СУПЕРКОМПЬЮТЕРНЫЕ КЛАСТЕРНЫЕ КОНФИГУРАЦИИ "СКИФ"

С.М. Абрамов, В.В. Анищенко, Н.Н. Парамонов

Суперкомпьютерные конфигурации "СКИФ" создаются в рамках 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 программы - "СКИФ"). Программа "СКИФ" выполняется (с учетом продления) в 2000-2004 гг. Государственные заказчики программы - Национальная академия наук Беларуси и Министерство образования и науки РФ. Головные исполнители программы - "Объединенный институт проблем информатики" (ОИПИ) НАН Беларуси и Институт программных систем РАН. В реализации программы принимают участие около 20 предприятий РБ и РФ. Система программных мероприятий включает 21 задание, которые предусматривают работы по созданию элементной базы, базовых конструктивных модулей, системного программного обеспечения (ПО) и законченных прикладных систем.

Главная цель Программы: возрождение компьютерной отрасли двух стран, промышленное производство семейства программно-совместимых установок с широким спектром производительности - до триллионов операций в секунду.

Важнейший практический результат четырех лет выполнения Программы - выпуск 12 образцов кластерных установок, начиная от двух "Первенцев" с пиковой производительностью 20 GFlops в секунду до суперкомпьютера "СКИФ К-500" 717 GFlops (рис. 1).

Производительность пиковая (и на задаче Linpack)	**716.8 (474.2) Gflops**
Тип процессора	**Intel Xeon 2.8 Ghz**
Число вычислительных узлов	**64**
Число процессоров	**128**
Оперативная память	**64 × 2 = 128 GB**
Дисковая память	**64 × 60 = 3840 GB**
Системная сеть	**3D-тор, 4×4×4, SCI, D336**
Вспомогательные сети	**2 × Gigabit Ethernet**
Конструктив узла (форм-фактор)	**1U**
Компоновка: 4 шкафа 20U с вычислительными узлами, 2 шкафа 20U со вспомогательным оборудованием (UPS, NAS, KVM и др.)	
Дополнительно: сервисная сеть RS-485 (On/Off, Reset, сериальная консоль), NAS 800 GB RAID-5, UPS.	

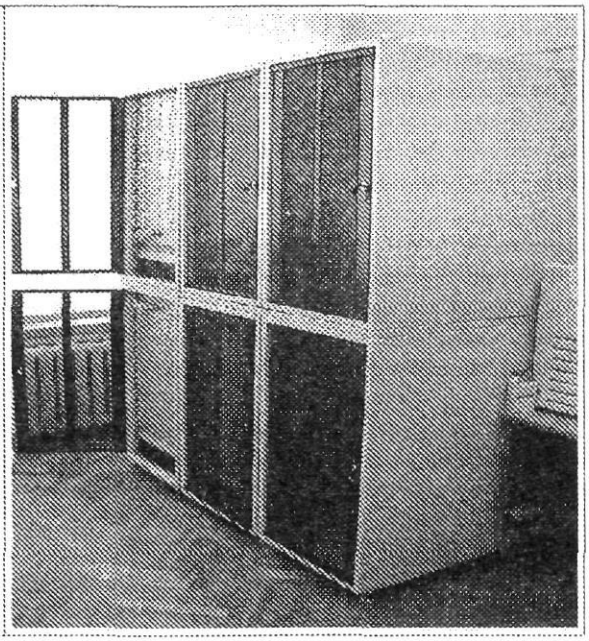

Рис. 1 Суперкомпьютерная конфигурация "СКИФ К-500" и ее основные технические характеристики.

В данной работе основное внимание уделено описания суперкомпьютерных конфигураций семейства "СКИФ" - установок, которые либо включены в список пятисот самых мощных компьютеров в мире (Top-500), либо, как мы надеемся, будут включены в этот список в ближайшее время.

Суперкомпьютер "СКИФ К-500", сентябрь 2003 года

Кластер "СКИФ К-500" создан в 2003 году для отработки принципов построения моделей суперкомпьютеров "СКИФ" с сверхвысокой производительностью, что приближается к терафлопному диапазону.

Осенью 2003 г. "СКИФ К-500" на тесте LinPack показал производительность 425.2 GFlops и был включен под номером 407 в 22-ой выпуск списка 500 самых производительных компьютерных систем в мире Top-500. Позже в результате тщательной настройки этот показатель был доведен до 474.2 GFlops.

Включение "СКИФ К-500" в список Top-500 означает достижение в 2003 году важного политического эффекта Программы - Республика Беларусь и Россия наравне с США, Японией и еще несколькими странами стали обладателями критической суперкомпьютерной технологии, повысив престиж Союзного государства, как разра-

ботчика таких технологий (см. сайты, отражающие ход выполнения и реализации программы Союзного государства "СКИФ" - http://www.skif.bas-net.by и http://skif.pereslavl.ru).

"СКИФ К-500" предназначен для решения научно-технических задач и особенно эффективен при решении задач с интенсивным межузловым обменом. Структурная схема "СКИФ К-500" представлена на рис. 2.

Пиковая производительность каждого узла - 11,2 Gflops. В кластере "СКИФ К-500" используется сеть SCI с топологией трехмерного тора 4x4x4 (связи L0, L1, L2 на рис. 2): блоки по 16 узлов в четырех отдельных шкафах соединены в двухмерный тор (L0, L1), шкафы связаны с помощью 16 колец линий (L2) третьего измерения тора. При передаче MPI-сообщений сеть SCI обеспечивает малую задержку (3 мкс) и высокую скорость обмена - 263 Мбайт/с.

Вспомогательная сеть Gigabit Ethernet предназначена для загрузки программ, данных, управления и мониторинга: топология "звезда", к коммутатору, соединенному с управляющей ПЭВМ подключены 3 коммутатора остальных шкафов. Вычислительные узлы одного шкафа подключаются к своему коммутатору.

Вторая вспомогательная сеть Gigabit Ethernet соединяет дополнительную дисковую память (NAS-сервер, RAID5, 800 Гбайт) с коммутаторами каждого из 4-х шкафов, с управляющей ПЭВМ и с внешней локальной сетью.

Отдельный интерфейс Gigabit Ethernet предназначен для доступа к управляющей ЭВМ из внешней локальной сети.

Сервисная сеть (на базе RS-485) в "СКИФ К-500" предназначена для выполнения с управляющей ПЭВМ на произвольном вычислительном узле операций включения/выключения электропитания, аппаратного сброса, взаимодействия с узлом в консольном режиме: изменение BIOS, управление загрузкой ОС, любые команды ОС, чтение (в т.ч. и "посмертно") консольных сообщений.

Сервис KVM обеспечивает подключение любого из вычислительных узлов к сервисному набору из клавиатуры, видеомонитора, "мыши" (для сервисного обслуживания).

На "СКИФ К-500" установлено ПО в составе: ОС LINUX RED HAT с поддержкой SMP, MPI 1.2 фирмы Scali (SSP 3.0.1), система мониторинга и управления Flame, компиляторы C, C++, Fortran, пакеты для параллельного программирования (включая Т-систему), прикладные библиотеки и пакеты.

Суперкомпьютер "СКИФ К-1000"

В сентябре 2004 года планируется завершение разработки суперкомпьютерной установки "СКИФ К-1000", которая создается с целью комплексной реализации принципов построения моделей суперкомпьютеров "СКИФ" Ряда 2 кластерного уровня с массовым параллелизмом сверхвысокой производительности (триллионы операций в секунду) на базе новых перспективных 64-разрядных процессоров и новых перспективных сетевых средств для интеграции вычислительных узлов.

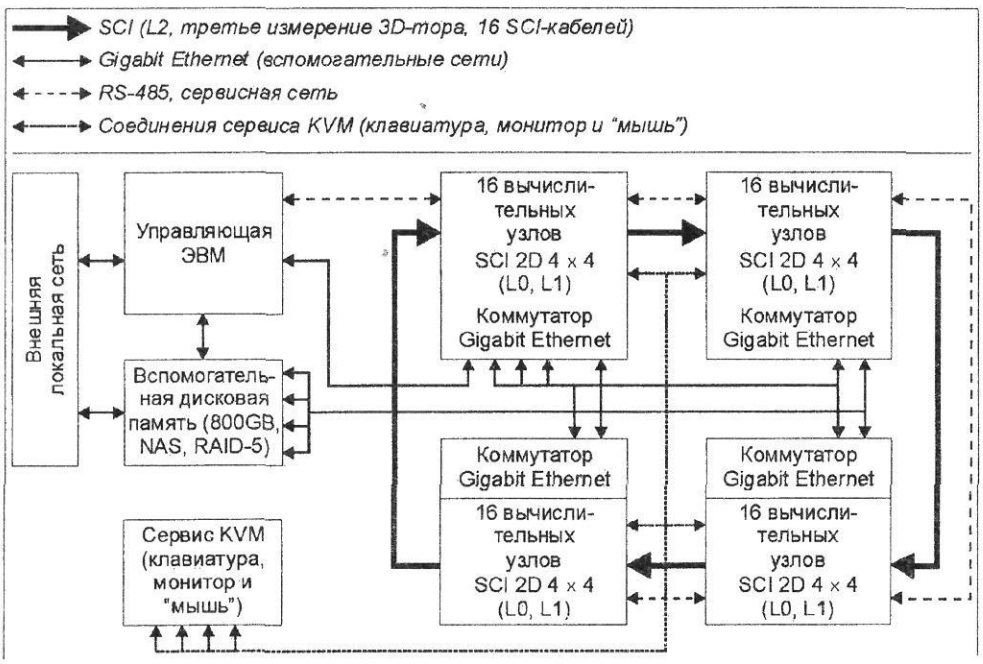

Рис.2. Структурная схема "СКИФ К-500"

Суперкомпьютер "СКИФ К-1000" является старшей моделью семейства "СКИФ". Технические параметры суперкомпьютера "СКИФ К-1000" должны обеспечивать его включение в Top-500 в 2004-2005 годах. Ниже рассмотрены основные технические требования к суперкомпьютеру "СКИФ К-1000".

Суперкомпьютер "СКИФ К-1000" является кластером из 200 вычислительных двухпроцессорных узлов на базе 64-разрядных процессоров, всего 400 процессоров AMD Opteron 248. 2,2Ghz, 1MB L2 (не хуже). Производительность "СКИФ К-1000" на тесте LinPack должна быть не менее 1.1 TFlops.

Каждый узел "СКИФ К-1000" должен содержать двухпроцессорную системную плату SMP 2 x AMD Opteron(tm) 248 (2.2Ghz, 1MB L2), основную память (4 Гбайт), жесткий диск (не менее 80 Гбайт), два интерфейса Gigabit Ethernet, адаптер Infiniband 4x, потребляемая мощность - не более 400 ВА.

В отличии от большинства установок семейства "СКИФ", в суперкомпьютере "СКИФ К-1000" для органи-

Производительность пиковая (и на задаче Linpack – оценка)	**1.76 (~1.1) Tflops**
Тип процессора **AMD Opteron 248, 2.2Ghz**	Системная сеть **Infiniband 4x, Fat Tree, FBB 10 Gbit/sec**
Число вычислительных узлов **200**	
Число процессоров **400**	Вспомогательные сети **2 × Gigabit Ethernet**
Оперативная память **200 × 4 = 800 GB**	Дополнительно **сервисная сеть (On/Off, Reset, сериальная консоль)**
Дисковая память **200 × 80 = 16 000 GB**	
Конструктив узла (форм-фактор) **1U**	

Рис. 3. Ожидаемые технические характеристики суперкомпьютера "СКИФ К-1000".

зации MPI-взаимодействия для параллельных вычислений будет использована не аппаратура SCI, а сеть на базе адаптеров и коммутаторов Infiniband. Коммутаторы Infiniband должны обеспечивать работоспособность всех вычислительных узлов в топологии Fat Tree с обеспечением FBB 10 Gbit/sec.

Как и в "СКИФ К-500", в суперкомпьютере "СКИФ К-1000" предполагается иметь две вспомогательные сети Gigabit Ethernet:

- для загрузки программ, данных, управления и мониторинга;
- для подключения дополнительной дисковой памяти.

В "СКИФ К-1000" будет использоваться сервисная сеть, по своим функциональным возможностям подобная сервисной сети "СКИФ К-500": обеспечение для любого вычислительного узла операции включение/выключение питания, аппаратного сброса (reset), возможности организации с управляющей ЭВМ консольного доступа к любому вычислительному узлу на этапах загрузки и конфигурирования BIOS, загрузки и работы ОС, "посмертного" чтения последних консольных сообщений ОС.

Дополнительная дисковая память должна быть реализована с использованием современных типовых средств файловых серверов (с применением RAID-технологии).

Программное обеспечение суперкомпьютера "СКИФ К-1000" включает:

- ядро ОС Linux, адаптированное для работы на суперкомпьютерах семейства "СКИФ";
- параллельная файловая система PVFS-SKIF;
- система очередей задач (PBS-SKIF) для кластерного уровня суперкомпьютеров семейства "СКИФ";
- система мониторинга FLAME (Functional Active Monitoring Environment);
- распределенная программная система отладки MPI-программ (ПС TDB);
- ядро T-системы (версия OpenTS);
- компилятор TGCC языков TC/T++;
- свободно распространяемые параллельные библиотеки и приложения,
- адаптированные для работы на суперкомпьютерах "СКИФ".

Заключение

В рамках программы "СКИФ" создан научно-технический задел, необходимый для комплексной реализации завершающего этапа программы, включая разработку образцов суперкомпьютеров триллионного диапазона производительности. Достижение этой цели позволит создать суперкомпьютерный центр, обеспечивающий расчеты в интересах различных предприятий и учреждений стран-участниц Союзного государства.

Успешное выполнение всех мероприятий Программы позволит в сравнительно короткие сроки при относительно небольших затратах выйти на собственный путь развития конкурентоспособной высокопроизводительной вычислительной техники, уровень которой будет соответствовать прогнозируемым требованиям со стороны широкой категории пользователей.