Реализация региональной атмосферной модели на вычислительных системах гибридной архитектуры

В. М. Степаненко, Д. Н. Микушин

Численное моделирование атмосферных процессов является основным инструментом решения задач прогноза погоды и оценки климатических изменений. Успешное решение этих задач во многом определяется эффективностью использования ресурсов многопроцессорных вычислительных комплексов. В настоящей работе представлено два варианта адаптации региональной атмосферной модели NH3D/CMM, развиваемой в НИВЦ МГУ, для гибридных многопроцессорных многоядерных вычислительных систем. В первом варианте реализована двумерная декомпозиция расчетной области и явная передача сообщений по стандарту MPI, а во втором – OpenMP и распределение вычислений для устройств на базе архитектуры Cell Broadband Engine с поддержкой обратной совместимости с MPI. Представлена эффективность этих вариантов модели на суперкомпьютере СКИФ-МГУ "Чебышев" и сервере IBM QS22 компании Т-Платформы.

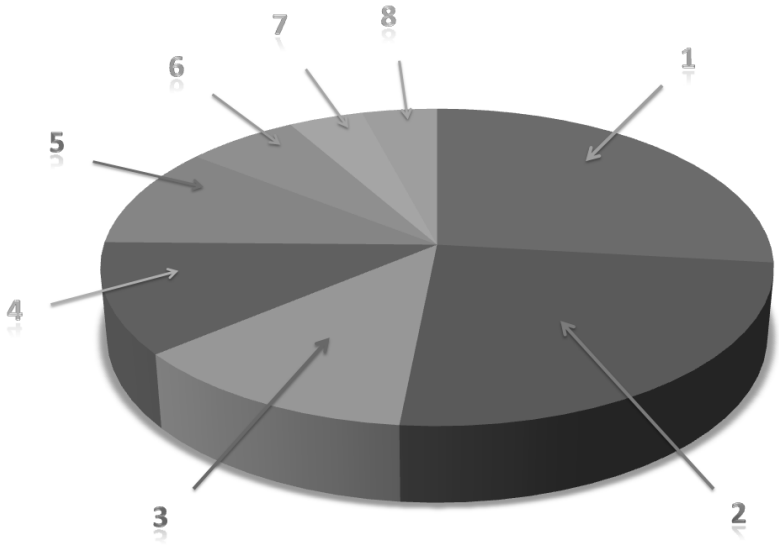
1. Введение

Региональные (мезомасштабные) атмосферные модели являются эффективным инструментом решения многих задач физики атмосферы, в особенности, прогноза погоды и климатических изменений. Они основаны на численном решении системы уравнений гидротермодинамики атмосферы, а также уравнений, описывающих процессы в деятельном слое суши (океана), который включает почву, биоту, водные объекты, ледники и урбанизированные территории. Процессы, не разрешаемые на численной сетке явно, учитываются с помощью параметризаций. Качество прогнозов атмосферных моделей совершенствуется посредством повышения пространственного разрешения, выбора более точных численных схем и усложнения параметризаций подсеточных процессов. Развитие модели в любом из перечисленных направлений приводит к существенному росту вычислительных затрат. Поэтому ведущие региональные модели нацелены на использование многопроцессорных вычислительных систем.

Большинство кластерных систем в настоящее время имеют один или два уровня распределения вычислений, основанные на использовании распределённой памяти или на комбинации общей и распределённой (т.н. «гибридная» двухуровневая архитектура). Для этой архитектуры реализованы многие атмосферные модели (например, региональные модели *MM5* и *WRF*), и они демонстрируют хорошую масштабируемость вплоть до нескольких сотен процессоров. Однако простое наращивание количества процессорных узлов часто оказывается недостаточно эффективным. Высокая стоимость и энергоёмкость традиционных однородных вычислительных систем заставляет искать принципиально иные технологические решения. Одно из перспективных направлений состоит в повышении количества уровней гибридности за счёт оснащения систем специализированными вычислителями типа *CellBE* или графических ускорителей. Сокращение поддержки некоторых второстепенных функций традиционных процессоров, упрощение системной логики, и, как следствие, возможность упаковки большого числа «легковесных» ядер позволяет получить высокоэффективное устройство для решения вычислительных задач. Данные разработки используются в суперкомпьютерах Roadrunner и Tsubame, и , по-видимому, доля таких систем в дальнейшем будет возрастать.

В НИВЦ МГУ развивается региональная атмосферная модель, основная на последовательном коде модели *NH3D* [1]. В исходный вариант модели включена параметризация переноса коротковолновой и длинноволновой радиации в атмосфере, модель гидротермодинамики водоемов и блок переноса атмосферной примеси [2]. Результаты профилирования последовательной версии региональной модели представлены на рис. 1. Модель применяется в задачах моделирования мезомасштабных циркуляций атмосферы над гидрологически неоднородной территорией [3] и в образовательном процессе на географическом факультете и

факультете ВМК МГУ. Дальнейшее развитие модели предполагает увеличение размерности конечно-разностной сетки, включение схемы деятельного слоя суши Института вычислительной математики РАН, тонкий учет процессов выпадения, сублимации и испарения снега в атмосфере и на подстилающей поверхности.



26% **1** – Уравнение для температуры и радиационный блок
24% **2** – Эллиптическое уравнение для геопотенциала
12% **3** – Уравнение неразрывности
11% **4** – Расчёт диффузионных слагаемых в уравнениях импульса и скаляров
10% **5** – Уравнение движения
6% **6** – Вычисление адвективных слагаемых
4% **7** – Интегрирование уравнения для влажности воздуха
4% **8** – Параметризация деятельного слоя суши и водоёмов

Рис. 1. Результаты профилирования региональной атмосферной модели

В настоящей работе представлены некоторые результаты реализации двух параллельных версий модели. Первая версия создается для «традиционных» суперкомпьютеров с сочетанием общей и распределенной памяти. На настоящем этапе реализован ряд блоков модели с применением стандарта *MPI*, этому посвящен п. 2. Вторая версия основана на пакете примитивов *GeoPhyCell*, предназначенном для распределения вычислений в существующих программах для систем с общей памятью и *Cell Broadband Engine Architecture* (*CBEA*) с обратной совместимостью с интерфейсом MPI. На данный момент пакет позволяет оснащать поддержкой параллельности широкий класс численных алгоритмов гидродинамики, таких как схемы уравнений переноса скалярных или векторных величин и вычислительно эквивалентные им. В п. 3 показана производительность распределённых вычислений на примере численного решения задачи о переносе пассивной примеси с использованием численной схемы “чехарда” для сервера *IBM QS22 Blade*, оснащенного парой высокопроизводительных карт *PowerXCell 8i*.

2. Реализация алгоритма для вычислительных систем с распределенной памятью

Применяемые в модели алгоритмы делятся на две группы: алгоритмы решения эволюционных уравнений и алгоритмы решения эллиптического уравнения для геопотенциала.

В модели *NH3D* для решения эволюционных уравнений используется явная схема «чехарда» с центральными разностями по пространству. Она содержит локальную зависимость вычислений

по данным, что позволяет эффективно использовать декартово разбиение области расчета между процессами. При этом для вычисления прогностических переменных в каждой подобласти производится обмен данных только на границах с соседними подобластями (процессами). Объем обмениваемых данных пропорционален ширине шаблона разностной схемы. В случае центральных разностей на границах подобластей необходимо обменивать сечения массива шириной 1.

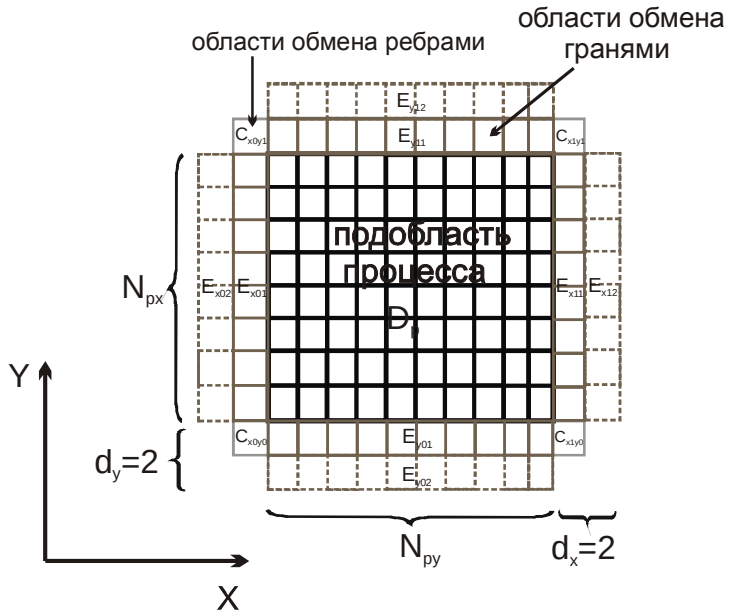


Рис. 2. Схема обменов данными процесса с процессами-соседями при XY-разбиении области

Отметим, что при разбиении расчетной области между процессами каждый процесс хранит только блоки массивов, находящиеся в его подобласти. Это снимает ограничение на размер расчетной сетки, определяемое объемом оперативной памяти одного процессора.

С точки зрения эффективности важным является выбор размерности разбиения. Ранее было показано [4], что двумерное и трехмерное разбиение позволяют получить большее ускорение, чем одномерное. В то же время, трехмерное разбиение не дает существенного выигрыша по сравнению с двумерным *XY*-разбиением даже в задачах, в которых размер сетки по вертикальной координате сравним с размером по горизонтальным координатам [5]. Поэтому для параллельной реализации региональной модели в настоящей работе использовано XY-разбиение.

Таблица 1. Характеристики двух схем транспонирования массивов в параллельной реализации двумерного БПФ

Схема	Общее количество посылаемых и получаемых сообщений на каждом процессе	Общее количество посылаемых и получаемых элементов на каждом процессе	Ограничение на количество процессов
1-й способ	$4N_{px}N_{py}$	$4N_xN_yN_s/\left(N_{px}N_{py}\right)$	$N_{px}N_{py}<N_s$
2-й способ	$2\left(2N_{px}+N_{py}\right)$	$6N_xN_yN_s/\left(N_{px}N_{py}\right)$	$N_{px}=N_{py}<N_s$

Для фильтрации вычислительной моды схемы «чехарда» используется фильтр по времени Асселина, который не требует обмена данными между процессами. Фильтрация ложных волн, возникающих из-за нелинейной неустойчивости, производится пространственным фильтром, размер шаблона которого вдоль каждой из координат равен 5. Это приводит к необходимости

обмена на границе подобластей сечениями массивов шириной 2. В результате, схема обменов принимает следующий вид (рис. 2).

Помимо эволюционных уравнений для компонент скорости, температуры, влажности воздуха и других скалярных переменных модель содержит также эллиптическое уравнение для геопотенциала. Решение конечно-разностного аналога этого уравнения осуществляется в три этапа. На первом этапе производится двумерное быстрое преобразование Фурье (БПФ) правой части уравнения, затем производятся прогонки по вертикальной координате, и к полученному решению применяется обратное БПФ. Параллельная реализация БПФ включает транспонирование массивов, которое может быть осуществлено двумя способами. Первый способ (рис. 3, верхний ряд) заключается в том, что исходное *XY*-разбиение массива между процессами транспонируется в σ-разбиение (σ – вертикальная координата модели), затем в каждом горизонтальном слое σ-разбиения соответствующий процесс производит БПФ по последовательному алгоритму. Второй способ учитывает (рис. 3, нижний ряд), что двумерное БПФ реализуется как последовательность одномерных БПФ по координатам *x* и *y*. Характеристики двух способов приведены в таблице 1 (N_x, N_y, N_s - размерности расчетной сетки, N_{px}, N_{py} - количество процессов по соответствующим осям).

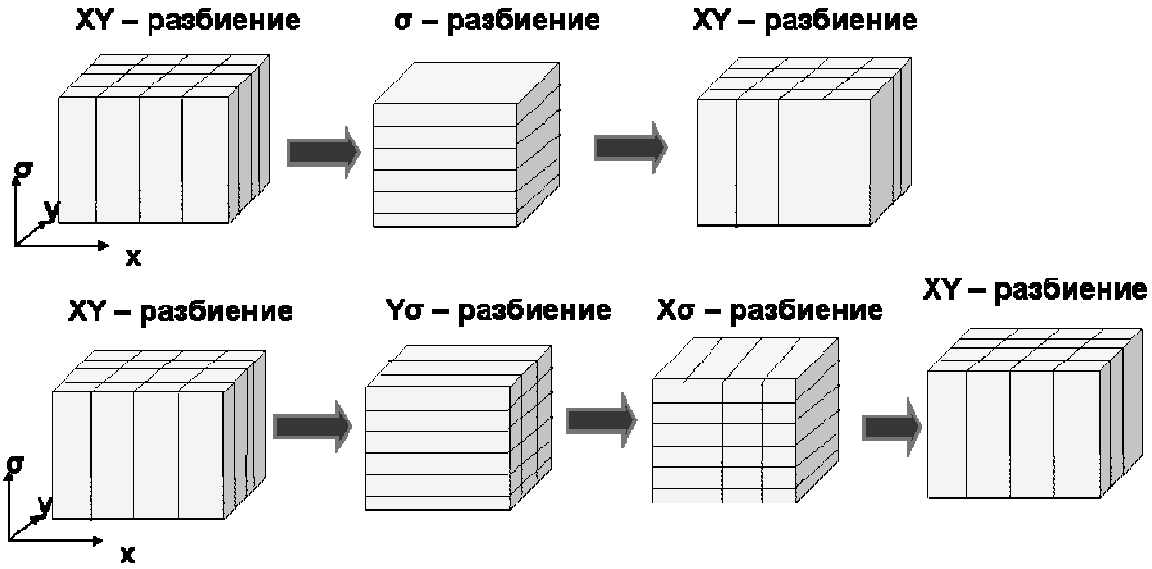


Рис. 3. Две схемы транспонирования массивов в параллельной реализации двумерного БПФ

Из таблицы видно, что общий объем пересылаемых элементов больше для второго способа, но при этом общее количество сообщений в нем существенно меньше. Эти два обстоятельства не позволяют без эксперимента сравнить время выполнения двух описанных способов. В то же время, важным аргументов в пользу второго способа является то, что для него существенно слабее ограничение на возможное количество процессов (в первом способе общее количество процессов ограничивается количеством σ-уровней модели, которое в региональных атмосферных моделях составляет всего несколько десятков). Таким образом, первый способ может быть рекомендован для моделей с большим количеством уровней по вертикали, например, для вихреразрешающих моделей пограничного слоя атмосферы (Глазунов, 2007).

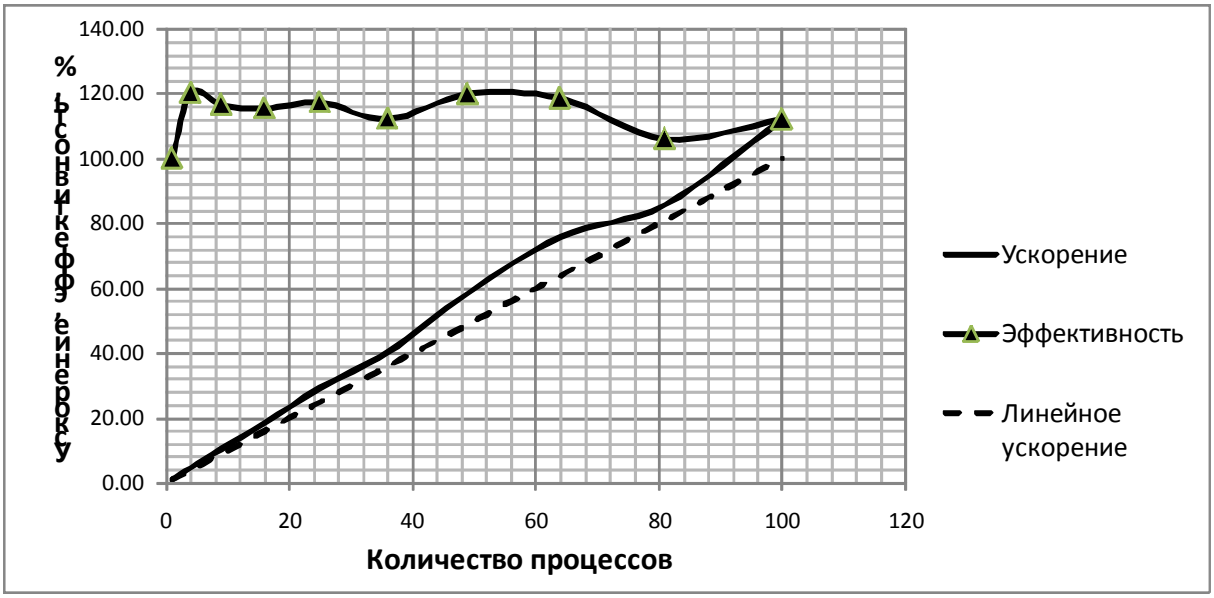


Рис. 4. Ускорение алгоритма явной схемы уравнений движения региональной атмосферной модели на суперкомпьютере СКИФ-МГУ «Чебышев»

На момент написания статьи с применением стандарта MPI реализованы две описанные схемы транспонирования и численная схема трех уравнений движения атмосферной модели. Они протестированы на суперкомпьютерах СКИФ-МГУ «Чебышев» и MVS-50000 (МСЦ РАН). На рисунке 4 показано ускорение алгоритма решения уравнений движения на компьютере СКИФ-МГУ. Таким образом, достигнуто сверхлинейное ускорение, что по всей видимости, вызвано более эффективным использованием кэш-памяти при уменьшении подобластей (размеров массивов на каждом процессе; размерность численной сетки во всех экспериментах была одинаковой). Такое ускорение наблюдалось при распределении вычислений «один процесс на один узел», в случае запуска нескольких процессов в пределах одного узла ускорение было меньше линейного.

3. Пакет для распределения вычислений GeoPhyCell

Конструктивно процессоры архитектуры *CellBE* в большей степени похожи на кластеры с распределённой памятью, чем на многоядерные процессоры [6]. Помимо основного мастер-процессора *Cell* содержит 8 независимых специализированных вычислителей (сопроцессоров), каждый из которых имеет небольшой объём локальной памяти (256 Кб). Доступ к основной оперативной памяти осуществляется через мастер-процессор и шину обмена посредством набора специальных команд, напоминающих интерфейс *MPI*. Однако приёмы *MPI*-программирования к *Cell* малоприменимы из-за ограниченных размеров локальной памяти. В большинстве случаев транспонирование данных невозможно осуществить полностью, и процесс вычислений делится на этапы, в ходе которых каждый сопроцессор обрабатывает не одну порцию данных, а сотни тысяч. Таким образом, по сравнению с MPI в этом случае отпадает необходимость обмена граничными данными между узлами и несколько усложняется алгоритм транспонирования.

В реализованной на настоящий момент версии GeoPhyCell поддерживается одномерное разбиение данных с возможностью двойной буферизации (double buffering), когда очередная порция данных загружается в память сопроцессора одновременно с ведением расчётов и пересылкой результатов по предыдущей порции (рис. 5). Для сеток с большим количеством узлов запланирована поддержка трёхмерного разбиения, которое за счёт отсутствия обменов между сопроцессорами фактически сводится к одномерному.

Для тестового эксперимента по переносу примеси от мгновенного источника заданным потоком ветра на сетке 25 млн. точек производительность при использовании 16 сопроцессоров в 13-15 раз превосходит производительность в случае использования одного сопроцессора и в 3-5 раз – вычисления на двух ядрах *AMD Athlon* 64 X2.

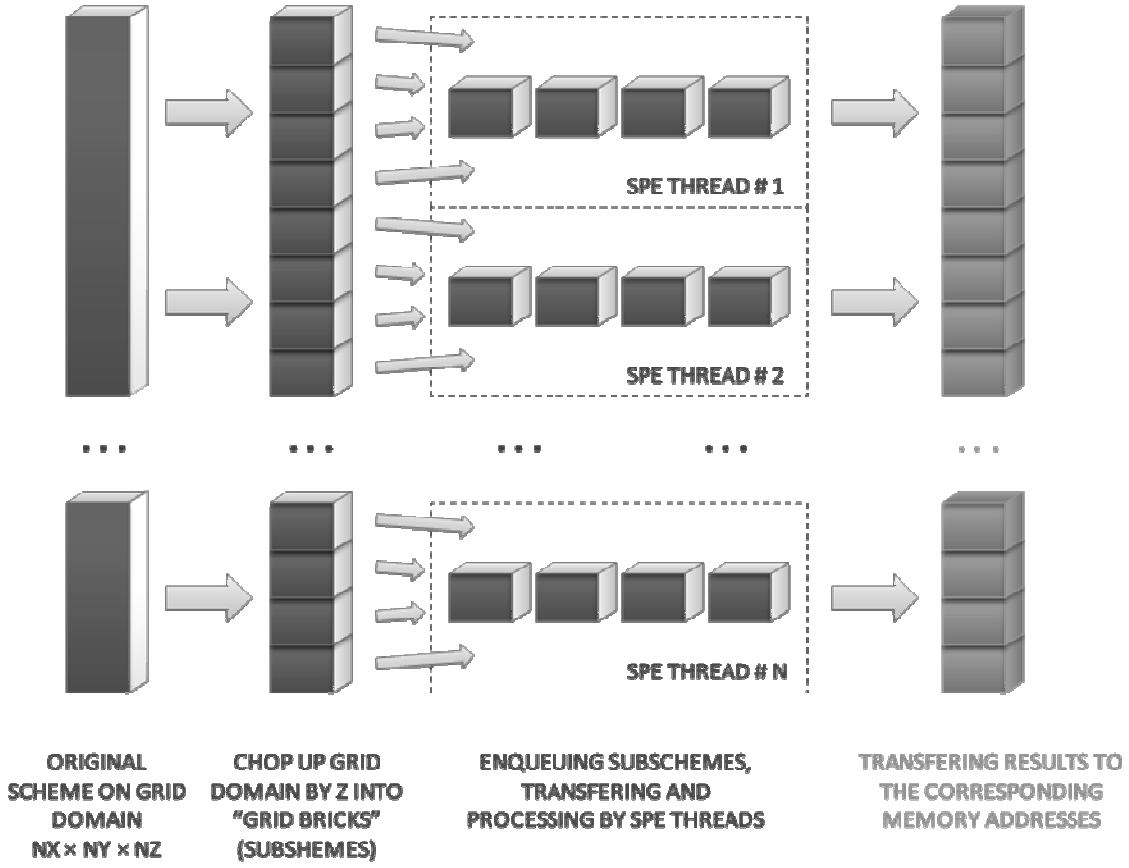


Рис. 5. Параллельные вычисления для *CellBE* с одномерным распределением расчётной области

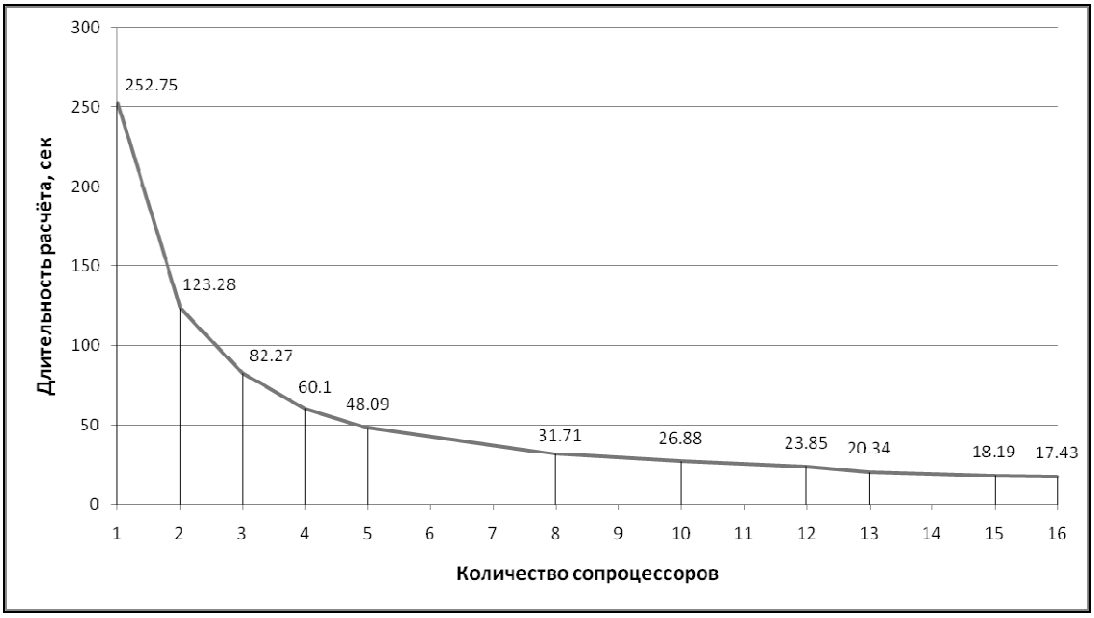


Рис. 6. Производительность GeoPhyCell на сервере *IBM QS22 Blade* в тестовом эксперименте на сетке $25*10^6$ точек

4. Заключение

В настоящей работе представлено два варианта адаптации региональной атмосферной модели *NH3D*, развиваемой в НИВЦ МГУ, для гибридных многопроцессорных многоядерных вычислительных систем. В первом варианте реализована двумерная декомпозиция расчетной области и явная передача сообщений по стандарту *MPI*. На примере решения трех уравнений движения показана высокая эффективность программы, в частности, продемонстрировано суперлинейное ускорение на суперкомпьютере СКИФ-МГУ «Чебышев». Кроме того, реализованы две схемы паралелльного решения эллиптического уравнения для геопотенциала с помощью быстрого преобразования Фурье. Второй вариант основан на пакете примитивов *GeoPhyCell*, предназначенном для распределения вычислений в существующих программах для систем с общей памятью и *Cell Broadband Engine Architecture* (*CBEA*) с обратной совместимостью с интерфейсом MPI. На данный момент пакет позволяет оснащать поддержкой параллельности широкий класс численных алгоритмов гидродинамики, таких как схемы уравнений переноса скалярных или векторных величин и вычислительно эквивалентные им. Ускорение распределённых вычислений в задаче переноса пассивной примеси на сервере *IBM QS22 Blade* составило 13-15 при использовании 16 сопроцессоров.

Литература

1. Miranda, P. M. A., and I. N. James. Non-linear three-dimensional effects on gravity wave drag: Splitting flow and breaking waves. Quart. J. R. Met. Soc., Vol. 118, 1992, pp. 1057-1082.
2. Степаненко В. М., Микушин Д. Н. Численное моделирование мезомасштабной динамики атмосферы и переноса примеси над гидрологически неоднородной поверхностью. – Вычислительные технологии, 2008, т. 13, вып. 3, с. 103-110.
3. Степаненко В. М., П. М. Миранда, В. Н. Лыкосов. Численное моделирование мезомасштабного взаимодействия атмосферы и гидрологически неоднородной суши. – Вычислительные технологии, 2006, т. 11, вып. 3, с. 118-127.
4. Danilkin E. A., and A. V. Starchenko. Numerical method for solution of the equations of an atmospheric boundary layer aerothermodynamics on MCS with the distributed memory// Abstracts of the International Conference on Environmental Observations, Modeling and Information Systems ENVIROMIS-2008, 28 June – 5 July 2008, Tomsk, Russia, p. 59.
5. Глазунов А. В. Численное моделирование сдвиговой турбулентности с использованием параллельных вычислений на компьютерах с распределенной памятью // Параллельные вычислительные технологии: Труды международной научной конференции (29 января — 2 февраля 2007 г., г. Челябинск). -Челябинск: Изд-во ЮУрГУ. - 2007. -Т. 2. -сс. 179-192
6. Abraham Arevalo et al, Programming the Cell Broadband Engine™: Architecture Examples and Best Practices, IBM Red-Books, Aug. 2008