Результаты оценочного тестирования отечественной высокоскоростной коммуникационной сети Ангара

А.А. Агарков, Т.Ф. Исмагилов, Д.В. Макагон, А.С. Семенов, А.С. Симонов
АО «НИЦЭВТ»

В статье представлены результаты сравнительного оценочного тестирования 36-узлового вычислительного кластера «Ангара-К1», оснащенного адаптерами коммуникационной сети Ангара, и суперкомпьютера МВС-10П с сетью InfiniBand 4x FDR, установленного в МСЦ РАН.

Ключевые слова: высокоскоростная сеть, интерконнект, Ангара, InfiniBand FDR, HPCG, HPL, NPB, ПЛАВ

1. Введение

По статистике списка TOP500 самых мощных суперкомпьютеров мира ([1], ноябрь 2015) можно заметить, что большинство представленных в нем систем используют коммерческие сети InfiniBand и Gigabit Ethernet. Однако суперкомпьютеры из первой десятки — Tianhe-2 (Китай), Cray (США), IBM Blue Gene/Q (США), K Computer (Япония) — используют собственные («заказные») коммуникационные сети, и только одна система использует коммерческую сеть InfiniBand. «Заказные» сети не поставляются отдельно от вычислительной системы, а коммерческие сети далеко не всегда подходят для эффективной реализации систем с высокими требованиями по масштабируемости, надежности и производительности.

Сеть Ангара [2–8] — первая российская высокоскоростная коммуникационная сеть на основе СБИС маршрутизатора. СБИС маршрутизатора коммуникационной сети является разработкой АО «НИЦЭВТ» и выпущена по технологии 65 нм. Сеть поддерживает топологию «многомерный тор» (возможны варианты от 1D- до 4D-тор), режим прямого доступа к памяти удаленных узлов RDMA, технологию GPUDirect, все стандартные средства программирования (библиотека MPI, технология OpenMP, библиотека SHMEM, стек протоколов TCP/IP). Коммуникационная сеть Ангара совместима с процессорами x86, «Эльбрус», а также ускорителями GPU, FPGA. Сеть Ангара отличается высокой пропускной способностью линков и низкими задержками передачи, которые соответствуют мировому уровню.

В настоящий момент в АО «НИЦЭВТ» собран 36-узловой кластер «Ангара-К1» с коммуникационной сетью Ангара. Данная работа посвящена оценочному тестированию сети Ангара в составе данного кластера и сравнению результатов с суперкомпьютером МВС-10П, оснащенного сетью Mellanox InfiniBand 4x FDR.

Предварительное сранительное оценочное тестирование кластера «Ангара-К1» и суперкомпьютера МВС-10П описано в работе [8]. В настоящей работе приведены результаты, полученные после настройки и некоторой оптимизации библиотеки MPI для сети Ангара и кластера «Ангара-К1». Кроме того, в дополнение к тестам, описанным ранее, представлены результаты оценочного тестирования основных коллективных операций: барьерной синхронизации, операций Allreduce и Alltoall.

Статья построена следующим образом. В разделе 2 дано описание использованного оборудования и программного обеспечения. В разделе 3 приведены результаты тестирования простых коммуникационных операций и коллективных операций. В разделе 4 представлены результаты производительности тестов HPL и HPCG, как наиболее распространенных тестов суперкомпьютеров. В разделе 5 приведены результаты тестов NPB, представляющих собой набор часто встречающихся в практике задач и охватывающих широкий диапазон требований к коммуникационной сети. В разделе 6 приведен анализ производительности большого суперкомпьютерного приложения — модели предсказания погоды ПЛАВ.

2. Вычислительные системы

В АО «НИЦЭВТ» собран кластер «Ангара-К1», состоящий из 36 узлов: 24 узла с двумя процессорами Intel Xeon E5-2630 (по 6 ядер, 2.3 ГГц) и 12 узлов с одним процессором Intel Xeon E5-2660 (8 ядер, 2.2 ГГц). Память каждого узла — 64 ГБ. Узлы объединены сетью Ангара с топологией 3D-тор 3×3×4. В исследовании использовалась библиотека MPI, основанная на MPICH версии 3.0.4, а также собственная реализация библиотеки SHMEM, соответствующая версии 1.0 стандарта OpenSHMEM [9].

Сопоставление результатов проводилось с суперкомпьютером МВС-10П, установленным в МСЦ РАН и включающим 207 узлов, в каждом узле по два процессора Intel Xeon E5-2690 (по 8 ядер, 2.9 ГГц) и 64 ГБ памяти. Узлы объединены сетью InfiniBand 4x FDR, топология — жирное дерево, 1:1. Во время тестирования использовалась библиотека Intel MPI 14.1.0.030 в режиме «как есть», то есть без дополнительной настройки.

В таблице 1 приведена сводная характеристика узлов кластера «Ангара-К1» и суперкомпьютера МВС-10П. В таблице 2 приведены основных характеристики используемых суперкомпьютеров.

Необходимо отметить, что используемые в обеих вычислительных системах процессоры фирмы Intel относятся к одному поколению Sandy Bridge, однако в узле МВС-10П находятся два процессора с частотой 2.9 ГГц, которая значительно выше, чем частота процессоров в кластере «Ангара-К1». Поэтому узел МВС-10П значительно мощнее любого из двух типов узлов кластера «Ангара-К1», что необходимо учитывать особенно при сравнении систем на прикладных тестах. Для адекватного сравнения прикладных тестов на обеих системах используется по 8 ядер; на каждом узле это значение соответствует максимальному количеству ядер в узле типа B кластера «Ангара-К1». При использовании большего числа ядер на B-узле режим работы задачи на вычислительном узле может меняться из-за использования технологии Hyper-Threading, адекватное сравнение в этом случае, например, с МВС-10П вряд ли возможно.

Таблица 1. Параметры вычислительных узлов, используемых в кластере «Ангара-К1» и суперкомпьютере МВС-10П.

Параметр	**«Ангара-К1»**		**МВС-10П**
	Узел типа A	**Узел типа B**	
Процессор	2×Intel Xeon E5-2630	Intel Xeon E5-2660	2×Intel Xeon E5-2690
Тактовая частота процессора, ГГц	2.3	2.2	2.9
Количество ядер в узле	12	8	8
Размер кэша L3, МБ	15	20	20
Память узла, ГБ	64	64	64
Пиковая производительность узла, Гфлопс	221	141	371

Таблица 2. Параметры вычислительных систем

Параметр	**«Ангара-К1»**	**МВС-10П**
Количество узлов	24×A, 12×B	207 (36)
Общая пиковая производительность, Тфлопс	6.988	76.838 (13.356)
Сеть	Ангара 3D-тор 3×3×4	InfiniBand 4x FDR Fat Tree 1:1

3. Коммуникационные операции

В данном разделе представлены результаты оценочного тестирования базовых коммуникационных операций, которые часто используются в прикладных задачах.

3.1. Задержка передачи сообщений

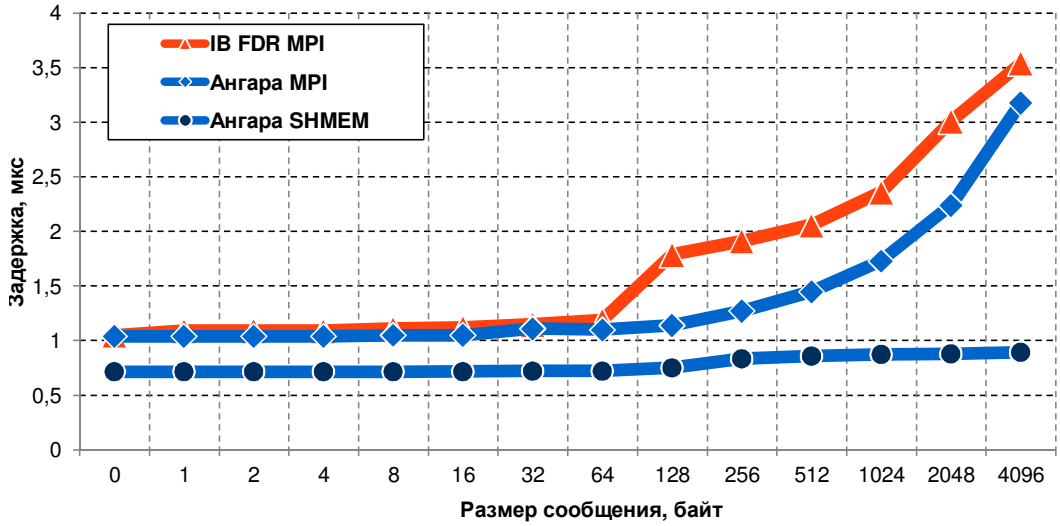


Рис. 1. Задержка передачи сообщения в сетях Ангара и InfiniBand FDR в зависимости от размера сообщения.

Одной из важнейших характеристик сети является задержка на передачу сообщений между двумя соседними узлами. Задержка измерялась при помощи теста osu_latency из пакета OSU Micro-Benchmarks, версия 5.1 [9]. Результаты измерений приведены на рисунке 1.

При использовании библиотеки MPI задержка на сети Ангара меньше, чем на сети InfiniBand FDR при размерах сообщений от 8 байт до 4 Кбайт. Использование библиотеки SHMEM позволяет уменьшить задержку на сети Ангара на 31% в сравнении с результатом, полученным при использовании библиотеки MPI. При этом задержка на передачу коротких сообщений с использованием SHMEM составляет 685 наносекунд, из которых 40 наносекунд уходят на обработку пакета в маршрутизаторе, 89 наносекунд — на передачу по линку, а

оставшиеся 556 наносекунд — это суммарная задержка на инжекцию и эжекцию сообщения через PCI Express. Сеть Ангара имеет торовую топологию, важным параметром является задержка на каждый шаг (хоп), которая составляет 129 наносекунд.

3.2. Коллективные операции

Следующими важными характеристиками сети являются время выполнения коллективных операций — барьерной синхронизации, операций Allreduce и Alltoall. В качестве MPI-реализации тестов выбраны тесты Barrier, Allreduce и Alltoall из пакета IMB (Intel MPI Benchmarks), версия 4.1. Для кластера «Ангара-К1» разработана собственная реализация тестов Barrier и Allreduce на сообщениях в 8 байт с использованием библиотеки SHMEM.

На кластере «Ангара-К1» во всех тестах в статье использовалось следующее правило выбора узлов: для заданного числа узлов выбирались узлы типа B, в случае их нехватки (для конфигураций от 16 узлов) добавлялись узлы типа A. Также для запуска всех тестов использовалось 8 процессов на каждом вычислительном узле.

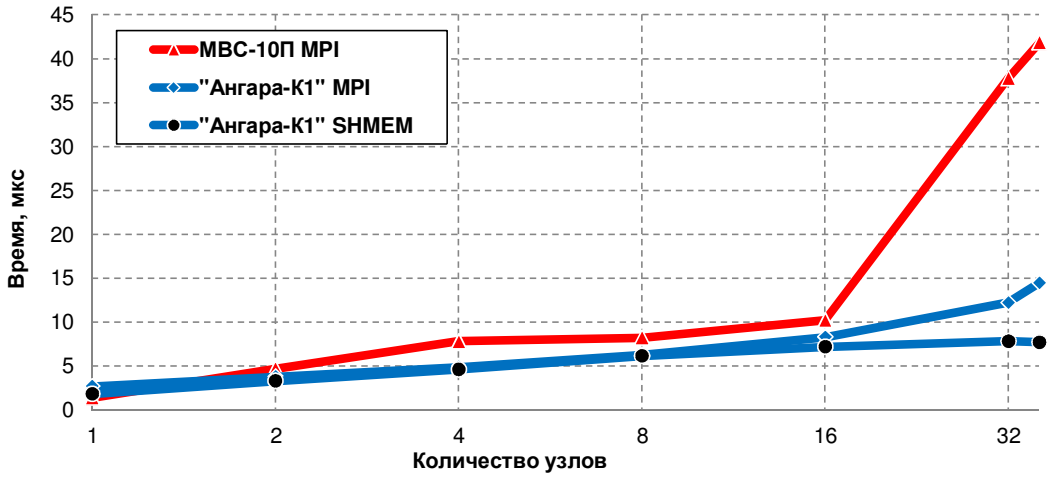


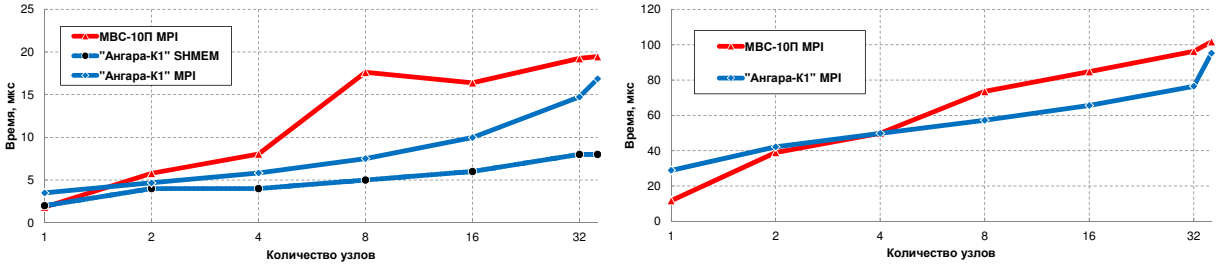
Рис. 2. Время выполнения барьерной синхронизации на кластере «Ангара-К1» и суперкомпьютере МВС-10П. Запуск осуществлялся с использованием 8 процессов на узел.

Время выполнения барьерной синхронизации для обеих систем приведено на рисунке 2. С использованием библиотеки MPI кластер «Ангара-К1» показывает лучшие результаты, чем МВС-10П. Существенный проигрыш МВС-10П на 32 и 36 узлах может быть обусловлен неоптимальным выбором алгоритма реализации функциии MPI_Barrier. На кластере «Ангара-К1» результаты с использованием библиотеки SHMEM значительно превосходят результаты, полученные при использовании библиотеки MPI.

Результат выполнения теста Allreduce приведен на рисунке 3. На рисунке 3a можно видеть график зависимости времени выполнения Allreduce на сообщениях размером 8 байт от числа узлов. Тест Allreduce на коротких сообщениях имеет те же особенности, что и тест Barrier. Время выполнения Allreduce с использованием библиотеки MPI на одном узле на кластере «Ангара-К1» в 1.7 раза хуже, чем на МВС-10П. Это может объясняться тем, что на МВС-10П установлены более мощные узлы. С другой стороны, использование библиотеки SHMEM позволяет нивелировать это отставание, что говорит о возможности оптимизации внутриузловых коммуникаций в библиотеке MPI.

На рисунке 3b представлены графики зависимости времени выполнения Allreduce на сообщениях размером 16 Кбайт в зависимости от числа узлов. На четырех и более узлах результаты на сети Ангара превосходят результаты, полученные на МВС-10П.

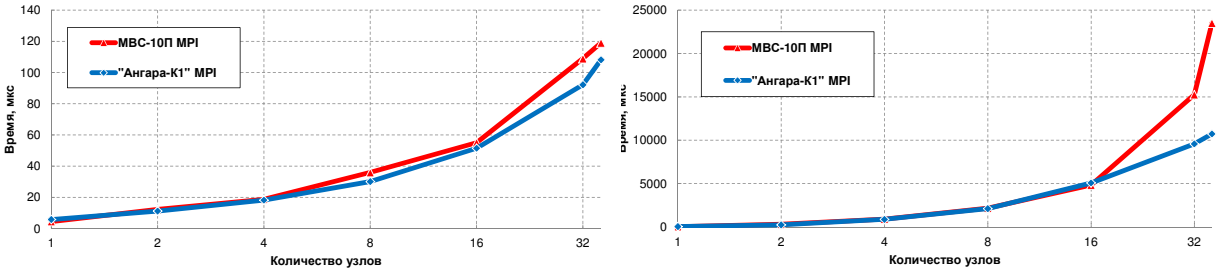
Результаты, полученные на сети Ангара с использованием библиотеки SHMEM, значи-



(a) Allreduce, сообщения 8 байт (b) Allreduce, сообщения 16384 байт

Рис. 3. Время выполнения операции Allreduce на кластере «Ангара-К1» и суперкомпьютере МВС-10П в зависимости от числа узлов на сообщениях в 8 байт (a) и 16384 байт (b). Запуск осуществлялся с использованием 8 процессов на узел.

тельно лучше, чем при использовании библиотеки MPI. С одной стороны, это демонстрация того, что библиотека SHMEM значительно ближе к аппаратуре и вносит меньше накладных расходов, чем библиотека MPI; с другой стороны, это показывает возможность оптимизации библиотеки MPI для сети Ангара.



(a) Alltoall, сообщения 8 байт (b) Alltoall, сообщения 16384 байт

Рис. 4. Время выполнения операции Alltoall на кластере сетью Ангара и суперкомпьютере МВС-10П в зависимости от числа узлов на сообщениях в 8 байт (a) и 16384 байта (b). Запуск осуществлялся с использованием 8 процессов на узел.

Результат выполнения теста Alltoall приведен на рисунке 4. На рисунке 4a представлен график зависимости времени выполнения Alltoall на сообщениях размером 8 байт от числа узлов. Тест Alltoall на коротких сообщениях на сети Ангара работает незначительно быстрее, чем на InfiniBand FDR.

На 36 узлах при размере сообщений 16384 байт (см. рисунок 4b) кластер «Ангара-К1» превосходит МВС-10П в 2.3 раза. Столь существенный выигрыш может быть объяснен не только особенностью сети Ангара, но и неудачными настройками и выбором алгоритма Alltoall на МВС-10П.

4. Тесты HPL и HPCG

Тест HPL (High-Performance LINPACK) [11] используется для ранжирования суперкомпьютеров в списке Top500. Тест LINPACK представляет собой решение СЛАУ $Ax = f$ методом LU-разложения, где A — плотнозаполненная матрица. Тест HPL — это реализация LINPACK на языке C для суперкомпьютеров с распределенной памятью.

Тест HPCG [12] — относительно новый тест, предназначенный для дополнения теста HPL. В основе HPCG лежит решение линейных уравнений с разреженной матрицей большой размерности при помощи итерационного метода сопряженных градиентов с многосе-

точным предобуславливателем. В отличие от HPL тест HPCG обеспечивает стрессовую нагрузку подсистемы памяти вычислительных узлов и коммуникационной сети, представляя значительный класс современных суперкомпьютерных приложений.

Использование тестов HPL и HPCG для оценки суперкомпьютеров позволяет показать верхний и нижний диапазоны реальной производительности для большинства прикладных задач.

Запуск теста HPL проводился с использованием 8 MPI-процессов на узел. На кластере «Ангара-К1» на HPL получено 85% от пиковой производительности в расчете на 8 используемых в узле ядер, см. таблицу 3. Тест HPL содержит мало коммуникационных обменов и не представляет интереса с точки зрения сети, поэтому сравнительного измерения на МВС-10П не проводились.

Для тестирования на обеих вычислительных системах использовалась оптимизированная реализация HPCG [13], разработанная в АО «НИЦЭВТ». Данная реализация включает оптимизации уровня вычислительного узла, в том числе изменение формата хранения разреженной матрицы и векторизацию. Выполнена также оптимизация межпроцессных обменов для сети Ангара при помощи использования библиотеки SHMEM вместо библиотеки MPI. Все проведенные оптимизации являются допустимыми с точки зрения спецификации теста HPCG. По сравнению с базовой версией теста HPCG 2.4 оптимизированная версия дает выигрыш на 36 узлах кластера практически в 2 раза.

Результаты выполнения теста приведены в таблице 3; пиковая производительность рассчитывалась исходя из вычислительной мощности 8 используемых ядер на каждой вычислительной системе. В HPCG межузловой обмен используется в двух функциях: обмен между соседними процессами при вычислении произведения разреженной матрицы на вектор и вычислении предобуславливателя, а также в функции вычисления скалярного произведения (редукция). За счет того, что размер передаваемых в тесте сообщений между соседними процессами небольшой, а задержка передачи коротких сообщений с использованием библиотеки SHMEM в сети Ангара значительно меньше, чем с использованием библиотеки MPI (см. рисунок 1), получен значительный выигрыш в производительности теста при замене библиотеки MPI на библиотеку SHMEM в обеих функциях. В итоге, с использованием SHMEM на кластере «Ангара-К1» удалось получить значительно более высокую производительность по отношению к пиковой в сравнении с достигнутой на МВС-10П.

Таблица 3. Результаты выполнения тестов HPL и HPCG

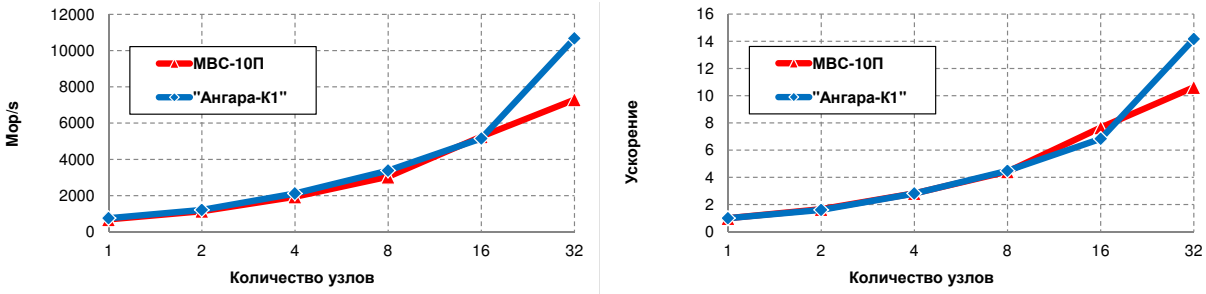
		«Ангара-К1»	**МВС-10П**
HPL	Тфлопс	4.44	–
	% пиковой	85 %	–
HPCG MPI	Гфлопс	279	363
	% пиковой	5.3 %	5.4 %
HPCG SHMEM	Гфлопс	342	–
	% пиковой	6.5 %	–

5. Тесты NPB

Набор тестов NAS Parallel Benchmarks (NPB) [14] является одним из самых распространенных тестов вычислительных систем. Тесты NPB включают в себя ряд синтетических задач и псевдоприложений, эмулирующих реальные приложения в области гидро- и аэродинамики. Сравнение вычислительных систем проводилось на MPI реализации теста версии 3.3.1. Для всех тестов NPB размеры обрабатываемых данных разбиваются на классы, расположенные в порядке увеличения объема обрабатываемых данных: S, A, B, C, D, E. Для тестирования выбран класс C, так как задачи этого класса уже достаточно большие, чтобы обеспечить необходимый параллелизм, но при этом влияние сети является заметным на 32-х узлах. На узле использовалось 8 MPI-процессов, максимальное количество используемых узлов — 32, оно связано с требованием, что количество используемых процессов должно быть степенью двойки.

На рисунках 5-9 показано сравнение производительности и ускорения тестов IS, FT, CG, MG и LU на кластере «Ангара-К1» и суперкомпьютере МВС-10П в зависимости от количества вычислительных узлов. На кластере «Ангара-К1» также, как и в других тестах, для заданного числа узлов выбирались узлы типа B, в случае их нехватки (для конфигураций от 16 узлов) добавлялись узлы типа A.

Графики расположены в порядке уменьшения доли коммуникаций в общем времени выполнения теста. Каждый следующий тест менее требователен к коммуникационной сети, чем предыдущий. Так как узлы МВС-10П более производительны по сравнению с узлами кластера «Ангара-К1», то практически на всех тестах производительность МВС-10П выше. Поэтому для лучшего сравнения сетей кроме реальной производительности используется характеристика полученного ускорения выполнения задачи с увеличением количества вычислительных узлов.



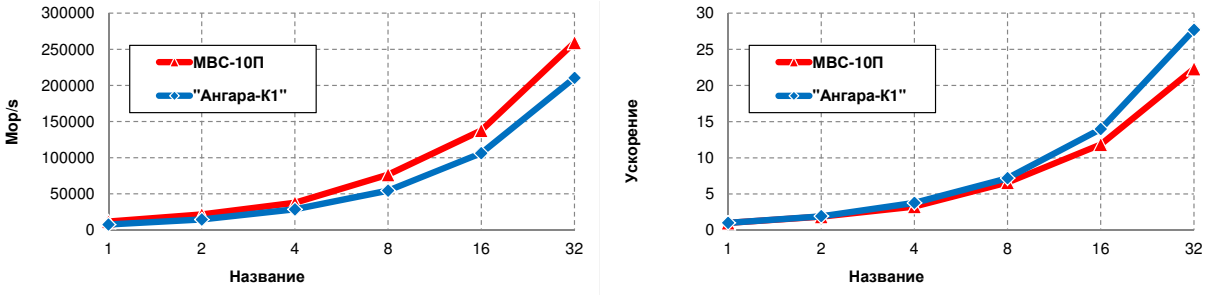
(a) IS производительность

(b) IS ускорение

Рис. 5. Результаты выполнения теста NPB IS на кластере «Ангара-К1» в сравнении с суперкомпьютером МВС-10П.

Тест IS выполняет распределенную сортировку N целых чисел. Доля локальных вычислений в этом тесте мала в сравнении с долей обменов. Производительность теста ограничена временем выполнения обменов типа «все-всем». Так как размер задачи фиксирован, то с ростом числа MPI процессов размер передаваемых сообщений уменьшается. При запуске на 32 узлах средний размер передаваемого сообщения составляет 8192 байта. Это единственный тест, на котором благодаря сети Ангара удается получить лучшую реальную производительность на 32-х узлах, чем на МВС-10П (см. рисунок 5). Полученный результат хорошо соотносится с результатами теста Alltoall (см. рисунок 4).

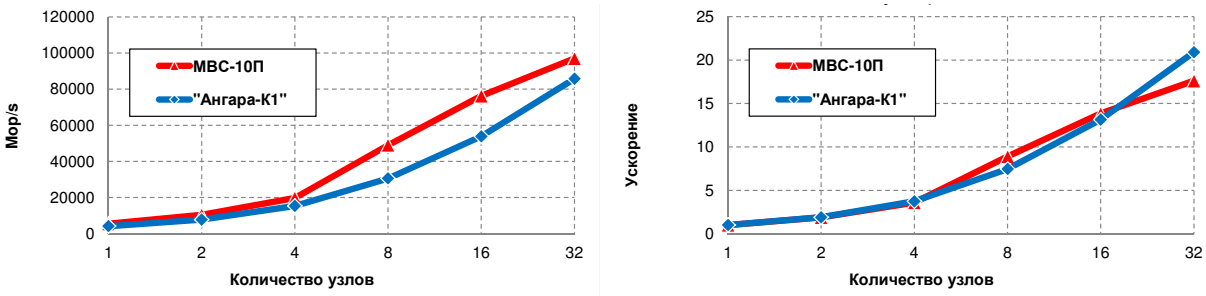
Тест FT заключается в нахождении решения уравнения в частных производных $\frac{\partial u(x,t)}{\partial t} = \alpha\nabla^2 u(x,t)$ при помощи быстого прямого и обратного преобразования Фурье. В отличие от теста IS в тесте FT существенно увеличивается время локальных вычислений. Но этом тесте также имеются передачи значительных объемов данных от каждого MPI-процесса каждо-



(a) FT производительность (b) FT ускорение

Рис. 6. Результаты выполнения теста NPB FT на кластере «Ангара-К1» в сравнении с суперкомпьютером МВС-10П.

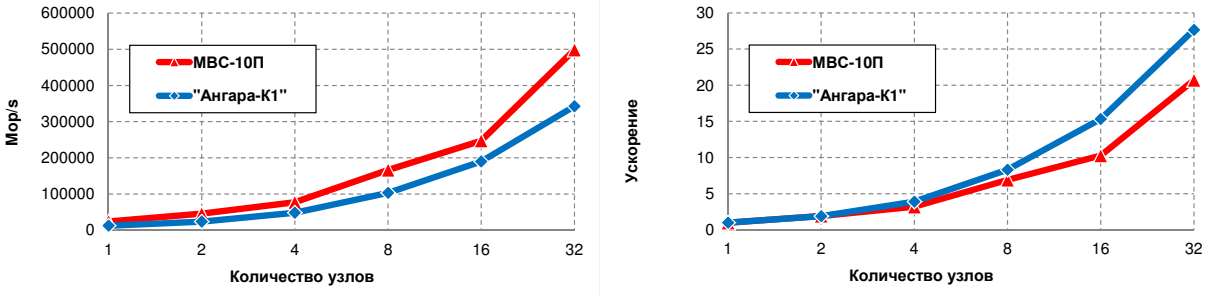
му. На 32-х узлах средний размер сообщений составляет 32768 байт. Несмотря на проигрыш в абсолютной производительности, кластер с сетью Ангара показывает существенно лучшее ускорение: на 32-х узлах производительность на кластере «Ангара-К1» увеличилась в 28 раз против 22-х на МВС-10П. Этот результат также хорошо соотносится с результатами, полученными на тесте Alltoall (см. рисунок 4).



(a) CG производительность (b) CG ускорение

Рис. 7. Результаты выполнения теста NPB CG на кластере «Ангара-К1» в сравнении с суперкомпьютером МВС-10П.

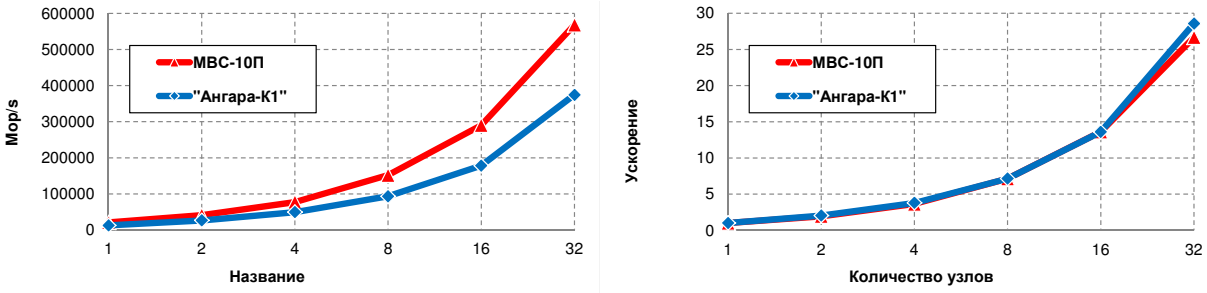
В тесте CG методом сопряженных градиентов вычисляется наименьшее собственное значение разреженной матрицы. Основная часть обменов проходит внутри небольших групп MPI-процессов. На 32-х узлах (см. рисунок 7) кластера «Ангара-К1» наблюдается значительное преимущество ускорения по сравнению с МВС-10П.



(a) MG производительность (b) MG ускорение

Рис. 8. Результаты выполнения теста NPB MG на кластере «Ангара-К1» в сравнении с суперкомпьютером МВС-10П.

Тест MG представляет собой приближенное решение трехмерного уравнения Пуассона в частных производных на заданной сетке с периодическими граничными условиями. В этом тесте имеется некоторая интенсивность коммуникаций, однако производительность кластера «Ангара-К1» уже существенно ниже по сравнению с МВС-10П: роль вычислительного узла в этом тесте уже очень велика. Начиная с 4-х узлов ускорение на кластере с сетью Ангара выше, чем на МВС-10П (см. рисунок 8).



(a) LU производительность

(b) LU ускорение

Рис. 9. Результаты выполнения теста NPB LU на кластере «Ангара-К1» в сравнении с суперкомпьютером МВС-10П.

В тесте LU решается система уравнений с равномерно разреженной блочной треугольной матрицей методом симметричной последовательной верхней релаксации, к которой приводят трехмерные уравнения Навье-Стокса. Как можно видеть на рисунке 9b, ускорение на кластере с сетью Ангара практически совпадает с полученным на МВС-10П. Это вызвано тем, что доля коммуникаций теста LU мала в сравнении с долей локальных вычислений.

6. Тестирование на модели ПЛАВ

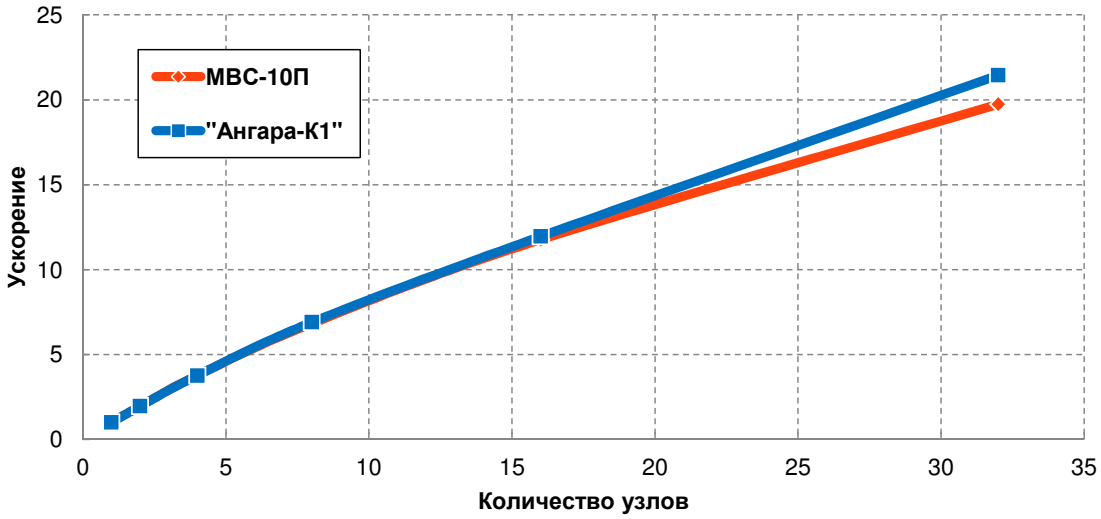


Рис. 10. Сравнение кластера «Ангара-К1» и суперкомпьютера МВС-10П на модели ПЛАВ.

Модель атмосферы ПЛАВ (ПолуЛагранжева, основана на уравнении Абсолютной заВихренности) [15,16] является основной моделью глобального среднесрочного прогноза погоды в России с 2010 года. Она включает в себя блок решения уравнений динамики атмо-

сферы, алгоритмы параметризаций процессов подсеточного масштаба (крупномасштабных осадков, глубокой конвекции, коротковолновой и длинноволновой радиации, пограничного слоя атмосферы, торможения гравитационных волн рельефом, модель многослойной почвы). Для распараллеливания ПЛАВ применяется сочетание библиотеки MPI и технологии OpenMP.

Модель ПЛАВ с разрешением 0.5625 градусов по долготе, переменным шагом по широте от 0.26 до 0.9 градуса, 50 уровнями по вертикали (размерности сетки 640×400×50) запущена на кластере «Ангара-К1» и суперкомпьютере МВС-10П. Данное разрешение выбрано значительно меньшим применяемого в реальных расчетах для демонстрации возможностей сетей на требуемом числе узлов. Для оценки ее производительности использовалось время, затраченное на получение прогноза на 400 часов вперед. При запуске на каждом узле кластера задействовалось 8 ядер: 4 MPI-процесса и 2 OpenMP треда на каждый MPI-процесс. При получении графика ускорения выполнения модели ПЛАВ на кластере «Ангара-К1» использовалось то же правило выбора узлов, что и для тестов NPB: для заданного числа узлов выбирались узлы типа B, в случае их нехватки (для конфигураций от 16 узлов) добавлялись узлы типа A.

На рисунке 10 показано сравнение ускорения, достигнутого при выполнении модели ПЛАВ, на кластере «Ангара-К1» и на суперкомпьютере МВС-10П. Время расчета прогноза ПЛАВ на одном узле кластера «Ангара-К1» составило 4697 секунд, на 32-х узлах — 217 секунд, на одном узле суперкомпьютера МВС-10П — 3411 секунд, на 32-х узлах — 173 секунды. Один из факторов, ограничивающих производительность модели ПЛАВ, — транспонирование матрицы, которое ведет к сложному для сети коммуникационному шаблону обменов «все со всеми». Таким образом, сеть Ангара обеспечивает на 9.6% более высокое ускорение при расчете ПЛАВ по сравнению с InfiniBand: 21.6 раз против 19.7 на МВС-10П. При заданном разрешении модель перестает масштабироваться после 32-х узлов на обеих вычислительных системах из-за отсутствия достаточного параллелизма.

7. Заключение

В статье представлены результаты сравнительного оценочного тестирования 36-узлового вычислительного кластера «Ангара-К1», оснащенного адаптерами коммуникационной сети Ангара, и суперкомпьютера МВС-10П с сетью InfiniBand 4x FDR, установленного в МСЦ РАН.

Оценочное тестирование проведено при помощи тестов разных уровней: простых коммуникационных операций, широкоизвестных тестов оценки производительности суперкомпьютеров HPL и HPCG, набора тестов NPB уровня прикладных задач, охватывающих широкий диапазон требований к коммуникационной сети, а также на модели предсказания погоды ПЛАВ.

Тест измерения задержки передачи сообщения с использованием библиотеки MPI показывает превосходство сети Ангара над сетью InfiniBand FDR при размерах сообщения от 8 байт до 4 Кбайт. Использование библиотеки SHMEM на сети Ангара позволяет получить значительно лучшие характеристики.

Для теста HPL показана возможность получения на кластере «Ангара-К1» необходимой реальной производительности. Тест HPCG предъявляет значительно более высокие требования к подсистеме памяти и коммуникационной сети, чем тест HPL. Для исследования использовалась оптимизированная авторами данной работы версия теста. Применение библиотеки MPI на кластере «Ангара-К1» в сравнении с МВС-10П позволило получить на данном тесте одинаковый уровень производительности по отношению к теоретической пиковой, а использование библотеки SHMEM позволило значительно улучшить результат на кластере «Ангара-К1».

Рассматриваемый набор тестов NPB включает тесты LU, MG, FT, CG, IS, рассматрива-

ется класс C. Каждый узел МВС-10П более производителен по сравнению с узлом кластера «Ангара-К1», поэтому общая производительность МВС-10П на тестах NPB опережает «Ангара-К1». Однако по ускорению кластер «Ангара-К1» на всех тестах с интенсивными коммуникациями опережает МВС-10П с сетью InfiniBand 4x FDR. Кроме того, на тесте сортировки целых чисел IS за счет лучшей реализации операции MPI_Alltoallv кластер «Ангара-К1» достиг на 32-х узлах более высокой производительности, чем суперкомпьютер МВС-10П с сетью InfiniBand 4x FDR.

Модель прогноза погоды ПЛАВ на небольшой расчетной сетке на кластере «Ангара-К1» показала ускорение, превышающее на 9.6% полученное на суперкомпьютере МВС-10П.

Дальнейшие работы включают в себя детальный анализ производительности тестов NPB, оптимизацию тестов NPB и прикладных задач при помощи библиотеки SHMEM, а также оптимизацию библиотеки MPI для сети Ангара.

В настоящее время ведется разработка второго поколения высокоскоростной коммуникационной сети Ангара, что показывает важный для пользователей факт, что пользователям при смене поколений оборудования не придется переучиваться и привыкать к новой технологии, оптимизированные под сеть Ангара программы будут также эффективнее работать при использовании сети Ангара-2.

Авторы статьи выражают благодарность Михаилу Андреевичу Толстых и Ростиславу Фадееву за помощь в исследовании производительности модели прогноза погоды ПЛАВ.

Литература

1. Top500 Supercomputing Sites. URL: `Top500.org` (дата обращения: 21.02.2016).

2. Макагон Д.В., Сыромятников Е.Л. Сети для суперкомпьютеров // Открытые системы. — 2011. — №7. — С. 33-37.

3. Корж А.А., Макагон Д.В., Жабин И.А., Сыромятников Е.Л. Отечественная коммуникационная сеть 3D-тор с поддержкой глобально адресуемой памяти для суперкомпьютеров транспетафлопсного уровня производительности. // Параллельные вычислительные технологии (ПаВТ'2010): труды международной научной конференции (29 марта-2 апреля 2010 г., г. Уфа). Челябинск: Издательский центр ЮУрГУ, 2010. — С. 227-237.
URL: `http://omega.sp.susu.ac.ru/books/conference/PaVT2010/full/134.pdf` (дата обращения: 29.04.2015).

4. Симонов А.С., Жабин И.А., Макагон Д.В. Разработка межузловой коммуникационной сети с топологией «многомерный тор» и поддержкой глобально адресуемой памяти для перспективных отечественных суперкомпьютеров. // Научно-техническая конференция «Перспективные направления развития вычислительной техники», ОАО «НИЦЭВТ», 2011.

5. Симонов А.С., Макагон Д.В., Жабин И.А., Щербак А.Н., Сыромятников Е.Л., Поляков Д.А. Первое поколение высокоскоростной коммуникационной сети «Ангара» // Наукоемкие технологии. — 2014. — Т. 15, №1. — С. 21-28.

6. Слуцкин А.И., Симонов А.С., Жабин И.А., Макагон Д.В., Сыромятников Е.Л. Разработка межузловой коммуникационной сети ЕС8430 «Ангара» для перспективных суперкомпьютеров // Успехи современной радиоэлектроники. — 2012. — №1.

7. Жабин И.А., Макагон Д.В., Симонов А.С. Кристалл для Ангары // Суперкомпьютеры. — Зима-2013. — С. 46-49.

8. Агарков А.А., Исмагилов Т.Ф., Макагон Д.В., Семенов А.С., Симонов. А.С. Предварительные результаты оценочного тестирования отечественной

высокоскоростной коммуникационной сети Ангара // Параллельные вычислительные технологии (ПаВТ'2016): труды международной научной конференции (28 марта – 1 апреля 2016 г., г. Архангельск). Челябинск: Издательский центр ЮУрГУ, 2016. – С. 42–53.

9. OpenSHMEM Application Programming Interface, Version 1.0, 31 January 2012. URL: http://openshmem.org/site/sites/default/site_files/openshmem_specification-1.0.pdf (дата обращения: 29.11.2015).

10. OSU Micro-Benchmarks. URL: http://mvapich.cse.ohio-state.edu/benchmarks/ (дата обращения: 29.11.2015).

11. Intel MPI Benchmarks. URL: https://software.intel.com/en-us/articles/intel-mpi-benchmarks (дата обращения: 29.11.2015).

12. High-Performance LINPACK. URL: http://www.netlib.org/benchmark/hpl/ (дата обращения: 29.11.2015).

13. M. Heroux, J. Dongarra, P. Luszczek. HPCG Technical Specification. Sandia Report SAND2013-8752. Printed October 2013. URL: https://software.sandia.gov/hpcg/doc/HPCG-Specification.pdf (дата обращения: 10.06.2015).

14. А.А. Агарков, А.С. Семенов, А.С. Симонов. Оптимизация теста HPCG для суперкомпьютеров с сетью «Ангара» // Суперкомпьютерные дни в России: Труды международной конференции (28–29 сентября 2015 г., г. Москва). — 2015. — С. 294–302.

15. NAS Parallel Benchmarks. URL: https://www.nas.nasa.gov/publications/npb.html (дата обращения: 29.11.2015).

16. Толстых М.А. Глобальная полулагранжева модель численного прогноза погоды. М, Обнинск: ОАО ФОП, 2010. 111 стр.

17. Толстых М.А., Мизяк В.Г. Параллельная версия полулагранжевой модели ПЛАВ с горизонтальным разрешением порядка 20 км // Труды Гидрометеорологического научно-исследовательского центра Российской Федерации. — 2011. — No 346. — С. 181–190.

Performance Evaluation of the «Angara» Interconnect

A.A. Agarkov, T.F. Ismagilov, D.V. Makagon, A.S. Semenov, A.S. Simonov

AO «NICEVT»

The paper presents performance evaluation results of 36-nodes cluster with «Angara» interconnect compared with MVS-10P supercomputer of JSCC RAS with InfiniBand 4x FDR interconnect.

Keywords: interconnect, «Angara», InfiniBand FDR, NPB, HPCG, HPL, SLAV

References

1. Top500 Supercomputing Sites. URL: `Top500.org` (accessed: 21.02.2016).

2. Makagon D.V., Syromyatnikov E.L. Seti dlya superkomp'yuterov [Supercomputers Interconnect]. Otkrytyye sistemy. SUBD. [Open Systems. DBMS]. — 2011. — N 7. —P. 33–37.

3. Korzh A.A., Makagon D.V., Zhabin I.A., Syromyatnikov E.L. Otechestvennaya kommunikatsionnaya set' 3D-tor s podderzhkoy global'no adresuyemoy pamyati dlya superkomp'yuterov transpetaflopsnogo urovnya proizvoditel'nosti [Russian 3D-torus Interconnect with Support of Global Address Space Memory]. Parallelnye vychislitelnye tekhnologii (PaVT'2010): Trudy mezhdunarodnoj nauchnoj konferentsii (Ufa, 29 marta – 2 aprelya 2010) [Parallel Computational Technologies (PCT'2010): Proceedings of the International Scientific Conference (Ufa, Russia, March, 29 – April, 2, 2010)]. Chelyabinsk, Publishing of the South Ural State University, 2010. P. 527–237.
URL: `http://omega.sp.susu.ac.ru/books/conference/PaVT2010/full/134.pdf` (accessed: 29.04.2015).

4. Simonov A.S., Zhabin I.A., Makagon D.V. Razrabotka mezhuzlovoy kommunikatsionnoy seti s topologiyey «mnogomernyy tor» i podderzhkoy global'no adresuyemoy pamyati dlya perspektivnykh otechestvennykh superkomp'yuterov [Development of the Multi-Dimensional Torus Topology Interconnectwith Support of Global Address Space Memory for Advanced National Supercomputers]. Nauchno-tekhnicheskaya konferentsiya «Perspektivnyye napravleniya razvitiya vychislitel'noy tekhniki» (Moskva, 28 iyunya) [Scientific and Technical Conference «Advanced Directions of the Computers Development Technology]. — Moscow: JSC «Concern «Vega», 2011. — P. 17–19

5. Simonov A.S., Makagon D.V., Zhabin I.A., Shcherbak A.N., Syromyatnikov E.L., Polyakov D.A. Pervoye pokoleniye vysokoskorostnoy kommunikatsionnoy seti «Angara» [The First Generation of Angara High-Speed Interconnect]. Naukoyemkiye tekhnologii [Science Technologies]. — 2014. — V. 15, N 1. — P. 21–28.

6. Slutskin A.I., Simonov A.S., Zhabin I.A., Makagon D.V., Syromyatnikov E.L. Razrabotka mezhuzlovoy kommunikatsionnoy seti YES8430 «Angara» dlya perspektivnykh superkomp'yuterov [Development of ES8430 Angara Interconnect for Future Russian Supercomputers]. Uspekhi sovremennoy radioelektroniki [Progress of the Modern Radioelectronics]. — 2012. — N 1. — P. 6–10.

7. Zhabin I.A., Makagon D.V., Simonov A.S. Kristall dlya Angary [Angara Chip] // Superkomp'yutery [Supercomputers]. —Winter-2013. — P. 46–49.

8. OpenSHMEM Application Programming Interface, Version 1.0, 31 January 2012. URL: http://openshmem.org/site/sites/default/site_files/openshmem_specification-1.0.pdf (accessed: 29.11.2015)

9. OSU Micro-Benchmarks. URL: http://mvapich.cse.ohio-state.edu/benchmarks/ (accessed: 29.11.2015).

10. Intel MPI Benchmarks. URL: https://software.intel.com/en-us/articles/intel-mpi-benchmarks (accessed: 29.11.2015).

11. High-Performance LINPACK. URL: http://www.netlib.org/benchmark/hpl/ (accessed: 29.11.2015).

12. M. Heroux, J. Dongarra, P. Luszczek. HPCG Technical Specification. Sandia Report SAND2013-8752. Printed October 2013. URL: https://software.sandia.gov/hpcg/doc/HPCG-Specification.pdf (accessed: 10.06.2015).

13. Agarkov A.A., Semenov A.S., Simonov A.S. Optimizaciya testa HPCG dlya superkomp'yuterov s set'yu «Angara» [Optimized Implementation of HPCG Benchmark on Supercomputer with "Angara"Interconnect]. // Superkomp'yuternye dni v Rossii: Trudy mezhdunarodnoj konferencii (28-29 sentyabrya 2015 g., g. Moskva) [Russian Supercomputing Days: Proceedings of the International Conference (Moscow, Russia, September 28-29, 2015.)]. — 2015. — C. 294-302.

14. NAS Parallel Benchmarks. URL: https://www.nas.nasa.gov/publications/npb.html (accessed: 29.11.2015).

15. Tolstykh M.A. Global'naya polulagranzheva model' chislennogo prognoza pogody [Global Semi-Lagrangian Model Numerical Weather Prediction Model]. M, Obninsk: OAO FOP, 2010. P. 111.

16. Tolstykh M.A., Mizyak V.G. Parallel'naya versiya polulagranzhevoj modeli PLAV s gorizontal'nym razresheniem poryadka 20 km [Parallel Implementation of the Semi-Lagrangian Model SLAV with Resolution about 20 km] // Trudy Gidrometeorologicheskogo nauchno-issledovatel'skogo centra Rossijskoj Federacii [Proceedings of the Hydrometeorological Center of Russian Federation]. 2011. No 346. P. 181-190.