

**Научно-исследовательский центр
Электронной Вычислительной Техники**

**Отчет по экспериментальной проверке 16-узлового кластера
ЭКСС-КУ с топологией 2-мерный тор (ЭКСС-1)**

(НИЦЭВТ, январь 2001)

Москва - 2001

Аннотация

Оценка эффективности принятых технических решений была проведена по результатам экспериментального исследования кластера ЭКСС-1. Исследования проводились на двух классах оценочных программ: задачи пакета NPВ 2.3, разработанного в подразделении NAS Центра Эймса NASA, тесты программ MPI пакета Pallas РМВ 2.2.

СОДЕРЖАНИЕ

Введение. Описание ЭКСС-1.....	4
1. Прикладные бенчмарки – задачи пакета NAS NPB 2.3.....	5
2. Результаты выполнения тестов NAS NPB 2.3 на Linux-кластере ЭКСС-1	9
3. Тестовый пакет программ MPI - Pallas PMB 2.2.....	34
3.1 Состав пакета	34
3.2. Методология тестов	35
3.2.1 Параметры, управляющие ПТ-MPI	36
3.2.2. Коммуникаторы, активные процессы	37
3.2.3 Размеры сообщений	38
3.2.4 Инициализация буферов.....	38
3.2.5 Подготовительная фаза.....	38
3.2.6 Синхронизация	38
3.2.7 Измеряющий тест.....	38
3.3 Описания тестов ПТ-MPI.....	40
3.3.1 Классификация тестов	40
3.3.1.1 Тесты Отдельных Передач.....	40
3.3.1.2 Тесты Параллельных Передач.....	40
3.3.1.3 Коллективные тесты	41
3.3.2 Описание тестов Отдельных Передач.....	41
3.3.2.1 Тест PingPong	42
3.3.2.2 Тест PingPing	43
3.3.3 Описание тестов Параллельных Передач	44
3.3.3.1 Тест Sendrecv	45
3.3.3.2 Тест Exchange	46
3.3.4 Описание Коллективных тестов	47
3.3.4.1 Тест Reduce.....	47
3.3.4.2 Тест Reduce_scatter	48
3.3.4.3 Тест Allreduce	48
3.3.4.4 Тест Allgather.....	48
3.3.4.5 Тест Allgatherv.....	48
3.3.4.6 Тест Alltoall.....	49
3.3.4.7 Тест Bcast.....	49
3.3.4.8 Тест Barrier	49
4. Результаты выполнения тестов PMB 2.2 на Linux-кластере ЭКСС-1. .	50
Литература	71

Введение. Описание ЭКСС-1

ЭКСС-1 - шестнадцатиузловой ЭКСС-КУ с топологией 2-мерного тора в следующей конфигурации:

- 16 узлов в конструктиве 3U, каждый узел содержит: 2-х процессорную системную плату SuperMicro P6 DBE с чипсетом 440 BX, два микропроцессора Pentium III 800 МГц, оперативную память объемом 512 Мбайт, жесткий диск IDE объемом 10 Гбайт, флоппи-диск, встроенный сетевой адаптер Fast Ethernet, высокоскоростной адаптер SCI D311/D312 (PCI-32 33 МГц);
- Коммутатор сети Fast Ethernet.
- Управляющая машина в составе: однопроцессорная системная плата, микропроцессор Pentium III 600 МГц, оперативная память объемом 256 Мбайт, жесткий диск IDE объемом 20 Гбайт, флоппи-диск, CD, сетевой адаптер Fast Ethernet, видеокарта AGP, монитор 17", клавиатура, манипулятор "мышь".

Программное обеспечение для обеспечения работы 16-узловой ЭКСС-КУ

- Операционная система Linux (RedHat 6.2 с ядром 2.2.16).
- Средства инсталляции операционной системы Linux на кластере.
- Средства инсталляции программного обеспечения Scali (SSP-2.1.0).
- Драйвер сетевых адаптеров SCI (ScaSCI).
- Программы реализации сетевого интерфейса пользователя нижнего уровня (API SISCO).
- Библиотеку программ MPI (ScaMPI 2.1), реализующих прикладной интерфейс передачи сообщений высокого уровня по стандарту MPI 1.2 для языков программирования Фортран и Си.
- Систему конфигурирования кластера ScaConf.

Были проведены исследования эффективности разработанного кластера:

- на задачах пакетов NPВ 2.3 и специализированный для проведения профилирования набор тестов NPВ 2.3/мод
- на тестах MPI (Pallas PMB-2.2).

1. Прикладные бенчмарки – задачи пакета NAS NPB 2.3.

Задачи пакета NPB 2.3 (Numerical aerospace simulation (NAS) Parallel Benchmark версии 2.3), а также отчеты с их описанием и результатами счета на различных системах доступны на WEB-сайте по адресу <http://www.nas.nasa.gov/NAS/NPB/>. NAS – подразделение Центра Эймса NASA. Задачи относятся к области вычислительной гидродинамики (computational fluid dynamics, CFD).

Пакет содержит восемь задач, пять из которых (EP, FT, MG, IS, CG) являются задачами-ядрами (kernel benchmarks), а три (LU, SP, BT) – моделями прикладных задач (application benchmarks). Задачи были выбраны после оценки множества больших прикладных программ класса CFD, решаемых в Центре Эймса NASA.

Задачи пакета NPB 2.3 содержат значительно больше вычислений, чем использовавшиеся ранее бенчмарки, например такие, как Livermore Loops или Linpack, поэтому они более приемлемы для оценки параллельных машин. С другой стороны, эти задачи относительно просты, что позволяет ставить эти задачи на новых вычислительных системах без значительных усилий и задержек.

Программы составлены таким образом, что можно задавать необходимую вычислительную сложность, а также степень их распараллеливания на вычислительные процессы, т.е. их дробление. Предусмотрены четыре класса сложности: W-для рабочих станций и персональных машин; A, B и C – для мультипроцессорных систем (с нарастающей вычислительной сложностью от A к C). Предусмотрен вариант пропуска задач на одном процессоре, а также на произвольно заданном количестве процессоров. Производится автоматический контроль правильности выполнения вычислений, а также автоматическое измерение времени счета, для чего используется процедура MPI_Wtime и определение производительности на задаче в миллионах операций с плавающей точкой в секунду (Mflops). При этом оценивается производительность всей системы в целом, а также производительность одного процессора.

В таблице 9 приведены общие сведения о размерности задач с вычислительной сложностью A, B и C, количество выполняемых операций с плавающей точкой и производительность двух классических супермашин на этих задачах [9].

При распараллеливании задач пакета следует учитывать дополнительные условия, накладываемые на количество процессоров, которое можно при этом использовать:

- EP - произвольное число процессоров;
- IS, CG, FT, MG, LU - один или степень 2-х (1, 2, 4, 8, 16...);
- SP, BT - один или квадрат целого (1, 4, 9, 16 ...).

Каждая из задач имеет свою отличительную особенность с точки зрения использования вычислительных ресурсов.

EP.

Данная задача позволяет оценивать наибольшую производительность на задачах с интенсивными вычислениями с плавающей точкой двойной точности. Межпроцессорные взаимодействия незначительны. Вместе с тем, в ней применяется датчик случайных чисел, использующий 8-байтовую целочисленную арифметику. Если процессор не поддерживает аппаратно такую

целочисленную арифметику, то применяется датчик, использующий арифметику с плавающей точкой двойной точности (это имеет место, например, для Pentium II). В этом случае оценка производительности на данной задаче, которую выдает программа, оказывается несколько заниженной.

Таблица 1-1. Размерность и количество операций с плавающей точкой задач пакета NPB 2.3.

Производительность эталонных супермашин, достигаемая на этих задачах.

Наименование	Abb.	Class A			Class B			Class C
		Размер	Операций ($\times 10^9$)	Cray Y-MP/1 Mflops	Размер	Операций ($\times 10^9$)	Cray C-90/1 Mflops	Размер
Embarassingly Parallel	EP	2^{28}	26.68	211	2^{30}	100.9	689	2^{32}
Multigrid	MG	256^3	3.905	176	256^3	18.81	557	512^3
Conjugate Gradient	CG	14000	1.508	127	75000	54.89	447	150000
3-D FFT PDE	FT	$256^2 \times 128$	5.631	196	512×256^2	71.37	645	512^3
Integer Sort	IS	$2^{23} \times 2^{19}$	0.7812	68	$2^{25} \times 2^{21}$	3.150	244	$2^{27} \times 2^{27}$
LU solver (appl. CFD)	LU	64^3	64.57	194	102^3	319.6	711	162^3
Pentadiagonal Solver (appl. CFD)	SP	64^3	102.0	216	102^3	447.1	648	162^3
Block Tridiagonal Solver (appl. CFD)	BT	64^3	181.3	229	102^3	721.5	705	162^3

MG.

Упрощенный метод вычислений на множестве сеток. Характерны высокоструктурированные взаимодействия между далеко отстоящими друг от друга процессорами, тестируются как короткие, так и длинные межпроцессорные взаимодействия. Для задач класса А максимальный объем сообщений – около 540 Кбайт.

CG.

Метод сопряженных градиентов используется для итерационного вычисления приближения наименьшего собственного значения большой, разреженной симметричной положительно определенной матрицы. Эта задача типична для вычислений на неструктурированных сетках. Тестируются нерегулярные взаимодействия между далеко отстоящими друг от друга процессорами, включая неструктурированное умножение матрицы на вектор (обращение к памяти по индекс-вектору).

FT.

Применение быстрого преобразования Фурье при решении трехмерного дифференциального уравнения с частными производными. Эта задача существенна для многих приложений, использующих спектральные методы. Является тяжелым тестом на производительность при обменах между далеко отстоящими друг от друга процессорами. Используется датчик случайных чисел, как в задаче EP.

IS.

Сортировка большого объема целых чисел. Эта задача существенна для программ, использующих “метод частиц”. Тестируется скорость целочисленных вычислений и производительность межпроцессорных взаимодействий. Используется датчик случайных чисел, как в задаче EP.

LU.

Нахождение конечно-разностного решения 3-х мерной системы уравнений Навье-Стокса для сжимаемой жидкости или газа. Применяется метод LU-разложения, с использованием алгоритма SSOR. Доля межпроцессорных взаимодействий мала. Передается большое количество коротких сообщений (не более 40 байт). Значительный коэффициент переиспользования данных в кеше.

SP.

Неявный алгоритм нахождения конечно-разностного решения 3-х мерной системы уравнений Навье-Стокса для сжимаемой жидкости или газа. Скалярная пентадиагональная схема. Большая гранулированность взаимодействий.

BT.

Неявный алгоритм нахождения конечно-разностного решения 3-х мерной системы уравнений Навье-Стокса для сжимаемой жидкости или газа. Блок-диагональная схема. Большая гранулированность взаимодействий.

Перечисленные задачи обладают следующими особенностями выполнения на мультипроцессорных системах. В зависимости от этих особенностей эти восемь задач можно разбить на следующие подгруппы.

Подгруппа 1 - EP

Вычислительный тест, коммуникаций практически нет, используется для оценки производительности системы на операциях с числами с плавающей точкой (тест также зависит от эффективности реализации 8-байтовой целой арифметики, используемой в датчике случайных чисел).

Подгруппа 2 - BT, MG, CG

Вычислительные тесты с небольшой долей коммуникационных потерь (менее 20% для 32 процессоров). Эффективность вычислений с увеличением количества используемых процессоров меняется незначительно. Таким образом, сбалансированной компенсации потерь на коммуникации нет. Поскольку доля вычислений велика, то потери на коммуникации не так заметны.

Подгруппа 3 - LU, SP

Вычислительные тесты, доля коммуникаций средняя (около 30% для 32 процессоров). Тесты с компенсирующими друг друга увеличением эффективности вычислений и затратами на коммуникации. Тест LU обычно на практике используется для оценки эффективности работы реализаций MPI при передаче коротких сообщений (оценка задержек передачи сообщений в сети).

Подгруппа 4 - FT, IS

Тест FT - вычислительный, IS - параллельная сортировка большого массива целых чисел. Доля затрат в обоих тестах на коммуникации подавляющая, может

в несколько раз превосходить затраты на вычисления. Тест IS обычно используется на практике для оценки уровня пропускной способности сети.

Числовые данные для перечисленных тестов, полученные на 16-узловом SCI-кластере ЭКСС-1 приведены далее в таблицах.

Отметим, что наивысший уровень производительности на задачах этого пакета (имеются сведения о задачах LU, FT и SP) был получен в [1] для кластера на базе микропроцессора Alpha 21264 (667 МГц) и сети Quadrics. Например, производительность на один процессор на задаче LU для 16 процессорной конфигурации составляет 345 Mflops.

2. Результаты выполнения тестов NAS NPВ 2.3 на Linux-кластере ЭКСС-1

















Для указания используемых при прогоне теста двухпроцессорных узлов кластера применена схема занятости узлов со следующими обозначениями:

- – узел не используется;
- ▣ – в узле задействован ОДИН процессор;
- – в узле задействованы ДВА процессора;

Таблица 1. Задача LU(H), кластер ЭКСС-1, класс А.

Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1	<pre> ■□□□ □□□□ □□□□ □□□□ </pre>	1253,18	95,20	95,20
1(2)	<pre> ■▣□□ □□□□ □□□□ □□□□ </pre>	632,97	188,47	94,24
	<pre> ■□□■ □□□□ □□□□ □□□□ </pre>	633,08	188,44	94,22
	<pre> ■□□□ □□□□ □□□□ □□□■ </pre>	633,13	188,24	94,21
1(4)	<pre> ■□□■ □□□□ □□□□ ■□□■ </pre>	318,53	374,52	93,63
	<pre> □□□■ □□□■ □□□■ ■□□□ </pre>	319,29	373,63	93,41
	<pre> ■□□□ ■□□□ □□□■ □□□■ </pre>	318,58	374,46	93,62
	<pre> ■□□■ □□□□ □□□□ □□□□ </pre>	318,38	374,70	93,67
1(8)	<pre> ■□□■ ■□□■ □□□□ □□□□ □□□□ </pre>	161,36	739,30	92,41
	<pre> □□□□ □□□□ ■□□■ ■□□■ ■□□■ </pre>	174,05	685,40	85,68
1(16)	<pre> ■□□■ ■□□■ ■□□■ ■□□■ </pre>	66,52	1793,48	112,09

Таблица 2. Задача LU(F), кластер ЭКСС-1, класс А.

Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1		1253,18	95,20	95,20
2		937,47	127,25	63,63
		1096,89	108,76	54,38
		948,71	125,75	62,87
		933,80	127,75	63,88
		947,05	125,97	62,98
		947,84	125,86	62,93
		919,14	129,79	64,90
		939,63	126,96	63,48
		946,43	126,05	63,02
		943,81	126,40	63,20
		943,47	126,44	63,22
		950,17	125,55	62,78
		937,63	127,23	63,62
		949,82	125,60	62,80
		946,02	126,10	63,05












		937,38	127,27	63,63
4		474,56	251,39	62,85
8		227,21	525,06	65,63
		227,72	523,88	65,49
		226,77	526,07	65,76
		254,88	468,04	58,51
16		78,85	1513,04	94,57
		84,94	1404,46	87,94
		79,14	1507,47	94,22
		78,45	1520,66	95,04
32		39,70	3004,88	93,90

Таблица 3. Задача ВТ (Н), кластер ЭКСС-1, класс А.












Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1		2020,35	83,29	83,29
1(4)		528,92	318,17	71,54
		536,24	313,82	78,46
		530,10	317,46	79,36
1(9)		244,57	688,10	76,46
		229,55	733,11	81,46
		235,41	714,85	79,43
1(16)		141,52	1189,10	74,32
		140,58	1196,26	74,77
		142,49	1181,06	73,82
		141,78	1186,93	74,18

Таблица 4. Задача ВТ (F), кластер ЭКСС-1, класс А.












Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1		2020,35	83,29	83,29
4		829,33	202,92	50,73
9		388,64	433,01	43,30
16		219,62	766,24	47,89
		227,02	741,29	46,33
		227,06	741,15	46,32
		219,65	766,15	47,88
		229,07	734,64	45,91
25		170,09	989,40	38,05
		168,14	1000,87	38,45
		164,08	1025,60	39,45

Таблица 5. Задача SP(H), кластер ЭКСС-1, класс А.

Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1		1431,69	59,38	59,38
1(4)		394,68	215,39	53,85
1(9)		182,52	465,76	51,75
		177,52	478,86	53,21
		173,47	490,05	54,45
1(16)		106,40	798,97	49,94
		106,04	801,69	50,11

Таблица 6. Задача SP(F), кластер ЭКСС-1, класс А.









Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
4		1431,69	59,38	59,38
9		227,72	306,10	30,61
16		168,57	504,31	31,52
		170,97	497,22	31,08
		175,72	483,78	30,24
		175,25	485,08	30,32
25		139,08	611,24	23,51
		146,20	581,48	22,36

Таблица 7. Задача EP (H), кластер ЭКСС-1, класс А.

Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1		383,43	1,40	1,40
1(2)		191,94	2,80	1,40
1(4)		96,16	5,58	1,40
1(8)		48,48	11,07	1,38
1(16)		24,84	21,61	1,35

Таблица 8. Задача EP (F), кластер ЭКСС-1, класс А.





Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
2		193,75	2,77	1,39
4		97,49	5,51	1,38
8		49,15	10,92	1,37
16		25,54	21,02	1,31
32		12,89	41,67	1,30

Таблица 9. Задача MG (H), кластер ЭКСС-1, класс А.


















Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1		62,60	62,18	62,18
		62,39	62,39	62,39
		65,93	59,04	59,04
1(2)		31,40	123,94	61,97
		31,48	123,65	61,83
		31,31	124,31	62,15
		31,16	124,93	62,46
1(4)		17,96	216,69	54,17
		18,19	214,03	53,51
		18,06	215,55	53,89
		18,03	215,84	53,96
1(8)		8,58	453,82	56,73
		8,52	457,05	57,13
		8,50	457,65	57,21
1(16)		4,94	787,96	49,25
		4,94	787,60	49,22
		4,95	786,32	49,14

Таблица 10. Задача MG (F), кластер ЭКСС-1, класс А.

Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
2	■□□□ □□□□ □□□□	43,69	89,08	44,54
	□□□□ □□□□ □□□■	43,97	88,52	44,26
	□□□□ □□□□ □■□□	51,08	76,21	38,10
4	■□□□ □□□□ □□□□	25,81	150,79	37,70
	■□□□ ■□□□ □□□□	25,67	151,65	37,91
	■□□■ □□□□ □□□□	25,65	151,75	37,94
	■□□□ □□□□ □□□■	25,80	150,88	37,72
8	■□□□ □□□□ □□□□	11,76	331,01	41,38
	□□□■ □□■□ ■□□□	11,76	331,08	41,38
	■□□□ □■□□ □□□■	11,80	329,93	41,24
	■□□□ ■□□□ □□□□	11,74	331,58	41,45
16	■□□□ ■□□□ □□□□ □□□□	6,87	566,78	35,42
	□□□■ □□■□ □□■□ □□■□	6,86	567,76	35,49
	■□□□ ■□□□ □□□□ □□■□	6,92	562,52	35,16
32	■□□□ ■□□□ ■□□□ ■□□□	4,29	908,11	28,38
	■□□□ ■□□□ ■□□□ ■□□□	4,26	913,12	28,54
	■□□□ ■□□□ ■□□□ ■□□□	4,32	900,60	28,14

Таблица 11. Задача FT(H), кластер ЭКСС-1, класс А.


































Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1		86,78	82,24	82,24
		86,60	82,41	82,41
		91,27	78,19	78,19
1(2)		55,23	129,22	64,61
		55,22	129,23	64,62
		55,14	129,43	64,72
1(4)		29,99	238,00	59,50
		29,89	238,79	59,70
		29,79	239,58	59,89
		29,75	239,88	59,87
1(8)		15,48	461,16	57,64
		15,50	460,32	57,54
		15,45	461,84	57,73
1(16)		8,12	879,20	54,95
		8,04	887,54	55,47
		8,05	886,78	55,42

Таблица 12. Задача FT(F), кластер ЭКСС-1, класс А.

Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
2	■□□□ □□□□ □□□□ □□□□	68,38	104,36	52,18
	□□□□ □□□□ □□□□ □□□■	68,64	103,97	51,98
	□□□□ □□□□ □□□□ □■□□	75,65	94,34	47,17
4	■□□□ □□□□ □□□□ □□□□	38,42	185,73	46,43
	■□□□ ■□□□ □□□□ □□□□	38,28	186,43	46,61
	■□□■ □□□□ □□□□ □□□□	38,24	186,61	46,65
	■□□□ □□□□ □□□□ □□□■	37,04	192,68	48,17
8	■□□□ □□□□ □□□□ □□□□	20,56	347,03	43,38
	□□□■ □□□■ □□□■ □□□■	20,57	346,92	43,37
	■□□□ □■□□ □□□■ □□□■	20,84	342,52	42,82
	■□□□ ■□□□ □□□□ □□□□	20,68	345,15	43,14
16	■□□□ ■□□□ □□□□ □□□□	11,32	630,30	39,39
	□□□■ □□□■ □□□■ □□□■	11,33	629,69	33,36
	■□□□ ■□□□ □□□■ □□□■	11,31	630,79	39,42
32	■□□□ ■□□□ ■□□□ ■□□□	5,95	1199,81	37,49
	■□□□ ■□□□ ■□□□ ■□□□	6,01	1187,28	37,10
	■□□□ ■□□□ ■□□□ ■□□□	6,00	1189,73	37,18

Таблица 13. Задача IS(H), кластер ЭКСС-1, класс А.

Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1		10,62	7,9	7,9
1(2)		6,86	12,24	6,12
		6,86	12,22	6,11
		6,84	12,26	6,13
		6,87	12,21	6,10
		6,92	12,13	6,06
1(4)		4,04	20,78	5,20
		4,02	20,88	5,22
		4,03	20,83	5,21
		4,05	20,69	5,17
		4,05	20,73	5,18
		4,04	20,75	5,19
		4,05	20,72	5,18
		4,03	20,80	5,20
		4,18	20,06	5,01
		4,05	20,71	5,18
		4,05	20,73	5,18








1(8)		2,35	35,69	4,46
		2,36	35,60	4,45
		2,35	35,72	4,47
		2,34	35,80	4,47
(1)16		1,39	60,20	3,76
		1,40	59,85	3,74
		1,40	59,75	3,73

Таблица 14. Задача IS(F), кластер ЭКСС-1, класс А.










Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1		10,61	7,91	7,91
2		9,03	9,29	4,65
		9,01	9,31	4,66
		10,23	8,20	4,1
4		5,72	14,67	3,67
8		3,52	23,82	2,98
		3,51	23,90	2,99
16		2,26	37,07	2,32
		2,23	37,66	2,35

Таблица 15. Задача CG(H), кластер ЭКСС-1, класс А.


















Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1		29,00	51,60	51,60
		28,93	51,72	51,72
		29,58	50,59	50,59
1(2)		14,28	104,80	52,40
		14,31	104,54	52,27
		14,23	105,20	52,60
		14,31	104,56	52,28
1(4)		8,25	181,41	45,35
		8,27	180,95	45,24
		8,25	181,41	45,35
		8,24	181,57	45,39
1(8)		4,91	304,80	38,10
		4,74	315,41	39,43
		4,75	315,10	39,39
1(16)		3,43	436,06	27,25
		3,51	426,00	26,62
		3,28	455,60	28,47

Таблица 16. Задача CG(F), кластер ЭКСС-1, класс А.

Число процессоров	Используемые узлы	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
2		19,88	75,26	37,63
		19,08	78,44	39,22
		20,35	73,53	36,77
4		11,31	132,29	33,07
		11,33	132,06	33,01
		11,25	133,06	33,27
		11,10	134,83	33,71
8		6,91	216,43	27,05
		6,83	218,94	27,37
		6,80	220,02	27,50
		6,85	218,49	27,31
16		4,73	316,69	19,79
		4,82	310,69	19,42
		4,50	332,68	20,79
32		4,91	304,64	9,52
		4,13	362,42	11,33






		3,33	449,88	14,06
		5,01	298,80	9,34
		3,36	445,04	13,91
		4,49	333,06	10,41
		3,56	420,59	13,14

Таблица 17. Задача IS(H), кластер ЭКСС-1, класс В.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1	40,51	8,28	8,28
1(2)	26,35	12,73	6,37
1(4)	16,03	20,94	5,23
1(8)	9,23	36,34	4,54
1(16)	5,49	61,12	3,82

Таблица 18. Задача IS(F), кластер ЭКСС-1, класс В.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1	40,51	8,28	8,28
2	33,02	10,16	5,08
4	21,79	15,40	3,85
8	13,37	25,09	3,14
16	8,02	41,86	2,67
32	5,47	61,36	1,92

Таблица 19. Задача FT(H), кластер ЭКСС-1, класс В.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1(8)	211,58	435,07	54,38
1(16)	115,55	796,64	49,79

Таблица 20. Задача FT(F), кластер ЭКСС-1, класс В.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
16	168,84	545,22	34,08
32	85,96	1070,93	33,47

Таблица 21. Задача MG(H), кластер ЭКСС-1, класс В.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1(16)	22,47	866,01	54,13

Таблица 22. Задача MG(F), кластер ЭКСС-1, класс В.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
16	32,64	596,32	37,27
32	20,35	956,13	29,88

Таблица 23. Задача CG(H), кластер ЭКСС-1, класс В.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1(4)	500,43	109,32	27,33
1(8)	199,43	274,33	34,29
1(16)	123,51	442,95	27,68

Таблица 24. Задача CG(F), кластер ЭКСС-1, класс В.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
4	822,18	66,54	16,64
8	297,10	184,14	23,02
16	183,29	298,48	18,65
32	96,14	569,04	17,78

Таблица 25. Задача VT(H), кластер ЭКСС-1, класс В.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1(16)	556,28	1262,29	78,89

Таблица 26. Задача VT(F), кластер ЭКСС-1, класс В.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
16	837,65	838,27	52,39
25	570,12	1231,64	47,37
25	565,51	1241,69	47,76

Таблица 27. Задача SP(H), кластер ЭКСС-1, класс В.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1(16)	415,78	853,84	53,36

Таблица 28. Задача SP(F), кластер ЭКСС-1, класс В.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
16	650,27	545,95	34,12
25	490,21	724,20	27,85

Таблица 29. Задача IS(H), кластер ЭКСС-1, класс С.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1(4)	64,71	20,74	5,18
1(8)	37,00	36,27	4,53
1(16)	21,56	62,26	3,89

Таблица 30. Задача IS(F), кластер ЭКСС-1, класс С.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
8	53,20	25,23	3,15
16	31,53	42,57	2,66
32	21,47	62,52	1,95

Таблица 31. Задача FT(H), кластер ЭКСС-1, класс С.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1(16)	658,00	602,42	37,65

Таблица 32. Задача FT(F), кластер ЭКСС-1, класс С.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
32	738,80	536,53	16,77

Таблица 33. Задача MG(H), кластер ЭКСС-1, класс С.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1(16)	156,10	997,40	62,34

Таблица 34. Задача MG(F), кластер ЭКСС-1, класс С.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
16	217,10	717,13	44,82
32	128,44	1212,17	37,88

Таблица 35. Задача CG(H), кластер ЭКСС-1, класс С.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
1(16)	408,77	350,68	21,92

Таблица 36. Задача CG(F), кластер ЭКСС-1, класс С.

Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
16	623,20	230,02	14,38
32	263,48	544,04	17,00

Таблица 37. Задача BT(F), кластер ЭКСС-1, класс С.

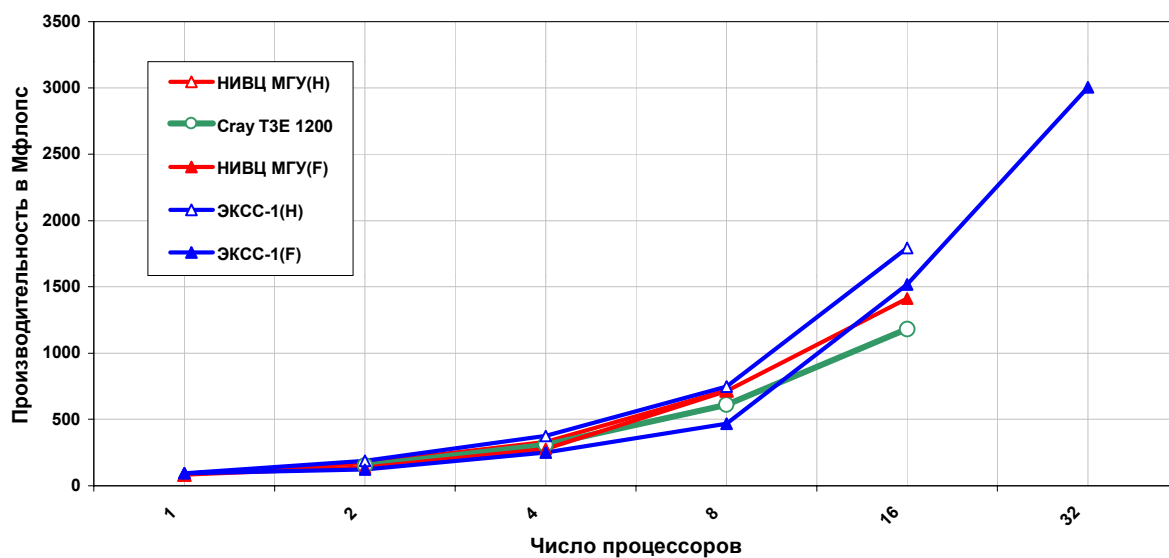
Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
25	2246,57	1275,85	49,07

Таблица 38. Задача SP(F), кластер ЭКСС-1, класс С.

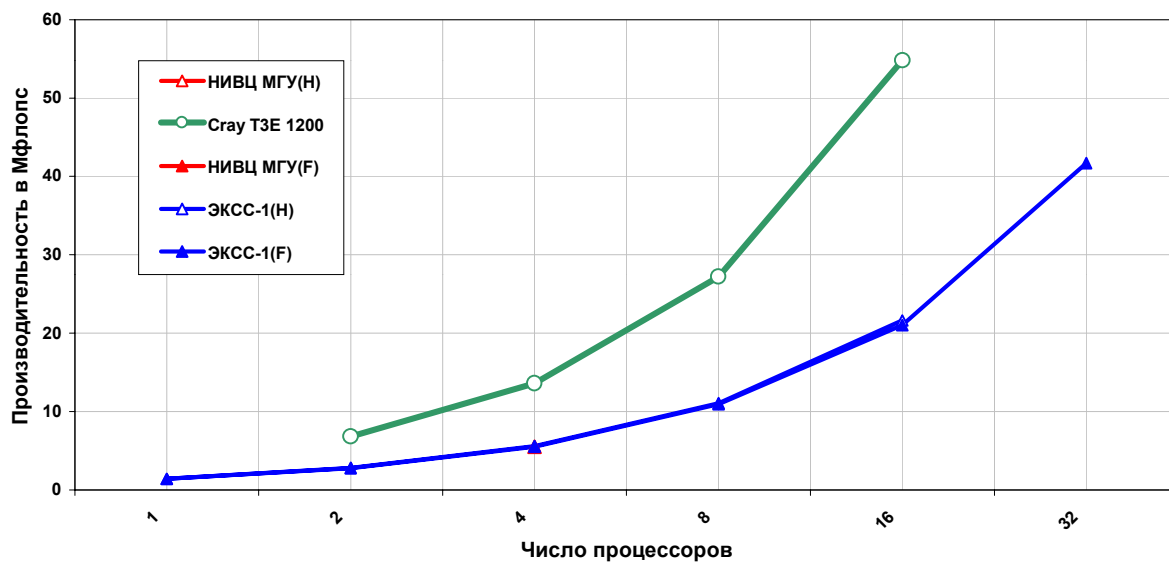
Число процессоров	Время решения (сек)	Общая производительность (mflops)	Производительность на один процессор (mflops)
25	1753,44	827,00	31,81

На представленных ниже графиках произведен сравнительный анализ производительности ЭКСС-1 с другими высокопроизводительными системами.

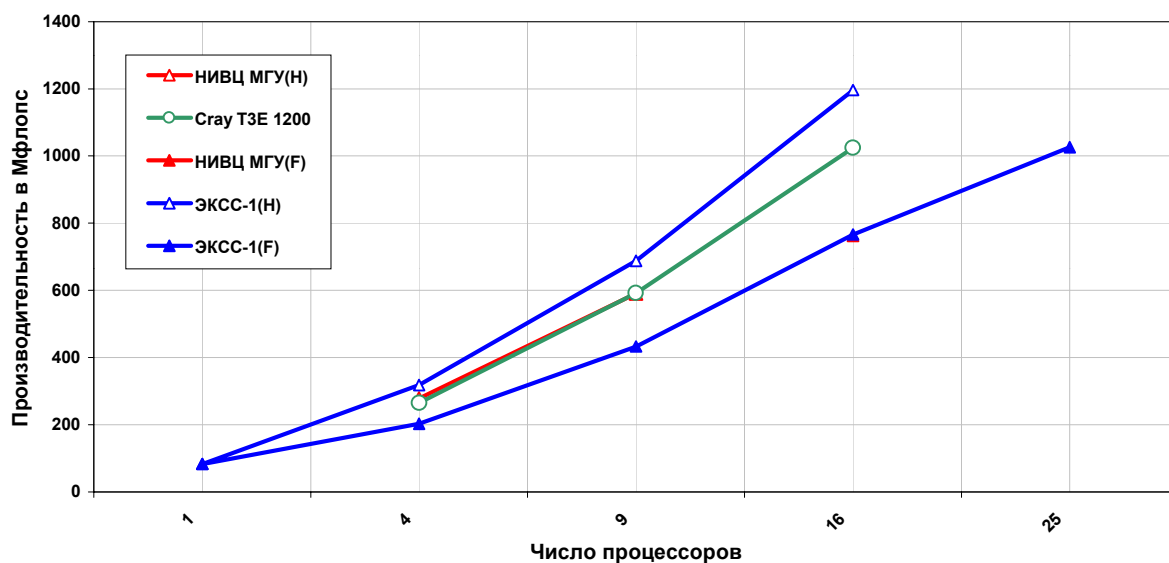
LU (A) : общая производительность



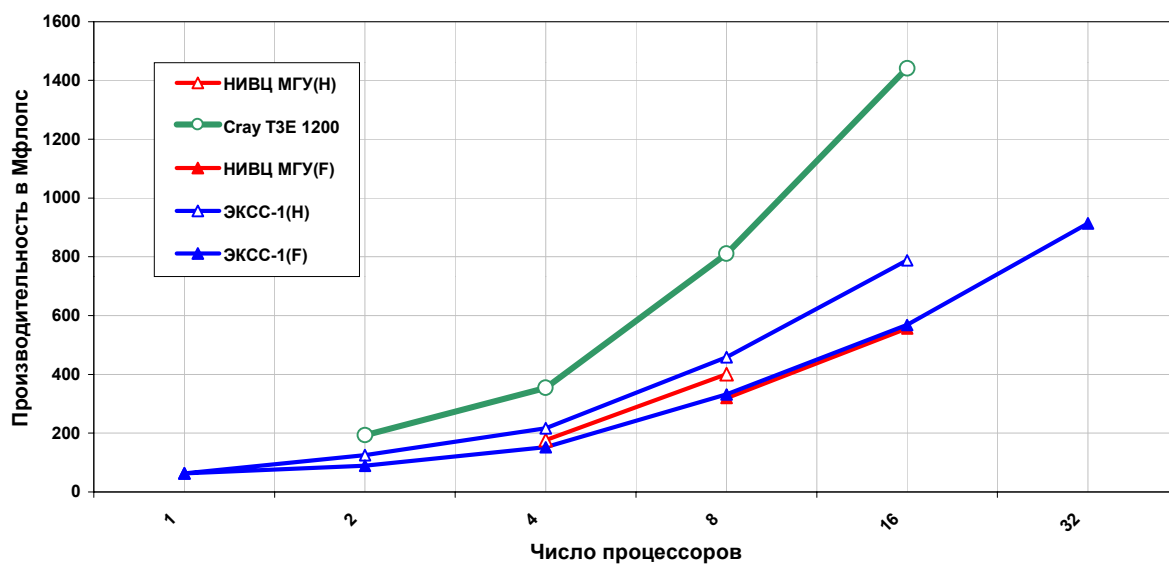
EP(A) : общая производительность



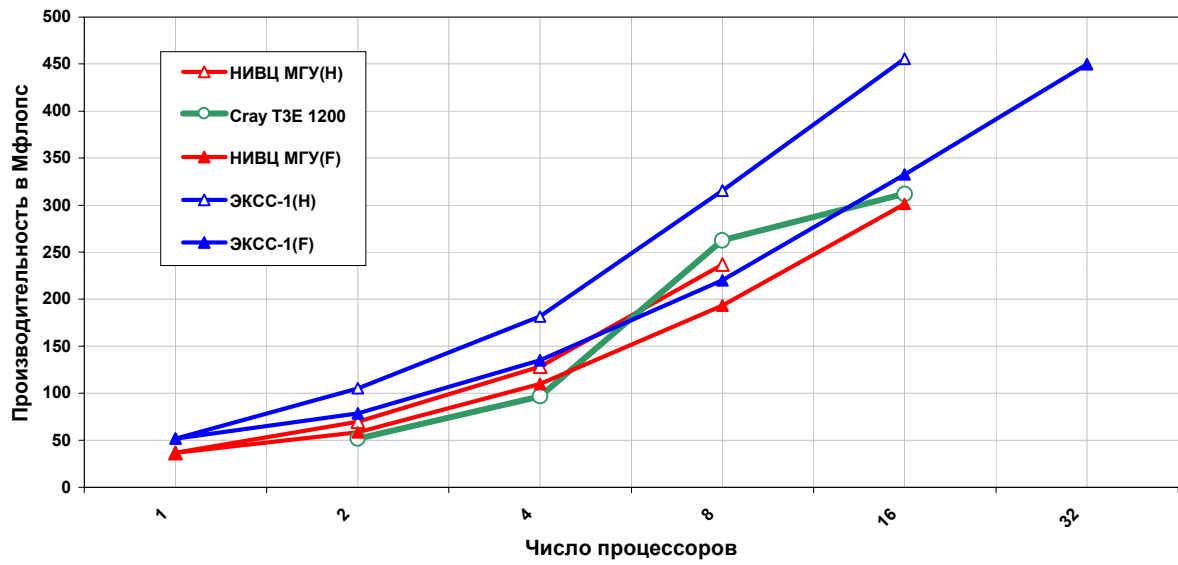
BT(A) : общая производительность



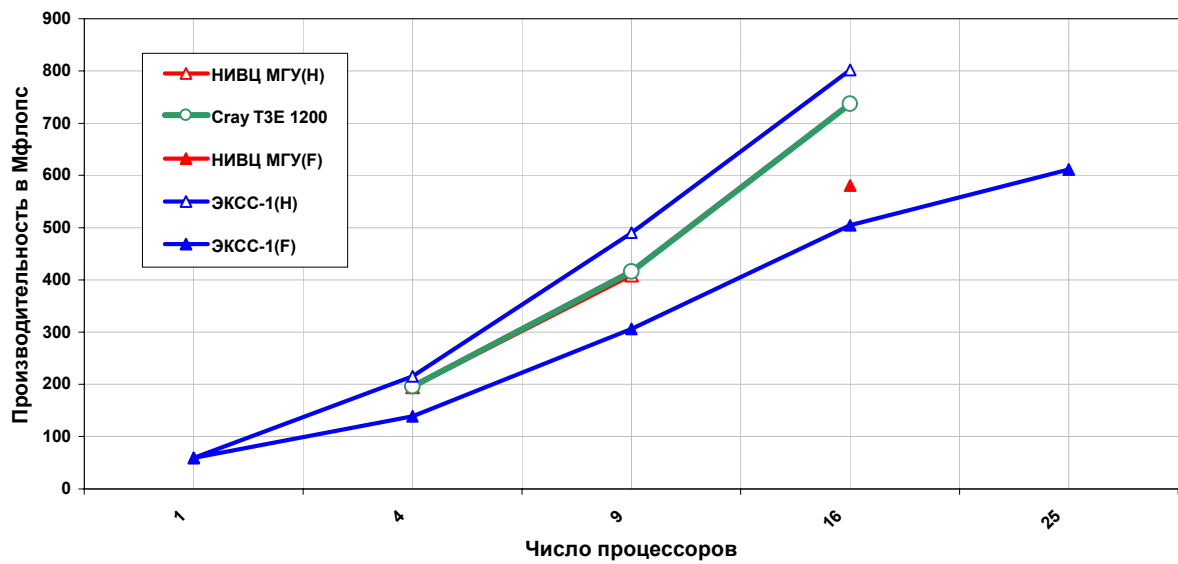
MG (A) : общая производительность



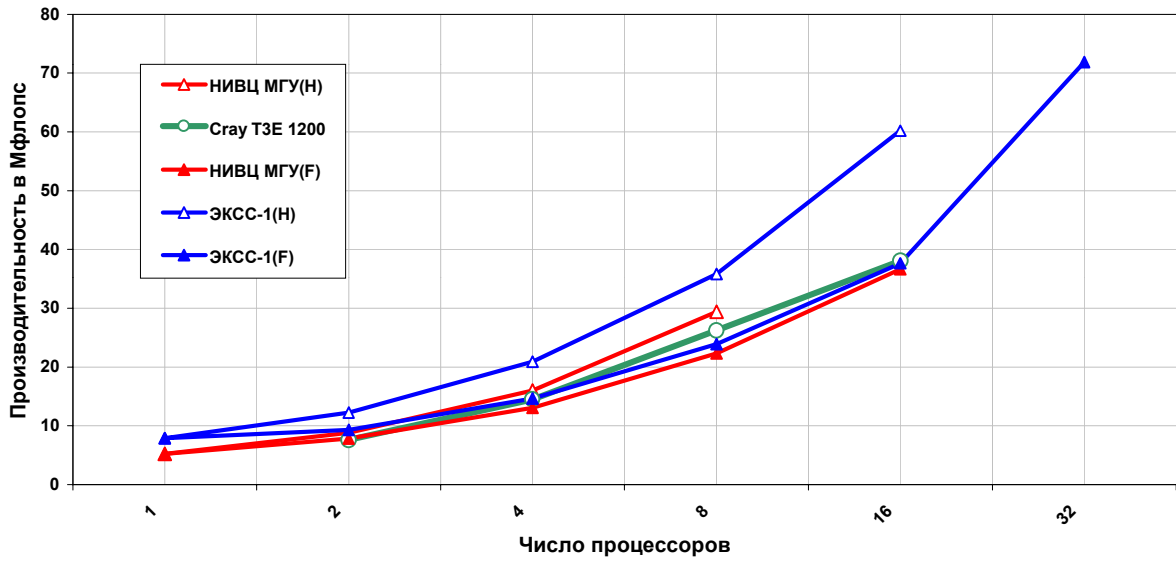
CG (A) : общая производительность



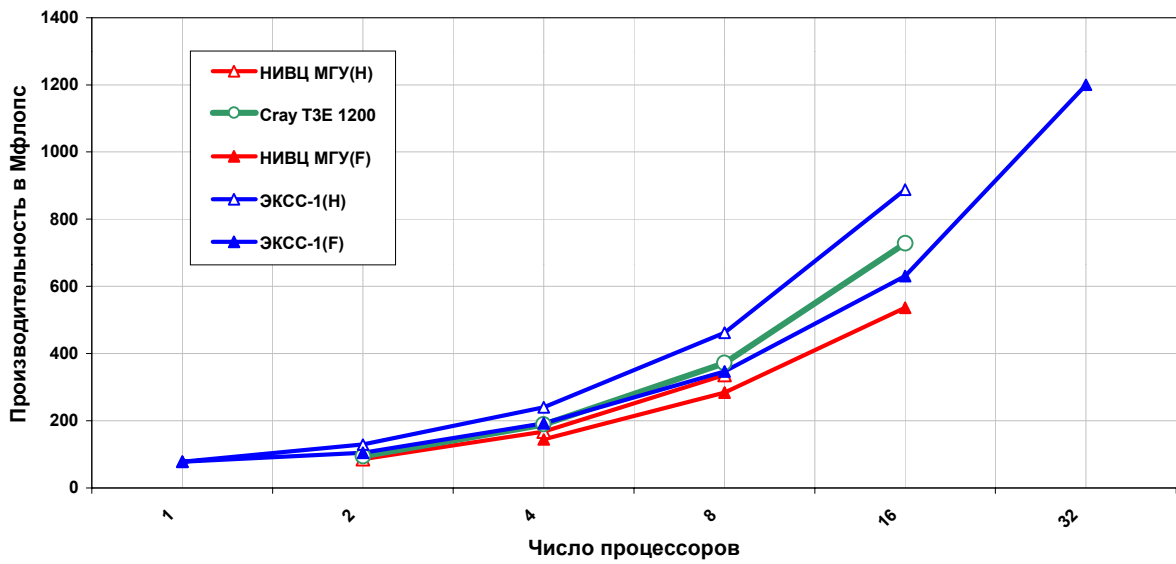
SP(A) : общая производительность



IS (A) : общая производительность



FT (A) : общая производительность



3. Тестовый пакет программ MPI - Pallas PMB 2.2

В этой части документа представляется пакет тестов MPI (ПТ-MPI), основанный на части PMB-MPI1 пакета тестов PMB 2.2.[2] (Известно, что пакет PMB 2.2 состоит из 3 частей: PMB-MPI1, PMB-EXT, PMB-IO.)

В ПТ-MPI реализован краткий набор MPI-тестов.[3]

С помощью одного исполнимого файла можно запустить на выполнение все тесты или какое-либо их подмножество из командной строки. Такие величины, как количество измерений (число повторений тестов), размер сообщений, выбор многогруппового режима выполнения или выполнение теста одной группой из всех процессов, являются программными параметрами.

В ПТ-MPI предусмотрены стандартная и дополнительная конфигурации. В стандартной конфигурации, все параметры, отмеченные далее, жестко фиксированы и не должны изменяться. Для определенных систем бывает необходимым изменить таблицу результатов, в этом случае пользователь может по своему выбору менять определенные параметры.

Такие величины, как минимальное количество процессов P_{min} и максимальное количество процессов P может быть задано пользователем в командной строке. Тесты выполняются P_{min} , $2P_{min}$, $4P_{min}$, ... $2^x P_{min} < P$ и P процессами.

3.1 Состав пакета

Текущая версия программного пакета ПТ-MPI содержит следующие тесты:

1. PingPong
2. PingPing
3. Sendrecv
4. Exchange
5. Bcast
6. Allgather
7. Allgatherv
8. Alltoall
9. Reduce
10. Reduce_scatter
11. Allreduce
12. Barrier

ПТ-MPI предоставляет возможность выполнять все тесты в одной группе процессов - в дальнейшем будем называть такие тесты *первичными* тестами, или в нескольких группах процессов - а такие тесты будем называть *мульти-тестами*. Имена мульти-тестов совпадают с именами первичных тестов, однако, в отчете о результатах тестирования будут выводиться на экран имена первичных тестов с предлогом Multi- . Например, при запуске PingPong на 4 процессах в режиме работы нескольких групп процессов, будет сформировано $N/2$ групп (по 2 процесса в каждой), все процессы будут одновременно выполнять PingPong.

Названия мульти-тестов при выводе результатов выглядят следующим образом:

1. Multi-PingPong
2. Multi-PingPing
3. Multi-Sendrecv
4. Multi-Exchange
5. Multi-Bcast
6. Multi-Allgather

7. Multi-Allgather
8. Multi-Alltoall
9. Multi-Reduce
10. Multi-Reduce_scatter
11. Multi-Allreduce
12. Multi-Barrier

3.2. Методология тестов

В этой части объясняется методология тестов.

Некоторые управляющие параметры жестко закодированы (подобно заданию числа процессов для выполнения тестов), некоторые же управляющие параметры устанавливаются в центральном заголовочном файле с помощью директив препроцессора. Для управления ПТ-MPI существуют стандартный и дополнительный режимы. В стандартном режиме все конфигурационные величины предварительно определены и не должны изменяться. Это гарантирует совместимость результирующих таблиц в стандартном режиме. В дополнительном режиме пользователь может устанавливать эти параметры по своему собственному выбору. Например, этот режим может быть использован для изменения максимального и/или минимального размеров сообщений.

Следующий граф показывает обзор потока управления внутри ПТ-MPI.

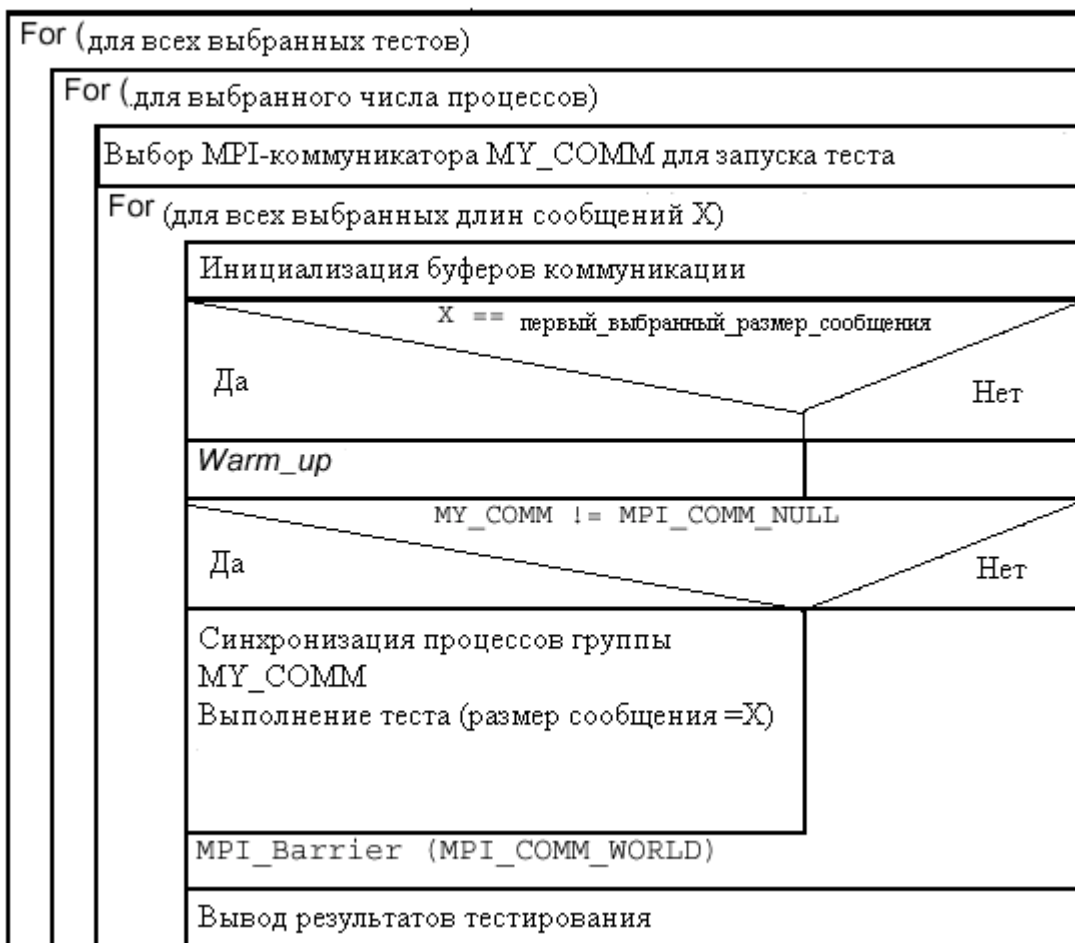


Рис.1 Поток управления в ПТ-MPI

3.2.1 Параметры, управляющие ПТ-MPI

Существует 9 параметров, управляющие работой пакета ПТ-MPI.

Они представлены в файле определений settings.h. ПТ-MPI предоставляет две установки параметров: стандартную и дополнительную. Полный список параметров ПТ-MPI и их объяснение представлены ниже в Таблице 3.2.1-1.

Таблица 3.2.1-1

Параметр (стандартный режим)	Значение
PMB_OPTIONAL	Устанавливается, когда должны быть задействованы пользовательские установки (дополнительный режим)
MINMSGLOG (0)	Размер данных = $\max(\text{unit}, 2^{\text{MINMSGLOG}})$, где unit=размер типа float для вычислительных MPI-операций и unit =1 для обычных MPI-операций.
MAXMSGLOG (22)	Самый большой размер сообщений = $2^{\text{MAXMSGLOG}}$. Используемые размеры 0, 2^i ($i = \text{MINMSGLOG}, \dots, \text{MAXMSGLOG}$).
MSGPERSAMPLE (1000)	Максимальное число повторений выполнения тестов.
OVERALL_VOL (40 Мбайт)	Предельное общее количество переданных байтов сообщений в ходе повторений выполнения теста. Для всех размеров данных < OVERALL_VOL число повторений уменьшается так, чтобы общее переданное количество байтов сообщений было не более чем OVERALL_VOL. Это предотвращает ненужные повторения для сообщений большого размера. Число повторов (n_sample)= $= \min(\text{MSGPERSAMPLE}, \max(1, \text{OVERALL_VOL}/x)), (x>0), x\text{-размер сообщения в байтах.}$
N_WARMUP (2)	Количество предварительных прогонов теста перед непосредственным выполнением.
N_BARR (2)	Количество выполнений MPI-операции MPI_Barrier для синхронизации процессов, выполняемой внутри теста непосредственно перед MPI-операцией передачи данных.
TARGET_CPU_SECS	Параметр в ПТ-MPI не задействован, он используется в PMB-EXT, PMB-IO.
MSGS_NONAGR	Параметр в ПТ-MPI не задействован, он используется в PMB-EXT, PMB-IO.

3.2.2. Коммуникаторы, активные процессы

Управление коммуникаторами производится в каждом шаге выбора коммуникатора MY_COMM на рис.1. Если коммуникатор MY_COMM существует, то этот коммуникатор освобождается.

Субкоммуникаторы формируются из группы, состоящей из процессов глобального коммуникатора MPI_COMM_WORLD:

{0,...,Q-1} (случай первичных тестов)

{0,...,Q-1}, {Q,...,2Q-1},... -(случай мультитестов)

Все процессы, которые принадлежат определенной группе(-ам) (кроме глобальной группы MPI_COMM_WORLD), называются *активными* процессами, соответствующий им коммуникатор называется MY_COMM, он используется в MPI-функциях, определяющих

тест. Все остальные процессы называются *неактивными* (*пассивными*) процессами, им соответствует коммуникатор – `MPI_COMM_NULL`.

3.2.3 Размеры сообщений

Значения минимального и максимального размеров сообщений устанавливаются в файле `settings.h`.

3.2.4 Инициализация буферов

Коммуникационные буферы динамически распределяются как `void*`. Чтобы назначить буферу содержимое, необходимо выполнить приведение типов. С одной стороны приведение типов данных обязательно для вычислительных тестов, с другой стороны оно облегчает проверку результатов.

Тип буфера в ПТ-MPI устанавливается с помощью `typedef assign_type` в файле `settings.h`. В данный момент выбран тип `float` для ПТ-MPI. Значения устанавливаются с помощью макроопределения:

```
#define BUF_VALUE(rank,i) ( 0,1*(rank+1)+(float)(i) )
```

В каждой инициализации коммуникационные буферы выглядят как типизованные массивы и инициализируются следующим образом:

```
((assign_type*)buffer)[i] = BUF_VALUE(rank,i);
```

где `rank`-номер вызывающего процесса в группе, определяемой коммуникатором `MY_COMM`.

3.2.5 Подготовительная фаза

Перед запуском измеряющего теста, выбранный тест выполняется `N_WARMUP` раз (опред. в `settings.h`) на основе максимального размера сообщения. Это должно скрывать обычные расходы на инициализацию системы передачи сообщений.

3.2.6 Синхронизация

Для синхронизации процессов перед выполнением измеряющего теста MPI-функция `MPI_Barrier(MPI_COMM)` (рис.1) выполняется `N_BARR` раз (опред. в `settings.h`).

3.2.7 Измеряющий тест

Из-за неточности измерений времени каждый тест выполняется повторно. Число повторений теста определяется макроопределением `MSGSPERSAMPLE` (константой, которая определена в файле `settings.h`). Чтобы избежать чрезмерных затрат времени выполнения в случае передачи сообщений большой длины, устанавливается верхняя граница повторений, равная `OVERALL_VOL / X` (`OVERALL_VOL` определена в файле `settings.h`, `X`-размер сообщения).

В конце концов, число повторений выполнения для всех тестов при передаче сообщений длиной `X` байт определяется следующим образом:

$$n_sample = MSGSPERSAMPLE \text{ (для } X=0\text{)}$$

$n_sample = \max(1, \min(MSGSPERSAMPLE, OVERALL_VOL/X))$ (для $X > 0$)

Ключевое измерение производится по следующей схеме:

```
for ( i=0; i<N_BARR; i++ ) MPI_Barrier(MY_COMM)
```

```
time = MPI_Wtime()
```

```
for ( i=0; i<n_sample; i++ )
```

```
    выполнение MPI-образца
```

```
time = (MPI_Wtime()-time)/n_sample
```

Выполнение MPI-образца представляет собой чистый образец теста без каких-либо дальнейших вызовов функций. Для тестов `Bcast` и `Reduce` необходим `root`-процесс для их операций. Номер `root`-процесса изменяется в цикле с каждой итерацией i : $i\%$ (Номер процесса в группе). Значение аргумента коммуникатора в MPI-функциях тестов, составляющих образец теста, определяется в 3.2.2.

3.3 Описания тестов ПТ-MPI

В этой части рассматриваются все тесты пакета. Объясняются основные принципы тестов.

3.3.1 Классификация тестов

Для более ясного представления набора тестов, тесты в ПТ-MPI делятся на следующие типы: тесты Отдельных Передач, тесты Параллельных Передач и Коллективные тесты. В тестах различного типа формат выводимых результатов выполнения теста выглядит по-разному. Данная классификация тестов представлена в Таблице 3.3.1-1.

Таблица 3.3.1-1. Классификация тестов

ПТ-MPI		
Тесты Отдельных Передач	Тесты Параллельных Передач	Коллективные тесты
PingPong	Sendrecv	Bcast
PingPing	Exchange	Allgather
	Multi-PingPong	Allgatherv
	Multi-PingPing	Alltoall
	Multi-Sendrecv	Reduce
	Multi-Exchange	Reduce_scatter
		Allreduce
		Barrier
		Multi-версии вышеперечисленных тестов

3.3.1.1 Тесты Отдельных Передач

В тестах этого класса отдельное сообщение пересылается между двумя процессами. В этих тестах отсутствует параллельная активность по передаче сообщений среди процессов. В данном случае следует ожидать лучших результатов по передаче сообщений.

В тесте PingPong осуществляется последовательная посылка сообщения от 0-го процесса 1-му процессу, и затем от 1-го процесса 0-му процессу. В тесте PingPing сообщения одновременно пересылаются между процессами. Тесты Отдельных Передач выполняются только с двумя активными процессами.

В случае PingPing величина пропускной способности относится к передаче отдельного сообщения. Пропускная способность в тесте PingPing в 2 раза меньше, чем в PingPong, что вызвано встречным трафиком сообщений в тесте PingPing.

3.3.1.2 Тесты Параллельных Передач

В тестах этого типа работа каждого процесса происходит параллельно с работой других процессов в сети. Тест измеряет эффективность передачи сообщений при глобальной загрузки сети.

При определении пропускной способности в расчет берется общее количество байт передаваемых и получаемых сообщений. Пропускная способность масштабируется

определенными коэффициентами. Времена передачи в тестах этого типа приводятся без масштабирования.

3.3.1.3 Коллективные тесты

Этот класс тестов содержит все коллективные тесты согласно MPI-стандарту. Здесь определяется качество реализации MPI.

Для простоты в этот класс включены Multi-версии этих тестов.

Отметим, что коллективные тесты (а именно Reduce, Allreduce, Reduce_scatter) не являются чисто тестами на передачу сообщений, но зависят от эффективной реализации определенных числовых операций.

3.3.2 Описание тестов Отдельных Передач

Эта часть описывает тесты Отдельных Передач подробнее. Каждый тест выполняется с сообщениями различной длины (от 0 до 4Мб в стандартном режиме). Времена передачи данных определенной длины усредняются по количеству повторений теста и по количеству процессов в группе. Здесь мы рассмотрим экземпляр теста со следующими исходными данными: x - определенная длина сообщения, основной MPI-тип данных для всех сообщений MPI_BYTE.

Пропускная способность определяется в Мбайт/сек= 2^{20} байт/сек по следующей формуле:

Пр.способность = $(x/2^{20}) * (time/10^6) = x/(time * 1.048576)$, где time в мксек.

3.3.2.1 Тест PingPong

Классический тест для измерения времени и пропускной способности передачи отдельного сообщения между 2 процессами. Измерения теста показаны на рис.2 между двумя стрелками.

МРІ операции, исп.в тесте	MPI_Send, MPI_Recv
Тип данных МРІ	MPI_BYTE
Отчет	time = $\Delta t/2$ (в мксек) пр.способность = $X/(1.048576 * \text{time})$

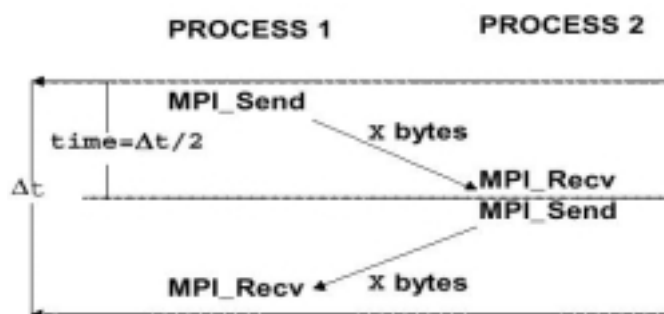


Рис.2 Обмен сообщениями между двумя процессами в тесте PingPong

3.3.2 Тест PingPing

Также как и PingPong тест PingPing измеряет времена передачи сообщений и пропускную способность, но при передаче встречных сообщений. Два процесса обмениваются друг с другом сообщениями (MPI_Isend/MPI_Recv/ MPI_Wait), с одновременной выдачей MPI_Isend.

МРІ операции, исп.в тесте	MPI_Isend/MPI_Wait, MPI_Recv
Тип данных МРІ	MPI_BYTE
Отчет	time = Δt (в мксек) пр.способность = $X / (1.048576 * \text{time})$

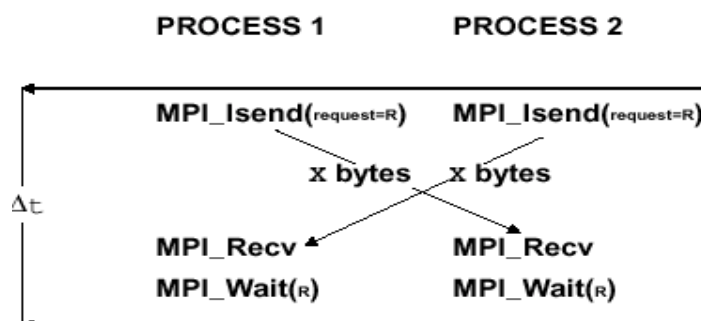


Рис.3 Обмен сообщениями между двумя процессами в тесте PingPing

3.3.3 Описание тестов Параллельных Передач

Эта часть объясняет подробнее тесты Параллельных Передач. Каждый тест выполняет передачи сообщений различной длины x байт (от 0 до 4Мб в стандартном режиме). Времена передачи данных определенной длины усредняются по количеству повторений прогонов теста и по количеству процессов в группе.

Основной тип данных для всех сообщений MPI_BYTE.

При вычислении пропускной способности берется в расчет общее число $nmsg$ входящих в и исходящих сообщений из конкретного процесса. В тесте Sendrecv конкретный процесс посылает и получает x байт, $nmsg = 2$. В тесте Exchange $nmsg = 4$. $Nmsg$ – является масштабирующим коэффициентом при расчете пропускной способности.

Значения пропускной способности вычисляются в Мбайт/сек= 2^{20} байт/сек по следующей формуле:

Пр.способность = $nmsg * (x/2^{20}) / (time/ 10^6) = nmsg * x / (time * 1.048576)$, где time в мксек.

3.3.3.1 Тест Sendrecv

Процессы формируют периодическую цепь связи. Каждый процесс посылает соседу справа и получает данные от соседа слева в цепи. Два сообщения приходятся на один процесс (одно приходит, другое выходит).

Тест Sendrecv измеряет времена передачи сообщений и пропускную способность.

MPI операции, исп.в тесте	MPI_Sendrecv
Тип данных MPI	MPI_BYTE
Отчет	time = Δt (в мксек) пр.способность = $2X/(1.048576 * \text{time})$

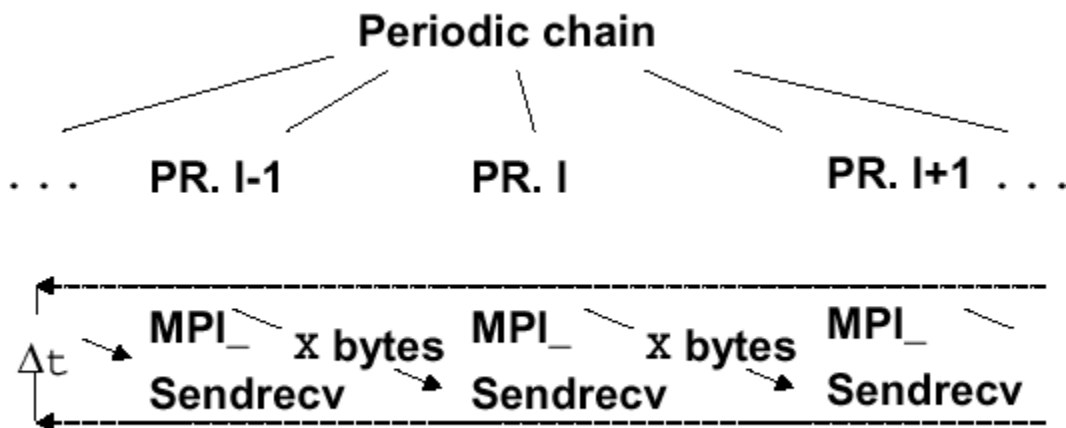


Рис 4. Обмен сообщениями между процессами в тесте Sendrecv.

3.3.3.2 Тест Exchange

Группа процессов составляет периодическую цепь, каждый процесс посылает данные соседу справа и соседу слева, получает данные от соседа слева и от соседа справа.

Коэффициент $nmsg=4$ (в каждом экземпляре теста каждого процесса: 2 сообщения входят, 2 сообщения исходят).

Тест Exchange измеряет времена передачи сообщений и пропускную способность.

MPI операции, исп.в тесте	MPI_Isend/ MPI_Waitall, MPI_Recv
Тип данных MPI	MPI_BYTE
Отчет	time = Δt (в мксек) пр.способность = $4X/(1.048576 * time)$

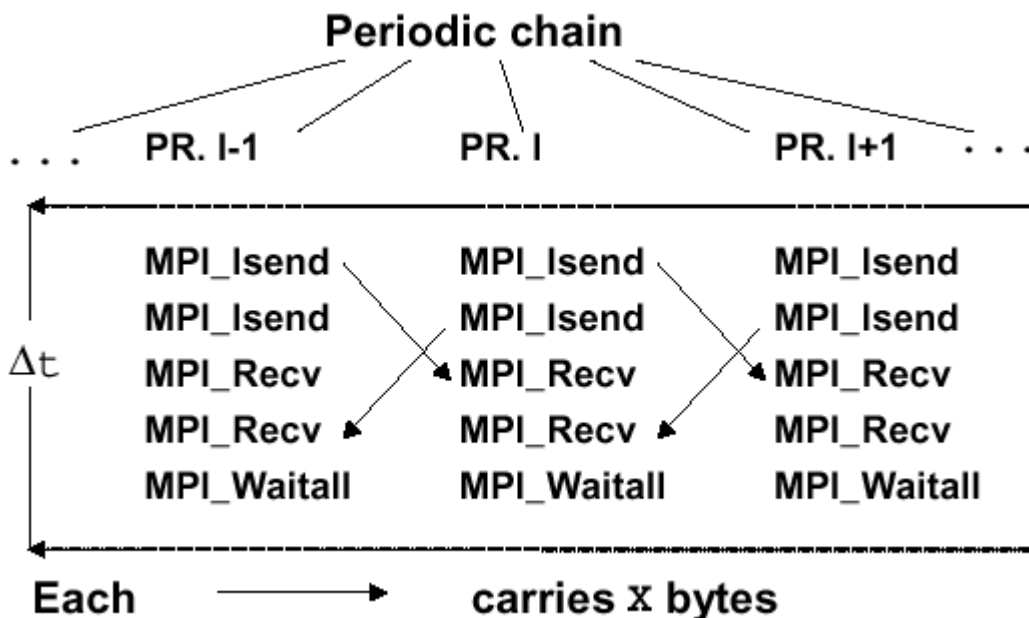


Рис.5. Обмен сообщениями между процессами в тесте Exchange.

3.3.4 Описание Коллективных тестов

Эта часть объясняет подробнее Коллективные тесты. Каждый тест выполняет передачи сообщений различной длины x байт (от 0 до 4Мб в стандартном режиме). Времена передачи данных определенной длины усредняются по количеству повторений прогонов теста и по количеству процессов в группе.

Основной тип данных для всех сообщений MPI_BYTE для тестов только передачи сообщений и MPI_FLOAT для вычислительных тестов.

3.3.4.1 Тест Reduce

Тест функции MPI_Reduce. Производит вычисления над вектором данных длины $L = X/\text{sizeof}(\text{float})$ из элементов типа float Root-процесс меняется циклически.

MPI операции, исп.в тесте	MPI_Reduce (op=MPI_SUM)
Тип данных MPI	MPI_FLOAT
Отчет	time = Δt (в мксек)

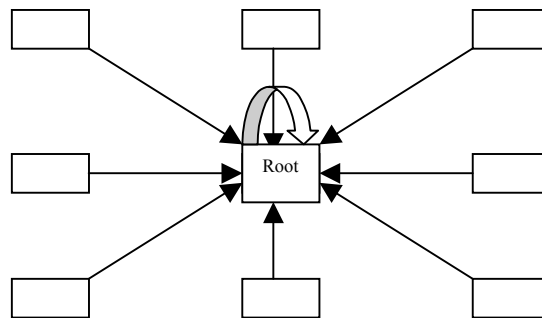


Рис.6 Сбор и обработка данных root-процессом.

3.3.4.2 Тест Reduce_scatter

Тест функции MPI_Reduce_scatter. Производит вычисления над вектором данных длины $L = X/\text{sizeof}(\text{float})$ из элементов типа float.

На стадии scatter, L элементов по возможности равномерно распределяются между активными процессами. А именно,

$L = r*nr+s$ ($s = L \bmod nr$), где nr - число процессов.

Процесс номер i получает r+1 элементов, когда $i < s$, и r элементов, когда $i \geq s$.

МРІ операции, исп.в тесте	MPI_Reduce_scatter (op=MPI_SUM)
Тип данных МРІ	MPI_FLOAT
Отчет	time = Δt (в мксек)

3.3.4.3 Тест Allreduce

Тест функции MPI_Allreduce. Производит вычисления над вектором данных длины $L = X/\text{sizeof}(\text{float})$ из элементов типа float.

МРІ операции, исп.в тесте	MPI_Allreduce (op=MPI_SUM)
Тип данных МРІ	MPI_FLOAT
Отчет	time = Δt (в мксек)

3.3.4.4 Тест Allgather

Тест функции MPI_Allgather. Каждый процесс посылает X байт и получает собранные $X*(\text{число процессов})$ байт.

МРІ операции, исп.в тесте	MPI_Allgather
Тип данных МРІ	MPI_BYTE
Отчет	time = Δt (в мксек)

3.3.4.5 Тест Allgatherv

Тест функции MPI_Allgatherv. Каждый процесс посылает X байт и получает собранные $X*(\text{число процессов})$ байт. Функционально подобен Allgather, однако с функцией MPI_Allgatherv.

МРІ операции, исп.в тесте	MPI_Allgatherv
Тип данных МРІ	MPI_BYTE
Отчет	time = Δt (в мксек)

3.3.4.6 Тест Alltoall

Тест функции MPI_Alltoall. Каждый процесс посылает X *(число процессов) байт (X байт на каждый процесс) и получает X *(число процессов) байт (X байт от каждого процесса).

MPI операции, исп.в тесте	MPI_Alltoall
Тип данных MPI	MPI_BYTE
Отчет	time = Δt (в мксек)

3.3.4.7 Тест Bcast

Тест функции MPI_Bcast. Root-процесс передает X байт всем процессам. Root-процесс изменяется циклически.

MPI операции, исп.в тесте	MPI_Bcast
Тип данных MPI	MPI_BYTE
Отчет	time = Δt (в мксек)

3.3.4.8 Тест Barrier

Обмен данными не производится.

MPI операции, исп.в тесте	MPI_Barrier
Тип данных MPI	-
Отчет	time = Δt (в мксек)

4. Результаты выполнения тестов PMB 2.2 на Linux-кластере ЭКСС-1.

Таблица 39 . Тест Alltoall(N), кластер ЭКСС-1 (2активных из 16 процессов).

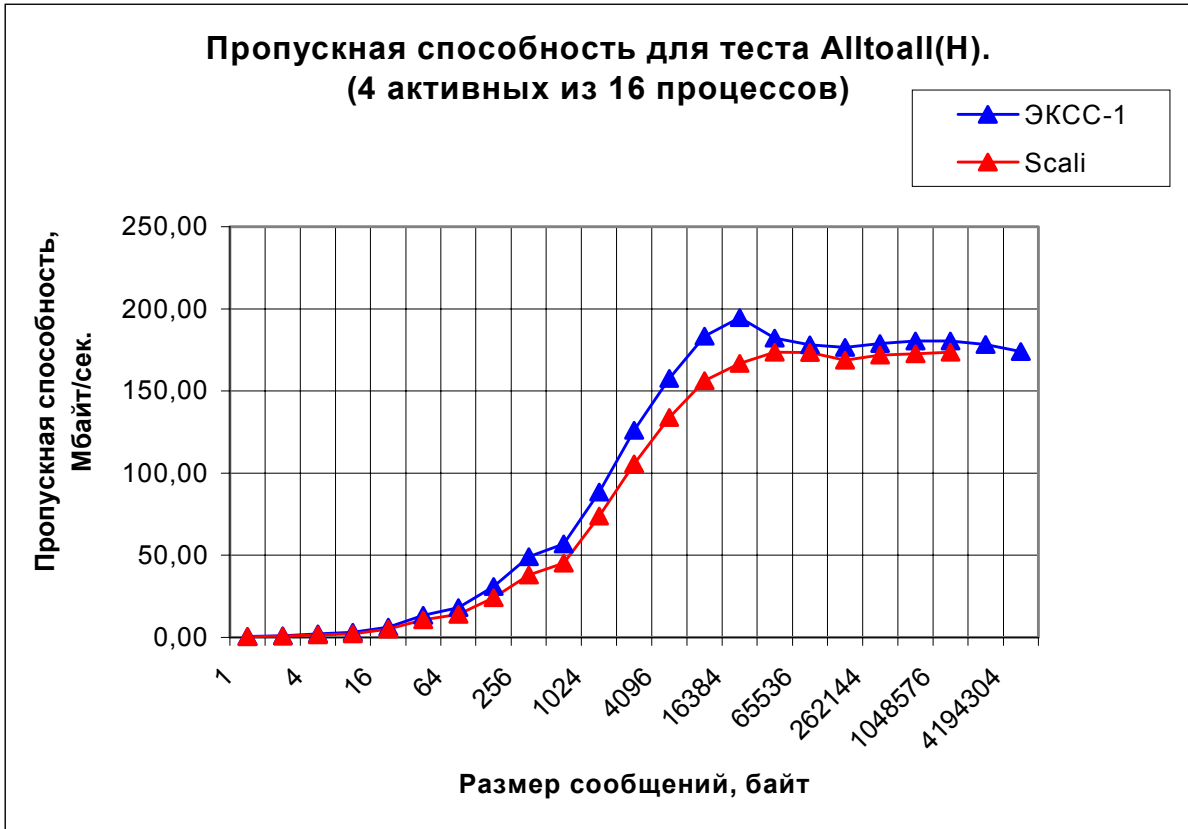
Длина сообщений, байт	Число повторений	ЭКСС-1		Scali	
		Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.	Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.
0	1000	8,16	0,00	8,16	0
1	1000	10,33	0,37	14,85	0,26
2	1000	10,36	0,74	14,57	0,52
4	1000	10,18	1,50	14,57	1,05
8	1000	13,51	2,26	17,52	1,74
16	1000	13,37	4,57	17,51	3,49
32	1000	12,49	9,77	16,93	7,21
64	1000	18,52	13,18	24,99	9,77
128	1000	21,63	22,57	28,74	16,99
256	1000	26,90	36,30	34,73	28,12
512	1000	37,22	52,48	47,21	41,37
1024	1000	51,67	75,60	68,44	57,08
2048	1000	74,69	104,60	95,39	81,90
4096	1000	126,13	123,88	156,41	99,90
8192	1000	223,19	140,02	274,25	113,95
16384	1000	417,41	149,73	505,22	123,71
32768	1000	820,13	152,41	976,62	127,99
65536	640	1923,14	130,00	1931,46	129,44
131072	320	4173,68	119,80	4249,74	117,65
262144	160	8666,06	115,39	8771,45	114,01
524288	80	17474,18	114,45	17449,74	114,61
1048576	40	34790,14	114,98	35041,03	114,15
2097152	20	68234,59	117,24		
4194304	10	135701,74	117,91		



Результаты выполнения теста Alltoall(H) на кластере ЭКСС-1 и кластере Scali в случае 2 активных из 16 процессов.

Таблица 40 . Тест Alltoall(N), кластер ЭКСС-1 (4 активных из 16 процессов).

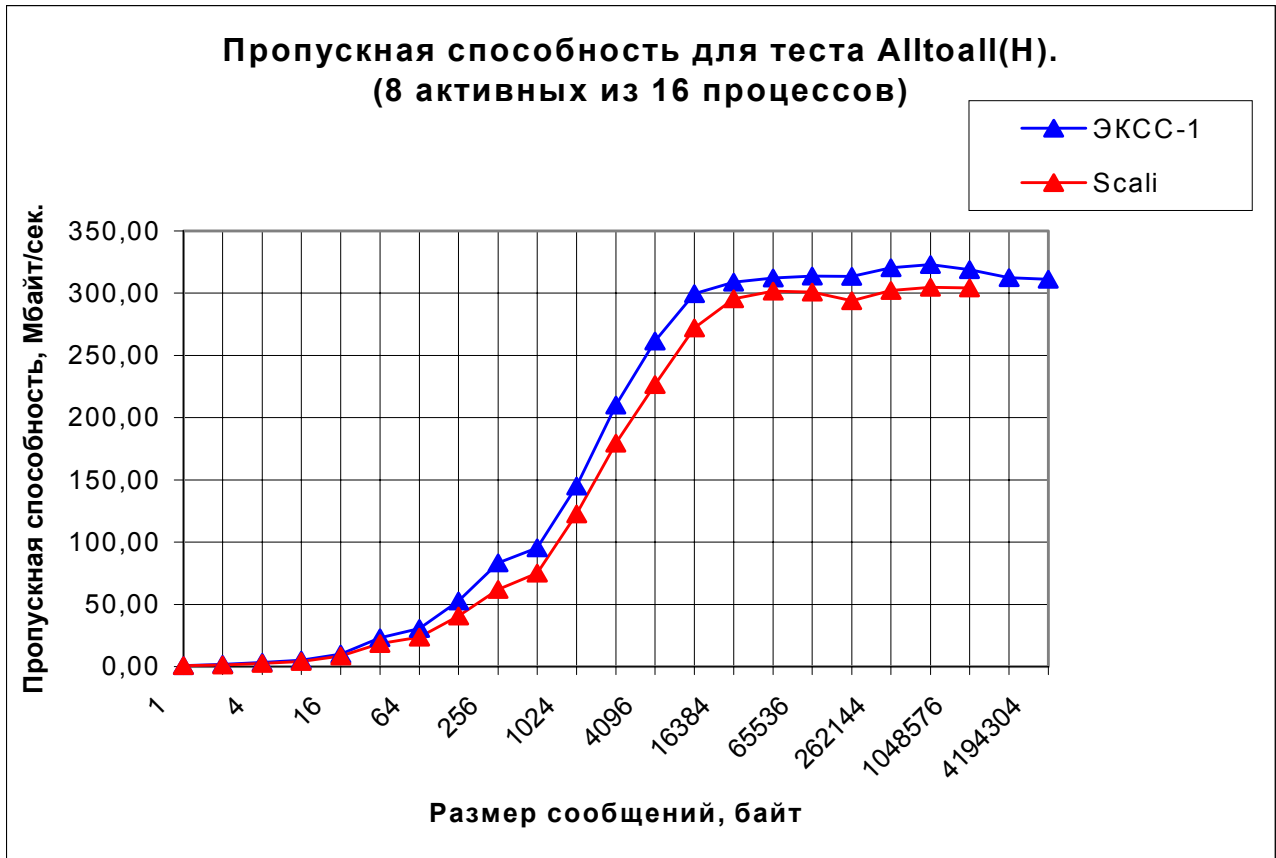
Длина сообщений, байт	Число повторений	ЭКСС-1		Scali	
		Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.	Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.
0	1000	16,38	0	20,58	0
1	1000	29,77	0,51	39,84	0,38
2	1000	29,61	1,03	39,51	0,77
4	1000	29,56	2,06	39,44	1,55
8	1000	39,05	3,13	56,12	2,18
16	1000	38,99	6,26	48,43	5,04
32	1000	36,06	13,54	45,54	10,72
64	1000	54,02	18,08	69,56	14,04
128	1000	62,91	31,05	80,81	24,17
256	1000	79,53	49,12	102,98	37,93
512	1000	137,37	56,87	173,16	45,12
1024	1000	176,69	88,43	211,62	73,84
2048	1000	247,79	126,11	295,63	105,71
4096	1000	396,46	157,65	467,07	133,81
8192	1000	681,46	183,43	799,78	156,29
16384	1000	1284,85	194,58	1499,23	166,75
32768	1000	2745,4	182,12	2880,94	173,55
65536	640	5613,82	178,13	5768,86	173,34
131072	320	11331,09	176,51	11848,5	168,80
262144	160	22370,33	178,81	23281,66	171,81
524288	80	44336,8	180,44	46361,39	172,56
1048576	40	88697,28	180,39	92185,92	173,56
2097152	20	179421,52	178,35		
4194304	10	367740,88	174,04		



Результаты выполнения теста Alltoall(H) на кластере ЭКСС-1 и кластере Scali в случае 4 активных из 16 процессов.

Таблица 41 . Тест Alltoall(N), кластер ЭКСС-1 (8 активных из 16 процессов).

Длина сообщений, байт	Число повторений	ЭКСС-1		Scali	
		Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.	Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.
0	1000	26,97	0	35,87	0
1	1000	70,74	0,86	91,19	0,67
2	1000	70,05	1,74	90,31	1,35
4	1000	70,35	3,47	90,5	2,70
8	1000	95,88	5,09	115,62	4,22
16	1000	98,27	9,94	114,96	8,49
32	1000	84,7	23,06	105,24	18,56
64	1000	127,41	30,66	164,53	23,74
128	1000	148,53	52,60	192,22	40,64
256	1000	187,49	83,34	252,37	61,91
512	1000	327,59	95,39	416,47	75,04
1024	1000	430,91	145,04	510	122,55
2048	1000	594,94	210,11	696	179,60
4096	1000	955,69	261,59	1103,48	226,56
8192	1000	1668,74	299,63	1837,88	272,05
16384	1000	3239,1	308,73	3384,31	295,48
32768	1000	6406,24	312,20	6632,18	301,56
65536	640	12752,33	313,67	13295,7	300,85
131072	320	25515,93	313,53	27220,52	293,90
262144	160	49929,28	320,45	52956,4	302,14
524288	80	99100,91	322,90	104975,37	304,83
1048576	40	200756,27	318,79	210404,35	304,18
2097152	20	409730,32	312,40		
4194304	10	822787,57	311,14		



Результаты выполнения теста Alltoall(H) на кластере ЭКСС-1 и кластере Scali в случае 8 активных из 16 процессов.

Таблица 42 . Тест Alltoall(N), кластер ЭКСС-1 (16 активных из 16 процессов).

Длина сообщений, байт	Число повторений	ЭКСС-1		Scali	
		Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.	Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.
0	1000	35,69	0	42,81	0
1	1000	154,83	1,58	199,36	1,22
2	1000	153,19	3,19	198,09	2,46
4	1000	154,01	6,34	198,69	4,92
8	1000	212,48	9,19	266,63	7,33
16	1000	209,79	18,62	254,64	15,34
32	1000	184,88	42,26	229,73	34,01
64	1000	286,32	54,57	359,84	43,42
128	1000	324,88	96,19	424,97	73,53
256	1000	408,08	153,16	527,95	118,38
512	1000	728,69	171,54	922,05	135,57
1024	1000	940,19	265,90	1142,8	218,76
2048	1000	1313,22	380,74	1534,38	325,86
4096	1000	2137,19	467,90	2393,17	417,86
8192	1000	3765,92	531,08	4105,11	487,20
16384	1000	7151,79	559,30	7573,83	528,13
32768	1000	13744,7	582,04	14551,31	549,78
65536	640	27063,93	591,19	28292,96	565,51
131072	320	54644,36	585,60	58382,79	548,11
262144	160	106689,45	599,87	113090,97	565,92
524288	80	209215,44	611,81	222706,2	574,75
1048576	40	428433,88	597,53	446399,66	573,48
2097152	20	881491,39	580,83		
4194304	10	1758936,34	582,17		



Результаты выполнения теста Alltoall(F) на кластере ЭКСС-1 и кластере Scali в случае 16 активных из 16 процессов.

Таблица 43 . Тест Alltoall(F), кластер ЭКСС-1 (2активных из 32 процессов).

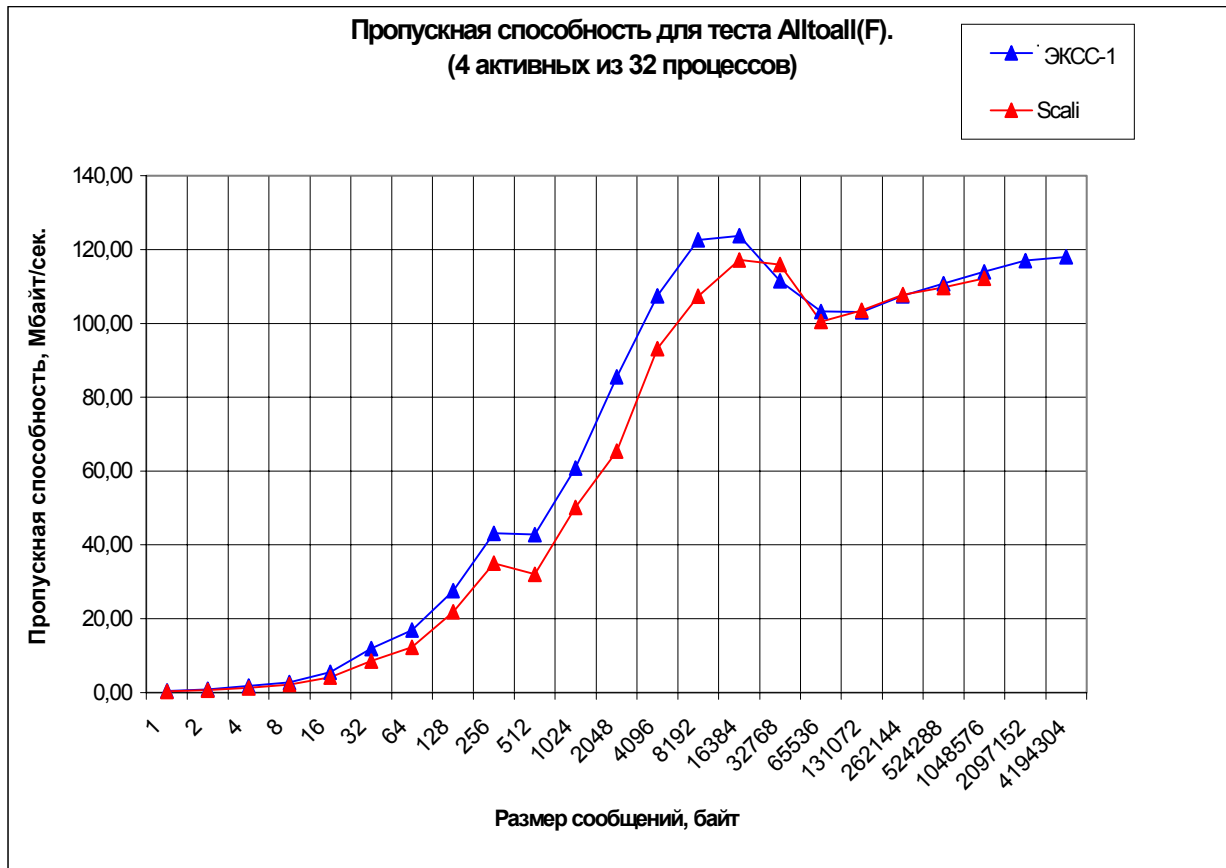
Длина сообщений, байт	Число повторений	ЭКСС-1		Scali	
		Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.	Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.
0	1000	3,14	0,00	4,96	0,00
1	1000	4,73	0,81	7,61	0,50
2	1000	4,74	1,61	7,71	0,99
4	1000	4,73	3,23	7,49	2,04
8	1000	5,11	5,97	8,2	3,72
16	1000	5,12	11,92	8,22	7,43
32	1000	5,19	23,52	8,45	14,45
64	1000	5,62	43,44	8,62	28,32
128	1000	6,56	74,43	9,77	49,98
256	1000	8,59	113,69	12,21	79,98
512	1000	12,38	157,76	15,65	124,80
1024	1000	18,4	212,30	24,15	161,75
2048	1000	29,78	262,34	37,03	210,98
4096	1000	54,32	287,65	73,61	212,27
8192	1000	103,67	301,44	145,88	214,22
16384	1000	199,9	312,66	269,45	231,95
32768	1000	442,75	282,33	578,01	216,26
65536	640	1723,62	145,04	1229,09	203,40
131072	320	4261,96	117,32	3903,56	128,09
262144	160	9678,24	103,32	9039,62	110,62
524288	80	20092,4	99,54	18487,11	108,18
1048576	40	40903,93	97,79	37256,26	107,36
2097152	20	81466,55	98,20		
4194304	10	157498,23	101,59		



Результаты выполнения теста Alltoall(F) на кластере ЭКСС-1 и кластере Scali в случае 2 активных из 32 процессов.

Таблица 44 . Тест Alltoall(F), кластер ЭКСС-1 (4 активных из 32 процессов).

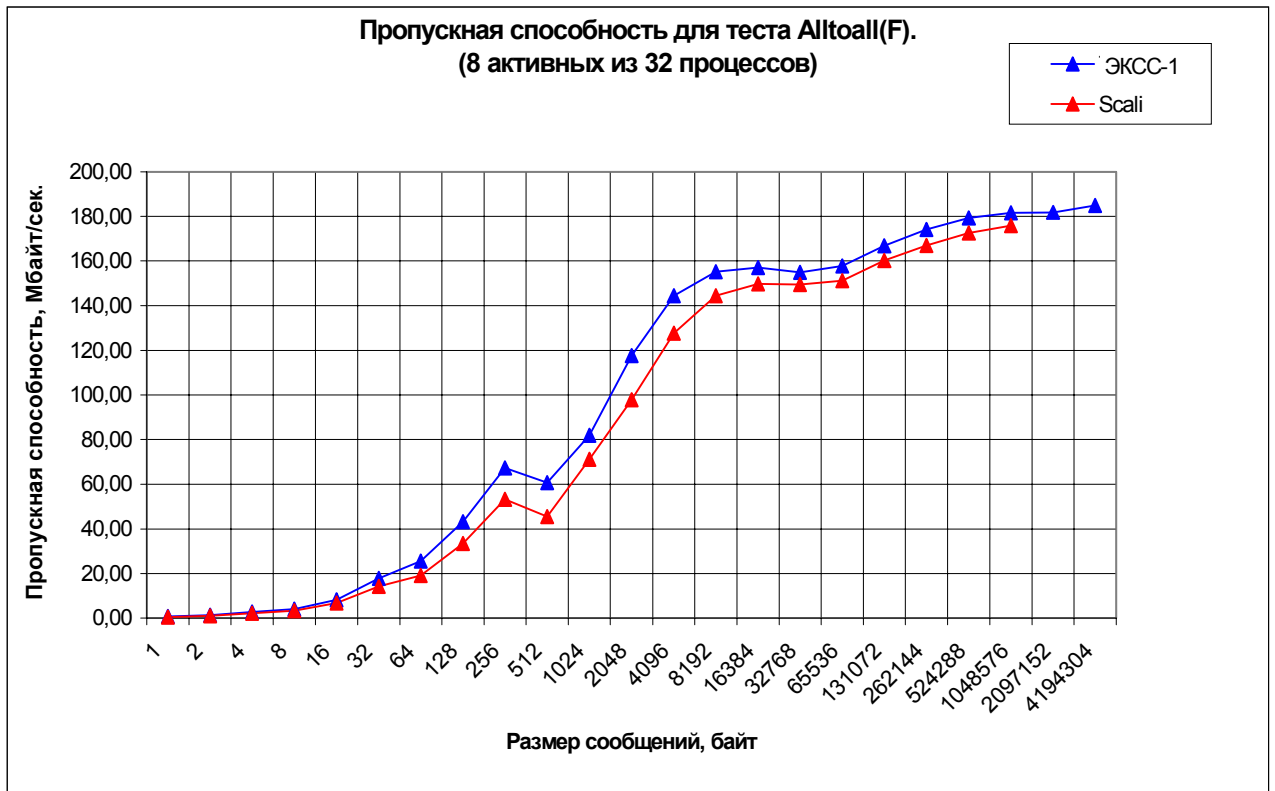
Длина сообщений, байт	Число повторений	ЭКСС-1		Scali	
		Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.	Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.
0	1000	12,79	0,00	17,45	0,00
1	1000	34,69	0,44	46,75	0,33
2	1000	35,1	0,87	46,87	0,65
4	1000	34,2	1,78	49,21	1,24
8	1000	44,35	2,75	56,89	2,15
16	1000	44,22	5,52	59	4,14
32	1000	40,86	11,95	56,77	8,60
64	1000	57,66	16,94	79,46	12,29
128	1000	70,73	27,61	89,54	21,81
256	1000	90,51	43,16	111,45	35,05
512	1000	182,49	42,81	243,74	32,05
1024	1000	256,78	60,85	311,67	50,13
2048	1000	365,41	85,52	478,26	65,34
4096	1000	581,47	107,49	670,58	93,20
8192	1000	1019,3	122,63	1164,33	107,36
16384	1000	2020,83	123,71	2132,96	117,21
32768	1000	4483,86	111,51	4312,38	115,95
65536	640	9687,94	103,22	9952,54	100,48
131072	320	19391,83	103,14	19315,72	103,54
262144	160	37210,02	107,50	37100,81	107,81
524288	80	72210,13	110,79	72915,22	109,72
1048576	40	140361,32	113,99	142624,02	112,18
2097152	20	273370,7	117,06		
4194304	10	542182,42	118,04		



Результаты выполнения теста Alltoall(F) на кластере ЭКСС-1 и кластере Scali в случае 4 активных из 32 процессов.

Таблица 45 . Тест Alltoall(F), кластер ЭКСС-1 (8 активных из 32 процессов).

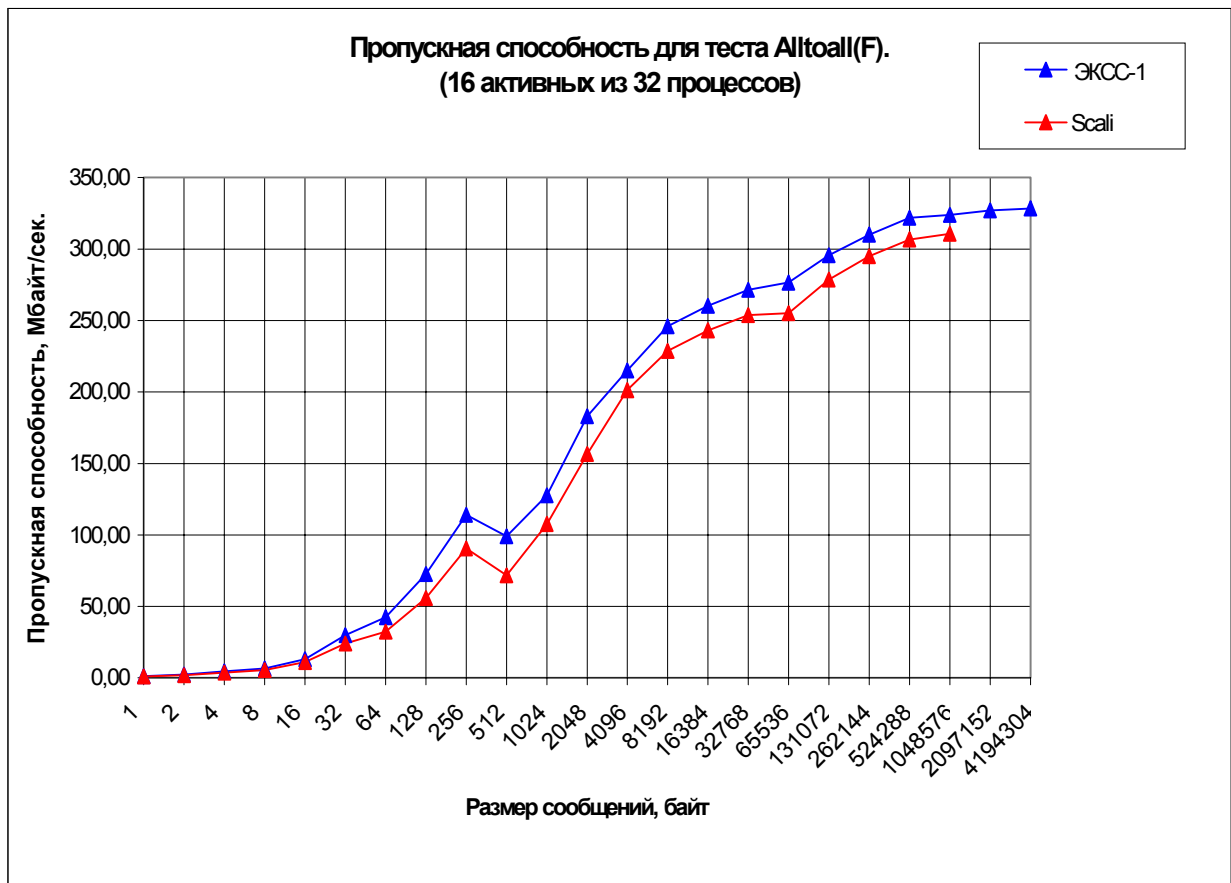
Длина сообщений, байт	Число повторений	ЭКСС-1		Scali	
		Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.	Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.
0	1000	21,19	0,00	27,89	0,00
1	1000	90,46	0,67	115,91	0,53
2	1000	88,77	1,38	118,21	1,03
4	1000	88,09	2,77	112,31	2,17
8	1000	119,7	4,08	145,91	3,35
16	1000	120,15	8,13	145,51	6,71
32	1000	109,42	17,85	138,06	14,15
64	1000	152,52	25,61	204,01	19,15
128	1000	181,07	43,15	233,15	33,51
256	1000	232,42	67,23	293,94	53,16
512	1000	515,13	60,66	687,86	45,43
1024	1000	762,98	81,92	879,1	71,10
2048	1000	1061,59	117,75	1277,22	97,87
4096	1000	1730,83	144,44	1957,92	127,69
8192	1000	3220,29	155,27	3461,28	144,46
16384	1000	6369,06	157,01	6674,03	149,83
32768	1000	12905,5	154,97	13378,25	149,50
65536	640	25358,51	157,74	26439,87	151,29
131072	320	47964,19	166,79	49898,59	160,33
262144	160	91915,46	174,07	95803,15	167,01
524288	80	178454,12	179,32	185503,89	172,50
1048576	40	352437,98	181,59	364112,46	175,77
2097152	20	704094,74	181,79		
4194304	10	1383780,04	185,00		



Результаты выполнения теста Alltoall(F) на кластере ЭКСС-1 и кластере Scali в случае 8 активных из 32 процессов.

Таблица 46 . Тест Alltoall(F), кластер ЭКСС-1 (16 активных из 32 процессов).

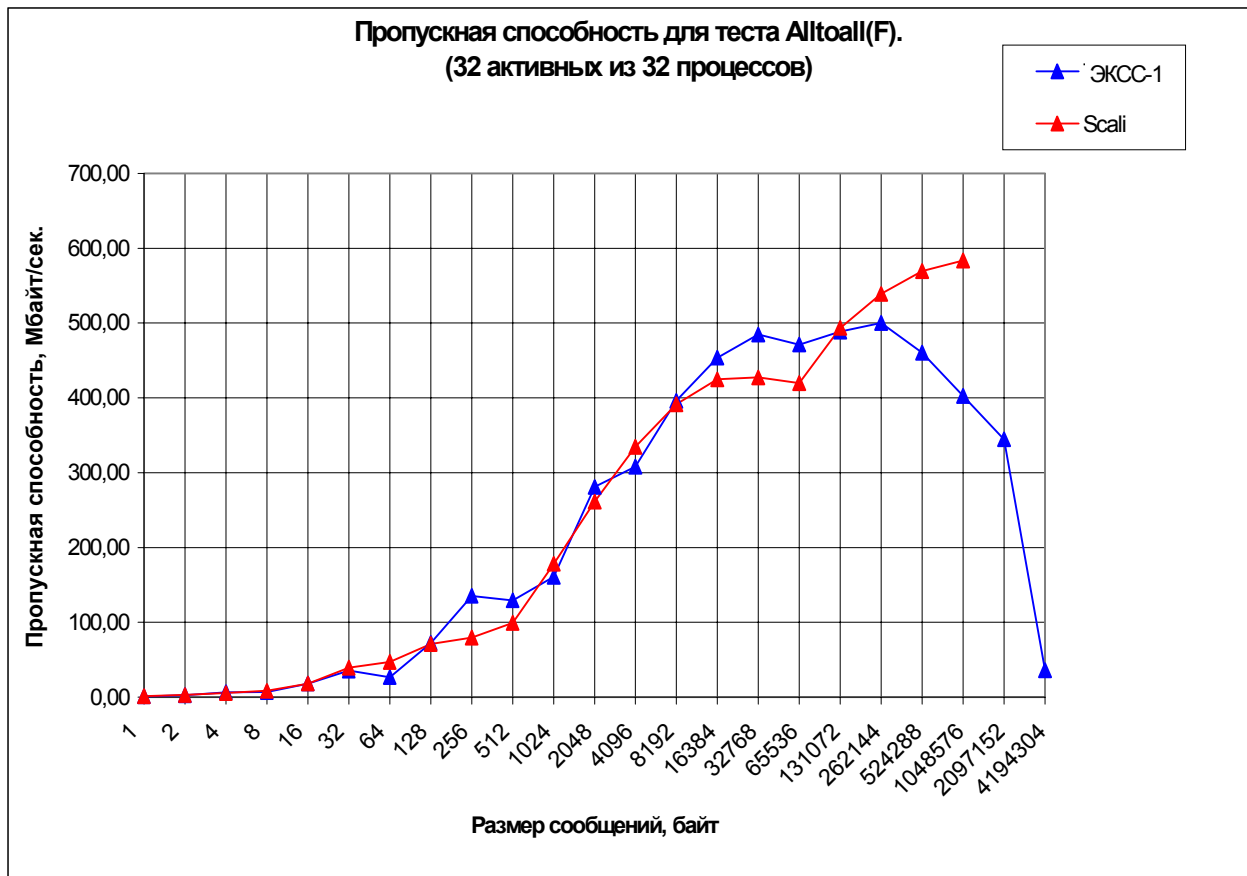
Длина сообщений, байт	Число повторений	ЭКСС-1		Scali	
		Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.	Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.
0	1000	31,12	0,00	43,57	0,00
1	1000	215,48	1,13	272,24	0,90
2	1000	213,88	2,28	273,46	1,79
4	1000	212,48	4,60	272,9	3,58
8	1000	299,35	6,52	359,7	5,43
16	1000	299,22	13,05	354,6	11,02
32	1000	262,18	29,80	324,76	24,06
64	1000	368,67	42,38	486,3	32,13
128	1000	429,56	72,75	558,91	55,91
256	1000	548,36	113,98	691,64	90,36
512	1000	1264,33	98,87	1742,39	71,74
1024	1000	1958,78	127,63	2320,97	107,71
2048	1000	2731,56	183,05	3190,4	156,72
4096	1000	4648,92	215,10	4967,9	201,29
8192	1000	8131,34	245,96	8743,93	228,73
16384	1000	15363,27	260,36	16449,12	243,17
32768	1000	29483,13	271,34	31512,21	253,87
65536	640	57860,09	276,53	62700,4	255,18
131072	320	108242,18	295,63	114799,98	278,75
262144	160	206328,91	310,18	217014,43	294,91
524288	80	397453,55	322,05	417187,09	306,82
1048576	40	790217,6	323,96	824218,84	310,60
2097152	20	1565697,41	327,01		
4194304	10	3117318,82	328,49		



Результаты выполнения теста Alltoall(F) на кластере ЭКСС-1 и кластере Scali в случае 16 активных из 32 процессов.

Таблица 47 . Тест Alltoall(F), кластер ЭКСС-1 (32 активных из 32 процессов).

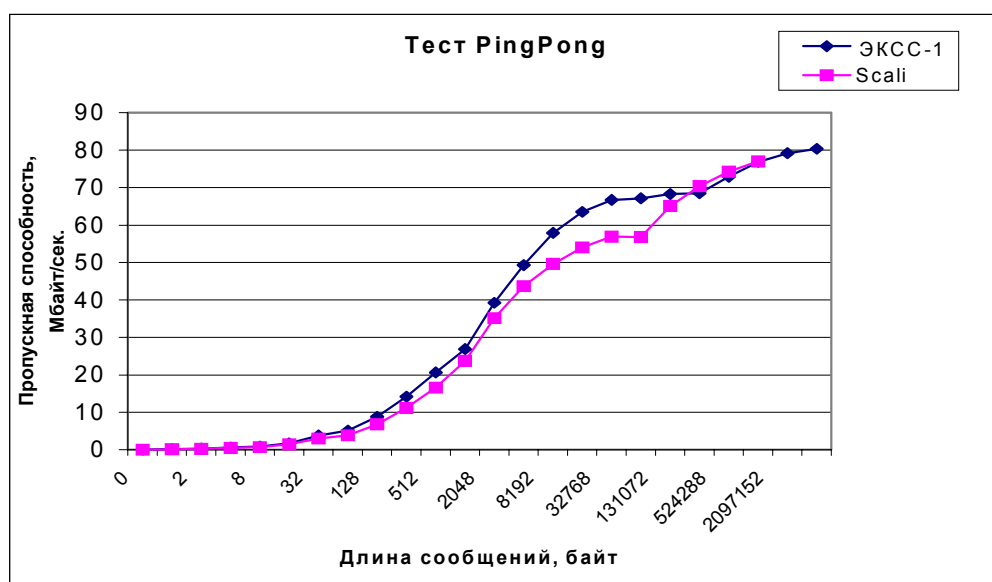
Длина сообщений, байт	Число повторений	ЭКСС-1		Scali	
		Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.	Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.
0	1000	49,55	0,00	60,14	0,00
1	1000	973,35	1,00	648,17	1,51
2	1000	701,34	2,78	657,53	2,97
4	1000	575,74	6,78	677,35	5,77
8	1000	1112,37	7,02	915,11	8,54
16	1000	851,03	18,36	861,38	18,14
32	1000	877,94	35,59	789,08	39,60
64	1000	2325,28	26,88	1325,5	47,15
128	1000	1725,18	72,46	1760,84	70,99
256	1000	1845,02	135,50	3132,57	79,81
512	1000	3856,79	129,64	5025,35	99,50
1024	1000	6225,9	160,62	5604,66	178,42
2048	1000	7113,78	281,14	7650,79	261,41
4096	1000	12975,19	308,28	11951,69	334,68
8192	1000	20196,93	396,10	20431,46	391,55
16384	1000	35281,75	453,49	37676,07	424,67
32768	1000	66029,81	484,63	74919,26	427,13
65536	640	135859,84	471,07	152415,8	419,90
131072	320	261995,81	488,56	259314,44	493,61
262144	160	511968,63	500,03	475230,88	538,69
524288	80	1112630,21	460,17	899186,83	569,40
1048576	40	2543428,4	402,61	1754529,51	583,63
2097152	20	5941833,89	344,67		
4194304	10	113418523,69	36,11		



Результаты выполнения теста Alltoall(F) на кластере ЭКСС-1 и кластере Scali в случае 32 активных из 32 процессов.

Таблица 48 . Тест PingPong(H), кластер ЭКСС-1.

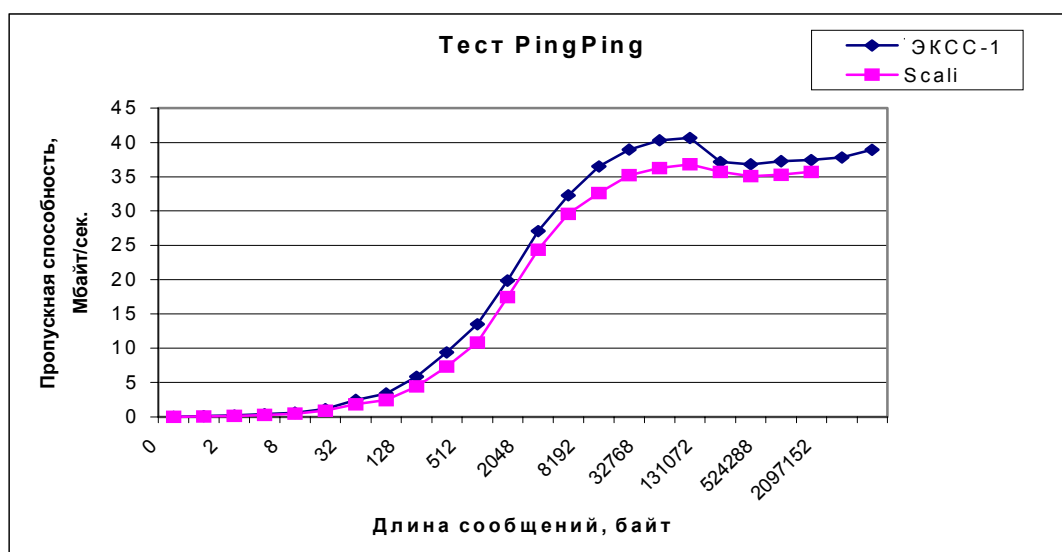
Длина сообщений, байт	Число повторений теста	ЭКСС-1		Scali	
		Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.	Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.
0	1000	5,21	0	6,47	0
1	1000	6,12	0,16	7,92	0,12
2	1000	6,14	0,31	7,92	0,24
4	1000	6,17	0,62	7,93	0,48
8	1000	8,92	0,86	10,86	0,7
16	1000	8,95	1,71	10,82	1,41
32	1000	8,07	3,78	10,23	2,98
64	1000	11,93	5,11	15,74	3,88
128	1000	13,89	8,79	17,93	6,81
256	1000	17,23	14,17	21,83	11,18
512	1000	23,63	20,66	29,41	16,6
1024	1000	36,33	26,88	41,21	23,7
2048	1000	49,74	39,27	55,55	35,16
4096	1000	79,2	49,32	89,39	43,7
8192	1000	134,86	57,93	157,51	49,6
16384	1000	245,91	63,54	288,95	54,08
32768	1000	468,58	66,69	549,08	56,91
65536	640	930,79	67,15	1100,6	56,79
131072	320	1831,45	68,25	1920,45	65,09
262144	160	3648,9	68,51	3552,37	70,38
524288	80	6859,99	72,89	6733,08	74,26
1048576	40	13005,08	76,89	12988,77	76,99
2097152	20	25250,22	79,21		
4194304	10	49761,35	80,38		



Результаты выполнения теста PingPong на кластере ЭКСС-1 и кластере Scali.

Таблица 49 . Тест PingPing(H), кластер ЭКСС-1.

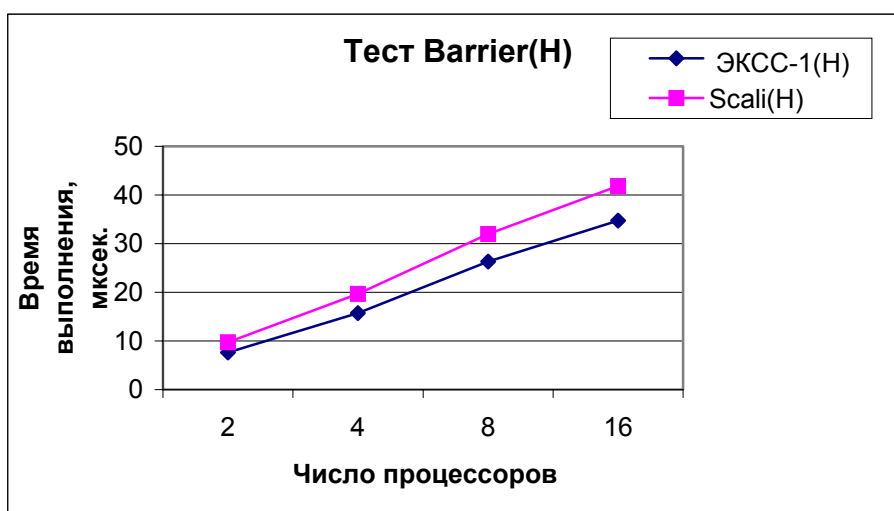
Длина сообщений, байт	Число повторений теста	ЭКСС-1		Scali	
		Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.	Время передачи сообщений, мксек	Пропускная способность, Мбайт/сек.
0	1000	8,87	0	11,76	0
1	1000	10,24	0,09	14,17	0,07
2	1000	10,13	0,19	14,16	0,13
4	1000	10,16	0,38	14,36	0,27
8	1000	13,34	0,57	16,81	0,45
16	1000	13,3	1,15	16,95	0,9
32	1000	12,42	2,46	16,7	1,83
64	1000	18	3,39	25	2,44
128	1000	20,97	5,82	27,66	4,41
256	1000	25,98	9,4	33,36	7,32
512	1000	36,14	13,51	44,96	10,86
1024	1000	49,17	19,86	56,02	17,43
2048	1000	72,17	27,06	80,14	24,37
4096	1000	120,98	32,29	132,07	29,58
8192	1000	214,09	36,49	239,32	32,64
16384	1000	400,99	38,97	443,57	35,23
32768	1000	774,97	40,32	861,79	36,26
65536	640	1538,07	40,64	1697,32	36,82
131072	320	3364,88	37,15	3499,56	35,72
262144	160	6787,5	36,83	7126,98	35,08
524288	80	13423,68	37,25	14168,3	35,29
1048576	40	26700,25	37,45	27997,88	35,72
2097152	20	52909,57	37,8		
4194304	10	102810,03	38,91		



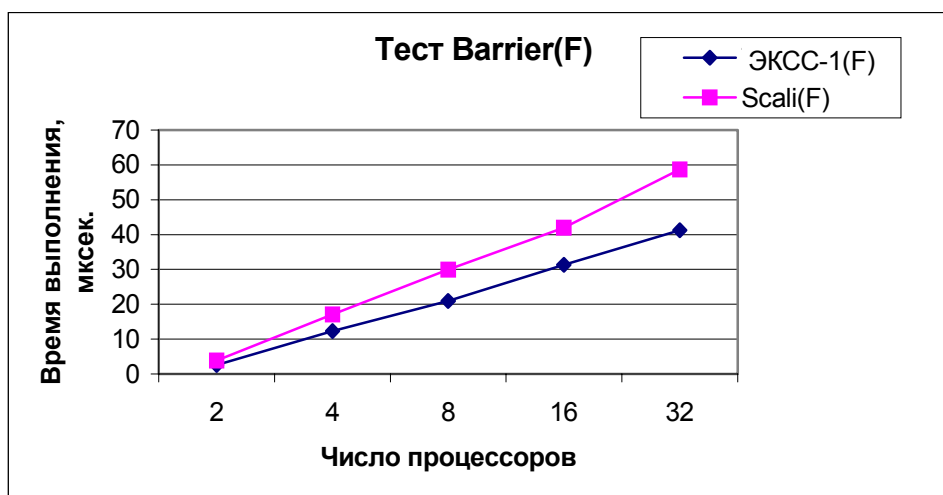
Результаты выполнения теста PingPing на кластере ЭКСС-1 и кластере Scali.

Таблица 50 . Тест Barrier, кластер ЭКСС-1, число повторений=1000000.

Число активных процессов	Время выполнения, мксек	Время выполнения, мксек	Время выполнения, мксек	Время выполнения, мксек
	ЭКСС-1(H)	Scali(H)	ЭКСС-1(F)	Scali(F)
2	7,62	9,71	2,55	3,83
4	15,72	19,69	12,31	17,08
8	26,3	31,95	20,9	29,99
16	34,72	41,81	31,37	42,01
32			41,27	58,71



Среднее время выполнение теста MPI_Barrier(H) на ЭКСС-1 и на кластере Scali.



Среднее время выполнение теста MPI_Barrier(F) на ЭКСС-1 и на кластере Scali.

Заключение

В данной работе на основе экспериментально полученных данных показана перспективность выбранного направления использования высокоскоростных сетевых технологий SCI при построении высокопроизводительных вычислительных кластеров.

Литература

- [1] *ORNL Compaq Alpha/ IBM SP evaluation*. www.epm.ornl.gov/~dunigan/alpha.
- [2] www.pallas.de
- [3] Message passing interface forum. *MPI: A Message-Passing Interface Standard*, June 1995. Version 1.1.