PSG INSTITUTE OF TECHNOLOGY AND APPLIED RESEARCH
NEELAMBUR ,COIMBATORE

FINAL YEAR PROJECT REVIEW - 3

DESIGN AND IMPLEMENTATION OF EFFICIENT PROCESSING ELEMENT ARRAY FOR CNN ACCELERATOR

Review Date: 10-05-2025

Project Guide:
Dr. M. Jayasanthi
Professor, ECE

Team members:
1.Nithila Shri R P – (715521106032)
2.Sarveshware S (715521106042)
3.Shanmuga Priya M- (715521106043)
4.Durai Murugan S - (715521106302)

PROBLEM STATEMENT

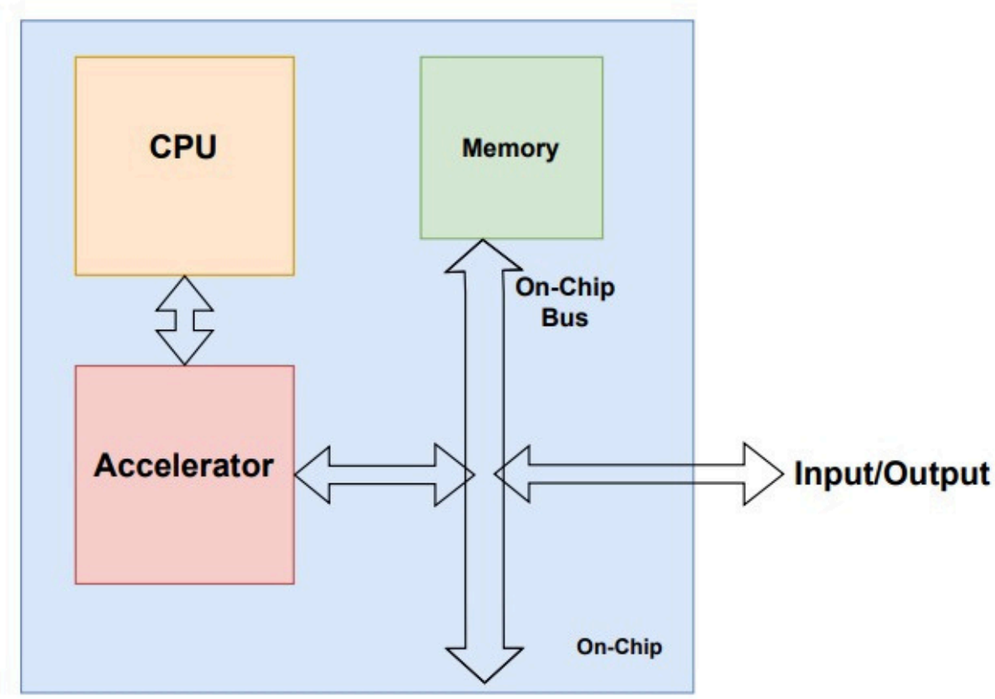
The project aims to address the computational complexity and power consuming challenges of deploying CNN accelerator on resource-constrained devices by optimizing their performance using a Processing Element (PE) Array architecture.

OBJECTIVES

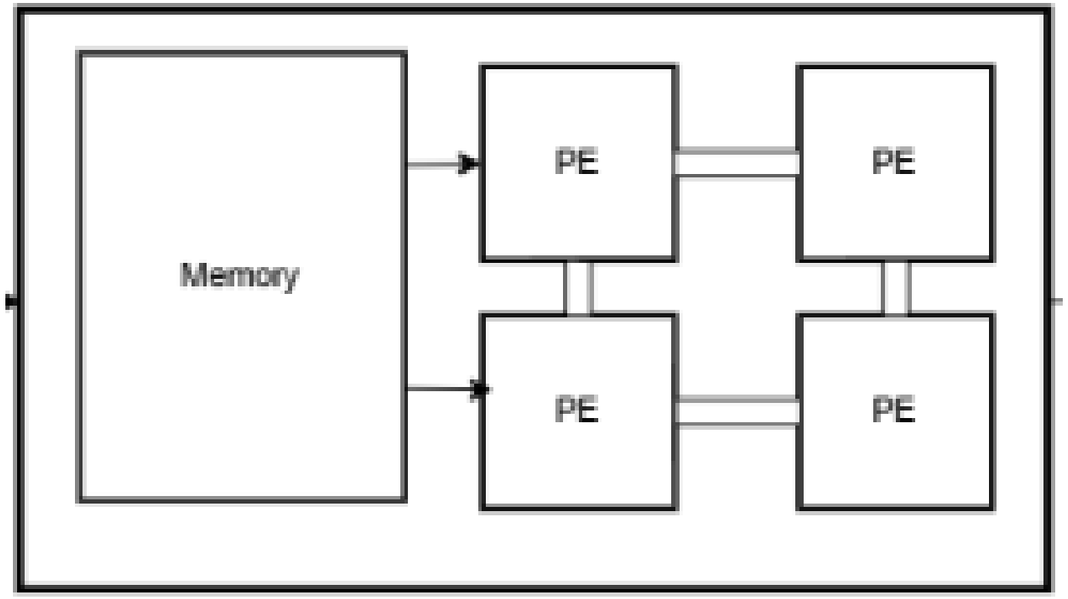
- To optimize data movement and computation.
- To minimize power consumption of processing element.
- To reduce the area of the PE array.

PROPOSED WORK

CNN Accelerator Chip

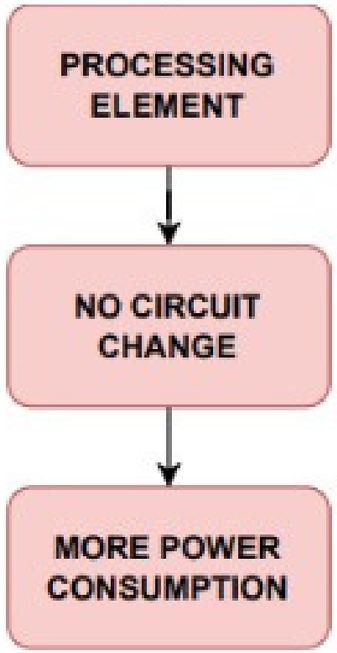


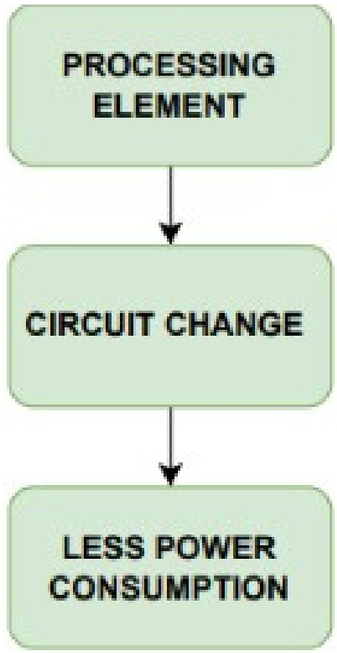
CNN Processing element array



PROPOSED WORK

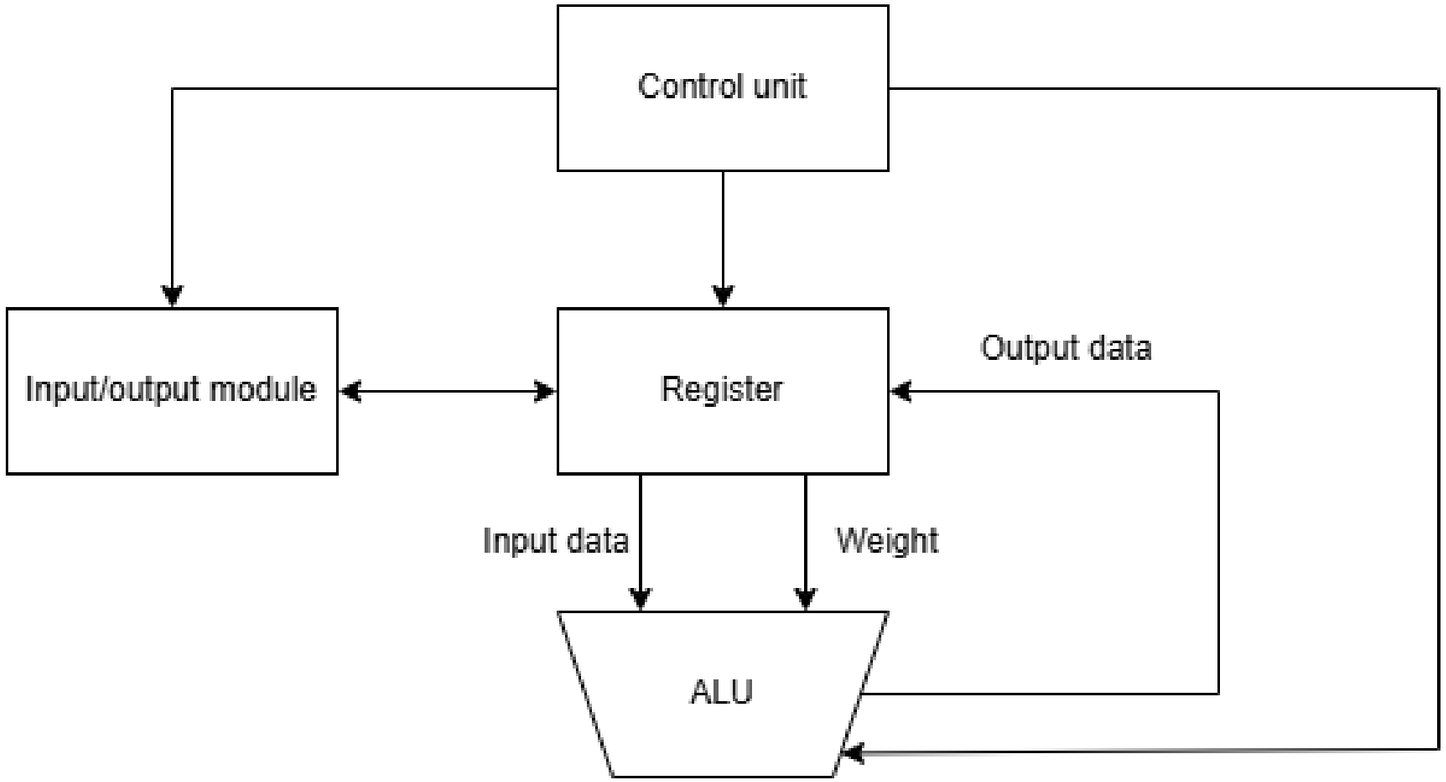
Idea Flow





PROPOSED WORK

Block diagram of a single processing element



ALU Operations :

's' input determines the operation performed by the ALU

- s = 3'b000 → Addition (In Hybrid Adder)
- s = 3'b001 → Multiplication (Dadda Multiplier)
- s = 3'b010 → Bitwise AND
- s = 3'b011 → Bitwise OR
- s = 3'b100 → Bitwise NOT of A
- s = 3'b101 → Bitwise NOT of B
- Default case → No operation → Output remains unchanged (C <= C).

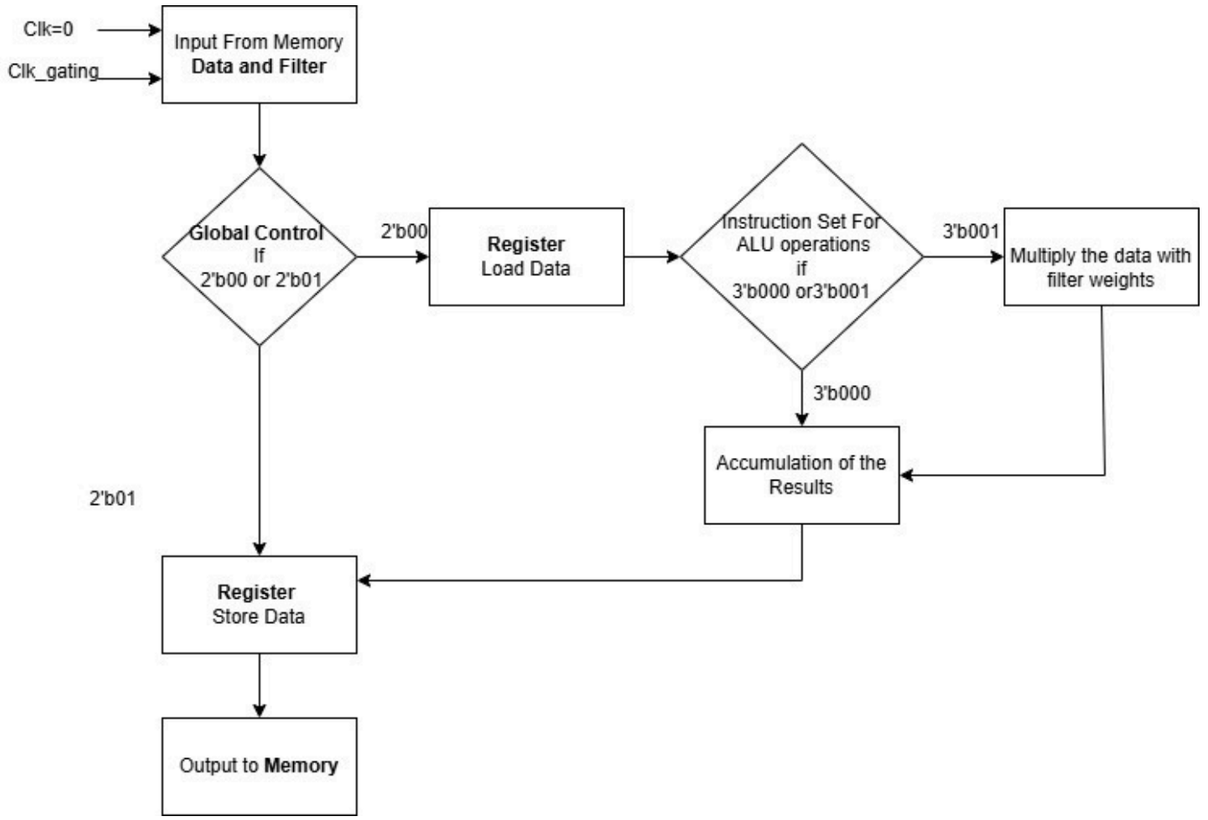
Register Operations :

'cui' signal controls how data flows in and out of Register

- cui = 2'b00 → Load input data
- cui = 2'b01 → Store input data
- cui= 2'b10 → Load output data
- cui= 2'b11 → Store output data

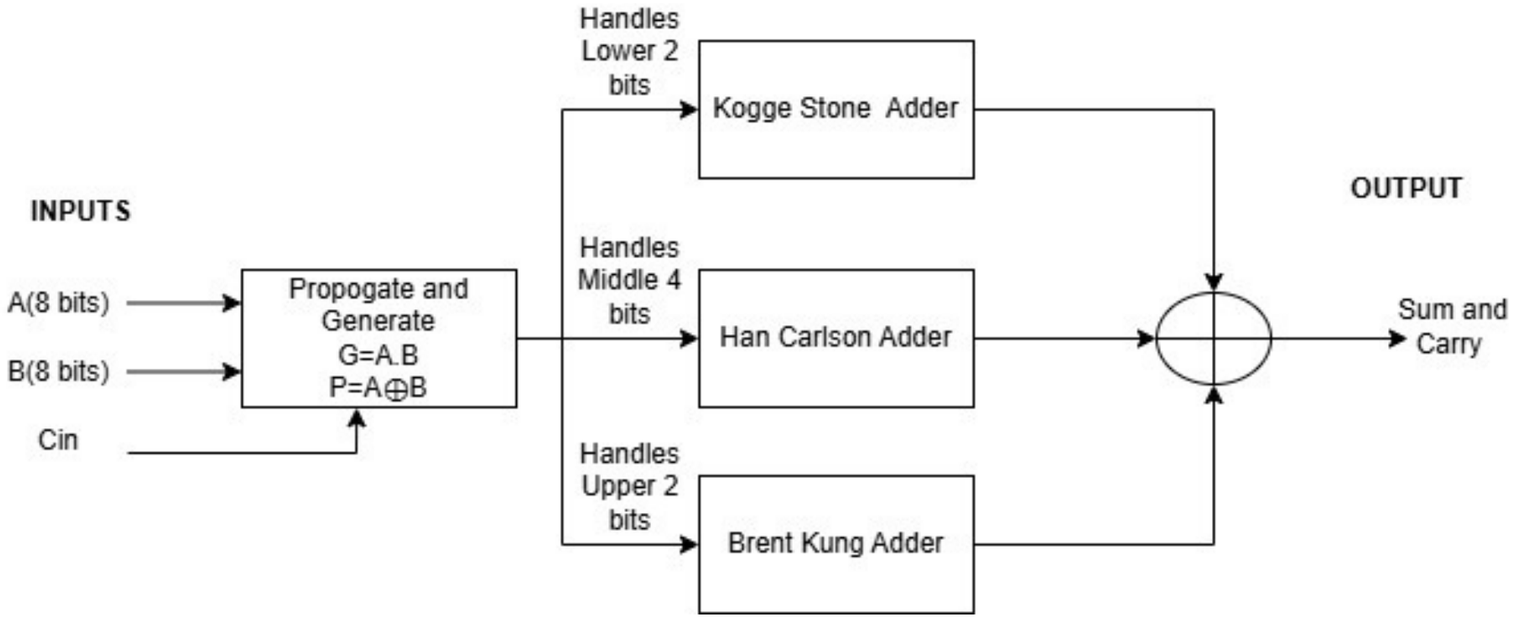
PROPOSED WORK

Convolution Operation Flow in PE Array



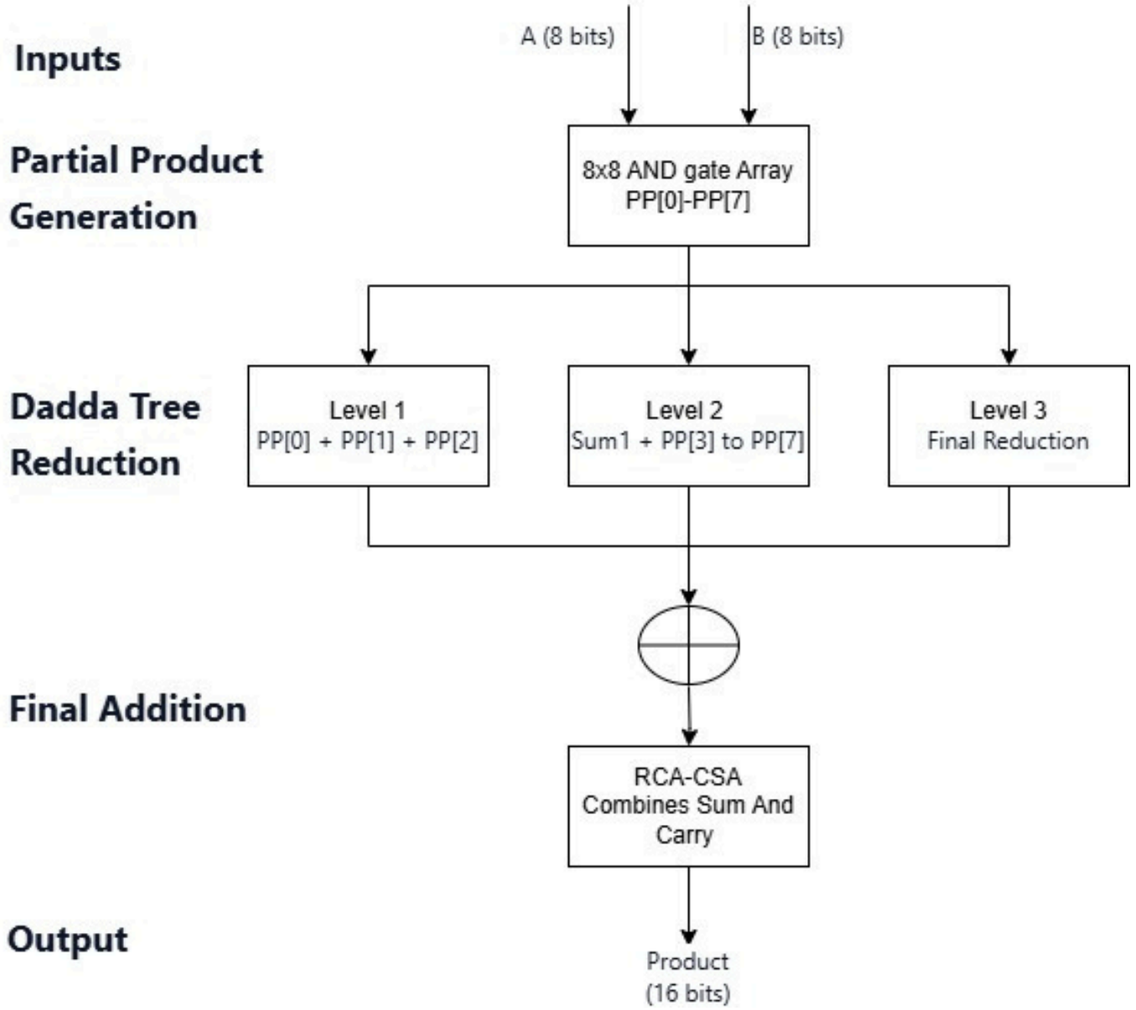
POWER MINIMIZATION METHODS

Block diagram of Hybrid Adder



POWER MINIMIZATION METHODS

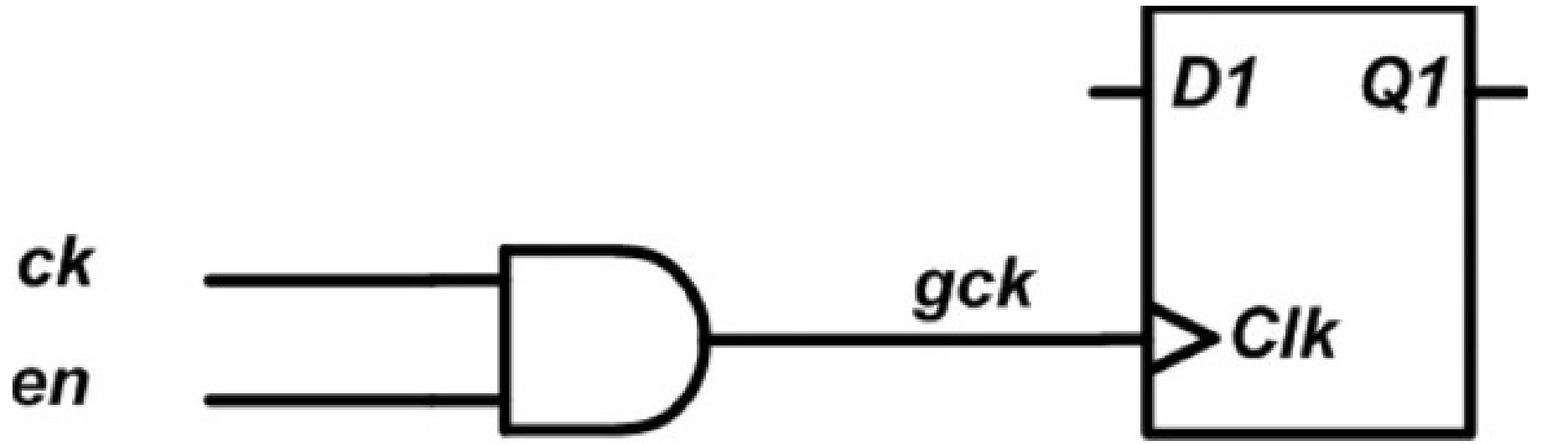
Block diagram of multiplier



- A Dadda multiplier is a hardware circuit that multiplies binary numbers using adders.
- The Dadda multiplier forms a partial product matrix using AND gates.
- Dadda multipliers are known for their efficiency, requiring fewer additions and logical operations compared to other multipliers, leading to lower power dissipation.

POWER MINIMIZATION METHODS

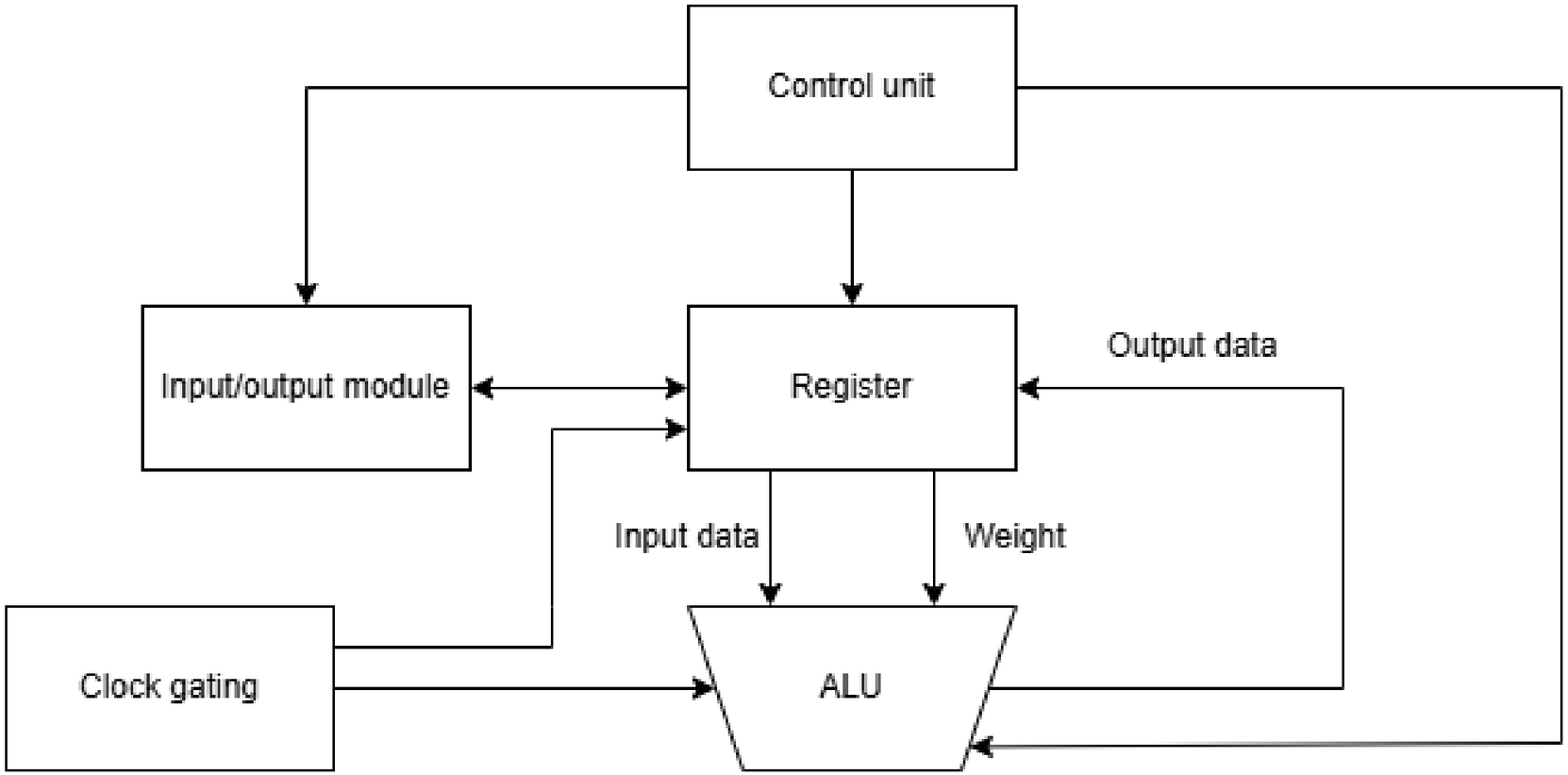
Clock gating- Minimizes the power



Clock gating is a technique that reduces power consumption by temporarily disabling the clock signal for parts of a circuit that aren't in use. It's a popular power management technique used in many synchronous circuits.

POWER MINIMIZATION METHODS

Clock gating- Single Processing Element



TOOLS USED

Xilinx Vivado:

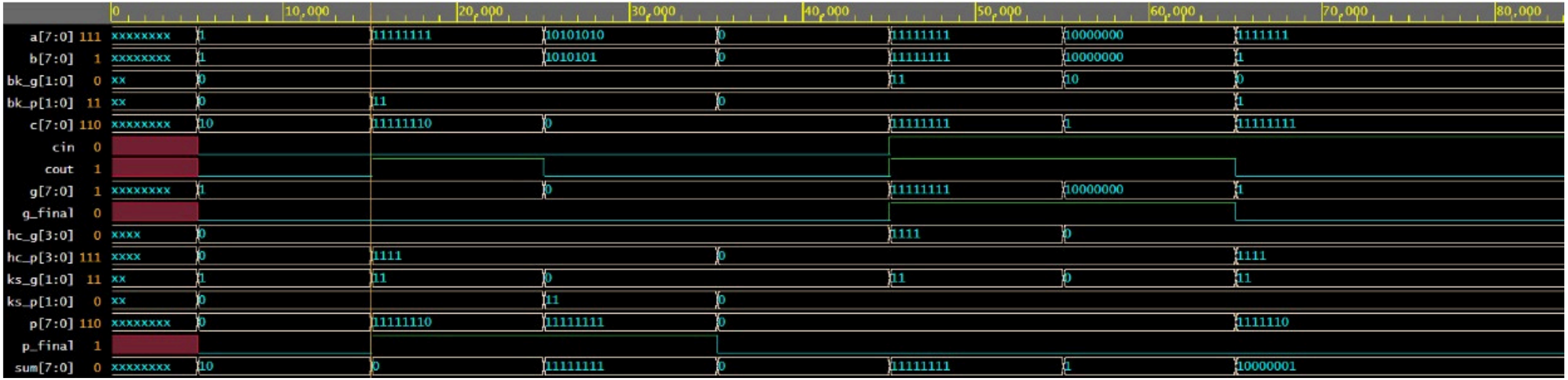
- Simulation
- Synthesis
- FPGA implementation

Cadence:

- Simulation
- Synthesis
- Power, area and timing report

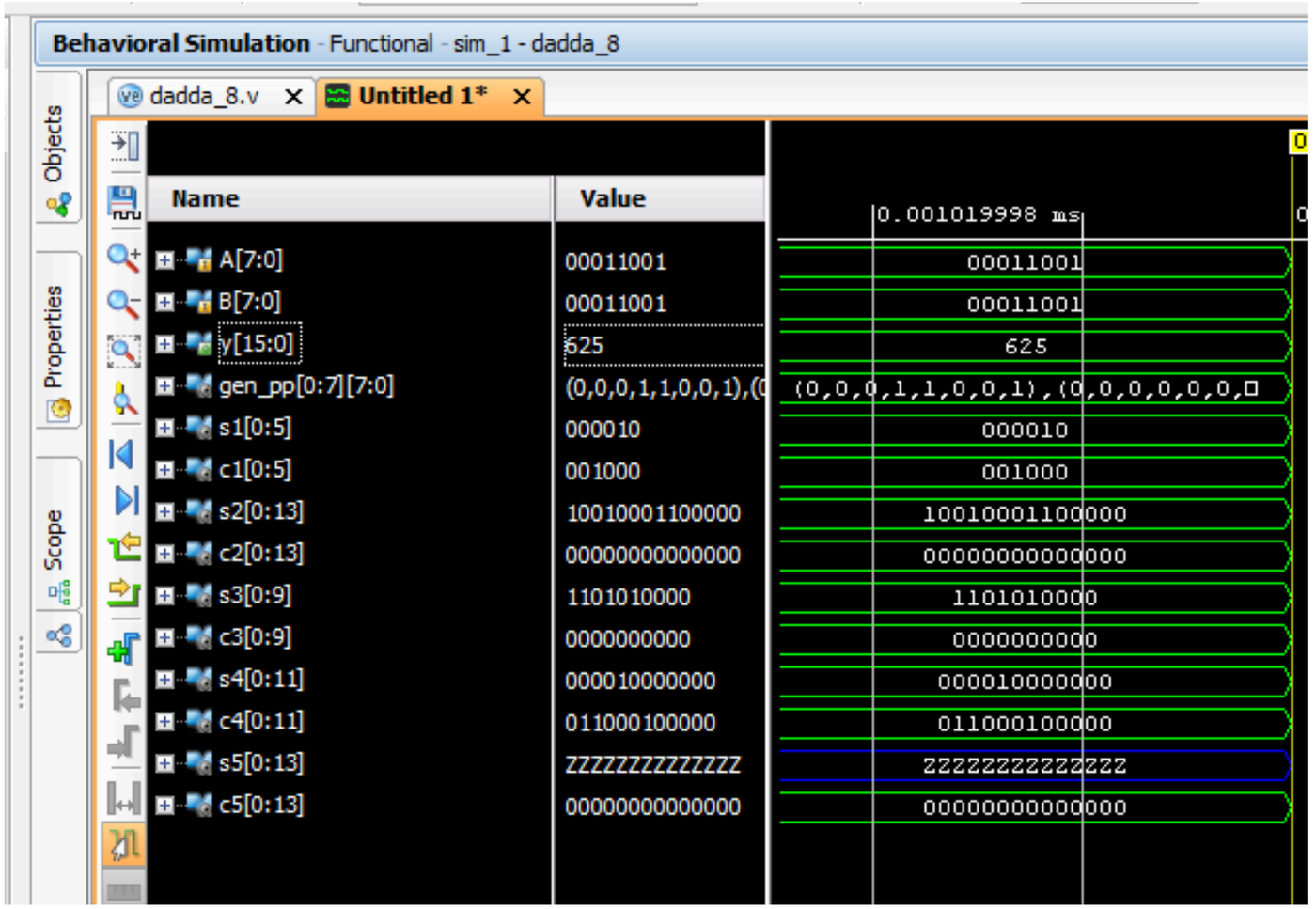
COMPLETED WORK WITH RESULTS

Simulation Output of Hybrid adder



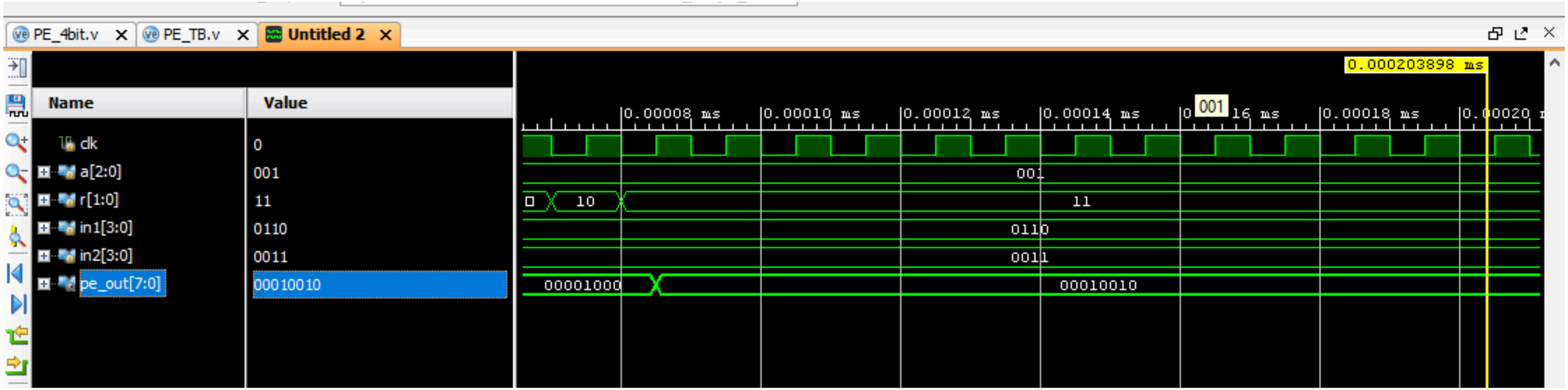
COMPLETED WORK WITH RESULTS

Simulation Output of Dadda Multiplier



COMPLETED WORK WITH RESULTS

Simulation Output of Processing element

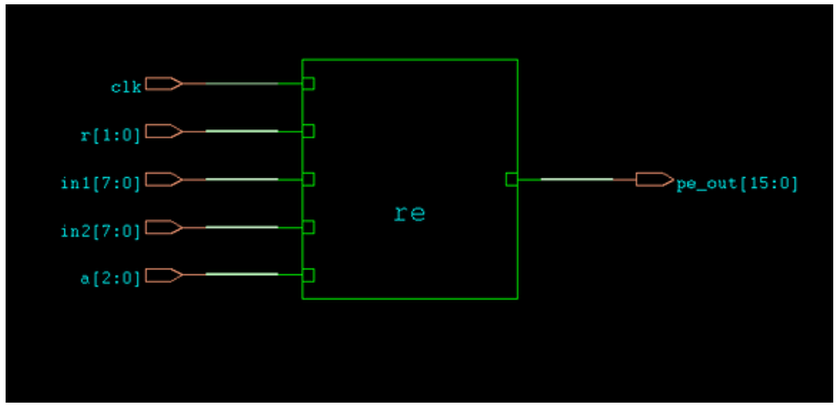


In the above simulation we have performed Addition of two 4-bit inputs and Multiplication of two 4-bit inputs.The operations are:

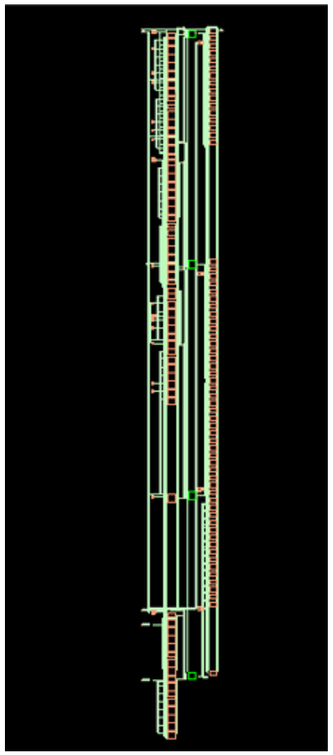
- Addition: in1=5 and in2=3 ----> pe_out=5+3=8
- Multiplication: in1=6 and in2=3 ----> pe_out=6*3=18

COMPLETED WORK WITH RESULTS

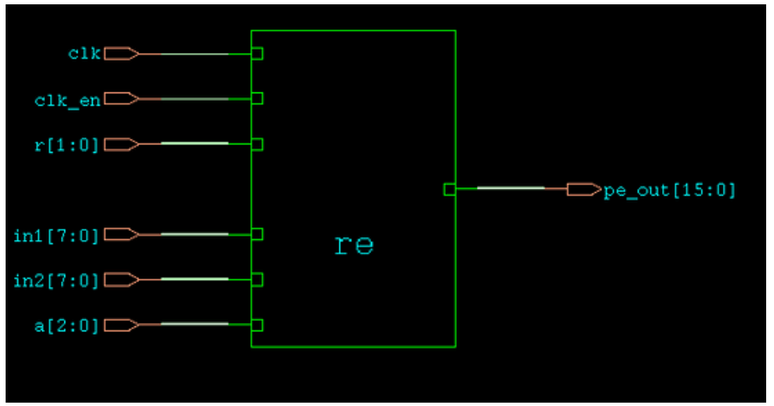
RTL Schematics



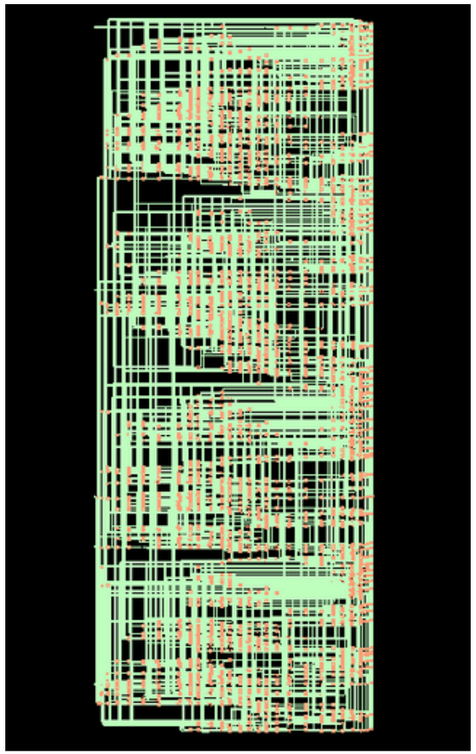
Normal Processing Element



Normal PE Array



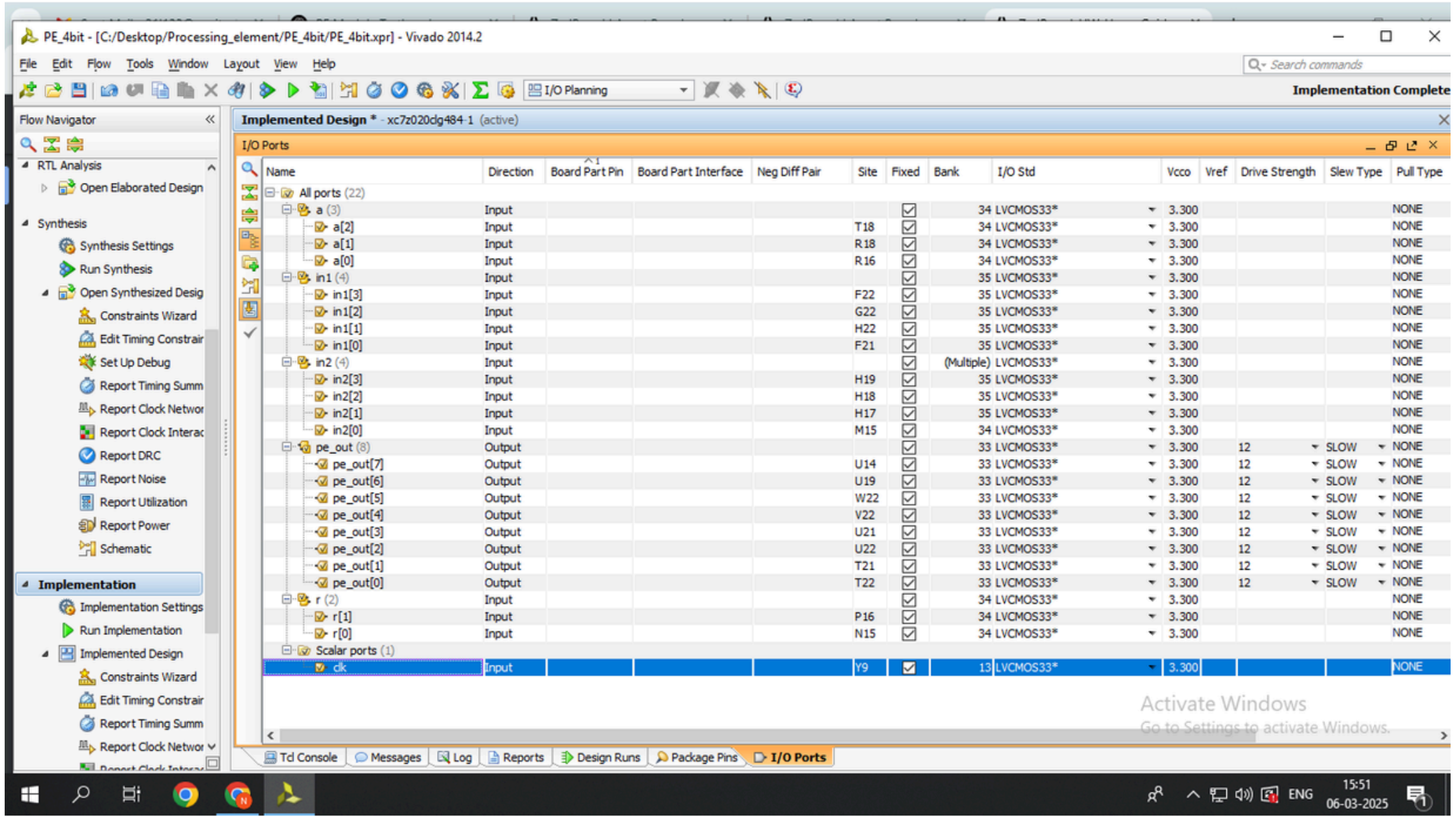
Low Power Processing Element



Low Power PE Array

COMPLETED WORK WITH RESULTS

FPGA Pin assignment diagram



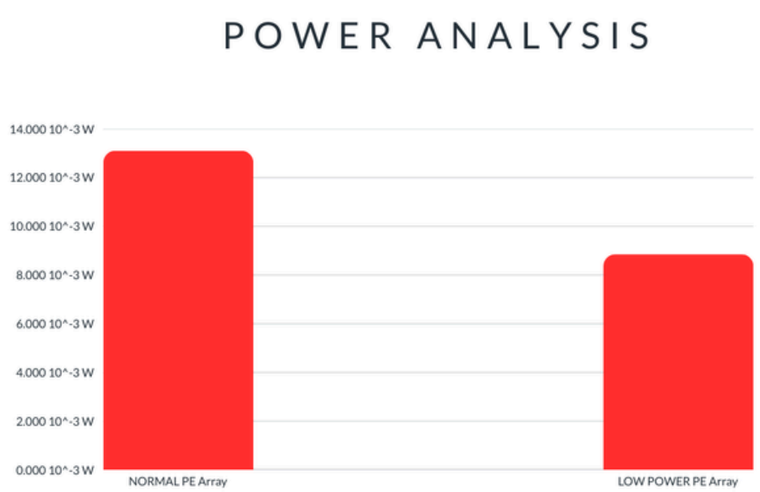
COMPLETED WORK WITH RESULTS

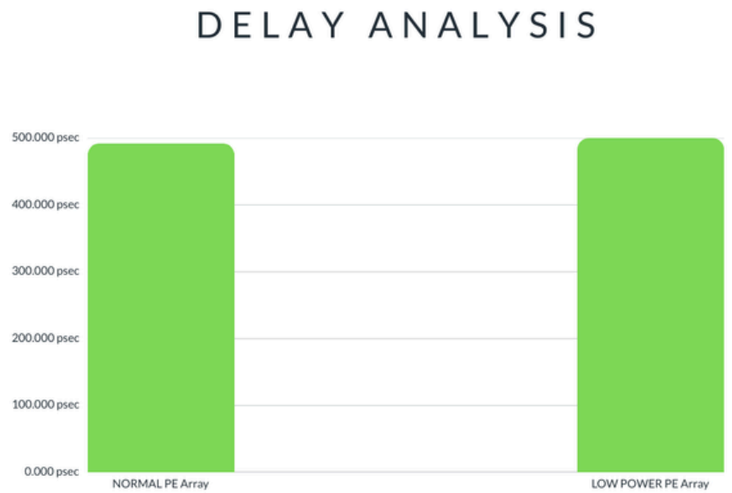
Comparison Table

	Actual Processing Element Array	**Optimized Processing Element Array**
Area	17328.469µm	16933.367 µm
Power	13.1807 milliwatts	8.84629 milliwatts
Delay	492 ps	500 ps

COMPLETED WORK WITH RESULTS

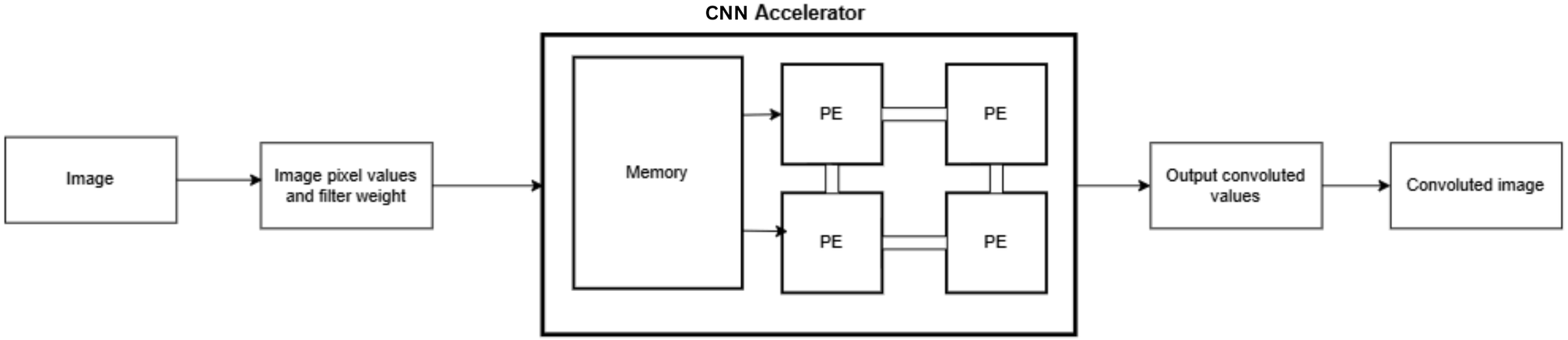
Graphical representation

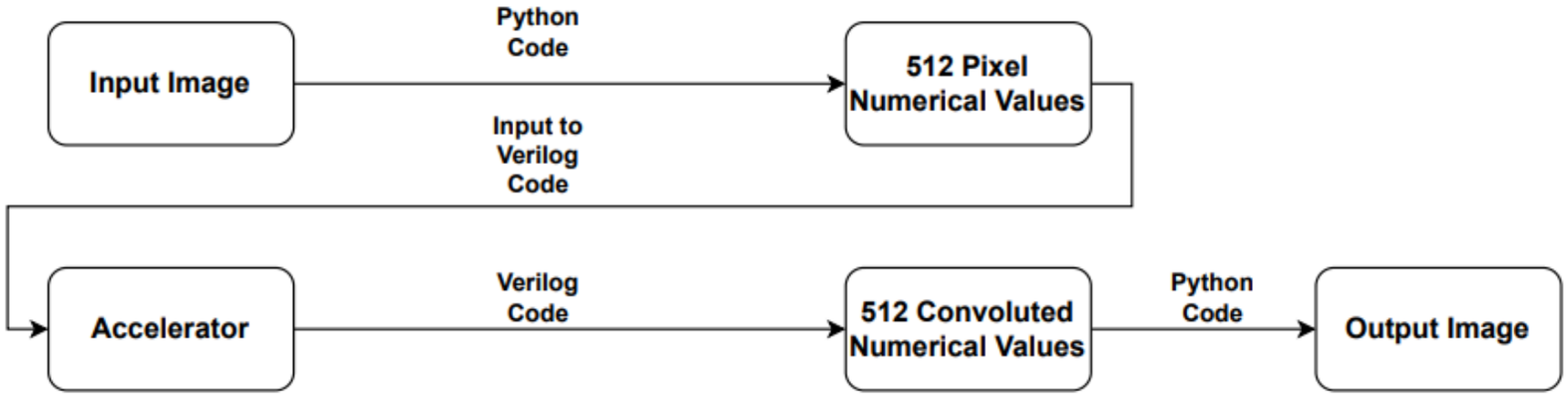




In summary, the optimized processing element array shows improvements in both area and power consumption. However, there is a slight increase in delay compared to the actual array, which might be acceptable depending on the specific performance requirements and trade-offs considered during the optimization process.

Real-World Use Cases of the Proposed CNN Accelerator





Real-World Use Cases of the Proposed CNN Accelerator

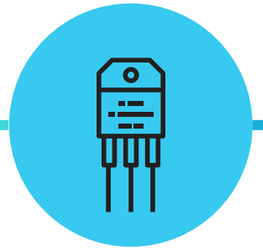
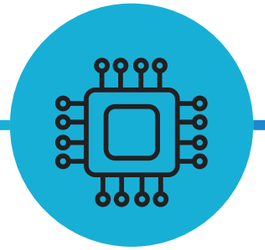
Simulation Output of PE Array

```
Smoothing Filter Output:
      3060
      3485
      3995
      4080
      3995
      3655
      3230
      3145
      3400
      3740
```

COMPLETED WORK WITH RESULTS

Application of Our Work-

TIMELINE

2024	JAN-FEB 2025	FEB 2025	APRIL-MAY2025	MAY-2025
				
Topic and Idea Pitching	1st Review : **Normal Architecture Design**	2nd Review : **Low Power Architecture Design**	3rd Review : **Report and Documentation**	**Completion of Project and Research Paper**

Real-World Use Cases of the Proposed CNN Accelerator

1. Edge AI Devices

Smart cameras, drones, robotics – Low-power, real-time object detection.

2. Medical Imaging

Portable diagnostic tools – Fast, energy-efficient analysis of X-rays/MRI scans.

3. Autonomous Vehicles

Lane/pedestrian detection – High-speed processing with minimal power.

4. IoT & Smart Farming

Crop disease monitoring – TinyML integration for low-cost sensors.

5. Surveillance Systems

24/7 anomaly detection – Power-efficient CCTV processing.

REFERENCES

- Li, T., Jiang, HL., Mo, H. et al. Approximate Processing Element Design and Analysis for the Implementation of CNN Accelerators. J. Comput. Sci. Technol. 38, 309–327 (2023).

- H. Xiao, H. Xu, X. Chen, Y. Wang and Y. Han, "Fast and High-Accuracy Approximate MAC Unit Design for CNN Computing," in IEEE Embedded Systems Letters, vol. 14, no. 3, pp. 155-158, Sept. 2022, doi: 10.1109/LES.2021.3137335.

- Fasih Ud Din Farrukh1, Tuo Xie1, Chun Zhang2, Zhihua Wang1, Fellow, IEEE 1 Institute of Microelectronics, Tsinghua University, Beijing 100084, China. 2 Research Institute of Tsinghua University in Shenzhen, ShenZhen 518055, China.

- Zhang, C.; Wang, X.; Yong, S.; Zhang, Y.; Li, Q.; Wang, C. An Energy-Efficient Convolutional Neural Network Processor Architecture Based on a Systolic Array. Appl. Sci. 2022, 12, 12633. https://doi.org/10.3390/ app122412633

- Y. Choi, D. Bae, J. Sim, S. Choi, M. Kim and L. -S. Kim, "Energy-Efficient Design of Processing Element for Convolutional Neural Network," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 64, no. 11, pp. 1332-1336, Nov. 2017, doi: 10.1109/TCSII.2017.2691771.

- Awad, O. M., Mahmoud, M., Edo, I., Zadeh, A. H., Bannon, C., Jayarajan, A., Pekhimenko, G., & Moshovos, A. (2020). *FPRaker: A processing element for accelerating neural network training* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2010.08065

REFERENCES

- Author(s). (2019, May). A solution to optimize multi-operand adders in CNN architecture on FPGA. In 2019 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. xx–xx). IEEE. https://doi.org/10.1109/ISCAS.2019.8702777

- Zhou, Y., Yan, J., Zhou, Y., Shao, Z., & Chen, J. (2024). Stochastic-binary hybrid spatial coding multiplier for convolutional neural network accelerator. IEEE Transactions on Nanotechnology, 23, 600–605. https://doi.org/10.1109/TNANO.2024.3444278

- El Khaili, M. (2014, October). 1354-bit processing unit design using VHDL structural modeling for multiprocessor architecture. IRACST – Engineering Science and Technology: An International Journal (ESTIJ), 4(5), 1354–1359. ISSN:

- Vinutha, C. R., Bharathi, M., & Divya, D. (n.d.). A survey on Brent-Kung, Han-Carlson and Kogge-Stone parallel prefix adders for their area, speed and power consumption

- Son, H., Na, Y., Kim, T., Al-Hamid, A. A., & Kim, H. (2021). CNN accelerator with minimal on-chip memory based on hierarchical array. In Proceedings of the 2021 18th International SoC Design Conference (ISOCC) (pp. 411–412). IEEE. https://doi.org/10.1109/ISOCC53507.2021.9613997

- Akin, B. (2019). FPGA-based CNN accelerator architecture (PE: processing element) [Figure]. ResearchGate.

- Son, H.-W., Al-Hamid, A. A., Na, Y.-S., Lee, D.-Y., & Kim, H.-W. (2024). CNN accelerator using proposed diagonal cyclic array for minimizing memory accesses. Computers, Materials & Continua, 77(2),. https://doi.org/10.32604/cmc.2023.038760

REFERENCES

- T. Yuan, W. Liu, J. Han and F. Lombardi, "High Performance CNN Accelerators Based on Hardware and Algorithm Co-Optimization," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 68, no. 1, pp. 250-263, Jan. 2021, doi: 10.1109/TCSI.2020.3030663.

- Kiningham, K., Graczyk, M., & Ramkumar, A. (n.d.). Design and analysis of a hardware CNN accelerator. Stanford University

- Shawahna, A., Sait, S. M., & El-Maleh, A. (2019). FPGA-based accelerators of deep learning networks for learning and classification: A review. IEEE Access, 7, 7823–7859. https://doi.org/10.1109/ACCESS.2018.2890150

- Sarkar, A. (2024). A novel FPGA-based CNN hardware accelerator: Optimization for convolutional layers using Karatsuba Ofman multiplier. arXiv. https://doi.org/10.48550/arXiv.2412.20393

- Munawar, M., Shabbir, Z., & Akram, M. (2023). Area, delay, and energy-efficient full Dadda multiplier. arXiv. https://doi.org/10.48550/arXiv.2307.05677

- Amin, M. A. A., Kartiwi, M., Yaacob, M., Hamidi, E. A. Z., Gunawan, T. S., & Ismail, N. (2022). Design of Brent Kung prefix form carry look-ahead adder. In 2022 8th International Conference on Wireless and Telematics (ICWT) (pp. 1–6). IEEE. https://doi.org/10.1109/ICWT55831.2022.9935137

- Sasireka, S., & Marimuthu, C. N. (2015). Implementation of Han-Carlson adder for error tolerant applications. International Journal of Electrical and Electronics Research, 3(4), 39–50. Retrieved from http://www.researchpublish.com

THANK YOU