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

Introduction

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

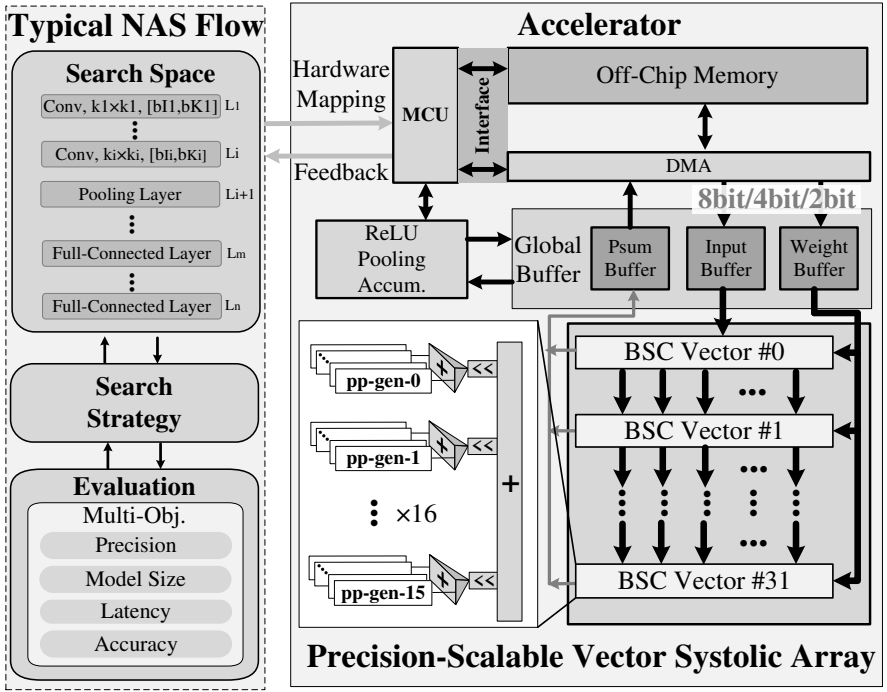
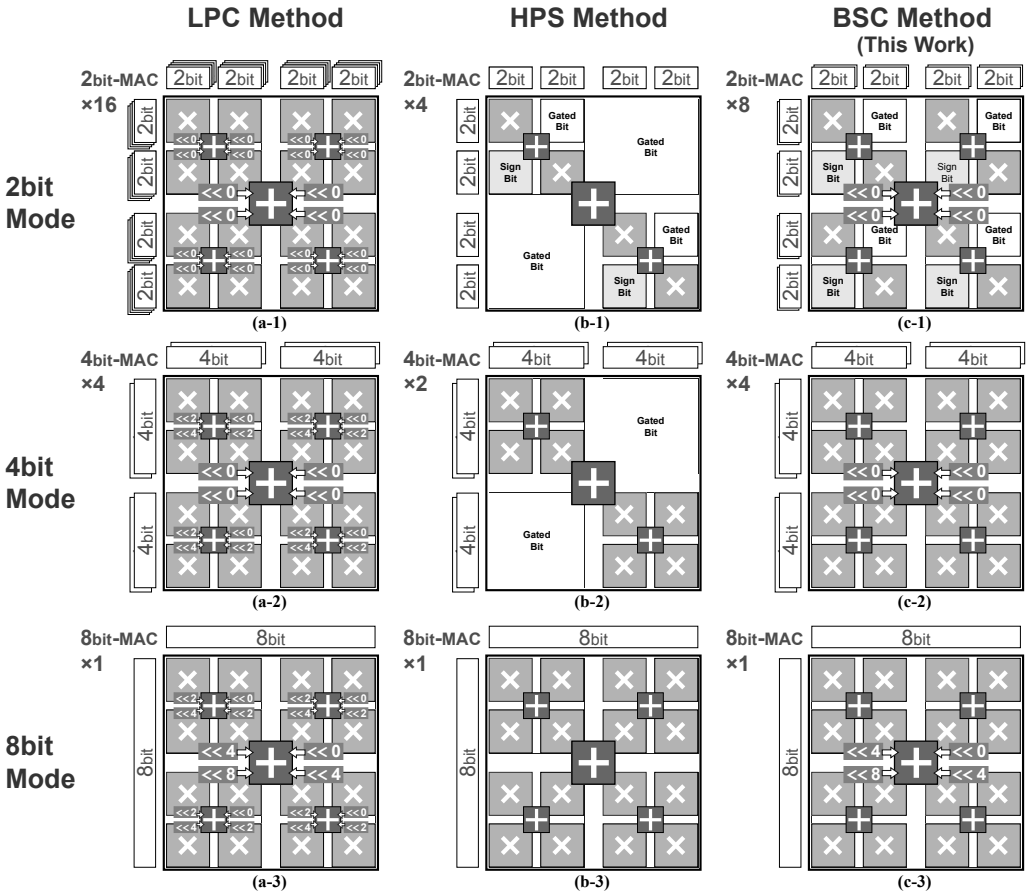


Fig. 1: Typical NAS flow with proposed multi-precision vector systolic array

BSC Multi-Precision Vector PE



Bit-split-and-combination (BSC) vector PE:

- Combines the advantages of LPC (BitFusion) and HPS(SubwordParallel), decent bandwidth & hardware utilization;
- Mainly 4-bit operation, tradeoff consumption and throughput;
- Adopt vectorized-multiplicator to reduce the number of accumulators and shifters, improve throughput and performance.

Precision-Scalable Vector Systolic PE Array

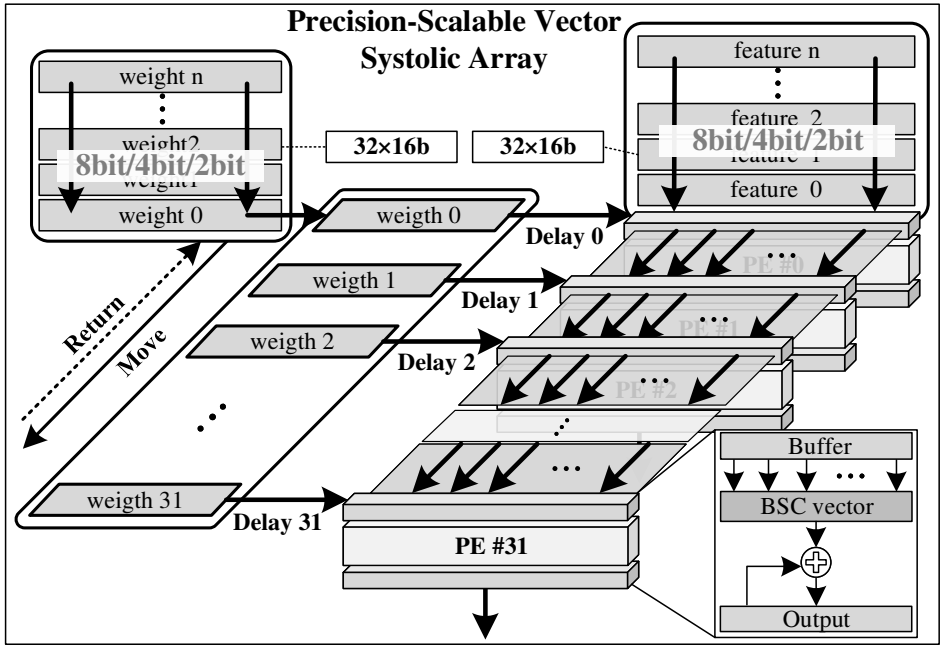


Fig. 3: BSC precision-scalable vector systolic PE array dataflow

Vector-Systolic PE Array:

- Multi-precision feature data transmits with vector length 32;
- Multi-precision Weight sent to the buffer of each PE after different clocks' delay;
- Multi-precision feature data transmits one-by-one with vector length 32;
- Outputs from PE array transmits to output buffer;

Experiment Results

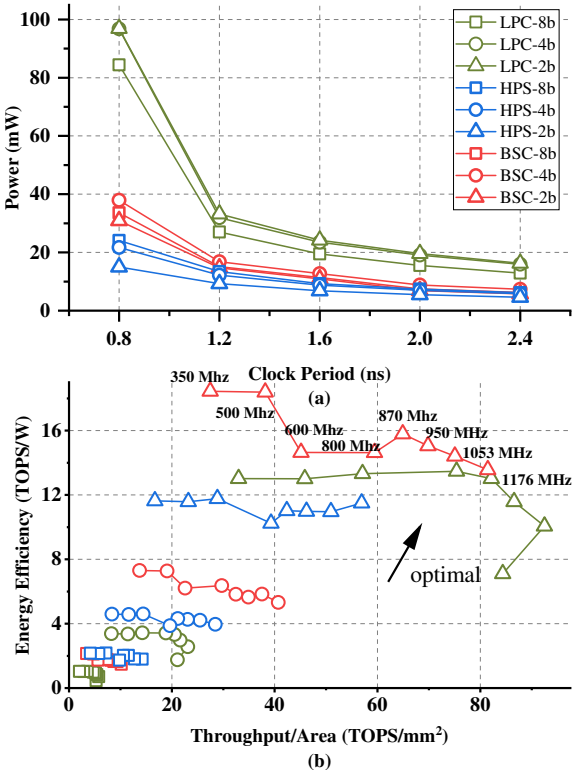


Fig. 4: Precision scalability comparison among BSC, LPC and HPS: (a) Energy vs. Delay; (b) Energy efficiency vs. Area efficiency

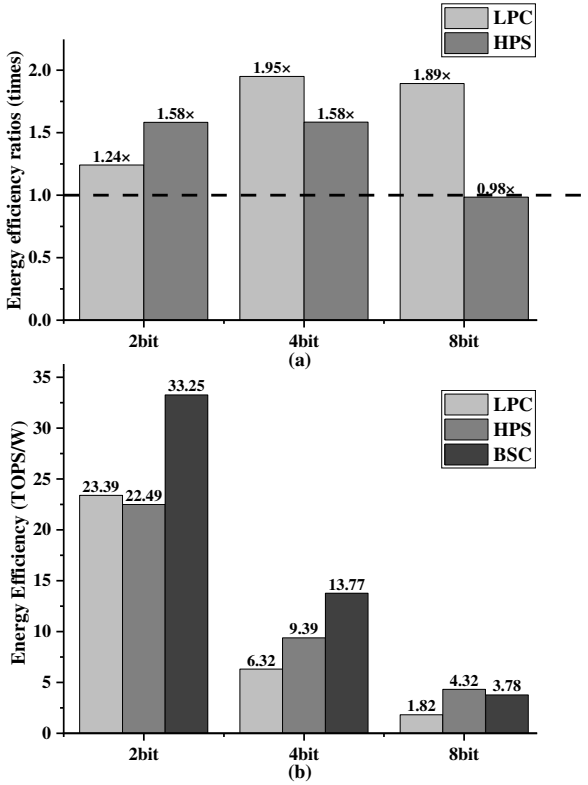


Fig. 5: Energy efficiency comparison of BSC, LPC and HPS: (a) Precision-scalable vector PE; (b) Vector systolic PE array

Table: 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

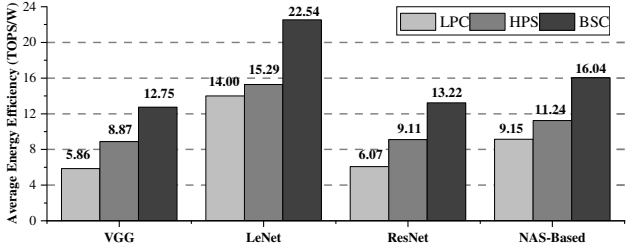


Fig. 6: Average energy efficiencies of precision-scalable vector systolic PE array with HPS, LPC and BSC vectors on multi-precision CNN benchmarks.

Conclusion & Future Work

- For better support for **NAS-Optimized Multi-Precision CNNs**, **the BSC vector systolic accelerator** is proposed.
- The proposed vector systolic BSC PE array achieves up to **22 TOPS/W (28nm SMIC)** in NAS-optimized multi-precision LeNet-5.

- **Tapeout Plan:**
  - 28nm TSMC Low Power;
  - 2.25mm^2 die area (only array with its SRAM, part of whole chip);
  - 16×16 BSC vector array with 144KB SRAM;
  - Scheduled in next two months (June 2022).

Thanks for your attention!

Q & A

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