A Reconfigurable Matrix Multiplication Coprocessor with High Area and Energy Efficiency for Visual Intelligent and Autonomous Mobile Robots

[1]Jipeng Wang, [1]Yi Zhan, [1]Zhaoxu Wang, [1]Zixuan Peng, [1]Jiarui Xu, [1]Bingqiang Liu, [1]Guoyi Yu, [2]Fengwei An, [1]Chao Wang, [1]Xuecheng Zou

[1]Huazhong University of Science and Technology, Wuhan, China

[2]Southern University of Science and Technology, Shenzhen, China

2021 IEEE Asian Solid-State Circuits Conference (A-SSCC) | 978-1-6654-4350-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/A-SSCC53895.2021.9634793

Matrix multiplication is an essential mathematical calculation in a wide range applications of signal processing, computer graphics and intelligent robots. The intelligent and autonomous robots involves various navigation algorithms (e.g. Extended Kalman Filter (EKF), reinforcement learning, A* and artificial potential field, etc.) [1-4] and deep neural network (DNN) algorithms (e.g. Darknet in YOLOv3), which all contain intensive matrix multiplications with different sizes and shapes. The emerging Intelligent and Autonomous Mobile Robots (I-AMRs) have put forward to a higher demand to efficient hardware acceleration of a comprehensive range of matrix multiplications as depicted in Fig. 1. Recent works have focused on the hardware acceleration of matrix multiplications optimized for a specified navigation or DNN algorithm [3]-[5], which cannot achieve high hardware utilization, high area and energy efficiency for the various matrix multiplications in I-AMRs.

In this paper, a Reconfigurable Matrix Multiplication (RMM) coprocessor is proposed with following features: i) a reconfigurable coprocessor architecture to accelerate matrix multiplications in versatile navigation and intelligent vision algorithms; ii) a reconfigurable systolic array (RSA) with novel matrix segmentation and mapping scheme for matrix multiplications of different sizes and shapes to achieve high hardware utilization and processing throughput; iii) a local memory buffer (LMB) organization with multiple banks and buffers for parallel processing and data reuse, and an address generation and decode (AGD) unit for the exploitation of data sparsity and symmetry, to increase overall area and energy efficiency.

The intelligent robot SoC consists of an Application Processing Unit (APU), a RMM coprocessor #0 as the navigation core, a RMM coprocessor #1 as the intelligent core and (*N*-2) RMM coprocessors for other specific applications as depicted in Fig. 2. The APU has a CPU core responsible for control, configuration and other calculations except the matrix multiplications accelerated in the RMM cores. Each RMM coprocessor is comprised of a Direct Memory Access (DMA) for data movement between off-chip memory and on-chip local memory, a 16-KB LMB with an AGD unit, a Finite State Machine (FSM) for top-level control and configuration, and a RSA with *x* rows and *y* columns of Processing Elements (PEs). The RSA can support a comprehensive range of matrix multiplications by reconfiguring the PEs and their interconnections according to the different sizes and shapes of multiplications. In the RSA, each PE contains a MAC with a northern input, two western inputs, a southern output and an eastern output, of which can be activated or disabled by MUX configuration to process the data flow in corresponding mode for different matrix multiplications.

The RMM coprocessor #0 & #1 are designed as the navigation and intelligent cores to exploit the diversity and data features in matrix multiplications of EKF and Darknet-19 algorithms as depicted in Fig. 1. In the computationally-intensive EKF algorithm that mainly contains matrix calculations of robot and landmarks' coordinates, there are actually 7 kinds of matrix multiplications of different sizes and shapes. Similarly, the different matrix multiplication calculations also exist in the Darknet-19 CNN algorithm. By flattening each convolution layer of Darknet-19, those three-dimensional convolution operations can mapped into 19 two-dimensional matrix multiplications. Hence, efficient matrix multiplication acceleration can be achieved by segmenting these matrix multiplications and mapping them into the RSAs of the navigation and intelligent cores. In addition, low power consumption can be achieved by exploiting the data sparsity and symmetry through the AGD in the LMB organization, as shown in Fig. 3.

The LMB organization has four data buffers for data buffering and reusing, of which each data buffer contains *L* SRAM banks. *L* is the larger size value of *x* and *y* of RSA to achieve the highest parallelism that is determined by the supporting modes of data flows on the RSA. As an example, for the 10 matrix multiplications of EKF algorithm, the corresponding switching network scheme and address coding scheme are implemented by the LMB configuration to enable the high-parallel processing and exploit the data reuse and sparsity. Observing a specific correlation of data flow in the EKF algorithm, e.g. the output C of multiplication (1) and the input A of multiplication (2), the data reuse can be realized by switching the I/O roles of the four data buffers and temporally storing the reused data in the local buffers. It's worth noting that the transpose of matrix can also be implemented by address decoding according to the data flow. The proposed AGD significantly reduce data computation as well as data movement by exploiting the properties of data structure and flow to maximize energy efficiency.

As a design case, the RMM navigation core has two working modes with respective matrix segmentation and mapping scheme as depicted in Fig. 4. Instead of designing a specific working mode for each type of EKF matrix multiplication, only two working modes are proposed to achieve an optimum trade-off between design complexity, control hardware overhead and average RSA hardware utilization, by considering the fact that the multiplication (8) occupies more than a half of total computation, i.e., 60%. In addition, to achieve an optimum trade-off between processing throughput, RSA hardware overhead and hardware utilization, RSA size *x* is determined by a suitable common factor of matrix B size *k* in mode 0 and matrix A size *n* in mode 1, by considering the idea that matrix segmentation should make the last segmented submatrix to fit the RSA as much as possible to maximize the RSA hardware utilization. Similarly, RSA size *y* is determined by a common factor of matrix B size *k* in mode 1 but cannot be much larger than matrix A size *n* in mode 0 according to the matrix segmentation and mapping scheme.

Figure 6 shows the implementation results for the proposed RMM coprocessor. Although it costs more slice LUTs, the RMM navigation core can support all 7 types of matrix multiplications in EKF algorithm at a 93.47% hardware utilization, while the benchmark design in [5] only support 1 type of multiplication. This RMM navigation core achieves a high precision (32-bits) but just occupies only 26-KB SRAM thanks to the exploitation of matrix's symmetry and sparsity. Benefiting from the LMB with multiple banks and buffers, the achieved peak performance is 924.5 MOPS, i.e., 1.09x higher than [5], while the core area is just 0.33 mm^2 and its area efficiency is 2.80 GOPS/mm^2. The overall energy efficiency is 1.98 GOPS/W, i.e., 4.83x higher than [5]. Figure 5 shows the algorithm flow chart of the EKF based navigation. Figure 7 shows the demo photo of the coprocessor for AMRs. In overall, the proposed RMM design can efficiently accelerate various matrix multiplications in a wide range of i-AMR algorithms, which is very competitive in a resource-constrained mobile robot platform.

Acknowledgement:

This work was jointly supported by National Key R&D Program of China (2019YFB1310001) and National Natural Science Foundation of China (61974053). Jipeng Wang and Yi Zhan contributed equally to this paper. Corresponding author: Chao Wang.

References:

[1] P. Geneva *et al*., "A Linear-Complexity EKF for Visual-Inertial Navigation with Loop Closures," *Intl. Conf. Robot. Autom.*, pp. 3535-3541, 2019.

[2] P. Geneva *et al*., "An Efficient Schmidt-EKF for 3D Visual-Inertial SLAM," *IEEE/CVF Conf. Comput. Vision Patt. Recog.*, pp. 12097-12107, 2019.

[3] Y. Kim *et al*., "A 0.55 V 1.1 mW Artificial Intelligence Processor with On-Chip PVT Compensation for Autonomous Mobile Robots," in *IEEE TCAS-I*, vol. 65, no. 2, pp. 567-580, Feb. 2018.

[4] N. Cao *et al*., "A 65-nm 8-to-3-b 1.0–0.36-V 9.1–1.1-TOPS/W Hybrid-Digital-Mixed-Signal Computing Platform for Accelerating Swarm Robotics," in *IEEE JSSC*, vol. 55, no. 1, pp. 49-59, Jan. 2020.

[5] D. T. Tertei *et al*., "FPGA Design of EKF Block Accelerator for 3D Visual SLAM," *Comput. Elect. Eng.*, vol. 55, pp. 123-137, 2016.

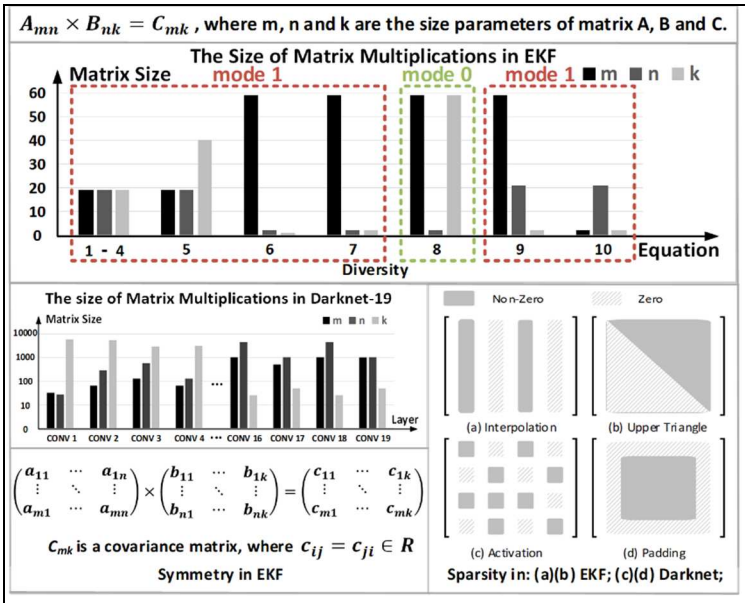

Fig. 1. Diversity and data features of matrix multiplications.

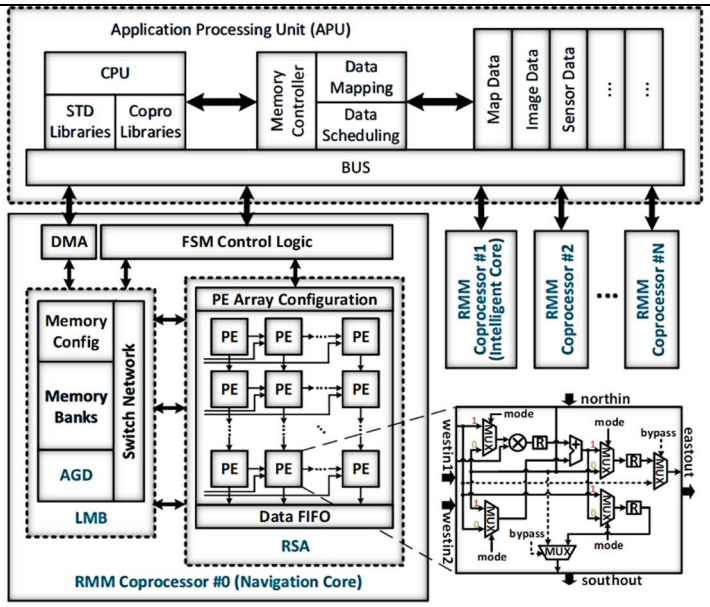

Fig. 2. Proposed RMM coprocessors in intelligent robot SoC.

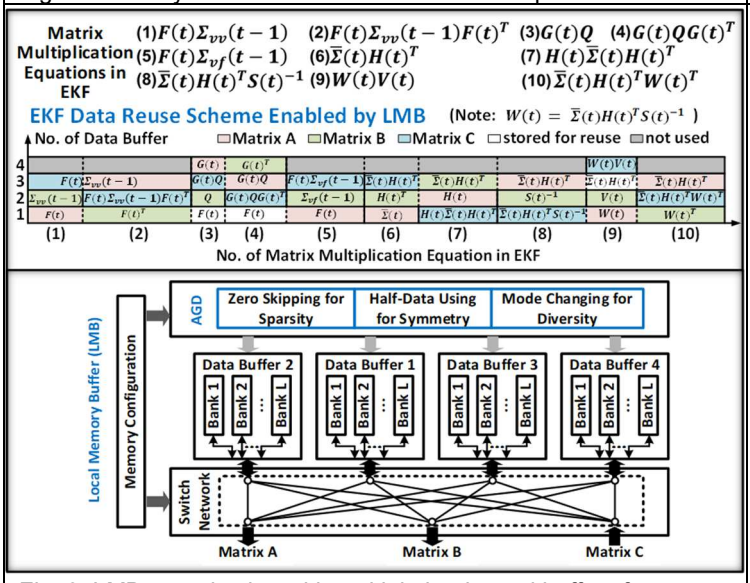

Fig. 3. LMB organization with multiple banks and buffers for parallel processing and data reuse.

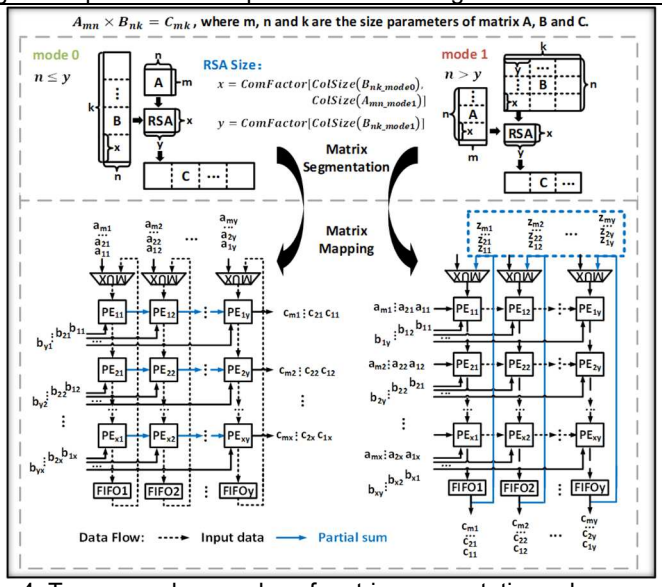

Fig. 4. Two examplary modes of matrix segmentation scheme and matrix mapping scheme for the RSA design of EKF acceleration.

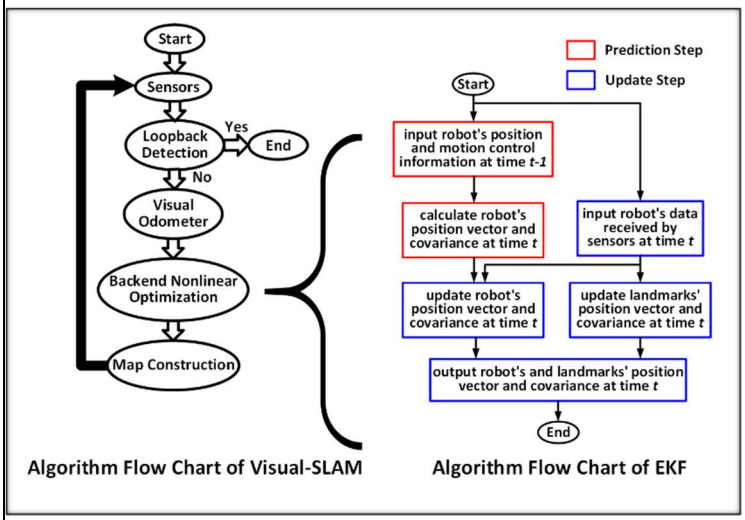

Fig. 5. Algorithm Flow Charts of EKF-SLAM. The EKF occupies most of the computation required by the Visual-SLAM (Simultaneous Localization and Mapping) algorithm.

	TCAS-I'18 [3] (Path Plan SoC)	JSSC'20 [4] (Path Plan SoC)	ELSEVIER CEE'16 [5] (EKF Accelerator Core)	This Work (EKF Navigation Core)
Process	65 nm	65 nm	N/A	65 nm*
FPGA Platform	N/A	N/A	Zynq 7020	Zynq 7000
Slice LUTs	N/A	N/A	7824	13586
Slice FFs	N/A	N/A	N/A	3793
DSPs	N/A	N/A	32	30
SRAM Size	891 KB	16 KB	94.5 KB	26 KB
Bit Precision	N/A	8 bits	32 bits	32 bits
Supply Voltage	1.2 V	1 V	1.8 V	1.8 V (1.2 V*)
Area	16 mm²	2 mm²	N/A	0.33 mm²*
Max Frequency	245 MHz	1.5 MHz	177 MHz	108 MHz (333 MHz*)
Power Consumption	151 mW	3.4 µW	1.302 W	0.467 W (14 mW*)
Hardware Utilization	N/A	N/A	99%	93.47%
Peak Performance	470k search/s	30.94 MOPS	442.5 MOPS	924.5 MOPS
Area Efficiency	29.38k search/s/mm²	15.47 MOPS/mm²	N/A	2.80 GOPS/mm²*
Energy Efficiency	320 nJ/search	9.1 TOPS/W	339.9 MOPS/W	1.98 GOPS/W (66.0 GOPS/W*)
Multisize Support	N/A	N/A	No	Yes (7 shapes and sizes)
Application	AMRs	Swarm Robotics	AMRs	I-AMRs

Fig. 6. Proposed RMM coprocessor characteristics and performance evaluation compared to the prior designs. Note: * represents the ASIC synthesis results.

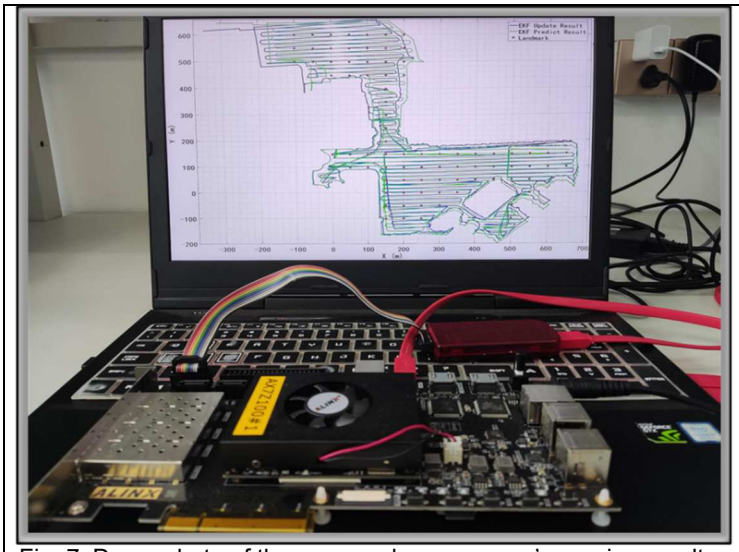

Fig. 7. Demo photo of the proposed coprocessor's running results for the localization and mapping of an office area in a company's building.