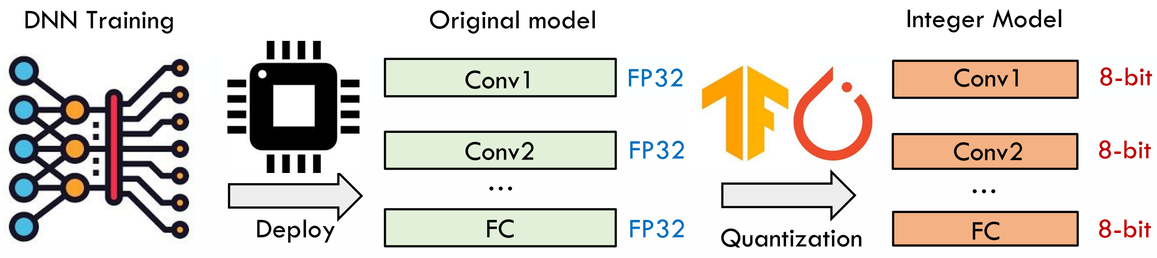
MSD: Mixing Signed Digit Representations for Hardware-efficient DNN Acceleration on FPGA with Heterogeneous Resources

Jiajun Wu, Jiajun Zhou, Yizhao Gao, Yuhao Ding, Ngai Wong, Hayden Kwok-Hay So

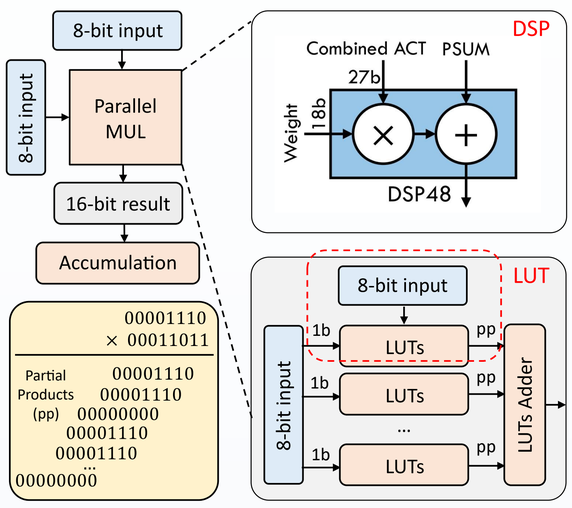
Department of Electrical and Electronic Engineering, University of Hong Kong

MOTIVATION

DNN Quantization



Quantized Multiplication on LUTs and DSPs



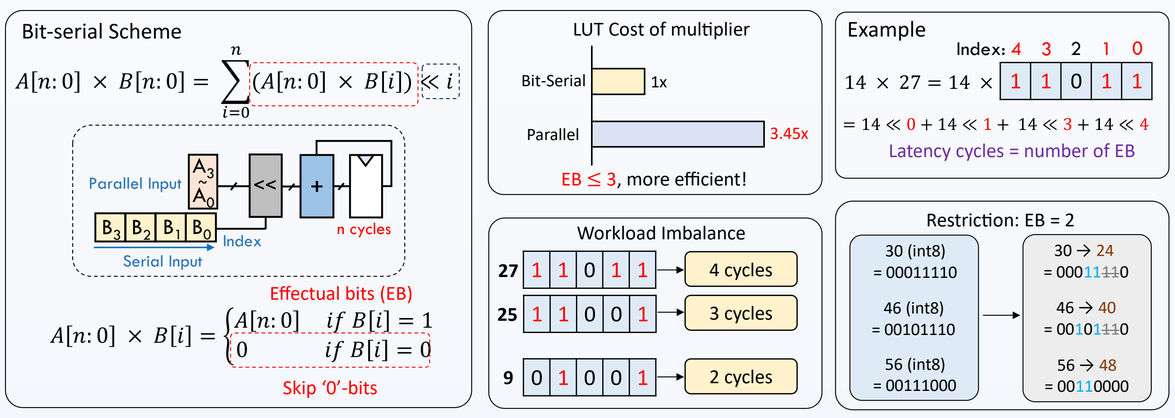
Bit-serial scheme with Bit-sparsity

```
  00001110
× 00011011
----------
  00001110
  00001110
  00001110
  00001110
```

Serial → Effectual partial products

- Only need to implement one module for bit-serial design
- Improve peak performance by deploying multiplication on both LUTs and DSPs
- Multiplication on LUTs: Bit-serial scheme with bit-sparsity optimization

Bit-serial Implementation and the Effectual Bits



- Make bit-serial scheme more efficient than conventional parallel design
- To solve the problem of workload imbalance in bit-serial scheme

MORE INFORMATION IS HERE!

Paper in our group site!

MSD open-source in GitHub

Our work passes the artifact evaluation process

METHODOLOGY

Restricted Signed-Digit Representation (RSD)

To solve the imbalance issue

30 (int8) = 00011110	30 → 24 = 00011110
46 (int8) = 00101110	46 → 40 = 00101110
56 (int8) = 00111000	56 → 48 = 00110000

Restriction: EB = 2

Large quantization error ☹

Our Approach: Signed-digit representation

Let $X[i]$ (i-th bit in the number with n-bit) expand to 0, 1 and -1 ($\bar{1}$):

$$X = \sum_{i=0}^{n-1} \left(X[i] \times 2^i\right), \qquad X[i] = \{0, 1, -1\}$$

Effectual bits (EB)

2's complement is a special case of signed-digit:

$$X = \sum_{i=0}^{n-1} \left(X[i] \times 2^i\right) \qquad X[i] = \begin{cases} \{0,1\} & if\ i \neq n-1 \\ \{0,-1\} & if\ i = n-1 \end{cases}$$

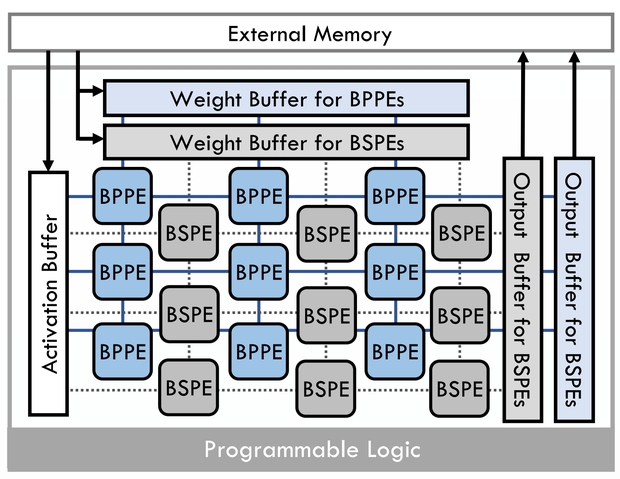
Restrict EB: Restricted signed-digit (RSD)

Original Numbers	2's complement $EB = 2$,	Error	RSD $EB = 2$	Error
30 (int8) = 00011110	30 → 24 = 00011110	6	30 = 32 - 2 = 0010000$\bar{1}$0	0
46 (int8) = 00101110	46 → 40 = 00101110	6	48 = 32 + 16 = 00110000	2
56 (int8) = 00111000	56 → 48 = 00010001	8 ☹	56 = 64 - 8 = 0100$\bar{1}$000 ☺	0

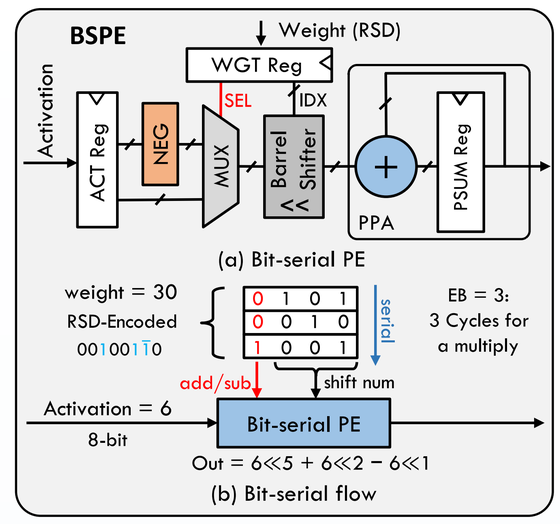
Smaller quantization errors compared with 2's complement!

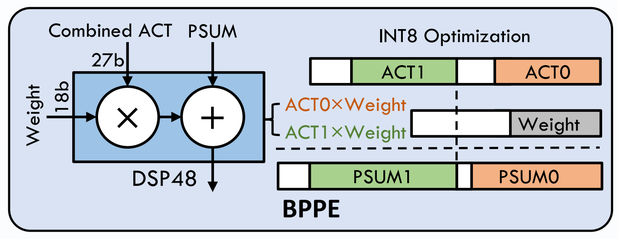
METHODOLOGY

Heterogeneous Architecture



BSPE: Bit-serial PE on LUTs. **BPPE**: Bit-parallel on DSPs





RSD Encode Under the EB = 2 Restriction

30 = 32 - 2 = 00100001̄0

48 = 32 + 16 = 00110000

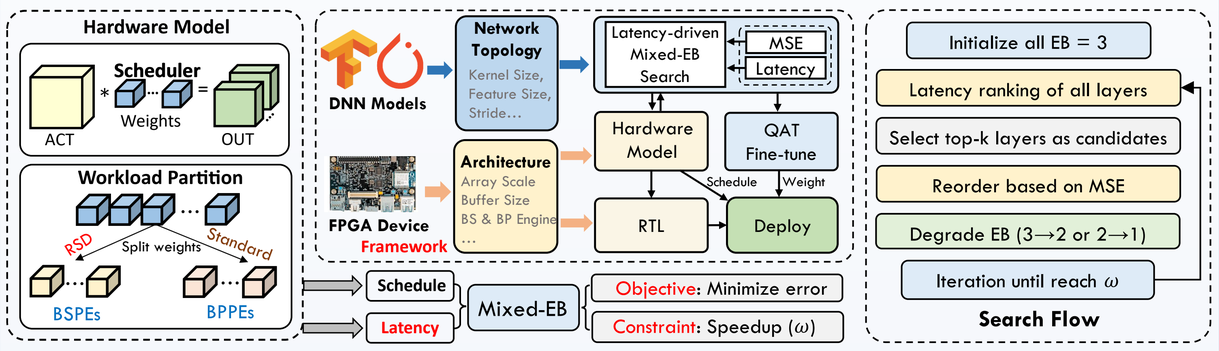
EB's position (index) encoding

Another SEL bit to indicate {1, -1}

0	1	0	1	
1	0	0	1	
0	1	0	1	
0	1	0	0	

- Bit-serial PE processes RSD-based weights, which need to be fine-tuned by QAT
- Bit-parallel PE processes standard weights. We need to balance the workloads

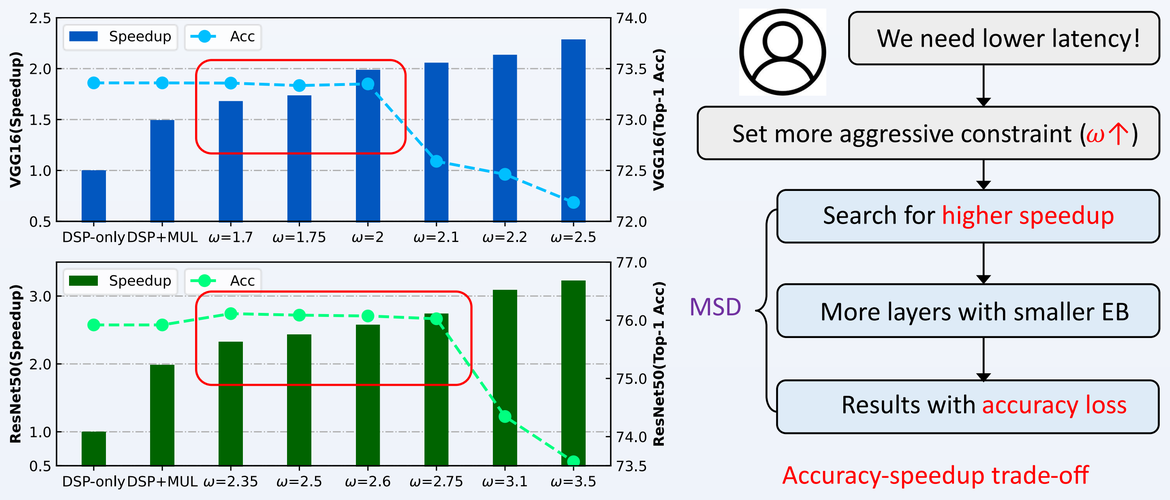
End-to-End Framework: Mixed-EB Quantization



With the scheduler and search algorithm, we set up the mixed-EB quantization scheme in which different layers have different restriction of EB, from 1 to 3.

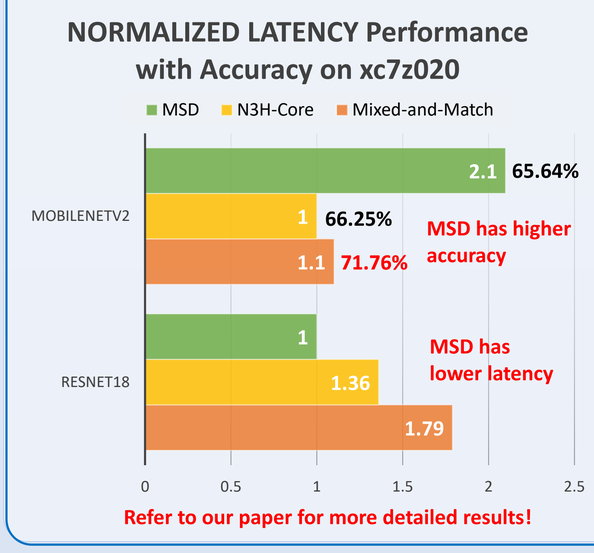
RESULTS

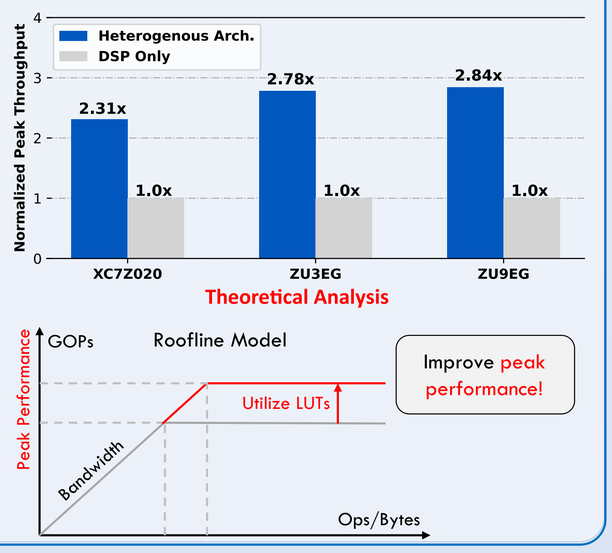
Accuracy-speedup Trade-off



- We can reach a balance between accuracy and speedup in the red box side.
- Also, out results show that the bit-serial with bit-sparsity scheme is more efficient than the conventional bit-parallel multiplier design, in terms of latency.

Comparison





CASR group, The University of Hong Kong, Hong Kong