ИНФОРМАТИКА

2006 октябрь-декабрь № 4

ПАРАЛЛЕЛЬНЫЕ АРХИТЕКТУРЫ

УДК 004.9

В.В. Анищенко[1], Д.Б. Жаворонков[2], В.И. Махнач[1], Н.Н. Парамонов[1], О.П. Чиж[1]

ПРОЕКТИРОВАНИЕ И РАЗРАБОТКА СЕМЕЙСТВА ТИПОВЫХ ПЕРСОНАЛЬНЫХ КЛАСТЕРОВ ТРИАДА

Рассматриваются общие тенденции развития суперкомпьютерных технологий. Приводятся концептуальные принципы создания семейства типовых персональных кластеров Триада и формулируются основные требования к их параметрам и характеристикам. Отражаются состояние дел по созданию первых моделей персональных кластеров и перспективы их практического использования.

Введение

Востребованность работ по развитию в странах Союзного государства направления суперкомпьютерных технологий определяется отставанием этих стран от ведущих мировых держав в развитии и применении новейших наукоемких информационных технологий, нацеленных на решение сложных задач машиностроения, биотехнологии, геологоразведки, контроля окружающей среды, транспорта и связи, государственных, коммерческих, военных и других приложений. Применение суперкомпьютерных систем и наукоемких программных продуктов актуально также в таких приложениях, как банки данных, информационно-аналитические системы, ситуационные центры управления, системы управления в реальном масштабе времени, боевые информационно-управляющие системы и др. Развитие суперкомпьютерных средств и технологий было главной задачей программы Союзного государства «Разработка и освоение в серийном производстве семейства высокопроизводительных вычислительных систем с параллельной архитектурой (суперкомпьютеров) и создание прикладных программно-аппаратных комплексов на их основе» (шифр «СКИФ», 2000–2004 гг.).

Создание в рамках программы «СКИФ» суперкомпьютерных конфигураций СКИФ К-500 и СКИФ К-1000 [1] позволило в 2003–2005 гг. развернуть в Объединенном институте проблем информатики Национальной академии наук Беларуси (ОИПИ НАН Беларуси) фронт работ по практическому использованию суперкомпьютерных технологий. Приобретение суперкомпьютерных конфигураций не по карману не только рядовым пользователям, но и крупным предприятиям. Для решения этой проблемы в ОИПИ НАН Беларуси был создан суперкомпьютерный центр коллективного пользования с возможностью удаленного доступа к его вычислительным ресурсам. Результаты комплексной реализации программы «СКИФ» позволили выйти на собственный путь развития конкурентоспособной высокопроизводительной вычислительной техники, уровень которой соответствует прогнозируемым требованиям со стороны широкой категории пользователей.

Тенденции развития информационных технологий диктуют необходимость широкого внедрения принципов параллельных вычислений для решения высокопроизводительных задач. Для этого требуется создание недорогих малогабаритных вычислительных комплексов, которые можно устанавливать в обычных рабочих помещениях или офисах. Переход на вычислительную технику с параллельной архитектурой вынуждает пользователей изменять весь привычный стиль взаимодействия с компьютерами. Меняется практически все: применяются другие языки программирования, видоизменяется большинство алгоритмов, от пользователей требуется предоставление многочисленных нестандартных и труднодобываемых характеристик решаемых задач, перестает быть дружественным интерфейс и т. п. Эти обстоятельства могут в значительной степени снизить эффективность использования новой и к тому же дорогой техники с параллельной архитектурой.

Изложенную проблему может решить появление сравнительно дешевых персональных кластеров, которые будут использоваться в офисах и лабораториях на рабочих местах научно-технических работников и других пользователей. Такая задача решается в ОИПИ НАН Беларуси и УП «НИИЭВМ» в рамках проекта «Исследование и разработка семейства типовых высокопроизводительных персональных аппаратно-программных вычислительных комплексов с параллельной архитектурой (персональных кластеров)» программы Союзного государства «Развитие и внедрение в государствах-участниках Союзного государства наукоемких компьютерных технологий на базе мультипроцессорных вычислительных систем» (шифр программы – «Триада»).

1. Общие тенденции развития суперкомпьютерных технологий

Магистральным путем развития современных суперкомпьютерных технологий является построение распределенных вычислительных систем с массовым параллелизмом. Как в научной среде, так и для нужд промышленности существует необходимость создания высокопроизводительной высоконадежной среды для разработки мобильных и масштабируемых приложений. С 1993 г. началась публикация списков Top500 самых мощных вычислительных систем в мире. В последнее время все настолько привыкли к постоянной борьбе за лидерство в списке Top500, что существенное внимание стали уделять лишь верхним строчкам. Места в списке в настоящее время во многом определяются бюджетом и политикой. По большому счету, необходимо лишь желание правительства создать или поддержать свой имидж как руководителя высокотехнологичного государства.

Основная же цель организации списков Top500 – предоставление надежного базиса для определения и отслеживания тенденций в области высокопроизводительных вычислений. Списки Top500 публикуются два раза в год.

Анализ данных, представленных в официальном рейтинге 26-й редакции Top500 [2], позволяет сделать вывод о динамичном обновлении списка. В эту редакцию списка входят 351 система 2005 г. выпуска (70,2 %); 94 системы, включая СКИФ К-1000, 2004 г. выпуска (18,8 %); 35 систем 2003 г. выпуска (7,0 %); 15 систем 2002 г. выпуска (3,0 %); 3 системы 2001 г. выпуска и по одной системе 2000 и 1999 гг. выпуска. Суперкомпьютер СКИФ К-1000 входит уже в четвертую подряд редакцию списка Top500, начиная с 24-й, что убедительно свидетельствует о перспективности принятых технических решений в рамках программы «СКИФ».

Существенными при реализации суперкомпьютерных технологий являются данные по архитектуре процессоров и суперкомпьютерных систем. Архитектура процессоров: скалярная – 97,2 %; векторная – 2,8 %. Архитектура суперкомпьютерных систем: кластерная – 360 систем (72,0 %); звездообразная – 36 систем (7,2 %); MPP – 104 системы (20,8 %).

Бурное развитие технологий HPC (High-Performance Computing) привело к естественной экспансии параллельной архитектуры (в основном кластерной) во все направления компьютерной отрасли: суперкомпьютеры, серверы, рабочие станции. Эта тенденция коснулась уже и самого массового звена средств вычислительной техники – персональных компьютеров. Как показал анализ тенденций развития компьютерных технологий, появились и уже становятся привычными термины «персональный суперкомпьютинг», «персональные кластеры» и даже «персональные суперкомпьютеры», создаются и соответствующие программно-технические средства [3, 4].

Кластерная революция поставила во главу угла соотношение цена/производительность, причем этот показатель приобретает особое значение для персонального суперкомпьютинга. Однако, чтобы персональные кластеры действительно смогли стать «персональными», они должны обеспечить, как минимум, все те возможности, благодаря которым «персоналки» и стали незаменимым вычислительным инструментом (и даже более чем инструментом) для массового пользователя:

– доступность по цене;
– дружественный интерфейс;
– развитое программное обеспечение (системное и прикладное);
– операционная среда ОС Microsoft Windows;

– небольшие габариты и, как следствие, возможность расположения непосредственно в рабочей зоне;

– включение непосредственно в розетку;

– небольшая потребляемая мощность;

– допустимый для офисных помещений уровень шума;

– круглосуточный режим работы без внешних устройств охлаждения.

Безусловно, на данном этапе развития компьютерных технологий персональные кластеры, донося до широкого пользователя возможности HPC, могут уступать по ряду показателей ПЭВМ с традиционной архитектурой. Но, как говорится, лед тронулся…

2. Принципы создания семейства персональных кластеров Триада

Основными принципами создания семейства персональных кластеров Триада являются:

– параллельная (кластерная) архитектура;

– программная совместимость с кластерами семейства СКИФ;

– работа с ОС Linux и ОС Microsoft Windows Compute Cluster Server 2003;

– использование перспективных конструкторско-технологических решений.

С учетом этих принципов предусматриваются следующие основные концептуальные технические характеристики семейства персональных кластеров Триада:

– реализация вычислительных узлов (ВУ) кластера на 64-разрядных платформах;

– конструктивная реализация на базовых конструктивных вычислительных модулях (БКВМ) кластеров семейства Триада;

– использование перспективных суперскалярных технологий обработки данных (многопотоковость, многоядерность, VLIW);

– использование новых внутренних шин (PCI-Express, Hyper-Transport);

– использование перспективных сетевых интерфейсов (Gigabit Ethernet, Infiniband, Myrinet и др.);

– использование энергосберегающих технологий для повышения плотности вычислительной мощности и уровня энергопотребления на единицу объема: в 1,5–2,0 раза больше по сравнению с БКВМ кластеров семейства СКИФ;

– уменьшение количества внешних связей между ВУ и сетевыми коммутаторами в 2,0–2,5 раза по сравнению с БКВМ кластерных систем семейства СКИФ;

– расширенные сервисные функции (мониторинг внутренней температуры, контроль работы системы вентиляции и др.);

– повышенная отказоустойчивость.

3. Архитектура персональных кластеров Триада

С учетом изложенных выше концептуальных принципов и областей применения в персональных кластерах Триада предусматривается архитектура с двухуровневой обработкой данных (рисунок).

Двухуровневая архитектура позволяет оптимизировать организацию параллельного счета задач как с крупноблочным (явным статическим или скрытым динамическим) параллелизмом, так и с конвейерным или мелкозернистым явным параллелизмом с большими потоками информации, требующими обработки в реальном режиме времени. Такая система включает:

– базовый (кластерный) архитектурный уровень;

– потоковый архитектурный уровень, реализующий модель потоковых вычислений.

Кластерный архитектурный уровень – это тесносвязанная сеть (кластер) ВУ, работающих под управлением ОС. Для организации параллельного выполнения прикладных задач на данном уровне должны использоваться:

– разработанная в рамках программы «СКИФ» для ОС Linux оригинальная система поддержки параллельных вычислений – Т-система, реализующая автоматическое динамическое распараллеливание программ;

– классические системы поддержки параллельных вычислений, обеспечивающие эффективное распараллеливание прикладных задач различных классов (как правило, задач с явным параллелизмом) MPI.

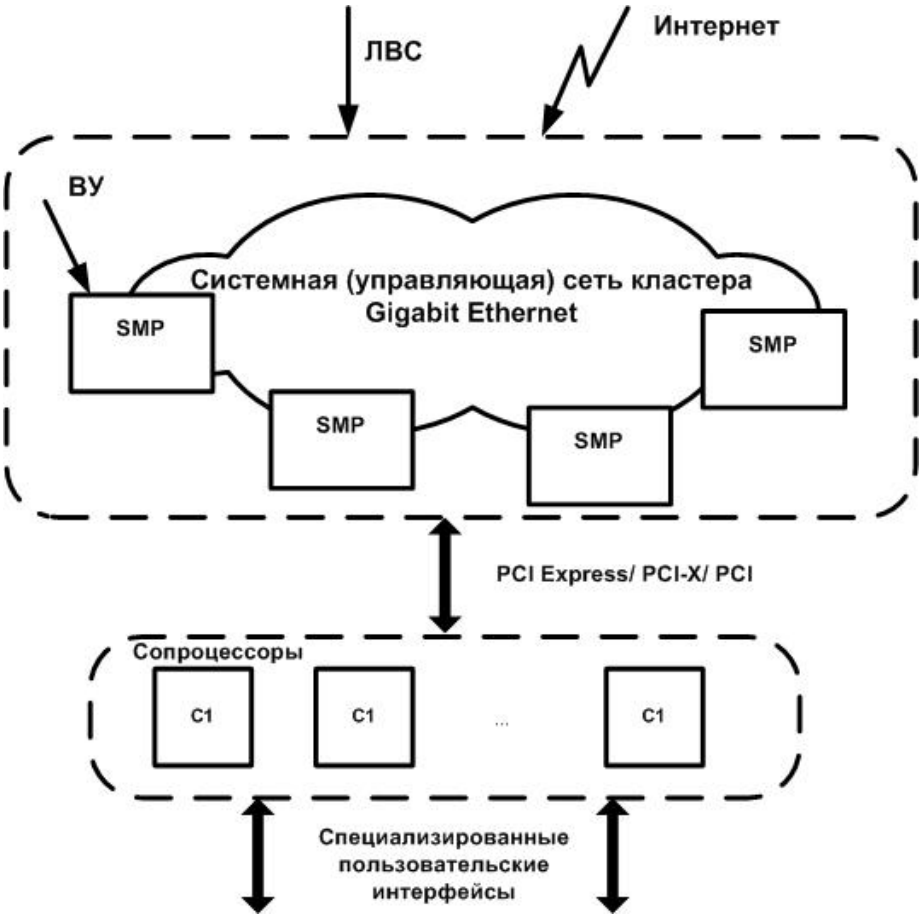

Рис. Двухуровневая архитектура персональных кластеров Триада

На кластерном уровне (КУ) с использованием Т-системы и MPI эффективно реализуются фрагменты со сложной логикой вычисления, с крупноблочным (явным статическим или скрытым динамическим) параллелизмом. Фрагменты же с простой логикой вычисления, с конвейерным или мелкозернистым явным параллелизмом, с большими потоками информации, требующими обработки в реальном режиме времени, на кластерных конфигурациях реализуются менее эффективно. Для организации параллельного исполнения задач с подобными фрагментами наиболее адекватна модель потоковых вычислений.

Кластерная архитектура является открытой и масштабируемой, т. е. не предъявляет жестких требований к программно-аппаратной платформе узлов кластера, топологии вычислительной сети, конфигурации и диапазону производительности суперкомпьютеров.

Базовая архитектура персональных кластеров, как и моделей семейства суперкомпьютеров СКИФ, реализуется на классических кластерах из ВУ на основе компонент широкого применения (стандартных микропроцессоров, модулей памяти, жестких дисков и материнских плат, в том числе с поддержкой SMP).

Системная сеть кластера объединяет узлы КУ в кластер. Данная сеть поддерживает масштабируемость КУ суперкомпьютера, а также пересылку и когерентность данных во всех ВУ КУ суперкомпьютера. Системная сеть кластера строится на основе специализированных высокоскоростных линков класса Gigabit Ethernet, Myrinet, InfiniBand и др., предназначенных для эффективной поддержки кластерных вычислений на уровне ОС и систем организации параллельных вычислений (Т-системы, MPI).

Вспомогательная сеть суперкомпьютера с протоколом TCP/IP объединяет узлы КУ в обычную локальную сеть TCP/IP LAN. Данная сеть может быть реализована на основе широко используемых сетевых технологий класса Fast Ethernet, Gigabit Ethernet и др. Она предназначена для управления системой, подключения рабочих мест пользователей, интеграции суперкомпьютера в локальную сеть предприятия и/или в глобальные сети. Кроме того, данный уровень может быть использован и системой организации параллельных кластерных вычислений (Т-системой, MPI) для вспомогательных целей. Основные потоки информации, возникающие при организации параллельных кластерных вычислений, передаются через системную сеть кластера.

В большинстве случаев аппаратура системной сети позволяет без ущерба для реализации кластерных вычислений поддержать на этой же аппаратуре реализацию сети TCP/IP. В этих случаях аппаратные части системной и вспомогательной сетей могут быть совмещены.

На потоковом уровне могут эффективно выполняться фрагменты прикладной проблемы с простой логикой вычисления, с конвейерным или мелкозернистым явным параллелизмом, с большими потоками информации, требующими обработки в реальном режиме времени. Потоковый уровень реализуется в виде специализированных процессоров. На потоковом уровне может быть эффективно реализован высокоскоростной потоковый обмен со стандартной компьютерной периферией и/или с нестандартными устройствами-датчиками, например с видеокамерами и другими приборами. Функции управления может выполнять один из ВУ кластера.

4. Базовое (системное) программное обеспечение персональных кластеров Триада

В семействе персональных кластеров Триада должна быть реализована возможность работы под управлением как ОС Linux, так и Windows Compute Cluster Server 2003 [5].

Использование ОС Windows Compute Cluster Server 2003 (CCS) фирмы Microsoft позволяет обеспечить на кластерных структурах привычную для пользователей ПЭВМ операционную среду Windows при выполнении высокопроизводительных вычислений.

Windows Compute Cluster Server 2003 может быть установлен с помощью стандартных технологий Windows. Microsoft Message Passing Interface (MS-MPI) полностью совместим со стандартами кодов MPI2. Интеграция с Active Directory обеспечивает сетевую безопасность, а использование Microsoft Management Console предоставляет привычный интерфейс администрирования и планирования.

Windows Compute Cluster Server 2003 поддерживает следующие базовые технологии:

– 64-разрядные компьютеры и ВУ кластеров;

– Message Passing Interface v2 (MPI2);

– сетевые технологии Gigabit Ethernet, Ethernet over RDMA, Infiniband и Myrinet.

Перечислим ключевые характеристики, подчеркивающие достоинства Windows Compute Cluster Server 2003 при организации высокопроизводительных вычислений.

Упрощенная организация и управление кластером. CCS обеспечивает быструю установку узлов и конфигурирование кластера, простую процедуру введения дополнительных узлов. Средства мониторинга и планирования CCS предоставляют масштабируемую среду управления, включая:

– автоматизированную установку (setup) с минимальными подсказками пользователя;

– удобное конфигурирование сети, дистанционные сервисы инсталляции, средства управления узлами кластера и обеспечения его информационной безопасности;

– интегрированный программный стек с встроенным планировщиком заданий, а также стек MPI для оптимизации межузловых обменов.

Естественная интеграция с Windows-инфраструктурой. CCS естественно интегрируется с существующей Windows-инфраструктурой, предоставляя средства и технологии управления кластером и обеспечения информационной безопасности. В частности, существующие средства используются следующим образом:

Active Directory – для упрощения аутентификации и установки ключей безопасности;

Microsoft Systems Management Server – для управления узлами кластера;

Remote Installation Services – в режиме дистанционной инсталляции узлов кластера;

Microsoft Operations Manager – для управления системой и заданиями;

Microsoft Management Console – в качестве системного (snap-in) инструментария.

Широкая поддержка приложений. Интегрированный программный стек обеспечивает поддержку рынка HPC, позволяя разработчику создавать широкий спектр приложений и инструментов.

Знакомая среда разработки. При разработке CCS-приложений используются опыт и знания, базирующиеся на технологиях Windows. Microsoft Visual Studio – наиболее распространенная в отрасли среда разработки, а Visual Studio 2005 включает поддержку HPC-приложений, например параллельный отладчик. CCS включает интегрированный MPI-уровень, базирующийся на спецификациях международного стандарта MPI2, что значительно упрощает адаптацию существующих параллельных приложений.

5. Основные требования к параметрам и характеристикам персональных кластеров Триада

Кластеры семейства Триада должны быть реализованы на БКВМ, разрабатываемых в рамках программы Союзного государства «Триада». Предполагается создание ряда БКВМ: Tower, стоечных и т. д.

Каждый БКВМ состоит из базового конструктивного модуля (БКМ) и базовых вычислительных модулей (БВМ). БВМ – это конструктивная реализация ВУ кластера. Для конструктивов Tower (БКВМ-Т) предполагается, что каждый узел кластера (БВМ) может содержать одно-, четырехпроцессорную системную плату SMP, микропроцессоры с тактовой частотой не ниже 1,6 ГГц, адаптеры для подключения к системной и вспомогательной сети, одну-две платы спецпроцессоров и иметь следующие параметры:

– емкость основной памяти не ниже 2 Гбайт;

– потребляемая мощность не более 400 Вт;

– теоретическая пиковая производительность каждого узла кластера не менее 3,2 Гфлопс, реальная производительность каждого опытного образца кластера будет уточнена в процессе разработки;

– дисковая память должна быть реализована с использованием современных типовых жестких дисков или средств файловых серверов, требования к параметрам дополнительной дисковой памяти должны быть уточнены в процессе разработки кластера;

– в кластерах должна быть обеспечена возможность подключения средств инженерного пульта, обеспечивающих управление ВУ, коммуникационными средствами и дисковой памятью;

– кластеры семейства Триада должны сохранять конструкцию, внешний вид и работоспособность в процессе и после воздействия на них следующих климатических факторов:

– пониженной рабочей температуры (10 °С);

– повышенной рабочей температуры (35 °С);

– повышенной относительной влажности воздуха (до 80 % при температуре 25 °С).

Кластеры семейства Триада служат для обработки открытой информации. При наличии конкретного потребителя возможно создание модификаций, предназначенных для обработки закрытой информации, имеющих средства защиты от несанкционированного доступа к хранимой и обрабатываемой информации, а также от несанкционированного доступа за счет побочных электромагнитных излучений.

Заключение

На первом этапе работ по созданию семейства персональных кластеров Триада (в четвертом квартале 2005 г.) проведен анализ направлений развития кластерных вычислительных систем, включая архитектурные и конструкторско-технологические решения, сформулированы основные принципы создания персональных кластеров. Выделены основные научно-технические решения суперкомпьютеров СКИФ, соответствующие принципам создания персональных кластеров, сформулированы основные технические характеристики персональных кластеров.

Результаты проведенной работы использованы в 2006 г. при разработке комплектов рабочей документации и создании экспериментальных образцов первых двух моделей семейства персональных кластеров Триада. Экспериментальные образцы будут использованы как высокопроизводительные вычислительные устройства в аппаратно-программном комплексе слежения в реальном масштабе времени за движущимися объектами и в распределенной телемедицинской системе реального времени по цифровой флюорографии. Создание этих проблемно-ориентированных программно-аппаратных комплексов на базе экспериментальных образцов позволит отработать основные технические принципы создания семейства персональных кластеров в реальных режимах функционирования.

Список литературы

1. Суперкомпьютерные конфигурации СКИФ / С.В. Абламейко [и др.]. – Минск: ОИПИ НАН Беларуси, 2005. – 170 с.

2. TOP500 Supercomputer Sites [Electronic resource]. – Mode of access: http://www.top500.org.

3. Tyan unleashes 16-core personal supercomputer [Electronic resource]. – Mode of access: http://www.regharrdware.co.uk/2006/06/07/tyan_unveils_typhoon.

4. Saturn Personal Clusters [Electronic resource]. – Mode of access: http://www.rocketcalc.com/?loc= saturn.

5. High-Performance Computing with Windows Compute Cluster Server 2003 [Electronic resource]. – Mode of access: http://www.microsoft.com/windowsserver2003/ccs/default.mspx.

Поступила 04.06.06

[1]*Объединенный институт проблем информатики НАН Беларуси, Минск, Сурганова, 6 e-mail: nick@newman.bas-net.by*

[2]*УП «НИИЭВМ», Минск, М. Богдановича, 155 e-mail: dmitry@niievm.by*

V.V. Anishchanka, D.B. Javoronkov, V.I. Mahnach, N.N. Paramonov, O.P. Tchij

DEVELOPMENT OF THE FAMILY OF TYPICAL PERSONAL CLUSTERS «TRIADA»

General principles of typical personal clusters «Triada» development are proposed. The architecture of clusters and system operating software are presented. The state of affairs on creation of first models of personal clusters and prospects of their practical use are reflected.