

УДК 519.6+681.3.012

МАТРИЧНОЕ ПРЕДСТАВЛЕНИЕ БЛОКОВ ДЕКОМПОЗИРОВАННЫХ АЛГОРИТМОВ СИСТОЛИЧЕСКОГО ТИПА

А. А. Тиунчик

A matrix model for uniting decomposed dependence graphs is proposed. The model allows one to solve the following problems: to detect similar dependence graphs; to construct unlimited dependence graphs for iterative algorithms on the base of a unique graph model; to construct unlimited dependence graphs for iterative algorithms on the base of a pair of graph models; to detect dependence graphs that can be used for it; to check an existence of a temporal coordination of united graphs.

Введение

Систолические процессоры – высокоэффективные специализированные многопроцессорные устройства, хорошо приспособленные для СВИС или FPGA реализации и использующие преимущества параллельной и конвейерной обработки информации. В настоящее время в [1–4] разработаны процедуры проектирования систолических процессоров для реализации алгоритмов, представленных как единое целое. К основным этапам проектирования относятся: представление реализуемого алгоритма в специальном виде, построение графовой модели, отображение этой модели на вычислительную структуру. Однако в ряде случаев построение графовой модели затруднительно в силу большой сложности или громоздкости реализуемого алгоритма. В этом случае целесообразно разбить исходный алгоритм на подалгоритмы и построить единую графовую модель из графов, представляющих эти подалгоритмы.

В работе [5] рассмотрено построение единой графовой модели для последовательного перемножения произвольного числа матриц. Единая модель построена на основе формализованной пространственной стыковки графовых моделей для отдельных умножений, однако задача выбора подходящих графовых моделей отдельных подалгоритмов решалась на основе перебора различных возможных комбинаций. В работе [6] рассмотрен более общий способ воссоединения графовых моделей в единую структуру, однако формализованная процедура нахождения оператора пространственного совмещения была предложена только для одномерного примера. Целью этой работы является разработка матричного аппарата представления структур обрабатываемых данных, которая позволяет: находить во множестве графовых моделей для реализации одного и того же алгоритма подобные; устанавливать существование аффинных операторов пространственной стыковки и быстро получать их матрицы для построения сложных алгоритмов на основе как единственной графовой модели, так и двух различных графовых моделей; формализованно устанавливать существование единого временного планирования для вычислительного устройства, реализующего весь алгоритм.

Keywords: *parallelization, systolic algorithm, iterative algorithm, space-time coordination*
2000 Mathematics Subject Classification: 65Y05, 68M07

© А. А. Тиунчик, 2002.

1. Основные понятия

Граф зависимостей (ГЗ) является графовой моделью алгоритма. Построение ГЗ основано на представлении исходного алгоритма однородными рекуррентными уравнениями [7]

$$x_i(v) = F_i(x_1(v - \varphi_{x_1}), x_2(v - \varphi_{x_2}), \dots, x_p(v - \varphi_{x_p})), \quad 1 \leq i \leq p, \quad p \in \mathbb{N}, \quad v \in V \subset \mathbb{Z}^m, \quad (1)$$

где x_1, x_2, \dots, x_p – переменные алгоритма, v – m -мерный вектор (точка m -мерной целочисленной решетки \mathbb{Z}^m); F_i – функция p переменных x_1, x_2, \dots, x_p , а $\varphi_{x_1}, \varphi_{x_2}, \dots, \varphi_{x_p}$ – m -мерные векторы. Предполагается, что векторы $\varphi_{x_1}, \varphi_{x_2}, \dots, \varphi_{x_p}$ являются локальными, т.е. компоненты векторов принадлежат множеству $\{0, 1, -1\}$. Область V называется областью вычислений.

Алгоритм (1) может быть представлен ГЗ, вершины которого идентифицируются с точками $v \in V$, а дуги – с векторами $\varphi_{x_1}, \varphi_{x_2}, \dots, \varphi_{x_p}$. Множество вершин графа совпадает с областью вычислений V . Множество дуг $E \subset V \times V$ характеризуется множеством векторов $\{\varphi_{x_1}, \varphi_{x_2}, \dots, \varphi_{x_p}\}$.

Вектор φ_{x_i} , соответствующий дуге для пересылки данного x_i из вершины v_1 в вершину v_2 , определяется координатами $\varphi_{x_i} = (v_2 - v_1)$. Если $(v - \varphi_{x_i}) \in V$, то вектор φ_{x_i} называется вектором зависимости. Если $(v - \varphi_{x_i}) \notin V$, то точка v называется точкой ввода; если $(v + \varphi_{x_i}) \notin V$, то точка v называется точкой вывода. Предполагается, что длина и направление соответствующего вектора ввода или вывода φ_{x_i} не определены.

Предполагается, что ГЗ является строго направленным, т.е. существует такой целочисленный вектор τ , $\tau \in \mathbb{Z}^m$, который образует острые углы со всеми векторами зависимостей φ_{x_i} :

$$\langle \tau, \varphi_{x_i} \rangle > 0, \quad 1 \leq i \leq p, \quad (2)$$

где $\langle \cdot, \cdot \rangle$ обозначает скалярное произведение. Вектор τ называется направляющим вектором ГЗ.

Пространственное отображение ГЗ на вычислительную структуру определяется обычно линейным оператором $\Pi: \mathbb{Z}^m \rightarrow \mathbb{Z}^r$ ($r = 0, 1$ или 2) [1–6]. Использование линейных операторов специального вида автоматически сохраняет исходную локальность векторов зависимости, что обеспечивает получение массивов с локальными межсоединениями. Таким образом, проектирование систолических процессоров (не имеющих глобальных межсоединений) предпочтительно осуществлять на основе ГЗ, представляющих собой отдельные подалгоритмы и имеющих только локальные дуги как внутри отдельных блоков, так и между этими блоками.

Необходимо отметить, что если ГЗ не ограничен или его размеры в направлениях, задаваемых некоторыми векторами ξ_1, ξ_2, \dots , неизвестны заранее (и могут даже определяться параметрами, генерируемыми в ходе реализации вычислительного процесса), то необходимо, чтобы подпространство, порожаемое векторами ξ_1, ξ_2, \dots , входило в ядро оператора Π . Такое отображение позволяет проектировать конечные систолические процессоры для реализации произвольного числа итерационных шагов алгоритма.

2. Матричное представление

Пусть заданный алгоритм состоит из отдельных информационно связанных подалгоритмов вида (1). Каждый подалгоритм представлен ГЗ, причем выходные промежуточные результаты одного ГЗ должны использоваться как начальные входные данные для последующих вычислений. Пространственное размещение вершин вывода может существенно отличаться от пространственного размещения вершин ввода. В силу этого возникает задача пространственного соединения и согласования отдельных блоков.

Типичный случай передачи данных – передача векторов, матриц и других упорядоченных структур. Естественно предположить, что и элементы структур, и соответствующие вершины ввода-вывода упорядочены в пространстве и образуют некоторые регулярные структуры.

Пусть области ввода-вывода являются $(m - 1)$ -мерными прямоугольными параллелепипедами. Пусть входные (выходные) вершины для элементов $s_{i_1, i_2, \dots, i_{m-1}}$ структуры S размещены в точках $(\nu_{p_0}, \nu_{p_1}, \nu_{p_2}, \dots, \nu_{p_{m-1}})$, где $\{p_0, p_1, p_2, \dots, p_{m-1}\}$ — некоторая перестановка чисел $\{0, 1, 2, \dots, m - 1\}$, $\nu_0 = c_0$, $\nu_k = c_k \pm i_k$, $1 \leq k \leq m - 1$, а $c_0, c_1, c_2, \dots, c_{m-1}$ — некоторые константы. Упорядоченные таким образом области ввода и вывода будем называть регулярно структурированными областями ввода (I -областями) и регулярно структурированными областями вывода (O -областями) соответственно. Предполагается, что эти области используются для передачи одних и тех же наборов данных; следовательно, их размеры должны быть равными.

Описание областей ввода и вывода должно устанавливать соответствие между элементами ввода или вывода и точками пространства, где и происходит ввод или вывод этих элементов. Как правило, описание областей ввода и вывода дается в очень сложной форме. Это затрудняет разработку формализованных средств пространственного преобразования этих областей. С другой стороны, число различных вариантов пространственного размещения I -областей и O -областей экспоненциально возрастает с ростом размерности пространства \mathbb{Z}^m , что существенно затрудняет решение задачи стыковки методом перебора.

Для решения задачи формализованного пространственного соединения ГЗ будем использовать следующее матричное представление областей ввода-вывода. Пусть вектор $d_k \in \mathbb{Z}^m$, $1 \leq k \leq m - 1$, имеет компоненты $(0, 0, \dots, 0, \pm 1, 0, \dots, 0)$, где знак ненулевой ν_{p_k} -й компоненты вектора d_k совпадает со знаком компоненты $\nu_k = c_k \pm i_k$, а вектор $d_0 \in \mathbb{Z}^m$ имеет компоненты $(0, 0, \dots, 0, \pm 1, 0, \dots, 0)$, где знак ненулевой ν_{p_0} -й компоненты должен быть определен следующим образом: d_0 направлен внутрь области вычислений в случае области входа; в противном случае d_0 направлен наружу. Чтобы описать I - или O -область, сформируем матрицу $D = (d_0^T, d_1^T, \dots, d_{m-1}^T)$, которая сгенерирована вектор-столбцами $d_0^T, d_1^T, \dots, d_{m-1}^T$. Таким образом, любая I - или O -область однозначно определена матрицей D . Очевидно, все векторы $d_0^T, d_1^T, \dots, d_{m-1}^T$ отличны от нуля и ортогональны. Следовательно, матрица D невырождена, и существует обратная матрица D^{-1} такая, что $DD^{-1} = D^{-1}D = E$. Более того, так как векторы $d_0^T, d_1^T, \dots, d_{m-1}^T$ единичны, то столбцы матрицы D образуют ортонормированную систему, а сама матрица D является ортогональной и, следовательно, $D^{-1} = D^T$.

Через D_I и D_O будем обозначать матрицы, определяющие I - и O -области соответственно. Через Φ обозначим матрицу, сформированную вектор-столбцами $\varphi_{x_1}^T, \varphi_{x_2}^T, \dots, \varphi_{x_p}^T$. Отметим, что матрица Φ может не быть квадратной.

Определение. Совокупность матриц D_I, D_O и Φ ГЗ G будем называть матричным представлением ГЗ G и обозначать $M(D_I, D_O, \Phi)$.

3. Нахождение подобных графов зависимостей

Определение. ГЗ с матричными представлениями $M(D_{I,1}, D_{O,1}, \Phi_1)$ и $M(D_{I,2}, D_{O,2}, \Phi_2)$ будем называть подобными, если существует такой линейный оператор F , что $FD_{I,1} = D_{I,2}$, $FD_{O,1} = D_{O,2}$, $F\Phi_1 = \Phi_2$.

Подобные ГЗ появляются в силу многообразия возможных вариантов подачи и пересылки начальных данных по ГЗ. Являясь, по существу (с точностью до линейного преобразования), одним и тем же ГЗ, подобные ГЗ для реализации одного и того же алгоритма усложняют задачу нахождения стыкуемых подграфов, увеличивая число возможных комбинаций. Таким образом, предварительное нахождение и устранение из рассмотрения подобных ГЗ может существенно упростить решение задачи нахождения стыкуемых подграфов.

Отметим, что в силу невырожденности матриц $D_{I,1}$, $D_{O,1}$, $D_{I,2}$ и $D_{O,2}$ матрица оператора F может быть однозначно найдена по этим матрицам, однако это же нельзя утверждать относительно Φ_1 и Φ_2 : в зависимости от их размерности могут быть как ситуации, когда оператор F не существует, так и ситуации, когда он определяется не единственным образом. Из сказанного выше вытекает истинность следующей леммы.

Лемма 1. В случае невырожденности матриц Φ_1 и Φ_2 для подобия ГЗ $M(D_{I,1}, D_{O,1}, \Phi_1)$ и $M(D_{I,2}, D_{O,2}, \Phi_2)$ необходимо и достаточно, чтобы $D_{I,2}D_{I,1}^{-1} = D_{O,2}D_{O,1}^{-1} = \Phi_2\Phi_1^{-1}$.

Отметим, что если получено n исходных ГЗ, то для нахождения всех подобных ГЗ в этом множестве требуется $S_n = \frac{n(n-1)}{2}$ сравнений. Однако процедуру перебора можно существенно упростить, если использовать следующие необходимые условия подобия.

Лемма 2. Для подобия ГЗ $M(D_{I,1}, D_{O,1}, \Phi_1)$ и $M(D_{I,2}, D_{O,2}, \Phi_2)$ необходимо, чтобы либо $D_{I,1} = D_{I,2}$, $D_{O,1} = D_{O,2}$, $\Phi_1 = \Phi_2$, либо $D_{I,1} \neq D_{I,2}$, $D_{O,1} \neq D_{O,2}$ и в случае невырожденности матриц Φ_1 и Φ_2 $\Phi_1 \neq \Phi_2$.

Лемма 3. Для подобия ГЗ $M(D_{I,1}, D_{O,1}, \Phi_1)$ и $M(D_{I,2}, D_{O,2}, \Phi_2)$ необходимо, чтобы $\frac{\det D_{I,2}}{\det D_{I,1}} = \frac{\det D_{O,2}}{\det D_{O,1}}$, а в случае невырожденности матриц Φ_1 и Φ_2 $\frac{\det D_{I,2}}{\det D_{I,1}} = \frac{\det D_{O,2}}{\det D_{O,1}} = \frac{\det \Phi_2}{\det \Phi_1}$.

Истинность этих лемм следует из единственности оператора F .

Процедура применения леммы 2 является вычислительно легкой, поскольку не требует вычислений, а сводится только к сравнению трех пар матриц. С другой стороны, процедуру применения леммы 2 можно дополнительно ускорить, если предварительно вычислить определители всех матриц (эти определители легко вычислимы в силу специальной структуры матриц, кроме того, они понадобятся также и при реализации процедуры леммы 3), так как из неравенства определителей следует и неравенство самих сравниваемых матриц (поэлементному сравнению подлежат только матрицы с одинаковыми определителями).

4. Пространственная стыковка отдельных ГЗ

ГЗ можно рассматривать как геометрический объект, на который можно действовать линейными или аффинными операторами. Два ГЗ будем называть пространственно состыкованными, если соединение любой пары соответствующих друг другу вершин ввода и вывода определяется одним и тем же вектором с координатами из множества $\{0, 1, -1\}$. Стыковка обеспечивает систолический способ воссоздания декомпозированного алгоритма. Нашей целью является разработка средств пространственной стыковки за счет изменения взаимного расположения вершин ввода и вывода.

Пусть A – аффинный оператор, $A: \mathbb{Z}^m \rightarrow \mathbb{Z}^m$, $Ax = Lx + l$, где L – линейный оператор, l – вектор переноса, $x \in \mathbb{Z}^m$. Линейные операторы и соответствующие им матрицы различать в обозначениях не будем. Граф G может быть преобразован оператором A в граф AG : преобразование может включать вращение и отражение L и перенос l . Знаком “*” в матричном представлении графа зависимостей будем обозначать произвольную матрицу (D_I , D_O или Φ), вид которой не имеет значения для дальнейшего изложения. Пусть G_1 и G_2 – ГЗ с O - и I -областями и матричными представлениями $M(*, D_O, *)$ и $M(D_I, *, *)$ соответственно. Информация из G_1 должна перекачиваться к пространственно присоединенному с ним графу AG_2 . Очевидно, если существует линейный оператор L такой, что O -область G_1 совпадает с I -областью LG_2 , т.е. $LD_I = D_O$, то всегда можно найти соответствующий вектор l , чтобы обеспечить пространственное соединение.

Теорема 1. Любой ГЗ с матричным представлением $M(*, D_O, *)$ может быть пространственно соединен с любым ГЗ с матричным представлением $M(D_I, *, *)$. Такое соединение может быть выполнено единственным способом.

Доказательство. Пусть линейное преобразование L обеспечивает пространственную стыковку I -области, описываемой матрицей D_I , к O -области, описываемой матрицей D_O , т.е. $LD_I = D_O$. Матрица D_O обратима, следовательно, L определяется единственным образом как

$$L = D_O D_I^{-1}. \quad (3)$$

Итак, преобразование A может быть определено для любых I - и O -области.

Матрицы D_O и D_I порождаются линейно независимыми единичными векторами, следовательно, произведение (3) определяет некоторое вращение гиперкуба на угол $\pi/2$ или π либо отражение относительно гиперплоскости, проходящей через одну из его граней. Оператор L должен сохранять внутреннюю структуру исходного ГЗ, т.е. сохранять локальность векторов зависимости и размещение областей ввода и вывода на границе области вычислений. Выполнение этих требований следует из геометрического смысла вращений и отражений. Теорема доказана.

Отметим, что изменение знака вектора d_0 в D_O или D_I на противоположный соответствует получению внутренней стыковки [8]: пересылаемые данные "отражаются" от границы ГЗ и возвращаются в исходную область вычислений.

Рассмотрим случай, когда исходный алгоритм разбит на множество идентичных подалгоритмов и требуется осуществить пространственную стыковку большого количества идентичных ГЗ. Существуют два подхода к решению задачи стыковки множества идентичных ГЗ: 1) построить ряд различных ГЗ для реализации одного и того же подалгоритма и попытаться выбрать из них такие, которые могут быть состыкованы естественным образом; 2) построить один ГЗ и попытаться найти алгебраические средства для таких преобразований этого ГЗ, которые удовлетворяли бы всем требованиям стыковки. Первый подход подробно исследован в [5]. Основным его недостатком в общем случае является высокая трудоемкость процедур построения и исследования различных графовых реализаций подалгоритмов. Рассмотрим возможности второго подхода.

Пусть G_1, G_2, \dots, G_Q – идентичные ГЗ, которые представляют Q подалгоритмов. Пусть A_n – аффинный оператор, $A_n: Z^m \rightarrow Z^m$, $A_n x = L_n x + I_n$, где L_n – линейный оператор, I_n – вектор переноса, $x \in Z^m$. Тогда граф G_n может быть преобразован операторами A_n в граф $A_n G_n$ (различные графы G_n могут быть преобразованы различными операторами A_n).

Линейные операторы L_n могут изменять пространственное размещение векторов зависимостей, что приводит к появлению большого числа типов вершин. В связи с этим обретает актуальность проектирование регулярной однородной цепи преобразованных ГЗ. Для построения такой цепи нежелательно использование большого числа различных операторов. Рассмотрим случай применения единственного оператора A .

Построение цепи пространственно связанных ГЗ с использованием единственного оператора A осуществляется следующим образом. Первый ГЗ G_1 остается неподвижным, второй ГЗ G_2 должен быть присоединен посредством оператора A . Предполагается, что третий ГЗ G_3 мог быть заранее связан с G_2 оператором A ; следовательно, присоединение G_1 к G_2 означает, что на G_3 действует оператор A^2 . Таким образом можно продолжать процедуру и получить цепь пространственно связанных ГЗ $G_1, AG_2, A^2G_3, A^3G_4, \dots, A^{Q-1}G_Q$, где n -я степень оператора A рассматривается как n последовательных преобразований оператором A :

$$\begin{aligned} A^n x &= (L + 1)^n x = \underbrace{L(L(L \dots L(Lx + 1) + 1 + \dots + 1) + 1) + 1} = \\ &= L^n x + L^{n-1}1 + L^{n-2}1 + L^{n-3}1 + \dots + L^3 1 + L^2 1 + L 1 + 1. \end{aligned}$$

Таким образом, аффинное преобразование оператором A^n сводится к линейному преобразованию L^n и параллельному переносу, задаваемому вектором $L^{n-1}1 + L^{n-2}1 + \dots + L 1 + 1$. Чтобы ограничить число различных операторов $L, L^2, L^3, \dots, L^{Q-1}$, будем предполагать, что существует число λ такое, что

$$L^\lambda = E, \quad \lambda \geq 1, \quad (4)$$

и λ мало (обычно $\lambda = 1$ или 2). Цепь ГЗ $G_1, AG_2, A^2G_3, A^3G_4, \dots$ при $L^\lambda = E$ будем называть L^λ -цепью.

Рассмотрим следующие задачи: 1) возможно ли объединить декомпозированный алгоритм (создать L^λ -цепь) на основе единственного ГЗ с матричным представлением $M(D_I, D_O, \Phi)$, если $\lambda \leq 2$, и 2) если это так, то какими свойствами должны обладать матрицы, входящие в это матричное представление?

Из теоремы 1 следует, что существует линейный оператор L_{OI} , обеспечивающий пространственную стыковку области ввода данных следующего ГЗ с областью вывода результатов исходного ГЗ, и линейный оператор L_{IO} , обеспечивающий пространственную стыковку области вывода предыдущего ГЗ с областью ввода исходного ГЗ. Так как все ГЗ идентичны, то все возможности пространственной стыковки могут быть установлены на основании исследования I - и O -областей одного ГЗ.

Теорема 2. L^λ -цепь при $\lambda \leq 2$ может быть построена тогда и только тогда, когда $D_O(D_I)^{-1} = D_I(D_O)^{-1}$.

Доказательство. Пусть D_I и D_O - матрицы описания I - и O -областей соответственно. Тогда $L_{IO}D_I = D_O$ и $L_{OI}D_O = D_I$,

$$L_{OI} = D_I(D_O)^{-1}, \tag{5}$$

$$L_{IO} = D_O(D_I)^{-1}. \tag{6}$$

Для построения L^λ -цепи должен использоваться один оператор. Таким образом, L_{IO} должен быть равен L_{OI} , а $D_O(D_I)^{-1} = D_I(D_O)^{-1}$. С другой стороны, если $D_O(D_I)^{-1} = D_I(D_O)^{-1}$, то $L_{IO} = L_{OI}$. Теорема доказана.

Теорема 3. Любая составленная из n ГЗ L^λ -цепь при $\lambda \leq 2$ может быть заключена в цилиндр, радиус которого не зависит от n .

Доказательство. Для доказательства этой теоремы достаточно показать, что любая L^λ -цепь при $\lambda \leq 2$ распространяется в направлении некоторого фиксированного вектора ξ .

Действительно, если $\lambda = 1$, то $L = E$, и вся цепь может быть построена путем параллельных переносов на вектор 1 . Вектор $\xi = 1$ должен входить в ядро оператора отображения всей цепи, при этом будет получен конечный систолический процессор для реализации произвольного числа шагов.

Пусть $\lambda = 2$. Так как $L^2 = E$, то для четного n

$$A^n x = E x + \underbrace{L1 + 1 + \dots + L1 + 1 + L1 + 1}_n = E x + \frac{n}{2}(L1 + 1),$$

а для нечетного n

$$A^n x = Lx + \underbrace{1 + L1 + \dots + L1 + 1}_n = Lx + \frac{n-1}{2}(L1 + 1) + 1.$$

Таким образом, L^λ -цепь распространяется в направлении вектора $L1 + 1$; вектор $\xi = L1 + 1$ должен входить в ядро оператора отображения. Теорема доказана.

При построении L^2 -цепи при $\lambda = 2$ используются не только исходные ГЗ, но и ГЗ, преобразованные линейным оператором L . В преобразованных ГЗ изменяется пространственное размещение векторов зависимости, что приводит к появлению новых типов вершин и, формально, новых ГЗ. Это позволяет рассмотреть еще один способ построения цепи состыкованных ГЗ, основанный на использовании двух различных исходных ГЗ.

Пусть даны ГЗ с матричным представлением $M(D_{I_1}, D_{O_1}, \Phi_1)$ и ГЗ с матричным представлением $M(D_{I_2}, D_{O_2}, \Phi_2)$. Пусть пространственная стыковка I -области второго ГЗ к O -области первого осуществляется оператором $L_{O_1 I_2}$, а I -области первого ГЗ к O -области второго - оператором $L_{O_2 I_1}$. Так как последовательное применение этих операторов к последующим ГЗ эквивалентно применению произведения операторов, то во избежание появления большого числа типов новых ГЗ естественно потребовать, чтобы выполнялось равенство $L_{O_1 I_2} L_{O_2 I_1} = E$.

Через L^2 -цепь обозначим цепь, построенную из ГЗ с матричными представлениями $M(D_{I_1}, D_{O_1}, \Phi_1)$ и $M(D_{I_2}, D_{O_2}, \Phi_2)$ на основе применения двух аффинных операторов $A_{O_2 I_1} : \mathbb{Z}^m \rightarrow \mathbb{Z}^m$, $A_{O_2 I_1} x = L_{O_2 I_1} x + l_{O_2 I_1}$ и $A_{O_1 I_2} : \mathbb{Z}^m \rightarrow \mathbb{Z}^m$, $A_{O_1 I_2} x = L_{O_1 I_2} x + l_{O_1 I_2}$, где $L_{O_2 I_1}$ и $L_{O_1 I_2}$ - линейные операторы, $l_{O_2 I_1}$ и $l_{O_1 I_2}$ - векторы переноса, $x \in \mathbb{Z}^m$;

$L_{O_2I_1}L_{O_1I_2} = L_{O_1I_2}L_{O_2I_1} = E$. Другими словами, L_*^2 -цепь строится параллельным переносом пар построенных состыкованных ГЗ с матричными представлениями $M(D_{I_1}, D_{O_1}, \Phi_1)$ и $M(D_{I_2}, D_{O_2}, \Phi_2)$. Такая процедура построения не противоречит требованию малого числа различных типов ГЗ (это число по-прежнему равно двум).

Теорема 4. L_*^2 -цепь из ГЗ с матричными представлениями $M(D_{I_1}, D_{O_1}, \Phi_1)$ и $M(D_{I_2}, D_{O_2}, \Phi_2)$ может быть построена тогда и только тогда, когда

$$D_{O_2}(D_{I_1})^{-1}D_{O_1}(D_{I_2})^{-1} = E. \quad (7)$$

Доказательство. Из теоремы 1 следует, что существуют линейные операторы $L_{O_2I_1}$ и $L_{O_1I_2}$, осуществляющие пространственную стыковку ГЗ с матричными представлениями $M(D_{I_1}, D_{O_1}, \Phi_1)$ и $M(D_{I_2}, D_{O_2}, \Phi_2)$:

$$L_{O_2I_1} = D_{O_2}(D_{I_1})^{-1}, \quad (8)$$

$$L_{O_1I_2} = D_{O_1}(D_{I_2})^{-1}. \quad (9)$$

Так как для построения L_*^2 -цепи требуется, чтобы $L_{O_2I_1}L_{O_1I_2} = L_{O_1I_2}L_{O_2I_1} = E$, то с учетом равенств (8) и (9) получаем соотношение (7).

С другой стороны, из соотношения (7) следует существование операторов $L_{O_2I_1}$ и $L_{O_1I_2}$, отвечающих требованию $L_{O_2I_1}L_{O_1I_2} = L_{O_1I_2}L_{O_2I_1} = E$. Теорема доказана.

Следствие. Для существования L_*^2 -цепи из ГЗ с матричными представлениями $M(D_{I_1}, D_{O_1}, \Phi_1)$ и $M(D_{I_2}, D_{O_2}, \Phi_2)$ необходимо, чтобы $\frac{\det D_{O_2} \det D_{O_1}}{\det D_{I_1} \det D_{I_2}} = 1$.

В заключение отметим, что при исследовании возможностей пространственной стыковки на основе анализа матричных представлений ГЗ матрица Φ не используется.

5. Временное согласование объединенного ГЗ

Временной режим работы проектируемого систолического процессора определяется функцией $t(v) = \langle \tau, v \rangle - \min_{v \in V} \langle \tau, v \rangle + t_0$, где τ — направляющий вектор ГЗ, t_0 — целочисленная неотрицательная константа. Нахождение непротиворечивого временного режима работы систолического устройства сводится к нахождению такого целочисленного вектора τ , для которого

$$\langle \tau, \varphi_{x_i} \rangle > 0, \quad (10)$$

для любого вектора зависимости φ_{x_i} , $1 \leq i \leq p$. Очевидно, что существование вектора τ , удовлетворяющего условию (10), равносильно существованию целочисленного решения τ системы линейных неравенств $\Phi\tau > 0$.

Теорема 5. Единая таймирующая функция для L^2 -цепи при $\lambda \leq 2$ существует тогда и только тогда, когда существует целочисленное решение τ системы линейных неравенств

$$B\tau > 0, \quad \text{где } B = \begin{pmatrix} (\Phi)^T \\ (D_I D_O^{-1} \Phi)^T \end{pmatrix}.$$

Истинность теоремы 5 следует из того, что направляющий вектор всего множества пространственно состыкованных ГЗ должен являться целочисленным решением систем неравенств $\Phi\tau > 0$, $L\Phi\tau > 0$, $L^2\Phi\tau > 0$, $L^3\Phi\tau > 0 \dots$, причем $L^2 = E$, $L = D_I D_O^{-1}$.

Аналогично устанавливается истинность и следующей ниже теоремы о существовании таймирующей функции для L_*^2 -цепи.

Теорема 6. Единая таймирующая функция для L_*^2 -цепи из ГЗ с матричными представлениями $M(D_{I_1}, D_{O_1}, \Phi_1)$ и $M(D_{I_2}, D_{O_2}, \Phi_2)$ существует тогда и только тогда, когда существует целочисленное решение τ системы линейных неравенств $B\tau > 0$, где $B =$

$$= \begin{pmatrix} (\Phi_1)^T \\ (D_{O_1} D_{I_2}^{-1} \Phi_2)^T \end{pmatrix}.$$

6. Пример

Рассмотрим построение систолического алгоритма для перемножения последовательности $N \times N$ -матриц

$$Y = A_1 \cdots A_3 \cdot A_2 \cdot A_1. \quad (11)$$

Эта процедура естественным образом представляется как последовательность умножений матрицы на матрицу: $X_1 = A_2 \cdot A_1$, $X_2 = A_3 \cdot X_1$, $X_3 = A_4 \cdot X_2, \dots, X_{l-1} = A_l \cdot X_{l-2}$. Через

$$Y = A \cdot X \quad (12)$$

обозначим некоторое промежуточное умножение матрицы на матрицу. Воспользуемся одной из систолических реализаций [5] операции (12).

Пусть даны матрицы $A = (a_{ip})$ и $X = (x_{pj})$, $1 \leq i, j, p \leq N$. Элементы матрицы $Y = (y_{ij})$, $1 \leq i, j \leq N$, формируются в соответствии с соотношениями $y_{ij} = \sum_{p=1}^N a_{ip}x_{pj}$. Реализация такого вычисления может быть представлена рекуррентными уравнениями

$$y(i, j, k) = a(i, j-1, k)x(i-1, j, k) + y(i, j, k-1), \quad (13)$$

$$a(i, j, k) = a(i, j-1, k), \quad x(i, j, k) = x(i-1, j, k), \quad 1 \leq i, j, k \leq N,$$

где ввод данных описывается присвоениями $x(0, j, k) = x_{kj}$, $a(i, 0, k) = a_{ik}$, $y(i, j, 0) = 0$, вывод осуществляется как $y_{ij} = y(i, j, N)$. Рекуррентные уравнения (13) позволяют построить ГЗ реализации (12): вершины ГЗ размещены в кубической области вычислений $\{(i, j, k) : 1 \leq i, j, k \leq N\}$. Векторами зависимости являются векторы $\varphi_x = (1, 0, 0)$, $\varphi_a = (0, 1, 0)$

и $\varphi_y = (0, 0, 1)$. Таким образом, $\Phi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Направляющим вектором этого ГЗ

является вектор $\tau = (\tau_1, \tau_2, \tau_3)$, где $\tau_1 \geq 1$, $\tau_2 \geq 1$, $\tau_3 \geq 1$.

Дадим матричное описание областей ввода и вывода. Векторами d_0 , d_1 и d_2 области ввода являются векторы $d_0^I = (1, 0, 0)$, $d_1^I = (0, 0, 1)$ и $d_2^I = (0, 1, 0)$ соответственно. Аналогично

$d_0^O = (0, 0, 1)$, $d_1^O = (1, 0, 0)$ и $d_2^O = (0, 1, 0)$. Таким образом, $D_I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$,

$$D_O = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

В работе [5] получено также и множество других рекуррентных уравнений для реализации итерационного шага (12). Приведем матричное описание ГЗ, представляющих эти алгоритмы. ГЗ, представляющий i -й алгоритм, будем обозначать Γ_{3i} , $1 \leq i \leq 16$. Соответствующие ему матрицы будем обозначать $D_{I,i}$, $D_{O,i}$ и Φ_i . Эти матрицы имеют следующий вид: $D_{I,1} =$

$$= D_{I,2} = D_{I,9} = D_{I,10} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad D_{I,3} = D_{I,4} = D_{I,11} = D_{I,12} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

$$D_{I,5} = D_{I,6} = D_{I,13} = D_{I,14} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \quad D_{I,7} = D_{I,8} = D_{I,15} = D_{I,16} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix},$$

$$D_{O,1} = D_{O,2} = D_{O,3} = D_{O,4} = D_{O,5} = D_{O,6} = D_{O,7} = D_{O,8} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad D_{O,9} = D_{O,10} =$$

$$= D_{O,11} = D_{O,12} = D_{O,13} = D_{O,14} = D_{O,15} = D_{O,16} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \end{pmatrix}, \quad \Phi_1 = \Phi_5 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\Phi_2 = \Phi_6 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \Phi_3 = \Phi_7 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \Phi_4 = \Phi_8 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \Phi_9 =$$

$$= \Phi_{13} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad \Phi_{10} = \Phi_{14} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad \Phi_{11} = \Phi_{15} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix},$$

$$\Phi_{12} = \Phi_{16} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Определители этих матриц легко вычислимы: $\det D_{I,1} = \det D_{I,2} = \det D_{I,7} = \det D_{I,8} = \det D_{I,9} = \det D_{I,10} = \det D_{I,15} = \det D_{I,16} = \det D_{O,9} = \det D_{O,10} = \det D_{O,11} = \det D_{O,12} = \det D_{O,13} = \det D_{O,14} = \det D_{O,15} = \det D_{O,16} = \det \Phi_2 = \det \Phi_3 = \det \Phi_6 = \det \Phi_7 = \det \Phi_9 = \det \Phi_{12} = \det \Phi_{13} = \det \Phi_{16} = -1$, $\det D_{I,3} = \det D_{I,4} = \det D_{I,5} = \det D_{I,6} = \det D_{I,11} = \det D_{I,12} = \det D_{I,13} = \det D_{I,14} = \det D_{O,1} = \det D_{O,2} = \det D_{O,3} = \det D_{O,4} = \det D_{O,5} = \det D_{O,6} = \det D_{O,7} = \det D_{O,8} = \det \Phi_1 = \det \Phi_4 = \det \Phi_5 = \det \Phi_8 = \det \Phi_{10} = \det \Phi_{11} = \det \Phi_{14} = \det \Phi_{15} = 1$.

Нахождение подобных ГЗ в данном случае требует $S_{16} = 120$ операций преобразования и сравнения трех пар матриц. Рассмотрим применение процедуры быстрого нахождения подобных ГЗ на примере нахождения ГЗ, подобных графу зависимостей ГЗ₁. В силу леммы 2 подобными ему могут быть только ГЗ₁₁, ГЗ₁₂, ГЗ₁₃, ГЗ₁₄, ГЗ₁₅, ГЗ₁₆. В силу леммы 3 подобными ГЗ₁ из этого множества могут быть только ГЗ₁₂ и ГЗ₁₃. Непосредственной проверкой выполнения условий леммы 1 убеждаемся в том, что только ГЗ₁₃ подобен ГЗ₁. Аналогичным путем устанавливаем, что парами подобных ГЗ являются ГЗ₁ и ГЗ₁₃, ГЗ₂ и ГЗ₁₄, ГЗ₃ и ГЗ₁₅, ГЗ₄ и ГЗ₁₆, ГЗ₅ и ГЗ₉, ГЗ₆ и ГЗ₁₀, ГЗ₇ и ГЗ₁₁, ГЗ₈ и ГЗ₁₂. Процесс нахождения пар подобных ГЗ отражен в таблице. Так как в правой колонке таблицы встречаются все ГЗ_{*i*}, $9 \leq i \leq 16$, то дальнейшая проверка смысла не имеет.

Нахождение подобных ГЗ на основании лемм 1-3.

Номера проверяемых ГЗ	Номера ГЗ, удовлетворяющих условию леммы 2	Номера ГЗ, удовлетворяющих условиям лемм 2 и 3	Номера подобных ГЗ (ГЗ, удовлетворяющих условию леммы 1)
1	11, 12, 13, 14, 15, 16	12, 13	13
2	11, 12, 13, 14, 15, 16	11, 14	14
3	9, 10, 13, 14, 15, 16	10, 15	15
4	9, 10, 13, 14, 15, 16	9, 16	16
5	9, 10, 11, 12, 15, 16	9, 16	9
6	9, 10, 11, 12, 15, 16	10, 15	10
7	9, 10, 11, 12, 13, 14	11, 14	11
8	9, 10, 11, 12, 13, 14	12, 13	12

При реализации всего алгоритма (11), определяемой соотношениями (12), результаты вычислений y_{ij} должны использоваться в качестве входных данных x_{kj} для следующего умножения матриц. Вершины для входных значений x_{ij} размещены в точках $(1, j, i)$, вершины вывода расположены в точках (i, j, N) . Рассмотрим возможность построения L^λ -цепи при $\lambda \leq 2$ для реализации (11) на основе единственного ГЗ.

Так как среди 16 исходных ГЗ выявлено 8 пар подобных, то можно ограничиться рассмотрением только первых восьми ГЗ_{*i*}, $1 \leq i \leq 8$. Более того, так как матрица Φ не оказывает влияния на возможность существования пространственной стыковки, то из рассмотрения можно временно исключить ГЗ₂, ГЗ₄, ГЗ₆ и ГЗ₈ (эти ГЗ имеют те же возможности для пространственной стыковки, что и ГЗ₁, ГЗ₃, ГЗ₅ и ГЗ₇ соответственно).

В соответствии с теоремой 2 L^λ -цепь при $\lambda \leq 2$ может быть построена только на основе ГЗ₁ (или ГЗ₂ и подобных им ГЗ₁₃ и ГЗ₁₄) и ГЗ₇ (или ГЗ₈ и подобных им ГЗ₁₁ и ГЗ₁₂).

Действительно, для ГЗ₁ $D_0 D_I^{-1} = D_I D_0^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$, для ГЗ₇ $D_0 D_I^{-1} = D_I D_0^{-1} = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{pmatrix}$, однако для ГЗ₃ $D_0 D_I^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} = D_I D_0^{-1}$ и для ГЗ₅ $D_0 D_I^{-1} = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{pmatrix} = D_I D_0^{-1}$.

Все допустимые варианты пространственного построения L^λ -цепи при $\lambda \leq 2$ (на основе ГЗ₁, ГЗ₇, ГЗ₂ и ГЗ₈) допускают единое таймирование в соответствии с теоремой 5.

Рассмотрим возможность построения L_*^λ -цепи. В силу следствия из теоремы 4 существуют две возможности построения L_*^λ -цепи: либо на основе ГЗ₁ (или ГЗ₂) и ГЗ₇ (или ГЗ₈), либо на основе ГЗ₃ (или ГЗ₄) и ГЗ₅ (или ГЗ₆). Проверка выполнения условий теоремы 4 показывает, что пространственное построение L_*^λ -цепи возможно только на основе ГЗ₃ (или ГЗ₄) и ГЗ₅ (или ГЗ₆). Проверка выполнения условий теоремы 6 показывает, что для L_*^λ -цепи на основе ГЗ₃ и ГЗ₅ и для L_*^λ -цепи на основе ГЗ₄ и ГЗ₆ возможно и единое таймирование.

Таким образом, матричное представление областей ввода и вывода обеспечило быстрое нахождение подобных графов зависимостей из 16 первоначально полученных, получение аффинных операторов для построения состыкованных графов зависимостей на основе как единственно-го ГЗ, так и на основе двух различных ГЗ, а также нахождение единого временного режима работы всей вычислительной структуры.

Литература

1. Воеводин В.В. Математические модели и методы в параллельных процессах. М.: Наука, 1986.
2. Kung S.-Y. VLSI Processor Arrays. Prentice-Hall Int., 1988.
3. Quinton P., Robert Y. Systolic algorithms and architectures. Prentice-Hall and Masson, 1989.
4. Lengauer C. Parallel Computing Technologies. Ed. N. Mirenkov. World Scientific, 1991. P. 32-46.
5. Соболевский П.И., Лиходед Н.А. Внешняя стыковка при реализации одношаговых итерационных процессов на систолических структурах. Мн., 1990 (Препринт / Ин-т математики АН Беларуси: 15(415)).
6. Tiountchik A.A. // Proc. Int. Workshop on Parallel Processing by Cellular Automata and Arrays - Parcella'96. Eds. R. Vollmar, W. Erhard, V. Jossifov. Berlin, 1996. P. 77-84.
7. Karp R.M., Miller R.E., Winograd S. The organization of computations for uniform recurrence equations // J. ACM. 1967. Vol. 14. P. 563-590.
8. Лиходед Н.А., Соболевский П.И. Стыковка базовых систолических вычислителей // Докл. АН БССР. 1989. № 5. С. 389-392.

Тиунчик Александр Александрович
 Институт математики НАН Беларуси
 г. Минск, ул. Сурганова, 11
 220072, Беларусь
 e-mail: aat@im.bas-net.by