IEEE Xplore® Digital Library

ISCAS 2020

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

CAS

High-Speed Power-Efficient Coarse-Grained Convolver Architecture using Depth-First Compression Scheme

Yi-Lin Wu, Yi Lu, and Juinn-Dar Huang

Institute of Electronics, National Chiao Tung University, Hsinchu, Taiwan

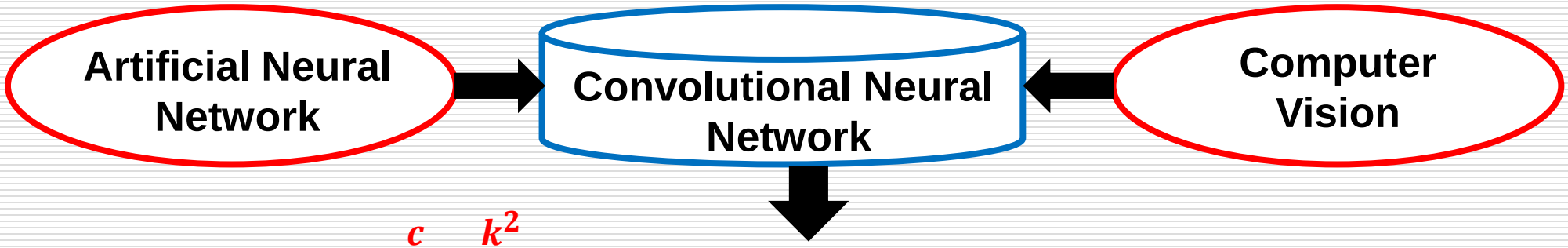
2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

國立交通大學
National Chiao Tung University

Outline

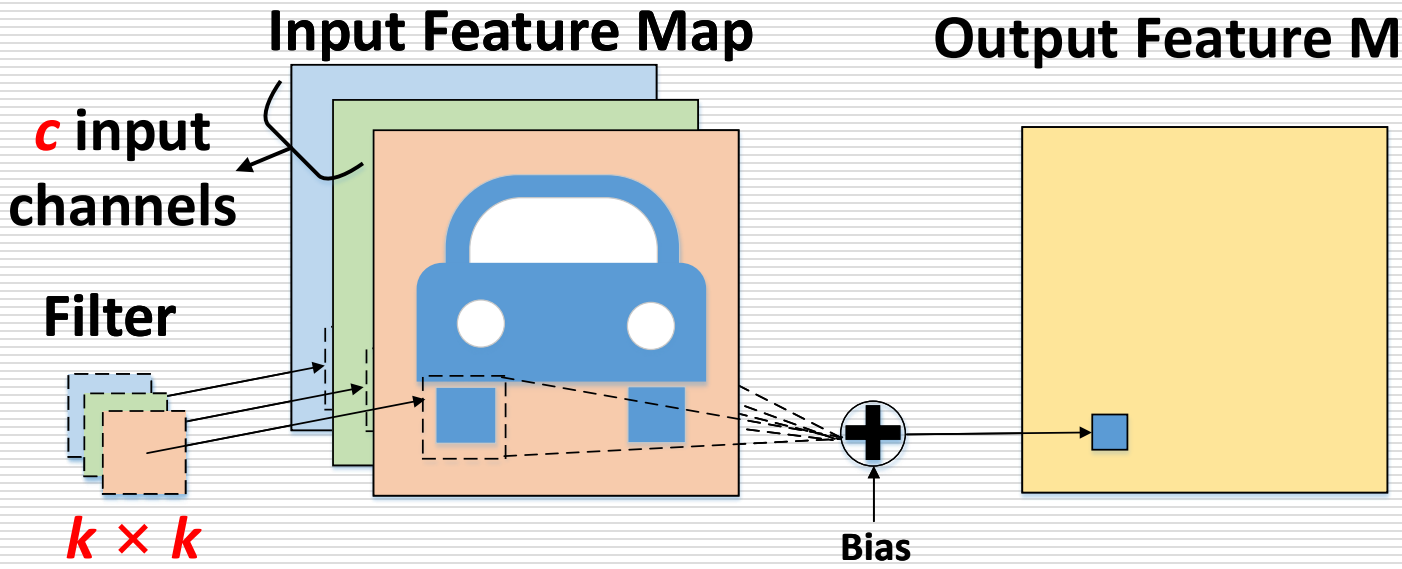
- Introduction
- Baseline convolver architecture
- Proposed convolver architecture
- Comparisons
- Experimental results
- Summary

Convolutional Neural Networks (CNNs)



$$Output = \left(\sum_{m=1}^{c}\sum_{n=1}^{k^2} Input_{m,n} \times Weight_{m,n}\right) + bias$$

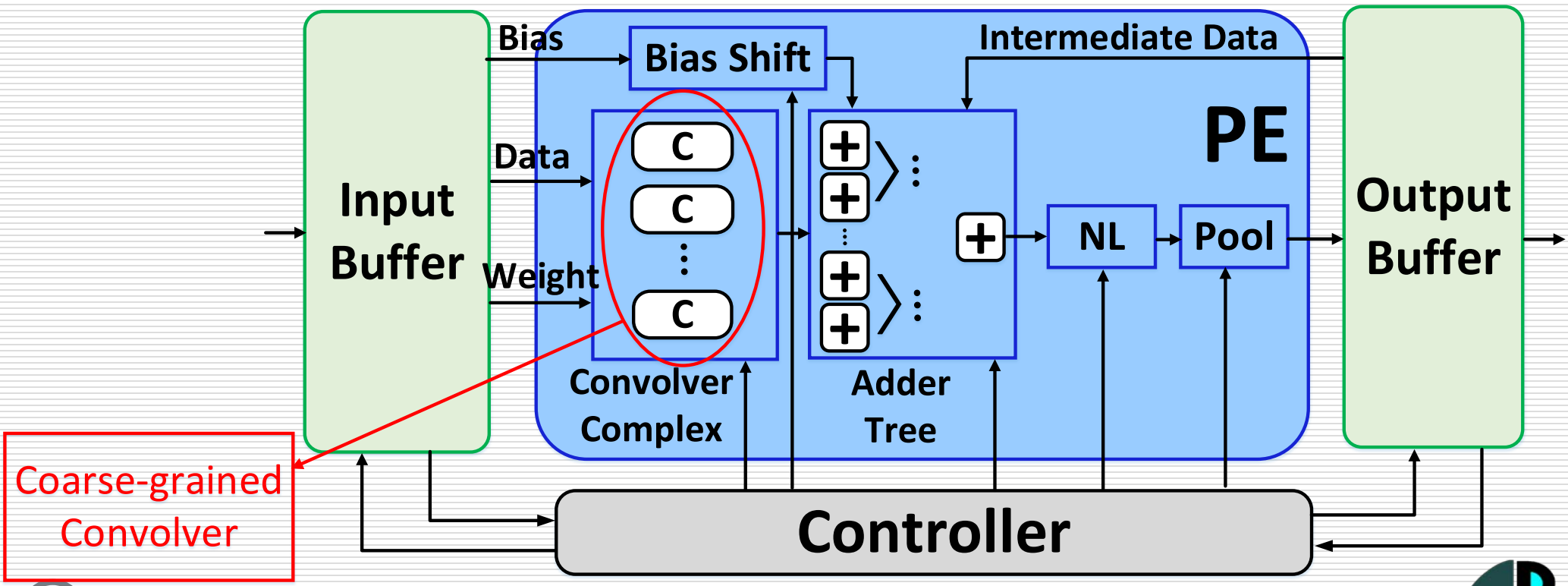
a convolution operation



CNN Hardware Accelerators

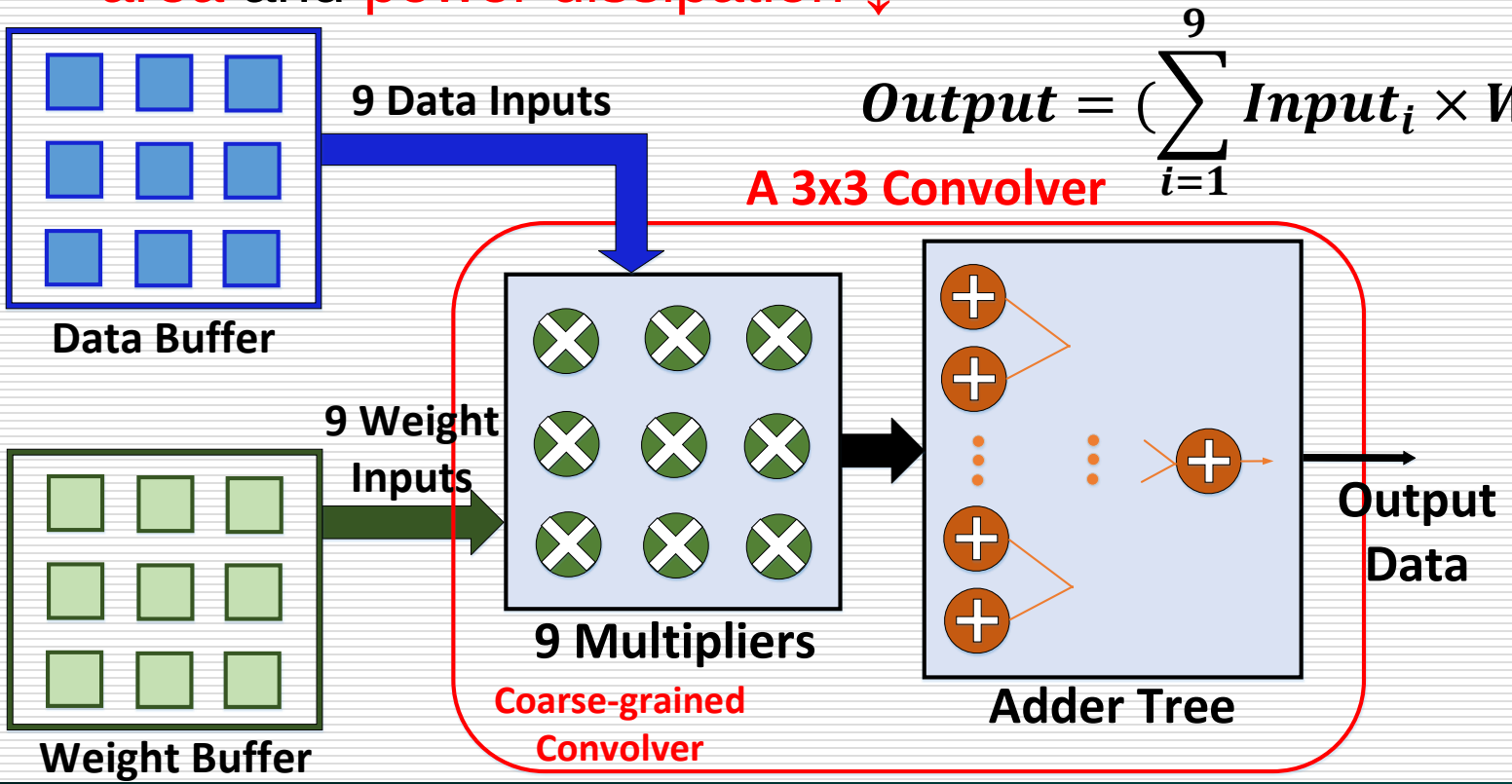
- Fine-grained architecture
  - a large set of simple PEs (e.g., Eyeriss)
- Coarse-grained architecture
  - fewer but more powerful PEs

[1] J. Qiu et al., ISFPGA 2016



Motivation – An Integrated Convolver

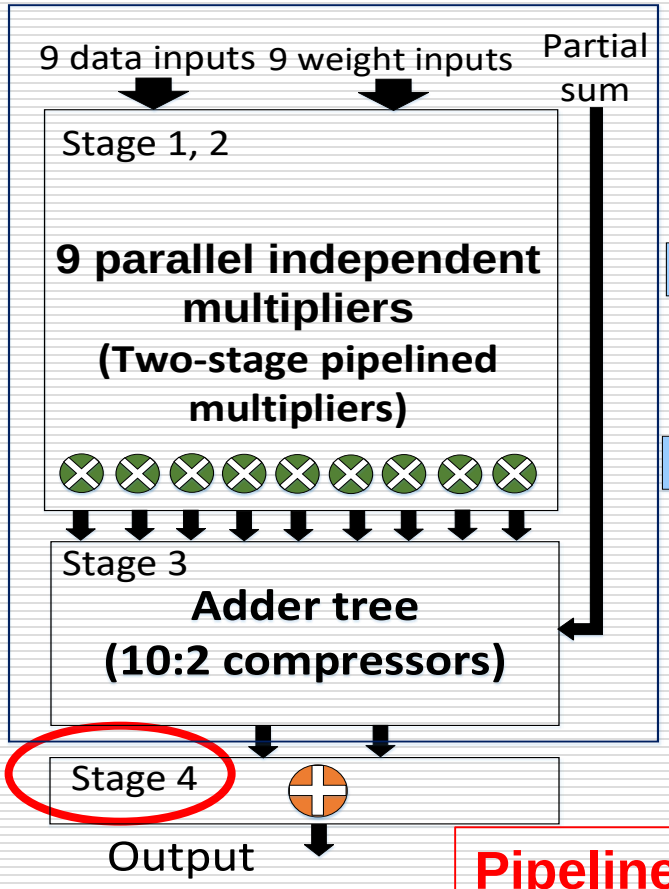
- Break the boundary between multipliers
  - global optimization (area/delay/power) across multipliers
- Eliminate internal carry-propagation adders (CPAs)
  - area and power dissipation ↓



Overview of Convolver Architectures

Previous work

Parallel independent multipliers + adder tree



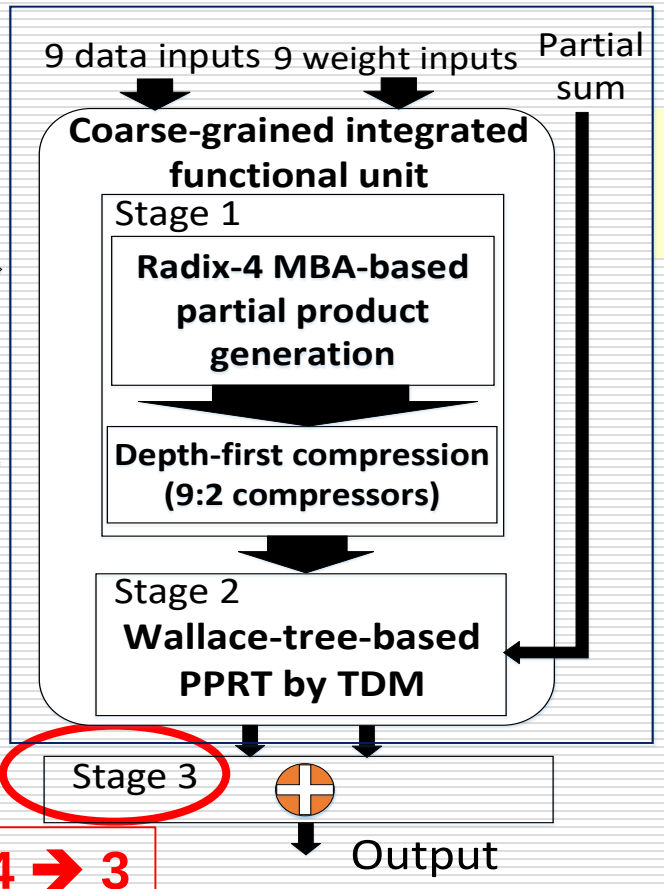
#CPA = 10

Repartition

Reorder

Proposed work

Coarse-grained integrated functional unit



#CPA = 1

Pipeline stages: 4 → 3
 # Pipeline registers: ↓

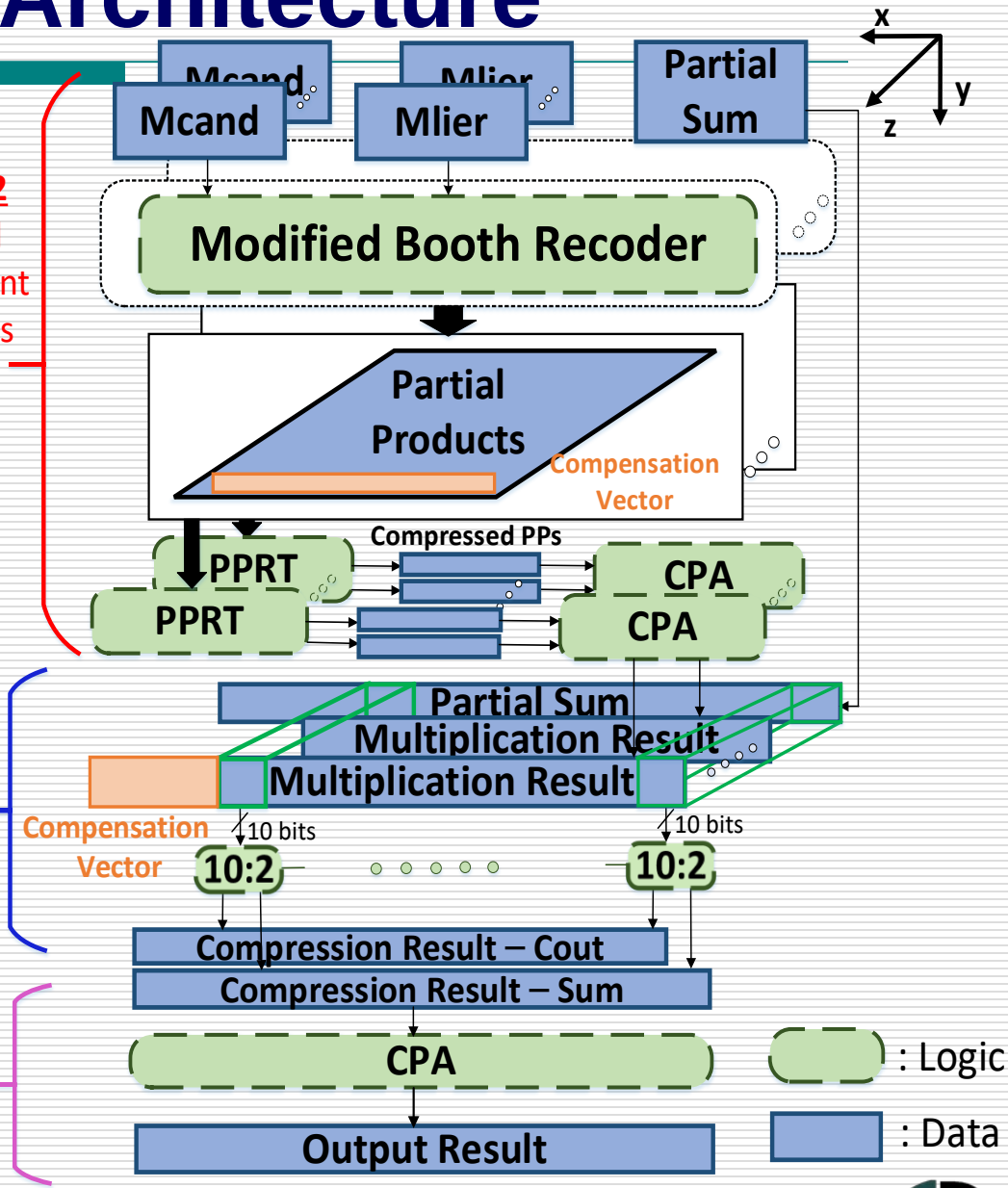
Example:
3x3 convolver

CPA

Two-stage pipelined multiplier

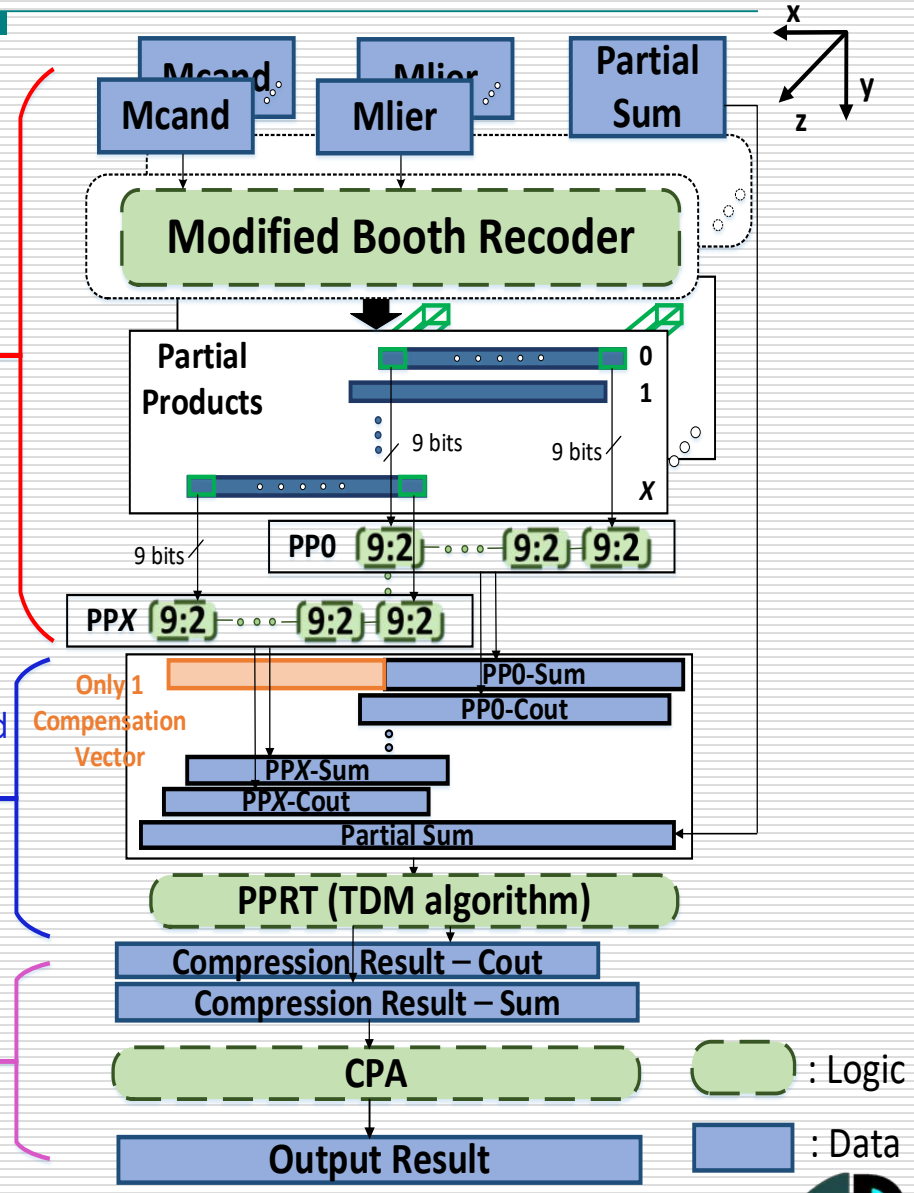
Baseline Convolver Architecture

- Stage 1&2 (along x-y plane)
  - 9 classical 2-stage pipelined multipliers
  - contains 9 CPAs and 9 **constant** compensation vectors
- Stage 3 (along z-axis)
  - an adder tree by 10:2 compressors
- Stage 4
  - 1 final CPA
- **Summary**
  - **4** pipeline stages
  - **10** CPAs
  - **10** compensation vectors



Proposed Convolver Architecture

- Stage 1 (along z-axis)
  - MBR + 9:2 compressors
  - **depth-first compression**: PPs of the same position across 9 multiplications
- Stage 2 (along x-y plane)
  - delay-minimized Wallace-tree
- Stage 3
  - final CPA
    - ➔ **only 1 CPA !**
- Only 3 pipeline stages
  - ➔ fewer pipeline registers
- Only 1 compensation vector
  - ➔ smaller area footprint



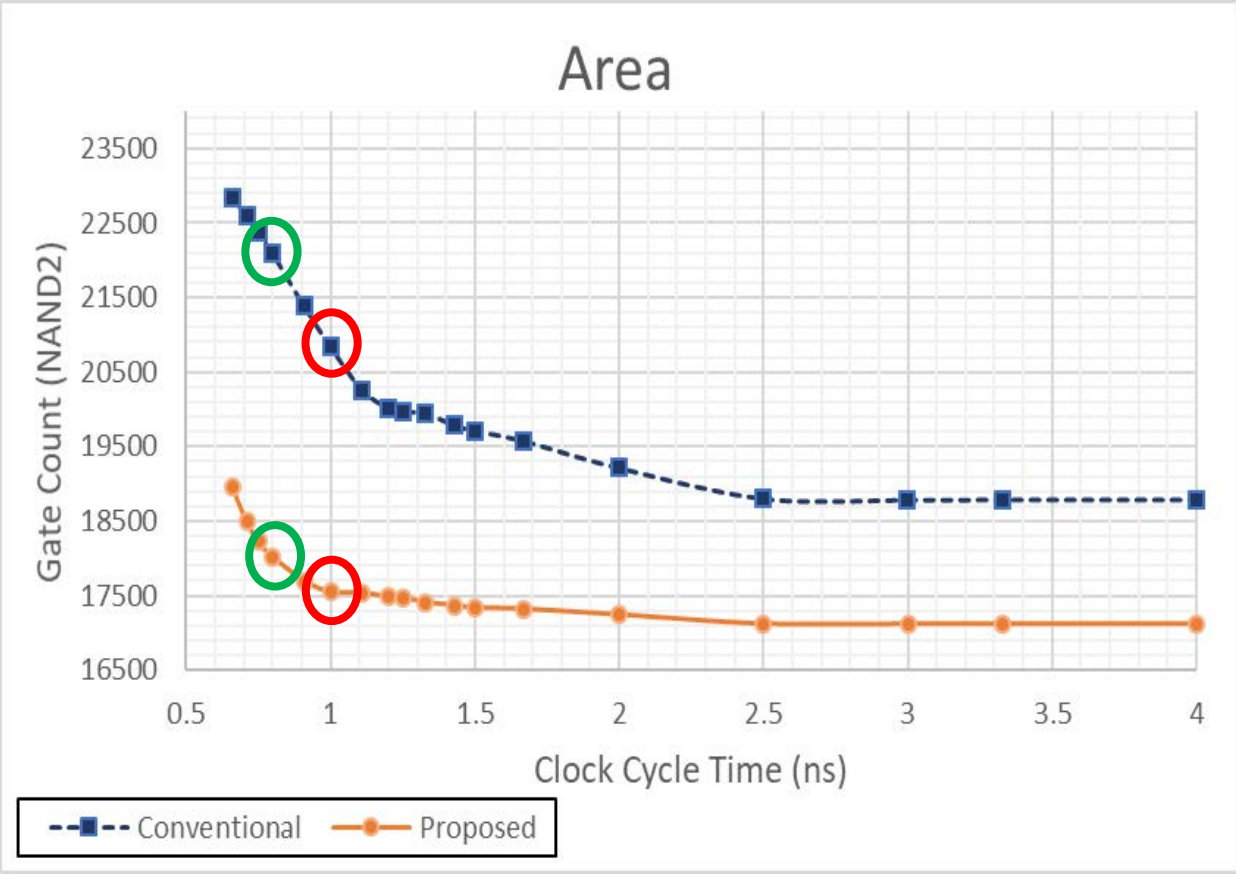
Comparisons between Two Convolvers

	Baseline	Proposed
Pipeline	4 stages	3 stages
Number of Register bits	1023~1213 (Due to retiming)	770 (36.5% less)
Number of CPAs	9 CPAs in internal multipliers + 1 CPA at the final stage = 10 CPAs	1 CPA at the final stage
Number of compensation vectors	9 compensation vectors in 9 internal multipliers + 1 compensation vector at the third stage	1 compensation vector at the second stage

Better !!!

Experimental Results (Area)

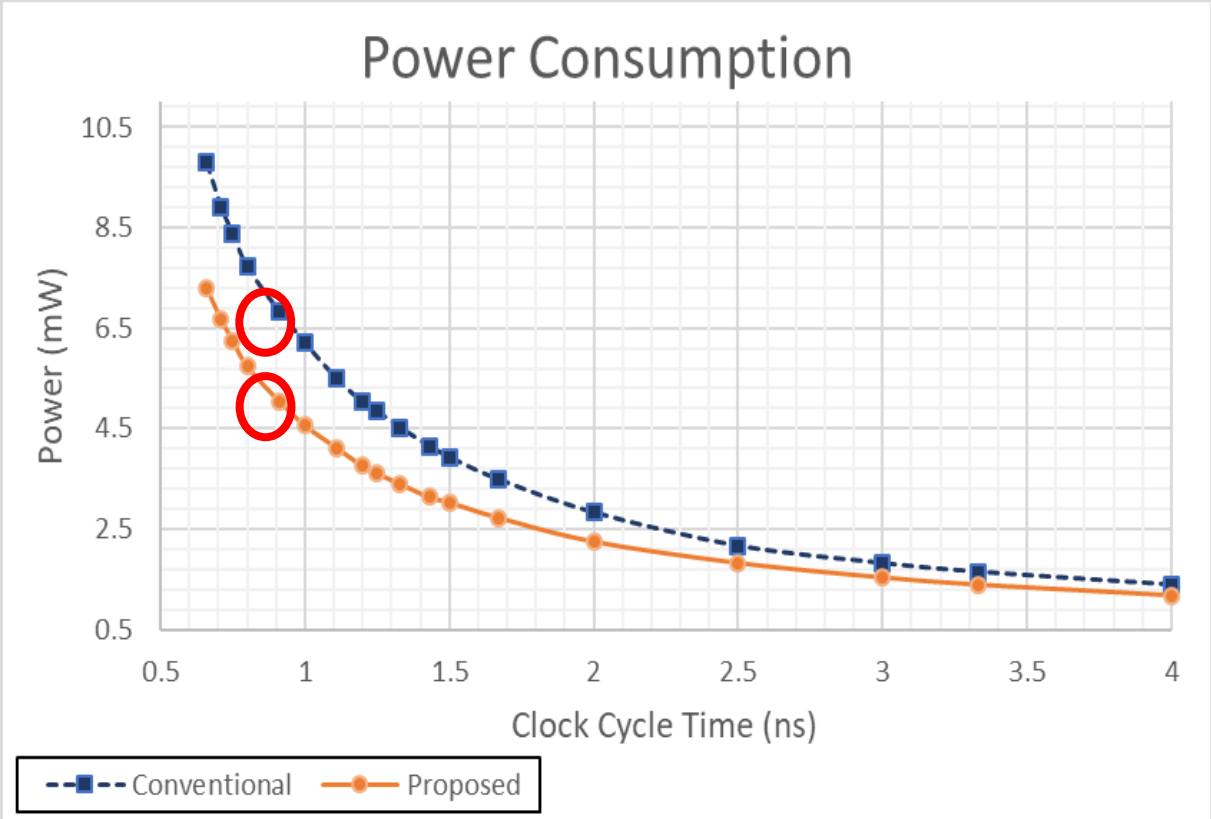
- Operating frequency up to 1.5 GHz
  - well-balanced pipeline
- Area reduction
  - 15.8% @ 1 GHz
    (20.8K ➔ 17.5K gates)
- Shorter cycle time
  - **better area reduction**
  - 18.5% @ 1.3 GHz
    (22.3K ➔ 18.2K gates)



- TSMC 40nm technology (0.9V, 25℃)
- Reported by Synopsys Design Compiler

Experimental Results (Power)

- Maximum power reduction
  - 26.5% @ 1 GHz
    (6.20mW ➔ 4.56mW)
- Shorter cycle time
  - **better power reduction**



- TSMC 40nm technology (0.9V, 25℃)
- Reported by Synopsys Design Compiler

Summary

- We proposed a coarse-grained highly-integrated convolver architecture
  - high-speed, area-efficient, low-power
- The depth-first compression scheme
  - global optimization across the multiplier boundary
  - no internal CPAs (10CPAs ➔ 1CPA)
  - 4 ➔ 3 well-balanced pipeline stages
- Experimental results show
  - area reduction 15.8% (@ 1GHz)
  - power reduction 26.5% (@ 1GHz)

Thank you