УДК 519.6:532.5:533.6.011 doi 10.26089/NumMet.v18r434

ИССЛЕДОВАНИЕ МАСШТАБИРУЕМОСТИ FLOWVISION НА КЛАСТЕРЕ С ИНТЕРКОННЕКТОМ АНГАРА

В. С. Акимов[1], Д. П. Силаев[2], А. С. Симонов[3], А. С. Семенов[4]

Исследуется масштабируемость вычислений задач газодинамики в программном комплексе FlowVision на кластере Ангара-К1 с интерконнектом Ангара. Рассматривались несколько тестовых задач, имеющих 260 тысяч, 5.5 млн и 26.8 млн расчетных ячеек. Вычисления во FlowVision проводились с использованием нового решателя систем линейных алгебраических уравнений, основанного на алгебраическом многосеточном методе AMG (Algebraic MultiGrid). Показано, что специальная технология FlowVision "Динамическая балансировка" позволяет существенно увеличить производительность вычислений, если особенности постановки расчетной задачи способствуют неравномерности загрузки процессоров. Кластер Ангара-К1 продемонстрировал отличные характеристики производительности и масштабируемости вычислений, не уступающие аналогам с интерконнектом 4x FDR Infiniband.

Ключевые слова: масштабируемость, FlowVision, CFD, газодинамика, кластер, суперкомпьютер, интерконнект, Ангара.

1. Введение. Развитие вычислительной техники происходит интенсивно, 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 — увеличивается пропускная способность памяти и совершенствуется интерконнект между вычислительными узлами. 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.

В современных суперкомпьютерах наиболее часто применяются коммерческие коммуникационные сети Mellanox Infiniband и Intel OmniPath. Кроме того, применяются коммуникационные сети типа IBM BlueGene/Q и Tofu, которые производятся для конкретных серий суперкомпьютеров и не поставляются отдельно от них. При этом в таких сериях могут создаваться как уникальные суперкомпьютеры из первой десятки Top500, так и небольшие системы для нужд промышленных организаций. Кроме того, интерес представляют европейские разработки в области высокоскоростных сетей, прежде всего это Bull Exascale Interconnect (BXI) [1] и Extoll [2].

Сеть Ангара [3, 4] — первая российская высокоскоростная коммуникационная сеть на основе СБИС маршрутизатора. СБИС маршрутизатор коммуникационной сети является разработкой Научно-исследовательского центра электронной вычислительной техники (НИЦЭВТ) и выпущен 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 и все стандартные средства программирования (библиотека MPI, технология OpenMP, библиотека SHMEM, стек протоколов TCP/IP). Коммуникационная сеть Ангара совместима с процессорами x86, Эльбрус и ARM, а также с ускорителями GPU и FPGA. В настоящий момент существуют два вычислительных кластера, оснащенных сетью Ангара: 32-узловой гибридный кластер в Объединенном институте высоких температур РАН (с топологией 4D-тор 4x2x2x2) и 36-узловой кластер Ангара-К1, установленный в здании НИЦЭВТ с топологией 3D-тор 4x3x3. Результаты оценочного тестирования кластера Ангара-К1 на тестах OSU, Intel MPI Benchmarks и NAS Parallel Benchmarks представлены в статье [5].

[1] ООО "Вычислительная инженерная платформа", ул. Юннатов, д. 18, 127083, Москва; инженер, e-mail: akimov@flowvision.ru

[2] ООО "Вычислительная инженерная платформа", ул. Юннатов, д. 18, 127083, Москва; ведущий разработчик, e-mail: silaev@flowvision.ru

[3] Научно-исследовательский центр электронной вычислительной техники (НИЦЭВТ), Варшавское шоссе, д. 125, 117587, Москва; первый зам. генерального директора, e-mail: simonov@nicevt.ru

[4] Научно-исследовательский центр электронной вычислительной техники (НИЦЭВТ), Варшавское шоссе, д. 125, 117587, Москва; зам. начальника отдела, e-mail: semenov@nicevt.ru

Одним из наиболее распространенных вариантов использования мощностей суперкомпьютеров являются инженерные расчеты в области гидро- и газодинамики. Со стороны пользователей CFD-кодов (Computational Fluid Dynamics) 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 определяется прежде всего выбором вычислительного метода и приемов при организации программного кода, от которых зависят показатели эффективности использования подсистемы памяти и интерконнекта.

В настоящей статье проводится исследование масштабируемости вычислений при решении нескольких задач газодинамики при помощи программного комплекса FlowVision на кластере Ангара-К1, оснащенного интерконнектом Ангара. Исследование выполняется в сравнении с другими суперкомпьютерами, использующими интерконнект Infiniband 4xFDR.

2. Программный комплекс FlowVision. Программный комплекс FlowVision — это многоцелевое решение для моделирования трехмерных течений жидкости и газа в технических и природных объектах, а также для визуализации этих течений методами компьютерной графики [6]. Моделируемые течения — это стационарные и нестационарные, сжимаемые и несжимаемые потоки жидкости и газа. FlowVision относится к программному обеспечению, использующему методы вычислительной гидрогазодинамики (CFD) и, в частности, метод конечных объемов (МКО). С использованием этих методов производится численное решение уравнений неразрывности, количества движения Навье–Стокса, энергии и др. При использовании МКО пространственная дискретизация решаемой задачи осуществляется путем разбиения расчетной области на небольшие соприкасающиеся объемы, представляющие собой ячейки расчетной сетки. Расчетная сетка во FlowVision является декартовой, ячейки сетки представляют собой гексаэдры. При этом имеется возможность производить локальное сгущение расчетной сетки в областях, где требуется более подробное разрешение особенностей геометрической модели или градиентов физических величин. Такое сгущение может быть проведено в локальном объеме, по поверхности геометрической модели или в зависимости от полей рассчитываемых переменных, в том числе в автоматическом режиме.

Неявные схемы аппроксимации, используемые во FlowVision, требуют решения разреженных систем линейных алгебраических уравнений (СЛАУ) с высокой точностью на системах с распределенной памятью. На решение СЛАУ расходуется значительная часть общих затрат машинного времени и оперативной памяти. В связи с этим, выбор эффективного метода решения СЛАУ является важной задачей и способен сократить время, требуемое на моделирование. Во FlowVision реализованы 3 различных решателя СЛАУ: алгебраический многосеточный метод (AMG) с агрегативным способом огрубления, AMG с селективным способом огрубления и TParFBSS, сочетающий предобусловливание типа неполного треугольного разложения и итерационную схему крыловского типа. Конкретный решатель СЛАУ выбирается адаптивно с помощью технологии AST (Aggregative AMG–Selective AMG–TParFBSS). Его выбор зависит от накопленной на предыдущих итерациях FlowVision истории решения СЛАУ конкретного типа. При получении всех результатов, представленных в данной статье, решение СЛАУ осуществлялось с помощью метода AMG с агрегативным способом огрубления.

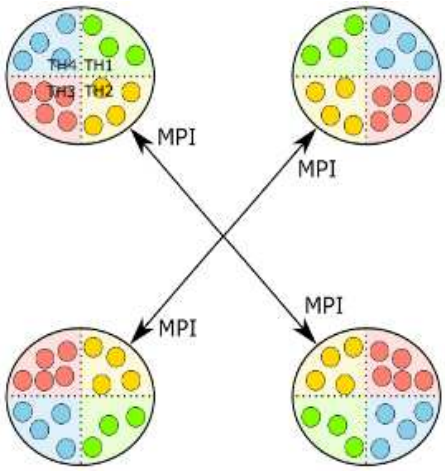


Рис. 1. Гибридная параллельная архитектура FlowVision

Для уменьшения времени счета проводимые вычисления требуется распараллеливать в соответствии с архитектурой используемой вычислительной техники. Современные вычислительные кластеры обладают архитектурой с распределенной памятью: с одной стороны, имеется набор вычислительных узлов, обмен данными между которыми осуществляется посредством интерконнекта, а с другой стороны, каждый узел представляет собой многопроцессорный (многосокетный) сервер с общим доступом к оперативной памяти. В рамках одного сокета доступ к памяти, обычно, является однородным (UMA — Uniform Memory Access), в то время как доступ к памяти соседнего сокета является неоднородным (NUMA — Non-Uniform Memory Access). Поэтому во FlowVision реализован гибридный подход к распараллеливанию вычислений, сочетающий в себе преимущества распараллеливания по MPI (Message Passing Interface) и по нитям (threads) (рис. 1) [7]. В рамках однородного доступа к памяти одного процессора, как правило, преимуществами

обладает метод распараллеливания по нитям, который, в том числе, позволяет использовать меньшее количество оперативной памяти. В то же время, для распараллеливания между процессорами необходимо использовать MPI. Поэтому реализация именно такой схемы изначально предлагается пользователям FlowVision: между процессорами распараллеливание происходит по MPI, а по ядрам процессора с использованием нитей (рис. 1). Однако в зависимости от архитектуры вычислительной сети и особенностей решаемых задач может оказаться более эффективным использование более чем одного MPI-процесса на процессор (особенно в случае большого количества ядер процессоров). В связи с этим, имеется гибкая возможность задания желаемых комбинаций MPI-процессов и нитей.

Спектр задач, решаемых во FlowVision, очень широк, и возможно бесконечное множество различных конфигураций расчетной сетки и количества процессоров, используемых для вычислений. Значительная доля промышленных задач в области гидро- и газодинамики имеют сложную геометрическую модель и требуют хорошего локального разрешения расчетной сеткой градиентов физических величин. При этом внутри расчетной области могут содержаться значительные объемы (например, твердых тел), не являющиеся расчетными. Также у поверхностей этих тел, зачастую, требуются локальные сгущения сетки и использование так называемой приповерхностной сетки. Кроме того, эти твердые тела могут менять положение в пространстве, при этом сетка динамически перестраивается в процессе расчета. Расчетную область от нерасчетной может отделять и свободная поверхность (поверхность раздела фаз), которая тоже динамически изменяется и способствует перестроению расчетной сетки. Объем вычислений, осуществляемый для ячеек около поверхности твердого тела, в приповерхностной сетке и около свободной поверхности, отличается от объема вычислений в ячейках, удаленных от этих поверхностей. Поэтому простой балансировки по количеству ячеек оказывается недостаточно. По этим причинам оказывается невозможным изначально организовать код таким образом, чтобы вычисления для всех вариантов постановок расчетных задач были бы одинаково хорошо сбалансированы и каждый MPI-процесс обрабатывал бы равноценный объем вычислений, особенно при наличии динамического перестроения сетки. Для решения этой проблемы FlowVision имеет собственный инструмент “Динамическая балансировка”, позволяющий существенно ускорить многопроцессорные вычисления за счет перераспределения ячеек в процессе расчета между MPI-процессами. Следует отметить, что “Динамическая балансировка” не просто уравнивает количество ячеек, обрабатываемых каждым MPI-процессом (что можно было бы сделать предварительно), а уравнивает время, затрачиваемое на вычисления на каждом из них (что можно сделать только в процессе расчета).

3. Исследования масштабируемости.

3.1. Технические характеристики использованных суперкомпьютеров. Исследования масштабируемости FlowVision проводились с использованием упомянутого выше кластера Ангара-К1, а также других суперкомпьютеров. Технические характеристики используемых суперкомпьютеров представлены в табл. 1.

На всех кластерах установлена система очередей SLURM, для запуска определенного количества MPI-процессов на каждый узел использовался параметр *ntasks-per-node*.

3.2. Тестовые задачи. При решении прикладных задач инженеры сталкиваются с множеством различных особенностей расчетных моделей. Эти задачи имеют различную геометрическую модель, количество расчетных ячеек, а также степень неоднородности расчетной сетки из-за наличия локальных сгущений сетки, нерасчетных объемов и др. Поэтому в рамках данной работы для исследования особенностей масштабирования вычислений использовались задачи различного типа и размерности. Основные особенности рассматриваемых тестовых задач сведены в табл. 2.

Задача обтекания каверны воздухом M219 Cavity case (рис. 2) является широко известным в литературе валидационным тестом [8] и представляет класс задач внешнего обтекания объектов, имеющих простую геометрическую форму. Как правило, расчетная сетка в таких задачах имеет локальные сгущения (рис. 2а), однако декомпозицию сетки по MPI-процессам удается провести таким образом, чтобы разница их загрузки не превышала 20%.

В тестовой задаче внешнего обтекания самолета (рис. 3), напротив, неравномерность загрузки процессоров может оказываться значительной (50% и более) из-за неоднородности расчетной сетки: наличия сложной геометрической модели (рис. 3а), сгущения сетки на поверхности (рис. 3б) и использования приповерхностной сетки.

В качестве теста с малым количеством расчетных ячеек и, кроме того, обеспечивающего отличную равномерность загрузки процессоров использовалась задача смешивания горячей и холодной воды в смесителе (рис. 4).

Таблица 1

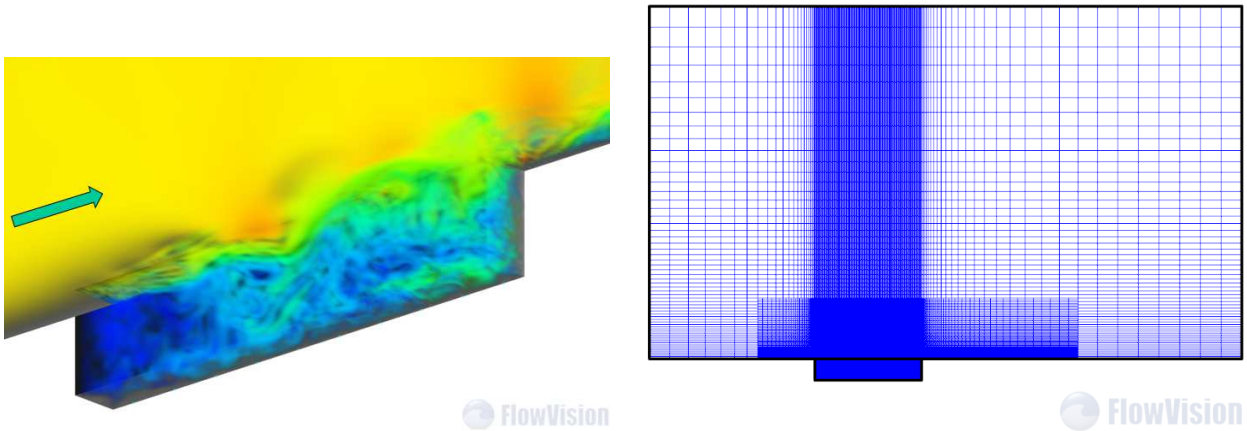
Технические характеристики суперкомпьютеров

Суперкомпьютер	Ангара-К1 (раздел A)	Ломоносов-2 (раздел compute)	Cluster Z
Процессор	Intel Xeon E5-2630, 2.30 GHz	Intel Xeon E5-2697v3, 2.6 GHz	Intel Xeon E5-2670, 2.6 GHz
Количество физических ядер процессора	6	14	8
Количество логических ядер при использовании Hyper-Threading (HT)	12	28	HT отключен
Кэш-память, МБ	15	35	20
Максимальная пропускная способность памяти, ГБ/с	42.6	68	51.2
Количество процессоров на узле	2	1	2
Количество оперативной памяти на узел, ГБ	64	64	64
Топология, интерконнект	Ангара, 3D-тор 4x3x3	Mellanox FDR InfiniBand (56 Гбит/с)	FDR InfiniBand (56 Гбит/с)
Реализация MPI	MPICH 3.0.4	OpenMPI 1.8.4	Intel MPI 5.1

Таблица 2

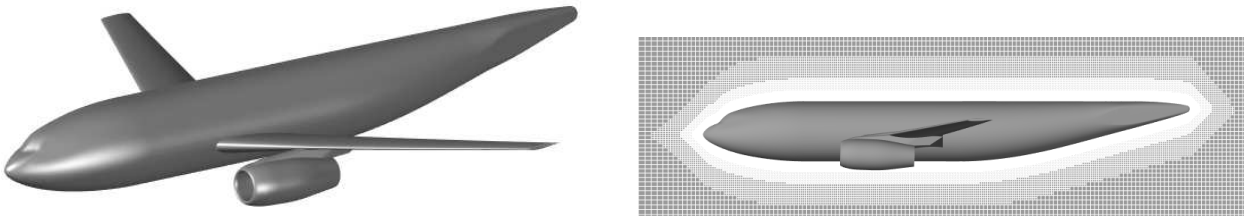
Характеристики тестовых задач

Тестовая задача	M219 Cavity case [8], рис. 2	Внешнее обтекание самолета, рис. 3	Смеситель, рис. 4
Постановка	Трехмерная	Трехмерная	Трехмерная
Моделируемые физические явления	Теплоперенос, движение	Теплоперенос, движение, турбулентность	Теплоперенос, движение, турбулентность
Количество ячеек расчетной сетки	5.5 млн	26.8 млн	260 тыс.
Приповерхностная сетка	Отсутствует	Есть	Отсутствует
Адаптация (сгущение) расчетной сетки	Локально в объеме	По поверхности	Отсутствует



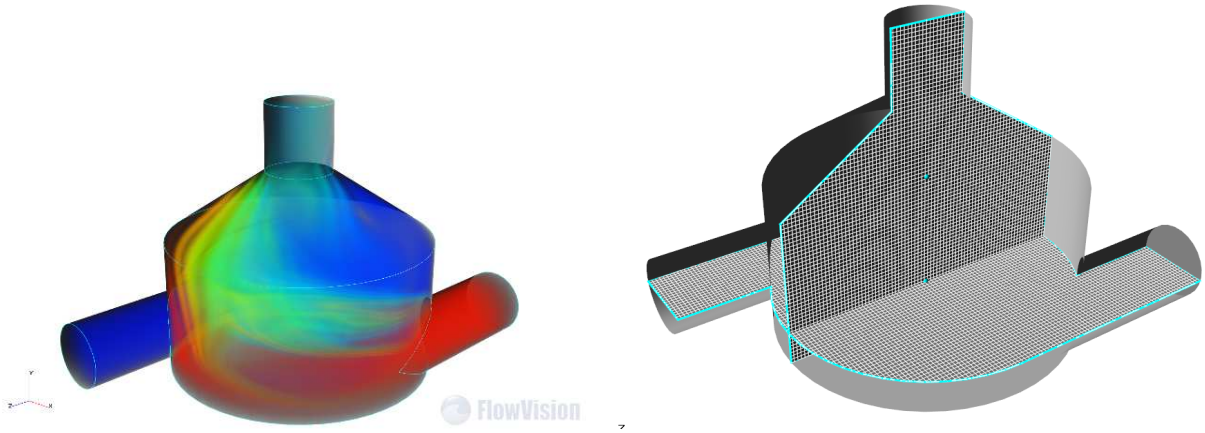
а) б)

Рис. 2. Задача M219 Cavity case: а) объемная визуализация скорости; б) расчетная сетка



а) б)

Рис. 3. Задача моделирования внешнего обтекания самолета:
а) геометрическая модель самолета; б) расчетная сетка



а) б)

Рис. 4. Задача смешивания горячей и холодной воды в смесителе:
а) объемная визуализация температуры; б) сечения расчетной сетки

3.3. Методика исследований и режимы запусков. Для осуществления исследований и получения результатов проводились запуски тестовых задач, представленных в табл. 2, в различных режимах и фиксировалось время вычисления контрольного шага для каждого из них. Под режимом запуска тестовой задачи в данной работе понимаются сведения о количестве задействованных расчетных узлов, количестве MPI-процессов, назначенных на каждый расчетный узел, и количестве нитей на каждый MPI-процесс (табл. 3).

Таблица 3

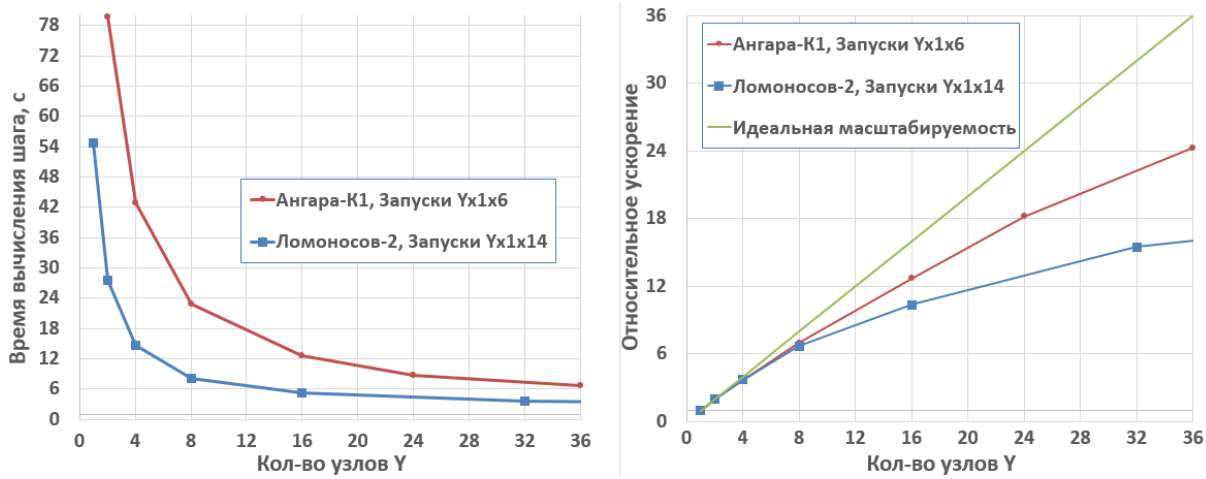
Режимы запуска параллельных задач

	Количество узлов (Nodes)	Количество MPI-процессов на узел (MPIs)	Количество нитей на каждый MPI-процесс (threads)
Параметр системы очередей SLURM или FlowVision	-N (параметр SLURM)	--ntasks-per-node (параметр SLURM)	threads (параметр FlowVision)
Пример	24	2	6
Обозначение для примера (Nodes x MPIs x threads)	24x2x6		

4. Результаты.

4.1. Масштабируемость вычислений при запусках по одному MPI-процессу на узел. На первом этапе исследовалась масштабируемость вычислений при запусках задачи M219 Cavity case с использованием одного MPI-процесса на узел и количества нитей, равного количеству физических ядер процессоров. Запуски проводились на двух суперкомпьютерах: Ангара-К1 и Ломоносов-2. Для всех исследуемых режимов тестовая задача запускалась на расчет с 150-го до 155-го шага и время вычисления фиксировалось для шага с номером 155. Следует отметить, что до 150-го расчетного шага течение га-

за уже развилось, все вспомогательные операции по построению расчетной сетки, ее адаптации, набору статистики и т.п. завершены.



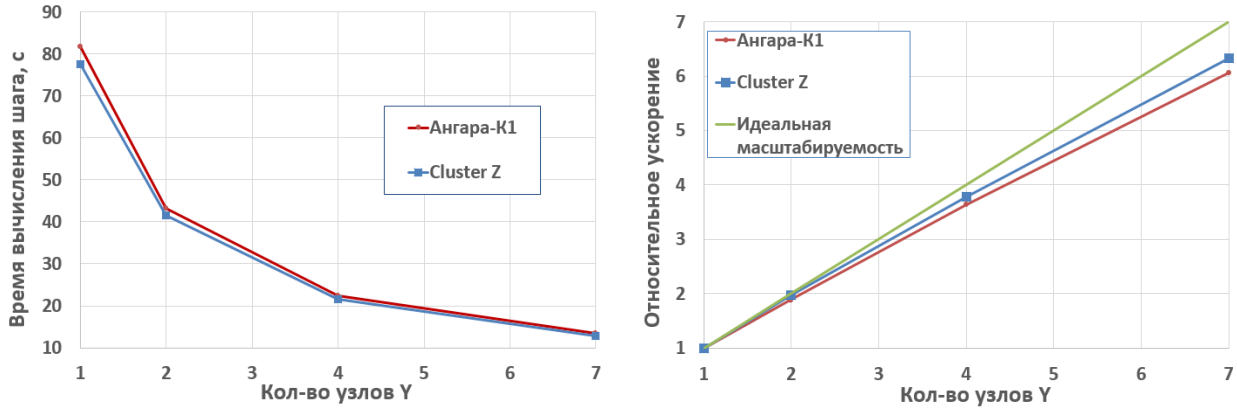
а) б)

Рис. 5. Масштабируемость вычислений при запусках по одному MPI-процессу на узел:
а) время вычисления шага; б) относительное ускорение

Зависимость времени вычисления шага от количества используемых узлов приведена на рис. 5а. На рис. 5б показано ускорение вычислений относительно показателей при одноузловом запуске.

Из этих рисунков следует, что время вычисления шага значительно ниже на суперкомпьютере Ломоносов-2 благодаря более современным процессорам (см. табл. 1). Этот фактор также способствует тому, что относительные временны́е затраты на MPI-обмены растут более интенсивно на Ломоносов-2 с увеличением количества узлов. Кроме того, из табл. 1 можно видеть, что процессоры Ломоносов-2 имеют в 2.3 раза большее количество ядер, в то время как максимальная пропускная способность памяти выше всего в 1.6 раза. Указанные факторы объясняют, почему масштабируемость Ломоносов-2 выглядит хуже.

4.2. Масштабируемость вычислений при запусках по два MPI-процесса на узел. Далее проводились запуски той же задачи по два MPI-процесса на узел кластеров Ангара-К1 и Cluster Z, т.е. по одному MPI-процессу на каждый физический процессор. Для корректности сравнения на обоих кластерах использовалось 6 нитей на MPI-процесс, т.е. проводились запуски в режиме Yx2x6. Сравнение результатов по времени вычисления шага и по относительному ускорению представлены на рис. 6.



а) б)

Рис. 6. Масштабируемость вычислений при запусках по два MPI-процесса на узел (запуски Yx2x6):
а) время вычисления шага; б) относительное ускорение

Из рис. 6 видно, что кластеры демонстрируют практически идентичные показатели производительности и масштабируемости вычислений.

4.3. Эффект использования Hyper-Threading (HT). На рис. 7 представлена кривая масштабируемости той же задачи по количеству нитей при запусках на 24 узла по два MPI-процесса на узел

кластера Ангара-К1. Можно видеть, что использование всех логических ядер HT позволяет получить прирост производительности всего на 4.5%.

С целью определить наиболее оптимальный способ использования логических ядер HT были проведены запуски при различных комбинациях MPIs и threads (табл. 3) при сохранении неизменным суммарного количества потоков на узел (MPIs x threads =24). Кроме того, для этих запусков неизменным оставалось количество узлов Nodes=24. Прирост производительности вычислений при этих запусках относительно запуска 24x2x6 представлен на рис. 8.

Как можно видеть из рис. 8, наиболее эффективным оказалось использование всех логических ядер HT за счет удвоения количества MPI-процессов (запуск 24x4x6), в то время, как запуски с использованием более восьми MPI-процессов на узел (т.е. четырех на физический процессор) показали ухудшение производительности вычислений.

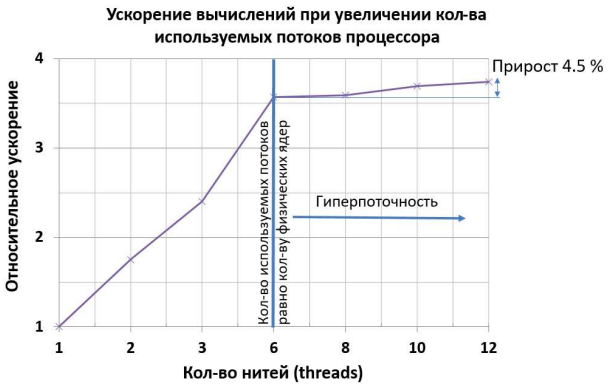


Рис. 7. Масштабируемость вычислений по количеству нитей

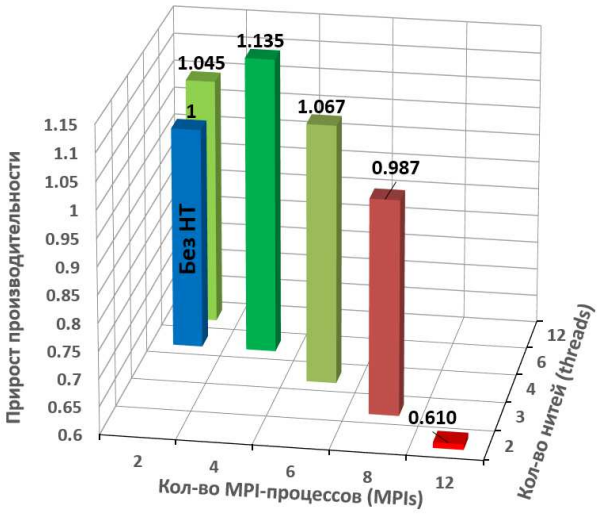
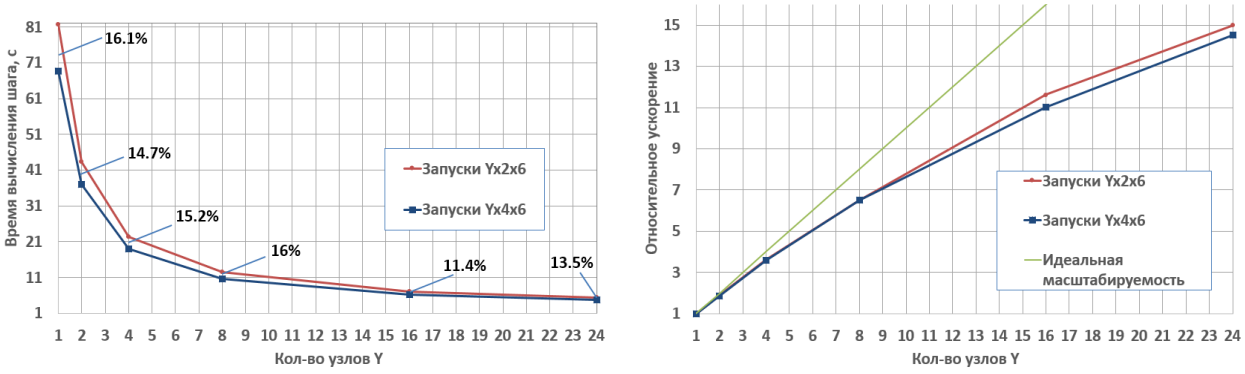


Рис. 8. Прирост производительности вычислений различных запусков по сравнению с запуском 24x2x6

Определенный интерес представляет вопрос, как выглядит преимущество найденного оптимального сочетания MPIs x threads = 4x6 при различном количестве узлов. На рис. 9 представлено сравнение кривых времени вычисления шага и относительного ускорения для запусков с использованием двух и четырех MPI-процессов на узел. Можно видеть, что время вычисления шага ниже на 11.4–16.1% в случае использования удвоенного количества MPI-процессов во всем диапазоне числа использованных узлов.



а) б)

Рис. 9. Сравнение масштабируемости вычислений при запусках по два и четыре MPI-процесса на узел: а) время вычисления шага; б) относительное ускорение

4.4. Масштабируемость задачи с малым числом ячеек и предел масштабируемости. Масштабируемость любого CFD-приложения имеет некоторые пределы, которые проявляются при определенном количестве расчетных ячеек, приходящихся на ядро процессора. С целью определения этих пределов в

данной работе рассматривалась задача с относительно малым количеством ячеек — 260 тысяч (см. табл. 2, задача “Смеситель”). Для всех исследуемых режимов запуска данной тестовой задачи фиксировалось время вычисления 8-го шага от начала расчета. На рис. 10 представлены кривые ускорения вычислений на кластерах Ангара-К1 и Ломоносов-2 при увеличении суммарного количества ядер относительно запуска на одно ядро процессора. Отметим, что в отличие от запусков, представленных на рис. 5, в данном случае на обоих кластерах на каждый узел запускалось по два MPI-процесса по 6 нитей (MPIs x threads = 2x6).

Из кривых на рис. 10 видно, что на обоих кластерах кривая масштабируемости имеет экстремум при количестве ячеек на ядро процессора, равном 1200. В данном случае кластер Ангара-К1 демонстрирует не только несколько лучшую кривую масштабируемости, но и лучшую “толерантность” к малому числу ячеек на ядро процессора в области, где возможны проявления эффектов кэш-памяти. Однако стоит отметить, что, как правило, при столь малом количестве ячеек на ядро процессора время, затрачиваемое на MPI-обмены, становится сравнимым с временем вычисления расчетного шага. Поэтому в подавляющем большинстве случаев при решении промышленных задач такая “глубокая” масштабируемость является коммерчески нецелесообразной. В целом, при желании выходить в область менее 5 тысяч ячеек на ядро пользователям FlowVision рекомендуется предварительно получить кривые, подобные рис. 10, на конкретном типе решаемой задачи и конкретной аппаратуре.

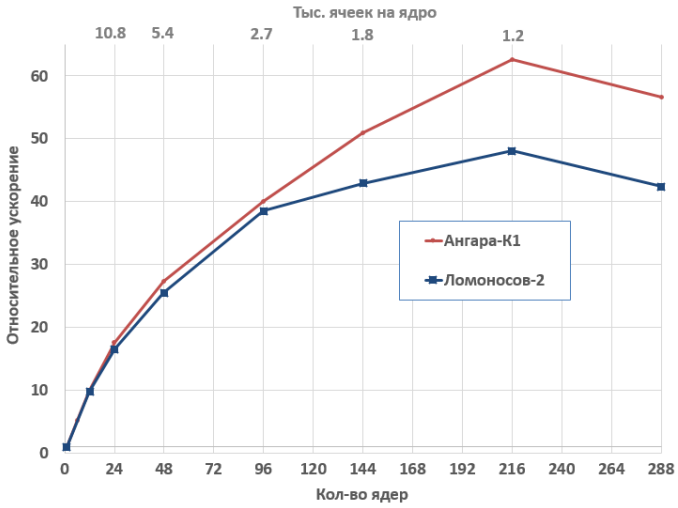
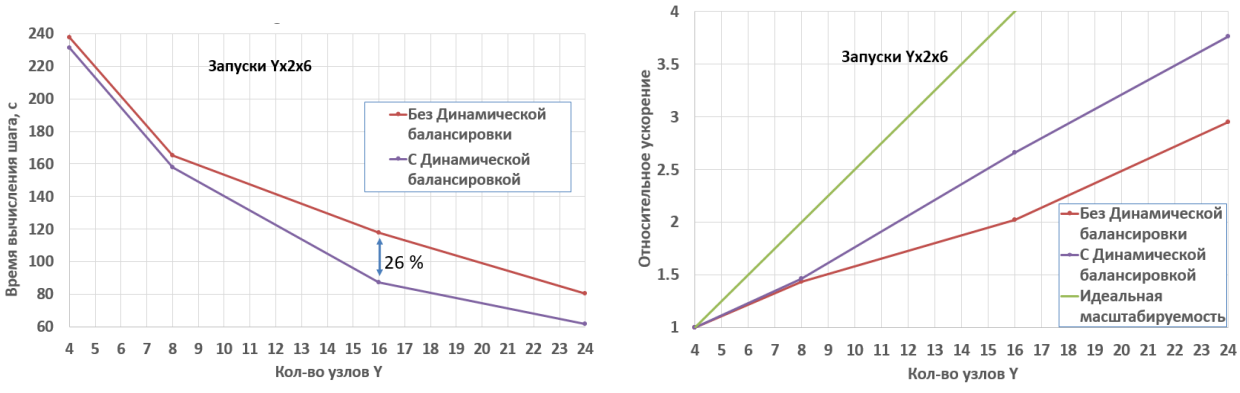


Рис. 10. Сравнение масштабируемости вычислений задачи с малым числом ячеек

4.5. Масштабируемость задачи с существенно неоднородной расчетной сеткой. Выше отмечено, что такие особенности задач, как наличие сложной геометрической модели, локальные сгущения расчетной сетки у поверхностей твердых тел и использование приповерхностной сетки в процессе расчета могут привести к значительной неравномерности загрузки MPI-процессов и, следовательно, процессоров. В частности, при осуществлении вычислений на кластере Ангара-К1 задачи внешнего обтекания самолета, представленной в табл. 2 и на рис. 3, с использованием 48 MPI-процессов (режим запуска 24x2x6) разница загрузки процессов составляет почти 300%. Такая разница является значительным дисбалансом и означает, что некоторые процессы большую часть времени ждут, пока остальные закончат вычисления.



а) б)

Рис. 11. Сравнение масштабируемости вычислений на кластере Ангара-К1 при запусках с использованием технологии “Динамическая балансировка” и без нее:
а) время вычисления шага; б) относительное ускорение

Для оценки эффективности влияния технологии “Динамическая балансировка” в таких задачах рассмотренная технология была включена на протяжении 10 расчетных шагов, начиная с 801-го. Данные снимались с 11-го по счету шага, когда динамическая балансировка уже сделала свою работу и уже не тратит машинное время на анализ существующей неравномерности. На рис. 11. представлено сравнение кривых времени вычисления 11-го по счету шага и относительного ускорения для запусков с использо-

ванием технологии “Динамическая балансировка” и без нее. Ускорение вычислений измерялось относительно запуска на 4 расчетных узла, так как при использовании меньшего количества узлов оказалось недостаточно оперативной памяти для осуществления вычислений.

Из рис. 11 можно видеть, что использование динамической балансировки позволяет значительно снизить время вычисления шага, а также значительно улучшить масштабируемость в области более 8 расчетных узлов.

Как было отмечено выше, сам процесс динамической балансировки занимает некоторое процессорное время, поэтому пользователям FlowVision не рекомендуется включать эту опцию на протяжении всего расчета. Активацию “Динамической балансировки” стоит включать на 5–10 шагов расчета при возникновении неравномерности загрузки процессов. При этом если в процессе расчета изменяется расчетная сетка, то выгодно настроить периодическое применение “Динамической балансировки”, для чего во FlowVision имеется специальный инструмент.

5. Выводы. В данной работе проведено исследование масштабируемости вычислений при решении нескольких задач газодинамики при помощи программного комплекса FlowVision на кластере Ангара-К1, оснащенного сетью Ангара с топологией 3D-тор. Задачи имеют различное количество ячеек расчетной сетки, а также используют различные возможности FlowVision, такие как моделирование турбулентности, адаптацию расчетной сетки в объеме и по поверхности, приповерхностную сетку.

Исследование позволило сформулировать следующие выводы.

1. Кластер Ангара-К1 с сетью Ангара продемонстрировал отличные характеристики производительности и масштабируемости вычислений, которые не уступают современным аналогам с интерконнектом 4x FDR Infiniband.

2. Удвоение количества MPI-потоков на узел за счет логических ядер HT позволяет снизить время, затрачиваемое на вычисления, на 11.4–16.1%.

3. Максимум кривой ускорения вычислений задачи с однородной расчетной сеткой, содержащей 260 тысяч ячеек, проявляется при выполнении на 216 ядрах, когда на каждое ядро приходится 1200 расчетных ячеек. При этом на кластере Ангара-К1 удается достичь более чем шестидесятикратного ускорения по сравнению с запуском данной задачи на одно ядро процессора. Необходимо заметить, что для эффективного использования оборудования рекомендуется, чтобы на каждое ядро приходилось не менее 5000 расчетных ячеек.

4. Специальная технология FlowVision “Динамическая балансировка” позволяет существенно увеличить производительность вычислений, если особенности постановки расчетной задачи способствуют неравномерности загрузки процессоров.

Статья рекомендована к публикации Программным комитетом Международной научной конференции “Суперкомпьютерные дни в России 2017” (http://russianscdays.org).

СПИСОК ЛИТЕРАТУРЫ

1. *Derradji S., Palfer-Sollier T., Panziera J.-P., et al.* The BXI interconnect architecture // Proc. IEEE 23rd Annual Symposium on High-Performance Interconnects. Washington, DC: IEEE Press, 2015. doi 10.1109/HOTI.2015.15.
2. *Fröning H., Nussle M., Litz H., et al.* On achieving high message rates // Proc. 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing. New York: IEEE Press, 2013. doi 10.1109/CCGrid.2013.43.
3. *Жабин И.А., Макагон Д.В., Поляков Д.А., Симонов А.С., Сыромятников Е.Л., Щербак А.Н.* Первое поколение высокоскоростной коммуникационной сети “Ангара” // Наукоемкие технологии. 2014. № 1. 21–27.
4. *Слуцкин А.И., Симонов А.С., Жабин И.А., Макагон Д.В., Сыромятников Е.Л.* Разработка межузловой коммуникационной сети ЕС8430 “Ангара” для перспективных российских суперкомпьютеров // Успехи современной радиоэлектроники. 2012. № 1. 6–10.
5. *Агарков А.А., Исмагилов Т.Ф., Макагон Д.В., Семенов А.С., Симонов А.С.* Результаты оценочного тестирования отечественной высокоскоростной коммуникационной сети Ангара // Тр. Международной конференции “Суперкомпьютерные дни в России”. М.: Изд-во Моск. ун-та, 2016. 626–639.
6. FlowVision Help. https://flowvision.ru/webhelp/fvru_30905. Cited November 14, 2017.
7. *Сушко Г.Б., Харченко С.А.* Экспериментальное исследование на СКИФ МГУ “Чебышев” комбинированной MPI+threads реализации алгоритма решения систем линейных уравнений, возникающих во FlowVision при моделировании задач вычислительной гидродинамики // Труды Международной конференции ПАВТ-2009. Челябинск: Изд-во УрГУ, 2009. 316–324.
8. *Henshaw M.J.C.* M219 cavity case: verification and validation data for computational unsteady aerodynamics // Tech. Rep. RTO-TR-26, AC/323(AVT)TP/19. QinetiQ, UK, 2002. 453–472.

Поступила в редакцию
07.11.2017

Scalability Study of FlowVision on the Cluster with Angara Interconnect

V. S. Akimov [1], D. P. Silaev [2], A. S. Simonov [3], and A. S. Semenov [4]

[1] *Limited Liability Company "Numerical Engineering Platform"; ulitsa Yunnatov 18, Moscow, 127083, Russia; Ph.D., Engineer, e-mail: akimov@flowvision.ru*

[2] *Limited Liability Company "Numerical Engineering Platform"; ulitsa Yunnatov 18, Moscow, 127083, Russia; Leading Developer, e-mail: silaev@flowvision.ru*

[3] *Scientific Research Center for Electronic Computer Technology; Varshavskoe shosse 125, Moscow, 117587, Russia; Ph.D., First Deputy General Director, e-mail: simonov@nicevt.ru*

[4] *Scientific Research Center for Electronic Computer Technology; Varshavskoe shosse 125, Moscow, 117587, Russia; Ph.D., Deputy Head of Department, e-mail: semenov@nicevt.ru*

Received November 7, 2017

Abstract: The scalability of computations in FlowVision CFD software on the Angara-C1 cluster equipped with Angara interconnect is studied. Several test problems with 260 thousand, 5.5 million and 26.8 million computational cells are considered. Computations in FlowVision are performed using a new solver of linear systems based on the algebraic multigrid (AMG) method. It is shown that the special FlowVision's technology named "Dynamic balancing" significantly improves the performance of computations if the peculiarities of the problem promote the non-uniform loading of CPUs. The Angara-C1 cluster demonstrates the excellent performance and scalability characteristics comparable with its analogues based on the 4x FDR Infiniband interconnect.

Keywords: scalability, FlowVision, CFD, gas dynamics, cluster, supercomputer, interconnect, Angara.

References

1. S. Derradji, T. Palfer-Sollier, J.-P. Panziera, et al., "The BXI Interconnect Architecture," in *Proc. IEEE 23rd Annual Symp. on High-Performance Interconnects, Santa Clara, USA, August 26–28, 2015* (IEEE Press, Washington, DC, 2015), doi 10.1109/HOTI.2015.15

2. H. Fröning, M. Nussle, H. Litz, et al., "On Achieving High Message Rates," in *Proc. 13th IEEE/ACM Int. Symp. on Cluster, Cloud, and Grid Computing, Delft, The Netherlands, May 13–16, 2013* (IEEE, New York, 2013), doi 10.1109/CCGrid.2013.43

3. I. A. Zhabin, D. V. Makagon, D. A. Polyakov, et al., "First Generation of Angara High-Speed Interconnection Network," Naukoemkie Tekhnol., No. 1, 21–27 (2014).

4. A. I. Slutskin, A. S. Simonov, I. A. Zhabin, et al., "Development of the ES8430 Angara Interconnect for Future Russian Supercomputers," Usp. Sovr. Radioelektron., No. 1, 6–10 (2012).

5. A. A. Agarkov, T. F. Ismagilov, D. V. Makagon, et al., "Performance Evaluation of the Angara Interconnect," in *Proc. Int. Conf. on Russian Supercomputing Days, Moscow, Russia, September 26–27, 2016* (Mosk. Gos. Univ., Moscow, 2016), pp. 626–639.

6. FlowVision Help. https://flowvision.ru/webhelp/fvru_30905. Cited November 14, 2017.

7. G. B. Sushko and S. A. Kharchenko, "Experimental Investigation of MPI+threads Implementation of the Algorithm for Solving Systems of Linear Equations Occurring in FlowVision when Solving CFD Problems Using the Chebyshev Supercomputer of Moscow University," in *Proc. Int. Conf. on Parallel Computational Technologies, Nizhnii Novgorod, Russia, March 30–April 3, 2009* (South Ural State Univ., Chelyabinsk, 2009), pp. 316–324.

8. M. J. C. Henshaw, "M219 Cavity Case: Verification and Validation Data for Computational Unsteady Aerodynamics," Tech. Rep. RTO-TR-26, AC/323(AVT)TP/19, QinetiQ, UK, 453–472 (2002).