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

А. И. Свитенков, В. Г. Маслов, А.В. Бухановский

Введение. Распространенным подходом к приближенному решению уравнения Шредингера является т.н. одноэлектронное приближение, в котором каждый электрон рассматривается как независимый, движущийся в усредненном поле ядер атомов и других электронов. Он используется в методах Хартри-Фока, функционала плотности (DFT) или приближении сильносвязанных электронов. Уточнение решения другими методами за счет учета корреляционной энергии электронов приводит к критическому росту вычислительной сложности и практически 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 квантовой химии [2]. Однако при традиционной постановке задачи происходит кубический рост вычислительной сложности с увеличением числа атомов, что делает ее неприменимой для систем, содержащих порядка 1000 атомов и более. В этой связи в настоящее время рассматриваются линейно масштабируемые методы, значительно расширяющие границы применения квантовой химии [3]. Однако увеличение размеров систем до десятков и сотен тысяч атомов требует параллельной реализации указанных алгоритмов. В работе [3] с помощью линейно масштабируемого метода “Divide-and-conquer” (DC) для квантово-химического уравнения Хартри-Фока моделировалась электронная структура молекулярных соединений типа графена и графана.

Специфика выбранных соединений позволяет наблюдать важные эффекты, связанные со сходимостью итерационного процесса самосогласования. Самосогласованный характер решения является особенностью большинства квантово-химических методов, 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]. В настоящей работе исследуется этот вопрос. Установлено, что 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 количества итераций самосогласования, требуемых для сходимости. Относительно полного времени решения задачи Хартри-Фока, таким образом, наблюдается гиперускорение, для объяснения такого эффекта потребовалось рассмотреть свойства уравнения Хартри-Фока и его решений.

Постановка задачи. В методе Хартри-Фока гамильтониан молекулярной системы в уравнении Шредингера заменяется приближенным одноэлектронным аналогом – т.н. называемым фокианом. В таком случае приближенное уравнение Шредингера, которое должно быть решено относительно волновых функции ψk(r) в действительной области пространства в некоторой окрестности неподвижных центров атомных ядер, приобретает вид [5]:

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

Здесь V – оператор эффективного потенциала, в котором движется электрон, и который можно представить в виде суммы V=V0+Vd+Vx; V0 описывает вклад взаимодействия электрона и атомных ядер, Vd и Vx – взаимодействие с остальными электронами системы:

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

При разложении по выбранному набору базисных функций k-я орбиталь представляет собой линейную комбинацию базисных функций φi с неизвестными коэффициентами Cki [6]:

$$\psi_k = \sum_{i=1}^{M} C_{ki}\varphi_i \tag{3}$$

относительно которых уравнение Хартри-Фока может быть записано так:

$$\sum_{j=1}^{N} F_{ij}C_{kj} = \varepsilon_k C_{ki}$$

Здесь 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:

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

Из уравнений (2) видно, что F, в свою очередь, зависит от P. Таким образом, уравнения (2)–(4) формируют т.н. самосогласованную задачу, решение которой сводится к итерационному процессу с последовательным уточнением матрицы плотности и соответствующей матрицы фокиана до достижения сходимости.

Из уравнения (3) видна кубическая сложность предлагаемого алгоритма относительно размера матриц P и F. Однако важно, что при использовании базиса сильно локализованных функций внедиагональные элементы матрицы плотности довольно быстро затухают с расстоянием, поэтому ~N, а не ~N2 – реальное число отличных от нуля матричных элементов [7]. То же относится к матрице фокиана в этом представлении, соответственно трудоемкость всех действий с такими матрицами ниже O(N3). Это свойство отражает локальный характер квантовой механики; оно так или иначе используется всеми линейно масштабируемыми алгоритмами решения задачи Хартри-Фока.

Алгоритм DC и его параллельная реализация. В алгоритме DC общая матрица плотности строится на основе решения, полученного не для всей системы, а для некоторых перекрывающихся фрагментов. Размер буферной зоны устанавливается посредством задания величины отсечения Sth интегралов перекрывания базисных функций, значения меньше которой считаются нулевыми. Атомы, базисные функции которых не перекрываются, считаются непосредственно не взаимодействующими [8].

Для каждой подобласти отдельно рассчитывается субматрица плотности. Общим условием для всей рассчитываемой системы является только энергия уровня Ферми. У подобласти полученной субматрицы исключаются из рассмотрения все «углы» (рис. 1).

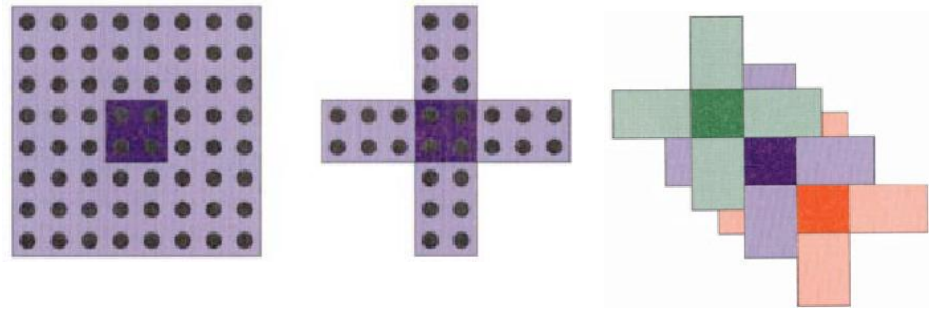


Рис. 1. Матрица плотности для центральной области. Из полной матрицы плотности (слева) в дальнейших вычислениях используется только «крестообразная» часть (справа) [6].

Полная матрица плотности составляется с использованием крестообразных блоков. Неперекрывающиеся области (центральные, на рис.1 – затемнены) суммируются с единичным весом, а перекрывающиеся– с весом 0.5.

Согласно методу DC, диагонализация гамильтониана и вычисление субматрицы плотности для каждого из фрагментов могут производиться совершенно независимо. Однако с уменьшением фрагмента до размеров, при которых буферная и центральная области становятся соизмеримыми, особенно сильно возрастают накладные расходы по сбору и раздаче матриц. В табл. 1 приведены примеры значений числа атомов во фрагменте молекулы графена для нескольких величин порогов отсечки.

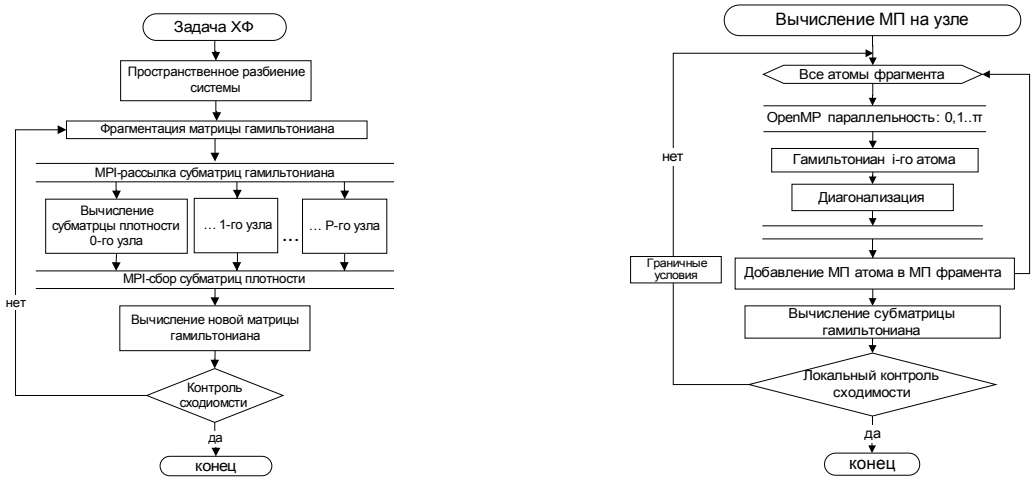
Таблица 1. Отношение размеров буферных и центральных областей для различных величин DC отсечения.

Sth	**Размер центральной области**	**Размер фрагмента**	**Относительный размер буферной зоны**
1e-4	81	358	4.42
1e-3	81	349	4.31
1e-3	240	453	1.88
1e-4	240	509	2.12
1e-4	429	814	1.9
1e-3	429	740	1.72

Объем передаваемой информации быстро возрастает с уменьшением размера фрагмента, что делает алгоритм малоэффективным при соответствующем увеличении числа узлов. Снизить объем накладных расходов предлагается за счет модификации параллельного алгоритма DC: осуществлять процесс самосогласования для каждого фрагмента локально, при фиксированных значениях элементов матрицы

плотности, соответствующих буферной области. Такую итерацию назовем локальной, под глобальной итерацией будет подразумеваться обычая для DC-алгоритма операция по уточнению матрицы гамильтониана.

На рис. 2(а) приведена блок-схема параллельной версии DC-алгоритма. Блоки вычисления субматриц плотности на узлах включают в себя последовательность действий, приведенную на рис. 2(б). Если итерации внешнего цикла не выполняются, то блок-схемы описывают параллельный вариант исходного алгоритма DC. В противном случае рис. 2(а) относится к модифицированному варианту, а рис. 2(б) отражает итерационный процесс, выполняемый локально.



(а) (б)

Рис. 2. Схема параллельного исполнения алгоритма DC

Оценим параллельное ускорение модифицированного и исходного вариантов алгоритма DC:

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

α, β – параметры сложности алгоритма, N – число атомов в системе, Π – число узлов, π – число вычислительных ядер на узле. Линейность алгоритма DC связана только с операцией диагонализации матрицы, в то время как обновление матрицы гамильтониана имеет квадратичную сложность с очень малым коэффициентом. Поскольку нелинейность проявляется для систем ~105 атомов, в модель введено квадратичное слагаемое. Третье слагаемое в знаменателе отвечает за накладные расходы: d – длина сообщения, приходящаяся на один атом системы, c – число соседей для каждого атома, κ – геометрический фактор, связывающий площадь фрагмента с длиной его границы, τ – скорость передачи данных между узлами; K – среднее число локальных итераций, приходящихся на одну глобальную.

Идея модификации алгоритма DC состоит в следующем: локальные итерации позволяют существенно повысить скорость сходимости глобального процесса самосогласования, тогда K>1 и накладные расходы снижаются в соответствующее число 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 разных размеров. Запуски производились на суперкомпьютере “Ломоносов” (МГУ) на 128, 256 и 512 узлах (8 вычислительных ядер на узел). В табл. 2 приведены результаты измерений полного времени решения задачи, времени вычисления одной глобальной итерации и эффективности использования ресурсов. Данные о времени приведены через знак дроби для случая 128, 256 и 512 узлов соответственно. Под коэффициентом эффективности подразумевается коэффициент, вычисляемый как (t1/t2)·(N2/N1), где t1 и t2 времена решения задачи самосогласования (третий столбец) или времена выполнения одной итерации (пятый столбец) для меньшего числа узлов N1 и большего N2. Отсутствие данных по ускорению объясняется невозможностью проведения последовательного расчета для систем таких размеров.

Таблица 2. Результаты измерений времени решения задач Хартри-Фока на 128/256/512 узлах для молекул графена различных размеров

N	Полное время выполнения (с)	Эффективность	Время выполнения одной итерации (с)	Эффективность
20802	2118/1469/1048	0.72/0.70	3.0/1.9/1.3	0.79/0.73
46202	4892/2839/1847	0.86/0.77	7.9/4.1/2.7	0.96/0.76
98562	32346/12835/7513	1.26/0.85	16.5/8.9/4.9	0.93/0.9

Из табл. 2 видно, что для системы максимального размера при переходе от 128 к 256 ядрам наблюдается гиперэффективность использования вычислительных ресурсов – 1.26. Эффект наблюдается при измерении полного времени решения задачи. При измерении времени только одной итерации максимально достигаемая эффективность меньше 1. Механизм возникновения такого эффекта в общих чертах следующий. Скорость сходимости процесса самосогласования для фрагментов молекулярной системы существенно зависит от числа атомов, поэтому с уменьшением размера фрагментов при определенных условиях выигрыш по времени от ускорения сходимости может превысить возрастающие затраты на пересылку данных. Однако такое качественное описание не позволяет определить эффективность использования именно модифицированного алгоритма DC. Сходимость процесса самосогласования для молекулы целиком может оказаться лучшей, чем для фрагментов, что приведет к проигрышу относительно исходного алгоритма. Если рассматривать модель (5) как ускорение относительно одной глобальной итерации, полное ускорение запишется как S0=S·n0(nl·nG)-1, где n0 –число итераций до сходимости в немодифицированном алгоритме DC, nl – число локальных итераций до сходимости, nG – число глобальных итераций. В модели для S принимается K=nl. Коэффициент при S в правой части выражения будем называть коэффициентом гиперэффективности.

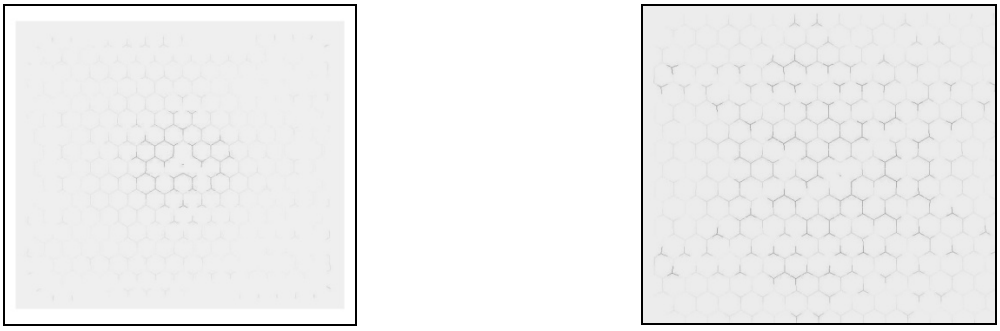
Модель сходимости процесса Хартри-Фока. Для теоретической оценки ускорения предлагаемой параллельной модификации алгоритма DC необходимо обсудить зависимость числа необходимых итераций от размера молекулы. Этот процесс приближенно достаточно легко можно описать из следующих общих соображений. Известно,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

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

Здесь G(q,m) есть функция Грина уравнения Пуассона. Это уравнение может быть решено итерационно, в результате чего будет получен ряд n-кратных интегральных слагаемых и остаточный член, порождаемые первым и вторым слагаемым в уравнении (6) соответственно. Сложность вычисления каждого слагаемого пропорциональна размеру области интегрирования в степени кратности интеграла (интеграл по поверхности возмущения масштабируется как 1). Если область интегрирования всегда соответствует области решения задачи – числу атомов в молекуле, то 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 очертить 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 тот факт, что сложность алгоритма – неполиномиальная.

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, составляла 10 % от точной величины и удерживалась до повторного достижения самосогласования. Полученная матрица плотности по модулю вычиталась из точной (рис. 3, точечное возмущение находится в центре изображений).



а) б)

Рис. 3. Пример локального и нелокального отклика для молекулы графена, содержащей 572 атома, при пороге DC отсечения а) 10-3 и б) 10-4

Разностная картина, наблюдаемая на рис.3(а), считается локальным откликом: края графенового листа практически не подсвечены, отклик на рис.3(б) нелокален, т.к. 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 их постоянное число с момента увеличения молекулы до размеров, превышающих область отклика. Для параметра DC-отсечения 10-3 описанное изменение поведения соответствует молекуле графена 10×10 бензольных колец (всего 282 атома), для параметров отсечения 2·10-4 и 1·10-4 экспоненциальный рост числа итераций наблюдается при увеличении стороны квадратного листа графена до 30 и 40 бензольных колец соответственно.

Согласно сказанному, модель зависимости числа итераций n0 от размера молекулы может быть записана так:

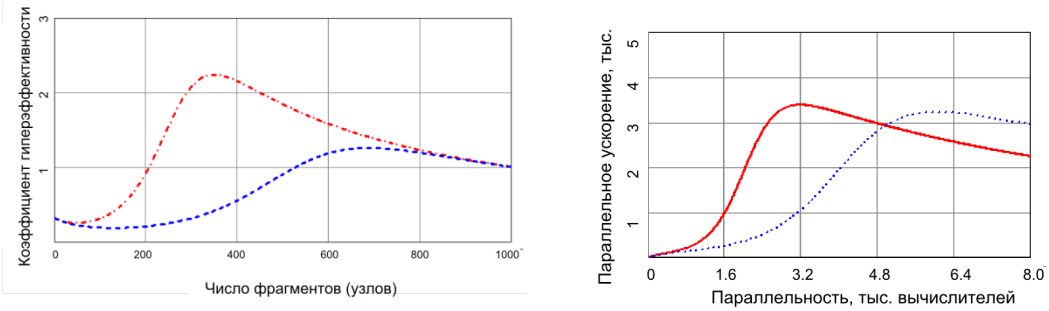
$$n_0(N) = \begin{cases} n, ae^{N/N_0} < n \\ ae^{N/N_0}, N < N_0 \\ a, N \geq N_0 \end{cases} \quad (7)$$

где N0 – размер молекулы, при котором наблюдается изменение поведения. Он, как и коэффициент a, определялся экспериментально. Первое из выражений в фигурных скобках введено для задания корректного поведения нашей модели вблизи нуля. Модель коэффициента гиперэффективности требует также знания поведения коэффициента nG, которое не исследовалось теоретически. Для числа фрагментов ~ 100 эксперимент показал слабую зависимость nG от N, которая линейно аппроксимировалась.

Итоговая модель ускорения имеет вид:

$$S_0 = S\frac{n_0(N)}{n_l(N)\cdot n_G(N)}, \quad n_G(N) = b + cN \quad (8)$$

На рис. 4(а) представлены графики для коэффициента гиперэффективности, построенные в соответствии с предлагаемой моделью; на рис. 4(б) – графики ускорения для времени решения задачи самосогласования Хартри-Фока на базе моделей (5) и (8).



а) б)

Рис. 4. Модели коэффициента гиперэффективности (а) и параллельного ускорения (б) решении задачи Хартри-Фока для квадратных листов графена 98562 атома (красная кривая) и 46202 атома (синяя)

Заключение. Предложенная модификация алгоритма DC приводит к снижению накладных расходов, пересчитанных на одну итерацию самосогласования. Однако эффективность по отношению к полному времени решения задачи самосогласования демонстрирует более сложное поведение, связанное с зависимостью числа итераций самосогласования от размера молекулы. Проведенные измерения показали высокую эффективность, и даже гиперэффективность, использования вычислительных ресурсов относительно базы, взятой при запуске на 128 узлах (8 ядер на узле). В ходе исследования этого вопроса была построена модель сходимости процесса

Хартри-Фока, на основе которой получена модель полного ускорения (рис. 4), позволяющая оценить вклад в ускорение от собственно распараллеливания, и от увеличения скорости сходимости с уменьшением размеров параллельного фрагмента. Данная модель может быть использована в дальнейшем для планирования вычислительного процесса на современных суперкомпьютерах.

ЛИТЕРАТУРА:

1. D. E. Bernholdt "Scalability of correlated electronic structure calculations on parallel computers: A case study of the RI-MP2 method" // Parallel Computing 26, p.945-963 (2000)
2. E. Degoli, S. Ossicini "Engineering Quantum Confined Silicon Nanostructures: Ab-Initio Study of the Structural, Electronic and Optical Properties"// V. 58, p.203-279 (2009)
3. H. Nakai, M. Kobayashi "Linear-scaling electronic structure calculation program based on divide-and-conquer method", V. 4, p.1145-1150 (2011)
4. I. Duchemina, F. Gygi "A scalable and accurate algorithm for the computation of Hartree–Fock exchange" // Computer Physics Communications V. 181, 5, p.855-860 (2010)
5. R. Alizadegan, K. J. Hsia, T. J. Martinez "A divide and conquer real space finite-element Hartree–Fock method" // THE JOURNAL OF CHEMICAL PHYSICS 132, 034101 (2010)
6. S. Goedecker, "Linear scaling electronic structure methods" // Rev. Mod. Phys. 71, No.4, 1085–1123 (1999)
7. C. Bolliger, "Linear Scaling Electronic Structure Methods", July 2008, http://www.math.ethz.ch/~kressner/students/bolliger.pdf
8. L. Lin, J. Lu, L. Ying, R.Car, "Fast algorithm for extracting the diagonal of the inverse matrix with application to the electronic structure of metallic systems" // Commun. Math. Sci., Vol. 7, No. 3, pp. 755–777 (2009)