

신경회로망의 아날로그 VLSI 구현 시 나타나는 문제점 (Constraints on Implementation of Neural Networks with Analog VLSI Circuits)

오상훈*이영직**
(S. H. Oh, Y. Lee)

신경회로망을 아날로그 VLSI로 구현하는 것은 디지털 구현방법에 비하여 집적도와 신호처리 속도의 장점이 있는 반면에, 아날로그 신호의 저장방법, 시냅스를 구현한 곱셈기의 비선형성, 동작영역, zero offset, noise, gain의 변동 등의 문제가 존재한다. 여기서는, 이러한 문제들이 신경회로망을 구현한 아날로그 회로에서 어떤 형태로 나타나는지 알아보았다. 위와 같은 비이상적 요인들이 신경회로망의 성능에 미치는 영향이 파악되면 보다 더 신뢰성을 갖는 신경회로망 chip을 설계/제작할 수 있을 것이다.

I. 서 론

시냅스(synapse)에 의하여 연결된 많은 수의 뉴런들이 계층적으로 구성되어 있는 신경회로망 모델을 하드웨어로 구현하는 방법으로 반도체 회로를 이용하는 것이 각광받고 있다. 이때, 신호를 표시하고 처리하는 방법은 크게 디지털과 아날로그로 나뉜다. 디지털 회로 설계를 이용할 경우, precision을 원하는 만큼 조정 가능하며, 데이터

의 저장, 신호의 연결 및 multiplexing, 회로설계, 그리고 침 간의 연결이 쉽게 이뤄진다 [1]. 그렇지만, 디지털 방식은 집적도와 대량 병렬처리의 관점에서 아날로그 방식보다 못하다. 아날로그 회로설계 기법을 이용하여 신경회로망을 구현하면, 집적도와 신호처리 속도가 뛰어나다 [2]. 따라서, 아날로그 설계기법을 이용한 신경회로망의 구현이 많이 시도되었지만, 이 방법 역시 여러 가지 한계와 문제점을 내포하고 있다.

신경회로망을 아날로그 회로로 구현하면, 아날로그 신호의 저장방법, 곱셈기의 비선형성과 동작영역, weight 구현의 정확성, zero offset, noise

* 기초기술부 선임연구원

** 자동통역연구실 실장, 책임연구원

등의 문제가 존재한다 [1-4]. 따라서, 이 문제들을 고려하여 아날로그 회로가 설계되어야 하며, 이들은 신경회로망의 성능과 밀접한 관계가 있다.

여기서는 이러한 문제들이 신경회로망을 구현한 아날로그 회로에서 어떤 형태로 나타났는지 살펴보자 한다.

II. 본 론

신경회로망을 구현한 아날로그 회로에서 가장 많이 차지하는 부분은 역시 시냅스에 해당하는 곱셈기이다. 이 곱셈기는 저장된 weight 값과 입력되는 신호를 곱한 결과를 출력한다.

곱셈기의 출력신호들은 해당되는 것끼리 더해진 후 neuron body에 입력되며, neuron body는 입력을 sigmoidal 변환한 신호를 출력한다. 시냅스와 neuron body가 아날로그 회로로 구현될 때 나타나는 문제는 다음과 같이 몇 가지 형태로 정리된다.

첫째, 시냅스를 구현한 아날로그 곱셈기에 관련된 것으로, weight 저장 방법에 따른 weight decay 문제, 곱셈기의 비선형성 및 gain의 변동, zero offset, accuracy, resolution, 그리고 clock feedthrough noise 등을 들 수 있다.

둘째, neuron body를 구현한 증폭기에 관련된 것으로, 역시 gain과 offset 그리고 clock feedthrough noise 이다.

셋째, 시냅스를 구현한 곱셈기와 neuron body를 구현한 증폭기는 모든 신호의 동작영역이 제한되어 있다.

이러한 문제들은 신경회로망을 구현하는 회로와 공정의 종류에 따라 나타나는 정도가 다르다. 이

러한 문제들에 대하여 구체적으로 알아보자.

1. 시냅스 구현 곱셈기의 비이상적 요소

가. Weight Decay

시냅스를 아날로그 곱셈기로 구현할 때, weight를 저장하는 device의 선택이 중요하다. MOS capacitor와 floating gate가 weight의 저장 device로 많이 사용되며, 이를 device에 전하(charge)로 저장된 weight는 누설 때문에 시간에 따라 감소하게 된다.

이러한 weight decay는 Gaussian perturbation보다 신경회로망의 성능을 나타내는 mean squared error에 더 큰 영향을 미치는 것이 시뮬레이션으로 확인되었다 [5].

먼저, MOS capacitor에 weight가 저장된 경우의 decay에 대하여 설명하겠다.

Capacitor에 저장된 전하는 transfer에 이용된 switching MOS 트랜지스터의 역방향 바이어스 p-n 접합을 통하여 substrate로 누설된다 [6].

이때, decay constant(τ) 혹은 leakage rate는 공정의 특성과 weight 저장 capacitor의 C 값에 의해 결정되며, weight refresh time의 결정에도 영향을 미친다. 그 예를 보면, Massengill은, MOSIS 2 μ m double-poly CMOS parameter set에 따른 시뮬레이션 결과, C가 600 fF이고, I_{leak} 는 $5 * 10^{-14} \text{ A}$ 이어서 $-5V < V_w < 5V$ 인 경우 1.5%인 150mV까지 decay를 허용하여 refresh time이 1.8s가 되도록 하였다 [7]. 실험적으로 측정한 leakage rate는 2 μ m MOSIS 공정

$C=2\text{pF}$ 에서 34mV/s 인 경우도 있으며 [1], $3\mu\text{m}$ SACMOS 공정 $C=320\text{fF}$ 에서 2mV/ms 인 경우도 있다 [3]. 또, $1.25\mu\text{m}$ CMOS 공정에서 1%의 decay에 1.5s 가 걸린다는 보고도 있다[6]. C 에 저장된 weight 전압은 시간에 따라

$$V_w(t) = V_w e^{-t/\tau} \quad (1)$$

와 같이 변하며, Lee 등은 MOSIS의 $2\mu\text{m}$ CMOS 공정으로 제작한 시냅스 cell에서 $\tau \approx 50\text{s}$ 을 측정하였다 [8]. Decay 허용정도에 따라 weight의 accuracy(N bit)는

$$N = \log_2 \frac{V_w}{V_w - V_w(t)} \quad (2)$$

와 같이 결정될 수 있다. 이 식에 따르면 $\tau \approx 50\text{s}$ 인 경우 8 bit accuracy를 위해서 최대한 0.195sec의 refresh time이 필요하다 [8]. 한편, 구현된 아날로그 곱셈기에서 V_w 가 reference 전압 (V_{wr})과 함께 $V_w - V_{wr}$ 형태로 사용될 때는, V_{wr} 을 외부에서 인가시키지 말고 V_w 와 같은 형태로 C 에 저장하는 것이 decay의 영향을 줄일 수 있다 [2].

Weight를 oxide에 둘러싸인 floating gate에 저장하는 방법은 Fowler-Nordheim tunneling, hot-electron injection, 혹은 UV mechanism을 이용한다 [8-10]. 이 방법들을 이용하여 floating gate에 전하가 저장되면 그만큼 MOS의 임계전압 (V_{TH})이 변하게 되므로, floating gate에 저장된 전하가 weight 전압의 역할을 하게 된다.

Floating gate는 oxide에 둘러싸여 있으므로 τ 는 아주 클 것이다 ($\tau \approx 50,000\text{s}$) [9]. 그렇지만,

floating gate에 전하가 저장된 MOS의 V_{TH} 는 oxide와 interface에 잡힌 전하의 재배치 때문에 초기에 drift가 생긴다 [8, 11]. 또, 온도가 올라갈 수록 drift는 증가하며, 125°C 이상에서는 저장된 전하의 열전자 방출 때문에 charge-loss drift도 발생한다 [11]. 따라서, weight를 floating gate에 저장하더라도 동작온도가 올라가면 여전히 decay 문제가 존재한다.

나. Gain의 변동

MOS를 이용하여 설계된 곱셈기는 트랜지스터들의 특성이 이상적이지 않기 때문에 원하지 않는 동작을 하게 된다 [11]. 그 중의 하나가 곱셈기의 gain이다. $2\mu\text{m}$ MOSIS 공정으로 제작된 회로상에서 gain을 결정하는 μCox 의 표준편차를 조사한 결과 $\sigma(\mu\text{C}_{\text{ox}}) = 0.5\mu\text{A/V}^2$ 인 보고가 있다 [1].

$1.2\mu\text{m}$ CMOS 공정에서 multiplier의 gain은 대략 10%의 변동이 있다고 생각되어진다 [4]. Gowda 등은 $2\mu\text{m}$ MOSIS 공정으로 제작된 transconduntance amplifier 시냅스에서 weight 전압이 2V인 경우 conduntance의 분포를 조사하여 Gaussian으로 볼 수 있음을 보였다 [12].

다. Zero Offset 전압

$2\mu\text{m}$ MOSIS 공정으로 제작된 곱셈기의 구성 트랜지스터에서 $\sigma(V_{TN}) = 10\text{mV}$, $\sigma(V_{TP}) = 15\text{mV}$ 라고 보고되었다 [1]. 이러한 V_{TH} 와 V_{TP} 의 차이 외에, device ratio와 parasitic capacitance의 mismatch 때문에 입력과 weight 전압이 0이더라도 곱셈기의 출력이 0이 아닌 경우가 있다 [3,

12]. 이것을 바로 잡기 위해 필요한 것이 zero offset 전압이다.

Lansner 등은 시냅스 chip을 만들어 특성측정 결과 zero offset 전압 $|V_{\text{ofs}}| \leq 16\text{mV}$ 라고 보고하였다 [13]. Zero offset은 $1.2\mu\text{m}$ CMOS 공정에서 $\pm 5\sim 10\%$ 의 변동이 있으며, 평균이 0인 Gaussian 분포로 가정할 수 있다 [4].

Multiplier가 입력으로 인가된 전압(V_j)을 바로 사용하지 않고 전류(I_j)로 변환하여 사용하는 경우에는, V_j 에서 I_j 변환에 따른 offset도 존재한다. 그 예로, Lont 등은 $3\mu\text{m}$ SACMOS 공정으로 제작된 회로에서 이 입력 offset이 최악의 경우 50mV 로 나타난다고 보고하였다 [3].

라. 비선형성(Nonlinearity)

아날로그 회로로 구성된 곱셈기는 곱셈의 결과가 선형적으로 변하지 않고, 다소 왜곡되는 특성을 보인다. 그렇지만, 이 비선형성은 임의로 변하는 것이 아니라, 회로의 특성에 의해 고정된다[14].

Kub 등은 곱셈기의 비선형 정도를 harmonic distortion의 측정으로 다음과 같이 알아보았다 [2] (표 1).

〈표 1〉 곱셈기의 종류에 따른 THD
(Total Harmonic Distortion)

곱셈기 종류	입력전압 (V_x)	weight 전압(V_w)	THD
Differential-pair	$\approx 7.5\text{V}$	1V_{pp}	$<0.5\%$
Gilbert	$-0.5\text{V} < V_x - V_{\text{xr}} < 1.0\text{V}$	1V_{pp}	$<1.5\%$
Modified Gilbert	$-0.5\text{V} < V_x - V_{\text{xr}} < 0.75\text{V}$	1V_{pp}	$<1\%$
	$-0.5\text{V} < V_x - V_{\text{xr}} < 0.5\text{V}$	1.5V_{pp}	$<2\%$

Differential-pair 곱셈기는 $V_x \approx 7.5\text{V}$ 에서 1V_{pp} 의 V_w 인 경우 0.5% 이내의 harmonic distortion을 보인다.

Gilbert 곱셈기의 경우는 $-0.5\text{V} < V_x - V_{\text{xr}} < 1.0\text{V}$ 에서 1V_{pp} 의 V_w 에 대하여 1.5% 이내의 total harmonic distortion을 보인다.

Modified Gilbert 곱셈기의 경우, $-0.5\text{V} < V_x - V_{\text{xr}} < 0.75\text{V}$ 에서 1V_{pp} 의 V_w 에 대하여 1% 이내, $-0.5\text{V} < V_x - V_{\text{xr}} < 0.5\text{V}$ 에서 1.5V_{pp} 의 V_w 에 대하여 2% 이내의 total harmonic distortion을 보인다고 보고되었다.

Transconductance amplifier 시냅스에서는 비선형성이 입력되는 I_j 에 크게 의존한다 [8]. Lansner 등은 비선형성을

$$D \xrightarrow{\text{def}} \max \zeta \frac{|f(\zeta) - \zeta|}{|\zeta|_{\max}} \quad (3)$$

로 정의하고, 자신들이 제작한 시냅스 chip의 특성을 측정하여 시냅스 비선형도가 $D_{wy} \leq 3\%$ (5bit accuracy)임을 얻었다 [13].

마. Accuracy와 Resolution

곱셈기의 accuracy는 weight를 얼마만큼 정확하게 구현하였는가로 본다면, resolution은 곱셈기의 출력에서 구분될 수 있는 ΔW 라고 볼 수 있다 [13]. 따라서, resolution은 곱셈기의 gain 변동이 얼마만큼 작은가에 관련되어 있다. Lansner 등은 실제 회로상에서 $V_{wres} \leq 2\text{mV}$ 로서 2V 의 weight 범위에서 10bit의 resolution을 얻을 수 있다고 보았다 [13].

바. Noise

아날로그 곱셈기에 저장되는 weight는 noise에 의해 영향을 받는다. Switching MOS 트랜지

스터의 noise는

$$V_n = \sqrt{kT/C} \quad (4)$$

로 나타나며, $C=320fF$ 에서 $V_n \approx 114\mu V$ 정도이다 [3]. 또, Lont 등은 $W=L=3\mu m$, $C=320fF$, $C_{GD}=0.3fF$ 인 회로에서 최악의 경우 $35mV$ 정도의 clock feedthrough에 의한 error 전압이 발생함을 보였다 [3]. 이외에, subthreshold conduction과 interface의 trap에 의한 전하손실도 무시할만 하지만 noise로 작용한다 [6]. 이러한 noise들 때문에 weight가 저장될 때 Gaussian 형태의 perturbation이 일어난다고 볼 수 있다 [5].

2. Neuron Body 구현

Neuron body는 Gilbert 곱셈기의 sigmoid 특성을 이용하거나 증폭기를 이용한다 [4, 15]. 어느 방법으로 구현하든 곱셈기와 마찬가지로 gain의 변동이 구현된 neuron body마다 나타난다.

Lansner 등은 neuron chip에서 sigmoid transfer 특성이 얼마만큼 원하는 값에서 벗어나는지 측정한 결과 $D_g \leq 2\%$ 를 얻었다 [13]. Gowda 등은 $2\mu m$ MOSIS 공정으로 제작된 neuron body의 gain의 분포를 조사한 결과 Gaussian으로 볼 수 있다고 보고하였다 [12]. 따라서, Gain의 변동은 1을 평균으로 한 Gaussian 분포로 가정할 수 있다 [4].

아날로그로 구현한 neuron body에서는 offset 문제도 있다. Transconductance 증폭기를 이용하여 neuron body를 구현한 경우, V_T 가 \pm

$10mV$ 의 변동을 나타내어 neuron body의 input offset 전압은 대략 $25mV$ 라는 보고가 있다 [15]. 또, 여기서는 증폭기의 출력전류가 양인 경우의 한계전류가 음인 경우의 한계 전류보다 6% 정도 크게 측정되었다. 일반적으로 증폭기의 출력전류는 $\pm 20\%$ 의 변동이 있다고 본다.

Exponential pulse decay modulation (EPDM) 곱셈기와 같이 곱셈 결과가 capacitor에 전하로 저장되는 경우에는 각 곱셈기에서 neuron body의 입력 capacitor로 연결된 switching MOS의 clock feedthrough noise도 neuron body의 동작에 영향을 미친다 [7].

3. 동작영역의 제한

위에서 설명한 아날로그 회로로 구현된 시냅스 혹은 neuron body는 그 동작영역이 제한을 받는다 [14]. 특히, 곱셈기는 동작영역을 크게 잡으면 비선형성이 증가한다. 대부분의 곱셈기는 $1V$ 정도 이내에서 곱하기를 수행한다 [7, 13]. Massengill은 EPDM 곱셈기를 구성하여 $5V$ 이내에서 사용하였다 [7]. 또, weight의 동작영역이 입력 값에 의해 $\sqrt{i_x/k}$ 로 제한되는 시냅스도 있으며 [3, 8], Gilbert 곱셈기에서는 $\max(V_3, V_4) > \min(V_1, V_2)$ 의 제한을 벗어나면 비선형성이 증가한다 [15].

Sigmoid 특성을 이용하는 neuron body는 gain에 따라 출력의 영역이 정해진다 [12]. 또, neuron body가 current source로 동작하여 I_{out} 을 출력하게 해주는 V_{out} 의 동작영역도 제한된다 [15].

III. 고 칠

신경회로망을 아날로그 전자회로로 구현하는 데는 이상과 같은 많은 제약조건이 있으나, 이들이 신경회로망의 성능에 얼마나 영향을 미치는가는 아직 이론적 분석이 없이 시뮬레이션으로만 확인되고 있다.

여기서 설명한 부분들을 기초로 하여 아날로그 전자회로로 구현된 시냅스와 neuron body에 관한 수학적 모델이 수립되면, 아날로그 전자회로로 구현된 신경회로망의 성능분석이 가능하다. 따라서, 회로설계시 이 분석을 근거로 하여 보다 더 신뢰성 있는 칩을 만들 수 있을 것이다.

참 고 문 헌

- [1] B. Linares-Barrance, E. Sanchez-Sinencio, A. Rodriguez-Vazquez, and J. L. Huertas, "A CMOS analog adaptive BAM with on-Chip learning and weight refreshing," *IEEE Trans. Neural Networks*, vol. 4, pp. 445-455, May 1993.
- [2] F. J. Kub, K. K. Moon, I. A. Mack, and F. M. Long, "Programmable analog vector matrix multipliers," *IEEE J. Solid State Circuits*, vol. 25, pp. 207-214, Feb. 1990.
- [3] J. B. Lont and W. Guggenbuhl, "Analog CMOS implementation of a multilayer perceptron with nonlinear synapses," *IEEE Trans. Neural Networks*, vol. 3, pp. 457-465, May 1992.
- [4] B. K. Dolenko and H. C. Card, "The effects of analog hardware properties on back-propagation networks with on chip learning," *Proc. ICNN'93 San Francisco*, vol. I, pp. 110-115, Mar. 1993.
- [5] D. B. Mundie and L. W. Massengill "Weight decay and resolution effects in feed forward artificial neural networks," *IEEE Trans. Neural Networks*, vol. 2, pp. 168-170, Jan. 1991.
- [6] D. B. Schwartz, R. E. Howard, and W. E. Hubbard, "A programmable analog neural network chip," *IEEE J. Solid-State Circuits*, vol. 24, pp. 313-319; Apr. 1989.
- [7] L. W. Massengill, "A dynamic CMOS multiplier for analog VLSI based on exponential pulse-decay modulation," *IEEE J. Solid-State Circuits*, vol. 26, pp. 268-276, Mar. 1991.
- [8] B. W. Lee and B. J. Sheu, *Hardware Annealing in Analog VLSI Neuro-computing*. Kluwer Acad. Publ., Massachusetts, 1991.
- [9] R. G. Benson and D. A. Kerns, "UV-activated conductances allow for multiple time scale learning," *IEEE Trans. Neural Networks*, vol. 4, pp. 434-440, May 1993.
- [10] G. Gauwenberghs, C. F. Neugebauer, and A. Yariv, "Analysis and verification of an analog VLSI incremental outer-product learning system," *IEEE Trans. Neural Networks*, vol. 3, pp. 488-497, May 1992.
- [11] E. Sackinger and W. Guggenbuhl, "An analog trimming circuit based on a floating-gate device," *IEEE J. Solid-State Circuits*, vol. 23, pp. 1437-1440, Dec. 1988.
- [12] S. M. Gowda, B. J. Sheu, and J. Choi, "Testing of programmable analog neural network processors," *Proc. IEEE 1992 Custom Integrated Circuits Conf.*, pp. 17.1-17.1.4, 1992.
- [13] J. A. Lansner and T. Lehmann, "Analog CMOS chip set for neural net works with arbitrary topologies," *IEEE Trans. Neural Networks*, vol. 4, pp. 441-444, May 1993.
- [14] L. W. Massengill and D. B. Mundie, "An analog neural hardware implementation using charge-injection multipliers and neuron-specific gain control," *IEEE Trans. Neural Networks*, vol. 3, pp. 354-362, May 1992.
- [15] C. Mead, *Analog VLSI and Neural Systems*. Addison-Wesley, 1989.