2022 ASPDAC

A Reconfigurable Inference Processor for Recurrent Neural Networks Based on Programmable Data Format in a Resource-Limited FPGA

Jiho Kim, Kwanyoung Park, Taehwan Kim

KOREA AEROSPACE UNIVERSITY

01 Bio

Ji-Ho Kim

- A 1st year M.S. student in electronics and information engineering at Korea Aerospace University

- Education

  - Korea Aerospace University, South Korea, B.S. (2021)

- Research interest

  - Low-power & resource-efficient deep-neural network inference / training processor
  - Reconfigurable architecture
  - Embedded systems design

02 Outline

- Motivations
- Contributions
- Inference & verification results
- Demonstration video
- Conclusion

03 Motivations & Contributions

1. Motivations

- Need to support various RNN models of different types
  - No omnipotent model always outperforms the others
- Supporting only a single data format is not efficient
  - Different data distributions for the activation vectors

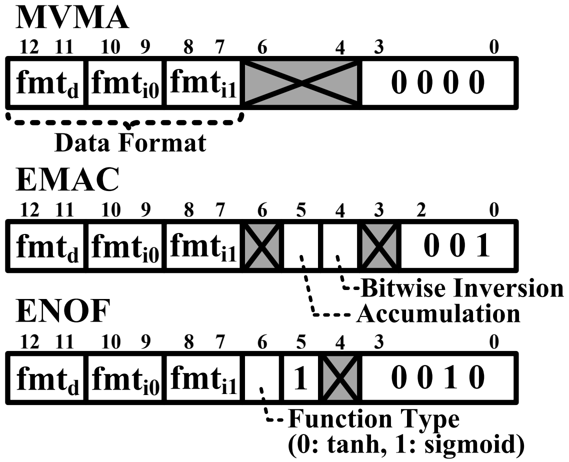
2. Contributions

- **RNN-specific instruction-set processor** providing dataflow reconfigurability
- **Resource-efficient architecture** based on a single array of multipy-accumulate units
- **Programmable data format** to achieve wide range / high precision

1.89MOP/LUT and **263.95GOP/J** in a low-cost FPGA system
The functionality has been verified under a fully-integrated inference system

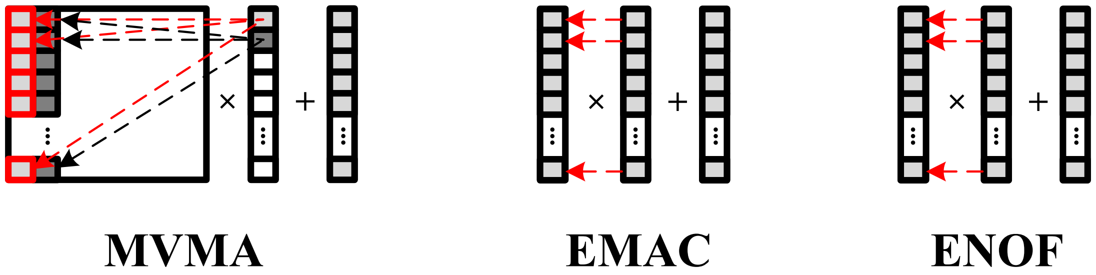
04 Resource Efficient RNN-Specific Instruction-Set Processor

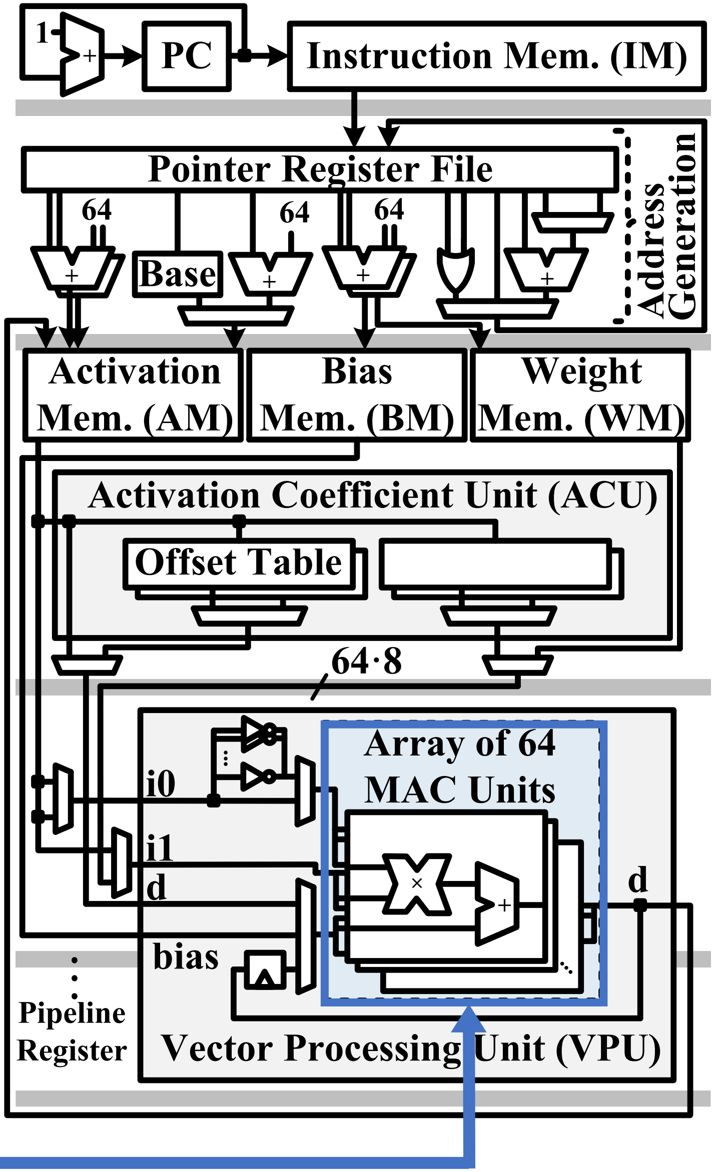
1. RNN-specific instruction-set directly supports primitive vector ops:



- **MVMA** : **M**atrix-**V**ector **MA**C
- **EMAC** : **E**lementwise **MAC**
- **ENOF** : **E**lementwise **NO**n-linear **F**unction

2. SIMD instruction-set processor
   - 64 scalar operations per cycle through 6 pipeline stages
3. VPU has been designed based on a single array of MAC units
   - Every vector operation can be performed with high resource eff.



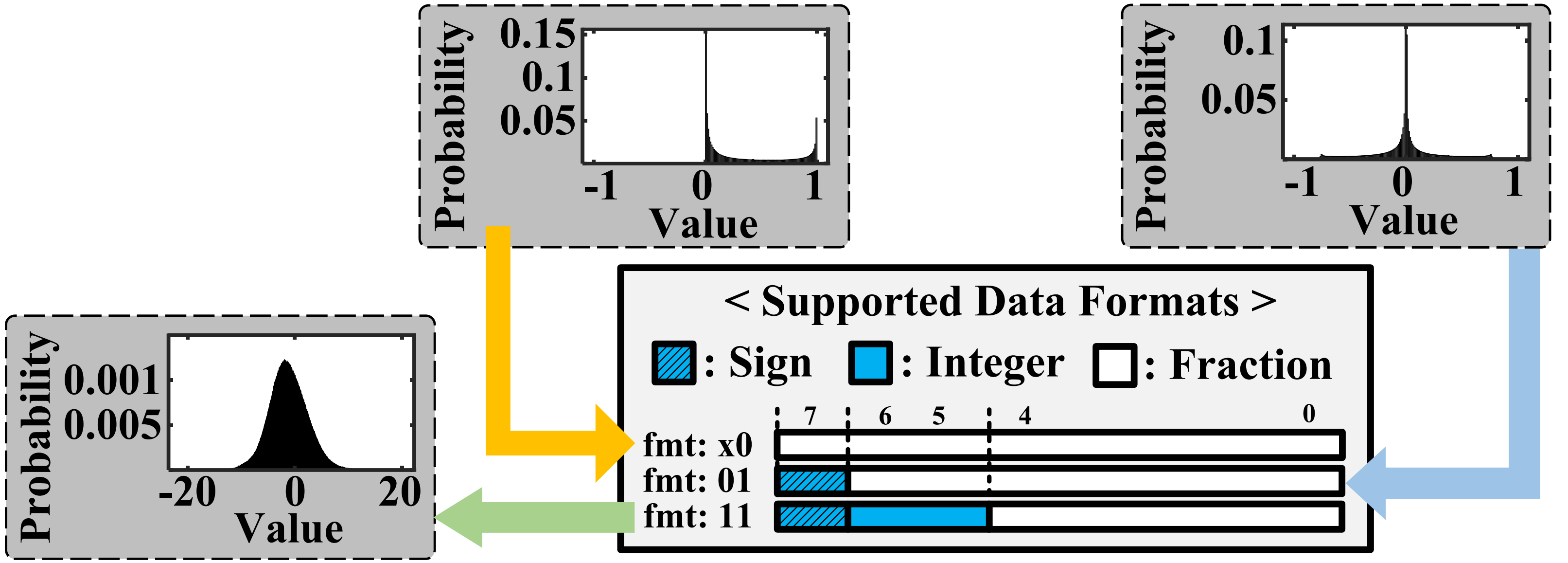


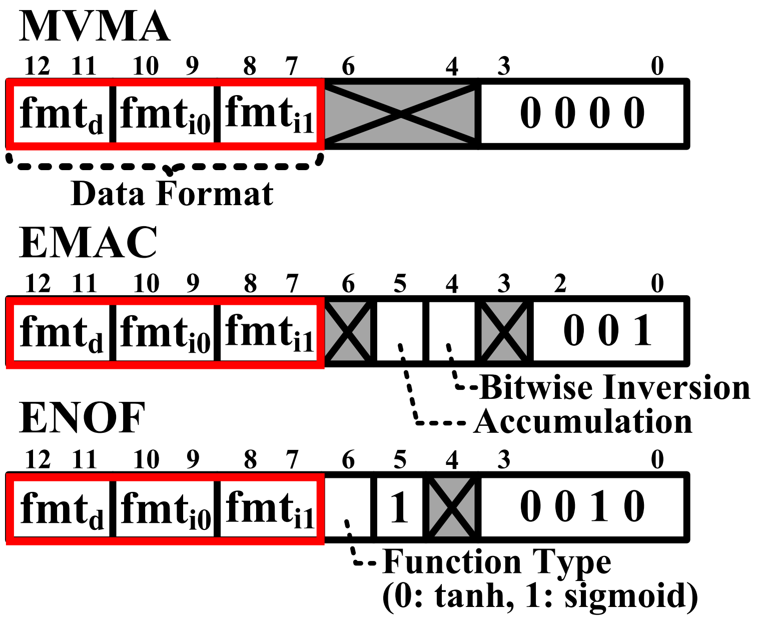
05 Programmable Data Format

- Data format of each vector operand is **programmable**

  → Achieve **wide range** / **high precision**

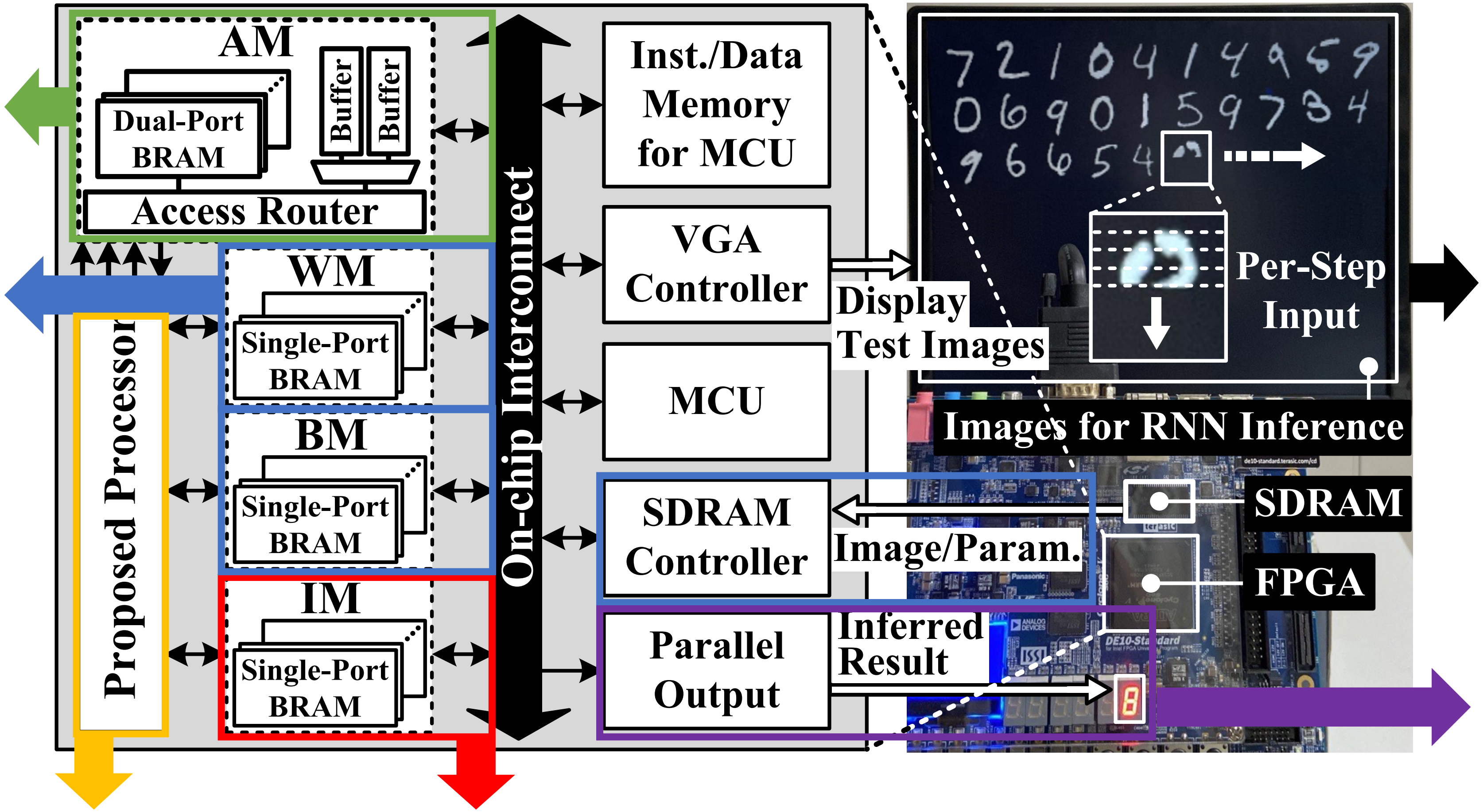
- Data size in the proposed processor is as **small** as 8 bits, while the inference results are **maintained**.





06 Overall Inference System

< A Prototype Inference System & Verification environment >



③ Load input data to the proposed processor

② Weight and bias into WM, BM from SDRAM

Sequential MNIST dataset

Display the Inference result

④ The proposed processor performs the inference by running the program

① MCU preloads the data of the dataflow description program

07 Implementation Results

- Implemented in a **low-cost resource-limited** FPGA device

		This work	**[1]**	**[2]**
FPGA device		Cyclone V	Cyclone V	Virtex-7
RNN model type		Reconfigurable	Reconfigurable	Only LSTM
Data format		Programmable 8bit	Fixed 16bit	Fixed 16bit
Resource usage	LUT	12K	18K	23K
	BRAM	801K	1620K	5940K
	DSP	64	64	142
Inference speed (GOP/s)		22.66	23.00	10.90
Resource Efficiency (MOP/s/LUT)		**1.89**	**1.28**	**0.46**
Energy Efficiency (GOP/J)		**263.95**	**166.31**	**7.10**

[1] J. Kim, J. Kim, and T.-H. Kim, “AERO: a 1.28 MOP/s/LUT reconfigurable inference processor for recurrent neural networks in a resource-limited FPGA,” *MDPI Elect.*, vol.10, no.11, pp.1249-1263, Apr. 2021.

[2] W. Zhang, F. Ge, C. Cui, Y. Yang, F. Zhou, and N. Wu, “Design and implementation of LSTM accelerator based on FPGA,” in *Proc. Int’l Conf. Comm. Tech.*, IEEE, Oct.2020, pp. 1675-1679.

08 Verification Results

- Proposed processor has been Verified for several **different models & datasets**

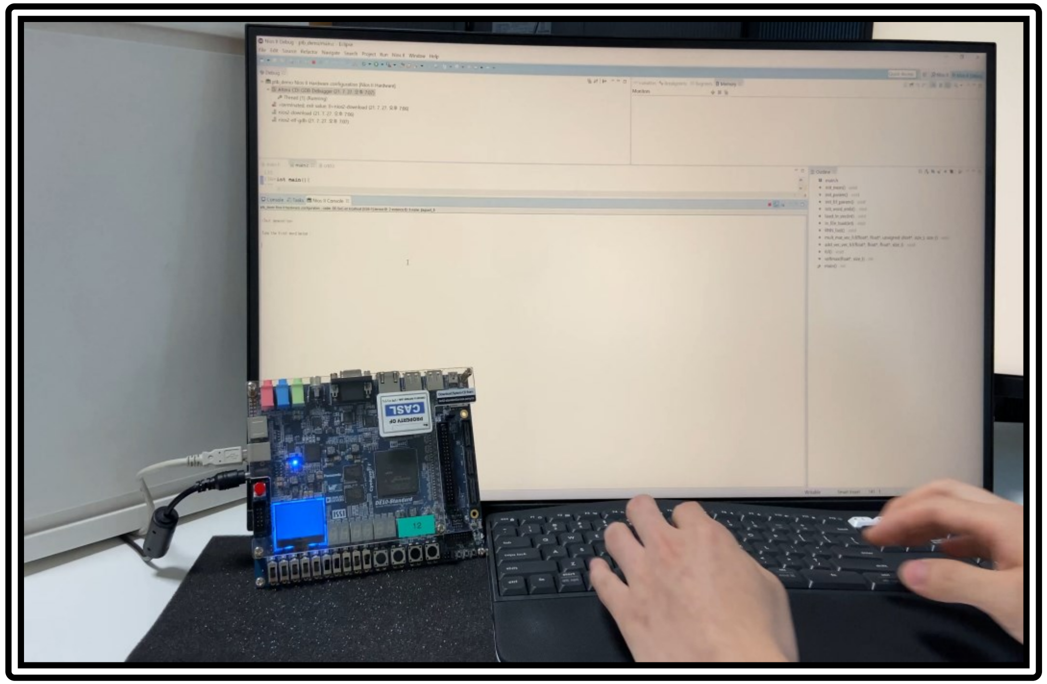
Dataset	Sequential MNIST	
RNN model	LSTM	Peephole GRU
Inference Performance	98.65%	98.17%
Performance loss a)	0.22%	0.61%
Per-step latency	13.64μs	8.56μs

Dataset	Penn Treebank	
RNN model	Peephole LSTM	Bidirectional GRU
Inference Performance	95.17 ppl	2.74 ppl
Performance loss a)	0.73 ppl	0.89 ppl
Per-step latency	11.95μs	20.51μs

a) The performance loss has been obtained by the difference from the inference performance achieved with the 32-bit floating-point data

09 Demonstration Video

: QR code of the demonstration video



	Sequential Image Classification	Natural Language Processing
Dataset	Sequential MNIST	PennTree bank
Model Size	LSTM-128H-64X	GRU-128H-128X-
Inference result	98.65%	97.41 ppl

10 Conclusion

- An efficient inference processor for RNN is designed and implemented in FPGA
- Proposed processor is designed to be reconfigurable for various models and perform every vector operation consistently utilizing a single array of MAC units
- Programmable data format is supported with the size of 8 bits, while achieving a near-floating-point inference results
- Resource efficiency and energy efficiency are 1.89 MOP/LUT and 263.95 GOP/J, respectively
- 4.1 times higher resource efficiency than the previous state-of-the-art result
- The functionality has been verified successfully for several different models under a fully-integrated inference system

Thank You!

Presenter: Ji-Ho Kim
Presenter’s E-mail: jhkim_ms@kau.kr
Corresponding author’s E-mail: taehwan.kim@kau.ac.kr