스토캐스틱 컴퓨팅 기반의 임베디드용 AI 프로세서

제 23회 대한민국 반도체 설계대전

김정은, 고광현, 정영우, 정원식

Seoul National University of Science & Technology

Electronic Engineering

SEOULTECH

CONTENTS

디자인 소개

스토캐스틱 컴퓨팅 기반의 임베디드용 AI 프로세서

면적 ↓ 파워 ↓ Fault-tolerance ↑ 정확도 ↑

창의성
확률적인 연산 회로를 프로세서에 적용

사업성
인공지능 가속기를 포함한 **내고장형의 경량** 프로세서를 on-chip으로 구현

기술성
높은 정확도의 스토캐스틱 연산기와 AI 가속기 및 페리페럴을 구현

완성도
삼성 28nm 공정을 통해 **칩 레이아웃 검증 및 제작 완료**

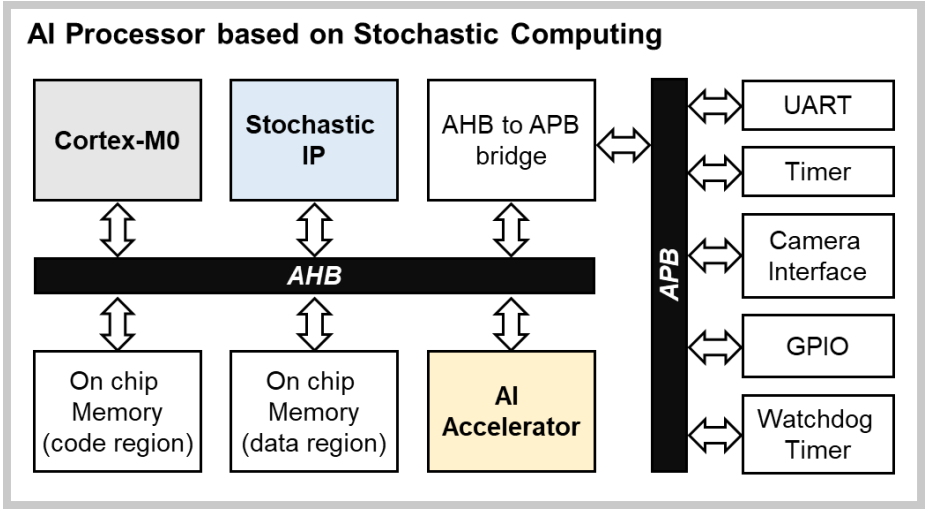
기존의 연산 회로는 복잡하고 오류에 취약
→ **확률적인 연산 회로를 이용하여 해결**

디자인 소개

- **Stochastic Computing** (SC) - 확률론적인
  - ✓일반적인 이진수를 [0,1] 범위의 **확률** 값으로 인코딩
    1. 랜덤 수와의 **크기 비교** ex) $\mathbf{100_{(2)}}$ vs 7, 5, 2, 6, 1, 0, 4, 3
    2. 숫자 **'1'의 비율**로 근사 → 1의 개수 / 시퀀스의 총 길이

이진수 $\mathbf{100_{(2)}}$ → 시퀀스 **00101101** → 확률 **4/8**

- **설계 구조 및 목표**



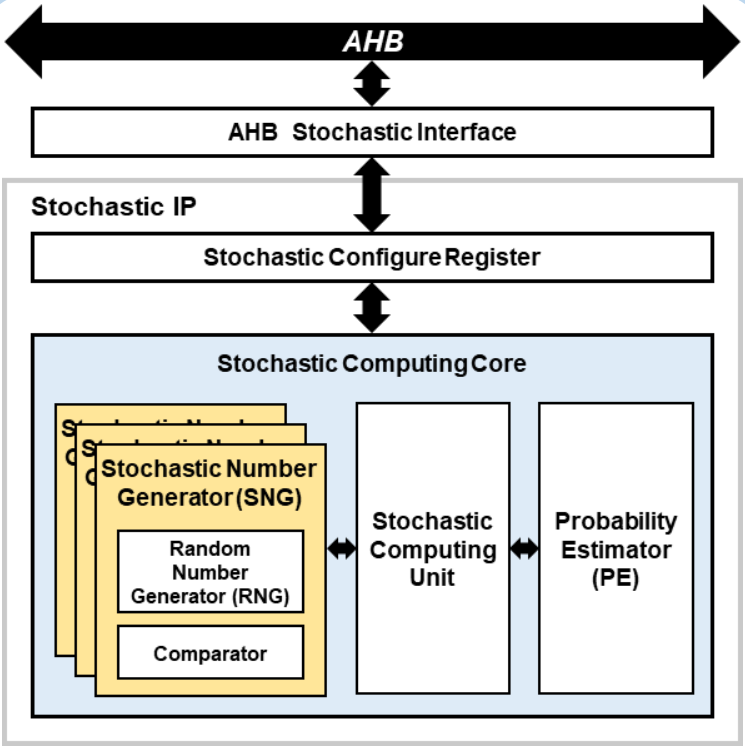
[스토캐스틱 컴퓨팅 기반 AI 프로세서의 전체 구조도]

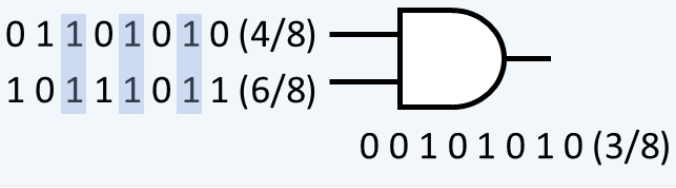
항목	설계 목표
코어 사양	Arm 사의 Cortex-M 시리즈
동작 주파수	50MHz
메모리	코드 영역: 16KB, 데이터 영역: 128KB
인터페이스	AHB, APB
스토캐스틱 IP	정확도 90% 이상
인공지능 가속기	**k-NN 알고리즘** 적용
페리페럴	카메라 I/F, 시리얼 I/F, GPIO, 타이머

[초기 설계 목표 및 사양]

성능이득 분석(1) : 면적 & 파워

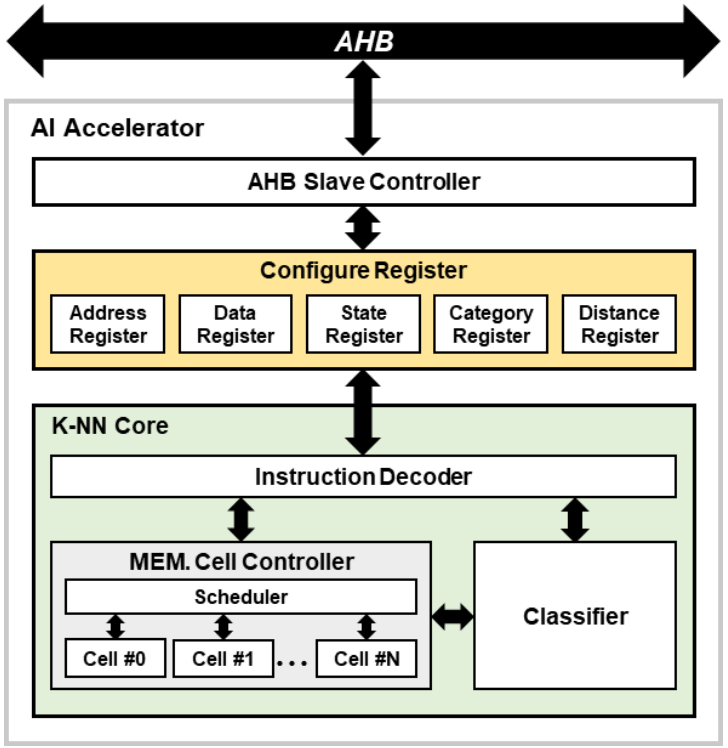
Stochastic IP

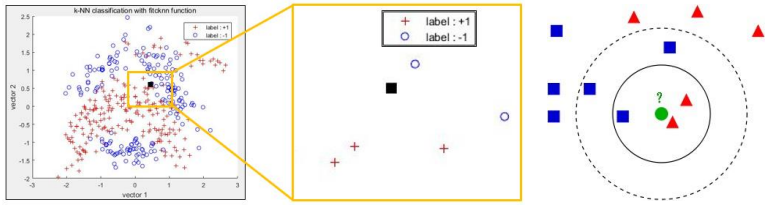




[두 사건이 동시에 '1'일 확률 = P1×P2]

AI Accelerator

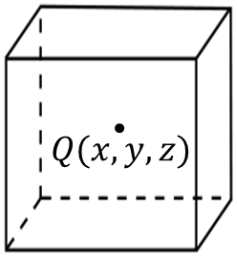




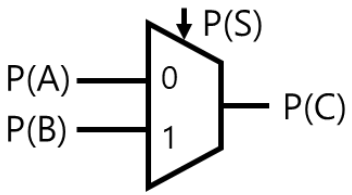
[k-NN 알고리즘: Manhattan distance]

성능이득 분석(1) : 면적 & 파워

$$
\begin{aligned}
Q = {} & xyzv_0 + xyzv_3 + xyzv_5 \\
& + xyzv_6 + yzv_1 + xzv_2 \\
& + xyv_4 + xzv_7 + yzv_7 \\
& + xyv_7 + xv_6 + yv_5 + zv_3 \\
& + v_7 - xyzv_1 - xyzv_2 \\
& - xyzv_4 - xyzv_7 - xzv_3 \\
& - yzv_3 - xyv_5 - yzv_5 \\
& - xyv_6 - xzv_6 - xv_7 \\
& - yv_7 - zv_7
\end{aligned}
$$

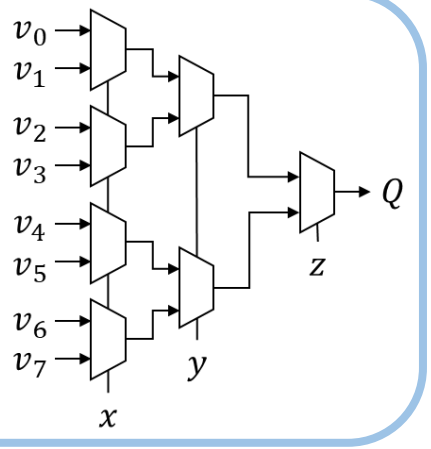


[Tri-linear interpolation 연산식[1]]

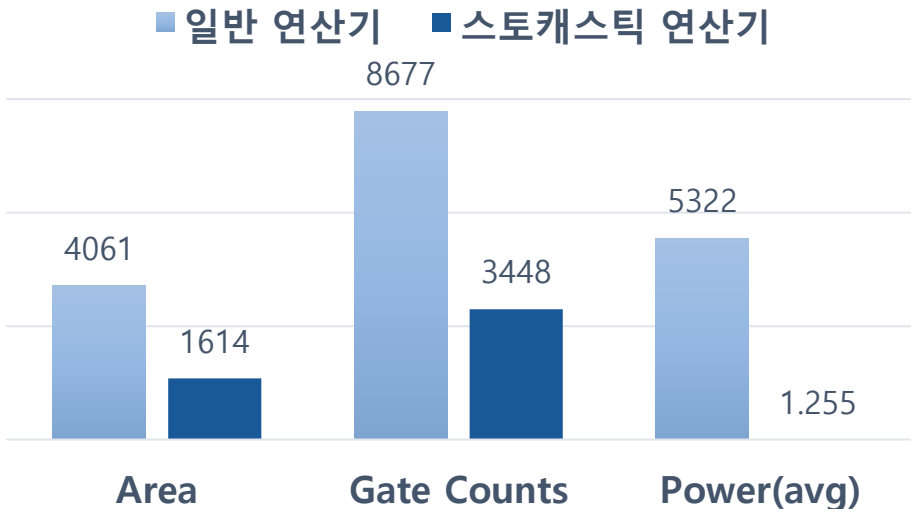


P(C) = (P(A)×(1-P(S))
+(P(B)×P(S))

∴P(S)=0.5인 경우,
P(C)=0.5(P(A)+P(B))



[SC 덧셈기 기반의 Tri-linear interpolation 연산기[2]]



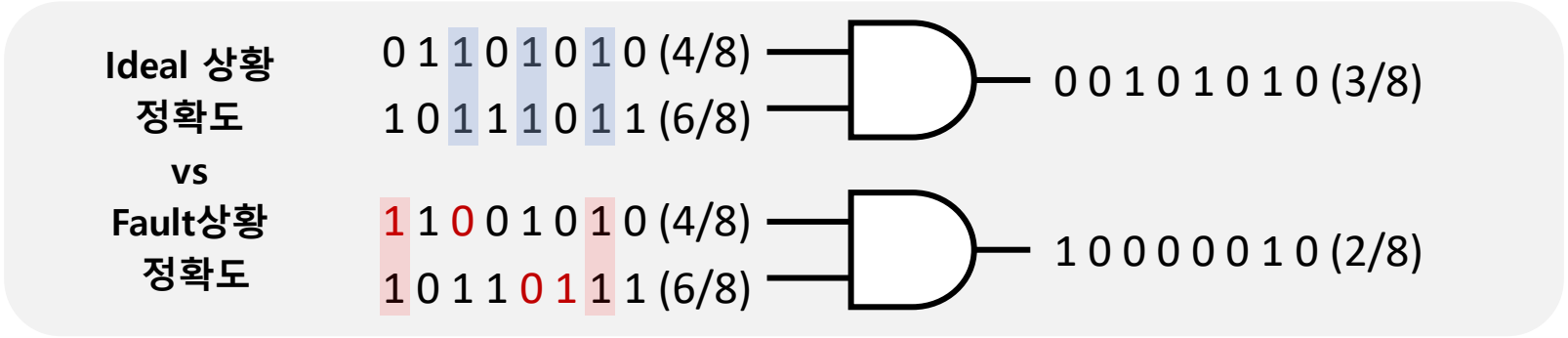
* 8bit 연산 기준
*Samsung 28nm, Design Compiler

		일반 연산기	스토캐스틱 연산기
Latency		1cycle	32cycles (8배 가속)
Area		$4{,}061\mu m^2$	$1{,}614\mu m^2$
Gate counts		8,677	3,448
power	Max	$5.3956mW$	$1.2243 \times 10^{-3}mW$
	Min	$1.0172 \times 10^{-3}mW$	$8.4204 \times 10^{-7}mW$
	avg	$5.322 \times 10^{-3}mW$	$1.255 \times 10^{-6}mW$

[1]Y. Shicai et.al "An high efficient and speed algorithm of Ray Casting in volume rendering," ICCE, 2011, pp. 1027-1030.
[2]Stochastic computing 기술동향 칼럼 from IDEC

성능이득 분석(2) : Fault-tolerance

- 시퀀스의 전체 길이 중 **'1'의 비율**을 컴퓨팅에 사용
  ex) (1,0,0,0) = (0,1,0,0) = (0,1,0,0,0,1,0,0) = 1/4 (시퀀스의 길이 길수록 fault에 강함)
- 시퀀스의 각 비트는 **동일한 weight**를 가짐
  ex) 3/256 → 4/256 or 2/256 vs 01000001 → 11000001



- **Fault-tolerance 분석**

디자인 결과

- **정확도 분석**: 이미지 엣지 검출 연산(Sobel 필터 연산)

P_1	P_2	P_3
P_4	P_5	P_6
P_7	P_8	P_9

(a) Sobel 필터 마스크

(b) 원본 이미지

$$|G| = |(P_1 + 2P_2 + P_3) - (P_7 + 2P_8 + 2P_9)| + |(P_3 + 2P_6 + P_9) - (P_1 + 2P_4 + P_7)|$$

(c) C코딩 필터 연산

(d) 스토캐스틱 필터 연산

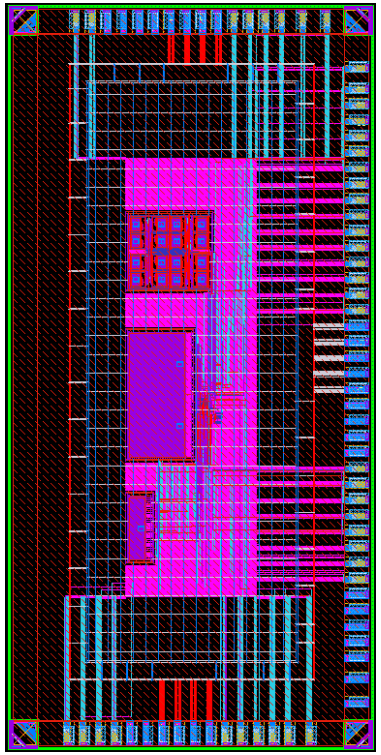
단위 (%)	(c)	(d)
평균 정확도	100%(가정)	92.4%

항목	설계 목표	목표 달성
코어 사양	Arm사의 Cortex-M 시리즈	O
동작 주파수	50MHz	O
메모리	코드 영역: 16KB / 데이터 영역: 128KB	O
인터페이스	AHB, APB	O
스토캐스틱 IP	정확도 92.4%	O
인공지능 가속기	k-NN (k-Nearest Neighbor) 알고리즘 적용	O
페리페럴	카메라 I/F, 시리얼 I/F, GPIO, 타이머	O
치수 (설계 면적)	2mm × 4mm	O

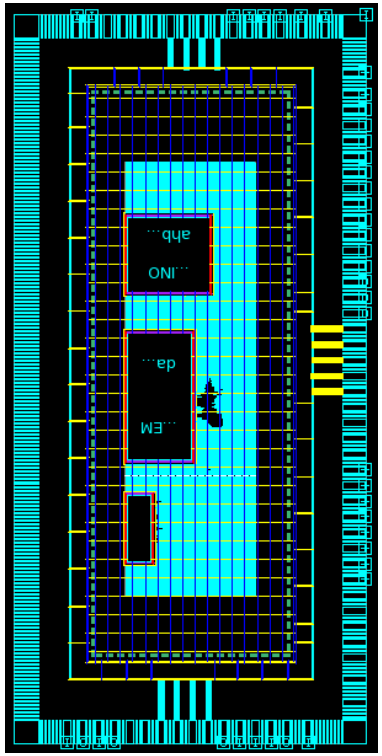
[디자인 결과 및 목표 달성 여부]

디자인 결과

- 레이아웃 디자인

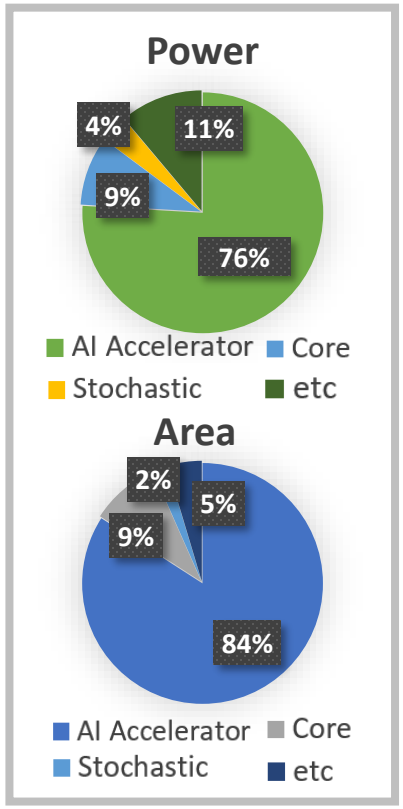


[칩 레이아웃(Virtuoso)]



[칩 레이아웃(ICC)]

항목	칩 사양
공정	Samsung 28nm CMOS
파워	Core: 1.0V
	IO: 1.8V
동작 주파수	50MHz
Gate Counts	1062K @ 50MHz
동작 사양	동작 온도:-40℃~125℃
	코어 전압: 0.9V~1.1V
메모리	code region: 16KB
	data region: 128KB
	AI Accelerator: 1KB
치수	2mm × 4mm

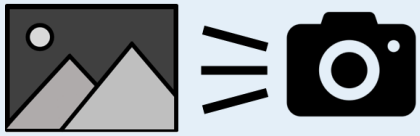


[프로세서 하위 모듈
면적 및 소비전력 비율]

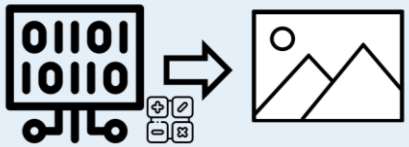
내고장형, 경량 회로 → Neural Networks, Image Processing과 같이
복잡하고 **많은 연산량**이 요구되는 응용 시스템에 적용 예상

데모 영상

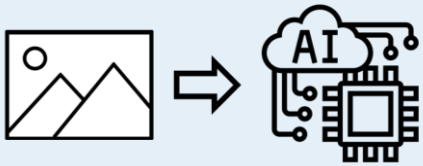
- **데모 순서**

1. 

카메라 모듈을 통한
촬영 및 **이미지 데이터 저장**

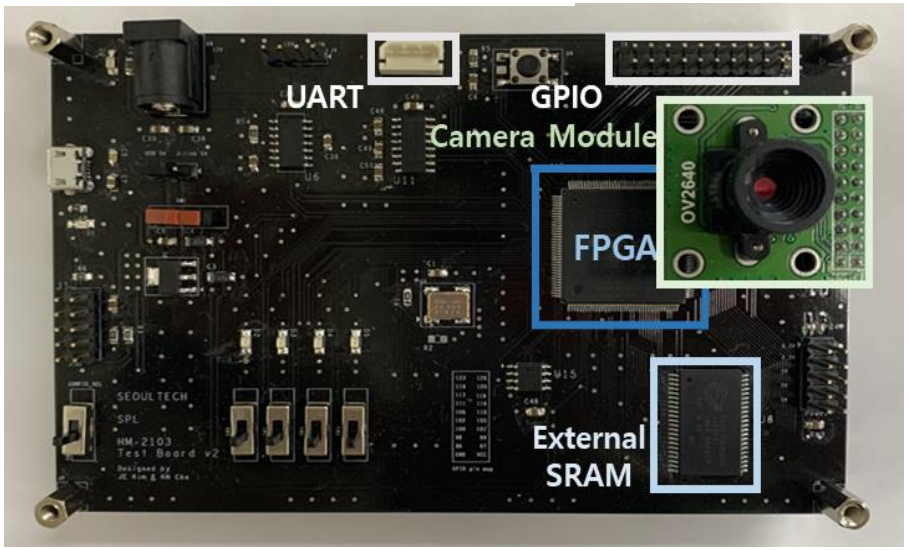
2. 

스토캐스틱 연산기를 통한
이미지 **엣지 검출 연산**

3. 

출력 이미지를 AI 가속기의
학습/인식 데이터로 활용

- **테스트 보드 사양**



[테스트 보드 이미지]

항목	테스트 보드 사양
FPGA	Altera MAX10 10M50SCE144C8G
입력 전압	5V / 12V
동작 주파수	50MHz
메모리	off-chip 512KB SRAM
카메라 모듈	OV2640 CMOS
페리페럴	카메라 I/F, 시리얼 I/F (UART), GPIO


김정은

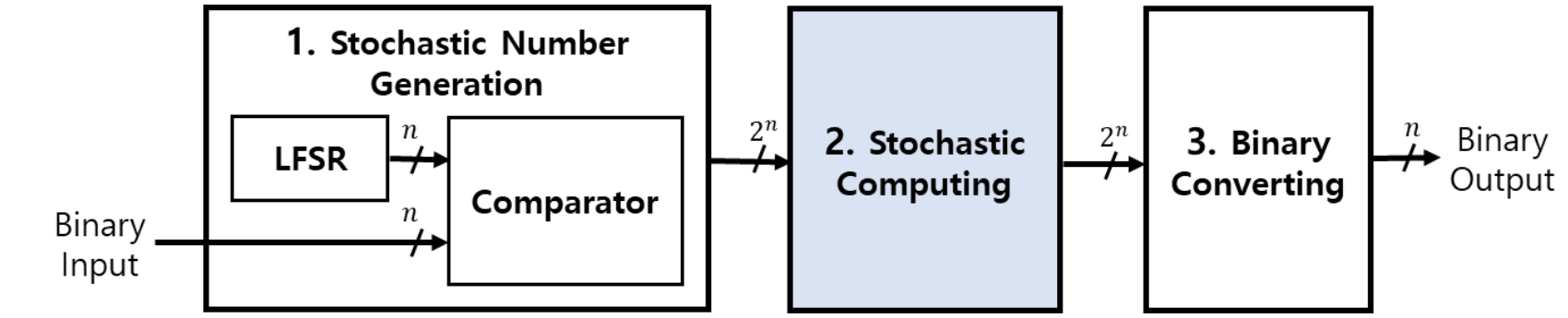

고광현

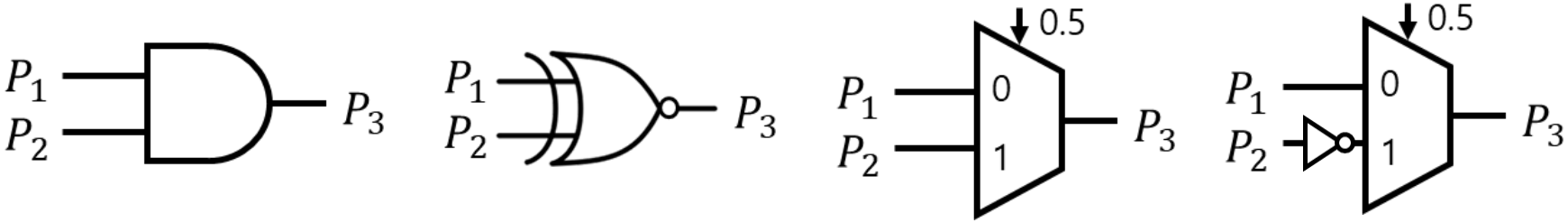

정영우


정원식

감사합니다

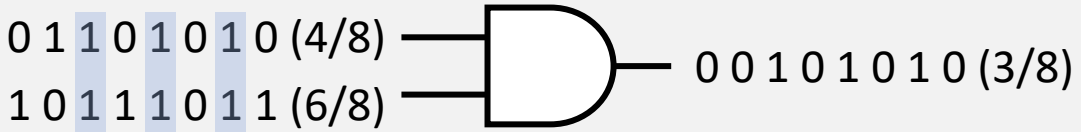
성능이득 분석(1) : 면적 & 파워





[Unipolar 곱셈기] [Bipolar 곱셈기] [덧셈기] [뺄셈기]

1. 두 개의 스토캐스틱 입력 시퀀스는 서로 **독립적**
2. 각 입력 시퀀스에서 **'1'이 나타날 확률**을 **P1, P2** 로 정의
3. AND gate의 출력이 '1'일 확률은, **두 사건이 동시에 '1'일 확률** = **P1×P2**



Backup slide (reference)

Area & Energy efficiency

W. Qian, X. Li, M. D. Riedel, K. Bazargan and D. J. Lilja, "An Architecture for Fault-Tolerant Computation with Stochastic Logic," in IEEE Transactions on Computers, vol. 60, no. 1, pp. 93-105, Jan. 2011, doi: 10.1109/TC.2010.202.

A 16×128 Stochastic-Binary Processing Element Array for Accelerating Stochastic Dot-Product Computation Using 1-16 Bit-Stream Length

Qian Chen, Yuqi Su, Hyunjoon Kim, Taegeun Yoo, Tony Tae-Hyoung Kim, and Bongjin Kim
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
50 Nanyang Avenue, Singapore, 639798
Email: {e170029, yuqi003, kimh0003}@e.ntu.edu.sg, {tgyoo, thkim, bjkim}@ntu.edu.sg

Abstract- This work presents 16×128 stochastic-binary processing elements for energy/area efficient processing of artificial neural networks. A processing element (PE) with all-digital components consists of an XNOR gate as a bipolar stochastic multiplier and an 8bit binary adder with 8× reg
sums. The PE array comprises 1
128 PEs cascaded in a single row
proposed dot-product unit is mini
of bit-streams required for minin
induced by the approximate stoc
dot-product operation requires th
16, which is two orders of magni
stochastic computation using MU
dot-product error is 6.9-to-1.5%
from the baseline stochastic meth
to-2048. A mean MNIST classifica
is 1.19% lower than 8b binary) us
The measured energy from a 65n
product, and the energy efficiency

Keywords- stochastic computation, product, multi-layer perceptron, im

I. INTROD

Stochastic computing [1-3] i
method based on probabilities of
many 1's in a bit-stream). S
stochastic computing are perfe
logic gates, it is suitable for mot
small footprint. Despite the po
has not been employed in pract
concerns in the latency and ene
the typical bit-stream length for

Stochastic Multiplier | Stochastic Add

	[7] ASSCC'16	[8] JSSC'19	Proposed
Computing Type	Analog (Deterministic)	Analog (Deterministic)	Digital (Stochastic)
Technology	28nm	65nm	65nm
Precision Control	Fixed (8b)	Fixed (6b/1b)	Reconfigurable
MAC Circuit Type	Analog	Analog In-Mem.	Digital
ADC/DAC Overhead	Embedded	Required	Not Required
Parallelism	No	16×	16×
Energy Efficiency	9.61TOPS/W	51.3TOPS/W	25.5TOPS/W @^{D}N=16
Energy FoM^{A}	208fJ	39.0fJ	39.3fJ @^{D}N=16
Area FoM^{B}	720.4μm^2	61.5μm^2	154.4μm^2
Accuracy^{C}	N/A	98% (CNN/4-layers)	96.1% @^{D}N=16 (MLP/3-layers)

^{A}Energy FoM=Energy/(# of inputs)×(# of dot-products)
^{B}Area FoM=Area/(# of inputs)×(# of dot-products) ^{C}MNIST dataset ^{D}Bit-stream length

978-3-9819263-4-7/DATE20

[레퍼런스 논문의 이미지]

Error Injection

Q. Chen, Y. Su, H. Kim, T. Yoo, T. T. -H. Kim and B. Kim, "A 16×128 Stochastic-Binary Processing Element Array for Accelerating Stochastic Dot-Product Computation Using 1-16 Bit-Stream Length," 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2020, pp. 678-681, doi: 10.23919/DATE48585.2020.9116349.

IEEE TRANSACTIONS ON COMPUTERS, VOL. 60, NO. 1, JANUARY 2011 93

An Architecture for Fault-Tolerant Computation with Stochastic Logic

Weikang Qian, *Student Member, IEEE*, Xin Li, Marc D. Riedel, *Member, IEEE*, Kia Bazargan, and David J. Lilja, *Fellow, IEEE*

Abstract—Mounting concerns over variability, defects, and noise motivate a new approach for digital circuitry: stochastic logic, that is to say, logic that operates on probabilistic signals and so can cope with errors and uncertainty. Techniques for probabilistic analysis of circuits and systems are well established. We advocate a strategy for synthesis. In prior work, we described a methodology for synthesizing stochastic logic, that is to say logic that operates on probabilistic bit streams. In this paper, we apply the concept of stochastic logic to a reconfigurable architecture that implements processing operations on a datapath. We analyze cost as well as the sources of error: approximation, quantization, and random fluctuations. We study the effectiveness of the architecture on a collection of benchmarks for image processing. The stochastic architecture requires less area than conventional hardware implementations. Moreover, it is much more tolerant of soft errors (bit flips) than these deterministic implementations. This fault tolerance scales gracefully to very large numbers of errors.

Index Terms—Stochastic logic, reconfigurable hardware, fault-tolerant computation.

1 INTRODUCTION

TABLE 2
Analysis of Error Distribution of the Gamma Correction Function

Added Error (%)	Produced Error (%)									
	> 5		> 10		> 15		> 20		> 25	
	Conv.	Stoc.	Conv.	Stoc.	Conv.	Stoc.	Conv.	Stoc.	Conv.	Stoc.
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	4.0	0.0	3.0	0.0	2.4	0.0	2.0	0.0	1.6	0.0
2	8.2	0.5	6.1	0.0	5.0	0.0	4.2	0.0	3.2	0.0
5	19.5	27.8	14.7	1.6	12.1	0.0	10.3	0.0	7.9	0.0
10	35.2	48.4	27.4	24.3	22.7	7.9	19.3	0.0	15.1	0.0
15	48.4	60.1	38.5	36.6	32.1	21.4	27.3	9.1	21.8	0.2

- **Multiplication.** Consider a two-input AND gate, shown in Fig. 2a. Suppose that its inputs are two

[레퍼런스 논문의 이미지]