Ph.D. Project Investigating Chiplet Interfaces for Efficient Near-Sensor Computing in Visual On-Device Intelligence

Peter Mbua
University of Florida
Gainesville, Florida
Email: mbua.esenju@ufl.edu

Christophe Bobda
University of Florida
Gainesville, Florida. 116200
Email: cbobda@ece.ufl.edu

***Abstract*—The ever-growing demand for intelligent devices capable of visual processing in real time at the edge requires a paradigm shift in computing architectures. Near-sensor computing offers a promising solution by bringing computation closer to the data source, enabling faster response times and reduced power consumption. However, traditional monolithic chip design struggles to meet the efficiency and flexibility demands of near-sensor visual intelligence tasks. This Ph.D. project investigates the transformative potential of chiplet interfaces in revolutionizing near-sensor computing for on-device visual intelligence. Chiplet technology offers a modular approach that enables the integration of heterogeneous cores and specialized hardware accelerators into a single package. Using this modularity, the project aims to achieve the following key objectives: *(1) design and explore novel chiplet interface architectures, (2) hardware-software co-design for chiplet-based near-sensor processing targeting visual data, (3) power efficiency exploration* and finally *real-world application validation*. The successful completion of this Ph.D. project is expected to yield significant contributions to the field of near-sensor computing. With expectations of proposing some novel chiplet interfaces for efficient visual on-device intelligence, the project has the potential to pave the way for a new generation of intelligent devices capable of real-time visual processing at the edge, with minimal reliance on cloud-based resources.**

I. Problem and Motivation

A. Problem

Previous research has shown the profound benefits of near-sensor computing in edge computing. Some of these advantages include real-time operations and limited security concerns. However, there is a trade-off for long battery life, especially when good quality is required [1], [2]. As ongoing research strives to bring some level of sensory data processing and inference closer to sensory units, there exist some wavering shortcomings like; limited energy and hardware overhead due to complex designs and algorithms [3], [4], [5].

Traditional monolithic chip design struggles to meet the efficiency and flexibility demands of near-sensor visual intelligence tasks [6]. These tasks require real-time processing at the edge of the network, but current designs are often *inefficient* - they may not be optimized for the specific needs of visual processing, leading to wasted power and processing resources. And, *inflexible* - they may not be easily adaptable to different visual intelligence applications, requiring custom designs for each task. The aforementioned drawbacks provide an avenue for exploring the benefit of chiplet design methodology while leveraging the strengths of chiplet interfaces to support near-sensor computing.

B. Motivation

The ever-growing demand for real-time and distributed processing at the edge has fueled significant interest in near-sensor computing [4]. This paradigm shifts processing power closer to data acquisition points, enabling faster response times and reduced data transmission burdens. However, traditional monolithic integration approaches struggle to meet the diverse computational needs and power constraints of near-sensor applications. This includes applications like: *autonomous systems*(e.g., self-driving cars), *Smart robotics*, and *real-time visual analytics* applications. By addressing the limitations of traditional chip design applicable to near-sensor computing, this research aims to *improve efficiency*, *enhance flexibility*, and *enable real-time processing at the edge*. The successful development of chiplet interfaces for near-sensor computing in visual on-device intelligence can lead to a new generation of intelligent devices with significant benefits across various sectors.

II. Background and related Work

Near-sensor computing, also sometimes referred to as in-sensor computing, is a computing paradigm that rethinks how data processing is handled, particularly for devices with sensors that generate a continuous stream of data. Near sensor computing addresses the limitations of cloud server for analysis [4]. These limitations include, *latency, power consumption* and *bandwidth limitations*. Figure 1 shows the concept of near-sensor computing and shows how it differs slightly from in-sensor computing.

Salient-based processing is a concept that aims to identify and enhance the most visually significant part of sensory data, known as salient regions [4]. Our initial exploration of this concept used temporal saliency (TS) and spatial saliency (SS) of sensory data to generate slaiency flags used by filters to

2576-2621/24/$31.00 ©2024 IEEE
DOI 10.1109/FCCM60383.2024.00056

prune visual data. The main difference between near-sensor

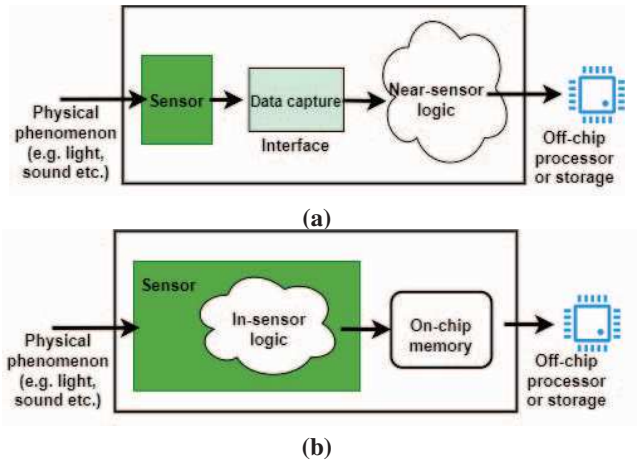

Fig. 1: In-sensor vs near-sensor processing. a. Near-sensor processing architecture. b. In-sensor processing architecture.

and in-sensor computing is the relative distance of the first level of sensory data processing to the sensory units.

The chiplet design methodology split a system into smaller chiplets, and then integrate heterogeneous or homogeneous chiplets through advanced packaging technology [7]. A chiplet is a functional integrated circuitry, designed independently, and thus we can have a third-party IP as a chiplet. Recent work [6], [8], [9] in chiplet technology research and development has developed emerging SoC interfaces such as universal chiplet interconnect express (UCIe), advanced interface bus (AIB) and OpenHandover (OH). These research progresses stems our research project. Also, we leverage on these related works [1], [3], [4], [5] for explorations in near-sensor computing.

III. Approach and Uniqueness

Here is a breakdown of our approach for this research project;

- Co-Designing Chiplet Interfaces and reconfigurable salient-based near-sensor circuit:
  - The salient-based near-sensor processing will be *reconfigurable* and designed using coarse-grained elements.
  - Exploration of Novel Communication Protocols: We will use the concepts of emerging chiplet interfaces as baselines.
- Exploring Heterogeneous Chiplet Integration for Specialized Processing:
  - Vision Processing Units (VPUs): By designing and researching hardware accelerators specifically for handling visual sensory data, we can harness potentially significant efficiency gains.
  - Neuromorphic Computing Cores: Investigate the feasibility of integrating neuromorphic chiplets to potentially offer power-efficient.

IV. Preliminary Results

We have explored near-sensor processing using salient-based processing. This approach sets the pace for light-weight near-sensor circuitry and still provides ”good enough” quality. Our approach is a two-stage hierarchical architecture consisting of saliency-based generators that feed a saliency filter. Preliminary results demonstrated significant energy and hardware optimization compared to previous work in this field. Specifically, the proposed architecture uses 0.24%, 0.60%, and 9% of the total processing elements used by examined related work. Similarly, the total on-chip power usage of 0.686W, which is 34.5 % and 5.3% less for the same baseline. This motivates the integration of our near-sensor logic into a system where a high-level algorithm can benefit from the data filtering.

V. Expected Results and Contributions

We envision the exploration of chiplet interface technology to improve near-sensor processing for visual intelligence. The project aims to achieve the following:

- A novel chiplet interface communication protocol.
- A chiplet adapter IP for visual data processing.
- Energy efficiency exploration and analysis of the realized architectures.
- Real-world application validation with a workable prototype.

The successful completion of this Ph.D. project is expected to yield significant contributions to the field of near-sensor computing. With expectations of proposing some novel chiplet interfaces for efficient visual on-device intelligence, the project has the potential to pave the way for a new generation of intelligent devices capable of real-time visual processing at the edge, with minimal reliance on cloud-based resources.

References

[1] H. J. Damsgaard, A. Ometov, and J. Nurmi, “Approximation opportunities in edge computing hardware: A systematic literature review,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–49, 2023.

[2] J. Park, E. Amaro, D. Mahajan, B. Thwaites, and H. Esmaeilzadeh, “Axgames: Towards crowdsourcing quality target determination in approximate computing,” *ACM SIGPLAN Notices*, vol. 51, no. 4, pp. 623–636, 2016.

[3] M. S. Park, C. Zhang, M. DeBole, and S. Kestur, “Accelerators for biologically-inspired attention and recognition,” in *Proceedings of the 50th Annual Design Automation Conference*, pp. 1–6, 2013.

[4] F. Zhou and Y. Chai, “Near-sensor and in-sensor computing,” *Nature Electronics*, vol. 3, no. 11, pp. 664–671, 2020.

[5] T. Wang, Y. Liang, X. Shen, X. Zheng, A. Mahmood, and Q. Z. Sheng, “Edge computing and sensor-cloud: Overview, solutions, and directions,” *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–37, 2023.

[6] T. Li, J. Hou, J. Yan, R. Liu, H. Yang, and Z. Sun, “Chiplet heterogeneous integration technology—status and challenges,” *Electronics*, vol. 9, no. 4, p. 670, 2020.

[7] X. Ma, Y. Wang, Y. Wang, X. Cai, and Y. Han, “Survey on chiplets: interface, interconnect and integration methodology,” *CCF Transactions on High Performance Computing*, vol. 4, no. 1, pp. 43–52, 2022.

[8] J. H. Lau, “Recent advances and trends in chiplet design and heterogeneous integration packaging,” *Journal of Electronic Packaging*, vol. 146, no. 1, p. 010801, 2024.

[9] D. D. Sharma, “High-performance, power-efficient three-dimensional system-in-package designs with universal chiplet interconnect express,” 2024.