АНАЛИЗ ТЕОРЕТИЧЕСКИХ ОЦЕНОК УСКОРЕНИЯ И ЭФФЕКТИВНОСТИ ДЛЯ ПАРАЛЛЕЛЬНЫХ АЛГОРИТМОВ РАСЧЕТА ЭЛЕКТРОННОЙ СТРУКТУРЫ МОЛЕКУЛ

А.М. Чернецов

Одной из сложнейших задач современности является квантово-химическая задача расчёта электронной структуры больших органических молекул, которые насчитывают десятки тысяч взаимодействующих электронов и ядер. Результаты расчётов используются, например, в фармакологии при решении задачи конструирования лекарств, в нанотехнологиях, при создании высокотемпературных сверхпроводников и др.

Расчёт электронной структуры сводится к решению уравнения Шредингера, которое описывает пространственное движение всех частиц системы, перемещающихся в силовом поле. Так как положение каждой частицы описывается тремя декартовыми координатами, то возникает уравнение в частных производных второго порядка в трёхмерном пространстве, решаемое аналитически лишь для очень малого класса систем, состоящих из одного или двух атомов. Для молекулярных систем с б**о**льшим числом атомов применяются трудоёмкие численные итерационные методы, которые могут быть реализованы только при использовании высокопроизводительных средств и параллельной обработки. Практическими результатами решения задачи расчёта электронной структуры являются взаимные расположения атомов в пространстве, создающих стабильные молекулы и относительные энергии этих молекул, а также временная зависимость между структурой молекул и их свойствами.

В общем случае размерность задачи расчёта молекулярной системы есть функция от числа атомов и размерности применяемого базиса в итерационном методе. Трудоёмкость задачи расчёта зависит от метода решения, размерности молекулярной системы и формы расчёта – с фиксированной или изменяемой геометрией положения атомов.

Расчеты электронной структуры гигантских молекул являются одними из самых сложных в современной науке и требуют использования высокопроизводительных вычислительных архитектур. Например, вычислительный эксперимент по изучению свойств молекулы октана C_8H_{18} (размер базиса 1468) на кластере HP, содержащем 1400 процессоров Itanium 2, потребовал более 23 часов работы. При этом загрузка процессоров составила 75%, а средняя производительность системы - 6,3 Tflops [1].

В современных квантово-химических исследованиях допустимыми временами расчёта считаются величины порядка секунд и минут, в исключительных случаях – часов. Однако, расчёт больших молекул, содержащих более 10^4 атомов, классическими методами даже с использованием высокопроизводительных систем, составляет порядка нескольких десятилетий. По мнению специалистов компании Intel компьютер, способный проводить квантовохимические расчеты любой сложности за допустимое время, должен иметь производительность не менее 10^{21} FLOPS, и такая мощность будет достигнута не ранее 2030 года [2].

Таким образом, актуальными являются исследования, направленные на организацию эффективных расчётов электронной структуры больших молекул на существующих сегодня высокопроизводительных архитектурах.

Большинство теорий для определения состояния многоэлектронной системы, имеющей минимальный уровень полной энергии, используют концепцию эффективного одноэлектронного гамильтониана (фокиана),

$$H = -\frac{1}{2}\nabla^2 + v(r)$$

который представляется в виде:

Здесь $\boldsymbol{v}$ – оператор потенциальной энергии, ∇^2 – оператор Лапласа, r – радиус-вектор.

При этом возникает задача на собственные значения:

$$H\psi_i = \varepsilon_i\psi_i$$

где i нумерует собственные значения и соответствующие собственные вектора. Таким образом, задача сводится к обработке матрицы размерности N, где N – размерность задачи, пропорциональная числу атомов. Основными методами расчетов электронной структуры молекул являются неэмпирические и полуэмпирические методы квантовой химии [3]. Неэмпирические методы имеют сложность расчета от $O(N^5)$ до $O(N!)$, а требования к памяти от $O(N^3)$.

Для расчета больших молекулярных систем, которые могут содержать от 10^3 до 10^7 атомов, целесообразно применение полуэмпирических методов квантовой химии в так называемом приближении нулевого дифференциального перекрывания [3], в общем случае имеющих асимптотическую сложность расчета $O(N^3)$. Центральным звеном при расчете молекулярных систем является решение симметричной задачи на собственные значения методом матричной диагонализации.

Для расчета физико-химических свойств нужны не сами собственные вектора, а матрица плотности *P*, являющаяся функцией от них. При нахождении матрицы плотности *P* начальное приближение матрицы формируется из исходного фокиана *F***.** Фокиан, в свою очередь, строится по определенным правилам на основе декартовых координат атомов, входящих в молекулу [3,4]. Сложность расчета фокиана для плотных матриц составляет $O(N^2)$.

Другой путь нахождения *P*, альтернативный решению задачи на собственные значения, – прямое "извлечение" *P* из фокиана. Одним из численных методов прямого нахождения матрицы плотности *P* является метод очистки (purification method) [5], разработанный еще в 1960 г. Однако в силу отсутствия в то время необходимых вычислительных ресурсов его применение было ограничено расчетом только небольших молекул. В 90-х гг. XX века на основе этого метода были разработаны различные модификации, позволяющие ускорить процесс вычислений.

Общая идея использования метода очистки для построения матрицы плотности основывается на итерационном разложении фокиана по непрерывно возрастающему полиному $P(x)$, $x \in [0,1]$ с зафиксированным экстремумом между 0 и 1, т.е. $P(0)=0$, $P'(1)=1$ и $P(0)= P'(1)=0$.

В одной из современных модификаций - методе Пальцера-Манолополиса [5,6] - итерационная формула для нахождения матрицы плотности *P* выглядит следующим образом:

$$P_{n+1} = \begin{cases} \frac{1}{1-c_n}\left((1-2c_n)\cdot P_n + (1+c_n)\cdot P_n^2 - P_n^3\right), c_n \le \frac{1}{2} \\ \frac{1}{c_n}\left((1+c_n)\cdot P_n^2 - P_n^3\right), c_n > \frac{1}{2} \end{cases}, c_n = \frac{tr(P_n^2 - P_n^3)}{tr(P_n - P_n^2)} \qquad (1)$$

где n – номер шага итерационного процесса, *P* - матрица плотности, tr – след матрицы, c_n - коэффициент. Процесс вычислений останавливается при достижении заданной точности по идемпотентности:

$$\sqrt{\frac{tr(P^2) - tr(P)}{tr(P)}} \le \varepsilon \qquad (2)$$

Из анализа выражения (1) следует, что основной вклад в вычислительную трудоемкость вносят матричные операции.

Поскольку при построении математической модели достаточно большого класса молекул возникают массивы данных разреженной структуры, то целесообразно использовать это свойство с целью повышения эффективности расчетов. Свойство разреженности используется для снижения требований к представлению в памяти данных очень больших размеров и сокращения вычислительной сложности алгоритмов обработки матриц.

Можно показать [4-7], что в общем случае матрицы фокиана *F* и плотности *P* являются разреженными. Рассмотрим разреженность матрицы блочно-трехдиагонального вида, которая соответствует описанию достаточно большого класса молекул, а именно полимеров и линейных несвёрнутых протеинов. На рис. 1 представлена блочно-трехдиагональная матрица (*block tridiagonal matrix*), где: $A1_i$, $A2_i$, $A3_i$ – матричные, в общем случае плотно заполненные, блоки размерности m_i; nbl-число блоков главной диагонали; $A1_i$ (i=1, nbl) – матрицы блоков главной диагонали, $A2_i$ и $A3_i$ (i=1, nbl-1) – соответственно матрицы блоков нижней поддиагонали и верхней наддиагонали. Количество блоков nbl может быть ограничено только доступными ресурсами памяти. Максимальная размерность мелких блоков(размер m определяется, исходя из квантово-химических свойств молекулярной системы) $m=\max\{m_i\}$.

$$A = \begin{bmatrix} A1 & A3 & 0 & \dots & 0 \\ A2 & A1 & A3 & \dots & \dots \\ 0 & A2 & A1 & \dots & \dots \\ 0 & \dots & \dots & \dots & A3 \\ 0 & \dots & \dots & A2 & A1 \end{bmatrix}_{N\times N}$$

Рис. 1. Структура блочно-трехдиагональной матрицы A

Для эффективной реализации метода Пальцера-Манолополиса в случае представления матрицы фокиана в форме блочно-трехдиагонального вида необходимо организовать эффективное выполнение операций сложений, умножения на число и умножение матриц.

При умножении двух блочно-трехдиагональных матриц *A* и *B* в матрице результата *R* появляется 4-ая ненулевая блочная диагональ, значениями которой можно пренебречь с точки зрения анализа (расчетов)

электронной структуры молекул. Таким образом, матрица *R* представляется в форме блочно-трехдиагональной. Для представления всех трех матриц *A*, *B*, *R* предлагается специальная схема хранения в виде трехмерных массивов. Тогда элементы матриц имеют 3 индекса (например, A_{ijk}), где первые два соответствуют положению элемента внутри блока, а последний - позиции блока в соответствующей диагонали.

В результате получим факторизованные выражения для блоков матрицы *R*:

$$R1_i = A2_{i-1} * B2_{i-1} + A1_i * B1_i + A2_i * B2_i$$

$$R2_i = A2_i * B1_i + A1_{i+1} * B2_{i+1}$$

$$R3_i = A1_i * B2_i + A2_{i+1} * B1_{i+1}$$

где Ai_j ,Bi_j – блочные матрицы i-х диагоналей исходных матриц *A* и *B*, j – позиция блока в диагонали, Ri_j – блоки результирующей матрицы.

Если учесть, что матрица фокиана *F* является симметричной и блоки на главной диагонали являются квадратными, то можно сэкономить память на хранении массивов блоков верхней наддиагонали и после упрощений получить выражения для R:

$$R1_i = A2_{i-1} * B2_{i-1}^T + A1_i * B1_i + A2_i^T * B2_i$$

$$R2_i = A2_i * B1_i + A1_{i+1} * B2_{i+1}$$

Факторизованные выражения для первых ($R1_1$ и $R2_1$) и последних ($R1_{nbl}$ и $R2_{nbl-1}$) блоков можно получить из общей формулы.

Вычислительная сложность параллельной обработки разреженных матриц может быть уменьшена за счет оптимального варианта организации обмена блоками [7].

В отличие от реализации метода для плотных матриц, нет необходимости рассылать в качестве начальных данных матрицу целиком. Теперь каждый процесс получает только те блоки матриц, которые необходимые для расчетов. По завершению вычислений происходит сбор всех результатов в главном процессе. Следовательно, можно выполнять множество распределенных расчетов. Во-первых, это расчет следа матрицы. Каждый процесс считает только свою часть, а затем главный процесс собирает результаты. Во-вторых, все линейные операции, такие как расчет коэффициентов c_n и преобразования матрицы *P* по формуле (1), можно также вести распределенно. Это позволяет в отдельно взятом процессе выполнять действия по умножению матриц только над теми блоками, которые назначены процессу. Процесс вычислений повторяется итерационно до достижения заданной точности матрицы плотности, определяемой выражением (2).

На рис. 2 представлена схема организации обменов между процессами по типу «точка-точка» (*MPI_Send*/*MPI_Recv*) некрайними блоками. Взаимодействие между различными процессами минимально: большая часть вычислений происходит независимо, а данные от других процессов требуются только для “пограничных” блоков. Более того, на каждой заданной итерации возможно заранее сделать все необходимые пересылки данных и уже затем производить умножение блоков. Для произвольного некрайнего блока в k-ом процессе необходимо получить блок A1 из процесса (k+1) и послать ему блок A2, а также послать блок A1 из процесса k в процесс (k-1) и получить блок A2 из процесса (k-1).

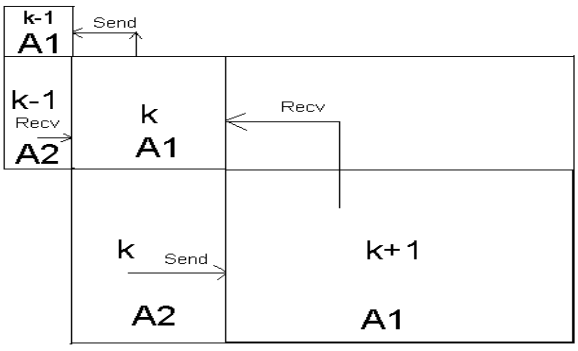


Рис.2. Схема пересылок между параллельными процессами

Использование описанных выше схем хранения и организации пересылок в параллельной программе повысило эффективность расчетов и увеличило допустимую размерность решаемой задачи. Так, при обработке плотных матриц в системе с объемом памяти 2 Гб максимальная размерность задачи составляет не более 8500, а для разреженных матриц – до 150000 [4].

Теперь оценим вычислительную сложность реализованных в [7] параллельных алгоритмов на множестве вычислителей p при использовании модели распределенной памяти со следующими предположениями, которые могут быть получены из формулы (1) для P_{n+1}.

Пусть вероятности того, что коэффициент $c_n \le ½$ и $c_n > ½$ - $P\{c_n \le ½\} = q$, $P\{c_n > ½\} = 1-q$; u –время аддитивной операции; v – время мультипликативной операции. Так как вычисления распределяются в среднем между вычислителями равномерно, то без ограничения общности положим, что каждый вычислитель обрабатывает одинаковое число блоков. В модели распределенной памяти необходимо учитывать время передач между вычислителями, поэтому введем функцию $F_{swap}(x)$ – взаимная передача x байт от одного процесса к другому в случае попарного вызова *MPI_Send*/*MPI_Recv*.

Тогда ускорение *S(p)* и эффективность *Q(p)* на *p* вычислителях оцениваются [7] как

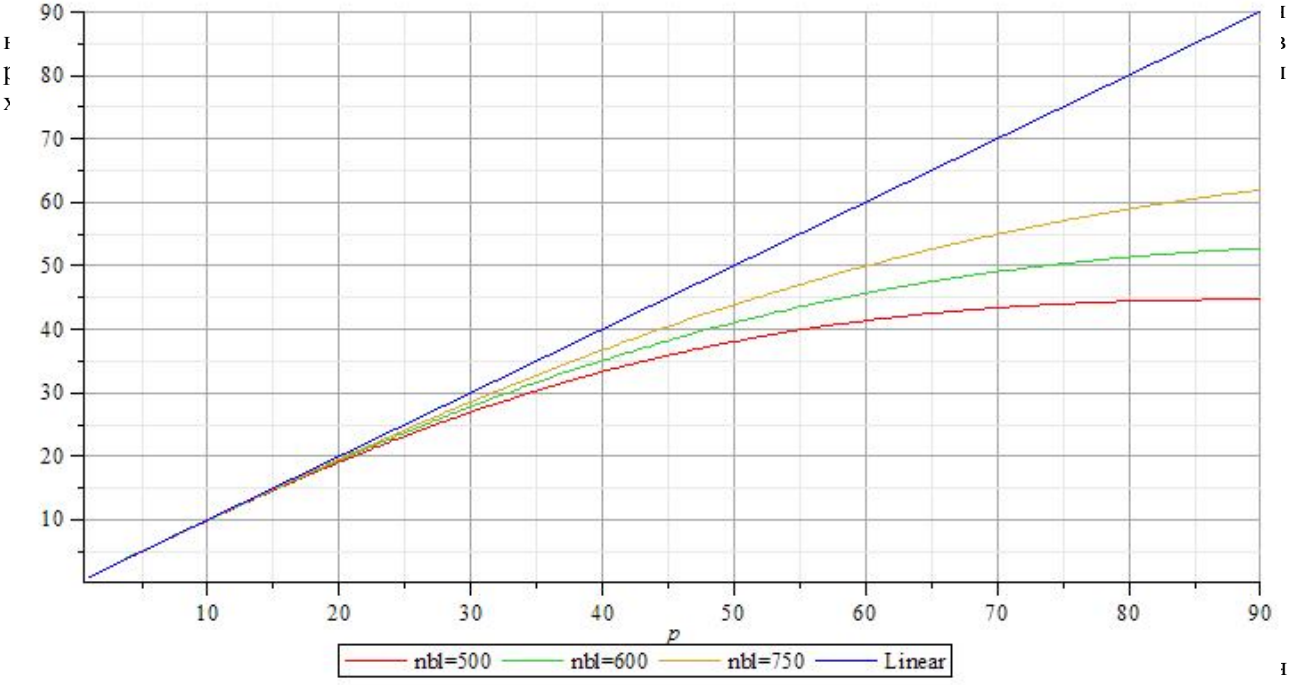
$$S(p) = \frac{q \cdot nbl \cdot m^2 \cdot [4 \cdot u + m \cdot nbl \cdot v] + (1-q) \cdot nbl \cdot m^2 \cdot [8 \cdot u + 10 \cdot m \cdot v]}{\frac{1}{p} \cdot nbl \cdot m^2 \cdot [8 \cdot u + 5 \cdot m \cdot v] + \frac{1}{p} \cdot (1-q) \cdot nbl \cdot m^2 \cdot [8 \cdot u + 10 \cdot m \cdot v] + 3 \cdot u \cdot m^2 \cdot p + 3 \cdot Fswap(m^2 \cdot 8) \cdot p}$$

$$Q(p) = \frac{S(p)}{p}$$

Ускорение (“эффект”), получаемое от использования разреженной формы хранения матрицы плотности, имеет вид:

$$S_{разр} = \frac{q \cdot (m \cdot nbl)^2 \cdot [4 \cdot u + m \cdot nbl \cdot v] + (1-q) \cdot (m \cdot nbl)^2 \cdot [4 \cdot u + 2 \cdot m \cdot nbl \cdot v]}{q \cdot nbl \cdot m^2 \cdot [8 \cdot u + 5 \cdot m \cdot v] + (1-q) \cdot nbl \cdot m^2 \cdot [8 \cdot u + 10 \cdot m \cdot v]} =$$

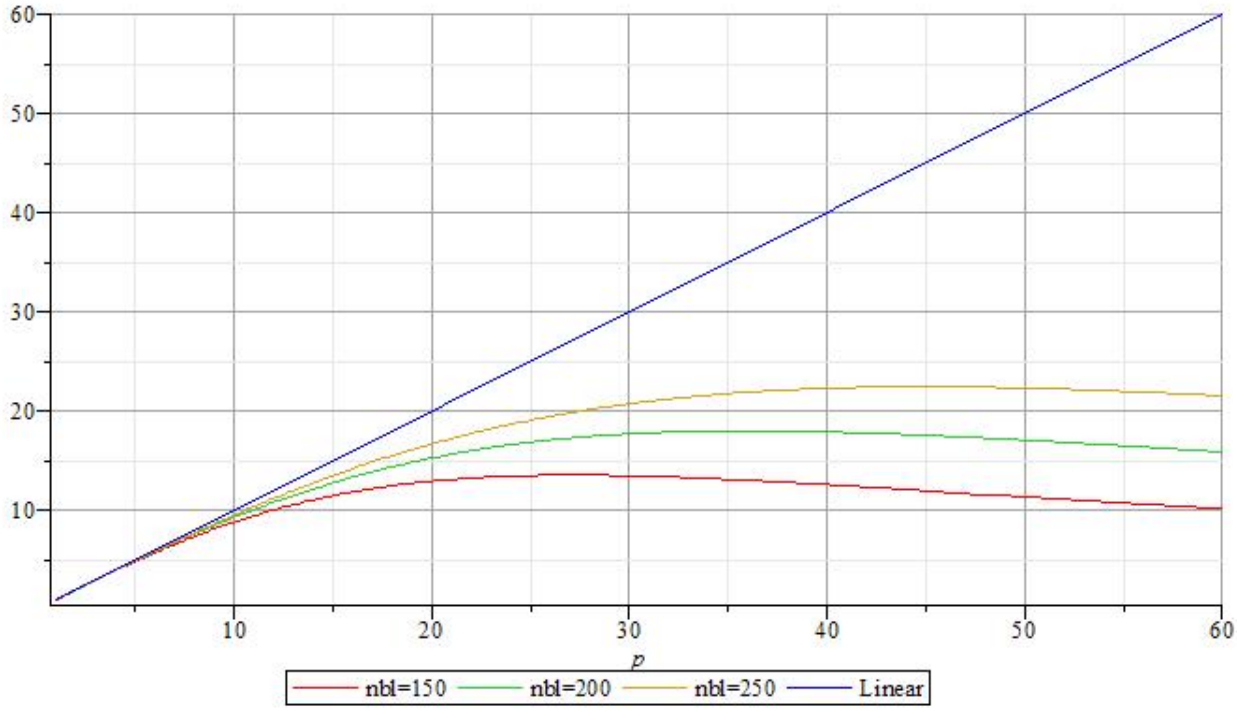
$$= \frac{q \cdot nbl \cdot [4 \cdot u + m \cdot nbl \cdot v] + (1-q) \cdot nbl \cdot [4 \cdot u + 2 \cdot m \cdot nbl \cdot v]}{q \cdot [8 \cdot u + 5 \cdot m \cdot v] + (1-q) \cdot [8 \cdot u + 10 \cdot m \cdot v]}$$



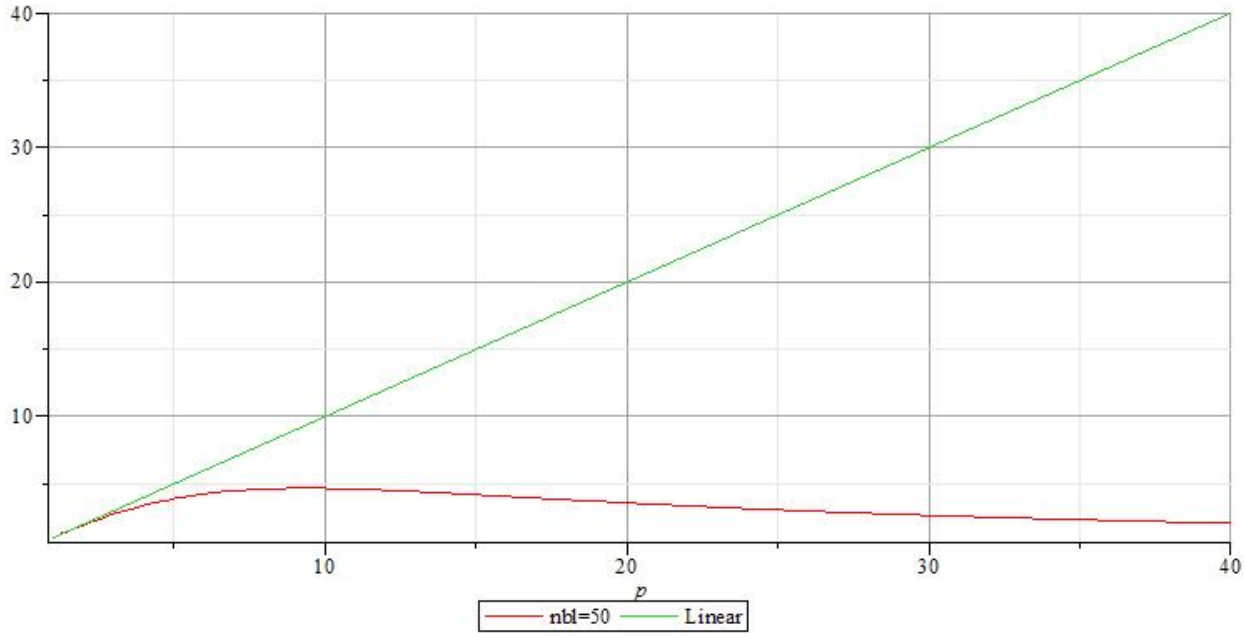
разных размерностей задачи и скоростей передачи данных.

На рис. 4а – 4в приведены графики зависимости ускорения от числа процессоров для разных размерностей задачи (а –125000,150000, 187500; б – 37500, 50000, 62500; в -12500) при фиксированной размерности блока 250. На рис. 4г приведены графики зависимости ускорения от числа процессоров для разных размерностей блока m =100;150;200;250 при фиксированном числе блоков 300.

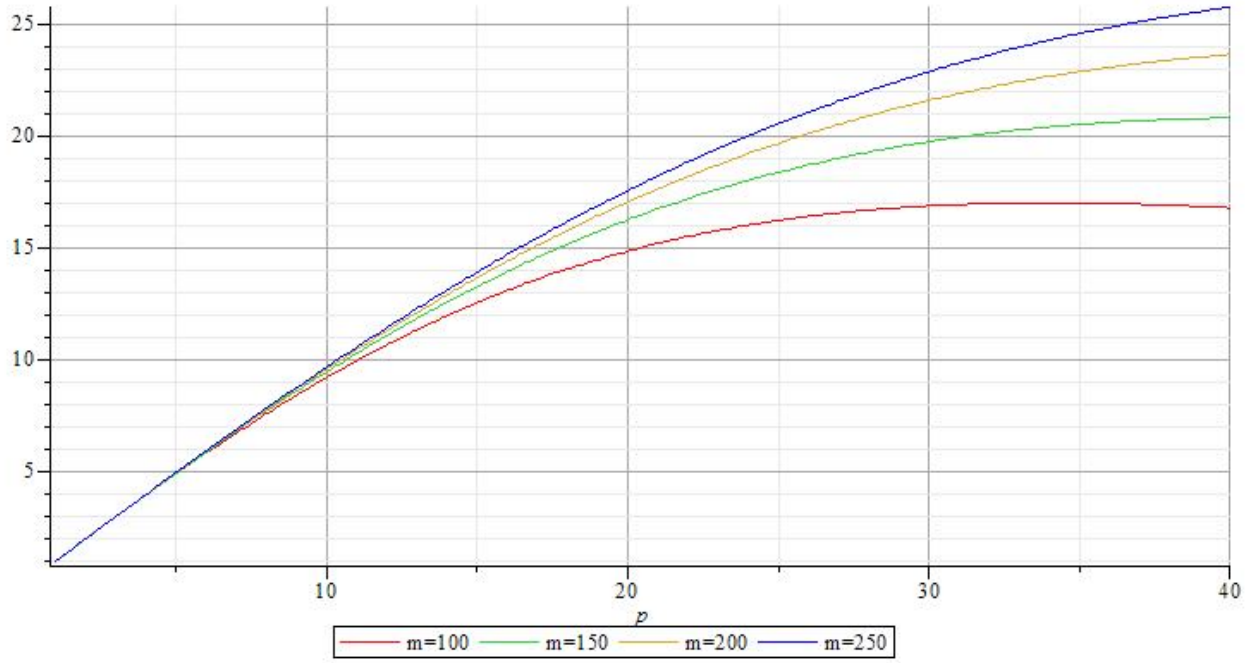
а)


60
50
40
30
20
10
10
20
30
40
50
60
p
nbl=150
nbl=200
nbl=250
Linear

б)


40
30
20
10
10
20
30
40
p
nbl=50
Linear

в)



г)

Рис. 4. Графики зависимости ускорения от числа процессоров при реализации параллельного алгоритма с учетом разреженности

Из приведенных графиков можно сделать следующие выводы:

- для любой размерности исходной задачи существует целесообразный диапазон изменения числа используемых процессоров p, в рамках которого возможно достижение максимального ускорения;
- чем больше размерность исходной задачи, тем а) шире диапазон изменения p и б) больше он сдвигается в сторону больших значений p;
- начиная с некоторого фиксированного числа процессоров (см., например, рис. 4в), ускорение снижается;
- при увеличении размерности блока при фиксированной исходной размерности задачи ускорение растет до определенного предела, а затем начинает снижаться. При этом происходит рост числа ненулевых элементов матрицы P.

Выведенная теоретическая оценка трудоемкости метода очистки позволяет для каждой заданной размерности задачи определить целесообразный диапазон использования процессоров p и возможное ускорение.

Теоретические оценки трудоемкости соответствуют ранее полученным экспериментальным данным [4,7,8] и позволяют давать прогнозы о достигаемых значениях ускорения и эффективности на конкретной вычислительной системе.

ЛИТЕРАТУРА:

1. B Lisa Pollack, Theresa L. Windus, Wibe A. de Jong, David A. Dixon. “Thermodynamic Properties of the C5, C6, and C8 n-Alkanes from ab Initio Electronic Structure Theory” // J. Phys. Chem. A, 2005, Vol. 109, № 31.
2. Stephen Pawlowski “Petascale Computing Research Challenges-A Manycore Perspective, 13th International Symposium on High-Performance Computer Architecture», 2007. URL: http://www2.engr.arizona.edu/~hpca/slides/2007_HPCA_Pawlowski.pdf.
3. Н.Ф. Степанов «Квантовая механика и квантовая химия», М., изд-ва “Мир” и “МГУ”, 2001, 518 с.
4. А.М. Чернецов, О.Ю. Шамаева «О параллельной реализации алгоритмов расчета электронной структуры больших молекул» // Вестник МЭИ, 2009, № 3., - С. 67-71.
5. A.M.N. Niklasson, C.J. Tymczak, M. Challacombe. “Trace resetting density matrix purification in O(N) self-consistent-field theory” // J.Chem.Phys., 2003. v.118, N15, p.1.
6. А.М. Чернецов, О.Ю. Шамаева, М.Б. Кузьминский «Распараллеливание в кластерах полуэмпирического квантово-химического метода Пальцера-Манолополиса для вычисления матрицы плотности сверхбольших молекул.» // Высокопроизводительные параллельные вычисления на кластерных системах. Материалы восьмого международного научно-практического семинара и всероссийской молодежной школы. Казань, 2008, С. 347-349.

7. А.М. Чернецов, О.Ю. Шамаева «Эффективные вычисления для расчета электронной структуры больших молекул» // Программные продукты и системы. - 2012. -№ 2., - С. 86-90.
8. А.М. Чернецов «Методы и программные средства организации эффективных вычислений для расчета молекулярных систем большой размерности» // Автореферат дисс. на соискание уч. степени канд. тех. наук, 2012.