Параллельный программный комплекс модели атмосферы для прогноза погоды и моделирования климата*

М.А. Толстых[1,2], Р.Ю. Фадеев[1], В.Г. Мизяк[2]

Институт вычислительной математики РАН[1], Гидрометцентр России[2]

Программный комплекс новой версии модели атмосферы ПЛАВ предназначен как для прогноза погоды, так и для моделирования изменений климата. Разработана система параллельного ввода-вывода, которая может подключаться как отдельный программный компонент, а также может задействовать некоторые из вычислительных процессов для выполнения операций чтения-записи. Система параллельного ввода-вывода внедрена в модель ПЛАВ и систему усвоения данных наблюдений атмосфере на основе ансамблевого фильтра Калмана. В программном комплексе модели атмосферы выполнены работы по повышению масштабируемости за счет увеличения количества нитей OpenMP, а также оптимизации обращений в оперативную память.

1. Введение

Менее десятка стран в мире развивают собственные технологии моделирования глобальной атмосферы (США, Англия, Канада, Франция, Япония, Германия, Австралия, Китай и Россия). Чем выше разрешение модели атмосферы (и, следовательно, размерность решаемой задачи), тем точнее мы можем описать рельеф земной поверхности (орографию) и его взаимодействие с атмосферным потоком. Более высокое разрешение модели атмосферы позволяет точнее описывать каскад энергии по спектру атмосферных движений, а в некоторых случаях перейти от параметрического к явному описанию атмосферных явлений. Это, в свою очередь, способствует уменьшению ошибок прогноза.

Численный прогноз погоды с высоким пространственным разрешением требует больших вычислительных ресурсов, в первую очередь в связи с тем, что оперативный прогноз налагает ограничение на допустимое время счета модели – не более 10–20 минут на расчет прогноза на 24 часа. Кроме того, в системе усвоения данных на каждом цикле необходимо рассчитывать порядка 100 краткосрочных прогнозов, что также требует больших вычислительных ресурсов. Размерность вычислительной области в современных глобальных моделях прогноза погоды составляет порядка 10^8 ($1000 \times 1000 \times 100$), что определяется необходимостью разрешения мезомасштабных синоптических процессов. Практические реализации таких моделей в мировых центрах используют порядка тысяч процессоров. Перспективные модели атмосферы должны будут иметь горизонтальное разрешение порядка нескольких километров. Развитие моделей численного прогноза погоды будет требовать эффективного использования десятков и сотен тысяч процессорных ядер. Поэтому программный комплекс прогностической модели атмосферы должен хорошо масштабироваться на компьютерах с массивно-параллельной архитектурой. Уменьшение ошибок прогноза может быть достигнуто за счет совершенствования описания (параметризации) процессов подсеточного масштаба, таких как приземный пограничный слой, и повышения пространственного разрешения. Однако повышение пространственного разрешения модели неизбежно влечет уменьшения шага интегрирования по времени.

Высокой скорости расчетов в задачах моделирования атмосферы можно достичь при использовании полунеявного полулагранжева подхода к дискретизации уравнений динамики. Полулагранжев подход [1] заключается в рассмотрении движения набора лагранжевых (перемещающихся вместе с воздухом) частиц, приходящих в конце шага по времени в узлы фиксированной вычислительной сетки и дискретизации уравнений на траекториях этих частиц. В полунеявной схеме интегрирования по времени линейные части уравнений динамики атмосферы, отвечающие за распространение быстрых инерционно-гравитационных волн, интегрируются по абсолютно устойчивой неявной схеме. Таким образом, шаг по времени в полунеявных полула-

* Работа выполнена при поддержке Программы фундаментальных исследований 43 Президиума РАН.

гранжевых моделях не ограничен условием устойчивости Куранта ни по скорости ветра, ни по скорости распространения инерционно-гравитационных волн. На практике, при сравнимой точности, шаг по времени полунеявных полулагранжевых моделей в 3-5 раз больше, чем в моделях, использующих другие численные методы.

В настоящее время полулагранжев подход применяется в большинстве оперативных моделей среднесрочного прогноза погоды, в т.ч. в моделях Европейского центра среднесрочного прогноза погоды и Английской метеослужбы, которые являются мировыми лидерами в области среднесрочного прогноза погоды.

Программный комплекс отечественной глобальной модели атмосферы ПЛАВ позволял эффективно использовать до примерно 1000 ядер (при горизонтальном разрешении около 20-25 км), однако, как мы видим, необходимо повышение масштабируемости программного комплекса модели. Работы в этом направлении представлены в настоящей статье. Во втором разделе представлено краткое описание модели ПЛАВ, третий раздел посвящен описанию системы параллельного ввода-вывода. В четвертом разделе описываются работы по повышению эффективности распараллеливания с помощью технологии OpenMP, а в разделе 5 - оптимизации обращений в оперативную память.

2. Глобальная модель атмосферы ПЛАВ

Основной моделью глобального среднесрочного прогноза погоды в России с 2010 года является глобальная полулагранжева модель атмосферы ПЛАВ (ПолуЛагранжева, основанная на уравнении Абсолютной заВихренности) [2]. Блок решения уравнений динамики атмосферы разработан в Институте вычислительной математики РАН и Гидрометцентре России. В модели ПЛАВ наряду с блоком решения уравнений динамики атмосферы собственной разработки в основном применяются алгоритмы параметризаций процессов подсеточного масштаба, разработанные под руководством Ж.-Ф.Желена возглавляемым Францией консорциумом по мезомасштабному прогнозу погоды ALADIN/LACE [3, 4]. В модель также включена отечественная параметризация крупномасштабных осадков [5] и модель многослойной почвы [6]. В современную версию модели также входят свободно распространяемые параметризации коротковолновой и длинноволновой радиации (CLIRAD [7] и RRTM [8] соответственно).

Оригинальными особенностями блока решения уравнений динамики атмосферы модели ПЛАВ являются применение конечных разностей четвертого порядка на несмещенной сетке для аппроксимации неадвективных слагаемых уравнений и использование вертикальной компоненты абсолютного вихря и дивергенции в качестве прогностических переменных. Существенным элементом для модели атмосферы, основанной на переменных «вертикальный компонент абсолютной завихренности – горизонтальная дивергенция», является быстрый и точный алгоритм восстановления компонент горизонтальной скорости ветра, описанный в [9]. Численные методы, применяемые в модели ПЛАВ, на тестовых задачах не уступают в точности спектральному методу решения уравнений динамики атмосферы, что было показано в [9].

Описание программной реализации модели на основе сочетания технологий MPI и OpenMP (гибридной технологии) описывается в [10]. Отметим, что гибридный подход был впервые в России применен к реальному сложному программному комплексу.

Код модели атмосферы ПЛАВ был проверен на масштабируемость на вычислительных системах РСК Торнадо и МВС-10п (установлена в Межведомственном суперкомпьютерном центре РАН). Разрешение современной версии модели составляет 0,225 градуса по долготе, по широте шаг сетки изменяется от 0,18 градуса в Северном полушарии до 0,25 градуса в южном, по вертикали - 51 неравномерно расположенных сигма-уровней. Размеры расчетной области составляют при этом 1600x866x51. На рисунке 1 приведено параллельное ускорение модели по отношению к времени счета на 54 процессорных ядрах. При расчетах использовалось 4 нити OpenMP. Общепринятой для моделей численного прогноза погоды и моделирования климата мерой эффективности является параллельное ускорение кода, равное 55-65 % от теоретического (при использовании тысяч процессоров).

Можно видеть, что модель ПЛАВ при данном пространственном разрешении эффективно масштабируется до 1152 ядер. При увеличении количества ядер от 432 до 864 наблюдается су-

перлинейное ускорение, что, по всей видимости, вызвано эффективным использованием кэш-памяти процессоров.

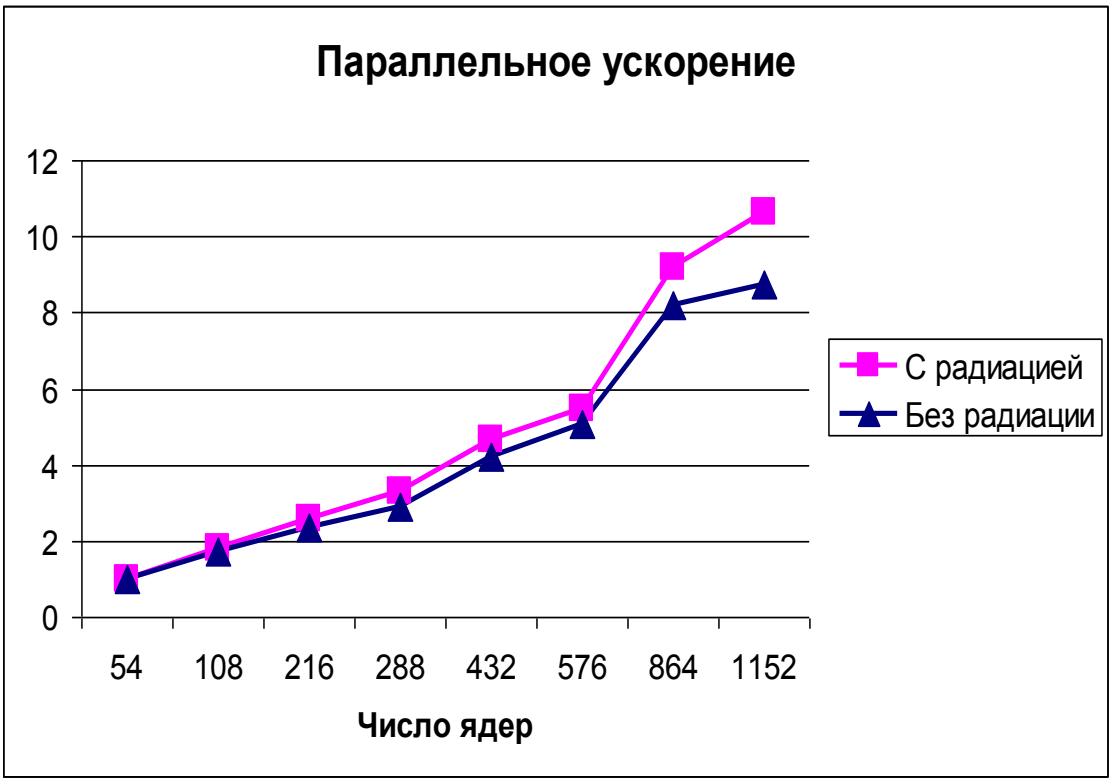

Рис. 1. Параллельное ускорение модели ПЛАВ по отношению к времени счета на 54 ядрах.

Технологии MPI и OpenMP использовались для параллельной реализации циклов по одной и той же координате – широте. Это ограничивало теоретический максимум количества используемых процессорных ядер количеством узлов сетки по широте. Реальный же максимум, с учетом большой ширины зависимости по данным в полулагранжевой модели ограничен величиной Nlat/4 (Nlat – число узлов сетки по широте), что для горизонтального разрешения используемой сетки по широте порядка 20 км дает значение максимума в 288 процессоров. С учетом использования четырех нитей OpenMP, это ограничивает число используемых процессорных ядер величиной около 1000. Кроме того, размещение рабочих массивов отдельно для каждой нити увеличивало необходимую для каждого параллельного MPI процесса память рабочих массивов, что, помимо физических ограничений на вычислительную систему, затрудняло локализацию обращений в память из заданного процессорного ядра. При высоком пространственном разрешении при использовании уже нескольких сотен процессоров в реальной оперативной технологии, предусматривающей вывод прогностической продукции каждый час модельного времени, производительность модели сильно снижалась. Это было вызвано блокирующей системой ввода-вывода, реализованной путем централизации операций чтения-записи на мастер-процессе.

3. Система параллельного ввода-вывода

3.1 Описание системы

Для новой версии глобальной модели атмосферы ПЛАВ с горизонтальным разрешением над территорией России около 20 км разработана система параллельного ввода-вывода, которая заменила собой алгоритм взаимодействия с файловой системой, основанный на мастер-процессе.

Разработанная система реализует возможность выполнения операций чтения-записи как вычислительными MPI процессами, так и дополнительными (не расчетными) процессами. По-

добный подход позволяет адаптировать систему ввода-вывода под особенности конкретной задачи. В случае относительно редкого обращения к файловой системе (запись промежуточных результатов и контрольных точек модели) используется некоторая часть вычислительных процессов. Далее такие вычислительные процессы мы будем называть гибридными. Использование дополнительных процессов, основной функцией которых является выполнение не блокирующих вычисления операций чтения и записи, становится актуальным, когда происходит частое обращение к дисковому пространству (например, отладка программного кода или тестирование модели). Промежуточный вариант, при котором операции ввода-вывода осуществляются как гибридными (вычислительными), так и дополнительными процессами может применяться в случае неоднородной вычислительной среды или в задачах с неоднородно распределенной структурой данных.

Система имеет 6 основных методов: инициализация, регистрация файла и данных в системе, чтение и запись данных и синхронный останов. Функция инициализации имеет следующий интерфейс:

```
call pio % init (type, tag_rage, pio_cw, local_cw, hybrid_np_min, hybrid_np_max, mem_tot, print_lev).
```

Здесь:

- `type` характеризует данный MPI-процесс (внешний или вычислительный);
- `tag_range` - целочисленный массив из двух элементов задающий пределы MPI тагов, доступных системе;
- `pio_cw` и `local_cw` - глобальный коммуникатор системы и локальный, соответствующий внешним или вычислительным процессам в зависимости от значения аргумента `type`;
- `hybrid_np_min, hybrid_np_max` - минимальное и максимальное число гибридных процессов;
- `mem_tot` и `print_lev` - общий размер доступной оперативной памяти и уровень диагностического вывода.

Следует отметить, что процедуре инициализации системы может предшествовать операция `MPI_Comm_split`, разделяющая глобальный коммуникатор на два: внешний и вычислительный.

Высокое пространственное разрешение современных численных моделей и большой объем расчетных данных означает необходимость производить чтение и запись не только глобального массива целиком, но и его отдельной части, которая может использоваться, например, в целях диагностики модели. Это означает, что данные для записи или чтения могут располагаться не на всех вычислительных узлах, а только на некоторой их части. Метод системы, ответственный за регистрацию массива данных имеет интерфейс, представленный ниже:

```
call pio % data_push (id, name, data, data_dim_range, save_dim_range, mem_use).
```

Здесь:

- `id` - уникальный идентификатор регистрируемых данных;
- `name` - идентификатор данных в файле;
- `data` - многомерный массив данных (двойной или одинарной точности);
- `data_dim_range` - массив, содержащий информацию о локальных пределах массива `data` (уникальных для данного MPI-процесса и дополнительных, которые могут использоваться в расчетном комплексе для вычисления производных, например);
- `save_dim_range` - массив, включающий глобальные пределы массива в файле;
- `mem_use` - ориентировочный объем оперативной памяти на момент выполнения операций чтения-записи с регистрируемым массивом данных (необязательный аргумент).

Важной особенностью разработанной системы ввода-вывода является динамическая адаптация числа ответственных за работу с файловой системой процессов, которая достигается за счет использования возможностей современного языка программирования Фортрана-2003. На рисунке 2 представлена диаграмма основных программных объектов реализованной системы.

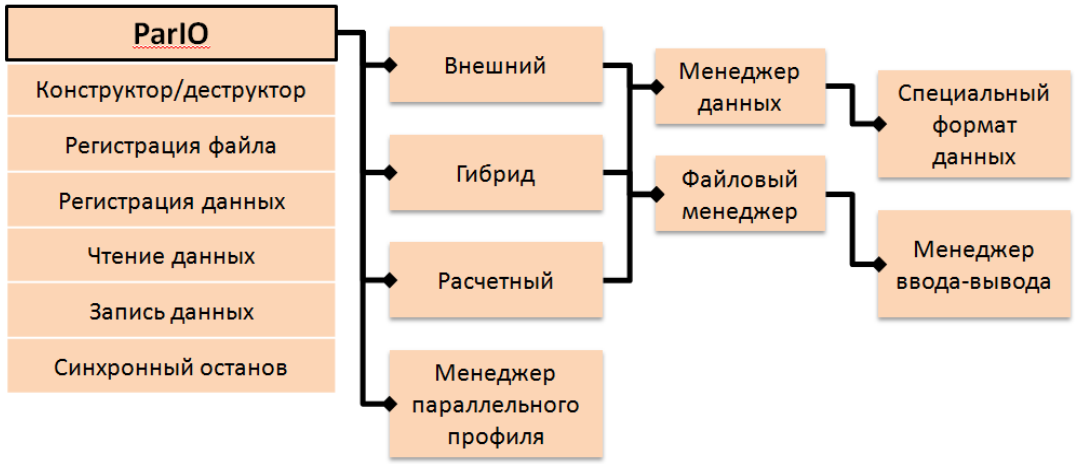

Рис. 2. Структура разработанной системы ввода-вывода: основные объекты и методы, доступные пользователю.

Для регистрируемого в системе ввода-вывода массива менеджером параллельного профиля (см. рисунок 2) создается профиль, который содержит информацию о взаимодействующих с файловой системой MPI процессах (внешний и гибридный) и вычислительных процессах, которые пересылают им (или получают от них) данные. Поскольку глобальные пределы индексов массива данных в файле (переменная `save_dim_range`) могут быть разными то параллельный профиль создается для каждого регистрируемого массива данных. Генерация профиля осуществляется с учетом объема доступной оперативной памяти и количества MPI процессов, которые содержат необходимые данные. Если общее число внешних процессов считается фиксированным, то число гибридных процессов может меняться в заданных пользователем пределах. Более того, MPI процесс расчетного коммуникатора для одних данных может относиться к числу вычислительных процессов, а для других - к гибридным.

Параллельный профиль системы ввода-вывода хранится в виде целочисленной квадратной матрицы M размерности *N*, где *N* - общее число MPI процессов (как вычислительных, так и дополнительных). Не равное нулю значение диагонального элемента матрицы M_{ii} указывает на то, что процесс *i* в глобальном MPI коммуникаторе обращается к файловой системе, а процессы отправляющие или получающие от него данные соответствуют ненулевым элементам матрицы в той же строке M_{ij}. Значение диагонального элемента матрицы M_{ii} определяет, в том числе, является ли соответствующий MPI процесс внешним (M_{ii}=1) или гибридным (M_{ii}=2).

Если число взаимодействующих с файловой системой MPI процессов становится больше количества вычислительных процессов с данными, которые надо прочитать или записать, то часть процессов переходит в режим ожидания. Подобная ситуация может реализоваться в том случае, когда размер массива для записи (чтения) относительно мал и, например, располагается целиком на одном MPI процессе. В предельном случае данные могут отвечать одной расчетной точке. Дополнительные MPI процессы в режиме ожидания распределяются между активными процессами системы ввода-вывода и ждут от них сигнала (об останове или другой операции). Эти взаимосвязи определяются столбцами матрицы M: не равные нулю элементы матрицы M_{ji} задают процессы *j* в режиме ожидания и зависимые от процесса *i*.

Фиксированный размер параллельного профиля позволяет распространять его всем MPI процессам глобального коммуникатора за одну операцию. Размер матрицы профиля не играет существенной роли, поскольку операция пересылки выполняется только один раз, а процедуры регистрации массива, обычно, выполняются гораздо реже, чем операции его чтения или записи на диск.

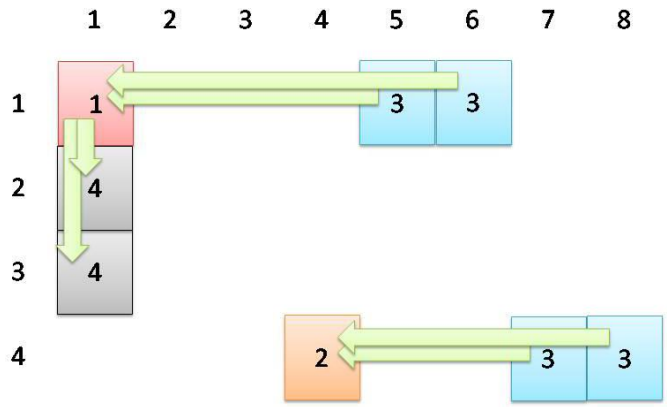

Рис. 3. Первые 4 строки матрицы M, соответствующей параллельному профилю системы ввода-вывода, при котором MPI процессы 1-3 являются внешними, 4 - гибридными, а 5-8 - вычислительными.

На рисунке 3 иллюстрируются первые 4 строки матрицы, соответствующей параллельному профилю системы ввода-вывода с 3 внешними процессами, один из которых является лидером (значение 1), а два остальных находятся в режиме ожидания (значение 4). Из 5 вычислительных MPI процессов один является гибридным (значение 2). Таким образом, матрица M содержит информацию о каждом MPI процессе и его взаимосвязи с другими MPI процессами. Отметим, что остальные 4 строки матрицы, равно как и незаполненные ячейки матрицы равны нулю.

Количество вычислительных процессов и их соответствие конкретному внешнему или гибридному процессу определяется из условия минимизации числа пересылок и балансировки объема передаваемых данных. Для этого каждому MPI процессу сопоставляется некоторый вес, характеризующий объем уникальных расчетных данных этого процесса. Процедура разделения вычислительных и гибридных узлов на заданное число групп с приблизительно равным весом проводится на основе алгоритма, в основе которого находится метод K-дерева [11].

Рисунок 4 иллюстрирует способ объединения локальных данных в группы для записи в параллельном режиме для некоторой задачи с размером глобального коммуникатора 3600 MPI процессов в случае неоднородно-распределенных данных и двумерной декомпозиции расчетной области. На рисунке 4 слева иллюстрируется вес каждого вычислительного и гибридного MPI процесса. На рисунке 4 справа цветом показан результат применения алгоритма разделения процессов на группы с приблизительно равным общим весом. В данном случае отклонение веса каждой группы от среднего по всем группам величины не превышает 8%.

Каждая группа на рисунке 4 (справа) представляет собой прямоугольник (в случае одномерной декомпозиции расчетной области это будет отрезок, а для трехмерной декомпозиции - параллелепипед). Прямоугольная (и выпуклая) структура конкатенированных данных позволяет производить запись за одну процедуру без использования условных операторов.

Методы системы ввода-вывода, ответственные за регистрацию файла, чтение и запись данных, имеют простой интерфейс:

```
call pio % file_push(name, create, nrec),
call pio % data_read(filename, id, recn).
```

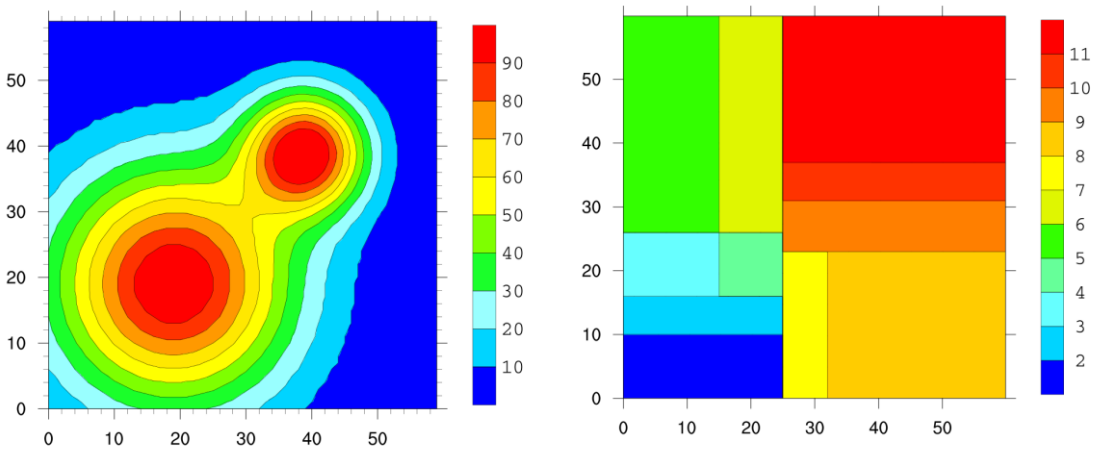

Рис. 4. Слева: вес каждого вычислительного и гибридного MPI процесса (общим числом 3600) в случае двумерной декомпозиции расчетной области. Права: цветом иллюстрируются те же MPI процессы, объединенные в 12 групп.

Здесь `name` - имя файла (он же уникальный идентификатор файла), логическая переменная `create` определяет создавать ли новый файл или записывать в конец уже имеющегося, `nrec` - задает число записей в файл (необходимо для эффективной работы библиотеки netCDF, например). В процедуре чтения данных аргумент `filename` отвечает названию файла, а `id` - уникальному идентификатору данных, `nrec` - номер записи (по времени, например).

В системе также реализована функция `stop`, которая отвечает за синхронный останов внешнего и расчетного коммуникаторов после окончания процедур чтения и записи.

Таким образом, разработанная система характеризуется достаточно простым интерфейсом доступных пользователю функций, но при этом обладает гибкостью в настройке и динамической адаптацией к формату, размеру и распределению расчетных данных.

Процедуры работы с файлами (открытие, создание и закрытие), чтения и записи данных реализованы в виде методов одного объекта (менеджер ввода-вывода на рисунке 2). Более того, функции чтения и записи оперируют данными, полученными от вычислительных узлов и объединенные в единый массив. Таким образом, пользователь имеет возможность производить чтение и запись удобными для него методами. Недостатком такого подхода является необходимость синхронизации метаданных, необходимых для выполнения операций чтения и записи (идентификаторы осей в netCDF, например), в случае использования гибридных процессов и не глобальных массивов.

Отметим, что система ввода-вывода была разработана для вычислительных моделей с прямоугольной (и выпуклой) структурой данных. Такой вид данных является существенным в двух процедурах: конкатенация данных (поиск соседних по данным MPI процессов) и их последующая запись. В то же время, система относительно легко может быть модернизирована для вычислительных моделей, использующих сетки с нерегулярной структурой. Так, например, новая версия глобальной модели атмосферы ПЛАВ использует редуцированную широтно-долготную сетку, в которой число узлов по долготе зависит от широты.

Разработанная система ввода-вывода позволяет оптимизировать процесс работы с файловой системой в зависимости от типа решаемой задачи: отладка кода и тестирование новых методов, оперативный прогноз погоды, работающий совместно с системой циклического усвоения метеоданных с шагом 6 часов и моделирование изменений климата на масштабах времени в десятки лет.

В настоящее время система параллельного ввода-вывода подключена к модели атмосферы ПЛАВ и верифицирована. Ведутся работы по оптимизации параметров системы, например, количества используемых для ввода-вывода ядер в различных конфигурациях модели (оперативный прогноз погоды и моделирование изменений климата).

3.2 Параллельный ввод/вывод в системе подготовки данных

Описываемая система параллельного ввода/вывода была также применена для организации чтения и записи в разрабатываемой перспективной системе подготовки начальных данных для модели ПЛАВ на основе Локального Ансамблевого Фильтра Калмана с преобразованием Ансамбля [12]. Это позволило несколько изменить организацию вычислений в параллельной программе, что в свою очередь привело к экономии задействованной памяти, увеличению масштабируемости, возможности применения двумерной декомпозиции расчётной области для вычислений с помощью MPI и улучшению балансировки загрузки вычислительных ресурсов.

Оригинальный алгоритм LETKF описан в [13]. Ключевым достоинством этого метода подготовки данных для прогностических моделей является полная независимость производимых операций в каждой точке модельного пространства от операций в остальных точках. Этот факт обуславливает возможность эффективного применения распараллеливания алгоритма по данным. Максимально возможное используемое для вычислений количество MPI процессов составляет произведение узлов горизонтальной сетки на количество уровней. Отсутствующий до недавнего времени параллельный ввод/вывод данных мешал эффективному применению этих вычислительных ресурсов для получения результата.

Параллельный ввод/вывод позволяет производить декомпозицию расчётной области на широтно-долготные прямоугольники ещё до начала этапа чтения полей прогнозов и получать свою порцию начальных данных каждым MPI процессом. Все дальнейшие вычисления и операции, включая запись полученных результатов в файлы, совершаются независимо от других MPI процессов. Таким образом, работа параллельной программы происходит без использования буферных массивов, необходимых для чтения и записи при последовательном вводе/выводе, что при максимальном разрешении модели атмосферы ПЛАВ 1600x866x51 для ансамбля из 40 прогнозов при 5 прогностических переменных способно сэкономить порядка 10^3 Гб памяти суммарно на всех используемых вычислительных узлах. Кроме того, все производимые операции не требуют барьеров и синхронизаций.

Для достижения оптимального баланса нагрузки на вычислительные узлы следует применять неравномерное распределение размеров широтно-долготных прямоугольников, обрабатываемых каждым MPI процессом. Общее количество получаемых данных различно из-за неравномерного распределения используемых в усвоении данных метеорологических наблюдений. Максимальное их количество имеется над Европой, Северной Америкой и некоторыми районами Азии. Практически полностью отсутствуют наблюдения над значительными частями мирового океана и в Антарктиде. Поэтому более насыщенные наблюдениями области должны быть меньшего размера, чем те, в которых количество наблюдений мало.

Сочетание оптимального разбиения размеров вычислительных подобластей и полностью параллельной работы программы ведут к увеличению масштабируемости. Таким образом, рост количества используемых вычислительных ресурсов будет приводить к пропорциональному уменьшению времени работы программы, что особенно важно в условиях возрастающего разрешения используемых полей и ограниченного времени, выделяющегося на работу оперативных версий приложений.

4. Повышение масштабируемости кода модели ПЛАВ

Программный комплекс полулагранжевой модели атмосферы ПЛАВ [2, 10] состоит из блока решения уравнений динамики атмосферы и набора параметризаций процессов подсеточного масштаба. В настоящее время общий объем кода с комментариями превысил 100000 строк. В свою очередь, в блоке решения уравнений динамики выделяются явные вычисления в сеточном пространстве, с заметным шаблоном зависимости по данным по горизонтали, а также вычисления в пространстве коэффициентов Фурье по долготе. В этом блоке имеются рекурсивные зависимости по широте, однако по волновым числам (по долготе) и по вертикальной координате зависимостей не имеется. В наборе параметризаций процессов подсеточного масштаба (солнечная радиация, вертикальная диффузия и пр.) расчеты ведутся независимо для любых точек горизонтальной сетке, однако во многих параметризациях имеются рекурсивные зависимости по вертикали.

Для разных блоков применяются различные подходы. В блоке явных вычислений динамики и наборе параметризаций процессов подсеточного масштаба, почти все программные модули были ранее организованы так, чтобы обрабатывать весь круг широты (при фиксированной широте). Такая организация вычислений была естественной для одномерной декомпозиции расчетной области по широте. Теперь все программные модули в указанных блоках организованы так, чтобы они могли обрабатывать произвольную часть круга широты. Для каждого процесса MPI организован цикл по нитям OpenMP, который является внешним по отношению к существующему циклу по широте. Каждая нить обрабатывает свою полосу долгот. Данная организация вычислений представлена на рисунке 5 (слева).

Для блока, выполняющего вычисления в пространстве коэффициентов Фурье по долготе (решения эллиптического уравнения, восстановление компонент скорости), дополнительно к распараллеливанию по полосе волновых чисел, обрабатываемой каждым MPI процессом независимо, ранее было реализовано распараллеливание по OpenMP по тем же полосам. Это также ограничивало максимально возможное количество используемых процессорных ядер. Чтобы повысить количество используемых процессорных нитей, вместо дополнительного распараллеливания цикла по волновым числам, OpenMP применяется теперь для распараллеливания циклов по вертикальной координате (рисунок 5, справа).

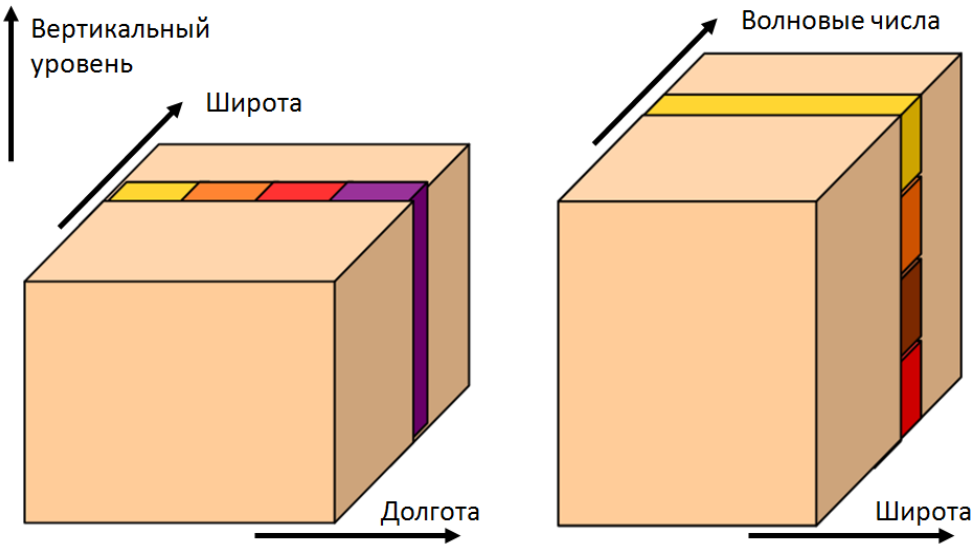

Рис. 5. Слева: распараллеливание по долготе в вычислениях «явной динамики» в случае 4 нитей OpenMP. Справа: распараллеливание по вертикальной координате в вычислениях в пространстве коэффициентов Фурье в случае 4 нитей OpenMP. Разные цвета соответствуют разным нитям OpenMP.

Выполнение данных работ уменьшило необходимый объем рабочих массивов, особенно заметный в блоке параметризаций процессов подсеточного масштаба. В этом блоке рабочие массивы для каждого MPI процесса ранее имели примерный объем *Nlon x Nlev x Nopenmp* x 100 (значения в современной версии модели: количество узлов сетки по долготе *Nlon*= 1600, количество уровней по вертикали *Nlev*=51). Теперь же независимо от количества нитей OpenMP этот объем составляет *Nlon x Nlev* x 100. Экономия достигнута и в других блоках модели, однако в блоке «явной» динамики объем рабочих массивов в несколько раз меньше, а в блоке вычислений в пространстве коэффициентов Фурье – еще в несколько раз меньше, чем в блоке «явной» динамики.

После реализации данных изменений в коде, модель была проверена на численном среднесрочном прогнозе погоды. Ошибки прогноза практически не изменились. Теперь программный комплекс модели может использовать как минимум 16 нитей OpenMP на один MPI процесс на тех же вычислительных узлах.

5. Оптимизация использования памяти в модели

Представленные в предыдущем разделе работы по повышению максимального количества используемых процессоров замедляют выполнение программы примерно в 2 раза при использовании четырех нитей OpenMP. Как и следовало ожидать, если не использовать OpenMP, скорость расчета по исходному и модифицированному коду совпадают. Причиной является неоптимальный доступ к данным. Действительно, исходный код модели на языке Фортран имел расположение индексов в глобальных массивах *(все долготы, все вертикальные уровни, широты для данного MPI-процесса),* а в локальных массивах блоках «явной» динамики и параметризаций - *(все долготы, все вертикальные уровни)*. Таким образом, при вызове любой из программных процедур этих блоков при реализации описанного в предыдущем разделе распараллеливания OpenMP по долготе осуществлялось копирование переменных, описывающих состояние атмосферы на текущем шаге по времени, в подмассив размерности (*долгота1:долгота2, все вертикальные уровни)* из глобального массива с упомянутой размерностью. Такое копирование должно осуществляться одновременно для каждой нити OpenMP (см. рисунок 4). Учитывая организацию памяти, все нити будут одновременно обращаться в маленький блок памяти, что и объясняет полученное замедление кода.

Для оптимизации доступа в память порядок индексов в глобальных массивах изменен на *(все вертикальные уровни, вся долготы, широты для данного MPI процесса)*. В части программных модулей (примерно 30 % от общего объема строк) изменен порядок индексов в локальных массивах, так, что внутренний векторизуемый цикл реализован по вертикальной координате. Изменены только модули, в которых циклы по вертикальной координате не имеют зависимостей и могут быть векторизованы. Таким образом, выполненные изменения не должны ухудшить хорошую векторизуемость исходного кода.

Выполненные изменения означают, что при вызове любой расчетной процедуры порции глобальных массивов передаются по адресу непосредственно, и неявное (либо явное) копирование в промежуточный массив не требуется. Изменение порядка индексов также позволило локализовать доступ память для данной нити OpenMP.

Блок параметризаций процессов подсеточного масштаба с такой организацией массивов был затем подключен к модели и успешно проверен на совпадение результатов модельных среднесрочных прогнозов погоды. В настоящее время проводится тестирование масштабируемости усовершенствованного программного комплекса модели.

Отметим, что в ранней версии модели ПЛАВ, при вертикальной размерности равной 28, векторизация по вертикали особого смысла не имела. В современной версии модели 51 уровень по вертикали, в ближайшие несколько лет планируется повысить количество уровней по вертикали до 90-100.

6. Выводы

Разработанная система параллельного ввода-вывода позволяет снизить время расчета оперативного прогноза погоды высокого разрешения с помощью модели атмосферы ПЛАВ при частой записи выходной прогностической продукции. Это достигнуто за счет выделения отдельных процессоров для обработки операций ввода-вывода и сжатия записываемой информации.

Применение разработанной системы параллельного ввода-вывода позволило также увеличить количество эффективно используемых вычислительных ядер при решении задачи подготовки начальных данных для модели атмосферы ПЛАВ, уменьшив при этом количество требуемой приложению памяти. За счёт этого для получения более качественных начальных данных можно использовать такие опции и параметры системы усвоения, которые требуют больших вычислительных затрат.

В ходе работ в блоках «явной» динамики модели ПЛАВ реализовано распараллеливание с помощью OpenMP по долготе, а для блока, выполняющего вычисления в пространстве коэффициентов Фурье по долготе (решения эллиптического уравнения, восстановление компонент скорости) - по вертикальной координате.

Таким образом, модифицированный программный комплекс полулагранжевой модели атмосферы ПЛАВ при данном разрешении 20-25 км теперь способен использовать вместо четырех нитей OpenMP максимально возможное количество нитей, определяемое архитектурой используемой вычислительной платформы. На данный момент проверена работа программного комплекса при количестве нитей 16. Максимально возможное число эффективно используемых процессов MPI пока что не изменилось и составляет 288 для размерности по широте 865. Это означает потенциальное повышение масштабируемости всего программного комплекса модели ПЛАВ при данном разрешении с 1000 до 4000 ядер.

В дальнейшем предполагается выполнить работы по замене прямого солвера эллиптического уравнения типа Гельмгольца на итерационный, и реализовать двумерную декомпозицию расчетной области. Это позволит повысить как минимум на порядок количество используемых процессов MPI.

Литература

1. Staniforth A., Côté J. Semi-Lagrangian integration schemes for atmospheric models. A review // Mon. Weather Rev. 1991. V. 119. P. 2206–2223.
2. Толстых М.А. Глобальная полулагранжева модель численного прогноза погоды. М, Обнинск: ОАО ФОП , 2010. 111 стр.
3. De Troch R., Hamdi R., van de Vyver H., Geleyn J.-F., Termonia P. Multiscale Performance of the ALARO-0 Model for Simulating Extreme Summer Precipitation Climatology in Belgium // J. Climate. 2013. V. 26 P. 8895-8915.
4. Geleyn J.-F., Bazile E., Bougeault P., Deque M., Ivanovici V., Joly A., Labbe L., Piedelievre J.-P., Piriou J.-M., Royer J.-F. Atmospheric parameterization schemes in Meteo-France's ARPEGE N.W.P. model // Parameterization of subgrid-scale physical processes, ECMWF Seminar proceedings. - Reading, UK: 1994. P. 385-402.
5. Кострыкин С.В., Эзау И.Н. Динамико-стохастическая схема расчета крупномасштабных осадков и облачности // Метеорология и гидрология. 2001. № 7. С. 23-39.
6. Володин Е.М., Лыкосов В.Н. Параметризация процессов тепло- и влагообмена в системе растительность - почва для моделирования общей циркуляции атмосферы. 1. Описание и расчеты с использованием локальных данных // Известия РАН. Физика атмосферы и океана. 1998. Т. 34, № 4. С. 453-465.
7. Tarasova T., Fomin B. The Use of New Parameterizations for Gaseous Absorption in the CLIRAD-SW Solar Radiation Code for Models // J. Atmos. and Oceanic Technology. 2007. V. 24, № 6. P. 1157–1162.
8. Mlawer E.J., Taubman S.J., Brown P.D., Iacono M.J. and Clough S.A.: RRTM, a validated correlated-k model for the longwave radiation// J. Geophys. Res. 1997. V. 102, N 16, 663-16, 682.
9. Tolstykh M.A., Shashkin V.V. Vorticity–divergence mass-conserving semi-Lagrangian shallow-water model using the reduced grid on the sphere // J. Comput. Phys. 2012. V. 231. P. 4205-4233.
10. Толстых М.А., Мизяк В.Г. Параллельная версия полулагранжевой модели ПЛАВ с горизонтальным разрешением порядка 20 км. // Труды Гидрометеорологического научно-исследовательского центра Российской Федерации. 2011, вып. 346. С. 181-190.
11. Bentley J.I. Multidimensional divide and conquer // Communications of the ACM. 1980. V 23, N 4. P 4-229.
12. Shlyaeva A.V., Tolstykh M.A., Mizyak V.G., Rogutov V.S. Local ensemble transform Kalman filter data assimilation system for the global semi-Lagrangian atmospheric model // Russ. J. Num. An. & Math. Mod. 2013. V 28, N 4. P 419-441.

13. Hunt B.R., Kostelich E.J., Szunyogh I. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter // Physica D: Nonlinear Phenomena, 2007. V. 230(1-2). P. 112–126.

Parallel program complex for numerical weather prediction and climate modeling

Mikhail Tolstykh, Rostislav Fadeev and Vassily Mizyak

Keywords: Parallel implementation of the global atmosphere model, Global atmosphere model, Numerical weather prediction, Climate changes modelling

The global atmosphere SL-AV model (Semi-Lagrangian, based on Absolute Vorticity equation) has been introduced into the operational practice at Hydrometeorological center of Russia in 2010. This allowed to reduce considerably the gap between Russia and the leading group of world prediction centers in medium-range weather forecasts. The new version of the SL-AV model is developed. This version has the horizontal resolution about 20 km over the Russia territory for numerical weather prediction, and it can be applied to the atmosphere forecast at time scales about ten days. This version is certified by Roshydromet recently. The same model version having coarser resolution is validated with a problem of climate change modelling using the AMIP2 protocol.

The program complex of the SL-AV model new version uses a combination of MPI and OpenMP parallel programming technologies and currently scales up to 1700 cores. To increase the file system efficiency, the parallel input-output system is developed that can be connected as a separate parallel program component. It can use if necessary some computing nodes for input/output operations. The parallel input-output system is introduced into the SL-AV model and observations data assimilation system based on the ensemble Kalman filter. Overall system performance is shown with the examples of numerical atmosphere modelling in various modes.

Also, the work on memory access optimization is carried out in the model program complex. This increased the parallelization efficiency.