Обзор системы К компьютер

•Hiroyuki Miyazaki •Yoshihiro Kusano •Naoki Shinjou •Fumiyoshi Shoji •Mitsuo Yokokawa •Tadashi Watanabe
FUJITSU Sci. Tech. J., Vol. 48, No. 3, pp. 255–265 (July2012),

RIKEN и Fujitsu совместно сотрудничали в разработке K компьютера, имея целью начало совместного использования осенью 2012 года в рамках инициативы инфраструктуры высокопроизводительных вычислений (HPCI), возглавляемой японским Министерством образования, культуры, спорта науки и технологии (MEXT). Так как K компьютер содержит более 80 000 вычислительных узлов, его построение с наиболее низким энергопотреблением и высокой надежностью было важным с точки зрения работоспособности. Эта статья описывает систему K компьютера и мероприятия, предпринятые для снижения энергопотребления и достижения высокой надежности и высокой доступности. Она также представляет результаты реализации этих мер.

1. Введение

Fujitsu активно разрабатывает и предлагает современные суперкомпьютеры уже более 30 лет с момента разработки в FACOM 230-75 APU -первого японского суперкомпьютера- в 1977году (рисунок 1). В результате проделанной работы, были развиты собственные аппаратные средства, включающие оригинальные процессоры, а также программное обеспечение и на этом пути наращивается техническая экспертиза в суперкомпьютинге.

Общий итог этой технической экспертизы был применен к разработке массивно параллельной вычислительной систему - K компьютера[note)] -который занял первое место в списке наиболее эффективных суперкомпьютеров в мире (2011).

K компьютер был разработан совместными усилиями RIKEN и Fujitsu как часть инициативы инфраструктуры высокопроизводительных вычислений (HPCI) осуществляемой японским Министерством образования, культуры, спорта науки и технологии (MEXT). Как следует из названия на японском языке "Кей", одной из целей этого проекта было достижение производительности 10^{16} операций с плавающей точкой в секунду (10PFLOPS). К тому же, K компьютер был разработан не только для достижения пиковой производительности в тестах, но и для обеспечения высокой эффективной производительности в приложениях, используемых в реальных исследованиях. Кроме того, чтобы обеспечить установку и работу всей системы в одном месте, нужно было снизить энергопотребление и обеспечить уровень надежности, который мог бы гарантировать общее функционирование масштабной системы.

Исходя из этого были установлены четыре целевые установки.

- Высокопроизводительные CPU для научных вычислений
- Новая архитектура интерконнекта для массивно параллельных вычислений
- Низкое энергопотребление
- Высокая надежность и доступность

CPU и архитектура интерконнекта описываются в деталях в других статьях этого специального выпуска[2),3)].

В данной статье мы представляем обзор компьютерной системы K, мероприятий, предпринятых для снижения энергопотребления и достижения высокой надежности и высокой доступности на системном уровне K компьютера, а также представляем результаты реализации этих мер.

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.

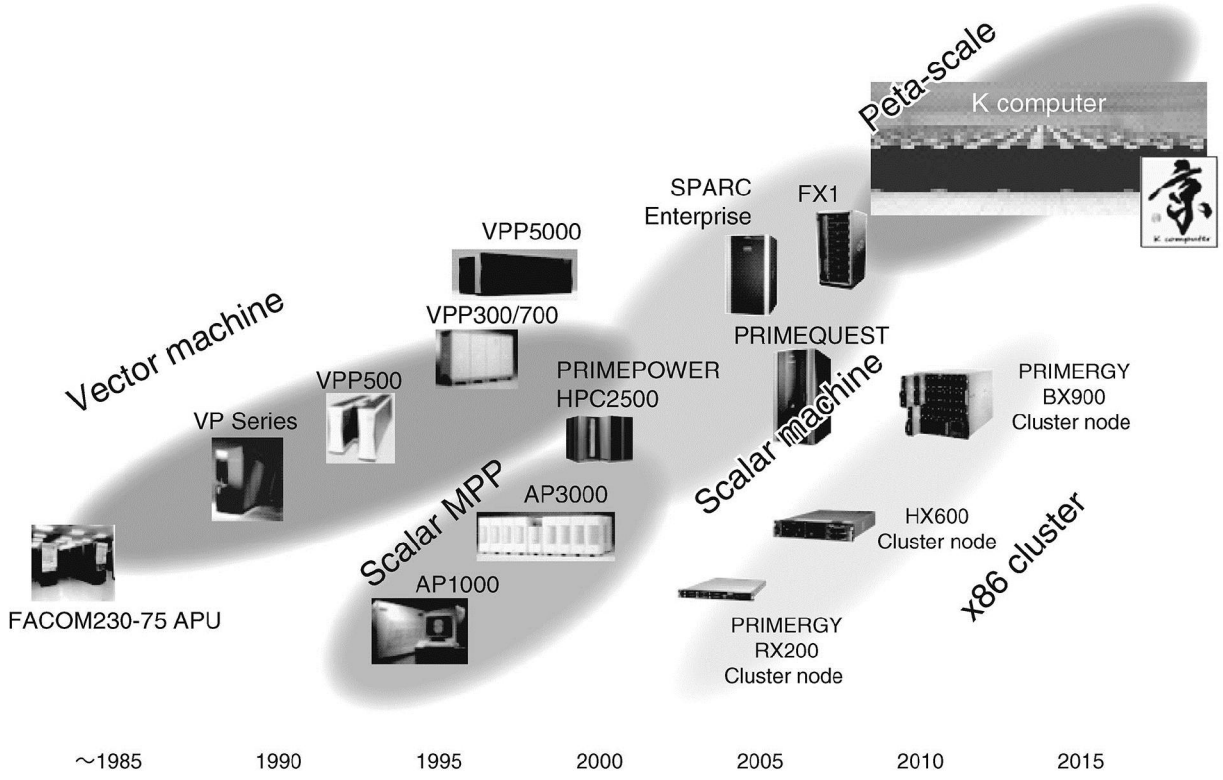

Рисунок 1
История разработки суперкомпьютеров в Fujitsu

2. Конфигурация вычислительных узлов в K компьютере

Вначале представим обзор вычислительных узлов, лежащих в основе компьютерной системы K. Вычислительный узел состоит из процессора, памяти и интерконнекта.

1) CPU

Мы разработали 8-ядерный процессор с теоретической пиковой производительностью 128GFLOPS, называемый "SPARC64 VIIIfx", как процессор для K компьютера (рис.2)[4]. SPARC64 VIIIfx выполнен с использованием передовой 45-нм технологии полупроводникового процесса Fujitsu, и имеет отношение производительности к энергопотреблению мирового уровня в 2,2GFLOPS/Вт достигаемого путем внедрения мероприятий по сокращению энергопотребления одновременно за счет технологического процесса и дизайна. Данный CPU применяет расширения для арифметических расчетов высокопроизводительных вычислений (HPC-ACE) применяемых для научных обработки данных и анализа. Он также использует а 6-MB/12-way раздел в качестве L2 кэша и использует архитектуру виртуального отдельного процессора с интегрированной многоядерностью (VISIMPACT, Virtual Single Processor by Integrated Multicore Architecture), эффективность которой ранее была продемонстрирована профессиональных технических компьютерных серверах Fujitsu's FX1. В результате этих особенностей SPARC64 VIIIfx достигает высокой производительности выполнения операций в области HPC. Кроме того, функции системного управления и управления доступом к памяти (MAC), которые ранее были реализованы на отдельных микросхемах, теперь интегрированы в CPU, что в результате приводит к высокой пропускной способности памяти и ее низкой латентности.

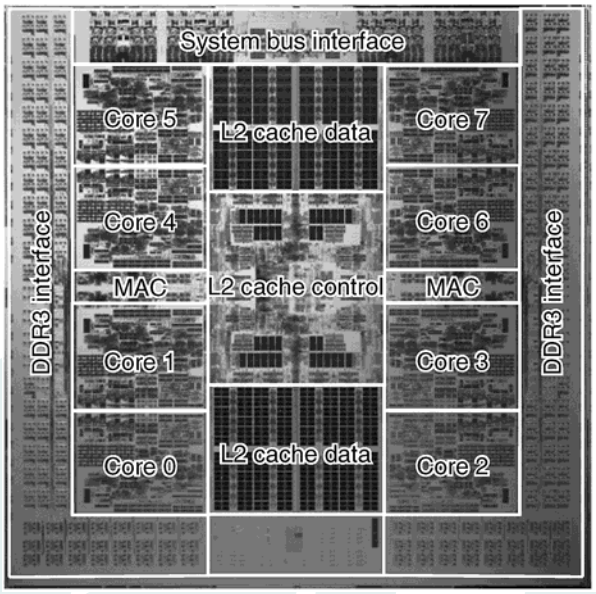

Рисунок 2
SPARC64 VIIIfx

2) Оперативная память.

В качестве основной памяти используется имеющаяся в продаже DDR3-SDRAM-DIMM. Модули DIMM являются широко используемым товаром, используемом в серверах и компьютерных кластерах, что обеспечивает ряд преимуществ. Например, это могут быть стабильные поставки от нескольких поставщиков примерно 700 000 модулей, которые можно получать в период производства около одного года со стабильным уровнем качества, причем могут быть выбраны модули с лучшими свойствами энергопотребления. Кроме того, 8-канальный интерфейс памяти на процессор обеспечивает пропускную способность памяти 64Гбайт/с, превышая показатели компьютеров конкурентов и обеспечивая высокую пропускную способность памяти, которая считается необходимой для научных вычислений.

3) Интерконнект

Для обеспечения вычислительной сети (интерконнекта) мы разработали и внедрили архитектуру интерконекта под названием "Tofu" для массивно параллельных вычислений с количеством узлов, превышающим 80 000 [5)]. Интерконнект Tofu (рисунок 3) представляет собой сеть прямого объединения, которая обеспечивает масштабируемые связи с низкой латентностью и высокой пропускной способностью для массивно параллельных групп процессоров и обеспечивает высокую производительность и доступность с помощью конфигурации сети 6D тора. Она использует алгоритм маршрутизации последовательности расширенной размерности которая позволяет представлять группы процессоров объединенными в 1–3 мерные торы, таким образом неисправные узлы становятся несуществующими для рассмотрения пользователя. Интерконнект Tofu также делает возможным выделение пользователю полной группы процессоров, объединенных в 1–3 мерный тор, даже если часть системы была выделена для использования.

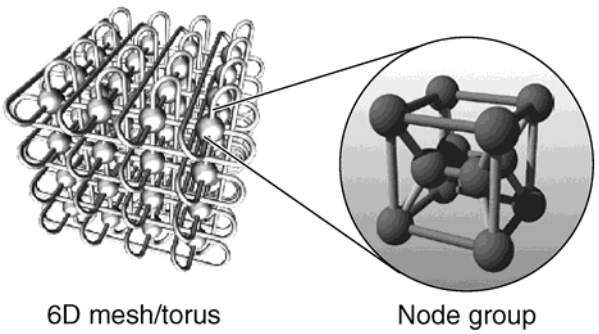

Рисунок 3
Интерконнект Tofu

Вычислительный узел К компьютера состоит из описанных выше CPU/DIMM и микросхемы контроллера интерконнекта (ICC). Поскольку интерконнект Tofu образует прямую сеть обмена данными, ICC включает в себя функцию ретрансляции пакетов между другими узлами (рисунок 4). Если выйдет из строя часть вычислительного узла, то будет затронута только работа, связанная с использованием этого вычислительного узла, в то время как при выходе из строя части маршрутизатора, связанного с этим вычислительным узлом, будут затронуты все вычислительные узлы, использующие вышедший из строя узел в качестве ретранслирующего. Уровень отказов частей маршрутизатора почти пропорционален количеству содержащихся в нем цепей и это требует делать его существенно более низким по сравнению с уровнем отказов частей узлов. Неисправности ICC микросхем по этой причине классифицированы в соответствии с их воздействием, которое определяет мероприятия, предпринимаемые во время возникновения неисправности. Конечный результат представляет собой механизм для продолжения функционирования системы в случае возникновения неисправности в части узла.

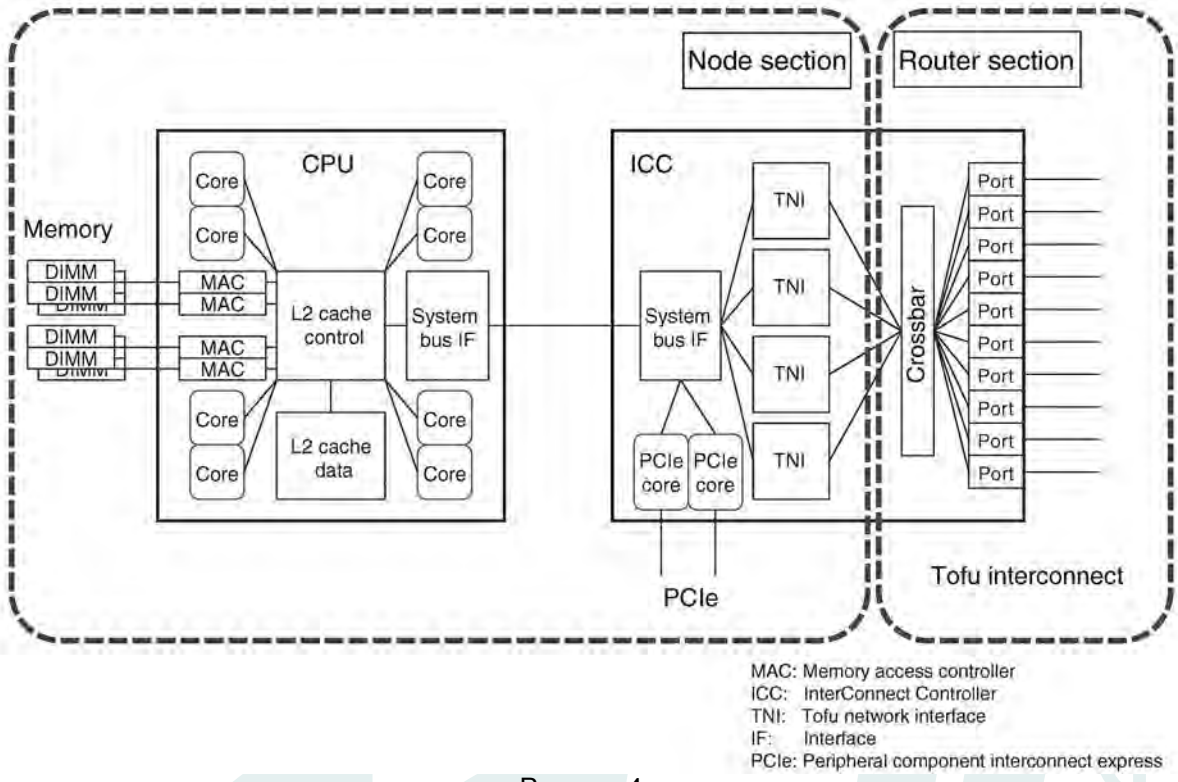

Рисунок 4
Концептуальная диаграмма узла и секции маршрутизации

3. Конфигурация стойки К компьютера

Вычислительные узлы К компьютера устанавливаются в специально разработанные стойки. Компьютерная стойка может содержать 24 системные платы (SBs), причем каждая содержит 4 вычислительных узла (рисунок 5), и 6 системных плат ввода/вывода (IOSBs), каждая из которых организовывает узел ввода/вывода для системной загрузки и доступа к файловой системе. Следовательно, всего 102 узла могут быть установлены в стойке.

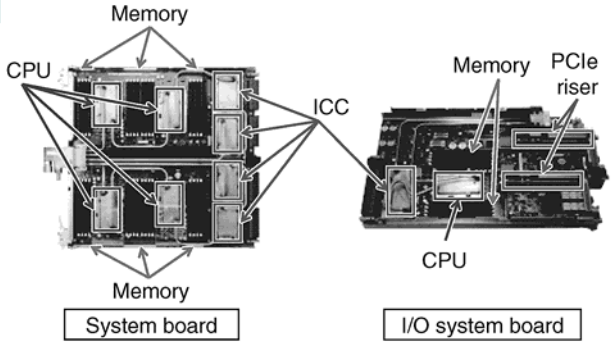

Рисунок 5
Системная плата и системная плата ввода/вывода

Для CPU/ICC используемых для вычислительных узлов и узлов ввода/вывода, а также и для бортовых устройств электроснабжения применяется водяное охлаждение. В частности, охлажденная вода подается в каждую SB и IOSB через трубы водяного охлаждения и шланги, установленные на стороне стойки. Водяное охлаждение используется для поддержания надежности системы, обеспечения внедрения высокой плотности и низкого энергопотребления (уменьшение тока утечки). В противоположность этому, модули DIMM установлен на SB и IOSB имеют воздушное охлаждение, так как они используют коммерчески доступные компонентов. Так как эти платы установлен горизонтально внутри стойки, воздух должен продуваться в горизонтальном направлении. Чтобы сделать возможным стойкам организовываться в конфигурации с высокой плотностью SB установлены по диагонали внутри стойки с образованием структуры, в которой воздух всасывается спереди, в то время как платы расположены под углом к потоку, который проходит через SB для охлаждения модулей DIMM с последующим выбросом через заднюю сторону (рисунок 6).

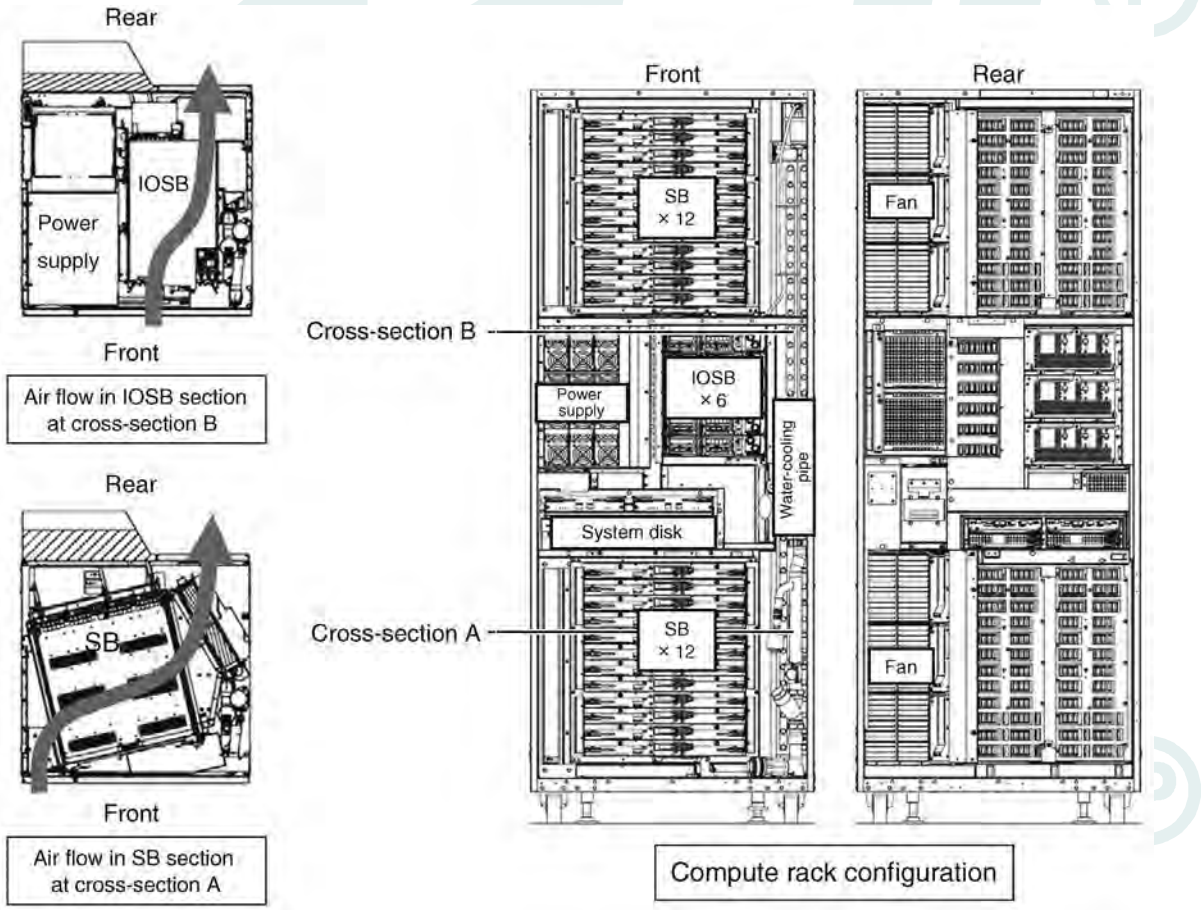

Рисунок 6
Конфигурация вычислительных стоек и направление воздушного потока

Шесть узлов ввода/вывода, установленных в каждую стойку классифицируются по применению:
- Узлы ввода/вывода загрузки (BIO): два узла
- Узлы локального ввода/вывода (LIO): три узла
- Узлы глобального ввода/вывода (GIO): один узел

Узлы BIO соединяются через интерфейс Fibre Channel до 8 Gbit/s (8G-FC) с диском системной загрузки (ETERNUS DX80) установленном в стойке. Эти узлы самостоятельно стартуют при включении питания и работают как сервер загрузки через Tofu для остальных 100 узлов в стойке.[6)] Причем один играет роль активной системы, в то время как второй выступает в качестве резервной системы. Если узел, выступающий в роли активной системы выходит из строя, резервный берет на себя его роль.

Узлы LIO осуществляют связь с локальными дисками (ETERNUS DX80) установленными в дисковых стойках, расположенными рядом с вычислительными стойками.

Узел GIO связывается с внешней глобальными системами хранения посредством интерфейса QDR InfiniBand. Эти системы хранения используют файловую систему Fujitsu Exabyte (FEFS)[7)] для обеспечения крупномасштабной, высокопроизводительной и высоконадежной файловой системы. Если узлы LIO или GIO в стойке выходят из строя, доступ к файловой системе может быть продолжен через путь к узлам LIO или GIO, смонтированным в соседней стойке.

Каждая стойка также содержит блоки питания (PSU), вентиляторы воздушного охлаждения, и обслуживающие процессоры (SP), все в избыточной конфигурации, следовательно, одиночный отказ не приводит к выходу из строя всей системы в стойке.

4. Общая конфигурация системы

Система К компьютер содержит итого 864 вычислительных стоек. Как показано на рис. 7, две ближайшие стойки соединяются кабелем Z-оси (объединяя 17 узлов включая узлы ввода/вывода). Существует одна дисковая стойка на каждые четыре вычислительные стойки и 45 стоек расположены вдоль Y-оси (36 вычислительные стойки + 9 дисковых стоек) и 24 стойки вдоль X-оси. К компьютер, следовательно имеет конфигурацию прямоугольного яруса из 24x45 стоек. Каждая дисковая стойка содержит 12 локальных дисков и каждый локальный диск с 2 ближайшими узлами LIO в X-направлении.

Существует нечто большее и совершенно необходимое в К компьютере, чем просто вычислительные стойки, дисковые стойки и параллельная файловая система - управляющий механизм, состоящий из внешних серверных аппаратных средств и эксплутационного/ управляющего программного обеспечения для управления системой и заданиями.[8)] Для замены аппаратных средств после их выхода из строя также необходимы работы по техническому обслуживанию, основанные на системе удаленной сигнализации и сервере обслуживания. Обзор всей системы К компьютера, установленной в RIKEN, приводится на рисунке 8.

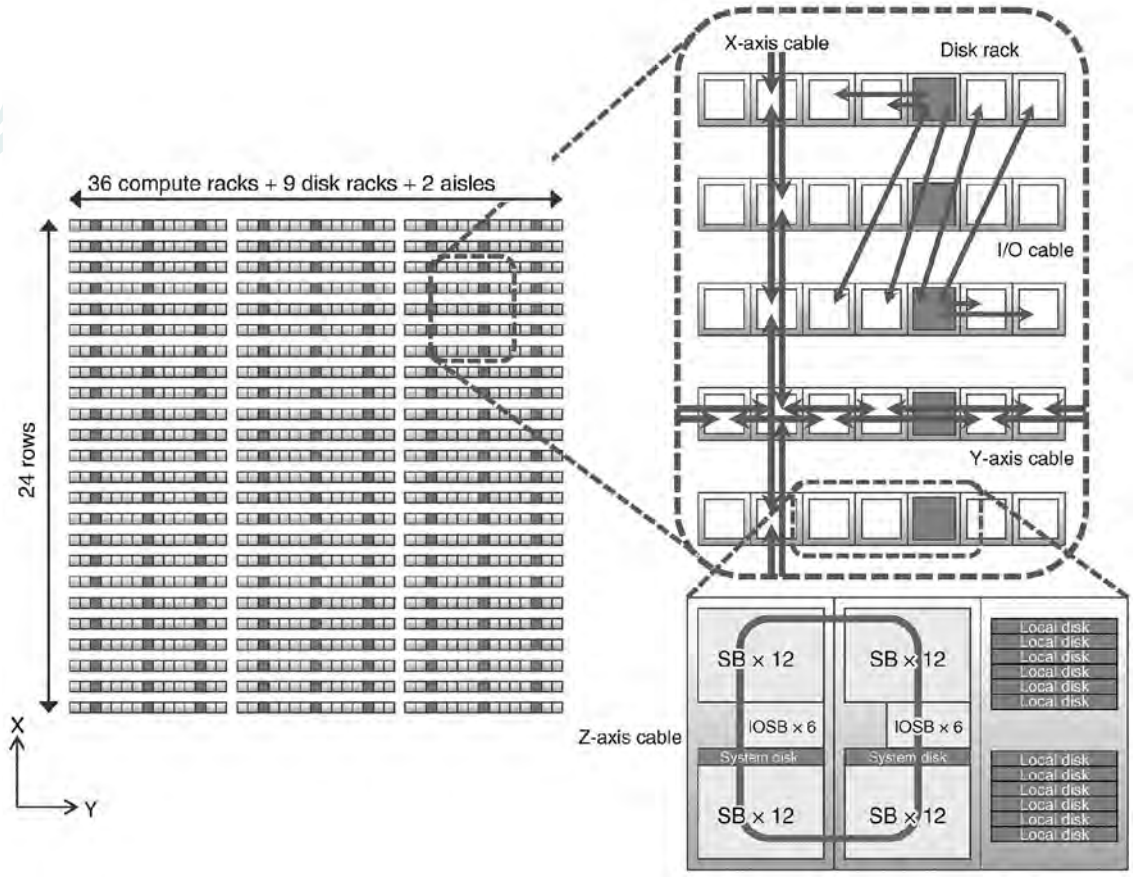

Рисунок 7
Конфигурация вычислительных стоек

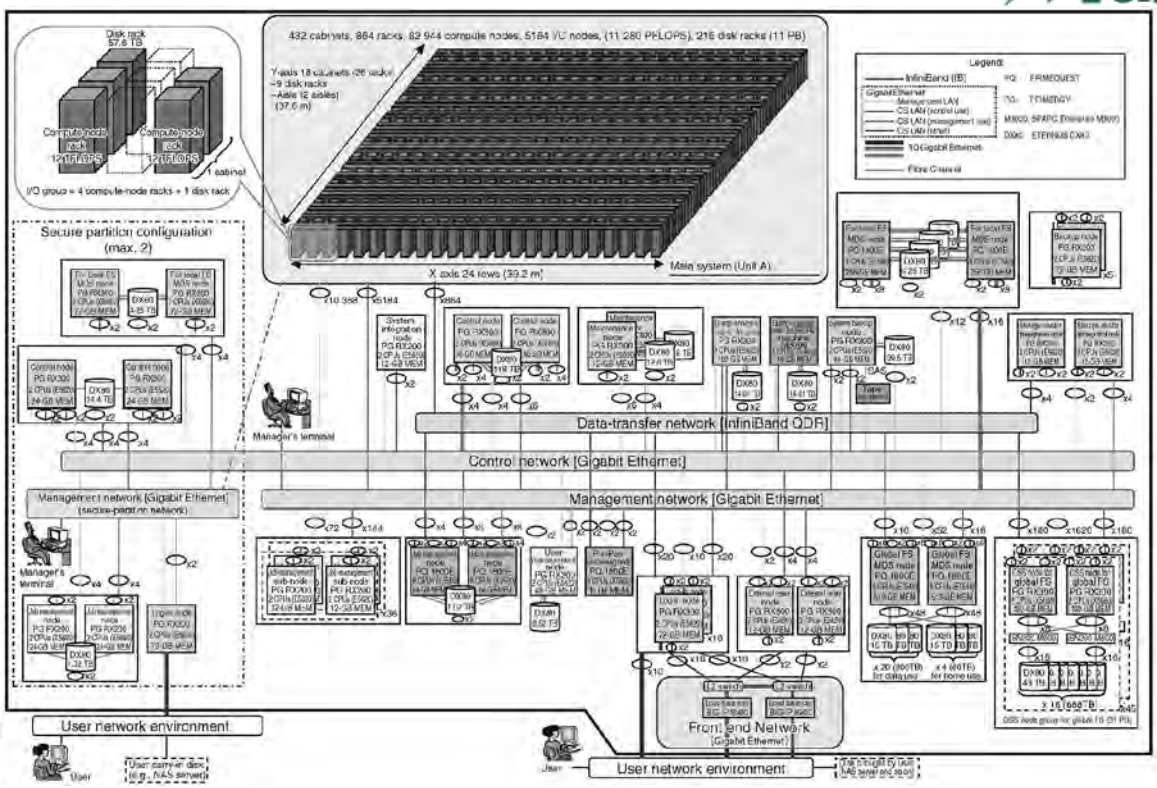

Рисунок 8
Обзор всей системы

5. Меры по снижению энергопотребления, предпринятые на системном уровне

Чтобы сократить энергопотребление всего К компьютера, первым шагом было внедрение энергичных мер по снижению энергопотребления на стадии проектирования CPU/ICC. Динамический ток был понижен путем использования стробирования, в то время как рабочая частота была снижена до 2.0ГГц, а токи утечки были значительно снижены путем применения транзисторов с высоким пороговым напряжением.[9)] Также было принято решение использовать водяное охлаждение, поскольку токи утечки резко возрастают при повышении температуры перехода. Это имело эффектом понижение температуры перехода с обычной 85°C до 30°C, что привело к значительному уменьшению тока утечки.

Использование конструкции стробирования для уменьшения мощности потребления БИС (LSI) делает колебания нагрузки весьма существенными. Это делает необходимым улучшение кратковременных (пиковых) характеристик в блоках питания. Система преобразования промежуточной шины была применена для улучшения кратковременных характеристик и эффективного энергоснабжения большого количества стоек с одновременным обеспечением реализации высокой плотности. В частности, кратковременные характеристики были улучшены путем расположения неизолированных пунктов нагрузки (POL) источников питания рядом с нагрузкой и запитывание на всей стойке было сделано более эффективным путем обеспечения каждой SB 48B источником от блоков питания в стойке. Монтаж шинного преобразователя изолированного типа на каждой SB делает возможным понижение подаваемого 48B напряжения и подавать питание на блоки питания POL. Этот подход снижает количество изолированных трансформаторов и достигает высокого кпд и высокой плотности систем питания.

6. Мероприятия по высокой надежности / высокой доступности на системном уровне

Принятие конфигураций систем-на-чипе (SOC) означает, что количество схем на микросхеме увеличивается, в то время как число компонентов вне микросхем уменьшается. Следовательно сбои в крупномасштабных микросхемах составляют бóльшую долю отказов системы, что означает, что снижение интенсивности отказов крупномасштабных микросхем имеет важное значение для стабильной работы К компьютера.

6.1 Мероприятия высокой надежности

В К компьютере надежность повышается за счет повышения отдельных компонентов и применения соответствующих способов использования этих компонентов. В связи с этим усилия по улучшению надежности системы выполняются в рамках непрерывности работы и доступности системы.

Как показано в таблице 1, предоставление основным компонентам избыточной конфигурации помогает обеспечить непрерывность функционирования и, следовательно, непрерывность работы, а замена неисправных компонентов в "горячем режиме" без остановки всей системы помогает обеспечить доступность системы.

Таблица 1
Высокая доступность за счет использования избыточной конфигурации и горячей замены

Основные компоненты	Избыточная конфигурация	Горячая замена
Блок питания стойки	Да	Да
Охлаждающие вентиляторы	Да	Да
Обслуживающие процессоры (SPs)	Да (дублирование/переключение)	Да
Системные платы (SBs/IOSBs)	Нет ⇒ Обход неисправности вдоль оси-B	Да
CPU/ICC	Нет ⇒ Доступность улучшена использованием водяного охлаждения, повторного исполнения, и корректирующего ошибки кода (ECC)	(горячая замена SB)
POL источники питания	Нет ⇒ Доступность улучшена использованием водяного охлаждения	(горячая замена SB)
Непосредственные конвертеры	Да	(горячая замена SB)
DIMMs	Нет ⇒ Восстановление данных с использованием дополнительных ECC	(горячая замена SB)
Системные диски	Да ⇒ Контроллер и блок питания: избыточный HDD: RAID5 конфигурация + горячий резерв	Да (модуль)

Более подробно: мероприятия в рамках обеспечения непрерывности работы включают применение избыточных аппаратных конфигураций, снижение числа компонентов, использующих конфигурации 1-узел/1-CPU, коррекцию ошибок с помощью кода коррекции ошибок (ECC) и повторов, а также понижение уровня полупроводниковых сбоев за счет снижения рабочей температуры путем применения водяного охлаждения БИС (LSIs) и POL блоков питания. Эти меры позволяют снизить вероятность аварии узла или задания из-за сбоев компонентов. Еще одной мерой, помогающей обеспечить непрерывность работы, являются использование электрических кабелей в интерконнекте Tofu.

6.2 Мероприятия высокой доступности

С точки зрения обеспечения высокой доступности было реализовано выравнивание координат Tofu внутри SB, следовательно, механизм обхода неисправного узла на основе 6D тора может быть расширен до обработки неисправных SB и горячего обслуживания. Это позволяет обходную маршрутизацию, несмотря на тот факт, что SB содержит только четыре узла. В частности, рисунок 9 показывает как четыре узла SB ограничены теми же координатами Y/B и объединены вдоль оси A и оси C на SB, следовательно, соединение вдоль оси B предлагает выход из SB.

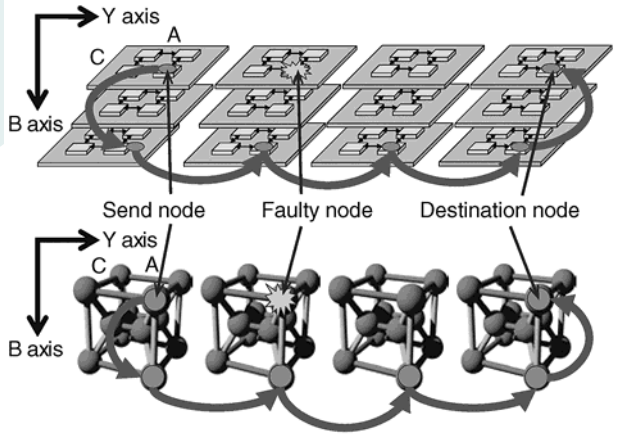

Рисунок 9
Концептуальная диаграмма обходных путей Tofu

Кроме того, обеспечивается избыточность путей ввода/вывода для трех типов узлов ввода/вывода (BIO, LIO, GIO), следовательно, по-прежнему можно использовать пути ввода/вывода на неисправных вычислительных узлах. Это достигается путем заимствования конфигурации, которая включает в себя альтернативные узлы ввода/вывода, имеющие общие направления подключения, которая позволяет вычислить группу узлов для доступа альтернативных узлов ввода/вывода через интерконнект Tofu (рисунок 10). Существует два узла BIO на вычислительную стойку в конфигурации с избыточными конфигурации активный/резервный, а также три узла LIO на вычислительную стойку, которые участвуют в избыточной конфигурации совместно с узлами LIO в другой вычислительной стойке, которые совместно используют локальные диски, установленные в дисковой стойке. Аналогично, единственный узел GIO в компьютерной стойке участвует в избыточной конфигурации совместно с узлом GIO другой вычислительной стойки.

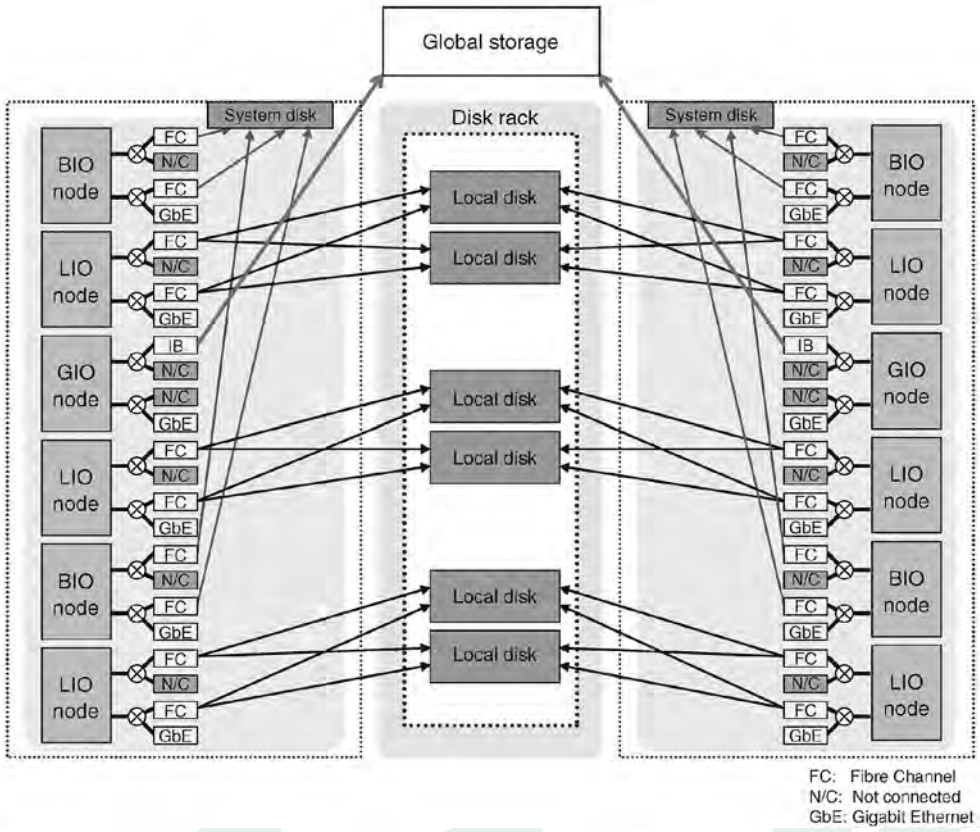

Рисунок 10
Избыточность путей ввода/вывода

7. Результаты тестирования

Хотя он пока находится в стадии конфигурации (на момент написания статьи -прим.перев.), K компьютер разработан в соответствии с целями, описанными в начале этой статьи, и достиг производительности 8,162PFLOPS на тесте HPC-LINPACK, используя только часть системы (672 стойки). Это достижение выиграло K компьютеру место № 1 в списке TOP500 объявленном в июне 2011.[10] Несмотря на то, что тест LINPACK выполнялся более чем 28 часов, все 68 544 узлов, участвовали в работе непрерывно без сбоя, демонстрируя высокую надежность системы и высокую доступность работы. Подробная информация о данных результатах тестирования, в также объявленные в ноябре 2011 года, приведены в таблице 2.

Таблица 2
Результаты TOP500/Green500

	June 2011	November 2011	Unit
TOP500 ranking	1	1	Rank
No. of cores	548 352	705 024	Units
Performance (Rmax)	8162.00	10 510.00	TFLOPS
Power	9898.56	12 659.9	kW
Green500 ranking	6	32	Rank
Power performance	824.56	830.18	MFLOPS/W

К компьютер также участвовал в тестировании HPC Challenge, который оценивает общую производительность суперкомпьютера. В этом тесте он получил номер 1 рейтинга во всех четырех категориях {Глобальный HPL, Глобального RandomAccess, EP STREAM (Триада) на систему, и Глобального БПФ} этого HPC Challenge Award (Class 1) объявленного в ноябре 2011 года. Кроме того, K компьютер был удостоен приза Gordon Bell Peak Performance, объявленного в ноябре 2011 года в знак признания его достижений в реальных приложениях. Эти результаты показывают, что K компьютер далеко не машина специализирующаяся на тесте LINPACK, обладает свойствами общего назначения, которые могут поддерживать широкий спектр приложений.

8. Заключение

Задачи разработки, установленные для K компьютера требуют для масштабной высокопроизводительной вычислительной системы проведения мероприятий по повышению значения надежности и высокой доступности начиная со стадии проектирования. RIKEN и Fujitsu провели работу для достижения цели пиковой производительности и эффективной производительности приложений и построения системы с высокой надежностью и высокой доступностью. K компьютер занял № 1 в списке TOP500 бенчмарка июня 2011 года, и должен остаться № 1 в списке, выходящем в следующем ноябре 2011 года (прим.перев. -так и было). Он также достигнет производительности LINPACK 10PFLOPS. Эти достижения показывают, что цели неуклонно выполняются, как это и планировалось. Чтобы ввести начинающийся полный комплекс услуг по мере завершения фазы настройки производительности подходящей к концу, мы будем оценивать эффективности реальных приложений и реальной работы системы.

Ссылки

1) M. Yokokawa et al.: The K computer: Japanese next-generation supercomputer development project. ISLPED, pp. 371–372 (2011).
2) T. Yoshida et al.: SPARC64 VIIIfx: CPU for the K computer. Fujitsu Sci. Tech. J., Vol. 48, No.3, pp. 274–279 (2012).
3) Y. Ajima et al.: Tofu: Interconnect for the K computer. Fujitsu Sci. Tech. J., Vol. 48, No.3, pp. 280–285 (2012).
4) T. Maruyama et al.: SPARC64 VIIIfx: A New-Generation Octocore Processor for Petascale Computing. IEEE Micro, Vol. 30, Issue 2, pp. 30–40 (2010).
5) Y. Ajima et al.: Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers. Computer, Vol. 42, No. 11, pp. 36–40 (2009).
6) J. Moroo et al: Operating System for the K computer. Fujitsu Sci. Tech. J., Vol. 48, No.3, pp. 295–301 (2012).
7) K. Sakai et al: High-Performance and Highly Reliable File System for the K computer. Fujitsu Sci. Tech. J., Vol. 48, No.3, pp. 302–309 (2012).
8) K. Hirai et al: Operations Management Software for the K computer. Fujitsu Sci. Tech. J., Vol. 48, No.3, pp. 310–316 (2012).
9) H. Okano et al.: Fine Grained Power Analysis and Low-Power Techniques of a 128GFLOPS/58W SPARC64TM VIIIfx Processor for Peta-scale Computing. Proc. VLSI Circuits Symposium, pp. 167–168 (2010).
10) TOP500 Supercomputing Sites. http://www.top500.org