LUCIA: A Scalable and Versatile 0.82 TOPS/W/mm^2 Chiplet Architecture for CNN in 22 nm FDSOI

Arturo Prieto[†], Kristoffer Westring[†], Alex Allfjord[†], Masoud Nouripayam[†], Sergio Castillo Mohedano[†], William Marnfeldt[†], Joachim Rodrigues[†‡], Linus Svensson[†], Per Andersson[†], and Victor Åberg[†]
[†]Department of Electrical and Information Technology, Lund University, Sweden
[‡]Department of EEI, System-on-Chip, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

***Abstract*—This paper presents LUCIA, an architecture tailored for multi-chiplet systems optimized for convolutional neural network (CNN) acceleration through configurable hardware engines and distributed workload mapping. The architecture accommodates multiple independent near-memory computing (NMC) units in close proximity to their respective memory banks, tightly coupled with a low-power RISC-V processor. Dedicated memory control units facilitate efficient intra-chiplet and inter-chiplet data transfers, reducing execution time by 34 % and consequently improving bandwidth utilization. Fabricated in 22 nm FDSOI technology, LUCIA achieves a peak clock frequency of 565 MHz and demonstrates an energy-area efficiency of 0.82 TOPS/W/mm^2, outperforming state-of-the-art implementations in comparable or more advanced technologies.**

***Index Terms*—Chiplet, convolutional neural network (CNN), edge artificial intelligence of things (AIoT), near-memory computing (NMC), SRAM.**

I. Introduction

In the artificial intelligence of things (AIoT) era, inference is increasingly performed on edge devices, driving the need for highly integrated system-on-chips (SoCs) that deliver high performance under strict memory and energy constraints. However, implementing large-scale designs on a single monolithic die exposes systems to significant yield risks [1]. To address this challenge, disaggregating monolithic designs into interconnected chiplets has emerged as a promising solution, enabling parallel and modular architectures that improve yield and scalability for edge-AI acceleration. Nevertheless, chiplet-based systems introduce new challenges in workload partitioning, scheduling, and inter-chiplet communication.

Manticore [2] and Occamy [3] introduce large chiplet-based architectures with many RISC-V cores for data parallelization, using a passive silicon interposer as the interconnection medium. As an alternative to a large cluster of RISC-V cores, the integration of tailored hardware acceleration units with low-power general-purpose platforms enables resource sharing, offering efficient solutions to process computation-intensive tasks on edge devices.

This work introduces a customized near-memory computing (NMC) architecture built on the framework of [4], purposefully engineered to exploit the parallelism and data locality of multi-chiplet systems. The proposed architecture processes convolutional neural network (CNN) models through modular workload distribution and task mapping strategies for high performance and energy-area efficiency.

The contribution of this work centers on a versatile and scalable hardware acceleration platform, where parallel processing with dedicated intra-chiplet data transfer units improves the energy efficiency per chip area. Furthermore, the proposed architecture includes an area-efficient interface for inter-chiplet communication, enabling low-cost communication for multi-chiplet systems bound together via interposer technology.

This work was financially supported by Vetenskapsrådet, SSF financed classIC Research Center, Chip JU REBECCA project (101097224), and BCDC by the Bavarian Ministry of Economic Affairs (Regional Development and Energy, RMF-SG20-3410-2-17-14)

II. LUCIA Implementation

LUCIA embeds NMC banks alongside memory macros of a RISC-V-based microcontroller SoC, enabling tightly coupled hardware acceleration, as shown in Fig. 1. This shared integration of compute and memory forms the architectural backbone of a scalable implementation for a chiplet architecture, where distributed NMC banks together with a dedicated NMC data transfer unit (DTU) minimize energy-costly data movement while enhancing parallelism. The system features three peripheral memory access (PerMA) modules, with 2-bit TX/RX interfaces for chiplet-to-chiplet (C2C) communication. The architecture is customizable at the RTL level, enabling a flexible configuration of NMC banks, processing element (PE) array and memory capacity, enabling the generation and verification of multiple configurations with minimal engineering effort. The presented work is realized as four NMC banks, where three banks are implemented with accurate PEs, while the fourth facilitates approximate multipliers (AMs), intended for a parallel research project. The NMC banks are configurable to operate either cooperatively or independently, allowing flexible task scheduling within the multi-bank architecture, and thereby supporting distributed workload scheduling.

A. Memory Sub-System

The on-chip SRAM is integrated as a tightly coupled memory, facilitating concurrent and low-latency access by the RISC-V and NMC units. A shared banking strategy for general-purpose operations (RISC-V) and hardware acceleration (NMC units) is defined. This doubles the number of banks per processor, resulting in reduced load capacitance and mitigated access contention [5]. The accelerator engines benefit from a unified memory pool, eliminating data duplication, and thereby reducing the total storage requirements. The proposed architecture offers substantial improvements in area efficiency, power consumption, and memory bandwidth utilization.

979-8-3195-0631-3/26/$31.00 ©2026 IEEE

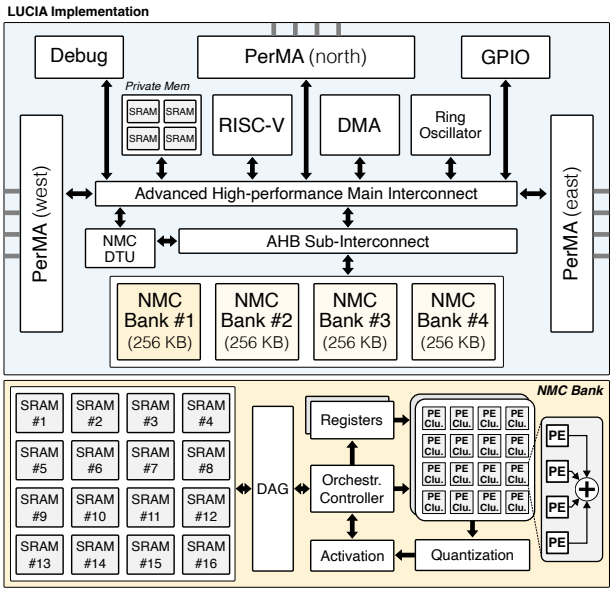


Fig. 1. High-level view of the LUCIA implementation with memory access units and embedded NMC units for hardware acceleration.

Furthermore, each bank of the memory sub-system is partitioned (customizable) into sixteen 16 kB SRAM macros, allowing high-throughput parallel access, which significantly increases aggregate bandwidth for hardware acceleration. The nominal memory bandwidth is preserved for the RISC-V core accesses and external interfaces to maintain system-level consistency. In contrast, the bandwidth available to NMC units is scaled by a factor of 16, i.e., from 32 bit/cycle to 512 bit/cycle, matching internal macro parallelism and supporting wide concurrent data transfers. This elevated communication rate imposes stricter demands on data orchestration and introduces control overhead, addressed by the NMC DTU. Therefore, a system-direct memory access (DMA) module retains the baseline 32 bit/cycle transfer rate for streamlined integration.

B. Near-Memory Computing Architecture

Each NMC bank integrates a dedicated data address generator (DAG) module that orchestrates fine-grained access to the SRAMs, coordinating an output-stationary scheduling strategy. Within each bank, two register banks supply input and weight data to a $4\times4\times2$ PE cluster matrix. Each PE performs four parallel multiply-accumulate (MAC) operations and locally accumulates partial sums, enabling independent output computation. The PEs use 8-bit precision for multiplication and 32-bit registers for accumulation. A high-throughput data scheduler ensures continuous parallel access to the input data, leveraging the enhanced memory bandwidth and tightly couples compute with memory. Each NMC bank processes workloads based on the parallel distribution of CNN models. For convolutional operations, the input channel data is distributed across the memory banks, with the scheduler optimized for local data reuse, and thus, optimizing parallelism while dramatically reducing data movement, see Fig. 2.

Inference-time quantization is applied in a pipelined fashion, followed by activation functions via a dedicated module

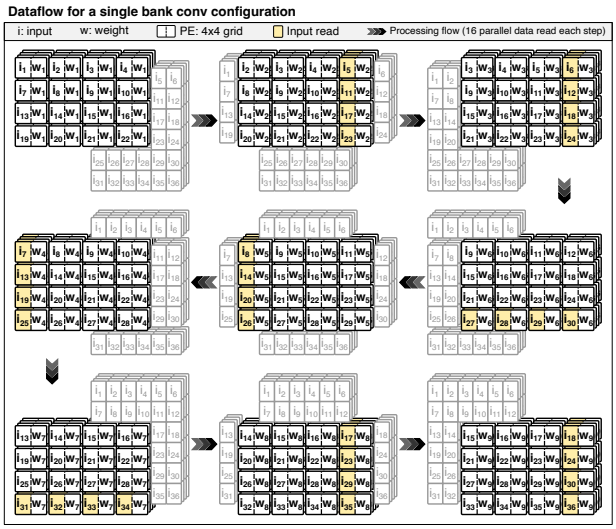


Fig. 2. Data scheduler for convolution operations with sixteen parallel memory accesses per bank.

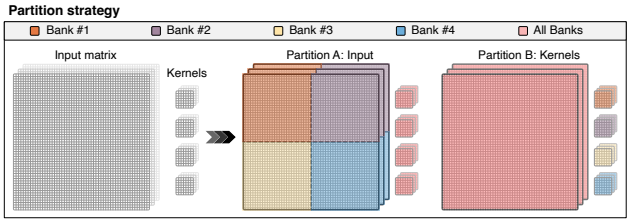


Fig. 3. Partition strategies with shared data between NMC banks for multi-bank computation.

that supports ReLU and other nonlinear functions through configurable lookup tables (LUT). The parameters of the latter are dynamically loaded during system initialization. A register-mapped control interface orchestrates the operation with start, stop, and configuration registers, which define the execution state of the individual NMC banks. Synchronization between the RISC-V core and NMC banks is achieved via an interrupt-driven mechanism enabling flexibility in workload configurations.

C. Versatile Workload Configuration

LUCIA is optimized for the acceleration of CNNs, which is crucial in AI applications that demand high efficiency and fast processing. The modular architecture design is optimized for chiplet-based computation strategies, which enables scalable extension to broader neural network (NN) workloads. Fig. 3 illustrates the NN partitioning strategy across the multi-bank system: Partition A distributes input activations across banks to exploit parallel data accesses, whereas Partition B assigns kernel weights to enable a concurrent computation. These partitioning schemes optimize compute density and memory bandwidth utilization, while minimizing external communication overhead. Multi-bank configurations offer scalability and flexibility for different NN architectures, offering increased parallelization opportunities.

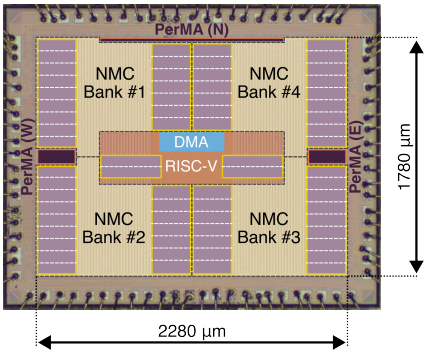
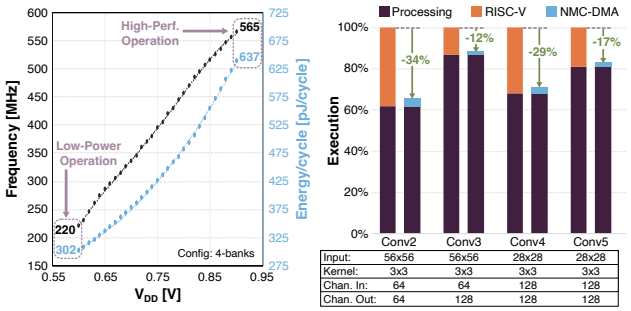


Fig. 4. Micrograph of the LUCIA chiplet and low-level modules.



(a) (b)

Fig. 5. (a) Measured operational frequency and energy response to the supply voltage (V_{DD}), and (b) VGG-16 convolutional layer tiles execution time with intra-chiplet data transfers comparison between RISC-V and NMC DTU.

D. Data Orchestration

Two optimized and configurable custom memory control units (NMC DTU and DMA) are implemented to provide a mechanism for managing data transfers and workload distribution across both inter-chiplet and intra-chiplet hierarchies. Specialized data organization units alleviate contention between NMC banks and peripheral systems, such as neighboring chiplets.

1) Intra-chiplet Data Orchestration: The realization of the NMC architecture requires a tailored data layout over the banked memory structure, introducing an additional overhead. Furthermore, the data organization is workload dependent and differs between data type, e.g., activations, weights, and biases. This is mitigated by the DMA supporting multiple address generation schemes, which makes data transfers to a bank "transparent" from a programming perspective. The NMC DTU, situated by the AHB sub-interconnect, enables high throughput data movement between NMC banks. It is programmed with memory pointers, data dimensions, and access sequencing, guaranteeing conflict-free transfers.

2) Inter-chiplet Data Orchestration: Three PerMA modules are integrated to provide a low-latency interface for adjacent C2C communication. The PerMAs operate as lightweight, autonomous external communication endpoints, eliminating the need for register-mapped configuration and thereby reducing control overhead. In conjunction with the DMA, they enable efficient off-chip data streaming and support a 2-bit transmitter–receiver C2C interface that balances throughput with external communication complexity.

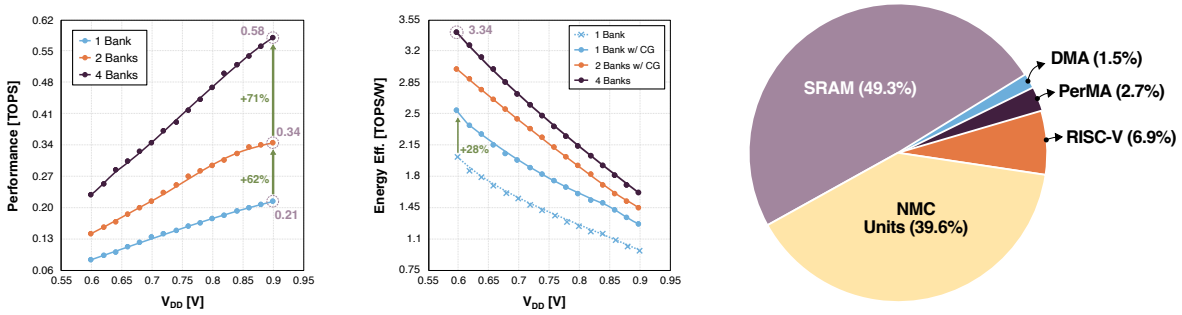
III. Measurements

LUCIA is fabricated in GlobalFoundries 22 nm FDSOI technology, occupying a core area of 2.28 mm × 1.78 mm. A custom-designed PCB serves as an integration platform, providing high-speed I/O interfaces, debug access, and power delivery for all parts of the design. A micrograph is shown in Fig. 4, with an emphasis on the spatial organization of the NMC banks. The architecture is centered around the RISC-V core, surrounded by memory macros (shown in purple) and data orchestration modules. The three PerMAs are placed on the north, west, and east sides of the chip for external communication while the DMA engine is strategically placed to facilitate high-throughput internal data movement. Chip measurements were conducted at ambient temperature. The measured power reduction of the NMC bank with approximate PEs was scaled to ensure a fair comparison. Convolutional layers are used as benchmark workloads to assess performance and energy efficiency across varying configurations, i.e., from single-bank to multi-bank NMC processing. A voltage-frequency characterization is performed for the four-bank configuration, as shown in Fig. 5a, revealing two distinct operating regimes: a low-power mode optimized for energy-per-cycle, and a high-performance mode supporting elevated throughput with operation up to 565 MHz. The execution time of four VGG-16 convolutional layer tiles following the partition of input data distributed across the four NMC banks and intra-chiplet data transfer between layers, is shown in Fig. 5b. Inter-bank data transfer using the custom NMC DTU unit efficiently reduces Conv2 execution time by 34 % compared to transferring data between banks using the RISC-V core.

Performance evaluation for standalone processing activity, shown in Fig. 6a, is conducted across multiple active-bank configurations, with each MAC operation counted as two arithmetic operations. The frequency is characterized for each configuration of active banks, where LUCIA achieves a peak throughput of 0.58 TOPS under full parallel activation of four NMC banks. Energy efficiency, detailed in Fig. 6b, reaches 3.34 TOPS/W in the low-power operating regime with four active banks. To further reduce power consumption, the NMC units employ fine-grained clock gating (CG), achieving a 28 % improvement in energy efficiency during single-bank operation. Additionally, measured leakage power indicates the impact of power gating (PG) on inactive banks. Power gating three NMC banks, while keeping active the memory macros to retain stored data valid, offers an estimated 15 % gain in energy efficiency. The area composition of LUCIA is illustrated in Fig. 6c, where SRAM blocks account for approximately 50 % of the total silicon footprint. This includes both the shared memory sub-system used for hardware acceleration and the private memory regions allocated to the RISC-V core.

IV. Discussion

A summary of relevant state-of-the-art (SoA) implementations is provided in Table I, encompassing both single-chip and multi-chiplet architectures. Two notable single-chip



(a) (b) (c)

Fig. 6. Design modularity evaluation as a function of active NMC banks and their impact on (a) performance, (b) energy, and (c) area.

TABLE I
MEASUREMENT RESULTS AND COMPARISON WITH SOA.

Source	ISSCC'22 [6]	ESSCIRC'23 [7]	VLSI'22 [8]	CICC'25 [9]	This work
Technology	65 nm	16 nm	22 nm	4 nm	22 nm
Design	SNCPU	Clust. + Acc.	MCP	NMC	NMC
Configuration	1-chip	1-chip	4-chip	16-chip	1-chip
Core Area [mm^2]	4.5	16 *	11.1	16	4.1
Voltage [V]	0.5 - 1	0.6 - 0.8	0.6 - 0.9	0.8 - 1.1	0.6 - 0.9
Max. Freq. [MHz]	400	530	492	500	565
Int Precision	8	2 to 8	16	8	8
Memory [MB/chip]	0.15	6.5	2.5	16	1
PEs/chip	100	36	1024	128	512
Performance [TOPS]	0.08	0.38 (int8)	1.07	2 ‡	**0.58**
Energy Eff. [TOPS/W]	1.82 †	2.9 † (int8)	7.19 †	8.7 ‡	**3.34** †
Energy-Area Eff. [TOPS/W/mm2]	0.4	0.18 (int8)	0.65	0.54 ‡	**0.82**

* Silicon die area † @ Low V_{DD} ‡ Dense computation mode

solutions are considered: the systolic neural CPU (SNCPU) in [6], which offers a reconfigurable architecture capable of operating as a multi-core RISC-V processor or a systolic CNN accelerator; and the Siracusa platform in [7], which integrates a RISC-V cluster with a tightly coupled accelerator for efficient AI processing. Multi-chiplet systems are represented by the NetFlex multi-chiplet package (MCP) in [8], designed for scalable networked inference, and the SPGEMM chiplet-based accelerator in [9], targeting sparse matrix multiplication workloads. While these works address the growing computational demands of modern AI models, LUCIA advances the architectural paradigm by explicitly mapping compute-intensive workloads onto a multi-bank topology. LUCIA leverages spatial tile partitioning of memory banks, tightly integrated with NMC units to achieve scalable acceleration with reduced data movement and improved memory locality.

LUCIA achieves an energy–area efficiency of 0.82 TOPS/W/mm^2, significantly surpassing comparable SoA implementations fabricated in equal or more advanced technology nodes. This result underscores the effectiveness of the proposed distributed architecture in delivering high throughput with superior energy efficiency per unit area. Looking toward future 2.5D multi-chiplet integration, the modular architecture supports scalable deployment across multiple chiplets, each tightly coupled to NMC banks. Interposer-based integration will further enhance C2C communication density, reduce communication latency, and unlock additional performance gains through high-bandwidth and low-power interconnects.

V. CONCLUSION

In this article, we present LUCIA, a foundational chiplet architecture suitable for parallel processing implementations, with NMC acceleration units integrated in the memory banks of a RISC-V-based platform. LUCIA is manufactured using 22 nm FDSOI technology and allows the configuration of partitioned workloads in multiple memory banks with enhanced bandwidth. The measurement results show an energy-area efficiency of 0.82 TOPS/W/mm^2, with a performance of 0.58 TOPS.

ACKNOWLEDGMENT

The authors are grateful for support from Codasip for providing the RISC-V IP, and GlobalFoundries for silicon fabrication.

REFERENCES

[1] P. Iff, M. Besta, M. Cavalcante, T. Fischer, L. Benini and T. Hoefler, "HexaMesh: Scaling to Hundreds of Chiplets with an Optimized Chiplet Arrangement," *60th ACM/IEEE Design Automation Conference (DAC)*, 2023, pp. 1-6.

[2] F. Zaruba, F. Schuiki and L. Benini, "Manticore: A 4096-Core RISC-V Chiplet Architecture for Ultraefficient Floating-Point Computing," in *IEEE Micro*, vol. 41, no. 2, pp. 36-42, 2021.

[3] G. Paulin et al., "Occamy: A 432-Core 28.1 DP-GFLOP/s/W 83% FPU Utilization Dual-Chiplet, Dual-HBM2E RISC-V-Based Accelerator for Stencil and Sparse Linear Algebra Computations with 8-to-64-bit Floating-Point Support in 12nm FinFET," *IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2024, pp. 1-2.

[4] A. Prieto et al., "NMCu-CNN: A Scalable Near-Memory Computing Co-Processor on a RISC-V MCU in 22nm FDSOI," in *IEEE Workshop on Signal Processing Systems (SiPS)*, 2025, pp. 1-5.

[5] L. Benini, E. Flamand, D. Fuin and D. Melpignano, "P2012: Building an ecosystem for a scalable, modular and high-efficiency embedded computing accelerator," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2012, pp. 983-987.

[6] Y. Ju and J. Gu, "A 65nm Systolic Neural CPU Processor for Combined Deep Learning and General-Purpose Computing with 95% PE Utilization, High Data Locality and Enhanced End-to-End Performance," *IEEE International Solid-State Circuits Conference (ISSCC)*, 2022, pp. 1-3.

[7] M. Scherer et al., "Siracusa: A Low-Power On-Sensor RISC-V SoC for Extended Reality Visual Processing in 16nm CMOS," *IEEE 49th European Solid State Circuits Conference (ESSCIRC)*, 2023, pp. 217-220.

[8] T. Chou et al., "NetFlex: A 22nm Multi-Chiplet Perception Accelerator in High-Density Fan-Out Wafer-Level Packaging," *IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2022, pp. 208-209.

[9] S. R. Srinivasa et al., "A 68 TOPS/W, 256MB SRAM Sparse GEMM Accelerator Tiled Across 16, 4nm Near Memory Compute (NMC) Chiplets Disaggregated 2.5D System," *IEEE Custom Integrated Circuits Conference (CICC)*, 2025, pp. 1-3.