Исследование масштабируемости вычислений в программном комплексе FlowVision на суперкомпьютерах Ломоносов и Ломоносов-2*

В.С. Акимов, А.А. Ющенко

ООО «ТЕСИС»

В данной работе исследуется масштабируемость вычислений задач гидро- газодинамики в программном комплексе FlowVision на суперкомпьютерах Ломоносов и Ломоносов-2. Представлены результаты масштабируемости вычислений по количеству ядер одного процессора для задачи, имеющей около 0,4млн. расчетных ячеек, а также масштабируемости по процессорам задачи с числом ячеек более 5млн. Проведен сравнительный анализ результатов, полученных при использовании двух суперкомпьютеров Ломоносов и Ломоносов-2. Получены кривые относительных затрат времени на процессы MPI-обмена при вычислениях на различном количестве процессоров на каждом из суперкомпьютеров. Даны рекомендации, обеспечивающие максимально эффективное использование вычислительных ресурсов современного суперкомпьютера Ломоносов-2 при решении задач гидро- газодинамики.

1. Введение

Развитие вычислительной техники происходит ежедневно, производители предлагают все более и более совершенные устройства, а вычислительные центры оборудуются с применением более современных технологий. С одной стороны растет количество ядер процессоров, с другой – расширяется шина памяти и совершенствуется интерконнект между процессорами (табл.1). Таким образом, перед инженерами компаний, занимающихся инсталляцией суперкомпьютерных комплексов, стоит нелегкая задача: обеспечить максимальную возможность полного раскрытия потенциала современной вычислительной техники в рамках многопроцессорного кластера. Тем временем конечный результат оценивается производительностью вычислений и экономической целесообразностью.

Одним из наиболее распространенных вариантов использования мощностей суперкомпьютеров являются инженерные расчеты в области гидро- газодинамики. Со стороны пользователей CFD-кодов спрос на повышение производительности вычислений всегда будет актуальным. Спектр решаемых задач давно вышел за пределы однопроцессорных вычислений, поэтому скорость счета, в основном, определяется возможностью многократно ускорять расчет посредством использования большого количества ядер и процессоров. Такая возможность называется масштабируемостью вычислений и зависит, прежде всего, от характеристик памяти, интерконнекта и умения программного кода этим пользоваться.

2. Исследования масштабируемости

Исследования масштабируемости вычислений проводились посредством расчетов тестовых задач в программном комплексе (ПК) FlowVision на суперкомпьютере Ломоносов-2 - одном из самых современных в мире и первом в мире, использующем топологию Flattened Butterfly на Infiniband. Также проводились сравнения с результатами, полученными на суперкомпьютере Ломоносов. Основные характеристики вычислительных мощностей приведены в таблице 1.

* Работа выполнена с использованием ресурсов суперкомпьютерного комплекса МГУ имени М.В. Ломоносова

Таблица 1. Технические характеристики суперкомпьютеров

Суперкомпьютер	Ломоносов (раздел gpu)	Ломоносов-2 (раздел compute)
Процессор	Intel(R) Xeon(R) CPU E5630 @ 2.53GHz	Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80GHz
Количество физических ядер процессора	4	10
Количество логических ядер при использовании Hyper-Threading (HT)	8	20
Кэш-память, МБ	12	25
Максимальная пропускная способность памяти, Гб/с	25,6	59,7
Количество процессоров на узел	2	1
Топология, интерконнект	QDR InfiniBand	Flattened Butterfly на InfiniBand

2.1 Тестовая задача

В качестве тестовой задачи было выбрано моделирование процесса смешивания холодной и горячей воды в смесителе (рис.1). Основные особенности задачи сведены в таблицу 2.

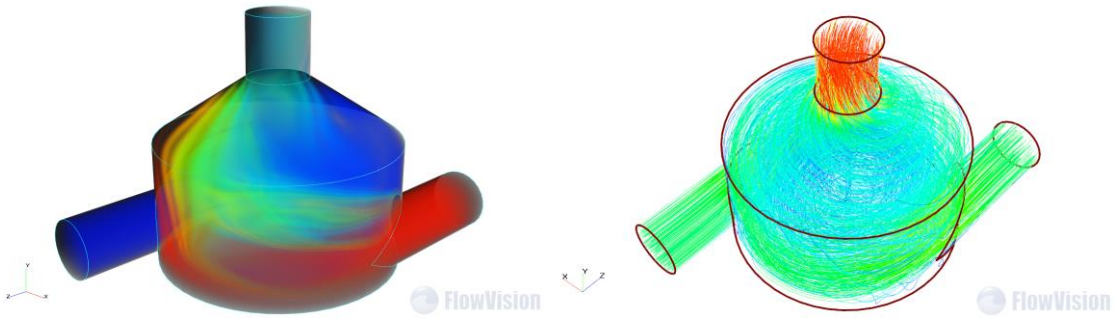


Рис.1. Задача смешивания горячей и холодной воды в смесителе

Таблица 2. Характеристики тестовой задачи

Задача смешивания горячей и холодной воды в смесителе	
Постановка	Трехмерная
Моделируемые физические явления	Теплоперенос (конвекция и теплопроводность), движение, турбулентность
Количество ячеек расчетной сетки, шт.	Малая размерность: 410108; большая размерность: 5172649
Адаптация расчетной сетки	отсутствует

Для наилучшей равномерности распределения нагрузки по процессорам расчетная сетка равномерна по всем направлениям и локальные сгущения отсутствуют. На рис.2 представлена визуализация распределения сетки с количеством ячеек около 5,2 млн. по 15 процессорам, а на рис.3 – по гиперячейкам для режима запуска на 15 процессоров по 20 нитей. В среднем, на одну нить в данном случае приходится по 12 гиперячеек.

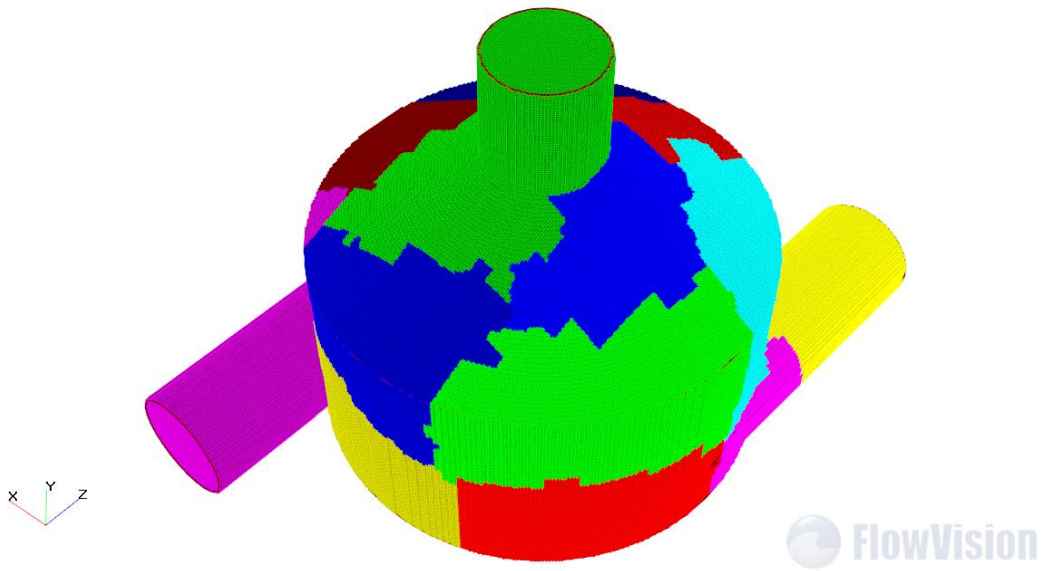


Рис. 2. Распределение расчетной сетки по процессорам

Рис. 3. Структурирование расчетной сетки по гиперячейкам

2.2 Масштабируемость по потокам

2.2.1 Однопроцессорный запуск

На первом этапе были проведены исследования скорости вычисления и масштабируемости по количеству потоков в пределах одного вычислительного процессора. Количество элементов расчетной сетки модели на этом этапе исследований составляет 410108 ячеек. При расчете данной задачи вспомогательные операции по построению сетки, её распределению, набору статистики и т.п. происходят в течение первых трех шагов, начиная с 4-го шага, время счета i-го шага незначительно изменяется от запуска к запуску. Поэтому, в качестве критерия скорости вычисления было выбрано время вычисления 5-го шага по времени. На рис.4. представлена зависимость относительного ускорения вычислений пятого шага при использовании различного кол-ва потоков, иными словами масштабируемость по потокам.

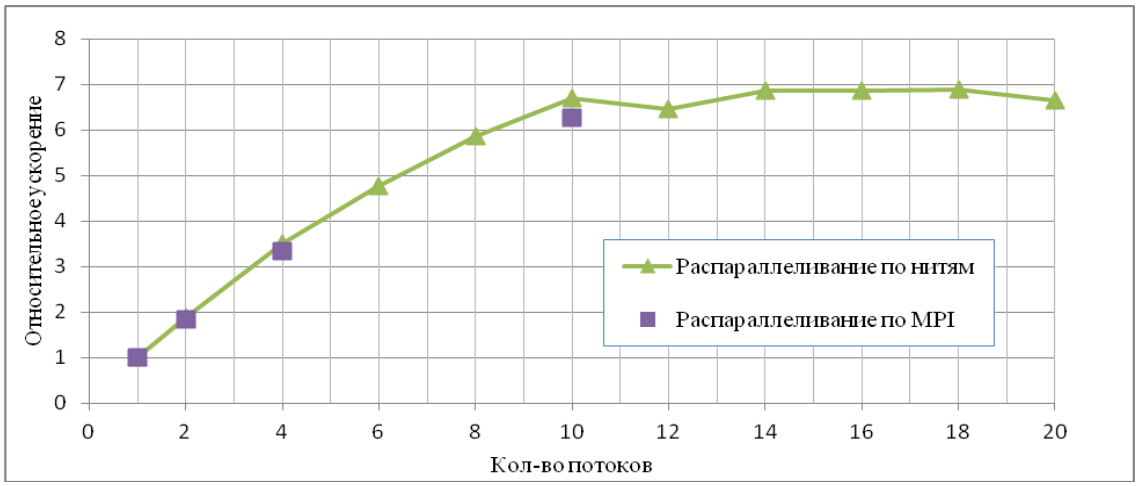
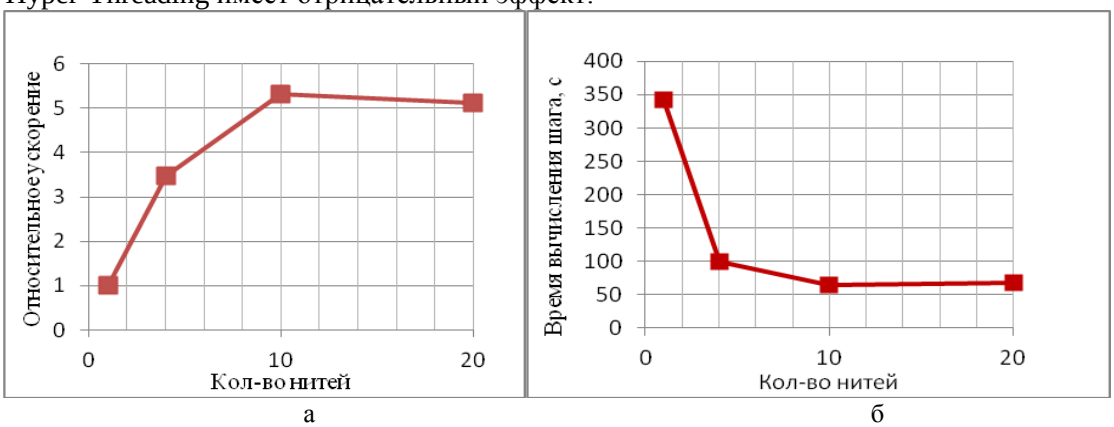


Рис.4. Масштабируемость вычислений по количеству потоков (нити, MPI-процессы) одного процессора. 410108 ячеек

Из результатов, представленных на рис.4 можно видеть, что увеличение количества используемых потоков в пределах кол-ва физических ядер (10шт на процессор, см. табл.1) дает значительное ускорение вычислений, а использование дополнительных потоков за счет логических ядер Hyper-Threading не дает значительного прироста в ускорении вычислений. При этом для данной задачи максимум кривой ускорения наблюдается при использовании 16 потоков, что соответствует количеству ячеек на поток около 26 тысяч. Стоит отметить, что прирост скорости вычислений при использовании 16 потоков составляет всего 5% относительно 10 потоков, в то время как использование более 18 приводит к замедлению скорости вычислений. Использование распараллеливания вычислений внутри процессора по MPI-процессам ожидаемо оказывается немного менее выигрышным, чем по нитям. Использование MPI на логических ядрах Hyper-Threading очевидно даст худшую масштабируемость, т.к. распараллеливание по MPI менее эффективно, чем по нитям. Продемонстрировать это не удалось, т.к. программное обеспечение кластера блокирует запуск более чем десяти MPI на процессор.

2.2.1 Восьмипроцессорный запуск

Далее проводились исследования для той же задачи, но с количеством расчетных ячеек равным 5172649 при запуске на 8 процессоров с разным кол-вом потоков. В этом случае, между процессорами применялось распараллеливание по MPI, а внутри процессора (между ядрами) - по нитям. Так же, как и на предыдущем этапе, данные по скорости вычисления снимались с 5-го шага по времени. Из представленных на рис.5. кривых масштабируемости видно, что оптимальным является использование всех физических ядер, а использование в полной мере технологии Hyper-Threading имеет отрицательный эффект.

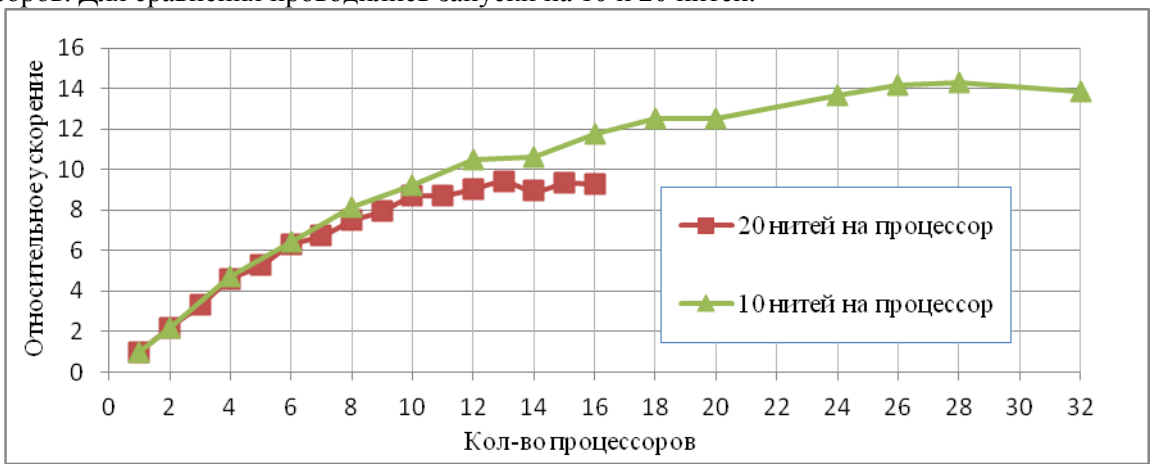


а б

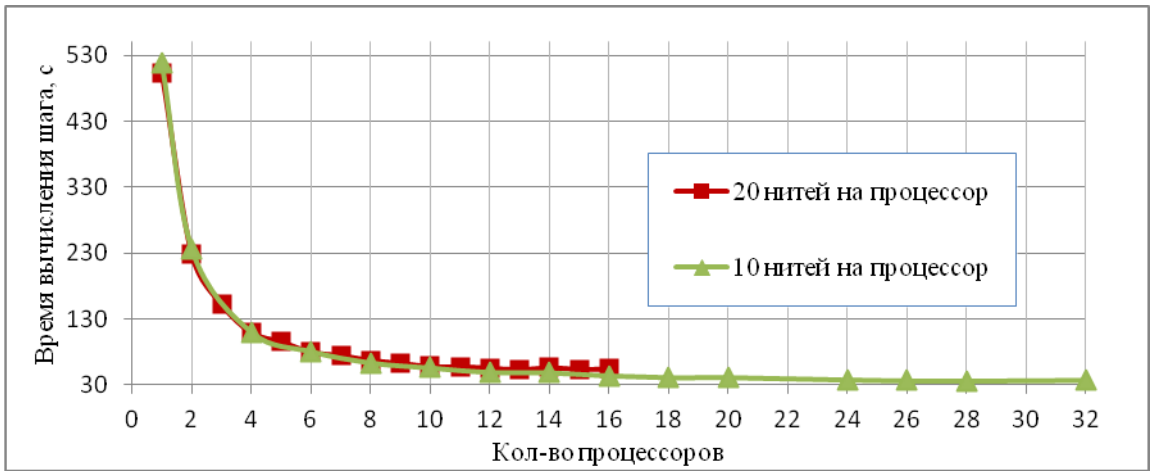
Рис.5. Масштабируемость вычислений по кол-ву нитей при запуске на 8 процессоров. 5172649 ячеек.
а – относительное ускорение; б – время вычисления шага

2.3 Масштабируемость по процессорам

На следующем этапе исследовалось ускорение вычислений в зависимости от кол-ва процессоров. Для сравнения проводились запуски на 10 и 20 нитей.



а



б

Рис.6. Масштабируемость вычислений по количеству процессоров. Запуски по 10 и по 20 нитей на процессор. 5172649 ячеек.
а – относительное ускорение; б – время вычисления шага

Из результатов, представленных на рис. 6 можно видеть, что в случае 20 нитей на процессор кривая масштабируемости выходит на “полку” уже при использовании более 12 процессоров, в то время как использование 10 физических ядер открывает возможности для значительного ускорения вычислений. Максимум ускорения при этом наблюдается в случае использования 28 процессоров, что соответствует 18,5 тыс. ячеек на нить, а время вычисления шага в этом случае составляет 36.199с. Такие различия между запусками по 10 и по 20 нитей на процессор связаны с перегруженностью шины памяти во втором случае, ведь она используется вдвое большим количеством потоков.

Для исключения возможности влияния неравномерной загрузки процессоров на результаты исследований были проведены тесты с включенной динамической балансировкой вычислений между процессорами. Включение этой опции может значительно ускорить вычисления в случае неравномерной загрузки ядер (неравномерная сетка, сложная геометрия, адаптация сетки по решению и пр.), но замедляет первые шаги расчета, так как на них происходит набор статистики и прочие вспомогательные операции. Поэтому, в данном случае, результаты снимались с 22-го шага по времени. На рис.7. представлено сравнение времени вычисления 22го шага с включенной и выключенной динамической балансировкой, а на рис. 8 – полное время счета 22х шагов для обоих случаев.

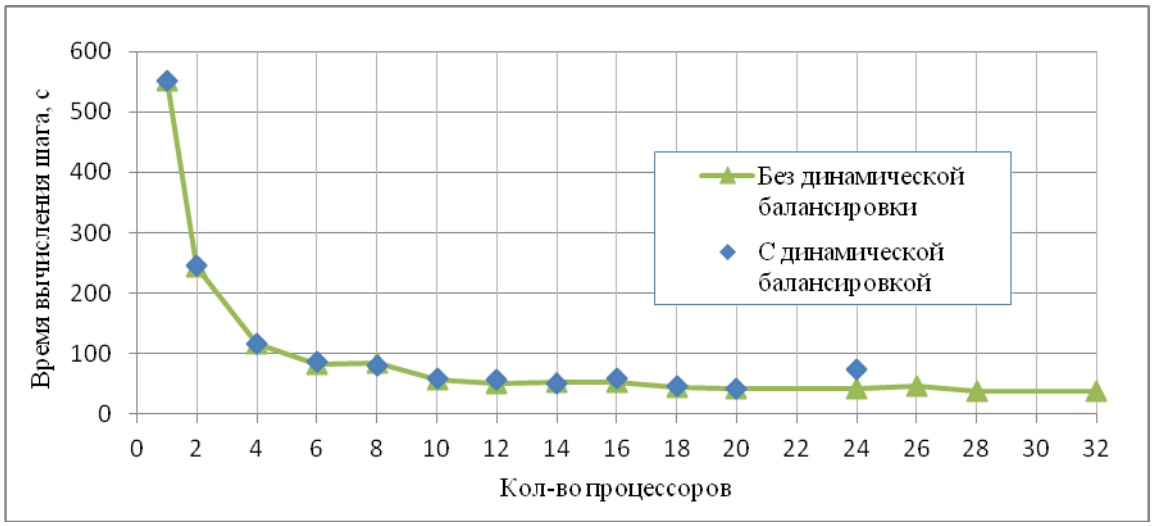


Рис. 7. Время вычисления шага при включенной и выключенной динамической балансировке. 5172649 ячеек.

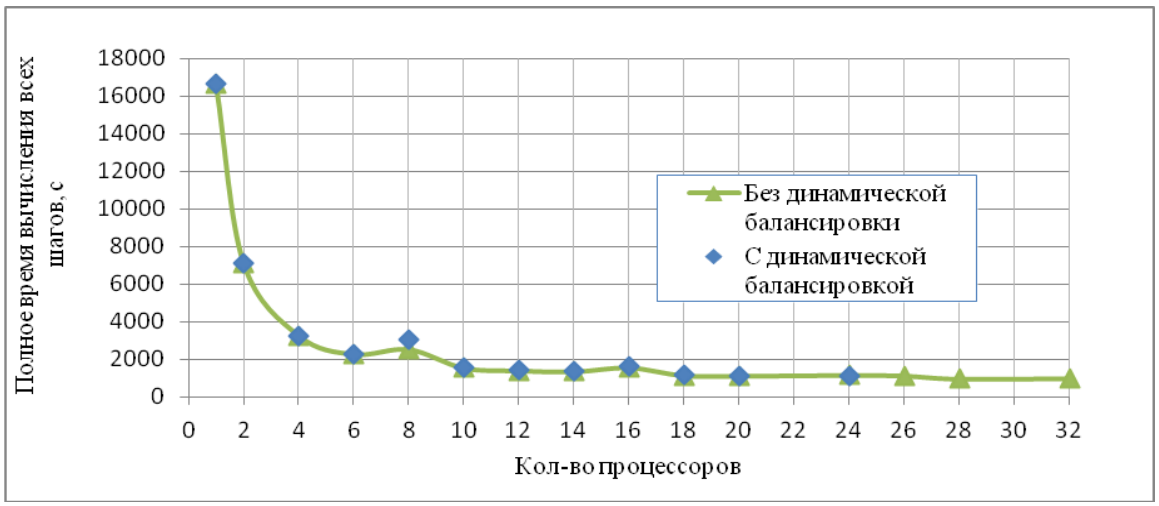


Рис. 8. Полное время вычисления при включенной и выключенной динамической балансировке. 5172649 ячеек.

Выше было отмечено, что сама постановка задачи и равномерная сетка должны обеспечить хорошую равномерность загрузки ядер, это хорошо подтверждается результатами, представленными на рисунках 7 и 8: балансировка не дает преимуществ в данном конкретном случае.

2.4 Сравнительный анализ масштабируемости на суперкомпьютерах Ломоносов и Ломносов-2

В рамках исследований были также произведены сравнения между суперкомпьютерами Ломоносов и Ломоносов-2. Совершенно очевидна разница между уровнем производительности процессоров, используемых на суперкомпьютере Ломоносов и Ломоносов-2, к тому же они имеют различное количество ядер. Поэтому время вычисления шага сравнивать нецелесообразно, но имеет смысл сравнить кривые ускорения. На рис.9 представлены кривые масштабируемости, полученные на двух суперкомпьютерах при запусках на различном количестве процессоров в режиме использования всех потоков, включая потоки на логические ядра Hyper-Threading. На рис. 10 представлены результаты по времени вычисления шага при тех же запусках, но по оси абсцисс отложено кол-во всех используемых нитей (потоков).

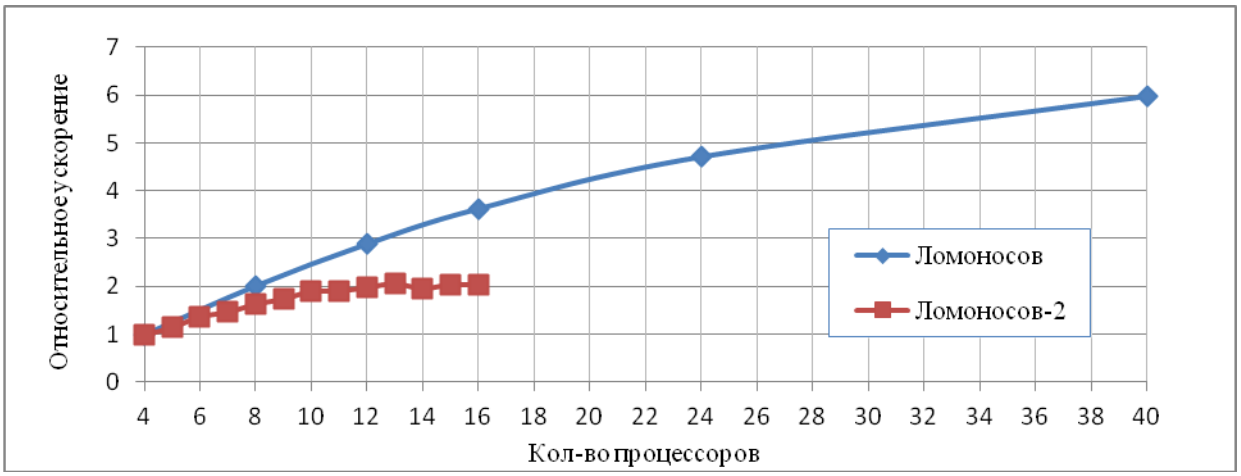


Рис. 9. Сравнение масштабируемости по процессорам на двух суперкомпьютерах. Используются все ядра, включая логические ядра Hyper-Threading. 5172649 ячеек.

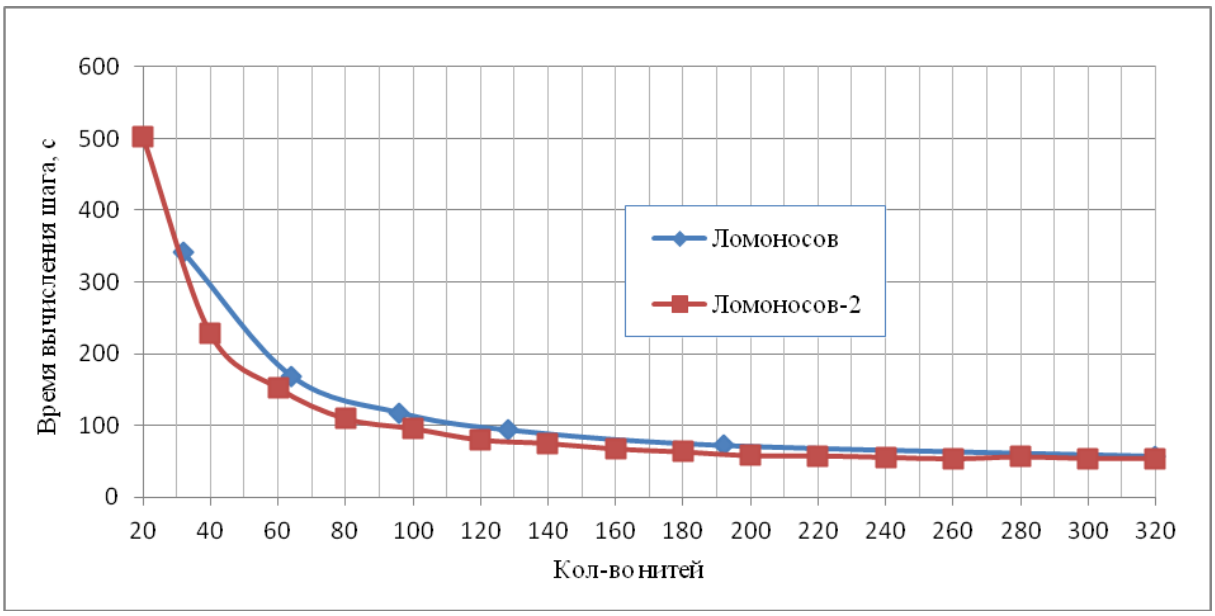
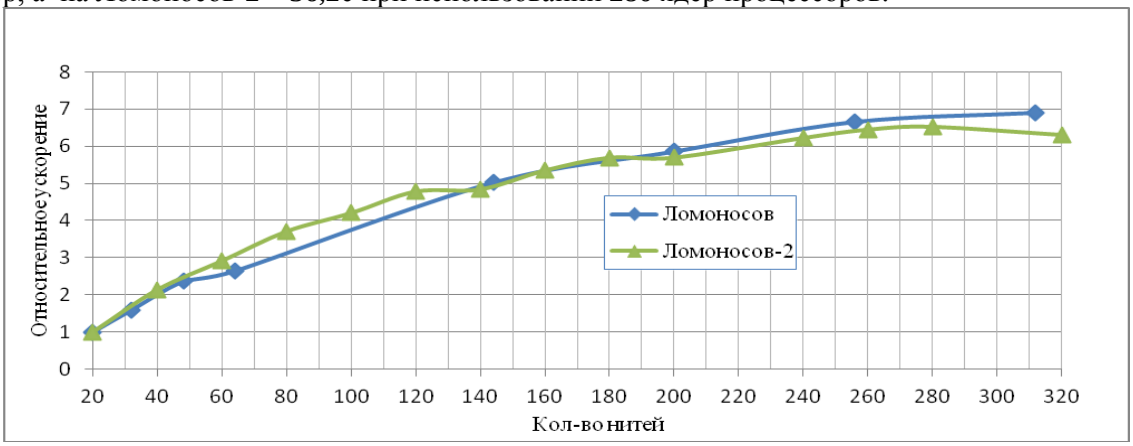


Рис. 10. Сравнение времени вычисления шага на двух суперкомпьютерах. Используются все ядра, включая логические ядра Hyper-Threading. 5172649 ячеек.

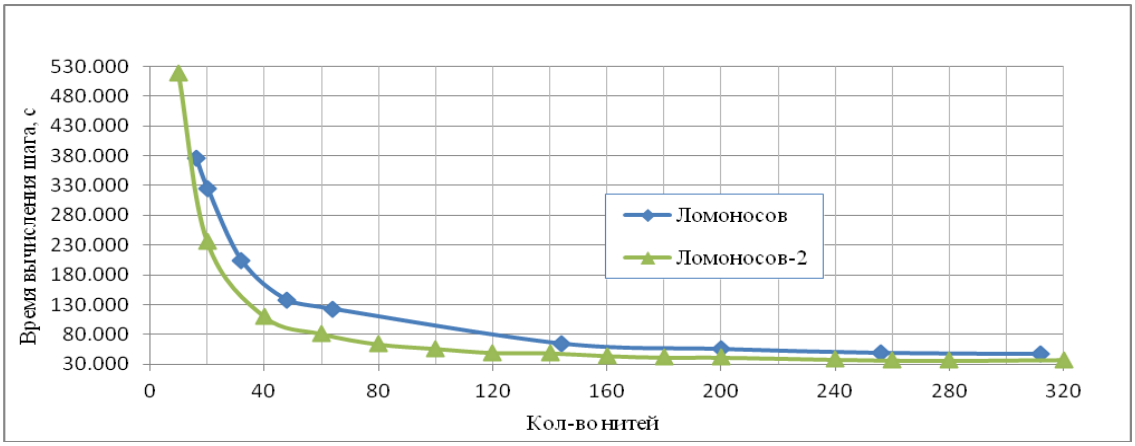
Из результатов, представленных на рис. 9 можно видеть, что ускорение происходит интенсивнее на суперкомпьютере Ломоносов. Этому можно найти объяснение, если сравнить количество ядер, объем Кэш-памяти и производительность шины памяти двух типов процессоров, используемых на этих суперкомпьютерах. Количество ядер отличается в 2,5 раза, объем Кэш-памяти в 2,08, а пропускная способность шины в 2,33 (см. табл.1). Таким образом, при современных тенденциях развития вычислительной техники скорость вычислений на CPU растет быстрее, чем скорость обменных операций (сетевых операций) (см. подробности ниже). По данным, представленным на рис. 10 видно, что при использовании большого количества потоков время вычисления на суперкомпьютере Ломоносов становится близким к времени вычисления на Ломоносов-2, а при 320 потоках отношение времен вычисления уже близко к единице: 57.1115с/54.2057с.

Далее были проведены сравнения масштабируемости вычислений на двух кластерах при запусках только на физические ядра, то есть количество нитей на процессор соответствовало количеству физических ядер: по 4 на процессор для суперкомпьютера Ломоносов и по 10 для суперкомпьютера Ломоносов-2. Таким образом, логические ядра Hyper-Threading при данном сравнении не использовались, хотя физически HT при этом не отключался. Ускорение в данном случае высчитывалось относительно времени вычисления шага при использовании 20 ядер и представлено на рис. 11а. В целом, на рис.11а можно видеть идентичное поведение кривых,

полученных для двух суперкомпьютеров, хотя в случае Ломоносов-2 кривая несколько раньше начинает загибаться (более 280 нитей). Этот эффект является также следствием соотношения характеристик процессоров, установленных на суперкомпьютеры, о чем написано выше. На рис. 11б представлено сравнение результатов по времени вычисления шага. Наименьшее время вычисления шага на суперкомпьютере Ломоносов составляет 47,02с при использовании 312 ядер, а на Ломоносов-2 – 36,2с при использовании 280 ядер процессоров.



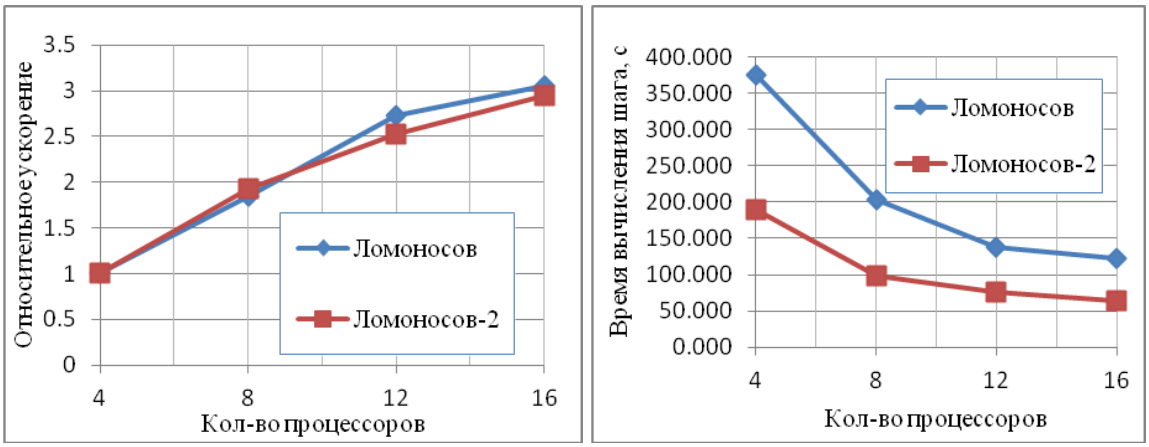
а



б

Рис.11. Сравнение масштабируемости по процессорам на двух суперкомпьютерах. Логические ядра HT не используются. 5172649 ячеек.
а – относительное ускорение; б – время вычисления шага

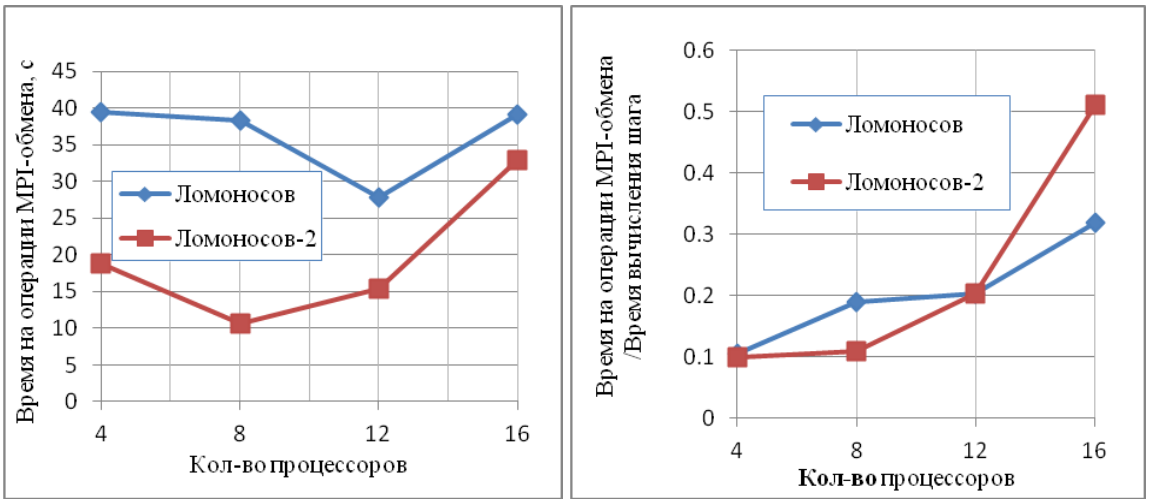
Пожалуй, наиболее корректным сравнением масштабируемости вычислений на двух различных суперкомпьютерах с различными процессорами и топологией, будет сравнение ускорения вычислений при запуске на одинаковое кол-во физических ядер. Так как процессора раздела gpu суперкомпьютера Ломоносов имеют 4 ядра (см. табл. 1), то запуски в данном сравнении производились в режиме по 4 нити на каждый процессор (4x4, 8x4, 12x4, 16x4) на обеих вычислительных машинах. Из результатов, представленных на рис. 12а можно видеть, что ускорение вычислений выглядит идентичным образом для обоих суперкомпьютеров, хотя время счета значительно ниже в случае использования Ломоносов-2 (рис.12б).



а б

Рис. 12. Сравнение масштабируемости по процессорам на двух суперкомпьютерах. Используются 4 ядра. 5172649 ячеек.
а – относительное ускорение; б – время вычисления

На рис. 13а представлены результаты по затратам времени на процессы MPI-обмена. Результаты снимались с 5-го шага по времени и осреднялись по количеству задействованных процессоров. Можно видеть, что эти временные затраты ниже в случае использовании суперкомпьютера Ломоносов-2, особенно при малопроцессорных запусках. Это является следствием более совершенной топологии суперкомпьютера Ломоносов-2. С другой стороны, из результатов, представленных на рис. 13б видно, что относительные затраты времени на MPI-обмен растут более интенсивно при использовании Ломоносов-2. Выше было отмечено, что, в плане исторического развития вычислительной техники, скорость вычислений на ядрах процессоров за счет увеличение числа ядер и тактовой частоты растет быстрее, чем скорость обменных операций. Поэтому доля времени, затраченного на процессы MPI-обмена, становится большей, особенно при многопоточных запусках на современных процессорах, хотя общее время вычислений значительно ниже (рис.12б). Например, в режиме запуска расчета на 16 процессоров по 4 нити отношение времени вычисления шага на суперкомпьютере Ломоносов ко времени вычисления на Ломоносов-2 составляет: 122,721с/ 64,4123с = 1,914.



а б

Рис. 13. Сравнение затрат времени на процессы MPI-обмена на двух суперкомпьютерах. Используются 4 ядра. 5172649 ячеек.
а – среднее время, затраченное на процессы MPI- обмена; б – относительные затраты времени на операции MPI-обмена

Проведенные исследования, к сожалению, не дают возможности раскрыть все преимущества современнейшей топологии, для этого необходимо запускать задачи с много большим количеством расчетных ячеек на значительно большее число процессоров. Исходя из проделан-

ных исследований, можно прогнозировать, что в режимах запуска на более чем 100 процессоров без использования логических ядер Hyper-Threading преимущества суперкомпьютера Ломоносов-2 будут очень значительными. Подобное исследование планируется к проведению в скором времени.

3. Выводы

1. Скорость вычислений на суперкомпьютере Ломоносов-2 значительно выше, чем на суперкомпьютере Ломоносов, особенно при использовании одинакового количества физических ядер на обоих кластерах. В режиме запуска на 16 процессоров по 4 нити время счета на Ломоносов-2 почти в 2 раза меньше. При сравнении скорости вычислений на оптимальных режимах запуска на обоих кластерах преимущество Ломоносов-2 составляет 30%, причем при меньшем количестве используемых ядер.
2. Для задач с количеством ячеек более 1млн оптимальным, с точки зрения скорости вычислений, можно считать гибридное распараллеливание по следующему алгоритму: Между процессорами по MPI (1 MPI-процесс на 1 процессор), а между ядрами процессора по нитям (количество нитей на процессор соответствует числу физических ядер). Оптимальное количество расчетных ячеек на ядро около 15-40 тысяч, но может значительно отличаться в зависимости от постановки задачи, используемых моделей и характеристик процессоров.
3. С точки зрения гидродинамических расчетов применение технологии Hyper-Threading на суперкомпьютере не оправдано и от неё лучше отказываться. Использование HT зачастую приводит к перегрузке шины памяти и, как следствие, замедлению времени вычислений. Получение преимущества от использования потоков HT требует специального исследования со стороны пользователя, требует определенного опыта работы с суперкомпьютерами. При этом даже если определить оптимальный режим запуска, преимущество использования HT раскрывается только при малопроцессорных запусках и не превышает 15%, а возможность получить замедление расчета достаточно велика.
4. При использовании всех ядер процессора, включая логические ядра HT, масштабируемость вычислений на суперкомпьютере Ломоносов лучше, чем масштабируемость вычислений на Ломоносов-2.
5. При использовании только физических ядер процессора масштабируемость вычислений на двух суперкомпьютерах выглядит идентично
6. Относительные затраты времени на MPI-обмен растут с увеличением кол-ва процессоров более интенсивно при использовании суперкомпьютера Ломоносов-2, чем при использовании Ломоносов. Это связано с тем, что скорости движения данных по каналам памяти и интерконнекту суперкомпьютера Ломоносов-2 не настолько выше в сравнении с Ломоносов, насколько выше вычислительная производительность.
7. Топология суперкомпьютера Ломоносов-2 выглядит более выигрышной, однако, для раскрытия всего потенциала современнейшей технологии необходимо использовать большое количество процессоров и задачу с большим количеством ячеек.

Литература

1. Харченко С.А. Влияние распараллеливания вычислений с поверхностными межпроцессорными границами на масштабируемость параллельного итерационного алгоритма решения систем линейных уравнений на примере уравнений вычислительной гидродинамики. Материалы международной научной конференции "Параллельные вычислительные технологии" (ПаВТ'2008), Санкт-Петербург, 28 января – 1февраля 2008 г. Челябинск, Изд. ЮУрГУ, 2008, с. 494-499.

2. Сушко Г.Б., Харченко С.А. Многопоточная параллельная реализация итерационного алгоритма решения систем линейных уравнений с динамическим распределением нагрузки по нитям вычислений. Труды международной научной конференции "Параллельные вычислительные технологии" (ПаВТ'2008), Санкт-Петербург, 28 января – 1 февраля 2008 г. Челябинск, Изд. ЮУрГУ, 2008, с. 452-457.

3. Воеводин Вл.В., Жуматий С.А., Соболев С.И., Антонов А.С., Брызгалов П.А., Никитенко Д.А., Стефанов К.С., Воеводин Вад.В. Практика суперкомпьютера "Ломоносов" // Открытые системы. - Москва: Издательский дом "Открытые системы", N 7, 2012. С. 36-39.
4. V. Sadovnichy, A. Tikhonravov, Vl. Voevodin, and V. Opanasenko "Lomonosov": Supercomputing at Moscow State University. In Contemporary High Performance Computing: From Petascale toward Exascale (Chapman & Hall/CRC Computational Science), pp.283-307, Boca Raton, USA, CRC Press, 2013.