Wenbo ZHANG, Bo DING, Shuai WEI, Qinrang LIU, Hong YU, Ke SONG, Wei GUO, Bo MEI, Rui ZHENG, 2026. WSC optimizer: an optimization tool for wafer-scale chip architecture exploration. *ENGINEERING Information Technology & Electronic Engineering*, 27(4):250008. https://doi.org/10.1631/ENG.ITEE.2025.0008

WSC optimizer: an optimization tool for wafer-scale chip architecture exploration

Key words: Wafer-scale chip; Hardware–software co-design; Chip layout; Design space exploration

Corresponding author: Shuai WEI
E-mail: weis0906@163.com
ORCID: https://orcid.org/0000-0002-4902-6645

Motivation

- Exponential growth in computing demand from large-scale language models (LLMs) outpaces traditional monolithic chip scaling, which hits 1-nm physical limits, causing low yield and high cost. Wafer-scale chip (WSC) with chiplet and advanced packaging is a promising solution but suffers from homogeneous architectures, inefficient resource use, and heavy reliance on manual design. Existing chiplet methods ignore core reuse and pipeline operations; automated design tools lack unified WSC modeling, failing to address multi-physical constraints, heterogeneous task adaptability, and efficient design space exploration, demanding a comprehensive WSC architecture exploration toolchain.

Main idea

- We propose a chiplet combination method and a hierarchical WSC construction method that integrates physical and design constraints. Through the design workflow of ***chiplet selection*** → ***component integration*** → ***wafer layout***, we standardize the logic for functional and interconnection adaptation, thereby providing a foundational framework that balances feasibility and flexibility for system-on-wafer (SoW) implementation.

Main idea

- We present a reinforcement learning-based evaluation and iterative optimization method. First, we develop mathematical models for chiplets and different applications, and compile physical parameters of chiplets and operator-specific attributes of applications to establish a standardized operator library. Second, to address the challenge of differentiated optimal scheduling across diverse layout results, we realize collaborative iteration of layout and scheduling within the reinforcement learning framework. This approach dynamically adapts to the matching requirements of chiplet and applications, efficiently explores optimal architecture configurations, and overcomes the trade-off between layout and scheduling exploration space.

Main idea

- We propose a collaborative hardware–software co-design methodology for WSC. We comprehensively account for the coupled relationships between bandwidth, latency, storage capacity, throughput, and task execution time. Through the coordinated design of operator clustering, task mapping, and pipeline scheduling, we achieve dynamic balance across multiple performance metrics. Meanwhile, we clarify the rationale for prioritizing static random access memory (SRAM) in WSC architectures: SRAM's low access latency and high reliability fulfill the system's demand for fast data response, and its tight integration with computing chiplets minimizes interconnection overhead.

Method

The integration of independent switch components stands as a core design decision: by decoupling routing/switching functions from compute chiplets, it not only avoids the rigidity limitations and high redesign costs of domain-specific integrated architectures, but also supports the direct reuse of general-purpose chiplets.

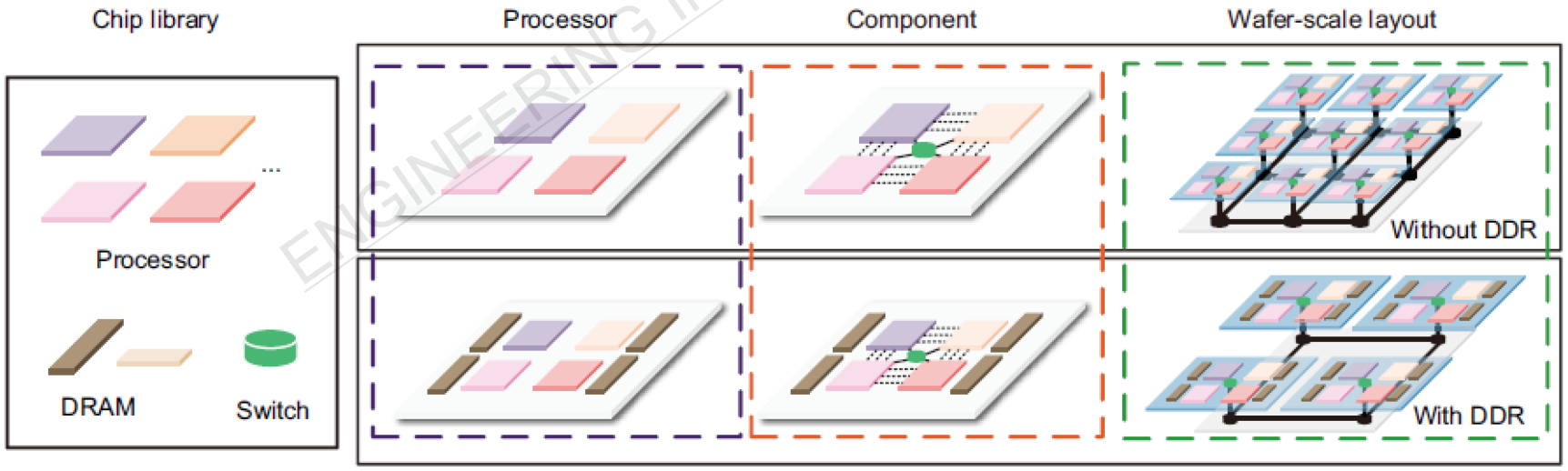

Fig. 2 An example of a hierarchical WSC design method

Method

This study aims to develop an efficient automated design framework for WSC architectures, enabling specific applications to achieve high performance. The overall framework is shown in Fig. 5.

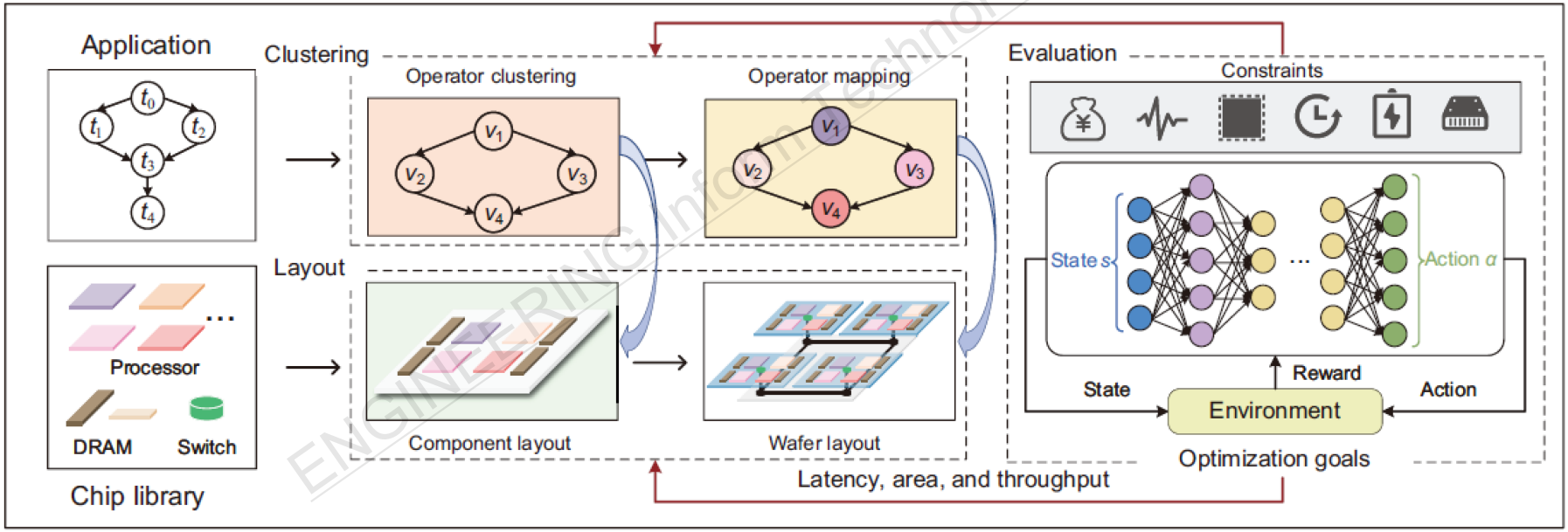

Fig. 5 A comprehensive optimization framework for operator clustering and task mapping in WSC systems

Method

The proposed reinforcement learning framework combines a graph convolutional network (GCN) with neural networks to optimize chip layout and scheduling.

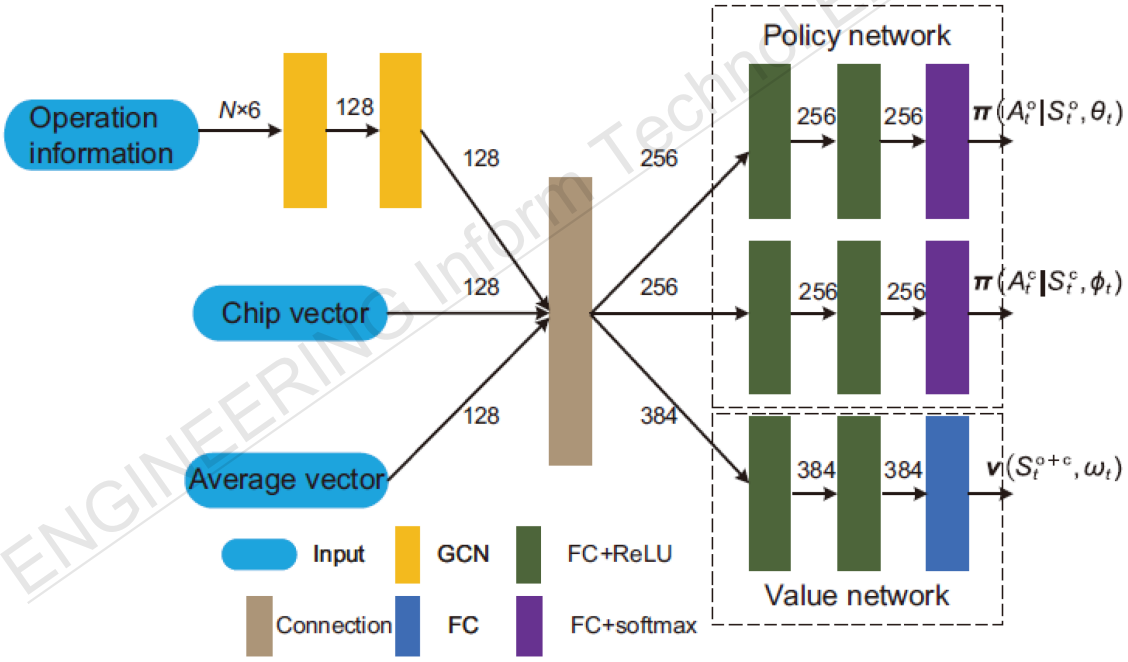

Fig. 7 Schematic of the reinforcement learning framework (FC: fully connected layer)

Method

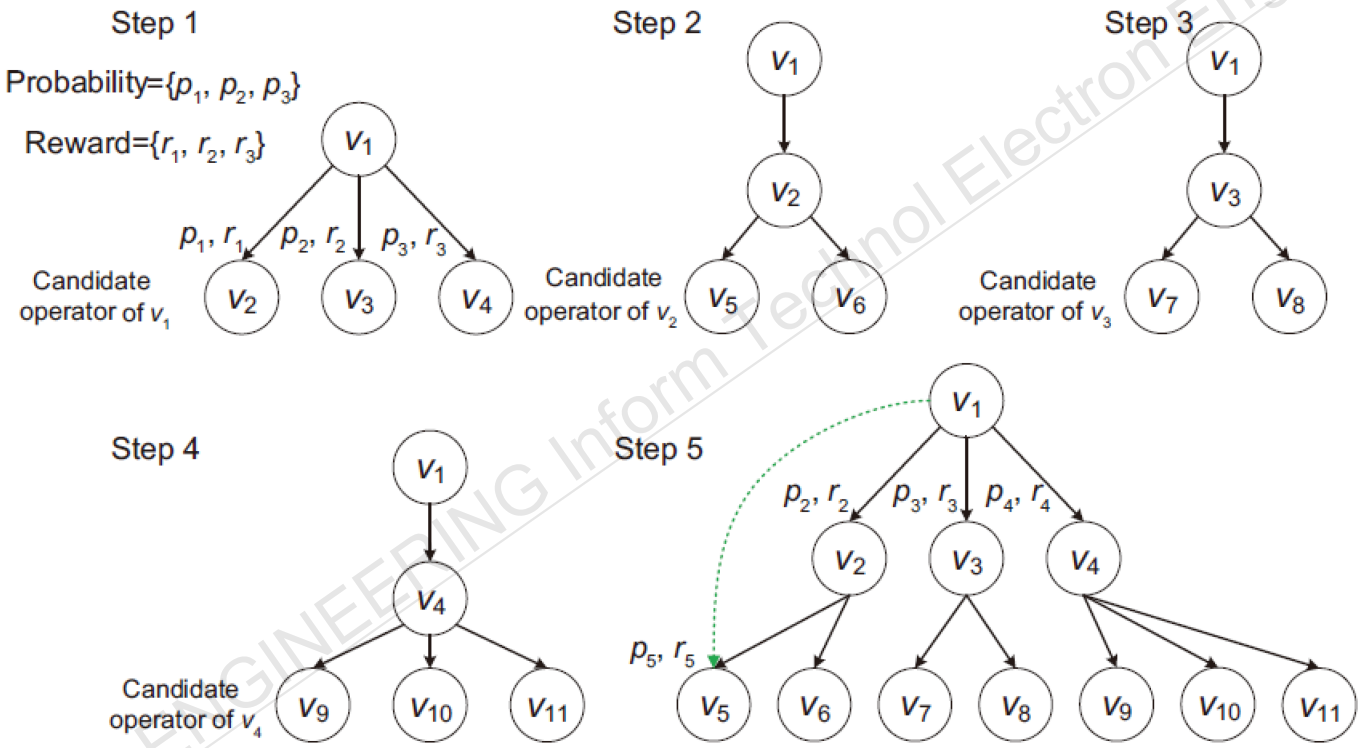

Fig. 8 An example of the operator search tree with probability and reward assignments

Results

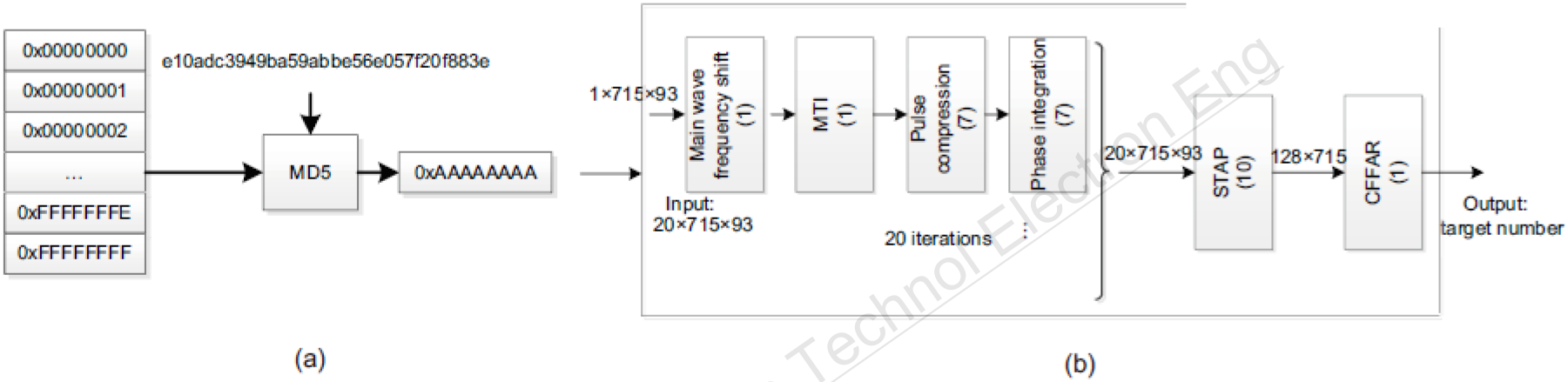

Fig. 9 An example of a decryption and encryption task (a) and a signal processing task (b)

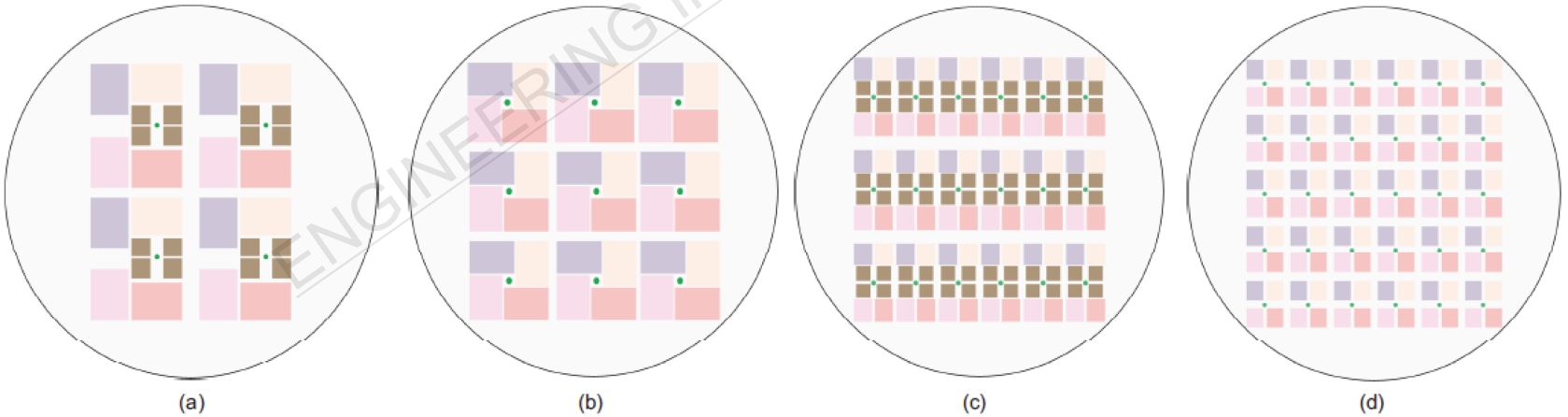

Fig. 10 Comparison of wafer layout for the MD5 task (a and b) and the signal processing task (c and d): (a) attached storage (MD5); (b) optimized layout (MD5); (c) attached storage (signal processing); (d) optimized layout (signal processing)

Results

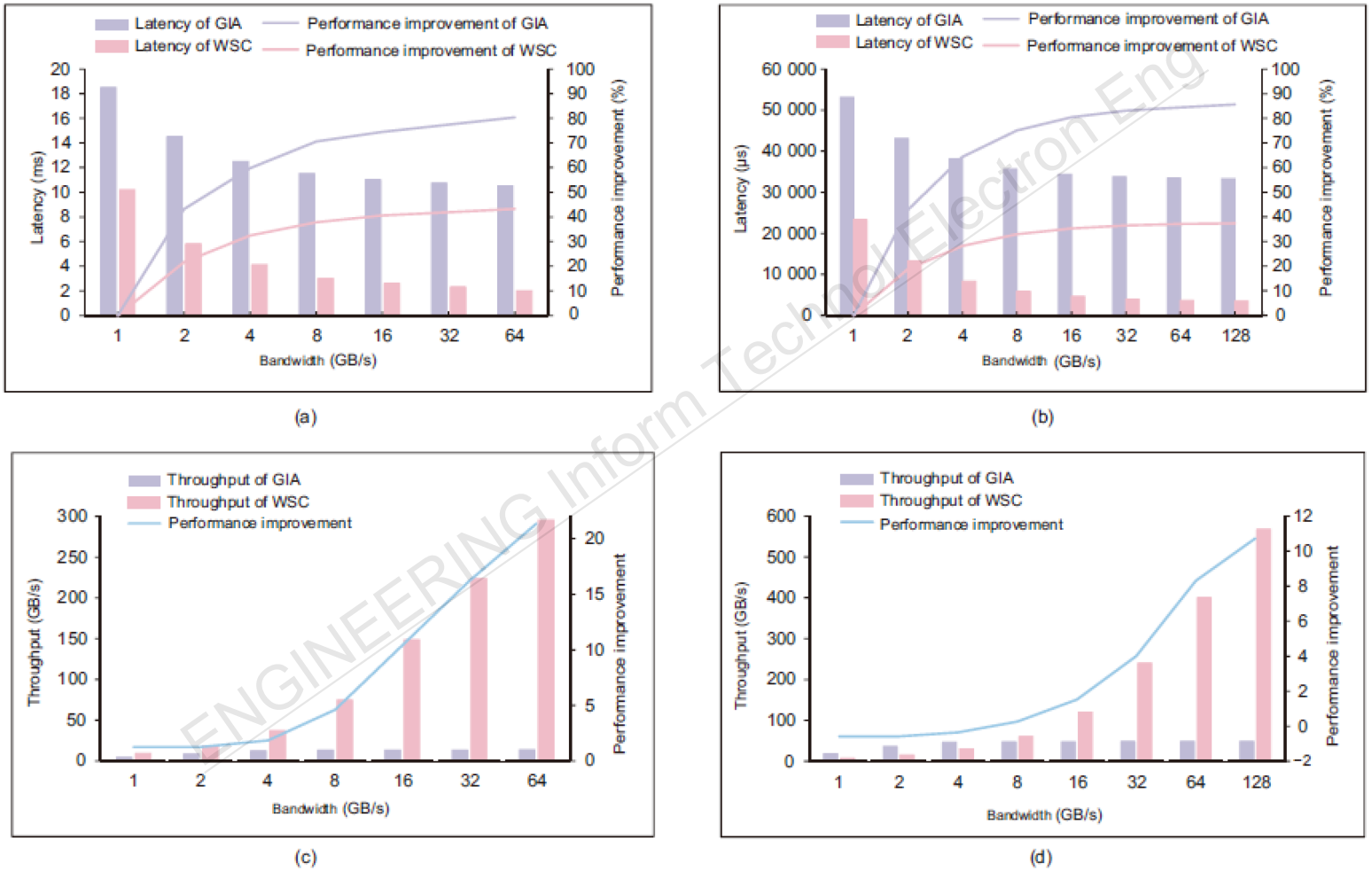

Fig. 11 Latency and throughput comparison of different design methods across MD5 procedure and signal processing procedure: (a) latency (MD5 procedure); (b) latency (signal processing procedure); (c) throughput (MD5 procedure); (d) throughput (signal processing procedure)

Results

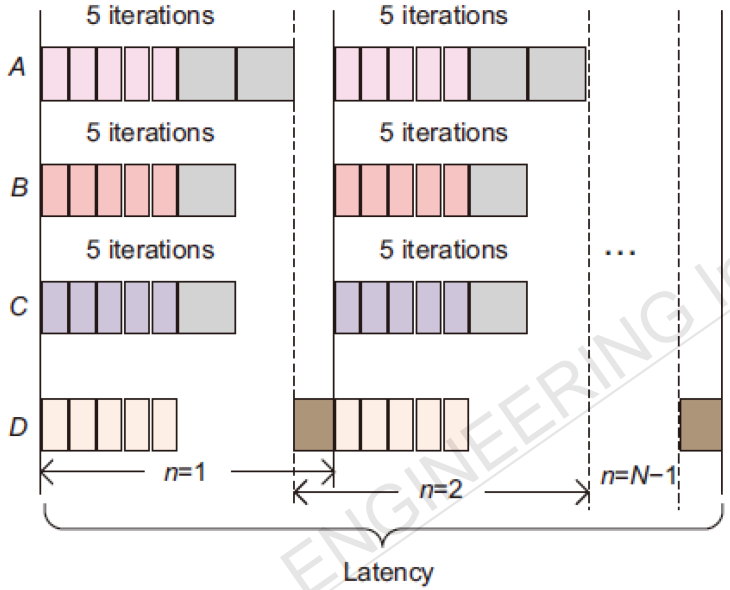

Fig. 12 Pipelined execution of the signal processing procedure

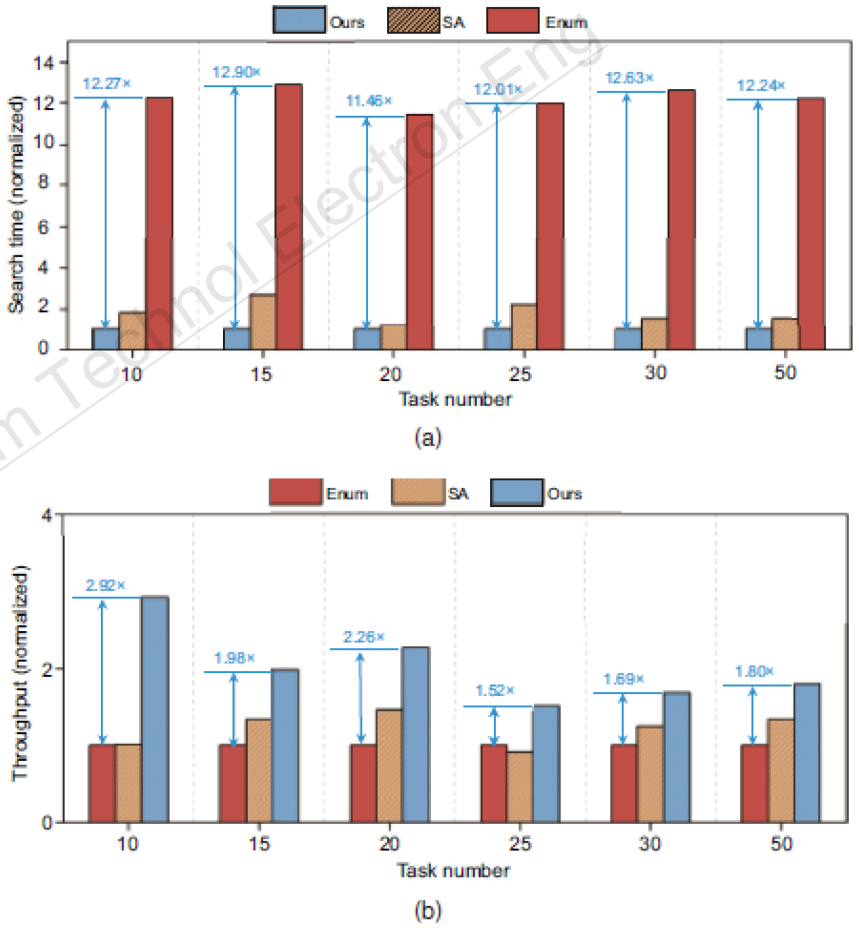

Fig. 13 Overall performance comparison of different optimization methods in multi-task scenarios: (a) search time comparison results in multi-task scenarios; (b) throughput comparison results in multi-task scenarios

Conclusions

The development of WSC design holds significant promise for the future. Our proposed hardware–software co-design methodology offers a comprehensive solution to address the complexities and high-performance requirements in modern microelectronics. We have systematically explored the WSC development process, focusing on optimizing the architecture space for hardware/software co-design. Despite challenges such as limited tools and core libraries in WSC design, we have made great progress by building a comprehensive core library and refining the design flow, which has greatly boosted design efficiency.

Wenbo ZHANG received the B.S. degree in information and communication engineering from Information Engineering University in 2018. He is currently pursuing the Ph.D. degree in Information Engineering University. His current research interests include software-defined system on wafer and cyberspace security.

Shuai WEI received the B.S. degree in computer science and technology, the M.S. degree in software engineering, and the Ph.D. degree in high performance computing and parallel compiling in 2005, 2008, and 2012, respectively. He is currently an associate professor with Information Engineering University. He has authored over 40 applied patents and published articles. His research interests include high-performance computing, cyber security, computer software network architecture, and machine learning.