Tofu: интерконнект для К компьютера

•Yuuichirou Ajima •Tomohiro Inoue •Shinya Hiramoto •Toshiyuki Shimizu
FUJITSU Sci. Tech. J., Vol. 48, No. 3, pp. 280–285 (July2012)),
перевод© ООО «Модуль-Проекты», http://www.mdl.ru, ссылки обязательны.

Интегрированный тор (Tofu, Torus fusion) является интерконнектом для массивных параллельных компьютеров, и он был разработан для построения К компьютера, который соединяет между собой более 80 000 узлов. Интерконнект Tofu обеспечивает высокую масштабируемость более чем 100 000 узлов, высокую производительность, высокую надежность и высокую доступность. Топологией сети является высокомасштабируемый шестимерный (сеточный) тор. Пропускная способность составляет 5 GB/s в каждом направлении. Каждый узел может осуществлять обмен в четырех направлениях одновременно. Трехмерная схема ранжированных карт улучшает доступность системы, а барьерный интерфейс Tofu (TBI, Tofu barrier interface) обрабатывает коллективный обмен данными с маленькой латентностью. Сетевые интерфейсы и маршрутизатор интерконнекта Tofu интегрированы в недавно разработанную микросхему, называемую контроллером интерконнекта (ICC, InterConnect Controller). Данная статья содержит обзор и характеристики микросхемы ICC, сеть шестимерного тора, высокопроизводительное и высоконадежное функционирование обмена данными и TBI.

1. Введение

Интегрированный тор (Tofu) является интерконнектом, разработанным для получения масштабируемости 100 000 узлов, что на два порядка выше, чем у существующих параллельных компьютеров, использующих непрямые связи. Он был разработан для построения К компьютера [note]), который соединяет 88 128 узлов. Для получения высокой производительности, высокой надежности и высокой доступности суперкомпьютера с беспрецедентно масштабным и массивной параллельностью были разработаны многие технологии для интерконнекта Tofu.

Этот документ излагает и описывает характеристики специализированной микросхемы, сетевые и коммуникационные функции интерконнекта Tofu. Вначале дается беглый обзор специализированной микросхемы, называемой контроллером интерконнекта (ICC, InterConnect Controller). Затем представляется сеть шестимерного тора. Потом мы знакомим с высокопроизводительными, высоконадежными функциями обмена данными. И, наконец, объясняется отличительная технология, называемая барьерным интерфесом Tofu (TBI, Tofu barrier interface).

2. Контроллер интерконнекта (ICC)

ICC является микросхемой, которая реализует интерконнект Tofu и соединяет процессоры SPARC64 соединением точка-точка. Микросхема ICC состоит из сетевого коммутатора Tofu (TNR, Tofu network router), четырех сетевых интерфейсов Tofu (TNIs, Tofu network interfaces), TBI и интерфейса PCI Express (Рисунок 1). TNR перемещает пакеты интерконнекта Tofu, TNIs являются интерфейсами для процессора, которые передают и принимают пакеты в/из сети и TBI обрабатывает коллективный обмен данными. PCI Express соединяет с внешними платами ввода/вывода и используется исключительно в узлах ввода/вывода. TNR производится с 10-портовыми связями Tofu и микросхема ICC использует связи (общим числом до 10) Tofu для связи с микросхемами ICC интегрированными в другие узлы. Характеристики ICC приводятся в Таблице 1.

note)
“K computer” - английское название, которое RIKEN использовал для суперкомпьютера в данном проекте начиная с июля 2010. “K” пришло из японского слова “Kei,” которое обозначает 10пета или 10 в 16й степени.

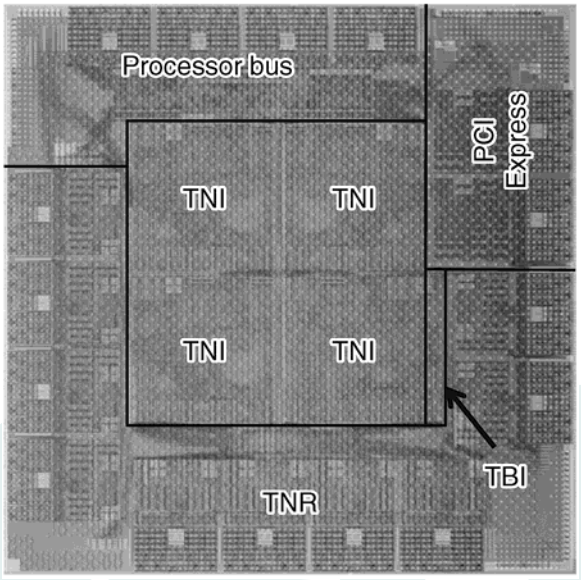

Рисунок 1
Микросхема ICC

Таблица 1
Характеристики ICC

Элемент	Описание
Кол-во одновременных обменов данными	4 соединения на передачу + 4 соединения на прием
Рабочая частота	312.5MHz
Пропускная способность коммутации	100GB/s
Скорость соединения	5GB/s в двух направлениях
Кол-во портов	10
Технологический процесс	65-nm CMOS
Размер кристалла	18.2mm x 18.1mm
Кол-во логических вентилей	48 миллионов вентилей
Кол-во ячеек SRAM	12 миллионов бит
Различные шины ввода/вывода Соединение Tofu Шина процессора PCI Express	 6.25Gb/s, 80lanes 6.25Gb/s, 32lanes 5Gb/s, 16lanes

3. Сеть шестимерного тора

3.1 Сетевая конфигурация

Позиционирование в сети шестимерного тора задается шестимерными координатами X, Y, Z, A, B и C. Оси X- и Y- являются координатными осями, которые соединяют стойки и длины осей X- и Y- соответствуют масштабу системы. Оси Z- и B- соединяют системные платы. Ось Z- имеет узел ввода/вывода в координате 0 и вычислительные узлы в координате 1 и выше. Оси B- соединяют три системные платы в кольцевую конфигурацию для гарантированной надежности (избыточности). Оси A- и C- являются координатными осями длиной 2 которые соединяют процессоры на каждой системной плате.

Полная сетевая топология является структурой с группами трехмерных ABC торов с размерами 2x3x2 соединяемыми трехмерными торами XYZ. Рисунок 2 представляет эту модель, демонстрирующую эту топологию.

3.2 Высокая масштабируемость

С использованием интерконнекта Tofu, число узлов может увеличиваться путем простого соединения кабелями и достигается высокая масштабируемость для более чем 100 000 узлов. Сеть шестимерного тора, согласно классификации, является сетью, называемой “непосредственной сетью” которая не использует внешних коммутаторов и характеризуется достаточно высокой степенью постоянства отношения среднего количества аппаратных средств на узел независимо от масштабов системы. Среднее количество аппаратных средств на узел, необходимых для интерконнекта Tofu включает только микросхему ICC и примерно 2.2 кабеля.

3.3 Порядок маршрутизации с расширенными размерностями

Пакеты направляются вдоль координатных осей в следующем порядке B, C, A, X, Y, Z, A, C и B. Первые оси маршрутизации ABC- могут иметь адресатами до 12 получателей, которые могут описываться для каждой команды передачи. Чтобы выбрать маршрут для исключения неисправности библиотека обмена данными использует пути направлений осей ABC-. Система уведомляет библиотеку обмена данными о местоположении сбойного узла в начале работы.

3.4 Ранжирование карт трехмерных торов

Для более легкой оптимизации коммуникационных шаблонов, использующих ближайшие соседние связи, интерконнект Tofu запасает одно-/двух-/трех- мерные пространства торов определяемого пользователем размера для представления пользователю. Для каждого процесса исполняемых пользовательских программ создаются различные значения рангов. Местоположение каждого процесса в пространстве описанного пользователем тора идентифицируется значением ранга. Когда определен трехмерный тор, система формирует три пространства, используя комбинации осей XYZ и одну из осей ABC. Система также назначает значение ранга для обеспечения выравнивания схемы "одним движением" в каждом пространстве. Рисунок 3 показывает пример присваивания значения ранга для случая, где трехмерный тор имеет размеры 8x12x6, определяемый пользователем.

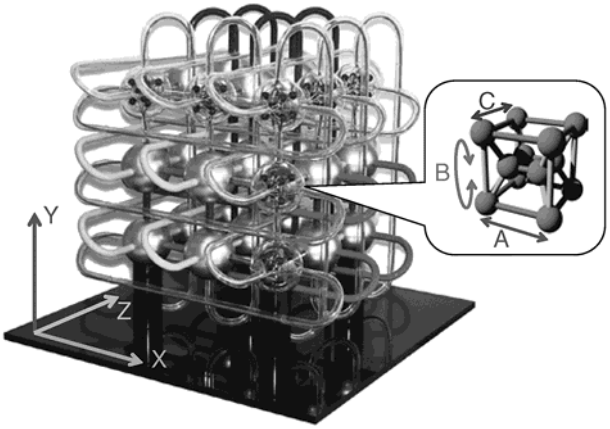

Рисунок 2
Модель топологии 6D тора

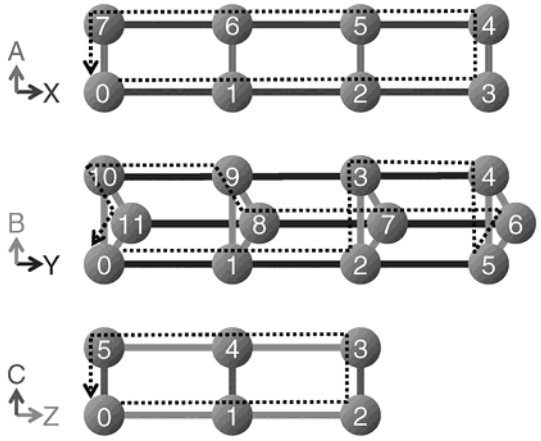

Рисунок 3
Пример ранжирования картрирования 3D тора

3.5 Высокая доступность

Даже при замене деталей (вышедших из строя системных плат) в процессе технического обслуживания, интерконнект Tofu позволяет непрерывное выполнение операций других системных плат, тем самым обеспечивая высокую доступность системы. Рисунок 4 демонстрирует пример присвоения значения ранга для обхода координаты, находящейся в процессе технического обслуживания или установки. В частности, ось B- используется в ранжировании картрирования трехмерного тора. Ось B- является кольцом из трех узлов и пространство, содержащее оси B- допускает быстрые схемы для обхода отдельного узла.[1)]

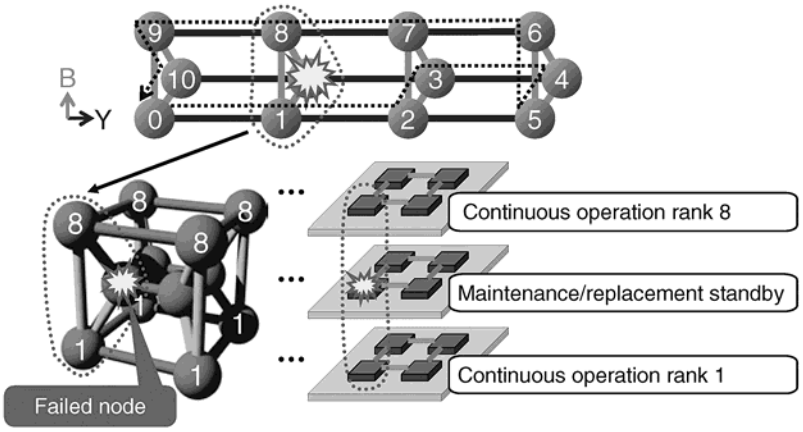

Рисунок 4
Присваивание значения ранга при замене вышедшей из строя системной платы
в процессе работ по тех.поддержке

4. Высокопроизводительный, высоконадежный обмен данными

4.1 RDMA обмен данными

Интерконнект Tofu поддерживает функции обмена данными на основе удаленного прямого доступа к памяти (RDMA, remote direct memory access). RDMA is a communication which accesses the specified memory address on the destination node and imposes a low reception processing load on the destination node. Интерконнект Tofu поддерживает механизм трансляции виртуальных адресов в физические, что делает возможным RDMA-доступ к пользовательской памяти при обмене данными. Механизм трансляции адресов включает в себя функции защиты областей памяти, а также поиск в кэше трансляции и таблице адресов трансляции.

4.2 Низколатентная, высокоэффективная пересылка

Интерконнект Tofu достигает сразу и низколатентную пересылку пакетов со значением около 0.1мкс на интервал связи (hop) виртуальной сквозной коммуникации, и высокоэффективную передачу данных с соотношением к теоретическому значению пропускной способности 90% или выше для пакетов с длиной до 2 KB. При виртуальной сквозной коммутации следующий интервал связи передачи стартует, когда получен заголовок пакета, что минимизирует латентность на интервал связи вне зависимости от длины пакета.

4.3 Одновременный обмен данными в четырех направлениях.

TNI интерконнекта Tofu выполняет прием и передачу одновременно. Четыре TNI работают независимо и каждый узел способен осуществлять параллельно передачу в четырех направлениях и прием от четырех источников.

Библиотека обмена данными назначает различные TNI для множественных асинхронных пересылок в зависимости от получателя. При синхронной пересылке библиотека пересылает данные через различные пути параллельно путем расщепления и назначения данных множеству TNI. Коллективный обмен данными использует множество TNI для достижения конвейерной пересылки в виртуальной топологии с ветвлением как у деревьев.

4.4 Виртуальные каналы

В интерконнекте Tofu задействовано четыре виртуальных канала: два для исключения тупиков маршрутизации и два для исключения тупиков запросов и ответов. Каждый порт получателя использует буфер размером 8KB на виртуальный канал, или 32KB в сумме, а деградация пропускной способности при перегрузке уменьшается посредством недавно разработанного алгоритма планирования виртуального канала.[2),3)]

4.5 Повторная передача уровня соединения

Соединение Tofu использует повторную передачу уровня соединения для коррекции битовых ошибок в соединении высокоскоростной пересылки для каждого интервала связи. По сравнению с ретрансляцией из конца в конец TCP/IP и InfiniBand, повторная передача уровня соединения значительно снижает деградацию производительности, вызываемую битовыми ошибками. Каждый передающий порт соединения Tofu имеет буфер передачи размером 8KB для передачи уровня соединения.

4.6 Разработка высокой надежности

SRAM и все пути сигналов в ICC защищены кодом коррекции ошибок и устойчивы к незначительным ошибкам, вызванным вторичными космическими лучами и тому подобными факторами. Сигналы для всех элементов управления, за исключением отладки (debugging), защищены битами четности, а для важных управляющих сигналов обнаружение неисправностей дополнено аппаратом мониторинга состояния, который определяет неправильные переходы состояний.

5. Барьерный интерфейс Tofu (TBI, Tofu barrier interface)

TBI является аппаратным модулем, который выполняет обработку коллективного обмена данными Barrier, Broadcast, Reduce и AllReduce в соответствующих узлах вместо программного обеспечения. Это выделяет гибкость, которая позволяет осуществлять многие алгоритмы в добавление к низкой латентности. TBI имеет восемь барьерных каналов и способен параллельно выполнять барьерную синхронизацию. Один канал используется для планирования синхронизации, а остальные семь используются библиотекой соединения.

5.1 Низколатентная обработка

Рисунок 5 показывает разницу между групповой обработкой обмена данными с использованием программного обеспечения и аппаратных средств (TBI). При обработке групповых обменов данными программным обеспечением и принимаемые и передаваемые данные проходят через основную память, что влечет за собой высокую латентность. Обработка обмена данными с использованием TBI не требует доступа к основной памяти и, таким образом, достигает низкой латентности.

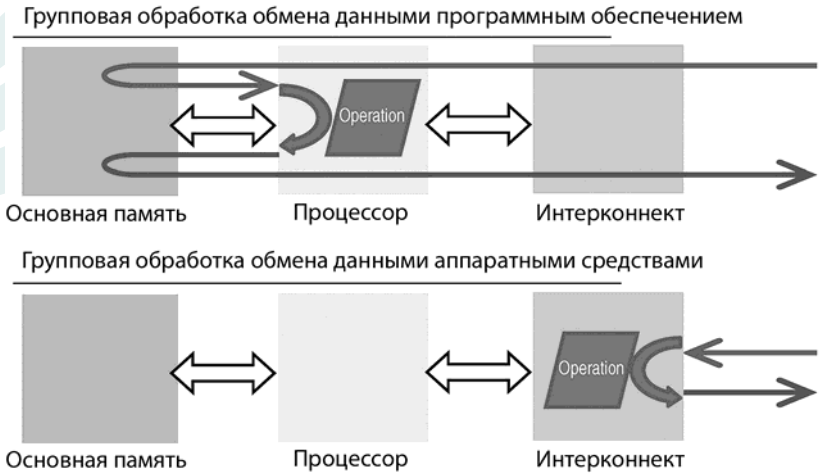

Рисунок 5
Разница между групповой обработкой обмена данными программным обеспечением и аппаратными средствами

5.2 Алгоритм обмена данными

Каждый узел снабжен 64 барьерами для приема, обработки и передачи. При использовании X барьеров, может быть реализован алгоритм обмена данными, который требует X раз для приема/передачи в каждом узле. Библиотека обмена данными использует соответствующий алгоритм обмена данными в зависимости от приложения. Например, алгоритм рекурсивного удвоения для N узлов предлагает низкую латентность, поскольку латентность пропорциональна log2N, однако требует большого количества барьеров, т.к. в каждом узле имеет место log2N раз приемов/передач. Алгоритм двойного кольца имеет высокую латентность, пропорциональную N, однако число приемов/передач равно только двум. В алгоритме обхода дерева латентность примерно удваивается по сравнению с рекурсивным удвоением, а количество приемов/передач равно пяти, что обеспечивает хороший баланс между низкой латентностью и экономией на потреблении барьеров.

5.3 Типы сокращения операций

Типами сокращения операций, поддерживаемыми TBI являются AND, OR, XOR, MAX и SUM для 64-битных целых чисел и SUM для чисел с плавающей запятой. Для получения одинакового результата с низкой задержкой вне зависимости от порядка операций, SUM с плавающей точкой использует оригинальный арифметический метод, в котором промежуточный результат представляется двумя 160-битными числами с плавающей точкой. Длина сообщения, поддерживаемая TBI составляет один элемент (скалярные данные).

5.4 Неустойчивость ОС (операционной системы)

Групповая обработка обмена данными программным обеспечением подвержена воздействию неустойчивости ОС. Неустойчивость ОС является флуктуацией в обработке между вычислительными процессами при параллельных вычислениях, которая вызывается прерываниями в вычислительных процессах из-за обработки переключений к процессу-демону и тому подобному. Типичное время прерывания длится от нескольких десятков микросекунд до нескольких миллисекунд, но при групповом обмене данными многие узлы ждут данные и задержка в обработке распространяется и влияет на многие узлы, что делает падение производительности более серьезным. В противоположность к этому, групповая обработка обмена данными аппаратными средствами имеет преимущество невосприимчивости к неустойчивости ОС.

6. Заключение

В статье представлен обзор и описание характеристик дизайна, сетевых и коммуникационных функций интерконнекта Tofu, который характеризуется масштабируемостью 100 000 узлов. Мы намерены и впредь повышать и развивать интерконнект Tofu, как интерконнект, который сочетает высокую масштабируемость, высокую производительность, высокую надежность и высокую доступность. Эти свойства становятся еще более важными для будущих exascale суперкомпьютеров.

Ссылки

1) Y. Ajima et al.: Tofu: A 6D Mesh/Torus interconnect for Exascale Computers. IEEE Computer, Vol. 42, No. 11, pp. 36–40 (2009).
2) Y. Ajima et al.: The Tofu Interconnect. The 19th Annual Symposium on High-Performance Interconnects, pp. 87–94 (2011).
3) Y. Ajima et al.: The Tofu Interconnect. IEEE Micro, Vol. 32, Issue 1, pp. 21–31 (2012).