DATE 22

DESIGN, AUTOMATION & TEST IN EUROPE

14 – 15 March 2022 · on-site event
16 – 23 March 2022 · online event

The European Event for Electronic
System Design & Test

A Precision-Scalable Energy-Efficient Bit-Split-and-Combination Vector Systolic Accelerator for NAS-Optimized DNNs on Edge

Authors: Kai Li , **Junzhuo Zhou** , Yuhang Wang , Junyi Luo , Zhengke Yang , Shuxin Yang ,
Wei Mao , Mingqiang Huang and Hao Yu*

Southern University of Science and Technology, China

SUSTech Southern University of Science and Technology

深港微电子学院 School of Microelectronics

Outline

- Introduction
  - NAS-Optimized Multi-Precision DNNs
  - Multi-Precision Neural Network Accelerator
- Bit-Split-and-Combination (BSC) multi-precision vector MAC
- Precision-scalable vector systolic PE array
- Experiment Results
- Conclusion

Outline

- Introduction
  - NAS-Optimized Multi-Precision DNNs
  - Multi-Precision Neural Network Accelerator
- Bit-Split-and-Combination (BSC) multi-precision vector MAC
- Precision-scalable vector systolic PE array
- Experiment Results
- Conclusion

NAS-Optimized Multi-Precision DNNs

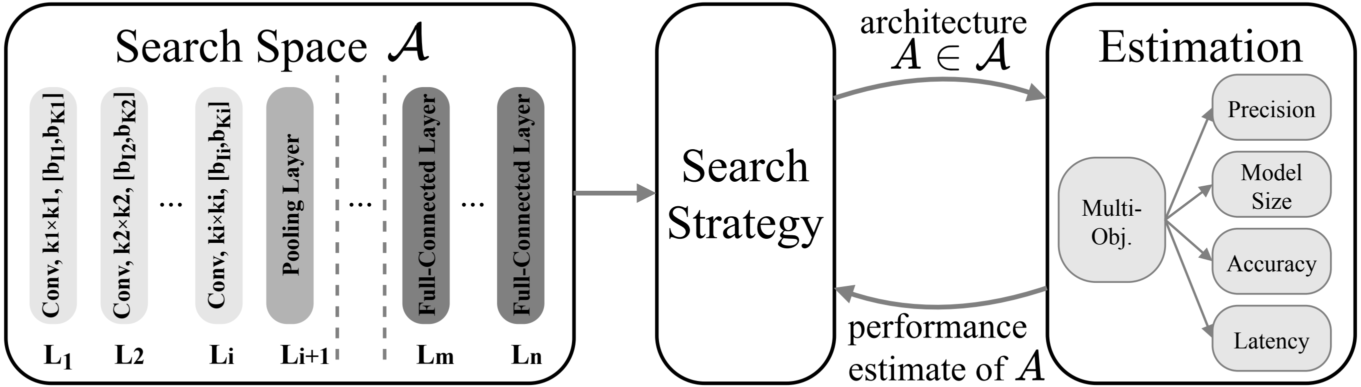


Fig. 1: Typical NAS workflow

Neural Architecture Search (NAS):

- Searches a neural network architecture for a customizable goal (maximize accuracy or meet latency constraints on particular hardware).
- A typical workflow of NAS can be divided into three aspects:
  - Search space is defined and constructed from variables such as convolutional kernel size and data precision.
  - Candidate network structures are identified through the search strategies.
  - Candidate networks are evaluated based on latency, accuracy, precision and size.
  - Subsequently, the next round of search is performed based on the feedback from the evaluation results.

NAS-Optimized Multi-Precision DNNs

Deep learning for edge computing:

deep neural network (DNN) 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**

⇩

Urgent need of energy-efficient multi-precision accelerator for NAS

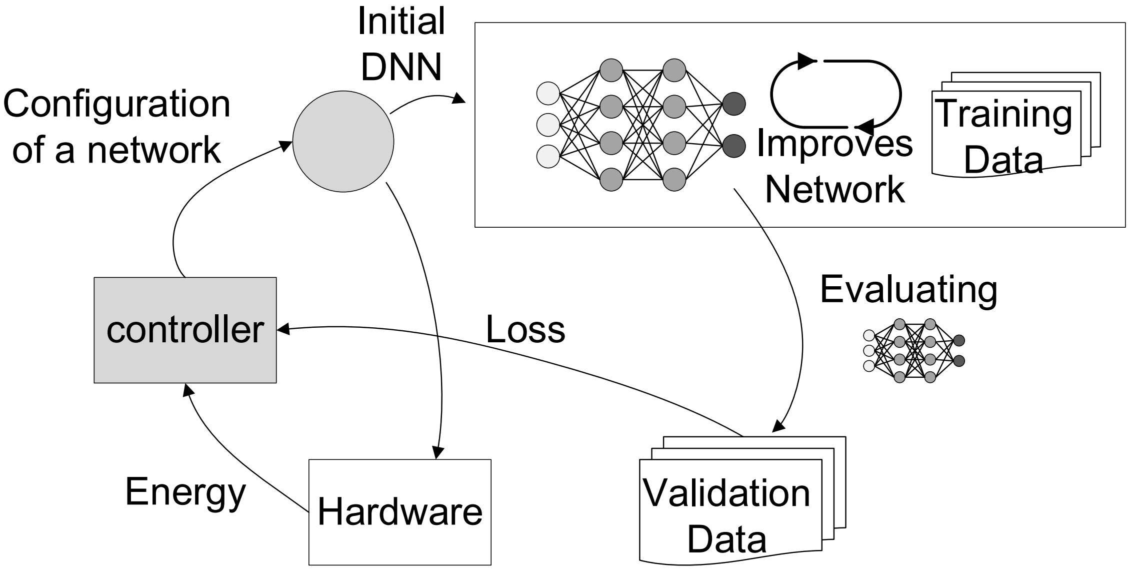


Fig. 2: Illustration of energy aware neural architecture search framework

Multi-Precision Neural Network Accelerator

Table 1: Evaluated NAS-based multi-precision CNN benchmarks.

CNN	Dataset	Model Weights	--- Proportion of --- 8-bits	4-bits	2-bits
VGG-16	CIFAR-10	138.0 MBytes	10.2%	89.8%	0%
LeNet-5	MNIST	0.5 MBytes	0%	55.0%	45.0%
ResNet-18	ImageNet	13.0 MBytes	5.5%	94.5%	0%
NAS-Based	-	-	21.8%	58.6%	19.6%

Note: NAS-Based summarized several VGG-16 models trained by NAS

4-bit operations: **>50% !**

Existing multi-precision design:

- Low-precision-combination (LPC): mainly 2-bit, large hardware cost, huge power consumption
- High-precision-split (HPS) : mainly 8-bit, poor throughput

Proposed Work:

- Bit-split-and-combination (BSC) vector PE: mainly 4-bit, tradeoff cost and throughput, better for NAS
- Precision-scalable vector systolic PE array: data reuse and energy efficient

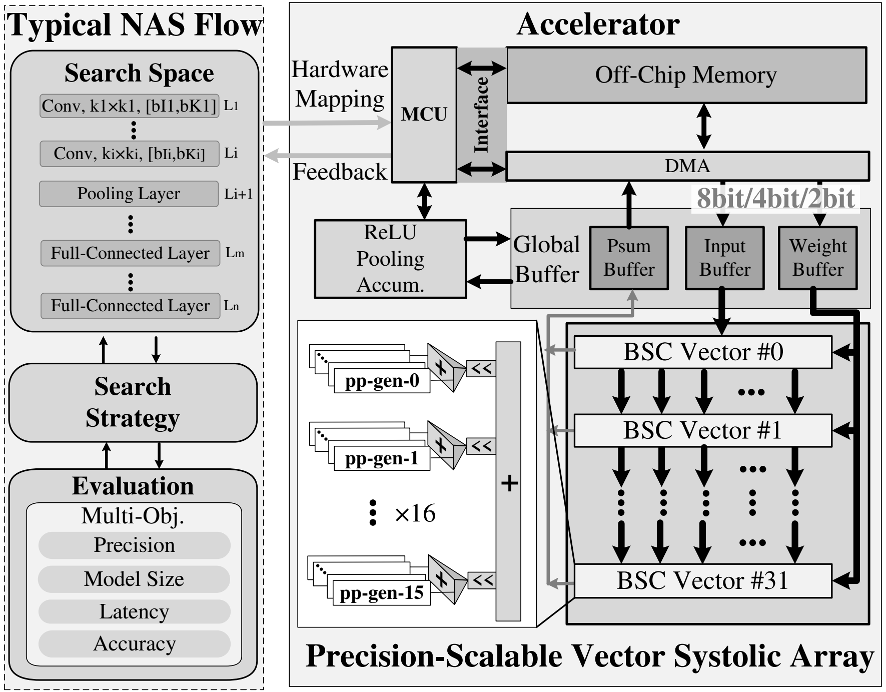


Fig. 3: Typical NAS flow with proposed multi-precision vector systolic array

Outline

- Introduction
  - NAS-Optimized Multi-Precision DNNs
  - Multi-Precision Neural Network Accelerator
- **Bit-Split-and-Combination (BSC) multi-precision vector MAC**
- Precision-scalable vector systolic PE array
- Experiment Results
- Conclusion

BSC multi-precision vector MAC

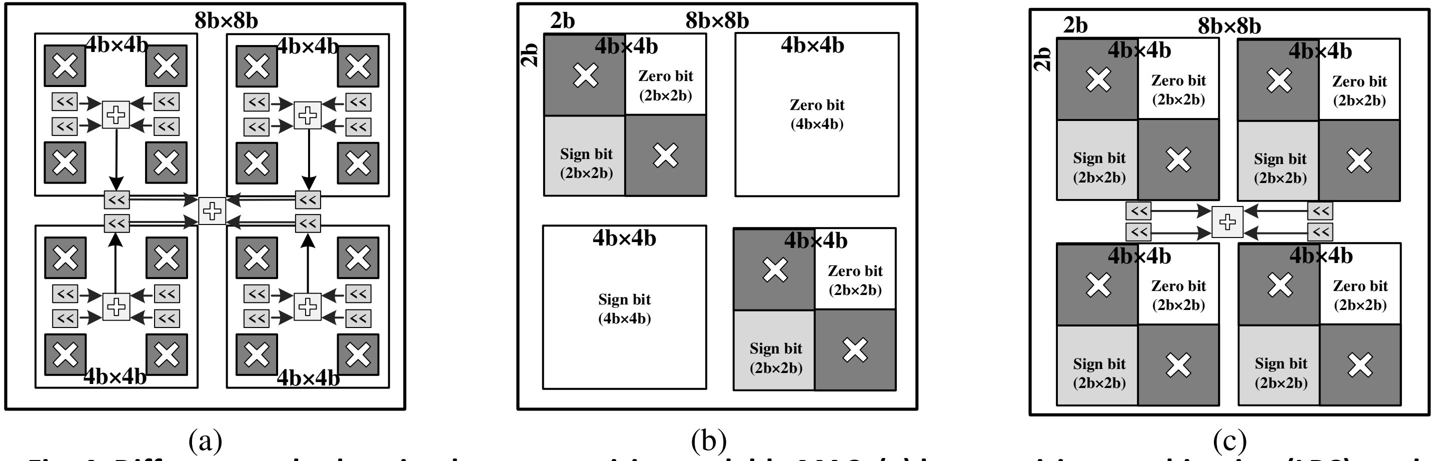


Fig. 4: Different methods to implement precision-scalable MAC: (a) low-precision-combination (LPC) method, (b) high-precision-split (HPS) method, (c) proposed bit-split-and-combination (BSC) method.

Method	Related works	Throughput (Ops.)			Bandwidth utilization (%)			Hardware utilization (%)			Feature
		2b×2b	4b×4b	8b×8b	2b×2b	4b×4b	8b×8b	2b×2b	4b×4b	8b×8b	
LPC	BitFusion, BitBlade, etc.	16	4	1	100	50	25	100	100	100	Large hardware cost Huge power consumption
HPS	Subword Parallel	4	2	1	100	100	100	25	50	100	Poor throughput performance
BSC	**Proposed**	**8**	**4**	**1**	**100**	**100**	**50**	**50**	**100**	**100**	**Tradeoff cost and throughput**

BSC multi-precision vector MAC

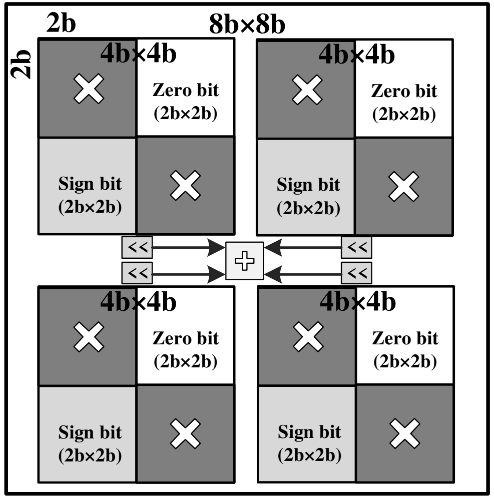


Fig. 4(c): BSC Method MAC

- 4-bit bit-split unit ×L
- Combine to 8-bit vector
- Share shifters (Area ↓)
- Throughput ↑
- Energy efficiency ↑

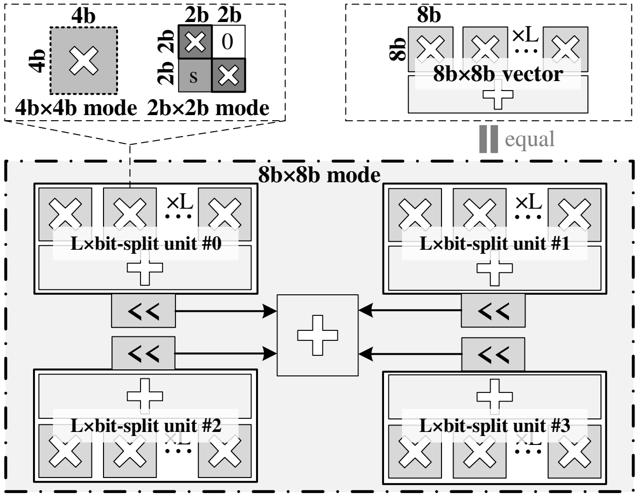


Fig. 5: BSC vector MAC with length L.

$$A_{n-bit} = \sum_{i=0}^{n} 2^{i} \times a_i \ (1)$$

$$B_{n-bit} = \sum_{j=0}^{n} 2^{j} \times b_i \ (2)$$

$$A \times B = \sum_{e=0,4,4,8} \left(\sum_{j=0}^{3} \sum_{i=0}^{3} 2^{i+j} \times a_i b_j \right) \times 2^{e*s} \ (3)$$

Note: if mode is 8-bit, then s=1, else s=0.

$$\boldsymbol{A_{vector}} \times \boldsymbol{B_{vector}}$$

$$= \sum_{l=1}^{L} \sum_{e=0,4,4,8} \left(\sum_{j=0}^{3} \sum_{i=0}^{3} 2^{i+j} \times a_i^l b_j^l \right) \times 2^{e*s}$$

$$= \sum_{e=0,4,4,8} \left(\sum_{l=1}^{L} \sum_{j=0}^{3} \sum_{i=0}^{3} 2^{i+j} \times a_i^l b_j^l \right) \times 2^{e*s} \ (4)$$

BSC multi-precision vector MAC

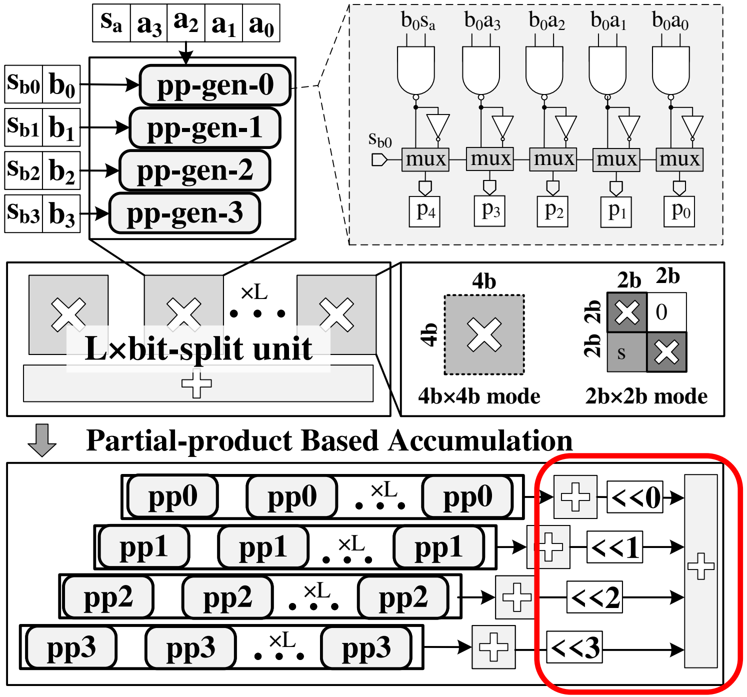


Fig. 6: Bit-split unit implementation with same shift partial-product accumulation.

$$\boldsymbol{A}_{vector} \times \boldsymbol{B}_{vector}$$

$$= \sum_{l=1}^{L} \sum_{e=0,4,4,8} \left(\sum_{j=0}^{3} \sum_{i=0}^{3} 2^{i+j} \times a_i^l b_j^l\right) \times 2^{e*s}$$

$$= \sum_{e=0,4,4,8} \left(\sum_{l=1}^{L} \sum_{j=0}^{3} \sum_{i=0}^{3} 2^{i+j} \times a_i^l b_j^l\right) \times 2^{e*s} \quad (4)$$

with same shift partial-product accumulation

$$\boldsymbol{A}_{vector} \times \boldsymbol{B}_{vector}$$

$$= \sum_{e=0,4,4,8} \left(\sum_{j=0}^{3} \sum_{l=1}^{L} \sum_{i=0}^{3} 2^{i+j} \times a_i^l b_j^l\right) \times 2^{e*s} \quad (5)$$

Sharing accumulators and shifters

- Hardware resources ↓
- Energy efficiency ↑

Outline

- Introduction
  - NAS-Optimized Multi-Precision DNNs
  - Multi-Precision Neural Network Accelerator
- Bit-Split-and-Combination (BSC) multi-precision vector MAC
- **Precision-scalable vector systolic PE array**
- Experiment Results
- Conclusion

Precision-scalable vector systolic PE array

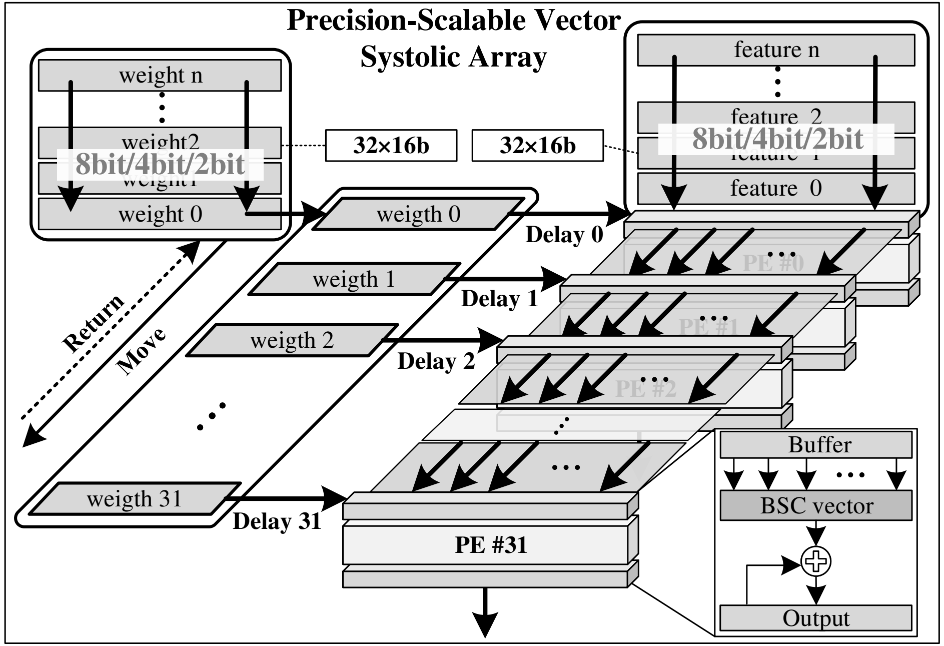


Fig. 7: BSC precision-scalable vector systolic PE array dataflow

- 32 BSC vector PEs
- Supports 1024 8bit×8bit, 4096 4bit×4bit or 8192 2bit×2bit MAC operations.
- Multi-precision feature data transmits with vector length 32
- Multi-precision Weight sent to the buffer of PE_0 to PE_31 after delay 0 clock to 31 clocks
- Outputs from PE array are transmits to psum buffer

The input data is reused efficiently!
(32 times reuse)

Precision-scalable vector systolic PE array

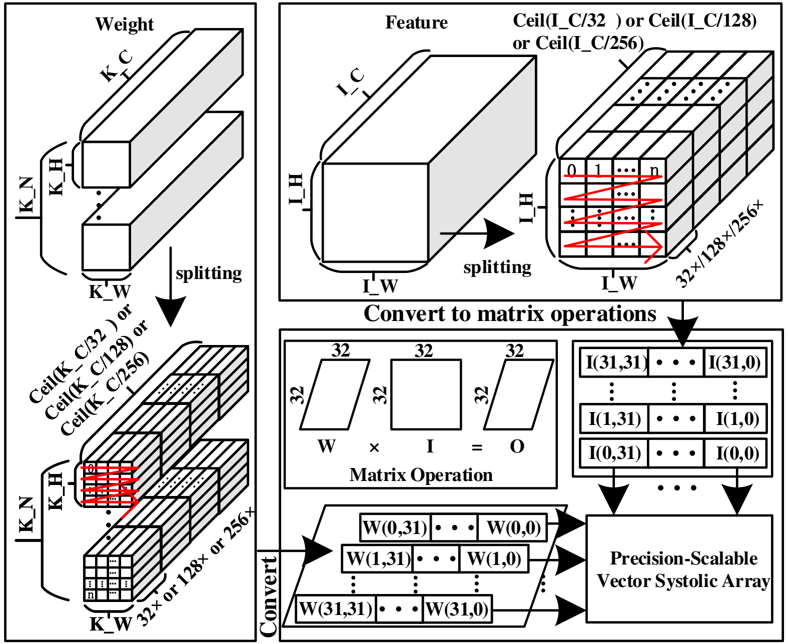


Fig. 8: Convolution and matrix mapping in vector systolic array.

CNN layer with:

- Weight data: W[K_N][K_C][K_W][K_H]
- Feature data: I[I_C][I_W][I_H]

Splitting and Mapping to matrix
Parallel with K_C and I_C channel

Precision-scalable vector systolic PE array:

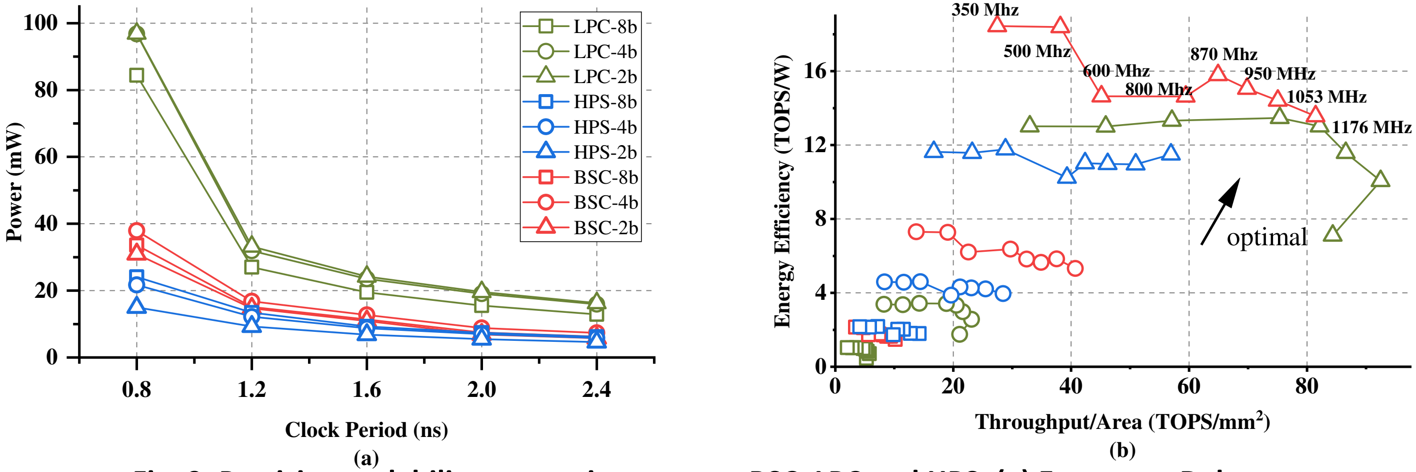
- 32×32 matrix operation

Note: 2-bit, 4-bit and 8-bit vector operation with 32, 128 and 256 parallelism, respectively.

Outline

- Introduction
  - NAS-Optimized Multi-Precision DNNs
  - Multi-Precision Neural Network Accelerator
- Bit-Split-and-Combination (BSC) multi-precision vector MAC
- Precision-scalable vector systolic PE array
- **Experiment Results**
- Conclusion

Experiment Results



(a)

(b)

Fig. 9: Precision scalability comparison among BSC, LPC and HPS: (a) Energy vs. Delay; (b) Energy efficiency vs. Area efficiency

Experiment set up:

- Synthesized by Synopsys Design Compiler
- Implemented under SIMC 28-nm 1V process
- PrimeTime PX are used to obtain the power
- VCS tools are applied in verification

BSC method has better energy & area efficiency!

Experiment Results

Comparison of Max Energy Efficiency:

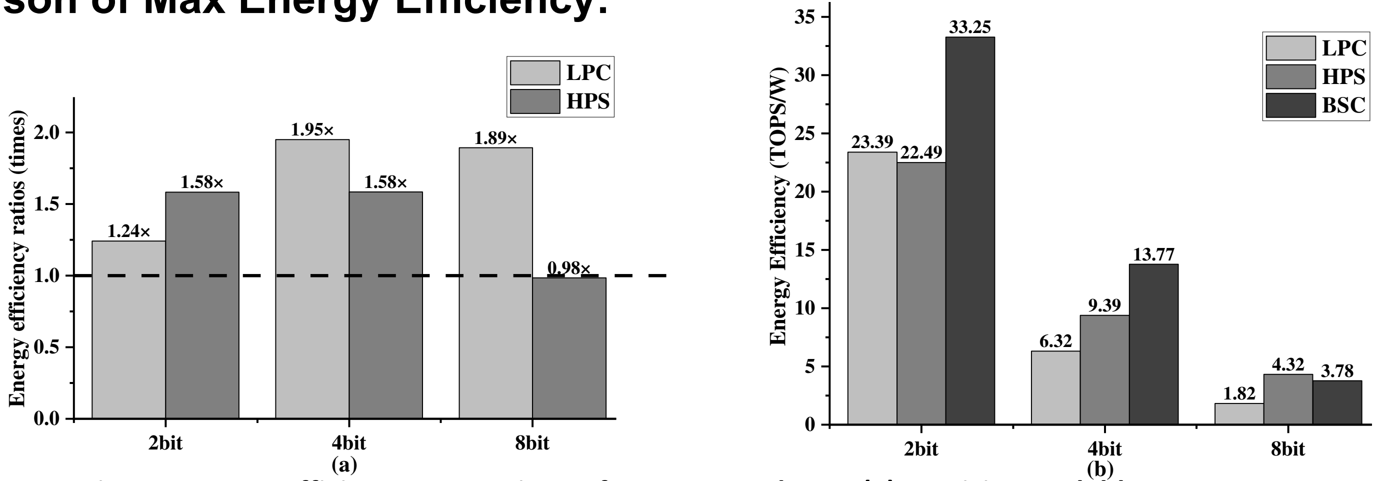


Fig. 10: Energy efficiency comparison of BSC, LPC and HPS: (a) Precision-scalable vector PE; (b) Vector systolic PE array

- 2× energy efficiency than LPC in 4-bit and 8-bit modes
- 1.6 × energy efficiency than HPS in 2-bit and 4-bit modes
- Systolic dataflow further improves energy efficiency

Experiment Results

Comparison of Multi-Precision Computation on NAS-CNNs:

Table 1: Evaluated NAS-based multi-precision CNN benchmarks.

CNN	Dataset	Model Weights	--- Proportion of --- 8-bits	4-bits	2-bits
VGG-16	CIFAR-10	138.0 MBytes	10.2%	89.8%	0%
LeNet-5	MNIST	0.5 MBytes	0%	55.0%	45.0%
ResNet-18	ImageNet	13.0 MBytes	5.5%	94.5%	0%
NAS-Based	-	-	21.8%	58.6%	19.6%

Note: NAS-Based summarized several VGG-16 models trained by NAS

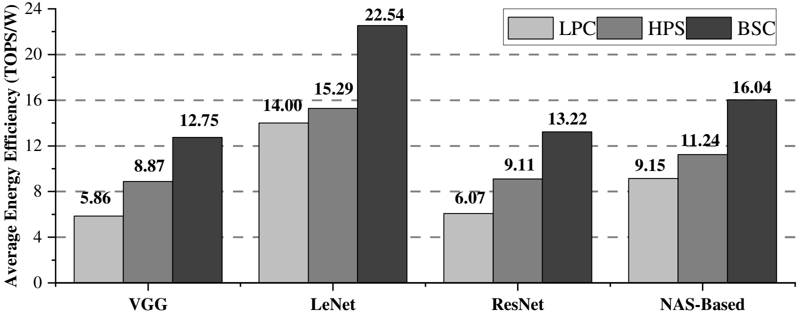


Fig. 11: Average energy efficiencies of precision-scalable vector systolic PE array with HPS, LPC and BSC vectors on multi-precision CNN benchmarks.

- 2.18× (LeNet) energy efficiency improvement
- Benefits from the **vector systolic architecture** and **high energy-efficient multi-precision BSC method**.

Outline

- Introduction
  - NAS-Optimized Multi-Precision DNNs
  - Multi-Precision Neural Network Accelerator
- Bit-Split-and-Combination (BSC) multi-precision vector MAC
- Precision-scalable vector systolic PE array
- Experiment Results
- **Conclusion**

Conclusion

- For better support for **NAS-Optimized Multi-Precision CNNs**, **the** **BSC vector systolic accelerator** with improved **energy-efficient** performance is proposed.
- The maximum energy efficiency of the proposed BSC vector PE is up to **1.95×** higher in 2-bit, 4-bit and 8-bit operations when compared with LPC and HPS PEs.
- The proposed vector systolic BSC PE array achieves up to **22.54 TOPS/W** in NAS-optimized multi-precision LeNet-5.
- The maximum improvement of average energy efficiency is **2.18×** higher than that of LPC PE array and HPS PE array.

Thanks for your attention!

Corresponding Author & PI
Prof. Hao YU
yuh3@sustech.edu.cn

SUSTech Southern University of Science and Technology

深港微电子学院
School of Microelectronics