УДК 519.68

ПРИНЦИПЫ ПОСТРОЕНИЯ СУПЕРКОМПЬЮТЕРОВ СЕМЕЙСТВА «СКИФ»

С.В. Абламейко[1], С.М. Абрамов[2], В.В. Анищенко[1], Н.Н. Парамонов[1], О.П. Чиж[1]
[1]Объединенный институт проблем информатики
Национальной академии наук Беларуси, Минск
[2]Институт программных систем Российской академии наук, Переславль-Залесский

Рассматриваются базовые архитектурные решения, основные принципы и решения в части базового ПО и аппаратных средств, а также подходы к созданию моделей суперкомпьютеров «СКИФ».

Введение

Главной целью 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» («СКИФ») [1-3] является возрождение компьютерной отрасли двух стран, промышленное производство ряда программно-совместимых моделей суперкомпьютеров с широким спектром производительности - до триллионов операций в секунду. Для достижения этой цели в рамках программы реализуется комплексный подход, включающий Концепцию создания моделей семейства суперкомпьютеров «СКИФ».

Концепция отражает основополагающие принципы создаваемых по программе суперкомпьютерных систем:

- базовые архитектурные решения;
- основные принципы и решения в части базового (общесистемного) ПО;
- основные принципы и решения в части аппаратных средств;
- идеологию создания моделей семейства суперкомпьютеров;
- основные принципы разработки конструкторской документации, проведения испытаний суперкомпьютерных систем и организации их серийного производства;
- общую схему реализации прикладных суперкомпьютерных конфигураций.

Основополагающими архитектурными принципами создания суперкомпьютерных конфигураций «СКИФ» являются:

- базовая кластерная архитектура;
- иерархические кластерные конфигурации (метакластеры);
- универсальная двухуровневая архитектура.

1. Базовая кластерная архитектура

Концепция создания моделей семейства суперкомпьютеров «СКИФ» базируется на масштабируемой кластерной архитектуре, реализуемой на ***классических кластерах из вычислительных узлов*** (см. рис. 1) ***на основе компонент широкого применения*** (стандартных микропроцессоров, модулей памяти, жестких дисков и материнских плат, в том числе с поддержкой SMP).

Кластерный архитектурный уровень – это тесносвязанная сеть (кластер) вычислительных узлов, работающих под управлением ОС Linux - одного из клонов широко используемой многопользовательской универсальной операционной системы

UNIX. Для организации параллельного выполнения прикладных задач на данном уровне используются:

разрабатываемая в рамках Программы 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, PVM, Norma, DVM и др. В семействе суперкомпьютеров «СКИФ» в качестве базовой классической системы поддержки параллельных вычислений выбран MPI, что не исключает использование других средств.

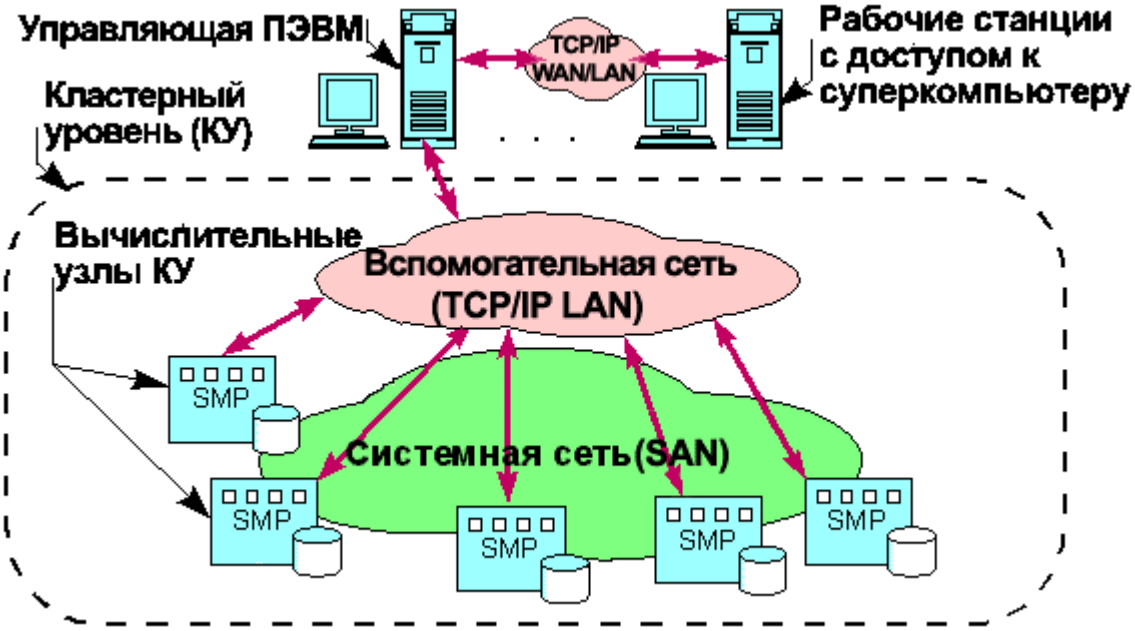

Рис. 1. Кластерная архитектура

На кластерном уровне 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 (data-flow).

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

Для организации взаимодействия вычислительных узлов суперкомпьютера в его составе используются различные сетевые (аппаратные и программные) средства, в совокупности образующие две системы передачи данных:

- системная сеть кластера (СС) или ***System Area Network (SAN)*** объединяет узлы кластерного уровня в кластер. Данная сеть поддерживает масштабируемость

кластерного уровня суперкомпьютера, а также пересылку и когерентность данных во всех вычислительных узлах кластерного уровня суперкомпьютера. Системная сеть кластера строится на основе специализированных высокоскоростных линков класса SCI, Myrinet, cLan, Infiniband и др., предназначенных для эффективной поддержки кластерных вычислений и соответствующей программной поддержки на уровне ОС Linux и систем организации параллельных вычислений (Т-система, MPI);

- вспомогательная сеть суперкомпьютера (ВС) с протоколом TCP/IP объединяет узлы кластерного уровня в обычную (TCP/IP) локальную сеть ***(TCP/IP LAN)***. Данная сеть может быть реализована на основе широко используемых сетевых технологий класса Fast Ethernet, Gigabit Ethernet и др. Данная сеть 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-система, MPI) для вспомогательных целей (основные потоки информации, возникающие при организации параллельных кластерных вычислений, передаются через системную сеть кластера).

Кластерные конфигурации на базе только вспомогательной сети TCP/IP без использования дорогостоящих специализированных высокоскоростных линков класса SCI могут быть реализованы в рамках семейства «СКИФ» в виде самостоятельных изделий ***(TCP/IP кластеры)***. Программное обеспечение таких кластеров – ОС Linux, Т-система и соответствующая реализация MPI. Реализация сравнительно недорогих TCP/IP кластеров на базе ***«масштабирования вниз»*** архитектурных решений «СКИФ» существенно расширяет область применения результатов реализации программы.

Кластерные конфигурации на базе только вспомогательной сети могут быть реализованы как на базовых конструктивах «СКИФ», так и путем кластеризации имеющихся у пользователей ПЭВМ ***(«персональные кластеры» или «супер ПЭВМ»)***.

2. Базовое (системное) программное обеспечение суперкомпьютеров

В качестве базовой операционной системы (ОС) в универсальном кластерном суперкомпьютере используется операционная система Linux. ОС Linux является одной из самых надежных, эффективных и перспективных операционных систем, которую сегодня многие коммерческие и государственные организации выбирают в качестве базовой для приложений и перспективных разработок в области параллельных вычислений. ОС Linux распространяется свободно (бесплатно) с исходными текстами. Это дает возможность модифицировать и вносить изменения, необходимые для реализации поставленной задачи.

Функциональные возможности ОС Linux и ее утилит развиваются огромной армией добровольных программистов-разработчиков, что обеспечивает непрерывность ее тестирования и корректировки ошибок в исходных текстах. Распространение ОС Linux не подвержено каким-либо ограничениям каких-либо стран или фирм. ОС Linux является открытой, то есть она реализована не только для платформ класса IBM PC, но и для многих других аппаратных платформ.

3. Иерархические кластерные конфигурации (метакластеры)

Отдельные кластеры могут быть объединены в единую кластерную конфигурацию – кластер высшего уровня или ***метакластер*** (Metacluster). Метакластерный принцип позволяет создавать распределенные метакластерные

конфигурации на базе локальных или глобальных сетей передачи данных. При этом, естественно, уменьшается степень связности подкластеров метакластерной конфигурации.

Системное программное обеспечение метакластера обеспечивает возможность реализации ***гетерогенных систем***, включающих подкластеры различной архитектуры на различных программно-аппаратных платформах.

Одним из перспективных программных продуктов, с использованием которого возможна реализация метакластерных конфигураций (по крайней мере, простой топологии типа point-to-point) на подкластерах с различными программно-аппаратными платформами, является ***IMPI*** (Interoperable Message Passing Interface). IMPI реализует стандартизованный протокол, обеспечивающий взаимодействие различных реализаций MPI. Это позволяет выполнять общую задачу на различной аппаратуре с использованием настраиваемых поставщиком (vendor-tuned) различных реализаций MPI на каждом узле кластерной конфигурации соответствующего уровня иерархии. Такая возможность полезна в случаях, когда объем вычислений задачи слишком велик для одной системы или когда разные части задачи оптимально выполнять на разных реализациях MPI.

IMPI определяет только протоколы, необходимые для взаимодействия различных реализаций MPI, а также может использовать собственные высокопроизводительные протоколы этих реализаций. Существуют свободно распространяемые (открытые) версии IMPI, например, на базе LAM/MPI.

Преимущества и цели реализации иерархической (метакластерной) архитектуры. Реализация архитектурных принципов иерархической организации суперкомпьютерных метакластерных конфигураций позволит решить важнейшие для создания моделей семейства суперкомпьютеров «СКИФ» задачи:

- обеспечение реально достижимой и экономически эффективной масштабируемости архитектурных решений. Это особенно важно для решения ключевой задачи программы: создание моделей суперкомпьютеров, позволяющих перекрыть широкий диапазон производительности и областей применения - от моделей суперкомпьютеров среднего класса (10-100 ГФлопс) до вычислительных систем с массовым параллелизмом сверхвысокой производительности (триллионы операций в секунду);
- создание единого информационного пространства участников программы, а, в перспективе, объединение научных сетей России и Беларуси, на базе распределенных сетевых суперкомпьютерных метакластерных конфигураций;
- обеспечение живучести суперкомпьютерных систем;
- объединение суперкомпьютерных конфигураций с разными архитектурными и программно-аппаратными платформами (гибридная метакластерная архитектура) в единую метакластерную суперкомпьютерную систему;
- создание глобальных сетевых конфигураций с гибридной метакластерной архитектурой терафлопового диапазона. Такие метакластеры могут быть созданы путем объединения кластеров «СКИФ» с другими существующими в РБ и РФ кластерными конфигурациями (например, в МГУ, Межведомственном суперкомпьютерном центре РФ и РАН и др.).

4. Универсальная двухуровневая архитектура

Для оптимизации организации на суперкомпьютерах «СКИФ» 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, Концепция предусматривает возможность реализации ***универсальной двухуровневой архитектуры*** суперкомпьютеров (см. рис. 2):

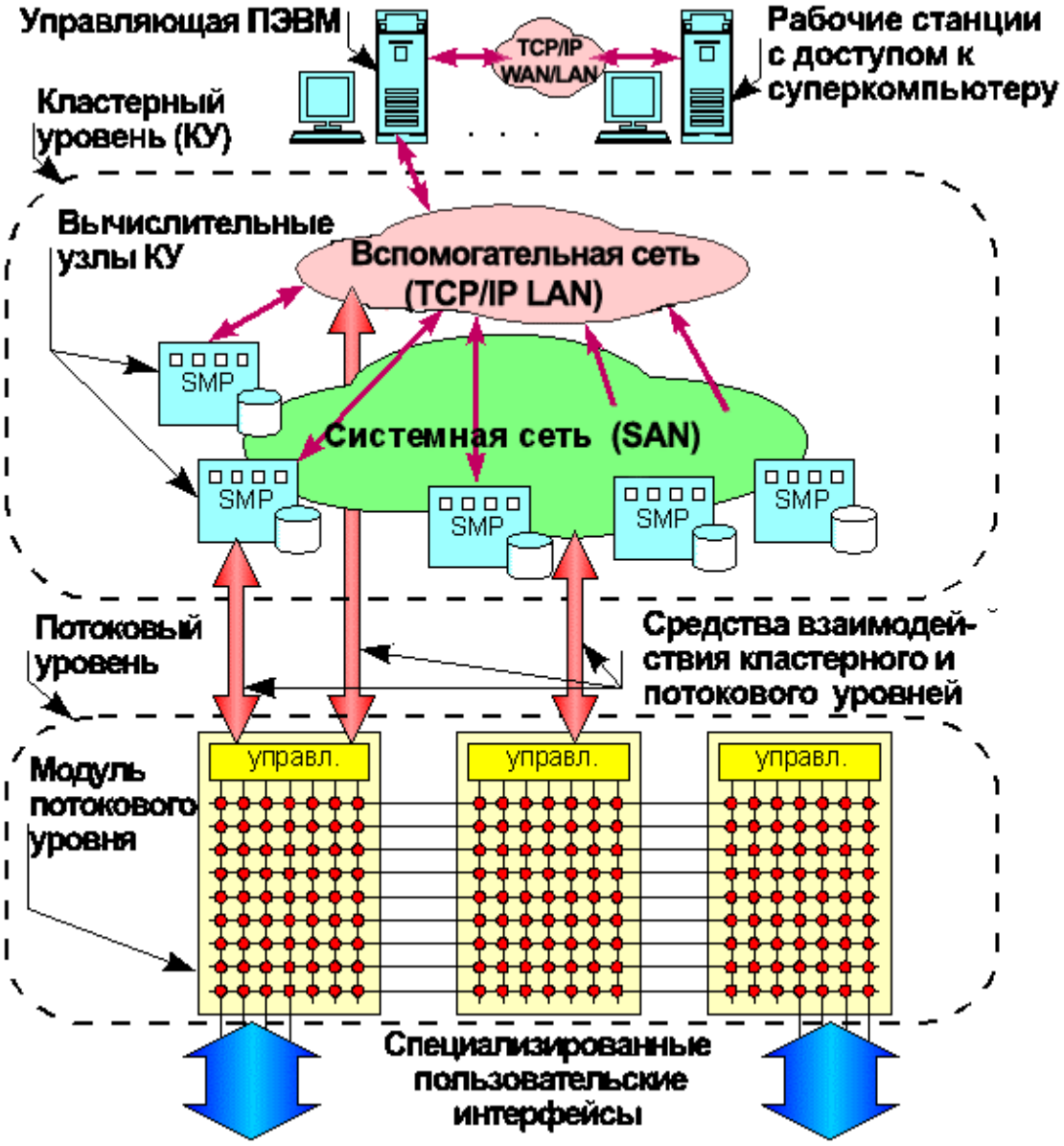

Рис. 2. Универсальная двухуровневая архитектура

- 1-й уровень - базовый (кластерный) архитектурный уровень;
- 2-й уровень - потоковый архитектурный уровень, реализующий модель потоковых вычислений (data-flow).

Концепция предусматривает реализацию потокового архитектурного уровня как на базе однородной вычислительной среды (ОВС) с использованием оригинальных СБИС ОВС, разрабатываемых в рамках программы, так и на базе других (альтернативных) структурных и технических решений (например, на базе нейроструктур, FPGA типа XILINX, ALTERA и др.). По сути, вычислительные модули потокового уровня являются сопроцессорами вычислительных ресурсов кластерной конфигурации.

Предпосылкой объединения двух программно-аппаратных решений (кластерного и потокового) для организации параллельной обработки в рамках одной вычислительной системы, является то, что эти два подхода своими сильными сторонами компенсируют недостатки друг друга. Тем самым, в общем случае, каждая прикладная проблема может быть разбита на:

- фрагменты со сложной логикой вычисления, с крупноблочным (явным статическим или скрытым динамическим) параллелизмом, эффективно реализуемые на кластерном уровне с использованием Т-системы и других (классических) систем поддержки параллельных вычислений;
- фрагменты с простой логикой вычисления, с конвейерным или мелкозернистым явным параллелизмом, с большими потоками информации, требующими обработки в реальном режиме времени, эффективно реализуемые на потоковом уровне.

5. Особенности архитектуры семейства суперкомпьютеров «СКИФ»

Предложенная многоуровневая схема реализации архитектурных принципов обладает рядом особенностей и преимуществ (по сравнению с аналогичными разработками), позволяющих достичь мировой уровень в суперкомпьютерной отрасли.

В части Т-системы - обеспечивается автоматическое динамическое распараллеливание программ, что освобождает программиста от большинства трудоемких аспектов разработки параллельных программ, свойственных различным системам ручного статического распараллеливания:

– обнаружение готовых к выполнению фрагментов задачи (процессов);

– их распределение по процессорам;

– их синхронизацию по данным.

Все эти (и другие) операции выполняются в Т-системе автоматически и в динамике (во время выполнения задачи). Тем самым при более низких затратах на разработку параллельных программ обеспечивается более высокая их надежность.

По сравнению с использованием распараллеливающих компиляторов, Т-система обеспечивает более глубокий уровень параллелизма во время выполнения программы и более полное использование вычислительных ресурсов мультипроцессоров. Это связано с принципиальными алгоритмическими трудностями (алгоритмически неразрешимыми проблемами), не позволяющими во время компиляции (в статике) выполнить полный точный анализ и предсказать последующее поведение программы во время счета.

Кроме указанных выше принципиальных преимуществ Т-системы перед известными сегодня методами организации параллельного счета, в реализации Т-системы имеется ряд технологических находок, не имеющих аналогов в мире:

– ***реализация понятия «неготовое значение»*** и поддержка корректного выполнения некоторых операций над неготовыми значениями. Тем самым поддерживается возможность выполнение счета в некотором процессе-потребителе в условиях, когда часть из обрабатываемых им значений еще не готова, т. е. не вычислена в соответствующем процессе-поставщике. Данное техническое решение обеспечивает обнаружение более глубокого параллелизма в программе;

– ***оригинальный алгоритм динамического автоматического распределения процессов по процессорам.*** Данный алгоритм учитывает особенности неоднородных распределенных вычислительных сетей. По сравнению с известными алгоритмами динамического автоматического распределения процессов по процессорам (например, с диффузионным алгоритмом и его модификациями), алгоритм Т-системы имеет

существенно более низкий трафик межпроцессорных передач. Тем самым, Т-система обеспечивает снижение накладных расходов на организацию параллельного счета и предъявляет менее жесткие требования к пропускной способности аппаратуры объединения процессорных элементов в кластер.

В части потокового уровня - архитектура вычислительных модулей потокового уровня позволяет использовать естественный параллелизм решаемой задачи вплоть до битового уровня, то есть уровня структуры обрабатываемых данных, а также позволяет строить конвейеры произвольной глубины. Потоковый уровень предоставляет возможность одновременной обработки множества независимых некогерентных потоков.

Фактически, при решении конкретной функции или самостоятельной задачи, на вычислительных модулях потокового уровня путем ввода соответствующей программы организуется ***спецпроцессор***, реализующий решаемую функцию или задачу с наибольшей эффективностью. На матрице модулей потокового уровня одновременно могут решаться несколько независимых задач и функций, причем механизм перезагрузки сегментов потокового уровня позволяет перезагружать часть матрицы без остановки выполнения еще незавершенных задач. Потоковый уровень обладает высокой гибкостью и перестраиваемостью, в частности, полной аппаратной и программной масштабируемостью, что позволяет строить на его основе вычислительные системы с большим быстродействием. Производительность матрицы модулей потокового уровня, теоретически, растет линейно с увеличением рабочей частоты поля и площади вычислительной матрицы.

Вычислительные модули потокового уровня позволяют создавать системы с высоким уровнем надежности и отказоустойчивости, эффективно реализовывать нейросетевые алгоритмы.

Заключение

Предложенные архитектурные принципы позволяют эффективно реализовывать любые виды параллелизма. ***Архитектура является открытой и масштабируемой***, то есть не накладывает жестких ограничений к программно-аппаратной платформе узлов кластера, топологии вычислительной сети, конфигурации и диапазону производительности суперкомпьютеров. Вычислительные системы, создаваемые на базе основополагающих концептуальных архитектурных принципов могут оптимально решать как классические вычислительные задачи математической физики и линейной алгебры, так и специализированные задачи обработки сигналов, моделирования виртуальной реальности, задачи управления сложными системами в реальном времени и другие приложения.

Литература

1. Разработка и опыт эксплуатации суперкомпьютеров семейства «СКИФ» / С.М. Абрамов, В.В. Анищенко, Н.Н. Парамонов, О.П. Чиж // Информационные системы и технологи. Мат. I междунар. конф. IST’2002 (5 – 8 ноября 2002 г.). Мн.: Изд-во БГУ, 2002. – Ч. 2. – С. 115-117.

2. Кластерные системы семейства суперкомпьютеров «СКИФ» / С.М. Абрамов, А.И. Адамович, М.Р. Коваленко и др. // Научный сервис в сети Интернет: Тр. Всерос. науч. конф. (22–27 сент. 2003 г., Новороссийск). – М.: Изд-во МГУ, 2003. – С.147-151.

3. Абламейко С.В., Абрамов С.М. Основные результаты суперкомпьютерной программы «СКИФ» Союзного государства // АКИИ’03: Третий расширенный семинар «Использование методов искусственного интеллекта в высокопроизводительных вычислениях и в аэрокосмических исследованиях». – М.: Физматлит, 2003. – С. 135-140.