УДК 004.942

А. И. Свитенков, Д. М. Спельников, В. Г. Маслов, А. В. Бухановский

ПАРАЛЛЕЛЬНОЕ РЕШЕНИЕ ЗАДАЧИ ХАРТРИ—ФОКА ДЛЯ МОЛЕКУЛЫ ГРАФЕНА: МАСШТАБИРУЕМОСТЬ И ГИПЕРЭФФЕКТИВНОСТЬ

Рассмотрен параллельный метод решения уравнения Хартри—Фока, основанный на алгоритме DC. Предложенный метод позволяет частично решать задачу самосогласования локально на вычислительных узлах, что служит сокращению накладных расходов.

Ключевые слова: *задача самосогласования, графен, параллельный алгоритм, уравнение Хартри—Фока, накладные расходы, квантовая химия, масштабируемость.*

Введение. Распространенным подходом к приближенному решению уравнения Шредингера является так называемое одноэлектронное приближение, при котором каждый электрон рассматривается как независимый, движущийся в усредненном поле ядер атомов и других электронов. Этот подход используется в методах Хартри—Фока, функционала плотности или приближении сильносвязанных электронов. Уточнение решения за счет учета корреляционной энергии электронов приводит к критическому росту вычислительной сложности и на практике возможно при моделировании молекулярных систем лишь из небольшого числа атомов [1].

Для моделирования свойств по возможности больших систем используются наиболее простые одноэлектронные *ab initio* или даже полуэмпирические приближения [2]. Однако традиционная постановка задачи и в этих случаях приводит к кубическому росту вычислительной сложности при увеличении числа атомов, что делает ее неприменимой для систем, содержащих 1000 атомов и более. В этой связи в настоящее время широко обсуждаются линейно масштабируемые методы, значительно расширяющие границы применения квантовой химии [3]. Однако увеличение размеров систем до десятков и сотен тысяч атомов дополнительно требует параллельной реализации указанных алгоритмов. В работе представлен опыт применения линейно масштабируемого метода “Divide-and-conquer” (DC) для квантово-химического уравнения Хартри—Фока. Моделированию подлежала электронная структура молекулярных соединений типа „графен“ и „графан“.

Специфика выбранных соединений позволяет наблюдать важные эффекты, связанные со сходимостью итерационного процесса самосогласования. При рассмотрении масштабируемости того или иного метода необходимо учитывать не только сложность выполнения одной итерации самосогласования, но и зависимость числа итераций от числа атомов [4].

Параллельный алгоритм, основанный на пространственной декомпозиции молекулярной системы, аналогичной декомпозиции в алгоритме DC, приводит не только к снижению времени выполнения одной итерации, но и к уменьшению числа итераций самосогласования, требуемых для сходимости. Относительно полного времени решения задачи Хартри—Фока, таким образом, наблюдается гиперускорение. Для объяснения описанного эффекта необходимо рассмотреть свойства уравнения Хартри—Фока и его решений.

Постановка задачи. Уравнение Хартри—Фока. В методе Хартри—Фока гамильтониан молекулярной системы в уравнении Шредингера заменяется приближенным одноэлектронным аналогом — фокианом. В таком случае приближенное уравнение Шредингера приобретает вид [5]:

$$-\frac{1}{2}\nabla^2{}_k\psi_k(\mathbf{r})+V\psi_k=\varepsilon_k\psi_k(\mathbf{r}). \quad (1)$$

Оно должно быть решено относительно волновых функции $\psi_k(\mathbf{r})$ в действительной области пространства в некоторой окрестности неподвижных центров атомных ядер. Здесь V — оператор эффективного потенциала, в котором движется электрон: $V=V_0+V_d+V_x$; V_0 описывает вклад взаимодействия электрона и атомных ядер; V_d и V_x описывают взаимодействие с остальными электронами системы:

$$\left.\begin{aligned} V_d\psi_j(\mathbf{r}) &\equiv \sum_i\left(\int\frac{|\psi_i(\mathbf{r}')|^2}{|\mathbf{r}-\mathbf{r}'|}dV'\right)\psi_j(\mathbf{r}),\\ V_x\psi_k(\mathbf{r}) &\equiv \sum_{j\neq k}\left(\int\frac{\psi_j^*(\mathbf{r}')\psi_k(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}dV'\right)\psi_j(\mathbf{r}). \end{aligned}\right\} \quad (2)$$

При разложении по выбранному набору базисных функций k-я орбиталь представляет собой линейную комбинацию базисных функций φ_i с неизвестными коэффициентами C_{ki} [6]:

$$\psi_k=\sum_{\mu=1}^{M}C_{ki}\varphi_i,$$

относительно которых уравнение Хартри—Фока может быть записано следующим образом:

$$\sum_{j=1}^{N}F_{ij}C_{kj}=\varepsilon_k C_{ki}, \quad (3)$$

где **F** — матрица фокиана в соответствующем базисном разложении. Выражение (3) представляет собой задачу на собственные числа и собственные векторы матрицы **F**. Матрица плотности **P** определяется выражением:

$$P_{ij}=\sum_{k=k_{occ}}C_{ki}C_{kj}. \quad (4)$$

Как видно, матрица фокиана, в свою очередь, зависит от матрицы плотности. Таким образом, уравнения (2)—(4) формируют так называемую самосогласованную задачу, решение которой сводится к итерационному процессу с последовательным уточнением матрицы **P** и соответствующей матрицы **F** до достижения сходимости.

Из уравнения (3) видна кубическая сложность предлагаемого алгоритма относительно размера матриц **P** и **F**, однако при использовании базиса сильно локализованных функций внедиагональные элементы матрицы плотности довольно быстро затухают с расстоянием. Это приводит к тому, что реальное число отличных от нуля матричных элементов оказывается ~N, а не ~N^2 [7]. То же относится и к матрице фокиана в этом представлении. Соответственно все действия с такими матрицами имеют трудоемкость ниже $O(N^3)$, что отражает локальный характер квантовой механики и так или иначе используется всеми линейно масштабируемыми алгоритмами решения задачи Хартри—Фока.

Алгоритм DC и его параллельная реализация. С помощью алгоритма DC общая матрица плотности строится на основе решения, полученного не для всей системы в целом, а для некоторых перекрывающихся фрагментов. Величина буферной зоны задается посредством величины отсечения S_{th} интегралов перекрывания базисных функций, значения меньше которой считаются нулевыми. Атомы, базисные функции которых не перекрываются, считаются не взаимодействующими непосредственно [8].

Для каждой подобласти проводится отдельный расчет субматрицы плотности. Объединяющим условием для всей рассчитываемой системы является только энергия уровня Ферми. У полученной субматрицы плотности для подобласти исключаются из рассмотрения все „углы“.

На рис. 1 приведена матрица плотности для центральной области. Из полной матрицы плотности (слева) в дальнейших вычислениях используется только „крестообразная“ часть (справа) [6].

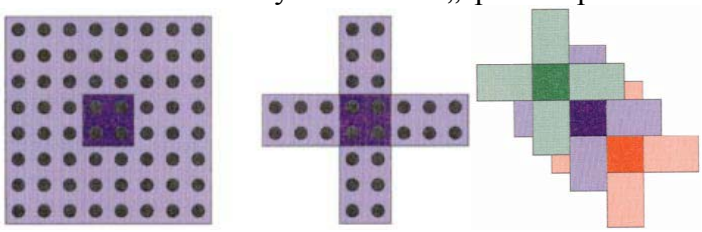

Рис. 1

Неперекрывающиеся области (центральные) суммируются с единичным весом, а перекрывающиеся — с весом 0,5.

Диагонализация гамильтониана и вычисление субматрицы плотности для каждого из фрагментов могут производиться совершенно независимо. Однако накладные расходы по сбору и раздаче матриц особенно сильно возрастают с уменьшением фрагмента до значений, при которых размеры буферной и центральной областей становятся соизмеримыми.

Объем передаваемой информации быстро возрастает при уменьшении размера фрагмента, что делает алгоритм малоэффективным при соответствующем увеличении количества узлов. В работе предпринята попытка снизить объем накладных расходов за счет модификации параллельного алгоритма DC. Предлагается выполнять процесс самосогласования для каждого фрагмента локально, при фиксированных значениях элементов матрицы плотности, соответствующих буферной области. Такая итерация далее будет называться локальной. Под глобальной итерацией далее будет подразумеваться обычная для DC алгоритма операция по уточнению матрицы гамильтониана.

На рис. 2, *а* представлена блок-схема параллельной версии DC алгоритма. Под блоком вычисления субматриц плотности на узлах понимается последовательность действий, приведенная на рис. 2, *б*. Если итерации внешнего цикла не выполняются, то блок-схемы описывают параллельный вариант исходного алгоритма DC, в противном случае — модифицированный вариант, а на рис. 2, *б* приведен итерационный процесс, выполняемый локально.

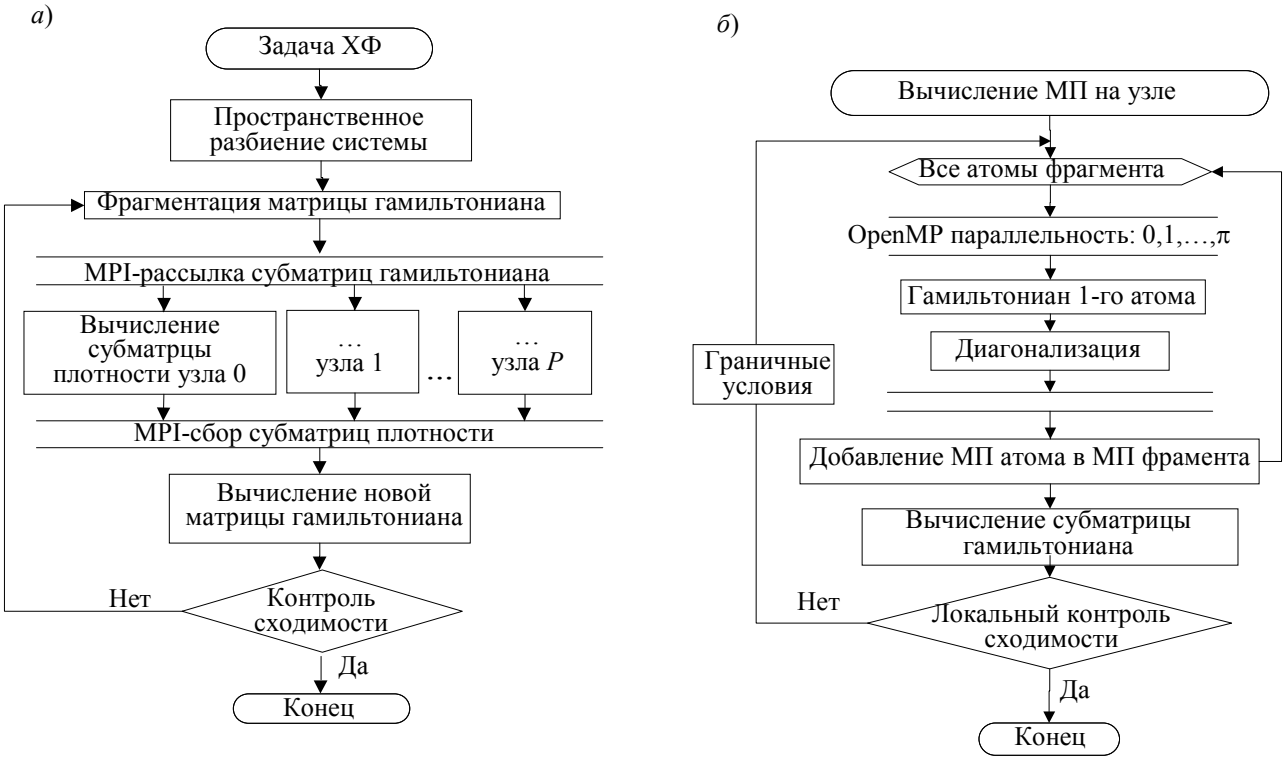

Рис. 2

Теоретически оценить параллельное ускорение модифицированного и исходного вариантов алгоритма DC позволяет выражение:

$$S = \frac{\alpha N + \beta N^2}{\frac{\alpha N}{\Pi\pi} + \frac{\beta N^2}{\Pi\pi} + \frac{d(1+c)}{\tau K}\left(N + \frac{c}{2}\sqrt{\kappa N \Pi}\right)}, \qquad (5)$$

где N — число атомов в системе, α и β — коэффициенты трудоемкости, Π — число узлов, π — число вычислительных ядер на узле. Операция диагонализации матрицы характеризуется линейностью алгоритма DC, в то время как обновление матрицы гамильтониана имеет квадратичную сложность с очень малым коэффициентом. Поскольку для систем размером $\sim 10^5$ атомов нелинейность проявляется, в модель введено квадратичное слагаемое. Третье слагаемое в знаменателе относится к накладным расходам: d — длина сообщения, приходящаяся на один атом системы, c — количество соседей для каждого атома, κ — геометрический фактор, связывающий площадь фрагмента с длиной его границы, τ — скорость MPI передачи данных; K — среднее число локальных итераций, приходящихся на одну глобальную.

Идея модификации алгоритма DC состоит именно в том, что локальные итерации позволяют существенно повысить скорость сходимости глобального процесса самосогласования, тогда $K > 1$, и накладные расходы снижаются в соответствующее количество раз. Случай $K = 1$ соответствует параллельному ускорению исходного алгоритма DC.

Измерение производительности. Параллельная производительность модифицированного алгоритма DC измерялась в ходе моделирования электронной плотности молекул графена различного размера. Запуски производились на суперкомпьютере МГУ „Ломоносов“ на 128, 256 и 512 узлах (8 вычислительных ядер на узел). В таблице приведены результаты измерения полного времени решения задачи, времени вычисления одной глобальной итерации и индексы эффективности использования ресурсов. Отсутствие данных по ускорению объясняется невозможностью проведения последовательного расчета для систем таких размеров.

Результаты измерений времени решения задач Хартри—Фока на 128/256/512 узлах для молекул графена различных размеров

Размер молекулы	Полное время выполнения, с	Эффективность	Время выполнения одной итерации, с	Эффективность
20802	2118/1469/1048	0,71/0,71	3,0/1,9/1,3	0,78/0,73
46202	4892/2839/1847	0,86/0,76	7,9/4,1/2,7	0,96/0,75
98562	32346/12835/7513	1,26/0,85	16,5/8,9/4,9	0,93/0,9

Из таблицы видно, что для системы максимального размера при переходе от 128 к 256 ядрам наблюдается гиперэффективность использования вычислительных ресурсов ~1,26. Эффект наблюдается для полного времени решения задачи, максимально достигаемая эффективность для одной итерации меньше единицы. Механизм возникновения обнаруженного эффекта в общих чертах следующий. Скорость сходимости процесса самосогласования для фрагментов молекулярной системы заметно зависит от числа атомов, поэтому при уменьшении размера фрагментов в определенных условиях выигрыш по времени от ускорения сходимости может превысить возрастающие затраты на пересылку данных. Однако такое качественное описание не позволяет ответить на важный вопрос об эффективности использования именно модифицированного алгоритма DC. Сходимость процесса самосогласования для молекулы целиком может оказаться лучшей, чем для фрагментов, это приведет к проигрышу относительно исходного алгоритма. Если рассматривать модель (5) как ускорение относительно одной глобальной итерации, полное ускорение запишется как $S_0 = S n_0 (n_l n_G)^{-1}$, где n_0 — число итераций до сходимости в немодифицированном алгоритме DC, n_l — число локальных итераций до сходимости,

n_G — число глобальных итераций. Коэффициент при S в правой части этого выражения будем называть коэффициентом гиперэффективности. В выражении (5) $K = n_l$.

Модель сходимости процесса Хартри—Фока. Для теоретической оценки ускорения предлагаемой параллельной модификации алгоритма DC необходимо рассмотреть зависимость числа необходимых итераций от размера молекулы. Этот непростой процесс приближенно можно описать из следующих общих соображений. Известно, что масштабируемость того или иного алгоритма определяется связанностью данных, характерной для решаемой задачи. Рассмотрим реакцию на точечное возмущение решения уравнения (1). В качестве такого практически точечного возмущения зафиксируем вариацию $\delta\psi$ на сферической поверхности σ некоторого малого радиуса. Решение будем искать во внешней области. Выражение (1) может быть представлено как уравнение Пуассона, если понимать его решение в виде итерационного процесса, в котором источник определяется видом ψ-функции, найденным на предыдущей итерации, тогда:

$$\psi_k(\mathbf{q}) = -\int \partial_n G(\mathbf{q},\mathbf{m})\delta\psi_k d\sigma_m + 2\int G(\mathbf{q},\mathbf{m})(\varepsilon_k - V)\psi_k d\Omega_m, \qquad (6)$$

где $G(\mathbf{q},\mathbf{m})$ — функция Грина уравнения Пуассона. Это уравнение может быть решено итерационно, в результате будет получен ряд n-кратных интегральных слагаемых и остаточный член, порождаемый первым и вторым слагаемыми в уравнении (6) соответственно. Сложность вычисления каждого из слагаемых пропорциональна размеру области интегрирования в степени кратности интеграла (интеграл по поверхности возмущения масштабируется как единица). Если область интегрирования всегда соответствует области решения задачи, конкретно — числу атомов в молекуле, то и решение задачи Хартри—Фока будет иметь экспоненциальную сложность, и таким же образом будет расти число итераций сходимости процесса самосогласования. Напротив, если отклик на точечное возмущение локален и можно определить область интегрирования, не зависящую от размера системы, то сложность будет постоянной. Более точное описание сходимости должно включать в себя учет убывания членов ряда и возможную локализацию областей интегрирования в них, этим объясняется отсутствие места для полиномиальной сложности.

Локальность квантово-химической модели, вводимая алгоритмом DC, имеет смысл локальности непосредственного взаимодействия атомных оболочек. В итерационном процессе самосогласования точечное возмущение может оказывать влияние, выходящее за пределы вводимой окрестности. Заметим, что оно может быть измерено непосредственно.

Для измерения окрестности релаксации точечного возмущения варьировались диагональные элементы матрицы плотности, относящиеся к выделенному атому вдалеке от границ молекулы графена. Вариация производилась после достижения сходимости процесса самосогласования, она составляла 10 % от точной величины и удерживалась до повторного достижения самосогласования. Полученная матрица плотности по модулю вычиталась из точной. На рис. 3 приведен пример локального и нелокального отклика для молекулы графена, содержащей 572 атома, при пороге DC отсечения 10^{-3} (*а*) и 10^{-4} (*б*). Точечное возмущение находится в центре изображений.

Разностная картина, наблюдаемая на рис. 3, *а*, считается локальным откликом: края графенового листа практически не подсвечены, напротив, на рис. 3, *б* отклик нелокален, поскольку возбуждение слабо затухает к краям, а картина отклика представляет собой, по-видимому, результат интерференции волновых функций, отраженных от края.

Измерения скорости сходимости подтвердили экспоненциальное возрастание числа итераций для случая нелокального отклика и его постоянство с момента увеличения молекулы до размеров, превышающих область отклика. Для параметра DC-отсечения 10^{-3} описанное изменение поведения соответствует молекуле графена размером 10×10 бензольных колец (всего 282 атома), для параметров отсечения $2\cdot10^{-4}$ и $1\cdot10^{-4}$ экспоненциальный рост числа

итераций наблюдается при увеличении стороны квадратного листа графена до 30 и 40 бензольных колец соответственно.

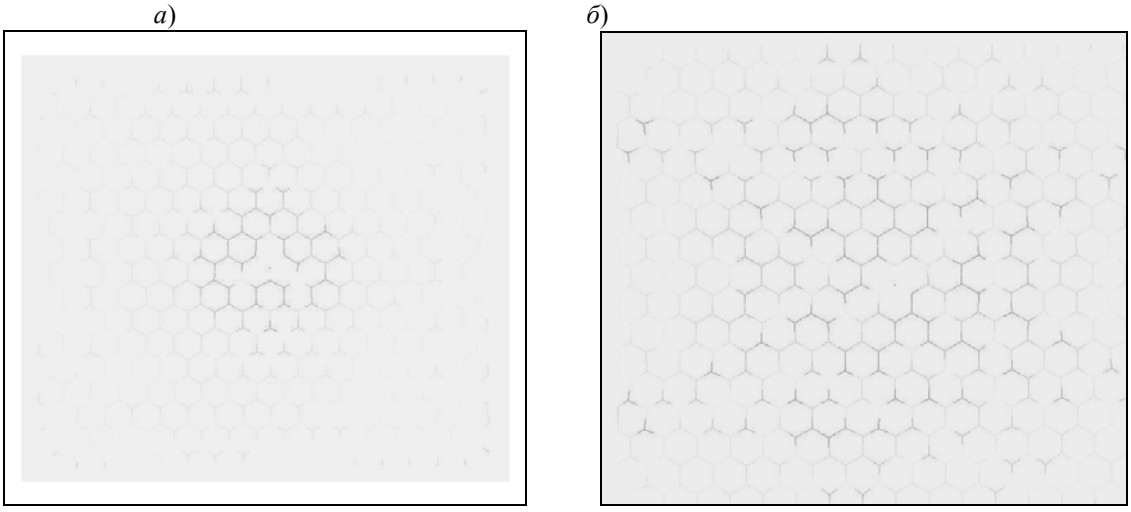

Рис. 3

В соответствии с такими предположениями модель зависимости числа итераций n_0 от размера молекулы может быть записана так:

$$n_0(N) = \begin{cases} n, ae^{N/N_0} < n, \\ ae^{N/N_0}, N < N_0, \\ a, N \geq N_0, \end{cases} \tag{7}$$

где N_0 — размер молекулы, при котором наблюдается изменение поведения; он, как и коэффициент α, определялся экспериментально. Первое из выражений в фигурных скобках введено для корректного поведения рассматриваемой модели вблизи нуля. Модель коэффициента гиперэффективности требует также знания поведения коэффициента n_G, которое не исследовалось теоретически. Эксперимент показал слабую зависимость n_G от числа фрагментов П~100, которая аппроксимировалась линейно.

Итоговая модель ускорения имеет следующий вид:

$$S_0 = S(N, \Pi)\frac{n_0(N)}{n_l(N/\Pi)n_G(\Pi)}, \quad n_G(\Pi) = b + c\Pi\,. \tag{8}$$

На рис. 4, *а* представлены графики коэффициента гиперэффективности, построенные в соответствии с предлагаемой моделью, *б* — ускорения решения задачи самосогласования Хартри—Фока для квадратных листов графена 98562 (*1*) и 46202 атома (*2*) на базе моделей (5) и (8). Коэффициент гиперэффективности, т.е. ускорение, которое было бы достигнуто только за счет изменения числа итераций до сходимости, принимает значения больше единицы в обоих рассмотренных случаях. Это, однако, не гарантирует наличия гиперускорения, которое проявляется только для случая графенового листа максимального размера. Более того, измеренная в ходе экспериментальных исследований эффективность не отвечает максимальному ускорению, которое могло бы быть получено. На рис. 4, *б* угол наклона секущей, проведенной из нуля к точке кривой для большего листа, превысит биссектрису координатного угла, только если точка будет располагаться вблизи максимума кривой. Для листа графена меньших размеров такая точка вообще отсутствует. Однако для обеих кривых найдется секущая, проведенная через две точки и имеющая „гиперэффективный“ угол наклона. Это означает, что проведенные измерения эффективности действительно ничего

не могут сообщить о реальном ускорении, а применение самого модифицированного алгоритма DC целесообразно только в ограниченном диапазоне числа пространственных фрагментов (область определения графика на рис. 4, *а*, соответствующая значениям, превышающим единицу).

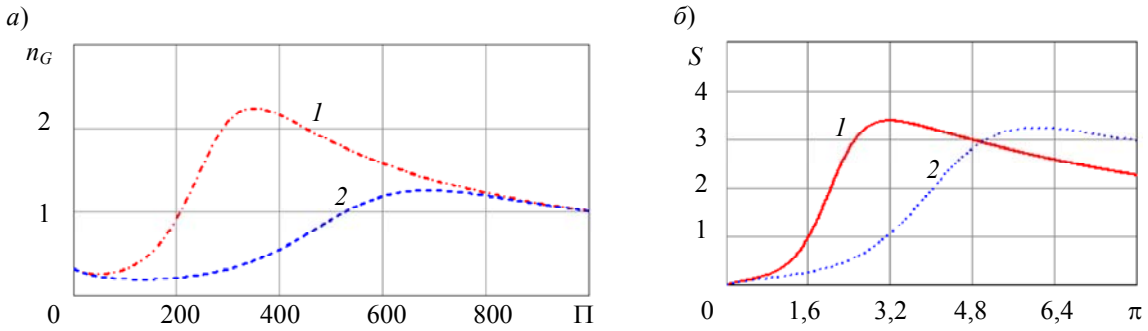

Рис. 4

Заключение. Предложенная модификация алгоритма DC позволяет снизить накладные расходы, пересчитанные на одну итерацию самосогласования. Однако эффективность по отношению к полному времени решения задачи самосогласования демонстрирует более сложное поведение, обусловленное зависимостью числа итераций самосогласования от размера молекулы. Проведенные измерения показали высокую эффективность и гиперэффективность использования вычислительных ресурсов относительно базы, взятой при запуске на 128 узлах (8 ядер на узле), однако это не является гарантией столь же высокой эффективности относительно гипотетического последовательного исполнения.

Работа выполнена в рамках контракта 07.514.11.4146 ФЦП „Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007—2013 годы“.

СПИСОК ЛИТЕРАТУРЫ

1. *Bernholdt D. E.* Scalability of correlated electronic structure calculations on parallel computers: A case study of the RI-MP2 method // Parallel Computing. 2000. Vol. 26. P. 945—963.
2. *Degoli E., Ossicini S.* Engineering Quantum Confined Silicon Nanostructures: Ab-Initio Study of the Structural, Electronic and Optical Properties. 2009. Vol. 58. P. 203—279.
3. *Nakai H., Kobayashi M.* Linear-scaling electronic structure calculation program based on divide-and-conquer method. 2011. Vol. 4. P. 1145—1150.
4. *Duchemina I., Gygi F.* A scalable and accurate algorithm for the computation of Hartree–Fock exchange // Computer Physics Communications. 2010. Vol. 181, N 5. P. 855—860.
5. *Alizadegan R., Hsia K. J., Martinez T. J.* A divide and conquer real space finite-element Hartree–Fock method // J. of Chem. Phys. 2010. Vol. 132. P. 034101.
6. *Goedecker S.* Linear scaling electronic structure methods // Rev. Mod. Phys. 1999. Vol. 71, N 4. P. 1085–1123.
7. Bolliger C. Linear Scaling Electronic Structure Methods. July 2008 [Электронный ресурс]: <http://www.math.ethz.ch/~kressner/students/bolliger.pdf>.
8. *Lin L., Lu J., Ying L., Car R.* Fast algorithm for extracting the diagonal of the inverse matrix with application to the electronic structure of metallic systems // Commun. Math. Sci. 2009. Vol. 7, N 3. P. 755—777.

Сведения об авторах

Андрей Игоревич Свитенков — Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики; Научно-исследовательский институт Наукоемких компьютерных технологий; инженер; E-mail: svitenkov@yandex.ru

Дмитрий Михайлович Спельников — Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики; Научно-исследовательский институт Наукоемких компьютерных технологий; младший научный сотрудник; E-mail: pilule@yandex.ru

Владимир Григорьевич Маслов — д-р физ.-мат. наук; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики; Центр „Информационные оптические технологии“; ведущий научный сотрудник; E-mail: maslov04@bk.ru

Александр Валерьевич Бухановский — д-р техн. наук; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Научно-исследовательский институт Наукоемких компьютерных технологий; директор НИИ НКТ;
E-mail: avb_mail@mail.ru

Рекомендована кафедрой высокопроизводительных вычислений

Поступила в редакцию 18.06.13 г.