Impact of Simulation Parameters on 3D NeuroSim Performance and Accuracy

Mohammadhossein Allahakbari

28/3/2025

Abstract

This report presents the experimental results obtained by varying key simulation parameters in Heterogeneous 3D NeuroSim. A baseline is established by using a VGG-8 model on the CIFAR-10 dataset and default parameters. We then analyze the impact on performance and accuracy when modifying:

1. **Memory Cell Type:** SRAM vs. RRAM.
2. **ADC Precision:** Different bit-precisions.
3. **ADC Sharing Degree:** Varying the number of columns sharing one ADC.
4. **Novel Mapping Parameter:** Enabling/disabling novel weight mapping.

1 Introduction

3D NeuroSim is used to simulate the hardware performance of neural network accelerators with heterogeneous 3D memory architectures. In the baseline configuration, the simulator was set with:

- Memory cell: RRAM
- ADC precision: 5 bits
- ADC sharing degree: 8 columns per ADC
- Novel Mapping: enabled

The baseline results and the effects of varying other parameters are discussed in subsequent sections.

2 Test Setup

All experiments were performed using a VGG-8 model on the CIFAR-10 dataset. The simulation metrics recorded many metrics but the focus of our study is on the following:

- **Test Accuracy:** Classification accuracy.
- **Hardware Performance:** Layer-by-Layer Read latency, total dynamic energy, and chip area.
- **Throughput:** Layer-by-Layer FPS.

Each experiment was run by only changing the parameter in question while leaving all the other parameters identical to the baseline.

3 Results and Analysis

3.1 RRAM vs. SRAM

This section compares the baseline results to the SRAM results.

Observations:

- **Test Accuracy:** Both RRAM and SRAM based architectures achieved identical results with 92% inference accuracy.
- **Latency:** RRAM layer-by-layer read latency is approximately 1.0362×10^6 ns, whereas SRAM latency is about 3.0463×10^6 ns.
- **Dynamic Energy:** Baseline total read dynamic energy is approximately 1.29545×10^7 pJ, compared to 5.34889×10^7 pJ in the SRAM run.
- **Throughput:** Baseline yields roughly 965 FPS versus approximately 328 FPS for SRAM.
- **Chip Clock Period and Area:** Baseline clock period is around 1.93 ns, while SRAM operates with a higher clock period (about 4.93 ns) and requires a larger chip area.

3.1.1 Dynamic Energy Comparison

Figure 1 illustrates the rescaled total read dynamic energy for both configurations (in units of 10^7 pJ) using a bar chart.

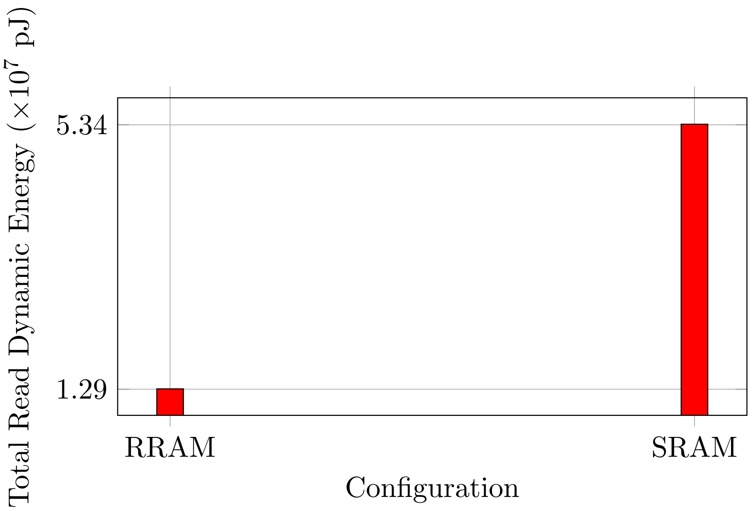

Figure 1: Total read dynamic energy for baseline vs. SRAM configurations.

Discussion: The comparison shows that the SRAM configuration has significantly higher dynamic energy consumption and increased latency, leading to reduced throughput compared to the baseline.

3.2 Impact of ADC Precision Variation

The ADC precision was varied to study its effect on both model accuracy and hardware performance. The baseline configuration uses a 5-bit ADC. Two additional tests were performed using 4-bit and 3-bit ADC precision. Lower ADC resolutions were not experimented with as the accuracy dropped to close to zero values.

Observations:

- **Test Accuracy:** The 3-bit configuration resulted in extremely low test accuracy (10%), the 4-bit ADC yielded an intermediate test accuracy of 56% and the baseline 5-bit configuration achieved 92% accuracy.
- **Read Latency:**
  - 3-bit ADC: 1.14262×10^6 ns
  - 4-bit ADC: 1.17485×10^6 ns
  - 5-bit ADC: $2.80909e \times 10^6$ ns
- **Dynamic Energy:**
  - 3-bit ADC: 2.09866×10^7 pJ
  - 4-bit ADC: 2.46563×10^7 pJ
  - 5-bit ADC: 3.19058×10^7 pJ
- **Throughput:**
  - 3-bit ADC: 875.18 FPS
  - 4-bit ADC: 851.175 FPS
  - 5-bit ADC: 355.987 FPS
- **Chip Area:**
  - 3-bit ADC: 1.07516×10^{-5} mm^2
  - 4-bit ADC: 1.22559×10^{-5} mm^2
  - 5-bit ADC: 2.04189×10^{-5} mm^2

3.2.1 Test Accuracy Comparison

Figure 2 shows a line graph that compares the test accuracy for the three ADC precision settings.

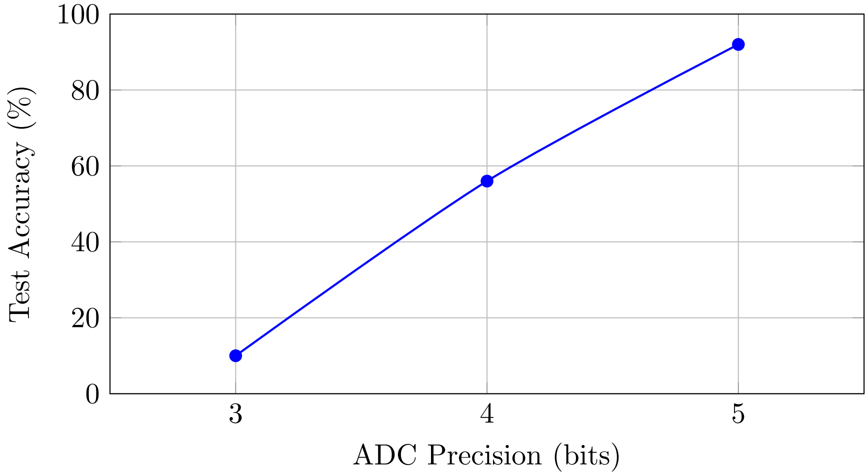

Figure 2: Test accuracy vs. ADC precision.

3.2.2 Dynamic Energy Comparison

Figure 3 illustrates the rescaled total read dynamic energy (in units of 10^7 pJ) for each ADC precision test using a bar chart.

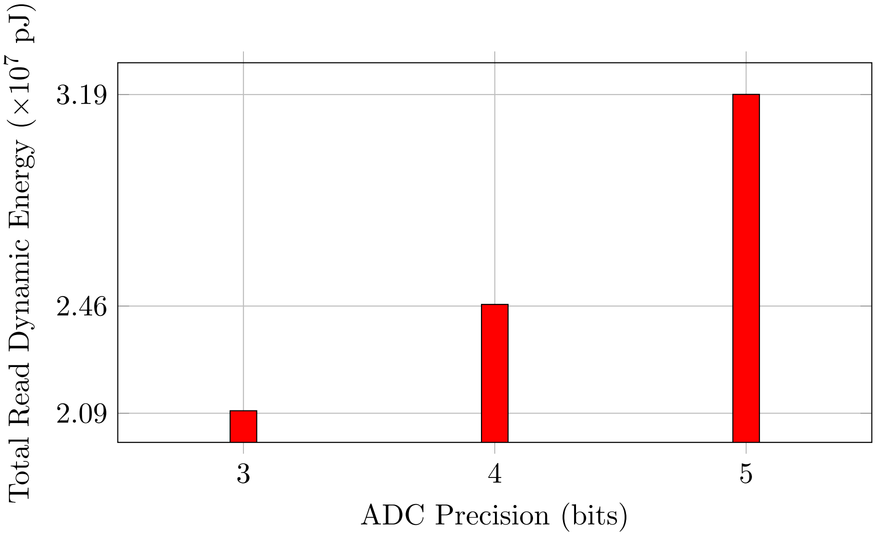

Figure 3: Total read dynamic energy vs. ADC precision.

Discussion: The line charts clearly show that reducing the ADC precision below the baseline (5-bit) negatively impacts the test accuracy while increasing the dynamic energy consumption. In particular, the 3-bit configuration leads to a severe accuracy drop (to 10%), making it unsuitable for the application. Although the 4-bit configuration performs better than the 3-bit case, it still underperforms compared to the 5-bit setting.

3.3 Impact of ADC Sharing Degree Variation

In addition to ADC precision, we investigated the effect of varying the ADC sharing degree. For these experiments, three sharing configurations were tested:

- **8 columns per ADC:** Baseline setting.
- **16 columns per ADC:** Higher sharing degree.
- **32 columns per ADC:** Max sharing degree.

Observations:

- **Test Accuracy:** All different sharing degrees acheived identical results with the accuracy of 92%, similar to the baseline.
- **Read Latency:**
  - 8-way sharing: 2.80909×10^6 ns
  - 16-way sharing: 1.31549×10^6 ns
  - 32-way sharing: 1.61826×10^6 ns
- **Dynamic Energy:**
  - 8-way sharing: 3.19058×10^7 pJ
  - 16-way sharing: 3.95113×10^7 pJ
  - 32-way sharing: 5.87941×10^7 pJ
- **Throughput:**
  - 8-way sharing: 355.987 FPS
  - 16-way sharing: 760.175 FPS
  - 32-way sharing: 617.949 FPS
- **Chip Area:**
  - 8-way sharing: 2.04189×10^{-5} mm^2
  - 16-way sharing: 1.25323×10^{-5} mm^2
  - 32-way sharing: $1.55235e \times 10^{-5}$ mm^2

3.3.1 Dynamic Energy Comparison vs. ADC Sharing Degree

Figure 4 illustrates the rescaled total read dynamic energy (in units of 10^7 pJ) for the different sharing configurations using a bar chart.

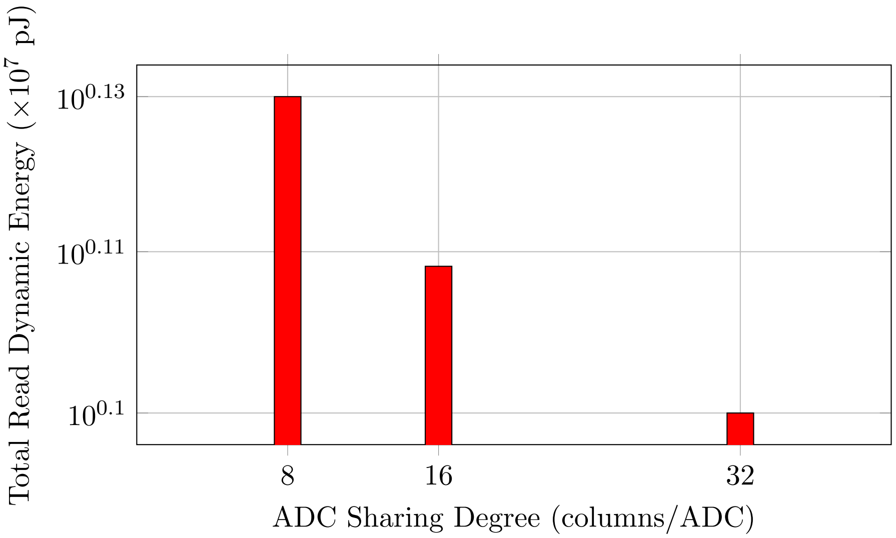

Figure 4: Total read dynamic energy vs. ADC sharing degree.

Discussion: The experiments compared ADC sharing configurations of 8, 16, and 32 columns per ADC. All configurations achieved a test accuracy of 92%. Notably, increasing sharing to 16 columns reduced the read latency from 2.81×10^6 ns (8-way) to 1.32×10^6 ns and nearly doubled the throughput to 760.175 FPS, albeit with a moderate rise in dynamic energy. In contrast, 32-way sharing slightly increased latency and decreased throughput while further raising energy consumption. Thus, 16-way sharing strikes an optimal balance between latency, energy, throughput, and chip area, as illustrated in Figure 4.

3.4 Impact of Novel Mapping Parameter

In this section we compare the baseline configuration (with `novelMapping` enabled) to a run using the conventional mapping method, i.e. with `novelMapping` disabled. Although both configurations achieve similar test accuracy (around 92%), there are notable differences in hardware performance.

Observations:

- **Test Accuracy:** Both configurations achieve approximately 92% accuracy.
- **Read Latency:** Both configurations achieve identical total read latency of 2.80909×10^6ns.
- **Dynamic Energy:** The baseline configuration (novelMapping enabled) exhibits a total read dynamic energy of about 3.19058×10^7 pJ, whereas the non-novel mapping run consumes approximately 3.55067×10^7 pJ which is 11% higher.
- **Throughput:** Both configurations achieve identical throughput of 355.987 FPS
- **Chip Area:** Both configurations require $2.04189e-05mm^2$ of chip area.

3.4.1 Dynamic Energy Comparison: Baseline vs. Non-novel Mapping

To illustrate the difference in energy consumption, Figure 5 shows a bar chart comparing the rescaled total read dynamic energy (in units of 10^7 pJ) for the two mapping strategies.

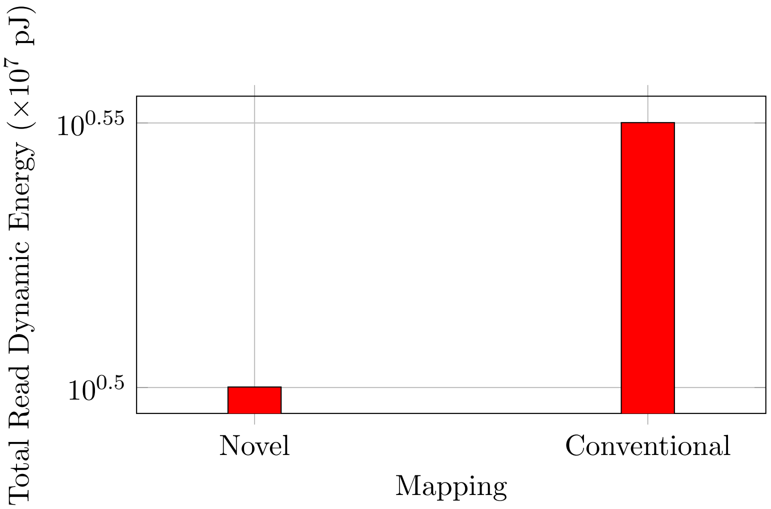

Figure 5: Total read dynamic energy for baseline (novelMapping enabled) vs. non-novel mapping configurations (rescaled, logarithmic scale).

Discussion: The bar chart indicates that while both configurations maintain similar test accuracy, the conventional (non-novel) mapping approach results in significantly higher dynamic energy consumption which is tied to the absence of data reuse in this configuration. This, in turn, implies higher latency and reduced throughput. Therefore, enabling the novel mapping parameter leads to improved energy efficiency and overall hardware performance, making it a favorable design choice.

4 Novel and Conventional Mapping Comparison

4.1 Conventional Mapping

In the conventional mapping approach, each 3D convolution kernel is unrolled into a vertical column, whereby the entire set of weights for a kernel (of dimension K×K×D) is concatenated into a single column within a large weight matrix. This method necessitates that, for every sliding window position during convolution, the same input feature-map (IFM) patch is repeatedly fetched from off-chip buffers and transferred via interconnects to the processing elements. Figure 6 illustrates how the kernels are unrolled.

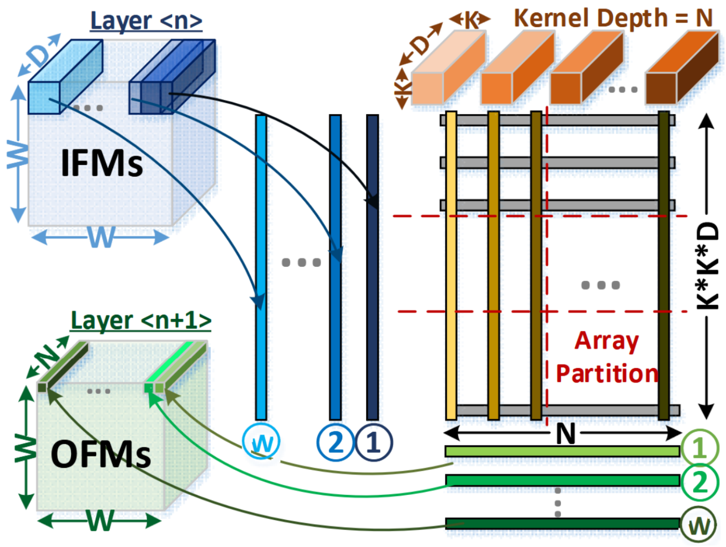

Figure 6: Conventional Mapping Scheme.

4.2 Novel Mapping

The novel mapping technique proposed in [2] partitions the kernel along its spatial dimensions. Instead of unrolling the entire 3D kernel into a single column, the weights corresponding to each spatial location within the kernel are allocated into separate sub-matrices. Each sub-matrix is then assigned to an individual processing element (PE). Under this scheme, the IFMs are distributed such that each PE receives only the specific input patch pertinent to its corresponding weight block. As the convolution operation advances, the reuse of IFMs is maximized through localized data transfers between adjacent PEs, where only the newly required activations are fetched from the higher-level buffers. Figure 7 illustrates how this scheme is implemented.

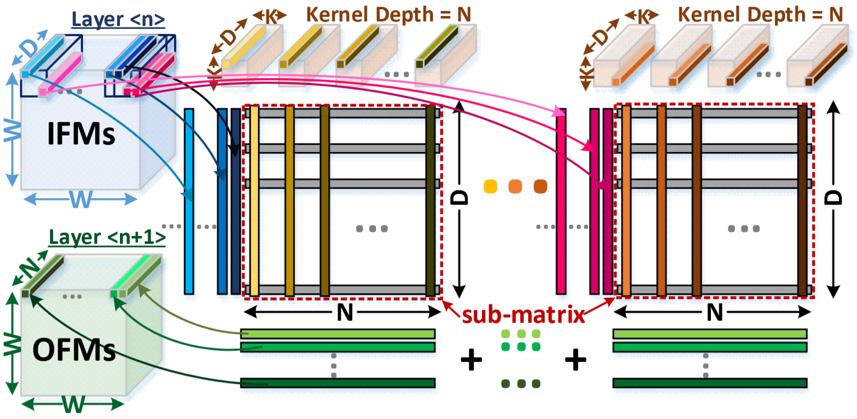

Figure 7: Novel Mapping Scheme.

4.3 Novel Mapping Benefits

As illustrated in 3.4, this spatially aware partitioning not only minimizes the distance and volume of data movement thereby reducing the latency and energy consumption of interconnects and buffers but also enhances overall system efficiency. The reduction in redundant data transfers leads to a 11% improvement in energy efficiency (as measured in TOPS/W) for networks such as VGG-8.

Conclusion

In summary, our experiments highlight the interplay between accuracy, energy consumption, and hardware performance in 3D NeuroSim under varying simulation parameters. RRAM-based designs offer lower latency, reduced dynamic energy, and higher throughput compared to SRAM, with no accuracy penalty. Increasing ADC precision improves classification accuracy but raises energy use and chip area, with a 5-bit ADC emerging as a practical trade-off. Adjusting the ADC sharing degree indicates that 16 columns per ADC balances energy consumption and throughput more effectively than 8 or 32 columns. Finally, enabling the novel mapping parameter reduces dynamic energy by approximately 11% while preserving the same accuracy and throughput as conventional mapping. These observations underscore the need for carefully selecting parameters based on the specific performance and efficiency requirements of neuromorphic hardware applications.

References

[1] X. Peng, W. Chakraborty, A. Kaul, W. Shim, M. S. Bakir, and S. Datta. Benchmarking monolithic 3d integration for compute-in-memory accelerators: Overcoming adc bottlenecks and maintaining scalability to 7nm or beyond. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020.

[2] X. Peng, R. Liu, and S. Yu. Optimizing weight mapping and data flow for convolutional neural networks on rram based processing-in-memory architecture. *IEEE*, 2019.

[3] X. Peng, R. Liu, and S. Yu. User manual of 3d neurosim v1.0, n.d. Accessed: March 20, 2025.