國立中山大學 工學院 聯合專題競賽與展示
COLLEGE OF ENGINEERING, NSYSU

積體電路設計研究所
Institute of Integrated Circuit Design

SUSTAINABLE DEVELOPMENT GOALS

9 產業、創新與基礎設施

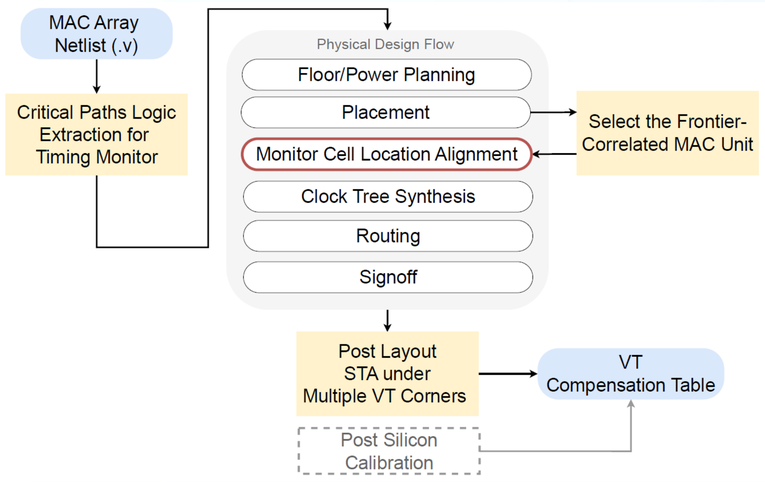
Monitor-Like Efficiency with Detector-Level Accuracy: Frontier-Aligned Timing Monitor for AI Accelerators

Advisor: Prof. Tong-Yu Hsieh **Student: Wei-Ji Chao, Tsung-Chun Chen, Chu-Cheng Chen**

Abstract

Systolic-array AI accelerators operating near threshold voltage face significant timing reliability challenges due to increased PVT sensitivity. While Razor flip-flops offer accurate bit-level detection, their area overhead limits scalability. Existing timing monitors are more efficient but lack granularity and adaptability. This work presents a frontier-aligned timing monitor that enables low-overhead, bit-level visibility. By analyzing post-layout delays in a 7nm systolic array, we identify a MAC unit highly correlated with the global bit-wise delay frontier. A co-located monitor path with tunable delay buffers enables PVT-aware calibration and precise alignment. Experimental results show an average delay error of 3.1% and area overhead as low as 0.1% in large arrays. The proposed design supports scalable, energy-efficient runtime approximation and adaptive voltage/frequency scaling (AVFS), offering a practical solution for fine-grained timing management in modern AI accelerators. This work has been accepted for oral presentation at the International Test Conference in Japan in December 2025.

Design Framework



Step 1: Extract Critical Paths of MAC Output Endpoints
Step 2: Select and Align Frontier-Correlated MAC Unit
Step 3: Post-Layout and Post-Silicon Calibration

Proposed Timing Monitor Architecture

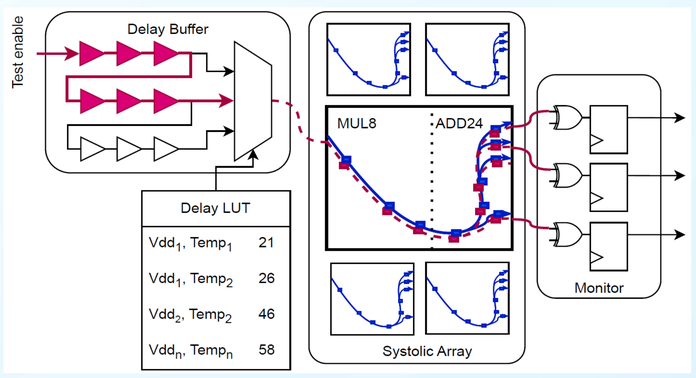


Figure 4. Proposed frontier-aligned timing monitor with VT-tunable buffer for delay compensation.

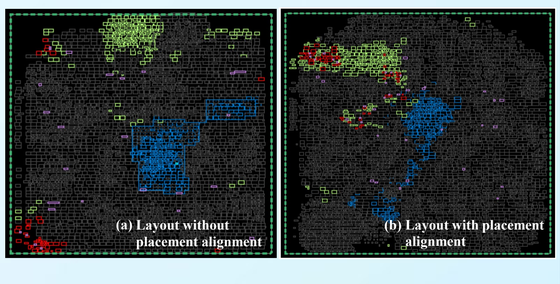


Figure 5. (a)–(b) layout comparison (green: target MAC; red: monitor).

Monitoring Mechanism Validation

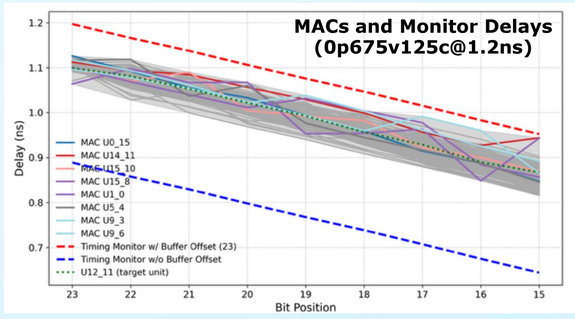


Figure 6. MACs and Monitor Delays. The red dashed line highlights the monitor delay with 23 buffers offset.

		4×4	8×8	16×16
Avg. Error	0.675V / -40°C	2.7%	3.3%	3.4%
	0.675V / 125°C	2.8%	3.1%	3.6%
	0.825V / -40°C	3.7%	3.3%	3.6%
	0.825V / 125°C	4.5%	3.2%	2.0%
	All-Corners	**3.4%**	**3.2%**	**3.1%**
Max. Error	0.675V / -40°C	7.2%	5.2%	6.5%
	0.675V / 125°C	4.9%	4.7%	6.3%
	0.825V / -40°C	5.2%	5.7%	5.7%
	0.825V / 125°C	8.3%	5.7%	3.8%
	All-Corners	**8.3%**	**5.7%**	**6.5%**
Area Overhead		**1.6%**	**0.4%**	**0.1%**

Table 1. Timing Error Rates and Area Overhead Across Voltage–Temperature Corners and Array Size

Experimental Setup

- **Design: 4×4 systolic array with 8-bit multiplier and 24-bit adder in each MAC**
- **Technology Node: Commercial 7-nm FinFET technology**
- **Temperature: -40 ˚C & 125 ˚C**
- **Voltage: 0.675 V & 0.825 V**
- **Clock Frequency: 833 MHz**

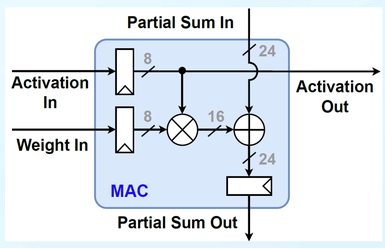


Figure 2. Architecture of a MAC unit

Delay Variation Analysis

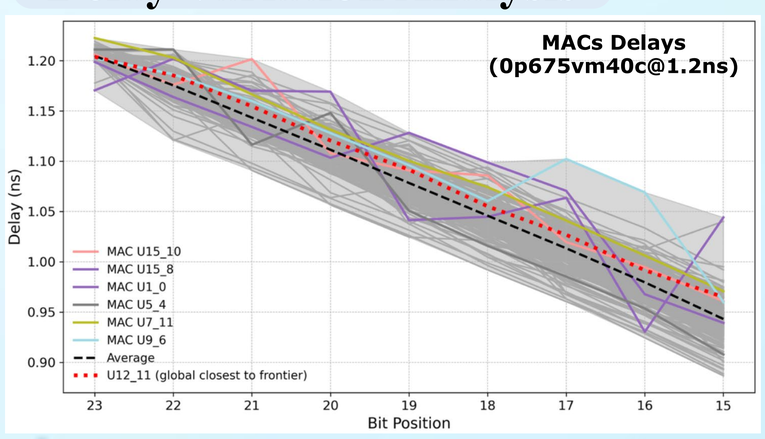


Figure 3. Output delay distribution across MAC units under aggressive timing (1.2 ns synthesis constraint).

Comparison

Category	Beyond Margin [1]	EFFORT [2]	Aging Comp. [3]	DSC-TRCP [4]	ISCPR [5]	Ours
Type	In-situ detector		Timing monitor			
Technology Node	28-nm	15-nm	45-nm	28-nm	22-nm	7-nm
Clock Frequency	42-MHz	67.5-MHz	-	420-MHz	300-MHz	1-GHz
Target Circuit	Systolic array	Systolic array	Single MAC unit and DCT/IDCT	Systolic array	54b×54b multiplier	Systolic array
Validation Stage	Post-silicon	Post-syn	Post-syn	Post-silicon	Post-silicon	Post-layout
Timing Granularity	Path-level	Bit-level	Bit-level	Path-level	Path-level	Bit-level
Area Overhead	1.31% of chip	5% of TPU	0.8% of DCT/IDCT	0.65% of chip	Low	1.6% of 4×4SA
Bit-Level Adaptivity	✗	✓	✓	✗	✗	✓
PVT-Aware Mechanism	Built-in tolerance		Guardband	Tunable delay		
Post-Silicon Calibration			Unable	Able		
Accuracy of Monitor (Target-Ideal)/Ideal	Achieve Negative Margin with Protection Mechanism			Max. 2.3% (Delay Margin)	Max. 4.5% (Voltage Margin)	Max. 8.3% Avg. 3.4% (Delay Margin)

Table 2. Comparison with State-of-the-Art Timing Error Detection and Monitoring Techniques

編號填寫處 F07