Формирование подсистем элементарных машин в вычислительных кластерах на базе составных коммутаторов

Перышкова Евгения Николаевна

e.peryshkova@gmail.com

Институт физики полупроводников им. А.В. Ржанова СО РАН,
Новосибирск, Россия

Организация коммуникационных сетей в современных вычислительных системах

- Коммуникационные сети с прямым соединением узлов (direct network)

  Cray Gemini, IBM PERCS, Fujitsu Tofu – многомерные торы

  СМПО-10G, Ангара – гиперкубы

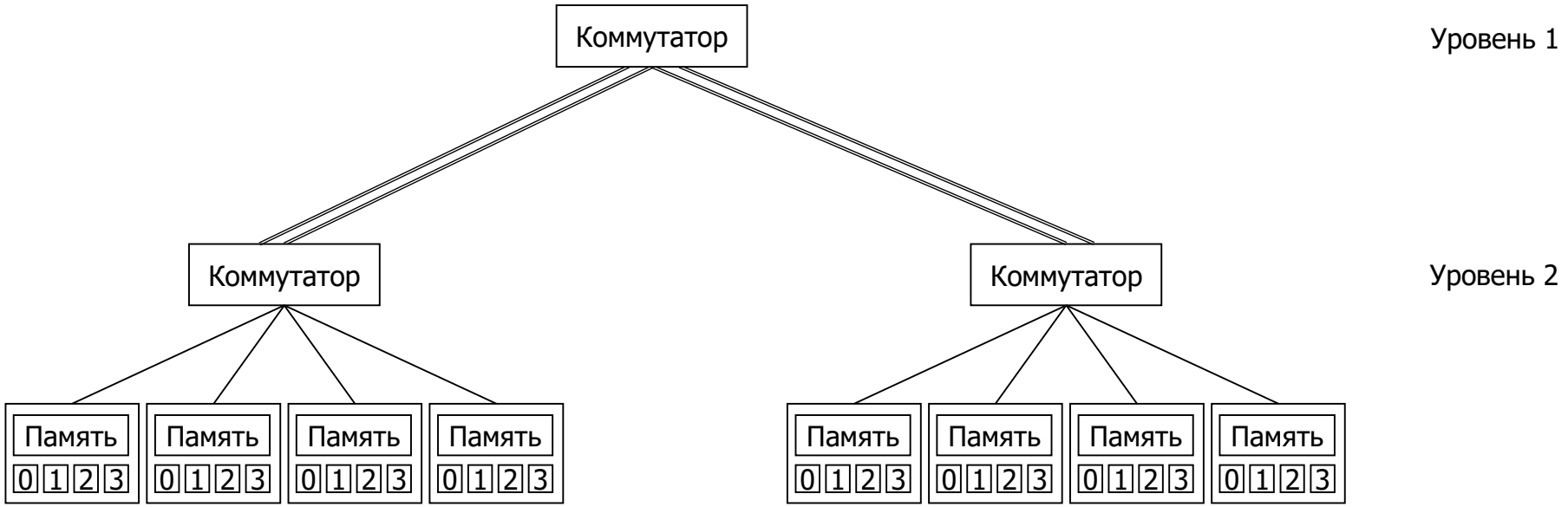
- Коммуникационные сети на базе составных коммутаторов (indirect network, switch-based network)

  Tianhe-2 (сеть TH Express-2)

  топология «толстое дерево» (fat tree, folded clos network) на базе стандарта InfiniBand

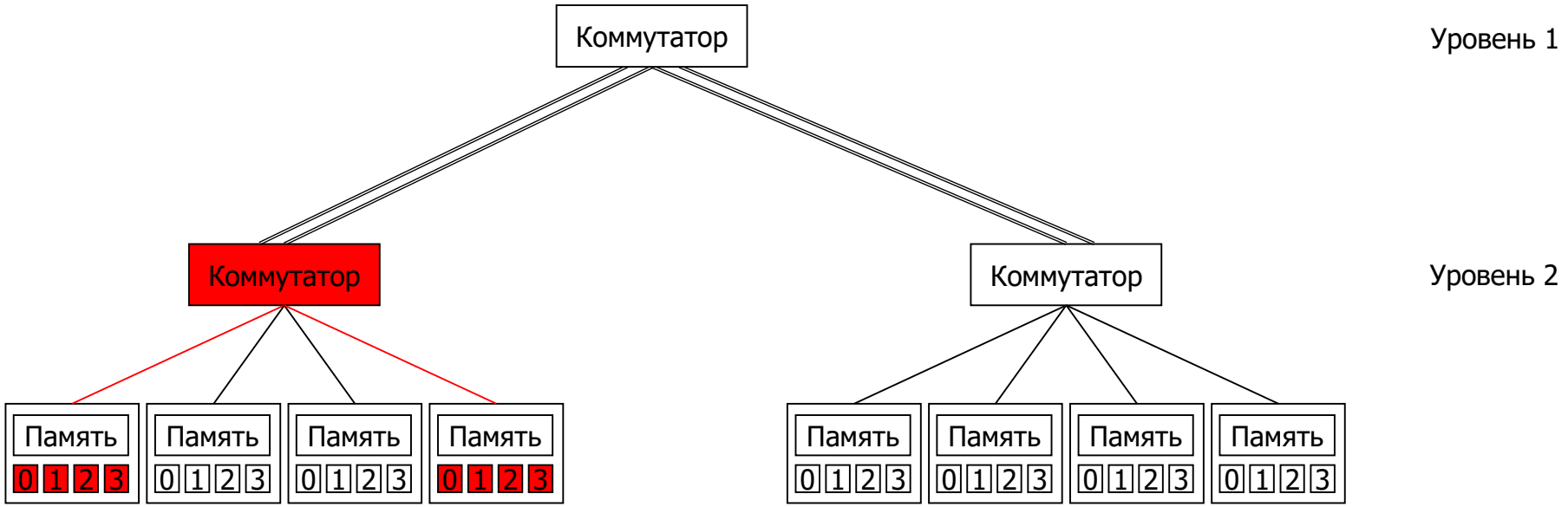
  наибольшее число высокопроизводительных систем списка Top500

Топология «толстое дерево» (fat tree, folded clos network)



Широкое распространение данной топологии обусловлено высокой пропускной способностью между элементарными машинами системы и одинаковым расстоянием между коммутаторами одного уровня

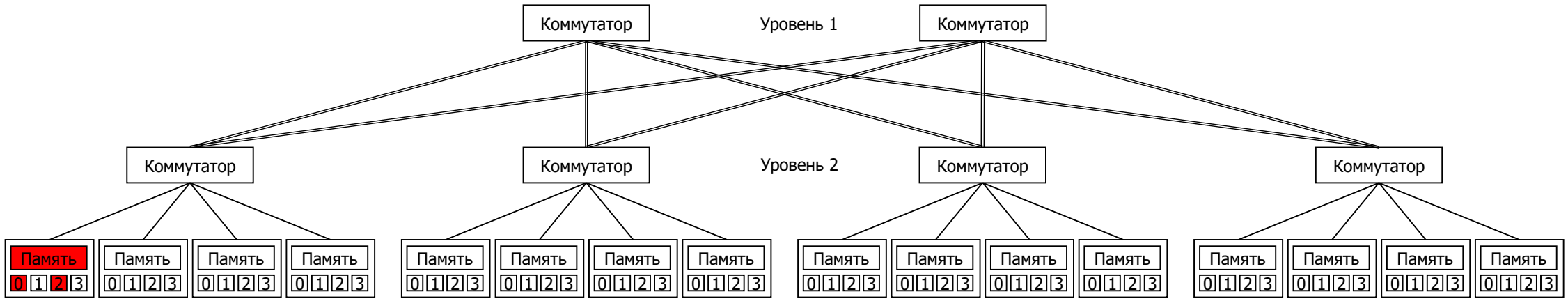
Топология «толстое дерево» (fat tree, folded clos network)



При одновременном использовании параллельными процессами каналов связи (сетевых адаптеров, коммутаторов на всех уровнях) возникает деградация их производительности (network contention)

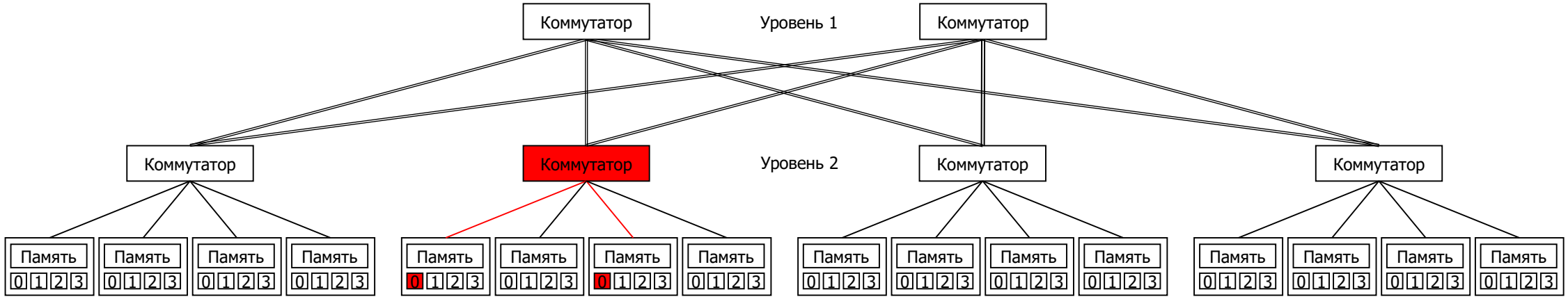
Цель работы: исследование алгоритмов формирования подсистем элементарных машин и оценка качества формируемых подсистем с точки зрения возникающей конкуренции за сетевые ресурсы.

Алгоритмы формирования подсистем элементарных машин



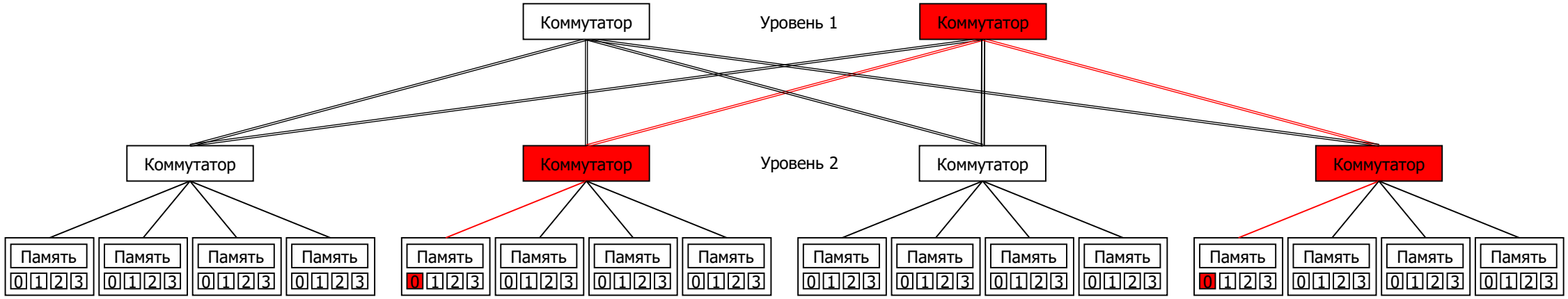
Разделяемый ресурс: взаимодействие через общую память ЭМ

Алгоритмы формирования подсистем элементарных машин



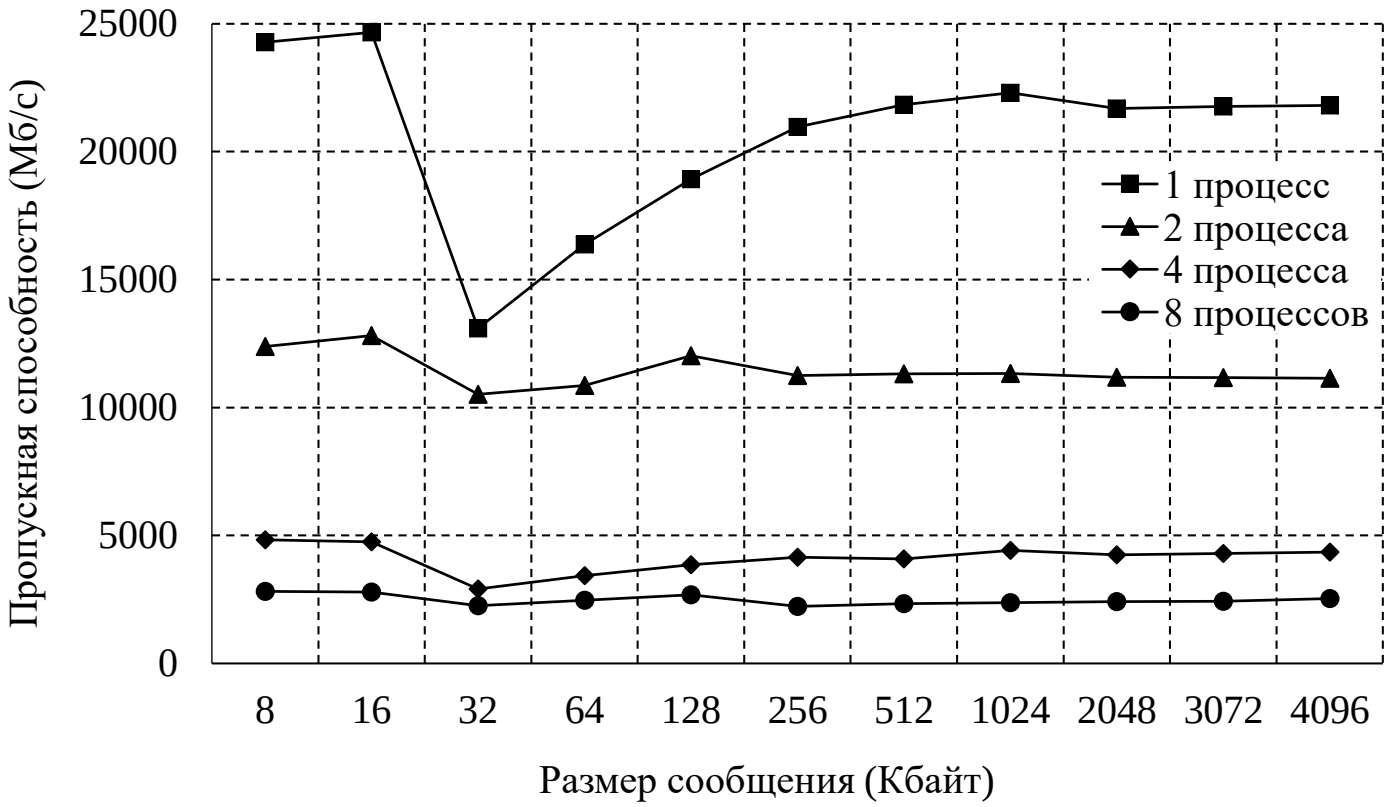
Разделяемый ресурс: взаимодействие через коммутатор 2 уровня

Алгоритмы формирования подсистем элементарных машин



Разделяемый ресурс: взаимодействие через коммутатор 1 уровня

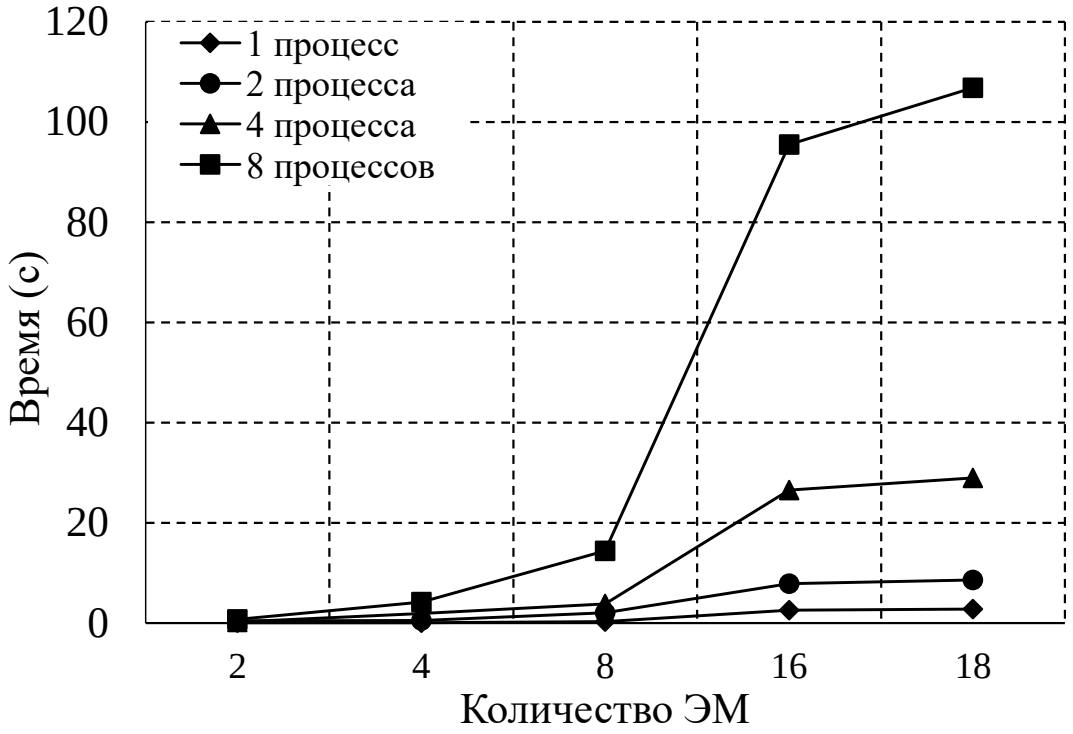
Конкуренция за сетевые ресурсы (network contention)



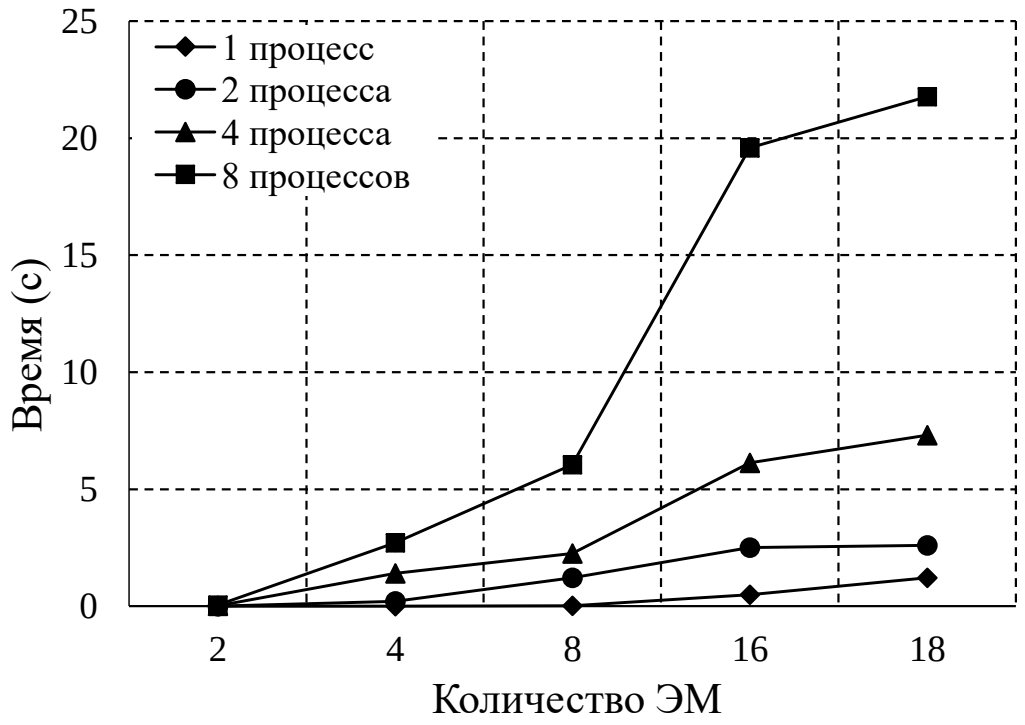
Передача сообщений (**MPI_Send** и **MPI_Recv**) между парой процессов на разных процессорных ядрах

Зависимость пропускной способности канала связи от размера передаваемого сообщения и количества процессов MPI-программы

Конкуренция за сетевые ресурсы (network contention)



размер сообщения 1 Мбайт



размер сообщения 64 Кбайт

Зависимость времени выполнения коллективной операции **MPI_Alltoall** от количества одновременно работающих процессов на элементарной машине.

Организация экспериментов

- В качестве тестовой задачи рассматривался тест IS (сортировка массива) из пакета NAS Parallel Benchmarks
- Использовались классы теста C и D, отличающиеся объемом обрабатываемых данных

- 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

Результаты экспериментов

Количество процессов	Количество вычислительных узлов	Процессов на узле	Время выполнения программы, с	Время выполнения коллективных обменов, с	Время выполнения двухсторонних обменов, с
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

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

Результаты экспериментов

Количество процессов	Количество вычислительных узлов	Процессов на узле	Время выполнения программы, с	Время выполнения коллективных обменов, с	Время выполнения двухсторонних обменов, с
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
	8	1	5,31	3,22	0,95
16	2	8	**5,18**	3,04	0,93
	4	4	5,19	0,98	0,94

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

Результаты экспериментов

Количество процессов	Количество вычислительных узлов	Процессов на узле	Время выполнения программы, с.	Время выполнения коллективных обменов, с	Время выполнения двухсторонних обменов, с
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
	8	1	61,86	17,39	16,86
16	2	8	**58,83**	15,03	16,71
	4	4	6360	3714	724

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

Выводы

- 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

Заключение

- 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.

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

Спасибо за внимание!

Перышкова Евгения Николаевна

e.peryshkova@gmail.com

Институт физики полупроводников им. А.В. Ржанова СО РАН,
Новосибирск, Россия