Доклады Национальной академии наук Беларуси

2015 **июль–август** **Том 59 № 4**

УДК 519.6

Н. А. ЛИХОДЕД, М. А. ПОЛЕЩУК

МЕТОД РАНЖИРОВАНИЯ ПАРАМЕТРОВ РАЗМЕРА БЛОКОВ ВЫЧИСЛЕНИЙ ПАРАЛЛЕЛЬНОГО АЛГОРИТМА

(Представлено членом-корреспондентом Л. А. Яновичем)

Белорусский государственный университет, Минск, Беларусь
likhoded@bsu.by; poleschuma@bsu.by

Исследуется задача получения макроопераций параллельного алгоритма, приводящих к меньшему числу обращений к глобальной памяти. Сформулированы и доказаны утверждения, позволяющие оценить объем коммуникационных операций, порождаемых разбиением множества итераций.

Ключевые слова: параллельные вычисления, распараллеливание алгоритмов, графический процессор, минимизации объема коммуникационных операций.

N. A. LIKHODED, M. A. PALIASHCHUK

METHOD OF RANKING TILES SIZE PARAMETERS OF A PARALLEL ALGORITHM

Belarusian State University, Minsk, Belarus
likhoded@bsu.by; poleschuma@bsu.by

A method for obtaining tiles of operations of a parallel algorithm is developed. Propositions for ranking tiles size parameters are stated and proved. Statements to assess the amount of communication operations generated by the partition of the set of iterations are stated and proved.

Keywords: parallel computing, parallelization of algorithms, GPU, minimization of communications.

Введение. Время решения задачи на современном компьютере во многом определяется степенью использования памяти с быстрым доступом. В качестве целевого компьютера, т. е. компьютера, на котором требуется реализовать параллельную версию алгоритма, будем рассматривать графические процессоры (GPU). При многопроцессорной обработке на GPU быстрым является процесс обращения к разделяемой памяти мультипроцессора, оперирующего некоторыми данными алгоритма, но не обращение к глобальной памяти GPU. Чем меньше число обращений к глобальной памяти, тем быстрее выполняется алгоритм.

Для реализации алгоритма на графическом процессоре множество операций алгоритма должно быть разбито на блоки, а блоки – на потоки (нити) вычислений. Множество операций алгоритма разбить на блоки можно путем тайлинга (тайлинг первого уровня), затем разбить блоки вычислений на потоки вычислений можно путем повторного применения тайлинга (тайлинг второго уровня) [1]. Тайлинг (tiling) – это преобразование алгоритма для получения макроопераций-тайлов [2; 3]; операции одного тайла выполняются атомарно, как одна единица вычислений, а обмен данными происходит массивами.

В данной работе предлагается метод, позволяющий получать тайлы вычислений с меньшим числом обращений к глобальной памяти. Используется анализ информационных зависимостей, порождающих коммуникационные операции. Сформулированы и доказаны утверждения, позволяющие ранжировать параметры размера тайлов. Предлагаемый метод можно также применять после некоторых предварительных оптимизирующих преобразований алгоритма. В частности, ранжировать параметры размера тайлов предлагаемым способом можно после преобразо-

ваний для улучшения параллелизма вычислений уровня тайлов и обеспечения одновременного начала шаблонных вычислений процессорами [4; 5]. Результаты исследований этой работы могут быть использованы для минимизации объема коммуникационных операций между параллельными вычислительными процессами, реализуемыми на суперкомпьютерах с распределенной памятью.

Предварительные сведения. Приведем необходимые для дальнейшего изложения сведения о формальном описании алгоритма и о тайлинге.

Будем считать, что алгоритм задан последовательной программой линейного класса [6]. Основную вычислительную часть такой программы составляют циклические конструкции; границы изменения параметров циклов задаются неоднородными формами, линейными по совокупности параметров циклов и внешних переменных. Предполагается, что в гнезде циклов имеется K выполняемых операторов S_β и используется L массивов a_l размерностей ν_l. Область изменения параметров циклов (область итераций) для оператора S_β и размерность этой области обозначим соответственно V_β и n_β.

Выполнение оператора S_β при конкретных значениях β вектора параметров цикла J будем называть операцией и обозначать $S_\beta(J)$. Зависимости (информационные связи) между операциями задаются функциями вида

$$\overline{\Phi}_{\alpha,\beta}(J) = \Phi_{\alpha,\beta}J + \Psi_{\alpha,\beta}N - \varphi^{\alpha,\beta}, \tag{1}$$

$$J \in V_{\alpha,\beta},\ N \in \mathbb{Z}^s,\ \Phi_{\alpha,\beta} \in \mathbb{Z}^{n_\alpha \times n_\beta},\ \Psi_{\alpha,\beta} \in \mathbb{Z}^{n_\alpha \times s},\ \varphi^{\alpha,\beta} \in \mathbb{Z}^{n_\alpha},$$

где $N \in \mathbb{Z}^s$ – вектор внешних переменных алгоритма; s – число внешних переменных. Функция зависимостей $\overline{\Phi}_{\alpha,\beta}(J)$ позволяет для операции $S_\beta(J)$ найти операцию $S_\alpha(I)$, от которой $S_\beta(J)$ зависит. Функции зависимостей являются удобным математическим аппаратом для описания информационных связей между операциями алгоритма (другие названия этих функций: покрывающие функции графа алгоритма [6], h-преобразования [7; 8]).

Вхождением (l, β, q) будем называть q-е вхождение массива a_l в оператор S_β. Индексы элементов l-го массива, связанных с вхождением (l, β, q), выражаются функцией $\overline{F}_{l,\beta,q}$ вида

$$\overline{F}_{l,\beta,q}(J) = F_{l,\beta,q}J + G_{l,\beta,q}N + f^{l,\beta,q},$$

$$J(j_1,...,j_{n_\beta}) \in V_\beta,\ N \in \mathbb{Z}^s,\ F_{l,\beta,q} \in \mathbb{Z}^{\nu_l \times n_\beta},\ G_{l,\beta,q} \in \mathbb{Z}^{\nu_l \times s},\ f^{l,\beta,q} \in \mathbb{Z}^{\nu_l}.$$

Пара вхождений $(l, \alpha, 1)$ и (l, β, q) порождает истинную зависимость $S_\alpha(I) \to S_\beta(J)$, если $S_\alpha(I)$ выполняется раньше $S_\beta(J)$; $S_\alpha(I)$ переопределяет (изменяет) элемент массива a_l, а $S_\beta(J)$ использует в качестве аргумента тот же элемент массива; между операциями $S_\alpha(I)$ и $S_\beta(J)$ этот элемент не переопределяется.

Пусть в гнезде циклов имеется Θ наборов выполняемых операторов. Под набором операторов будем понимать один или несколько операторов, окруженных одним и тем же множеством циклов. Операторы и наборы операторов линейно упорядочены расположением их в записи алгоритма. Обозначим: V^ϑ, $1 \le \vartheta \le \Theta$, – области изменения параметров циклов, окружающих наборы операторов, n^ϑ – размерность области V^ϑ, число циклов, окружающих ϑ-й набор операторов. Заметим, что если оператор S_β, принадлежит набору операторов с номером ϑ^β, то область V_β может быть уже области V^{ϑ^β}.

Как уже отмечалось, тайлинг – это преобразование алгоритма для получения макроопераций, называемых зерном вычислений, или тайлами. При тайлинге каждый цикл разбивается на два цикла: глобальный, параметр которого определяет на данном уровне вложенности порядок вычисления тайлов, и локальный, в котором параметр исходного цикла изменяется в границах одного тайла. Допускается вырожденное разбиение цикла, при котором все итерации относятся к глобальному циклу или все итерации относятся к локальному циклу.

Следующие величины и множества используются для формализации тайлинга.

$m_\zeta^\vartheta = \min\limits_{J(j_1,j_2,...,j_{n^\vartheta}) \in V^\vartheta} j_\zeta$, $M_\zeta^\vartheta = \max\limits_{J(j_1,j_2,...,j_{n^\vartheta}) \in V^\vartheta} j_\zeta$, $1 \le \zeta \le n^\vartheta$, – предельные значения изменения параметров циклов;

$r_1^{\vartheta},...,r_{n^{\vartheta}}^{\vartheta}$ – заданные натуральные числа, определяющие размеры тайла; r_ζ^{ϑ} обозначает число значений параметра j_ζ, приходящихся на один тайл ϑ-го набора операторов; r_ζ^{ϑ} может принимать фиксированное значение в пределах от 1 до $r_\zeta^{\vartheta,\max}$ включительно, где $r_\zeta^{\vartheta,\max} = M_\zeta^{\vartheta} - m_\zeta^{\vartheta} + 1$; если $r_\zeta^{\vartheta} = 1$, то цикл с параметром j_ζ является глобальным не разбиваемым; если $r_\zeta^{\vartheta} = r_\zeta^{\vartheta,\max}$, то цикл с параметром j_ζ является локальным не разбиваемым; если два набора операторов имеют общий цикл с параметром j_ζ, то $r_\zeta^{\vartheta 1} = r_\zeta^{\vartheta 2}$;

$Q_\zeta^{\vartheta} = \left\lceil (M_\zeta^{\vartheta} - m_\zeta^{\vartheta} + 1)/r_\zeta^{\vartheta} \right\rceil$, $1 \le \zeta \le n^{\vartheta}$, – число частей, на которые при формировании тайлов разбивается область значений параметра j_ζ цикла, окружающего ϑ-й набор операторов;

$V^{\vartheta,\mathrm{gl}} = \{J^{\mathrm{gl}}(j_1^{\mathrm{gl}},...,j_{n^{\vartheta}}^{\mathrm{gl}}) \mid 0 \le j_\xi^{\mathrm{gl}} \le Q_\xi^{\vartheta} - 1,\ 1 \le \zeta \le n^{\vartheta}\}$ – области изменения параметров глобальных, т. е. уровня тайлов, циклов;

$V^{\vartheta}_{J^{\mathrm{gl}}} = \{J(j_1,...,j_{n^{\vartheta}}) \in V^{\vartheta} \mid m_\zeta^{\vartheta} + j_\zeta^{\mathrm{gl}} r_\zeta^{\vartheta} \le j_\zeta \le m_\zeta^{\vartheta} - 1 + (j_\zeta^{\mathrm{gl}} + 1) r_\zeta^{\vartheta},\ \ 1 \le \zeta \le n^{\vartheta}\}$, $J^{\mathrm{gl}} \in V^{\vartheta,\mathrm{gl}}$, – области изменения параметров локальных (уровня операций тайлов) циклов при фиксированных значениях параметров глобальных циклов. Множество операций, выполняемых на итерациях множества $V^{\vartheta}_{J^{\mathrm{gl}}}$, будем также обозначать $V^{\vartheta}_{J^{\mathrm{gl}}}$. Множества $V^{\vartheta}_{J^{\mathrm{gl}}}$ называются тайлами.

Условия, характеризующие объем коммуникационных операций. Результаты, представленные в этом разделе, позволяют оценить объем коммуникационных операций между тайлами первого уровня (т. е. между блоками вычислений).

Пусть рассматривается гнездо циклов, для которого тайлинг допустим. Если заведомо известно, что $r_\zeta^{\vartheta} = 1$ (цикл с параметром j_ζ является глобальным не разбиваемым) или $r_\zeta^{\vartheta} = r_\zeta^{\vartheta,\max}$ (цикл с параметром j_ζ является локальным не разбиваемым), то разбиение итераций j_ζ не рассматривается. Обозначим через $(\Phi_{\alpha,\beta})_\zeta$ и $(\Psi_{\alpha,\beta})_\zeta$ строки матриц $\Phi_{\alpha,\beta}$ и $\Psi_{\alpha,\beta}$ с номером ζ, через $e_\zeta^{(n_\beta)}$ – вектор-строку размера n_β, у которой координата с номером ζ равна 1, а остальные координаты нулевые; обозначим еще $\eta_\zeta^{\alpha,\beta} = (\Psi_{\alpha,\beta})_\zeta N - \varphi_\zeta^{\alpha,\beta}$.

Л е м м а 1. *Пусть определение элемента некоторого массива a_l происходит на вхождении $(l, \alpha, 1)$ в левой части оператора S_α, а использование – на вхождении (l, β, q) в правой части оператора S_β, причем в окружении операторов S_α и S_β имеется цикл с параметром j_ζ, и выполняется условие*

$$(\Phi_{\alpha,\beta})_\zeta = e_\zeta^{(n_\beta)}. \tag{2}$$

Если

$$\eta_\zeta^{\alpha,\beta} = 0, \tag{3}$$

то определение и использование элемента массива данных происходит при одном и том же значении параметра цикла j_ξ^{gl}.

Если

$$0 < \left|\eta_\zeta^{\alpha,\beta}\right| < r_\zeta^{\vartheta}, \tag{4}$$

то определение и использование элемента массива данных происходит при одном и том же значении параметра цикла j_ξ^{gl} на $r_\zeta^{\vartheta} - \left|\eta_\zeta^{\alpha,\beta}\right|$ итерациях из каждых r_ζ^{ϑ} итераций j_ζ, а на $\left|\eta_\zeta^{\alpha,\beta}\right|$ итерациях из каждых r_ζ^{ϑ} итераций j_ζ определение и использование элемента массива данных происходит при разных значениях параметра цикла j_ξ^{gl}.

Если

$$\left|\eta_\zeta^{\alpha,\beta}\right| \ge r_\zeta^{\vartheta}, \tag{5}$$

то определение и использование элемента массива данных происходит при разных значениях параметра цикла j_ξ^{gl}.

Д о к а з а т е л ь с т в о. Докажем первое утверждение леммы. Функция зависимостей $I = \Phi_{\alpha,\beta}(J)$ определяет операцию $S_\alpha(I(i_1,...,i_{n_\alpha}))$ вычисления данного, требуемого операцией $S_\beta(J(j_1,...,j_{n_\beta}))$ в качестве аргумента. Из равенств (1)–(3) следует $i_\zeta = (\Phi_{\alpha,\beta})_\zeta J = e_\zeta^{(n_\beta)} J = j_\zeta$, поэтому значения параметров цикла j_ξ^{gl}, на которых определяется и используется элемент массива данных, совпадают.

Докажем второе утверждение леммы. Так как из определения тайлов имеем

$$j_\zeta = m_\zeta^\vartheta + j_\zeta^{gl} r_\zeta^\vartheta + \alpha_\zeta^\vartheta,$$

где $0 \le \alpha_\zeta^\vartheta < r_\zeta^\vartheta$,
то

$$i_\zeta = (\Phi_{\alpha,\beta})_\zeta J + \eta_\zeta^{\alpha,\beta} = e_\zeta^{(n_\beta)} J + \eta_\zeta^{\alpha,\beta} = m_\zeta^\vartheta + j_\zeta^{gl} r_\zeta^\vartheta + \alpha_\zeta^\vartheta + \eta_\zeta^{\alpha,\beta}.$$

Найдем значение параметра цикла i_ζ^{gl}, на котором определяется элемент массива данных ($\lfloor \cdot \rfloor$ обозначает ближайшее «снизу» целое число):

$$i_\zeta^{\mathrm{gl}} = \left\lfloor \frac{i_\zeta - m_\zeta^\vartheta}{r_\zeta^\vartheta} \right\rfloor = \left\lfloor \frac{j_\zeta^{\mathrm{gl}} r_\zeta^\vartheta + \alpha_\zeta^\vartheta + \eta_\zeta^{\alpha,\beta}}{r_\zeta^\vartheta} \right\rfloor = j_\zeta^{\mathrm{gl}} + \left\lfloor \frac{\alpha_\zeta^\vartheta + \eta_\zeta^{\alpha,\beta}}{r_\zeta^\vartheta} \right\rfloor. \tag{6}$$

Последний член равен нулю при $r_\zeta^\vartheta - \left|\eta_\zeta^{\alpha,\beta}\right|$ значениях α_ζ^ϑ, на остальных $\left|\eta_\zeta^{\alpha,\beta}\right|$ значениях α_ζ^ϑ последний член не равен нулю.

Для доказательства третьего утверждения теоремы достаточно заметить, что если выполняется условие (5), то последний член в равенстве (6) не является нулевым. □

В частном случае $\Psi = 0$ первые два утверждения теоремы 1 сформулированы в работе [9].

Л е м м а 2. *Пусть определение элемента некоторого массива a_l происходит при выполнении операции $S_\alpha(I(i_1,...,i_{n_\alpha}))$, а использование – при выполнении операции $S_\beta(J(j_1,...,j_{n_\beta}))$, причем в окружении операторов S_α и S_β имеется цикл с параметром j_ζ. Если выполняется условие*

$$(\Phi_{\alpha,\beta})_\zeta = \alpha_\zeta e_\zeta^{(n_\beta)}, \tag{7}$$

где $\alpha_\zeta \in \mathbb{Z}$, $\alpha_\zeta \ne 1$,
то для всех итераций j_ζ, удовлетворяющих условию

$$\left|(\alpha_\zeta - 1) j_\zeta + \eta_\zeta^{\alpha,\beta}\right| \ge r_\zeta^\vartheta, \tag{8}$$

определение и использование элемента массива данных происходит при разных значениях параметра цикла j_ζ^{gl}.

Д о к а з а т е л ь с т в о. Преобразуем выражение под знаком модуля в неравенстве (8):

$$(\alpha_\zeta - 1) j_\zeta + \eta_\zeta^{\alpha,\beta} = \alpha_\zeta j_\zeta + \eta_\zeta^{\alpha,\beta} - j_\zeta = \alpha_\zeta e_\zeta^{(n_\beta)} J + \eta_\zeta^{\alpha,\beta} - j_\zeta =$$

$$(\Phi_{\alpha,\beta})_\varsigma J + \eta_\zeta^{\alpha,\beta} - j_\zeta = (\Phi_{\alpha,\beta})_\varsigma J + (\Psi_{\alpha,\beta})_\zeta N - \varphi_\zeta^{\alpha,\beta} - j_\zeta = i_\zeta - j_\zeta.$$

Таким образом, если $\alpha_\zeta \ne 1$, то $\left|i_\zeta - j_\zeta\right| \ge r_\zeta^\vartheta$. Поэтому определение и использование элемента массива данных происходит при разных значениях параметра цикла j_ζ^{gl}. □

З а м е ч а н и е 1. Из леммы 2 следует, что при выполнении условия (7) число итераций j_ζ, для которых определение и использование элемента массива данных происходит при одном и том же значении цикла с параметром j_ζ^{gl}, не может быть большим (число r_ζ^ϑ, как правило, гораздо меньше количества всех итераций j_ζ). Основная причина этого заключается в том, что модуль разности $\left|j_\zeta - i_\zeta\right|$ не ограничен небольшим (в пределах нескольких единиц) числом. В случае выполнения условия

$$(\Phi_{\alpha,\beta})_\zeta \ne \alpha_\zeta e_\zeta^{(n_\beta)},$$

где $\alpha_\zeta \in \mathbb{Z}$,
модуль разности $\left|j_\zeta - i_\zeta\right|$ также не ограничен небольшим числом, так как зависит от J. Поэтому лишь для небольшого числа итераций j_ζ определение и использование элемента массива данных может происходить при одинаковых значениях цикла с параметром j_ζ^{gl}.

Пусть вхождение (l, β, q) в правую часть некоторого оператора порождает истинную зависимость, $\overline{\Phi}_{\alpha,\beta}$ – функция зависимостей. Обозначим

$$\rho_{l,\beta,q} = \operatorname{rank} F_{l,\beta,q},\ \rho_{l,\beta,q}^{\zeta} = \operatorname{rank}\begin{pmatrix} F_{l,\beta,q} \\ e_\zeta^{(n_\beta)} \end{pmatrix},\ \rho_{l,\beta,q}^{\Phi} = \operatorname{rank}\begin{pmatrix} F_{l,\beta,q} \\ \Phi_{\alpha,\beta} \end{pmatrix},\ \rho_{l,\beta,q}^{\Phi,\zeta} = \operatorname{rank}\begin{pmatrix} F_{l,\beta,q} \\ \Phi_{\alpha,\beta} \\ e_\zeta^{(n_\beta)} \end{pmatrix}.$$

Если вхождение (l, β, q) не порождает истинную зависимость, то по определению $\rho^{\Phi}_{l,\beta,q} = \rho_{l,\beta,q}$, $\rho^{\Phi,\zeta}_{l,\beta,q} = \rho^{\zeta}_{l,\beta,q}$. Отметим, что $\rho^{\Phi}_{l,\beta,q}$ и $\rho^{\Phi,\zeta}_{l,\beta,q} - 1$ отличаются не более чем на 1.

Обозначим $M^{\vartheta} = \max\limits_{\zeta}(M^{\vartheta}_{\zeta} - m^{\vartheta}_{\zeta}) + 1$ – наибольшее число итераций циклов, участвующих в получении тайлов. Для простоты записи будем использовать обозначение M без индекса ϑ, где будет подразумеваться набор операторов ϑ^{β} при упоминании оператора S_{β}. Отметим, что величина $Q^{\vartheta}_{\zeta} r^{\vartheta}_{\zeta}$ имеет порядок M. Определим термин «фиксированное данное массива» как конкретное, неизмененное содержимое соответствующей ячейки памяти. Следующая лемма позволяет оценить число используемых фиксированных данных на каждом вхождении (l, β, q) при фиксированных значениях циклов с параметрами j_{ζ} и j^{gl}_{ξ}, общее число используемых фиксированных данных на вхождении (l, β, q).

Л е м м а 3. *Число фиксированных данных, используемых на вхождении* (l, β, q) *в правой части оператора* S_{β}, *при фиксированном значении цикла с параметром* j_{ζ} *оценивается величиной* $O(M^{\rho^{\Phi,\zeta}_{l,\beta,q}-1})$.

Число фиксированных данных, используемых на вхождении (l, β, q), *оценивается величиной* $O(M^{\rho^{\Phi}_{l,\beta,q}})$.

Число фиксированных данных, используемых на вхождении (l, β, q), *при фиксированном значении глобального цикла с параметром* j^{gl}_{ξ} *оценивается величиной* $O(M^{\rho^{\Phi}_{l,\beta,q}})$, *если* $\rho^{\Phi,\zeta}_{l,\beta,q} - 1 = \rho^{\Phi}_{l,\beta,q}$, *и величиной* $O(r^{\vartheta}_{\zeta} M^{\rho^{\Phi}_{l,\beta,q}-1})$, *если* $\rho^{\Phi,\zeta}_{l,\beta,q} = \rho^{\Phi}_{l,\beta,q}$.

Д о к а з а т е л ь с т в о. На вхождении (l, β, q) фиксированное данное используется при фиксированном значении цикла с параметром j_{ζ} на итерациях подпространства итераций размерности $n_{\beta} - \rho^{\Phi,\zeta}_{l,\beta,q}$ [10; 11]. Тогда число фиксированных данных, используемых на вхождении (l, β, q) при фиксированном значении цикла с параметром j_{ζ}, т. е. число подпространств итераций размерности $n_{\beta} - \rho^{\Phi,\zeta}_{l,\beta,q}$ в пространстве размерности $n_{\beta} - 1$ есть $O(M^{n_{\beta}-1-(n_{\beta}-\rho^{\Phi,\zeta}_{l,\beta,q})}) = O(M^{\rho^{\Phi,\zeta}_{l,\beta,q}-1})$. Аналогично, число фиксированных данных, используемых на вхождении (l, β, q), есть число подпространств итераций размерности $n_{\beta} - \rho^{\Phi}_{l,\beta,q}$ в пространстве размерности n_{β} и для них справедлива асимптотическая оценка $O(M^{n_{\beta}-(n_{\beta}-\rho^{\Phi}_{l,\beta,q})}) = O(M^{\rho^{\Phi}_{l,\beta,q}})$.

Если выполняется условие $\rho^{\Phi,\zeta}_{l,\beta,q} - 1 = \rho^{\Phi}_{l,\beta,q}$, то оценка числа используемых на вхождении (l, β, q) фиксированных данных одинакова как для одного j_{ζ}, так и для любого числа итераций j_{ζ}. Поэтому число фиксированных данных при r^{ϑ}_{ζ} значениях цикла с параметром j_{ζ} оценивается величиной $O(M^{\rho^{\Phi,\zeta}_{l,\beta,q}-1}) = O(M^{\rho^{\Phi}_{l,\beta,q}})$. Если выполняется условие $\rho^{\Phi,\zeta}_{l,\beta,q} = \rho^{\Phi}_{l,\beta,q}$, то для каждого j_{ζ} (всего r^{ϑ}_{ζ} значений j_{ζ}) используются новые фиксированные данные. Так как $Q^{\vartheta}_{\zeta} r^{\vartheta}_{\zeta} = O(M)$, то r^{ϑ}_{ζ} может быть величиной порядка M и ее следует учесть в асимптотической оценке $O(r^{\vartheta}_{\zeta} M^{\rho^{\Phi,\zeta}_{l,\beta,q}-1})$ числа фиксированных данных. □

Оценка объема коммуникационных операций. В этом разделе для каждого вхождения (l, β, q) в правую часть оператора S_{β} оценивается объем коммуникационных операций чтения и соответствующих им операций записи, порождаемых разбиением итераций j_{ζ} при получении блоков вычислений.

Сделаем предположения, необходимые для практического использования результатов исследований: элементы матриц $\Phi_{\alpha,\beta}$, $\Psi_{\alpha,\beta}$ и векторов $\varphi^{\alpha,\beta}$ по модулю не превосходят нескольких единиц; $\left|\eta^{\alpha,\beta}_{\zeta}\right| = \left|\varphi^{\alpha,\beta}_{\zeta}\right| < r^{\vartheta}_{\zeta}$, если $(\Psi_{\alpha,\beta})_{\zeta} = 0$; $\left|\eta^{\alpha,\beta}_{\zeta}\right| \geq r^{\vartheta}_{\zeta}$, если $(\Psi_{\alpha,\beta})_{\zeta} \neq 0$.

Т е о р е м а. *Пусть вхождение* (l, β, q) *порождает истинную зависимость: определение некоторого данного происходит на вхождении* $(l, \alpha, 1)$ *в левой части оператора* S_{α}, *а использование – на вхождении* (l, β, q) *в правой части оператора* S_{β}, *причем в окружении операторов* S_{α} *и* S_{β} *имеется цикл с параметром* j_{ζ}.

Тогда при получении блоков вычислений для реализации алгоритма на графическом процессоре, разбиение итераций цикла j_{ζ} *порождает коммуникационные операции чтения и записи, объем которых имеет следующие оценки:*

1) если выполняются условия (2), (3), то не требуется ни операций чтения, ни операций записи;

2) если выполняются условия (2), (4), то требуется $O(Q_\zeta^\vartheta M^{\rho^{\Phi}_{l,\beta,q}-1})$ *операций чтения и операций записи;*

3) если условие (2) не выполняется, выполняется условие $\rho^{\Phi,\zeta}_{l,\beta,q}=\rho^{\Phi}_{l,\beta,q}$, *то требуется* $O(M^{\rho^{\Phi}_{l,\beta,q}})$ *операций чтения и операций записи;*

4) если условие (2) не выполняется, выполняется условие $\rho^{\Phi,\zeta}_{l,\beta,q}-1=\rho^{\Phi}_{l,\beta,q}$, *то требуется* $O(Q_\zeta^\vartheta M^{\rho^{\Phi}_{l,\beta,q}})$ *операций чтения и* $O(M^{\rho^{\Phi}_{l,\beta,q}})$ *операций записи.*

В случае, когда вхождение (l,β,q) *не порождает истинной зависимости (происходит обращение к входным данным) или цикл с параметром* j_ζ *имеется в окружении только оператора* S_β, *оценки следующие:*

5) если выполняется условие $\rho^{\Phi,\zeta}_{l,\beta,q}=\rho^{\Phi}_{l,\beta,q}$, *то требуется* $O(M^{\rho^{\Phi}_{l,\beta,q}})$ *операций чтения;*

6) если выполняется условие $\rho^{\Phi,\zeta}_{l,\beta,q}-1=\rho^{\Phi}_{l,\beta,q}$, *то требуется* $O(Q_\zeta^\vartheta M^{\rho^{\Phi}_{l,\beta,q}})$ *операций чтения.*

Д о к а з а т е л ь с т в о. Утверждения теоремы оценивают объем коммуникационных операций, порождаемых разбиением итераций цикла j_ζ, в зависимости от возможных значений координат строки $(\Phi_{\alpha,\beta})_\zeta$ и величины $\eta_\zeta^{\alpha,\beta}$ (шесть случаев, возможных на практике). Напомним, множество итераций цикла с параметром j_ζ разбивается на Q_ζ^ϑ частей.

Рассмотрим утверждение первого случая теоремы. Пусть вхождение (l,β,q) порождает истинную зависимость и выполняются условия (2), (3). Тогда выполняется равенство $\rho^{\Phi,\zeta}_{l,\beta,q}=\rho^{\Phi}_{l,\beta,q}$; из леммы 1 и леммы 3 следует, что в каждой из Q_ζ^ϑ частей все $O(r_\zeta^\vartheta M^{\rho^{\Phi}_{l,\beta,q}-1})$ фиксированных данных определяются и используются при одном и том же значении параметра цикла j_ζ^{gl}. Коммуникационных операций не требуется.

Во втором случае теоремы вхождение (l,β,q) порождает истинную зависимость, выполняются условия (2), (4). Тогда выполняется равенство $\rho^{\Phi,\zeta}_{l,\beta,q}=\rho^{\Phi}_{l,\beta,q}$; из леммы 1 и леммы 3 следует, что в каждой из Q_ζ^ϑ частей $\left|\eta_\zeta^{\alpha,\beta}\right|O(M^{\rho^{\Phi,\zeta}_{l,\beta,q}-1})$ данных, из общего числа $O(r_\zeta^\vartheta M^{\rho^{\Phi}_{l,\beta,q}-1})$ используемых в каждой части фиксированных данных, определяются и используются при разных итерациях цикла j_ζ^{gl}, причем независимо от значения r_ζ^ϑ величина $\left|\eta_\zeta^{\alpha,\beta}\right|$ не превышает нескольких единиц. Суммарный объем коммуникационных операций чтения и операций записи по всем Q_ζ^ϑ частям определяется оценкой $Q_\zeta^\vartheta O(M^{\rho^{\Phi,\zeta}_{l,\beta,q}-1})=O(Q_\zeta^\vartheta M^{\rho^{\Phi}_{l,\beta,q}-1})$.

Рассмотрим третий и четвертый случаи теоремы. В третьем случае выполняется условие $\rho^{\Phi,\zeta}_{l,\beta,q}=\rho^{\Phi}_{l,\beta,q}$, в четвертом – условие $\rho^{\Phi,\zeta}_{l,\beta,q}-1=\rho^{\Phi}_{l,\beta,q}$. Согласно лемме 3, число используемых фиксированных данных в каждой из Q_ζ^ϑ частей оценивается в третьем случае величиной $O(r_\zeta^\vartheta M^{\rho^{\Phi}_{l,\beta,q}-1})$, в четвертом – величиной $O(M^{\rho^{\Phi}_{l,\beta,q}})$. По условию, вхождение (l,β,q) порождает истинную зависимость, но условие (2) не выполняется. Поэтому в третьем случае нельзя утверждать, что в каждой из Q_ζ^ϑ частей только $O(M^{\rho^{\Phi,\zeta}_{l,\beta,q}-1})$ данных из общего числа $O(r_\zeta^\vartheta M^{\rho^{\Phi}_{l,\beta,q}-1})$ используемых (в каждой части) фиксированных данных определяются и используются при разных итерациях цикла j_ζ^{gl} (см. замечание 1); объемы коммуникационных операций чтения и операций записи по всем Q_ζ^ϑ частям следует оценить величиной $Q_\zeta^\vartheta O(r_\zeta^\vartheta M^{\rho^{\Phi}_{l,\beta,q}-1})=O(M^{\rho^{\Phi}_{l,\beta,q}})$. Аналогично, в четвертом случае объем коммуникационных операций чтения по всем Q_ζ^ϑ частям равен $Q_\zeta^\vartheta O(M^{\rho^{\Phi}_{l,\beta,q}})=O(Q_\zeta^\vartheta M^{\rho^{\Phi}_{l,\beta,q}})$. Объем коммуникационных операций записи в четвертом случае оценивается величиной $O(M^{\rho^{\Phi}_{l,\beta,q}})$ – числом фиксированных данных, используемых, согласно лемме 3, на вхождении (l,β,q) как на одной части (выполняется условие $\rho^{\Phi,\zeta}_{l,\beta,q}-1=\rho^{\Phi}_{l,\beta,q}$), так и на всех Q_ζ^ϑ частях.

В пятом и шестом случаях теоремы вхождение (l,β,q) не порождает истинной зависимости (происходит обращение к входным данным) или порождает истинную зависимость $S_\alpha(I)\to S_\beta(J)$, но цикл с параметром j_ζ имеется в окружении только оператора S_β. В пятом случае выполняется условие $\rho^{\Phi,\zeta}_{l,\beta,q}=\rho^{\Phi}_{l,\beta,q}$, в шестом – условие $\rho^{\Phi,\zeta}_{l,\beta,q}-1=\rho^{\Phi}_{l,\beta,q}$. Число используемых фиксированных

данных в каждой из Q_ξ^ϑ частей оценивается, согласно лемме 3, в пятом случае величиной $O(r_\xi^\vartheta M^{\rho^{\Phi}_{l,\beta,q}-1})$, в шестом – величиной $O(M^{\rho^{\Phi}_{l,\beta,q}})$. Тогда объем коммуникационных операций (только чтение) по всем Q_ξ^ϑ частям в пятом случае равен $Q_\xi^\vartheta O(r_\xi^\vartheta M^{\rho^{\Phi}_{l,\beta,q}-1}) = O(M^{\rho^{\Phi}_{l,\beta,q}})$, в шестом – $Q_\xi^\vartheta O(M^{\rho^{\Phi}_{l,\beta,q}}) = O(Q_\xi^\vartheta M^{\rho^{\Phi}_{l,\beta,q}})$. □

З а м е ч а н и е 2. Если вхождение (l,β,q) не порождает истинной зависимости, то ранее введены обозначения $\rho^{\Phi}_{l,\beta,q} = \rho_{l,\beta,q}$ и $\rho^{\Phi,\zeta}_{l,\beta,q} = \rho^{\zeta}_{l,\beta,q}$. Поэтому предположения пунктов 5 и 6 теоремы можно записать в виде $\rho^{\zeta}_{l,\beta,q} = \rho_{l,\beta,q}$ и $\rho^{\zeta}_{l,\beta,q} - 1 = \rho_{l,\beta,q}$, а объем операций чтения в виде $O(M^{\rho_{l,\beta,q}})$ и $O(Q_\xi^\vartheta M^{\rho_{l,\beta,q}})$ соответственно.

Приоритеты разбиения итераций циклов. Утверждения теоремы позволяют выяснить асимптотику суммарного объема коммуникационных операций, порождаемых разбиением множества итераций j_ζ на Q_ξ^ϑ частей, в зависимости от возможных значений координат строки $(\Phi_{\alpha,\beta})_\zeta$ и величины $\eta_\zeta^{\alpha,\beta}$. Полученные оценки позволяют ранжировать параметры размера тайлов для минимизации объема коммуникационных операций между блоками вычислений.

Для каждого вхождения (l,β,q) в правую часть оператора S_β, окруженного циклом с параметром j_ζ, утверждения теоремы определяют объемы коммуникационных операций чтения $\omega^{\zeta,R}_{l,\beta,q}$ и операций записи $\omega^{\zeta,W}_{l,\beta,q}$. Их величины могут принимать одно из значений: нуль (первый случай теоремы); $O(Q_\xi^\vartheta M^\tau)$, где $\tau \in Z, \tau \geq 0$, (второй, четвертый и шестой случаи теоремы); $O(M^\tau)$ (третий, четвертый и пятый случаи теоремы). Для координаты j_ζ обозначим через ω_ξ^R и ω_ξ^W сумму объемов коммуникационных операций чтения и операций записи по всем вхождениям (l,β,q): $\omega_\xi^R = \sum_{(l,\beta,q)} \omega^{\zeta,R}_{l,\beta,q}$, $\omega_\xi^W = \sum_{(l,\beta,q)} \omega^{\zeta,W}_{l,\beta,q}$. Определим также суммарный объем ω_ζ всех (чтение и запись) коммуникационных операций для координаты j_ζ : $\omega_\zeta = \omega_\xi^R + \omega_\xi^W$. Заметим, что каждая из величин ω_ξ^R, ω_ξ^W и ω_ζ равна нулю или оценивается величинами $O(Q_\xi^\vartheta M^\tau)$, $O(M^\tau)$ (вид оценки зависит от наличия членов с множителем Q_ξ^ϑ).

Определим через z_ξ^R, z_ξ^W и z_ζ приоритеты разбиения цикла с параметром j_ζ по чтению, по записи и суммарно по чтению и записи. Положим z_ξ^R равным –1, если $\omega_\xi^R = 0$; равным τ, если $\omega_\xi^R = O(M^\tau)$; равным $\tau + 0{,}5$, если $\omega_\xi^R = O(Q_\xi^\vartheta M^\tau)$; z_ξ^W и z_ζ определяются аналогично через значения ω_ξ^W и ω_ζ. Число 0,5 отражает зависимость Q_ξ^ϑ и M, что следует из равенства $M = O(r_\zeta^\vartheta Q_\xi^\vartheta)$. Приоритет –1 является самым высоким (разбиение координаты j_ζ не приводит к коммуникационным операциям), с ростом значения величин z_ξ^R, z_ξ^W и z_ζ приоритет убывает. Отметим, что если в реальных вычислениях на графическом процессоре скорости доступа к глобальной памяти по чтению и по записи (или их средние оценки) известны и они различаются на один или более порядок относительно M, следует вместо суммарного по чтению и записи приоритета z_ζ использовать приоритеты по чтению z_ξ^R и по записи z_ξ^W с учетом такого различия.

Приоритет (и по чтению, и по записи) разбиения цикла с параметром j_ζ определяет возможность уменьшения r_ζ^ϑ без увеличения оценки коммуникационных издержек. Уменьшать r_ζ^ϑ требуется, если блоки вычислений являются слишком большими для эффективной реализации на мультипроцессорах. Целая величина приоритета предоставляет такую возможность, а нецелая – не предоставляет. Нецелый приоритет (вида $\tau + 0{,}5$, где $\tau \in Z, \tau \geq 0$) отражает наличие множителя Q_ξ^ϑ в оценке объема коммуникаций. Уменьшение r_ζ^ϑ увеличит Q_ξ^ϑ (в силу $M = O(r_\zeta^\vartheta Q_\xi^\vartheta)$) и, следовательно, оценку числа коммуникационных операций. Реальное число коммуникаций также может увеличиться. Например, нарушение условия $\left|\eta_\zeta^{\alpha,\beta}\right| < r_\zeta^\vartheta$ по причине малости r_ζ^ϑ во втором случае теоремы приведет к коммуникациям на каждой итерации (вместо нескольких единиц) из r_ζ^ϑ итераций j_ζ (условие (5) леммы 1) в каждой из Q_ξ^ϑ частей. В случае целого приоритета Q_ξ^ϑ не входит в оценку объема коммуникационных операций и, следовательно, есть возможность уменьшения r_ζ^ϑ без роста оценки. При ранжировании (установлении соотношений размеров относительно друг друга) параметров размера блоков вычислений параллельного алгоритма, реализуемого на графическом процессоре, следует учитывать в оценках объема коммуникационных операций только слагаемые с множителем Q_ξ^ϑ.

Пример. Рассмотрим алгоритм прямого хода метода Гаусса решения систем линейных алгебраических уравнений:

do $k = 1, n - 1$
  do $i = k + 1, n$
    do $j = k + 1, n + 1$
      $a(i,j) = a(i,j) - \dfrac{a(i,k)}{a(k,k)}\, a(k,j)$
    enddo
  enddo
enddo

Матрицы $\Phi_{\alpha,\beta}$ и векторы $\varphi^{\alpha,\beta}$ в функциях зависимостей будем для наглядности помечать элементами массивов, фигурирующими на порождающих зависимости вхождениях. Например, матрицу $\Phi_{\alpha,\beta}$ и вектор $\varphi^{\alpha,\beta}$ для вхождения $a(i, k)$ обозначим $\Phi_{a(i,j),a(i,k)}$ и $\varphi^{a(i,j),a(i,k)}$.

Для второго вхождения массива a в оператор ($a(i, j)$ – использование прежнего значения обновляемого элемента) имеем

$$F_{1,1,2} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ \Phi_{a(i,j),a(i,j)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ \varphi^{a(i,j),a(i,j)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},\ \rho^{\Phi}_{1,1,2} = \operatorname{rank}\begin{pmatrix} F_{1,1,2} \\ \Phi_{a(i,j),a(i,j)} \end{pmatrix} = 3.$$

Объемы коммуникационных операций $\omega^{\zeta,R}_{1,1,2}$ и $\omega^{\zeta,W}_{1,1,2}$, $\zeta = 1, 2, 3$, оцениваются величинами $\omega^{1,R}_{1,1,2} = \omega^{1,W}_{1,1,2} = O(Q_1 M^{\rho^{\Phi}_{1,1,2}-1}) = O(Q_1 M^2)$ (второй случай теоремы), $\omega^{2,R}_{1,1,2} = \omega^{2,W}_{1,1,2} = 0$ (первый случай), $\omega^{3,R}_{1,1,2} = \omega^{3,W}_{1,1,2} = 0$ (первый случай).

Для вхождения $a(i, k)$ (использование столбца, содержащего ведущий элемент) получим

$$F_{1,1,3} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},\ \Phi_{a(i,j),a(i,k)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},\ \varphi^{a(i,j),a(i,k)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},\ \rho^{\Phi}_{1,1,3} = \operatorname{rank}\begin{pmatrix} F_{1,1,3} \\ \Phi_{a(i,j),a(i,k)} \end{pmatrix} = 2,$$

$\omega^{1,R}_{1,1,3} = \omega^{1,W}_{1,1,3} = O(Q_1 M^{\rho^{\Phi}_{1,1,3}-1}) = O(Q_1 M)$ (второй случай), $\omega^{2,R}_{1,1,3} = \omega^{2,W}_{1,1,3} = 0$ (первый случай), $\omega^{3,R}_{1,1,3} = O(Q_3 M^{\rho^{\Phi}_{1,1,3}}) = O(Q_3 M^2)$ и $\omega^{3,W}_{1,1,3} = O(M^{\rho^{\Phi}_{1,1,3}}) = O(M^2)$ (четвертый случай).

Для вхождения $a(k, j)$ (использование строки, содержащей ведущий элемент)

$$F_{1,1,4} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ \Phi_{a(i,j),a(k,j)} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ \varphi^{a(i,j),a(k,j)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},\ \rho^{\Phi}_{1,1,2} = \operatorname{rank}\begin{pmatrix} F_{1,1,4} \\ \Phi_{a(i,j),a(k,j)} \end{pmatrix} = 2,$$

$\omega^{1,R}_{1,1,4} = \omega^{1,W}_{1,1,4} = O(Q_1 M^{\rho^{\Phi}_{1,1,4}-1}) = O(Q_1 M)$ (второй случай), $\omega^{2,R}_{1,1,4} = O(Q_2 M^{\rho^{\Phi}_{1,1,4}}) = O(Q_2 M^2)$ и $\omega^{2,W}_{1,1,4} = O(M^{\rho^{\Phi}_{1,1,4}}) = O(M^2)$ (четвертый случай), $\omega^{3,R}_{1,1,4} = \omega^{3,W}_{1,1,4} = 0$ (первый случай).

Для вхождения $a(k, k)$ (использование ведущего элемента)

$$F_{1,1,5} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},\ \Phi_{a(i,j),a(k,k)} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},\ \varphi^{a(i,j),a(k,k)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},\ \rho^{\Phi}_{1,1,5} = \operatorname{rank}\begin{pmatrix} F_{1,1,5} \\ \Phi_{a(i,j),a(k,k)} \end{pmatrix} = 1,$$

$\omega^{1,R}_{1,1,5} = \omega^{1,W}_{1,1,5} = O(Q_1 M^{\rho^{\Phi}_{1,1,5}-1}) = O(Q_1)$ (второй случай), $\omega^{2,R}_{1,1,5} = O(Q_2 M^{\rho^{\Phi}_{1,1,5}}) = O(Q_2 M)$, $\omega^{2,W}_{1,1,5} = O(M^{\rho^{\Phi}_{1,1,5}}) = O(M)$, $\omega^{3,R}_{1,1,5} = O(Q_3 M^{\rho^{\Phi}_{1,1,5}}) = O(Q_3 M)$ и $\omega^{3,W}_{1,1,5} = O(M^{\rho^{\Phi}_{1,1,5}}) = O(M)$ (четвертый случай).

Объемы операций чтения, операций записи и суммарный объем коммуникационных операций для первой, второй и третьей координат оцениваются величинами $\omega^R_1 = O(Q_1 M^2)$, $\omega^W_1 = O(Q_1 M^2)$, $\omega^R_2 = O(Q_2 M^2)$, $\omega^W_2 = O(M^2)$, $\omega^R_3 = O(Q_3 M^2)$, $\omega^W_3 = O(M^2)$, $\omega_1 = O(Q_1 M^2)$, $\omega_2 = O(Q_2 M^2)$, $\omega_3 = O(Q_3 M^2)$. Первая, вторая и третья координаты имеют приоритеты для разбиения по чтению $z^R_1 = 2{,}5$, $z^R_2 = 2{,}5$, $z^R_3 = 2{,}5$, по записи $z^W_1 = 2{,}5$, $z^W_2 = 2$, $z^W_3 = 2$, по чтению и записи $z_1 = 2{,}5$, $z_2 = 2{,}5$, $z_3 = 2{,}5$.

Для ранжирования параметров r_1, r_2 и r_3 размера блоков вычислений графического процессора найдем объемы (для различия с вычисленными выше объемами будем помечать их индексом rank) операций чтения, операций записи и суммарный объем коммуникационных операций для первой, второй и третьей координат, учитывая только слагаемые с множителем Q_ζ: $\omega_1^{R,\mathrm{rank}} = O(Q_1 M^2)$, $\omega_1^{W,\mathrm{rank}} = O(Q_1 M^2)$, $\omega_2^{R,\mathrm{rank}} = O(Q_2 M^2)$, $\omega_2^{W,\mathrm{rank}} = 0$, $\omega_3^{R,\mathrm{rank}} = O(Q_3 M^2)$, $\omega_3^{W,\mathrm{rank}} = 0$, $\omega_1^{\mathrm{rank}} = O(Q_1 M^2)$, $\omega_2^{\mathrm{rank}} = O(Q_2 M^2)$, $\omega_3^{\mathrm{rank}} = O(Q_3 M^2)$. Первая, вторая и третья координаты имеют приоритеты для разбиения по чтению $z_1^{R,\mathrm{rank}} = 2{,}5$, $z_2^{R,\mathrm{rank}} = 2{,}5$, $z_3^{R,\mathrm{rank}} = 2{,}5$, по записи $z_1^{W,\mathrm{rank}} = 2{,}5$, $z_2^{W,\mathrm{rank}} = -1$, $z_3^{W,\mathrm{rank}} = -1$, по чтению и записи $z_1^{\mathrm{rank}} = 2{,}5$, $z_2^{\mathrm{rank}} = 2{,}5$, $z_3^{\mathrm{rank}} = 2{,}5$. При этом отсутствует возможность уменьшения r_1 без роста оценки коммуникационных издержек и по чтению, и по записи (нецелый приоритет $z_1^{R,\mathrm{rank}} = z_1^{W,\mathrm{rank}} = 2{,}5$); имеется возможность уменьшения r_2 и r_3 без роста оценки коммуникационных операций записи (целый приоритет $z_2^{W,\mathrm{rank}} = z_3^{W,\mathrm{rank}} = -1$). Поэтому уменьшать блоки вычислений следует в первую очередь за счет параметров размера r_2 и r_3.

Таким образом, в работе сформулированы и доказаны утверждения, позволяющие оценить объем коммуникационных операций, порождаемых разбиением множества итераций. Определены приоритеты параметров циклов, позволяющие ранжировать параметры размера тайлов первого уровня для минимизации объема коммуникационных операций. Приведено обоснование старшинства приоритетов на основании числа обращений к глобальной памяти графического процессора.

Список использованной литературы

1. *Baskaran, M.* Automatic C-to-CUDA code generation for affine programs / M. Baskaran, J. Ramanujam, P. Sadayappan // Proceedings of the Compiler Construction, 19th International Conference. Part of the Joint European Conferences on Theory and Practice of Software. – Paphos, Cyprus, March 2010.

2. *Xue, J.* Time-minimal tiling when rise is larger than zero / J. Xue, W. Cai // Parallel Computing. – 2002. – Vol. 28, N 5. – P. 915–939.

3. *Kim, D. G.* Parameterized tiling for imperfectly nested loops / D. G. Kim, S. Rajopadhye // Technical Report CS-09-101, Colorado State University, Department of Computer Science, February 2009. – 21 p.

4. Automatic parallelization of tiled loop nests with enhanced fine-grained parallelism on GPUs / P. Di [et al.] // 41st International Conference on Parallel Processing. – Pittsburgh, PA, USA, September 2012. IEEE Computer Society, 2012. – P. 350–359.

5. *Bandishti, V.* Tiling stencil computations to maximize parallelism / V. Bandishti, I. Pananilath, U. Bondhugula // Proceedings of Supercomputing. – Los Alamitos, CA, USA. IEEE Computer Society Press, 2012. – P. 40:1–40:11.

6. *Воеводин, В. В.* Параллельные вычисления / В. В. Воеводин, Вл. В. Воеводин. – СПб.: БХВ-Петербург, 2002. – 608 с.

7. *Feautrier, P.* Some efficient solutions to the affine scheduling problem. Part 1 / P. Feautrier // International J. of Parallel Programming. – 1992. – Vol. 21, N 5. – P. 313–348.

8. Automatic transformations for communication-minimized parallelization and locality optimization in the polyhedral model / U. Bondhugula [et al.] // Lecture notes in computer science. – 2008. – N 4959. – P. 132–146.

9. *Лиходед, Н. А.* Оценка объема коммуникационных операций параллельного зернистого алгоритма / Н. А. Лиходед, М. А. Полещук // Междунар. конгресс по информатике: информационные системы и технологии CSIST'2013, 4–7 ноября 2013 г., Минск, Беларусь. Бел. гос. ун-т. – Минск, 2013. – С. 377–381.

10. *Лиходед, Н. А.* Характеристика локальности параллельных реализаций многомерных циклов / Н. А. Лиходед // Докл. НАН Беларуси. – 2010. – Т. 54, № 1. – С. 26–32.

11. *Адуцкевич, Е. В.* К распараллеливанию последовательных программ: распределение массивов между процессорами и структуризация коммуникаций / Е. В. Адуцкевич, Н. А. Лиходед, А. О. Сикорский // Кибернетика и системный анализ. – 2012. – Т. 48, № 1. – С. 144–163.

Поступило в редакцию 30.03.2015