УДК 004.056.53:004.272

Р.Т. Файзуллин, Е.В. Щерба, Д.А. Волков

Схема реализации параллельных вычислений как инструмент защиты обрабатываемых данных

Предлагается схема разделения вычислений и хранения данных в центрах обработки данных, гарантирующая невозможность восстановления матрицы системы линейных уравнений на отдельных вычислительных узлах. Доказана применимость данного подхода в случае численного решения некоторых краевых задач для уравнений математической физики.
Ключевые слова: распределенные вычисления, разделение секрета, ЦОД, Грид-система.

Проблема защиты распределенных вычислений. Консолидация обработки и хранения больших массивов информации в центрах обработки данных – одно из самых перспективных направлений совершенствования корпоративных систем. Применение и внедрение ЦОД позволяют наиболее эффективно использовать коллективные вычислительные ресурсы, уменьшают общее число оборудования, снижают расходы на их поддержку [1, 2].

Необходимым элементом ЦОД является гарантированная защита информации. Основными направлениями защиты информации являются: защита от вирусных атак, обеспечение безопасности процесса взаимодействия информационных систем ЦОД с внешними источниками информации и, что наиболее важно, защита информации от несанкционированного доступа. Несмотря на все усилия у пользователей имеется определенная и обоснованная степень недоверия к уровню защиты ЦОД и к облачным вычислениям, основанная на угрозе атаки сговором (collusion attack). Даже наличие криптографических средств защиты информации и требуемых лицензий у поставщика облачных услуг не является для потенциальных потребителей достаточной гарантией защищенности процессов хранения и обработки информации.

Указанная практическая задача реализации защищенных распределенных вычислений (secure multi party computation, MPC) может быть решена с помощью различных схем разделения секрета (СРС). Традиционно в данных схемах присутствует постановщик задачи (клиент, input party, IP), распределяющий данные по вычислительным кластерам (computation parties, CP), и получатель результата (result party, RP) – зачастую клиент (IP). Классические СРС имеют ряд недостатков (высокие накладные расходы на коммуникации, отсутствие гарантии защищенности). Представляется возможным предложить такие СРС, в которых малая, но определяющая информативность часть секрета хранится или обрабатывается у клиента [3–5]. Данный подход позволяет добиться того, что в ЦОД хранятся данные, но не сама информация, и в каждом случае можно привести прозрачное для клиента доказательство того, что по большему массиву данных нельзя в принципе восстановить значимую информацию.

В качестве примера можно рассмотреть такую массовую задачу, как решение СЛАУ. Следует учесть, что структура матриц систем кодирует структуру моделируемых объектов, а задача декодирования и восстановления информации об объекте довольно проста. Поэтому возникает вопрос о правомерности передачи массовых вычислений в облако и об использовании неконтролируемых вычислительных ресурсов. В этом случае естественным выглядит использование распараллеливания вычислений в качестве инструмента разделения секрета со строгим доказательством сохранения секрета при использовании неполных данных.

Предлагаемые схемы параллелизации. Рассмотрим задачу решения большой системы линейных алгебраических уравнений $\mathbf{GX}=\mathbf{F}$, где квадратная матрица $\mathbf{G}$ не вырождена. Предположим, что некий достаточно большой главный минор $\mathbf{A}$ матрицы $\mathbf{G}$ также не вырожден. Тогда можно записать:

$$\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{pmatrix}=\begin{pmatrix}\mathbf{F}_1\\ \mathbf{F}_2\end{pmatrix}$$

и привести систему к виду

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B} \end{pmatrix}\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}=\begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2-\mathbf{C}\mathbf{A}^{-1}\mathbf{F}_1 \end{pmatrix}.$$

Если размерность матрицы $\mathbf{D}$ намного меньше, чем размерность $\mathbf{A}$, клиент может получать значения $\mathbf{X}_2$ на своей вычислительной системе, а наиболее трудоемкие операции по вычислению $\mathbf{A}^{-1}, \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ передавать на общедоступные вычислительные ресурсы.

Обратим внимание, что это далеко не искусственная задача. Например, расчет сложных гидравлических систем сводится к решению систем нелинейных алгебраических уравнений, которые описывают два закона Кирхгофа применительно к графу системы: закон сохранения массы в узлах и закон сохранения энергии вдоль цикла. Размерность $\mathbf{D}$ в этом случае равна числу линейно независимых циклов в графе, что намного меньше числа узлов [6]. Расчет $\mathbf{A}^{-1}$, $\mathbf{A}^{-1}\mathbf{B}$, $\mathbf{A}^{-1}\mathbf{F}_1$ можно произвести только один раз, если граф не изменяется во время эксплуатации системы, так как нелинейность системы определяется матрицами $\mathbf{C}$ и $\mathbf{D}$ и правой частью $\mathbf{F}_2$. В случае массовых расчетов распараллеливание происходит по вариантам, определяемым различными $\mathbf{C}$, $\mathbf{D}$, $\mathbf{F}_2$, т.е. заданием регуляторов расхода и давления и напорами.

Другая массовая задача – вычисление обратной матрицы – может быть решена с помощью формулы Фробениуса [7]:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1}=\begin{pmatrix} \mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{B}\mathbf{H}^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{H}^{-1} \\ -\mathbf{H}^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{H}^{-1} \end{pmatrix},$$

где $\mathbf{H}=\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ и $\mathbf{A}^{-1}$ существует.

Можно предложить много схем жесткого распараллеливания, но представляется возможным ввести простые правила:

1) два процессора передают данные промежуточных расчетов блоков на третий, если на третьем ранее не обрабатывался какой-либо из блоков матрицы;

2) вычисления следует распределять таким образом, чтобы на каждом процессоре в любой момент времени присутствовало не более чем два блока исходной или промежуточных матриц, причем не лежащих в одной «строке». Это позволяет избежать возможности попытки нахождения нормального решения, которое может дать некоторую информацию о точном решении.

Обратим внимание, что в этом случае мы получаем выигрыш в вычислениях в отличие от предыдущего случая, так как число операций для вычисления обратной матрицы $O(N^4)$, и, начиная с некоторого L, сравнимого с N, получается, что $O((N-L)^4)+O((N-L)^3)<O(N^4)$.

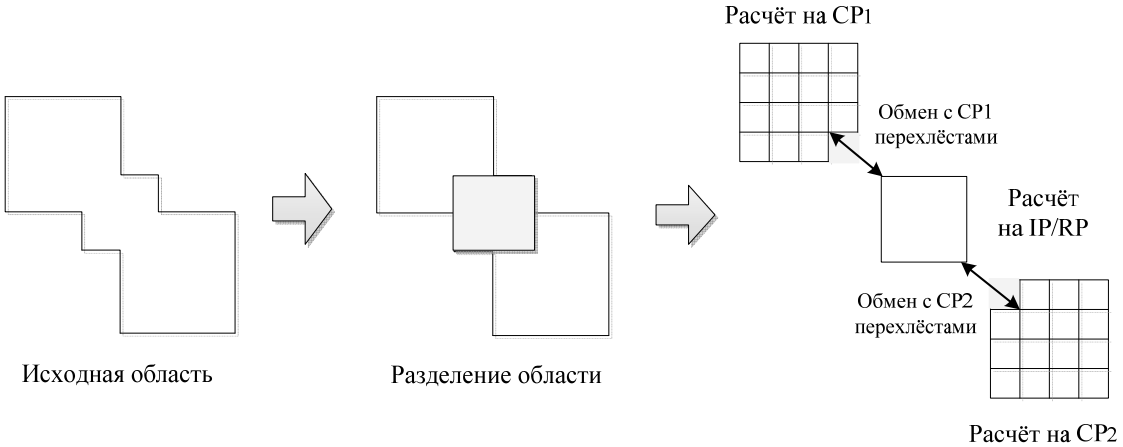

Рис. 1. Распределение вычислений

Если обратиться к итерационным методам и системам уравнений, ассоциированным с эллиптическими задачами, полученными с помощью метода конечных разностей или методом конечных элементов, то решение задачи разделения секрета можно получить с помощью альтернирующего метода Шварца. В этом случае на вычислительной системе клиента (IP) будут осуществляться опе-

рации, ассоциированные с границами подобластей (рис. 1). Основные вычислительные операции, например вычисления в областях сгущения расчетных точек, возлагаются на арендуемые кластеры (CP), а обмен осуществляется пересылкой граничных данных [8].

Общая организация схемы вычислений представлена на рис. 2. Обратим внимание на то, что в качестве управляющего сервера может и должен выступать компьютер клиента (IP/RP).

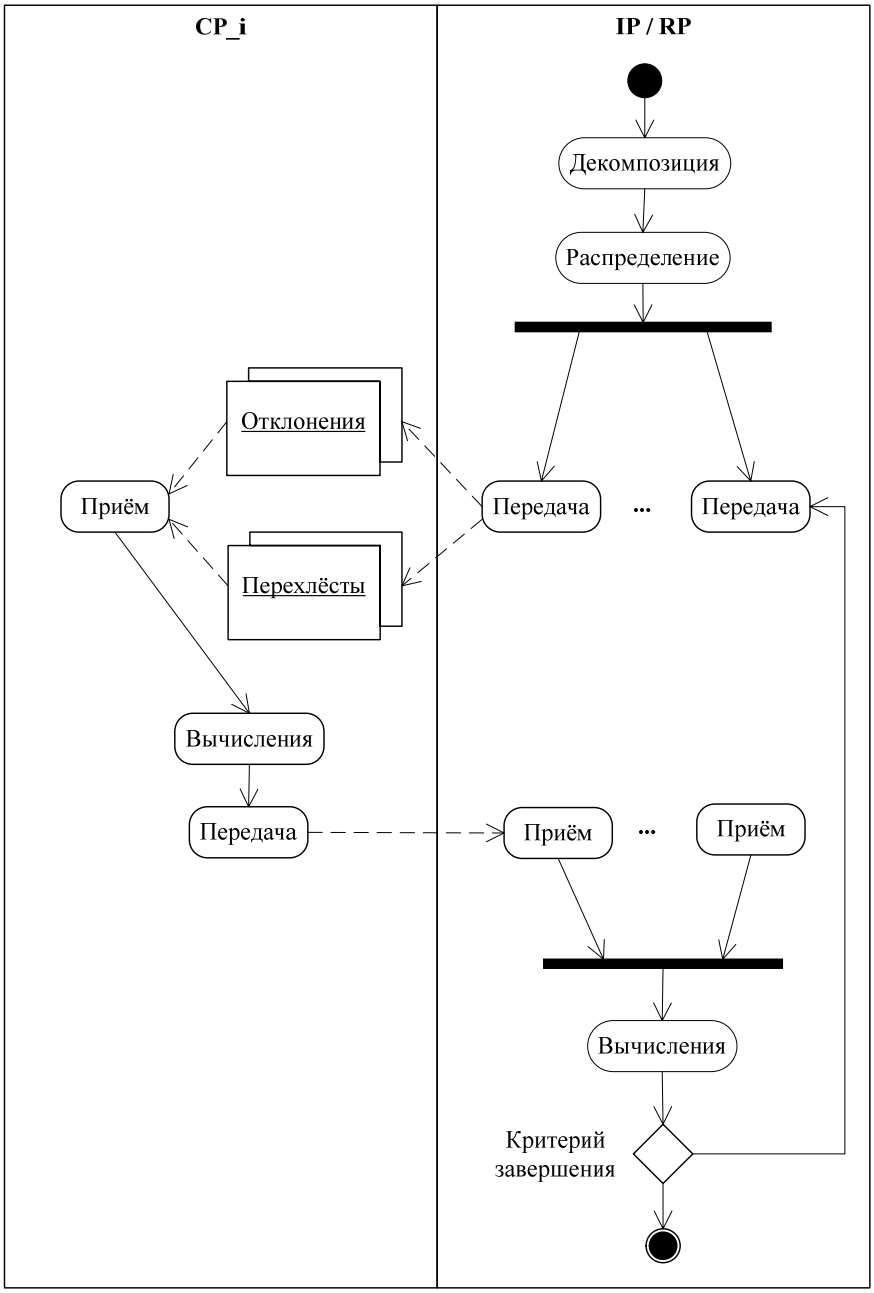

Рис. 2. UML-диаграмма управления вычислениями

Каким образом разделять области по вычислительным устройствам? На рис. 3 приведена постановка модельной задачи обтекания профиля с заданной линией разрыва потенциала.

Учитывая, что расчетные точки сгущаются в области, прилегающей к задней кромке и к точке торможения потоку, разделение областей можно выполнить так, как показано на рис. 4.

Реализация предложенной схемы возможна на базе вычислительной Грид-системы. Указанная Грид-система должна включать удаленные арендуемые кластеры (CP), сеть хранения данных SAN (Fibre Channel / 10 Gb Ethernet), сервер доступа и управления Грид-системой (IP/RP), а также вспомогательный локальный кластер для завершающих вычислений. Основную роль в такой системе играет сервер доступа и управления Грид-системой, отвечающий за постановку, декомпозицию,

диспетчеризацию и мониторинг задач. На текущем этапе исследований на основе разработанных схем разделения секрета ведётся разработка программного обеспечения и протоколов для управления сервером доступа, декомпозиции задач и взаимодействия элементов вычислительной системы.

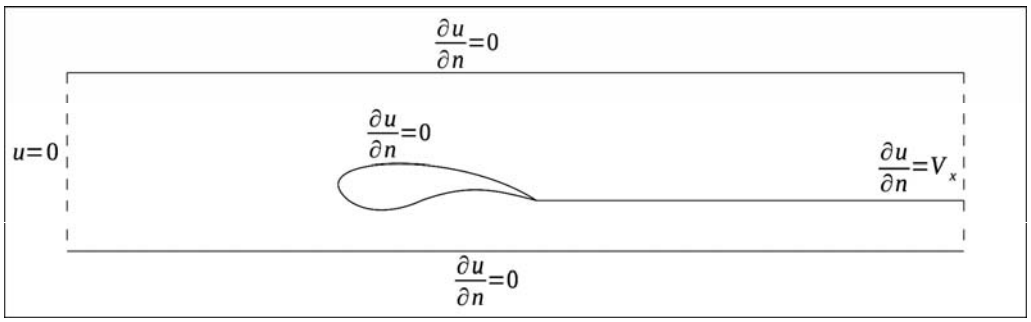

Рис. 3. Постановка задачи обтекания профиля в трубе

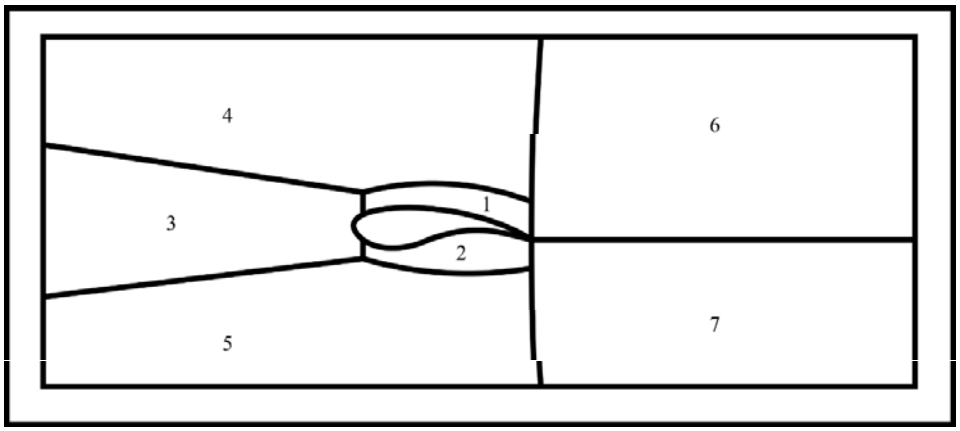

Рис. 4. Разделение областей ответственности между IP/RP (*1*, *2*) и CP (*3*–*7*)

На рис. 5 представлены результаты экспериментальных вычислений, отображающие эффективность вычислительных схем с пересылкой файлов границ. В качестве элементов вычислительной Грид-системы в ходе эксперимента были использованы не кластеры, а типовые бытовые вычислители. Видно, что с ростом размерности задачи эффективность растет и остается единственным способом решения задач большой размерности.

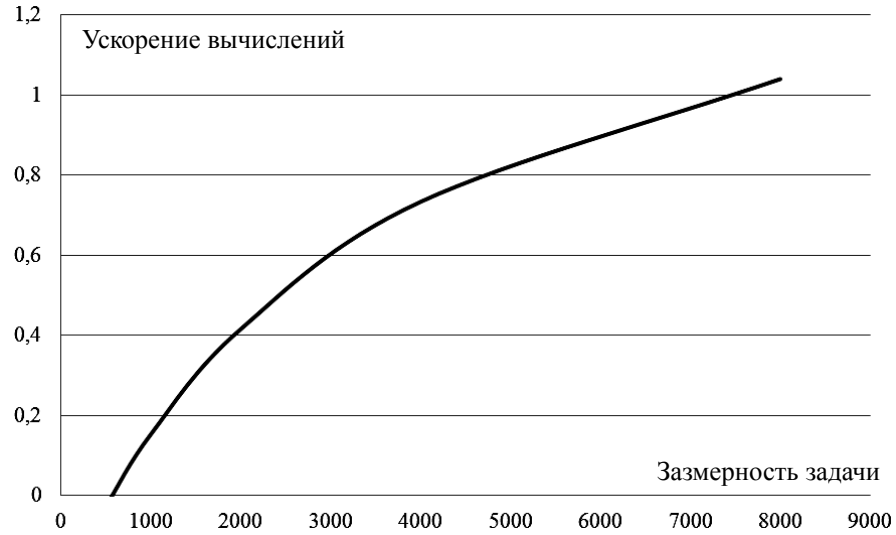

Рис. 5. Решение задачи Дирихле для уравнения Лапласа на расчетной сетке размера $N \times N$

Оценка трудоёмкости восстановления секрета. Естественным образом возникает вопрос о том, насколько надежно защищает предложенный способ разделения секрета от восстановления

секрета на основе знания данных, предоставляемых поставщикам облачных вычислений (СР). Представляется возможным использовать строго доказуемую неединственность решения, или единственность с точностью до широкого класса функций как инструмент в защите информации при решении краевых задач с привлечением неконтролируемых вычислительных ресурсов.

Указанный тезис можно пояснить на простейшем примере. Рассмотрим две односвязные области: $\Omega, \Xi \subset \Omega$ и задачу Дирихле для уравнения Лапласа:

$$\Delta u(x,y) = 0,\ (x,y) \in \Omega - \Xi, \quad u\big|_{\partial\Omega} = \varphi(x,y), \quad u\big|_{\partial\Xi} = \psi(x,y).$$

Решение этой задачи существует и единственно при достаточно слабых условиях на границы и рассматриваемые классы граничных функций. Предположим, что задача решается численно с помощью альтернирующего метода Шварца, где итерации последовательно проводятся на двух подобластях $\Theta_1, \Theta_2, \Theta_1 \cup \Theta_2 = \Omega - \Xi$, представляющих собой вложенные кольца. Кольцо Θ_1 примыкает к Ξ и контролируется вычислительным устройством *A*, а остальная часть Ω контролируется вычислительным устройством *B*. Возникает вопрос, можно ли обладая информацией о решении в области $\Omega - (\Theta_1 \cup \Xi)$, получить информацию о границе Ξ и значениях функции или ее производных на $\partial\Xi$? Если решение такой задачи будет неединственно или единственно с точностью до широкого класса функций, то подобную схему можно использовать для вычислений с арендой мощных вычислительных устройств и с гарантией сохранения секрета.

Рассмотрим три краевые задачи для уравнения Лапласа и покажем, насколько трудоемким является восстановление границы неизвестной области и значений функции на ней в том случае, когда известно решение на некоторой области, топологически эквивалентной кольцу. Тем обстоятельством, что функция сеточная и условной корректностью рассматриваемых задач будем пренебрегать. Покажем, что решение в каждом из этих случаев неединственно.

Рассмотрим задачу Коши для уравнения Лапласа:

$$\Delta u(x,y) = 0,\ (x,y) \in \Omega - \Xi,$$

где Ω – односвязная область, Ξ строго включена в Ω и ее граница не пересекается с границей Ω. Заданы

$$u\big|_{\partial\Omega} = \varphi(x,y), \quad u_n\big|_{\partial\Omega} = \psi(x,y).$$

Задача заключается в нахождении неизвестных функций $\zeta(s)$ и области Ξ:

$$u_{\partial\Xi} = \zeta(s), \Xi \subset \Omega.$$

Если выполнятся условия Неймана $\int\limits_{\partial\Omega} u_n = 0$, то область Ξ может быть пустым множеством, точнее, если сужение решения задачи Неймана на границу равно заданному значению: $v\big|_{\partial\Omega} = u_n\big|_{\partial\Omega} = \varphi$. Но оно может быть и не пустым, а любым, включенным в Ω. Таким образом, в этом, самом простом случае решение неединственно. Очевидно, что в случае выполнения условия Неймана, выполнения условий согласования между граничными значениями функции и ее нормальной производной область Ξ определяется неединственным образом. То есть вырезая из области Ω любое Ξ, не касающееся границы Ω, мы никак не влияем на решение вне Ξ и граничные условия.

Как поступать в более сложном случае, когда согласования между граничными условиями нет? Рассмотрим логарифмический потенциал по неизвестной площади и два нелинейных уравнения, выполняющихся на границе Ω, обозначим их *A*:

$$\sigma \iint\limits_{\Xi} \ln\left(\frac{1}{\sqrt{(x-x_s)^2 + (y-y_s)^2}}\right) ds = \varphi(x,y), \qquad \sigma \frac{\partial}{\partial n} \iint\limits_{\Xi} \ln\left(\frac{1}{\sqrt{(x-x_s)^2 + (y-y_s)^2}}\right) ds = \psi(x,y),$$

где σ – это неизвестная константа. Второе уравнение при выборе конкретного σ хорошо известно, это обратная задача гравиметрии определения формы области. Предположим, что решение второго уравнения единственно. Тогда, подставляя решение в первое уравнение, мы в общем случае не получим равенство. Надо будет подобрать такое σ и соответствующее ему решение, например, методом деления пополам, так, что оба уравнения обратятся в тождество. Таким образом, мы можем найти область и функцию $u_{\partial\Xi} = \zeta(s)$. Но обратим внимание на то, что, рассматривая разложения: $\varphi = \varphi_1 + \varphi_2, \psi = \psi_1 + \psi_2$, где слагаемые с индексом 1 отвечают согласованным краевым условиям задачи Неймана, мы можем получить решение в виде $u = u_1 + u_2$, где индекс *1* отвечает решению ранее

рассмотренной задачи Неймана, а индекс 2 отвечает решению задачи с помощью логарифмического потенциала. Рассматривая в качестве Ξ любую область Ξ_1, такую, что $\Xi \subset \Xi_1 \subset \Omega$, мы получим функцию, удовлетворяющую уравнению Лапласа в $\Omega - \Xi$ и заданным краевым условиям, что доказывает неединственность решения задачи.

Рассмотрим задачу продолжения функции, заданной на некоторой области K, гомеоморфной кольцу, через внутреннюю границу в область Ω, считая, что $\Delta u(x,y) = 0,\ (x,y) \in K,\ \Omega - \Xi$, где Ω – односвязная область, Ξ включена в Ω, и ее граница не пересекается с границей Ω. Значения функции в кольце известны: $u(x,y), (x,y) \in K$.

Задача заключается в нахождении неизвестной области Ξ и неизвестной нормальной производной на границе этой области:

$$u_n|_{\partial\Xi} = g(x,y), \Xi \subset \Omega \quad \text{с условием} \quad \int_{\partial\Xi} g(x,y)dl = 0.$$

Функция $V(x,y) = \frac{1}{\pi}\int_{\partial\Xi} g(x_s,y_s)\ln\left(\frac{1}{\sqrt{(x_s-x)^2+(y_s-y)^2}}\right)ds$ задает гармоническую функцию вне границы Ξ, удовлетворяющую условию Неймана на границе $V_n|_{\partial\Xi} = g(x,y)$.

Для любых двух замкнутых кривых в K, окаймляющих Ω, можно составить два интегральных уравнения первого рода, связывающих известные значения $u(x,y),\ (x,y) \in K$ и значения, индуцированные логарифмическим потенциалом простого слоя. Предположим, что решение этих уравнений существует и единственно вне зависимости от выбора кривых. Рассмотрим область $\Xi_1, \Xi \subset \Xi_1 \subset \Omega$. Функция V задает нормальную производную $V_n|_{\partial\Xi_1}$, с помощью которой можно также построить гармоническую функцию W вне и внутри Ξ_1. В самом деле, из-за отсутствия источников $\int_{\partial\Xi_1} V_n dl = 0$ и в качестве плотности можно взять $V_n|_{\partial\Xi_1}$. Внутри кольца $\Xi_1 - \Xi$ и вне его функция W совпадает с индуцированной ранее, так как их разность является гармонической функцией с нулевыми условиями Дирихле и Неймана на границе Ξ_1. В итоге мы получим неединственность Ξ при принятом допущении, что решение системы интегральных уравнений единственно задает Ξ.

Предположим, что на границе Ξ задано условие непротекания $u_n|_{\partial\Xi} = 0, \Xi \subset \Omega$. В этом случае построение потенциала V с помощью логарифмического потенциала простого слоя, как в предыдущем случае, невозможно. Как поступить в этом случае? Рассмотрим u решение смешанной задачи Неймана в кольце $\Omega - \Xi$

$$\Delta u(x,y) = 0,\ (x,y) \in \Omega - \Xi, \quad u|_{\partial\Omega} = \varphi(x,y),\ u_n|_{\partial\Xi} = 0$$

и покажем, что выбор Ξ не единствен. Функции u можно поставить в соответствие сопряженную гармоническую функцию v так, что $f = u + iv$ будет голоморфной. Если мы отступим от $\partial\Xi$ и выберем некоторую точку (x_0, y_0), лежащую во внутренности $\Omega - \Xi$, то через нее проходит Γ, линия уровня v, на которой также будет выполняться условие непротекания. Область, ограниченную Γ, можно рассматривать как Ξ_1, и u будет гармонической функцией в $\Omega - \Xi_1$.

Таким образом, практическая задача сводится только к методу подбора, который можно организовать, обладая некоторой априорной информацией и выбирая вид $\partial\Xi$ так, чтобы на внешней границе Ω происходила наиболее гладкая склейка с известным вне Ω решением и его производными. Оценим трудоемкость такого подхода. Если, например, сетка в подобласти $1 \cup 2$ представляет собой образ прямоугольной сетки размера $K \cdot N$, где K – число шагов сетки по нормали от Ξ, а N – число шагов сетки по поверхности, то число выборов «гладкой функции», описывающей $\partial\Xi$, будет равно $K \cdot 2^{N-1}$. Отсюда следует, что уже при $N \approx 80$ возникает комбинаторный взрыв и решение задачи продолжения методом подбора невозможно на современной и перспективной вычислительной технике.

Обратим также внимание на то, что неизвестными (секретом) являются не только Ξ и $\zeta(x,y)$ или $g(x,y)$, но и параметры сгущения или организации сетки внутри области Ω, что не только полностью исключает возможность решения задачи за приемлемое время, но и ставит вопрос об алгоритмической неразрешимости задачи *B*.

Заключение. Предложенные схемы организации защищенных параллельных вычислений (МРС), а также введенные простые правила позволяют решить вопрос о правомерности передачи массовых вычислений во внешние ЦОДы и об использовании неконтролируемых вычислительных ресурсов. Показано, что задача восстановления информации в этом случае сводится к задаче аналитического продолжения с неединственным решением. Реализация разработанных схем возможна на базе типовой Грид-системы и специализированного программного обеспечения.

Литература

1. BYTEmag. Центры обработки данных [Электронный ресурс]. – Режим доступа: http://www.bytemag.ru/articles/detail.php?ID=14070?ID=14070, свободный (дата обращения: 01.04.2014).

2. Об утверждении Стратегии развития отрасли информационных технологий в Российской Федерации на 2014–2020 годы и на перспективу до 2025 года [Электронный ресурс]. – Режим доступа: http://government.ru/docs/8024, свободный (дата обращения: 01.04.2014).

3. Файзуллин Р.Т. Приложение алгоритма префиксного кодирования массива данных в схеме разделения секрета / Р.Т. Файзуллин, Д.А. Сагайдак // Доклады ТУСУРа. – 2012. – № 1 (25), ч. 2. – С. 136–140.

4. Файзуллин Р.Т. Алгоритмы разделения секрета с использованием принципиально малой части в качестве ключа / Р.Т. Файзуллин, И.Р. Файзуллин, О.Т. Данилова // Вестник Тюм. гос. ун-та. – 2011. – Вып. 7. – С. 175–179.

5. Кручинин В.В. Подходы к созданию защищенного архива на основе разделения секрета / В.В. Кручинин, А.А. Шелупанов // Доклады ТУСУРа. – 2008. – № 2 (18), ч. 1. – С. 67–72.

6. Логинов К.В. Расчет, оптимизация и управление режимами работы больших гидравлических сетей / К.В. Логинов, А.М. Мызников, Р.Т. Файзуллин // Математическое моделирование. – 2006. – Вып. 18 (9). – С. 92–106.

7. Bodewig E. Matrix calculus. – Amsterdam: North-Holland, 1956. – 334 p.

8. Мещеряков Р.В. Критерий структурной сложности информационных систем // Труды СПИИРАН. – 2010. – № 3 (14). – С. 76–90.

Файзуллин Рашит Тагирович
Д-р техн. наук, профессор, зав. каф. комплексной защиты информации
Омского государственного технического университета (ОмГТУ)
Тел.: 8 (381-2) 21-77-02
Эл. почта: r.t.faizullin@mail.ru

Щерба Евгений Викторович
Канд. техн. наук, доцент каф. комплексной защиты информации ОмГТУ
Тел.: 8 (381-2) 21-77-02
Эл. почта: evscherba@gmail.com

Волков Данил Андреевич
Студент Омского государственного университета им. Ф.М. Достоевского
Эл. почта: volkovdanil91@gmail.com

Faizullin R.T., Shcherba E.V., Volkov D.A.

A scheme for the implementation of parallel computing as a protection mechanism of processed data

In this paper we propose a scheme of separation of computing and storage in data centers. The scheme guarantees the impossibility of restoring the matrix system of linear equations on individual compute nodes. The applicability of this approach in the case of numerical solution of some boundary value problems for equations of mathematical physics has been proved.

Keywords: distributed computing, secret sharing, DPC, Grid-system.