SCIENTIFIC PUBLICATIONS OF THE STATE UNIVERSITY OF NOVI PAZAR
SER. A: APPL. MATH. INFORM. AND MECH. vol. 1, 1 (2009), 21-28.

Optimizing Space-Time Parameters of Hexagonal Systolic Arrays

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

Abstract:
In this paper we synthesize a family of hexagonal arrays, $SA(r)$, that implement matrix multiplication. We have observed that the execution time of hexagonal array, that has minimal number of processing elements for a given problem size, can be reduced if the number of PEs is increased. Since the execution time and the number of PEs are two most important performance measures of the systolic array, we take their product AT^2, $AT^2 = \Omega_r(n)T_{exe}^2$, to compare the arrays from this family. With respect to this performance measure the best array is obtained for $r = [n/2]$, where n is a dimension of square matrices while r indicates the extension, in terms of rows, of the array that has minimal number of processing elements for a given problem size.

Keywords: systolic arrays, space-time parameters,

1 Introduction

High-performance, special purpose computer systems are typically used to meet specific application requirements or to off-load computations that are especially taxing to general purpose computers. As hardware cost and size continue to drop and processing requirements become well-understood, in areas such as signal and image processing, more special-purpose systems are being constructed. A group of researchers headed by H. T. Kung, has introduced the systolic concept for parallel architectures in the period of 1978-1982. The major features of adopting systolic arrays (SA) for special purpose processing architectures are: simple and regular design, concurrency and communications and balancing computation with the I/O.

Computational tasks can be conceptually classified into two families: compute-bound computations and I/O-bound computations. For example, matrix multiplication represents compute-bound computations. On the other hand, adding two matrices is I/O-bound task. Speeding up a compute-bound computation, may often be accomplished in a relatively simple and inexpensive manner, that is by the systolic approach, without increasing I/O requirements.

Manuscript received March 12, 2009 ; accepted May 25,2009.

I.Ž. Milovanović, E. I. Milovanović, T. I. Tokić, M. K. Stojčev, N. M. Stojanović are with the Faculy of Electronic Engineering, Niš, Serbia.

To handle matrix multiplication, hexagonal systolic array has been proposed by Kung and Leiserson [1, 2]. This array has been designed in an add-hoc manner. Therefore, research efforts in this area were directed toward the development of a general methodology for mapping high-level computations into hardware structures. Many such methodologies have been proposed in the last decade [2]-[17]. Most are based on the concept of dependence vectors to order in time and space the index points representing the algorithm. The ordered index points are represented by nodes in a dependence graph. This graph can be projected along defined directions to obtain the target architecture. In the case of matrix multiplication, there are 19 allowable projection directions. Ten of them give planar SAs which can be classified into three classes according to the interconnection pattern between the processing elements (PE). Among them are hexagonal SAs, and among them the array proposed in [1]. The array proposed in [1] is especially attractive since it can be used for fault-tolerant matrix multiplication with minimal hardware overhead (see, for example [18]-[21]). The arrays obtained by the systematic methodologies are not always optimal with respect to particular space, time or space-time parameters. Therefore a considerable research effort was directed towards optimizing some of them (see, for example [8]-[12], [16], [21]-[23]).

In [9] (see also [10, 11, 12]) a methodology for designing planar SAs with optimal number of PEs for a given problem size, that implements matrix multiplication algorithm, was developed. Hexagonal array obtained by that methodology has

$$\Omega(n) = n^2 \tag{1}$$

processing elements, where n is a dimension of square matrices. For that number of PEs the execution time has been minimized, and is equal to

$$T_{exe} = \begin{cases} 2n-1, & \text{if } n\text{mod}3 \neq 0. \\ 2n, & \text{if } n\text{mod}3 = 0. \end{cases} \tag{2}$$

Hexagonal array synthesized in [11] has the same number of PEs and T_{exe} as the one from [9], but its geometric and chip area were optimized. This array will be taken as a reference array in this paper. Since the execution time and the number of PEs are two most important performance measures of the SA, we take their product AT^2, $AT^2 = \Omega(n)T_{exe}^2$, to compare the arrays. AT^2 measure of the reference array is

$$AT^2 = \begin{cases} n^2(2n-1)^2, & \text{if } n\text{mod}3 \neq 0 \\ n^2(2n)^2, & \text{if } n\text{mod}3 = 0 \end{cases} \tag{3}$$

and it is not a minimal possible. In this paper we will synthesize a family of hexagonal arrays, try to balance between number of PEs and execution time, and find out the array with minimal AT^2 measure.

2 Main result

A bidirectional 1D SA that implements matrix-vector multiplication with optimal number of PEs was designed in [27]. In [25] (see also [26]) it was observed that if the number of

PEs in the array obtained in [27] is increased for one, the execution time is decreased for one time unit. It was proved that this procedure can be applied consecutively $n-1$ times, where n is a dimension of square matrix. This fact has inspired us to find out if this idea can be applied on hexagonal array for matrix multiplication.

We start from the array obtained in [11]. Without deterioring generality, we will assume that $n \bmod 3 \neq 0$. Number of PEs and execution time of this array are given by (1) and (2), respectively. If we increase the number of PEs for n in the direction of C data flow, $C = A \times B$, (see Fig. 1), it can be concluded that the execution time is decreased for one time unit. Of course, we have to reschedule input elements A and B to preserve the correctness of matrix multiplication.

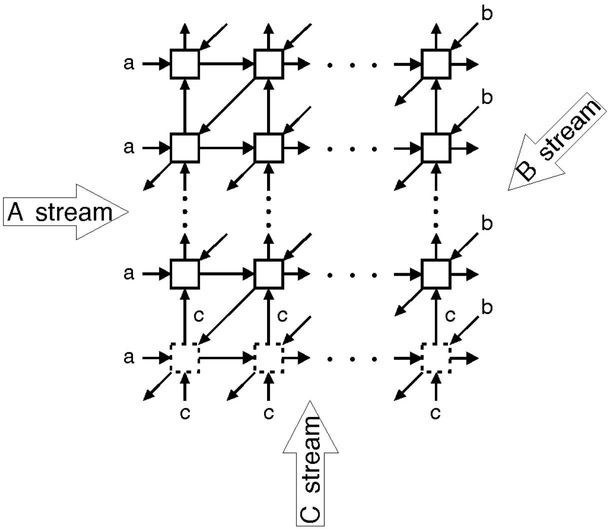

Fig. 1. Hexagonal SA expanded for n PEs in the direction of C data flow. Additional processing elements are denoted by dashed squares.

Actually, we have obtained hexagonal SA that has $\Omega(n) = n^2 + n$ PEs and $T_{exe}^{(1)} = 2n-2$. The question is whether this procedure can be applied consecutively up to the problem size n? Therefore we perform the following analysis. Suppose that we have expanded the reference array enough, such that first three rows in C stream begin with the computation simultaneously. Since the problem size is n, i.e. there are n inner products, the first row in C stream will be computed after n time units. Because of the SA topology and pipelining, in the $(n+1)$-th time unit, some elements from A and B streams would form partial product of some elements from the third row of C stream, which has already been computed, thus affecting the correctness of the computation. Therefore we conclude that the number of rows, s, in C stream that may begin with the computation simultaneously safely must be $s < 3$. Having this in mind we will derive the upper bound of the array extension.

Denote by r, $r \in N_o$, the number of rows, each of them containing n PEs, that have been added to the reference array in the direction of C stream, similarly as in Fig. 1. This means that now we have SA with $\Omega_p(n) = n^2 + rn$ PEs, and we expect to achieve the execution time of $T_{exe}^{(r)} = 2n - r - 1$ time units. Let us find out the upper bound for r. The lower bound is obviously $r = 0$, i.e. it is the unextended array from [11]. To achieve $T_{exe}^{(r)} = 2n - r - 1$ the deadline for the last row of C stream to begin with the computation is in the $(n - r - 1)$-th time instance. On the other hand, this requires that in the first time instance $s = 2r - n + 2$ rows in C stream begin with the computation. Thus, we obtain

$$s = \begin{cases} 1, & \text{if } 2r \leq n-1 \\ 2r - n + 2, & \text{if } 2r \geq n-1 \end{cases}. \tag{4}$$

According to the previously estimated condition $s < 3$, and (4), we conclude that the following must be satisfied

$$2r < n + 1,$$

i.e. the upper bound for r is

$$r = \left[\frac{n}{2}\right].$$

Thus we have proved the following result.

Theorem 1 *For each r,* $0 \leq r \leq [\frac{n}{2}]$ *there exists planar hexagonal systolic array with*

$$\Omega_r(n) = n(n + r)$$

processing elements, that implements matrix multiplication algorithm for time

$$T_{exe} = \begin{cases} 2n - r - 1, & \text{if } n \text{mod} 3 \neq 0 \\ 2n - r, & \text{if } n \text{mod} 3 = 0 \end{cases},$$

where n is a dimension of square matrices.

Corollary 1 *According to Theorem 1, for* $r = 0$ *the reference array from [11] is obtained.*

Remark 1 *Since the upper bound for r is* $[\frac{n}{2}]$*, it is not possible to design planar hexagonal array that performs matrix multiplication for time* $T_{exe} = n$*, i.e.* $T_{exe} = n + 1$*. This is quite unexpected having in mind the results from [26].*

According to the Theorem 1, a class of hexagonal arrays, $SA(r)$, can be synthesized. Let us find out which of these arrays has the minimal AT^2 measure. Therefore we consider the function

$$\varphi(r) = n(n + r)(2n - r - 1)^2, \quad 0 \leq r \leq \left[\frac{n}{2}\right]. \tag{5}$$

Without affecting the generality, we again consider the case $n\text{mod}3 \neq 0$. Graphic of the function $\varphi(r)$ is depicted in Fig. 2. $\varphi(r)$ is monotone decreasing and reaches its minimum for $r = [\frac{n}{2}]$. This implies that the following is valid.

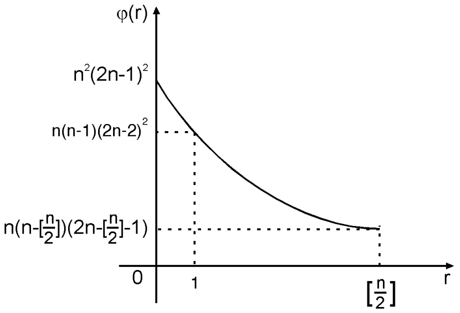

Fig. 2. Function $\varphi(r)$

Theorem 2 *Planar hexagonal systolic array with minimal AT^2 measure that implements matrix multiplication has the following features*

$$\Omega(n) = n(n + \left[\frac{n}{2}\right],$$
$$T_{exe} = \begin{cases} 2n - \left[\frac{n}{2}\right] - 1, & \text{if } n\text{mod}3 \neq 0 \\ 2n - \left[\frac{n}{2}\right], & \text{if } n\text{mod}3 = 0 \end{cases}.$$

Corollary 2 *It is not difficult to conclude that reference array from [11] has the worst AT^2 measure compared to the arrays obtained for $r = 1, 2, \ldots, [\frac{n}{2}]$.*

Data flow in hexagonal arrays for $n = 4$ and $r = 0, 1, 2$ are given in Figures 3, 4 and 5, respectively. The array in Fig. 5 is optimal with respect to AT^2 measure when $n = 4$.

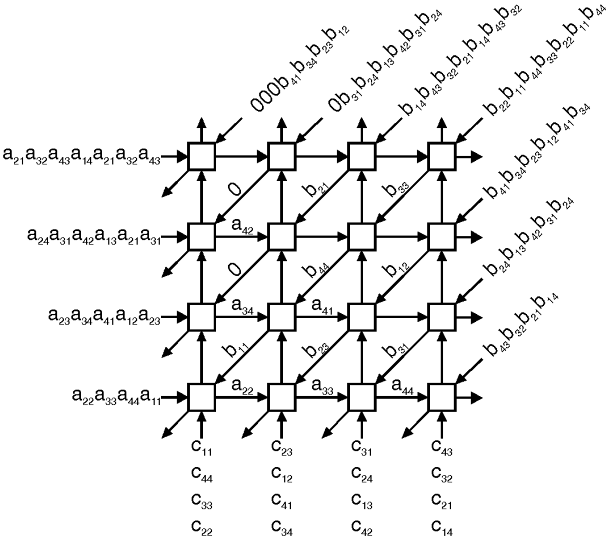

Fig. 3. Data flow in unextended array for $n = 4$

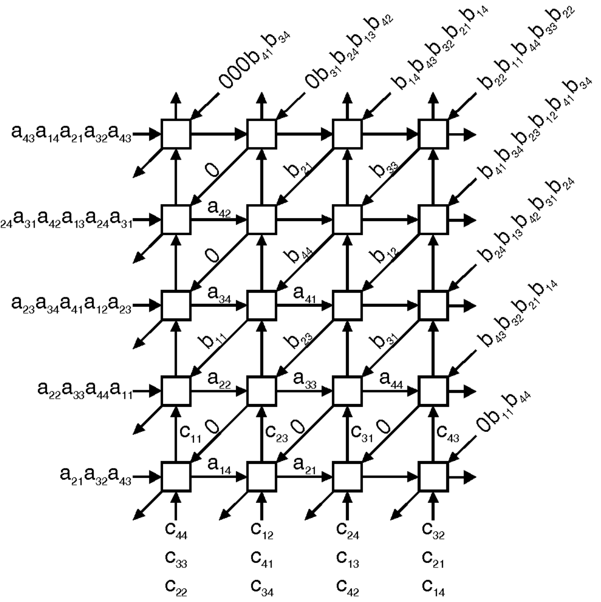
$000b_{41}b_{34}$
$0b_{31}b_{24}b_{13}b_{42}$
$b_{14}b_{43}b_{32}b_{21}b_{14}$
$b_{22}b_{11}b_{44}b_{33}b_{22}$
$b_{41}b_{34}b_{23}b_{12}b_{41}b_{34}$
$b_{24}b_{13}b_{42}b_{31}b_{24}$
$b_{43}b_{32}b_{21}b_{14}$
$0b_{11}b_{44}$
$a_{43}a_{14}a_{21}a_{32}a_{43}$
0
b_{21}
b_{33}
$a_{24}a_{31}a_{42}a_{13}a_{24}a_{31}$
a_{42}
0
b_{44}
b_{12}
$a_{23}a_{34}a_{41}a_{12}a_{23}$
a_{34}
a_{41}
b_{11}
b_{23}
b_{31}
$a_{22}a_{33}a_{44}a_{11}$
a_{22}
a_{33}
a_{44}
c_{11}
0
c_{23}
0
c_{31}
0
c_{43}
$a_{21}a_{32}a_{43}$
a_{14}
a_{21}
c_{44} c_{33} c_{22}
c_{12} c_{41} c_{34}
c_{24} c_{13} c_{42}
c_{32} c_{21} c_{14}

Fig. 4. Dataflow in the array expanded with one row ($r = 1$) of processing elements when $n = 4$

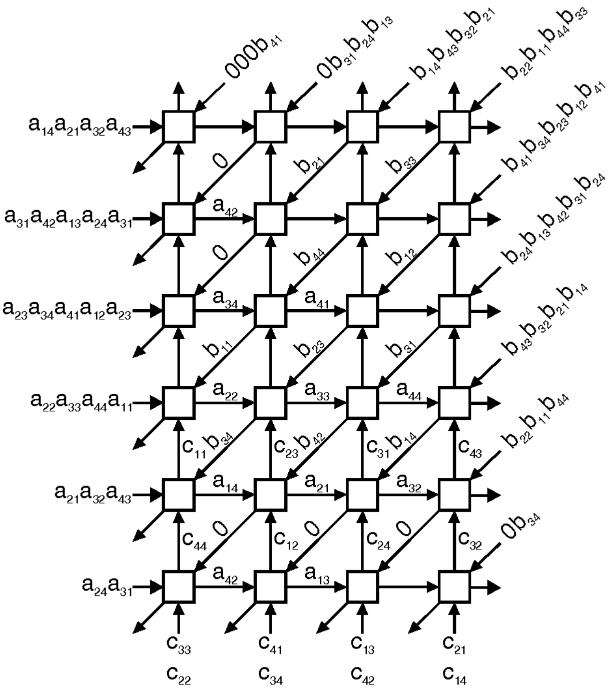

Fig. 5. Dataflow in the array expanded with two rows ($r = 2$) of processing elements when $n = 4$.

3 Conclusion

We have observed that the execution time of hexagonal array that implements matrix multiplication can be reduced if the number of PEs is increased. Therefore in this paper we have synthesized a family of systolic arrays, $SA(r)$. We have estimated the lower and upper bounds for r, i.e. $0 \leq r \leq [n/2]$. The number of PEs and execution time of matrix multiplication in the family $SA(r)$ are given in Theorem 1. Since the execution time and the number of PEs are two most important performance measures of the SA, we take their product AT^2, $AT^2 = \Omega_r(n)T_{exe}^2$, to compare the arrays from this family. With respect to this performance measure the best array is obtained for $r = [n/2]$.

References

[1] H. T. KUNG, C. E. LEISERSON, *Systolic arrays for VLSI, Introduction to VLSI systems*, (C. Mead, L.Conway, eds.), Addison-Wesley Ltd., Reading, MA, 1980.

[2] H. T. KUNG, *Why systolic architectures?*, Computer, 15 (1982), 37-46.

[3] M. CHEN, *A design methodology for synthesizing parallel algorithms and architectures*, J. Parallel Distributed Comput., (1986), 461-491.

[4] A. L. DECEGAMA, *Parallel processing architectures and VLSI hardware*, Prentice Hall, New Jersey, Vol. 1, 1989.

[5] J. A. FORTES, K. S. FU, B. W. WAH, *Systematic design approaches for algorithmically specified systolic arrays*, (V. Milutinović, ed.), Computer Architectures, Elsevier Ltd., New York, 1988, 454-494.

[6] H. V. JAGADISH, T. KAILATH, *A family of new efficient arrays for matrix multiplication*, IEEE Trans. Comput., 38 (1) (1989), 149-155.

[7] C. LANGAUER, *A view of systolic design*, In: Proc. International Conference Parallel Computing Technologies, (N. N. Mirenkov, ed.), Novosibirsk'91, World Scientific, Singapore, 1991, 32-46.

[8] G. -J. LI, B. W. WAH, *The design of optimal systolic arrays*, IEEE Trans. Comput., 34 (1) (1985), 66-77.

[9] I. Z. MILENTIJEVIĆ, I. Ž. MILOVANOVIĆ, E. I. MILOVANOVIĆ, M. K. STOJČEV, *The design of optimal planar systolic arrays for matrix multiplication*, Computers Math. Applic. 33, 6 (1997), 17-35.

[10] E. I. MILOVANOVIĆ, I. Z. MILENTIJEVIĆ, I. Ž. MILOVANOVIĆ, *Designing of processor–time optimal systolic array for matrix multiplication*, Comput. Artificial Intelligence, 16 (1) (1997), 1-11.

[11] M. P. BEKAKOS, E. I. MILOVANOVIĆ, I. Ž. MILOVANOVIĆ, I. Z. MILENTIJEVIĆ, *An efficient systolic array for matrix multiplication*, Proc. of the Fourth Hellenic European Conference on Computer Mathematics and its Applications (HERCMA'98), Athens'98, (E. A. Lipitakis, ed.), Vol.1 (1999), 298-317.

[12] E. I. MILOVANOVIĆ, G. V. MILOVANOVIĆ, I. Ž. MILOVANOVIĆ, D. MILOSAVLJEVIĆ, *Designing hexagonal systolic array by composite mappings*, Facta Univ. Ser. Math. Inform., Vol. 1 (1997), 1-11.

[13] D. I. MOLDOVAN, *On the design of algorithms for VLSI systolic arrays*, IEE Proc., 71 (1983), 113-120.

[14] W. SHANG, J. A. FORTES, *On time mapping of uniform dependence algorithms into lower dimensional processor arrays*, IEEE Trans. Parallel Distr. Syst., 3 (3) (1992), 350-363.

[15] S. G. SEDUKHIN, *The designing and analysing systolic algorithms and structures*, Programming, 2(1990), 20-40 (In Russian).

[16] S. G. SEDUKHIN, G. Z. KARAPETIAN, *Design of optimal systolic systems for matrix multiplication of different structures*, Report 85, Comput. Center Sibirian Division of USSR Academy of Science, Novosibirsk, 1990.

[17] C. R. WAN, D. J. EVANS, *Nineteen ways of systolic matrix multiplication*, Intern. J. Computer Math., Vol. 98 (1998), 39-69.

[18] M. O. ESONU, A. J. AL-KHALILI, S. HARIRI, D. AL-KHALILI, *Fault-tolerant design methodology for systolic array architecture*, IEE Proc. Comput. Digit. Tech., Vol. 141 (1) (1994), 17-28.

[19] C. N. ZHANG, T. M. BACHTIAR, W. K. CHOU, *Optimal fault-tolerant design approach for VLSI processors*, IEE Proc. Comput. Digit. Tech., Vol. 144 (1) (1997), 15-21.

[20] C. N. ZHANG, *Systematic design of systolic arrays for computing multiple time instances*, Microelectronic Journal, 23 (1992), 543-553.

[21] I. Ž. MILOVANOVIĆ, T. I. TOKIĆ, M. K. STOJČEV, E. I. MILOVANOVIĆ, N. M. NOVAKOVIĆ, *Mapping matrix multiplication algorithm onto optimal fault-tolerant systolic array*, Proc. 22nd Inter. Conf. Microelectronics (MIEL 2000), Niš'00, 711-714.

[22] M. GUŠEV, D. J. EVANS, *Nonlinear transformations of the matrix multiplication algorithm*, Inter. J. Comput. Math., Vol. 45 (1992), 1-21.

[23] D. J. EVANS, M. GUŠEV, *The magic of interlocking property: Fast systolic design*, Par. Algor. Appl., 10 (1997), 195-209.

[24] T. I. TOKIĆ, E. I. MILOVANOVIĆ, N. M. NOVAKOVIĆ, I. Ž. MILOVANOVIĆ, M. K. STOJČEV, *Matrix multiplication on non-planar systolic arrays*, Facta Univ. Ser. Electr. Energ., Vol. 13, 2 (2000), 157-165.

[25] O. B. EFREMIDIS, M. P. BEKAKOS, *A nonlinear approach to design processor-time optimal systolic arrays for matrix-vector multiplication*, Proc. of the Fourth Hellenic European Conference on Computer Mathematics and its Applications (HERCMA'98), Athens'98, (E. A. Lipitakis, ed.), Vol. 1 (1999), 327-345.

[26] N. NOVAKOVIĆ, E. MILOVANOVIĆ, M. STOJČEV, T. TOKIĆ, I. MILOVANOVIĆ, *Optimization of bidirectional systolic arrays for matrix-vector multiplication*, J. Electrotechn. Math., 4 (1999), 35-40.

[27] I. Ž. MILOVANOVIĆ, E. I. MILOVANOVIĆ, I. Z. MILENTIJEVIĆ, M. K. STOJČEV, *Designing of processor-time optimal systolic arrays for band matrix-vector multiplication*, Computers Math. Applic., Vol. 32, 2 (1996), 21-31.