КОММУНИКАЦИОННЫЕ СРЕДЫ В КЛАСТЕРНЫХ ПАРАЛЛЕЛЬНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМАХ

Солонин А.Н.

Кафедра ЭВМ ДонНТУ,
ninolos@ukrtop.com

Abstract

Solonin A,Communication environments in cluster parallel computing systems*. The existing modern means of organization of communication between workstations in a cluster are considered in the article. The qualitative analysis of these means is carried out, the basic criterions of a choice of communication environments are formulated at construction of computing cluster systems.*

Введение

Необходимость роста производительности вычислительных ресурсов ни у кого в современном мире не вызывает сомнений. Приобретение готовых суперкомпьютеров представляется в условиях нынешнего финансирования науки и образования маловероятным. Поэтому особое значение приобретают вопросы разработки распределенных параллельных вычислительных сред ([1, 2, 12, 13]), кластерных структур на базе сетей рабочих станций, выбора коммутационных систем кластеров.

В данной статье рассматриваются существующие современные средства организации связи между рабочими станциями в кластере. Проведен качественный анализ этих средств, сформулированы основные критерии выбора коммуникационных сред (КС) при построении кластерных вычислительных систем (ВС).

1. Реализация коммуникационных сред. КС на основе интерфейса SCI

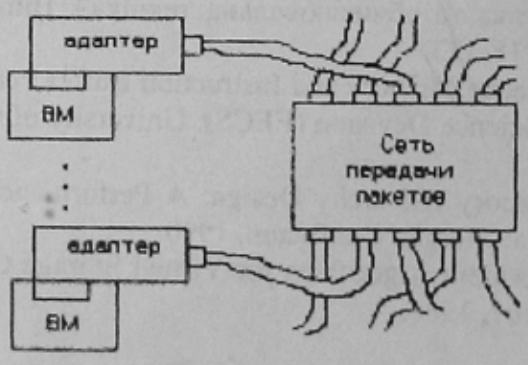

Рис.1 - Состав КС

На структурном уровне КС состоит из трех компонентов, как показано на рис.1:

- адаптеров, осуществляющих интерфейс между вычислительной машиной (ВМ) и сетью передачи пакетов;
- коммутаторов сети передачи пакетов;
- кабелей, служащих для подсоединения входных и выходных каналов (линков) адаптеров к портам коммутатора и соединения коммутаторов друг с другом.

Одной из наиболее новых и перспективных КС является КС на базе масштабируемого

когерентного интерфейса (Scalable Coherent Interface - SCI). SCI принят как стандарт ANSI/IEEE Std 1596-1992 [3]. Его развивают Apple, Dolphin Interconnect Solutions, SGI/CRAY, SUN, IBM и ряд других организаций. Микросхемы, реализующие этот стандарт, выпускаются серийно несколькими фирмами ([4,5]).

Стандарт SCI обеспечивает построение легкой в реализации, масштабируемой, эффективной в стоимостном аспекте коммуникационной среды для объединения процессоров и памятей, либо создания сети рабочих станций, либо для организации ввода/вывода суперЭВМ, высокопроизводительных серверов и рабочих станций современных микропроцессоров. Стандарт предусматривает создание пропускной способности не менее 1Гбайта/с для сосредоточенных систем и не менее 1Гбита/с для распределенных систем типа сети рабочих станций.

SCI - узлы защищены от механических воздействий и электромагнитных излучений, а также могут выниматься и вставляться без отключения питания. Структура системы с использованием SCI показана на рис. 2.

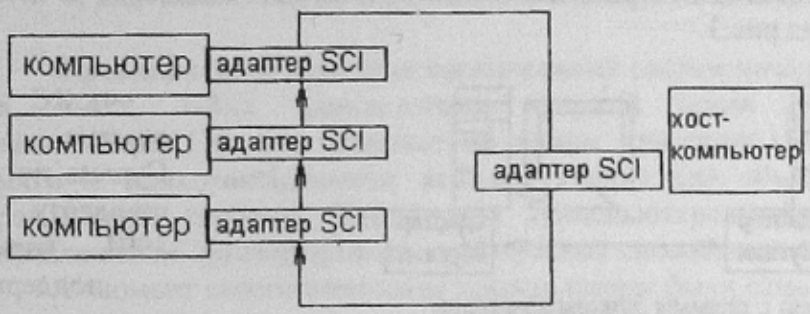

Рис.2 - Структура системы с использованием SCI

Каждый SCI - узел имеет входной и выходной каналы. Узлы связаны однонаправленными каналами "точка - точка" либо с соседним узлом, либо подключены к коммутатору. При объединении узлов должна обязательно формироваться циклическая магистраль (кольцо) из узлов, соединенных каналами " точка - точка", между входным и выходным каналами каждого узла. Один узел в кольце, называемый scrubber, выполняет функции: инициализации узлов кольца с установлением адресов, управления таймерами, уничтожения пакетов, не нашедших адресата. Этот узел помечает проходящие через него пакеты и уничтожает уже помеченные пакеты. В кольце может быть только один scrubber.

Ниже приведен набор оборудования, предлагаемый фирмой Dolphin, для построения SCI – кластеров:

- адаптер "шина Sbus – 2 <-> SCI" ;
- адаптер "шина PCI – 2 <-> SCI" ;
- коммутатор 4x4;
- внешний разветвитель EDU (external distribution unit);
- станционные кабели;
- линейные кабели.

Последние три позиции перечня устройств нуждаются в объяснении. Адаптеры и коммутаторы имеют входные и выходные линки SCI. Конструктивно целесообразно иметь один разъем, а не два отдельных для входного и выходного линков. Поэтому используются два типа кабелей: узловые, содержащие две разнонаправленные магистрали для входного и выходного линков, и линейные, состоящие из

однонаправленной магистрали. Для перехода от двунаправленного узлового кабеля к двум однонаправленным линейным кабелям используется внешний разветвитель EDU.

2. Коммуникационные среды MYRINET и Raceway

Логический протокол Myrinet был развит в Caltech Submicron Systems Architecture Project [6]. Эта среда стандартизирует формат пакета, способ адресации ВМ, набор управляющих символов протокола передачи пакетов. КС образуется адаптерами «шина компьютера – линк сети» и коммутаторами сети. Никаких требований к адаптеру, кроме собственно реализации протокола, среда Myrinet не предъявляет.

Каждый линк содержит пару однонаправленных каналов, образуя дуплексный канал. Пропускная способность линка составляет 80 Мбайт/с. Структуры вычислительных систем, построенных с использованием адаптеров и коммутаторов Myrinet показаны на рис.3.

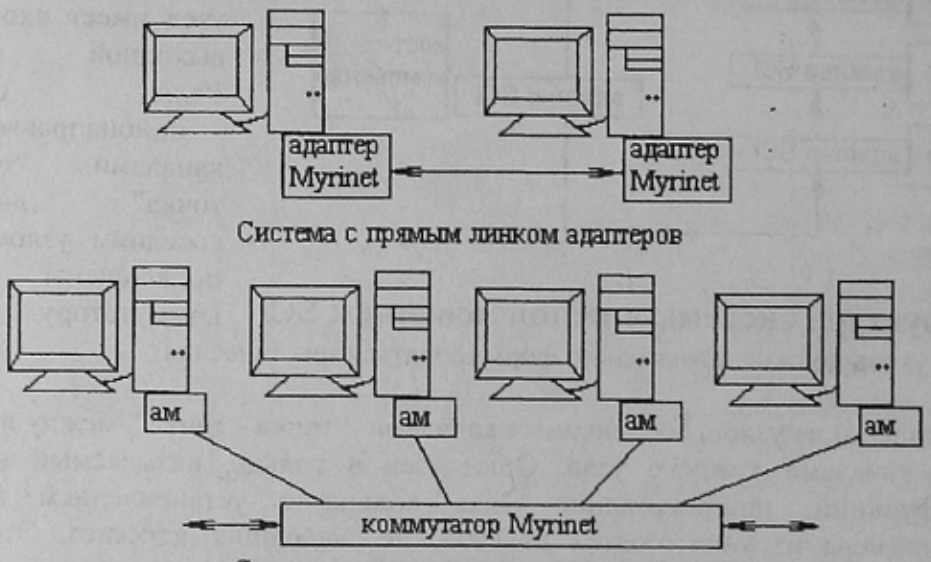

Рис.3 - Структуры ВС с использованием Myrinet

КС Raceway создана фирмой Cypress по открытому стандарту VITA 5-1994, изданному и поддерживаемому организацией VITA. КС Raceway обеспечивает пропускную способность на уровне 1 Гбайт/с. КС создается с использованием кристалла коммутатора Cypress CY7C965 Raceway Crossbar и соответствующих адаптеров порт коммутатора – шина ВМ. Адаптеры должны содержать в своем составе интерфейсные схемы, канал прямого доступа в память, таймеры и другие блоки, которые необходимы для управления коммутатором.

Коммутатор может одновременно соединять три входных порта с тремя различными выходными портами, что обеспечивает общую пропускную способность 480 Мбайт/с. Структуры ВС, создаваемые с использованием среды Raceway, такие же как в случае применения среды Myrinet.

Разница заключается в количестве портов коммутаторов, форматах передаваемых пакетов и в протоколах.

3. Коннектор шин PCI: SRC 3266 DE и Memory Channel фирмы DEC

SRC 3266 DE [7] предназначен для работы с 32 – разрядной шиной PCI на частотах вплоть до 66 МГц. Количество объединяемых кристаллов SRC 3266 может достигать 256. Линки, связывающие SRC кристаллы, имеют ширину 16 битов и тактируются сигналом 266 МГц. Таким образом, пиковая пропускная способность линка составляет 532 Мбайт/с. Допустимая длина линка до 3 метров.

Memory Channel представляет собой КС, развиваемую фирмой DEC для построения кластерных систем [8,9]. Каждый модуль кластера имеет Memory Channel – адаптер, подсоединенный к коммутатору. Коммутатор реализует парные соединения «точка – точка» или трансляционное соединение «точка – всем точкам». Максимальная пропускная способность 132 Мбайт/с.

4. Использование транспьютерных микропроцессоров в КС

Промышленное освоение параллельных систем началось с выпуска семейств T-2xx, T-4xx, T-8xx транспьютеров фирмой Inmos (концерн SGS Thomson). Транспьютеры [10, 14] содержат на одном кристалле собственно процессор, блок памяти и коммуникационные каналы (линки) для объединения транспьютеров в параллельную систему. Линки этих транспьютеров передают данные в каждом направлении по одному проводу с пропускной способностью 10 Мбит/с.

В момент своего появления транспьютеры были самыми производительными 32 – разрядными микропроцессорами. Однако несколько позже были созданы микропроцессоры в десятки раз более производительные при сохранении лидерства транспьютеров в области коммуникационных возможностей. Это привело к появлению систем, состоящих из транспьютероподобных вычислительных машин, в которых вычислительная компонента реализована на базе мощного вычислительного микропроцессора, а коммуникационная – на базе транспьютера. На основе транспьютерных линков разработан стандарт IEEE P1355, который предусматривает передачу данных со скоростями в диапазоне 10 Мбит/с – 1 Гбит/с.

5. КС на базе телекоммуникационных технологий

В предыдущих пунктах рассматривались КС, появившиеся в результате совершенствования вычислительных систем. Развитие цифровых сетей передачи данных, локальных и глобальных сетей также идет в направлении повышения пропускной способности и интеграции услуг. Однако в этой сфере существует функционирующее сетевое и оконечное оборудование, представляющее собой большую материальную ценность, для воспроизводства которой необходимы многие годы. Кроме того, есть сложившаяся организационная структура, терминология и т.д. Поэтому цифровые сети передачи данных имеют специфику, выражающуюся в стремлении к совместимости с уже существующим оборудованием.

Среди высокоскоростных сетей (с пропускной способностью более 100 Мбит/с) можно отметить быстрый Ethernet, поддерживающий форматы данных Ethernet и Token Ring, FDDI, ATM, Fibre Channel и другие [11].

6. Сравнительный анализ коммуникационных сред

Доступность для использования:

В достаточной степени апробированы телекоммуникационные системы цифровой передачи данных. Остальные среды выпускаются как экспериментальные (или в ограниченной серии) образцы на базе полузаказных вентильных матриц, а не как хорошо апробированные заказные СБИС. Хотя в последнее время ситуация для SCI и Myrinet начинает меняться.

Реализация протоколов:

Протокол SCI достаточно сложен, содержит большие потенциальные возможности по управлению трафиком. Однако использование этих возможностей предполагает развитое программное обеспечение, которое, скорее всего, не будет свободно распространяемым, что затруднит его модификацию и использование. SCI образует сеть с коммутацией пакетов.

Выгодно отличается своей простой концепцией и аппаратной реализацией протокола КС Myrinet. Однако набор средств для управления трафиком несколько ограничен. Myrinet образует сеть с передачей сообщений по методу прокладки пути (wormhole).

Среда Raceway также требует программного обеспечения для управления трафиком. Она использует коммутацию каналов.

Среда на базе коннекторов SRC3266 DE привлекательна прозрачностью мостов между объединяемыми шинами PCI. Среда реализует протокол шины PCI, что позволяет довольно быстро разрабатывать оригинальные КС.

Протокол для Memory Channel достаточно сложен и имеет развитые средства квитирования ([8], [10]), что, по мнению разработчиков, должно предотвращать потерю пакетов.

В транспьютерах поддержан уровень передачи сообщений. Предполагается программная реализация сетевых протоколов.

Протоколы телекоммуникационных систем цифровой передачи данных сложны, регламентированы большим количеством рекомендаций и стандартов. Программная реализация этих протоколов является собственностью фирм –производителей, что делает ее практически недоступной при построении вычислительных систем, в которых реализуется специализация межмодульных обменов.

Протоколы локальных сетей и сетевые операционные системы более доступны, по крайней мере по программным интерфейсам.

Пропускная способность и задержка:

Сведения о пропускной способности и задержках при передачах приведены в табл.1.

Таким образом, выбор КС определяется:

- пропускной способностью шины, к которой подсоединяется адаптер;
- пропускной способностью и стоимостью адаптера;
- пропускной способностью и стоимостью сети коммутаторов;
- диапазоном масштабируемости сети коммутаторов;
- задержками при установлении соединений;
- коммерческой доступностью коммуникационной среды;
- практической апробацией КС в существующих системах;

- наличием развитого программного обеспечения адаптера и операционных систем, в которых адаптеры работоспособны.

Таблица 1

Параметр	SCI	Myrinet	SRC3266DE	Raceway	MC2	F. Ethernet
ПкПС* адаптера «шина CPU – коммутатор сети»,Мбайт/с	От 96 до 102		132		132	12
Задержка адаптера,мкс	2 – 4					
ПкПС портов коммутатора, Мбайт/с	200	80	532	160	132	12
Задержка коммутатора,мкс	1		2			
Количество портов коммутатора	4		2		8	
Ширина инф-х + управляющих линий порта	16+2	8+2	16	32+5	16	4
Частота передачи данных, МГц	50	40	266	40	66	

*ПкПС – пиковая пропускная способность.

Конечно, эти характеристики достаточно условны и выбор не может быть однозначным. Необходимо также учитывать класс тех задач, которые предполагается решать на создаваемой вычислительной системе. Все же из всех рассмотренных КС можно выделить, как самые распространенные и удовлетворительные по большинству вышеуказанных критериев, следующие:

- КС на основе Fast Ethernet: дешевы, просты в использовании, надежны, широко распространены в локальных сетях;
- КС на основе SCI и Myrinet: относительно дороги, высокопроизводительны, перспективны для дальнейшего развития.

Заключение

Проведенный анализ показал, что коммуникационные среды для кластерных вычислительных систем в настоящее время все еще находятся на этапе развития. Наиболее перспективными по своим действительным и потенциальным возможностям являются КС на базе SCI и Myrinet. В период 2001 – 2003 г. в рамках договора о сотрудничестве между факультетом ВТИ и ISR, IPVR предполагается построение и исследование возможностей кластерных систем и их коммуникационных систем.

Литература

1. Святний В.А. Проблемы паралельного моделювання складных динамічних систем.- Наукові праці ДонДТУ, серія ІКОТ, вип. 6, Донецьк, 1999, С. 6-14.
2. Святний В.А., Солонін О.М., Надєєв Д.В., Степанов І., Ротермель К., Цайтц М. Розподілене паралельне моделююче середовище.- Наукові праці ДонДТУ. Серія

"Проблеми моделювання та автоматизації проектування динамічних систем". Випуск 29:-Донецьк, ДонДТУ, 2001. – С.229 – 234.
3. IEEE Computer Society. IEEE Standart for Scalable Coherent Interface (SCI), IEEE Std 1596-1992, New York, August, 1993.
4. Dolphin Interconnect Solutions, Inc. (http://www.dolphinics.com)
5. VITESSE Semiconductor Corporation (http://www.vitesse.com)
6. Microm, Inc. (http://www.myri.com).
7. Sebring System, Inc. (http://www.sebringring.com)
8. Корнеев В.В. Параллельные вычислительные системы. М. – Нолидж. 1999 312 с.
9. Бройнль Т. Паралельне програмування. К.-ВШ 1997- 358 с.
10. Корнеев В.В., Киселев А.В. Современные микропроцессоры. М. Нолидж. 1998. 237 с.
11. Биленко А. Телекоммуникации. Взгляд изнутри. "Chip News". 1998. №4(25), С.2 – 14.
12. Аноприенко А.Я, Святний В.А. Высокопроизводительные информационно – моделирующие среды для исследования, разработки и сопровождения сложных динамических систем. - Наукові праці ДонДТУ. Серія "Проблеми моделювання та автоматизації проектування динамічних систем". Випуск 29:-Донецьк, ДонДТУ, 2001.–С.346-368.
13. Хокни Р., Джосссхоуп К. Параллельные ЭВМ. Архитектура, программирование и алгоритмы. М. - РиС. 1986. 358с.
14. Богуславский Л.Б., Ляхов А.И. Методы оценки производительности многопроцессорных систем. М. – Наука. 1992. 217с.

Поступила в редакційну колегію 30. 11.2001 р.