INTERNATIONAL JOURNAL OF COMPUTERS
DOI: 10.46300/9108.2021.15.1 Volume 15, 2021

Latency Analysis in the 2-Dimensional Systolic Arrays for Matrix Multiplication

Halil Snopce
CST Faculty
SEE-University Tetovo,
North Macedonia
h.snopce@seeu.edu.mk

Azir Aliu
CST Faculty
SEE-University Tetovo,
North Macedonia
Azir.aliu@seeu.edu.mk

Received: January 25, 2021. Revised: March 9, 2021. Accepted: March 15, 2021. Published: March 22, 2021.

***Abstract*—This paper deals with the latency analysis in a two-dimensional systolic array for matrix multiplication. The latency for all possible connection schemes is discussed. In this way there is obtained the lower bound of the latency that can be achieved using such arrays.**

Keywords—Systolic arrays, data availability, matrix multiplication, fan-out, connection schemes, lower bound of latency.

I. Introduction

A systolic array is a computing network possessing with which a high parallelism can be achieved. These arrays are appropriate to solve problems using nested-loop algorithms; such is the problem of parallel solution of matrix-matrix multiplication. The most known properties of the systolic arrays are locality, pipeline-ability and regularity. More about systolic arrays see [7, 8, 9, 10].

In the literature there are known some systolic array designs for matrix multiplication with fixed input/output (I/O) bandwidth which is $2n$ (n is the size of a matrix), and with the latency which differs and has different values depending of the design. In [1] it is proposed a hexagonal array with the latency $3n$. Another array with the same latency is proposed in [2]. The optimization of this result with the new latency /equal with $2n$ is given in [3]. The array with the latency $5n\ 2$ is proposed /in [5]. Also in [6] is given an array with latency $3n\ 2$. Optimization of some methods using regular iterative algorithms is proposed in [4].The question which arises here is about the lower bound of the latency in designing systolic arrays. Intuitively that bound would be equal to n.

II. Matrix Multiplication on Systolic Arrays

Given two matrices A and B of type $n \times n$, it is a computational problem to find their product. If we denote the product of A and B by C, then the entries of C can be calculated by the following formula:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \qquad i, j = 1, \ldots, n$$

In fact, this example is the simplest case of parallel matrix multiplication because the matrices are taken to be quadratic. The systolic array is quadratic 2-D such that the elements of matrices A and B have to be fed appropriately into the Processors of the array (PEs). The array consists of $n \times n$ PEs. Each PE indexed by i and j, in each step k, adds a partial product $a_{ik} b_{kj}$ to the accumulated sum for every c_{ij}. In the figure 1 is presented (just the first two cycles of total 7 cycles) the systolic matrix multiplication of two quadratic matrices of order 3 where the elements of a matrix A are fed into the array by the rows and the elements of the matrix B are fed into the array by the columns. Hence, we have a movement in two directions. The elements of the resulting matrix C are stationary.

III. Definition of the Latency

Definition: The latency for the matrix product $A \times B$ is the time between the first entry of the elements of the matrices A and B, until the last element of the product is calculated.

In fig. 2 it is given the graphical representation of the latency (denoted by L), which in fact is the time needed for data to move along the critical path.

From fig. 2 it can be concluded that the latency can be divided into three parts and the total latency can be expressed as:

$$L = L_{de} + L_{dd} + L_{dp} \qquad (1)$$

where L_{de} is called the data entry time and it is the time between the entry of the first element and the entry of the last element of the critical path. The second part, L_{dd}, is called the data delay time and it is equal with the number of steps that the data delay before entering in the first PE of the array. And L_{dp} is called the data processing time and it is the time required for moving and processing of each data from each PE in the critical path.

IV. The Lower Bound of Latency

As we know, systolic arrays have a high degree of regularity and locality. Regularity means that the repetitiveness of the interconnections of just one or few PEs, it makes possible to draw the whole array. On the other hand, locality is both a space and time feature, and it means that each PE can only interact with its nearest surrounding neighbors, and any transaction from one PE to the next PE is completed in only one unit time delay.

Taking into the consideration these two features of systolic arrays, we can give a methodology for finding the lower bound of the latency ($L_{\min}$) for two-dimensional systolic arrays. Since the I/O bandwidth is fixed and it is equal to $2n$, and on the other hand the total number of the input elements is $2n^2$ (two matrices by n^2 elements), we get that the number of steps required for the data entry is $2n^2/2n = n$. Because in the first step (when the element is transferred to the systolic array) the element is immediately entered into the first PE, we have

$$L_{de} = n - 1.$$

Furthermore, because we want to calculate the minimal latency L_{dd} can be taken to be zero. Hence

$$L_{\min} = (n-1) + L_{dp}. \qquad (2)$$

Because of the features of regularity and locality, there are four different connection schemes for systolic arrays. The connection schemes depend on the so called fan-out (the number of inputs that can be connected to an output) of PEs. The fan-out can take values 1, 2, 4 and 8. It cannot take value greater than 8 because of the feature of space-locality. In figure 3 are given models of these four connections. We are also analyzing the case of triangular array with fan-out=3.

The simplest case is when the fan-out is 1, and it is presented in the figure 4.

In this case each element of the matrices A and B has to travel through n PEs, and there are necessarily n time cycles. Thus, $L_{dp} = n$ and finally:

$$L_{\min} = (n-1) + n = 2n - 1 \qquad (3)$$

In the case when the fan-out is equal to 2, we present two different schemes. The first one is presented in figure 5.

In this case, in each time cycle, the elements of the matrix A (respectively B) are distributed into 2 new PEs, and after k steps the total number of PEs through which the elements are passed is:

$$\overset{t=1}{1} + \overset{t=2}{2} + \overset{t=3}{2} + \ldots + \overset{t=k}{2} = 1 + 2 \cdot (k-1)$$

The elements of A (respectively B) must take part in n PEs. Therefore we have an additional condition for the value of k, which is:

$$1 + 2(k-1) \geq n \Rightarrow k \geq \frac{n+1}{2} \qquad (4)$$

From the inequality (4), we have that $L_{dp} \geq n/2 + 1/2$, and therefore:

$$L_{\min} = \left\lfloor (n-1) + \frac{n}{2} + \frac{1}{2} \right\rfloor = \left\lfloor \frac{3n-1}{2} \right\rfloor \qquad (5)$$

The second scheme with the fan-out 2 is presented in fig. 6. In this case, the number of PEs in which the elements of the matrix A (respectively B) pass in each time cycle is given by the formula:

$$\overset{t=1}{1} + \overset{t=2}{2} + \overset{t=3}{3} + \ldots + \overset{t=k}{k} = \frac{k(k+1)}{2}$$

Similarly as in previous case, we can put the condition $k(k+1)/2 \geq n$. After solving the quadratic inequality (taking only the solution with positive value) we have

$$k \geq \frac{\sqrt{1+8n} - 1}{2} \Rightarrow L_{dp} \geq \frac{\sqrt{1+8n} - 1}{2}$$

so,

$$L_{\min} = \left\lfloor n - 1 + \frac{\sqrt{1+8n} - 1}{2} \right\rfloor = \left\lfloor \frac{2n + \sqrt{1+8n} - 3}{2} \right\rfloor \qquad (6)$$

There are some other schemes with the fan-out being equal to 2 as well, but generally the value of $L_{\min}$ is the same like in the last obtained result.

In the case when the fan-out is equal to 3, we present the triangular systolic array given in the figure 7.

Similarly as in the second case with fan-out 2 we have the following distribution of PEs in which the elements of the matrix A (respectively B) pass in each time cycle:

$$\overset{t=1}{1} + \overset{t=2}{2} + \overset{t=3}{3} + \ldots + \overset{t=k}{k} = \frac{k(k+1)}{2}$$

We can conclude that it is the same as in the case of second option with fan-out 2, therefore the minimum latency will have the same value.

In the same manner like in the case with the fan-out=2, there are different schemes for the case with the fan-out=4. In our analysis it is considered the case given in figure 8, which is more appropriate design for the fan-out=4.

According to fig. 8, the number of PEs versus t will be:

$$\overset{t=1}{1}+\overset{t=2}{4}+\overset{t=3}{8}+\ldots+4(\overset{t=k}{k}-1)=1+4\left(1+2+\ldots+(k-1)\right)=$$
$$=1+4\cdot\frac{k(k-1)}{2}=1+2k(k-1)$$

From the condition $1+2k(k-1)\geq n$ (finding the positive solution of the quadratic inequality) we have that $k\geq\left(1+\sqrt{2n-1}\right)/2\Rightarrow L_{dp}\geq\left(1+\sqrt{2n-1}\right)/2$ and therefore:

$$L_{\min}=\left\lfloor n-1+\frac{1+\sqrt{2n-1}}{2}\right\rfloor=\left\lfloor\frac{2n+\sqrt{2n-1}-1}{2}\right\rfloor \qquad (7)$$

The last case is the array with the fan-out=8. The graphical representation of this case is given in figure 9.

From fig.9 we can conclude that the number of PEs visited each time by the elements of the matrix A (respectively B) is given by:

$$\overset{t=1}{1}+\overset{t=2}{8}+\overset{t=3}{16}+\ldots+8(\overset{t=k}{k}-1)=1+8\left(1+2+\ldots+(k-1)\right)=$$
$$=1+8\cdot\frac{k(k-1)}{2}=1+4k(k-1)$$

n PEs are reached when the obtained value is greater then n. Hence,

$$1+4k(k-1)\geq n\Rightarrow 4k^2-4k-n+1\geq 0\Rightarrow k\geq\frac{1+\sqrt{n}}{2}$$
$$\Rightarrow L_{dp}\geq\frac{1+\sqrt{n}}{2}$$
$$\Rightarrow L_{\min}=\left\lfloor n-1+\frac{1+\sqrt{n}}{2}\right\rfloor=\left\lfloor\frac{2n+\sqrt{n}-1}{2}\right\rfloor \qquad (8)$$

From the analysis done with all kinds of fan-outs, we can conclude that with systolic arrays it is impossible to achieve latency equal to n with bandwidth $2n$. The lower bound for the latency approximately is equal to $n+\sqrt{n}/2$, and this is obtained for the fan-out=8. So, the latency decreases when the fan-out increases. The only possibility to have the latency to be equal to n, is to achieve the result $L_{dp}=1$, which is impossible to be achieved in the case of systolic arrays for matrix multiplication.

REFERENCES

[1] S.Y. Kung, "VLSI Array Processor for Signal Processing," In Proc. Conf. Advanced Res. Integrated Circuits, 1980.

[2] G.J. Li and B.W. Wah, "The design of Optimal Systolic Arrays," IEEE Trans. On Computers, vol.C-34, pp. 66-77, 1985.

[3] A.K. Oudjida, S.Titri, M. Hamerlain, "Mapping Full Systolic Arrays for Matrix Product on XILINX's XC4000(E,EX) FPGAs", COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering, Vol.21 Iss:1, pp.69-81, 2002.

[4] J.C. Tsay and P.Y. Chang, "Design of Efficient Regular Arrays for Matrix Multiplication by Two step Regularization," IEEE tran. On Parallel and Distributed computing, 1995.

[5] H.V. Jagadish and Kailath, " A family of a new efficient arrays for matrix multiplication", IEEE Trans. On Computers, vol. 38, pp 149-155, January 1989.

[6] Benaini A. and Robert Y., " An even faster array for matrix multiplication", Parallel computing, vol. 12, pp 249-254, 1989.

[7] Snopce, H., Elmazi, L., Reducing the number of processors elements in systolic arrays for matrix multiplication using linear transformation matrix, Int. J. of Computers, Communications and Control, Vol. III (2008), Suppl. issue: Proceedings of ICCCC 2008, pp. 486-490.

[8] Snopce, H., Elmazi, L., the importance of using linear transformation matrix in determining the number of PEs in systolic arrays, proceedings of ITI 2008, p.p 885-892, Cavtat, Croatia.

[9] Snopce, H., Spahiu, I., some characteristics of systolic arrays, World Academy of Science, Engineering and Technology, Issue 64, April 2010, WASET 2010, Rome, Italy, pp. 245-251.

[10] M.P. Bekakos, Highly Parallel Computations-Algorithms and Applications, Democritus University of Thrace, Greece, pp. 139-209, 2001.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

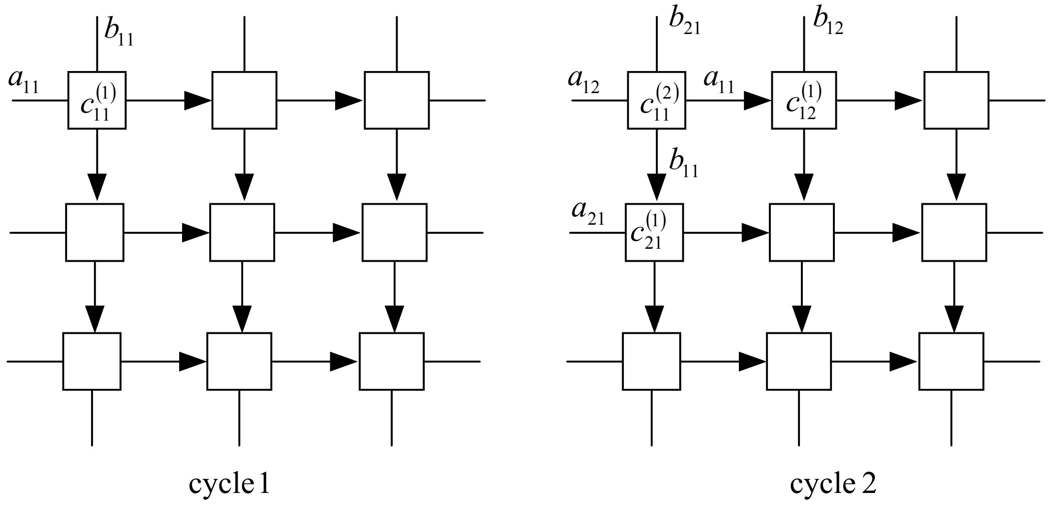


Fig. 1 first two cycles of matrix multiplication on systolic array

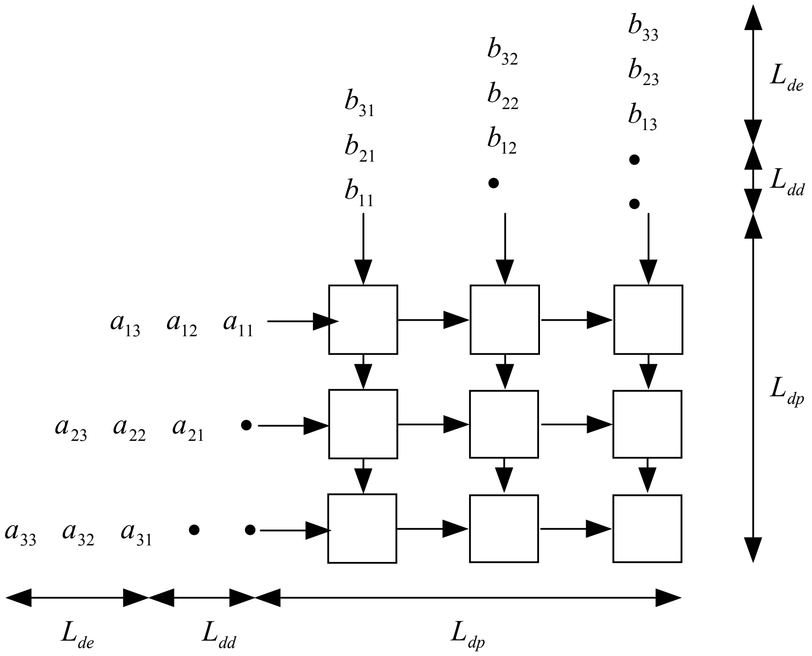
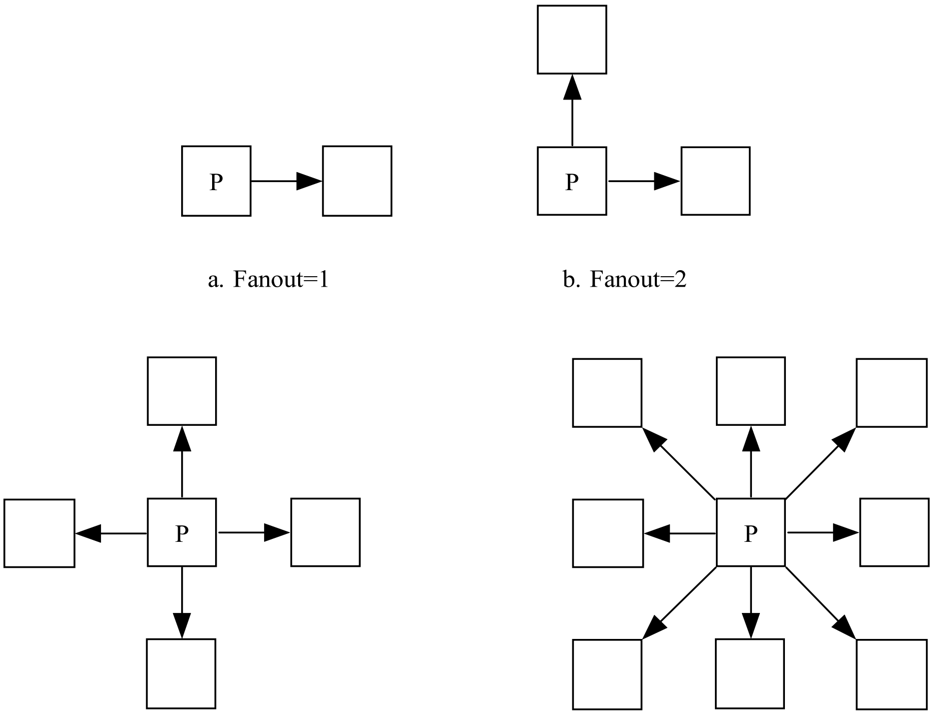


Fig. 2 Graphical representation



a. Fanout=1

b. Fanout=2

c. Fanout=4

d. Fanout=8

Fig.3 Connection schemes of systolic arrays

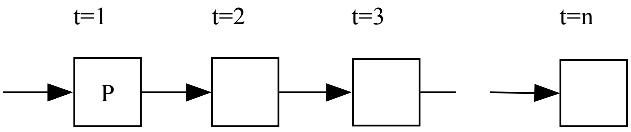


Fig.4 Data availability when fan-out=1

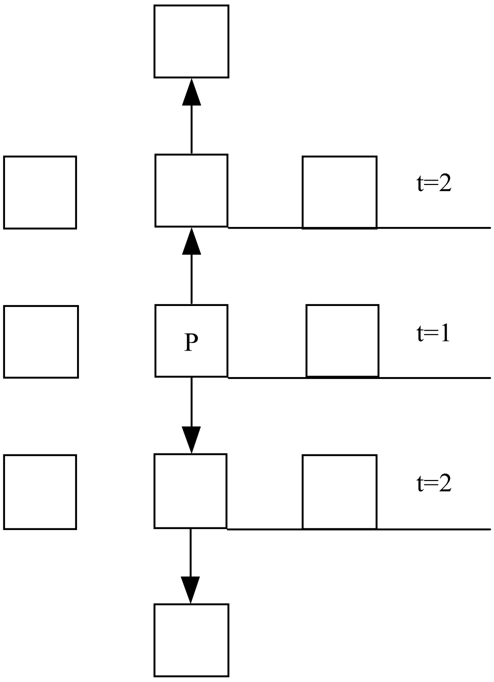


Fig. 5 Data availability when fan-out=2 (first option)

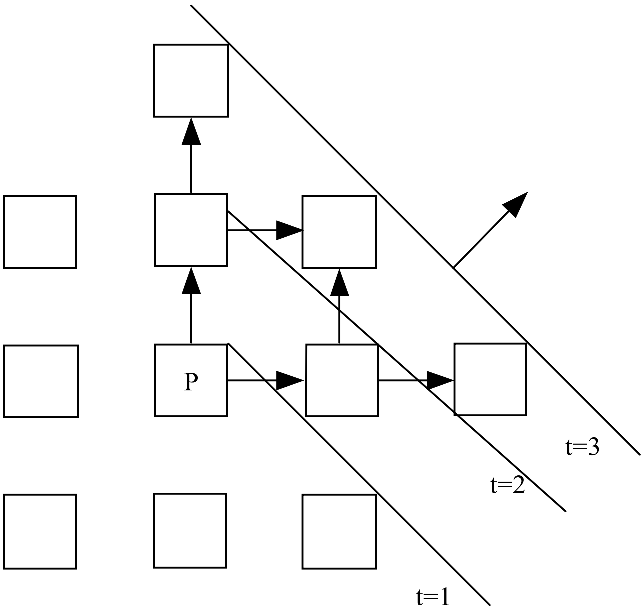


Fig. 6 Data availability when fan-out=2 (second option)

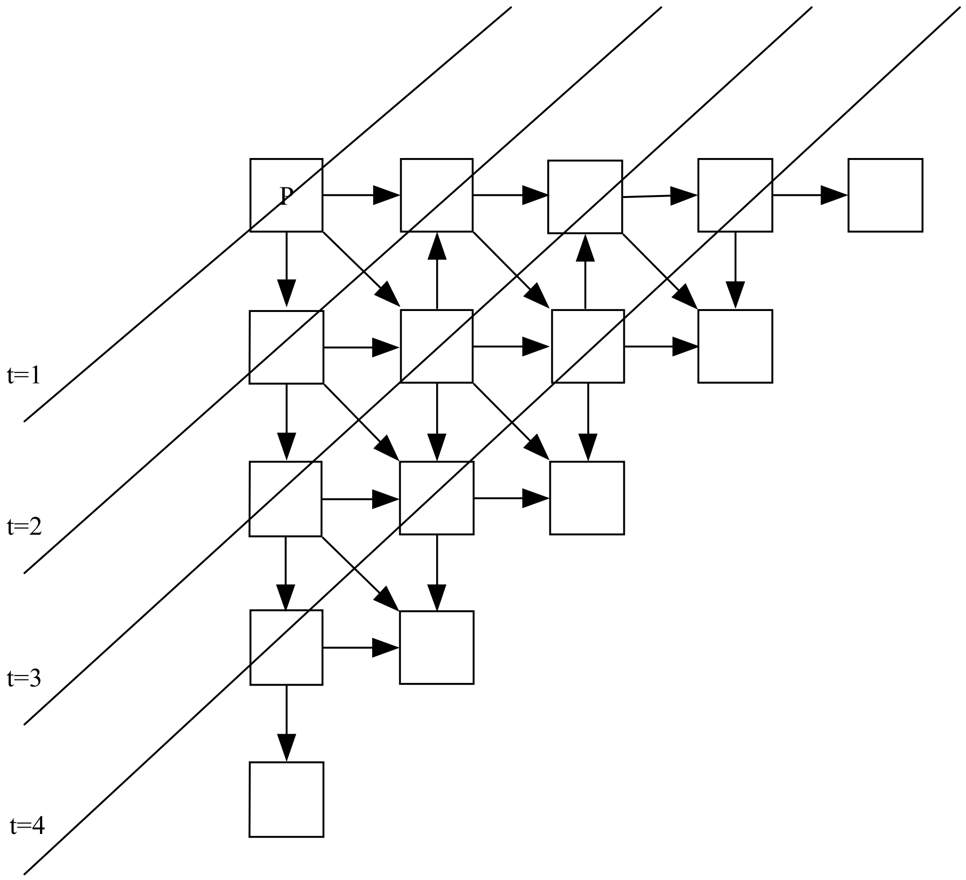


Fig. 7 Data availability when fan-out=3 (triangular array)

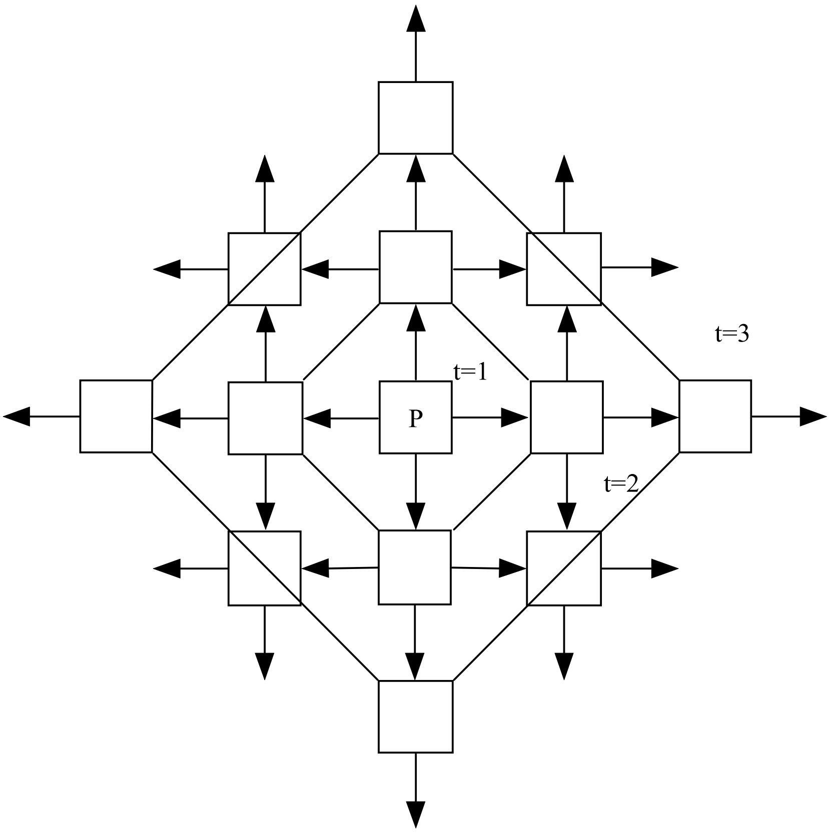


Fig. 8 Data availability when fan-out=4

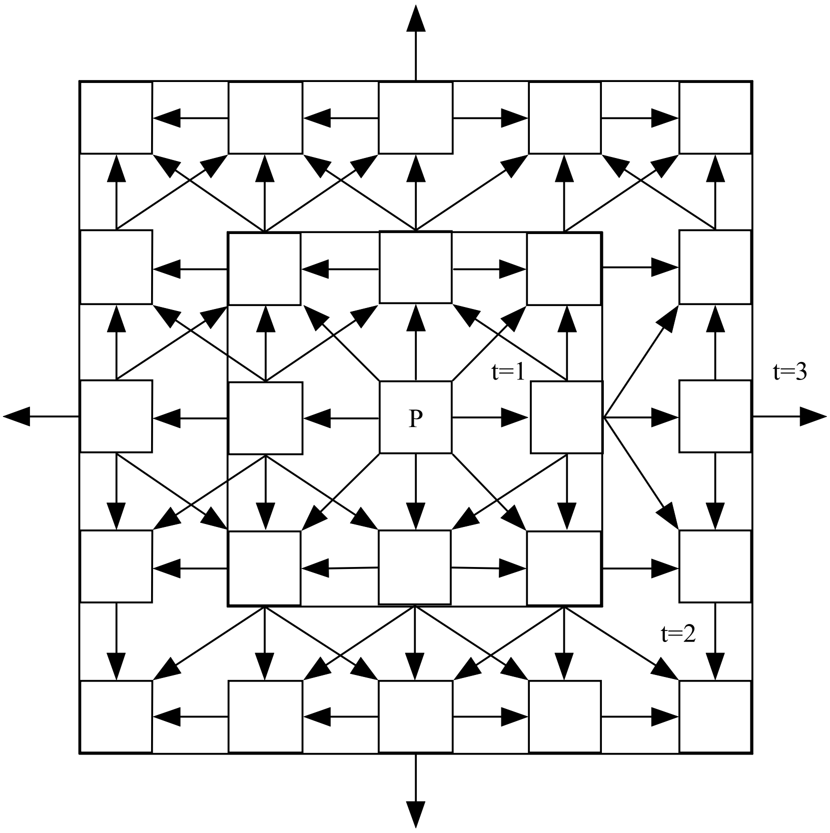


Fig. 9 Data availability when fan-out=8