Determining space parameters in systolic array design

I. Ž. Milovanović, E. I. Milovanović, M. K. Stojčev
T. I. Tokić and N. M. Stojanović

Abstract

This paper[1] addresses the problem of determining geometric and chip area of systolic arrays for one class of nested loop algorithms in advance, before the systolic array synthesis. These parameters are determined only according to a given projection direction and loop boundaries.

1 Introduction

VLSI technology has made possible the integration of circuits with hundreds of thousands of components into a single silicon chip. This high level of integration opens the way for massive parallel computations. Systolic processing constitutes a feasible solution for massive parallel computations. Its principles are compatible with VLSI technology characteristics. Since systolic arrays are highly regular, only algorithms with repetitive computations perform well on them. Algorithms with nested loops fall into this category.

Several systolic arrays (SA) can be designed that implement a given nested loop algorithm. They can differ in several aspects including array topology, number of the processing elements (PE), execution time, geometric and chip area, number of I/O pins etc. [1]-[5]. Therefore it might be important to determine these features in advance, before the systolic array synthesis. This enables the designer to adapt the synthesis process to the predefined requirements.

The objective of this paper is to determine the geometric and chip area of the systolic array that implements a three nested loop algorithm, before SA synthesis. Taking into account some properties of the systolic algorithm and with appropriate choice of a transformation matrix can minimize space parameters of the SA.

---

[1]Presented at the IMC "Filomat 2001", Niš, August 26–30, 2001
2000 Mathematics Subject Classification: 68W10
Keywords: Systolic array, linear transformations

2 Background

We will firstly describe the procedure for determining geometric and chip area of the SA. Suppose, without deteriorating the generality, that the computations in the algorithm are performed according to the following loop nest:

for $k := 1$ **to** N_3 **do**
  for $j := 1$ **to** N_2 **do**
    for $i := 1$ **to** N_1 **do**

This means that the inner computation space of the systolic algorithm is defined by

$$P_{int} = \{(i,j,k)| \quad 1 \leq i \leq N_1, 1 \leq j \leq N_2, 1 \leq k \leq N_3\}. \tag{1}$$

Also, suppose that the two-dimensional (2D) planar SA is synthesized according to the given projection direction vector $\vec{\mu} = [\mu_1 \mu_2 \mu_3]^T$ and the corresponding transformation matrix

$$T = \begin{bmatrix} \vec{\Pi} \\ S \end{bmatrix} = \begin{bmatrix} \vec{\Pi} \\ \vec{S}_1 \\ \vec{S}_2 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{bmatrix}, \tag{2}$$

where $\vec{\Pi}$ determines time schedule and S is the space mapping function determining PE locations and communication channels between them.

To obtain planar systolic array, the following conditions must be fulfilled

$$\mu_i \in \{-1, 0, 1\}, \qquad i = 1, 2, 3$$

and

$$\vec{\Pi} \cdot \vec{\mu} \neq 0.$$

Systolic array that implements a given systolic algorithm is obtained according to the following mapping

$$T : (P_{int}, D) \longrightarrow (P, \Delta) \tag{3}$$

where D is a dependency matrix of a given algorithm and

$$\Delta = [\Delta_t \ \Delta_s]^T, \quad \Delta_t = \vec{\Pi} \cdot D, \quad \Delta_s = S \cdot D,$$
$$P = \{[t \ x \ y]^T\}, \quad t = \vec{\Pi}[i \ j \ k]^T,$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = S \begin{bmatrix} i \\ j \\ k \end{bmatrix} \tag{4}$$

for all $[i \ j \ k]^T \in P_{int}$. The x-y positions of the PEs in the SA are determined according to (4), while communication links between them are determined by Δ_s.

Definition 1 *Geometric area, g_a, of a 2D systolic array is the area of the smallest convex polygon which bounds the PEs in the x-y plane.*

Denote by $T_{1i},\ i = 1, 2, 3$ the $(1, i)$-cofactor of matrix T. Geometric area of the SA can be determined according to the following equality

$$g_a = (N_1 - 1)(N_2 - 1)|T_{13}| + (N_1 - 1)(N_3 - 1)|T_{12}| + (N_2 - 1)(N_3 - 1)|T_{11}|. \quad (5)$$

Definition 2 *The chip area, c_a, of the 2D systolic array is obtained according to*

$$c_a = (\max\{x\} - \min\{x\} + 1)(\max\{y\} - \min\{y\} + 1),$$

where x and y are defined by (4).

In other words, chip area is the smallest rectangle that bounds the obtained SA in the x-y plane.

In practice, the chip area is determined from

$$c_a = \Big(1 + \sum_{j=1}^{3} |t_{2j}|(N_j - 1)\Big)\Big(1 + \sum_{j=1}^{3} |t_{3j}|(N_j - 1)\Big). \quad (6)$$

According to (5) and (6) one can see that g_a and c_a depend on the bounds of P_{int}, which we cannot affect, and a transformation matrix T. To minimize these parameters, it is obvious that we have to deal with transformation T. This is discussed in the next section.

3 Main result

It is common known fact that for a given projection direction $\vec{\mu} = [\mu_1\, \mu_2\, \mu_3]^T$ there are several valid transformations T that map systolic algorithm into SA implementation. Since space parameters of the SA directly depend on the transformation T, the goal is to reduce a set of valid transformations to those that yield to optimal or near optimal SA with respect to space parameters. Therefore we introduce the following conditions for matrix T:

- Matrix T must be nonsingular, i.e. $\det T \neq 0$;
- Projection direction vector $\vec{\mu}$ has to be orthogonal to the projection plane, i.e.
$$\vec{\mu} = (\vec{S}_1 \times \vec{S}_2)^T; \quad (7)$$
- To avoid conflicts in the SA, the following two conditions must not be satisfied simultaneously
$$\vec{\Pi}\vec{p}_1 = \vec{\Pi}\vec{p}_2 \quad \text{and} \quad S\vec{p}_1 = S\vec{p}_2$$
for each $\vec{p}_1 \neq \vec{p}_2$ from P_{int};

- To achieve near-neighbor communications between the PEs, the following must be satisfied

$$t_{ij} \in \{-1, 0, 1\}, \quad 2 \leq i \leq 3, \quad 1 \leq j \leq 3;$$

- If $\mu_1 = 1$, then the following equality must be satisfied

$$t_{22}t_{32} + t_{23}t_{33} = 0, \tag{8}$$

and, if $\mu_2 = \pm 1$, then

$$t_{21}t_{31} + t_{23}t_{33} = 0. \tag{9}$$

must be fulfilled.

Note that conditions (8) and (9) cannot be both satisfied. Also, in order to minimize space parameters the condition (7) has replaced the condition $\vec{\mu} \cdot \vec{S}_1 = 0$ and $\vec{\mu} \cdot \vec{S}_2 = 0$.

Besides reducing the set of valid transformations $\{T\}$, it is also important to observe if the systolic algorithm has some features that can help us to minimize space parameters. Namely, under certain conditions, it is possible to map a given algorithm into an equivalent one which is accommodated to a given projection direction $\vec{\mu}$.

Let k be an iterative index variable in systolic algorithm $\mathcal{A}$. Under this condition we involve the following definitions.

Definition 3 *If for some fixed j the ordering of computations in algorithm $\mathcal{A}$, can be performed over arbitrary permutations of index variables i and k, we say that $\mathcal{A}$ is an $\mathcal{A}(i,k)$ adaptable.*

Definition 4 *If for some fixed i the ordering of computations in algorithm $\mathcal{A}$, can be performed over arbitrary permutations of index variables j and k, we say that $\mathcal{A}$ is an $\mathcal{A}(j,k)$ adaptable.*

Remark 1 *If a given algorithm $\mathcal{A}$ satisfies both Definition 3 and 4, we say that $\mathcal{A}$ is adaptable.*

The above definitions enable us to introduce the following accommodation of a given algorithm.

Definition 5 *Suppose that a given algorithm is of type $\mathcal{A}(j,k)$. If $\vec{\mu} = [1 \ \mu_2 \ \mu_3]^T$ is an allowable projection direction, the mapping $H = (F, G)$, $H : (P_{int}, D) \mapsto (\bar{P}_{int}, \bar{D})$, that performs accommodation of algorithm $\mathcal{A}$ to a given projection direction, is defined by*

$$F = \begin{bmatrix} 1 & 0 & 0 \\ \mu_2 & 1 & 0 \\ \mu_3 & 0 & 1 \end{bmatrix}, \qquad G = \begin{bmatrix} 0 \\ g_2 \\ g_3 \end{bmatrix},$$

where g_2 and g_3 are the smallest integers determined such that for each $[i\ j\ k]^T \in P_{int}$, the following is valid

$$\mu_2 i + j + g_2 > 0, \quad \textit{and} \quad \mu_3 i + k + g_3 > 0.$$

The elements of $\bar{P}_{int}$ are obtained according to

$$[u\ v\ w]^T = F[i\ j\ k]^T + G.$$

Definition 6 *Suppose that a given algorithm is of type $\mathcal{A}(i,k)$. If $\vec{\mu} = [\mu_1 \pm 1\, \mu_3]^T$ is an allowable projection direction, mapping $H = (F,G)$, $H : (P_{int}, D) \mapsto (\bar{P}_{int}, \bar{D})$ is defined by*

$$F = \begin{bmatrix} 1 & \pm\mu_1 & 0 \\ 0 & 1 & 0 \\ 0 & \pm\mu_3 & 1 \end{bmatrix}, \qquad G = \begin{bmatrix} g_1 \\ 0 \\ g_3 \end{bmatrix},$$

where g_1 and g_3 are the smallest integers determined such that foe each $[i\, j\, k]^T \in P_{int}$ the following is valid

$$i \pm \mu_1 j + g_1 > 0 \quad \textit{and} \quad k \pm \mu_3 j + g_3 > 0.$$

The elements of $\bar{P}_{int}$ are obtained according to

$$[u\ v\ w]^T = F[i\ j\ k]^T + G.$$

Suppose that a given systolic algorithm $\mathcal{A}$ is of type $\mathcal{A}(j,k)$ and that for a given direction projection $\vec{\mu} = [1\, \mu_2\, \mu_3]^T$, transformation matrix has been determined according to the previously defined criteria. Then, the SA can be synthesized according to the following two mappings

$$(P_{int}, D) \overset{H}{\longmapsto} (\bar{P}_{int}, \bar{D}) \quad \text{and} \quad (\bar{P}_{int}, \bar{D}) \overset{T}{\longmapsto} (P, \Delta).$$

Space parameters g_a and c_a of the obtained SA depend only of elements of matrix M, $M = T \circ F$,

$$M = \begin{bmatrix} t_{11} + \mu_2 t_{12} + \mu_3 t_{13} & t_{12} & t_{13} \\ 0 & t_{22} & t_{23} \\ 0 & t_{32} & t_{33} \end{bmatrix}.$$

Since $M_{12} = M_{13} = 0$ and $M_{11} = t_{22}t_{33} - t_{23}t_{32} = 1$, according to (5) and (6) we have that

$$g_a = (N_2 - 1)(N_3 - 1) \tag{10}$$

and

$$c_a = [1 + (N_2 - 1)|t_{22}| + (N_3 - 1)|t_{23}|][1 + (N_2 - 1)|t_{32}| + (N_3 - 1)|t_{33}|]. \tag{11}$$

Comparing (10) and (5), i.e. (11) and (6), it can be easily concluded that the results obtained by the described procedure have been improved.

Similarly, if a given systolic algorithm $\mathcal{A}$ is of $\mathcal{A}(i,k)$ type and if $\vec{\mu} = [\mu_1 \pm 1\,\mu_3]^T$, we have

$$g_a = (N_1 - 1)(N_3 - 1)$$

and

$$c_a = [1 + (N_1 - 1)|t_{21}| + (N_3 - 1)|t_{23}|][1 + (N_1 - 1)|t_{31}| + (N_3 - 1)|t_{33}|].$$

It is interesting when a given algorithm $\mathcal{A}$ is both of $\mathcal{A}(i,k)$ and $\mathcal{A}(j,k)$ type, and when $\vec{\mu} = [1 \pm 1\,\mu_3]^T$. In that case we have

$$g_a = (N_3 - 1)\min\{N_1 - 1, N_2 - 1\} \quad \text{and} \quad c_a = N_3 \min\{N_1, N_2\}.$$

At the end, let us note that a great number of algorithms in scientific and technical computations have the features from Definition 5 and 6. For example matrix multiplication and matrix-vector multiplication.

References

[1] M. O. Esonu, J. Al-Khalili, S. Hariri, D. Al-Khalili, *Systolic arrays: How to chose them*, IEE Proc. **139:3** (1992), 179-188.

[2] C. N. Zhang, J. H. weston, Y.-F. Yan, *Determining object functions in systolic array designs*, IEEE Teans. Very Large Scale Integration (VLSI) Systems **2:3** (1994), 357-360.

[3] T. I. Tokić, I. Ž. Milovanović, D. M. Randjelović, E. I. Milovanović, *Determining VLSI array size for one class of nested loop algorithms*, In: Advances in Computer and Information Sciences (U. Güdükbay, T. Dagar, A. Gürsay, E. Gelembe, eds.), IDS Press, 1998, 389-396.

[4] C. N. Zhang, T. M. Bachtiar, W. K. Chou, *Optimal fault-tolerant design approach for VLSI array processors*, IEE Proc. Comput. Digit. Tech. **144:1** (1987), 15-20.

[5] P.-Z. Lee, Z.-M. Kedem, *Mapping nested loop algorithms into multidimensional systolic arrays*, IEEE Trans. Parallel Distrib. Syst. **1:1** (1990), 64-76.

Faculty of Electronic Engineering
University of Niš, Beogradska 14, 18000 Niš, Serbia
igor@elfak.ni.ac.yu