PIM-based AI accelerator and software stack for efficient LLM inference

Pablo Robin Guerrero
Undergraduate Student
École Polytechnique Fédérale de Lausanne (EPFL)

Under the supervision of
Prof. Jeehoon Kang and Haechan An
Concurrency & Parallelism Laboratory

School of Computing

KOREA ADVANCED INSTITUTE OF SCIENCE & TECHNOLOGY (KAIST)
한국과학기술원

Daejeon, South Korea, Summer 2024

1 Abstract

Optimizing Large Language Model (LLM) inference is a critical challenge due to the computational demands and complexity of these models. This research explores the use of AI accelerators and software stacks, specifically focusing on Process-In-Memory (PIM) architecture, as a promising solution to enhance inference performance. Given the evolving landscape of deep learning, where models increasingly require substantial computation, memory, and bandwidth, the efficient deployment of these resources becomes critical.

To address these needs, specialized accelerators are being explored to distribute the model executions to multiple hardwares. However, leveraging multiple heterogeneous accelerators, such as Neural Processing Units (NPUs) and PIMs, necessitates advanced software stack, including specialized compilers and runtime systems, bridging the high-level application and low-level hardware. Currently, there is a gap in compiler support capable of automatically generating efficient workflow across diverse accelerators. This research aims to overcome this gap by developing a heterogeneous and distributed deep learning compiler designed specifically for NPUs and PIMs, by focusing on the software and compiler aspects, seeking to optimize the performance of LLM inference.

2 Introduction

LLMs inference is computationally intensive, requiring optimized strategies to manage both the high compute and memory demands effectively. This research focuses on the integration of AI accelerators, specifically PIM architectures, to address the inefficiencies in current inference methodologies. It began with an extensive review of existing published papers, where key papers and tutorials on AI accelerators and PIM architectures were summarized to establish both strong theoretical and practical foundations. Then, comparative analyses were conducted to evaluate different methodologies, particularly in the context of PIM-based accelerators and software support, highlighting their advantages in handling the bandwidth-intensive computations required for LLM inference.

A key aspect of this research was the comparative analysis of various LLM inference approaches, focusing on how PIM and NPU architectures could be utilized to optimize different computational tasks. NPUs are primarily focused on efficiently handling General Matrix Multiply (GEMM) operations, which are central to many AI workloads. However, they often struggle with bandwidth-intensive tasks such as General Matrix-Vector Multiply (GEMV) operations, which can lead to underutilization of computational resources. In contrast, PIM architectures leverage internal memory bandwidth to more effectively manage GEMV operations, thereby improving overall system efficiency. By combining the strengths of both NPU and PIM architectures, the research explores how these accelerators can complement each other to optimize the full range of computational tasks required for LLM inference. This comparative analysis also provided essential background context by reviewing architectures such as NeuPIMs and AttAcc.

The NeuPIMs architecture introduces micro-architectural innovations like dual row buffers and sub-batch interleaving, which are designed to improve memory bandwidth utilization and parallel processing capabilities. These features make NeuPIMs particularly effective for handling the compute and memory demands of LLM inference. On the other hand, the AttAcc architecture is designed specifically to optimize the attention layers in LLMs, reducing data movement within key-value (KV) matrices and thereby addressing one of the primary bottlenecks in LLM inference.

Both theoretical and practical approach were complemented by hands-on implementations, including the reproduction of existing evaluations to validate the effectiveness of PIM in accelerating LLM tasks. The research was conducted through an iteration-based workflow, characterized by a cycle of comment-review-comment-review. This iterative process ensured thorough exploration and refinement of ideas. The findings were then used to draft proposals for further exploration of PIM-enabled systems in LLM applications. The goal is to develop high-performance solutions that can overcome the bottlenecks inherent in LLM inference, with a particular focus on optimizing both hardware and software components.

3 Work Done

My internship was focused on experiencing the early stages of the research process, with a strong emphasis on making tangible contributions rather than simply studying papers or tutorials. The work involved a deep exploration of PIM-based technology aimed at optimizing large-scale model inference, particularly for LLMs. This included a comprehensive paper review, comparative analysis of PIM-based architectures and software stacks, direct communication with authors for clarification of complex concepts, reproduction of experimental results, and the development of new research proposals based on identified gaps.

3.1 Paper Summarization

A critical initial phase of this research involved the systematic review and summarization of key papers in the domain of PIM-based AI accelerator architecture and software stacks. The papers reviewed include NeuPIMs, AttAcc, SpecPIM, and PIM-DL, each addressing distinct challenges within the context of LLMs inference:

NeuPIMs enhances the integration of neuromorphic computing within memory modules to mitigate the inefficiencies associated with data movement between memory and processors in traditional architectures. This approach primarily aims to reduce energy consumption and latency, offering a promising solution to the von Neumann bottleneck.

AttAcc introduces a novel DRAM-based PIM architecture designed to optimize the performance of memory-bound attention layers in Transformer models. By addressing the memory capacity constraints related to storing key-value matrices, AttAcc presents a targeted solution for accelerating Transformer-based inference.

SpecPIM examines the potential of speculative inference on PIM-enabled systems, with a focus on reducing latency in autoregressive decoding. This is achieved through the co-exploration of architecture and dataflow, balancing the computational load between a draft model and the target model to optimize performance.

PIM-DL focuses on the optimization of deep learning workloads across distributed memory systems. By co-optimizing algorithmic and system-level aspects, PIM-DL aims to enhance the applicability of commodity DRAM-PIMs for deep learning tasks, ensuring efficient execution across neural network layers.

The iterative review process applied in summarizing these papers played a key role in identifying the specific problems each work addressed, and in developing a nuanced understanding of the solutions proposed within the context of PIM-based architecture for efficient LLM inference.

3.2 Comparing NeuPIMs and AttAcc

A comparative analysis was conducted to critically evaluate the NeuPIMs and AttAcc architectures. This analysis sought to systematically compare the strengths and limitations of each approach within a unified framework. The **NeuPIMs** architecture offers significant reductions in data movement by integrating computation directly within memory. However, it exhibits limitations in terms of flexibility and scalability, particularly when applied to large transformer models which result from the poor parallelism implementation of multiple devices workload.
In contrast, **AttAcc** provides a specialized solution for optimizing attention layers in Transformer models. By leveraging the inherent strengths of DRAM-based PIM for memory-intensive tasks, AttAcc demonstrates improved performance for specific inference tasks. Nonetheless, its reliance on DRAM-based architecture may pose challenges when scaling to more general-purpose applications.

This comparative evaluation was essential in understanding the trade-offs between computational efficiency and memory usage in different PIM architectures, contributing to a more comprehensive perspective on their applicability in LLM inference.

3.3 Communication with Authors

To address ambiguities and gain deeper insights into the methodologies discussed in the reviewed papers, direct communication with the authors was established. These interactions were essential for clarifying complex concepts and resolving gaps in the available information. The correspondence was conducted in a manner that was clear, concise, and focused on obtaining specific technical clarifications, thereby enhancing the overall understanding of the research.

3.4 Paper Reproduction

The reproduction of experiments detailed in the NeuPIMs papers was undertaken as a practical component of the research. This involved executing and debugging the code provided by the authors, in order to critically analyze the results. Through this process, key limitations were identified, such as the lack of multi-device support in the NeuPIMs architecture, which presents a significant challenge for scaling the system in real-world applications. The reproduction of these experiments provided empirical evidence of the potential and limitations of these PIM architectures.

3.5 Research Proposal

Building on the insights gained from the paper review, comparative analysis, and experimental reproduction, new research proposals were investigate. The process of identifying and articulating research problems was central to this task, with a focus on addressing the gaps and limitations observed in existing PIM architectures. For example, the lack of multi-device support in NeuPIMs led to the proposal of enhancements aimed at improving scalability in PIM systems. The development of these proposals followed an iterative approach, with initial broad explorations being refined through guidance from senior researchers, ultimately focusing on the most promising research directions.

4 Future Work

The research conducted thus far has provided valuable insights into the optimization of Large Language Model (LLM) inference through the use of AI accelerators, particularly Process-In-Memory (PIM) architectures. Building on these findings, several avenues for future work can be pursued to further enhance the effectiveness and applicability of these technologies.

4.1 Advancing Heterogeneous System Integration

Future research should focus on the integration of multiple heterogeneous accelerators, including Neural Processing Units (NPUs) and PIMs, to create more unified and efficient systems. This includes developing advanced compilers and runtime systems capable of seamlessly managing workload distribution across different hardware. Exploring strategies for dynamic scheduling and load balancing could also improve performance and resource utilization.

4.2 Enhancing Compiler and Software Support

The current gap in compiler support for heterogeneous accelerators highlights the need for further development in this area. Future work should aim at creating a more sophisticated compilation framework that can automatically optimize and translate high-level model representations into efficient code for diverse accelerators. This includes developing techniques for better inter-device communication and optimizing memory access patterns.

4.3 Exploring New PIM-based Architectures

While NeuPIMs and AttAcc represent significant advancements, there is potential for exploring new PIM architectures that address their limitations. Research could investigate novel microarchitectural innovations to improve scalability and flexibility, such as advanced multi-device coordination strategies and adaptive memory management techniques. Additionally, the development of hybrid PIM architectures that combine the strengths of different PIM approaches could offer enhanced performance and versatility.

4.4 Expanding Real-World Applications

Further research should focus on applying PIM-based architectures to a broader range of real-world applications beyond LLMs. This includes evaluating the effectiveness of these architectures in different domains such as computer vision, reinforcement learning, and other high-performance computing tasks. Understanding how PIMs can be tailored to specific application needs will provide insights into their broader utility and potential impact.

4.5 Improving Experimental Reproducibility

Efforts should be made to improve the reproducibility of experimental results in PIM research. This involves developing standardized benchmarks, providing detailed experimental methodologies, and ensuring that code and data are publicly available for verification. By enhancing reproducibility, the research community can more effectively validate findings and build on existing work.

4.6 Addressing Power and Energy Efficiency

As PIM architectures continue to evolve, there is a growing need to address power and energy efficiency. Future research should explore ways to optimize the power consumption of PIM-based systems, particularly in large-scale deployments. This could involve developing energy-aware algorithms, power-efficient hardware designs, and techniques for minimizing energy overheads associated with memory operations.

By pursuing these avenues, future work can further advance the field of AI accelerators and PIM architectures, leading to more efficient and scalable solutions for LLM inference and beyond. The continued exploration of these topics will contribute to the development of high-performance systems that can meet the growing demands of modern AI applications.

5 Conclusion

The research presented in this paper highlights the potential of PIM architectures in optimizing LLM inference. By systematically evaluating existing PIM-based technologies like NeuPIMs and AttAcc, this study underscores the advantages of integrating computation directly within memory to address the bottlenecks associated with traditional AI accelerators. The comparative analysis and experimental reproduction conducted as part of this research provide valuable insights into the strengths and limitations of current PIM architectures, particularly in handling memory-intensive operations.

One of the key findings of this research is the significant impact that heterogeneous accelerator integration, including both NPUs and PIMs, can have on improving the efficiency and scalability of LLM inference. The development of a deep learning compiler tailored to such heterogeneous systems is identified as a crucial step toward realizing the full potential of these architectures.

While the research has made important contributions to the field, it also opens several avenues for future work. Enhancing the scalability of PIM systems, improving compiler support for diverse accelerators, and expanding the applicability of PIM technologies to other domains are all critical areas that warrant further exploration. Additionally, the importance of power and energy efficiency in the development of next-generation AI accelerators cannot be overstated.

In conclusion, this study serves as a foundational step towards the development of high-performance, scalable solutions for LLM inference, leveraging the unique capabilities of PIM architectures. The continued exploration and refinement of these technologies will be essential in meeting the growing computational demands of modern AI applications.

Acknowledgments

The research was conduct at the Concurrency & Parallelism (CP) Laboratory of KAIST.

I am deeply grateful to have had the opportunity to work at the CP Laboratory of KAIST. I would like to extend my heartfelt thanks to Haechan for his invaluable feedback, continuous support throughout this internship and giving me such amazing insights.

I am also sincerely thankful to Prof. Jeehoon Kang and everyone at the CP Laboratory for warmly welcoming me this summer and providing me with the chance to conduct this research. Your guidance and hospitality made this experience truly enriching.

References

[1] Guseul Heo, Sangyeop Lee, Jaehong Cho, Hyunmin Choi, Sanghyeon Lee, Hyungkyu Ham, Gwangsun Kim, Divya Mahajan, Jongse Park. 2024. *NeuPIMs: NPU-PIM Heterogeneous Acceleration for Batched LLM Inferencing*.

[2] Jaehyun Park, Jaewan Choi, Kwanhee Kyung, Michael Jaemin Kim, Yongsuk Kwon, Nam Sung Kim, Jung Ho Ahn. 2024. *AttAcc! Unleashing the Power of PIM for Batched Transformer-based Generative Model Inference*.

[3] Cong Li, Zhe Zhou, Size Zheng, Jiaxi Zhang, Yun Liang, Guangyu Sun. 2024. *SpecPIM: Accelerating Speculative Inference on PIM-Enabled System via Architecture-Dataflow Co-Exploration*.

[4] Cong Li, Zhe Zhou, Yang Wang, Fan Yang, Ting Cao, Mao Yang, Yun Liang, Guangyu Sun. 2024. *PIM-DL: Expanding the Applicability of Commodity DRAM-PIMs for Deep Learning via Algorithm-System Co-Optimization*.