Appendix: Artifact Description/Artifact Evaluation

Artifact Description (AD)

I. Overview of Contributions and Artifacts

A. Paper's Main Contributions

This paper proposes M^{3}XU, which is a multi-mode matrix processing unit that supports IEEE 754 single-precision and complex 32-bit floating-point computations. M^{3}XU consists of (1) additions of logic to feed different parts of inputs, (2) extensions to the arithmetic units to support exact FP32 precisions, and (3) slight extensions to accumulators to accumulate numbers in double-precision formats. M3XU enables MXUs to handle standard FP32 floating-point numbers and FP32C complex numbers adeptly, achieving their theoretical throughput under current memory technologies and with relatively minor area overhead. The following is a list of the main contributions of the paper.

C_1 HDL implementation of M^{3}XU.
C_2 Performance emulation framwork for M^{3}XU.
C_3 Applications using M^{3}XU.

B. Computational Artifacts

Artifact Persistance:
https://doi.org/YY.YYYY/zenodo.0XXXXX

A_1 **`SystemVerilog`**: HDL source code for M^{3}XU evaluation.
A_2 **`cutlass`**: Matrix multiplication benchmark for performance emulation.
A_3 **`fft`**: Fast Fourier Transform benchmark for case study application.
A_4 **`nebula`**: Neural Network benchmark for case study application.
A_5 **`snapMRF`**: Magnetic Resonance Fingerprinting benchmark for case study application.
A_6 **`knn`**: K-Nearest Neighbors benchmark for case study application.

Artifact ID	Contributions Supported	Related Paper Elements
A_1	C_1	Table 3 Figure 5 (a), 5 (b)
A_2	C_2	Table 2, 4 Figure 4, 5(c), 5(d)
A_3-A_6	C_3	Figures 6-9

II. Artifact Identification

A. Computational Artifact A_1

Relation To Contributions

Provided the source code for:

- HDL for hardware synthesize result.
- Setting up performance emulation framwork for M^{3}XU.
- Performance emulation of case study application using M^{3}XU.

Expected Results

- Area energy consumption of M^{3}XU.
- Raw perfromance data including throughput and latency of M^{3}XU.
- End to end latency of application M^{3}XU.

Algorithm A should be faster than Algorithms C and B in all GPU scenarios.

Expected Reproduction Time (in Minutes)

- 5 hrs for HDL synthesization.
- 120 hrs for M^{3}XU microbenchmark on all input sizes.
- 24 hrs for case study applications.

The expected computational time of this artifact on GPU X is 20 min.

Artifact Setup (incl. Inputs)

Hardware: Nvidia A100 DGX Station. Or equivlent system installed with Nvidia A100 PICE GPUs.
Software:

- Linux kernel version 5.4.0-81-generic
- CUDA 11.4, driver 470.57.0
- Synopsis design compiler with the 45nm FreePDK45 library
- Nvidia cutlass

Datasets: Input data are synthetic by framework or provided.
Installation and Deploymen: Configuration and build commands are provided.

Artifact Execution

- follow ReadME.md for cutlass installation and M^{3}XU integration.
- Configure licenses for Synopsis design compiler.
- Makefiles are provided for each case study applications.

B. Computational Artifact A_2

Relation To Contributions

Provided the source code for:

- HDL for hardware synthesize result.
- Setting up performance emulation framwork for M^{3}XU.
- Performance emulation of case study application using M^{3}XU.

Expected Results

- Area energy consumption of M^{3}XU.
- Raw perfromance data including throughput and latency of M^{3}XU.
- End to end latency of application M^{3}XU.

Algorithm A should be faster than Algorithms C and B in all GPU scenarios.

Expected Reproduction Time (in Minutes)

- 5 hrs for HDL synthesization.
- 120 hrs for M^{3}XU microbenchmark on all input sizes.
- 24 hrs for case study applications.

The expected computational time of this artifact on GPU X is 20 min.

Artifact Setup (incl. Inputs)

Hardware: Nvidia A100 DGX Station. Or equivlent system installed with Nvidia A100 PICE GPUs.
Software:

- Linux kernel version 5.4.0-81-generic
- CUDA 11.4, driver 470.57.0
- Synopsis design compiler with the 45nm FreePDK45 library
- Nvidia cutlass

Datasets: Input data are synthetic by framework or provided.
Installation and Deploymen: Configuration and build commands are provided.

Artifact Execution

- follow ReadME.md for cutlass installation and M^{3}XU integration.
- Configure licenses for Synopsis design compiler.
- Makefiles are provided for each case study applications.

C. Computational Artifact A_3-A_6

Relation To Contributions

Provided the source code for:

- HDL for hardware synthesize result.
- Setting up performance emulation framwork for M^{3}XU.
- Performance emulation of case study application using M^{3}XU.

Expected Results

- Area energy consumption of M^{3}XU.
- Raw perfromance data including throughput and latency of M^{3}XU.
- End to end latency of application M^{3}XU.

Algorithm A should be faster than Algorithms C and B in all GPU scenarios.

Expected Reproduction Time (in Minutes)

- 5 hrs for HDL synthesization.
- 120 hrs for M^{3}XU microbenchmark on all input sizes.
- 24 hrs for case study applications.

The expected computational time of this artifact on GPU X is 20 min.

Artifact Setup (incl. Inputs)

Hardware: Nvidia A100 DGX Station. Or equivlent system installed with Nvidia A100 PICE GPUs.
Software:

- Linux kernel version 5.4.0-81-generic
- CUDA 11.4, driver 470.57.0
- Synopsis design compiler with the 45nm FreePDK45 library
- Nvidia cutlass

Datasets: Input data are synthetic by framework or provided.
Installation and Deploymen: Configuration and build commands are provided.

Artifact Execution

- follow ReadME.md for cutlass installation and M^{3}XU integration.
- Configure licenses for Synopsis design compiler.
- Makefiles are provided for each case study applications.

Artifact Analysis (incl. Outputs)