Systolic Array for 3x3 Matrices Multiplication of Integer Data type

Ayyappa Koppuravuri
Electronics and Communication Engineering(ECE)
International Institute of Information Technology Bangalore(IIITB)
Bangalore, India
Ayyappa.Koppuravuri@iiitb.ac.in

***Abstract*—This document describes about a Systolic Array which can be used for 3x3 Matrices Multiplication of integer data type. As an alternative of pipeline structures a systolic array can be used which has Processing elements.This model consists of a few delay blocks. These are very useful for operations like Convolutions.**

***Index Terms*—Systolic Array, Convolution, pipeline structures, Processing Elements**

I. Introduction

Matrix Multiplication can be particularly be done by many algorithms. In this paper matrix multiplication for 3x3 matrices in integer data type is specially discussed using the Systolic array which is a hardware structure used for operating matrix multiplication fastly as well as effeciently. The Processing elements in a Systolic Array are programmed to have a same operation.

Systolic Array has mainly two different types of Components triggered by the same clock. They are:

A. Processing Elements(PE)

A Processing Element is a part of an hardware structure called systolic arrays alternatively called as cells. Processing Elements perform a common operation which are generally simple, but for different kinds of input. These cells will have memory banks for holding the data after each computation in the Processing Element.

The Processing Element in this particular systolic array has three inputs **c**, **b**, and **c** which has outputs **a' = a**, **b' = b**, and **c' = c + a * b**. The operation **c + a * b** can be considered as the common operation programmed in these processing elements. The outputs **a'**, and **b'** are the outcome of the property of the **PE** acting as a hardware Register holding data for a clock cycle. These operations are briefly described using a block diagram in **Fig. 1**.

B. Delay block

A Delay block is just a group of flipflops for holding data for a clock cycle which is triggered to all the Delay blocks and Processing Elements at a time. These are used to make the Systolic Array symmetrical. In this case the Systolic Array is **5x5** structured. The inputs to the delay block are **a**, and **b**. The outputs of the delay block are **a' = a**, and **b' = b**. The schematic of delay block is shown in **Fig. 2**

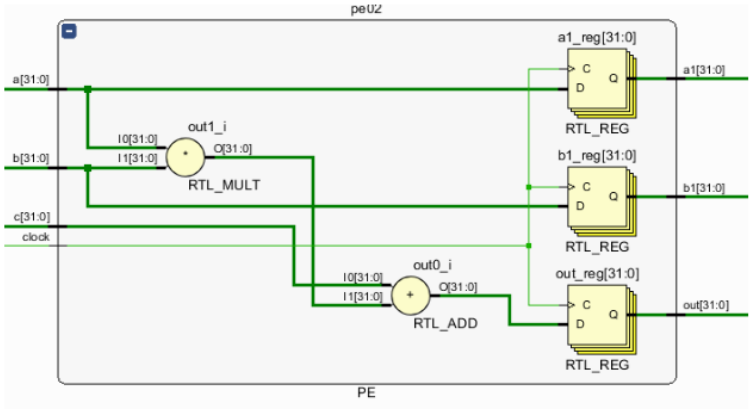

Fig. 1. Internal Structure of the PE used in the Systolic Array

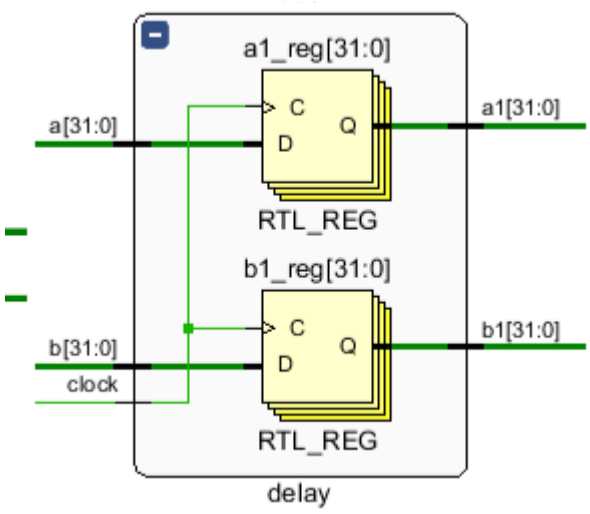

Fig. 2. Schematic of Delay Block used in the Systolic Array

II. Systolic Array

The systolic array used for matrix multiplication of 3x3 matrices of integer data types have inputs $A_{3x3}, and B_{3x3}$.

The whole process process takes **8** clock cycles of duration, In which each row of the Matrix A is passed to a00 , a10, and a20 and the first column of Matrix B

is passed to b00 , b01, and b02. The second row of the Matrix A is passed to a10 , a20, and a30, and the second column of the Matrix B is passed to b01 , b02 , and b03 respectively. Similarly, The third of row of the Matrix A is passed to a20 , a30 , and a40, and the third column of the Matrix B is passed to b02 , b03 and b04 respectively in the first consecutive clock cycles.Other values remains as zero.

The output of the Matrices Multiplication namely D is collected at c35, c45, c55, c35 , and c45 of the systolic array. D00, D01, D02, D10, and D20 are collected at c55, c45, c35, c54, and c53 indices of the systolic array at 6^{th} clock cycle of the process.Similarly, The indices D12, D11, D21 of the output matrix are collected at the indices c45, c55, and c54 of the systolic array at 7^{th} clock cycle of the process. Lastly, the index D22 of the Matrix D is collected at the index c55 of the systolic array at 8^{th} clock cycle of the process.

The Structure of the model we are using for Matrix multiplication is gicen in **Fig. 4** along with appropriate indexing used for above description.

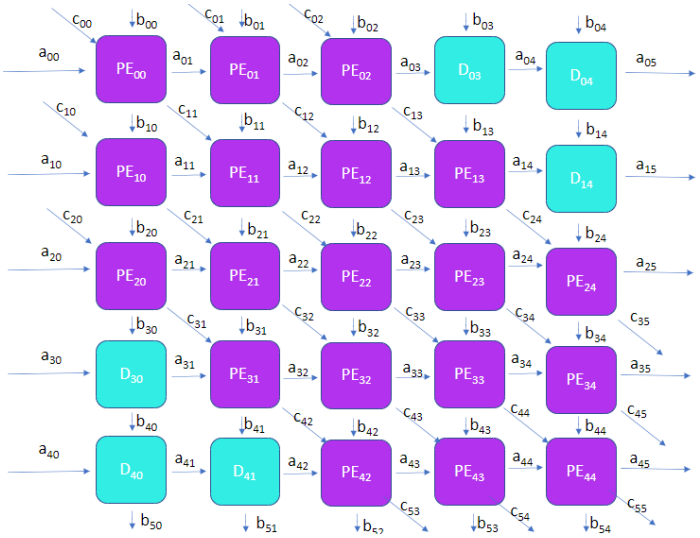

Fig. 4. Schematic of Systolic Array

The wave for an example is given as **Fig. 3**. we can also see the data moment at a top level by observing the waveform.

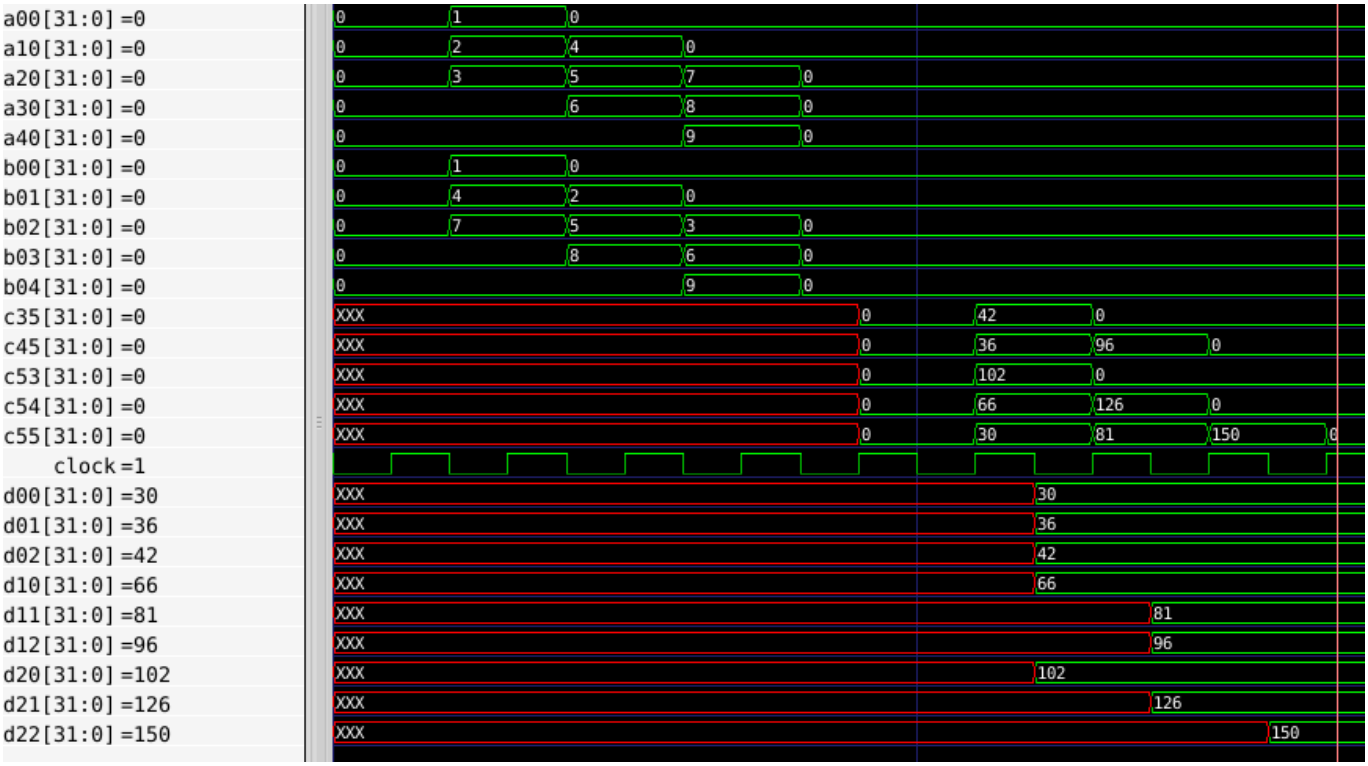

Fig. 3. Waveform showing the external data flow of the systolic array

where A = $\begin{bmatrix} A00 & A01 & A02 \\ A10 & A11 & A12 \\ A20 & A21 & A22 \end{bmatrix}$,B = $\begin{bmatrix} B00 & B01 & B02 \\ B10 & B11 & B12 \\ B20 & B21 & B22 \end{bmatrix}$

, and

D = $\begin{bmatrix} D00 & D01 & D02 \\ D10 & D11 & D12 \\ D20 & D21 & D22 \end{bmatrix}$

REFERENCES

REFERENCES

[1] C. Bagavathi MTech, O. Saraniya ME, PhD, in Deep Learning and Parallel Computing Environment for Bioengineering Systems, 2019
[2] https://youtu.be/vADVh1ogNo0