УДК 519.612

ПАРАЛЛЕЛЬНАЯ РЕАЛИЗАЦИЯ ИТЕРАЦИОННОГО АЛГОРИТМА РЕШЕНИЯ НЕСИММЕТРИЧНЫХ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ С ЧАСТИЧНЫМ СОХРАНЕНИЕМ СПЕКТРАЛЬНОЙ/СИНГУЛЯРНОЙ ИНФОРМАЦИИ ПРИ ЯВНЫХ РЕСТАРТАХ

С. А. Харченко[1]

Предложена параллельная реализация итерационного алгоритма SOFGMRES(m) с частичным сохранением информации при явных рестартах. В алгоритме имеется важная степень свободы — произвольное начальное подпространство. Из обоснования сходимости алгоритма SOFGMRES(m) следует, что начальное подпространство при его правильном выборе можно рассматривать как дополнительное предобусловливание, поскольку оно уменьшает обобщенную обусловленность матрицы на подпространстве и ускоряет сходимость алгоритма. Результаты экспериментов показывают надежность, алгебраическую и параллельную эффективность предложенного алгоритма по сравнению с классическими алгоритмами типа подпространств Крылова. Работы выполнена в рамках государственного контракта 02.514.11.4125 с Министерством образования и науки РФ. Статья рекомендована к печати программным комитетом Международной научной конференции “Научный сервис в сети Интернет: суперкомпьютерные центры и задачи” (http://agora.guru.ru/abrau).

Ключевые слова: параллельный итерационный алгоритм, явные рестарты, обусловленность на подпространстве, дополнительное предобусловливание.

Введение. Моделирование задач вычислительной аэро- и гидродинамики в настоящее время является актуальной проблемой многих отраслей промышленности. Параллельная версия программного комплекса FlowVision [1, 2], разрабатываемого в ООО “ТЕСИС”, используется для моделирования широкого круга индустриальных задач, включая задачи с подвижными телами и свободными поверхностями для областей со сложной 3D-геометрией. В программном комплексе FlowVision используются неявные схемы аппроксимации, что приводит к необходимости решать соответствующие системы линейных уравнений. Для адекватного воспроизведения тонких физических эффектов в геометрически сложных трехмерных областях требуются подробные расчетные сетки, содержащие от сотен тысяч до сотен миллионов расчетных ячеек. Моделирование подобных задач требует огромных вычислительных ресурсов и может быть проведено только на самой современной мощной параллельной вычислительной технике.

При моделировании задач гидродинамики с использованием неявных схем аппроксимации системы линейных уравнений приходится решать многократно с достаточно высокой точностью, причем этот этап вычислений является одним из самых ресурсоемких. По этой причине повышение эффективности параллельных алгоритмов решения систем линейных уравнений является одним из способов повышения эффективности использования вычислительных ресурсов при моделировании.

Современные эффективные параллельные итерационные алгоритмы решения несимметричных систем линейных уравнений включают в себя построение предобусловливания и соответствующие итерации, как правило крыловского типа. В симметричном случае традиционно используется итерационный алгоритм CG [3], в несимметричном — это такие итерационные алгоритмы, как алгоритм Ланцоша [4], QMR [5] или BiCGSTAB [6]. Все упомянутые итерационные алгоритмы являются алгоритмами крыловского типа с короткими рекуррентными соотношениями. В несимметричном случае возможно также использование алгоритма GMRES [7] с полными ортогонализациями векторов и его аналогов. Существенным преимуществом алгоритма GMRES без рестартов является его высокая устойчивость по отношению к ошибкам округления в несимметричном случае, однако высокие затраты памяти и вычислений на ортогонализацию векторов делают его неконкурентоспособным. Для уменьшения затрат памяти и вычислений при ортогонализации векторов было предложено несколько модификаций этого алгоритма с неявными рестартами, таких как IRGMRES [8]. В настоящей статье рассматривается параллельная реализация алгоритма SOFGMRES [9] с явными рестартами и сохранением части информации при рестартах.

[1] ООО “ТЕСИС”, ул. Юннатов, 18, 127083, Москва; Вычислительный центр им. А. А. Дородницына РАН, ул. Вавилова, 40, 119333, Москва; науч. сотр., e-mail: skh@tesis.com.ru

1. Алгоритм SOFGMRES(m). Рассмотрим систему линейных уравнений

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{1}$$

где $A \in \mathbb{R}^{N\times N}$ — заданная невырожденная и, возможно, незнакоопределенная несимметричная матрица, $\boldsymbol{b} \in \mathbb{R}^N$ — заданный вектор правой части и $\boldsymbol{x} \in \mathbb{R}^N$ — неизвестный вектор. В дальнейшем будем предполагать, что предобусловливатель и ненулевое начальное приближение, если таковые имеются, уже учтены соответствующим образом в A и $\boldsymbol{b}$.

Основой алгоритма SOFGMRES(m) являются матричные соотношения [9] вида

$$AY_k = W_k R_k, \quad Y_k^{\mathrm{T}} Y_k = I_k, \quad W_k^{\mathrm{T}} W_k = I_k, \tag{2}$$

где $Y_k \in \mathbb{R}^{N\times k}$, $W_k \in \mathbb{R}^{N\times k}$, $R_k \in \mathbb{R}^{k\times k}$ и R_k — верхняя треугольная матрица. Соотношения (2) будем называть *матричными соотношениями в QR-форме*. Пусть $Y_k^{\perp}$ — дополнение к Y_k до ортонормированного базиса в $\mathbb{R}^N$ и $W_k^{\perp}$ — дополнение к W_k до ортонормированного базиса в $\mathrm{Span}\,(A)$.

Предположим, что для некоторого k, $0 \leqslant k < N$, имеют место соотношения (2). Обозначим через $y_{k+i} \in \mathbb{R}^N$, где $i = 1,\ldots,m$ и $k+m < N$, дополнительный набор направлений, такой, что матрица $\left[Y_k, y_{k+1}, \ldots, y_{k+m}\right]$ является матрицей с ортонормированными столбцами, т.е.

$$\left[Y_k, y_{k+1}, \ldots, y_{k+m}\right]^{\mathrm{T}} \left[Y_k, y_{k+1}, \ldots, y_{k+m}\right] = I_{k+m}.$$

Рассмотрим QR-разложение матрицы $\left[AY_k, \boldsymbol{b}, Ay_{k+1}, \ldots, Ay_{k+m}\right]$, где $\boldsymbol{b}$ — правая часть из (1), в соответствии с построениями работы [9]. Имеют место соотношения

$$\left[AY_k, \boldsymbol{b}, Ay_{k+1}, \ldots, Ay_{k+m}\right] = P_{k+m+1} S_{k+m+1}, \quad P_{k+m+1}^{\mathrm{T}} P_{k+m+1} = I_{k+m+1}, \tag{3}$$

где $P_{k+m+1} \in \mathbb{R}^{M\times(k+m+1)}$ — матрица с ортонормированными столбцами и $S_{k+m+1} \in \mathbb{R}^{(k+m+1)\times(k+m+1)}$ — верхняя треугольная матрица. Выделяя из соотношений (3) $(k+1)$-й столбец, отвечающий правой части $\boldsymbol{b}$, для $i = 0,\ldots,m$ матричные соотношения (3) можно переписать в виде

$$AY_{k+i} = P_{k+i+1} H_{k+i}, \quad Y_{k+i}^T Y_{k+i} = I_{k+i}, \quad P_{k+i+1}^T P_{k+i+1} = I_{k+i+1}, \quad \boldsymbol{b} = P_{k+i+1}\boldsymbol{\gamma}_{k+i+1}, \tag{4}$$

где $Y_{k+i} = \left[Y_k, y_{k+1}, \ldots, y_{k+i}\right]$, матрица P_{k+i+1} содержит первые $(k+i+1)$ столбцов матрицы P_{k+m+1}, первые $(k+1)$ компоненты вектора $\boldsymbol{\gamma}_{k+i+1} \in \mathbb{R}^{k+i+1}$ есть первые $(k+1)$ элементов $(k+1)$-го столбца матрицы S_{k+m+1}, а остальные компоненты нулевые, верхняя хессенбергова матрица H_{k+i} составлена из первых $(k+i+1)$ столбцов матрицы S_{k+m+1}, исключая $(k+1)$-й столбец.

Матрица H_{k+i} в соотношениях (4) верхняя хессенбергова, поэтому существует набор вращений Гивенса $G_j \in \mathbb{R}^{(k+i+1)\times(k+i+1)}$, $j = 1,\ldots,k+i$, $G_j = \begin{bmatrix} I_{j-1} & & 0 \\ & c_j & s_j \\ & -s_j & c_j \\ 0 & & I_{k+i-j} \end{bmatrix}$, $c_j^2 + s_j^2 = 1$, такой, что

$$H_{k+i} = Q_{k+i} \begin{bmatrix} R_{k+i} \\ 0 \end{bmatrix}, \tag{5}$$

где $Q_{k+i} = G_1 * \cdots * G_{k+i}$ — ортогональная матрица, которая представляет собой произведение матриц вращений Гивенса, и R_{k+i} — верхняя треугольная матрица. Равенство (5) является QR-разложением матрицы H_{k+i}. Заметим, что по построению первые k вращений Гивенса — единичные матрицы. Обозначим

$$W_{k+i} = P_{k+i+1} Q_{k+i} \begin{bmatrix} I_{k+i} \\ 0 \end{bmatrix}. \tag{6}$$

Тогда первое равенство из (4) можно представить в QR-форме: $AY_{k+i} = W_{k+i}R_{k+i}$. Соотношения (4) будем называть *рекуррентными матричными соотношениями алгоритма* SOFGMRES(m).

В табл. 1 представлено описание алгоритма SOFGMRES(m). Из описания алгоритма SOFGMRES(m) следует, что на каждой итерации этого алгоритма необходимо произвести одно умножение вектора невязки на предобусловленную матрицу и две ортогонализации наборов векторов. Очевидно также, что по сравнению с алгоритмом GMRES требуется хранить два набора векторов вместо одного.

Таблица 1

Алгоритм SOFGMRES(m)

Шаг 0.	Инициализация:	$x_0 = 0$; Пусть имеют место соотношения (4) при $i = 0$;
For $i = 1, \ldots, m$ Do		
Шаг 1.	Решение локальной задачи минимизации:	$\left\| H_{k+i-1} z_{i-1} - \begin{bmatrix} \gamma_{k+1} \\ 0 \end{bmatrix} \right\| \to \min\limits_{z_{i-1}}$;
Шаг 2.	Вычисление вектора редуцированной невязки:	$t_{i-1} = \begin{bmatrix} \gamma_{k+1} \\ 0 \end{bmatrix} - H_{k+i-1} z_{i-1}$;
Шаг 3.	Вычисление вектора текущей невязки:	$r_{i-1} = P_{k+i} t_{i-1}$;

Шаг 4.	Вычисление нового приближения к решению:	If $\|r_{i-1}\| \leqslant \varepsilon$ then
		$x_{i-1} = Y_{k+i-1} z_{i-1}$;
		Stop
		End If

Шаг 5.	Ортогонализация:	текущий вектор невязки r_{i-1} ортогонализуем к предыдущим векторам направлений; получаем новый нормированный вектор y_{k+i}, ортогональный направлениям $y_1, \ldots, y_{k+i-1}$, и соотношение $r_{i-1} = Y_{k+i} \left[g_{1,k+i}, \ldots, g_{k+i,k+i} \right]^{\mathrm{T}}$; (7)
Шаг 6.	Умножение предобусловленной матрицы A на вектор направления y_{k+i}:	$\widehat{p}_{k+i} = A y_{k+i}$;
Шаг 7.	Ортогонализация:	для текущего вектора $\widehat{p}_{k+i}$ достраиваем QR-разложение (4); обозначим $P_{k+i+1} = \left[P_{k+i}, p_{k+i+1} \right]$,; $Y_{k+i} = \left[Y_k, y_{k+1}, \ldots, y_{k+i} \right]$, $H_{k+i} = \begin{bmatrix} & & h_{1,k+i+1} \\ & H_{k+i-1} & \vdots \\ & & h_{k+i,k+i+1} \\ 0 & & h_{k+i+1,k+i+1} \end{bmatrix}$;
End For		

2. Обоснование сходимости алгоритма SOFGMRES(m). Приведем без доказательств основные утверждения из теории сходимости алгоритма SOFGMRES(m), подробные доказательства представлены в работе [9].

Лемма 1. *Для сингулярных чисел матриц A и R_{k+i} для всех $0 \leqslant i \leqslant m$ имеют место неравенства теоремы Коши о разделении.*

Лемма 2. *Для всех $0 < i \leqslant m$ имеет место соотношение $r_{i-1} \perp \mathrm{Span}\,(W_{k+i-1})$.*

Теорема 1. *Пусть первые k векторов направлений в матричных соотношениях* (6) *получены проведением k итераций алгоритма Арнольди с матрицей A и начальным вектором $\boldsymbol{b}/\|\boldsymbol{b}\|$. Пусть на всех итерациях внутреннего цикла алгоритма SOFGMRES(m) имеют место неравенства $g_{k+i,k+i} \neq 0$ для всех $0 < i \leqslant m$. Тогда полученные после m итераций внутреннего цикла алгоритма SOFGMRES матричные соотношения можно рассматривать как матричные соотношения алгоритма Арнольди.*

Лемма 3. *Имеют место неравенства* $\|r_i\| \leqslant \|r_{i-1}\|\sqrt{1-\theta_{i-1}^2\psi_{i-1}^2}$, *где* $\theta_{i-1} = \dfrac{\left|(Ar_{i-1}, r_{i-1})\right|}{\left\|r_{i-1}\right\|^2}$, $\psi_{i-1} = \left\|\left\{R_{k+i}\right\}_{*,k+i}\right\|^{-1}$ *и* $\theta_{i-1}\psi_{i-1} \leqslant 1$.

Определение. Пусть имеют место матричные соотношения в QR-форме (2). Обозначим

$$\Theta_A(Y_k) = \min_{\substack{w\perp \mathrm{Span}\,(W_k)\\ \|w\|\neq 0}} \frac{\left|(Aw,w)\right|}{\|w\|^2}, \quad \Psi_A(Y_k)^{-1} = \max_{\substack{y\perp \mathrm{Span}\,(Y_k)\\ \|y\|\neq 0}} \frac{\|Ay\|}{\|y\|}, \quad \eta_A(Y_k) = \Theta_A(Y_k)^*\Psi_A(Y_k).$$

Величину $\eta_A(Y_k)$ будем называть *обобщенным обратным числом обусловленности матрицы A на подпространстве* $\mathrm{Span}\,(Y_k)$.

Лемма 4. *Монотонность характеристик подпространств: если* $\mathrm{Span}\,(Y) \subset \mathrm{Span}\,(Z)$, *то*

$$\Theta_A(Y) \leqslant \Theta_A(Z), \quad \Psi_A(Y) \leqslant \Psi_A(Z), \quad \eta_A(Y) \leqslant \eta_A(Z).$$

Лемма 5. *Имеют место неравенства* $0 \leqslant \eta_A(Y_k) \leqslant 1$.

Теорема 2. *Имеют место неравенства*

$$\|r_i\| \leqslant \|r_{i-1}\|\sqrt{1-\eta_A^2\left([Y_k, y_{k+1}, \ldots, y_{k+i-1}]\right)} \leqslant \|r_{i-1}\|\sqrt{1-\eta_A^2(Y_k)}\,.$$

3. Сохранение части спектральной/сингулярной информации при рестартах. Из теоремы 1 следует, что если начальное подпространство отсутствует ($k = 0$) и на итерациях не было обрыва в вычислениях (равенство нулю коэффициента $g_{k+i,k+i}$ из (7)), то алгоритм SOFGMRES(m) математически эквивалентен алгоритму GMRES. При этом в алгоритме SOFGMRES(m) требуется в два раза больше памяти для хранения векторов и вдвое бóльшие затраты на ортогонализацию векторов. С другой стороны, в алгоритме SOFGMRES(m) в выборе начального подпространства имеется произвол. С точки зрения оценки теоремы 2, для быстрой сходимости алгоритма SOFGMRES(m) лучшим является такое начальное подпространство $\mathrm{Span}\,(Y_k)$, для которого обратное число обусловленности $\eta_A(Y_k)$ матрицы A на подпространстве $\mathrm{Span}\,(Y_k)$ максимально близко к единице.

Существуют различные способы нахождения подходящего начального подпространства $\mathrm{Span}\,(Y_k)$. Базовым способом нахождения такого подпространства является следующий: для небольшого m проводим цикл итераций алгоритма SOFGMRES(m), строим новое приближение к решению, из полученного нового набора направлений оставляем некоторые линейные комбинации векторов построенного набора, остальные направления отбрасываем, получаем новое начальное подпространство для последующих циклов итераций и т.д. Ясно, что фильтрация части направлений ухудшает сходимость по сравнению с полным сохранением подпространств (по существу, по сравнению с алгоритмом GMRES без рестартов). С другой стороны, существует естественный способ фильтрации новых направлений, при котором число сохраняемых направлений минимально и потери скорости сходимости из-за фильтрации векторов также минимальны.

Предлагаемый способ фильтрации основывается на оценке теоремы 2 [9]. Зададим два порога λ_*^+, $0 < \lambda_*^+ < 1$, и σ_*, $\sigma_* > 1$. Будем строить два поднабора $\widetilde{Y}_{l_1}$ и $\widetilde{\widetilde{Y}}_{l_2}$ линейных комбинаций текущих векторов $y_{k+1}, \ldots, y_{k+m}$, по возможности раздельно обеспечивающих выполнение неравенств

$$\lambda_A^+\Big(\big[Y_k, \widetilde{Y}_{l_1}\big]\Big) \geqslant \lambda_*^+, \quad \sigma_A\Big(\big[Y_k, \widetilde{\widetilde{Y}}_{l_2}\big]\Big) \leqslant \sigma_*. \tag{8}$$

Обозначим $R_{k+m} = \begin{bmatrix} R_k & \widehat{R}_m^{(12)} \\ 0 & \widehat{R}_m^{(22)} \end{bmatrix}$. Для обеспечения второго неравенства из (8) вычисляем правые сингулярные пары $\left\{\sigma_j, v_j\right\}$, $j = 1, \ldots, m$, матрицы $\begin{bmatrix} \widehat{R}_m^{(12)} \\ \widehat{R}_m^{(22)} \end{bmatrix}$ и полагаем параметр l_2 равным числу сингулярных чисел этой матрицы, превышающих заданный порог σ_*. Набор векторов блока $\widetilde{\widetilde{Y}}_{l_2}$ строим как линейную комбинацию векторов $y_{k+1}, \ldots, y_{k+m}$ с коэффициентами, равными компонентам соответствующего сингулярного вектора v_j. Такой выбор сохраняемого подпространства объясняется тем, что если бы имело место равенство $m = N - k$, то при таком выборе, очевидно, в дополнительном подпространстве

не осталось бы сингулярных чисел, превышающих порог σ_*. Для обеспечения первого неравенства из (8) вычислим матрицу $T_m = \left[w_{k+1}, \ldots, w_{k+m}\right]^{\mathrm{T}} \left[y_{k+1}, \ldots, y_{k+m}\right] \left(\widehat{R}_m^{(22)}\right)^{\mathrm{T}}$.

Для симметризованной матрицы ${T_m}^+ = \dfrac{T_m + T_m^{\mathrm{T}}}{2}$ вычислим все собственные пары этой матрицы $\left\{\lambda_j, x_j\right\}$, $j = 1, \ldots, m$. Полагаем параметр l_1 равным числу собственных значений матрицы ${T_m}^+$, меньших заданного порога λ_*^+. Базис подпространства $\mathrm{Span}\left(\widetilde{Y}_{l_1}\right)$ строим как ортонормированную линейную комбинацию векторов $y_{k+1}, \ldots, y_{k+m}$ с коэффициентами $\left(\widehat{R}_m^{(22)}\right)^{-1} x_j$ для соответствующих собственных векторов x_j. Такой выбор базиса подпространства $\mathrm{Span}\left(\widetilde{Y}_{l_1}\right)$ объясняется рассмотрением предельного случая $m = N - k$. В этом случае имеет место тождество ${T_m}^+ = \left(W_k^{\perp}\right)^{\mathrm{T}} A_+ W_k^{\perp}$, а для построенного таким образом $\mathrm{Span}\left(\widetilde{Y}_{l_1}\right)$ будет выполнено первое неравенство из (8).

Для обеспечения одновременного выполнения неравенств (8) в силу монотонности характеристик подпространств (лемма 4) теперь достаточно вычислить единый ортонормированный базис $\widehat{Y}_l$, такой, что $\left(\mathrm{Span}\left(\widetilde{Y}_{l_1}\right) \cup \mathrm{Span}\left(\widetilde{\widetilde{Y}}_{l_2}\right)\right) \subset \mathrm{Span}\left(\widehat{Y}_l\right)$.

Для минимизации числа сохраненных направлений при большом количестве рестартов, в алгоритме SOFGMRES иногда целесообразно проводить так называемые двойные и даже тройные фильтрации, при которых в фильтрации участвуют не только текущие направления, но и группы направлений, сохраненных на предыдущих циклах алгоритма.

4. Комбинированная MPI+threads параллельная реализация ортогонализации векторов на основе преобразований Хаусхолдера. С точки зрения параллельной реализации основными вычислительными операциями алгоритма SOFGMRES являются умножение матрицы системы уравнений на вектор, умножение предобусловливателя на вектор (эти две операции в алгоритме SOFGMRES для простоты изложения сведены к одной), а также ортогонализация векторов и операция вычисления линейной комбинации векторов направлений. Параллельные алгоритмы умножения матрицы системы уравнений и предобусловливателя на вектор подробно рассмотрены в других работах [10–12], поэтому здесь мы остановимся только на параллельной реализации алгоритмов, связанных с ортогонализациями векторов.

Хорошо известно [13], что с точки зрения минимизации ошибок округления ортогонализацию векторов целесообразно производить неявно с использованием преобразований Хаусхолдера, как это делается в случае вычисления QR-разложения прямоугольной матрицы.

Пусть для начала для некоторой прямоугольной матрицы $C \in \mathbb{R}^{N \times k}$, $k << N$, построено QR-разложение с применением преобразований Хаусхолдера в виде произведения

$$C = \left(\prod_{i=1}^{k} G_i\right) \begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{9}$$

где $R \in \mathbb{R}^{k \times k}$ — квадратная верхняя треугольная матрица, $G_i \in \mathbb{R}^{N \times N}$ — ортогональные матрицы преобразования Хаусхолдера, имеющие вид $G_i = I_N + \tau_i u_i u_i^{\mathrm{T}}$, τ_i — некоторое число, $u_i \in \mathbb{R}^N$ — некоторый вектор, первые $(i-1)$ компонент которого нулевые. Расширение QR-разложения для расширенной на один столбец матрицы $[C, c] \in \mathbb{R}^{N \times (k+1)}$, как это хорошо известно, можно вычислить применением в соответствующем порядке транспонированных преобразований Хаусхолдера к столбцу c и последующим добавлением из результата нового столбца в матрицу R и построением нового преобразования Хаусхолдера.

В контексте ортогонализации векторов из QR-разложения (9) можно явно выделить ортонормированные векторы базиса подпространства $\mathrm{Span}\,(C)$ в виде произведения $p_j = \left(\prod_{i=1}^{k} G_i\right) \begin{bmatrix} e_j \\ 0 \end{bmatrix}$.

В случае параллельных вычислений разные части матрицы C обычно расположены на разных процессорах. В этом случае QR-разложение имеет смысл вычислять в распределенном виде способом, отличным от (9). Пусть теперь прямоугольная матрица $C \in \mathbb{R}^{N \times k}$, $k << N$, имеет вид

$$C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}, \tag{10}$$

где матрицы C_1 и C_2 имеют соответственно N_1 и N_2 строк, $N = N_1 + N_2$. Построим ее QR-разложение на основе преобразований Хаусхолдера с явным использованием блочной структуры матрицы C. Для этого

строим на основе преобразований Хаусхолдера QR-разложения блочных строк матрицы C:

$$C_1 = Q_1 R_1, \quad C_2 = Q_2 R_2. \tag{11}$$

Дополнительно к (11) на основе преобразований Хаусхолдера вычислим QR-разложение

$$\begin{bmatrix} R_1 \\ R_2 \end{bmatrix} = \begin{bmatrix} \Theta_1 \\ \Theta_2 \end{bmatrix} R. \tag{12}$$

Объединяя соотношения (10)–(12), можно получить соотношение $C = QR$, где матрица

$$Q = \begin{bmatrix} Q_1 {*} \Theta_1 \\ Q_2 {*} \Theta_2 \end{bmatrix} \tag{13}$$

имеет ортонормированные столбцы, а матрица R — верхняя треугольная. Тем самым получено QR-разложение матрицы C. Во всех соотношениях (11)–(13) при написании факторов QR-разложения имеется в виду неявное мультипликативное представление Q-фактора в виде произведения преобразований Хаусхолдера вида (9).

Рассмотрим теперь организацию вычислений в случае добавления в матрицу C одного столбца $c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$. На основе преобразований Хаусхолдера строим QR-разложения блочных строк расширенной матрицы $\big[C, c\big] \in \mathbb{R}^{N \times (k+1)}$:

$$\big[C_1, c_1\big] = \big[Q_1, q_1\big] \begin{bmatrix} R_1 \, r_1 \\ 0 \;\; \widetilde{r}_1 \end{bmatrix}, \quad \big[C_2, c_2\big] = \big[Q_2, q_2\big] \begin{bmatrix} R_2 \, r_2 \\ 0 \;\; \widetilde{r}_2 \end{bmatrix}. \tag{14}$$

На основе преобразований Хаусхолдера достраиваем QR-разложение (12):

$$\begin{bmatrix} R_1 \, r_1 \\ R_2 \, r_2 \\ 0 \;\; \widetilde{r}_1 \\ 0 \;\; \widetilde{r}_2 \end{bmatrix} = \begin{bmatrix} \Theta_1 \, \theta_1 \\ \Theta_2 \, \theta_2 \\ 0 \;\; \widetilde{\theta}_1 \\ 0 \;\; \widetilde{\theta}_2 \end{bmatrix} \begin{bmatrix} R \, r \\ 0 \; \widetilde{r} \end{bmatrix}. \tag{15}$$

Из соотношений (14) и (15) следует равенство $\big[C, c\big] = \widetilde{Q} \begin{bmatrix} R \, r \\ 0 \; \widetilde{r} \end{bmatrix}$, где

$$\widetilde{Q} = \begin{bmatrix} Q_1 {*} \Theta_1 \; Q_1 {*} \theta_1 + q_1 {*} \widetilde{\theta}_1 \\ Q_2 {*} \Theta_2 \; Q_2 {*} \theta_2 + q_2 {*} \widetilde{\theta}_2 \end{bmatrix} = \begin{bmatrix} Q_1 \, q_1 \;\; 0 \;\; 0 \\ 0 \;\; 0 \; Q_2 \, q_2 \end{bmatrix} \begin{bmatrix} \Theta_1 \, \theta_1 \\ 0 \;\; \widetilde{\theta}_1 \\ \Theta_2 \, \theta_2 \\ 0 \;\; \widetilde{\theta}_2 \end{bmatrix}, \tag{16}$$

причем $\widetilde{Q}$ — матрица с ортонормированными столбцами. Заметим также, что в QR-разложении (15) в матрице с левой стороны равенства число строк вдвое превышает число столбцов. Из формулы (16) следует, что для умножения на Q-фактор QR-разложения требуется сначала входной вектор умножить на Q-фактор разложения (15), переставить компоненты результата, а затем локально раздельно умножить на Q-факторы первой и второй блочных строк.

По аналогии с описанной выше техникой можно организовать распределенное вычисление QR-разложения прямоугольной матрицы и в случае, когда эта матрица разбивается на число блочных строк, большее двух, например на t блочных строк, $t > 2$. В этом случае в соответствующем совместном QR-разложении типа (15) число строк будет в t раз больше, чем число столбцов.

Пусть теперь имеется параллельный компьютер с неоднородным доступом к памяти. А именно, пусть имеется набор узлов с распределенной памятью и в каждом узле имеется несколько многоядерных процессоров, работающих на общей памяти. На компьютере с такой неоднородной архитектурой естественно организовать комбинированные параллельные вычисления, в которых обмены и синхронизации по распределенной памяти осуществляются на основе MPI (Message Passing Interface), а по общей — на основе какого-либо из механизмов работы с потоками, например на основе технологии Intel TBB [14]. Для такого компьютера имеется MPI+threads реализация алгоритма решения систем линейных уравнений, например [12]. При этом предполагается, что число MPI-процессов равно числу процессоров, а каждый

MPI-процесс порождает число потоков, равное числу ядер процессора [12]. Дополнительные вычисления, связанные с ортогонализацией векторов на основе преобразований Хаусхолдера, можно организовать для алгоритма SOFGMRES(m) следующим образом.

Распределение по MPI-процессам матрицы системы уравнений порождает распределение по MPI-процессам строк матрицы, для которой необходимо вычислить QR-разложение на основе преобразований Хаусхолдера. Все строки этой матрицы для текущего MPI-процесса разбиваются на число частей, равное числу потоков у этого MPI-процесса. QR-разложение своей части строк матрицы вычисляется потоками независимо. Затем для достроения единого QR-разложения достраивается объединяющий QR для текущего MPI-процесса. Аналогично для распределенной памяти: сначала все MPI-процессы достраивают свои части QR-разложений с учетом разбиения по блочным строкам, а затем R-части QR-разложений собираются, например на нулевом процессе, и достраивается заключительный объединяющий QR. При умножении на Q-фактор информация движется в обратном направлении: сначала производится умножение на Q-фактор объединяющих QR-разложений, затем после обмена, если это необходимо, локальными переставленными частями вектора результата производится умножение на локальный Q-фактор этого MPI-процесса/потока вычислений.

5. Начальное подпространство как дополнительное предобусловливание для ускорения сходимости алгоритма SOFGMRES(m). Из теории сходимости алгоритма SOFGMRES(m) следует, что чем ближе обобщенное обратное число обусловленности на начальном или построенном подпространстве к единице, тем быстрее сходимость. При этом не имеет значения, каким именно образом было найдено соответствующее начальное подпространство. В предыдущем разделе описан естественный итерационный способ построения нужного подпространства. В данном разделе рассмотрим возможные неитерационные способы построения подпространства и способы, основанные на предыстории вычислений.

При решении системы линейных уравнений с одной и той же матрицей и последовательно возникающими правыми частями естественно использовать финальное подпространство, построенное для предыдущей правой части, как начальное подпространство при решении системы со следующей правой частью. Если правых частей много, то можно последовательно расширять начальное подпространство вплоть до такой скорости сходимости, при которой достигается баланс между затратами памяти и вычислений на ортогонализацию большого числа векторов и количеством итераций.

В инженерных приложениях часто возникает ситуация, когда на последовательности шагов по времени необходимо решать системы уравнений одного и того же типа со слабо меняющимися матрицами коэффициентов. Традиционно в таких случаях для уменьшения затрат на некотором числе шагов по времени делаются попытки сохранить предобусловливание. Оказывается, что вне зависимости от сохранения и/или смены предобусловливания можно сохранять и преобразовывать подпространство.

Пусть Y_k — подпространство, построенное для предыдущей матрицы и предыдущего предобусловливания. Преобразуем его к виду, не зависящему от предыдущего предобусловливания, умножением на матрицу LU-разложения: $\widetilde{Y}_k = (LU)Y_k$. В этом виде физический смысл полученного набора векторов — это некоторое подпространство полей искомой физической величины, например давления, если решалось уравнение для давления. Этот набор полей снова может быть преобразован в новый предобусловленный вид обратным преобразованием $\widetilde{\widetilde{Y}}_k = (L_{\text{new}}U_{\text{new}})^{-1}\widetilde{Y}_k$, где $(L_{\text{new}}U_{\text{new}})^{-1}$ — новый предобусловливатель для новой системы уравнений. Полученный набор векторов $\widetilde{\widetilde{Y}}_k$ можно использовать как базис для построения нового начального подпространства для новой предобусловленной матрицы.

Для того чтобы подобное сохранение начального подпространства имело содержательный смысл, необходимо, чтобы некоторые поля физических величин, полученные в качестве промежуточной информации, не сильно и желательно непрерывно менялись при изменении задачи.

Если при последовательном решении систем уравнений одного и того же типа произошло некоторое небольшое изменение геометрии задачи и изменилась расчетная сетка, то даже и в этом случае имеется содержательная возможность для сохранения информации о предыдущем подпространстве при решении задачи на новом шаге по времени. Для этого аналогично предыдущему преобразуем подпространства из вида, зависящего от предобусловливания, в вид, имеющий конкретный физический смысл. Как правило, в задачах с изменяющейся геометрией всегда имеется возможность пересчета с помощью процедуры типа интерполяции всех величин, имеющих физический смысл, с одной расчетной сетки на другую. Применим эту процедуру к построенному набору полей физической величины. В результате получим некоторый набор полей физической величины на новой сетке. Применим обратное преобразование набора векторов для приведения его к новому предобусловливателю, как это описано в предыдущем разделе. Полученное подпространство можно использовать как начальное при итерационном решении с помощью алгоритма SOFGMRES(m) системы уравнений с физической величиной на новой расчетной сетке.

6. Результаты численных экспериментов. Численные эксперименты по решению системы линейных уравнений проводились с симметричной положительно определенной матрицей порядка $N = 1\,135\,596$, число ненулевых элементов Nz в верхнем треугольнике матрицы равно 75 651 236. Тестовая система уравнений возникает при моделировании перемещений напряженно-деформированного состояния для задачи со сложной 3D-геометрией. Выбор в качестве тестовой матрицы симметричной положительно определенной матрицы обусловлен необходимостью провести прямое сравнение алгоритма SOFGMRES(m) с традиционными итерационными алгоритмами (табл. 2).

Таблица 2
Сравнение итерационных методов в режиме 1MPI × 4threads

Метод	Число итераций	Время, сек.
CG	480	104
Алгоритм Ланцоша	531	252
SOFGMRES(m)	525	260

Заполнение предобусловливателя для тестовой задачи составляло 185% от заполнения матрицы. В алгоритме SOFGMRES(m) использовались следующие параметры: размер базового цикла $m = 10$, нижний порог фильтрации $\lambda = 0.001$, верхний порог фильтрации $\sigma = 2.0$, количество циклов итераций до двойной фильтрации — 10, количество двойных фильтраций до тройной фильтрации — 3. При фильтрации направлений после любого цикла сохранялось как минимум одно направление, соответствующее наименьшему собственному значению матрицы ${T_m}^+$. Критерием остановки итераций во всех экспериментах было относительное уменьшение нормы невязки для нулевого начального приближения в 10^9 раз. Расчеты проводились на процессоре Intel Core I7 920 (Nehalem), 4 ядра, частота 2.67 GHz.

В алгоритме SOFGMRES(m) число сохраненных векторов после окончания итераций $k = 13$, максимальное число векторов для хранения данных в процессе итераций равно 58.

Предобусловливатель и число итераций в табл. 3 не зависят от режима вычислений и равно 525 благодаря специальной технологии организации параллельных вычислений, описанной в работе [12].

Таблица 3
Масштабируемость алгоритма SOFGMRES(m) в различных режимах вычислений

Режим, MPI × threads	1 × 1	2 × 1	1 × 2	4 × 1	2 × 2	1 × 4
Время, сек.	642	354	431	231	259	260

Для проведения эксперимента по выяснению свойств подпространства как дополнительного предобусловливания была выбрана пара матриц $A^{(I)}$ и $A^{(II)}$, возникающих при моделировании задач вычислительной гидродинамики на двух последовательных шагах по времени, число неизвестных задачи $N = 3\,675\,240$, $N_z = 86\,429\,074$. Нумерация, блочное разбиение задачи и параметры предобусловливания были неизменными во всех экспериментах. Сравнивались сходимость алгоритма SOFGMRES(m) для двух последовательных правых частей для обеих матриц с сохранением подпространства после первой правой части, а также сходимость для второй матрицы с сохраненным и преобразованным начальным подпространством после решения первой.

1) Пара $\left(A^{(I)}, A^{(I)}\right)$. Для первой правой части сходимость алгоритма SOFGMRES(m) достигнута после 263 итераций, время — 342 с. Для второй правой части с сохраненным предобусловливанием и начальным подпространством размерности 6 сходимость достигнута за 86 итераций, время — 127 с.

2) Пара $\left(A^{(II)}, A^{(II)}\right)$. Для первой правой части сходимость алгоритма SOFGMRES(m) достигнута после 423 итераций, время — 560 с. Для второй правой части с сохраненным предобусловливанием и начальным подпространством размерности 6 сходимость достигнута за 80 итераций, время — 122 с.

3) Пара $\left(A^{(I)}, A^{(II)}\right)$. Для первой правой части сходимость алгоритма SOFGMRES(m) достигнута после 263 итераций, время — 344 с. Начальное подпространство для второй матрицы в алгоритме SOFGMRES(m) размерности 6 построено перевычислением финального подпространства, полученного проведением итераций для первой матрицы. Сходимость достигнута за 126 итераций, время — 193 с.

Таким образом, в результате сохранения подпространства время итераций уменьшилось с 560 до 193 секунд, и это несмотря на то, что начальное пространство строилось для другой матрицы.

Заключение. Предложена параллельная реализация итерационного алгоритма SOFGMRES(m) для компьютеров с неоднородным доступом к памяти. Численный эксперимент показал, что даже при малом размере базового цикла ($m = 10$) и при огромном числе рестартов в алгоритме (52 рестарта) с точки зрения сходимости алгоритм SOFGMRES(m) ведет себя так, как будто рестартов итераций нет совсем. Число итераций и время вычислений для тестовой задачи почти те же, что и у алгоритма Ланцоша (на одну итерацию в алгоритме Ланцоша приходится два умножения на предобусловленную матрицу), а время вычислений всего приблизительно в 2.5 раза выше, чем у классического алгоритма CG, применимого только в симметричном положительно определенном случае. Эксперименты также показали, что с точки

зрения масштабируемости вычислений на тестовом компьютере более эффективно распараллеливание по MPI, чем по нитям вычислений. Это, возможно, связано с архитектурными особенностями процессора Intel Core I7 920 (Nehalem) и с особенностями использования памяти процессора в различных режимах вычислений.

Проведенные эксперименты с итерационным алгоритмом SOFGMRES(m) с произвольным начальным подпространством и фильтрацией подпространств показывают его высокую эффективность по сравнению с традиционными алгоритмами типа подпространств Крылова. Как показывают эксперименты, дополнительная степень свободы этого алгоритма — начальное подпространство — при правильном выборе подпространства позволяет значительно (до 4 раз и выше) ускорить сходимость по сравнению с классическими алгоритмами типа подпространств Крылова.

СПИСОК ЛИТЕРАТУРЫ

1. *Aksenov A., Dyadkin A., Pokhilko V.* Overcoming of barrier between CAD and CFD by modified finite volume method // Proc. 1998 ASME Pressure Vessels and Piping Division Conference. San Diego, ASME PVP. 1998.
2. *Aksenov A.A., Kharchenko S.A., Konshin V.N., Pokhilko V.I.* FlowVision software: numerical simulation of industrial CFD applications on parallel computer systems // Parallel CFD 2003 Conference. Book of Abstracts. Moscow, 2003. 280–284.
3. *Тыртышников Е.Е.* Краткий курс численного анализа. Москва: ВИНИТИ, 1994.
4. *Paige C.C., Saunders M.A.* LSQR: An algorithm for sparse linear equations and sparse least squares // ACM Transactions on Mathematical Software. 1982. **8**, N 1. 43–71.
5. *Freund R.W., Nachtigal N.M.* QMR: a quasi-minimal residual method for non-Hermitian linear systems // Numer. Math. 1991. **60**. 315–339.
6. *Van der Vorst H.A.* Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems // SIAM J. Sci. Stat. Comput. 1992. **13**, N 2. 631–644.
7. *Saad Y., Schultz M.H.* GMRES: A generalized minimum residual algorithm for solving non-symmetric linear systems // SIAM J. Sci. Stat. Comput. 1986. **7**. 856–869.
8. *Morgan R.B.* Implicitly restarted GMRES and Arnoldi methods for nonsymmetric systems of equations // SIAM J. Matrix Anal. Appl. 2000. **21**, N 4. 1112–1135.
9. *Харченко С.А., Еремин А.Ю.* Новые алгоритмы типа GMRES(k) с рестартами и анализ их свойств сходимости на основе QR-формы матричных соотношений // Зап. науч. семин. ЛОМИ. 2000. **268**. 190–241.
10. *Дядькин А.А., Харченко С.А.* Алгоритмы декомпозиции области и нумерации ячеек с учетом локальных адаптаций расчетной сетки при параллельном решении систем уравнений в пакете FlowVision // Тр. Международной научной конференции “Научный сервис в сети Internet: многоядерный компьютерный мир”. Москва, 2007. 201–206.
11. *Харченко С.А.* Влияние распараллеливания вычислений с поверхностными межпроцессорными границами на масштабируемость параллельного итерационного алгоритма решения систем линейных уравнений на примере уравнений вычислительной гидродинамики // Тр. Международной научной конференции “Параллельные вычислительные технологии (ПаВТ’2008)”, Санкт-Петербург, 28 января–1 февраля 2008 г. Челябинск: Изд-во ЮУрГУ, 2008. 494–499.
12. *Сушко Г.Б., Харченко С.А.* Экспериментальное исследование на СКИФ МГУ “Чебышев” комбинированной MPI+threads реализации алгоритма решения систем линейных уравнений, возникающих во FlowVision при моделировании задач вычислительной гидродинамики // Тр. Международной научной конференции “Параллельные вычислительные технологии (ПаВТ’2009)”, Нижний Новгород, 30 марта–3 апреля 2009 г. Челябинск: Изд-во ЮУрГУ, 2009. 316–324.
13. *Walker H.F.* Implementation of the GMRES method using householder transformations // SIAM J. on Sci. and Stat. Comp. 1988. **9**, N 1. 152–163.
14. Intel Threading Building Blocks Tutorial-1.6 / Intel Corp. 2007.

Поступила в редакцию
30.10.2010