

主 题:	[ISOC 2020] Decision Notification (Paper ID=39)	
发件人:	"Microsoft CMT" <email@msr-cmt.org>	2020-8-18 4:52:50
收件人:	"Jiayi Liu" <3170105617@zju.edu.cn>	

Dear Author,

We are pleased to inform you that your submitted paper is accepted by the Technical Program Committee for presentation at ISOC 2020.

The information on your submitted paper is as follows:

Id of submission: 39

Title of submission: A Novel Scheme to Map Convolutional Networks to Network-on-Chip with Computing-In-Memory Nodes

Status of submission: Accept (Oral)

Track of submission: Machine Learning and AI

Presentation Type: Oral

Please visit the following Microsoft CMT site in order to check the detailed review result regarding your submitted paper.

<https://cmt3.research.microsoft.com/2020ISOC>

*** Camera-Ready Paper Submission & Author Registration Deadline ***

: Friday, September 4, 2020 (KST, GMT+9)

Acceptance of your paper is made with the understanding that at least one author will attend the conference to present the paper. If a paper is not presented during a session, it causes severe disruption to the session. To avoid such interruptions, we would greatly appreciate it if you would inform us and withdraw your paper as soon as possible in the event that you or a co-author cannot attend the conference to present the paper. In order for the author's paper to be presented at the conference and included in the conference proceedings, at least one author of each accepted paper must register for the conference at a full registration rate no later than September 4, 2020 (author and early registration deadline).

Electronic submission of the camera-ready version of manuscripts is now open and will be due by September 4, 2020. Please note that your camera-ready paper should be a 2-page IEEE-conference-paper format explained on the homepage (www.isoc.org). In addition, an IEEE Electronic IEEE Copyright Form(eCF) should be submitted as you submit the final manuscript.

Detailed instructions regarding the preparation and submission of the camera-ready version of your paper are available on the conference website from Wednesday, August 19th.

If you have any questions concerning the submission of your paper, please check the conference website or contact us.

On behalf of the ISOC 2020 Technical Program Committee, we thank you for your contribution and look forward to seeing you at the conference in Yeosu, Korea.

Sincerely,

Jongsun Park
Technical Program Committee Chair
The 17th International SoC Design Conference (ISOC 2020)
E-mail: jongsun@korea.ac.kr

ISOC 2020 Secretariat
E-mail: secretary@isoc.org

Microsoft respects your privacy. To learn more, please read our [Privacy Statement](#).

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

A Novel Scheme to Map Convolutional Networks to Network-on-Chip with Computing-In-Memory Nodes

Jiayi Liu

College of Information Science and Electronic Engineering
Zhejiang University
Zhejiang, China
3170105617@zju.edu.cn

Kejie Huang

College of Information Science and Electronic Engineering
Zhejiang University
Zhejiang, China
huangkejie@zju.edu.cn

Abstract— **Computing-In Memory (CIM)** has been widely used to accelerate the inferencing speed of deep learning. **Network-on-Chips (NoCs)** are usually used together with CIM to enable the versatile ability of the hardware. This paper proposes a bandwidth aware mapping scheme to minimize both hops and bandwidth requirement. The simulation results show that the proposed scheme could reduce the hops and bandwidth requirements by more than 33.57% and 46.13%, respectively.

Keywords; *in-memory computing; mapping; NoC; NSGAI*

I. INTRODUCTION

The Computing In-Memory (CIM) is one of the most promising techniques in the post Moore's Law era to break the power and memory walls. The Resistive Non-Volatile Memory (RNVN) like RRAM has shown great potential in low power massive parallel in-memory computing. However, most of the existing works only focus on the core level design. The load-store scheme will be inefficient for CIM, because of the high write power and slow write speed of the emerging RNVN. Therefore, new flexible interconnecting architectures and mapping strategies should be developed to meet the various requirements of neural networks. Adopting Network-on-Chip (NoC) in CIM for high parallelism and scalability has attracted wide interests from both industry and academia [1].

In this paper, we propose a novel scheme to automatically map a Convolutional Neural Network (CNN) to a 2-D mesh NoC. The processing node is based on the RRAM-based $N \times N$ CIM array. To reduce the latency, power consumption, and bandwidth of routers, and maximize the parallelism, the weights of each layer are duplicated, partitioned, and placed to simplify the dataflow in NoC. The results show that our scheme could greatly the number of hops and the bandwidth requirement.

II. FRAMEWORK

NoC has been widely proposed for versatile interconnect for CIM [4]. As shown in Fig. 1, RRAM based Processing Elements (PEs) taken from [2-3] are adopted as the processing node in NoC which consists of an $N \times N$ MAC array. PEs are interconnected by the routers to enable various neural network connections. In this section, we describe the mapping strategies, mainly about how to compute CNN based on NoC and the mapping algorithm. In our scheme, the weights and connections of CNN are mapped to PEs and the paths of routers, respectively.

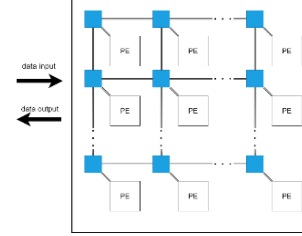


Figure 1. The architecture of NoC

A. The mapping strategies

1) *Weight copy*: The duplication of weights is to maximize the parallel computing, and reduces the delay caused by the synchronization between layers. According to the input data amount of each layer, the number of copies of the weights m_n can be determined by:

$$\frac{t_i}{m_i} = \frac{t_j}{m_j} \quad (1)$$

where t_i is the computing time, which is positively correlated to the amount of input data.

2) *Mapping order*: As shown in Fig. 2, the weights mapping to PEs in one layer can be expressed as the longitudinal-direction expansion and the horizontal expansion, representing the partial sum and other situations, respectively. This is to simplify the PE splitting and merging.

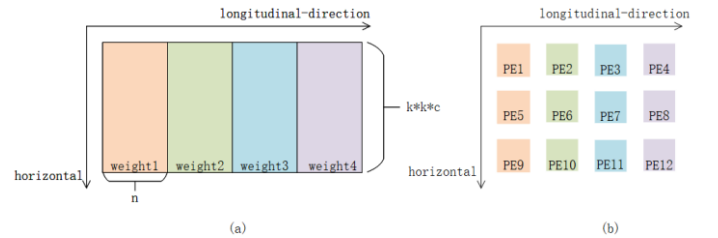


Figure 2. Mapping the weights to PEs

Supposing the input of the n^{th} layer in the network is $W_n \times H_n \times c_n$, the output is $W'_n \times H'_n \times c'_n$, and the size of the kernel is $k_n \times k_n \times c_n \times n_n$. The number of PEs is $b_n \times b'_n$, where b_n and b'_n are the required PEs in horizontal and the longitudinal direction, respectively. They are defined as:

主 题: [ISOC 2020] Decision Notification (Paper ID=39)

发件人: "Microsoft CMT" <email@msr-cmt.org>

2020-8-18 4:52:50

收件人: "Jiayi Liu" <3170105617@zju.edu.cn>

Dear Author,

We are pleased to inform you that your submitted paper is accepted by the Technical Program Committee for presentation at ISOC 2020.

The information on your submitted paper is as follows:

Id of submission: 39

Title of submission: A Novel Scheme to Map Convolutional Networks to Network-on-Chip with Computing-In-Memory Nodes

Status of submission: Accept (Oral)

Track of submission: Machine Learning and AI

Presentation Type: Oral

Please visit the following Microsoft CMT site in order to check the detailed review result regarding your submitted paper.

<https://cmt3.research.microsoft.com/2020ISOC>

*** Camera-Ready Paper Submission & Author Registration Deadline ***
: Friday, September 4, 2020 (KST, GMT+9)

Acceptance of your paper is made with the understanding that at least one author will attend the conference to present the paper. If a paper is not presented during a session, it causes severe disruption to the session. To avoid such interruptions, we would greatly appreciate it if you would inform us and withdraw your paper as soon as possible in the event that you or a co-author cannot attend the conference to present the paper. In order for the author's paper to be presented at the conference and included in the conference proceedings, at least one author of each accepted paper must register for the conference at a full registration rate no later than September 4, 2020 (author and early registration deadline).

Electronic submission of the camera-ready version of manuscripts is now open and will be due by September 4, 2020. Please note that your camera-ready paper should be a 2-page IEEE-conference-paper format explained on the homepage (www.isoc.org). In addition, an IEEE Electronic IEEE Copyright Form(eCF) should be submitted as you submit the final manuscript.

Detailed instructions regarding the preparation and submission of the camera-ready version of your paper are available on the conference website from Wednesday, August 19th.

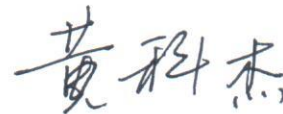
If you have any questions concerning the submission of your paper, please check the conference website or contact us.

On behalf of the ISOC 2020 Technical Program Committee, we thank you for your contribution and look forward to seeing you at the conference in Yeosu, Korea.

Sincerely,

Jongsun Park
Technical Program Committee Chair
The 17th International SoC Design Conference (ISOC 2020)
E-mail: jongsun@korea.ac.kr

ISOC 2020 Secretariat
E-mail: secretary@isoc.org



Microsoft respects your privacy. To learn more, please read our Privacy Statement.

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052