**Multistage Neural Network**

Name: Nirmal Kumar Sedhumadhavan

Unityid: nsedhum@ncsu.edu

StudentID: 200483323

1/(delay.area)

1.8879 \* 10-9 (ns-1.um-2)

Logic Area: 14895.7341

(um2)

Memory: 260 MBytes

Delay (ns to run provided provided example).

Clock period:7 ns

# cycles”:5080

Delay (TA provided example. TA to complete):

1/(delay.area) (TA)

**Abstract**

This Project involves designing hardware for a multistage neural network. The first layer consists of a convolution layer which has 3\*3 weights window. The input data for performing convolution is read from an input SRAM and the weights data is read from another SRAM. Once convolution is performed, the RELU activation function is used which limits the lowest value to 0 and highest value to 127.Once RELU activation function is performed, we pass the intermediate output data to the max pooling layer which has a window size of 4. This output data is stored in an output SRAM which is finally tested for correctness.

First for convolution, we fetch a 4\*4 matrix from i/p sram while parallelly fetching the 3\*3 matrix for weights out of which we get 4 convolution results from. We then perform RELU on these 4 results and then directly perform max pooling on these 4 results. Once we’re done with one 4\*4 matrix we decide to move horizontally or vertically based on the row counter and the column counter. My design achieved a clock period of 7 ns and an area of 14895.7341 um2.

**1. Introduction**

1.A multi-layer neural network consisting of convolution layer with RELU as activation function and max pooling layer has been designed and implemented in Verilog.

2. Four SRAMs are used, one for input SRAM, one for weight SRAM, one for scratchpad SRAM and one for output SRAM.

3. Four Convolutions are performed at a time as we fetch a 4\*4 matrix from input SRAM.

4. Once we perform 4 convolutions, we perform RELU activation on the 4 convolution outputs.

5.These 4 RELU outputs are then fed to the max pooling layer which produces the maximum of the 4 values as output.

6.This output is then stored to the output SRAM.

7.The 4\*4 input window is then moved horizontally and the same steps 3,4,5,6 are repeated.

8.Once horizontal movement in done, we move the 4\*4 input window vertically and the same steps 3,4,5,6 are repeated.

**2.Microarchitecture**

My design makes use of the following for the convolution process

->An 8-bit multiplier for the multiplication part of the convolution process

->A 20-bit accumulator for the addition part of the convolution process

->4 comparators are used for implementing RELU

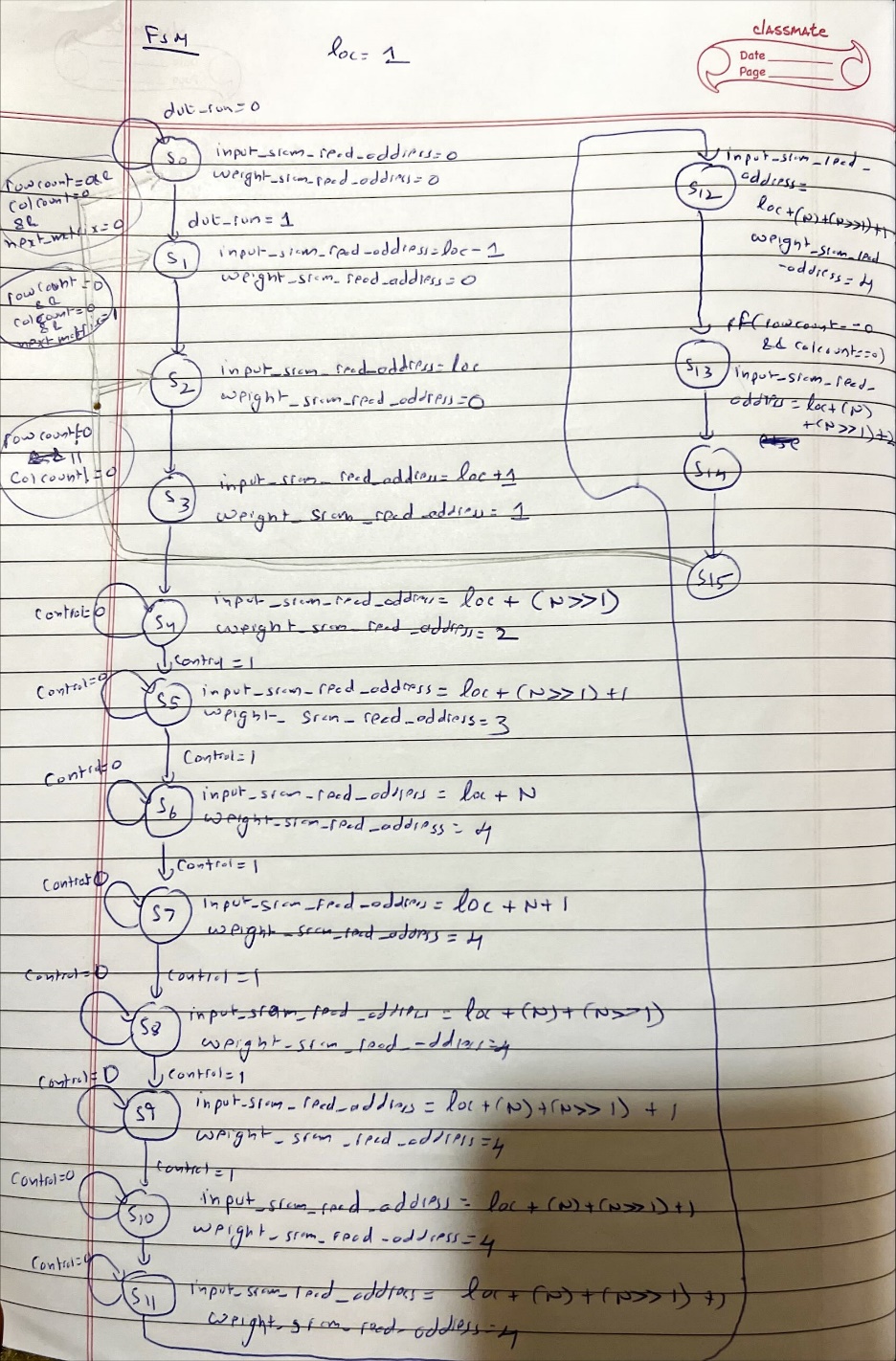
->3 comparators are used for implementing Max pooling

**3. Interface Specification**

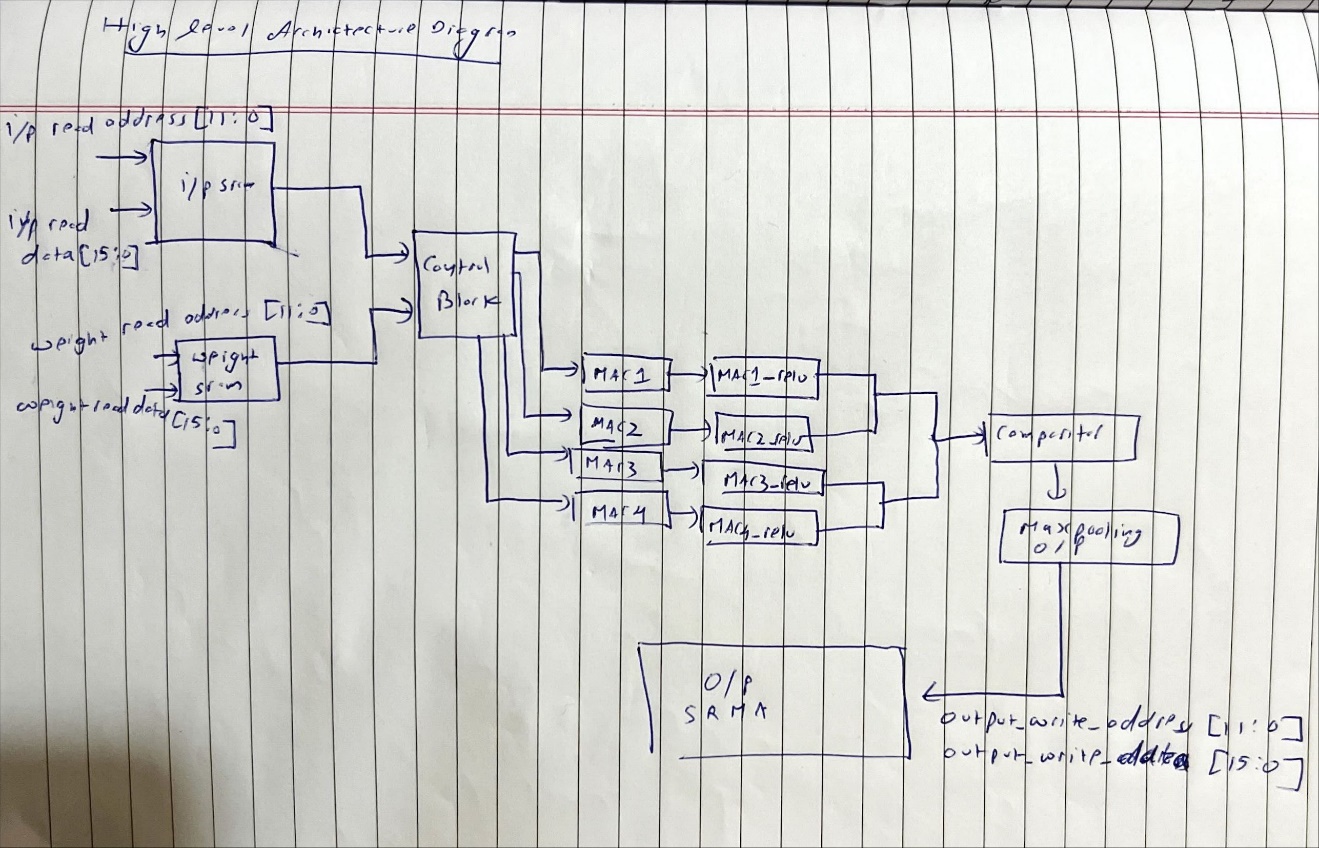
|  |  |  |
| --- | --- | --- |
| **Signal** | **Width** | **Function** |
| dut\_run | 1 | used to indicate when to start process |
| dut\_busy | 1 | Goes high once dut\_run goes high and stays high until process is done |
| reset\_b | 1 | Used for resetting to initial state |
| clk | 1 | Used as clock signal for design |
| input\_sram\_read\_address | 12 | Used to set which address to read from with regards to input SRAM |
| input\_sram\_read\_data | 16 | Reads input SRAM data |
| output\_sram\_write\_addresss | 12 | Used to set which address to write to with regards to output SRAM |
| output\_sram\_write\_data | 16 | Writes output SRAM data |
| weights\_sram\_read\_address | 12 | Used to set which address to read from with regards to weight SRAM |
| weights\_sram\_read\_data | 16 | Reads weight SRAM data |
| Starting\_position | 12 | Used to set the starting point of the 4\*4 input window we are using |
| Matrix\_size | 8 | Used to store size of matrix |
| Row\_counter | 12 | Used to determine which row we are at |
| Column\_counter | 12 | Used to determine which column we are at |
| Row\_max | 12 | Used to determine end of matrix |
| Column\_max | 12 | Used to determine end of matrix |
| i1 - i16 | 8 | Used to store the 4\*4 input matrix window values |
| k1 - k9 | 8 | Used to store the 3\*3 kernel matrix values |
| Conv1 - Conv4 | 20 | Used to store values of convolution computations of the 4\*4 input window |
| Relu1 - Relu4 | 20 | Used to store values of RELU computations of the 4\*4 input window |
| Max, Max1, Max2 | 20 | Used to store max pooling values |
| Store | 8 | Used to store the value of first half of max pooling value to be stored in SRAM |
| write\_address\_counter | 20 | Used to determine which address to write to next with respect to output SRAM |
| Write\_flag | 1 | Used to determine whether to write in upper 8 bits or lower 8 bits of the output SRAM location |
| Matrix\_size\_flag | 1 | Used to determine whether to read size of matrix or not |
| Next\_matrix\_flag | 1 | Used as a flag to indicate if there is another matrix in the input stream |

**4. Technical Implementation**

**FSM**

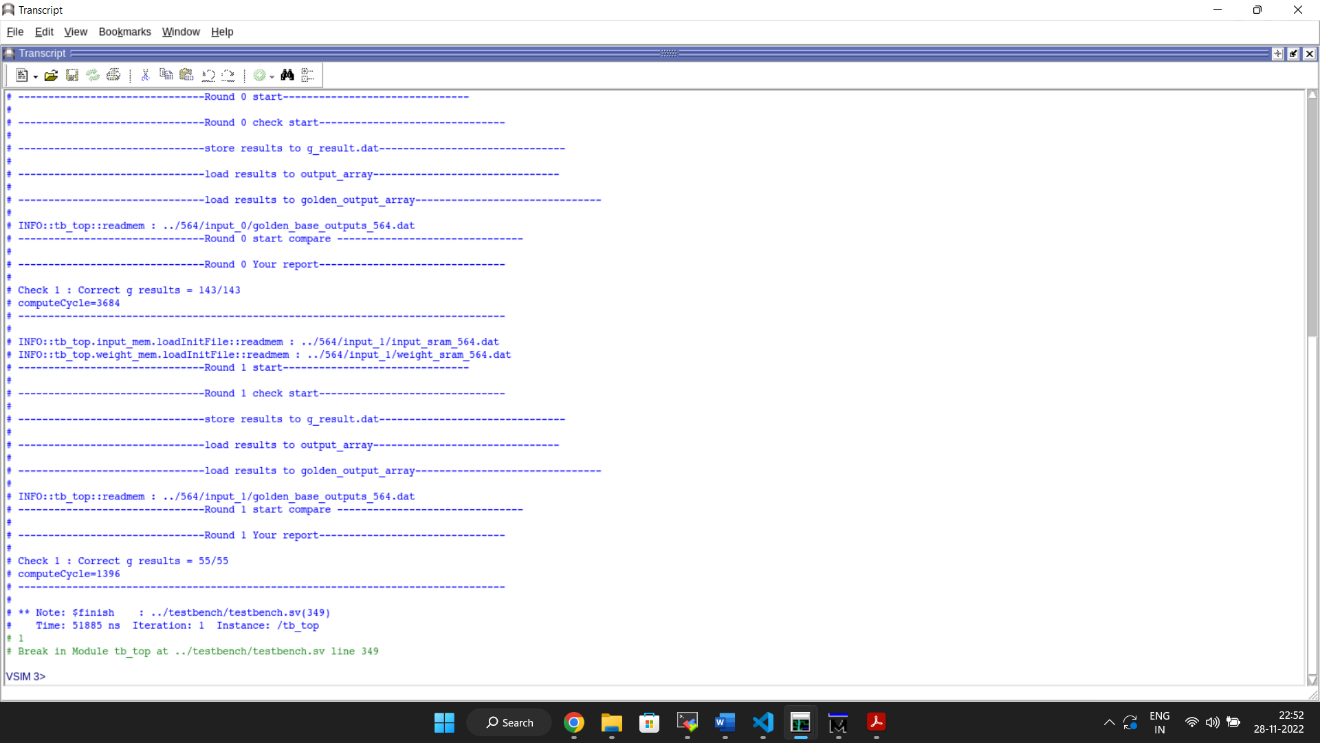


**High Level Architecture**

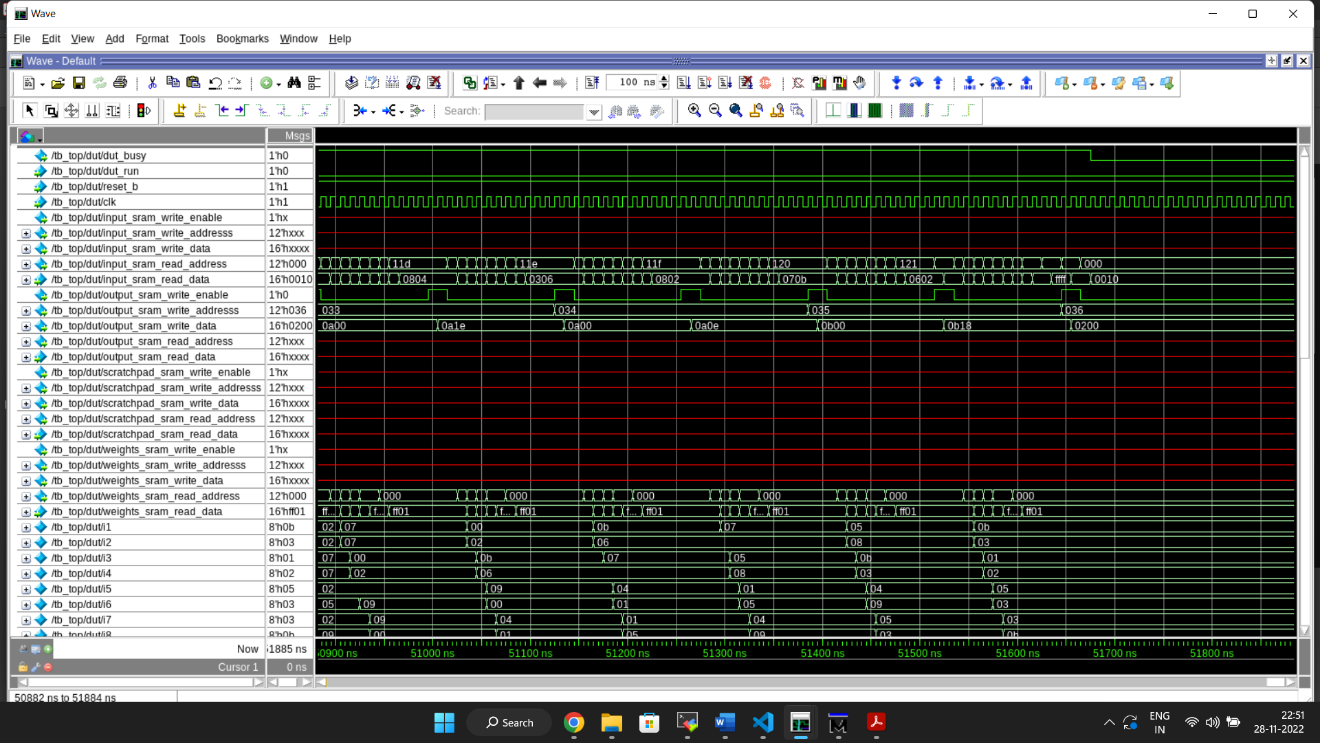


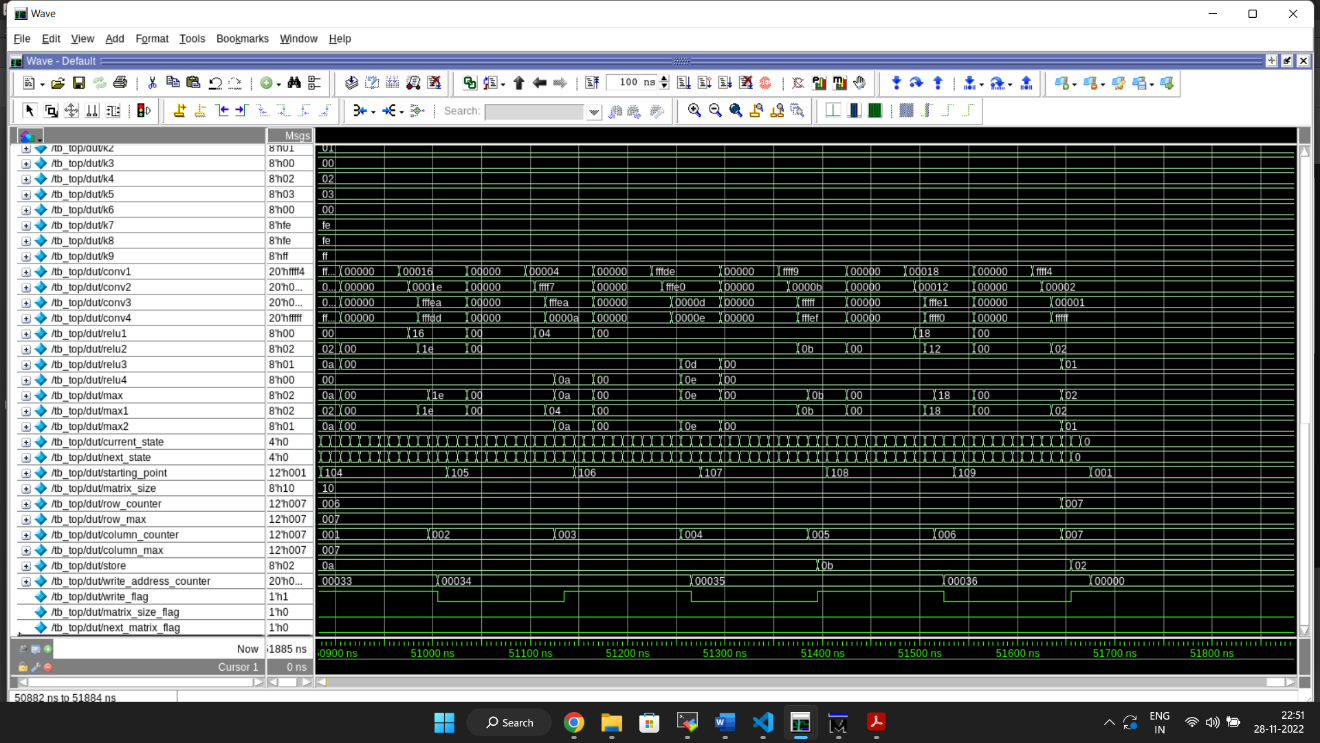
**5. Verification**

The design was run on modelsim using a testbench and the outputs were compared to the golden model outputs.



Waveform





**Timing\_max\_slow**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Report : timing

-path full

-delay max

-max\_paths 1

Design : MyDesign

Version: S-2021.06-SP3

Date : Mon Nov 28 23:18:34 2022

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Operating Conditions: slow Library: NangateOpenCellLibrary\_PDKv1\_2\_v2008\_10\_slow\_nldm

Wire Load Model Mode: top

Startpoint: i3\_reg[2] (rising edge-triggered flip-flop clocked by clk)

Endpoint: conv1\_reg[17]

(rising edge-triggered flip-flop clocked by clk)

Path Group: clk

Path Type: max

Point Incr Path

-----------------------------------------------------------

clock clk (rise edge) 0.0000 0.0000

clock network delay (ideal) 0.0000 0.0000

i3\_reg[2]/CK (DFF\_X1) 0.0000 0.0000 r

i3\_reg[2]/Q (DFF\_X1) 0.6120 0.6120 f

U6823/ZN (XNOR2\_X2) 0.3276 0.9396 f

U1769/ZN (AND2\_X2) 0.2497 1.1892 f

U1759/ZN (INV\_X4) 0.1773 1.3665 r

U1265/ZN (OAI22\_X1) 0.1744 1.5409 f

U8223/CO (FA\_X1) 0.6205 2.1613 f

U8485/S (FA\_X1) 0.9266 3.0879 r

U7703/ZN (XNOR2\_X1) 0.3659 3.4538 r

U6833/ZN (XNOR2\_X2) 0.3598 3.8136 r

U8530/CO (FA\_X1) 0.4245 4.2381 r

U6451/ZN (NAND2\_X2) 0.0958 4.3339 f

U6449/ZN (NAND2\_X2) 0.1251 4.4590 r

U6447/ZN (NAND2\_X2) 0.0687 4.5277 f

U2885/ZN (NAND2\_X2) 0.1270 4.6547 r

U2953/ZN (XNOR2\_X2) 0.2928 4.9475 r

U2952/ZN (XNOR2\_X2) 0.3184 5.2659 r

U2951/ZN (XNOR2\_X2) 0.3302 5.5961 r

U6681/ZN (NOR2\_X2) 0.0889 5.6850 f

U6680/ZN (NOR2\_X2) 0.1270 5.8120 r

U6677/ZN (NOR2\_X2) 0.0570 5.8690 f

U1884/ZN (NAND2\_X2) 0.2088 6.0778 r

U7154/ZN (NAND3\_X2) 0.0889 6.1668 f

U7302/ZN (NAND3\_X2) 0.1365 6.3033 r

U7305/ZN (XNOR2\_X2) 0.2859 6.5891 r

U7304/ZN (NAND2\_X2) 0.0737 6.6628 f

U7303/ZN (NAND2\_X2) 0.1018 6.7646 r

conv1\_reg[17]/D (DFF\_X1) 0.0000 6.7646 r

data arrival time 6.7646

clock clk (rise edge) 7.0000 7.0000

clock network delay (ideal) 0.0000 7.0000

clock uncertainty -0.0500 6.9500

conv1\_reg[17]/CK (DFF\_X1) 0.0000 6.9500 r

library setup time -0.1854 6.7646

data required time 6.7646

-----------------------------------------------------------

data required time 6.7646

data arrival time -6.7646

-----------------------------------------------------------

slack (MET) 0.0000

1

**Timing\_max\_slow\_holdfixed**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Report : timing

-path full

-delay max

-max\_paths 1

Design : MyDesign

Version: S-2021.06-SP3

Date : Mon Nov 28 23:18:47 2022

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Operating Conditions: slow Library: NangateOpenCellLibrary\_PDKv1\_2\_v2008\_10\_slow\_nldm

Wire Load Model Mode: top

Startpoint: i3\_reg[2] (rising edge-triggered flip-flop clocked by clk)

Endpoint: conv1\_reg[17]

(rising edge-triggered flip-flop clocked by clk)

Path Group: clk

Path Type: max

Point Incr Path

-----------------------------------------------------------

clock clk (rise edge) 0.0000 0.0000

clock network delay (ideal) 0.0000 0.0000

i3\_reg[2]/CK (DFF\_X1) 0.0000 0.0000 r

i3\_reg[2]/Q (DFF\_X1) 0.6120 0.6120 f

U6823/ZN (XNOR2\_X2) 0.3276 0.9396 f

U1769/ZN (AND2\_X2) 0.2497 1.1892 f

U1759/ZN (INV\_X4) 0.1773 1.3665 r

U1265/ZN (OAI22\_X1) 0.1744 1.5409 f

U8223/CO (FA\_X1) 0.6205 2.1613 f

U8485/S (FA\_X1) 0.9266 3.0879 r

U7703/ZN (XNOR2\_X1) 0.3659 3.4538 r

U6833/ZN (XNOR2\_X2) 0.3598 3.8136 r

U8530/CO (FA\_X1) 0.4245 4.2381 r

U6451/ZN (NAND2\_X2) 0.0958 4.3339 f

U6449/ZN (NAND2\_X2) 0.1251 4.4590 r

U6447/ZN (NAND2\_X2) 0.0687 4.5277 f

U2885/ZN (NAND2\_X2) 0.1270 4.6547 r

U2953/ZN (XNOR2\_X2) 0.2928 4.9475 r

U2952/ZN (XNOR2\_X2) 0.3184 5.2659 r

U2951/ZN (XNOR2\_X2) 0.3302 5.5961 r

U6681/ZN (NOR2\_X2) 0.0889 5.6850 f

U6680/ZN (NOR2\_X2) 0.1270 5.8120 r

U6677/ZN (NOR2\_X2) 0.0570 5.8690 f

U1884/ZN (NAND2\_X2) 0.2088 6.0778 r

U7154/ZN (NAND3\_X2) 0.0889 6.1668 f

U7302/ZN (NAND3\_X2) 0.1365 6.3033 r

U7305/ZN (XNOR2\_X2) 0.2859 6.5891 r

U7304/ZN (NAND2\_X2) 0.0737 6.6628 f

U7303/ZN (NAND2\_X2) 0.1018 6.7646 r

conv1\_reg[17]/D (DFF\_X1) 0.0000 6.7646 r

data arrival time 6.7646

clock clk (rise edge) 7.0000 7.0000

clock network delay (ideal) 0.0000 7.0000

clock uncertainty -0.0500 6.9500

conv1\_reg[17]/CK (DFF\_X1) 0.0000 6.9500 r

library setup time -0.1854 6.7646

data required time 6.7646

-----------------------------------------------------------

data required time 6.7646

data arrival time -6.7646

-----------------------------------------------------------

slack (MET) 0.0000

1

**Timing\_min\_fast\_holdcheck**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Report : timing

-path full

-delay min

-max\_paths 1

Design : MyDesign

Version: S-2021.06-SP3

Date : Mon Nov 28 23:18:42 2022

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Operating Conditions: fast Library: NangateOpenCellLibrary\_PDKv1\_2\_v2008\_10\_fast\_nldm

Wire Load Model Mode: top

Startpoint: write\_address\_counter\_reg[10]

(rising edge-triggered flip-flop clocked by clk)

Endpoint: write\_address\_counter\_reg[10]

(rising edge-triggered flip-flop clocked by clk)

Path Group: clk

Path Type: min

Point Incr Path

--------------------------------------------------------------------------

clock clk (rise edge) 0.0000 0.0000

clock network delay (ideal) 0.0000 0.0000

write\_address\_counter\_reg[10]/CK (DFF\_X1) 0.0000 0.0000 r

write\_address\_counter\_reg[10]/Q (DFF\_X1) 0.0619 0.0619 r

U12163/ZN (OAI21\_X1) 0.0198 0.0817 f

write\_address\_counter\_reg[10]/D (DFF\_X1) 0.0000 0.0817 f

data arrival time 0.0817

clock clk (rise edge) 0.0000 0.0000

clock network delay (ideal) 0.0000 0.0000

clock uncertainty 0.0500 0.0500

write\_address\_counter\_reg[10]/CK (DFF\_X1) 0.0000 0.0500 r

library hold time 0.0006 0.0506

data required time 0.0506

--------------------------------------------------------------------------

data required time 0.0506

data arrival time -0.0817

--------------------------------------------------------------------------

slack (MET) 0.0311

1

**6. Results Achieved**

Clock Period:7 ns

Compute Cycles:5080

Area: 14895.7341 um2

Performance: 1.8879 \* 10-9 ns-1.um-2

**7. Conclusion**

A multi-layer convolutional neural network was successfully designed and synthesized which reads values from SRAM and computes value to write the output in SRAM. The design results were matched with the golden output. The clock achieved was 7 ns with an area of 14895.7341 um2.Tradeoff between clock and area was kept in mind while doing the simulation. Design before coding was thoroughly practiced. Working on this complex design taught various topics such as SRAM interfacing, how to handle synthesis issues, and optimizing the design to achieve good performance overall keeping the clock and area constraints in mind.