УДК 004.72, 004.94 DOI: 10.14529/cmse150304

ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ ПОДСЕТИ КОЛЛЕКТИВНЫХ ОПЕРАЦИЙ СЕТИ «АНГАРА»[1]

А.В. Мукосей, А.С. Семенов, А.С. Симонов

В ОАО «НИЦЭВТ» разрабатывается высокоскоростная коммуникационная сеть «Ангара» с топологией «многомерный тор». Для исследования и оценки производительности разрабатываемой сети при большом количестве используемых узлов создана параллельная потактовая имитационная модель сети. Сеть «Ангара» имеет аппаратную поддержку двух коллективных операций — broadcast и reduce. В статье описана реализация коллективных операций в имитационной модели, и представлены результаты оценки их производительности при помощи модели. Оценки производительности получены на базовых тестах broadcast и reduce, а также на прикладных задачах — умножение разреженной матрицы на вектор и численное решение нелинейного уравнения теплопроводности.

Ключевые слова: имитационное моделирование, «Ангара», многомерный тор, коммуникационная сеть, коллективные операции.

Введение

В настоящее время суперкомпьютеры содержат сотни тысяч вычислительных ядер. Эффективность одновременной работы ядер на задачах с интенсивным обменом данными между ними (задачи моделирования, задачи на графах и нерегулярных сетках, вычисления с использованием разреженных матриц) в основном определяется производительностью коммуникационной сети, соединяющей вычислительные узлы высокоскоростными каналами связи (линками).

В ОАО «НИЦЭВТ» разрабатывается высокоскоростная коммуникационная сеть «Ангара» с топологией «многомерный тор» [1–6]. В 2013 году выпущен кристалл маршрутизатора этой сети [7], на его основе в 2015 году ожидается построение суперкомпьютера.

Для некоторых прикладных задач требуется эффективное выполнение коллективных коммуникационных операций, в которых задействовано сразу много вычислительных узлов. В сети «Ангара» реализована аппаратная поддержка двух коллективных операций — broadcast и reduce [8]. Для этого добавлена виртуальная подсеть, состоящая из двух виртуальных каналов с особыми правилами маршрутизации. Виртуальная подсеть имеет топологию дерева, наложенную на «многомерный тор».

Для оценки производительности на тестовых программах и для исследования новых архитектур на языке Charm++ создана параллельная потактовая имитационная модель разрабатываемой коммуникационной сети [4]. Однако поддержка коллективных операций в модели отсутствовала.

Имитационное моделирование позволяет исследовать характеристики сетей и суперкомпьютеров, состоящих из большого числа узлов, что особенно важно для коллективных операций, эффективная реализация которых начинает проявляться при большом числе используемых узлов. Имитационное моделирование большого числа узлов важно из экономических соображений, так как большой суперкомпьютер-макет построить дорого по экономическим соображениям.

[1] Статья рекомендована к публикации программным комитетом Международной научной конференции «Параллельные вычислительные технологии – 2015».

Целью данной работы является реализация подсети коллективных операций в имитационной модели, а также демонстрация производительности подсети коллективных операций при помощи разработанной модели. Статья организована следующим образом. Во втором разделе описывается архитектура поддержки коллективных операций в коммуникационной сети. Третий раздел посвящен реализации поддержки коллективных операций в потактовой имитационной модели сети. В четвертом разделе описываются оценки производительности коллективных операций, полученные при помощи имитационной модели сети. В заключении перечисляются основные результаты работы и планы дальнейших исследований.

1. Коллективные операции в маршрутизаторе коммуникационной сети «Ангара»

Коллективные операции используются для обмена данными между несколькими узлами системы. Их относят к основным примитивам взаимодействия вычислительных процессов в большинстве стандартов параллельного программирования, ориентированных на выполнение на системах с распределенной памятью (MPI [9], Shmem [10], PGAS-языки — UPC [11], X10 [12]); они могут составлять значительную часть коммуникационных обменов в процессе работы [13]. Реализация коллективных операций с использованием операций типа «точка-точка» имеет ряд недостатков, таких как большая доля дублирующего трафика, плохая масштабируемость [14], поэтому их аппаратная поддержка способствует повышению масштабируемости параллельной программы (см., например, [15]).

Высокоскоростная коммуникационная сеть «Ангара» с топологией «многомерный тор» поддерживает детерминированную и адаптивную передачу пакетов, неблокирующие записи, чтения, атомарные операции, отказоустойчивость на канальном уровне и обход отказавших каналов и узлов. В рамках данной сети реализована аппаратная поддержка двух коллективных операций — broadcast и reduce [8]. Для этого добавлена виртуальная подсеть, состоящая из двух виртуальных каналов с отдельными буферами и специальными правилами маршрутизации. Виртуальная подсеть имеет топологию дерева (рис. 1), построенного в торе. Выбирается корневой узел, от которого строится дерево с учетом порядка измерений: X, Y, Z, W (это позволяет предотвратить возможные дедлоки). В построенном дереве существует два направления движения: от корня и к корню. Каждому направлению соответствует свой виртуальный канал. В системе могут быть транзитные узлы, в них процессоры не посылают и не получают данных.

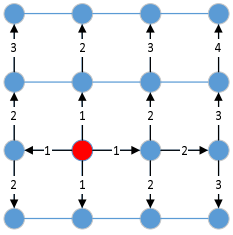


Рис. 1. Виртуальная подсеть коллективных операций на примере 2D-решетки. Стрелками обозначено направление движения от корня, цифрами обозначены этапы обхода дерева

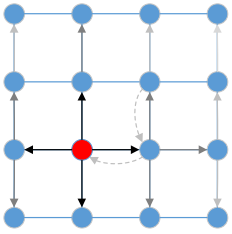


Рис. 2. Схема выполнения операции broadcast не из корневого узла на примере 2D-решетки. Пунктирной линией выделен путь до корня. Сплошной — распространение от корня

При выполнении операции broadcast каждый узел при получении пакета от узла, находящегося выше по дереву, рассылает его всем узлам, находящимся ниже по дереву (см. рис. 1). При инжекции пакета в сеть не в корневом узле сначала генерируется запрос на broadcast, который посылается корню (рис. 2).

При выполнении операции reduce (или варианта all reduce) узел ожидает пакеты от хоста, если узел не транзитный, и всех узлов, находящихся ниже по дереву; выполняет над ними указанную в пакете коммутативную ассоциативную бинарную операцию и отправляет готовый результат вверх к корню. В текущей реализации поддерживаются операции максимума, минимума и суммы целых чисел. Операция reduce в корне завершается аппаратной отправкой (без эжекции) результата заданному узлу посредством операции «точка-точка» (broadcast для all reduce).

Для определения направления к корню и от корня на каждом узле задается таблица маршрутизации коллективной подсети, при этом указываются следующие поля: направления на узлы ниже и выше по дереву; является ли узел транзитным или корневым. Для задания корректного дерева должны выполняться следующие критерии:

- корень ровно один;
- если в каком-то узле выставлено направление вниз по дереву, то в этом направлении должен находиться принадлежащий дереву узел, в котором направление вверх по дереву выставлено противоположным данному;
- направления на узлы ниже по дереву могут быть только: 1) по измерениям, следующим за направлением вверх по дереву в рамках порядка направлений, 2) по направлению, противоположному направлению вверх по дереву.

В рамках сети можно задавать различные пересекающиеся деревья. Поддерживается 16 деревьев, каждому соответствует свой идентификатор TreeId, по которому производится выборка из таблицы маршрутизации при принятии решения по маршрутизации пакета. Одновременно маршрутизатор поддерживает до 16 различных пакетов reduce по каждому TreeId. Каждый reduce, выполняющийся по данному дереву, имеет свой идентификатор ReduceId (от 0 до 15).

С точки зрения прикладного программиста, базовые версии коллективных операций — односторонние асинхронные операции. Процессор не блокируется после отправки сообщения, а результат записывается в память без активного участия принимающей стороны, это позволяет совмещать ожидание окончания операции со счетом. Для того, чтобы узнать, что коллективная операция завершилась и результат доступен вычислительным узлам, существуют механизмы синхронизации, основанные на коллективных операциях.

2. Реализация коллективных операций в параллельной имитационной модели

Для оценки производительности и исследования новых архитектур высокоскоростной коммуникационной сети разработана и используется параллельная потактовая имитационная модель [4]. Модель разработана на языке Charm++ и позволяет моделировать на вычислительном кластере конфигурации с большим количеством моделируемых узлов сети. Модель маршрутизатора устроена достаточно гибко и имеет большое количество конфигурационных параметров. Это позволяет подстраивать модель для различных типов сетей.

В модели реализованы следующие топологии: тор произвольной размерности, сеть Кэли, сеть Клоса.

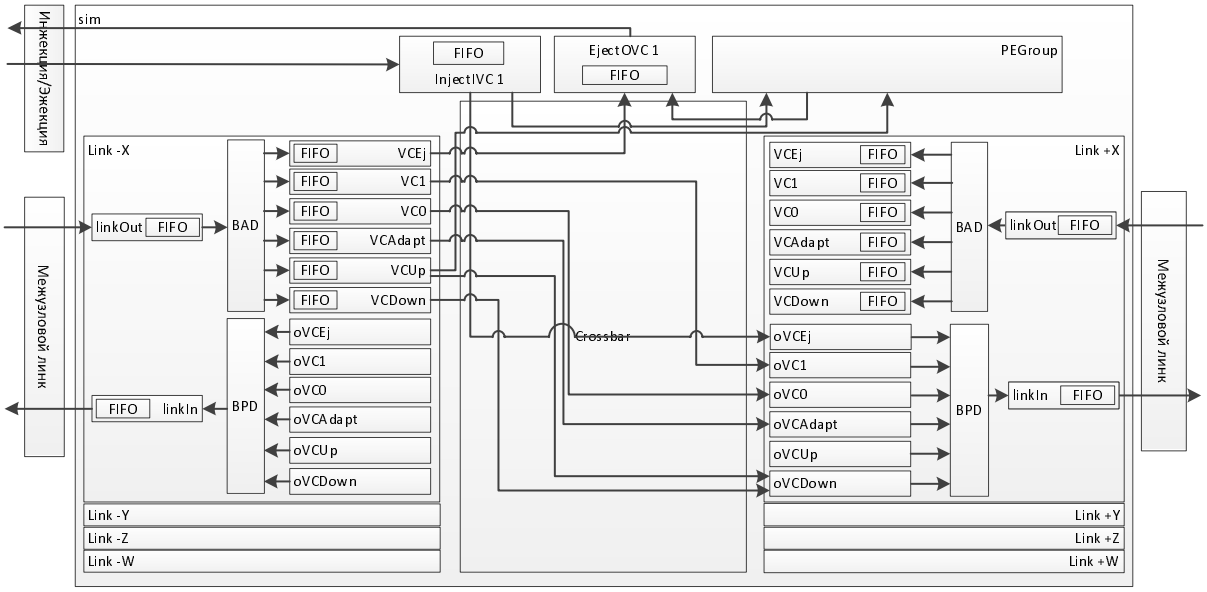


Рис. 3. Общая схема маршрутизатора коммуникационной сети с топологией «многомерный тор», реализованная в имитационной модели сети

На рис. 3 представлена общая схема маршрутизатора рассматриваемой сети, реализованная в имитационной модели.

Маршрутизатор имеет два типа входов: межузловые и инжекционные. Аналогично, имеется два типа выходов: межузловые и эжекционные. Межузловые входы (выходы) соединяются с выходами (входами) других узлов соответственно, посредством физических каналов, так называемых межузловых линков. Считается, что помехи отсутствуют. Инжекционные (эжекционные) каналы, так называемые процессорные, служат для связи процессора с маршрутизатором. Каждый процессор соединяется с маршрутизатором посредством одного или нескольких инжекционных и такого же количества эжекционных каналов. Каналы представляют собой FIFO-буфера, доступные только для чтения или только для записи. Минимальная длина передаваемых данных в сети — 128 бит (флит данных).

Для поддержки виртуальных подсетей в маршрутизаторе предусмотрены виртуальные каналы (VC). Виртуальные каналы представляют собой FIFO-буфера с блоком маршрутизации. Каждый пакет из линка через блок анализа данных BAD попадает в заданный типом пакета FIFO-буфер.

Для реализации подсети коллективных операций добавлены два виртуальных канала: для движения к корню (VCUp) и для движения к листьям (VCDown), а также блок PEGroup для эжекции, инжекции и обработки коллективных пакетов.

2.1. Виртуальные каналы VCUp и VCDown

За основу виртуальных каналов VCUp и VCDown взят детерминированный виртуальный канал (VCDet). При этом изменена маршрутизация, а также добавлен механизм выдачи копий пакета. Маршрутизация осуществляется по данным из таблицы маршрутизации и по заголовку пакета. Таблица маршрутизации заполняется на этапе инициализации модели. В таблице 16 строк, по строке на каждое дерево. У строк есть следующие поля: $TreeId$

— номер дерева подсети коллективных операций, $isRoot$ — определяет является ли узел корнем, $toRoot$ — направление вверх к корню, Pe — участвует ли узел в коллективной операции или он транзитный, $Dirs$ — направления вниз по дереву, Dir_sum — количество направлений (нужно для контроля выданных пакетов). Процесс прохождения пакета по виртуальным каналам коллективных операций состоит из следующих этапов (рис. 4):

- ожидание приема всего пакета в буфер виртуального канала;
- чтение таблицы маршрутизации, анализ головного флита пакета;
- в зависимости от виртуального канала и полученной выше информации составляется список направлений на передачу: детям, в вычислительный узел или к корню.

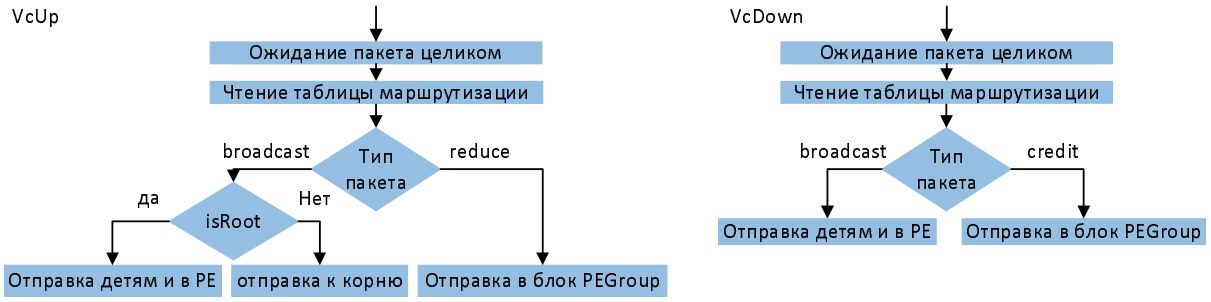


Рис. 4. Логика работы виртуальных каналов VCUp и VCDown

2.2. Блок PEGroup

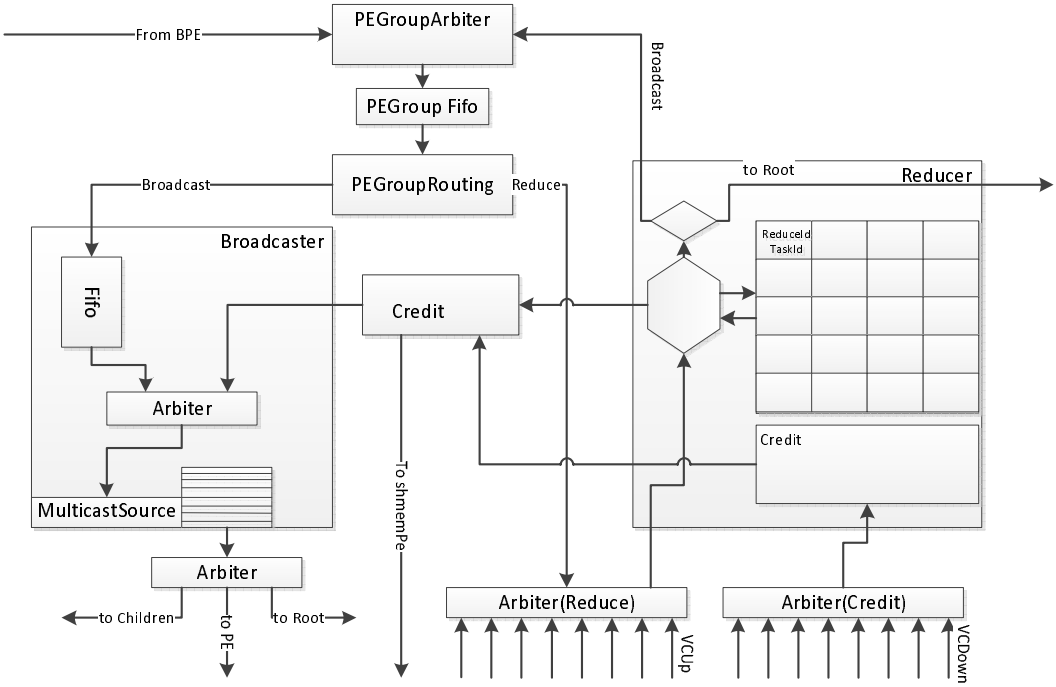


Рис. 5. Общая схема блока PEGroup

Блок PEGroup (рис. 5) предназначен для приема/передачи пакетов типа broadcast, reduce-пакетов и кредитных пакетов. Он разделен на три основные части: Broadcaster — блок рассылки broadcast пакетов, Reducer — блок для пакетов reduce и блок анализа кредитных пакетов. Инжектированные пакеты коллективных операций от процессора сначала попадают в очередь PEGroup Fifo блока PEGroup. После чего блок маршрутизации, ана-

лизируя пакеты, отправляет их в соответствующий блок. Reduce- и кредитные пакеты, пришедшие из виртуальных каналов, попадают в блок Reducer.

2.3. Блок Reducer

Блок Reducer принимает пакеты типа reduce. Над данными из пакета типа reduce блок Reducer производит заданную операцию (определяется по данным из головного флита пакета), сохраняет значения у себя в памяти для дальнейших операций или отправляет дальше по сети. Для каждого TreeId и ReduceId в блоке Reducer имеется изначально равный нулю счетчик принятых пакетов. Когда в Reducer из виртуального канала приходит пакет reduce, то его процесс прохождения по блоку Reducer будет следующим:

- Из заголовочного флита пакета считываются TreeId и ReduceId.
- Увеличивается на 1 значение счетчика принятых пакетов по TreeId и ReduceId.
- Считываются значения из памяти по TreeId и ReduceId (если это первый пакет, то ничего не происходит).
- Производится заданная арифметическая операция над каждым флитом данных из вновь пришедшего пакета и соответствующим значением из памяти (если это первый пакет, то ничего не происходит).
- Полученные значения вновь записываются в память по TreeId и ReduceId (если это первый пакет, то в качестве значений берутся флиты данных пакета).
- Если пришедший пакет был последним, то полученный результат выдается в блок маршрутизации. В кредитном блоке увеличивается на 1 значение счетчика выданных результирующих пакетов reduce по TreeId.

Если узел не был корневым, то маршрутизатор выставит запрос к корню вверх по дереву. Если узел корневой, то будет выставлен запрос на передачу в очередь PEGroup Fifo, для рассылки результата операции редукции (reduce убрать).

2.4. Блок анализа кредитных пакетов

Блок анализа кредитных пакетов создан для контроля количества одновременно обрабатываемых reduce пакетов в блоке Reducer; допустимо 16 операций по одному дереву. Блок отправляет кредитный пакет в PE и всем своим детям. Отправка детям происходит каждый раз, когда из узла было отправлено вверх по дереву или в процессор 8 пакетов с результатом редукции для каждого фиксированного дерева. Отправка в PE осуществляется каждые 128 отправленных пакетов с результатом редукции по всем TreeId или каждые 4096 тактов. В процессор отправляются значения всех счетчиков выполненных операций редукции по всем TreeId. Также блок анализирует приходящие кредитные пакеты. Каждый такой пакет означает, что узел-родитель может принять еще 8 пакетов по данному TreeId. Кредитный блок осуществляет передачу пакетов через блок Broadcaster.

2.5. Блок Broadcaster

Блок Broadcaster берет поочередно пакеты из очереди PEGroup Fifo и кредитного блока, анализирует заголовки пакетов и таблицу маршрутизации и принимает решение о маршрутизации: вниз по дереву, вверх к корню или на инжекцию в PE.

В модели реализован механизм построения временных диаграмм прохождения пакета по маршрутизатору. Для каждого реализованного блока с помощью этого механизма

построены временные диаграммы. Все диаграммы соответствую временным диаграммам аппаратной реализации. Пример диаграммы прохождения пакета по виртуальному каналу VCUp приведен на рис. 6.

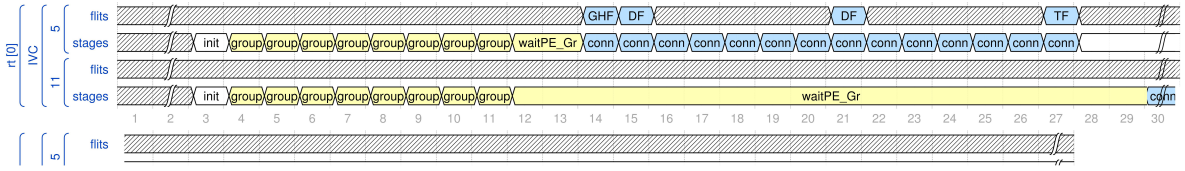


Рис. 6. Временная диаграмма прохождения пакета по каналу VCUp

3. Оценка производительности коллективных операций на базовых тестах и прикладных задачах

Оценка производительности коллективных операций проводилась в два этапа. На первом этапе исследовались базовые операции broadcast и all reduce, имеющие непосредственную аппаратную поддержку в подсети коллективных операций. Задача второго этапа — демонстрация производительности небольших прикладных тестов-задач. Выбраны следующие задачи:умножение разреженной матрицы на вектор, численное решение задачи с нелинейным уравнением теплопроводности.

Оценки производительности были получены при помощи имитационной модели сети, которая описана в третьем разделе. Параметры модели сети соответствуют значениям параметрам СБИС маршрутизатора сети «Ангара», выпущенного в 2013 году и приведены в таблице. Необходимые вычисления проводились на процессоре, параметры которого также приведены в таблице.

Таблица

Параметры оцениваемого суперкомпьютера

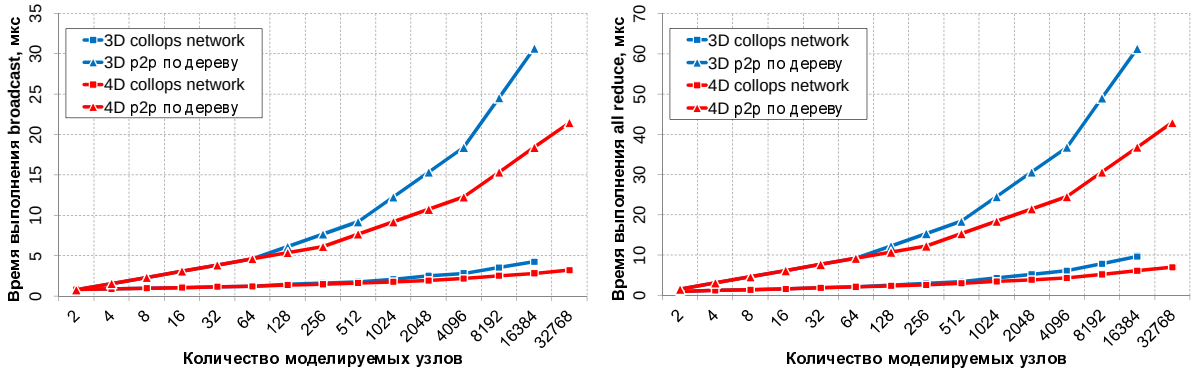
Параметр	**Значение**
Процессор	Intel Xeon E5-2660
Тактовая частота процессора, ГГц	2.2
Количество ядер в узле	8
Размер кэша L3, МБ	20
Пиковая производительность узла, Гфлопс	140.8
Тактовая частота маршрутизатора, МГц	500
Интерфейс процессора с сетью	PCIe gen2 x16
Задержка на инжекцию (эжекцию) пакета через PCIe, нс	300
Задержка передачи по линку, нс	80
Топология сети	3D/4D-тор

3.1. Базовые тесты broadcast и all reduce

На этапе тестирования базовых операций оценивалась производительность базовых операций broadcast и all reduce. Тесты заключались в посылке одного пакета максимального размера в 16 флитов (256 байт) при помощи операций broadcast и all reduce для разного количества узлов моделируемой сети.

На рис. 7 представлены результаты выполнения операций broadcast и all reduce. Синим и красным показано время выполнения (в мкс) коллективной операции для сети с топологией 3D- и 4D-тор соответственно. Квадратными маркерами отмечено время выполнения коллективной операции, реализованной с помощью сети коллективных операций, треугольными маркерами — при помощи коммуникационных операций «точка-точка» по тому же дереву, что и в подсети коллективных операций, но полученное аналитически.

Для операции broadcast в сети с топологией 3D-тор выигрыш для 8 узлов (тор 2x2x2) составляет 2,39 раз, а для 8096 узлов (тор 16x16x32) — 6,97 раз. Для топологии 4D-тор выигрыш для 8096 узлов (тор 8x8x8x16) составляет 6,18 раз. Превосходство аппаратной поддержки коллективных операций обусловлено отсутствием накладных расходов на эжекцию/инжекцию пакетов. Сети с топологией 4D-тор обладают меньшим диаметром по сравнению с сетями 3D-тор при одинаковом количестве узлов, поэтому для сетей с топологией 4D-тор выигрыш меньше. Для операции all reduce разница в производительности примерно такая же. Таким образом, реализация операций с помощью коллективной подсети дает значительный выигрыш по сравнению с использованием операций «точка-точка».



а) Операция broadcast б) Операция all reduce

Рис. 7. Время выполнения операций (в микросекундах) а broadcast и б all reduce

3.2. Умножение разреженной матрицы на вектор

Операция умножения разреженной матрицы на вектор лежит в основе метода сопряженных градиентов CG. CG — метод нахождения локального минимума функции на основе информации о ее значениях и градиенте. Один из способов распараллеливания этой операции заключается в следующем [16]: пусть дана разреженная матрица A и вектор x, требуется найти $y = Ax$. Строки матрицы блоками равномерно распределяются по узлам:

$$A = (A_1, A_2, \ldots, A_{np})^T \Rightarrow Ax = (A_1x, A_2x, \ldots, A_{np}x)^T, \quad (1)$$

где A_i — строки матрицы A, хранящиеся на i-ом узле. При умножении строк матрицы на вектор получаются результирующие части искомого вектора y: $y_j = A_jx$, где $j = 1, 2, \ldots, np$, np — количество вычислительных узлов. Для сборки результирующего век-

тора y на каждом из вычислительных узлов для последующих вычислений понадобится коллективная операция all gather.

Для реализации all gather через операцию «точка-точка» используется алгоритмом «рекурсивное удвоение» [17]. Алгоритм заключается в следующем: сначала все узлы попарно обмениваются друг с другом, затем пары обмениваются попарно таким образом, что каждый узел пары обменивается данными с другим узлом пары и т.д. Схема работы алгоритма изображена на рис. 8.

Для реализации all gather через коллективные операции используется broadcast — все узлы рассылают свою часть вектора, инициируя операцию broadcast. Используются все 16 деревьев, корни которых равномерно распределены по системе.

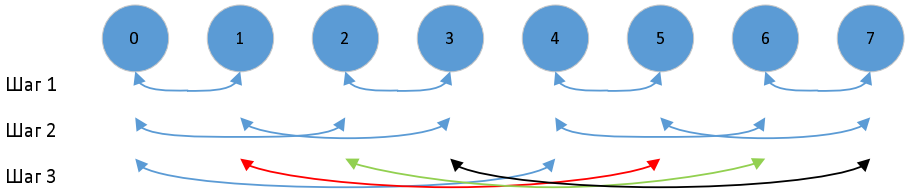


Рис. 8. Схема работы алгоритма «рекурсивное удвоение» для системы из 8 узлов

Для расчетов выбрана матрица с $d = 10$ ненулевыми элементами в строке и размером $N = 20480000$. Измерение времени умножения строк матрицы на плотный вектор проводилось на реальном процессоре Intel Xeon E5-2660 с использованием тестовой программы на C/OpenMP отдельно для каждого количества строк матрицы, попадающих на вычислительный узел при увеличении количества узлов сети и дроблении задачи. Производительность на одном узле составляет 591 Мфлопс. Время выполнения операции all gather получено с помощью имитационной модели сети.

Оценка точности имитационного моделирования проводилась для операции all gather, реализованной при помощи операций «точка-точка». Время выполнения all gather, полученное при помощи имитационной модели, сравнивалось со временем выполнения этой операции на тестовом кластере, параметры которого приведены в таблице, для 2, 4 и 8 узлов. Максимальная разница времен выполнения составляет 8,96 %.

На рис. 9 видно, что использование аппаратных коллективных операций дало максимальный прирост производительности 28 % на 2048 узлах для сети 4D-тор по сравнению с использованием коллективных операций, реализованных с помощью операций «точка-точка».

3.3. Задача с нелинейным уравнением теплопроводности

Рассматривается двухмерная нелинейная задача теплопроводности:

$$\frac{\partial u}{\partial t} = \sigma(x_1, x_2, t)\left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right) + f(x_1, x_2, t) \tag{2}$$

$$t\epsilon[0, t_k], x_1\epsilon[a, b], x_2\epsilon[c, d], \forall x_1, x_2, t : \sigma \geqslant 0 \tag{3}$$

Явная разностная схема:

$$\frac{U_{j,k}^{n+1} - U_{j,k}^n}{\Delta t} = \sigma_{j,k}^n \left(\frac{U_{j+1,k}^n - 2U_{j,k}^n + U_{j-1,k}^n}{h_j^2} + \frac{U_{j,k+1}^n - 2U_{j,k}^n + U_{j,k-1}^n}{h_k^2}\right) + f_{j,k}^h \tag{4}$$

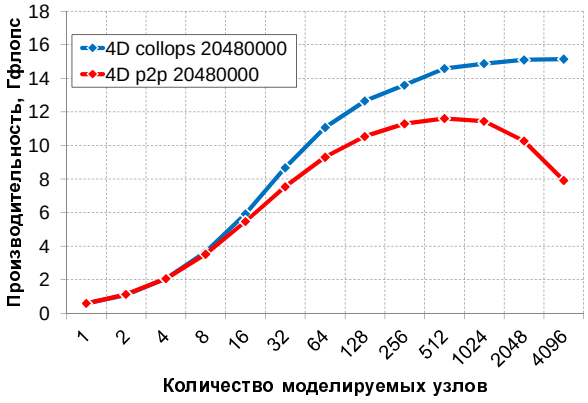


Рис. 9. Производительность задачи умножения разреженной матрицы на вектор ($N = 20480000$, $d = 10$) в зависимости от количества вычислительных узлов для сети с топологией 4D-тор

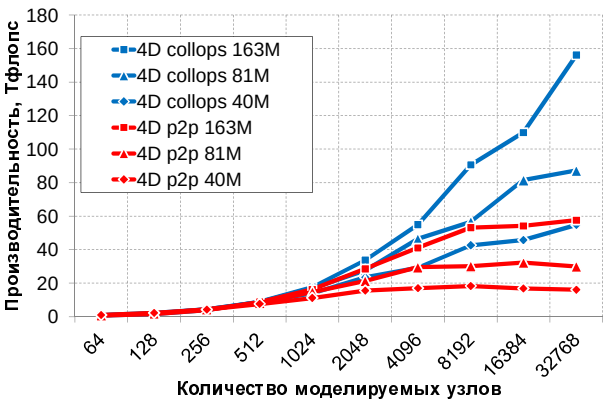


Рис. 10. Производительность двухмерной задачи теплопроводности на различных сетках в зависимости от количества вычислительных узлов

Условие устойчивости:

$$\Delta t \leqslant \frac{1}{2\sigma(x_1, x_2, t)\left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right)} \leqslant \frac{1}{2max[\sigma(x_1, x_2, t)]\left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right)} \tag{5}$$

Двухмерная область распределяется поровну по вычислительным узлам. В явной схеме для вычисления значения в ячейке на следующем слое по времени необходимы 9 значений ячеек с предыдущего слоя (см. рис. 11). Для этого на каждый узел требуется область больше на единицу в каждую сторону, чтобы была возможность рассчитать границы области. После каждого расчета области необходим обмен гранями между узлами и вычисление оптимального шага по времени. Для вычисления временного шага, требуется вычислить максимум коэффициента температуропроводности (усл. устойчивости 5). Вычисление максимума происходит при помощи коллективной операции all reduce. Общая схема параллельной реализации задачи представлена на рис. 12.

Производительность вычисляется по формуле: $\frac{N_{op}}{T_S + T_{max}} * 10^{-12}$ [Тфлопс], где N_{op} — количество операций с плавающей точкой. Рассматривались задачи с общим количеством ячеек, равным 40960000, 81920000 и 163840000 (на рис. 10 обозначены 40M, 81M и 163M соответственно). Время расчета области (T_S) измерялось на реальном процессоре Intel Xeon E5-2660 с использованием тестовой программы на C/OpenMP отдельно для каждого размера области, приходящейся на вычислительный узел при увеличении количества узлов сети и дроблении задачи. Производительность на одном узле составляет приблизительно 11 Гфлопс. Время вычисления максимума коэффициента температуропроводности (T_{max}) получено при помощи имитационного моделирования операции all reduce.

На данной задаче (см. рис. 10) аппаратная поддержка коллективных операций в сети 4D-тор сказывается, когда узлов в сети более 64. На большом числе вычислительных узлов (32768) производительность варианта с использованием подсети коллективных операций превышает производительность варианта с реализацией на основе операций «точка-точка» в 3 раза для задачи с 40M ячеек в сетке, для остальных размеров задач выигрыш составляет 2,7 раз.

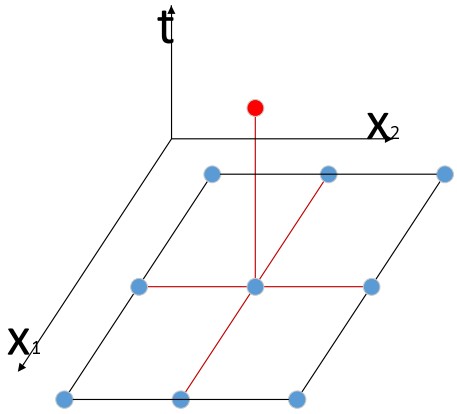


Рис. 11. Шаблон явной схемы двухмерной нелинейной задачи теплопроводности

Рис. 12. Схема параллельной реализации явной схемы для нелинейного уравнения теплопроводности.

Заключение

В данной работе описана реализация поддержки подсети коллективных операций в параллельной потактовой иммитационной модели сети «Ангара». С использованием имитационной модели получены демонстрационные оценки производительности базовых коллективных операций и небольших прикладных задач.

Оценки времени выполнения базовых коллективных операций reduce и broadcast показали значительный выигрыш (более 2-х раз для 8 узлов, более 6-ти раз для 8192 узлов) при использовании аппаратной подсети коллективных операций относительно программной реализации этих операций при помощи операции «точка-точка». С увеличением количества вычислительных узлов разница в производительности увеличивается.

Оценка производительности прикладных задач проводилась на имитационной модели на примере задачи умножения разреженной матрицы на вектор и задачи численного решения явной схемы нелинейного уравнения теплопроводности. Для задачи умножения матрицы на вектор выбраны следующие параметры: 20480000 — размер матрицы с 10 ненулевыми элементами в строке, задача теплопроводности рассматривалась на сетках с количеством ячеек 40960000, 81920000 и 163840000.

На задаче умножения разреженной матрицы на вектор, реализованной с помощью операции all gather, выигрыш от использования аппаратной поддержки коллективных операций по сравнению с реализацией на основе операции «точка-точка» небольшой, растет с увеличением количества вычислительных узлов и максимально составляет 28 % на 2048 узлах для сети с топологией 4D-тор.

На задаче численного решения явной схемы нелинейного уравнения теплопроводности на тысячах и десятках тысяч вычислительных узлов производительность варианта с использованием подсети коллективных операций превышает производительность варианта с реализацией на основе операций «точка-точка» от 2,7 до 3 раз. При увеличении количества вычислительных узлов разница в производительности растет.

Выполненная работа позволила получить оценки производительности тестов с использованием аппаратной поддержки коллективных операций в коммуникационной сети с то-

пологией «многомерный-тор» и показала возможность получения высокой производительности с использованием таких коллективных операций для выбранных задач. При этом выбраны конкретные размеры демонстрационных задач.

С использованием представленных в данной статье результатов планируется детальное исследование производительности коллективных операций с целью разработки новой архитектуры подсети коллективных операций для следующего поколения маршрутизатора сети «Ангара».

Литература

1. Макагон, Д.В. Сети для суперкомпьютеров / Д.В. Макагон, Е.Л. Сыромятников // Открытые системы. СУБД. — 2011. — № 7. — С. 33–37.
2. Корж, А.А. Отечественная коммуникационная сеть 3D-тор с поддержкой глобально адресуемой памяти для суперкомпьютеров транспетафлопсного уровня производительности / А.А. Корж, Д.В. Макагон, И.А. Жабин, Е.Л. Сыромятников // Паралельные вычислительные технологии (ПаВТ'2010): Труды международной научной конференции (Уфа, 29 марта – 2 апреля 2010 г.). — 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).
3. Симонов, А.С. Разработка межузловой коммуникационной сети с топологией «многомерный тор» и поддержкой глобально адресуемой памяти для перспективных отечественных суперкомпьютеров / А.С. Симонов, И.А. Жабин, Д.В. Макагон // Научно-техническая конференция «Перспективные направления развития вычислительной техники» (Москва, 28 июня). — Москва: ОАО «Концерн «Вега», 2011. — С. 17–19.
4. Эйсымонт, Л.К. Моделирование российского суперкомпьютера «Ангара» на суперкомпьютере / Л.К. Эйсымонт, А.С. Семенов, А.А. Соколов, А.С. Фролов, А.Б. Шворин // В сборнике «Суперкомпьютерные технологии в науке, образовании и промышленности» под редакцией академика В.А. Садовничего, академика Г.И. Савина, чл.-корр. РАН Вл.В. Воеводина. — Москва: Издательство Московского университета, 2009. — С. 145–150.
5. Симонов, А.С. Первое поколение высокоскоростной коммуникационной сети «Ангара» / А.С. Симонов, Д.В. Макагон, И.А. Жабин А.Н. Щербак, Е.Л. Сыромятников, Д.А. Поляков // Наукоемкие технологии. — 2014. — Т. 15, № 1. — С. 21–28.
6. Слуцкин, А.И. Разработка межузловой коммуникационной сети ЕС8430 «Ангара» для перспективных суперкомпьютеров / А.И. Слуцкин, А.С. Симонов, И.А. Жабин, Д.В. Макагон, Е.Л. Сыромятников // Успехи современной радиоэлектроники. — 2012. — № 1. — С. 6–10.
7. Жабин, И.А. Кристалл для Ангары / И.А. Жабин, Д.В. Макагон, А.С. Симонов // Суперкомпьютеры. — Зима-2013. — С. 46–49.
8. Макагон, Д.В. Реализация аппаратной поддержки коллективных операций в маршрутизаторе высокоскоростной коммуникационной сети с топологией «многомерный тор» / Д.В. Макагон, Е.Л. Сыромятников, С.И. Парута, А.А. Румянцев // Успехи современной радиоэлектроники. — 2012. — № 1. — С. 11–15.
9. Message Passing Interface Forum, MPI: A Message-Passing Interface Standard, 1995.

URL: http://www.mpi-forum.org/docs/mpi-1.1/mpi-11-html/node64.html (дата обращения: 29.04.2015).

10. Feind, K. Shared Memory Access (SHMEM) Routines. Cray Research, 1995. / K. Feind. URL: https://cug.org/5-publications/proceedings_attendee_lists/1997CD/S95PROC/303_308.PDF (дата обращения: 29.04.2015).
11. Wiebel, F. UPC Collective Operations Specifications. — 2003. / E. Wiebel, D. Greenberg, S. Seidel. URL: http://upc.gwu.edu/docs/UPC_Coll_Spec_V1.0.pdf (дата обращения: 29.04.2015).
12. Saraswat, V. X10 Language Specification. — 2011. / V. Saraswat, B. Bloom, I. Peshansky, O. Tardieu, D. Grove. URL: http://dist.codehaus.org/x10/documentation/languagespec/x10-latest.pdf (дата обращения: 29.04.2015).
13. Fox, G. Solving Problems on Concurrent Processors / G. Fox, M. Johnson, G. Lyzenga, S. Otto, J. Salmon, D. Walker // General techniques and regular problems. — V. 1, — Prentice-Hall Inc., 1998. — P. 592.
14. Bala, V. CCL: a Portable and Tunable Collective Communication Library for Scalable Parallel Computers / V. Bala, J. Bruck, R. Cypher, P. Elustondo, H. Ching-Tien, S. Kipnis, M. Snir // Parallel and Distributed System — 1995. — V. 6, — P. 154–164. DOI: 10.1109/71.342126.
15. Almasi, G. Efficient Implementation of Allreduce on BlueGene/L Collective Network / G. Almasi, G. Dozsa, C. Erway, B. Steinmacher-Burow // Recent Advances in Parallel Virtual Machine and Message Passing Interface. — 2005. — P.57–66. DOI: 10.1007/11557265_12.
16. Пожилов, И.А. Прогнозирование масштабируемости задачи умножения разреженной матрицы на вектор при помощи модели коммуникационной сети / И.А. Пожилов, А.С. Семенов, Д.В. Макагон // Вестник УГАТУ. — 2012. — Т. 16, № 6 (51). — С. 158–163.
17. Thakur, R. Optimization of Collective Communication Operations in MPICH. / R. Thakur, R. Rabenseifner, W. Gropp. URL: http://www.mcs.anl.gov/~thakur/papers/ijhpca-coll.pdf (дата обращения: 29.04.2015).

Мукосей Анатолий Викторович, ОАО «НИЦЭВТ» (Москва, Российская Федерация), mukav@mail.ru.

Семенов Александр Сергеевич, к.т.н., ОАО«НИЦЭВТ» (Москва, Российская Федерация), alxdr.semenov@gmail.com.

Симонов Алексей Сергеевич к.т.н., с.н.с., ОАО«НИЦЭВТ» (Москва, Российская Федерация), simonov@nicevt.ru.

Поступила в редакцию 10 апреля 2015 г.

Bulletin of the South Ural State University
Series "Computational Mathematics and Software Engineering"
2015, vol. 4, no. 3, pp. 40–55

DOI: 10.14529/cmse150304

SIMULATION OF COLLECTIVE OPERATIONS HARDWARE SUPPORT FOR «ANGARA» INTERCONNECT

A.V. Mukosey, JSC «NICEVT» (Moscow, Russian Federation) mukav@mail.ru,
A.S. Semenov, JSC «NICEVT» (Moscow, Russian Federation)
alxdr.semenov@gmail.com,
A.S. Simonov, JSC «NICEVT» (Moscow, Russian Federation) simonov@nicevt.ru

JSC NICEVT develops the Angara high-speed interconnect with multi-dimensional torus topology. To evaluate the performance of the interconnect on a large number of nodes a cycle-accurate simulator is used. Angara interconnect supports two types of collective operations: broadcast and reduce. The paper describes the implementation of collective operations in the simulator and presents an early performance evaluation. Performance benchmarks include some basic broadcast and all-reduce tests, as well as several well-known computational applications, specificaly, sparse matrix-vector multiplication and numerical solution of the nonlinear heat conduction equation.

Keywords: Keywords: simulation, Angara inteconnect, multi-dimensional torus, collectives operations.

References

1. Makagon D.V., Syromyatnikov E.L. Seti dlya superkomp'yuterov [Supercomputers Interconnect]. Otkrytyye sistemy. SUBD. [Open Systems. DBMS]. 2011. No. 7. P. 33–37.
2. 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).
3. 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

4. Eysymont L.K., Semenov A.S., Sokolov A.A., Frolov A.S., Shvorin A.B. Modelirovaniye rossiyskogo superkomp'yutera «Angara» na superkomp'yutere [Simulation of Angara Russian Supercomputer on the Supercomputer]. V sbornike «Superkomp'yuternyye tekhnologii v nauke, obrazovanii i promyshlennosti» pod redaktsiyey akademika V.A. Sadovnichego, akademika G.I. Savina, chl.-korr. RAN Vl.V. Voyevodina [Edited by Member of RAS V.A. Sadovnichy, Member of RAS G.I. Savin, Corresponding member of RAS V.V.Voevodin]. Moscow: Publishing of Moscow State University, 2009. P. 145–150.
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. Vol. 15, No. 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. No. 1. P. 6–10.
7. Zhabin I.A. Kristall dlya Angary [Angara Chip] / I.A. Zhabin, D.V. Makagon, A.S. Simonov Superkomp'yutery [Supercomputers]. Winter-2013. P. 46–49.
8. Makagon D.V., Syromyatnikov E.L., Paruta S.I., Rumyantsev A.A. Realizatsiya apparatnoy podderzhki kollektivnykh operatsiy v marshrutizatore vysokoskorostnoy kommunikatsionnoy seti s topologiyey «mnogomernyy tor» [The Implementation of Collective Operations Hardware Support in High Speed Interconnect with Multi-Dimensional Torus Topology]. Uspekhi sovremennoy radioelektroniki [Progress of the Modern Radioelectronics]. 2012. No. 1. P. 11–15.
9. Message Passing Interface Forum, MPI: A Message-Passing Interface Standard. 1995. URL: `http://www.mpi-forum.org/docs/mpi-1.1/mpi-11-html/node64.html` (accessed: 29.04.2015).
10. Feind K. Shared Memory Access (SHMEM) Routines. Cray Research. 1995. URL: `https://cug.org/5-publications/proceedings_attendee_lists/1997CD/S95PROC/303_308.PDF` (accessed: 29.04.2015).
11. Wiebe F., Wiebel E., Greenberg D., Seide S. UPC Collective Operations Specifications. 2003. URL: `http://upc.gwu.edu/docs/UPC_Coll_Spec_V1.0.pdf` (accessed: 29.04.2015).
12. Saraswat V., Bloom B., Peshansky I., Tardieu O., Grove D. X10 Language Specification. 2011. URL: `http://dist.codehaus.org/x10/documentation/languagespec/x10-latest.pdf` (accessed: 29.04.2015).
13. Fox G., Johnson M., Lyzenga G., Otto S., Salmon J., Walker D. Solving Problems on Concurrent Processors // General techniques and regular problems. Vol. 1, Prentice-Hall Inc., 1998. P. 592.
14. Bala V., Bruck J., Cypher R., Elustondo P., Ching-Tien H., Kipnis S., Snir M. CCL: a Portable and Tunable Collective Communication Library for Scalable Parallel Computers // Parallel and Distributed System. 1995. Vol. 6, P. 154–164. DOI: 10.1109/71.342126.
15. Almasi G., Dozsa G., Erway C., Steinmacher-Burow B. Efficient Implementation of Allreduce on BlueGene/L Collective Network // Recent Advances in Parallel Virtual Machine and Message Passing Interface. 2005. P. 57–66 DOI: 10.1007/11557265_12.
16. Pozhilov I.A., Semenov A.S., Makagon D.V. Prognozirovaniye masshtabiruyemosti zadachi

umnozheniya razrezhennoy matritsy na vektor pri pomoshchi modeli kommunikatsionnoy seti [Scalability Prediction of the Sparse Matrix-Vector Multiplication Using the Interconnection Network Simulator]. Vestnik UGATU. 2012. Vol. 16, No. 6 (51). P. 158–163.

17. Thakur R., Rabenseifner R., Gropp W. Optimization of Collective Communication Operations in MPICH. URL: `http://www.mcs.anl.gov/~thakur/papers/ijhpca-coll.pdf` (accessed: 29.04.2015).

Received April 10, 2015.