“Scalable Systolic Array Multiplier Optimized by Sparse Matrix”

Presented By:
Banda Anusha (MT2022504)
Dantu Nandini Devi (MS2022007)
Mahati Basavaraju(MS2022016)
Vasanthi D R (PH2021504)

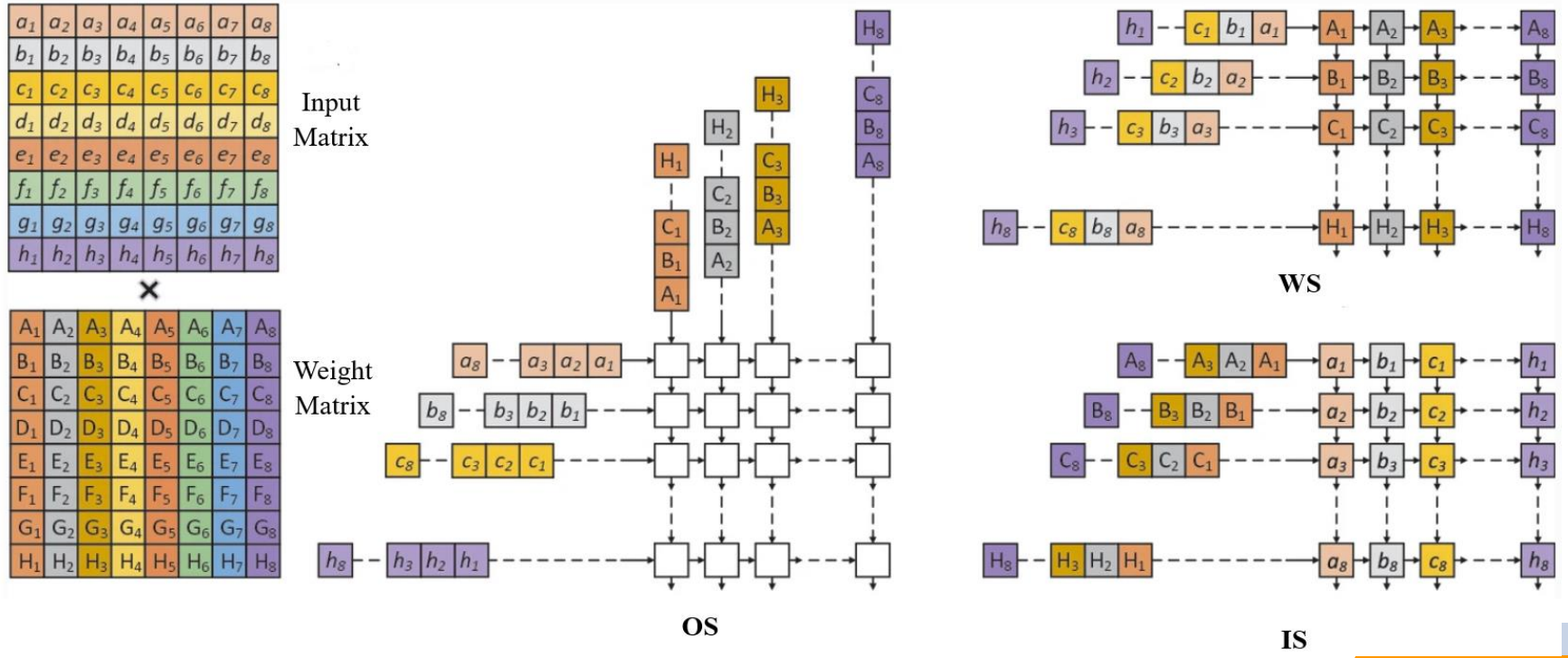
AGENDA

- Objective
- Introduction
- Application
- Paper Implementation & Simulation Results
- Proposed Method
- Implementation Results
- Future Work & Conclusion
- References

Objective

Sparse matrix multiplication optimized by scalable systolic array has been considered, which was compared with traditional systolic array for application in neural networks .

INTRODUCTION – SYSTOLIC ARRAY



Cont.

SPARSE MATRIX MULTIPLICATION

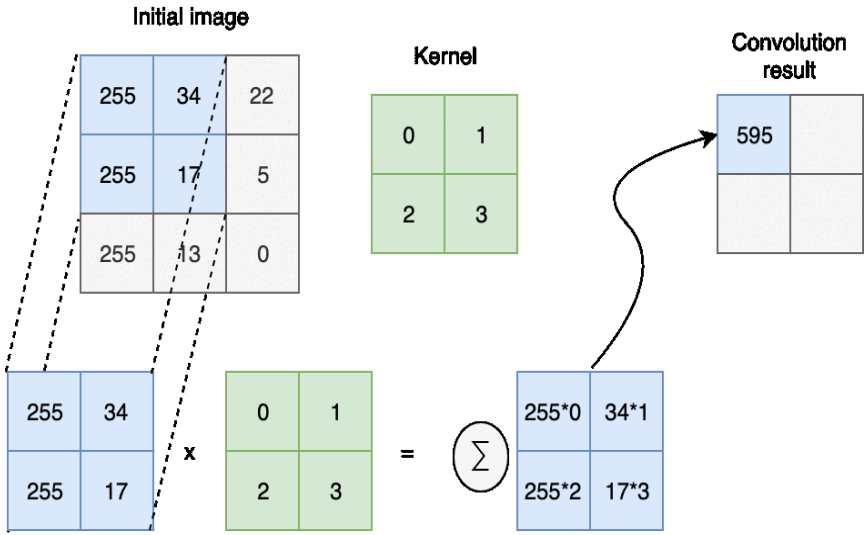
$$\begin{bmatrix} 0 & 0 & 3 & 0 & 4 \\ 0 & 0 & 5 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 6 & 0 & 0 \end{bmatrix}$$

Row	0	0	1	1	3	3
Column	2	4	2	3	1	2
Value	3	4	5	7	2	6

APPLICATION - CONVOLTUTION NEURAL NETWORK

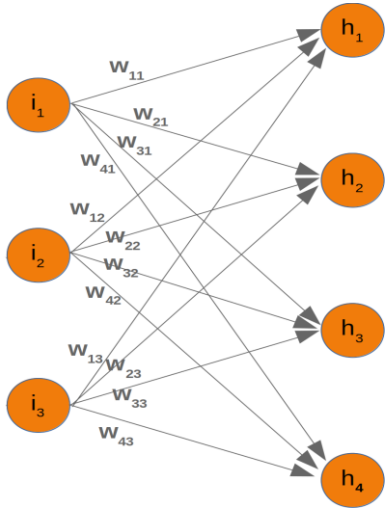
Convolution Layer

Process : Convolution

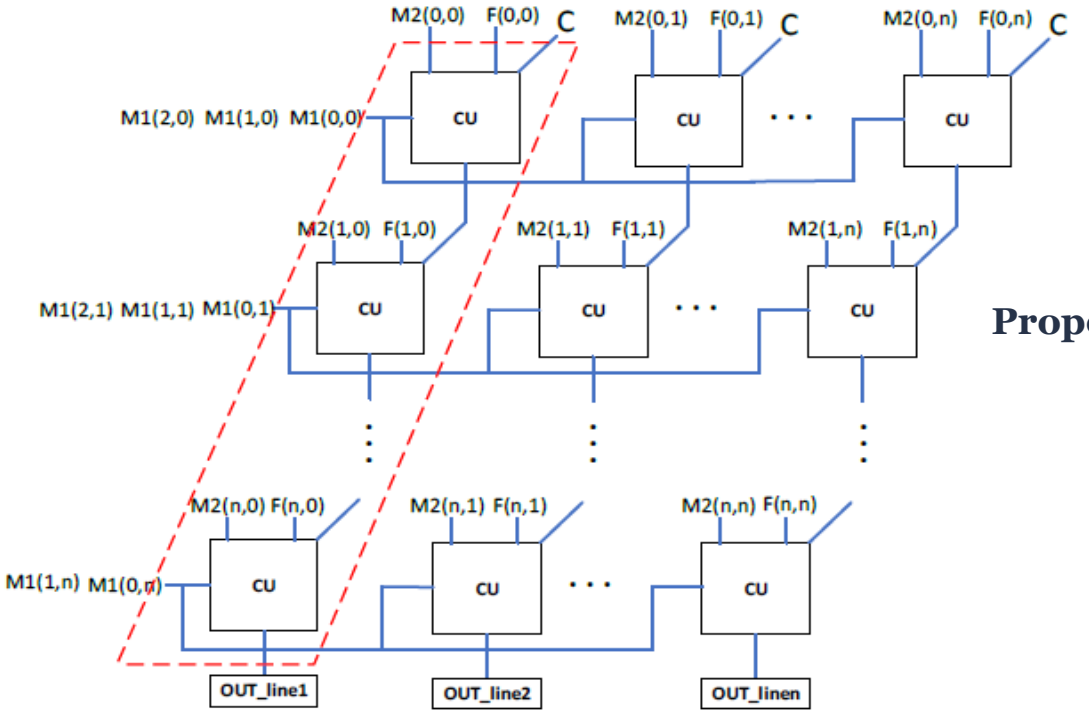


Fully-Connected Layer

Process : Matrix Multiplication

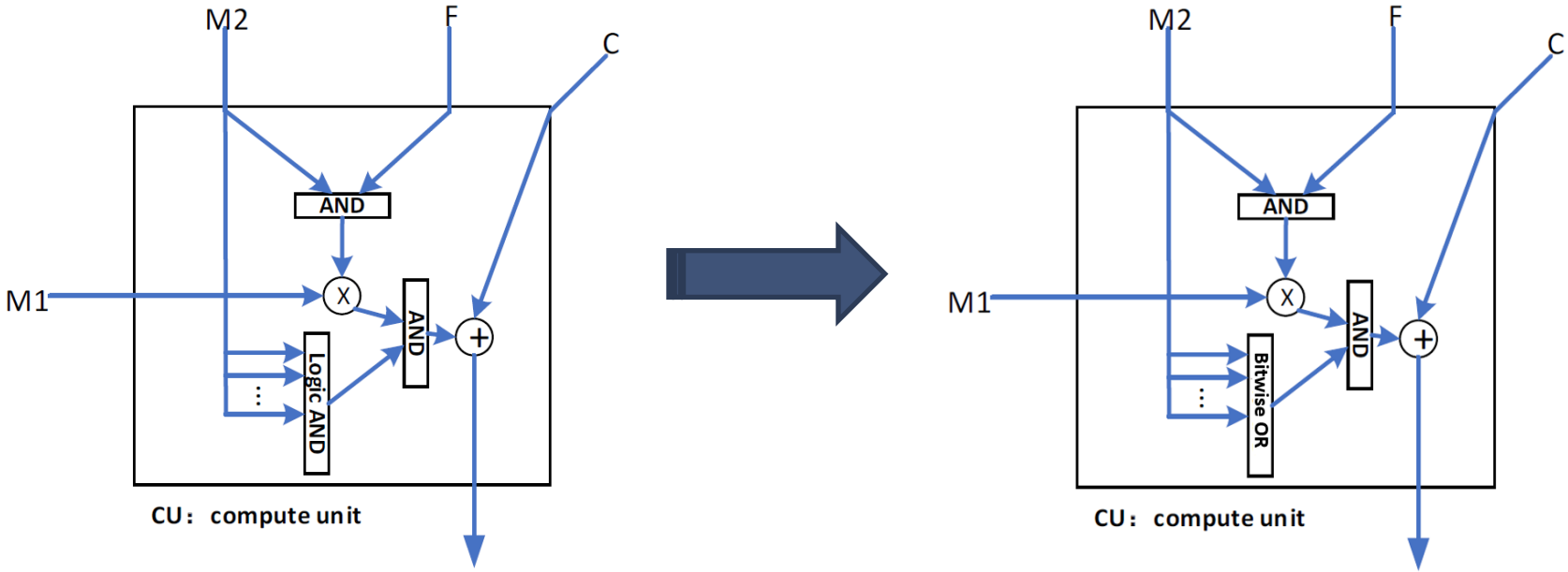


PAPER IMLEMENTATION

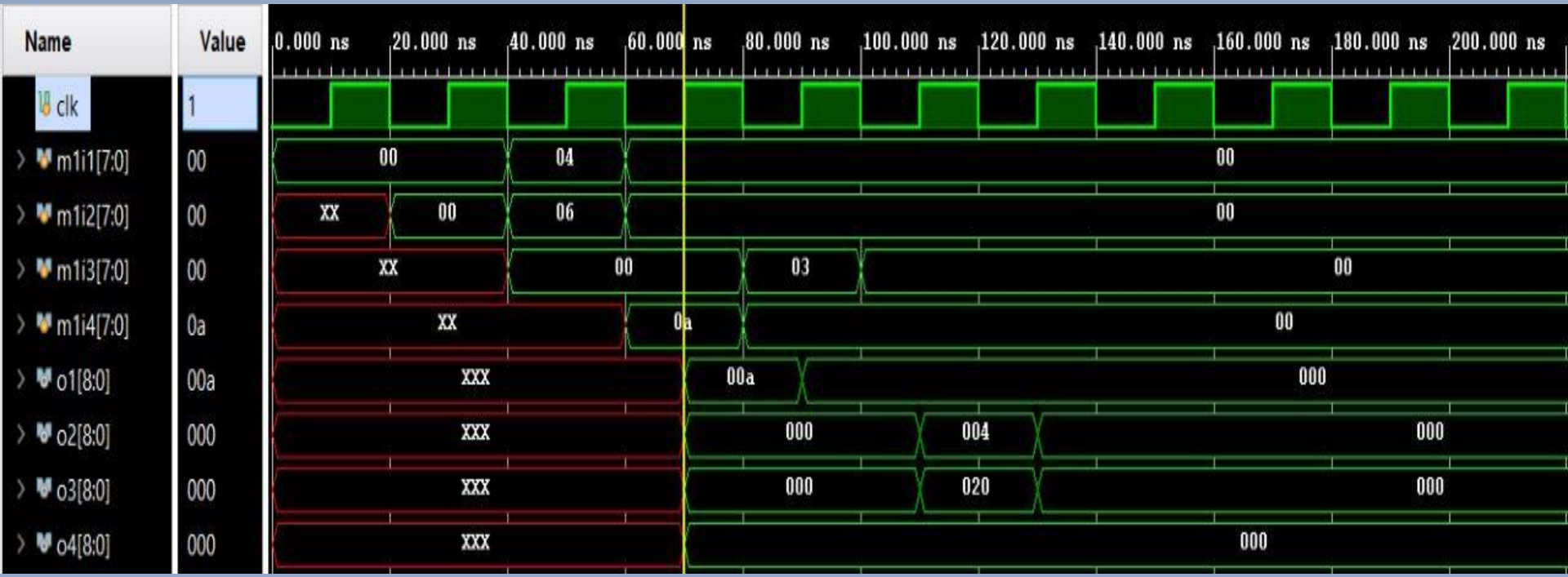


Proposed Systolic Array Circuit

Structure of Computation Unit



SIMULATION RESULTS



FPGA Results

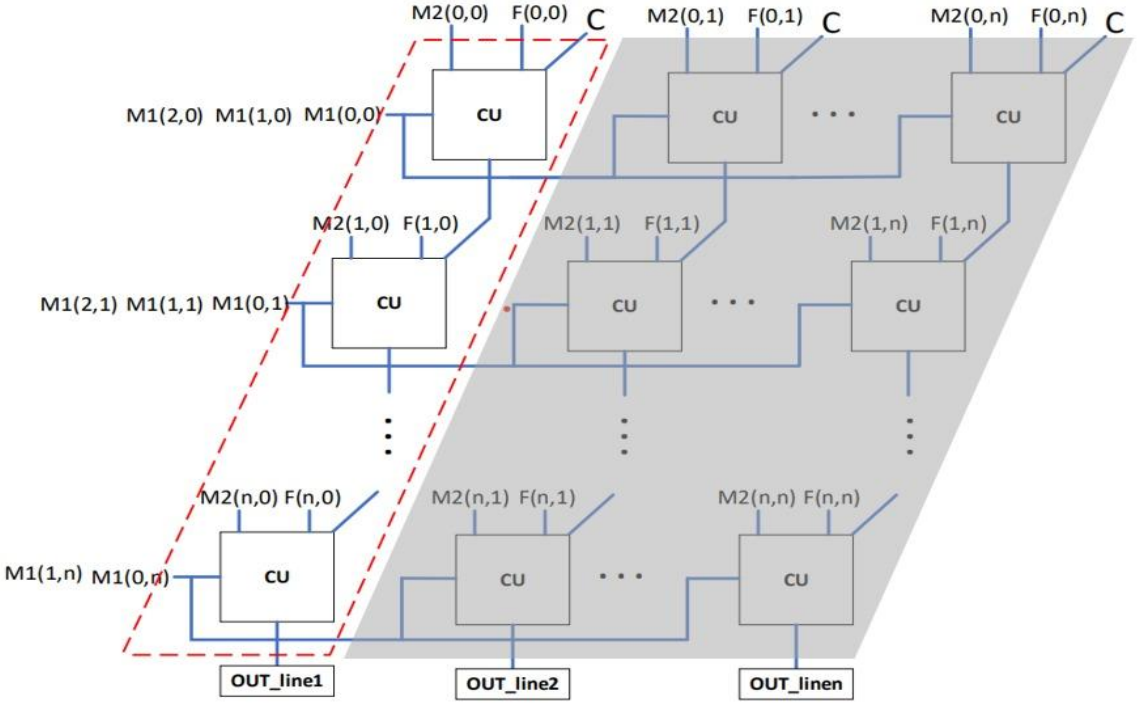
Multiplier Structure	Delay (ns)	LUT
Traditional systolic array	7.874	732
Optimized sparse matrix	6.827	658

ASIC Results

Multiplier Structure	Delay (ps)	AREA (um^2)
Traditional systolic array	4817	5187.146
Optimized sparse matrix	2767	4368.024

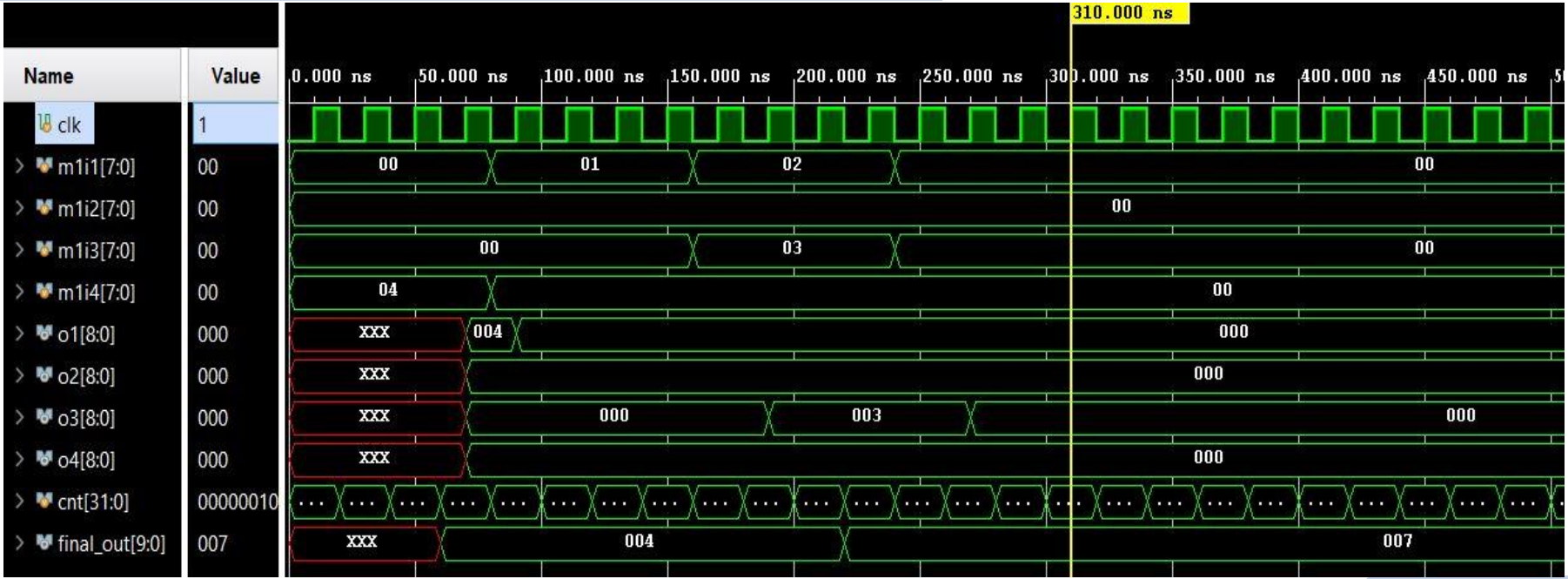
Convolution using Paper Implementation

Systolic Array Structure for Convolution

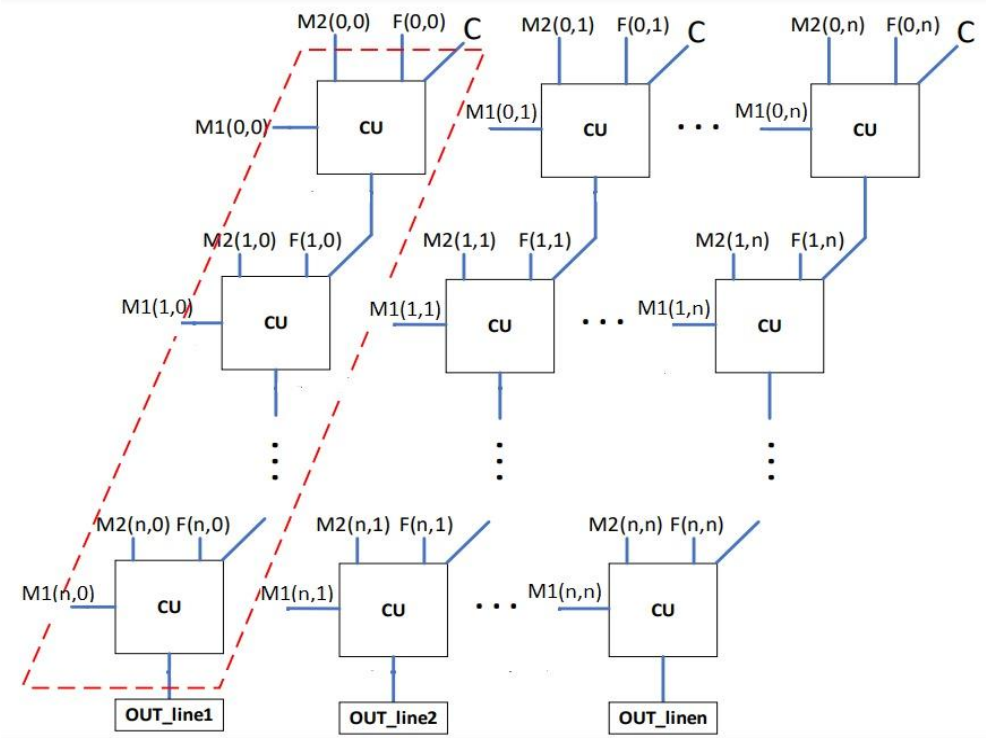


Simulation Results

N x N cycles

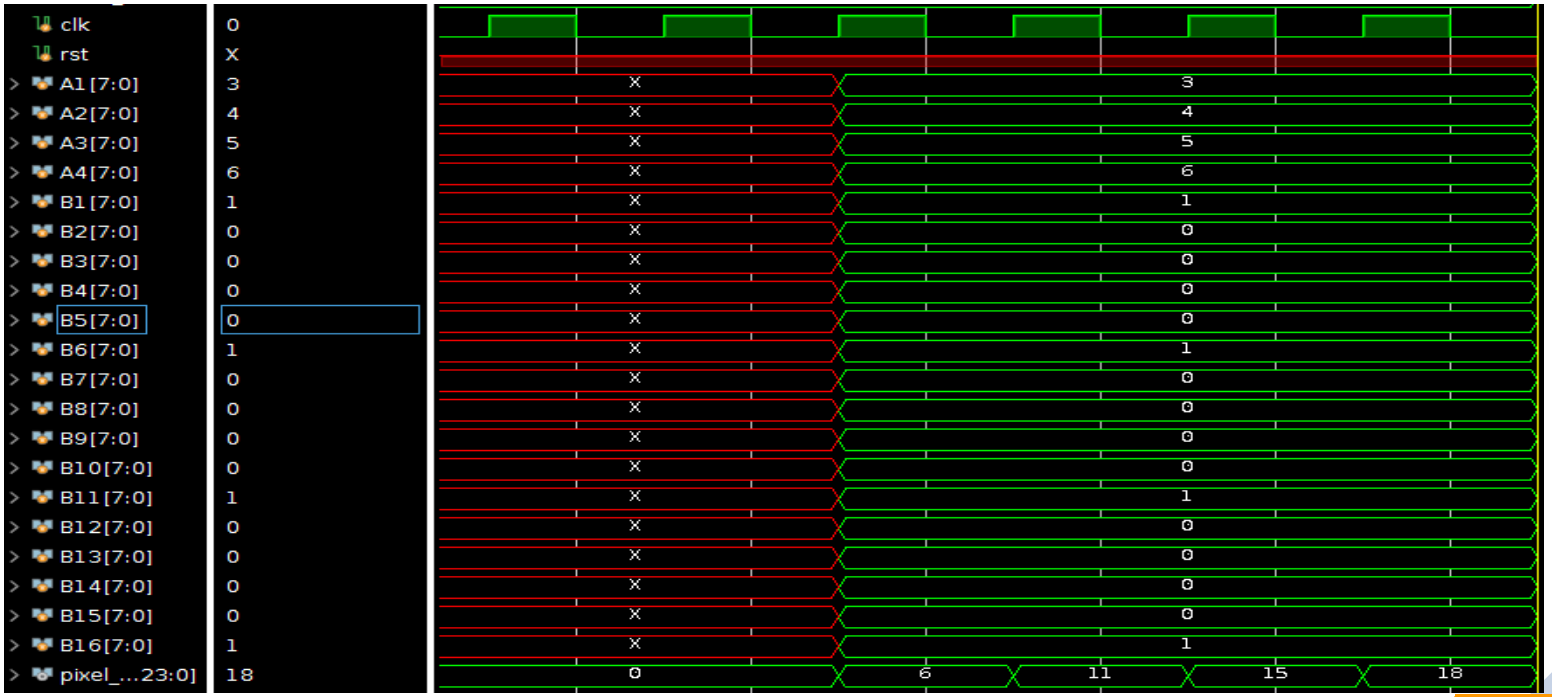


Proposed Method



Simulation Results

N cycles



Implementation Results

Original Image

Gaussian Image

FUTURE WORK

- Replace adders and multipliers with their efficient counterparts.
- Include Compressed Sparse Row.

CONCLUSION

- AND gate in paper is replaced by Bitwise OR to get required results.
- Convolution is also implemented with same architecture and a new one.
- Significant improvement in delay and number of clock cycles has been found with respect to traditional structure.

REFERENCES

[1] R. Jia, T. Xu and Y. Chang, "Scalable Systolic Array Multiplier Optimized by Sparse Matrix," *2021 IEEE 14th International Conference on ASIC (ASICON)*, Kunming, China, 2021, pp. 1-4, doi: 10.1109/ASICON52560.2021.9620326.

[2] Asgari B , Hadidi R , Kim H . Proposing a Fast and Scalable Systolic Array for Matrix Multiplication, 2020 IEEE 28th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM). IEEE, 2020.

[3] Petkov N. Systolic arrays for matrix I\/O format conversion[J]. Electron. Lett. 1988.

[4] Tang L, Cai G, Zheng Y, et al. A resource and performance optimization reduction circuit on FPGAs[J]. IEEE Transactions on Parallel and Distributed Systems, 2020, 32(2): 355-366.

[5] R. W. Means, "A new two-dimensional systolic array for image processing and neural network applications," *IJCNN-91-Seattle International Joint Conference on Neural Networks*, Seattle, WA, USA, 1991, pp. 925 vol.2-, doi: 10.1109/IJCNN.1991.155582.


thank you