A 22nm Resource-Frugal Hyper-Heterogeneous Multi-Modal System-on-Chip Towards In-Orbit Computing

Quan Cheng[1,2], Qiufeng Li[2], Weirong Dong[2], Mingtao Zhang[1], Ruilin Zhang[1], Mingqiang Huang[2], Hao Yu[2], Yiyu Shi[3], Hiromitsu Awano[1], Takashi Sato[1], Longyang Lin[2], Masanori Hashimoto[1]

[1]Kyoto University, Kyoto, Japan, [2]Southern University of Science and Technology, Shenzhen, China, [3]University of Notre Dame, USA

Integrating artificial intelligence (AI) into in-orbit computing offers significant benefits, but current satellites face challenges in processing large sensor data volumes due to limited communication and computing resources, resulting in high latency [1]. Intelligent Early Discard (IED) [2] addresses this by filtering irrelevant data early, optimizing bandwidth and data usage. However, this demands high-performance onboard computing for efficient data preprocessing and AI acceleration [3, 4]. Additionally, Space radiation, including solar energetic particles and cosmic rays, can cause Single Event Upsets (SEUs) in satellite systems [5], risking mission failure and increasing reliability demands [6, 7]. To tackle these challenges, we propose a resource-frugal hyper-heterogeneous System-on-Chip (SoC) architecture for in-orbit computing. The SoC features two modes: (1) a specialized computation engine for AI acceleration, and (2) a multi-core mode with dual-core lock-step (DCLS) and vector computing for efficient, fault-tolerant data processing (Fig. 1). This resource-frugal architecture enables full sharing of Processing Elements (PEs) and memories for dynamic workload allocation, enhancing in-orbit performance by processing IED data directly on the satellite and reducing costly data transmission to Earth.

The architecture in Fig. 2 illustrates the proposed SoC for in-orbit computing, featuring 8 Processing Cores (PCs). Each PC includes two 3-stage RISC-V cores with DCLS checkers at each pipeline stage (fetch, decode, execute), which can be toggled on or off for fault tolerance and flexibility. When disabled, the RISC-V cores operate independently. When enabled, DCLS checkers log errors, triggering pipeline rollback for recovery and user notification. Also, these cores are optimized for vector computing, supporting inter-PC systolic data flow among PEs, each equipped with 16 INT8 multiply-accumulates (MACs). The system's data paths are interconnected via crossbar switches, enabling efficient data movement between cores and memory blocks. These memory blocks can be configured as either Instruction/Data Tightly-Coupled Memories (I/DTCMs) or activation/weight memories (A/WMEMs), optimizing data storage and retrieval. Memory access is handled via 32-bit interfaces for the RISC-V cores and JTAG, and 128-bit interfaces for DMA and PEs.

As shown in Fig. 3, the SoC operates in two modes: (1) computation engine mode and (2) multi-core mode. In computation engine mode (left), the SoC functions as a single AI engine, with only PC#1 active as the controller. PEs in each PC form a systolic array for efficient computation and data reuse by up to 16 times, while 16 inter-PC PEs create a pipeline for high-throughput processing. The compute scheduler coordinates tasks, directing data flow through PEs in sequence (#1, #2, etc.), maximizing the computation efficiency in a linear configuration. If the inter-PC path carries activations, the intra-PC path carries weights, and vice versa. In contrast, in multi-core mode (right), all PCs are active, and each PE is controlled by its respective RISC-V core, with enhanced vector computing capabilities. The cores can execute custom instructions and communicate via interconnected buses, providing parallelism and scalability. Mode switching is controlled by PC#1 with a one-cycle delay. Notably, with shared PEs and memory blocks for resource-frugal computing in both modes, the standalone computation engine design can be extended to support multi-core mode with only a 12% increase in area, primarily from other RISC-V cores.

As shown in Fig. 4, the proposed SoC offers efficient mode-specific memory management (left). Memory blocks consist of two primary types of memory: I/DTCM and A/WMEM, each with specific roles in handling data and instructions for computation. In addition, the partition boundaries can be tuned for individual applications, allowing flexible memory allocation. For memory blocks outside PCs, 4 external memory blocks are aggregated into a large macro with a 128-bit access width supporting high-throughput computation in the computation engine. In multi-core mode, memory blocks are shared among the PCs as I/DTCM, enabling efficient parallel data processing across cores. Additionally, memory blocks are fully shared between the computation engine and multi-core modes to minimize data movement. In multi-core mode, the SoC preprocesses raw data and stores it in I/DTCM blocks. When engine mode is activated, these I/DTCM blocks can be reconfigured as A/WMEM, enabling direct use of preprocessed data with zero data movement. Remote sensing image processing for in-orbit applications requires flexibility to handle diverse needs and ensure data quality through key preprocessing steps (top-right). These include radiometric calibration to correct sensor imperfections, atmospheric correction to offset degradations, resizing to a 3x1024x1024 standard, and normalization for analysis. Multi-core processing with vector computing are used for efficient parallel handling of tasks. Object detection follows in a computation engine framework, streamlining data processing for remote sensing models like SuperYOLO [8]. SuperYOLO integrates RGB and IR inputs, with IR imagery providing critical data in adverse conditions to boost detection accuracy. The system detects target objects in the imagery and transmits the target area to Earth if detected. In experiments using the VEDAI dataset, our design's IED strategy reduces unnecessary (non-target area) data transmission by 98.63%. Also, to enhance hardware efficiency and reliability, we quantize SuperYOLO based on INT8, achieving 72.21% mAP50 accuracy with the boundary-aware activation function (BReLU [9]). BReLU enhances reliability by capping values, so deviations near the boundary don't affect the output. Results show accuracy improves from 0 to 63.78% as fault rates increase from 0 to 3% (bottom-right).

On the other hand, in-orbit computing can be affected by radiation effects, such as SEUs, making it crucial to evaluate the reliability of the SoC through neutron (up to 392 MeV) and alpha particle radiation experiments (Fig. 5), where high-energy neutrons have similar impact as protons in space [10]. To better assess the reliability of each component, we split the superYOLO application into two parts: convolution, pooling, and activation functions run in the computation engine mode, while the remaining computations run in the multi-core mode. Also, the soft error rate of memory blocks are tested as a reference. The memory blocks exhibit cross-sections of 4.07268E-15 cm^2/bit under neutron exposure and 1.70871E-11 cm^2/bit under alpha particle exposure. Furthermore, to categorize errors by severity, four events are defined (bottom-middle). The experiments indicate that, with DCLS protection, 85.39% of errors in RISC-V cores are detected and corrected, nearly reducing the cDUE cross-section by an order of magnitude, significantly enhancing system reliability. Besides, some errors go undetected because certain data paths, like the vector computing path designed for high-throughput processing, operate independently and bypass the DCLS checkers. Also, the errors detected in the Fetch stage account for the majority, representing 68.4% of the total errors in the cores. Periodic error scans to determine optimal reboot intervals can improve system reliability based on memory and event cross-sections from radiation experiments.

Fig. 6 summarizes the proposed SoC and compares it with prior arts. This work improves the latency by 17.77X in vector computing under multi-core mode, with a power consumption of 3.73-118.05 mW at 123-1225 MHz frequency. The peak power efficiency is 17.18 TOPS/W at 0.6V and 340MHz in computation engine mode. During neural network (NN) evaluations for SuperYOLO, our design achieves 14.72 TOPS/W. Meanwhile, in real remote sensing applications, our design improves the power efficiency by 2.3X to 2.5X at similar process nodes, attributed to the meticulously organized inter-PC data flow that seamlessly aligns with our hardware architecture to support up to 16 times of data reuse. Besides, in multi-core mode, our design improves the energy efficiency by 1.1-343X thanks to vector computing. Fig. 7 shows the 22nm CMOS die micrograph.

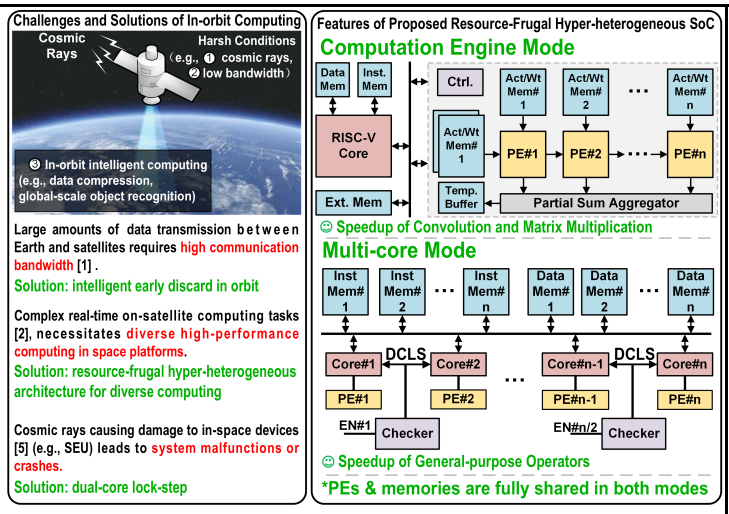


Fig. 1. Challenges of in-orbit computing (left), and the features of proposed SoC (right).

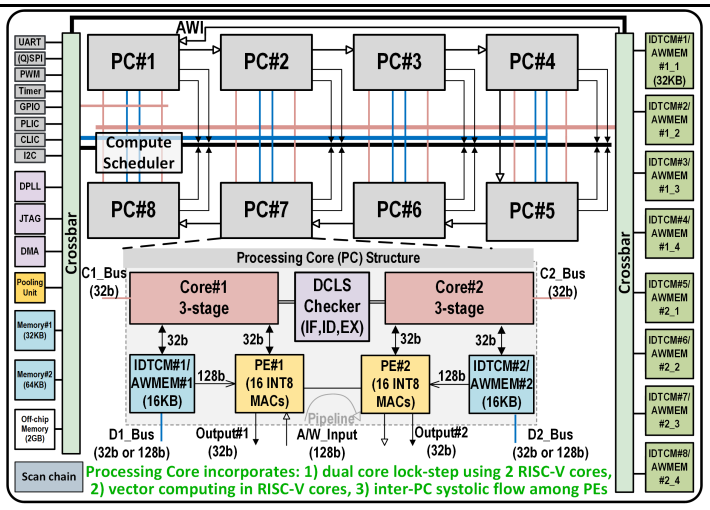


Fig. 2. System-on-Chip architecture and details of processing core.

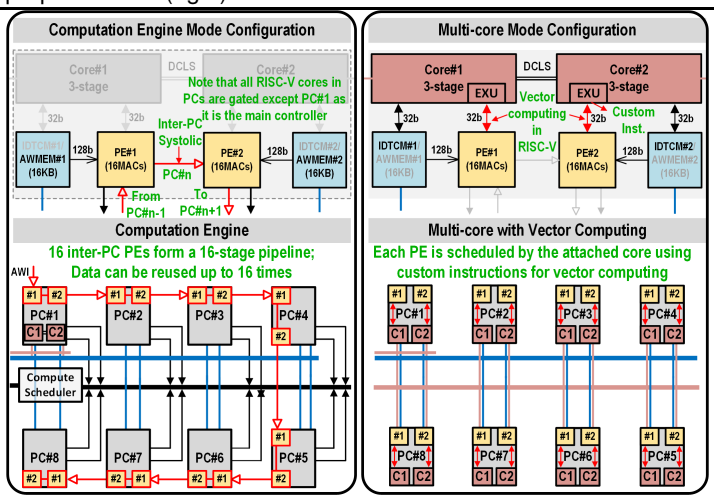


Fig. 3. Reconfiguration of computation engine mode (left) and multi-core mode (right).

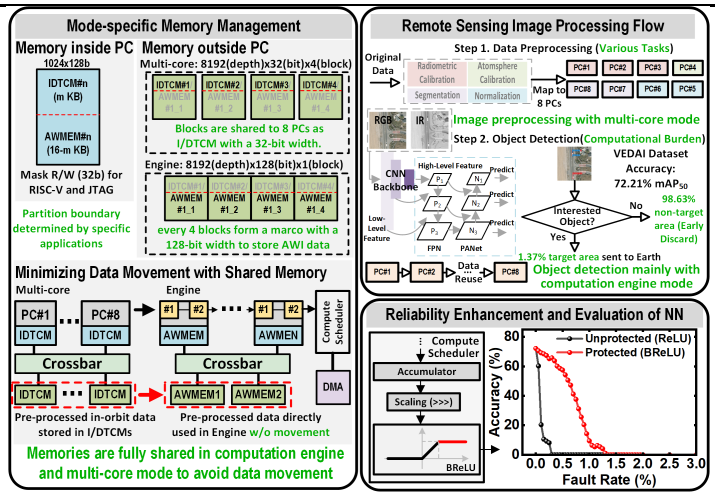


Fig. 4. Mode-specific memory management (left), computation flow in remoting sensing (top-right), and reliability enhancement and evaluation of neural network using BReLU (bottom-right).

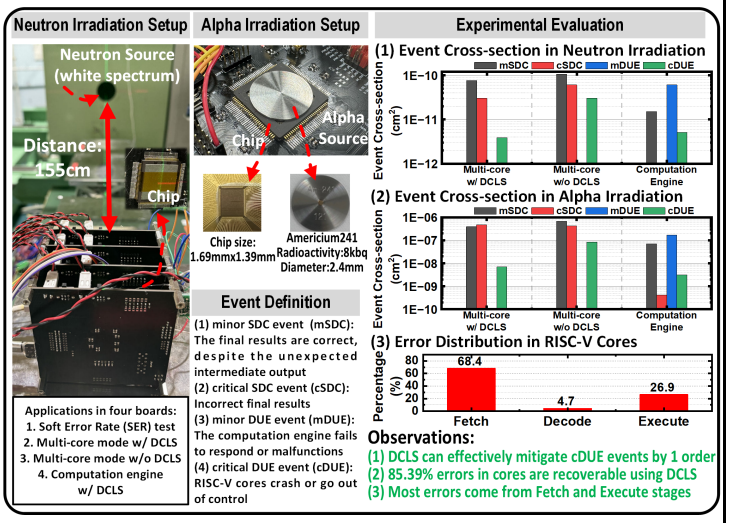
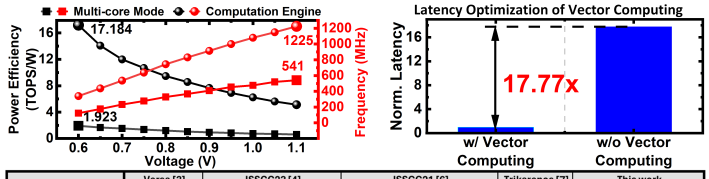
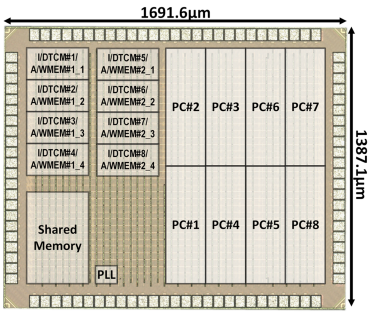


Fig. 5. Neutron and Alpha irradiation experiments and evaluation.



	Versa [3]	ISSCC23 [4]	ISSCC21 [6]	Trikarenos [7]	This work
Process	28nm	22nm FDX	12nm	28nm	22nm
Architecture	36x cortex-M4F	1x RV32IMCFXpulp+ 16x RV32IMCFXpupnn+ RBE Accel.	1x Cortex-R52 for control + 8x Cortex-A76 for application + 1x 3D GPU for application + 22 cores for image recognition	3x Ibex RISC-V	**(1) 16x RV32IM cores w/ Vector Computing or (2) 2x RV32IM Cores + 1x Computation Engine**
Radiation-hardening	-	-	DCLS	TCLS	DCLS & BReLU
Supply [V]	0.55-1.0	0.5-0.8	0.8 (core) 0.85 (CPU, GPU)	0.9	0.6-1.1V
Frequency	12-510MHz	420	1.8GHz Cortex-A76 1GHz Safety Processor	250-270MHz	123-541MHz (RISC-V) 340-1225MHz (Engine)
Area [mm²]	12	18.7	230.46	2	2.35
Power [mW]	7.9-811.2	12.8 - 123	4400	2.4-19.7	3.73-50.07 (Multi-core) 7.73-118.05 (Engine)
Memory	1.27MB	1152KB	~41.6MB	256KB	608KB
Precision	FP32	INT2/4/8/16/32(RISC-V) INT2-8(RBE)	INT8	-	INT8/16/32(RISC-V) INT8 (PE)
Peak Peformace [TOPS]	0.0119	0.637@RBE2x2	60.4	666 DMIPS	0.6272 (Engine) 0.0306(Multi-core)
Peak Power Efficiency [TOPS/W]	0.0364	1.66@INT in RISC-V 12.4@RBE2x2 6.38@RBE Mixed ResNet20 5.83@RBE4x4 ResNet18	13.8 6.1 @ VGG16	0.0056	**17.184 (Engine) 1.923 (Multi-core) 14.72 @ SuperYOLO**

Fig. 6. Measurement results and comparison with state-of-the-art designs.



	Specification
Process	22nm CMOS
Capability	1. Multi-core w/ Vector Computing 2. Computation Engine 3. Memory and PE Reuse 4. Dual-core lock-step for RISC-V 5. BReLU for NN
Supply Voltage [V]	0.6-1.1
Frequency [MHz]	123-541MHz @ RISC-V 340-1225MHz @ Engine
Chip Area [mm^2]	2.35
Core Area [mm^2]	1.60
Power [mW]	3.73-50.07 @ Multi-core 7.73-118.05 @ Engine
On-chip Memory [KB]	608
MACs	256 @ INT8
Precision	INT8/16/32 (RISC-V) INT8 (Engine)
Peak Performance [TOPS]	0.6272 @ Engine 0.0306 @ Multi-core
Area Efficiency [$TOPS/mm^2$]	0.392 @ Engine 0.0020 @ Multi-core
Peak Energy Efficiency @ 0.6V [TOPS/W]	17.184 @ Computation Engine* 1.923 @ Multi-core Mode* 14.72 @ SuperYOLO**

*50% sparsity, 100% PE utilization;
**21.36/50.45% sparsity for weight/activation, VEDAI dataset

Fig. 7. Die micrograph (left), and chip summary (right).

Acknowledgements:

This work was supported in part by the Grant-in-Aid for Scientific Research (S) from Japan Society for the Promotion of Science (JSPS) under Grant 24H00073, by JST CREST, Japan, under Grant JPMJCR19K5; the National Natural Science Foundation of China under Grant 62274081; Grant 2023QN10X177.

References:

[1] G. Furano et al., "Towards the Use of Artificial Intelligence on the Edge in Space Systems: Challenges and Opportunities," in IEEE Aerospace and Electronic Systems Magazine, vol. 35, no. 12, pp. 44-56, 1 Dec. 2020, doi: 10.1109/MAES.2020.3008468.

[2] B. Denby and B. Lucia, "Orbital Edge Computing: Machine Inference in Space," in IEEE Computer Architecture Letters, vol. 18, no. 1, pp. 59-62, 1 Jan.-June 2019, doi: 10.1109/LCA.2019.2907539.

[3] S. Kim et al., "Versa: A 36-Core Systolic Multiprocessor With Dynamically Reconfigurable Interconnect and Memory," in IEEE Journal of Solid-State Circuits, vol. 57, no. 4, pp. 986-998, April 2022, doi: 10.1109/JSSC.2022.3140241.

[4] F. Conti et al., "22.1 A 12.4TOPS/W @ 136GOPS AI-IoT System-on-Chip with 16 RISC-V, 2-to-8b Precision-Scalable DNN Acceleration and 30%-Boost Adaptive Body Biasing," 2023 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023, pp. 21-23, doi: 10.1109/ISSCC42615.2023.10067643.

[5] R. Harboe-Sorensen et al., "Observation and analysis of Single Event Effects on-board the SOHO satellite," RADECS 2001. 2001 6th European Conference on Radiation and Its Effects on Components and Systems (Cat. No.01TH8605), Grenoble, France, 2001, pp. 37-43, doi: 10.1109/RADECS.2001.1159256.

[6] K. Matsubara et al., "4.2 A 12nm Autonomous-Driving Processor with 60.4TOPS, 13.8TOPS/W CNN Executed by Task-Separated ASIL D Control," 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2021, pp. 56-58, doi: 10.1109/ISSCC42613.2021.9365745.

[7] M. Rogenmoser and L. Benini, "Trikarenos: A Fault-Tolerant RISC-V-based Microcontroller for CubeSats in 28nm," 2023 30th IEEE International Conference on Electronics, Circuits and Systems (ICECS), Istanbul, Turkiye, 2023, pp. 1-4, doi: 10.1109/ICECS58634.2023.10382727.

[8] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li and Q. Du, "SuperYOLO: Super Resolution Assisted Object Detection in Multimodal Remote Sensing Imagery," in IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-15, 2023, Art no. 5605415, doi: 10.1109/TGRS.2023.3258666.

[9] J. Zhan, R. Sun, W. Jiang, Y. Jiang, X. Yin and C. Zhuo, "Improving Fault Tolerance for Reliable DNN Using Boundary-Aware Activation," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 41, no. 10, pp. 3414-3425, Oct. 2022, doi: 10.1109/TCAD.2021.3129114.

[10] J. Baggio, V. Ferlet-Cavrois, H. Duarte and O. Flament, "Analysis of proton/neutron SEU sensitivity of commercial SRAMs-application to the terrestrial environment test method," in IEEE Transactions on Nuclear Science, vol. 51, no. 6, pp. 3420-3426, Dec. 2004, doi: 10.1109/TNS.2004.839135.