Multiprocessor for Matrix Multiplication

-using Systolic Arrays

JIAN CHEN

31008658

Matrix Multiplication

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ b_{41} & b_{42} & b_{43} & b_{44} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & c_{34} \\ c_{41} & c_{42} & c_{43} & c_{44} \end{bmatrix}$$

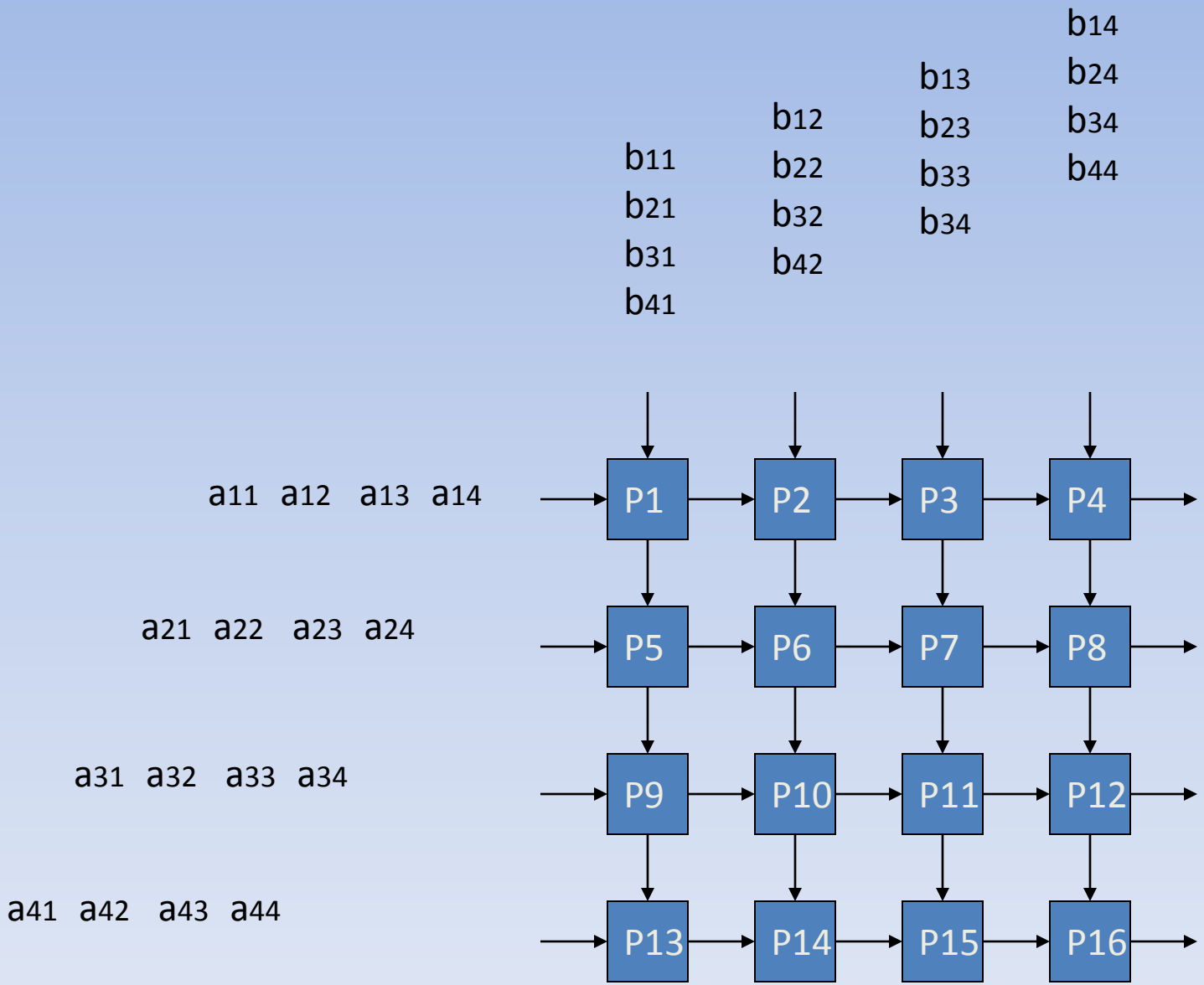
$$c_{11} = a_{11} \bullet b_{11} + a_{12} \bullet b_{21} + a_{13} \bullet b_{31} + a_{14} \bullet b_{41}$$
$$c_{12} = a_{11} \bullet b_{12} + a_{12} \bullet b_{22} + a_{13} \bullet b_{32} + a_{14} \bullet b_{42}$$
$$c_{13} = a_{11} \bullet b_{13} + a_{12} \bullet b_{23} + a_{13} \bullet b_{33} + a_{14} \bullet b_{43}$$
$$c_{14} = a_{11} \bullet b_{14} + a_{12} \bullet b_{24} + a_{13} \bullet b_{34} + a_{14} \bullet b_{44}$$
$$c_{21} = a_{21} \bullet b_{11} + a_{22} \bullet b_{21} + a_{23} \bullet b_{31} + a_{24} \bullet b_{41}$$
$$c_{22} = a_{21} \bullet b_{12} + a_{22} \bullet b_{22} + a_{23} \bullet b_{32} + a_{24} \bullet b_{42}$$
$$c_{23} = a_{21} \bullet b_{13} + a_{22} \bullet b_{23} + a_{23} \bullet b_{33} + a_{24} \bullet b_{43}$$
$$c_{24} = a_{21} \bullet b_{14} + a_{22} \bullet b_{24} + a_{23} \bullet b_{34} + a_{24} \bullet b_{44}$$
$$c_{31} = a_{31} \bullet b_{11} + a_{32} \bullet b_{21} + a_{33} \bullet b_{31} + a_{34} \bullet b_{41}$$

What Is a Systolic Array?

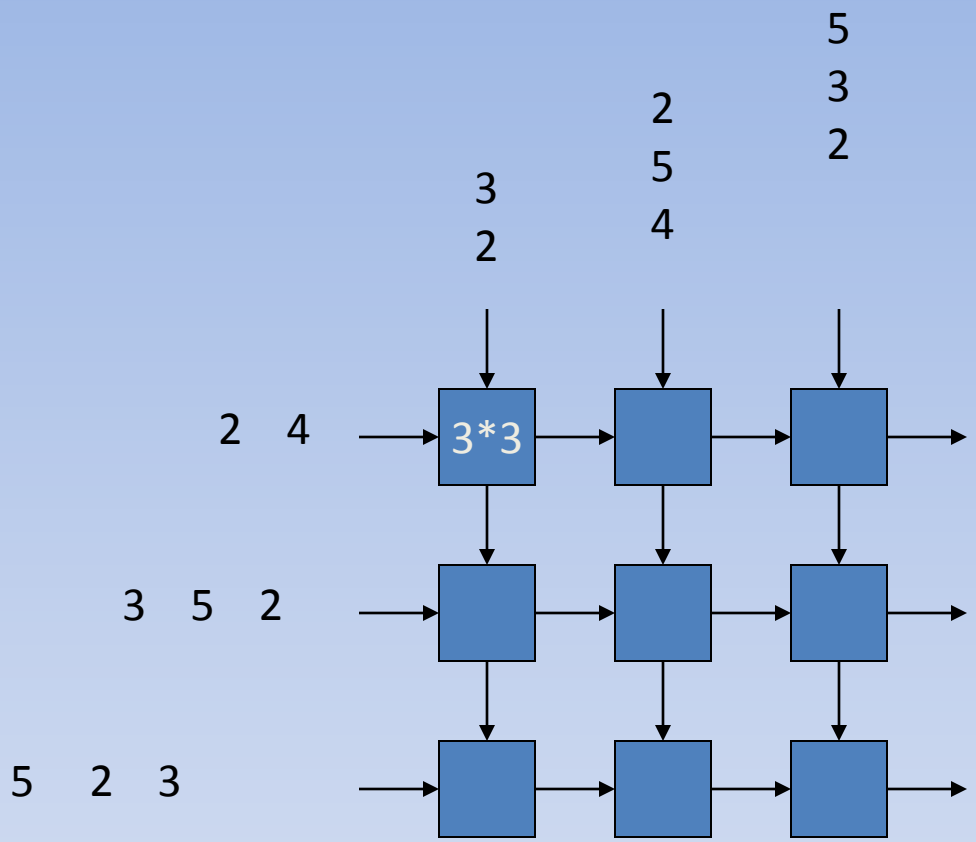
- A specialized form of parallel computing.
- Multiple processors connected by short wires.
- Unlike many forms of parallelism which lose speed through their connection.
- Cells(processors), compute data and store it independently of each other.

Systolic Unit(cell)

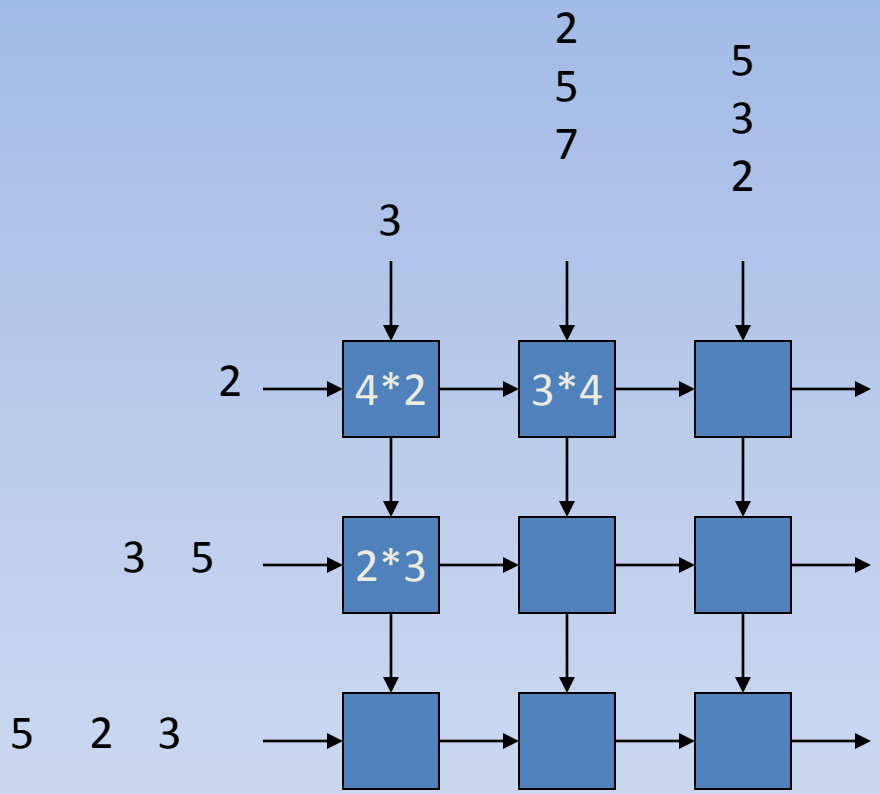
- Each unit is an independent processor.
- Every cell has some registers and an ALU.
- The cells share information with their neighbors, after performing the needed operations on the data.
- Performed parallel &pipelining


b11
b21
b31
b41
b12
b22
b32
b42
b13
b23
b33
b34
b14
b24
b34
b44
a11 a12 a13 a14
a21 a22 a23 a24
a31 a32 a33 a34
a41 a42 a43 a44
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

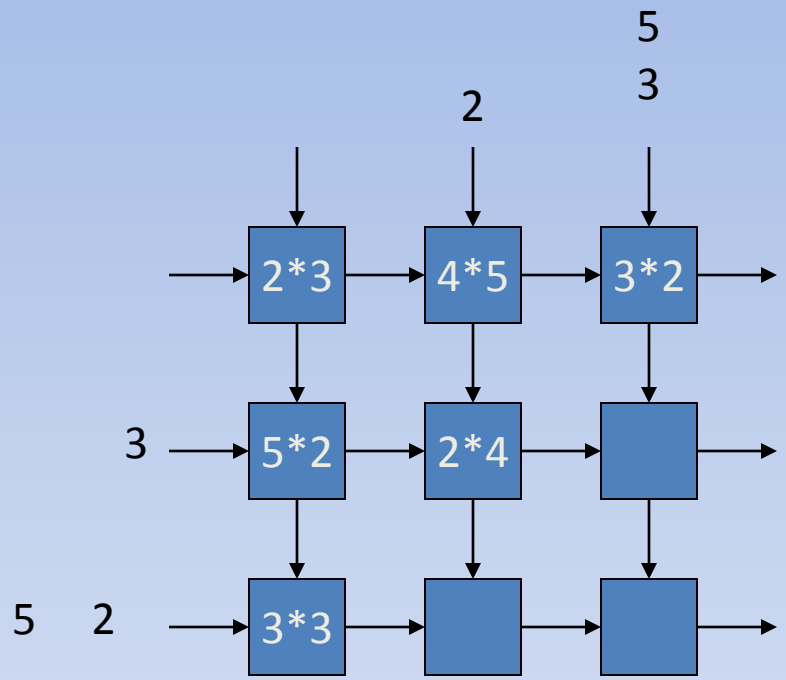
Clock tick: 1



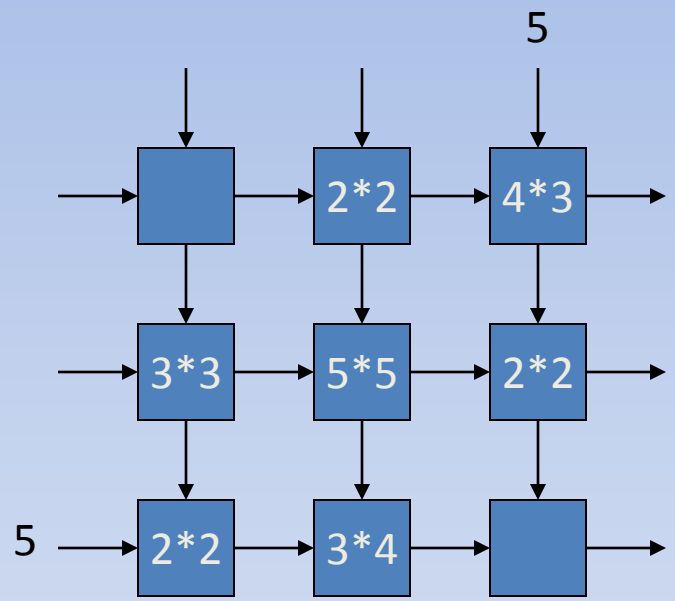
Clock tick: 2



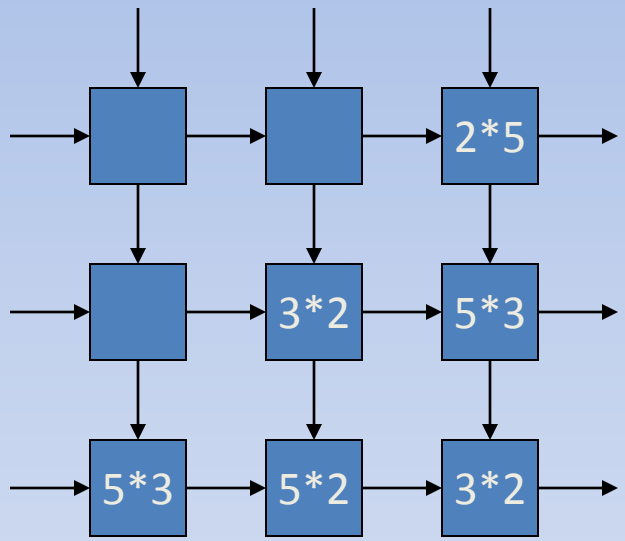
Clock tick: 3



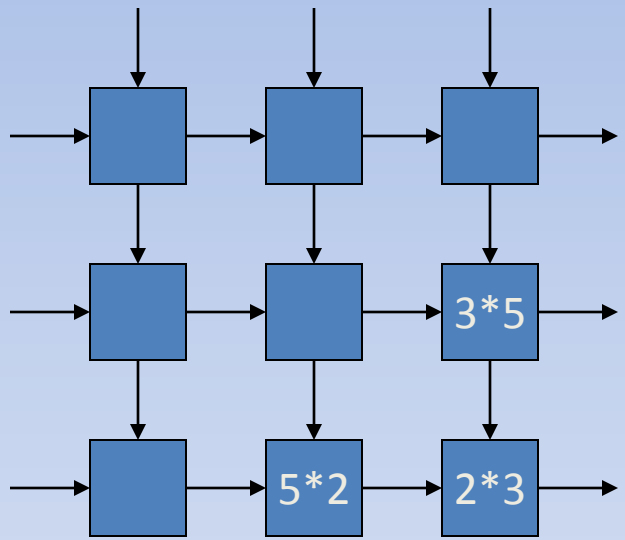
Clock tick: 4



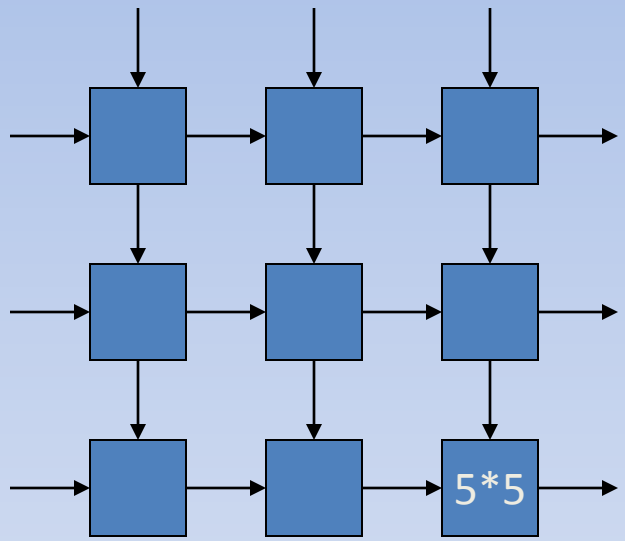
Clock tick: 5



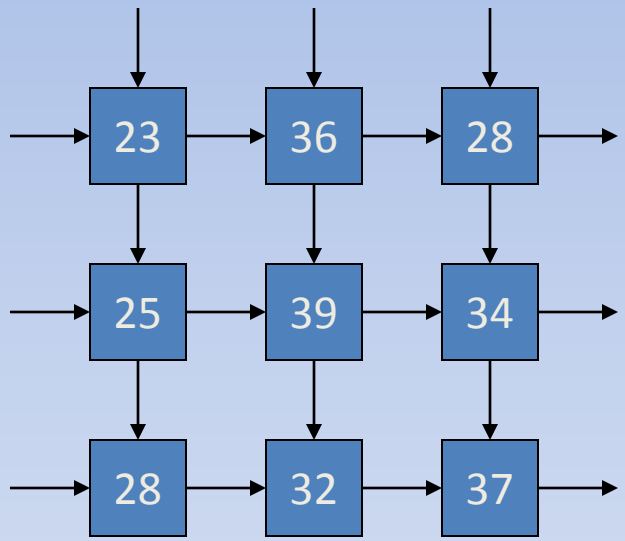
Clock 6

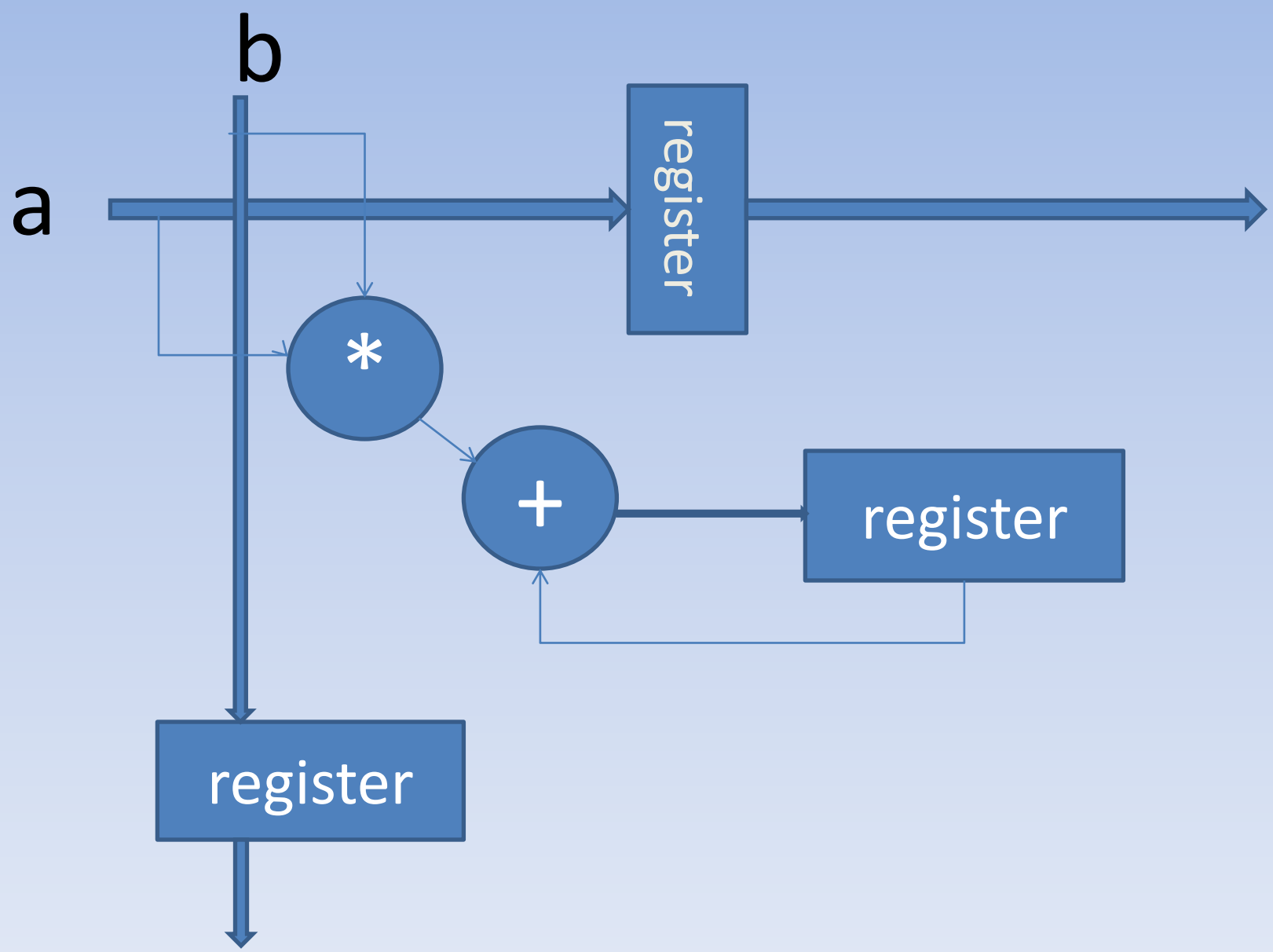


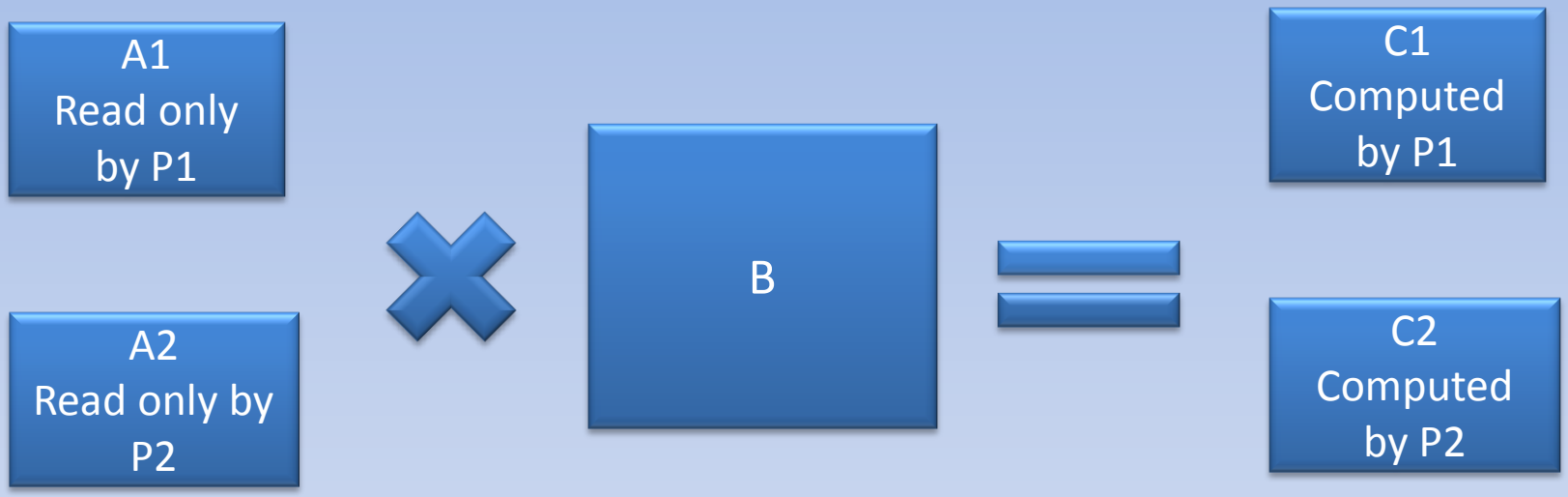
Clock tick: 7



Clock cycle :2n + 1




b
a
register
*
+
register
register


A1
Read only
by P1
A2
Read only by
P2
B
C1
Computed
by P1
C2
Computed
by P2

Loading Blocking

1st version the input matrices are generated by Linear Feedback Shift

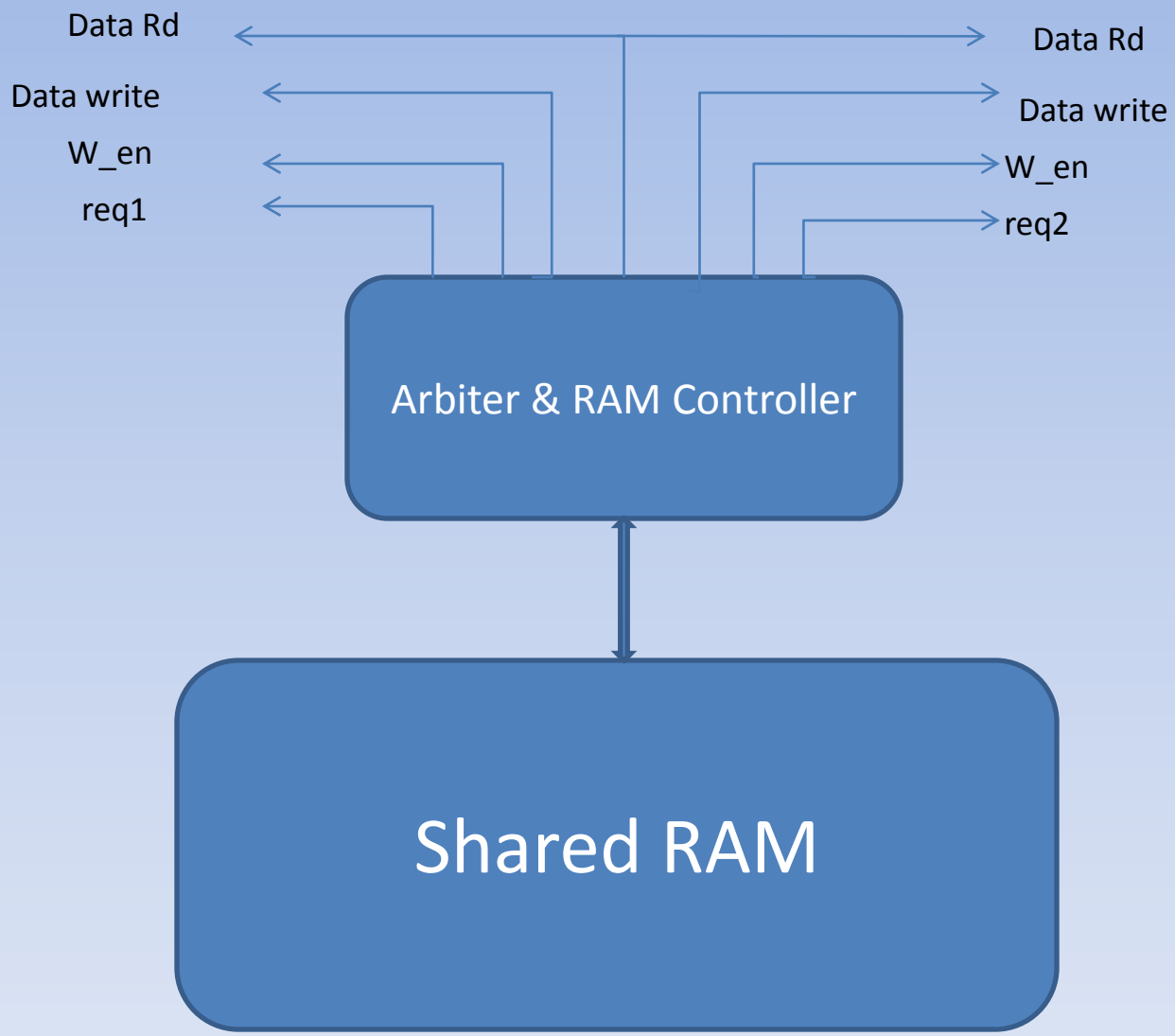
2ND version multi-ports ROM

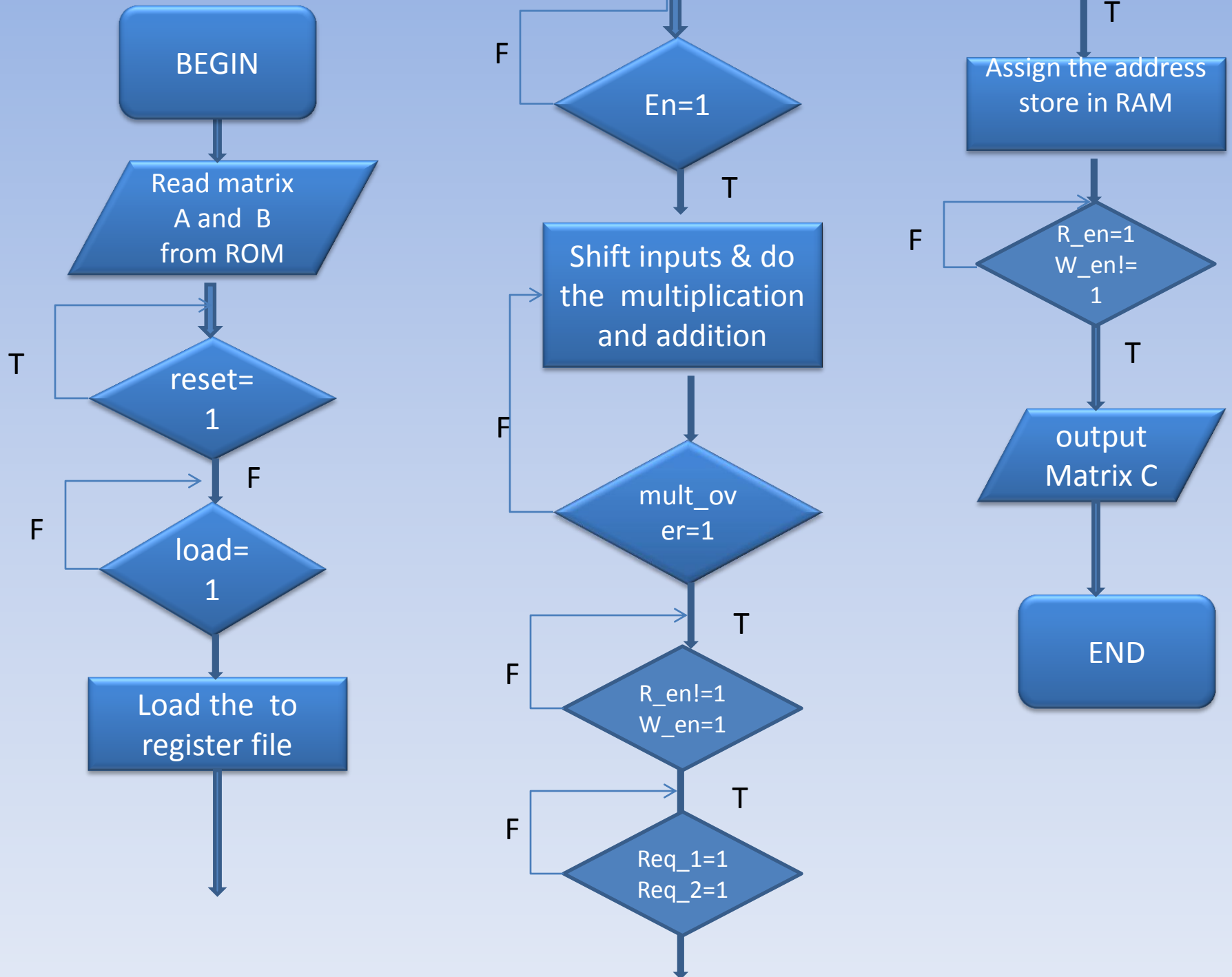
Systolic-Array

I modify the code which developed in the 1st part, and make sure matrix is 3 by 6;The length of input is 8 bits [-128 127],and the length of output elements is 16 bits [-32768 32767].

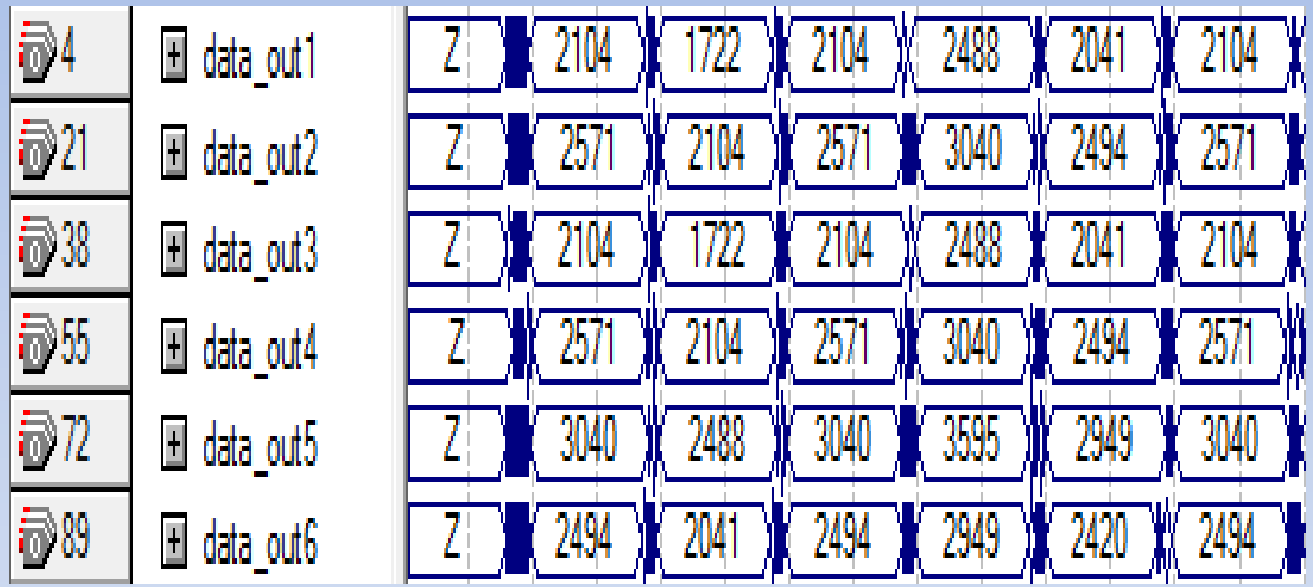
Output Blocking

RAM Controller &Arbiter & multi-port RAM


Data Rd
Data write
W_en
req1
Data Rd
Data write
W_en
req2
Arbiter & RAM Controller
Shared RAM


BEGIN
Read matrix A and B from ROM
reset= 1
T
F
load= 1
F
Load the to register file
En=1
F
T
Shift inputs & do the multiplication and addition
mult_ov er=1
F
T
R_en!=1 W_en=1
F
T
Req_1=1 Req_2=1
F
T
Assign the address store in RAM
R_en=1 W_en!= 1
F
T
output Matrix C
END

Simulation


4
data_out1
Z
2104
1722
2104
2488
2041
2104
21
data_out2
Z
2571
2104
2571
3040
2494
2571
38
data_out3
Z
2104
1722
2104
2488
2041
2104
55
data_out4
Z
2571
2104
2571
3040
2494
2571
72
data_out5
Z
3040
2488
3040
3595
2949
3040
89
data_out6
Z
2494
2041
2494
2949
2420
2494

Compare

6 by 6(input 8 bits Output 16 bits)	Version I(with LFSR)	Version II(with ROM)
Total LE	3371(Total memory 1164 bits)	2949
Power Dissipation	208.2mW	199.85mW
Fmax	145.65MHz	206MHz
Worest-case Tco	18.5ns	11.372ns

4*4 matrix multiplication	Systolic Array	Parallel with sequence
Fmax	312MHz	170MHz
Worst-case Tco	11ns	10.5ns
Power dissipation	232mW	216mW