Hardware Accelerator for Vision Transformer (ViT)

Indo-U.S. Science & Technology Forum
Catalyzing Indo-U.S. Science & Technology Cooperation
IUSSTF

USC Viterbi School of Engineering

Shashank Nag[1], Gourav Datta[2], Souvik Kundu[2], Peter A. Beerel[2]

[1] Department of Electrical Engineering, Indian Institute of Technology Madras

[2] Ming Hsieh Department of Electrical and Computer Engineering – Systems, Viterbi School of Engineering, University of Southern California

Introduction

- Computer Vision applications have traditionally used CNN based model architectures
- With the success of Transformers [1] in NLP tasks, self attention based architectures were explored for vision tasks as well.
- Vision Transformer (ViT) [2] proposes to apply self attention to 16 x 16 patches of images, and use the transformer encoder model.
- Hardware accelerators designed for vision tasks are optimised for CNNs, and are not suitable for transformer based models.
- We seek to look at hardware accelerators designed for transformer models (NLP based) and design an accelerator for ViT adapting from those.

Computations involved in ViT

- The ViT model is primarily the encoder of the transformer. Patch embeddings fed as input to model.
- Involves Multi-head self attention(MHA)/ residual blocks, feedforward (FF) blocks, layer normalisations and softmax layers.
- *L* such encoder layers are stacked to form the model.
- The MHA and FF operations are effectively comprised of Matrix Multiplication and Matrix Vector Multiplication operations.

Review

- Wang et al. [3] proposes hardware accelerator for vision transformer models. However, it involves a single PE unit without any emphasis on scheduling schemes.
- Hardware accelerators for NLP based transformers have been designed [4] – [7], with some specifically targeting MHA layers. Optimised designs proposed for non linear units.
- Many of these do not exploit potential for concurrent computations.
- We adopt the idea of having a granular pipeline between two processing blocks from [7], and seek to parallelise operations across heads – maintaining a high HUE.

Proposed Architecture (Scheduling)

- ViT inference when run over a series of images has an advantage over NLP transformers – input sequence length is typically constant.
- We propose a PE block based architecture with granular pipeline, for Multi-head Self Attention computations.
- Separate optimised block for Softmax.
- Two PE blocks, each having k PE units for computing all the heads concurrently.
- Operations scheduled among the two blocks at a granular level, for maximum hardware utilisation.

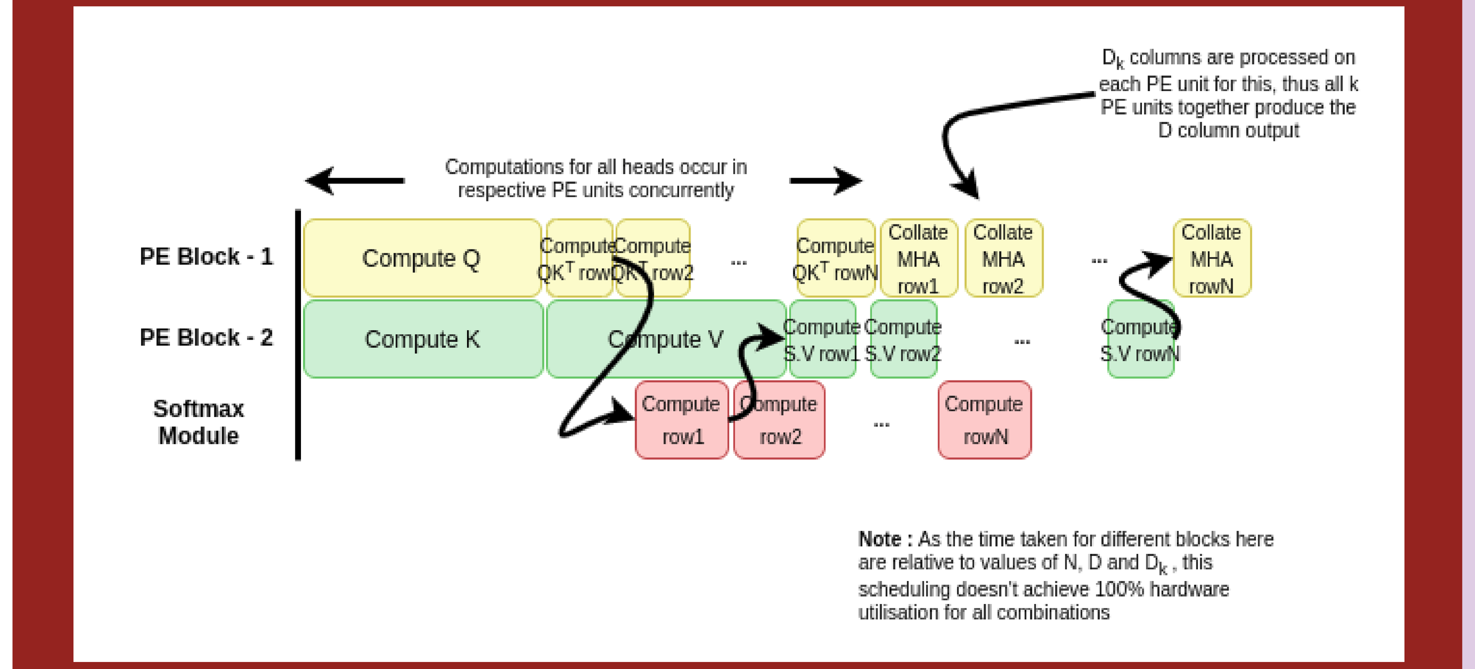

Fig. 1 : Proposed scheduling scheme among the PE blocks.

Proposed Architecture (Design)

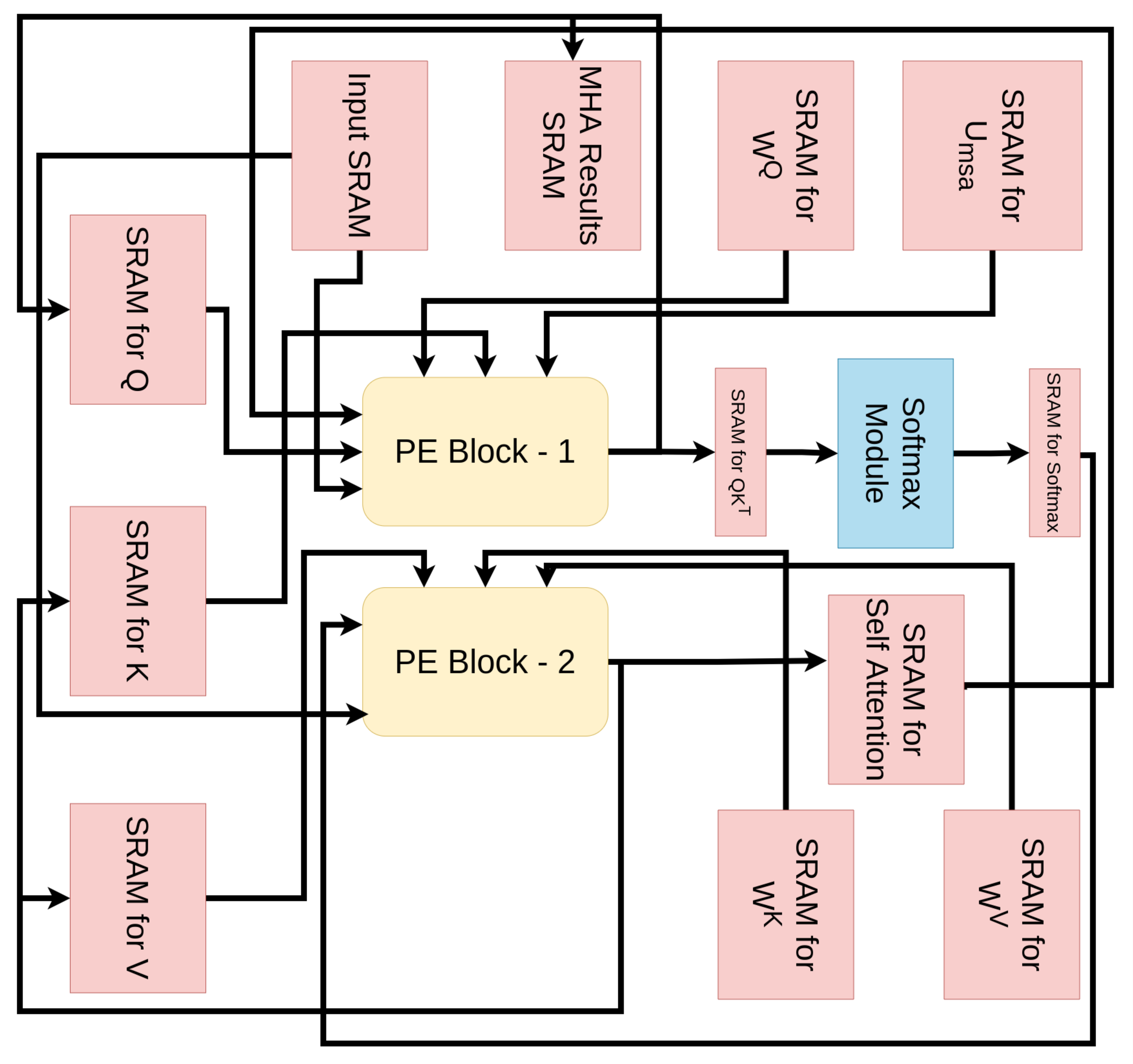

Fig. 2 : Proposed Overall Architecture.

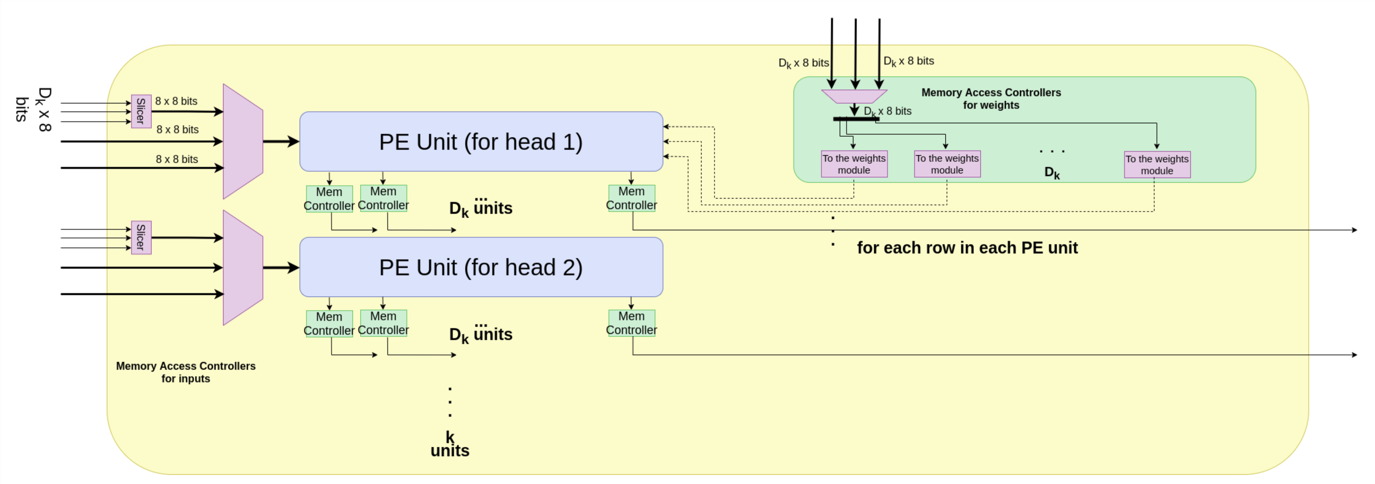

Fig. 3 : A PE Block

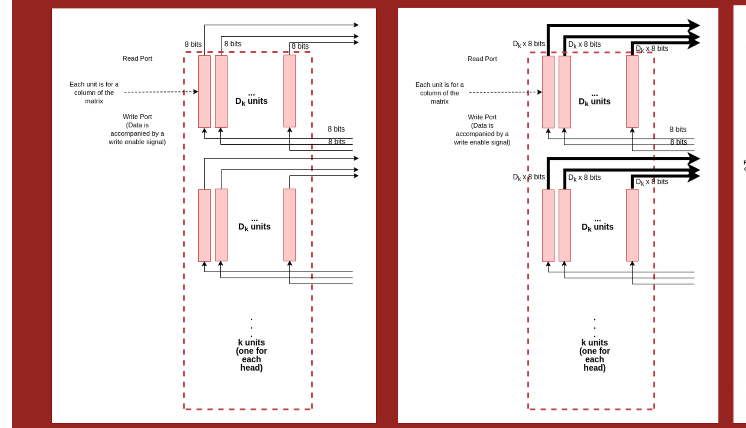

Fig. 4 : Memory Access Patterns

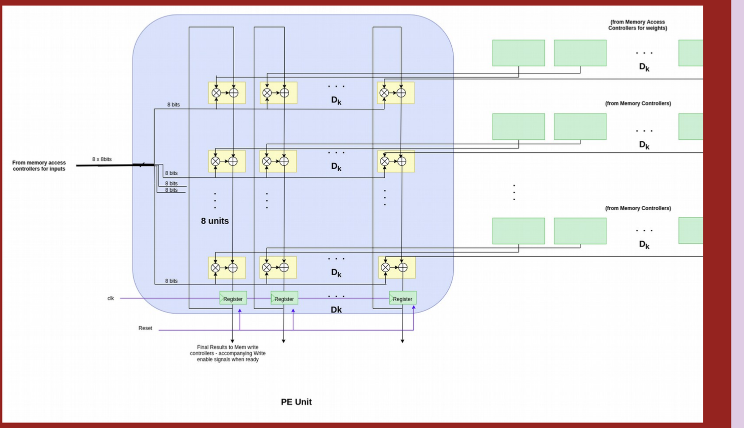

Fig. 5 : A PE Unit

Discussion

- Proposed architecture optimally schedules operations that could be done concurrently, using 2 PE blocks; could potentially achieve better latency than [3]-[5] which have a single PE block.
- Higher hardware utilisation efficiency achieved over [6] which uses separate units for the computations.

Future Work / Ongoing Steps

- Extend the design to the other layers involved in the transformer model, and other vision transformer models.
- Benchmark the implemented design for comparison with the state-of-the-art, and PiM designs like [7].

Acknowledgements

This work was carried out as a part of the IUSSTF – Viterbi Summer Research Experience Program 2022. Special thanks to the Indo – U.S. Science and Technology Forum and the Viterbi School of Engineering, University of Southern California for supporting this program.

References

[1] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, I. Guyon et al., Eds., vol. 30. Curran Associates, Inc., 2017.
[2] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale"
[3] H.-Y. Wang and T.-S. Chang, "Row-wise accelerator for vision transformer," 05 2022
[4] S. Lu et al., "Hardware accelerator for multi-head attention and positionwise feed-forward in the transformer," in 2020 IEEE 33rd International System-on-Chip Conference (SOCC), 2020, pp. 84–89
[5] J. Park et al., "Optimus: Optimized matrix multiplication structure for transformer neural network accelerator," in MLSys, 2020
[6] B. Li et al., "Ftrans: energy-efficient acceleration of transformers using fpga," in Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design, 2020, pp. 175–180
[7] X. Yang et al., "Retransformer: Reram-based processing-in-memory architecture for transformer acceleration," in Proceedings of the 39th International Conference on Computer-Aided Design, ser. ICCAD '20