南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

An Energy-Efficient Bit-Split-and-Combination Systolic Accelerator for NAS-Based Multi-Precision Convolution Neural Networks

presented by Liuyao Dai

Authors:

Liuyao Dai, Quan Cheng, Yuhang Wang, Gengbin huang, Junzhuo Zhou, Kai Li, Wei Mao, and Hao Yu

Southern University of Science and Technology

Shenzhen, China

Outline

- Introduction
- Bit-Split-and-Combination MAC
- Multi-precision systolic accelerator
- Systolic dataflow
- Experiment
- Conclusion

Outline

- **Introduction**
- **Bit-Split-and-Combination MAC**
- **Multi-precision systolic accelerator**
- **Systolic dataflow**
- **Experiment**
- **Conclusion**

Introduction

Deep learning for edge computing：

convolutional neural network (CNN) models are becoming more complex with larger parameters

Neural Architecture Search (NAS) can search for optimized multi-precision neural network models：

- negligible loss of accuracy
- energy efficiency

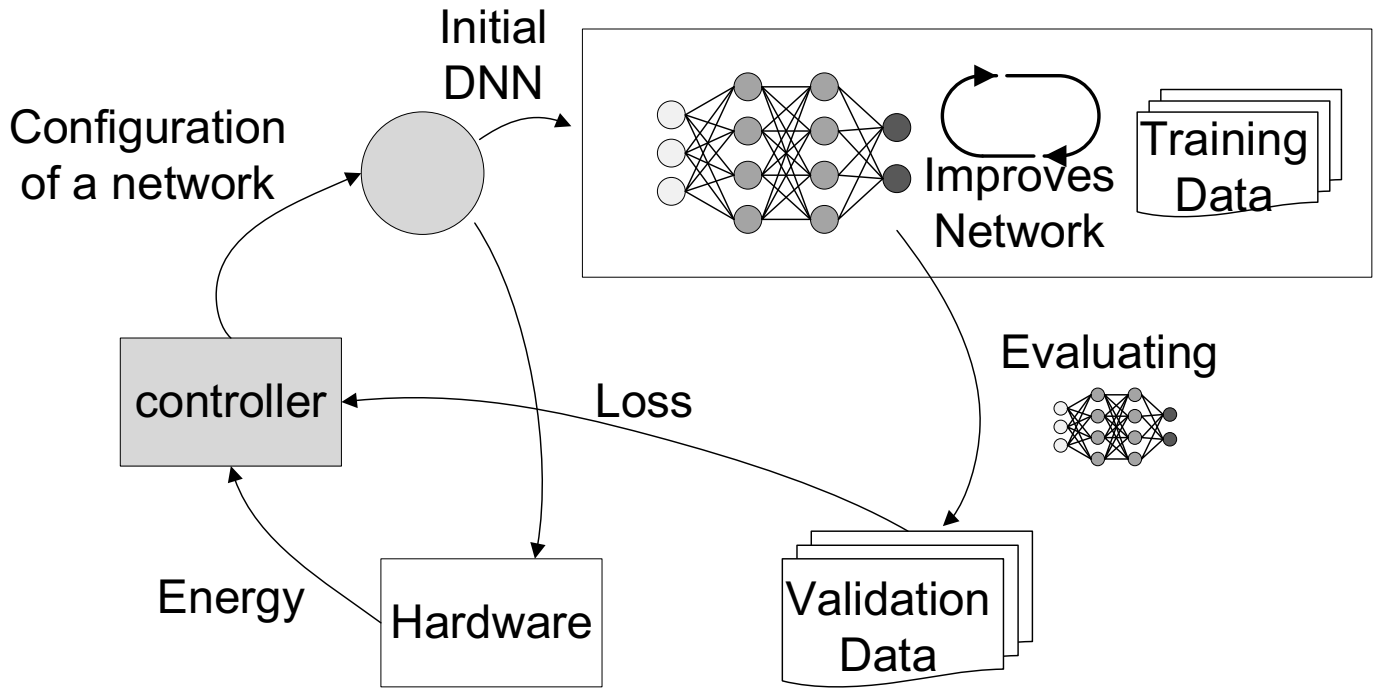


Fig. 1 Illustration of energy aware neural architecture search framework

Introduction

Existing multi-precision multiply-accumulate (MAC) disadvantages:

- Bottom-up low-precision-combination (LPC)
  large hardware cost
  huge power consumption
- Top-down high-precision-split (HPS)
  poor throughput performance

Proposed Work:

- Bit-split-and-combination (BSC) method
  tradeoff cost and throughput
- Multi-precision systolic dataflow
  data reuse and energy efficient

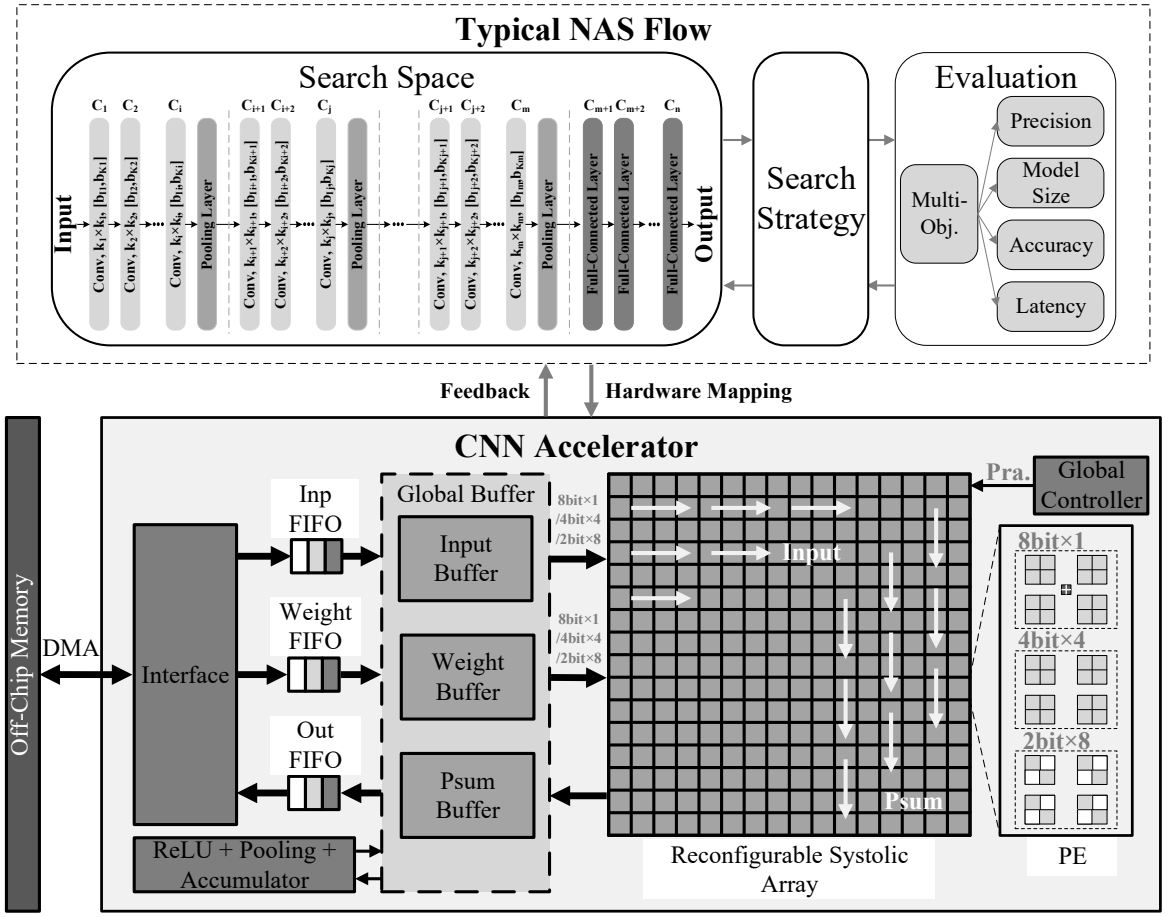


Fig. 2 Typical NAS flow with proposed multi-precision systolic architecture

Outline

- Introduction
- **Bit-Split-and-Combination MAC**
- Multi-precision systolic accelerator
- Systolic dataflow
- Experiment
- Conclusion

Bit-Split-and-Combination MAC

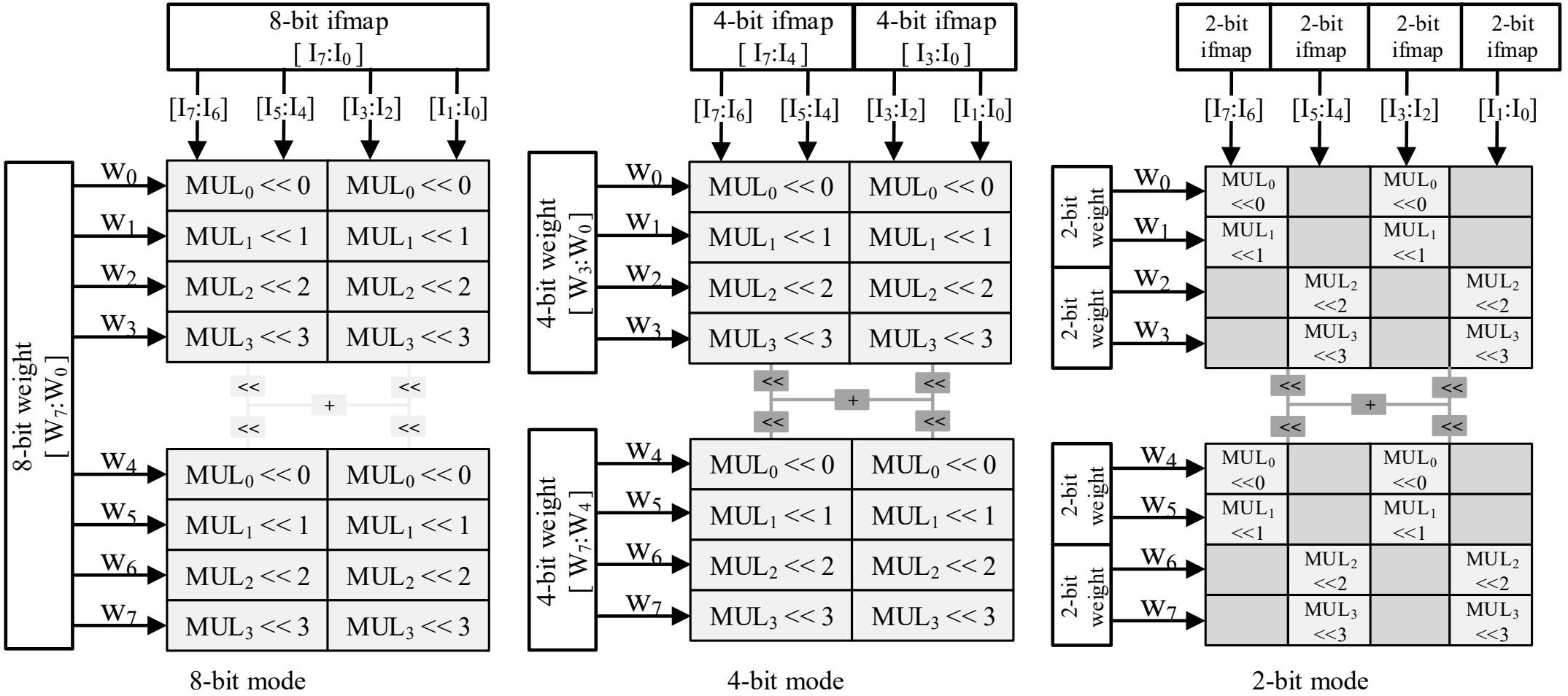


Fig. 3 BSC MACs for multi-precision operations including 8-bit, 4-bit and 2-bit modes

Outline

- Introduction
- Bit-Split-and-Combination MAC
- **Multi-precision systolic accelerator**
- Systolic dataflow
- Experiment
- Conclusion

Systolic PE Array

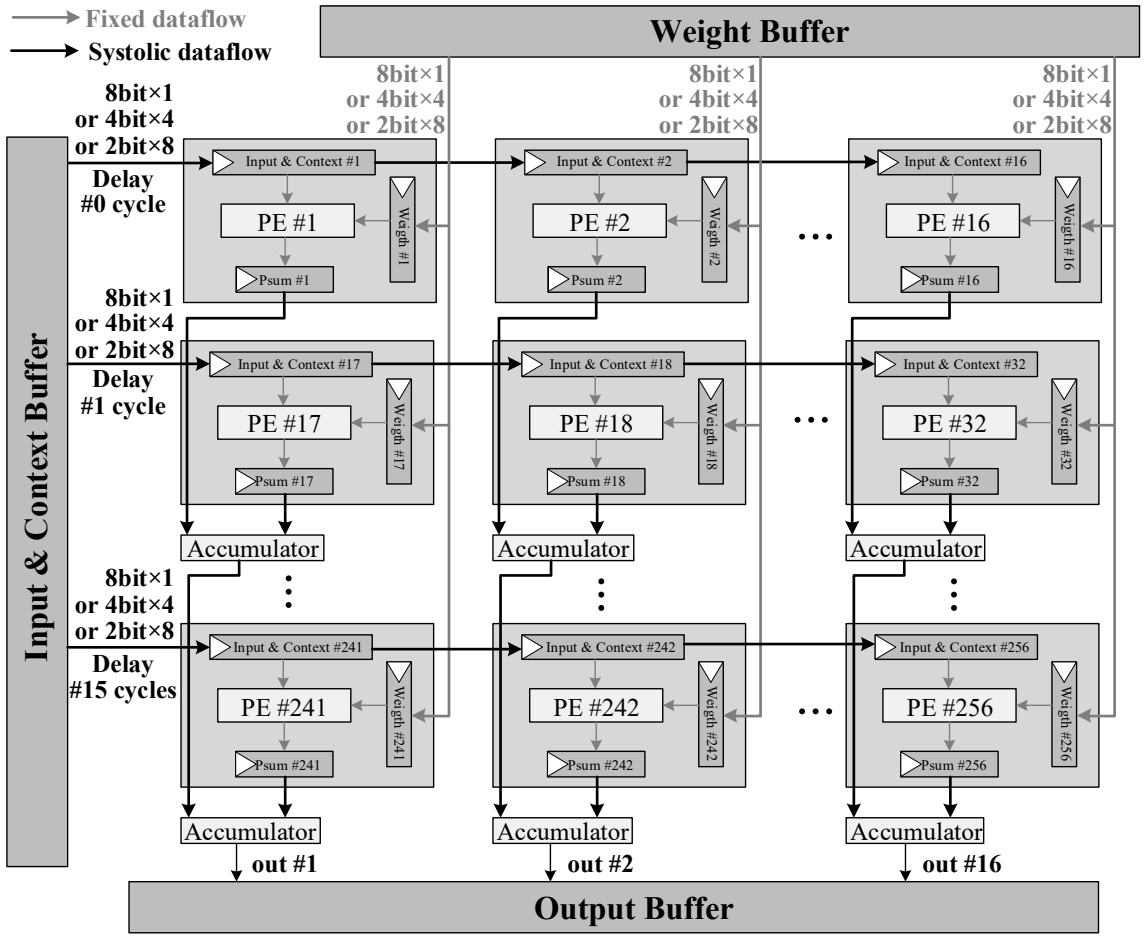


Fig. 4 Multi-precision systolic dataflow of PE array

- Weights are stored in PE array fetched from weight buffer.
- Input activations flow from PE_1 to PE_{256} sequentially.
- Outputs from PE array are transmitted to psum buffer.

Outline

- Introduction
- Bit-Split-and-Combination MAC
- Multi-precision systolic accelerator
- **Systolic dataflow**
- Experiment
- Conclusion

Data reuse in the systolic dataflow

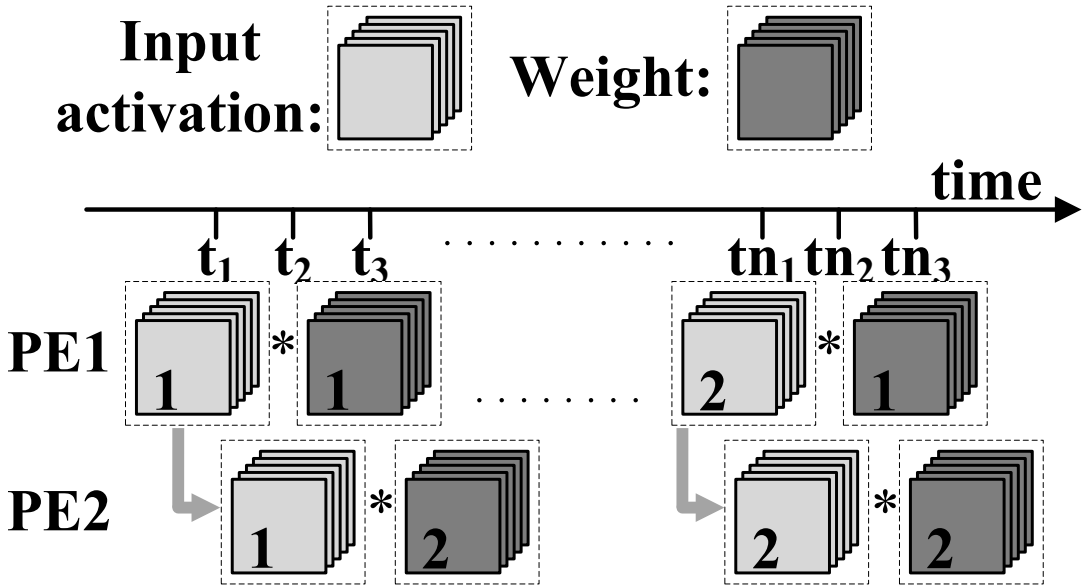


Fig. 5 Data reuse in the processing of the systolic dataflow

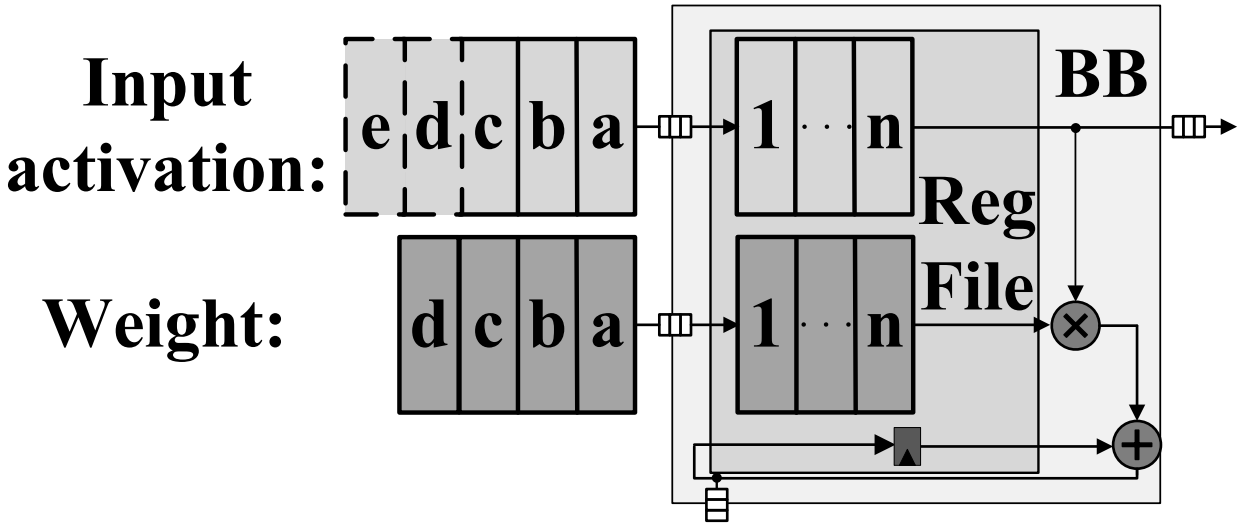


Fig. 6 Data reuse in BSC BB

- input activations from different input channels are input to the same PE in different cycles
- different filters are stored in different PEs

- Input activation is reused K times in BSC BB
- Weight is reused E^2 times in BSC BB

Outline

- Introduction
- Bit-Split-and-Combination MAC
- Multi-precision systolic accelerator
- Systolic dataflow
- **Experiment**
- Conclusion

MAC Performance Comparison

Precision	HPS(TOPS/W)	LPC(TOPS/W)	BSC(TOPS/W)	BSC/HPS	BSC/LPC
2 bit	12.6	32.7	27.36	2.4	0.82
4 bit	6.29	8.16	13.68	2.4	1.64
8 bit	3.14	2.04	3.42	1.2	1.64

Table 1 Energy efficient comparison of multi-precision mac units

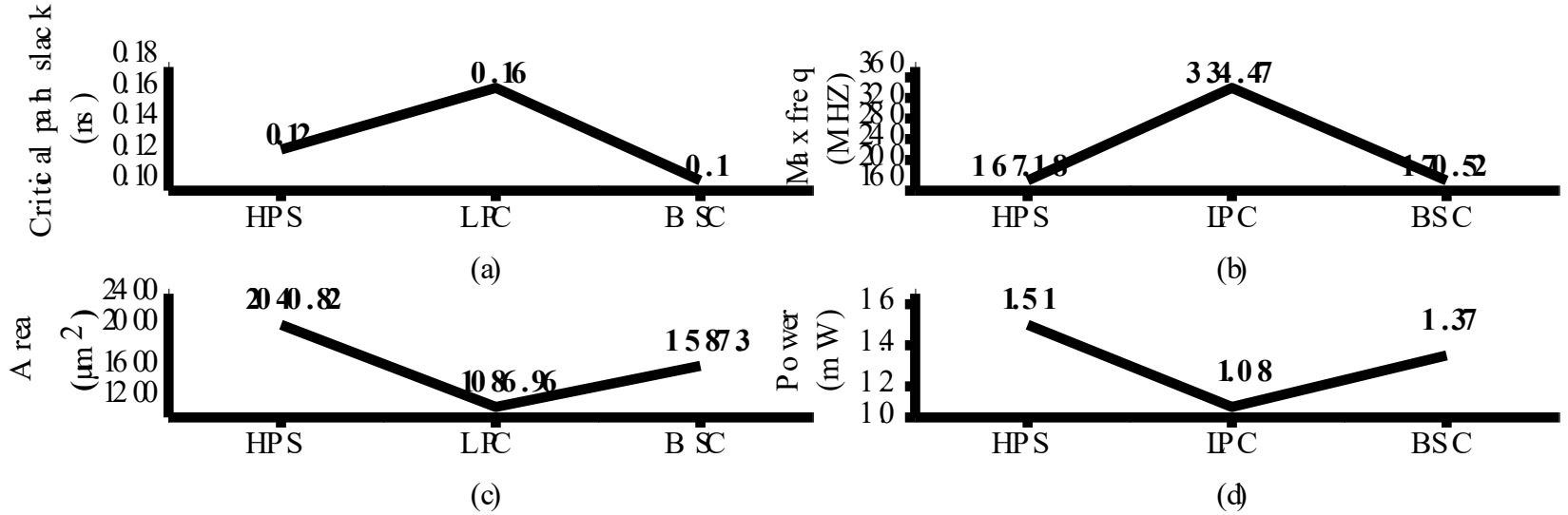


Fig. 7 Performance metrics comparison between traditional HPS, LPC and the BSC MAC units

- the proposed BSC MAC unit in this work has the characteristics of low power consumption, low hardware cost and fast calculation speed.

System Performance Comparison

CNN	Dataset	Model Weights	2bit/4bit/8bit proportion
VGG-16	CIFAR-10	138.0 MBytes	0%/89.8%/10.2%
ResNet-18	ImageNet	13.0 MBytes	0%/94.5%/5.5%
LeNet-5	MNIST	0.5 MBytes	45.0%/55.0%/0%

Table 2 Evaluated CNN benchmarks

	Gemmini	Bit-serial	Bit-fusion	BSC
Technology	FinFET 16 nm	28 nm	28 nm	28 nm
Cores	256 PEs	4096 SIPs	512 fusion units	256 PEs
On-chip (Memory)	64 KB	2 MB eDRAM 16 KB SRAM	181.5 KB	180 KB
Chip area (mm^2)	0.467	1.40	/	1.43
Frequency	500 MHZ	980 MHZ	500 MHZ	500 MHZ

Table 3 Evaluated accelerators

System Performance Comparison

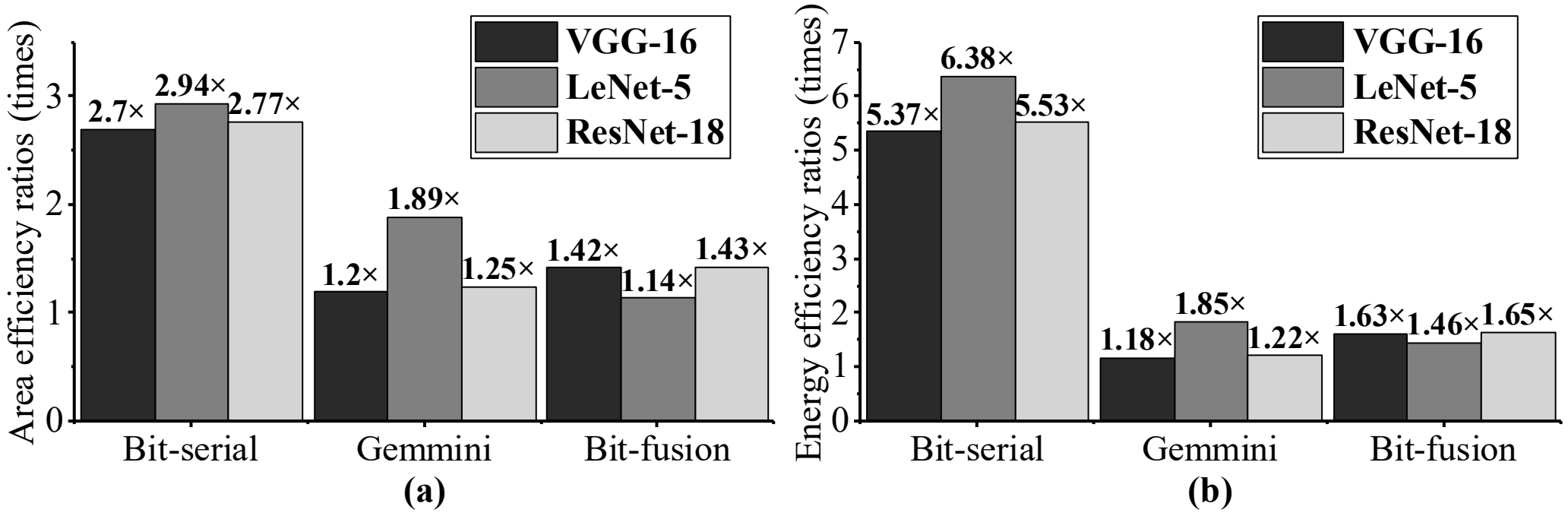
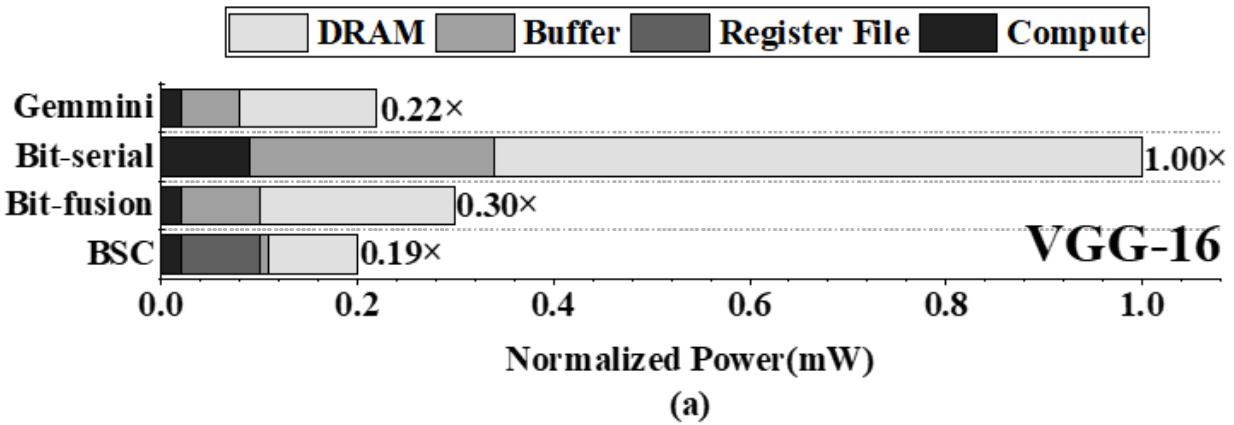
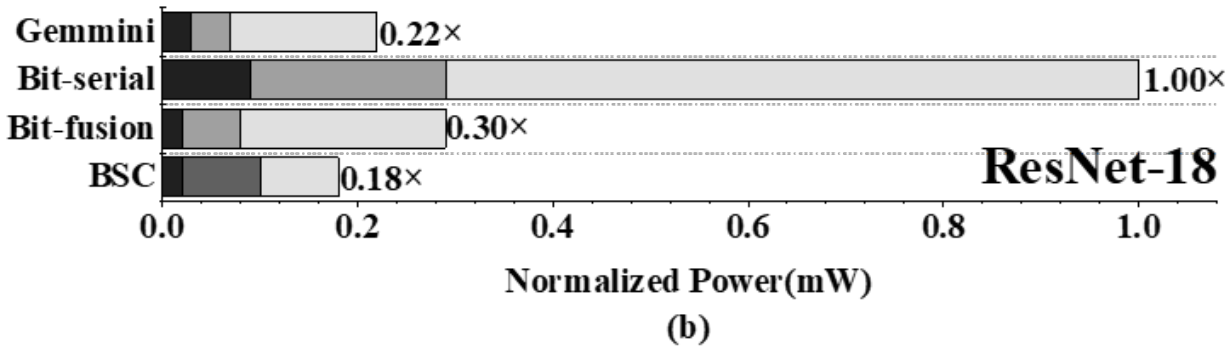


Fig. 8 Improvement ratios of the BSC systolic accelerator to Gemmini, Bit-fusion and Bit-serial on multi-precision CNN benchmarks: (a) Area efficiency, (b) Energy efficiency.

- Area efficiency performance: Compared with Bit-fusion, at most 1.43× ratios are achieved owing to the heavy additional logics of Bit-fusion.
- Energy efficiency performance： Compared with Gemmini and Bit-serial, at most 1.85× and 6.38× ratios are achieved owing to the proposed work supporting both input activations and weights for multi-precision operations.

System Performance Comparison





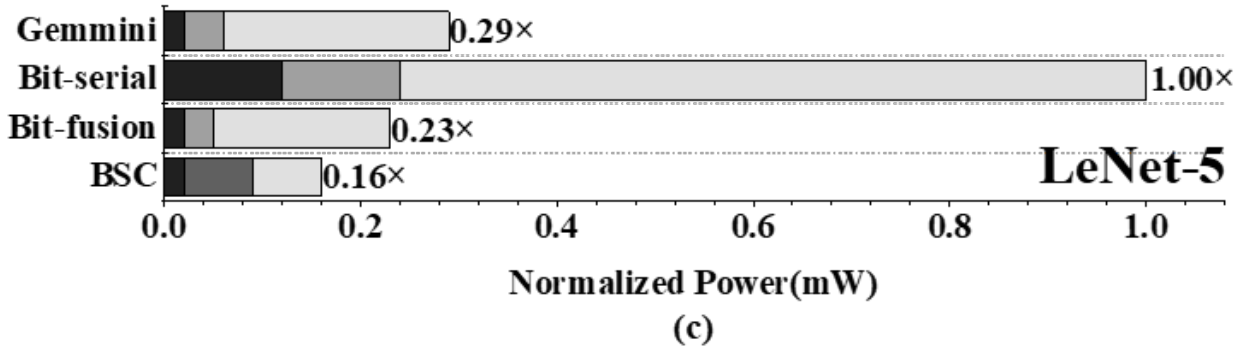


Fig. 9 Power breakdown of Gemmini, Bit-serial, Bit-fusion and the proposed BSC accelerator under different benchmarks:(a)VGG-16, (b)ResNet-18, (c)LeNet-5.

- the proposed accelerator provides flexibility in both inputs and weights bit-width, leading to the highest reduction ratio by 6.38×, 1.85× and 1.65× compared with Bit-serial, Gemmini and Bit-fusion.

Outline

- Introduction
- Bit-Split-and-Combination MAC
- Multi-precision systolic accelerator
- Systolic dataflow
- Experiment
- **Conclusion**

Conclusion

In this paper, an energy-efficient multi-precision systolic accelerator is designed

- The reconfigurable architecture supports NAS-based CNNs with 2-8 various bit-widths.
- Compared with the state-of-the-art accelerators Gemmini, Bit-serial and Bit-fusion on the multi-precision CNN benchmarks, the proposed BSC accelerator achieves at least 1.18×, 5.37× and 1.46× energy efficiency.
- For area efficiency, the improvement ratio of 1.2×, 2.7× and 1.14× are achieved at least.
- The results show the proposed work is of great potential for multi-precision edge-computing

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Thanks for your attention!