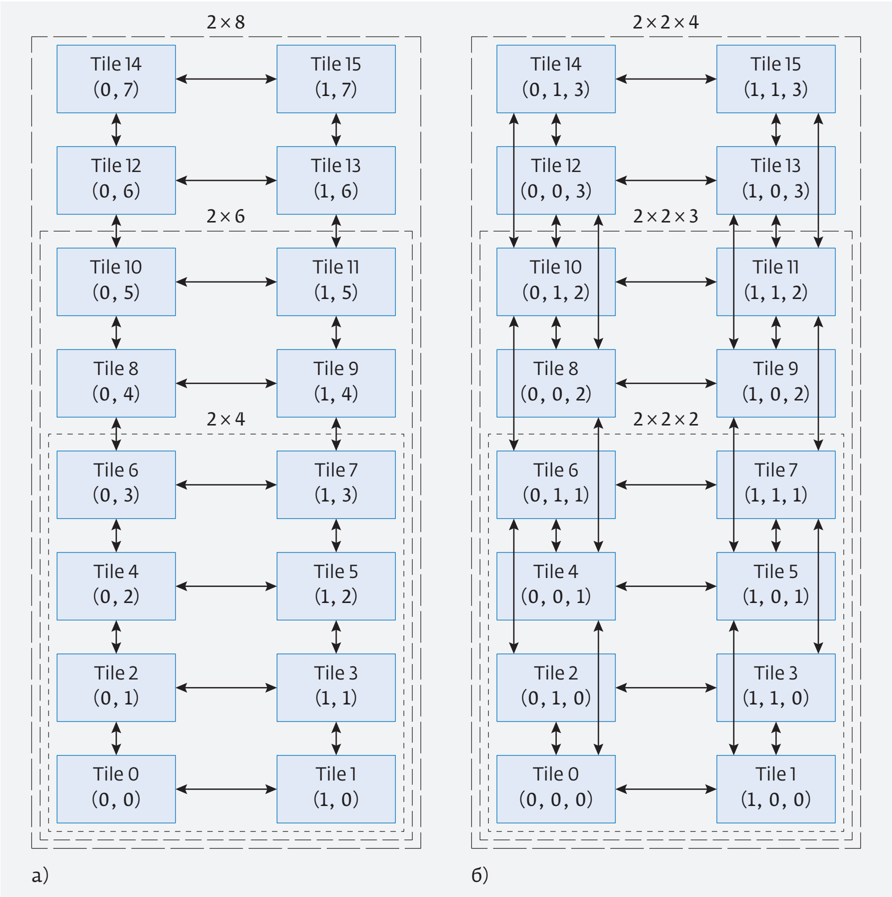

Рис. 2. Схемы исследуемых сетей на кристалле: а – сети с топологией 2d mesh; б – сети с топологией 2d torus-mesh

топология 2d torus-mesh, в которой стороны 2d mesh сети замыкаются линками вдоль одной из осей. Статическую маршрутизацию в сети на кристалле с топологией 2d torus-mesh удобно рассматривать по трем осям X-Y-Z, хотя сама сеть и реализуется на плоскости. Каждому узлу присваивается координата (X, Y, Z), пакет из узла-отправителя сначала перемещается вдоль оси X, пока она не совпадет с координатой X узла-получателя, затем вдоль оси Y и, наконец, вдоль оси Z. Как и в случае статической маршрутизации X-Y в 2d mesh сети, такая маршрутизация не допускает взаимных блокировок пакетов. Стоит отметить, что использование маршрутизации только по двум осям X-Y в 2d torus-mesh сети может привести к взаимной блокировке пакетов при перемещении по оси, вдоль которой замыкается 2d mesh, и требует введения дополнительных виртуальных каналов, поэтому маршрутизация X-Y-Z обладает преимуществом. Схематично сети на кристалле с топологией 2d torus-mesh для 8-, 12- и 16-ядерных процессоров представлены на рис. 2б и имеют размеры 2×2×2, 2×2×3 и 2×2×4 соответственно. Для всех трех конфигураций параметр bisectional width составляет 4.

В рамках данной работы были исследованы сети на кристалле с топологиями 2d mesh и 2d torus-mesh для 8-, 12- и 16-ядерных процессоров. Схематично представленные на рис. 2 сети были построены с учетом задержек в линиях связей на базе Verilog-моделей коммутаторов, используемых в процессорах семейства «Эльбрус». Разработанное тестовое окружение позволяет создавать поток пакетов с произвольным темпом от абонентов сети и равномерным распределением по адресатам, а также измерять задержки доставки пакетов и эффективную пропускную способность сети на кристалле. В следующих разделах приведены результаты этих измерений.

ЗАДЕРЖКИ В СЕТИ

Первая часть исследования касается задержек доставки пакетов в сети на кристалле. Необходимо отметить, что в исследуемых сетях учитываются все физические характеристики технологического процесса 16 нм, необходимые для получения максимально точных результатов, а также наличие пересинхронизаторов пакетов в интерфейсах процессорных ядер. Приведенные ниже результаты учитывают не только блокировки пакетов в сети, но и внутренние задержки коммутаторов, задержки дополнительных регистровых станций для длинных линков, задержки пересинхронизаторов.

На рис. 3 приведен график зависимости времени прохождения пакетов после ввода в сеть. По горизонтальной оси указана величина нагрузки в сети – темп, с которым новые пакеты помещаются в сеть. Все абоненты формируют пакеты с одинаковым темпом от 0,1 (один пакет за 10 процессорных тактов) до 1 (каждый такт). По вертикальной оси приведена средняя задержка пакета в процессорных тактах от момента помещения в сеть до момента выдачи устройству-получателю.

При минимальной нагрузке 0,1 (1 пакет за 10 тактов от каждого абонента) время прохождения пакета определяется исключительно структурой задержек в сети и соответствует среднему времени прохождения одиночного пакета – от 12 до 15 тактов в зависимости от топологии. Дальнейший рост нагрузки приводит к возникновению блокировок при пересылке пакетов между узлами и дополнительной задержки из-за этих блокировок. Чем больше абонентов в сети и чем меньше ее связность, тем быстрее

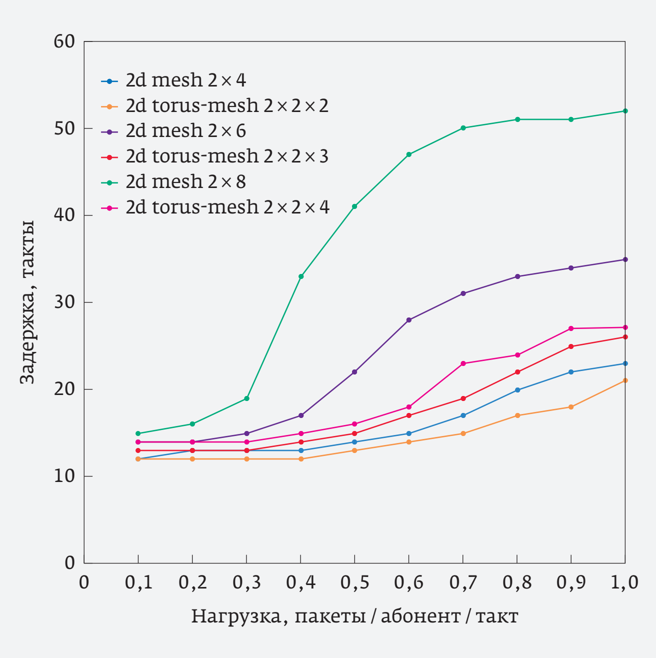

Рис. 3. График зависимости времени прохождения пакетов после помещения в сеть от нагрузки

нарастает среднее время прохождения пакета между абонентами. Так, для сети 2d mesh 2 × 8 наблюдается резкий рост задержек уже при нагрузке 0,4, которые достигают 52 тактов при максимальной нагрузке. Для сети 2d mesh 2 × 6 средняя величина задержек удваивается при нагрузке 0,6 и достигает 35 тактов при нагрузке 1. Сети с топологиями 2d mesh 2 × 4 и 2d torus-mesh демонстрируют плавный рост задержек, которые не превышают 27 тактов.

Блокировки внутри сети могут вызвать блокировку выдачи новых пакетов в сеть, так как коммутаторы буферизуют только ограниченное число пакетов. При такой нагрузке время нахождения пакета в сети перестает увеличиваться, что и показывает график на рис. 3, но увеличивается время с момента планируемой посылки пакета до его выдачи получателю. Этот график представлен на рис. 4.

На этом графике все топологии демонстрируют резкий рост общих задержек при достижении нагрузки насыщения, превышающей эффективную пропускную способность сети. Для сети 2d mesh 2 × 8 насыщение достигается при нагрузке 0,3, для сети 2d mesh 2 × 6 – при нагрузке 0,5. Сеть 2d torus-mesh 2 × 2 × 4 достигает насыщения при нагрузке 0,6, остальные топологии – при нагрузках в районе 0,7. Стоит отметить, что вследствие накопления пакетов в устройствах-отправителях из-за блокировок по выдаче в сеть, крутизна графиков зависит от общего числа пакетов, которое используется при моделировании. Чем больше требуется передать пакетов, тем круче будут графики.

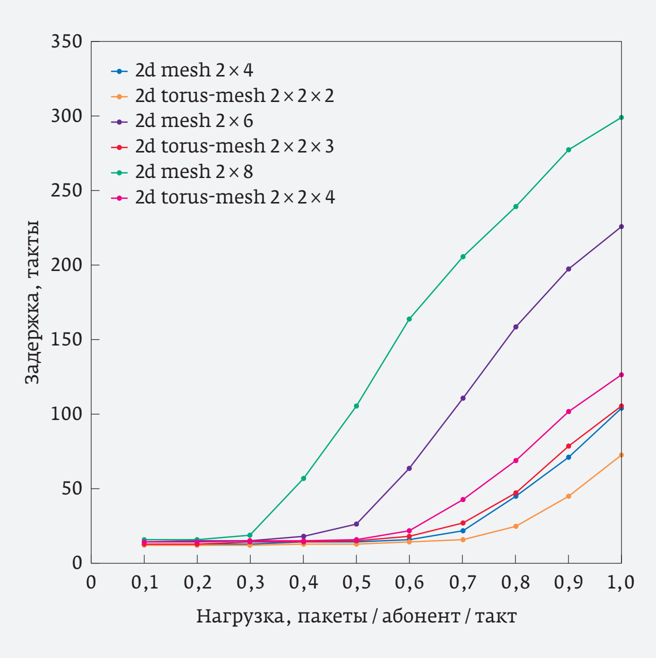

Рис. 4. График зависимости общей задержки пакетов в сети от нагрузки

ПРОПУСКНАЯ СПОСОБНОСТЬ СЕТИ

Для исследования масштабируемости пропускной способности сети на кристалле варьировалось число абонентов, отправляющих пакеты с максимальным возможным темпом и случайным распределением по всем абонентам-получателям. По горизонтальной оси графика на рис. 5 отмечено число абонентов-отправителей, по вертикальной оси указана пропускная способность сети в пакетах за такт.

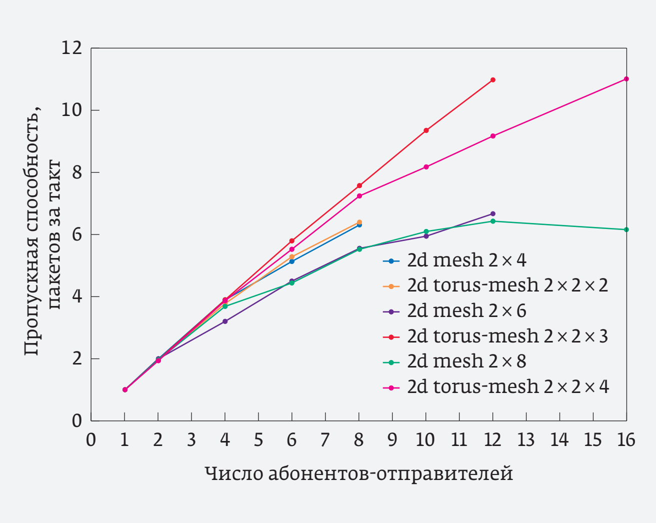

Рис. 5. График зависимости пропускной способности сети от числа абонентов-отправителей

Обе топологии сети на кристалле 8-ядерного процессора демонстрируют очень близкие графики масштабируемости пропускной способности с достижением максимального значения 6,3–6,4 пакетов за такт, что соответствует 80% от пиковой пропускной способности. Данный результат говорит об отсутствии необходимости в использовании топологии 2d torus-mesh с большей связностью при построении 8-ядерного процессора. Графики масштабируемости для большего числа процессорных ядер демонстрируют преимущество топологии 2d torus-mesh перед топологией 2d mesh. Максимальная пропускная способность сети 2d torus-mesh достигает примерно 11 пакетов за такт и составляет 90 и 70% от пиковой для 12- и 16-ядерного процессоров соответственно. Максимальная пропускная способность сети 2d mesh остается на уровне 6,1–6,7 пакетов за такт и составляет 56 и 38% от пиковой для 12- и 16-ядерного процессоров соответственно. Полученные результаты позволяют сделать вывод, что для 16-ядерного высокопроизводительного процессора топология 2d torus-mesh обладает безусловным преимуществом, а при проектировании 12-ядерного процессора выбор топологии должен основываться в большей степени на требованиях к мощности и площади кристалла, так как реализация 2d torus-mesh требует больших аппаратных затрат.

* * *

Таким образом, в статье были рассмотрены подходы к проектированию сетей на кристалле для многоядерных процессоров с архитектурой «Эльбрус» шестого поколения. Исследовались характеристики и масштабируемость сетей с топологиями 2d mesh и 2d torus-mesh, учитывающими особенности и ограничения физического проектирования топологии кристалла. Модели сетей были построены на базе Verilog-описания реальных коммутаторов и учитывали все необходимые физические характеристики технологического процесса. Результаты исследования показали, что для построения 8- и 12-ядерных процессоров подходящей топологией сети на кристалле является 2d mesh, в то время как для высокопроизводительного 16-ядерного процессора необходимо использовать топологию 2d torus-mesh, обеспечивающую большую пропускную способность.

ЛИТЕРАТУРА

1. **Suggs D., Subramony M., Bouvier D.** The AMD "Zen 2" Processor // IEEE Micro. 2020. V. 40. № 2. PP. 45–52.
2. **Tam S. M.** et al. Skylake-SP: A 14nm 28-core xeon® processor // 2018 IEEE International Solid-State Circuits Conference (ISSCC). – IEEE, 2018. PP. 34–36.
3. **Aingaran K.** et al. M7: Oracle's next-generation Sparc processor // IEEE Micro. 2015. V. 35. № 2. PP. 36–45.
4. **Kostenko V. O.** et al. Elbrus-8C: The Latest Yield from MCST and MIPT Collaboration // 2015 International Conference on Engineering and Telecommunication (EnT). – IEEE, 2015. PP. 67–68.
5. **Kozhin A. S.** et al. The 5th generation 28nm 8-core VLIW Elbrus-8C processor architecture // 2016 International Conference on Engineering and Telecommunication (EnT). – IEEE, 2016. PP. 86–90.
6. **Альфонсо Д. М., Деменко Р. В., Кожин А. С.** и др. Микроархитектура восьмиядерного универсального микропроцессора «Эльбрус-8С» // Вопросы радиоэлектроники. 2016. № 3. С. 6–13.
7. **Jerger N. E., Peh L.-S.** On-chip networks. Synthesis Lectures on Computer Architecture. 2009. 141 p.
8. **Кожин А. С., Сахин Ю. Х.** Коммутация соединений процессорных ядер с общим кэшем третьего уровня микропроцессора «Эльбрус-4С+» // Вопросы радиоэлектроники. 2013. № 3. С. 5–14.