ELSEVIER

Microelectronics Journal **29** (1998) 313–321
© 1998 Elsevier Science Limited
Printed in Great Britain. All rights reserved
0026-2692/98/$19.00

PII:S0026-2692(97)00043-8

Refereed paper

A quantitative analysis of wiring lengths in 2D and 3D VLSI

Aleksandar Milenkovic* and Veljko Milutinovic

Department of Computer Engineering, School of Electrical Engineering, University of Belgrade, POB 35-54, 11120 Belgrade, Serbia, Yugoslavia

Performance and cost of the widely used submicron 2D VLSI technology are primarily determined by interconnection delays and on-chip area. One of the possibilities for overcoming this problem is the use of the innovative 3D VLSI. In this structure, shortening of interconnection wires can be achieved, resulting in better performance and packing density. This analysis assumes an existing 3D channel routing methodology, based on the 2D channel routing methodology for standard-cell VLSI. The interconnection wire length in 3D and 2D structures is compared for several examples of systolic arrays. The experiments show that the average interconnection wire length in 3D structures is from 20 to 50% of the average interconnection wire length in 2D structures, depending on the number of active layers in 3D VLSI. © 1998 Elsevier Science Ltd. All rights reserved.

1. Introduction

As the silicon VLSI technology approaches its fundamental scaling limit of about $0.2\,\mu m$, the concept of three-dimensional (3D) integration has been proposed to enhance packing density and speed performance [1]. Figure 1 shows a typical structure of 3D chips. A 3D VLSI chip consists of a stack of active (silicon) layers made possible by the silicon on insulator technology. In this technology, devices are first fabricated on an active layer using a 2D process. A passive (silicon dioxide) layer is then grown on the top of the active layer and planarized by etching before another polysilicon layer is deposited on the top of the insulator and recrystallized by laser beams. The new silicon layer can then be used to fabricate devices using

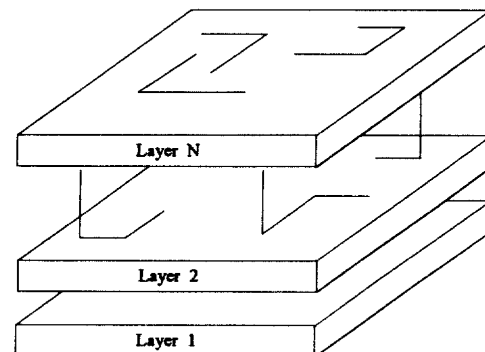


Fig. 1. A 3D VLSI chip structure.

*Author to whom correspondence should be addressed. E-mail: emilenka@etf.bg.ac.yu.

the 2D process. By repeating these steps, many active layers can be packed into one chip.

The concept of systolic arrays has been introduced as a solution for applications that require extensive throughput. A systolic system is based on a set of interconnected processing elements (PEs), each capable of performing some simple operation. In systolic arrays, information flows between cells in a pipelined fashion, and communication with the outside world occurs only at the boundary cells. Algorithms suitable for implementation can be found in many applications, such as digital signal and image processing, linear algebra, pattern recognition, linear and dynamic programming, and graph problems [2].

In this research, 3D structures of the systolic array for matrix multiplication are proposed and compared with corresponding 2D structures, considering the distribution of interconnection wire lengths.

The rest of the paper is organized as follows. Section 2 defines the problem statement. Section 3 summarizes the existing solutions—only the issues of interest for this research. Section 4 introduces the proposed solutions. Section 5 summarizes the conditions and assumptions of this research. Performance and complexity comparison results are given in Section 6. Section 7 concludes.

2. Problem statement

The main problem in this research is to compare 3D structures of systolic arrays and classical 2D structures, considering the distribution of interconnection wire lengths, average interconnection wire length and on-chip VLSI area. The concept of 3D systolic arrays has been introduced to solve some inherent limitations to the speed, extensibility and partitionability of 2D systolic arrays. A 3D systolic array can be implemented by 3D VLSI or by 3D packing of 2D VLSI. Some of the advantages of the 3D systolic arrays are due to the 3D architecture, and the others are due to the 3D structures. In this paper a quantitative analysis of the advantages due to 3D structure is considered.

This type of research is important since the interconnection delays and on-chip area primarily determine chip complexity and performance. Gate delays decrease with scaling, whereas interconnection delays remain constant. Consequently, the circuit speed becomes dominated by interconnection delays rather than device delays [3–5].

3. Existing solutions

There are many research papers on 3D VLSI; however, the authors are not aware of any paper which includes a quantitative analysis of wiring lengths. Although 3D chips have been fabricated in laboratories during the previous decade [6–8], no 3D computer-aided-design (CAD) is available to ease the complex design effort, and to make possible the comparison of different approaches to 3D VLSI. An analytical analysis of advantages and disadvantages of 3D systolic arrays has been shown in [9].

One approach to 3D routing for standard-cell VLSI design is proposed in [10]. The routing methodology is based on the 2D channel routing methodology; thus, it is called 3D channel routing methodology. In the proposed solution, the 3D routing problem is solved by decomposing it into two subproblems: (a) intra-layer routing; and (b) inter-layer routing. The intra-layer routing represents the interconnections of terminals on each active layer, and it is a 2D channel routing problem, while the inter-layer routing represents the interconnections of terminals on different active layers. Tong [10] has shown that the inter-layer routing can be transformed into a 2D channel routing problem. The terminals on the two channels of the intra-layer routing layers that have to be

connected are mapped onto the boundaries of the 2D channel on the inter-layer, in order to be interconnected.

4. Proposed solutions

The essence of the proposed research is to contribute to the domain of 3D VLSI interconnection wire length analysis. Consequently, the research includes several experiments based on systolic arrays for matrix multiplication. Each experiment includes classical 2D implementations and one or two 3D implementations. Two main placement policies are used: (a) each processing element or a group of processing elements is placed in a single active layer; and (b) each processing element is placed across all active layers.

In experiment 1 three implementations of the systolic array with four 4-bit processing elements are considered: (a) 2D implementation; (b) 3D implementation with two active layers; and (c) 3D implementation with four active layers. The 2D implementation of a systolic array optimized to perform matrix multiplication (**C**=**A**×**B**) and corresponding data streams are shown in Fig. 2. The 3D implementations with

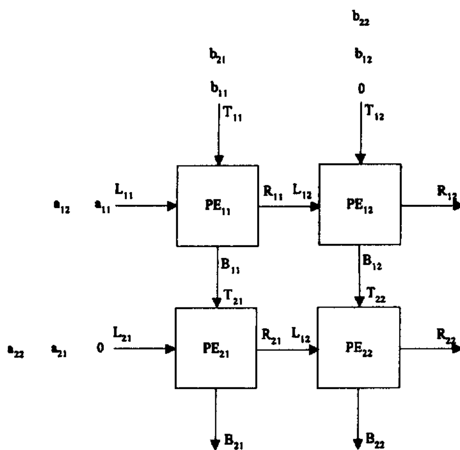


Fig. 2. The block diagram of the systolic array with four processing elements.

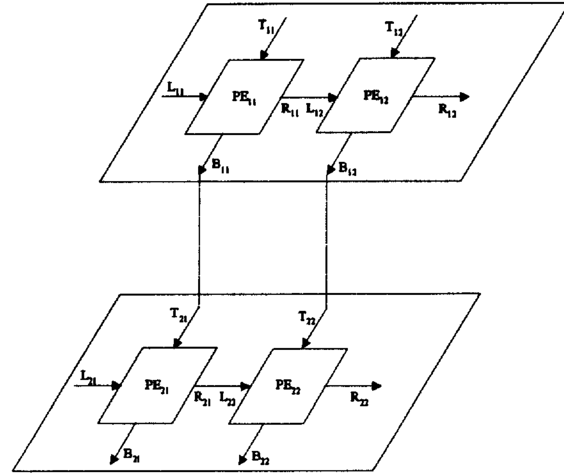


Fig. 3. The proposed placement of the systolic array in the

two and four active layers are shown in Figs 3 and 4, respectively. A block diagram of the processing element is given in Fig. 5. The processing element includes a multiplier (×), an adder (+), an accumulator register (REG_AC), a multiplexer (MUX), and two transfer registers (REG_B, REG_R). Processing element PE_{ij} performs an additive multiply operation which is described with the following formulae:

$$B_{ij} \leftarrow T_{ij},\ R_{ij} \leftarrow L_{ij},\ \mathrm{REG_AC} \leftarrow \mathrm{REG_AC} + T_{ij} \times L_{ij}$$

Experiment 2 is the same as experiment 1, except that 8-bit processing elements are used instead of 4-bit processing elements.

Experiment 3 follows the second placement policy. Two implementations are considered: (a) 2D implementation which is shown in Fig. 2; and (b) 3D implementation with two active layers, which is shown in Fig. 6. In this experiment, the processing element is organized in two pipeline stages; its block diagram is shown in Fig. 7. The first stage performs multiplication and the second stage performs addition. For 3D implementation, the first stage is placed in the

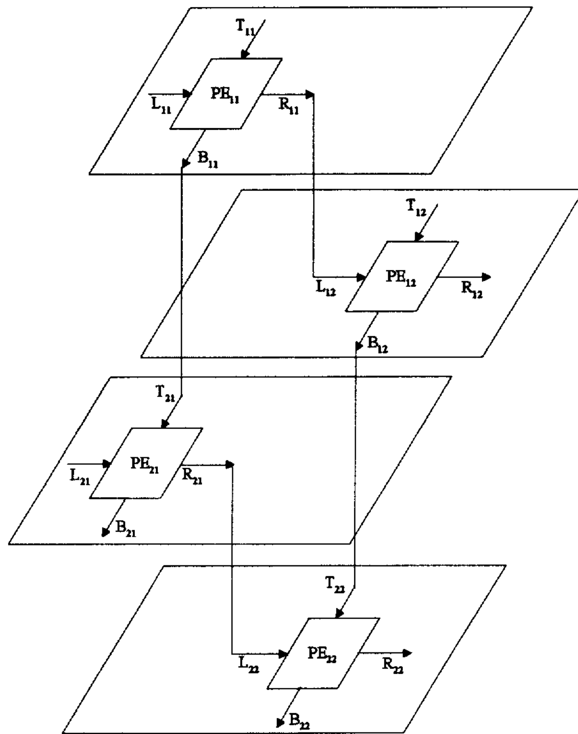


Fig. 4. The proposed placement of the systolic array in the 3D structure with four active layers.

first active layer, and the second pipeline stage is placed in the second active layer.

5. Conditions and assumptions

Generally, under the term 'conditions' we assume some characteristics of the real environment. Under the term 'assumptions' we assume simplifications which make the analysis possible and/or less complex without negative impacts on the generality of the results.

In our research, 2D VLSI chips and each active layer of 3D VLSI chips are designed using Tanner Research tools (NetTran, GateSim, L-Edit) [11–13] for 2D VLSI design, running on low-cost workstations. The Standard Cell Place and Route facility of the L-Edit is used for placement and routing. The output of the L-Edit program represents the file in CIF format that contains mask geometry information for each mask level.

The quantitative analysis of interconnection wire lengths is done using the originally developed program package called MARS [14]. Package MARS includes: (a) the program which extracts

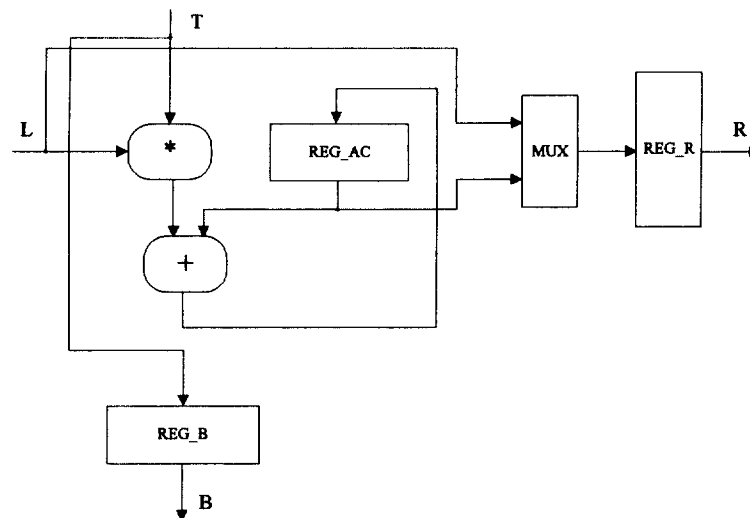


Fig. 5. The block diagram of an additive multiply processing element.

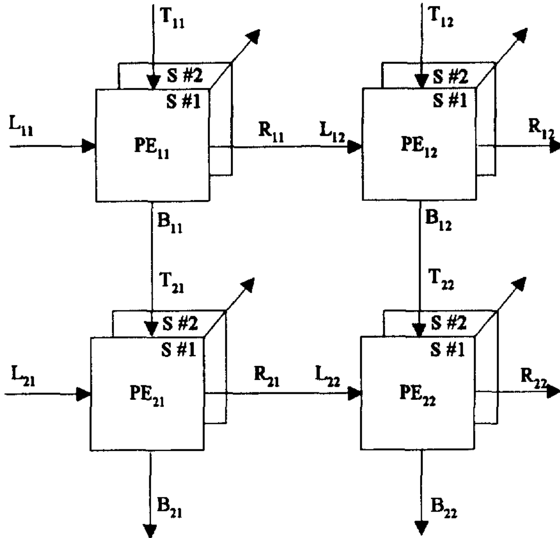


Fig. 6. The proposed placement in experiment 3.

geometric primitives relevant for interconnections (input/output terminals of standard cell and geometric primitives from interconnection layers); (b) the program which instances the CIF symbols, by applying defined transformations; and (c) the program which determines the sets of connected geometric primitives, extracts interconnections and calculates their lengths in CIF units.

It is assumed that a 3D interconnection wire length is between two values, **min** and **max**. **Min** is equal to the minimum of 2D interconnection wire lengths in 3D structures, and **max** is equal to the maximum of 2D interconnection lengths in 3D structures.

A 3D chip area is given as:

$$N \cdot \max_i(A_i), \ i = 1, 2, \ldots N$$

where N represents the number of active layers and A_i represents the on-chip area (only core without I/O pads) of the ith active layer.

6. Results

Distribution of the interconnection wire lengths, the on-chip area and the average interconnection wire length are measured for all experiments considered.

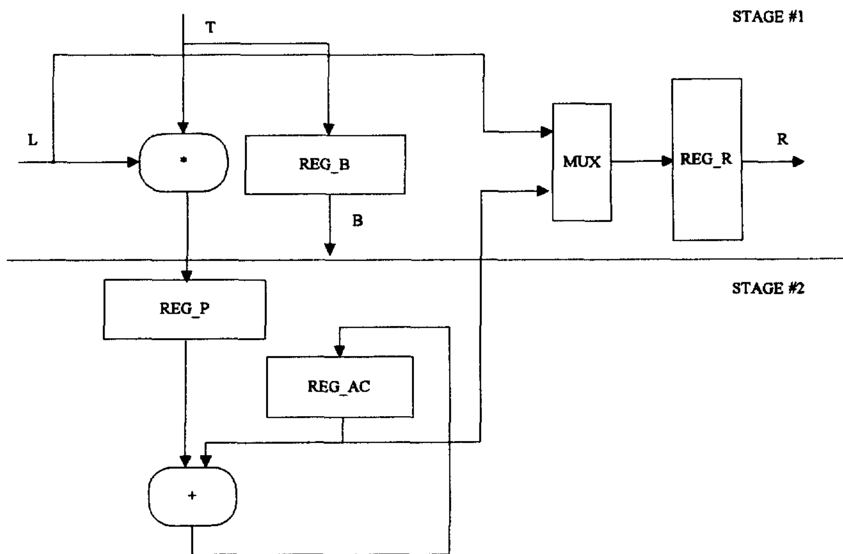


Fig. 7. Block diagram of the processing element with two pipeline stages.

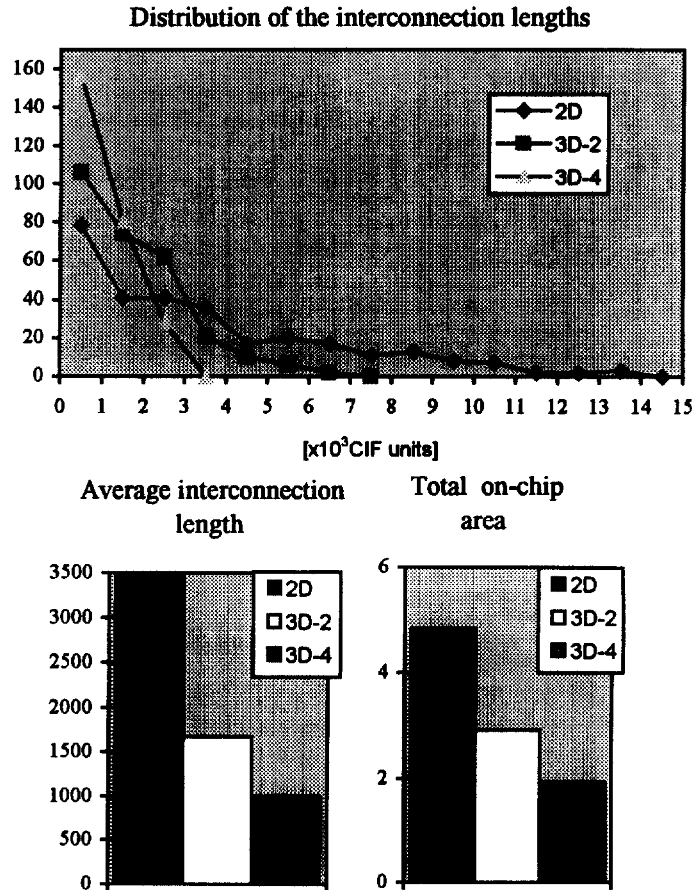


Fig. 8. The results of experiment 1.

The results of experiment 1 are given in Fig. 8. As a unit of measure for wiring length the CIF unit is used (CIF unit=0.5λ). The average wiring length for 3D-2 (3D with two active layers) implementation is 48% of the average wiring length for 2D implementation, and the average wiring length for 3D-4 implementation is 29% of the average wiring length for 2D implementation. Overall areas for 3D-2 and 3D-4 implementations are 60 and 40% of the area for 2D implementation, respectively.

The results of experiment 2 are given in Fig. 9. Average wiring length for 3D-2 and 3D-4 implementations are 43 and 18% of the average wiring length for 2D implementation, respectively. The overall on-chip areas for 3D-2 and 3D-4 implementations are 45 and 20% of the on-chip area for 2D implementation, respectively.

The results of experiment 3 are given in Fig. 10. The average wiring length for 3D implementation is 45% of the average wiring length for 2D implementation. The overall on-chip area for 3D-2 implementation is 60% of the total on-chip area needed for 2D implementation.

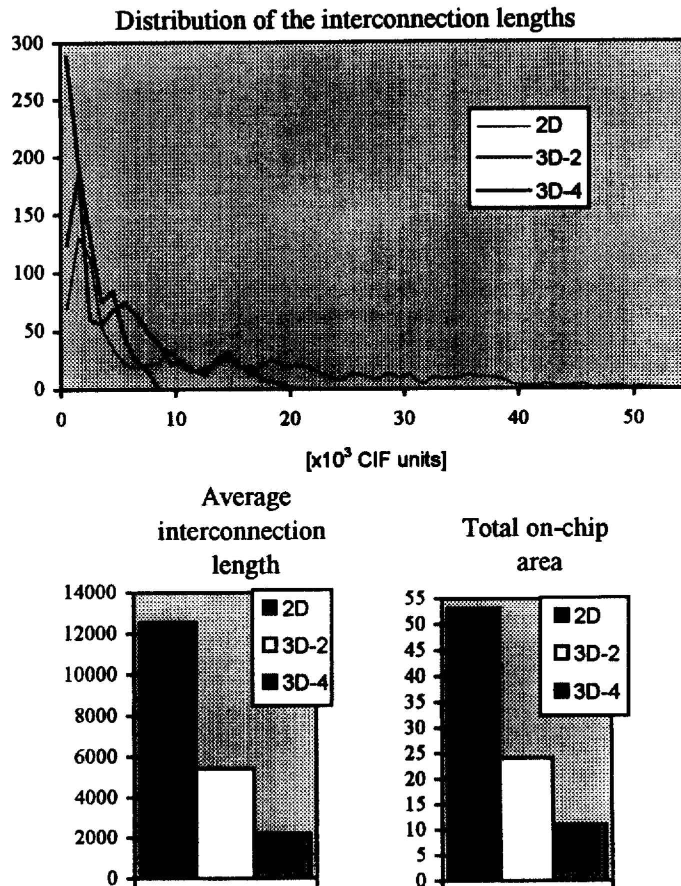


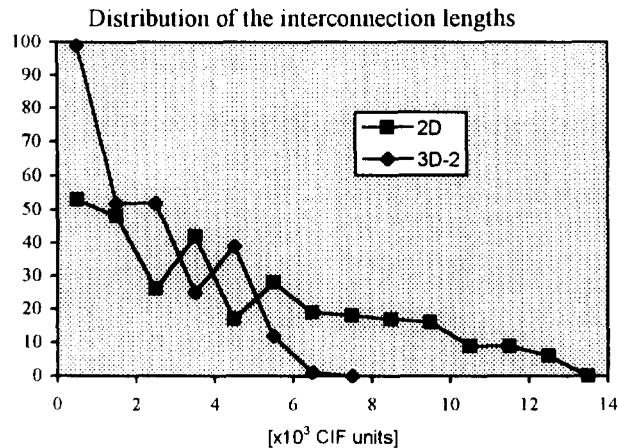
Fig. 9. The results of experiment 2.

For 3D implementations in two active layers the average interconnection length is between 43 and 48% of the average interconnection length for 2D implementations, while the overall on-chip area is between 45 and 60% of the on-chip area for 2D implementations; the same conclusion holds for all experiments conducted. For 3D implementations in four active layers, the average interconnection length is between 18 and 30% of the average interconnection length for 2D implementations, while the overall on-chip area is between 20 and 40% of the on-chip area for 2D implementations.

At first glance, it appears that the percentage of improvement depends on the number of interconnections in the circuit: with the growth of the number of interconnections, the improvement measured by the average interconnection length shortening and on-chip area reducing also increases.

7. Conclusion

A quantitative analysis of the interconnection lengths for 2D and 3D implementations of the systolic arrays is presented. The interconnection wire lengths are compared using existing tools for 2D VLSI and the originally developed program package MARS, which performs



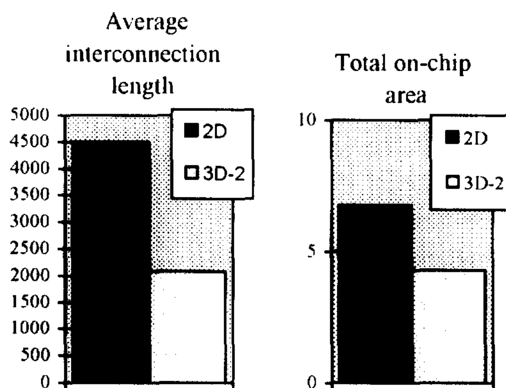


Fig. 10. The results of experiment 3.

analysis of the interconnection wire lengths. Parameters of the comparison are the distribution of the interconnection wire lengths, the average interconnection wire length and the total on-chip area. These parameters can be used as inputs for a detailed post-layout timing analysis.

The results of the experiments show a considerable improvement in interconnection length shortening and on-chip area reduction. Interconnection length shortening and on-chip area reduction lead to higher performance and lower complexity. We believe that progress in 3D technologies (3D VLSI and 3D packing of 2D VLSI) will make the 3D systolic arrays more competitive in terms of cost. To understand fully the advantages of 3D systolic arrays, further research work is needed in the area of 3D architecture and 3D technologies. One area that needs to be investigated further is the development of efficient algorithms for 3D routing and placement methodology in 3D structures.

Acknowledgements

The authors are grateful to Jelica Protic for her useful comments on earlier drafts of this paper, and to Leon Alkalaj of Jet Propulsion Labs for

his help and suggestions during the embryonic stages of this research.

References

[1] Kokubu, A. Three dimensional device project, *Preconference Tutorial of the 13th Int. Symp. on Computer Architecture*, ACM Press, Tokyo, Japan, June 1986.

[2] Fortes, J. and Wah, B. Systolic arrays—from concept to implementation, *IEEE Comput.*, 205 (1987) 12–17.

[3] Mead, C. and Conway, L., *Introduction to VLSI Systems*, Addison-Wesley, Reading, MA, 1980.

[4] Weste, N. and Kamran Eshraghian, *Principles of CMOS VLSI Design*, Addison-Wesley, Reading, MA, 1985.

[5] Kung, S.Y. *VLSI Array Microprocessors*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[6] Grinberg, J., Nudd, G. and Etchells, D. A cellular VLSI architecture, *IEEE Comput.*, 171 (1984) 69–80.

[7] Malhi, S.D.S., Davis, H.E., Stierman, R.J., Bean, K.E., Driscoll, C.C. and Chatterjee, P.K. Orthogonal chip mount—a 3D hybrid wafer scale integration technology, *Technical Digest, Int. Electron Devices Meeting*, 1987, pp. 104–106.

[8] Akasaka, Y., Nishimura, T. and Nakata, H. Trends in three-dimensional integration, *Solid State Technol.*, 28 (1988) 81–89.

[9] Lakhani, S., Wang, Y., Milenkovic, A. and Milutinovic, V. 2D matrix multiplication on a 3D systolic array, *Microelectronics J.*, 271 (1996) 11–22.

[10] Tong, C.C. and Wu, C. Routing in a three-dimensional chip, *IEEE Trans. Comput.*, 441 (1995) 106–117.

[11] *SchemLib—A Shematic Library Condensed Manual*, Tanner Research, Pasadena, CA, 1987.

[12] *GateSim—A Gate Level Simulator Condensed Manual*, Tanner Research, Pasadena, CA, 1987.

[13] *L-Edit User's Manual*, Tanner Research, Pasadena, CA, 1990.

[14] Milenkovic, A. A quantitative analysis of wiring lengths in 2D and 3D VLSI implementation of 2D systolic arrays, Belgrade University Master Thesis, 1997.