THE CHIPS TO SYSTEMS CONFERENCE
63
SHANGHAI JIAO TONG UNIVERSITY
SPONSORED BY:
CEDA
IEEE Council on Electronic Design Automation
sigda
acm

Adaptive Multiply-Accumulate Unit with Scalable Critical Path for Aggressive Voltage Underscaling DNN Accelerators

Tongjing Wu[1], Siting Liu[2], Tong Li[1], Hui Wang[1], Zhigang Mao[1], Jie Han[3], Honglan Jiang[1]

[1] School of Integrated Circuits, Shanghai Jiao Tong University, Shanghai, China

[2] School of Information Science and Technology, ShanghaiTech University, Shanghai, China

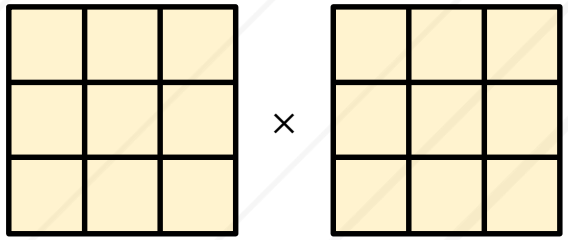
[3]Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada

Adaptive Multiply-Accumulate Unit with Scalable Critical Path for Aggressive Voltage Underscaling DNN Accelerators

Contents

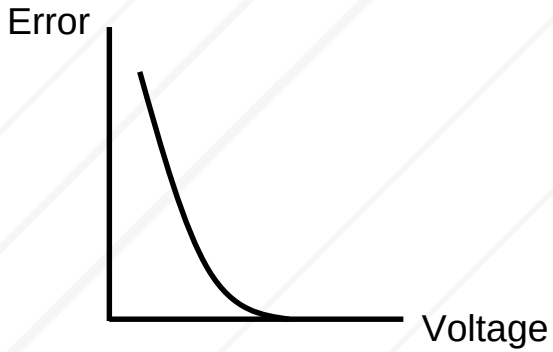
- Motivations
- ADA Design
- ADA-Plus Design
- Experiment Results
- Conclusions

Motivation



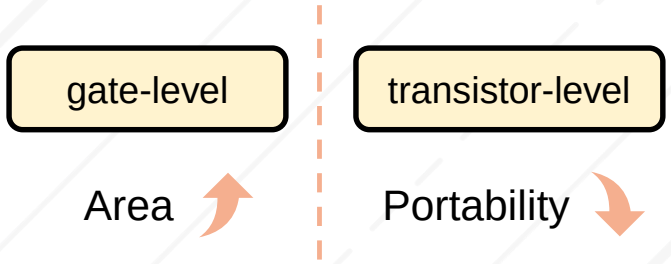
Matrix multiplication dominates DNN workloads

Efficient matrix multiplication is critical to the power-performance tradeoff in deep learning accelerators



Voltage underscaling is promising yet error-prone

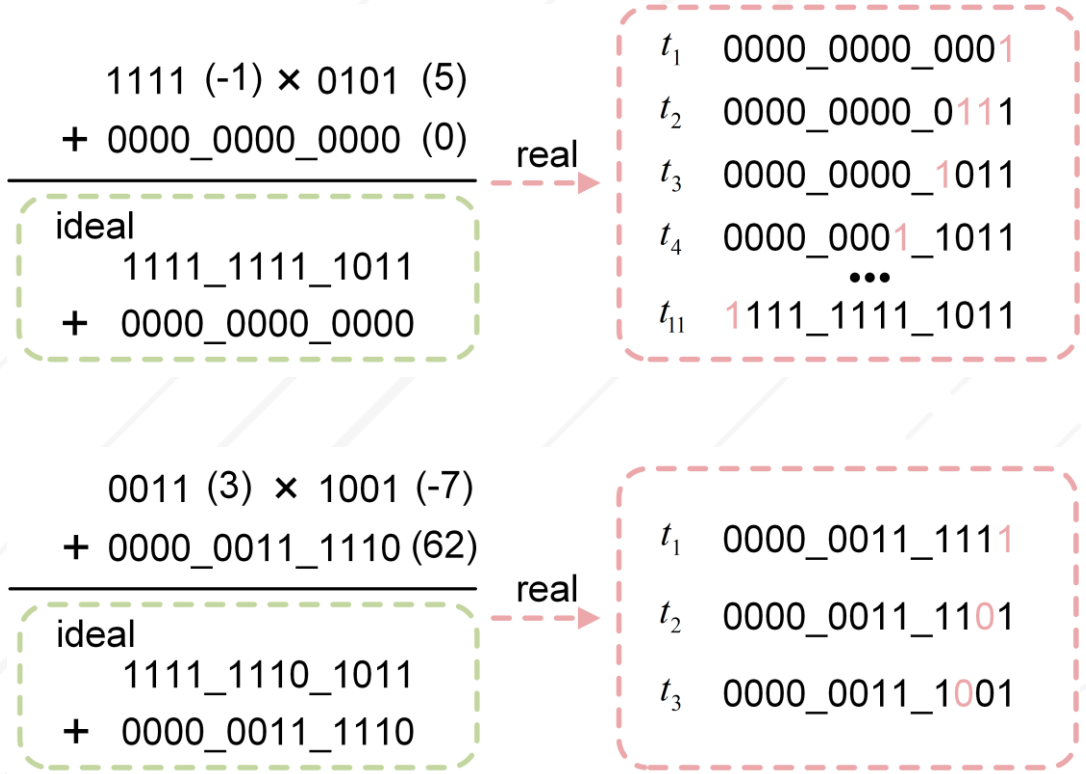
Voltage underscaling significantly reduces power consumption but introduces timing errors



Existing EDaC techniques face limitations

Prior error detection and correction (EDaC) approaches face a key trade-off: gate-level designs incur high area overhead, while transistor-level designs suffer from limited portability

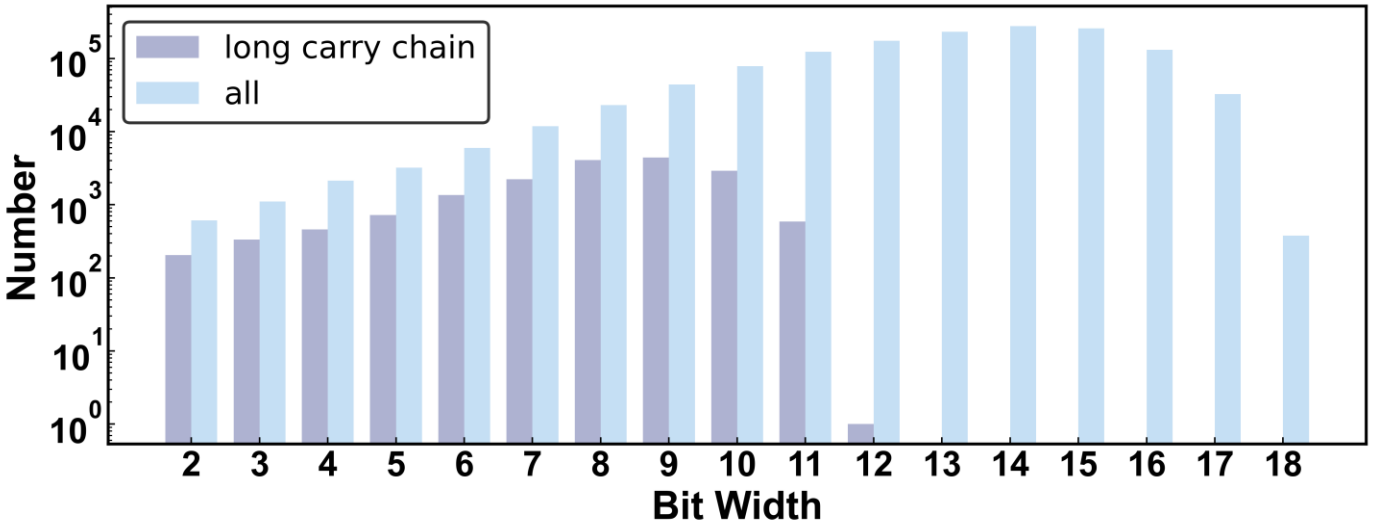
Motivation



4×4 multiplication + 12-bit addition synthesized MAC unit

- The actual carry behaviors deviate from the ideal cases
- **Reason**: the MAC netlist is flattened by the synthesis tool and the multiplier and adder are merged
- Logic Optimization + Voltage Underscaling ➔ Non-ideal MAC Carry Chain
- **Solution**: Reduce the bit width

Motivation



Distribution of the required bit widths for accumulation outputs in a convolution layer

- 1st Conv Layer
- 6-layer CNN (CIFAR-10)
- 16×16 systolic array

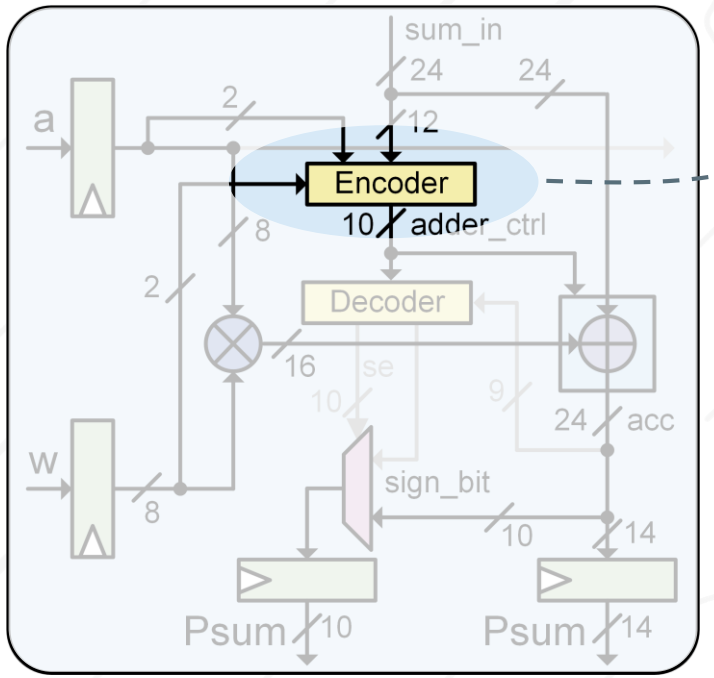
Dynamic bit-width adjustment is essential

- **Long Carry Chains**: Primarily occur at smaller bit widths (≤ 12 bits) due to sign inversions
- **Total Distribution**: The vast majority of MAC results require > 12 bits
- Statically fixing bit width to 12 bits (to avoid timing violations) leads to **massive overflow**

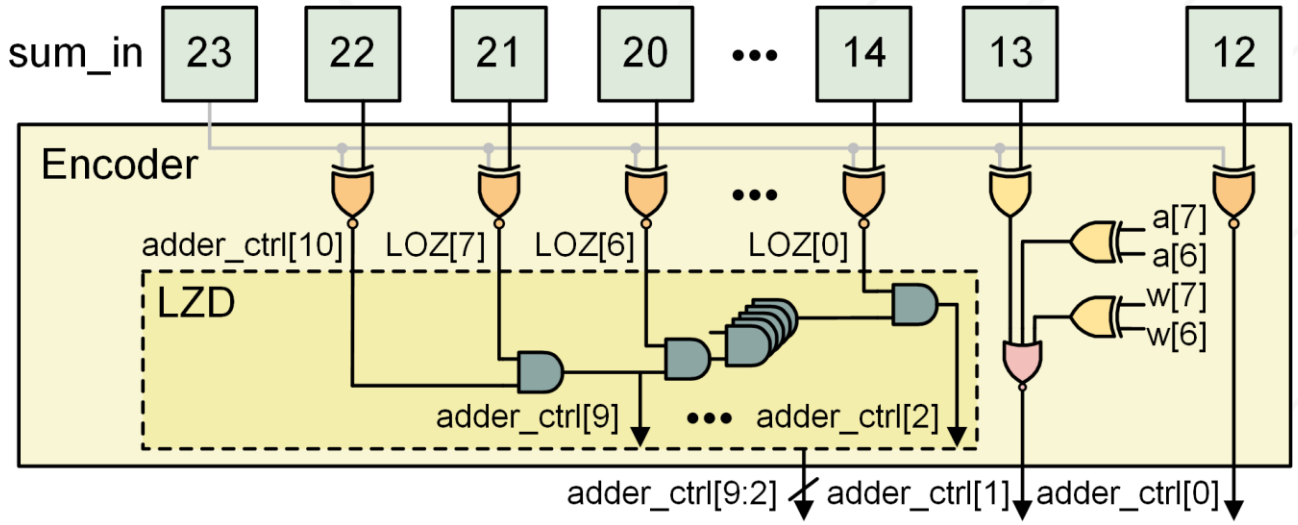
THE CHIPS TO SYSTEMS CONFERENCE
63
SPONSORED BY:
CEDA
IEEE Council on Electronic Design Automation
sigda
acm

ADA Design

Encoder

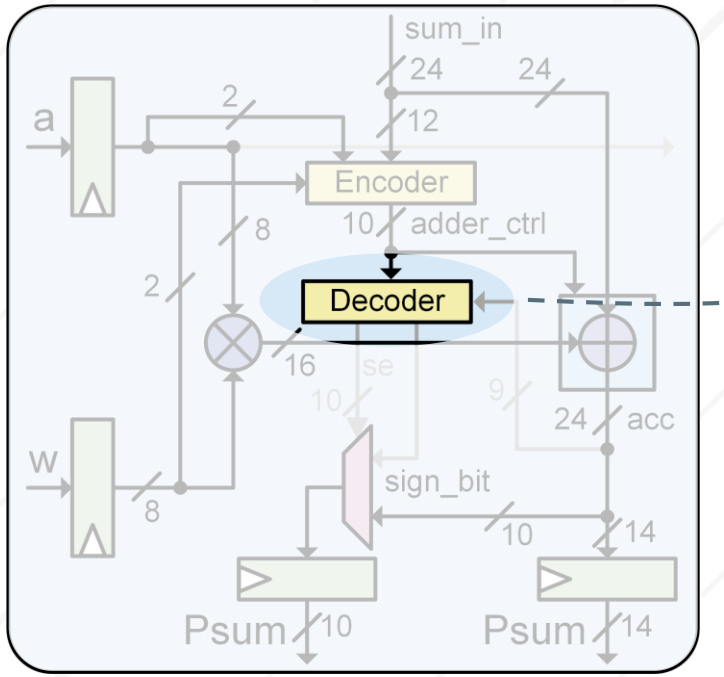


The overview of ADA

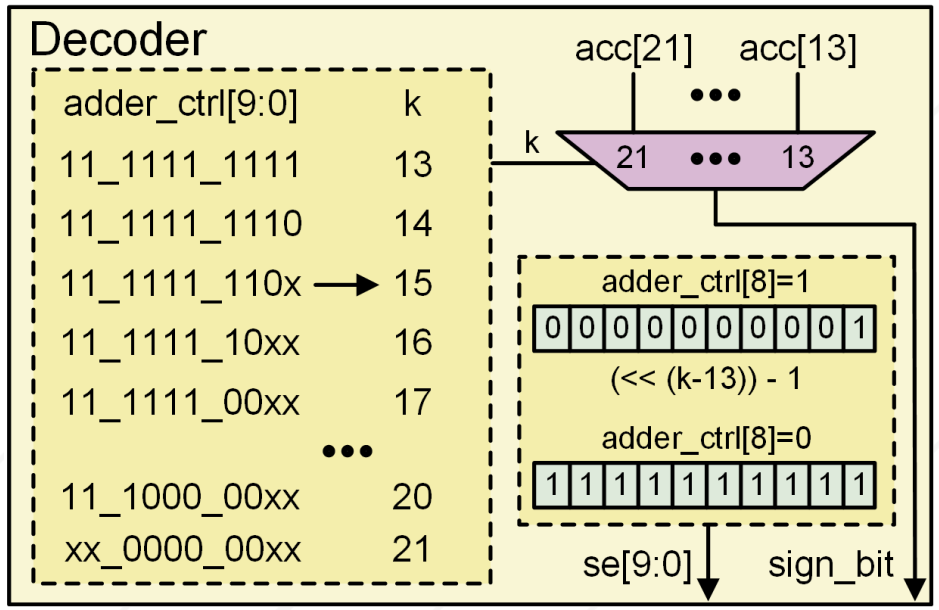


- Detects leading ones/zeros in sum_in to generate FA control signals (adder_ctrl)
- Performs bit-wise XNOR between the sign bit (sum_in[23]) and specific data bits (sum_in[22:12])
- Uses cascaded AND gates to propagate signals; a detected leading bit forces subsequent controls to 0

Decoder

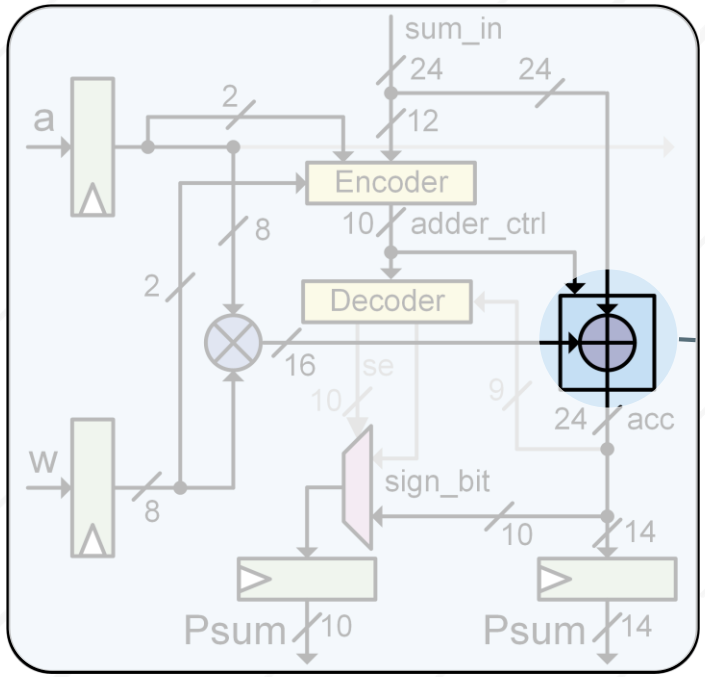


The overview of ADA

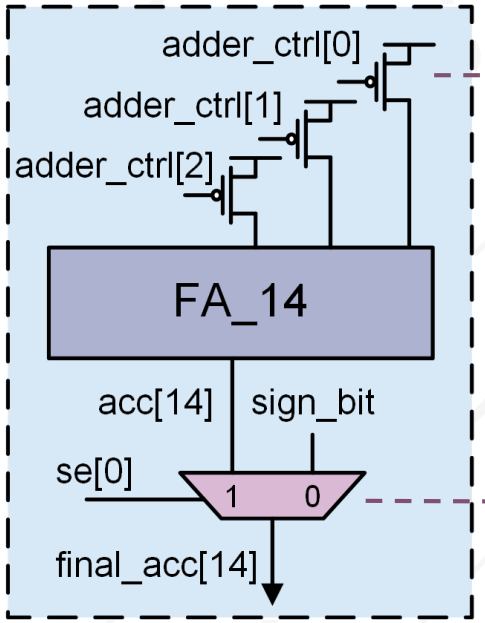


- Locate Boundary: Decodes adder_ctrl to find the highest active FA index (k)
- Extract Sign: Uses a MUX to select acc[k] as the dynamic sign_bit
- Mask & Bypass: Generates mask (se) to assign sign_bit to upper bits, bypassing disabled FAs to **shorten the critical path**

Adder control



The overview of ADA



Power gating

- FAs 0-13: Always ON.
- FAs 14+: Dynamically activated

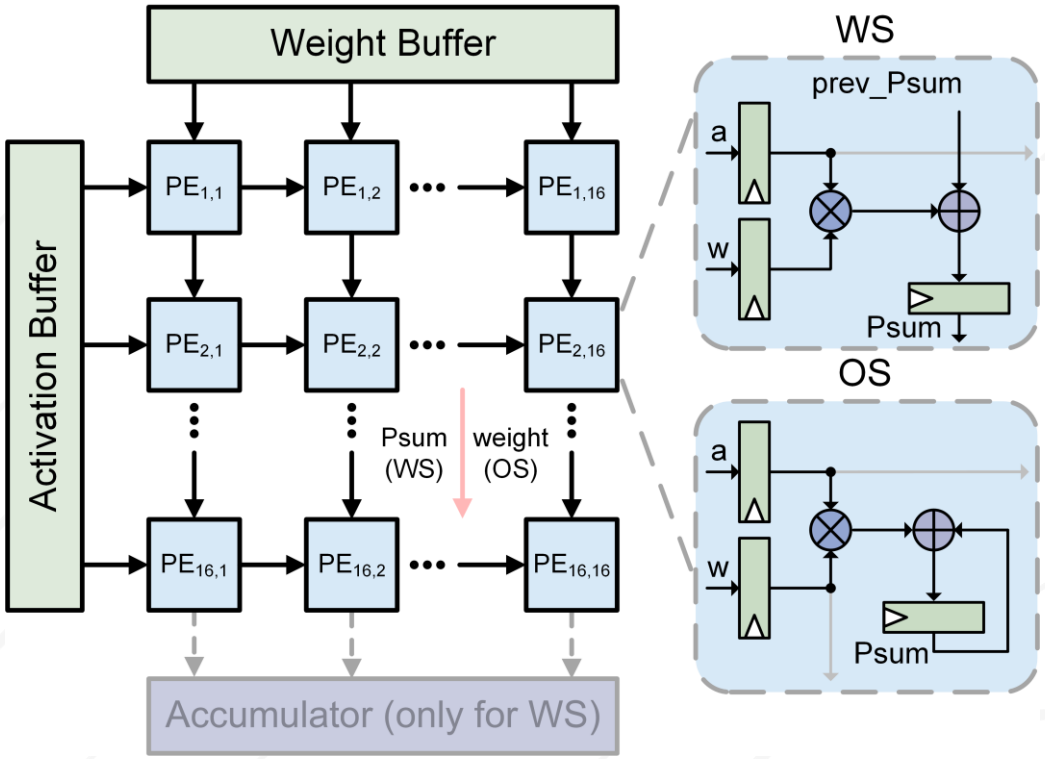
Delay bypassing

- Drives PMOS networks to power-gate FAs 14+
- MUX logic bypasses inactive FAs, directly feeding sign_bit to final_acc
- **Saves power** by turning off unnecessary FAs
- **Reduces delay** by directly assigning the sign_bit to inactive higher bits

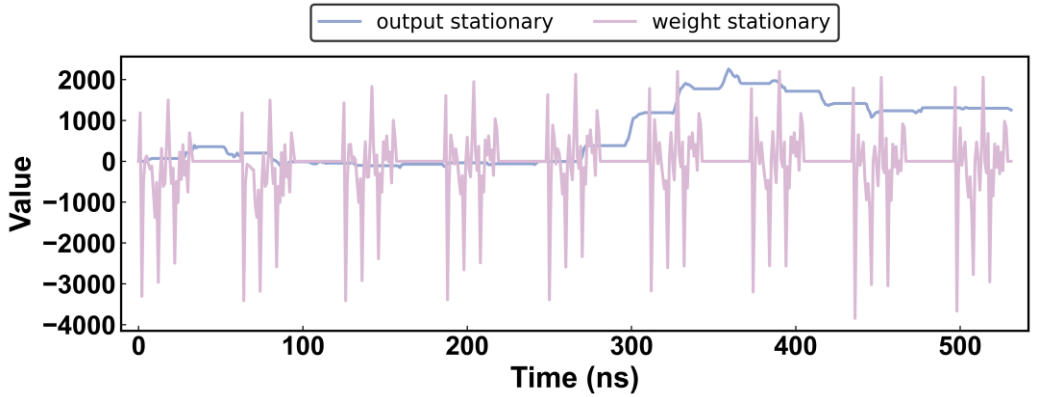
THE CHIPS TO SYSTEMS CONFERENCE
63
SPONSORED BY:
CEDA
IEEE Council on Electronic Design Automation
sigda
acm

ADA-Plus Design

Observation



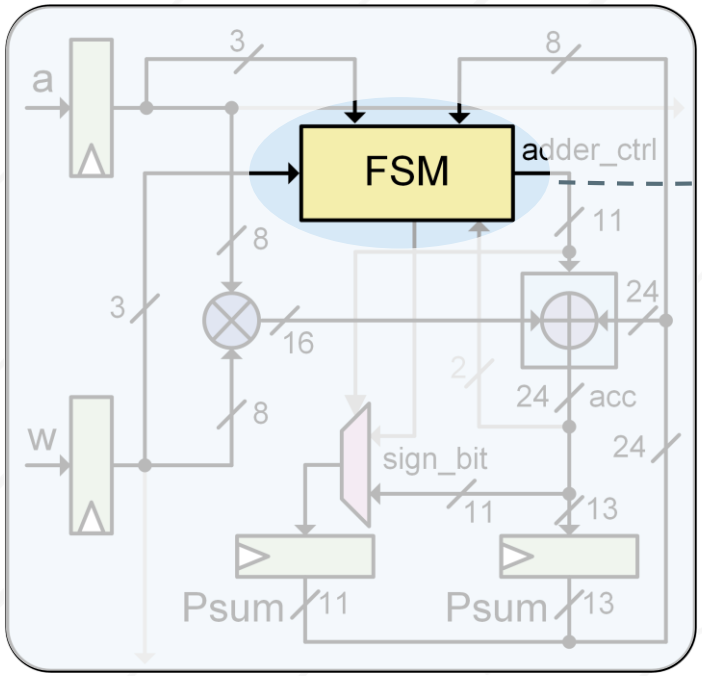
The schematic of systolic array and dataflow



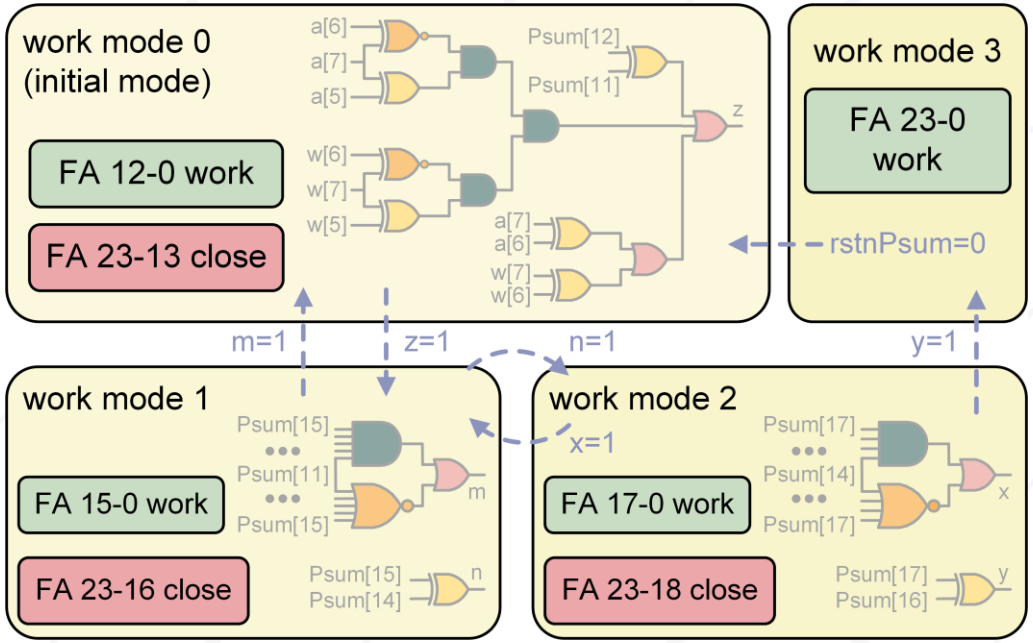
Trend of accumulation results in WS and OS PE

- **WS Mode**: High fluctuations. Requires expensive per-cycle LZD (ADA)
- **OS Mode:** Continuous, gradual transitions due to strong temporal correlation
- Proposed Design: ADA-Plus embeds a small **FSM** into the MAC to track and adjust bit-width gradually

FSM

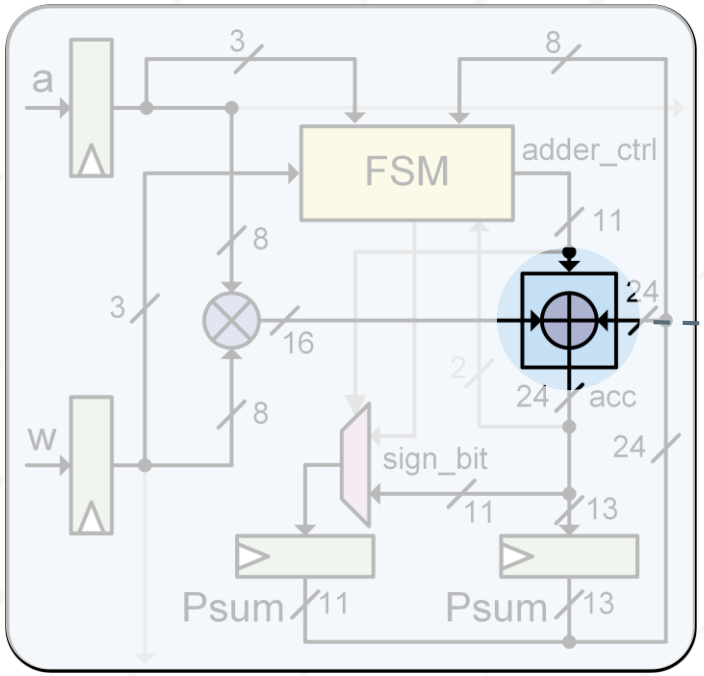


The overview of ADA-Plus

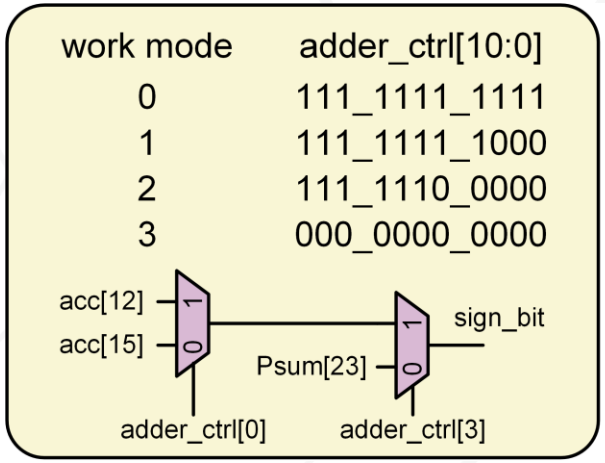


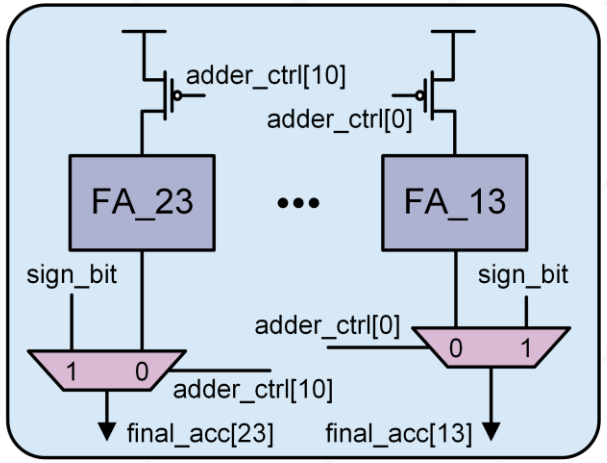
- **Core Logic**: Tracks Psum boundaries to scale FAs dynamically
- Upward Transitions: Triggers activate more FAs when Psum MSBs indicate overflow risks
- Downward Transitions: Triggers fall back to lower modes to shorten critical path when Psum margins are safe

Adder control



The overview of ADA-Plus



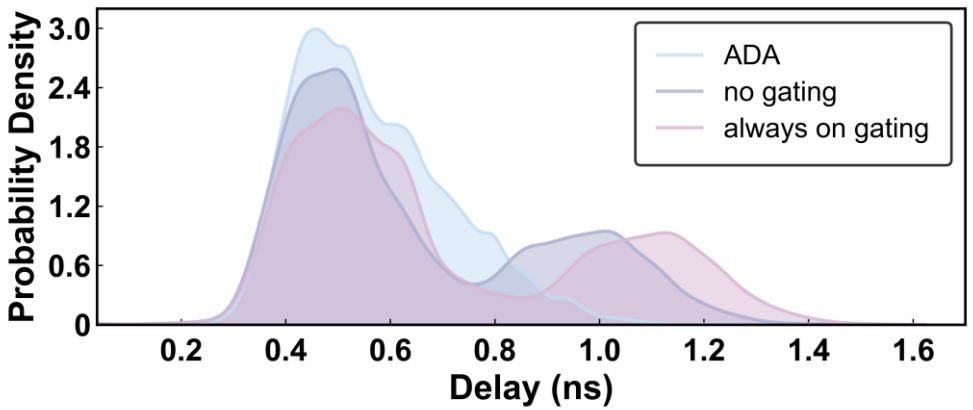


- Unified Control: adder_ctrl handles both FA power-gating and output MUXing
- Mode-Driven Sign: Smartly shifts the sign_bit source based on the active FSM state
- Dual Benefits: **Maximizes** energy savings (PMOS off) while **minimizing** critical path delay (MUX bypass)

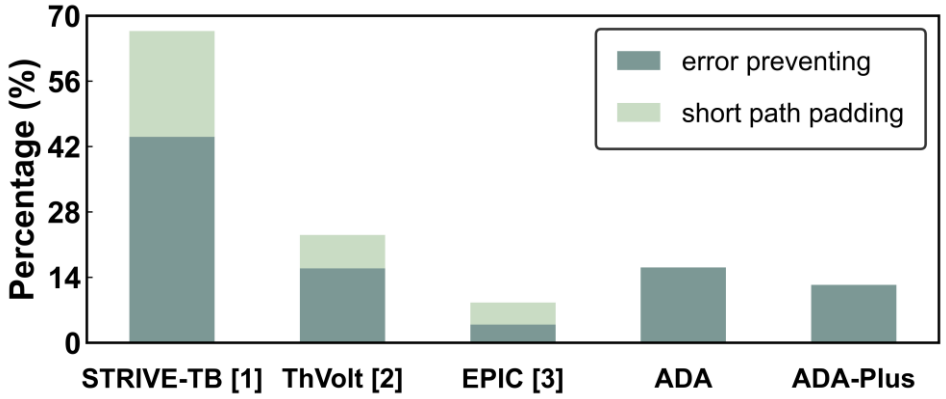
THE CHIPS TO SYSTEMS CONFERENCE
63
SPONSORED BY:
CEDA
IEEE Council on Electronic Design Automation
sigda
acm

Experiment Results

Delay & Area



KDE curves of MAC output delay



Area overhead of voltage underscaling schemes

- Negligible penalty: PMOS switches add **minimal** base delay
- Tail truncation: Extreme delays eliminated
- Result: Critical path significantly **shortened**

- Zero padding: No short-path buffers required
- ADA: **16.1%** overhead
- ADA-Plus: **12.4%** overhead

Area overhead **30%-75%**

Robustness

corner	temp	Power reduction (%)		
		ADA (OS)	ADA-Plus (OS)	ADA (WS)
tt	-40℃	19.59	18.24	16.55
	-25℃	19.35	17.82	15.46
	-125℃	25.64	25.32	21.94
ss	-40℃	17.68	16.35	14.97
	-25℃	20.99	19.74	17.73
	-125℃	27.02	26.05	**24.53**
ff	-40℃	19.81	16.66	15.15
	-25℃	16.43	15.59	10.98
	-125℃	**29.19**	**28.96**	24.48

Minimum operating voltage and power reduction results across PVT corners for the proposed MAC-based systolic arrays

Setup

- Technology Node: 28 nm CMOS standard cell library (Nominal 0.9 V)
- Baseline MAC: 8-bit inputs and 24-bit accumulation
- Workload: 8×8 systolic array running a $784 \times 16 \times 16 \times 10$ MLP on the MNIST dataset
- Method: VCS_XA AMS simulation

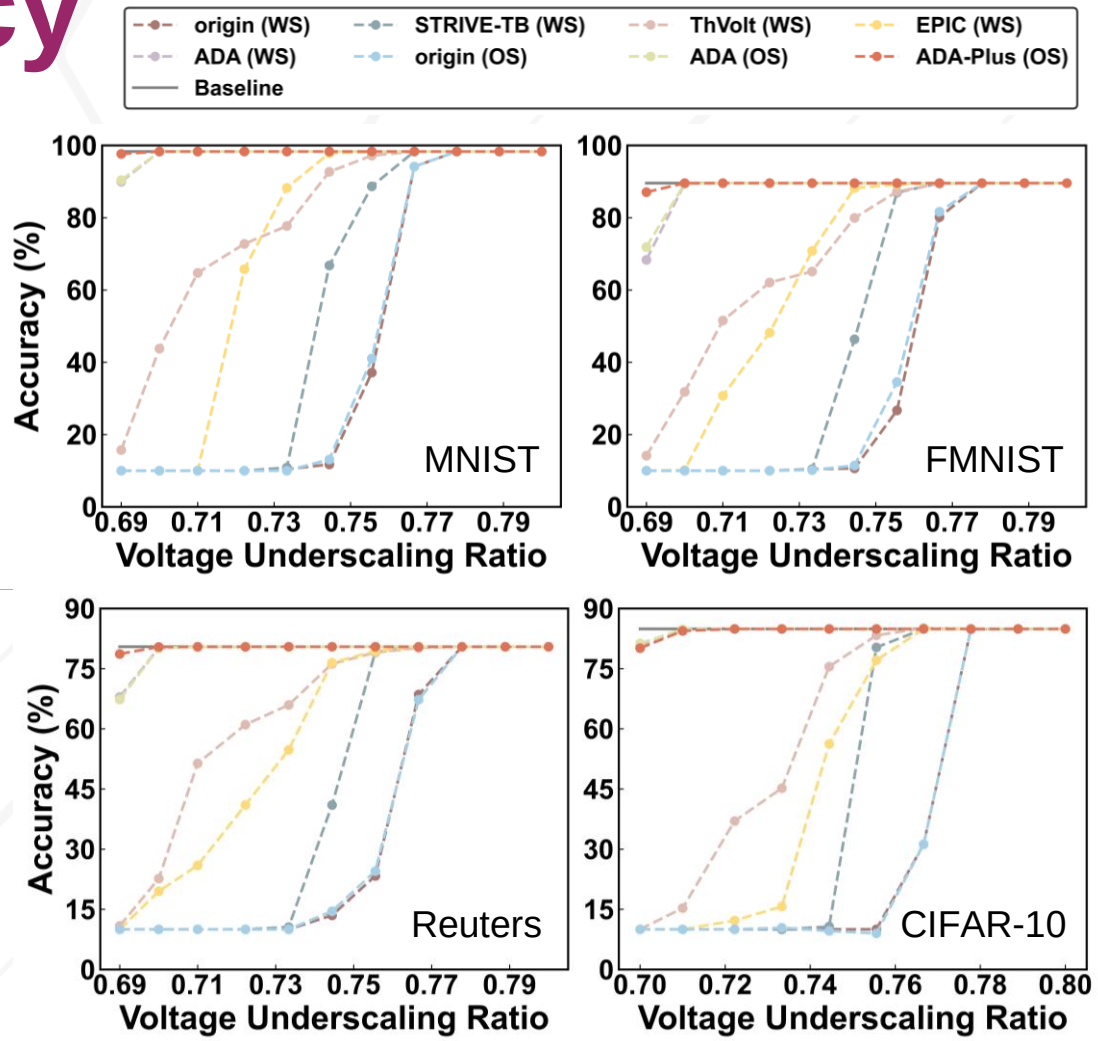
Result

- ADA: Up to **29.19%** power reduction
- ADA-Plus: Maintains accuracy at a lower voltage and achieves up to **28.96%** power reduction

Inference Accuracy

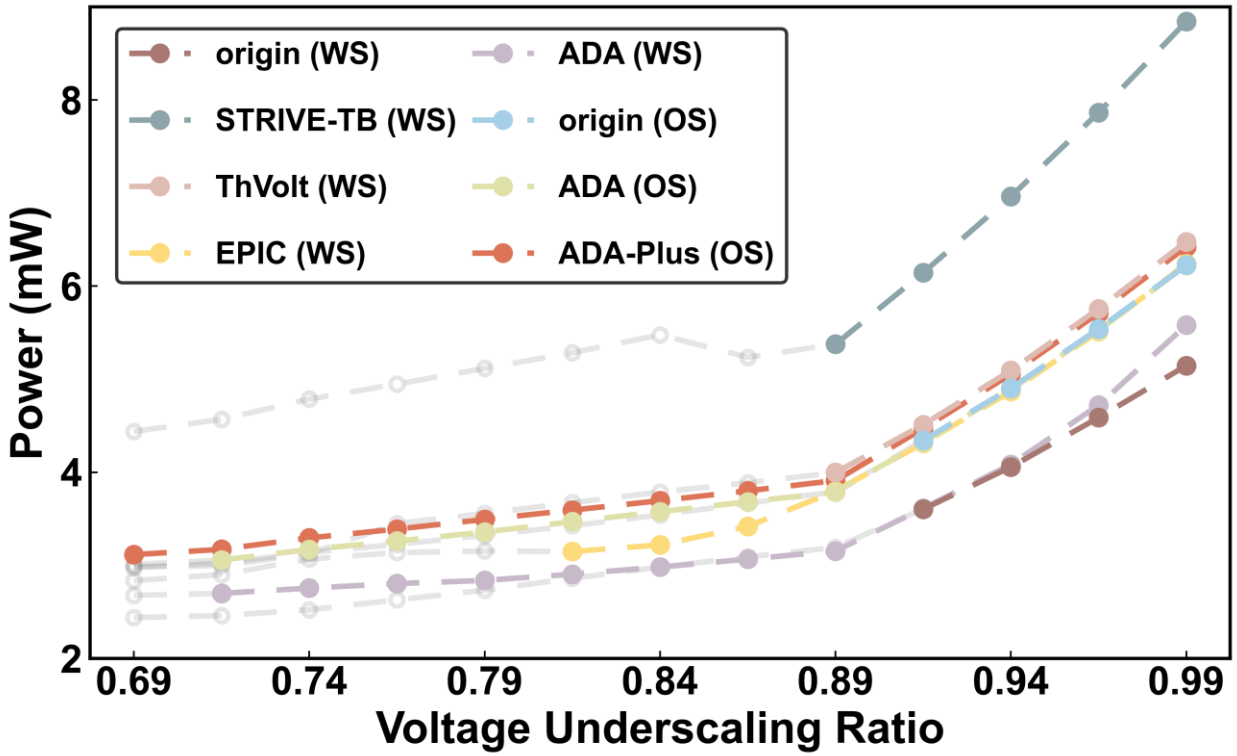
Benchmarks			Acc (%)
Dataset	Architectures		
MNIST	MLP	$784 \times 256 \times 256 \times 10$	98.30
FMNIST		$784 \times 256 \times 256 \times 10$	89.56
Reuters		$2048 \times 256 \times 256 \times 46$	80.45
CIFAR-10	CNN	CONV: (32, 32, 32) × (32, 32, 32) CONV: (16, 16, 64) × (16, 16, 64) CONV: (8, 8, 128) × (8, 8, 128) FC: 2048×10	84.9

- 16×16 systolic array, 1GHz
- Error-Free: Down to 0.71 (MLP) and 0.72 (CNN)
- Ultimate Limits: Functional down to **0.69** (MLP) / **0.70** (CNN)



Inference accuracy of systolic array on DNN benchmarks with voltage scaling

Power evaluation



Power of systolic arrays consisting of the considered MAC designs under different voltage underscaling ratios

- Massive Total Savings: Up to **~50.9%** power reduction compared to the original array at nominal voltage
- vs. Baseline at V_{min}: ADA-based designs save **15.4% ~ 19.3%** power
- vs. Prior EDaC: Consumes up to **48.4%** less power than prior EDaC baselines

SPONSORED BY:
CEDA IEEE Council on Electronic Design Automation
sigda acm

Conclusions

Conclusions

- **ADA framework**: An adaptive circuit dynamically adjusting MAC critical paths based on input bit-widths under voltage underscaling
- **Low area overhead**: Implemented purely with standard cells, adding only **16.1%** area overhead to the baseline
- **Robust inference**: Sustains error-free DNN operations down to **0.71** voltage ratios, achieving **50.9%** power savings
- **ADA-Plus optimization**: Leverages output stationary dataflow to predict required bit-widths, reducing ADA's area by **23%**
- **State-of-the-art efficiency**: Achieves lowest voltage scaling limits and **14.2%–48.4%** power reduction compared to prior designs

References

[1] Noel Daniel Gundi, Zinnia Muntaha Mowri, Andrew Chamberlin, Sanghamitra Roy, and Koushik Chakraborty. 2023. STRIVE: Enabling Choke Point Detection and Timing Error Resilience in a Low-Power Tensor Processing Unit. In 2023 60th ACM/IEEE Design Automation Conference (DAC). 1–6. doi:10.1109/DAC56929.2023.10247879.

[2] Jeff Zhang, Kartheek Rangineni, Zahra Ghodsi, and Siddharth Garg. 2018. Thundervolt: enabling aggressive voltage underscaling and timing error resilience for energy efficient deep learning accelerators. In Proceedings of the 55th Annual Design Automation Conference. 1–6.

[3] Tongjing Wu, Xiaolu Hu, Tong Li, Siting Liu, Hui Wang, Weifeng He, Zhigang Mao, and Honglan Jiang. 2025. EPIC: Error PredIction and Correction for Power-Efficient Voltage Underscaling Multiply-Accumulate Unit. In 2025 62nd ACM/IEEE Design Automation Conference (DAC). 1–7. doi:10.1109/DAC63849.2025.11132540.

SPONSORED BY:

Thanks for listening