Программное обеспечения управления работой К компьютера

• Kouichi Hirai • Yuji Iguchi • Atsuya Uno • Motoyoshi Kurokawa
FUJITSU Sci. Tech. J., Vol. 48, No. 3, pp. 310–316 (July2012)

Суперкомпьютерные системы неуклонно растут в масштабах (количество процессорных ядер и числа узлов) как ответ на постоянно растущий спрос на вычислительные мощности.
Программное обеспечение управления работой является ключом к эксплуатации таких ультра-крупномасштабных систем в стабильном режиме и предоставления пользователям среды высокопроизводительных вычислений, имеющей высокую степень загрузки.
Fujitsu ранее разработала программное обеспечение, называемое "Parallelnavi" для управления унифицированными операциями на своих суперкомпьютерных системах с 3000-узлами, однако унифицированно для управления ультра-крупномасштабной системой, подобной К компьютеру с масштабом превышающим 80 000 узлов, компания продвинула свое развитие технологий управления операциями.
В данной статье рассматривается программное обеспечение для управления работой, разработанное для К компьютера с концентрацией внимания на функциях достижения стабильной работы ультра-крупномасштабной системы и планировщиком заданий для обеспечения высокопроизводительного вычислительного окружения.

1. Введение

Для обеспечения стабильной работы системы и высокопроизводительной вычислительной среды с использованием управления работой ультра-крупномасштабной системы подобной К компьютеру [note 1)], были установлены следующие цели.

- Снижение издержек, связанных с программным обеспечение управления работой
- Обеспечение дружелюбного пользователю вывода суммированные огромные объемы информации
- Избежать падения производительности выполнения из-за взаимного влияния пакетных заданий

Для достижения этих целей и эффективного извлечения высокой вычислительной производительности, предоставляемой аппаратным обеспечением, мы расширили наши усилия в области развития в технологиях управления системным программным обеспечением.

Эта статья описывает наш подход к достижению этих целей в программном обеспечении управления работой К компьютера и описывает особенности этого программного обеспечения.

2. Конфигурация программного обеспечения управления работой

Программное обеспечение для управления работой К компьютера в целом подразделяется на функции управления системой и планировщиком задач, которые кратко излагаются ниже.

1) Функции управления системой

Эти функции снабжают системного администратора представлением управления работой в виде единой системы.

Они позволяют администратору управлять информацией о конфигурации системы и контролировать рабочее состояние всех узлов, составляющих систему, обеспечивая при этом высокую доступность системы в условиях аномальных событий.

Они также позволяют администратору устанавливать новое программное обеспечение и осуществлять поддержку существующее программное обеспечение на основе применения патчей и других средств.

2) Планировщик заданий

Эта функция обеспечивает среду выполнения для пакетных заданий, запускаемых несколькими пользователями для достижения совместного использования суперкомпьютерной системы.

В следующих разделах представляются особенности этих функций управления системой, а также планировщик заданий.

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.

3. Функции управления системой

3.1 Работа с ультра-крупномасштабной системой

Функции управления системой постоянно следят за рабочим состоянием системы на предмет возникновения неисправностей.

По мере роста масштабов системы, однако, сетевые нагрузки, вызываемые регулярным мониторингом и связанной с ним обработкой на компьютерах, становятся слишком большими, чтобы их можно было продолжать игнорировать.

По этой причине функции управления системой включают следующие меры, позволяющие не отставатьот продолжающегося роста суперкомпьютерной системы.

- Распределение нагрузки обработки мониторинга с использованием иерархической структуры
- Снижение шума системы с использованием интерконнекта Tofu

Эти и другие меры, такие как отображение суммированного рабочего состояния узлов системы были разработаны для того, чтобы справиться с ультра-крупномасштабной системой.

3.2 Распределение нагруски с помощью иерархической структуры

Для обеспечения системного администратора окружением оперативного управления, имеющей единое представление системы, необходимо исключить сосредоточения нагрузки на персональном компьютере администратора (управляющем PC) или выделенном сервере.

Если каждый узел в среде из 80,000 узлов просто бы отправил данные мониторинга по сети управляющему компьютеру, то легко может быть исчерпана мощность промежуточных сетевых путей или самого управляющего компьютера.

Также можно ожидать, что количество узлов системы возрастает со временем, так что становятся необходимыми некоторые формы распределения нагрузки.

Программное обеспечение для управления работой К компьютера использует иерархическую архитектуру программного обеспечения, описываемую ниже, для достижения распределения нагрузки и параллельной обработки в управлении функционированием.

Как показано на Рисунке 1, эта иерархическая структура делит систему на группы узлов и выделяет суб-узел управления заданиями в каждой группе для управления узлами, расположенными в этой группе.

При наличии суб-узлов управления заданиями регламентируется обработка операций управления в рамках соответствующих им групп узлов распределяя нагрузку выполнения операций управления в масштабах всей системы.

Такое распределение нагрузки может быть легко достигнуто, даже если число узлов увеличивается после первоначального развертывания системы путем простого добавления дополнительных групп узлов, что делает возможной высокую степень расширения системы.

Кроме того, позволяя мониторинг и управление узлами во время обычных операций, такая иерархическая структура также выступает в качестве платформы для работ по техническому обслуживанию, таких как установка систем и запуск приложений исправлений, а также в качестве платформы для любого вида управляющих операций между узлами, таких как инициация и терминация параллельной программы.

Использование нескольких подузлов управления заданиями таким образом, распределяет и распараллеливает обработку операций управления позволяя запустить в пределах одной секунды параллельную программу, которая будет работать на 3000 узлов, по сравнению с несколькими секундами, требующимися для этого в предыдущих системах.

Таким образом, эта иерархическая структура может принести пользу не только для управления работой, но и для других процессов, таких как производительность запуска ультра-крупномасштабных параллельных программ, так как это может помочь сократить различия во временах обработки операций, которые не могут быть проигнорированы в ультра-крупномасштабной системе.

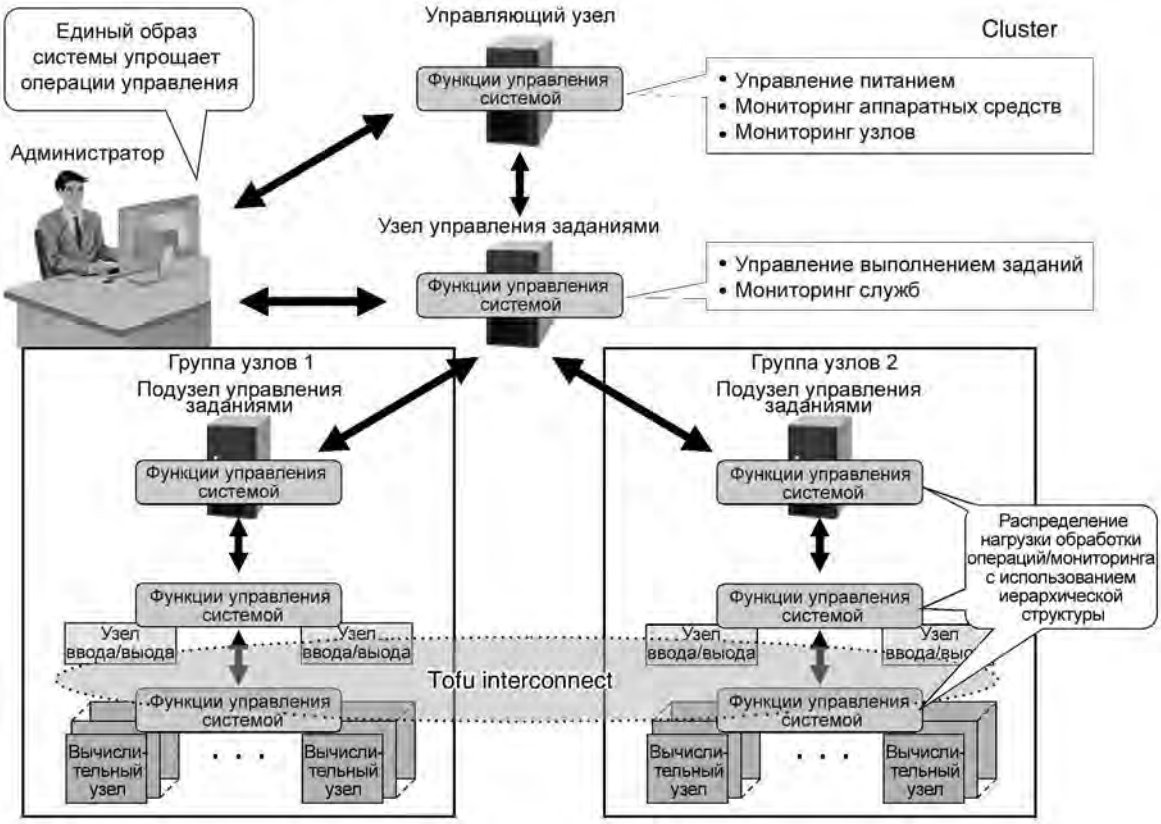

Рисунок 1
Распределение нагрузки с помощью иерархической структуры

3.3 Объединение информации о работе узлов с использованием интерконнекта Tofu

The monitoring of node status and the execution of various types of controls by operations management software is generally carried out using the following scheme: a daemon program running on a node receives an interrupt from the network such as a Gigabit Ethernet, gathers information on the node in question, and transfers that information to the management PC over the same network.

Мониторинг за состоянием узла и выполнение различных типов регулировок программным обеспечением управления работой обычно проводят по следующей схеме: программа-демон работает на узле, получает прерывание от сети, например от Gigabit Ethernet, собирает информацию на обсуждаемом узле, и передает эту информацию на управляющий PC через ту же сеть.

Тем не менее, работа такой схемы на ультра-крупномасштабной системе может привести к системному шуму, вызванного сетевыми прерываниями и работой демона, и этот шум может значительно ухудшить производительность выполнения параллельной программы.

Таким образом, вместо того, чтобы собирать и возвращать информацию по возникновению сетевого прерывания, программное обеспечение для управления работой на К компьютере использует схему, в которой каждый узел независимо и периодически сохраняет в памяти свои собственные состояния об операциях, которые могут быть получены извне с помощью дистанционного прямого доступа к памяти (RDMA)[note2)] с использованием интерконнекта Tofu.

Такой мониторинг и сохранение состояния работы узла на каждом узле выполняется на уровне ядра системы, насколько это возможно, чтобы значительно уменьшить образование шума системы.

Эта и другие контрмеры по отношению к шуму системы, включенные в программное обеспечение для управления работой снизили уровень шумов в системе до 1/100, по отношению к кластеру PC, который генерирует шум системы примерно раз в каждую миллисекунду.

3.4 Другие функции

В дополнение к указанным выше мерам по сокращению накладных операциям управления для ультра-крупномасштабной системы, следующие функции также предоставляются как функции управления работой для достижения стабильного функционирования системы.

note2) Функция коммуникации интерконнекта Tofu.

1) Система высокой доступности

Эта функция позволяет выполняющемуся в настоящий момент заданию продолжить выполнение, даже если определенные узлы в системе управления операциями должны выйти из строя.

Она таким образом гарантирует непрерывность выполнения заданий даже во время возникновения неисправностей системы.

Однако, даже с этой функцией не было бы возможным запустить новые задания в случае, когда выходят из строя определенные управляющие узлы, и, по этой причине, также поддерживается конфигурация с резервированием для управляющих узлов, чтобы гарантировать непрерывную работу всей системы.

2) Дружественное представление суммированного огромного объема информации

Если состояния более чем 80 000 узлов, находящихся под управлением, должно были быть просто отображено в виде один-узел на строку в терминальном режиме командной строки, то результатом, очевидно, будет массовый показ более чем в 80 000 строк.

В качестве альтернативы эта функция отображает сводку общесистемной информации, такой, например, как число узлов, находящихся в настоящее время в различных состояниях.

Если необходима более детальная информация, также предоставляются различные варианты.

Например, пользователь может выбирать отображение информации, указав определенную группу узлов или определенный узел.

4. Планировщик заданий

4.1 Подоплека разработки

Работа ранее разработанных Fujitsu планировщиков для суперкомпьютерных систем имеют следующие особенности.

- Предотвращение взаимных помех между заданиями путем предварительного выделения вычислительных ресурсов
- Принятие концепции "параллельных заданий", в которой параллельная программа, работающая на множестве узлов рассматривается как единое задание
- Поддержка различных требований суперкомпьютерных центров коллективного пользования

При разработке планировщика заданий для К компьютера было принято решение включить эти функции и добавть два важных направления развития: распределение нагрузки для поддержки ультра-крупномасштабной системы и улучшение коэффициента использования системы.

4.2 Поддержка ультрамасштабной системы

Число заданий, обрабатываемых планировщиком резко возрастает по мере расширения системы.

Ожидается, что К компьютер будет обрабатывать более 1,000,000 заданий одновременно.

Планировщик заданий для К компьютера выполняет этапы обработки для управления работой от принятия задания до его завершения (т.е. обработка принятия, обработка приоритизации выполнения, выделение вычислительных ресурсов, собственно выполнение задания и процесс терминации).

Эти шагт выполняются параллельно для сокращения общего времени обработки.

Планировщик заданий также выполняет обработку для определения в каком порядке, в какое время и на каких узлах должны выполняться задания (процесс, называемый "планированием")

Вычислительная сложность для определения оптимального решения в этой задаче увеличивается в соответствии с масштабом системы, создавая высокую стоимость обработки.

Планировщик заданий К компьютера сокращает это вычислительное время, выполняя процесс выбора злов выполнения параллельно.

В результате описанных выше мероприятий, данная работа планировщика достигает производительности ввода заданий на уровне 3 миллисекунд по сравнению с 0,4 секунды требующихся для представления одного задания с использованием планировщика заданий, разработанного сторонней компанией (PBSpro), используемого для кластеров PC (по результатам исследований Fujitsu).

Кроме того, как и в случае функций управления системой, планировщик заданий для К компьютера включает функции для работы с большим количеством информации, связанной с выполнением задания.

Например, он может отображать информацию в обобщенном виде, чтобы предотвратить одновременный вывод массивного объема данных, и, как показано на Рисунке 2, это дает возможность пользователю упростить отображение информацию путем выбора какую информацию отображать подробно, указания полей сортировки и т.д.

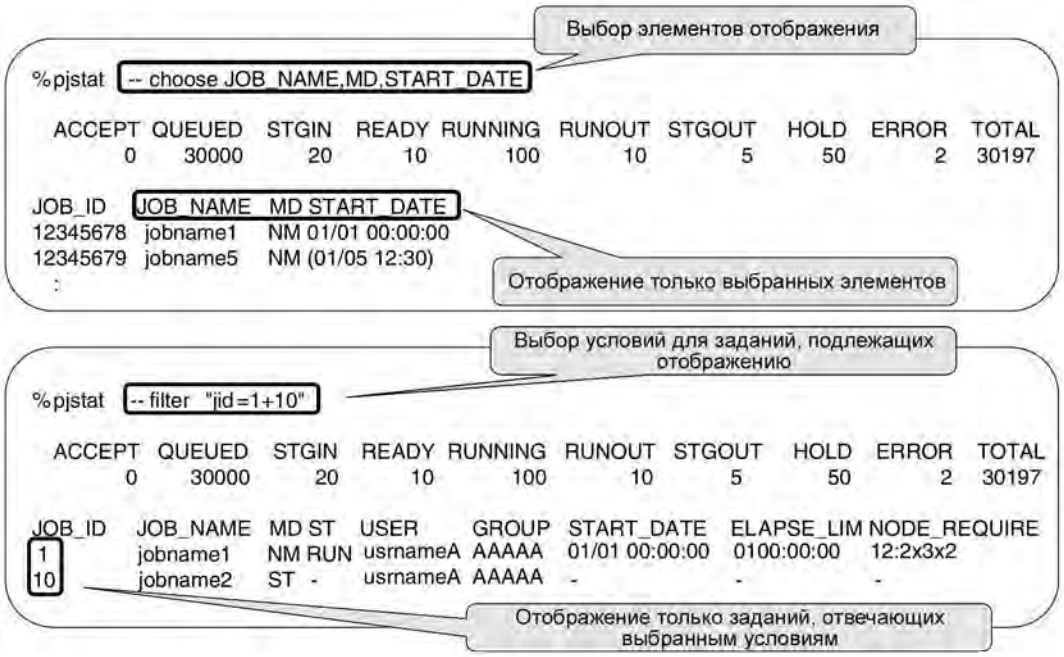

Рисунок 2
Примеры выбора и отображения информации о заданиях

4.3 Улучшение коэффициента использования системы

В центрах коллективного пользования задания разнообразных масштабов - от простых заданий, требующих только один узел, - до ультра-масштабных параллельных заданий, требующих десятки тысяч узлов - работают одновременно.

В такой среде, простое выполнение заданий в порядке, в котором они были поставляются, означает, что если масштаб необходимый следующему заданию превышает количество свободных узлов (узлов, на которых в настоящее время не выполняются задания), то задание будет ждать, таким образом ухудшая коэффициент использования всей системы.

Чтобы решить эту проблему, планировщик заданий для К компьютера поддерживает планирование с обратным заполнением, которое повышает коэффициент использования при наличии мелких заданий выполняемых перед крупномасштабными заданиями в течение периода, когда вычислительные ресурсы свободны до того момента времени, когда планируется начать крупномасштабное задание.

Кроме того, следующие меры при предварительной распределения вычислительных узлов для ожидающих рабочих мест с целью повышения коэффициента использования, принимая характеристики межсоединения Tofu во внимание.

- Для упращения предварительного выделения группы соседних узлов, пока не распределенные вычислительные узлы (свободные узлы) настраиваются таким образом, чтобы быть по возможности близкорасположенными (соприкасающимися).
- При выполнении такого рода настройки, возможные узлы (для запускаемых заданий) ищутся среди различных трехмерных образов с использованием интерконнекта Tofu так, чтобы увеличивать концентрацию соприкасающихся узлов среди всех свободных.

При раскрытии многих свободных узлов таким образом, становится более легким последоующее выделение узлов, которое эффективно повышает коэффициент использования.

4.4 Работа с конфликтами ввода/вывода (файловая конвейеризация)

По мере достижения системой "ультра-больших" масштабов, возрастает вероятность того, что производительнгость выполнения заданий упадет из-за конфликтов ввода/вывода среди одновременно выполняющихся заданий.

По этой причине К компьютер поддерживает "функции файловой конвейеризации", которые передают файлы, необходимые для выполнения заданий в локальную файловую систему до начала выполнения работы и которая собирает файлы, содержащие результаты выполнения задания.

Обычная функция конвейеризации файлов выполняет передачу файлов как часть выполняемого задания (синхронная конвейеризация), но при такой технике вычислительные ресурсы, необходимые для выполнения задания, остаются зарезервированными даже во время передачи файлов, в результате чего происходят потери ценных ресурсов.

Функция конвейеризации в К компьютере, напротив, отделяет передачу файла от выполнения задания, как показано на рисунке 3, тем самым предотвращая такие потери вычислительных ресурсов.

Другими словами, она использует наличие узлов ввода/вывода, независимых от вычислительных узлов, (особенность конфигурации этой системы) и выполняет передачу файлов на такие узлы ввода/вывода асинхронно с выполнением задания (асинхронная конвейеризация).

Выполнение передачи файлов до начала выполнения задания, таким образом, позволяет вычесть время передачи файлов из непосредственного времени выполнения задания, что повышает загруженность системы.

4.5 Улучшение работоспособности системы

Политика эксплуатации суперкомпьютерных центров коллективного пользования в Японии сильно разнится от центра к центру.

Существует, следовательно, потребность в планировщике заданий, которые могут поддерживать различные политики эксплуатации.

Политики эксплуатации, настраиваемые для центров коллективного пользования, в целом можно разделить на два типа.

- Отдельные политики применительно к каждой работе для определения пределов и ограничений, принимаемых во время выполнения задания (объем памяти и число процессоров, которое может использоваться, максимальное время выполнения и т.д.)
- Общая политика выполнения задания определяющая, например, какие задания имеют приоритет выполнения

Планировщик заданий К компьютера определяет различные ограничения и ограничения для индивидуальных политик с помощью функции под названием "задания ACL" и определяет общую политику выполнения заданий с использованием функций под названием "политика планирования".

Разработка этих функций упрощает применение политик эксплуатации.

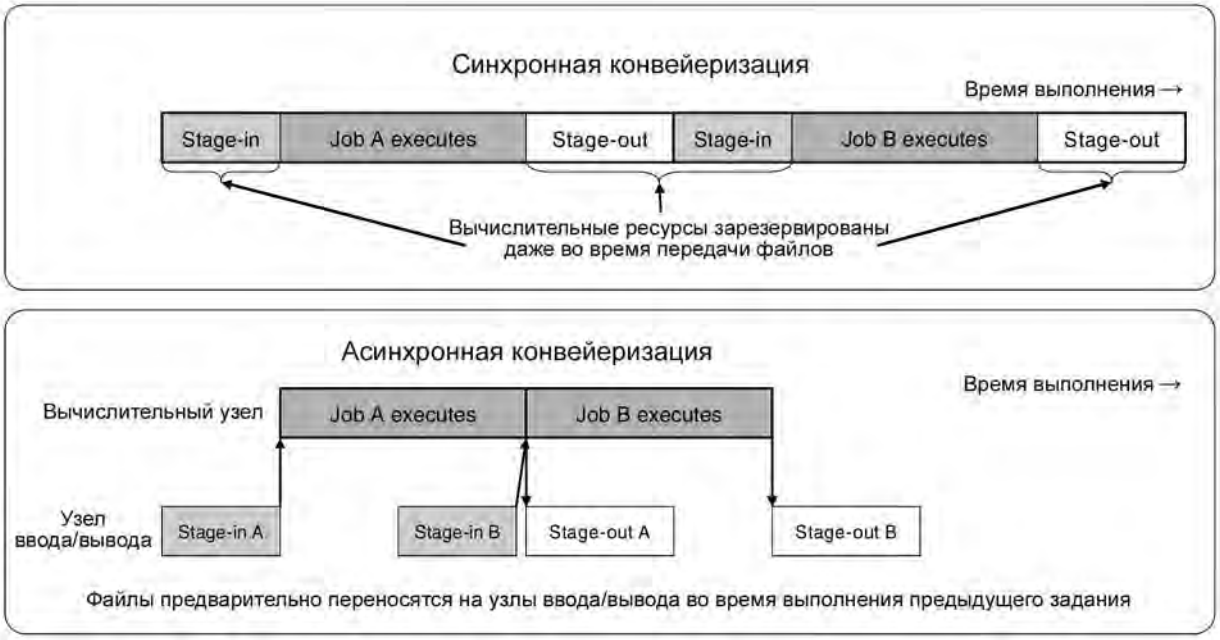

Рисунок 3
Использование асинхронной конвейеризации улучшают коэффициент использования

5. Заключение

Мы разработали программное обеспечение для управления работой К компьютера, чтобы ответить на вопросы, уникальные для ультра-крупномасштабной системы и для поддержки дальнейшего расширения масштабов системы в ближайшие годы.

Данная статья сосредоточена на усилиях, которые мы, таким образом, приняли для обеспечения работы ультра-крупномасштабной системы.

Дополнительные новые подходы будут необходимы для дальнейшего улучшения работоспособности системы.

Например, существует необходимость для инструментария, который может имитировать поведение системы при изменении параметров системы так, что системный администратор может выполнять настройку системы и планирование емкости.

Сочетания такого инструментария с программным обеспечением управления работой должно содействовать достижению стабильного и эффективного функционирования системы.