ФОРМИРОВАНИЕ ПОДСИСТЕМ ЭЛЕМЕНТАРНЫХ МАШИН В ВЫЧИСЛИТЕЛЬНЫХ КЛАСТЕРАХ НА БАЗЕ СОСТАВНЫХ КОММУТАТОРОВ

Е. Н. Перышкова

Институт физики полупроводников им. А. В. Ржанова СО РАН

УДК 004.75

В настоящее время коммуникационные сети большинства высокопроизводительных вычислительных систем построены на базе составных коммутаторов. Такой подход к организации коммуникационной сети позволяет обеспечить высокую пропускную способность и одинаковое количество промежуточных коммутаторов между элементарными машинами, подключенных к коммутаторам одного уровня. Однако, при одновременном использовании параллельными процессами каналов связи из-за конкуренции за разделяемые ресурсы (сетевой контроллер, канал связи, порт коммутатора) возникает деградация производительности (network contention. Большинство алгоритмов формирования подсистем элементарных машин, реализуемых в системах управления ресурсами (IBM LoadLeveler, Altair PBS Pro, SLURM, TORQUE), не учитывают возможного падения производительности сетевой подсистемы при одновременном использовании ее компонента параллельными процессами программы. Целью работы является исследование алгоритмов формирования подсистем элементарных машин и оценка качества формируемых подсистем с точки зрения возникающей конкуренции за сетевые ресурсы.

Ключевые слова: параллельное мультипрограммирование, организация функционирования, вычислительные системы.

Введение. Современные высокопроизводительные вычислительные системы (ВС) можно разделить на 2 класса, в зависимости от организации их коммуникационных сетей:

1. ВС на базе коммуникационных сетей с прямым соединением узлов (direct network);

2. ВС с коммуникационными сетями на базе составных коммутаторов (indirect network, switch-based network).

К первому классу ВС, как правило, относятся проприетарные системы со структурами сетей типа: многомерные торы (Cray Gemini, IBM PERCS, Fujitsu Tofu) [1-3], гиперкубы (СМПО-10G, Ангара). Второй класс сетей образован наибольшим числом высокопроизводительных систем списка Top500 – ВС на базе коммутаторов с ограниченным числом портов. Для формирования ВС с большим количеством элементарных машин (ЭМ) выполняется многоуровневое объединение множества подобных коммутаторов. Примером может служить топология «толстое дерево» (fat tree, folded clos network) [4] на базе коммутаторов стандарта InfiniBand и система Tianhe-2 (сеть TH Express-2). Широкое распространение данной топологии обусловлено высокой пропускной способностью между ЭМ системы и одинаковым расстоянием между коммутаторами одного уровня (в смысле количества промежуточных коммутаторов на пути от одного до другого).

При одновременном использовании параллельными процессами каналов связи (сетевых адаптеров, коммутаторов на всех уровнях) возникает деградация их производительности (network contention). Например, при одновременной передачи сообщений всеми ядрами од-

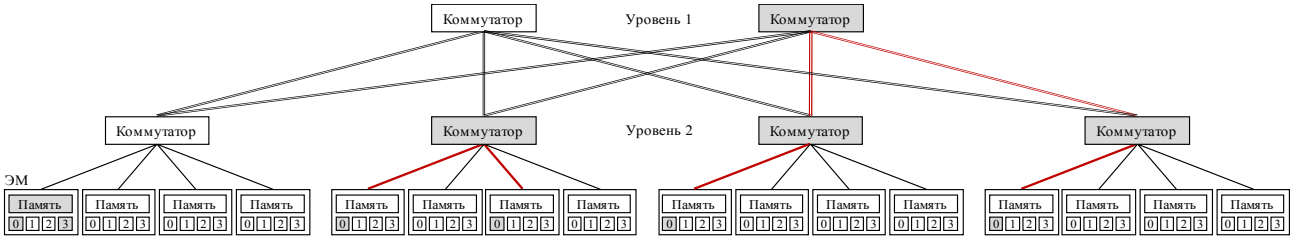

Рис. 1. Вычислительная система из 16 четырех ядерных ЭМ (топология «толстое дерево»)

ной ЭМ ядрам других машин, возникает конкуренция за сетевой контроллер и соответствующий канал и порт коммутатора. Следствием этого является падение пропускной способности сети. Целью работы является исследование алгоритмов формирования подсистем элементарных машин и оценка качества формируемых подсистем с точки зрения возникающей конкуренции за сетевые ресурсы.

1. Алгоритмы формирования подсистем. В настоящее время наиболее распространенная техническая реализация ЭМ – многопроцессорный SMP/NUMA-узел. На рис. 1 представлен пример ВС из 16 четырех ядерных элементарных машин, объединенных сетью с топологией «толстое дерево».

Время передачи информационного сообщения зависит от того какие процессорные ядра участвуют в операции обмена. Рассмотрим 3 возможных случая на рис. 1:

1. Взаимодействующие ядра находятся в одной ЭМ и передают сообщения через их общую память (на рис. 1 – два ядра первой ЭМ);

2. Взаимодействующие ядра находятся на разных ЭМ, подключенных к одному коммутатору второго уровня. Сообщения передаются через коммуникационную сеть (на рис. 1 – два ядра, реализующие обмен через второй коммутатор второго уровня);

3. Взаимодействующие ядра находятся на разных ЭМ, подключенных к разным коммутаторам второго уровня. Для передачи сообщений требуется совершить транзитный проход через коммутатор первого уровня (на рис. 1 – два ядра, реализующие обмен через второй коммутатор первого уровня).

Продемонстрировать возникающую конкуренцию за сетевые ресурсы можно на примере параллельной программы, реализующий парный вызов каждым процессорным ядром функции MPI_Send и MPI_Recv. Зависимость пропускной способности каналов связи от количества одновременно работающих на ЭМ процессов представлены на рис. 2. Программа запускалась на двух ЭМ по 8 процессорных ядер на каждом.

При разработке параллельных программ для ВС значительное место по частоте использования и суммарному времени выполнения занимают коллективные операции обменов информации. Наиболее популярным видом коллективной операции является трансляционно-циклический обмен («каждый-всем», all-to-all). В таких обменах участвуют все ветви параллельной программы, что приводит к деградации производительности каналов связи.

В тестовой параллельной программе реализован вызов каждым процессом функции MPI_Alltoall. На рис. 3 продемонстрировано зависимость времени передачи сообщения размером 1 мегабайта (*а*) и 64 килобайта (*б*), в зависимости от количества одновременно работающих процессов на ЭМ.

2. Организация экспериментов. В качестве тестовой задачи рассматривался тест IS (сортировка чисел) из пакета NAS Parallel Benchmarks, как реализующая основные схемы информационных обменов, возникающих во многих алгоритмах. Использовались классы

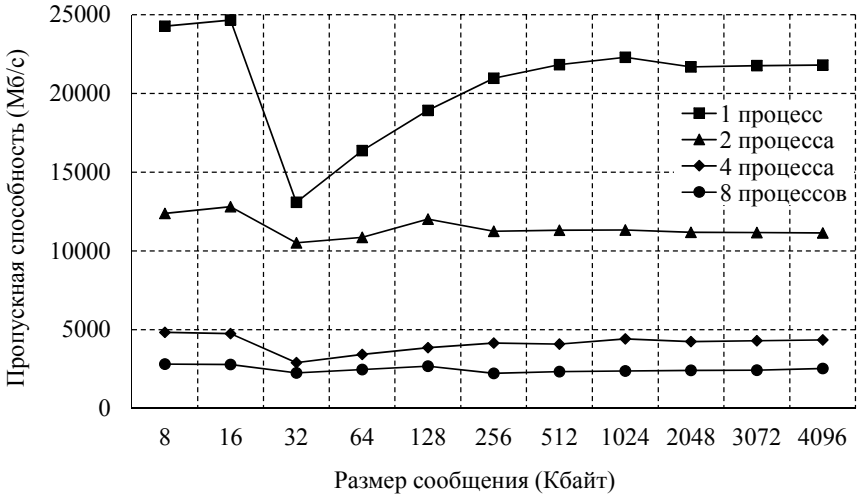

Рис. 2. Зависимость пропускной способности канала связи от размера передаваемого сообщения и количества процессов MPI-программы, выполняющихся на ЭМ (разделяющих канал связи, тест ping-pong)

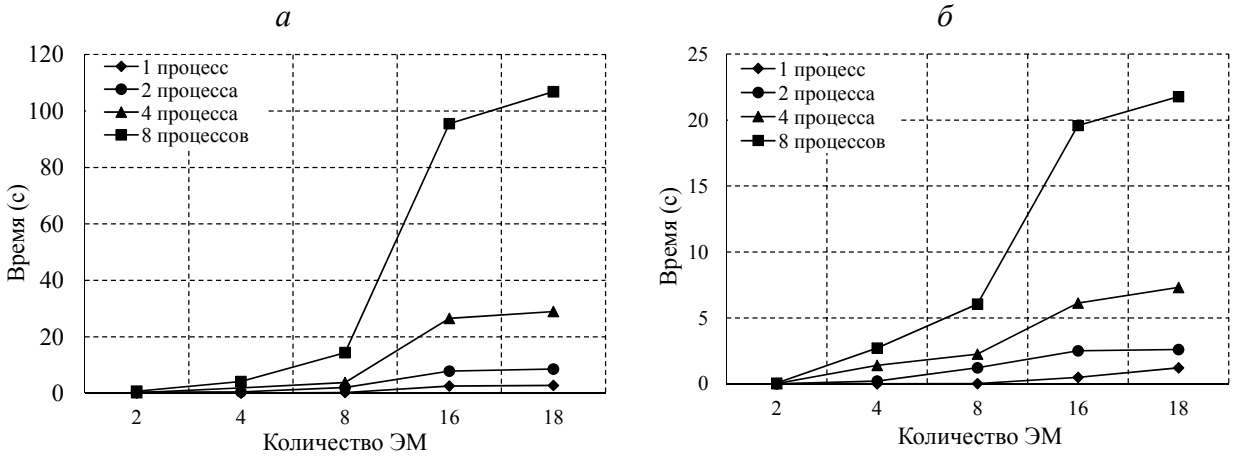

Рис. 3. Зависимость времени выполнения коллективной операции Alltoall от количества одновременно работающих процессов на элементарной машине. Размер сообщения: *а* – 1 Мбайт; *б* – 64 Кбайт

теста C и D, отличающиеся в объеме обрабатываемых данных. Для класса C используется размерность задачи 2^{27}, для класса D размерность задачи составляет 2^{29}.

Тестовые программы модифицированы для измерения общего времени их выполнения и времени пребывания в функциях MPI.

Экспериментальная часть работы выполнена на вычислительных кластерах с коммуникаторами и сетевыми адаптерами стандартов InfiniBand QDR и Gigabit Ethernet.

Рассматриваются эвристические алгоритмы формирования подсистемы из P процессорных ядер. Предполагается, что ЭМ системы – многопроцессорные вычислительные узлы с q процессорными ядрами. Алгоритмом ранга k назовем алгоритм, который выделяет с каждой ЭМ по k ядер. Таким образом, для формирования подсистемы из P процессорных ядер необходимо $\lceil P/k \rceil$ ЭМ.

Экспериментально исследованы q различных алгоритмов формирования подсистем – алгоритмы рангов 1, 2, …, q. Для каждого алгоритма и тестовой программы проанализировано общее время выполнения и время выполнения функции MPI, при возникающей конкуренции за использование сетевых ресурсов.

3. Результаты экспериментов. Временные характеристики теста IS класс C на вычислительном кластере с коммуникаторами и сетевыми адаптерами стандарта Gigabit Ethernet для различных конфигураций подсистем представлены на рис. 4.

Количество процессов	Количество вычислительных узлов	Процессов на узле	Время выполнения программы	Время выполнения коллективных обменов	Время выполнения двухсторонних обменов
4	1	4	14,11	5,99	1,32
	2	2	20,74	14,36	1,23
	4	1	18,12	12,1	1
8	1	8	13,31	6,72	1,92
	2	4	17,3	13,4	1
	4	2	15,48	12,3	1,07
	8	1	10,79	7,78	1,13
16	2	8	16,58	13,72	1,78
	4	4	13,58	11,96	0,95
	8	2	10,6	9,13	0,88
	16	1	30,2	28,26	0,71
32	4	8	13,9	12,89	1,13
	8	4	9,83	9,06	0,67
	16	2	30,83	30,09	0,65
64	8	8	15,15	14,73	0,54
	16	4	48,77	48,41	3,12

Рис. 4. Временные характеристики теста IS класс C из пакета NAS Parallel Benchmarks на подсистемах различных конфигураций (адаптер стандарта Gigabit Ethernet)

Временные характеристики теста IS класс C на вычислительном кластере с коммуникаторами и сетевыми адаптерами стандарта InfiniBand QDR для различных конфигураций подсистем представлены на рис. 5.

Количество процессов	Количество вычислительных узлов	Процессов на узле	Время выполнения программы	Время выполнения коллективных обменов	Время выполнения двухсторонних обменов
4	1	4	5,19	0,98	0,94
	2	2	5,09	1	0,94
	4	1	4,81	0,78	0,94
8	1	8	2,8	0,63	0,67
	2	4	6,54	2,5	1,41
	4	2	2,96	0,75	0,68
16	2	8	5,31	3,22	0,95
	4	4	5,18	3,04	0,93

Рис. 5. Временные характеристики теста IS класс C из пакета NAS Parallel Benchmarks на подсистемах различных конфигураций (адаптер стандарта InfiniBand QDR)

Временные характеристики теста IS класс D на вычислительном кластере с коммуникаторами и сетевыми адаптерами стандарта InfiniBand QDR для различных конфигураций подсистем представлены на рис. 6.

Количество процессов	Количество вычислительных узлов	Процессов на узле	Время выполнения программы	Время выполнения коллективных обменов	Время выполнения двухсторонних обменов
4	1	4	6360	3714	724
	2	2	109,12	14,23	25,76
	4	1	99,83	11,01	18,49
8	1	8	7121	3506	514
	2	4	112,31	19,19	31,82
	4	2	55,88	11,3	13,73
16	2	8	61,86	17,39	16,86
	4	4	58,83	15,03	16,71

Рис. 6. Временные характеристики теста IS класс D из пакета NAS Parallel Benchmarks на подсистемах различных конфигураций (адаптер стандарта InfiniBand QDR)

На рис. 4 – 6 выделено минимальное время выполнения программы для каждого количества процессоров.

Заключение. В данной работе выполнено исследование алгоритмов формирования подсистем элементарных машин и оценка качества формируемых подсистем с точки зрения возникающей конкуренции за сетевые ресурсы. Показана деградация производительности каналов связи при различных вариантах формирования подсистем вследствие разделения каналов связи несколькими процессами.

Направлением дальнейших работ – создание алгоритмов формирования подсистем элементарных машин с учетом загруженности каналов связи.

Список литературы

1. Alverson R., Roweth D., Kaplan L. The Gemini System Interconnect // Proc. 18th IEEE Symposium on High Performance Interconnects. Washington, DC: IEEE Press, 2010. 83–87.

2. Chen D., Eisley N.A., Heidelberger P., Senger R., et al. The IBM Blue Gene/Q interconnection network and message unit // Proc. 2011 International Conference for High Performance Computing, Networking, Storage and Analysis. New York: ACM Press, 2011. doi 10.1145/2063384.2063419.

3. Ajima, Y., Inoue, T., Hiramoto, S., Shimizu, T., Takagi, Y. The Tofu Interconnect. IEEE Micro 32(1), 2012. 21–31.

4. Корнеев В.В. Вычислительные системы / В. В. Корнеев. – М.: Гелиос АРВ, 2004. – 512 с.

Перышкова Евгения Николаевна – инженер Института физики полупроводников им. А. В. Ржанова СО РАН; 630090, Новосибирск; e-mail: peryshkova@isp.nsc.ru;