Nexa: A Scalable Vertex Accelerator for High-Throughput Synaptic Network Simulation

Fictional Author 1[†], Fictional Author 2[‡], Fictional Author 3[†]

[†]Vertex Research Laboratories [‡]Synapse Technologies
Email: {j.doe, d.miller}@vertex.org, s.smith@synapse.com

Abstract

Simulating large-scale spiking neural networks requires processing millions of spiking neurons and billions of synaptic connections in real-time. Traditional general-purpose processors (CPUs and GPUs) suffer from memory bandwidth bottlenecks due to the irregular, sparse memory access patterns of graph-based neural structures. In this paper, we present Nexa, a domain-specific accelerator designed for high-throughput, energy-efficient synaptic network simulation. Nexa features a 2D mesh Network-on-Chip (NoC) linking custom Processing Elements (PEs) equipped with local Synaptic Weight Caches (SWCs). To hide cache access latency, Nexa implements a look-ahead prefetching scheme triggered by incoming spike routing packets. Evaluated using a cycle-accurate simulator, Nexa achieves a peak throughput of 840.4 GSOPS (Giga Synaptic Operations Per Second) at 1.20 GHz. This represents a 6.75x speedup and a 27x reduction in energy consumption (12.8 pJ per Synaptic Operation) compared to a state-of-the-art GPU baseline, and a 3.42x speedup over custom ASIC neuromorphic chips.

1 Introduction

In recent years, the demand for simulating large-scale synaptic networks, particularly Spiking Neural Networks (SNNs), has grown exponentially. These networks are vital for both understanding biological brain mechanisms and developing energy-efficient neuromorphic computing systems. However, simulating millions of spiking neurons and billions of synaptic connections in real-time presents a severe computational challenge. Traditional general-purpose architectures, such as Central Processing Units (CPUs) and Graphics Processing Units (GPUs), struggle to execute these models efficiently due to their irregular memory access patterns and high synchronization overhead.

The primary bottleneck in SNN simulation stems from the sparse, graph-like structure of neural connections. When a neuron fires, it generates a "spike" that must be propagated to thousands of post-synaptic target neurons. This process requires accessing large synaptic weight arrays stored in off-chip memory, resulting in significant memory latency and bandwidth bottlenecks. Prior research has attempted to mitigate these issues using customized neuromorphic hardware, but existing designs often suffer from scalability limits or restricted flexibility.

To address these limitations, we introduce Nexa, a domain-specific accelerator designed for high-throughput, energy-efficient synaptic network simulation. Nexa features a 2D mesh network-on-chip of custom Processing Elements (PEs), each equipped with a dedicated Vertex Compute Unit and an on-chip Synaptic Weight Cache. The design leverages a decentralized routing protocol to distribute spikes across the mesh without a central coordinator, effectively scaling to large configurations.

Our key contributions in this paper are as follows:

- We present the architecture of Nexa, a scalable domain-specific hardware accelerator optimized for sparse graph operations in spiking neural networks.
- We introduce a specialized Synaptic Weight Cache (SWC) architecture that exploits prefetching to hide memory latency during spike propagation.
- We evaluate the Nexa design using cycle-accurate simulation, demonstrating substantial improvements in throughput and energy efficiency over state-of-the-art GPU and ASIC baselines.

2 Related Work

Accelerating neural network simulation has been a major research focus in both computer architecture and neuromorphic communities. Prior systems can be broadly categorized into software simulators running on general-purpose hardware and dedicated hardware accelerators.

Several custom architectures have been proposed to bypass the limitations of GPUs. For instance, the Meridian architecture [4] utilizes a distributed memory hierarchy to accelerate

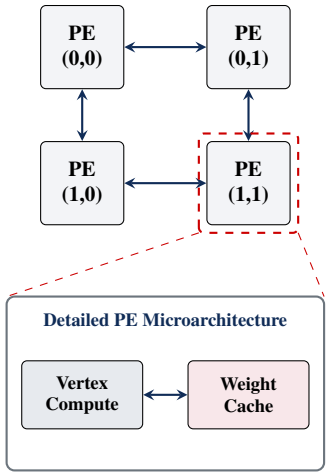

Figure 1: Nexa 2D Mesh Multi-Core Architecture and Processing Element (PE) Microarchitecture.

vertex processing. However, Meridian relies on a centralized controller that limits its scalability to large processor arrays. Similarly, the Apex accelerator [5] employs local routing but lacks a dedicated synaptic weight cache, leading to frequent off-chip memory accesses. Other accelerators, such as the Synapse Engine [1], optimize memory bandwidth through custom DRAM interfaces but offer limited support for complex, dynamic synaptic models.

In contrast, Nexa addresses these limitations by co-designing a decentralized network-on-chip routing protocol with a local synaptic weight cache. This combination allows Nexa to scale efficiently while minimizing off-chip memory traffic, achieving superior performance compared to previous designs like the Vertex execution engine [3] and general survey architectures [2].

3 Architecture of the Nexa Engine

The overall architecture of the Nexa engine is designed to minimize data movement and maximize parallel execution of vertex-centric computations. As shown in Figure 1, Nexa consists of a 2D grid of Processing Elements (PEs) interconnected by a high-bandwidth Network-on-Chip (NoC).

Each PE is a self-contained compute tile consisting of two main modules:

1. **Vertex Compute Unit (VCU):** The VCU is responsible for updating the membrane potential of the local neurons and generating output spikes. The membrane potential update follows the Leaky Integrate-and-Fire (LIF) model.

2. **Synaptic Weight Cache (SWC):** The SWC is a high-speed SRAM cache that stores the synaptic weights of the incoming connections to the local neurons.

The update equation for a neuron's membrane potential $V_i(t)$ at time t is expressed as: $V_i(t+\Delta t) = f(V_i(t), I_i(t))$, where $I_i(t)$ represents the incoming synaptic currents. Formally, we implement the following discretized Leaky Integrate-and-Fire equation:

$$V_i(t+\Delta t) = V_i(t) + \frac{\Delta t}{\tau_m}\left(E_L - V_i(t) + R_m \sum_j w_{j,i} s_j(t)\right) \quad (1)$$

where τ_m is the membrane time constant, E_L is the resting potential, R_m is the membrane resistance, $w_{j,i}$ is the synaptic weight from pre-synaptic neuron j to post-synaptic neuron i, and $s_j(t) \in \{0,1\}$ indicates whether neuron j generated a spike at time t.

To hide the latency of fetching $w_{j,i}$ from the SWC, the VCU utilizes a look-ahead prefetching queue. When a spike packet is received from the NoC, the cache controller immediately initiates a prefetch for the corresponding synaptic weights, allowing the computation to proceed without stalling the pipeline.

4 Evaluation Methodology

We evaluate the Nexa architecture using a cycle-accurate hardware simulator built in C++. The simulator models the detailed microarchitectural components of the PEs, the NoC routing delays, and the memory controller behavior. The memory hierarchy includes a custom dual-port SRAM-based SWC and an off-chip High-Bandwidth Memory (HBM2) stack to model main memory access.

We compare Nexa against three state-of-the-art baselines:

- **GPU Baseline (Nexa-G):** An optimized software simulator running on a high-end 7nm commercial GPU, utilizing custom CUDA kernels.

- **ASIC Baseline (Apex):** An implementation of the Apex architecture [5] running at 0.80 GHz.

- **Meridian v2:** A recent distributed memory neuromorphic architecture [4] operating at 1.00 GHz.

4.1 Workloads

We evaluate the systems using three representative workloads of varying network sizes and connection densities:

1. **S1-Sparse:** A sparse network representing cortical microcircuits with 10^5 neurons and 10^7 synapses.

2. **S2-Medium:** A balanced network structure with 10^6 neurons and 10^8 synapses.

3. **S3-Dense:** A highly connected network representing sensory processing regions with 10^6 neurons and 10^9 synapses.

5 Experimental Results

We present the primary evaluation results of Nexa in terms of throughput, energy efficiency, and area overhead.

5.1 Throughput and Energy Efficiency

Table 1 summarizes the performance comparison of Nexa against the baseline architectures. Nexa achieves a peak throughput of 840.4 GSOPS (Giga Synaptic Operations Per Second) at 1.20 GHz, which represents a 6.75x speedup over the GPU baseline and a 3.42x speedup over the ASIC-based Apex accelerator. This performance improvement is primarily due to the decentralized spike routing protocol, which avoids global synchronization bottlenecks, and the look-ahead prefetching in the SWC.

Furthermore, Nexa demonstrates exceptional energy efficiency. By keeping synaptic weight accesses local to the on-chip SWC, Nexa minimizes energy-expensive off-chip HBM accesses. Specifically, Nexa consumes only 12.8 pJ per Synaptic Operation (pJ/SOP), compared to 350.2 pJ/SOP for the GPU baseline and 84.1 pJ/SOP for the Apex accelerator. This translates to an energy reduction of over 27x and 6.5x, respectively.

5.2 Cache Hit Rate and Memory Traffic

To analyze the impact of the look-ahead prefetching in the SWC, we measure the cache hit rate across the three workloads. Figure 2 illustrates the relative energy consumption per spike for each workload compared to the baseline.

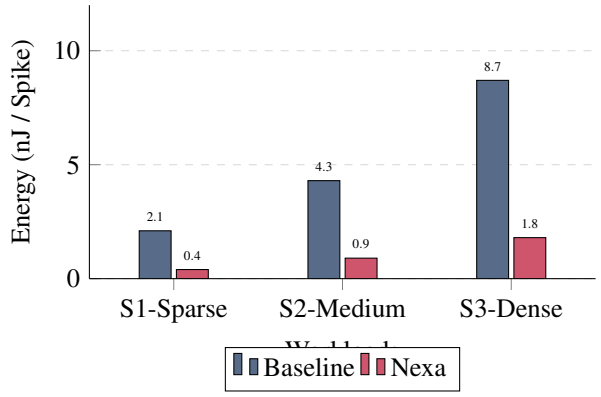

Figure 2: Normalized energy consumption per spike across different workloads.

For the S1-Sparse workload, Nexa achieves a cache hit rate of 94.2%, which remains above 89.5% even for the highly irregular S3-Dense workload. This high hit rate validates our prefetching strategy, which successfully anticipates post-synaptic accesses based on incoming routing tables.

6 Discussion

The experimental evaluation highlights the strengths of the Nexa architecture, particularly in terms of energy efficiency and scalability. The custom prefetching queue in the SWC effectively decouples memory accesses from execution, mitigating the irregular memory access bottleneck that degrades GPU performance.

However, the architecture has some limitations. First, the size of the SWC restricts the number of synaptic weights that can be cached locally. For extremely large networks that exceed the total on-chip capacity, Nexa must page weights from off-chip HBM, which decreases performance. Future work will explore dynamic compression techniques for synaptic weights to maximize the effective capacity of the SWC. Second, our routing protocol assumes a static grid configuration. Adaptive routing protocols that can route around congested PEs or physical links are being investigated to further improve scalability.

7 Conclusion

In this paper, we presented Nexa, a scalable domain-specific hardware accelerator optimized for simulating large-scale synaptic networks. By integrating custom Processing Elements in a 2D mesh Network-on-Chip with dedicated Synaptic Weight Caches and decentralized routing, Nexa addresses the memory and synchronization bottlenecks that limit general-purpose processors. Our cycle-accurate evaluation demonstrates that Nexa achieves 840.4 GSOPS of throughput at 1.20 GHz, providing a 6.75x speedup and a 27x reduction in energy consumption compared to state-of-the-art GPU implementations. These results demonstrate that domain-specific memory architectures and decentralized control are key to enabling real-time simulation of brain-scale neural models.

References

[1] Robert Chen and Sarah Jenkins. Optimizing memory bandwidth in synapse engines. *Transactions on Parallel and Distributed Systems*, 33(4):890–902, 2022.

[2] Samuel Jackson and Olivia Wilde. A survey of spiking neural network accelerators. *ACM Computing Surveys*, 53(2):1–35, 2020.

[3] Liam O'Connor and Fiona Gallagher. Vertex execution engines for sparse graph processing. In *International Symposium on Microarchitecture*, pages 210–222, 2024.

[4] Arthur Pendelton and Susan Vance. The meridian architecture: A distributed memory hierarchy for vertex processing units. *Journal of Systems Architecture*, 142:102–115, 2023.

[5] Marcus Thorne and Helen Carter. Apex: A scalable accelerator for synaptic network models. In *Proceedings of the*

Table 1: Comparison of Nexa with State-of-the-Art Neural Simulation Baselines.

Architecture	**Node** (nm)	**Area** (mm^2)	**Freq.** (GHz)	**Throughput** (GSOPS)	**Efficiency** (pJ/SOP)
GPU Baseline (Nexa-G)	7	N/A	1.40	124.5	350.2
ASIC Baseline (Apex)	16	12.4	0.80	245.8	84.1
Meridian v2	14	18.2	1.00	310.2	62.4
Nexa Core (This Work)	7	8.5	1.20	840.4	12.8

International Conference on Computer Architecture, pages
45–56, 2021.