공학석사 학위논문

InGaZnO 박막 트랜지스터를 이용한 3T1C 전하 저장형 심층 인공 신경망 가속기

2023년 02월

서울대학교 대학원

재료공학부

강 민 승

InGaZnO 박막 트랜지스터를 이용한 3T1C 전하 저장형 심층 인공 신경망 가속기

지도 교수 김 상 범

이 논문을 공학석사 학위논문으로 제출함
2023년 01월

서울대학교 대학원
재료공학부
강 민 승

강민승의 공학석사 학위논문을 인준함
2023년 01월

위 원 장 박 민 혁 (인)

부위원장 김 상 범 (인)

위 원 강 기 훈 (인)

초 록

인공지능 알고리즘은 이미지 인식, 자연어 처리 등의 분야에서 괄목할 만한 발전을 이루고 있지만, 인공지능 구조의 복잡성, 전력 소모, 학습 기간의 증가를 전통적인 폰 노이만 컴퓨팅 구조가 따라가지 못하는 상황이다. 폰 노이만 병목 현상을 극복하기 위해 비휘발성 메모리를 crossbar 형태로 제작하여 학습 과정에서의 행렬-벡터 곱연산을 가속하려는 시도가 있었지만, 가중치 갱신의 비선형성, 비대칭성에 의해 현재로는 추가 연구가 필요하다. 반면 Si CMOS와 커패시터를 이용하면 이상적인 가중치 갱신이 가능하지만, 휘발성이라는 단점이 있다. 본 연구에서는 낮은 누설전류 수준을 가지는 amorphous InGaZnO 박막 트랜지스터를 활용하여 3T1C 구조의 시냅스 소자를 제시하였다. a-IGZO 트랜지스터는 n-type만 존재하기 때문에 가중치 갱신 과정이 비선형적일 수 있지만 소자가 인공지능 학습 알고리즘이 의도하는 가중치로 수렴할 수 있었고, 제작한 가중치 갱신 모델과 실험을 통해 소자의 비이상적인 특성들을 개선할 수 있는 방법을 제시하였다. 또한, 낮은 누설전류에 의한 10,000분 이상의 가중치 보존 시간 상수를 확인하였으며, 5×10^7의 가중치 갱신 사이클 동안 시냅스 소자가 변화 없이 동작하는 것 또한 확인하였다. 본 연구에서 제시된 3T1C 소자와 적합한 알고리즘이 결합한다면 인공지능을 저전력, 고속으로 학습할 수 있을 것으로 기대한다.

주요어 : InGaZnO 박막 트랜지스터, 심층신경망 연산 가속기, 전하저장형 시냅스, 낮은 누설전류, 가중치 갱신의 선형성과 대칭성
학 번 : 2021-28307

목 차

List of Tables

List of Figures

1. 서 론

1.1 심층신경망 가속을 위한 시냅스 소자의 필요성

인공지능 알고리즘은 이미지 인식, 자연어 처리 등의 복잡한 작업을 특화할 수 있는 장점이 있고, [16], [42]와 같은 고도화된 알고리즘의 발달로 미래 산업에서의 중요성이 증가하고 있다. 새롭게 개발되는 심층 인공 신경망들은 높은 분류 정확도를 가지지만 더 많은 가중치 저장과 연산이 요구된다. 1998년의 LeNet-5[30]은 10^6개 이하의 파라미터를 사용했지만, 최신 신경망들은 10^{14}개 이상의 파라미터를 사용하며, 이에 따라 학습 과정에서의 소비 전력과 비용이 기하급수적으로 증가하고 있다[37, 57]. 심층 인공 신경망의 연산 중 행렬 벡터 곱 연산이 (Matrix-Vector Multiplication, MVM) 가장 큰 비중을 차지하는데[5], 통용되는 폰 노이만(von Neumann) 컴퓨팅 구조는 메모리와 연산 장치 사이의 병목 현상 때문에 전력 효율과 연산 속도 측면에서 소프트웨어의 발전 방향과 적합하지 않다.

Crossbar array 구조의 resistive processing unit (RPU)를 사용하는 새로운 컴퓨터 아키텍처는 인공 신경망 구동에서의 폰 노이만 컴퓨팅 구조의 한계를 극복하기 위해 제안되었다[54]. RPU 아키텍처는 뉴런 신호를 전달하는 서로 평행한 행과 열 방향의 금속 선과, 선의 교차점마다 시냅스 역할을 하는 메모리 소자로 이루어진다. 금속 선들은 뉴런의 입출력을 위해 존재하며, 행 방향의 금속 선은 입력 전압 신호가 주입되고 열 방향은 전류 신호가 출력된다. 교차점의 메모리 소자는 시냅스의 가중치를 전기전도도로 저장한다. 옴의 법칙과 키르히호프의 법칙에 의해 각 행에 전압 신호를 주입하면 입력 전압 벡터와 전기전도도 행렬의 곱 연산 결과가 각 열에 전류의 합으로 나타나고, 하나의 crossbar array가

심층 신경망의 두 층 사이에서 일어나는 연산을 한 번에 수행하게 된다. 이렇게 crossbar array 구조의 RPU는 가중치 저장과 MVM 연산이 한 공간에서 일어나며 기존 병목 현상에서 벗어나 뛰어난 전력 효율과 처리 속도를 가질 수 있다.

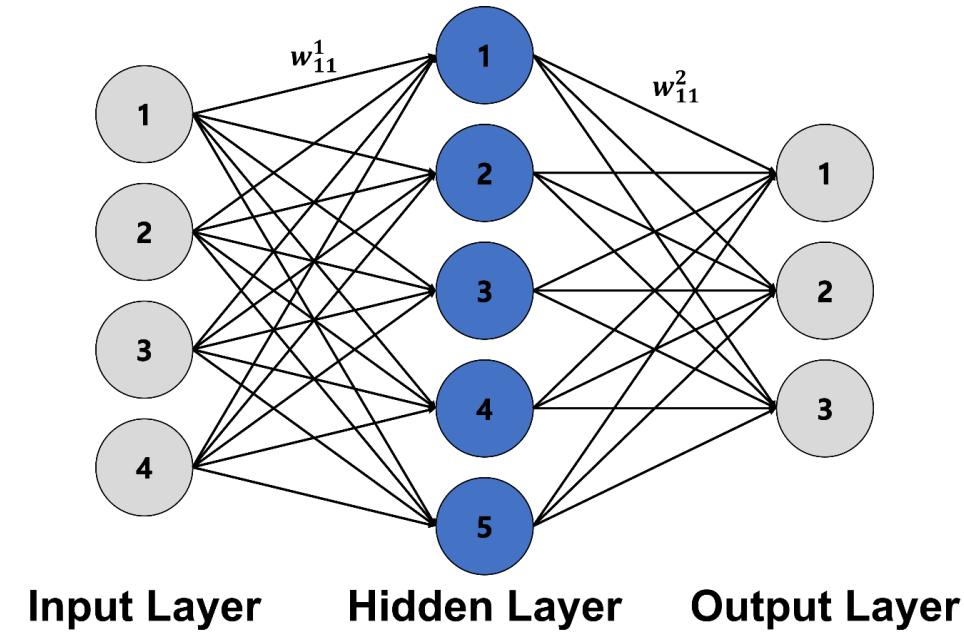

Figure 1.1 심층 신경망 구조. 원은 뉴런을, 실선은 시냅스 가중치를 의미한다.

RRAM (Resistive Random Access Memory)[58], PRAM (Phase change Random Access Memory)[44], MRAM (Magnetic Random Access Memory)[23], FeFET (Ferroelectric Field Effect Transistor)[22]과 같은 차세대 비휘발성 메모리가(Non-Volatile Memory, NVM) 시냅스 가중치 저장 장치로써 제안되었고, 외부 컴퓨터에서 학습된 네트워크를 업로드 한 추론용(inference-only) 어레이로 제작된 바가 있다[25]. 그러나 학습 과정까지 가속할 수 있는 완전한 심층 인공 신경망 가속기 제작을 위해서는 온 칩 학습(on-chip

learning)이 가능해야만 한다. On-chip learning을 위해서는 표현할 수 있는 가중치 단계가 많아야 하며 시냅스 가중치 갱신이 선형적이고 대칭적이어야 한다. 또한, 학습 과정에서 소자를 읽고 쓰는 과정이 안정적이어야 하므로 내구성과 아날로그 상태 retention이 일정 수준 이상이어야 한다[15]. 많은 NVM이 가지는 문제인 비선형적, 비대칭적 시냅스 가중치 갱신은 특히 학습된 신경망의 정확도를 크게 떨어트리는 것으로 알려져 있다[15, 21, 55, 59].

이런 차세대 비휘발성 메모리의 한계 때문에 crossbar array의 시냅스 소자를 CMOS 트랜지스터와 커패시터를 통해 표현하고자 하는 시도가 있다[26, 33, 35]. CMOS는 성숙한 기술이며, 빠른 동작 속도와 linear하고 symmetric 한 가중치 갱신이 가능하다는 점에서 이점을 가지지만, DRAM이 64 ms마다 refresh 과정을 거치는 것처럼 CMOS 트랜지스터는 off current가 크기 때문에 시간에 따라 커패시터에 저장된 정보가 누실 된다는 단점이 있다. [33]에서처럼 비교적 작은 규모의 신경망을 학습시키는 데는 빠른 주기로 커패시터가 갱신되기 때문에 retention에 의한 문제가 발생하지 않지만, 실용적인 거대한 신경망들에 적용되기는 어렵다. 커패시터의 용량을 늘려 동일 누설 전류에 대해 전압 감소를 낮출 수 있지만 가중치 갱신 에너지 소모와 소자 면적이 증가한다는 상충관계가 있다.

따라서 현재의 CMOS 기술과 미래의 이상적인 비휘발성 기술을 연결하는 중간다리 역할이 필요하며, 본 논문에서는 amorphous InGaZnO (a-IGZO) 박막 트랜지스터(Thin Film Transistor, TFT)와 커패시터를 이용한 3T1C 전하 저장형 시냅스를 이로 제안하는 바이다. a-IGZO TFT는 누설 전류가 작아 CMOS 기반 시냅스와 다르게 작은 커패시터 용량으로도 합리적인 데이터 retention time을 가질 수 있으며, 동시에 NVM 소자보다 많은 가중치 단계를 표현할 수 있고 선형적, 대칭적인

갱신이 가능하다. 3T1C a-IGZO TFT 시냅스는 on-chip learning을 위한 하드웨어 가속기뿐만 아니라 다른 비휘발성 메모리 어레이와 함께 사용하는 보조 어레이 역할[14, 31] 또한 수행할 수 있을 것으로 기대한다.

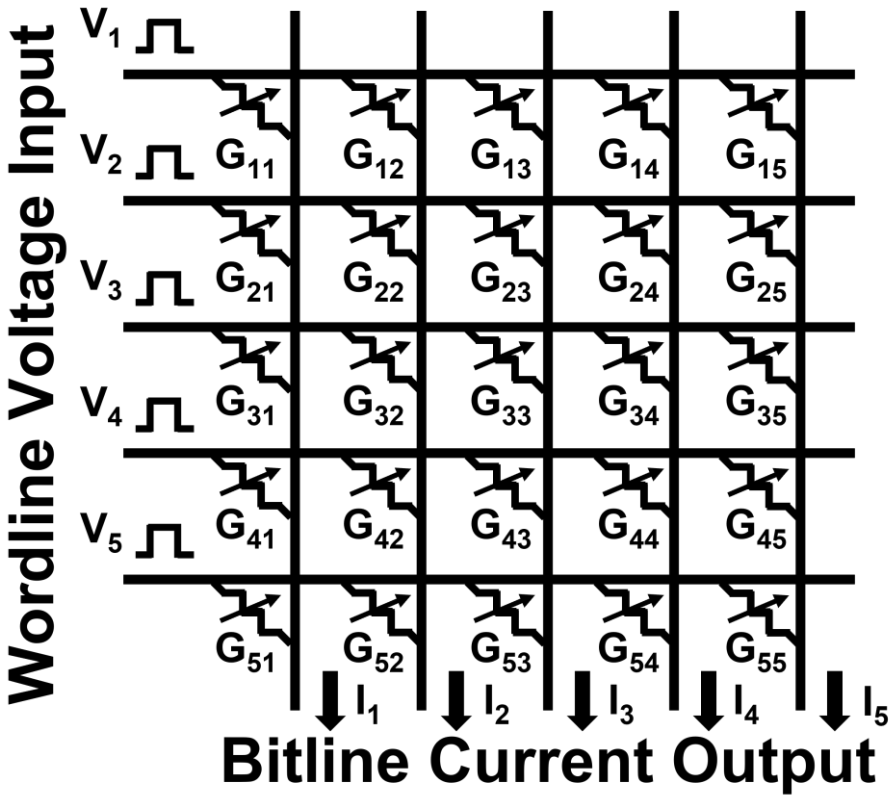

Figure 1.2 Crossbar 어레이 구조. 가로 방향 wordline으로는 전압 input이, 세로 방향 bitline으로는 전류 output이 출력된다.

2. 문헌 조사

2.1 비휘발성 메모리를 이용한 시냅스 소자

심층 신경망 연산 가속기는 추론만 가능한 array와 on-chip learning까지 가능한 array로 나뉜다. 비휘발성 메모리를 이용한 추론용 array는 메모리의 retention 성능과 일관성 있는 데이터 읽기가 가능해야 하며, on-chip learning을 위해서는 추가로 큰 on/off ratio, 선형적이고 대칭적인 가중치 갱신, 좋은 cycling endurance 특성이 요구된다[15]. 이러한 조건을 만족하지 못하면 최대 정확도 감소 등 제약이 있다. 특히 가중치 갱신의 비선형성, 비대칭성이 학습에 가장 큰 문제로 알려졌다. 심층 신경망 학습 과정상 많은 수의 potentiation과 depression 명령이 가해지는데, 가중치 갱신이 비선형적이라면 갱신에 따라 특정 가중치로 수렴해버려 의도치 않은 방향으로 학습되기 때문이다[15, 20].

PRAM은 작동 원리가 명확하며 CMOS 기술과 호환되는 장점을 가진 상용화된 제품이 있을 정도로[17] 성숙한 비휘발성 소자이다. PRAM은 좋은 retention 특성을 가져서 추론용 array로는 알맞은 성질을 가진다. 그러나 weight를 증가시키는 과정인 potentiation은 가능하지만, weight를 감소시키는 depression 과정은 아날로그 프로그래밍이 불가능한 단점이 있다. PRAM에서 저항을 높이는 방향의 쓰기는 상변화 물질을 녹인 뒤 급랭하는 방식이고, 이 과정이 점진적으로 일어날 수 없어 전기전도도 변화가 급격하다[45]. 이 문제 때문에 PRAM을 시냅스 소자로 사용할 때는 두 PRAM 소자를 하나의 시냅스 소자로 활용하지만[27] 동작 속도 감소 및 면적 증가의 문제가 발생한다. 이외에도 상을 변화시킬 때의 큰 에너지 소모, 상 간의 부피 차이로 인한 내구성 저하와 시간에 따른 resistance drift 현상 또한 극복해야 할 점이다[43].

RRAM은 metal-insulator-metal 구조로 insulator에 전기 전도성이 있는 통로가 형성된 정도에 따라 정보를 저장한다. 구조가 간단해 제작이 용이하고 면적이 $4F^2$로 작으며 선택 소자 없이도[52] array 동작이 가능하다는 장점이 있는 소자이다. 다만 대체로 depression 과정이 급격하며, 전도성이 있는 통로가 형성되는 과정이 stochastic 하다는 점에서 on-chip learning 용 소자로는 적합하지 않다. Bilayer 구조의 RRAM으로 선형적이고 연속적인 시냅스 갱신을 이룬 연구도 있지만[53], 제작이 복잡하며 여전히 표현할 수 있는 가중치 수가 작다는 한계가 있다.

MRAM은 동작 속도가 다른 비휘발성 메모리에 빠르다는 장점이 있지만 on/off ratio가 작은 문제가 있다. 따라서 표현할 수 있는 시냅스 가중치의 범위가 작고, 심층 신경망 학습에 제한이 된다. 이를 해결하고자 [23]에서는 시냅스 가중치가 0과 1로만 나뉘는 이진 신경망(Binary Neural Network, BNN)에서 MRAM을 이용하였지만, 가중치의 정보 손실에 의한 신경망의 최대 추론 정확도에 한계가 생기는 문제가 있다.

이처럼 PRAM, RRAM, MRAM을 비롯한 비휘발성 메모리 소자들은 on-chip learning을 위한 array에서의 사용하기에는 각각 한계점이 많다. 메모리의 성질에 최적화된 주변 회로를 이용하거나 복잡한 읽기, 쓰기 과정을 통해 신경망의 정확도 하락을 어느 정도 막을 수 있지만, 면적, 전력 소모 등을 희생해야 하므로 on-chip learning에 적합한 시냅스 소자가 필요하다[48]. 특히 write and verify와 같은 소자에 저장된 전도도 상태를 알아야 하는 보정법은 MVM 연산의 병렬성을 잃는 치명적인 단점이 있다.

2.2 Si-CMOS 회로 기반의 on-chip learning

NVM 소자를 사용한 심층 신경망 가속기는 on-chip learning에는 한계가 있기 때문에 CMOS 트랜지스터와 커패시터를 이용한 전하 저장형 시냅스가 제안된 바 있다[33]. 2개의 update 트랜지스터의 source나 drain, 그리고 read 트랜지스터의 gate가 커패시터의 상단 전극에 연결된 3T1C 구조이다. 기본적인 개념은 DRAM과 동일하지만, 정보 갱신과 읽는 방식에 차이가 있다. 아날로그 정보를 저장해야 하므로 2개의 update 트랜지스터가 필요하다. 하나의 NMOS update 트랜지스터로 potentiation을 하는 경우를 생각하면, update 트랜지스터의 source 전압은 커패시터에 저장된 전하에 의해 결정되고, 저장된 정보에 따라 update 트랜지스터에 흐르는 전류 크기가 변하기 때문에 NMOS는 전류 주입에 적합하지 않고, 전류 방출에만 사용할 수 있다. 따라서 potentiation 용 PMOS, depression 용 NMOS로 update 트랜지스터가 2개 필요하다. 또한, DRAM은 저장 커패시터를 방전시켜 저장되어 있던 전하량을 측정해 읽을 때마다 정보 손실이 일어나지만 3T1C 시냅스는 커패시터 전압을 read 트랜지스터의 gate에 가해 아날로그 정보를 read transistor에 흐르는 전류로 읽을 수 있다. 이때, 커패시터의 전압과 read 트랜지스터의 전류가 일대일 대응이 되기 위해서는 read 트랜지스터의 drain에 적절히 작은 전압을 가해 트랜지스터가 linear region에서 동작해야 한다.

Si 트랜지스터를 이용한 전하 저장형 시냅스는 [33]에서와 같이 update 트랜지스터가 saturation region에서 동작하는 한 선형적인 가중치 갱신이 가능하다. 다만 Si 트랜지스터를 통한 누설 전류가 크기 때문에 데이터 retention에는 문제가 있고, 정보 누실을 보정하는 과정이나 비휘발성 정보 저장 장치의 보조가 필요하다. Ambrogio *et al.*의 연구에

서는 PRAM array와 3T1C array를 함께 사용하였다[1]. CMOS 3T1C가 휘발성 메모리이기 때문에 선형적, 대칭적 가중치 갱신이 가능한 3T1C array에서 우선 학습을 한 뒤, 임계점에 도달하면 커패시터의 정보를 비휘발성 메모리인 PRAM으로 옮기는 방식이다. 비휘발성 메모리와 CMOS 트랜지스터의 장점만을 이용할 수 있는 연산 가속기이지만, 학습 과정은 모두 CMOS 3T1C에서 일어나기 때문에 학습 과정의 retention 문제에서 벗어나지 못한다. 간단한 구조의 신경망에서의 MNIST handwritten dataset 학습과 같은 부하가 적은 작업에 대해서는 학습 주기가 짧기 때문에 커패시터 전하량 갱신이 자주 일어나고 retention 문제에 큰 영향을 받지 않지만, 더 복잡한 네트워크와 학습 작업의 경우 가중치 갱신 간 시간이 길기 때문에 나쁜 retention에 의한 정확도 하락이 나타날 수 있다[33].

2.3 InGaZnO 박막 트랜지스터를 활용한 시냅스 소자

IGZO는 산화물 기반 반도체의 대표적인 물질로, 비정질 상에서도 높은 mobility를 가지는 반도체로 유지된다는 점에서 장점이 있다. 일반적으로 반도체는 결정질이 아닐 때 mobility가 크게 감소하지만, In 원자의 구형 5s 오비탈이 크고 등방이기 때문에 비정질 상이여도 오비탈의 겹침이 있어 mobility 저하가 크지 않다[24]. 비정질 상에서도 전기적으로 선호되는 특성을 가진다는 장점 때문에 저온 공정이 가능해 디스플레이의 구동 회로에서 박막 트랜지스터(Thin Film Transistor, TFT) channel 물질로 이용된다[8].

a-IGZO를 비롯해 비정질 반도체는 자연적으로 dopant state로 작용하는 산소 공핍(V_O)이 많아 별도의 doping 과정 없이도 n-type 반도체이다. 다만, Si CMOS처럼 ion implantation 등을 통해 p-type으로 사용

할 수 없어 n-type 트랜지스터만 제작 가능하다[24]. V_O 농도는 a-IGZO channel 증착 시의 산소 분압 등으로 조절이 가능하며, carrier 농도가 높을 시에는 depletion mode 트랜지스터로 동작한다.

a-IGZO TFT는 작은 누설 전류를 가진다. Si CMOS의 누설 전류는 pA/μm 수준이지만[33, 39], a-IGZO TFT는 yA/μm 수준의 누설 전류까지 보고된 바가 있다[41]. 이는 IGZO 물질이 가시광선이 흡수되지 않을 정도의 높은 bandgap energy를 가지기 때문이다. 낮은 누설 전류를 가진다는 특성 때문에 최근 IGZO TFT를 이용하여 메모리 소자를 제작하려는 연구가 많다. Sekine *et al.*의 연구에서는 DRAM처럼 커패시터에 전하를 저장하되, 읽기 과정에서 커패시터를 방전시켜 저장되어 있던 전하량을 확인하는 것이 아닌 gate가 커패시터에 연결된 읽기 트랜지스터를 이용해 non-destructive readout이 가능한 2T1C 메모리 소자를 제안하였다[38, 41]. IGZO TFT의 누설 전류가 작기 때문에 높은 온도에서도 장시간 정보 retention이 가능하였다. 이에 더해 write 트랜지스터의 source와 read 트랜지스터의 gate 사이에 존재하는 기생 커패시터를 저장 공간으로 활용하는 2T0C 메모리 소자가 제안된 바 있으며[40], 수직 적층형 a-IGZO 2T0C 소자 또한 Duan *et al.*의 연구에서 제시되었다[12]. 커패시터 용량이 작음에도 불구하고 기존 DRAM의 64 ms 주기의 refresh보다 긴 300 s의 retention time이 보고된 바 있다.

Dual gate 트랜지스터와 커패시터를 이용한 시냅스와 뉴런 소자 또한 Hu *et al.*의 연구에서 보고되었다[18, 19]. 시냅스 소자는 2T1C 구조이며, 커패시터의 전극이 read 트랜지스터의 bottom gate에 연결되고, top gate에 전압 입력 신호를 가할 수 있도록 하였다. 커패시터에 저장된 전압과 전압 입력 신호에 의해 read 트랜지스터에 흐르는 전류의 크기가 달라지고, 합산된 전류가 inverter 기반의 뉴런 소자가 전압 신호를 출

력하도록 한다. On-chip learning이 가능하였으며, $<10^{-18}$ A/μm 수준의 낮은 누설 전류 수준에 의해 커패시터 전압이 0.1 V 감소하는 데 4시간이 필요한 수준의 retention 성능을 달성하였다. 시냅스 소자 2개로 AND, OR 논리 gate 구현 및 4X6 array에서의 TETRIS 패턴 인식 작업이 가능하였지만, 추론 동작이 수백 μs 정도로 느리며, 디지털 형태의 입출력만 처리할 수 있다는 한계점이 있다.

IGZO는 NMOS만 존재한다는 한계를 극복한 6T1C 시냅스 소자 또한 제안된 바 있다[51]. NMOS는 전하를 방전시키는 데에 적합하기 때문에 potentiation과 depression을 서로 반대되는 방향으로 전하를 방전시키는 방식으로 가중치를 갱신한다. 가중치 갱신을 위한 트랜지스터 4개와 읽기를 위한 커패시터의 상, 하단 전극에 gate가 연결된 트랜지스터 2개, 정보 저장을 위한 커패시터로 이루어진 구조이다. 실제로 전류가 흐르는 두 update 트랜지스터의 source는 항상 접지된 상태이기 때문에 일정한 gate-source 전압 차가 유지되며, 트랜지스터가 일정한 전류를 흘리는 saturation mode에서 동작할 수 있도록 한다. 저장된 시냅스 가중치에 무관하게 saturation mode에서 동작하는 update 트랜지스터에 의해 선형적이고 대칭적인 갱신이 가능하다. 커패시터의 양 전극에 연결된 트랜지스터가 모두 켜져야 갱신이 일어나기 때문에 별도의 선택 소자 없이 array 구동이 가능하지만, update 트랜지스터의 기생 커패시턴스에 의해 half select 된 시냅스 소자의 커패시터에 저장된 전하에 변화가 생긴다는 한계가 있다.

다만 bias stress에 대한 stability (positive, negative bias stability, PBS, NBS)는 해결되어야 할 과제이다[47]. 기본적으로 V_O가 많고, channel과 gate insulator 사이의 defect 들에 gate bias에 의해 carrier trapping이 일어나고, Flash 메모리의 threshold voltage가 변화하는 원

리와 동일하게 channel에 가해지는 전압을 바꾸는 효과가 일어나 트랜지스터의 전기적 특성이 변한다[13]. Top gate TFT를 기준으로, positive bias가 오래 가해지면 electron trapping이 일어나 threshold voltage가 증가하며 반대로 negative bias는 threshold voltage를 낮춘다. 일반적으로 IGZO는 n-type 반도체이고, hole 농도가 작기 때문에 NBS의 영향이 적지만, TFT에 빛이 가해지면 (PBIS, NBIS) electron-hole pair가 생성되어 bias stress가 가속화되는 효과가 있다[29]. 수소 원자 또한 bias stress에 따라 TFT 성질을 변화시킨다. 공정 과정 중 channel에 포함되었던 H 원자가 강한 전압에 의해 결합을 끊고 gate insulator 쪽으로 확산하며 트랜지스터의 특성이 변화한다. Charge trapping은 가역적인 반응이지만, H 원자의 확산은 결합의 변화가 있기 때문에 비가역적인 변화를 야기한다[9]. 현재 디스플레이 구동에 사용되는 IGZO TFT는 bias stress에 의한 threshold voltage 변화를 sensing하고 보정해주는 회로가 있지만[11], 집적도 높은 메모리 소자에 IGZO TFT를 사용하기 위해서는 반드시 해결되어야 하는 문제이다.

3. 실험 및 분석 방법

3.1 단일 트랜지스터 및 커패시터 제작

본 논문에서 3T1C에 포함되는 트랜지스터는 top gate staggered 구조로 thermal SiO_2 위에 제작하였다. Source, drain 금속으로는 Tungsten을 이용하였으며, a-IGZO는 In : Ga : Zn = 1 : 1: 1 비율로 sputtering 기법을 이용해 증착하였다. Gate insulator는 HfO_2를 atomic layer deposition (ALD)를 통해 증착하였다. 실험에 사용한 트랜지스터의 channel dimension은 2 μm x 5 μm이며, 유전막 두께는 10 또는 15 nm이다. Source, drain과 gate가 수직으로 겹친 구조이고, 그 길이는 0.25 μm이다. 공정 과정은 Figure 3.1.1와 같으며, 소자 구조는 Figure 3.1.2와 같다.

제작된 트랜지스터의 transfer curve와 output curve는 각 Figure 3.1.3과 Figure 3.1.4와 같다. 본 실험에서는 모두 동일한 channel dimension을 가지는 트랜지스터만을 사용하였기 때문에 별도의 normalization 과정 없이 I_{DS}가 10^{-11} A가 되는 전압을 threshold voltage로 정의하였다. Subthreshold swing은 drain 전류가 10^{-11} A에서 10^{-10} A가 되는데 필요한 전압 차이로 계산하였다. Carrier mobility는 수식 (1)을 통해 계산하였다. Channel 물질이 polycrystalline이 아닌 amorphous 상이기 때문에 웨이퍼 위 많은 트랜지스터가 좋은 균일도를 가지는 것을 확인할 수 있었다.

$$\mu_{FE}^{Sat} = \frac{2L}{WC_{ox}}\left(\frac{d\sqrt{I_{DS}}}{dV_g}\right)^2 \quad (1)$$

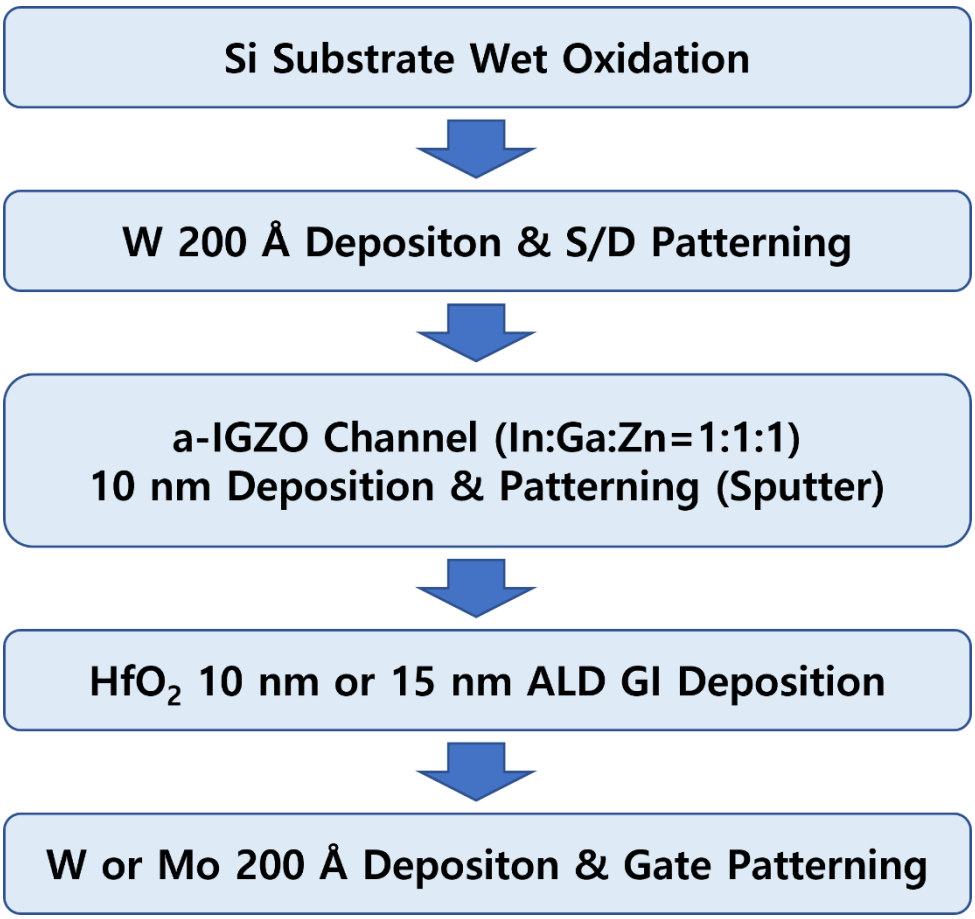

Figure 3.1.1 IGZO 트랜지스터 공정 순서

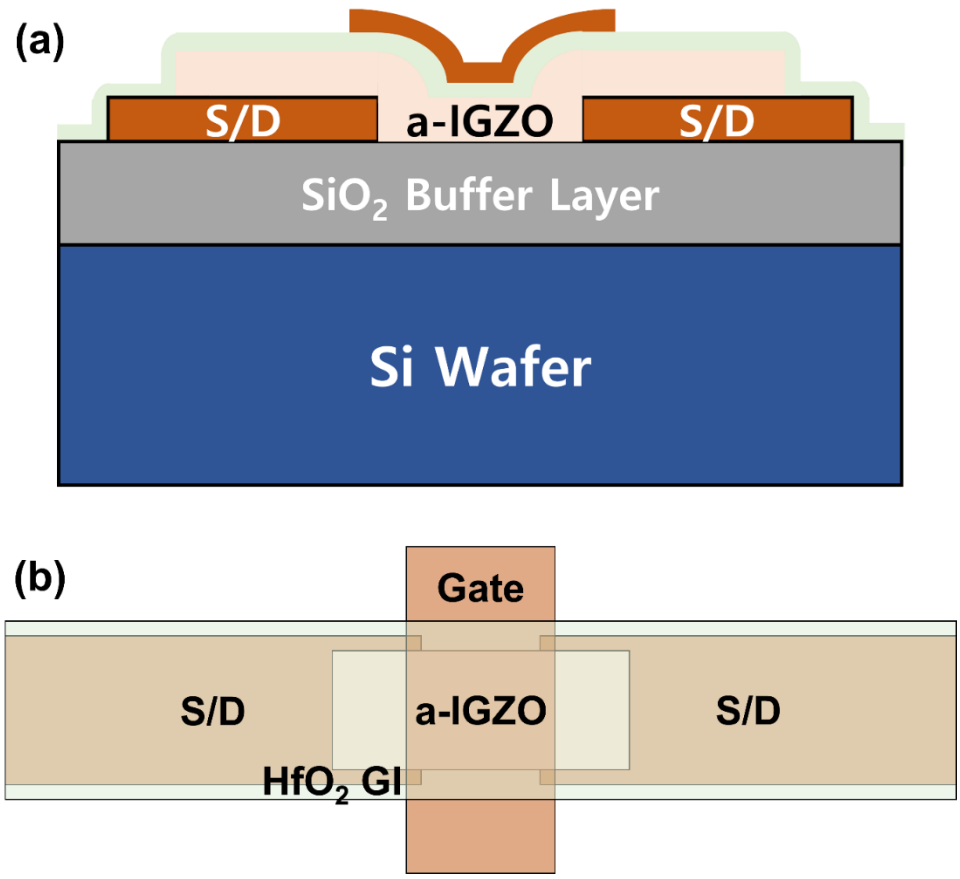

Figure 3.1.2 IGZO 트랜지스터 구조. (a)는 단면, (b)는 top view이다.

표 1 웨이퍼 별 트랜지스터 전기적 성질

	Top gate	Gate oxide	S.S. (mV/dec)	V_{th} (V)	I_{on}* (μA)	Mobility ($cm^2/V/s$)
Wafer 1	W	HfO_2 10 nm	112.9 ± 4.4	-0.58 ± 0.19	2.1 ± 0.63	6.02 ± 0.41
Wafer 2	Mo	HfO_2 15 nm	99.6 ± 1.9	0.33 ± 0.02	0.025 ± 0.0097	0.31 ± 0.17

*V_{DS} = 1.5 V, V_{GS} = 1.0 V

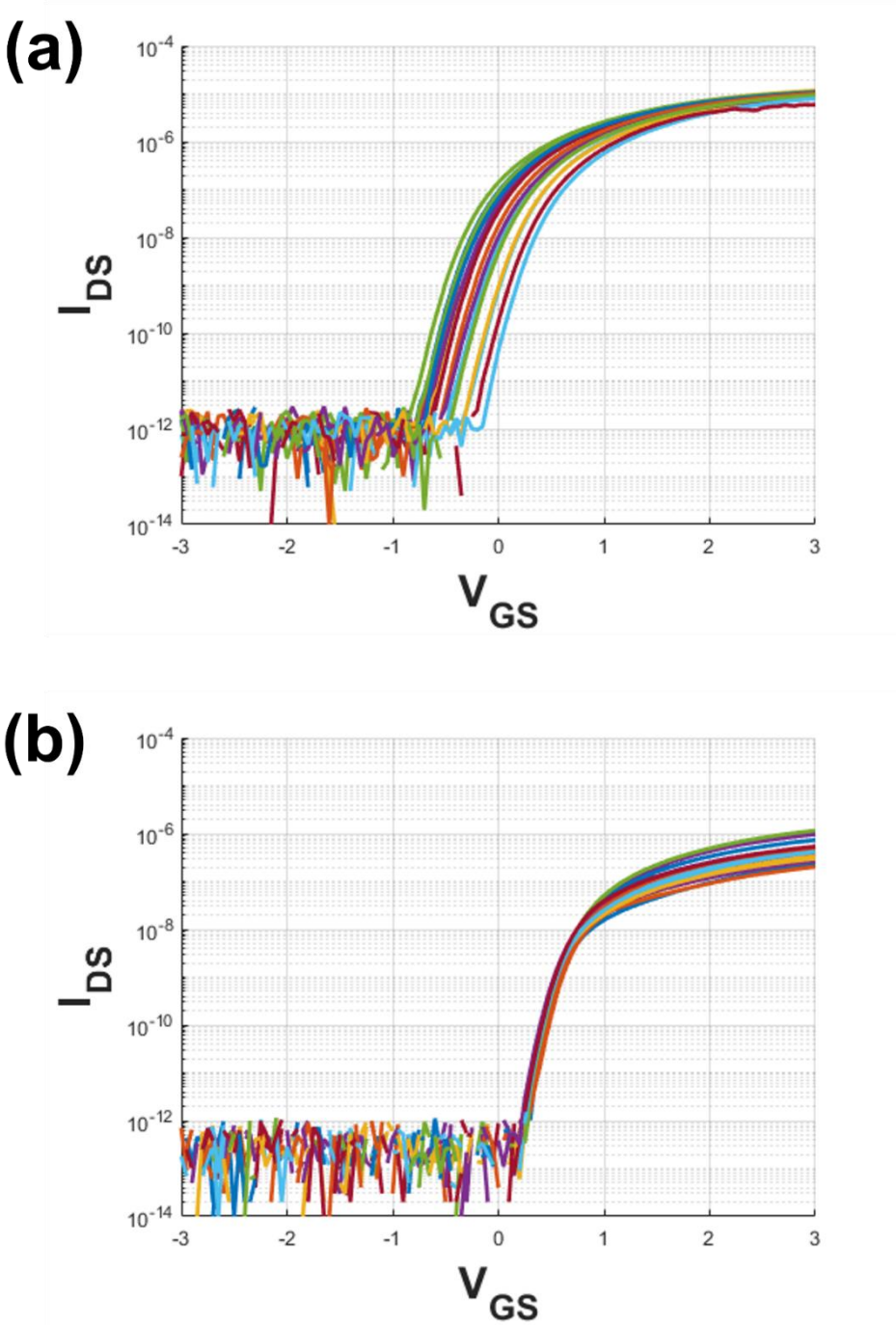

Figure 3.1.3 제작된 트랜지스터의 transfer curve. (a)는 Wafer1, (b)는 Wafer 2의 측정 결과이다.

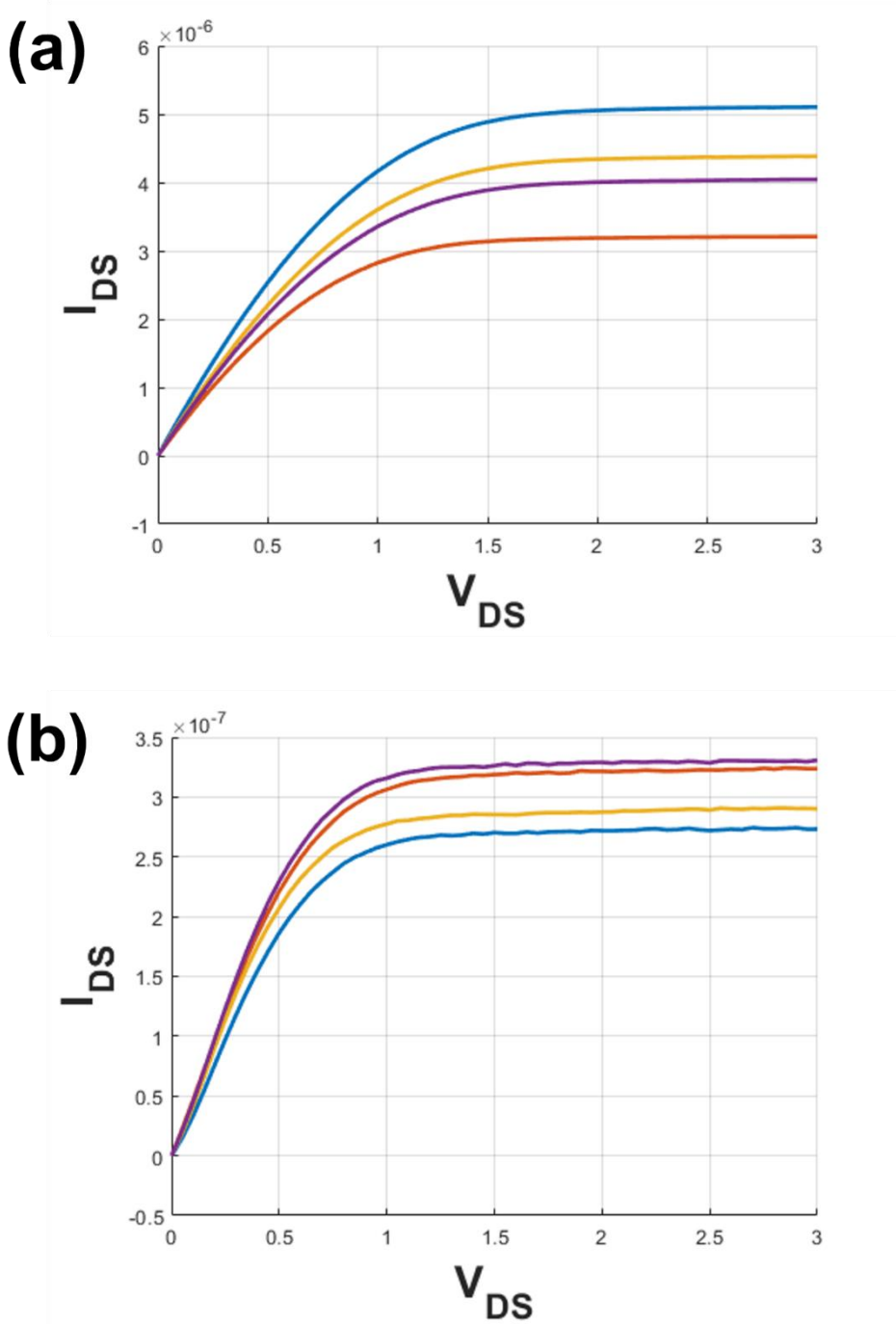

Figure 3.1.4 제작된 트랜지스터의 output curve. (a)는 Wafer 1, (b)는 Wafer 2의 결과이다.

가중치 저장을 위한 커패시터는 high-k dielectric 물질인 HfO_2를 사용하여 metal-insulator-metal (MIM) 구조로 제작하였다. 커패시터에 전하를 저장하는 장치의 경우, 누설 전류가 동일하다는 가정하에 커패시터의 용량이 클수록 수식 (2)에 의해 전압 정보의 손실이 적기 때문에 HfO_2를 사용하였다. 제작한 커패시터는 100 μm x 100 μm의 면적을 가지며, ALD로 증착한 HfO_2 두께는 10 – 15 nm이다. HP4284A LCR meter로 용량을 측정한 결과 −3 – 3 V 전압 범위에서 Figure 3.1.6과 같이 110 – 160 pF 정도이다.

$$\Delta V = I_{leakage} \times \Delta t / C_{storage} \quad (2)$$

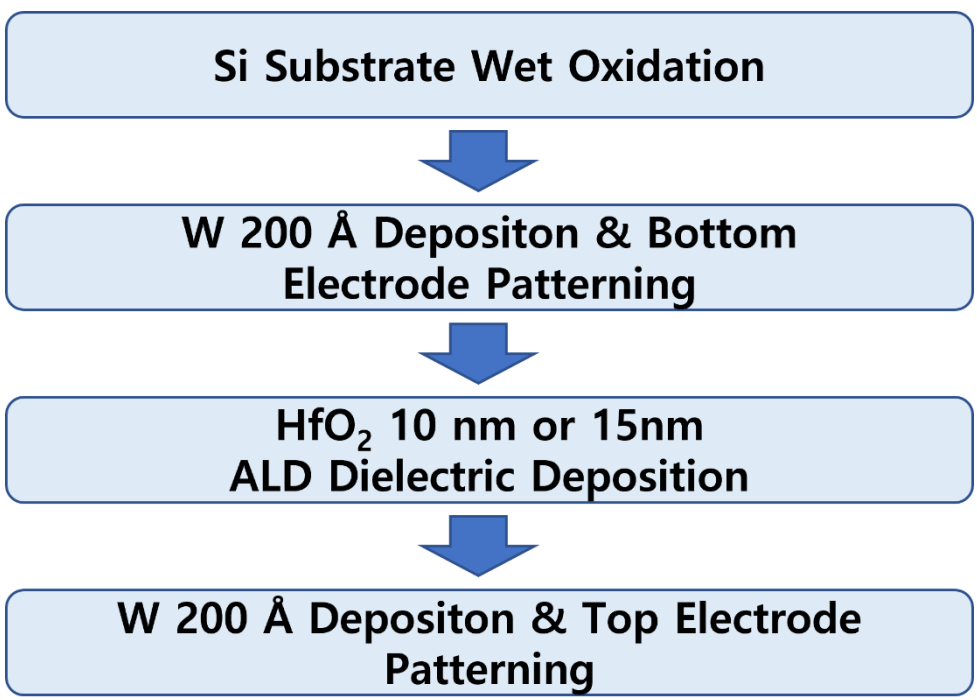

Figure 3.1.5 커패시터 공정 순서

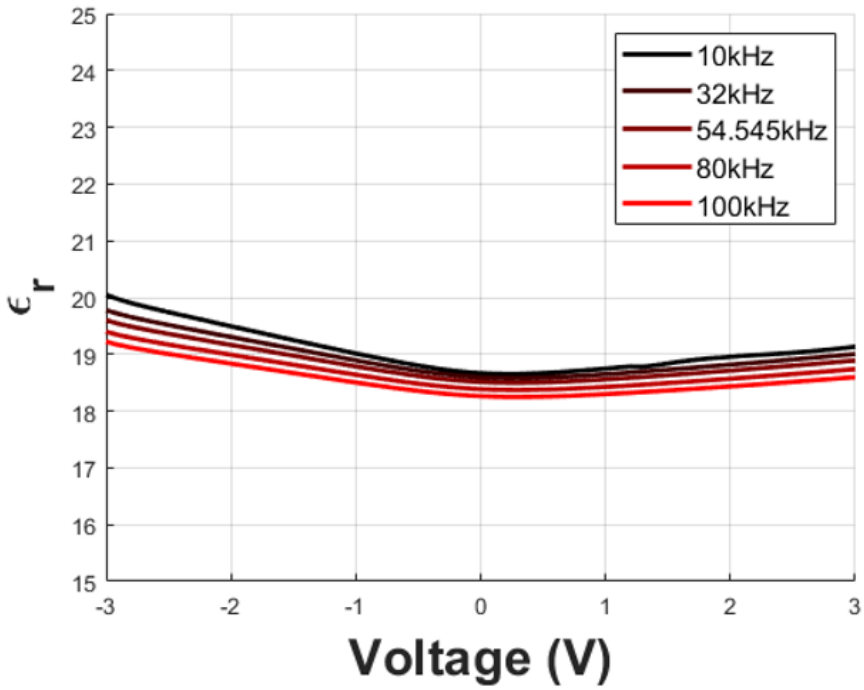

Figure 3.1.6 커패시터 C-V 측정 결과. HfO_2 10 nm, 100x100 μm^2 커패시터를 측정하였다. 흔히 알려진 HfO_2의 relative permittivity인 18-20 사이 값을 가짐을 통해 유전막이 정상적으로 증착된 것을 확인하였다.

3.2 3T1C 구조 및 동작

3T1C의 구조는 [33]의 CMOS 3T1C와 유사한 구조이다. 2개의 시냅스 가중치 갱신용 트랜지스터(update 트랜지스터)와 1개의 읽기 트랜지스터(read 트랜지스터), 그리고 전하를 저장하는 커패시터로 이루어진다. 시냅스 가중치를 증가시키는 potentiation 트랜지스터(N1)의 drain은 V_{DD}와, source는 커패시터의 상단 전극과 연결되어 있으며, 가중치를 감소시키는 depression 트랜지스터(N2)의 drain은 커패시터의 상단 전극과, source는 GND에 연결되었다. Read 트랜지스터의 gate가 커패시터의 상단 전극과 연결되어 커패시터의 전압을 읽을 수 있도록 하였다. Read 트랜지스터의 drain에는 작은 전압을 가해 linear region에서 동작할 수 있도록 하였다.

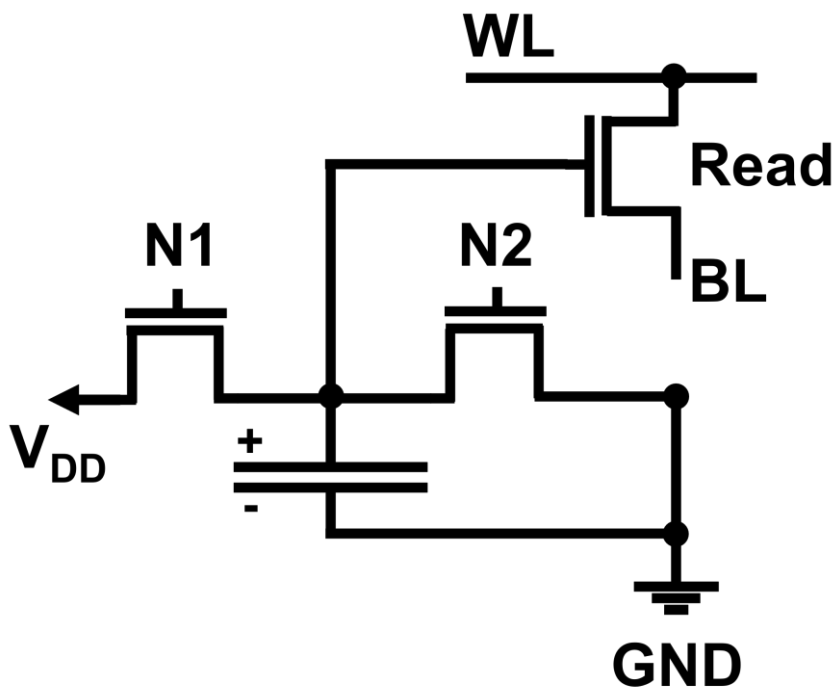

Figure 3.2.1 3T1C 회로도. 두 개의 update 트랜지스터, 하나의 read 트랜지스터, 하나의 커패시터로 이루어졌다.

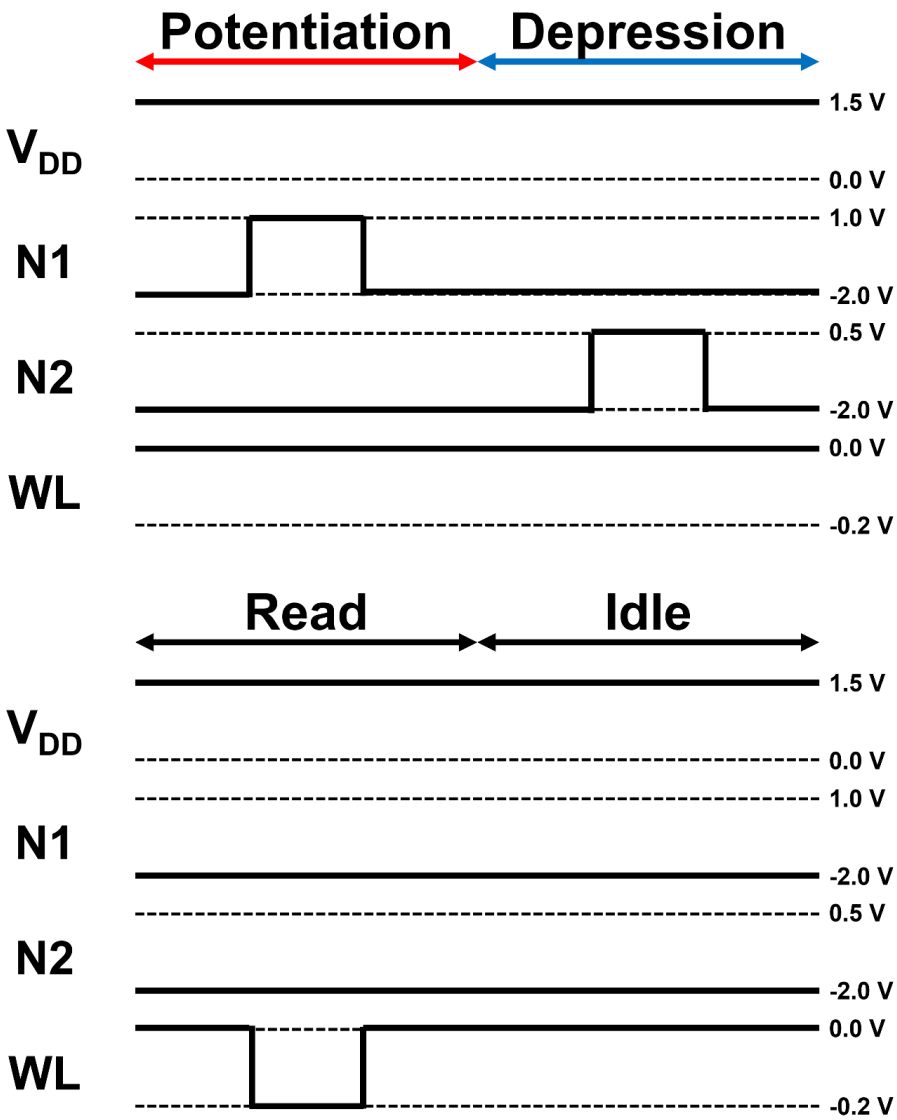

Figure 3.2.2 3T1C 동작 방식의 예시. 회로 동작은 크게 potentiation, depression, read, idle (data retention)으로 나눌 수 있다.

다만, 3T1C array를 제작하였을 때 random access가 가능하도록 select 해주는 소자가 필요하다. 이 문제는 Figure 3.2.3과 같이 update transistor의 gate에 AND gate를 추가하여서 해결할 수 있다. AND gate의 두 input이 동시에 들어왔을 때만 update transistor의 gate에 신호가 전달되어 array 내 특정 소자만 가중치를 갱신할 수 있다. AND gate를 구성하기 위해 추가 transistor가 필요하다는 단점이 있지만, a-IGZO는 CMOS BEOL 공정에 호환되기 때문에[2, 3] wafer 위에 CMOS 회로를 제작한 뒤, 그 위에 3T1C 소자를 수직 적층하는 방식으로 소자 면적을 최소화할 수 있다.

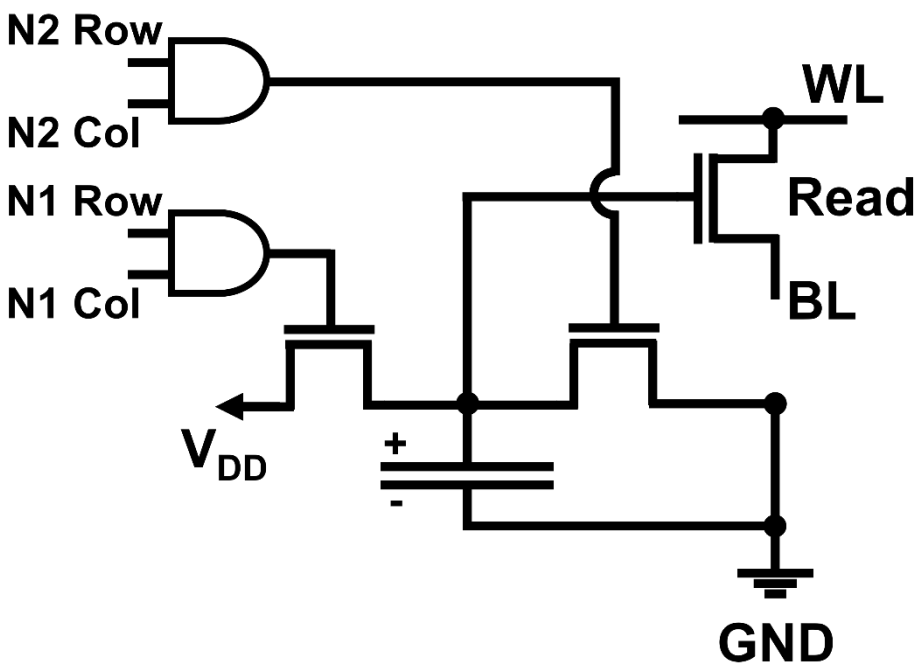

Figure 3.2.3 어레이에서의 3T1C 회로도. 선택 소자로 AND gate를 사용한다.

3.3 심층신경망 가속을 위한 시냅스 소자의 필요성

3T1C 시냅스 소자 및 array 측정은 Arduino DUE micro controller unit (MCU)로 구동되는 PCB 주변 회로로 진행하였다. 측정 환경은 Figure 3.3.1과 같다. 3T1C array 측정을 위해서는 많은 수의 pad가 연결되어야 하므로 45개의 probe가 일렬로 나열된 probe card를 사용하여 측정하였다.

MCU에 미리 시냅스 가중치 갱신, 가중치 retention, cycling endurance test 등의 모든 동작 시나리오를 프로그래밍한 뒤, 실험 조건을 제어 컴퓨터로 MCU와 주변 장비들에 명령을 내리면 조건에 맞추어 PCB 회로와 3T1C 소자에 적절한 전압 신호가 가해지고, 전류 신호를 읽어오는 방식이다. Update 트랜지스터에는 전압 신호가 가해지며, 읽기 과정에서는 read 트랜지스터의 drain에 전압을 가해준 뒤 흐르는 전류를 PCB의 적분기 회로와 MCU의 ADC로 10-bit 정확도를 가지는 정수로 출력한다(Figure 3.3.1). PCB 주변 회로에는 소자에 가해질 전압을 출력하는 power supply가 연결되어 있으며, MCU가 가하는 전압 신호에 따라 3T1C의 각 트랜지스터에 OFF 또는 ON 전압을 가한다. MCU가 가할 수 있는 전압 신호의 최소 길이는 약 100 ns이기 때문에 이보다 더 짧은 전압 신호를 가할 때는 81110A, 81115A 신호 발생기를 사용하였다. PCB 주변 회로에서 나오는 신호를 트리거로 사용하여 짧은 전압 신호를 3T1C 트랜지스터에 가할 수 있다.

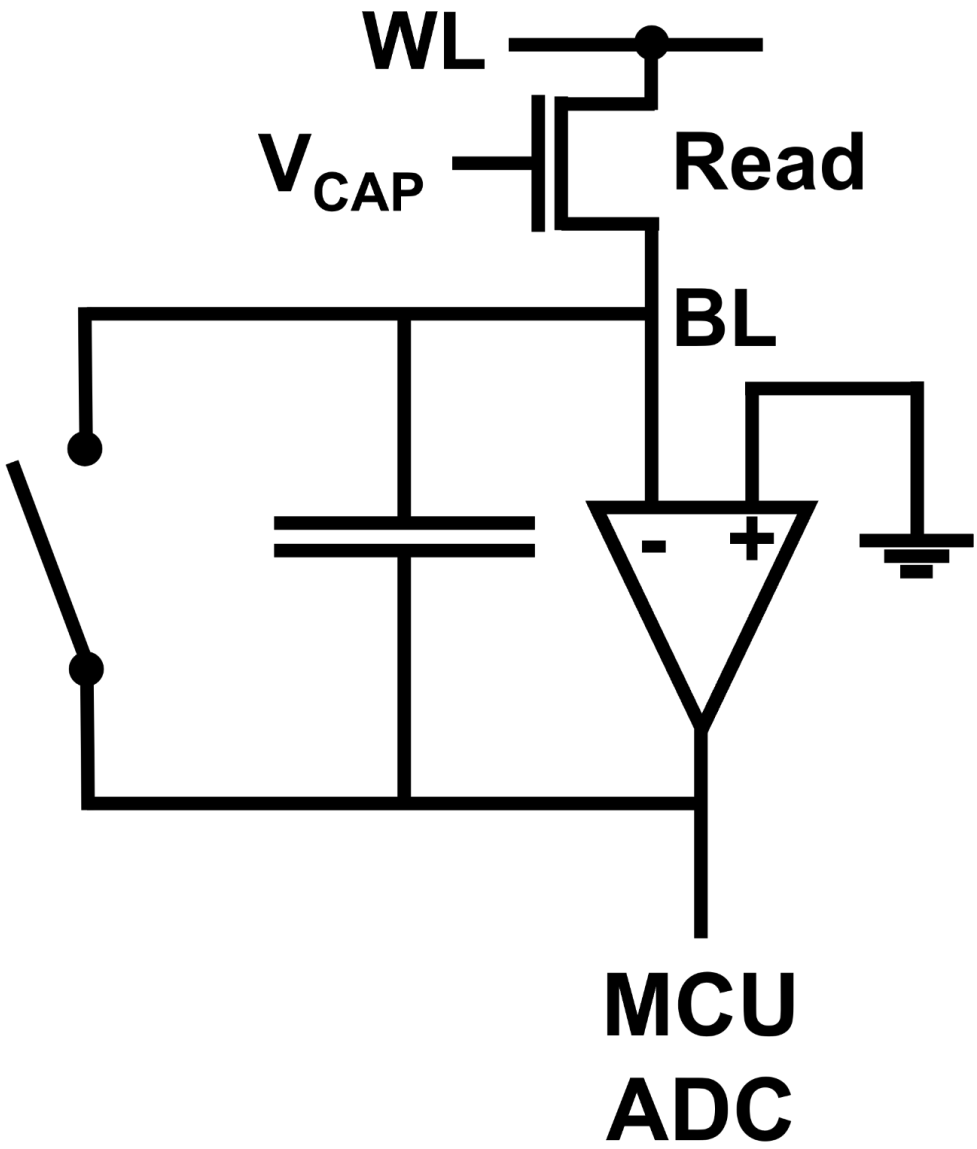

Figure 3.3.1 3T1C read 회로. OP amplifier가 BL을 GND로 잡아주는 동시에 적분 커패시터와 함께 read 동작 또한 수행한다.

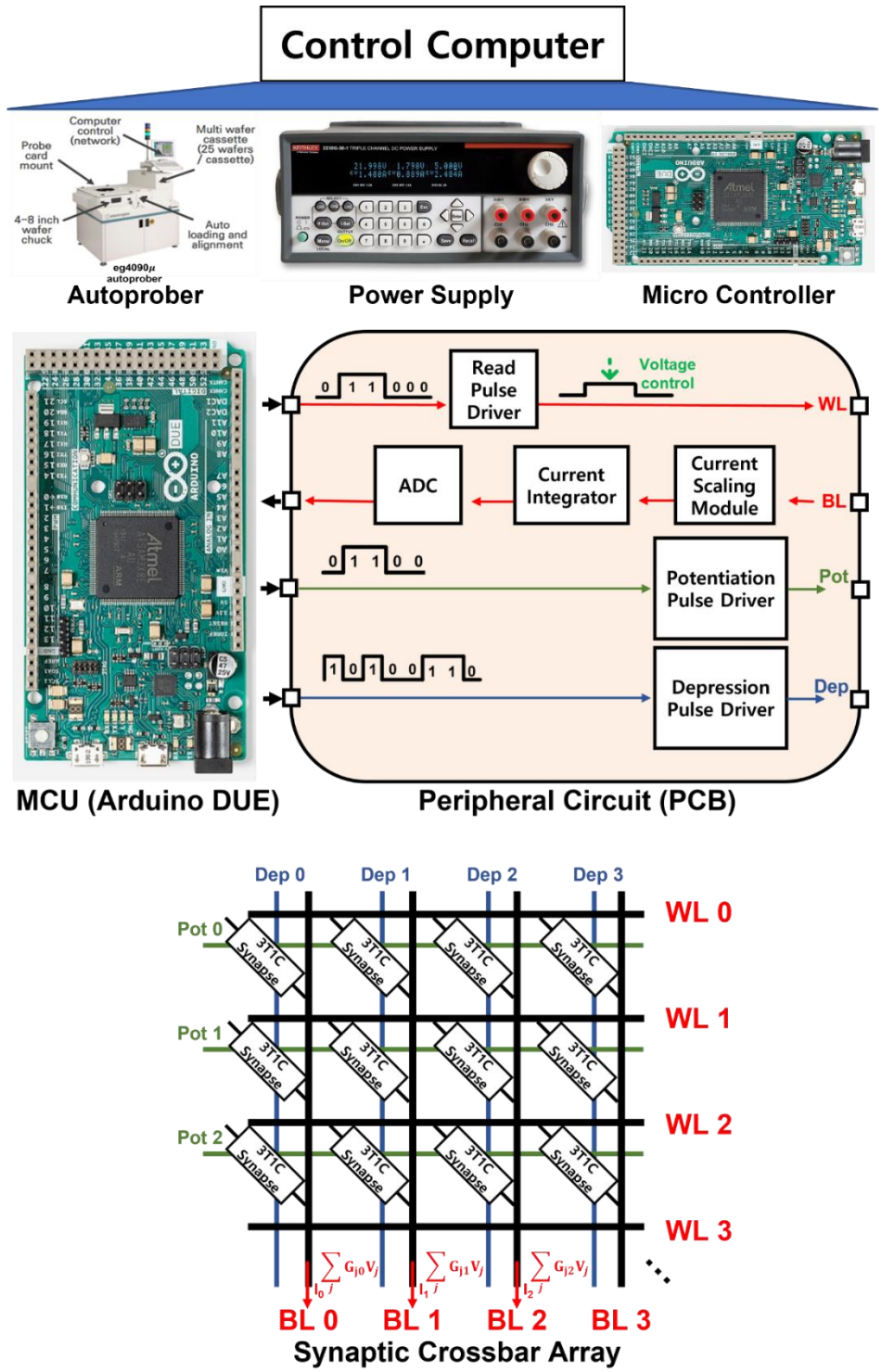

Figure 3.3.2 MCU, PCB를 이용한 주변회로와 측정 환경.

4. 결과 및 논의

4.1 3T1C 가중치 갱신 모델링

4.1.1 모델링 방식

선호되는 가중치 갱신 특성을 가질 수 있는 IGZO TFT로 제작된 3T1C 시냅스 소자이지만, 비이상적인 특성이 존재할 수 있다. 우선 커패시터에 전류를 주입하는 과정인 potentiation은 NMOS만으로 이루어진 소자에서는 완벽히 선형적일 수 없다. N1 트랜지스터의 source 전압은 커패시터의 전압이고, 커패시터에 저장된 전압에 따라 N1 트랜지스터의 V_{DS}와 V_{GS}가 변한다. 이처럼 potentiation에 따라서 N1 트랜지스터가 켜지는 정도가 변하고, 가중치 갱신량이 달라져 비선형적인 가중치 갱신이 일어날 수 있다. 또한, 반대 방향의 갱신 과정인 depression도 일정 구간에서는 비선형적인 갱신이 일어날 수 있다. 트랜지스터가 saturation 영역에서 동작하면 저장된 커패시터 전압에 무관하게 N2 트랜지스터에 일정 전류가 흘러 선형적으로 갱신되지만, 가중치가 충분히 작아져 N2 트랜지스터의 V_{DS}가 V_{GS}보다 작아진다면 트랜지스터가 linear 영역에서 동작해 비선형적인 갱신이 일어날 수 있다. 이런 이상적이지 않은 특성들을 파악, 개선하고, 최종적으로는 NeuroSim[7]이나 PyTorch와 같은 인공 신경망 시뮬레이터로 학습 정확도를 파악하기 위해서는 모델링이 필요하다. 따라서 본 연구에서는 유한요소법(Finite Element Method, FEM)을 이용한 근사 모델과 수식적으로 풀어낸 모델 두 가지로 3T1C의 가중치 갱신을 평가하였다.

모델링은 흔히 알려진 NMOS drain 전류 식인 수식 (3), (4)를 이용하였다. μ_n는 전자의 mobility, C_{ox}는 gate insulator의 단위 면적 당 커패시턴스, λ는 saturation region에서의 비이상적 전류를 표현하기 위한

파라미터이다.

$$I_{DS} = \mu_n C_{ox} \frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right](1 + \lambda V_{DS})$$

$$(V_{GS} > V_T, V_{DS} \leq V_{GS} - V_T) \quad (3)$$

$$I_{DS} = \frac{1}{2}\mu_n C_{ox} \frac{W}{L}(V_{GS} - V_T)^2(1 + \lambda V_{DS})$$

$$(V_{GS} > V_T, V_{DS} > V_{GS} - V_T) \quad (4)$$

FEM 모델링의 결과는 Figure 4.1.1.1과 같았다. 3T1C는 동작마다 하나의 트랜지스터만 동작하기 때문에 간단한 in-house MATLAB 모델을 제작하였다. 외부 전압 조건과 커패시터 전압에 따라 수식 (3), (4)를 이용하여 update 트랜지스터에 흐르는 전류를 계산한 뒤, 흐르는 전하량만큼 커패시터의 전압을 갱신하는 과정을 반복하였다. 전류를 계산할 때 필요한 파라미터들은 Figure 3.1.6에서 측정한 값을 사용하였다. 예측하였던 것과 같이 potentiation은 전반적으로 비선형성이 드러났고, depression은 낮은 커패시터 전압 영역에서만 비선형적인 가중치 갱신이 일어남을 확인하였다.

수식적으로 해를 구하는 방식 또한 시도하였다. Read 트랜지스터가 완벽하게 선형적인 V_G-I_{DS} 관계를 보인다고 가정하였을 때 가중치의 선형성은 커패시터 전압이 시간에 따라 얼마나 선형적으로 변화하는지를 분석하면 된다. Potentiation은 특이하게 수식 (5)에 의해 외부 조건인 V_{DD}, $V_{N1,ON}$ 조건에 의해서 saturation 또는 linear 영역에서 동작할지가 결정된다. 따라서 potentiation 전 구간을 하나의 수식으로 나타낼 수 있었고, 이는 수식 (6), (7)을 풀어서 얻을 수 있었다. 수식 (9)에서는 가중치 갱신량이 가중치에 대한 이차식 형태를 가진다는 것을 확인할 수 있었으며, 수식 (10)을 통해서는 saturation region에서의 potentiation

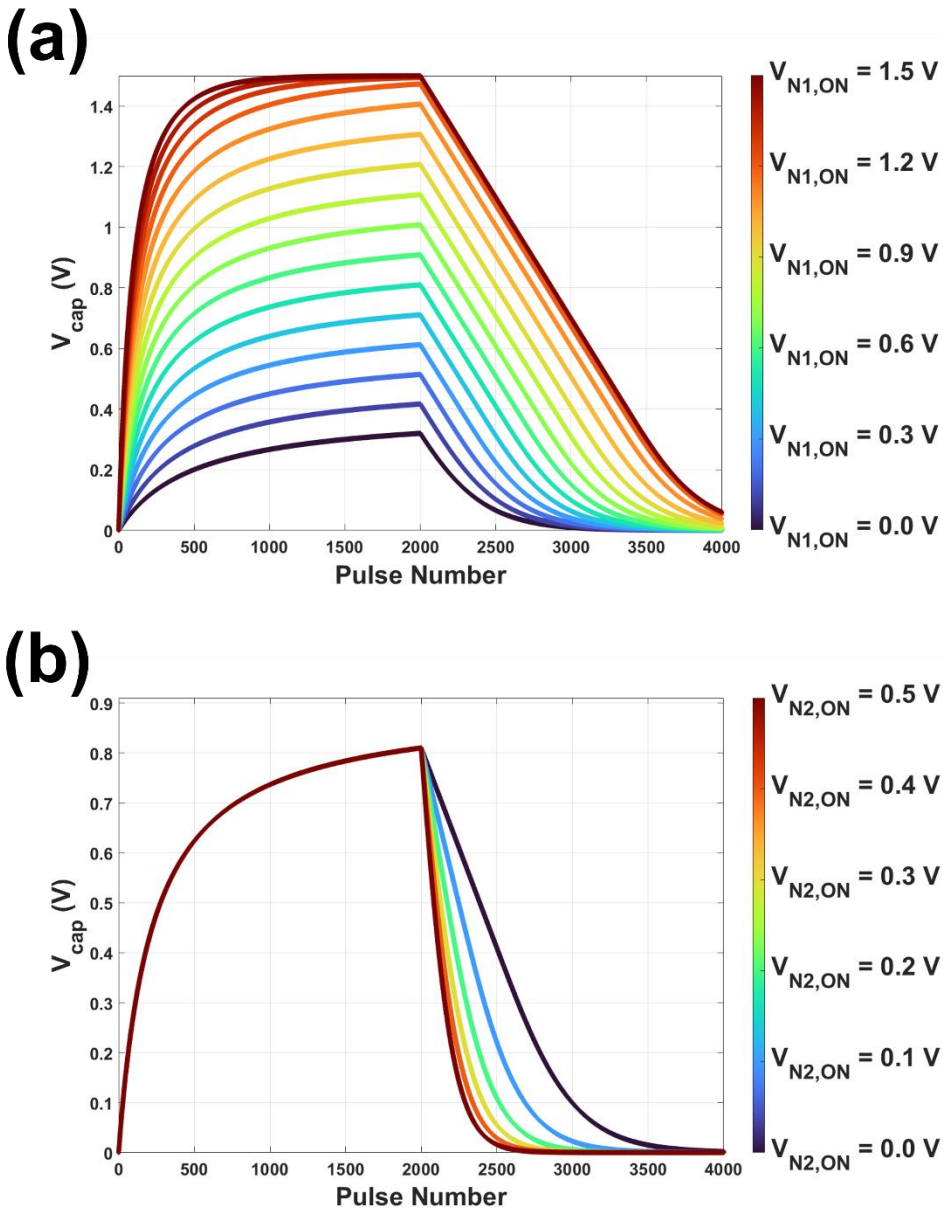

Figure 4.1.1.1 FEM 모델에서의 전압에 따른 가중치 갱신. (a)는 N1 트랜지스터의 gate에 가해지는 전압, (b)는 N2 트랜지스터의 gate에 가해지는 전압에 따른 potentiation-depression 경향성 변화 결과이다.

가중치 변화 양상을 예측할 수 있었다. 매우 작은 값인 λ는 무시하였다.

$$\begin{aligned} V_{DS} - V_{GS} + V_T &= (V_{DD} - V_{cap}) - (V_{N1,ON} - V_{cap}) + V_T \\ &= V_{DD} - V_{N1,ON} - V_T \end{aligned} \quad (5)$$

$$\frac{dI_{DS}\left(V_{cap}(t)\right)}{dV_{cap}(t)} \approx K\left(V_{cap}(t) - (V_G - V_T)\right) \quad (6)$$

$$V_{cap}(t) = \left(\frac{1}{C_{str}}\right)\int_0^t I_{DS}\left(V_{cap}(t')\right)dt' \quad (7)$$

$$C_{str}V''_{cap}(t) = KV'_{cap}(t)\left(V_{cap}(t) - (V_G - V_T)\right) \quad (8)$$

$$\frac{dG}{dn} \propto \frac{dV_{cap}(t)}{dt} = \frac{K}{2C_{str}}\left(V_G - V_T - V_{cap}(t)\right)^2 + D \quad (9)$$

$$V_{cap}(t) = (V_G - V_T) - \frac{1}{\frac{K}{2C_{str}}t + \frac{1}{V_G - V_T}} \quad (10)$$

Depression도 앞선 방법과 동일한 방식으로 수식적인 해를 구할 수 있었다. 다만 depression 과정은 가중치에 따라 N2 트랜지스터가 동작하는 영역이 달라지기 때문에 두 구간으로 나누어 해를 구했다. Saturation 영역에서 동작할 때는 항상 동일한 전류가 흐르기 때문에 간단하게 수식 (11)과 같은 해를 보이며, 커패시터 전압이 $V_G - V_T$보다 작아지는 구간부터는 수식 (12)-(16)과 같은 방식으로 해를 구할 수 있었다.

$$V_{cap}(t) = V_{max} - \frac{K}{2C_{str}}(V_G - V_T)^2 t \quad (11)$$

$$\frac{dI_{DS}\left(V_{cap}(t)\right)}{dV_{cap}(t)} \approx K\left((V_G - V_T) - V_{cap}(t)\right) \quad (12)$$

$$V_{cap}(t) = (V_G - V_T) - \left(\frac{1}{C_{str}}\right)\int_0^t I_{DS}\left(V_{cap}(t')\right)dt' \quad (13)$$

$$C_{str}V''_{cap}(t) = -KV'_{cap}(t)\left(V_{cap}(t) - (V_G - V_T)\right) \quad (14)$$

$$\frac{dV_{cap}(t)}{dt} = \frac{K}{2C_{str}}\left(V_G - V_T - V_{cap}(t)\right)^2 - \frac{K}{2C_{str}}(V_G - V_T)^2 \quad (15)$$

$$V_{cap}(t) = \frac{2(V_G - V_T)exp(-\frac{K}{C_{str}}(V_G - V_T)t)}{exp\left(-\frac{K}{C_{str}}(V_G - V_T)t\right) + 1} \quad (16)$$

Depression 또한 potentiation과 마찬가지로 이차식의 갱신량-가중치 관계식을 가졌으며, 수식 (10), (11), (16)을 이용해 potentiation-depression 과정을 예측한 결과는 Figure 4.1.1.2와 같다. FEM 모델과 수식적인 해가 동일한 가중치 갱신 양상을 가졌고, 모델링이 문제없이 되었다는 것을 확인하였다.

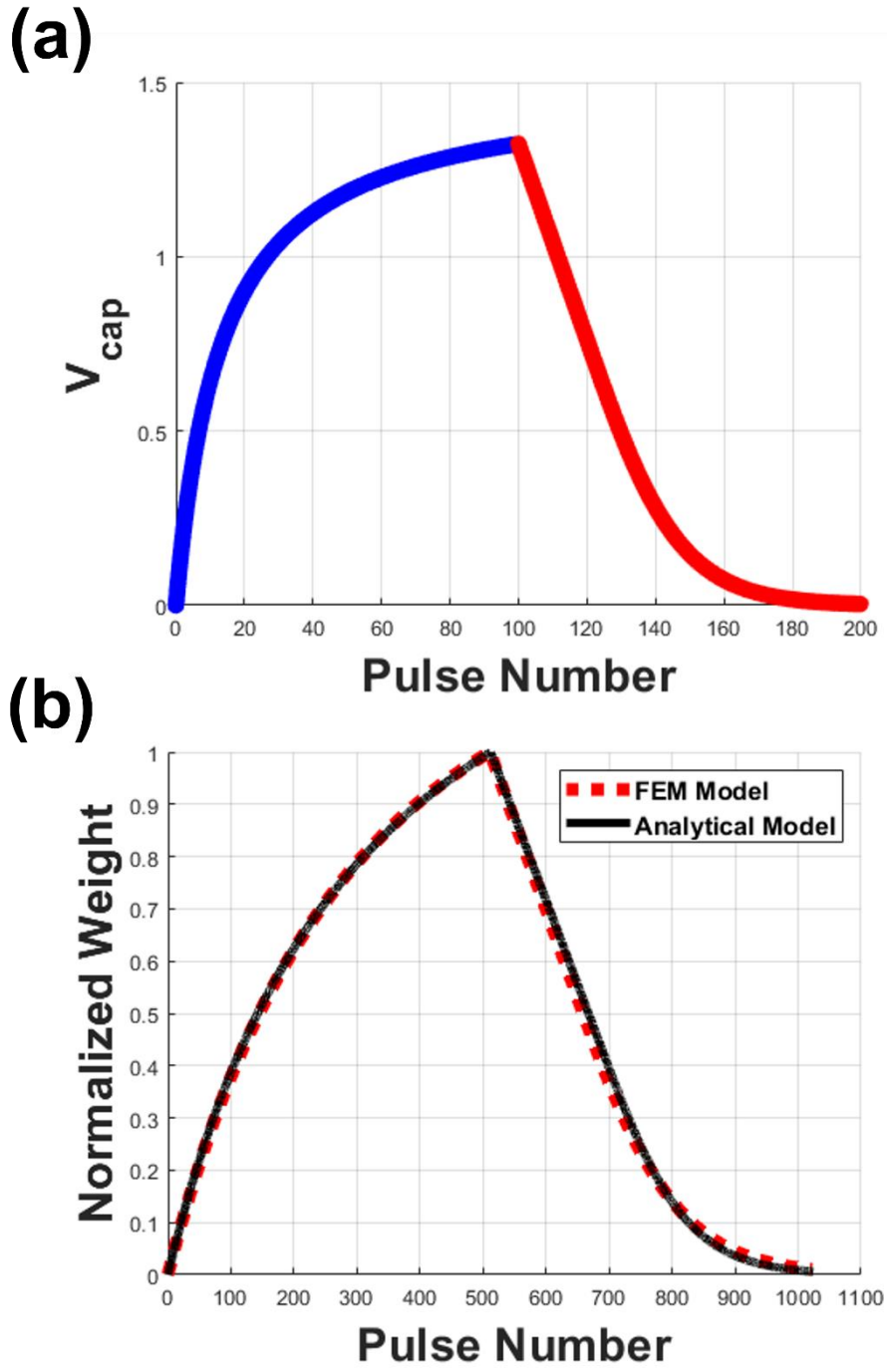

Figure 4.1.1.2 수식적인 모델의 결과. (a)는 수식적인 해를 통해 예측한 potentiation-depression 결과이며, (b)에서 수식적인 해와 FEM 모델이 일치하는 것을 확인하였다.

4.1.2 가중치 갱신의 선형 대칭성 평가 방법

최적의 3T1C 동작 전압 조건 탐색을 위해서는 가중치 갱신의 비선형성, 비대칭성의 정량적인 평가 방법이 필요하다. [6, 7, 32]과 같이 연구나 소자에 따라서 평가 방법이 다르지만, 대체로 지수함수 기반의 모델을 이용한다. 다만, 시냅스 소자에 따라 비선형성이 나타나는 물리적인 이유가 다르기 때문에 3T1C에 한해서는 새로운 선형 대칭성 평가 방법이 필요하다.

앞서 수식적으로 구한 해인 수식 (9)와 수식 (15)는 Brivio *et al.*의 연구와[4] 비슷한 형태이지만, 두 식은 시냅스 소자의 동작 영역에서의 비선형성, 비대칭성을 평가하기에는 알맞지 않다. 3T1C 가중치 갱신의 평가 지표가 될 수 있는 수식이 필요하며, 본 연구에서는 표준편차 기반의 수식 (17)과 수식 (18)을 통해 가중치 갱신의 선형성과 갱신량을 평가하였다.

$$\boldsymbol{NL = \frac{1}{< \Delta ADC >}\sqrt{\frac{1}{N}\Sigma(\Delta ADC - < \Delta ADC >)^2} \times 100\%} \qquad \textbf{(17)}$$

$$\boldsymbol{\alpha = \frac{< \Delta ADC >}{n}} \qquad \textbf{(18)}$$

NL은 가중치 갱신의 선형성을 평가하는 지표이며, α 는 가한 programming pulse 당 가중치 갱신의 양을 나타낸다. NL이 0에 가까울수록 평균에서 벗어나지 않는 선형적인 갱신을 의미하며 α가 클수록 한 programming pulse가 일으키는 가중치 갱신이 크다. $< \Delta ADC >$는 사용하는 커패시터 전압 구간에서의 평균적인 가중치 갱신을 의미하며, N은 읽은 지점들 수, n은 읽은 지점 사이에 가한 pulse 수를 의미한다. 일반적으로 통계 기반의 선형성 모델은 가중치 갱신의 경향을 정확하게 파악

하기에 불리하지만, IGZO 3T1C의 경우 potentiation, depression 모두 가중치 갱신이 정확하게 모델링 되기 때문에 이러한 평가법을 사용해도 문제가 없다.

Potentiation과 depression의 비대칭성은 수식 (19), (20)으로 선형성의 대칭성과 갱신량의 대칭성을 평가하였다.

$$\boldsymbol{max}\left(\frac{\boldsymbol{NL_{pot}}}{\boldsymbol{NL_{dep}}}, \frac{\boldsymbol{NL_{dep}}}{\boldsymbol{NL_{pot}}}\right) \quad (\mathbf{19})$$

$$\boldsymbol{max}\left(\frac{\boldsymbol{\alpha_{pot}}}{\boldsymbol{\alpha_{dep}}}, \frac{\boldsymbol{\alpha_{dep}}}{\boldsymbol{\alpha_{pot}}}\right) \quad (\mathbf{20})$$

4.2 3T1C 가중치 갱신

4.2.1 측정 결과와 모델링 비교

가중치 갱신 실험은 우선 N2 트랜지스터를 충분히 긴 시간 동안 켜서 커패시터 전압을 0 V로 만든 뒤, potentiation과 depression 순서대로 진행하였고, 시냅스 가중치는 read 트랜지스터에 흐르는 전류를 analog to digital converter (ADC)를 통해 디지털화하여 측정하였다. 커패시터에 저장된 전압은 ADC 측정값으로부터 간접적으로 계산하였다. Figure 4.2.1.1과 같이 읽기 조건과 동일한 V_D, V_S에서의 read 트랜지스터의 transfer curve를 측정한 뒤, transfer curve에서 변환된 I_{DS}에 대응되는 전압으로 커패시터에 저장된 전압을 추측하였다. I_{DS} 값은 ADC 측정값을 수식 (21)를 통해 변환하였다. 측정한 transfer curve 데이터에 해당하는 전류 값이 없는 경우는 선형 보간법을 통해 커패시터 전압을 산출하였다. C_{int}와 V_{int}는 적분기 커패시터의 커패시턴스와 전압이며, ADC가 0–3.3 V를 10–bit precision으로 변환하기 때문에 수식 (21)와 같이 계산하였다.

$$\boldsymbol{I_{DS}} = \frac{\boldsymbol{Q}}{\boldsymbol{t}} = \frac{\boldsymbol{Q_{int}}}{(\boldsymbol{read\ time})} = \boldsymbol{C_{int}} \times \frac{\boldsymbol{V_{int}}}{(\boldsymbol{read\ time})}$$
$$= \boldsymbol{C_{int}} \times (\boldsymbol{ADC\ value}) \times \frac{\frac{\boldsymbol{3.3}}{\boldsymbol{1023}}}{(\boldsymbol{read\ time})} \qquad (\boldsymbol{21})$$

측정 결과는 Figure 4.2.1.2와 같다. FEM 모델이 예측한 가중치 갱신의 경향과 실제 측정된 potentiation–depression 경향이 상당 부분 일치하는 것을 확인하였다. 모델과 측정 사이의 차이는 update 트랜지스터들의 산포, read 트랜지스터의 비선형성 등의 효과에 의해 나타났다고 추정하였다.

IGZO TFT는 n-type만 존재하기 때문에 예상하였던 것처럼 potentiation 과정에서 약간의 비선형성을 확인할 수 있었고, depression도 높은 커패시터 전압 구간에서는 선형적인 갱신이 일어나지만 낮은 전압에서는 비선형적인 갱신이 일어나는 것을 확인하였다.

Figure 4.2.1.3에서는 potentiation의 $V_{cap}-dV_{cap}$을 나타냈다. 수식 (9)이 예측한 이차식의 경향이 나타났으며, 이차식을 fitting 했을 때 V_G-V_T의 값 또한 실험에 사용한 소자와 동작 조건에 일치하였다.

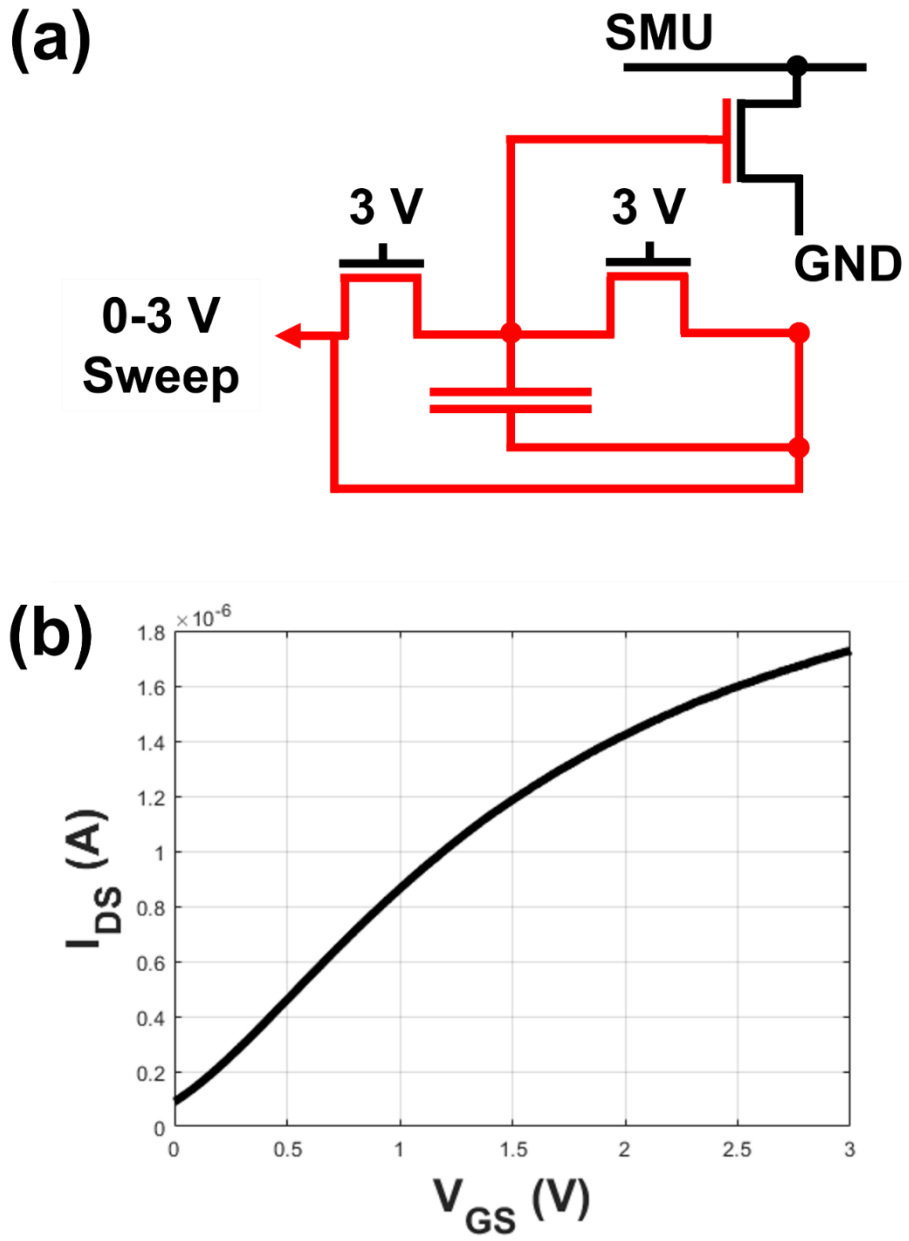

Figure 4.2.1.1 Read 트랜지스터 (a)측정 방법과 (b)결과. N1과 N2를 충분히 키면 V_{DD}/GND 전압이 온전히 read 트랜지스터의 gate에 전달될 수 있다.

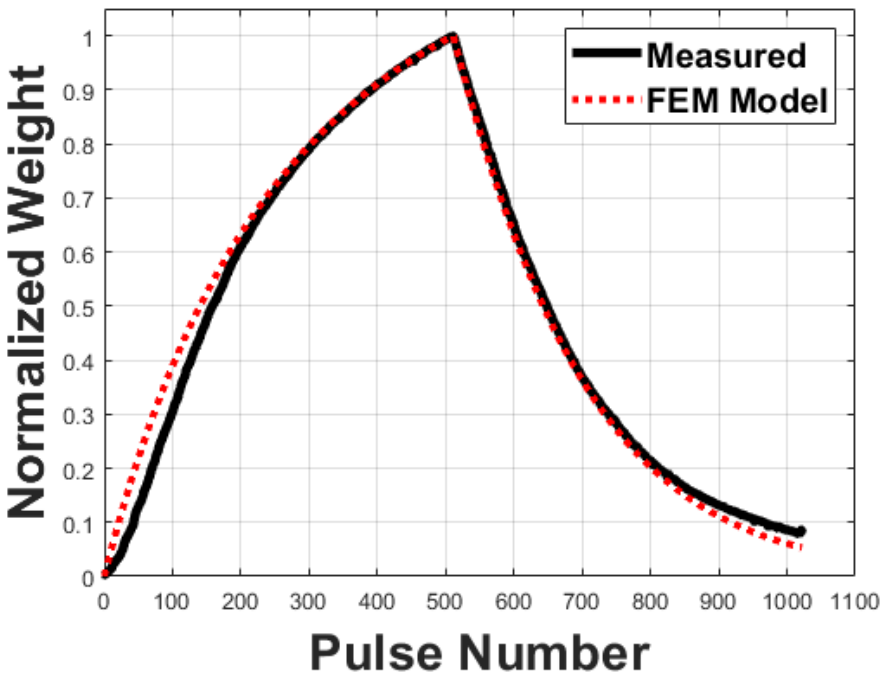

Figure 4.2.1.2 가중치 갱신의 측정값과 FEM 모델의 비교

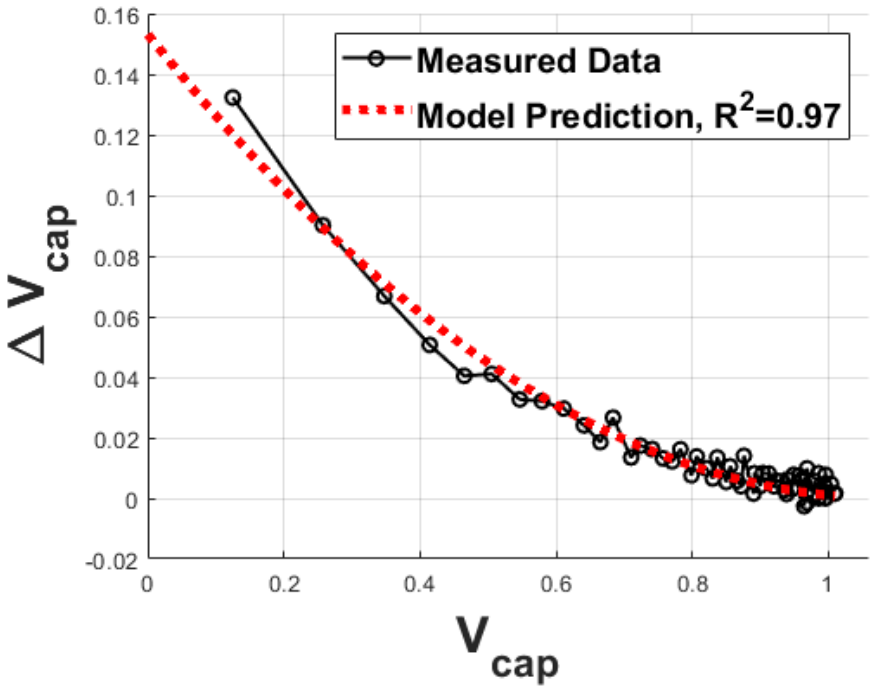

Figure 4.2.1.3 가중치 갱신 변화의 측정값과 FEM 모델의 비교

4.2.2 시냅스 소자의 고속 동작

비휘발성 메모리가 아닌 트랜지스터를 이용하여 시냅스를 제작했을 때의 장점 중 하나는 가중치 갱신 동작이 고속으로 가능하다는 것이다. IGZO TFT는 수 ns 수준으로 동작할 수 있다는 사실이 보고된 바 있으며[50], 최적화를 통해 더 빠른 동작 또한 가능할 것이다. 또한, 보다 작은 면적의 커패시터를 사용하기 위해서는 한 번의 갱신 때 N1, N2 트랜지스터에 흐르는 전류량을 자유롭게 조절할 수 있어야 한다. 신호발생기로 ns 수준의 가중치 갱신 실험 결과는 Figure 4.2.2.1과 같다. 8 ns 갱신 펄스까지 시냅스가 정상 동작하였고, 이를 통해 3T1C 시냅스의 고속 동작 가능성뿐만 아니라 scalability, 저전력 동작 또한 확인할 수 있었다.

표 2 3T1C 소자 동작 속도 및 전력 소비 예상

	MCU Peripheral	FPGA Peripheral*
Program Speed	1 μs	< 5 ns
Read Speed	10 - 20 ms	< 8 μs
Program Energy	< 1 pJ	< 1 fJ
Read Energy	< 10 nJ	< 3 pJ

*FPGA와 tape-out 주변회로 chip을 사용한 측정

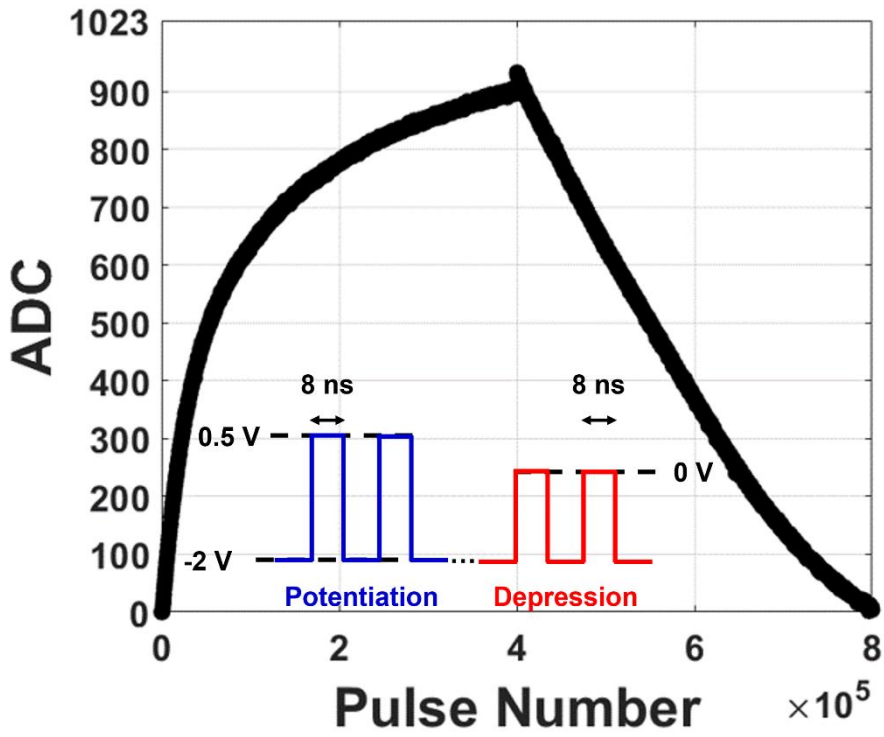

Figure 4.2.2.1 ns 수준의 시냅스 동작

4.2.3 전압 조건에 따른 가중치 갱신

수식 (9)에서는 항상 커패시터 전압 변화가 전압의 이차식에 비례하는 것으로 계산이 되지만, 실제 시냅스 동작에서는 비교적 선형적인 가중치 갱신이 측정되었다. 특히 Figure 4.1.1.1의 FEM 시뮬레이션 결과에서도 전압 조건에 따라서 시냅스의 거동이 달라지는 것을 확인할 수 있었다. 높은 학습 정확도를 가지는 3T1C array 구동을 위해서는 전압 조건에 따른 시냅스 가중치 갱신의 경향을 파악, 최적화하는 과정이 필요하다.

Figure 4.2.3.1은 update 트랜지스터 동작 전압 조건에 따른 가중치 갱신의 선형성과 대칭성을 나타냈다. Potentiation은 N1 gate 전압이 증가할수록 선형성은 향상되었으며, α 는 증가하는 경향을 보였다. Depression도 N2 gate 전압 증가에 따라 α 가 증가하였지만, potentiation과 반대로 비선형성을 나타내는 NL은 높은 gate 전압이 가해질수록 증가하였다.

동작 전압 상승에 따른 α 증가는 수식 (3), (4)으로 설명 가능하였다. 당연하지만 트랜지스터의 gate 전압이 클수록 더 큰 I_{DS}가 흐르기 때문에 한 갱신 신호 당 변하는 ADC 값이 커진 것이다. Potentiation에서 선형성 향상은 Figure 4.2.3.2로 설명할 수 있다. 전압 갱신량은 수식 (9)에서 증명하였듯이 커패시터 전압에 대한 이차식의 관계를 가지는데, N1 트랜지스터의 동작 전압이 상승하면 V_G-V_T가 증가해 동일한 전압 범위에서는 선형성이 향상된다. Depression의 선형성이 potentiation과 반대의 경향을 가지는 이유는 N2 트랜지스터가 linear mode로 동작하는 전압 구간이 달라지기 때문으로 설명하였다. N2 트랜지스터의 동작이 saturation에서 linear 영역으로 바뀌는 경계는 수식 (4)에서 V_G-V_T이다. 동작 전압이 증가하면 V_G-V_T가 증가하여 더 높은 전압 범위부터 linear 영역에서 동작하고, 선형적인 갱신이 일어나는 saturation 영역에

서 동작하는 범위가 줄어 선형성이 나빠지는 것이다. 다만 depression 과정이 대체로 linear 하기 때문에 potentiation만큼의 변화는 없는 것으로 추정되었다.

실험 결과를 통해 선형적인 갱신을 위해서는 potentiation은 높은 gate 전압, depression은 낮은 gate 전압에서 동작해야 한다는 결론을 내릴 수 있었다. 다만 이 경우 선형성과 선형성의 대칭성(수식 (19))은 개선할 수 있지만, 갱신량의 대칭성(수식 (20))은 악화되는 trade-off가 있다. 이는 Figure 4.2.3.3과 같이 커패시터의 하단 전극, V_{DD}, potentiation 전압을 향상시키는 것으로 해결할 수 있다. Potentiation 과정의 전압은 모두 동일한 양이 상승하였기 때문에 갱신에 변화가 없지만, N2 트랜지스터에서는 V_D가 크게 증가한 효과이기 때문에 커패시터에 저장된 전압에 무관하게 항상 saturation 영역에서 동작하게 할 수 있다. Depression 트랜지스터에서 더 이상 선형성을 위해 낮은 전압을 사용할 필요가 없어지기 때문에 자유롭게 갱신량의 대칭성도 해결된다. 커패시터 하단 전극 전압의 조절로 read 트랜지스터의 가장 선형적인 구간을 사용할 수 있다는 부가 효과도 얻을 수 있다.

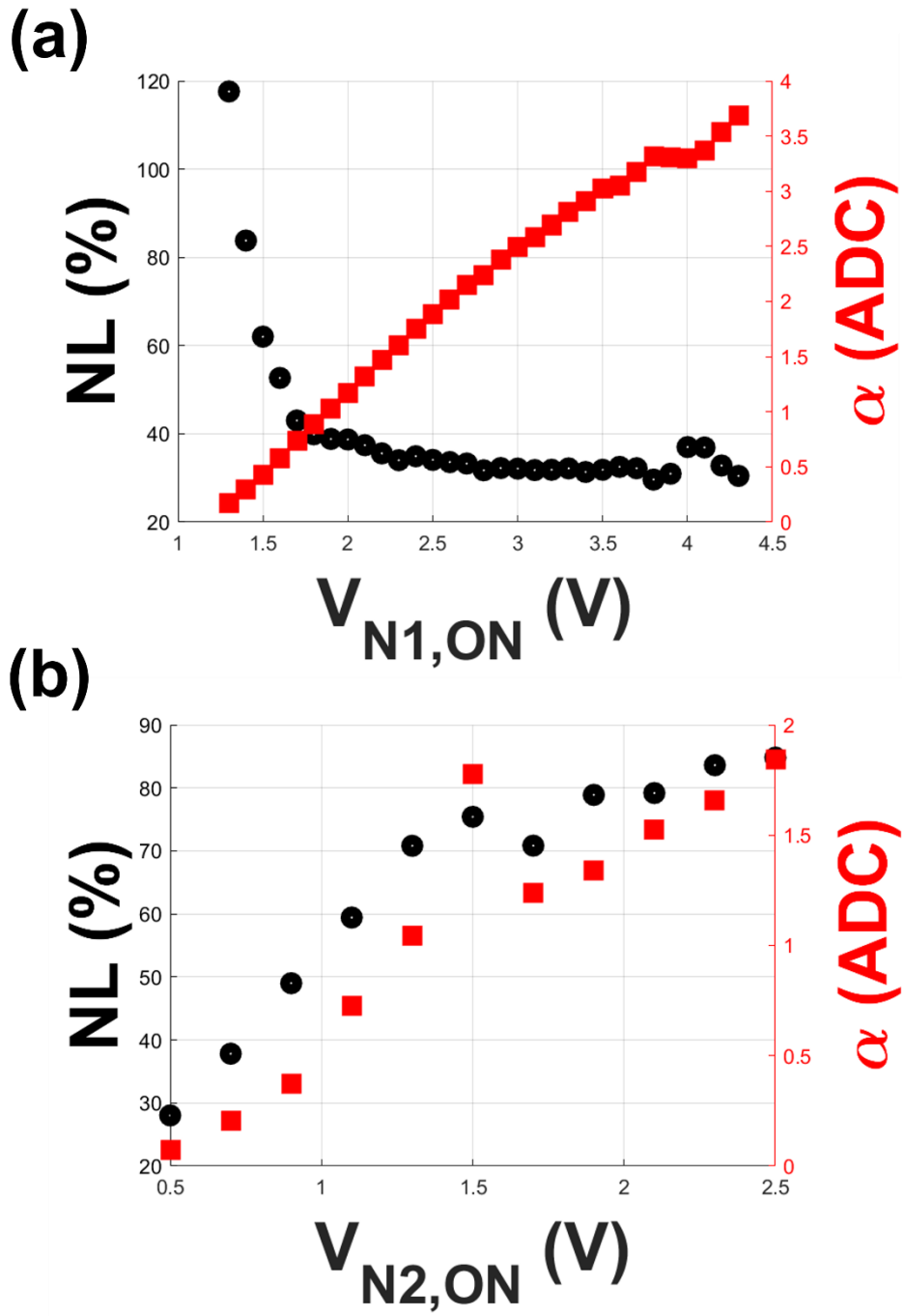

Figure 4.2.3.1 전압에 따른 가중치 갱신의 선형성. (a)는 N1 gate 전압에 따른 potentiation 변화, (b)는 N2 gate 전압에 따른 depression 변화이다.

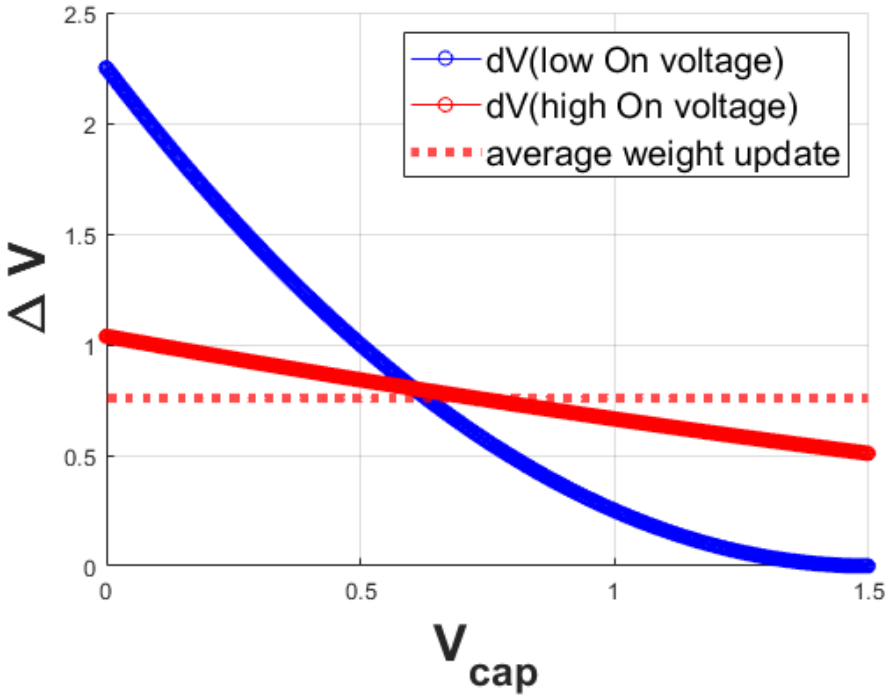

Figure 4.2.3.2 높은 전압에서의 선형적 갱신 해석. 낮은 N1 gate 전압(blue)에 비해 높은 N1 gate 전압(red)이 평균 가중치 갱신량이 같을 때 더 선형적인 가중치 갱신이 일어난다.

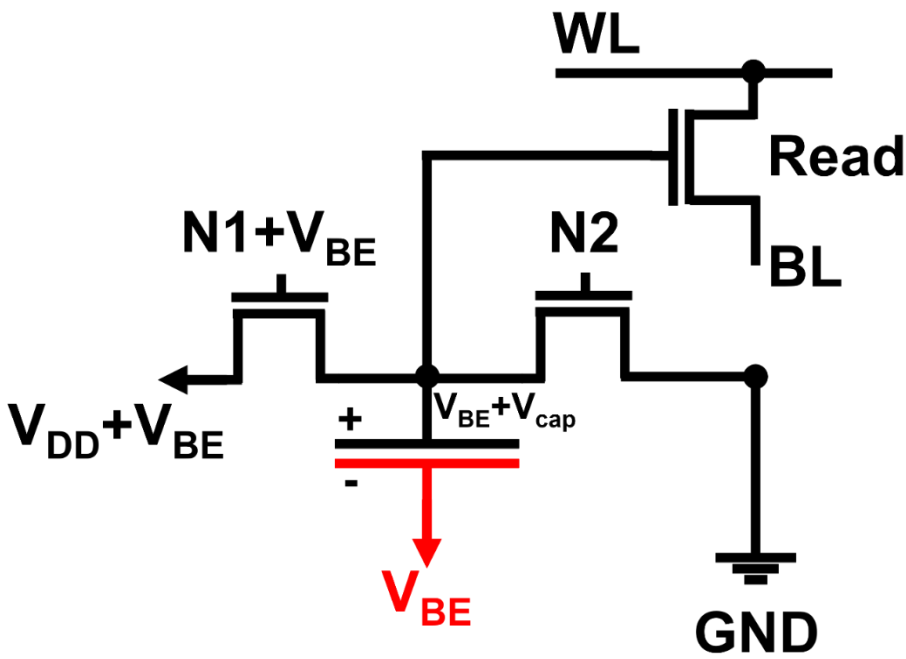

Figure 4.2.3.3 커패시터 하단 전극 boosting. 하단 전극에 V_{BE}가 가해지면 상단 전극에는 $V_{BE}+V_{cap}$이 가해져 N2 트랜지스터의 source-drain 사이 더 큰 전압이 걸리고 saturation region에서 동작하게 된다.

4.2.4 시냅스 산포 평가

3T1C 시냅스 간 산포는 device-to-device (D2D) 산포와 cycle-to-cycle (C2C) 산포 두 종류로 분류할 수 있다. 인공 신경망에서의 아날로그 연산은 fault-tolerant한 것으로 알려졌지만, 산포가 일정 수준을 넘는 경우 on-chip learning의 최종 정확도에 영향을 준다. 특히 D2D는 산포가 보정되는 방향으로 학습이 되며 최종 정확도에 큰 영향을 주지 않지만, C2C 산포 증가는 정확도 하락의 원인이 되는 것으로 알려져 있다[15].

D2D, C2C 산포는 동일 wafer 위의 소자들을 비교하였다. C2C 산포는 한 소자에서 10회 반복 측정해 평가하였고 D2D 산포는 한 die 위의 25개 소자를 비교하였다. Figure 4.2.4.1에서 시냅스가 일관성 있게 동작하는 것을 확인하였다. 낮은 ADC 값에서는 산포가 상대 표준편차가 큰 것으로 나타나는데, 이는 ADC 회로의 noise에 의한 것으로 판단된다. 시냅스 상의 C2C 산포는 무시 가능하며, 커패시터 전압에 무관하게 존재하는 주변 회로의 noise에 의해 발생하였기 때문에 Figure 4.2.4.1(b)와 같은 경향성이 나타났다. 이와 같은 부분은 실제 tape-out 주변 회로를 사용하면 개선이 있을 것으로 기대된다. Figure 4.2.4.2에서 소자 간 산포는 C2C에 비해 큰 것을 확인하였다. 그러나 D2D 산포 또한 낮은 ADC 영역을 제외한다면 상대 표준편차 20% 이하의 일관성 있는 potentiation-depression이 일어났음을 Figure 4.2.4.2(b)에서 볼 수 있다. 소자가 모두 동작하며, D2D 산포가 심각하지 않기 때문에 학습 능력에는 영향이 없을 것으로 판단하였다. D2D 산포는 공정 최적화를 통해 추가 개선이 가능하다.

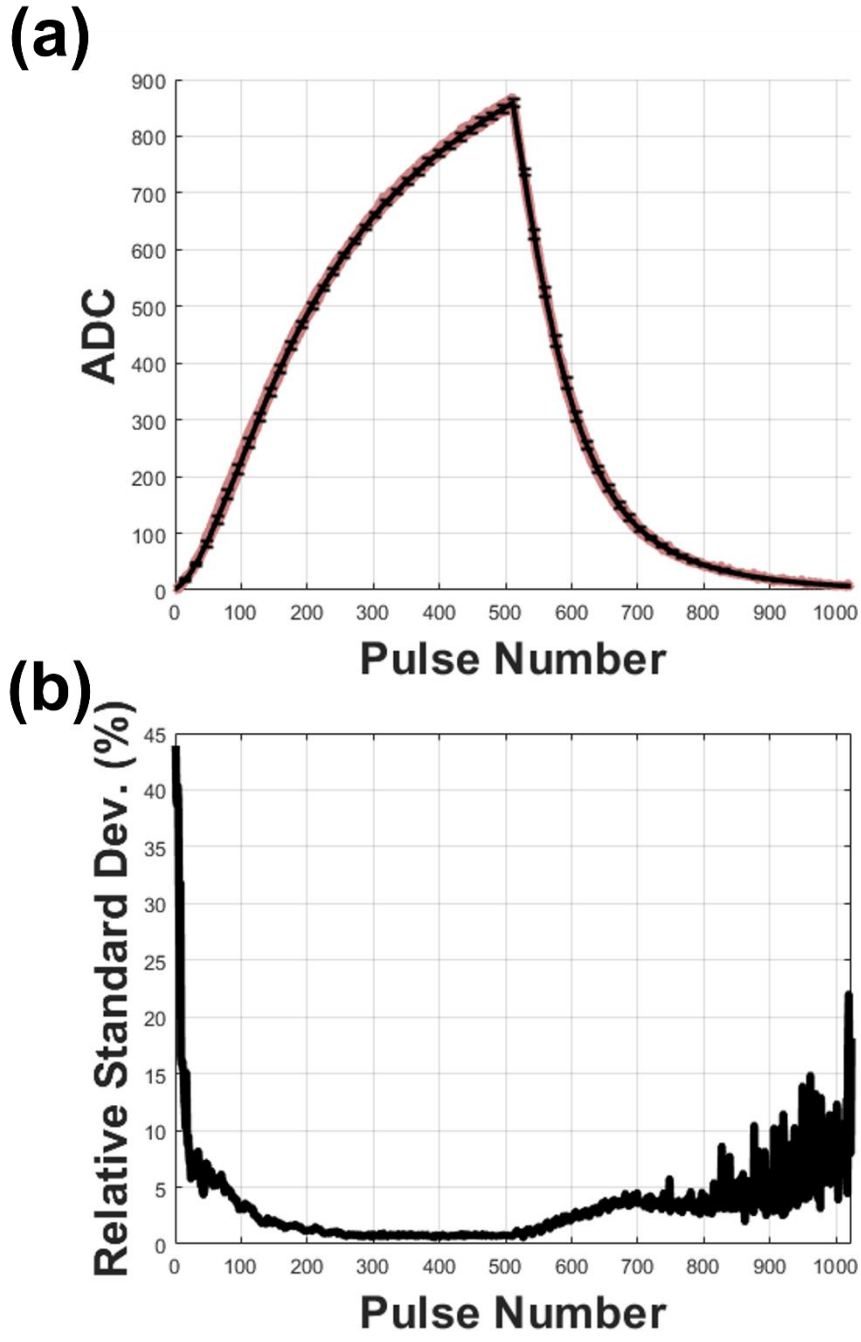

Figure 4.2.4.1 시냅스 소자 가중치 갱신의 cycle-to-cycle 산포. (a)는 10회의 potentiation-depression 과정을 나타낸다. (b)는 pulse number에 따른 상대 표준편차를 나타낸다.

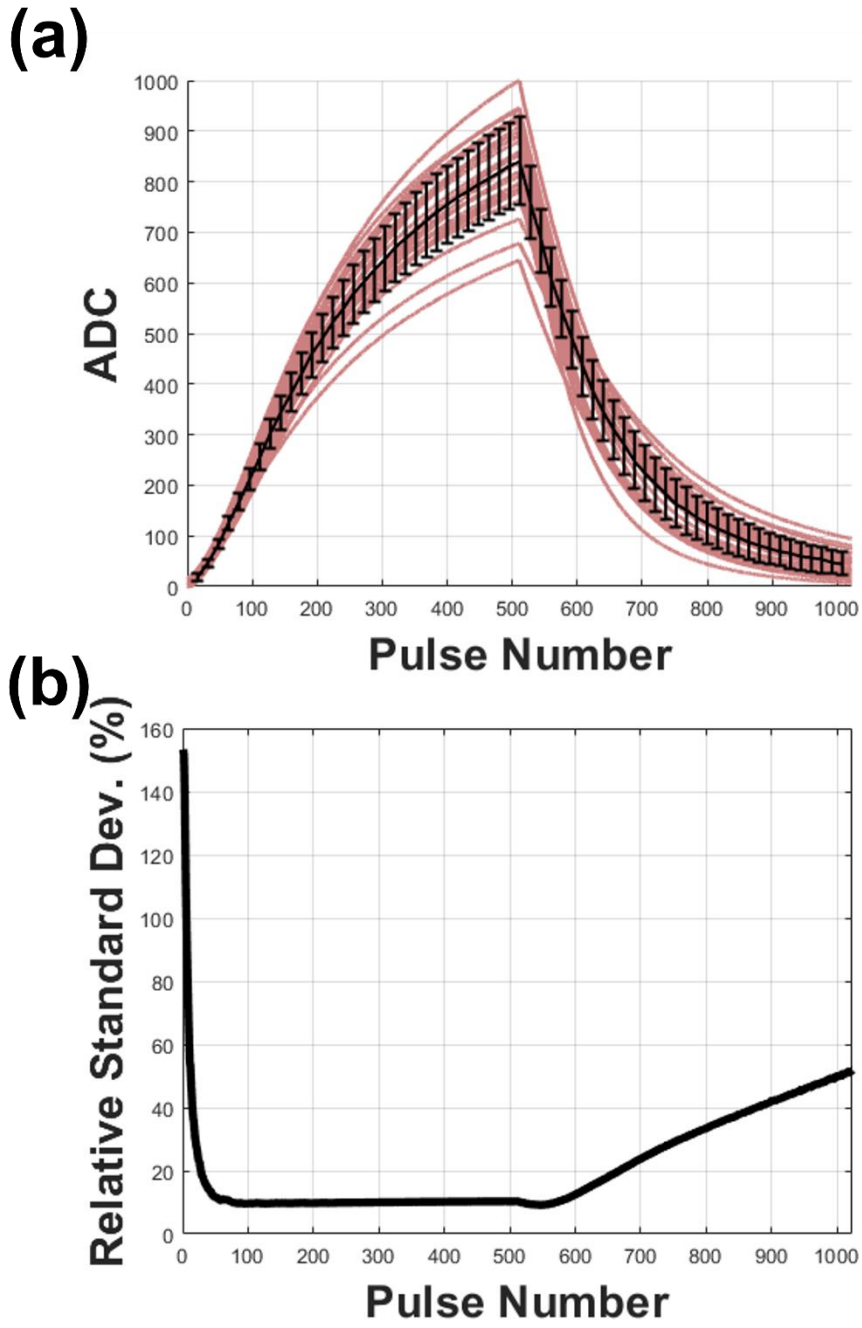

Figure 4.2.4.2 시냅스 소자 가중치 갱신의 device-to-device 산포. (a)는 25개의 device 간 산포를 나타낸다. (b)는 pulse number에 따른 상대 표준편차를 나타낸다.

4.2.5 목표 가중치 도달 능력 확인

시냅스가 인공 신경망 내에서 의도한 대로 동작할 수 있는지 검증하기 위해 시냅스의 가중치가 목표하는 값으로 수렴하는지를 확인하였다. 목표하는 ADC 값을 설정한 뒤, 현재 ADC 값과 목표를 비교하여 가중치를 갱신하는 과정을 반복하여 목표에 수렴할 수 있는지를 확인하였다. 시냅스를 동작한 방식은 Figure 4.2.5.1와 같다. 확률적인 update는 Gokmen *et al.*의 연구에서 입력값과 역전파된 오차 값의 곱을 별도의 연산 장치 없이 계산하고자 사용되었던 방법이며[15], 본래라면 N1, N2의 gate에 연결된 AND gate의 두 입력 값이 모두 도달할 때 update가 되는 방식이지만, 본 연구에서는 AND gate를 구현할 수 없었기 때문에 N1, N2의 gate에 직접 MCU 소프트웨어 상에서 계산된 확률적 update 신호를 인가하였다.

실험 결과는 Figure 4.2.5.2와 같았으며, 가중치의 시작점과 목표에 무관하게 수렴하였다. Potentiation의 경우 Figure 4.2.1.2에서 확인하였던 바와 같이 가중치가 높을수록 포화되는 현상이 있다. Potentiation과 depression으로 동일한 ADC 차이를 갱신하는데 필요한 update cycle 수를 비교했을 때, 적은 ADC 갱신은 potentiation과 depression 간 차이가 없었던 반면, 더 큰 간격의 갱신에서는 potentiation이 depression 보다 더 느리게 수렴하였다. 그러나 수렴 속도에 큰 차이는 없었으며, 모든 실험에서 목표에 수렴하였기 때문에 3T1C의 시냅스 소자로의 가능성을 검증할 수 있었다.

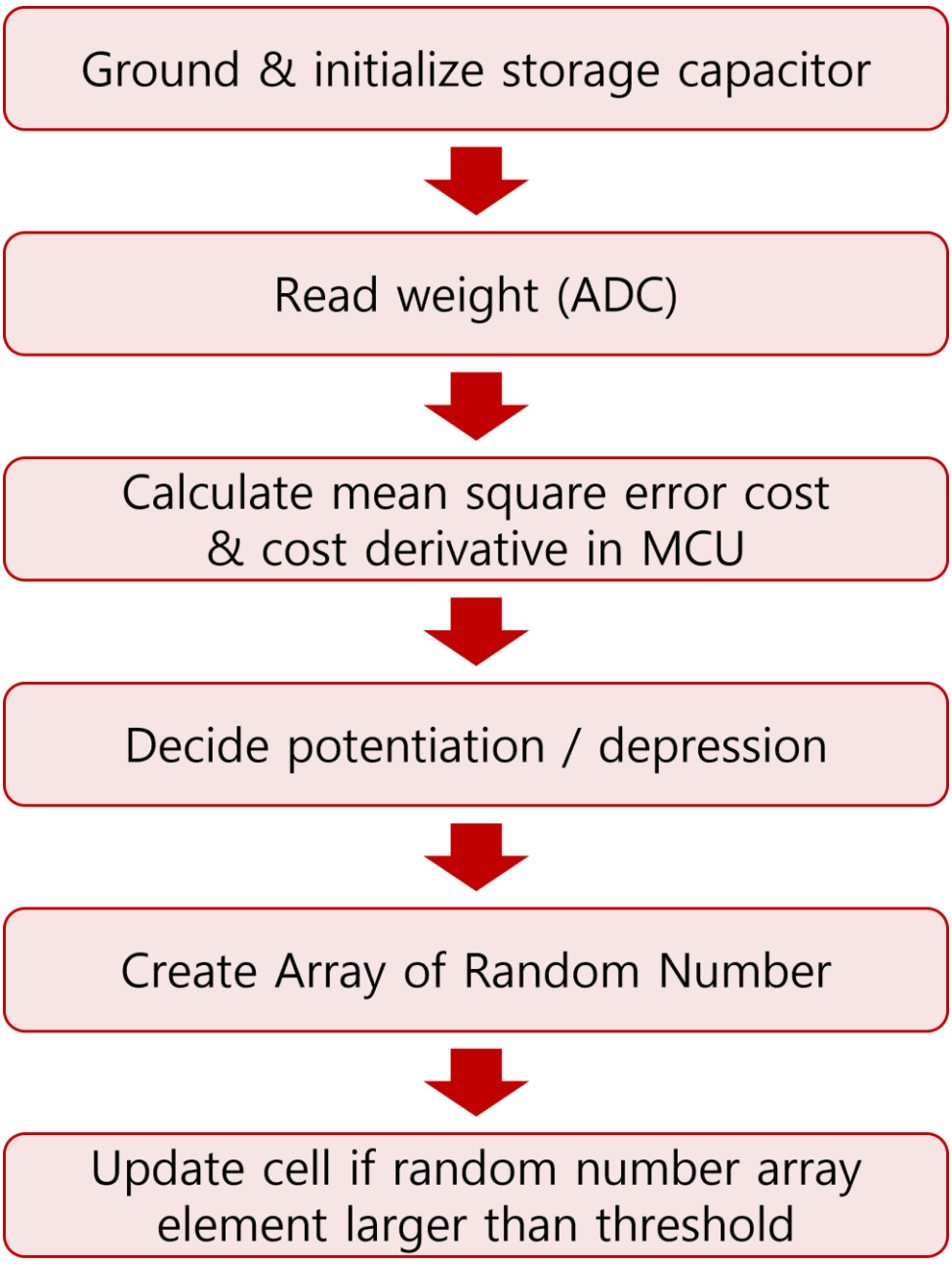

Figure 4.2.5.1 목표 가중치 도달 실험 방법.

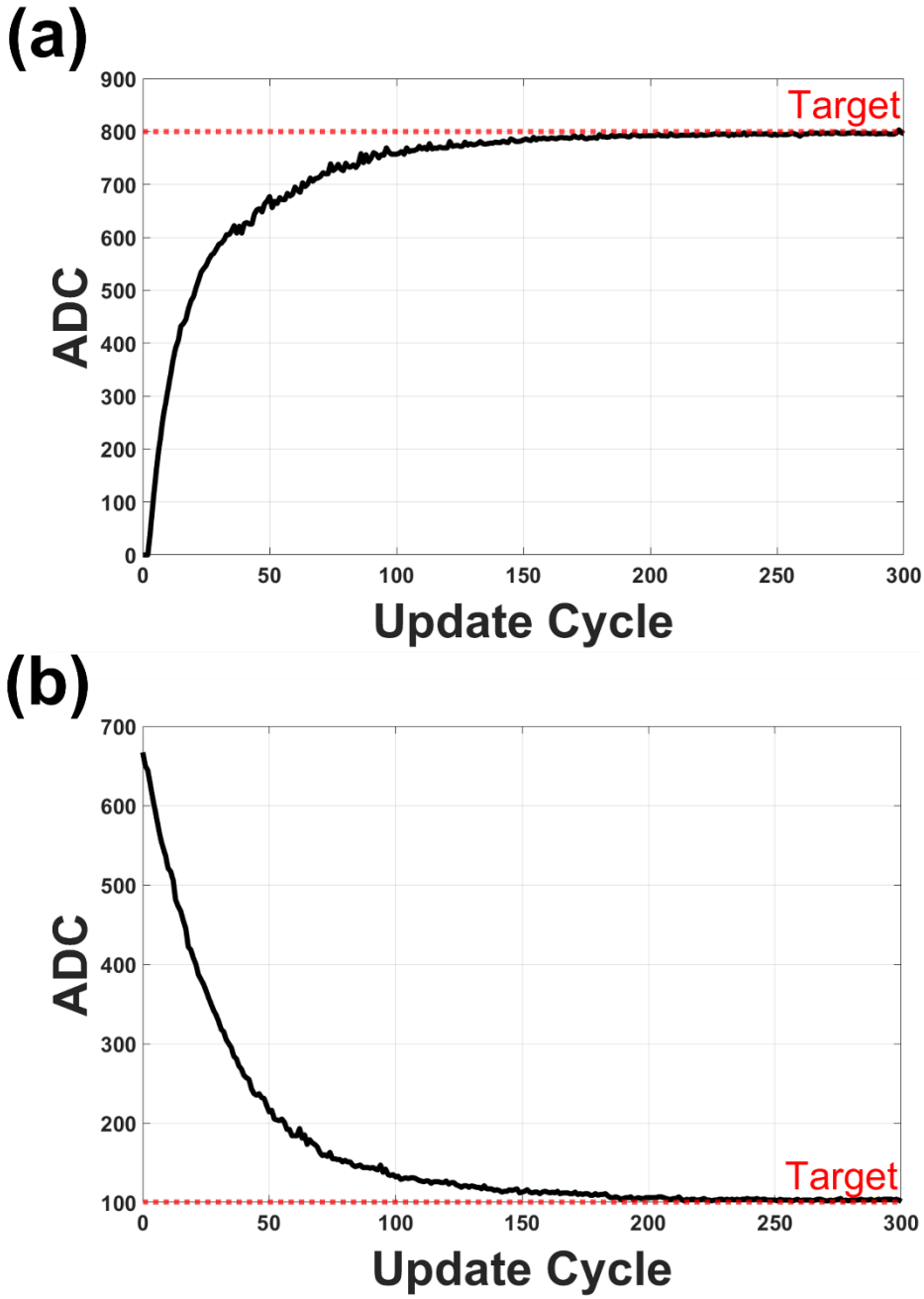

Figure 4.2.5.2 목표 가중치 도달 실험 결과. (a)는 potentiation을 사용한 가중치 갱신, (b)는 depression을 사용한 가중치 갱신이다.

4.3 가중치 retention

4.3.1 가중치 retention 실험 결과

거대한 신경망을 학습시킬 때나 장시간 추론을 하는 경우 시냅스 소자의 retention이 중요하다. Retention 실험은 시냅스 가중치를 특정 수준으로 update 한 뒤, 일정 간격으로 가중치 값을 읽어 ADC 값의 변화를 분석하였다. 가중치 감소는 간단한 RC 회로에서의 방전 모델인 exponential decay 모델을 사용할 수도 있지만, 실험 결과는 수식 (22)이 더 잘 표현하였기 때문에 수식 (22)를 분석에 이용하였다[46]. β는 1 이하의 상수로 exponential을 extended exponential로 만드는 역할을 하며[34] retention 성능은 시간 상수 τ의 크기로 평가할 수 있다.

$$\boldsymbol{ADC(t) = A \times exp\left(-\left(\frac{t}{\tau}\right)^{\beta}\right)} \quad \textbf{(22)}$$

Retention 실험의 결과는 Figure 4.3.1.1과 같다. 수식 (22) 모델이 누설 전류에 의한 가중치 변화를 잘 모델링 하는 것을 확인할 수 있었다. β 값 또한 [46]의 연구와 비슷한 값을 가졌다. 누설 전류 수준이 작은 a-IGZO TFT를 사용하였기 때문에 retention 시간 상수가 약 10,220분 수준으로 Si-CMOS 기반 3T1C에 비해 뛰어났다. ADC 출력값을 커패시터 전압으로 환산한 결과, 3T1C에서의 총 누설 전류 수준은 $<10^{-15}\mathrm{A}/\mu m$ 정도이다. Si MOSFET에 비해서는 훌륭한 수치이지만, 이전 보고된 값들에 비해서는 떨어지는데, 이는 3T1C에 사용되는 트랜지스터가 여러 개이며, 커패시터 등 추가로 누설 전류가 발생할 수 있는 통로가 있기 때문으로 추측하였다.

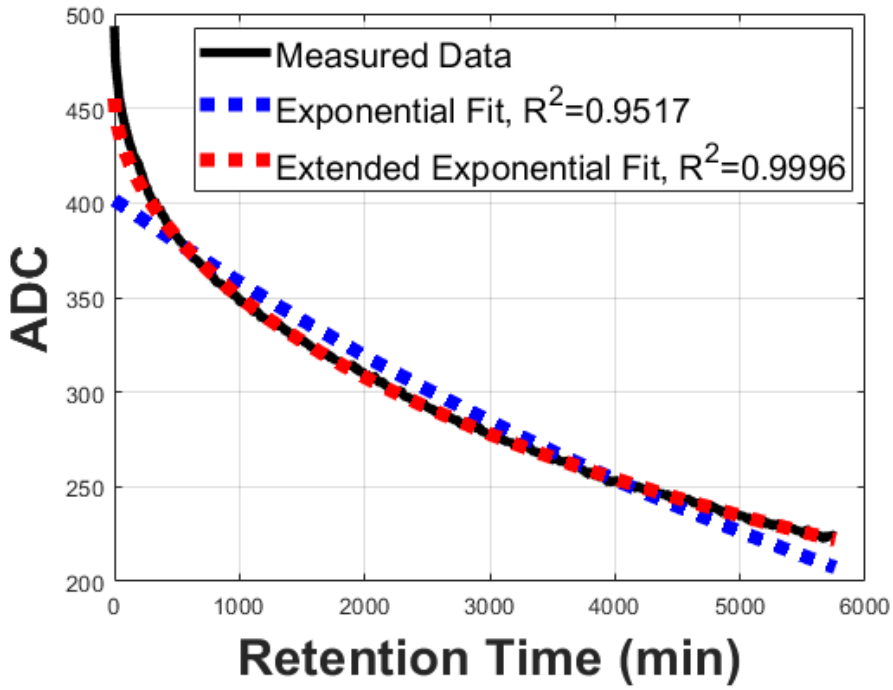

Figure 4.3.1.1 가중치 retention 실험 결과. 파란 선은 일반 exponential decay 모델, 빨간 선은 extended exponential decay 모델을 사용하여 평가한 결과이다.

4.3.2 5T1C와 3T1C의 retention 차이

5T1C(6T1C) 구조는 a-IGZO TFT는 NMOS만 존재한다는 단점을 보완하기 위해 제시된 시냅스 회로이다[51]. NMOS는 전류 주입에 부적합해 potentiation이 비선형적이게 되고, 가중치 갱신이 비대칭적으로 일어나기 때문에 potentiation 과정을 전류를 방전하는 구조로 변경하였다. Figure 4.3.2.1에서와 같이 potentiation에서는 N1, N2 트랜지스터를, depression에서는 N3, N4 트랜지스터를 사용하며 실제 전류가 흐르는 N2, N4 트랜지스터에는 V_{GS}가 일정하게 유지되도록 하였다. 이런 특징에 의해 3T1C에서의 depression처럼 potentiation도 선형적일 수 있으며, 대칭적인 가중치 갱신이 가능하기 때문에 높은 신경망 학습 정확도를 기대할 수 있다. 또한, N1과 N2 (N3와 N4)가 동시에 켜져야 갱신이 일어나기 때문에 별도의 AND gate 회로가 없어도 array에서 사용할 수 있다는 장점이 있다. 다만 커패시터가 floating 상태이기 때문에 커패시터 전압을 읽기 위해서는 N3 트랜지스터를 켜서 커패시터의 하단 전극을 $V_{DD}/2$로 정의하는 과정이 필요하다.

많은 장점을 가지는 5T1C 시냅스 소자이지만, 가중치 읽기 과정의 N3를 키는 과정에서 기생 커패시턴스에 의해 가중치가 저장된 커패시터의 전압이 변하는 문제가 있다. Figure 4.3.2.2에서 retention 실험 중 읽기 빈도에 따라 달라지는 retention 시간 상수를 확인할 수 있었다. a-IGZO TFT로 누설 전류가 흐르는 경우는 가중치 감소가 수식 (22)을 따라야 하지만 기생 커패시턴스에 의해 가중치가 손상되기 때문에 exponential decay 식을 따랐다. 신경망의 학습, 추론 과정에서 읽기 과정은 빈번하게 일어나는데, 과정마다 가중치가 변화하면 정확도가 떨어질 수밖에 없기 때문에 치명적인 문제이다. 또한, 읽기 과정뿐만 아니라 array 내에서 가중치 갱신을 하는 경우 선택되지 않은 소자들은 update

트랜지스터 중 하나만 켜지게 되는 half-select 상태에 있게 되는데, 이 과정 또한 동일하게 커패시터 전압에 간섭을 일으켜 문제가 될 수 있다. 반면 3T1C는 가중치가 저장되는 커패시터의 하단 전극이 항상 GND로 정의되기 때문에 읽기 과정에서 update 트랜지스터가 켜질 이유가 없으며, array 구동 또한 N1, N2 트랜지스터에 연결된 AND gate 바탕으로 이루어지기 때문에 5T1C에서와 같은 문제가 발생할 수 없다. 5T1C에서는 읽기 과정의 빈도와 retention 시간 상수 사이의 관계가 읽기 과정에 의해 병렬적인 누설전류 path가 있는 것처럼 모델링이 되지만, 읽기 빈도를 바꾸어 가며 retention 실험을 한 결과 3T1C에서는 시간 상수와의 경향성을 찾을 수 없었다.

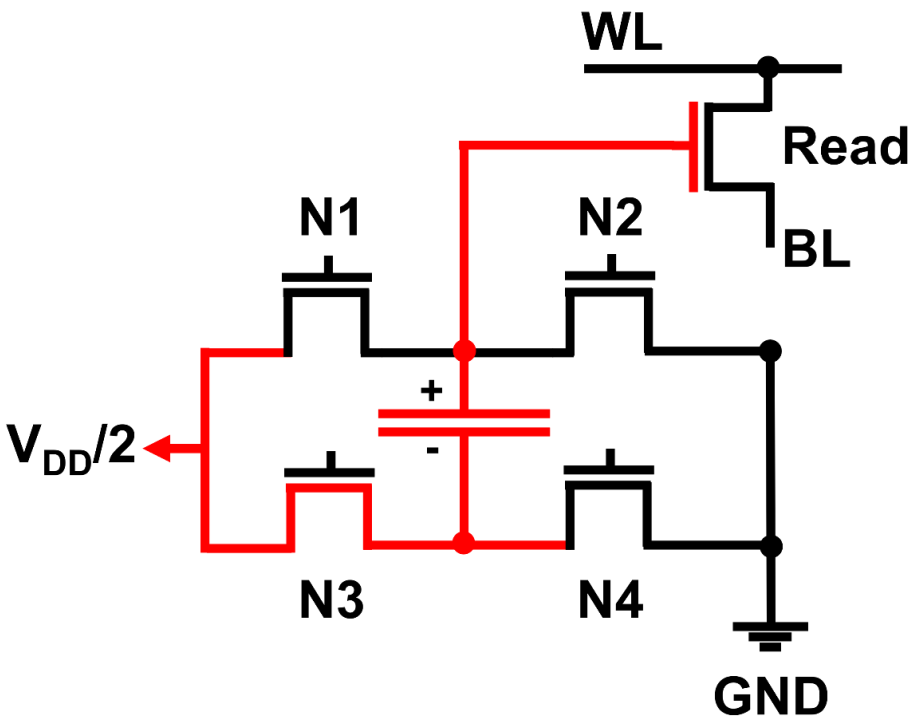

Figure 4.3.1.2 5T1C 회로도와 읽기 방식. Read 과정에는 빨간 선을 따라 read 트랜지스터의 gate에 $V_{DD}/2+V_{cap}$의 전압이 가해진다.

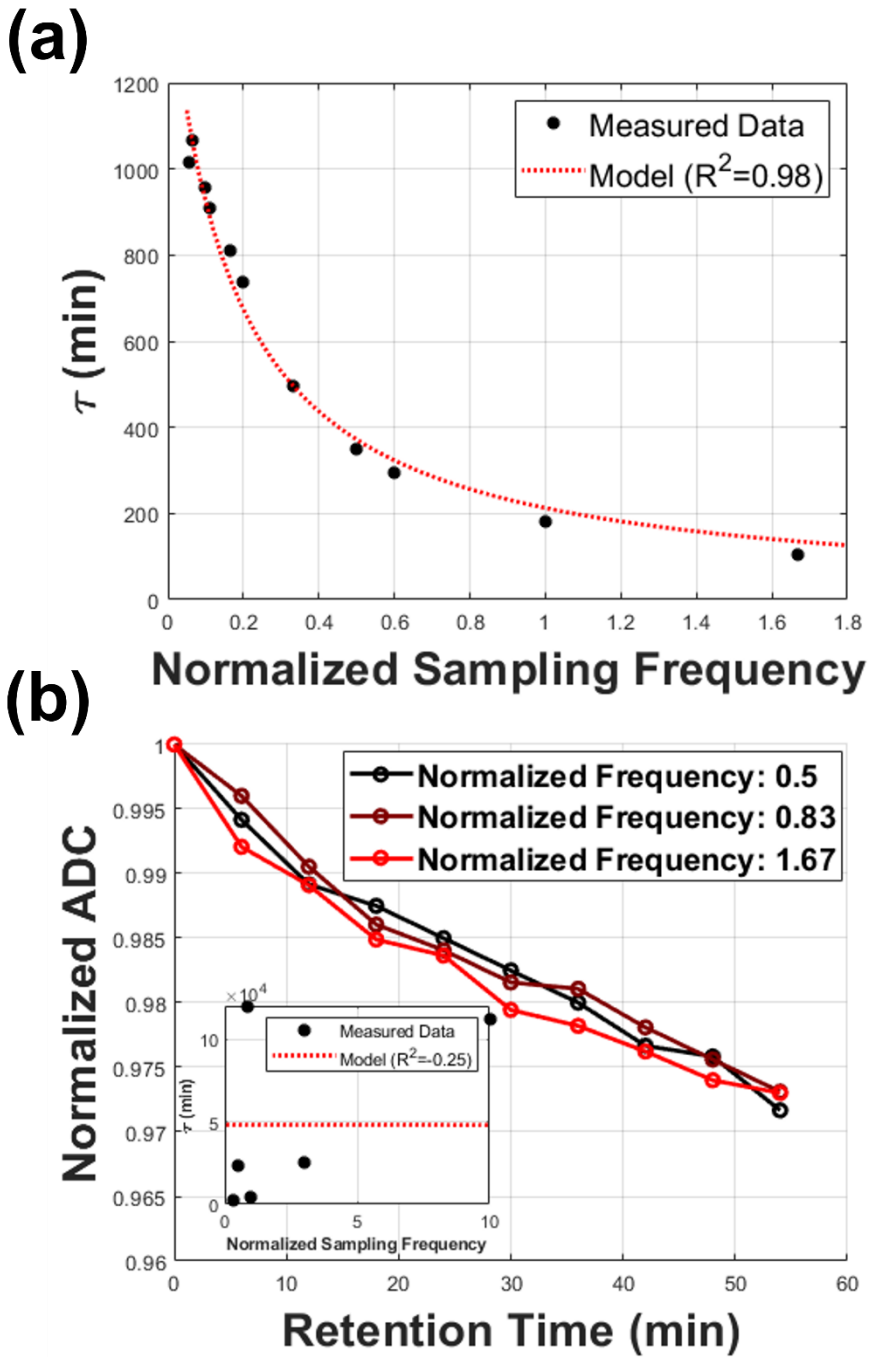

Figure 4.3.2.2 5T1C와 3T1C의 read 빈도에 따른 retention 성능 차이. (a)는 5T1C의 결과, (b)는 3T1C의 결과이다. (a)는 read 빈도에 일정한 반비례 관계를 보였다.

4.3.3 Retention 실험 전후 시냅스 성능 평가

비정질 산화물 기반 TFT는 bias stress에 큰 영향을 받는 것으로 알려져 있다. 3T1C 내 개별 트랜지스터들은 retention 실험 동안 bias stress를 받게 된다. N1과 N2는 항상 음의 전압으로 꺼진 상태이기 때문에 NBS가 가해지며, 커패시터에는 양의 전압이 저장되어 있기 때문에 read 트랜지스터는 PBS가 가해진다. Bias stress에 의해서 트랜지스터의 전기적 성질이 변화하고, 3T1C의 가중치 갱신에 변화가 생길 수 있기 때문에 retention 실험 전후로 각 트랜지스터의 transfer curve를 비교하였다.

Read 트랜지스터는 Figure 4.2.1.1와 동일한 방식으로 측정하였으며, N1과 N2 트랜지스터는 Figure 4.3.3.1과 같이 측정하였다. 각 트랜지스터에 직접적으로 전압을 가할 수 있는 방식이 없기 때문에 이처럼 간접적으로 실험하였다. 측정하지 않는 트랜지스터는 양의 gate 전압을 크게 걸어주고, 측정하고자 하는 트랜지스터의 gate 전압을 sweep해서 transfer curve를 측정하였다. Sweep하는 전압이 V_{th}와 비슷한 정도일 때는 직렬로 연결된 트랜지스터들에 흐르는 전류가 작기 때문에 V_{DD} node에 가하는 전압이 모두 측정하고자 하는 트랜지스터에 전달되지만, 더 큰 전류에서는 V_{DD} node에 가해준 전압이 두 트랜지스터에 나뉘어 가해지기 때문에 V_{th}는 측정이 가능하지만 on current는 정확한 분석이 어렵다.

실험 결과는 Figure 4.3.3.2와 같다. 총 세 가지 경우: 1) retention 실험 이전, 2) 빛을 쪼이며 retention 실험한 이후, 3) 빛을 쪼이지 않으며 retention 실험한 이후에 대해서 측정한 결과이다. NBS가 가해지는 N1, N2 트랜지스터의 경우 빛이 쪼여지는 경우는 음의 방향으로 V_{th}가 이동하는 반면 빛이 쪼이지 않는 경우 retention 실험 전후 차이가 미미하였

다. 이는 a-IGZO 내 hole carrier 농도가 작아 hole trapping이 일어나기 어렵지만, 빛이 쪼여지는 경우 electron-hole pair가 생성되어 V_{th}가 변한 것으로 설명하였다[29]. Update 트랜지스터의 V_{th}가 작아지면 동일 동작 전압에서 더 많은 전류가 흐를 수 있고, 세밀한 가중치 갱신이 불가할 수 있지만, 빛만 막는다면 문제를 방지할 수 있다. Read 트랜지스터의 PBS에 의한 전기적 성질 변화는 electron trapping에 일어나는 현상이기 때문에 빛의 유무와 무관하게 일어났다. 이전 연구에서 보고된 바와 동일하게[10] V_{th}가 증가하였으며 on current가 감소하였다. Read 트랜지스터의 열화는 동일한 가중치가 다르게 읽히게 하므로 추가 연구가 필요할 수 있다.

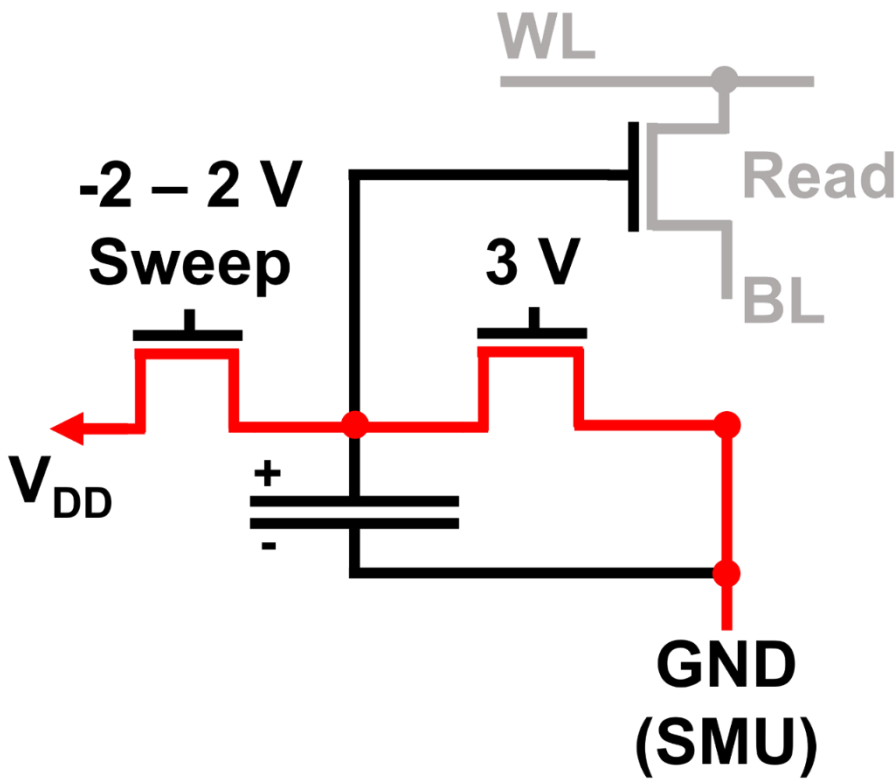

Figure 4.3.3.1 3T1C 소자 내 N1, N2 트랜지스터 transfer curve 측정 방법. N2를 측정할 때는 그림에서 N1과 N2에 가하는 전압을 바꾸어 가하였다.

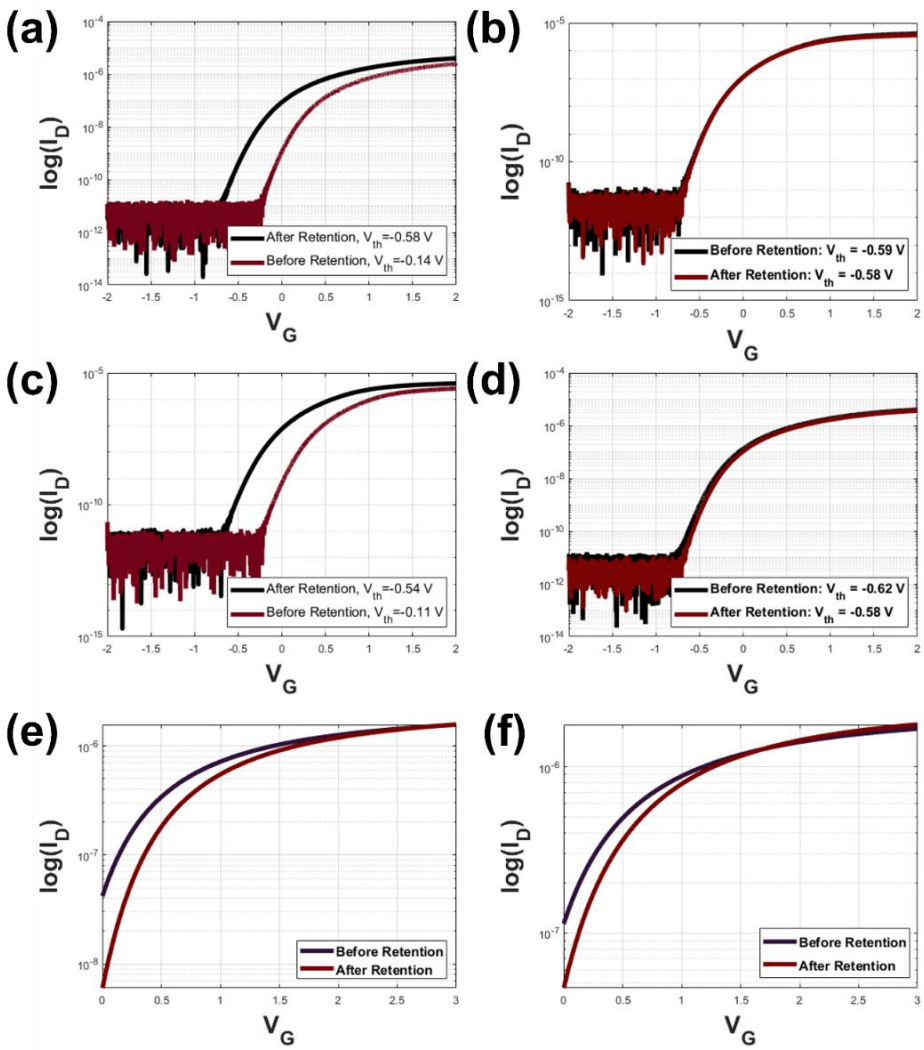

Figure 4.3.3.2 N1, N2 트랜지스터의 bias stress에 대한 안정성. (a)와 (b)는 N1, (c)와 (d)는 N2, (e)와 (f)는 read 트랜지스터의 결과이다. (a), (c), (e)는 빛이 있는 상태에서의 bias stress에 대한 변화, (b), (d), (f)는 어두운 상태에서의 bias stress에 대한 변화이다.

4.4 Cycling endurance

4.4.1 Endurance 실험 방법 및 결과

추론 array와 달리 on chip learning을 위해서는 많은 epoch 동안 시냅스 소자 갱신이 가능해야 하므로 좋은 내구성이 필요하다. Cycling endurance 실험은 4.2.2에서의 potentiation-depression 과정을 반복하였다. 실험에 사용하는 PCB 주변 회로의 한계상 read 과정이 ms 수준으로 오래 걸리기 때문에 endurance 실험에서는 매 potentiation-depression cycle마다 읽는 것이 아닌, 10의 거듭제곱과 같은 특정 시점에서만 읽고, 이외에는 update만 하였다. 이미 4.3.3에서 빛을 쪼이지 않았을 때 N1과 N2의 negative bias stress에 대한 안정성이 뛰어나다는 사실을 확인하였기 때문에 endurance 실험에서는 빛을 차단하였다.

측정 결과는 Figure 4.4.1.1과 같다. 5×10^7 cycle (cycle 당 potentiation 2,000번, depression 1,500번) 동안 endurance 실험을 하였으며, 시냅스 가중치 갱신 경향에 큰 변화 없이 잘 동작하는 것을 확인하였다. PRAM과 같이 재료를 물리적으로 변경하는 것이 아닌[56], 트랜지스터를 통해 전류를 흘리는 방식이기 때문에 DRAM과 같이 훌륭한 endurance 성능을 가지는 것을 확인할 수 있었다. Figure 4.4.1.2에서는 가중치 갱신의 선형성, 갱신량 등에 크지는 않지만 potentiation-depression cycle을 반복할수록 선형성, 대칭성이 개선되는 변화를 보였다. Cycling 상황에서 가장 취약한 트랜지스터 파악과 선형성, 대칭성 개선의 원인 분석을 위해 4.3.2와 동일한 방식으로 개별 트랜지스터들의 transfer curve를 측정하였다.

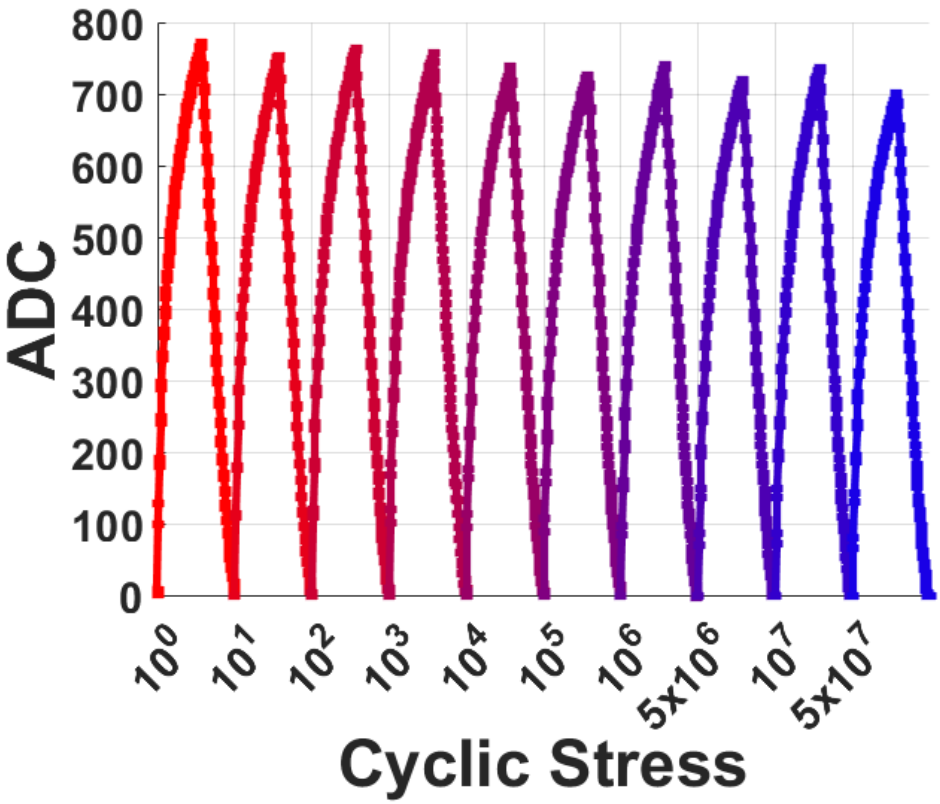

Figure 4.4.1.1 Cycling endurance 실험 결과. $5x10^7$ cycle의 potentiation-depression 동안 큰 변화 없이 동작하였다.

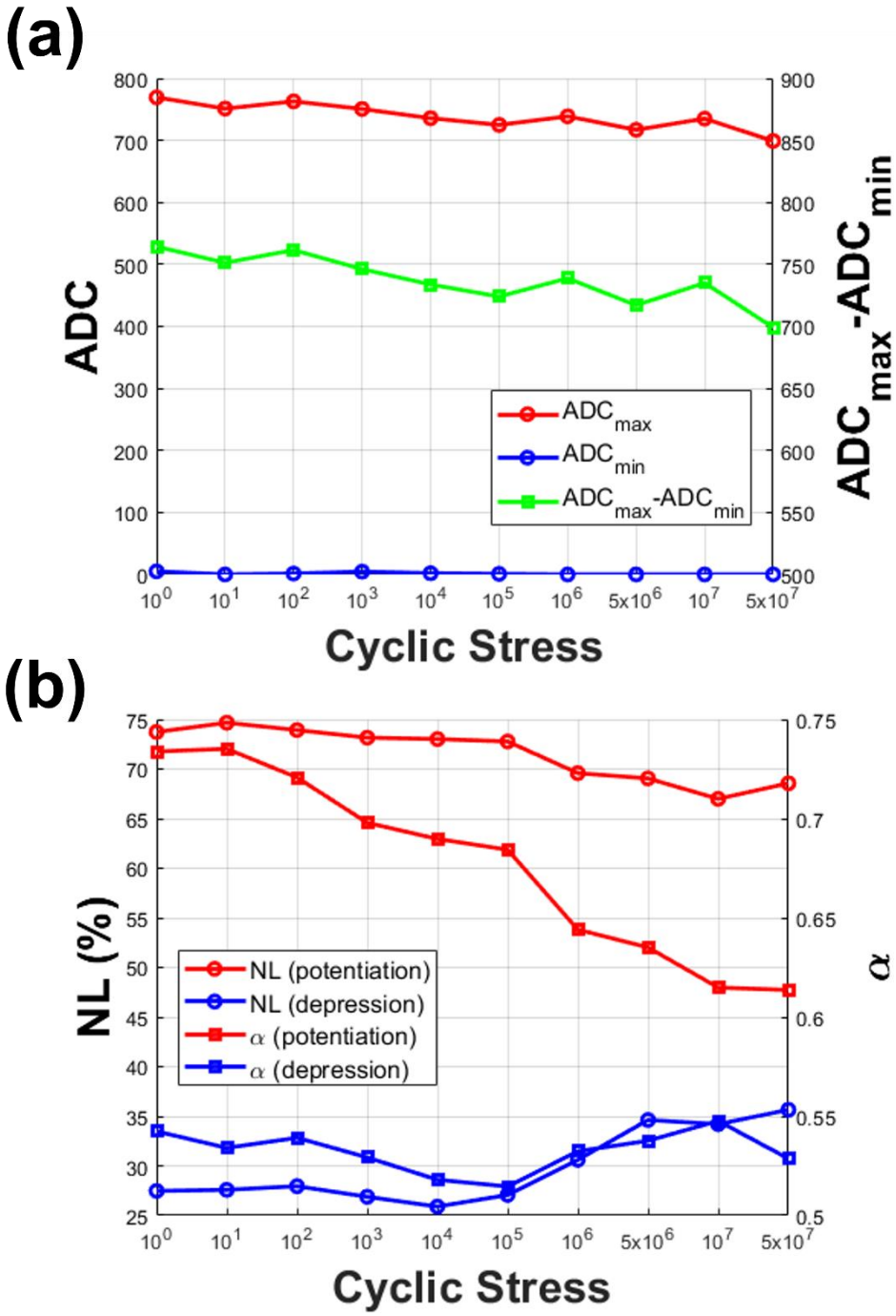

Figure 4.4.1.2 Cyclic stress가 시냅스에 미치는 영향. (a)는 ADC range, (b)는 가중치 갱신의 양과 선형성을 정리한 결과이다.

4.4.2 Cycling 부하가 시냅스 성능에 미치는 영향

일정한 bias stress가 가해지는 retention 실험과는 달리 cycling endurance 실험에서는 AC bias stress가 가해진다. N1, N2 트랜지스터는 on/off가 반복되며 read 트랜지스터에는 커패시터 충·방전에 의해 양의 AC bias stress가 가해진다. Figure 4.4.2.1은 Figure 4.4.1.2에서의 소자의 cycling endurance 실험 전후 개별 트랜지스터 측정 결과이다. AC stress가 가해졌지만, retention 실험 때와 동일하게 N1과 N2는 빛이 쪼여지지 않았기 때문에 V_{th} 변화가 미미하였다. 그러나 read 트랜지스터는 AC PBS의 영향에 의해 retention과 동일한 경향으로 V_{th}가 변화하였고, on-current의 열화 또한 확인하였다.

개별 트랜지스터 성능 측정을 통해 cyclic stress에 따른 가중치 갱신의 선형성, 대칭성 개선은 read 트랜지스터에 의한 효과라고 결론지을 수 있었다. Read 트랜지스터가 PBS에 노출되었을 때 전류 수준이 낮아지는 현상 이외에도 V_{th} 변화에 의해 낮은 커패시터 전압 영역이 subthreshold region에 가까워지며 transfer curve가 아래로 볼록해지는 변화가 일어나는데, 이것이 potentiation의 비선형성을 상쇄시킨 것으로 추측하였다. 낮은 커패시터 전압에서 N1 트랜지스터에 많은 전류가 흐르는 것이 potentiation의 비선형성에 큰 영향을 준다. Read 트랜지스터가 변하면 실제로 커패시터의 전압은 급격하게 갱신이 일어나지만 읽기 과정에서 상쇄되어 BL에 흐르는 전류는 기존보다 선형적으로 증가하는 것으로 결론지었다. 실제로 potentiation을 구간별로 나누어 비교하였을 때, cyclic stress에 따라 potentiation 초기에 일어나는 갱신량에 급격한 감소가 있음을 확인하였다. 따라서 적절한 read 트랜지스터 설계를 통해서도 더 선형적이고 대칭적인 가중치 갱신이 가능할 것이다.

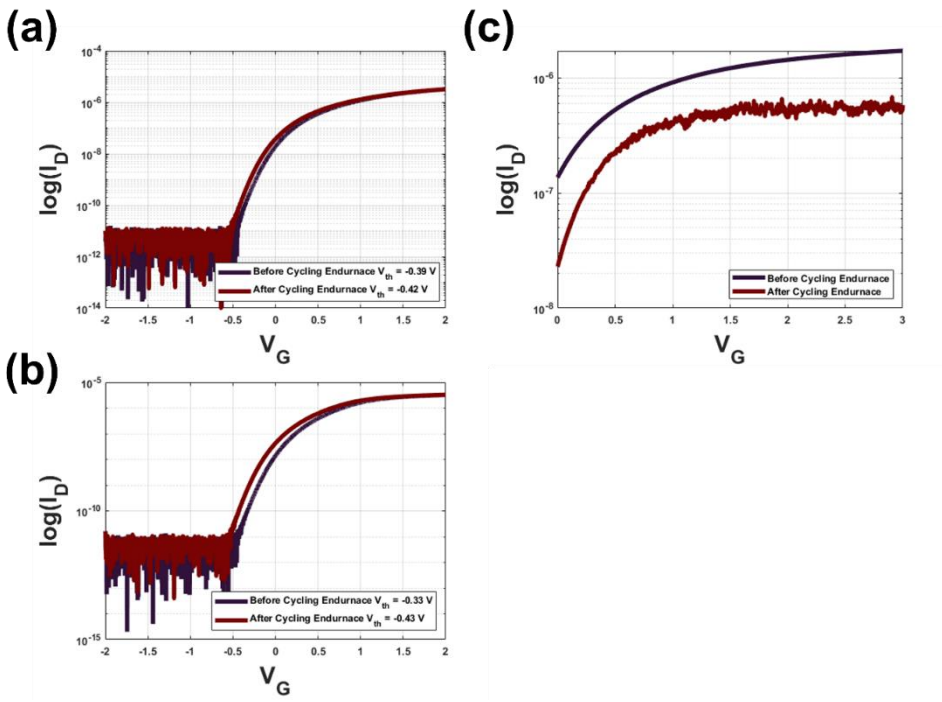

Figure 4.4.2.1 Cyclic stress 전후의 트랜지스터 성능 변화. (a)는 N1, (b)는 N2, (c)는 read 트랜지스터의 성능 변화를 나타낸 결과이다.

5. 결 론

본 연구에서는 현재 컴퓨팅이 심층 신경망 학습 시에 가지는 전력 소모, 연산 속도 측면에서의 한계를 극복할 수 있는 crossbar 형 심층 신경망 가속기에 적합한 가중치 소자로 a-IGZO TFT와 커패시터를 이용한 3T1C 시냅스를 제시하였다. PRAM, RRAM 등의 차세대 비휘발성 메모리에 비해 선형적이고 대칭적인 가중치 갱신을 제공하는 동시에, Si CMOS 기반의 시냅스에 비해 상대적으로 휘발성이 적은 장점을 가지기 때문에 대규모 신경망의 on-chip learning에 적합하다.

실험을 통해 고속의 가중치 갱신이 가능함을 확인하였으며, 가중치 retention과 endurance 성능 또한 뛰어남을 확인하였다. N-type만 존재하는 a-IGZO TFT의 특성상 potentiation 과정이 비선형적이었지만 동작 전압 조건의 최적화를 통해 이를 최적화할 수 있었으며, read 트랜지스터의 설계로 비선형성과 비대칭성을 상쇄할 수 있는 가능성을 보였다. Retention 또한 시간 상수가 10,000분 이상으로 뛰어났으며, 추후 공정 최적화로 성능 향상을 기대할 수 있었다. Cycling endurance도 트랜지스터 기반 시냅스 소자이기 때문에 훌륭하였다.

CMOS BEOL 공정에 호환되는 공정이기 때문에 Si-CMOS 공정과 수직 적층이 가능하다는 장점이 있지만, 3T1C는 선택 소자로 면적이 큰 AND gate를 사용해야 한다는 단점이 있다. 면적 감소를 위해 빠른 동작 속도를 포기하고 potentiation과 depression에 동일한 update 트랜지스터를 사용하는 2T1C 구조가 도움이 될 수 있다. 또는 AND gate 대신 dual gate update 트랜지스터를 통해 update 트랜지스터에서 선택 과정까지 일어날 수 있게 제작할 수 있다. 시냅스 동작에 따라 read 트랜지스터가 받는 bias stress에 대한 안정성도 반드시 필요하다. 이는 공정 최적화를 통한 막질 개선 등을 통해 이룰 수 있을 것으로 기대된다. 마

지막으로 potentiation과 depression의 비선형성, 비대칭성의 경우 zero-shifting[28]과 같은 3T1C에 최적화된 학습 알고리즘을 이용하면 학습 정화도 열화를 막을 수 있을 것으로 예상된다.

요약하자면, 본 연구에서는 트랜지스터를 사용하지만 속도, 전력 소모, 데이터 유지 능력 측면에서 합리적인 성능을 가지는 시냅스 소자를 제안하였다. 공정 최적화와 적절한 알고리즘 개발을 통해 3T1C의 on-chip learning array로의 성능을 향상시킨다면 현재 폰 노이만 기반 컴퓨팅의 한계를 극복하며, on-device AI 등 저전력, 고성능 AI 프로세서가 필요한 분야에서 사용될 수 있을 것으로 기대된다.

참고 문헌

[1] S. Ambrogio et al., Equivalent-accuracy accelerated neural-network training using analogue memory. Nature. 558, 60–67 (2018).

[2] H. Baba et al., in Technical Digest - International Electron Devices Meeting, IEDM (Institute of Electrical and Electronics Engineers Inc., 2021), vols. 2021-December, pp. 21.2.1-21.2.4.

[3] A. Belmonte et al., in Technical Digest - International Electron Devices Meeting, IEDM (Institute of Electrical and Electronics Engineers Inc., 2020), vols. 2020-December, pp. 28.2.1-28.2.4.

[4] S. Brivio, D. R. B. Ly, E. Vianello, S. Spiga, Non-linear Memristive Synaptic Dynamics for Efficient Unsupervised Learning in Spiking Neural Networks. Frontiers in Neuroscience. 15 (2021), doi:10.3389/fnins.2021.580909.

[5] G. W. Burr et al., Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element. IEEE Transactions on Electron Devices. 62, 3498–3507 (2015).

[6] C. C. Chang et al., Mitigating Asymmetric Nonlinear Weight Update Effects in Hardware Neural Network Based on Analog Resistive Synapse. IEEE Journal on Emerging and Selected Topics in Circuits and Systems. 8, 116–124 (2018).

[7] P. Y. Chen, X. Peng, S. Yu, in Technical Digest - International

Electron Devices Meeting, IEDM (Institute of Electrical and Electronics Engineers Inc., 2018), pp. 6.1.1–6.1.4.

[8] Y. Chen et al., An 18.6–μm–Pitch Gate Driver Using a–IGZO TFTs for Ultrahigh–Definition AR/VR Displays. IEEE Transactions on Electron Devices. 67, 4929–4933 (2020).

[9] Y. T. Chien et al., Performance Enhancement of InGaZnO Top–Gate Thin Film Transistor with Low–Temperature High–Pressure Fluorine Treatment. IEEE Electron Device Letters. 42, 1611–1614 (2021).

[10] M. H. Cho et al., Comparative Study on Performance of IGZO Transistors With Sputtered and Atomic Layer Deposited Channel Layer. IEEE Transactions on Electron Devices. 66, 1783–1788 (2019).

[11] S. Choi et al., Positive Bias Stress Instability of InGaZnO TFTs with Self–Aligned Top–Gate Structure in the Threshold–Voltage Compensated Pixel. IEEE Electron Device Letters. 41, 50–53 (2020).

[12] X. Duan et al., Novel Vertical Channel–All–Around (CAA) In–Ga–Zn–O FET for 2T0C–DRAM With High Density Beyond 4F2by Monolithic Stacking. IEEE Transactions on Electron Devices. 69, 2196–2202 (2022).

[13] R. Garcia et al., A compact drain current model for thin–film transistor under bias stress condition. IEEE Transactions on Electron Devices. 65, 1803–1809 (2018).

[14] T. Gokmen, W. Haensch, Algorithm for Training Neural Networks on Resistive Device Arrays. Frontiers in

Neuroscience. 14 (2020), doi:10.3389/fnins.2020.00103.

[15] T. Gokmen, Y. Vlasov, Acceleration of deep neural network training with resistive cross-point devices: Design considerations. Frontiers in Neuroscience. 10 (2016), doi:10.3389/fnins.2016.00333.

[16] K. He, X. Zhang, S. Ren, J. Sun, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (IEEE Computer Society, 2016), vols. 2016-December, pp. 770–778.

[17] T. Hirofuchi, R. Takano, A prompt report on the performance of intel optane DC persistent memory module. IEICE Transactions on Information and Systems. E103D, 1168–1172 (2020).

[18] Y. Hu, Y. Wang, T. Lei, F. Wang, M. Wong, Neuromorphic Implementation of Logic Functions Based on Parallel Dual-Gate Thin-Film Transistors. IEEE Electron Device Letters. 43, 741–744 (2022).

[19] Y. Hu, T. Lei, Y. Wang, F. Wang, M. Wong, An Artificial Neural Network Implemented Using Parallel Dual-Gate Thin-Film Transistors. IEEE Transactions on Electron Devices (2022), doi:10.1109/TED.2022.3201836.

[20] S. Huang, X. Sun, X. Peng, H. Jiang, S. Yu, in Proceedings of the 2020 Design, Automation and Test in Europe Conference and Exhibition, DATE 2020 (Institute of Electrical and Electronics Engineers Inc., 2020), pp. 1025–1030.

[21] R. Islam et al., Device and materials requirements for neuromorphic computing. Journal of Physics D: Applied Physics.

52 (2019), , doi:10.1088/1361-6463/aaf784.

[22] M. Jerry et al., in Technical Digest - International Electron Devices Meeting, IEDM (Institute of Electrical and Electronics Engineers Inc., 2018), pp. 6.2.1-6.2.4.

[23] S. Jung et al., A crossbar array of magnetoresistive memory devices for in-memory computing. Nature. 601, 211–216 (2022).

[24] T. Kamiya, K. Nomura, H. Hosono, Present status of amorphous In-Ga-Zn-O thin-film transistors. Science and Technology of Advanced Materials. 11 (2010), doi:10.1088/1468-6996/11/4/044305.

[25] F. Kiani, J. Yin, Z. Wang, J. Joshua Yang, Q. Xia, A fully hardware-based memristive multilayer neural network. Science Advances. 7 (2021), doi:10.1126/sciadv.abj4801.

[26] S. Kim, T. Gokmen, H. M. Lee, W. E. Haensch, in Midwest Symposium on Circuits and Systems (Institute of Electrical and Electronics Engineers Inc., 2017), vols. 2017-August, pp. 422–425.

[27] W. Kim et al., in Digest of Technical Papers - Symposium on VLSI Technology (Institute of Electrical and Electronics Engineers Inc., 2019), vols. 2019-June, pp. T66–T67.

[28] H. Kim et al., Zero-shifting technique for deep neural network training on resistive cross-point arrays. arXiv preprint arXiv:1907.10228.

[29] C. W. Kuo et al., On the Optimization of Performance and Reliability in a-InGaZnO Thin-Film Transistors by Versatile

Light Shielding Design. IEEE Transactions on Electron Devices. 68, 1654–1658 (2021).

[30] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. Proceedings of the IEEE. 86, 2278–2323 (1998).

[31] C. Lee, K. Noh, W. Ji, T. Gokmen, S. Kim, Impact of Asymmetric Weight Update on Neural Network Training With Tiki-Taka Algorithm. Frontiers in Neuroscience. 15 (2022), doi:10.3389/fnins.2021.767953.

[32] G. H. Lee et al., Effect of weight overlap region on neuromorphic system with memristive synaptic devices. Chaos, Solitons and Fractals. 157 (2022), doi:10.1016/j.chaos.2022.111999.

[33] Y. Li et al., in Digest of Technical Papers - Symposium on VLSI Technology (Institute of Electrical and Electronics Engineers Inc., 2018), vols. 2018-June, pp. 25–26.

[34] A. Lukichev, Physical meaning of the stretched exponential Kohlrausch function. Physics Letters, Section A: General, Atomic and Solid State Physics. 383, 2983–2987 (2019).

[35] Y. Luo, S. Yu, Accelerating Deep Neural Network In-Situ Training with Non-Volatile and Volatile Memory Based Hybrid Precision Synapses. IEEE Transactions on Computers. 69, 1113–1127 (2020).

[36] N. Lv et al., Suppression of the Short-Channel Effect in Dehydrogenated Elevated-Metal Metal- Oxide (EMMO) Thin-Film Transistors. IEEE Transactions on Electron Devices. 67, 3001–3004 (2020).

[37] A. Mehonic and A. J. Kenyon., Brain-inspired computing needs a master plan. Nature. 604.7905, 255-260 (2022).

[38] M. Oota et al., in Technical Digest - International Electron Devices Meeting, IEDM (Institute of Electrical and Electronics Engineers Inc., 2019), vols. 2019-December.

[39] K. Roy, S. Mukhopadhyay, H. Mahmoodi-Meimand, Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits. Proceedings of the IEEE. 91, 305–327 (2003).

[40] D. Saito et al., IGZO-Based Compute Cell for Analog In-Memory Computing - DTCO Analysis to Enable Ultralow-Power AI at Edge. IEEE Transactions on Electron Devices. 67, 4616–4620 (2020).

[41] Y. Sekine et al., (Invited) Success in Measurement the Lowest Off-state Current of Transistor in the World. ECS Transactions. 37, 77–88 (2019).

[42] M. Shoeybi et al., Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. (2019), arXiv:1909.08053v4.

[43] X. Sun et al., PCM-Based Analog Compute-In-Memory: Impact of Device Non-Idealities on Inference Accuracy. IEEE Transactions on Electron Devices. 68, 5585–5591 (2021).

[44] M. Suri et al., in Technical Digest - International Electron Devices Meeting, IEDM (2011).

[45] M. Suri et al., Physical aspects of low power synapses based on phase change memory devices. Journal of Applied

Physics (2012), vol. 112.

[46] M. Tsubuku et al., Analysis for Extremely Low Off-State Current in CAAC-IGZO FETs. ECS Transactions. 67, 17–22 (2015).

[47] H. Wang, M. Wang, D. Zhang, Q. Shan, Degradation of a-InGaZnO TFTs under Synchronized Gate and Drain Voltage Pulses. IEEE Transactions on Electron Devices. 65, 995–1001 (2018).

[48] P. Wang, S. Yu, Ferroelectric devices and circuits for neuro-inspired computing. MRS Communications. 10, 538–548 (2020).

[49] S. Wang et al., Resilience of Fluorinated Indium-Gallium-Zinc Oxide Thin-Film Transistor against Hydrogen-Induced Degradation. IEEE Electron Device Letters. 41, 729–732 (2020).

[50] Y. Wang et al., Amorphous-InGaZnO Thin-Film Transistors Operating beyond 1 GHz Achieved by Optimizing the Channel and Gate Dimensions. IEEE Transactions on Electron Devices. 65, 1377–1382 (2018).

[51] J. Won et al., Device-algorithm co-optimization for an on-chip trainable capacitor-based synaptic device with IGZO access transistor and retention-centric Tiki-Taka algorithm [Unpublished manuscript]. Department of Materials Science and Engineering, Seoul National University (2023)

[52] H. S. P. Wong et al., in Proceedings of the IEEE (Institute of Electrical and Electronics Engineers Inc., 2012), vol. 100, pp. 1951–1970.

[53] W. Wu et al., Improving Analog Switching in HfOx-Based Resistive Memory with a Thermal Enhanced Layer. IEEE Electron Device Letters. 38, 1019–1022 (2017).

[54] Q. Xia, J. J. Yang, Memristive crossbar arrays for brain-inspired computing. Nature Materials. 18 (2019), pp. 309–323.

[55] Y. Xiang et al., Impacts of State Instability and Retention Failure of Filamentary Analog RRAM on the Performance of Deep Neural Network. IEEE Transactions on Electron Devices. 66, 4517–4522 (2019).

[56] Y. Xie et al., Self-Healing of a Confined Phase Change Memory Device with a Metallic Surfactant Layer. Advanced Materials. 30 (2018), doi:10.1002/adma.201705587.

[57] X. Xu et al., Scaling for edge inference of deep neural networks. Nature Electronics. 1, 216–222 (2018).

[58] C. X. Xue et al., in Digest of Technical Papers - IEEE International Solid-State Circuits Conference (Institute of Electrical and Electronics Engineers Inc., 2020), vols. 2020-February, pp. 244–246.

[59] S. Yu, Neuro-Inspired Computing with Emerging Nonvolatile Memorys. Proceedings of the IEEE. 106, 260–285 (2018).

Abstract

3T1C Charge-Storage Type Synapse Using InGaZnO Thin-Film-Transistors for Deep Neural Network Acceleration

Minseung Kang
Materials Science and Engineering
The Graduate School
Seoul National University

Artificial intelligence (AI) has achieved remarkable progress in various fields such as image recognition and natural language processing. However, the complexity of emerging AI algorithms results in high power consumption and long training periods in conventional von-Neumann computing. While accelerating matrix-vector multiplication in crossbar arrays of nonvolatile memories has been suggested as a remedy for von-Neumann bottleneck issue, inherent nonidealities of nonvolatile memories, especially nonlinear and asymmetric weight updates, prevent its application. Si-CMOS and capacitor-based synapse may have linear, symmetric weight updates, but is volatile in nature. Amorphous InGaZnO thin film transistor 3T1C synapse circuit as a training accelerator is suggested in this work. Nonlinearity and asymmetry can be expected due to only n-type transistors existing for a-IGZO TFTs, but no issues were

found in weight updating in simulated learning schemes. Mitigation methods have also been suggested founded on weight update models and experiments. Outstanding retention performance of more than 10,000 min was measured as expected of low off current of a-IGZO TFTs. Synaptic operation did not experience significant changes after 5×10^7 weight update cycles. Combined with optimized learning algorithms, a-IGZO 3T1C synapse can be a candidate for low power, high-speed AI accelerator.

Keywords : InGaZnO TFTs, DNN accelerator, charge-storage type synapse, low off-current, linear and symmetric weight update
Student Number : 2021-28307

공학석사 학위논문

InGaZnO 박막 트랜지스터를 이용한 3T1C 전하 저장형 심층 인공 신경망 가속기

2023년 02월

서울대학교 대학원

재료공학부

강 민 승

InGaZnO 박막 트랜지스터를 이용한 3T1C 전하 저장형 심층 인공 신경망 가속기

지도 교수 김 상 범

이 논문을 공학석사 학위논문으로 제출함
2023년 01월

서울대학교 대학원
재료공학부
강 민 승

강민승의 공학석사 학위논문을 인준함
2023년 01월

위 원 장 박 민 혁 (인)

부위원장 김 상 범 (인)

위 원 강 기 훈 (인)

초 록

인공지능 알고리즘은 이미지 인식, 자연어 처리 등의 분야에서 괄목할 만한 발전을 이루고 있지만, 인공지능 구조의 복잡성, 전력 소모, 학습 기간의 증가를 전통적인 폰 노이만 컴퓨팅 구조가 따라가지 못하는 상황이다. 폰 노이만 병목 현상을 극복하기 위해 비휘발성 메모리를 crossbar 형태로 제작하여 학습 과정에서의 행렬-벡터 곱연산을 가속하려는 시도가 있었지만, 가중치 갱신의 비선형성, 비대칭성에 의해 현재로는 추가 연구가 필요하다. 반면 Si CMOS와 커패시터를 이용하면 이상적인 가중치 갱신이 가능하지만, 휘발성이라는 단점이 있다. 본 연구에서는 낮은 누설전류 수준을 가지는 amorphous InGaZnO 박막 트랜지스터를 활용하여 3T1C 구조의 시냅스 소자를 제시하였다. a-IGZO 트랜지스터는 n-type만 존재하기 때문에 가중치 갱신 과정이 비선형적일 수 있지만 소자가 인공지능 학습 알고리즘이 의도하는 가중치로 수렴할 수 있었고, 제작한 가중치 갱신 모델과 실험을 통해 소자의 비이상적인 특성들을 개선할 수 있는 방법을 제시하였다. 또한, 낮은 누설전류에 의한 10,000분 이상의 가중치 보존 시간 상수를 확인하였으며, 5×10^7의 가중치 갱신 사이클 동안 시냅스 소자가 변화 없이 동작하는 것 또한 확인하였다. 본 연구에서 제시된 3T1C 소자와 적합한 알고리즘이 결합한다면 인공지능을 저전력, 고속으로 학습할 수 있을 것으로 기대한다.

주요어 : InGaZnO 박막 트랜지스터, 심층신경망 연산 가속기, 전하저장형 시냅스, 낮은 누설전류, 가중치 갱신의 선형성과 대칭성
학 번 : 2021-28307

목 차

List of Tables

List of Figures

1. 서 론

1.1 심층신경망 가속을 위한 시냅스 소자의 필요성

인공지능 알고리즘은 이미지 인식, 자연어 처리 등의 복잡한 작업을 특화할 수 있는 장점이 있고, [16], [42]와 같은 고도화된 알고리즘의 발달로 미래 산업에서의 중요성이 증가하고 있다. 새롭게 개발되는 심층 인공 신경망들은 높은 분류 정확도를 가지지만 더 많은 가중치 저장과 연산이 요구된다. 1998년의 LeNet-5[30]은 10^6개 이하의 파라미터를 사용했지만, 최신 신경망들은 10^{14}개 이상의 파라미터를 사용하며, 이에 따라 학습 과정에서의 소비 전력과 비용이 기하급수적으로 증가하고 있다[37, 57]. 심층 인공 신경망의 연산 중 행렬 벡터 곱 연산이 (Matrix-Vector Multiplication, MVM) 가장 큰 비중을 차지하는데[5], 통용되는 폰 노이만(von Neumann) 컴퓨팅 구조는 메모리와 연산 장치 사이의 병목 현상 때문에 전력 효율과 연산 속도 측면에서 소프트웨어의 발전 방향과 적합하지 않다.

Crossbar array 구조의 resistive processing unit (RPU)를 사용하는 새로운 컴퓨터 아키텍처는 인공 신경망 구동에서의 폰 노이만 컴퓨팅 구조의 한계를 극복하기 위해 제안되었다[54]. RPU 아키텍처는 뉴런 신호를 전달하는 서로 평행한 행과 열 방향의 금속 선과, 선의 교차점마다 시냅스 역할을 하는 메모리 소자로 이루어진다. 금속 선들은 뉴런의 입출력을 위해 존재하며, 행 방향의 금속 선은 입력 전압 신호가 주입되고 열 방향은 전류 신호가 출력된다. 교차점의 메모리 소자는 시냅스의 가중치를 전기전도도로 저장한다. 옴의 법칙과 키르히호프의 법칙에 의해 각 행에 전압 신호를 주입하면 입력 전압 벡터와 전기전도도 행렬의 곱 연산 결과가 각 열에 전류의 합으로 나타나고, 하나의 crossbar array가

심층 신경망의 두 층 사이에서 일어나는 연산을 한 번에 수행하게 된다. 이렇게 crossbar array 구조의 RPU는 가중치 저장과 MVM 연산이 한 공간에서 일어나며 기존 병목 현상에서 벗어나 뛰어난 전력 효율과 처리 속도를 가질 수 있다.

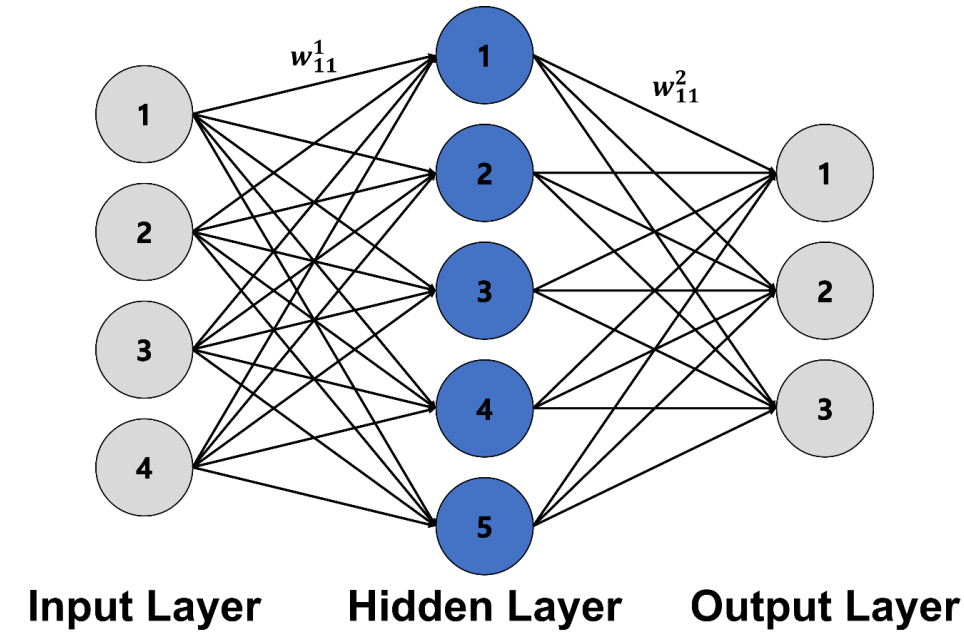

Figure 1.1 심층 신경망 구조. 원은 뉴런을, 실선은 시냅스 가중치를 의미한다.

RRAM (Resistive Random Access Memory)[58], PRAM (Phase change Random Access Memory)[44], MRAM (Magnetic Random Access Memory)[23], FeFET (Ferroelectric Field Effect Transistor)[22]과 같은 차세대 비휘발성 메모리가(Non-Volatile Memory, NVM) 시냅스 가중치 저장 장치로써 제안되었고, 외부 컴퓨터에서 학습된 네트워크를 업로드 한 추론용(inference-only) 어레이로 제작된 바가 있다[25]. 그러나 학습 과정까지 가속할 수 있는 완전한 심층 인공 신경망 가속기 제작을 위해서는 온 칩 학습(on-chip

learning)이 가능해야만 한다. On-chip learning을 위해서는 표현할 수 있는 가중치 단계가 많아야 하며 시냅스 가중치 갱신이 선형적이고 대칭적이어야 한다. 또한, 학습 과정에서 소자를 읽고 쓰는 과정이 안정적이어야 하므로 내구성과 아날로그 상태 retention이 일정 수준 이상이어야 한다[15]. 많은 NVM이 가지는 문제인 비선형적, 비대칭적 시냅스 가중치 갱신은 특히 학습된 신경망의 정확도를 크게 떨어트리는 것으로 알려져 있다[15, 21, 55, 59].

이런 차세대 비휘발성 메모리의 한계 때문에 crossbar array의 시냅스 소자를 CMOS 트랜지스터와 커패시터를 통해 표현하고자 하는 시도가 있다[26, 33, 35]. CMOS는 성숙한 기술이며, 빠른 동작 속도와 linear하고 symmetric 한 가중치 갱신이 가능하다는 점에서 이점을 가지지만, DRAM이 64 ms마다 refresh 과정을 거치는 것처럼 CMOS 트랜지스터는 off current가 크기 때문에 시간에 따라 커패시터에 저장된 정보가 누실 된다는 단점이 있다. [33]에서처럼 비교적 작은 규모의 신경망을 학습시키는 데는 빠른 주기로 커패시터가 갱신되기 때문에 retention에 의한 문제가 발생하지 않지만, 실용적인 거대한 신경망들에 적용되기는 어렵다. 커패시터의 용량을 늘려 동일 누설 전류에 대해 전압 감소를 낮출 수 있지만 가중치 갱신 에너지 소모와 소자 면적이 증가한다는 상충관계가 있다.

따라서 현재의 CMOS 기술과 미래의 이상적인 비휘발성 기술을 연결하는 중간다리 역할이 필요하며, 본 논문에서는 amorphous InGaZnO (a-IGZO) 박막 트랜지스터(Thin Film Transistor, TFT)와 커패시터를 이용한 3T1C 전하 저장형 시냅스를 이로 제안하는 바이다. a-IGZO TFT는 누설 전류가 작아 CMOS 기반 시냅스와 다르게 작은 커패시터 용량으로도 합리적인 데이터 retention time을 가질 수 있으며, 동시에 NVM 소자보다 많은 가중치 단계를 표현할 수 있고 선형적, 대칭적인

갱신이 가능하다. 3T1C a-IGZO TFT 시냅스는 on-chip learning을 위한 하드웨어 가속기뿐만 아니라 다른 비휘발성 메모리 어레이와 함께 사용하는 보조 어레이 역할[14, 31] 또한 수행할 수 있을 것으로 기대한다.

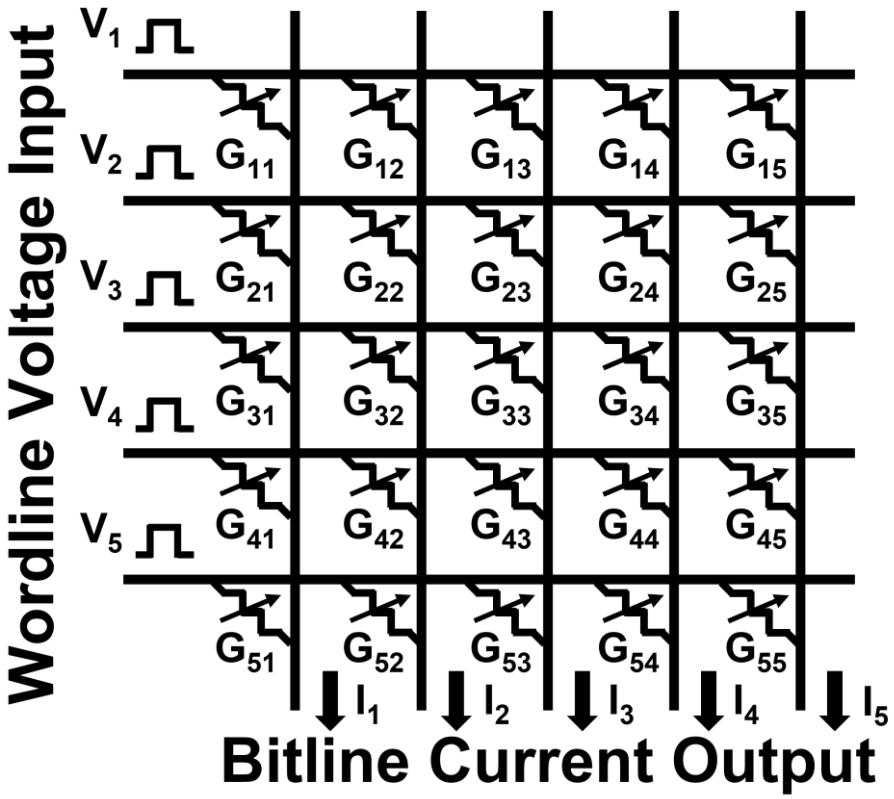

Figure 1.2 Crossbar 어레이 구조. 가로 방향 wordline으로는 전압 input이, 세로 방향 bitline으로는 전류 output이 출력된다.

2. 문헌 조사

2.1 비휘발성 메모리를 이용한 시냅스 소자

심층 신경망 연산 가속기는 추론만 가능한 array와 on-chip learning까지 가능한 array로 나뉜다. 비휘발성 메모리를 이용한 추론용 array는 메모리의 retention 성능과 일관성 있는 데이터 읽기가 가능해야 하며, on-chip learning을 위해서는 추가로 큰 on/off ratio, 선형적이고 대칭적인 가중치 갱신, 좋은 cycling endurance 특성이 요구된다[15]. 이러한 조건을 만족하지 못하면 최대 정확도 감소 등 제약이 있다. 특히 가중치 갱신의 비선형성, 비대칭성이 학습에 가장 큰 문제로 알려졌다. 심층 신경망 학습 과정상 많은 수의 potentiation과 depression 명령이 가해지는데, 가중치 갱신이 비선형적이라면 갱신에 따라 특정 가중치로 수렴해버려 의도치 않은 방향으로 학습되기 때문이다[15, 20].

PRAM은 작동 원리가 명확하며 CMOS 기술과 호환되는 장점을 가진 상용화된 제품이 있을 정도로[17] 성숙한 비휘발성 소자이다. PRAM은 좋은 retention 특성을 가져서 추론용 array로는 알맞은 성질을 가진다. 그러나 weight를 증가시키는 과정인 potentiation은 가능하지만, weight를 감소시키는 depression 과정은 아날로그 프로그래밍이 불가능한 단점이 있다. PRAM에서 저항을 높이는 방향의 쓰기는 상변화 물질을 녹인 뒤 급랭하는 방식이고, 이 과정이 점진적으로 일어날 수 없어 전기전도도 변화가 급격하다[45]. 이 문제 때문에 PRAM을 시냅스 소자로 사용할 때는 두 PRAM 소자를 하나의 시냅스 소자로 활용하지만[27] 동작 속도 감소 및 면적 증가의 문제가 발생한다. 이외에도 상을 변화시킬 때의 큰 에너지 소모, 상 간의 부피 차이로 인한 내구성 저하와 시간에 따른 resistance drift 현상 또한 극복해야 할 점이다[43].

RRAM은 metal-insulator-metal 구조로 insulator에 전기 전도성이 있는 통로가 형성된 정도에 따라 정보를 저장한다. 구조가 간단해 제작이 용이하고 면적이 $4F^2$로 작으며 선택 소자 없이도[52] array 동작이 가능하다는 장점이 있는 소자이다. 다만 대체로 depression 과정이 급격하며, 전도성이 있는 통로가 형성되는 과정이 stochastic 하다는 점에서 on-chip learning 용 소자로는 적합하지 않다. Bilayer 구조의 RRAM으로 선형적이고 연속적인 시냅스 갱신을 이룬 연구도 있지만[53], 제작이 복잡하며 여전히 표현할 수 있는 가중치 수가 작다는 한계가 있다.

MRAM은 동작 속도가 다른 비휘발성 메모리에 빠르다는 장점이 있지만 on/off ratio가 작은 문제가 있다. 따라서 표현할 수 있는 시냅스 가중치의 범위가 작고, 심층 신경망 학습에 제한이 된다. 이를 해결하고자 [23]에서는 시냅스 가중치가 0과 1로만 나뉘는 이진 신경망(Binary Neural Network, BNN)에서 MRAM을 이용하였지만, 가중치의 정보 손실에 의한 신경망의 최대 추론 정확도에 한계가 생기는 문제가 있다.

이처럼 PRAM, RRAM, MRAM을 비롯한 비휘발성 메모리 소자들은 on-chip learning을 위한 array에서의 사용하기에는 각각 한계점이 많다. 메모리의 성질에 최적화된 주변 회로를 이용하거나 복잡한 읽기, 쓰기 과정을 통해 신경망의 정확도 하락을 어느 정도 막을 수 있지만, 면적, 전력 소모 등을 희생해야 하므로 on-chip learning에 적합한 시냅스 소자가 필요하다[48]. 특히 write and verify와 같은 소자에 저장된 전도도 상태를 알아야 하는 보정법은 MVM 연산의 병렬성을 잃는 치명적인 단점이 있다.

2.2 Si-CMOS 회로 기반의 on-chip learning

NVM 소자를 사용한 심층 신경망 가속기는 on-chip learning에는 한계가 있기 때문에 CMOS 트랜지스터와 커패시터를 이용한 전하 저장형 시냅스가 제안된 바 있다[33]. 2개의 update 트랜지스터의 source나 drain, 그리고 read 트랜지스터의 gate가 커패시터의 상단 전극에 연결된 3T1C 구조이다. 기본적인 개념은 DRAM과 동일하지만, 정보 갱신과 읽는 방식에 차이가 있다. 아날로그 정보를 저장해야 하므로 2개의 update 트랜지스터가 필요하다. 하나의 NMOS update 트랜지스터로 potentiation을 하는 경우를 생각하면, update 트랜지스터의 source 전압은 커패시터에 저장된 전하에 의해 결정되고, 저장된 정보에 따라 update 트랜지스터에 흐르는 전류 크기가 변하기 때문에 NMOS는 전류 주입에 적합하지 않고, 전류 방출에만 사용할 수 있다. 따라서 potentiation 용 PMOS, depression 용 NMOS로 update 트랜지스터가 2개 필요하다. 또한, DRAM은 저장 커패시터를 방전시켜 저장되어 있던 전하량을 측정해 읽을 때마다 정보 손실이 일어나지만 3T1C 시냅스는 커패시터 전압을 read 트랜지스터의 gate에 가해 아날로그 정보를 read transistor에 흐르는 전류로 읽을 수 있다. 이때, 커패시터의 전압과 read 트랜지스터의 전류가 일대일 대응이 되기 위해서는 read 트랜지스터의 drain에 적절히 작은 전압을 가해 트랜지스터가 linear region에서 동작해야 한다.

Si 트랜지스터를 이용한 전하 저장형 시냅스는 [33]에서와 같이 update 트랜지스터가 saturation region에서 동작하는 한 선형적인 가중치 갱신이 가능하다. 다만 Si 트랜지스터를 통한 누설 전류가 크기 때문에 데이터 retention에는 문제가 있고, 정보 누실을 보정하는 과정이나 비휘발성 정보 저장 장치의 보조가 필요하다. Ambrogio *et al.*의 연구에

서는 PRAM array와 3T1C array를 함께 사용하였다[1]. CMOS 3T1C가 휘발성 메모리이기 때문에 선형적, 대칭적 가중치 갱신이 가능한 3T1C array에서 우선 학습을 한 뒤, 임계점에 도달하면 커패시터의 정보를 비휘발성 메모리인 PRAM으로 옮기는 방식이다. 비휘발성 메모리와 CMOS 트랜지스터의 장점만을 이용할 수 있는 연산 가속기이지만, 학습 과정은 모두 CMOS 3T1C에서 일어나기 때문에 학습 과정의 retention 문제에서 벗어나지 못한다. 간단한 구조의 신경망에서의 MNIST handwritten dataset 학습과 같은 부하가 적은 작업에 대해서는 학습 주기가 짧기 때문에 커패시터 전하량 갱신이 자주 일어나고 retention 문제에 큰 영향을 받지 않지만, 더 복잡한 네트워크와 학습 작업의 경우 가중치 갱신 간 시간이 길기 때문에 나쁜 retention에 의한 정확도 하락이 나타날 수 있다[33].

2.3 InGaZnO 박막 트랜지스터를 활용한 시냅스 소자

IGZO는 산화물 기반 반도체의 대표적인 물질로, 비정질 상에서도 높은 mobility를 가지는 반도체로 유지된다는 점에서 장점이 있다. 일반적으로 반도체는 결정질이 아닐 때 mobility가 크게 감소하지만, In 원자의 구형 5s 오비탈이 크고 등방이기 때문에 비정질 상이여도 오비탈의 겹침이 있어 mobility 저하가 크지 않다[24]. 비정질 상에서도 전기적으로 선호되는 특성을 가진다는 장점 때문에 저온 공정이 가능해 디스플레이의 구동 회로에서 박막 트랜지스터(Thin Film Transistor, TFT) channel 물질로 이용된다[8].

a-IGZO를 비롯해 비정질 반도체는 자연적으로 dopant state로 작용하는 산소 공핍(V_O)이 많아 별도의 doping 과정 없이도 n-type 반도체이다. 다만, Si CMOS처럼 ion implantation 등을 통해 p-type으로 사용

할 수 없어 n-type 트랜지스터만 제작 가능하다[24]. V_O 농도는 a-IGZO channel 증착 시의 산소 분압 등으로 조절이 가능하며, carrier 농도가 높을 시에는 depletion mode 트랜지스터로 동작한다.

a-IGZO TFT는 작은 누설 전류를 가진다. Si CMOS의 누설 전류는 pA/μm 수준이지만[33, 39], a-IGZO TFT는 yA/μm 수준의 누설 전류까지 보고된 바가 있다[41]. 이는 IGZO 물질이 가시광선이 흡수되지 않을 정도의 높은 bandgap energy를 가지기 때문이다. 낮은 누설 전류를 가진다는 특성 때문에 최근 IGZO TFT를 이용하여 메모리 소자를 제작하려는 연구가 많다. Sekine *et al.*의 연구에서는 DRAM처럼 커패시터에 전하를 저장하되, 읽기 과정에서 커패시터를 방전시켜 저장되어 있던 전하량을 확인하는 것이 아닌 gate가 커패시터에 연결된 읽기 트랜지스터를 이용해 non-destructive readout이 가능한 2T1C 메모리 소자를 제안하였다[38, 41]. IGZO TFT의 누설 전류가 작기 때문에 높은 온도에서도 장시간 정보 retention이 가능하였다. 이에 더해 write 트랜지스터의 source와 read 트랜지스터의 gate 사이에 존재하는 기생 커패시터를 저장 공간으로 활용하는 2T0C 메모리 소자가 제안된 바 있으며[40], 수직 적층형 a-IGZO 2T0C 소자 또한 Duan *et al.*의 연구에서 제시되었다[12]. 커패시터 용량이 작음에도 불구하고 기존 DRAM의 64 ms 주기의 refresh보다 긴 300 s의 retention time이 보고된 바 있다.

Dual gate 트랜지스터와 커패시터를 이용한 시냅스와 뉴런 소자 또한 Hu *et al.*의 연구에서 보고되었다[18, 19]. 시냅스 소자는 2T1C 구조이며, 커패시터의 전극이 read 트랜지스터의 bottom gate에 연결되고, top gate에 전압 입력 신호를 가할 수 있도록 하였다. 커패시터에 저장된 전압과 전압 입력 신호에 의해 read 트랜지스터에 흐르는 전류의 크기가 달라지고, 합산된 전류가 inverter 기반의 뉴런 소자가 전압 신호를 출

력하도록 한다. On-chip learning이 가능하였으며, $<10^{-18}$ A/μm 수준의 낮은 누설 전류 수준에 의해 커패시터 전압이 0.1 V 감소하는 데 4시간이 필요한 수준의 retention 성능을 달성하였다. 시냅스 소자 2개로 AND, OR 논리 gate 구현 및 4X6 array에서의 TETRIS 패턴 인식 작업이 가능하였지만, 추론 동작이 수백 μs 정도로 느리며, 디지털 형태의 입출력만 처리할 수 있다는 한계점이 있다.

IGZO는 NMOS만 존재한다는 한계를 극복한 6T1C 시냅스 소자 또한 제안된 바 있다[51]. NMOS는 전하를 방전시키는 데에 적합하기 때문에 potentiation과 depression을 서로 반대되는 방향으로 전하를 방전시키는 방식으로 가중치를 갱신한다. 가중치 갱신을 위한 트랜지스터 4개와 읽기를 위한 커패시터의 상, 하단 전극에 gate가 연결된 트랜지스터 2개, 정보 저장을 위한 커패시터로 이루어진 구조이다. 실제로 전류가 흐르는 두 update 트랜지스터의 source는 항상 접지된 상태이기 때문에 일정한 gate-source 전압 차가 유지되며, 트랜지스터가 일정한 전류를 흘리는 saturation mode에서 동작할 수 있도록 한다. 저장된 시냅스 가중치에 무관하게 saturation mode에서 동작하는 update 트랜지스터에 의해 선형적이고 대칭적인 갱신이 가능하다. 커패시터의 양 전극에 연결된 트랜지스터가 모두 켜져야 갱신이 일어나기 때문에 별도의 선택 소자 없이 array 구동이 가능하지만, update 트랜지스터의 기생 커패시턴스에 의해 half select 된 시냅스 소자의 커패시터에 저장된 전하에 변화가 생긴다는 한계가 있다.

다만 bias stress에 대한 stability (positive, negative bias stability, PBS, NBS)는 해결되어야 할 과제이다[47]. 기본적으로 V_O가 많고, channel과 gate insulator 사이의 defect 들에 gate bias에 의해 carrier trapping이 일어나고, Flash 메모리의 threshold voltage가 변화하는 원

리와 동일하게 channel에 가해지는 전압을 바꾸는 효과가 일어나 트랜지스터의 전기적 특성이 변한다[13]. Top gate TFT를 기준으로, positive bias가 오래 가해지면 electron trapping이 일어나 threshold voltage가 증가하며 반대로 negative bias는 threshold voltage를 낮춘다. 일반적으로 IGZO는 n-type 반도체이고, hole 농도가 작기 때문에 NBS의 영향이 적지만, TFT에 빛이 가해지면 (PBIS, NBIS) electron-hole pair가 생성되어 bias stress가 가속화되는 효과가 있다[29]. 수소 원자 또한 bias stress에 따라 TFT 성질을 변화시킨다. 공정 과정 중 channel에 포함되었던 H 원자가 강한 전압에 의해 결합을 끊고 gate insulator 쪽으로 확산하며 트랜지스터의 특성이 변화한다. Charge trapping은 가역적인 반응이지만, H 원자의 확산은 결합의 변화가 있기 때문에 비가역적인 변화를 야기한다[9]. 현재 디스플레이 구동에 사용되는 IGZO TFT는 bias stress에 의한 threshold voltage 변화를 sensing하고 보정해주는 회로가 있지만[11], 집적도 높은 메모리 소자에 IGZO TFT를 사용하기 위해서는 반드시 해결되어야 하는 문제이다.

3. 실험 및 분석 방법

3.1 단일 트랜지스터 및 커패시터 제작

본 논문에서 3T1C에 포함되는 트랜지스터는 top gate staggered 구조로 thermal SiO_2 위에 제작하였다. Source, drain 금속으로는 Tungsten을 이용하였으며, a-IGZO는 In : Ga : Zn = 1 : 1: 1 비율로 sputtering 기법을 이용해 증착하였다. Gate insulator는 HfO_2를 atomic layer deposition (ALD)를 통해 증착하였다. 실험에 사용한 트랜지스터의 channel dimension은 2 μm x 5 μm이며, 유전막 두께는 10 또는 15 nm이다. Source, drain과 gate가 수직으로 겹친 구조이고, 그 길이는 0.25 μm이다. 공정 과정은 Figure 3.1.1와 같으며, 소자 구조는 Figure 3.1.2와 같다.

제작된 트랜지스터의 transfer curve와 output curve는 각 Figure 3.1.3과 Figure 3.1.4와 같다. 본 실험에서는 모두 동일한 channel dimension을 가지는 트랜지스터만을 사용하였기 때문에 별도의 normalization 과정 없이 I_{DS}가 10^{-11} A가 되는 전압을 threshold voltage로 정의하였다. Subthreshold swing은 drain 전류가 10^{-11} A에서 10^{-10} A가 되는데 필요한 전압 차이로 계산하였다. Carrier mobility는 수식 (1)을 통해 계산하였다. Channel 물질이 polycrystalline이 아닌 amorphous 상이기 때문에 웨이퍼 위 많은 트랜지스터가 좋은 균일도를 가지는 것을 확인할 수 있었다.

$$\mu_{FE}^{Sat} = \frac{2L}{WC_{ox}}\left(\frac{d\sqrt{I_{DS}}}{dV_g}\right)^2 \quad (1)$$

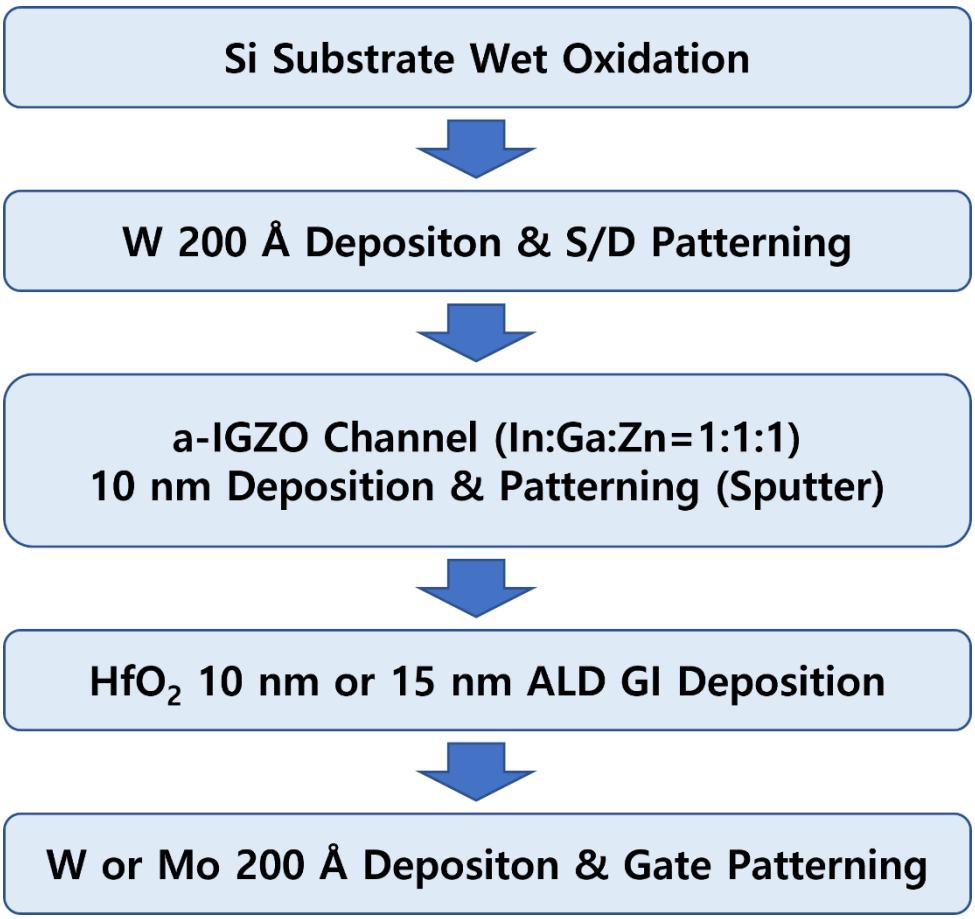

Figure 3.1.1 IGZO 트랜지스터 공정 순서

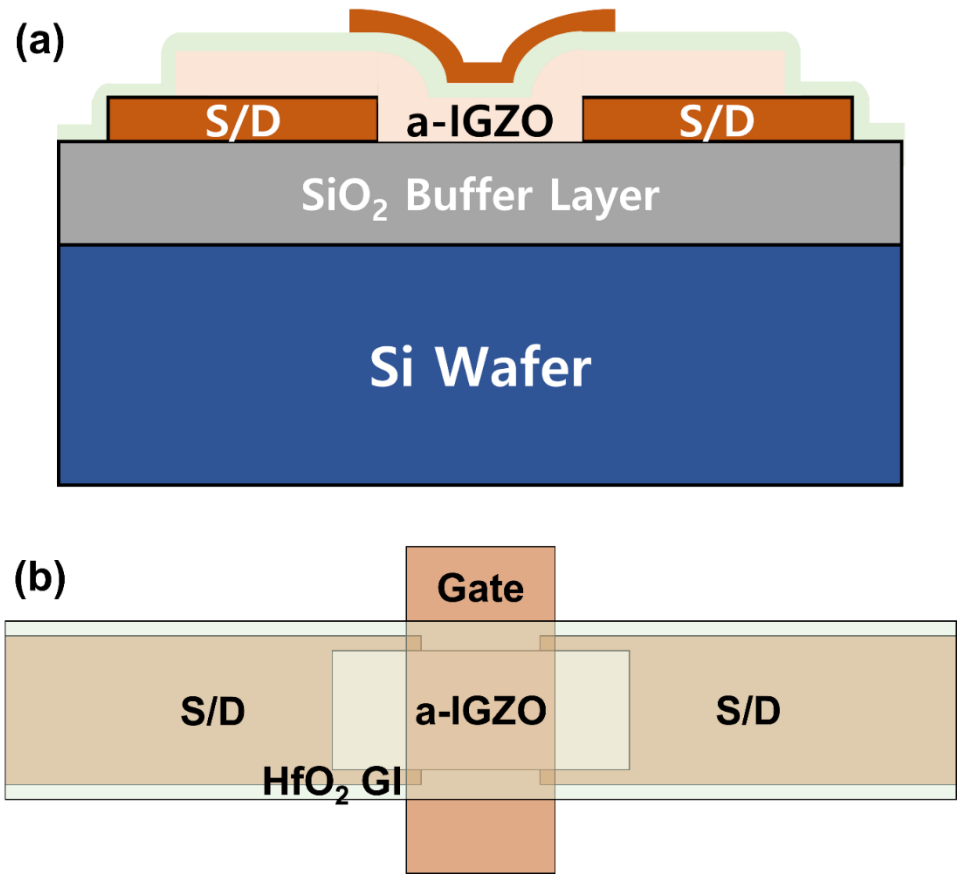

Figure 3.1.2 IGZO 트랜지스터 구조. (a)는 단면, (b)는 top view이다.

표 1 웨이퍼 별 트랜지스터 전기적 성질

	Top gate	Gate oxide	S.S. (mV/dec)	V_{th} (V)	I_{on}* (μA)	Mobility ($cm^2/V/s$)
Wafer 1	W	HfO_2 10 nm	112.9 ± 4.4	-0.58 ± 0.19	2.1 ± 0.63	6.02 ± 0.41
Wafer 2	Mo	HfO_2 15 nm	99.6 ± 1.9	0.33 ± 0.02	0.025 ± 0.0097	0.31 ± 0.17

*V_{DS} = 1.5 V, V_{GS} = 1.0 V

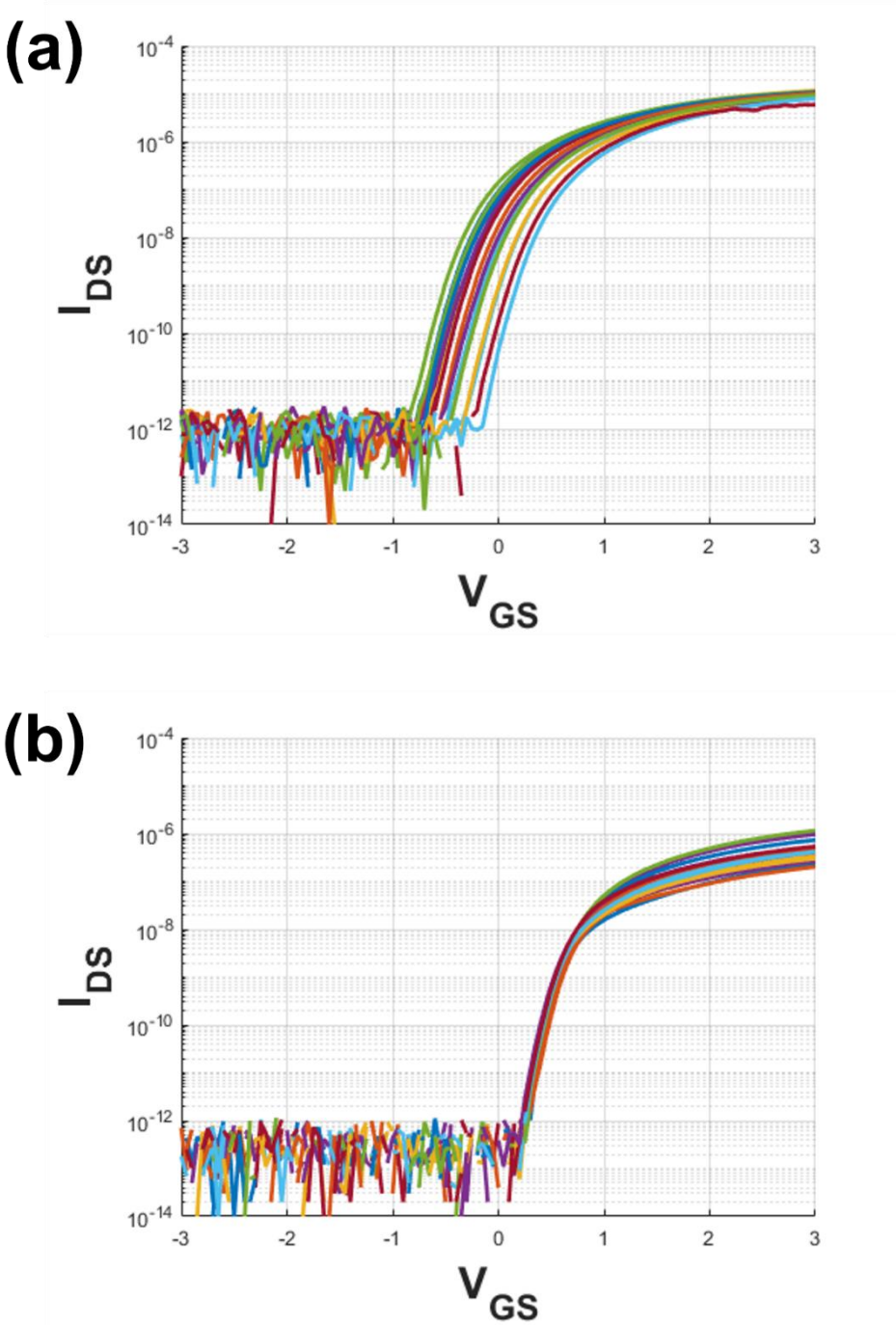

Figure 3.1.3 제작된 트랜지스터의 transfer curve. (a)는 Wafer1, (b)는 Wafer 2의 측정 결과이다.

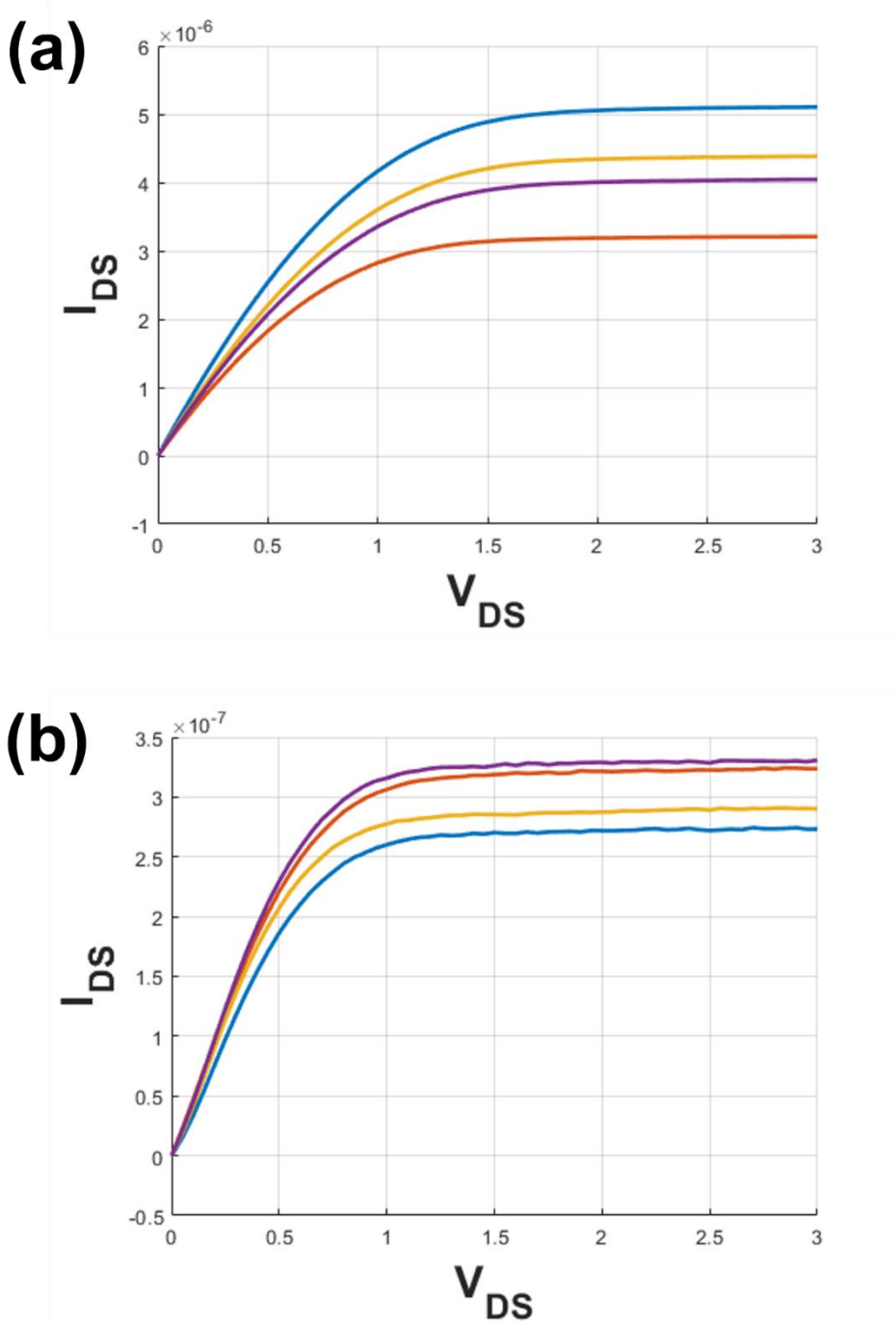

Figure 3.1.4 제작된 트랜지스터의 output curve. (a)는 Wafer 1, (b)는 Wafer 2의 결과이다.

가중치 저장을 위한 커패시터는 high-k dielectric 물질인 HfO_2를 사용하여 metal-insulator-metal (MIM) 구조로 제작하였다. 커패시터에 전하를 저장하는 장치의 경우, 누설 전류가 동일하다는 가정하에 커패시터의 용량이 클수록 수식 (2)에 의해 전압 정보의 손실이 적기 때문에 HfO_2를 사용하였다. 제작한 커패시터는 100 μm x 100 μm의 면적을 가지며, ALD로 증착한 HfO_2 두께는 10 – 15 nm이다. HP4284A LCR meter로 용량을 측정한 결과 −3 – 3 V 전압 범위에서 Figure 3.1.6과 같이 110 – 160 pF 정도이다.

$$\Delta V = I_{leakage} \times \Delta t / C_{storage} \quad (2)$$

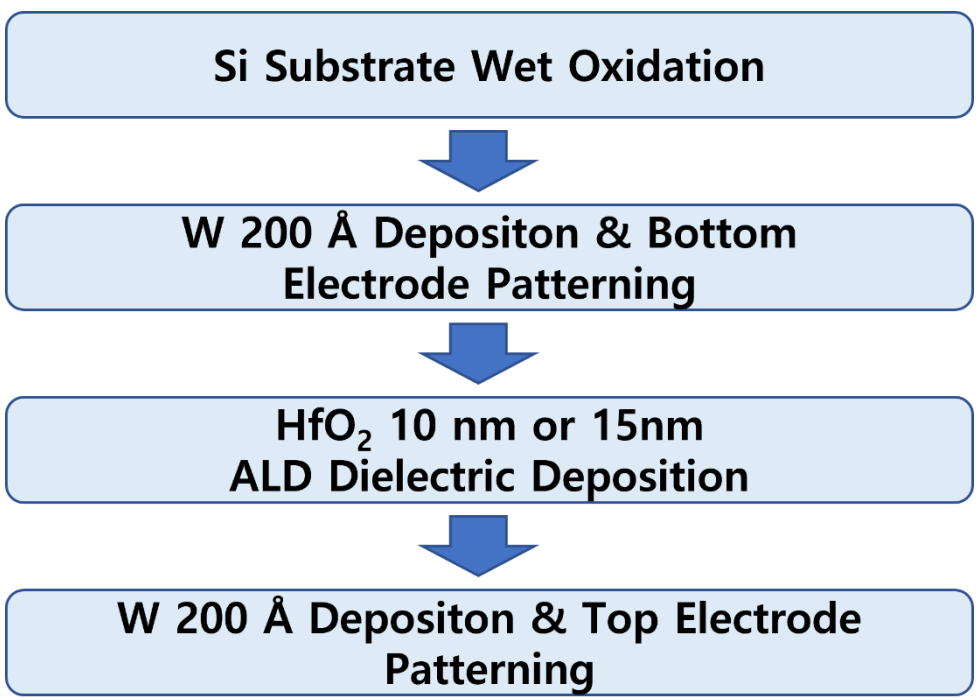

Figure 3.1.5 커패시터 공정 순서

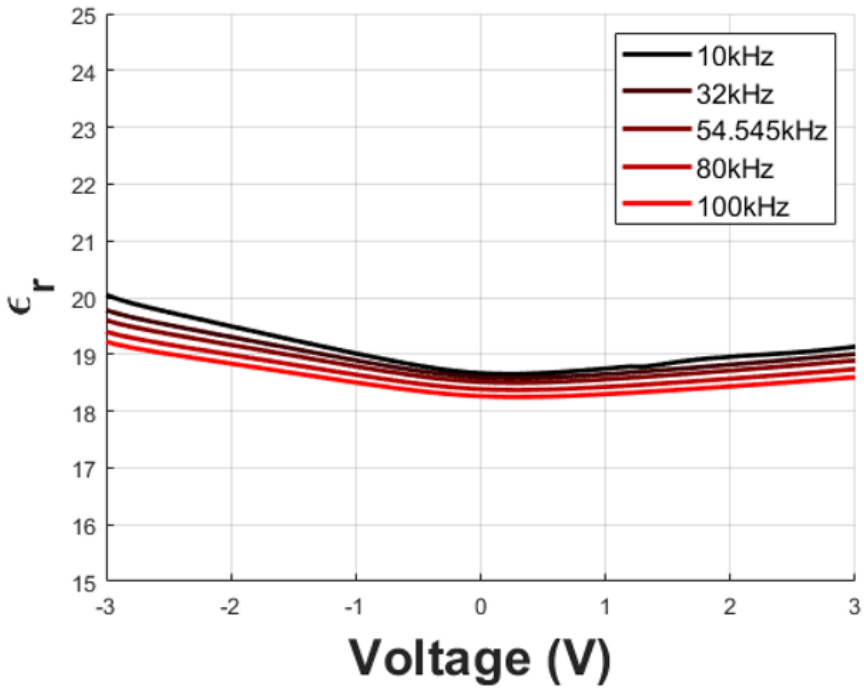

Figure 3.1.6 커패시터 C-V 측정 결과. HfO_2 10 nm, 100x100 μm^2 커패시터를 측정하였다. 흔히 알려진 HfO_2의 relative permittivity인 18-20 사이 값을 가짐을 통해 유전막이 정상적으로 증착된 것을 확인하였다.

3.2 3T1C 구조 및 동작

3T1C의 구조는 [33]의 CMOS 3T1C와 유사한 구조이다. 2개의 시냅스 가중치 갱신용 트랜지스터(update 트랜지스터)와 1개의 읽기 트랜지스터(read 트랜지스터), 그리고 전하를 저장하는 커패시터로 이루어진다. 시냅스 가중치를 증가시키는 potentiation 트랜지스터(N1)의 drain은 V_{DD}와, source는 커패시터의 상단 전극과 연결되어 있으며, 가중치를 감소시키는 depression 트랜지스터(N2)의 drain은 커패시터의 상단 전극과, source는 GND에 연결되었다. Read 트랜지스터의 gate가 커패시터의 상단 전극과 연결되어 커패시터의 전압을 읽을 수 있도록 하였다. Read 트랜지스터의 drain에는 작은 전압을 가해 linear region에서 동작할 수 있도록 하였다.

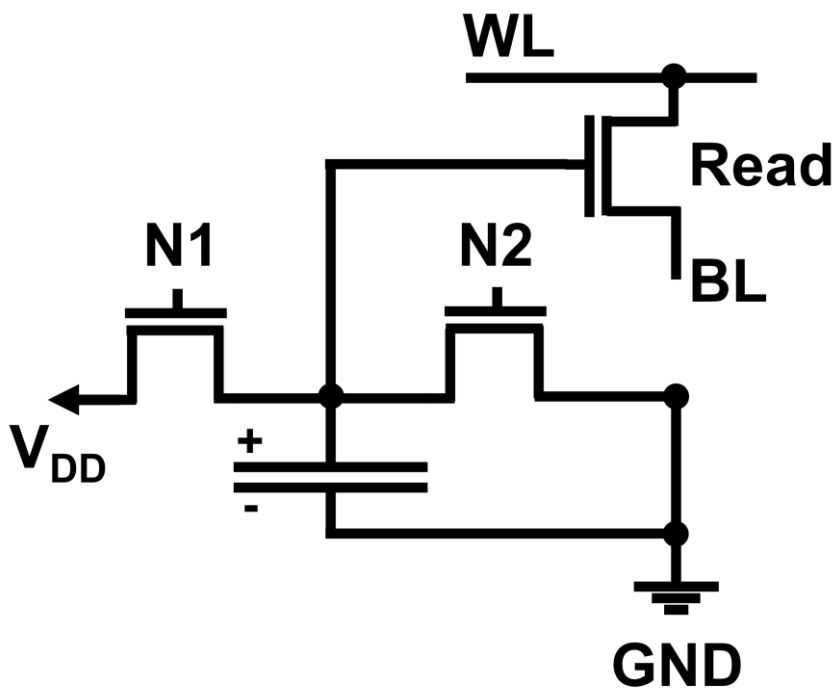

Figure 3.2.1 3T1C 회로도. 두 개의 update 트랜지스터, 하나의 read 트랜지스터, 하나의 커패시터로 이루어졌다.

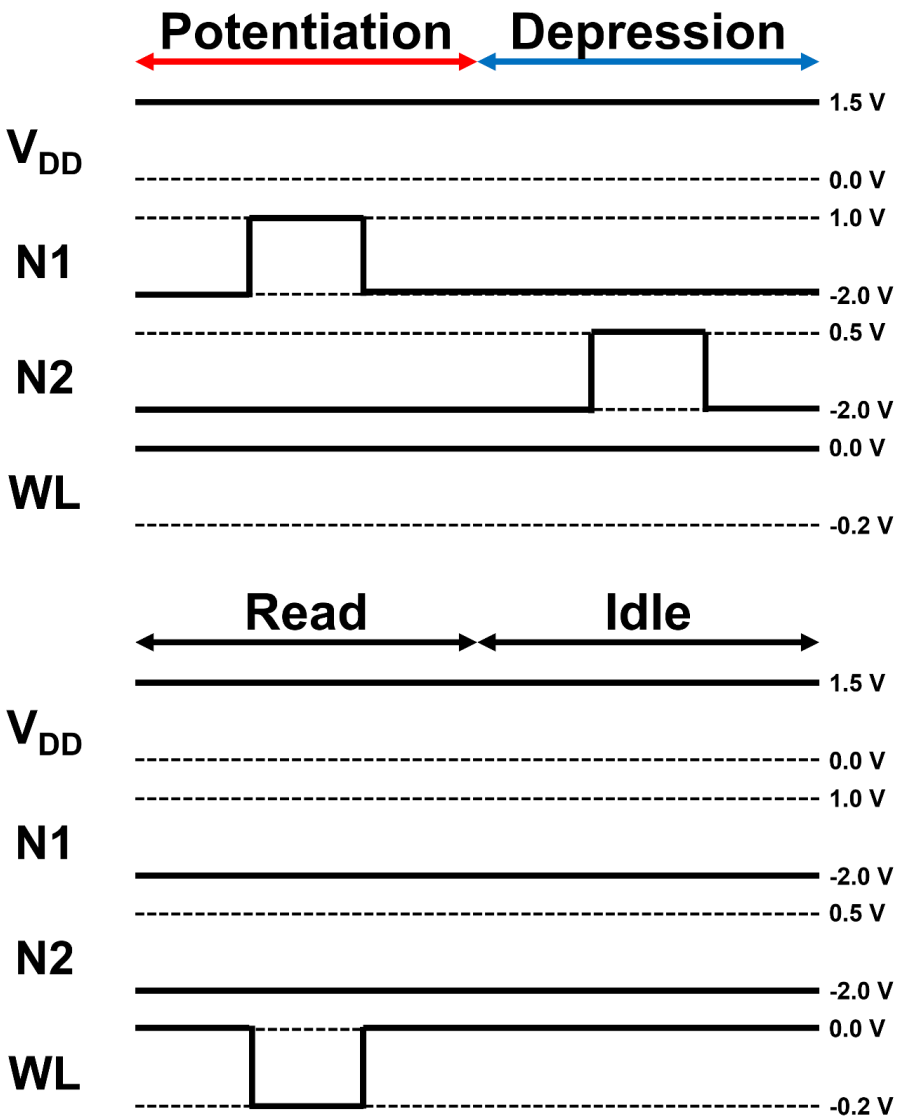

Figure 3.2.2 3T1C 동작 방식의 예시. 회로 동작은 크게 potentiation, depression, read, idle (data retention)으로 나눌 수 있다.

다만, 3T1C array를 제작하였을 때 random access가 가능하도록 select 해주는 소자가 필요하다. 이 문제는 Figure 3.2.3과 같이 update transistor의 gate에 AND gate를 추가하여서 해결할 수 있다. AND gate의 두 input이 동시에 들어왔을 때만 update transistor의 gate에 신호가 전달되어 array 내 특정 소자만 가중치를 갱신할 수 있다. AND gate를 구성하기 위해 추가 transistor가 필요하다는 단점이 있지만, a-IGZO는 CMOS BEOL 공정에 호환되기 때문에[2, 3] wafer 위에 CMOS 회로를 제작한 뒤, 그 위에 3T1C 소자를 수직 적층하는 방식으로 소자 면적을 최소화할 수 있다.

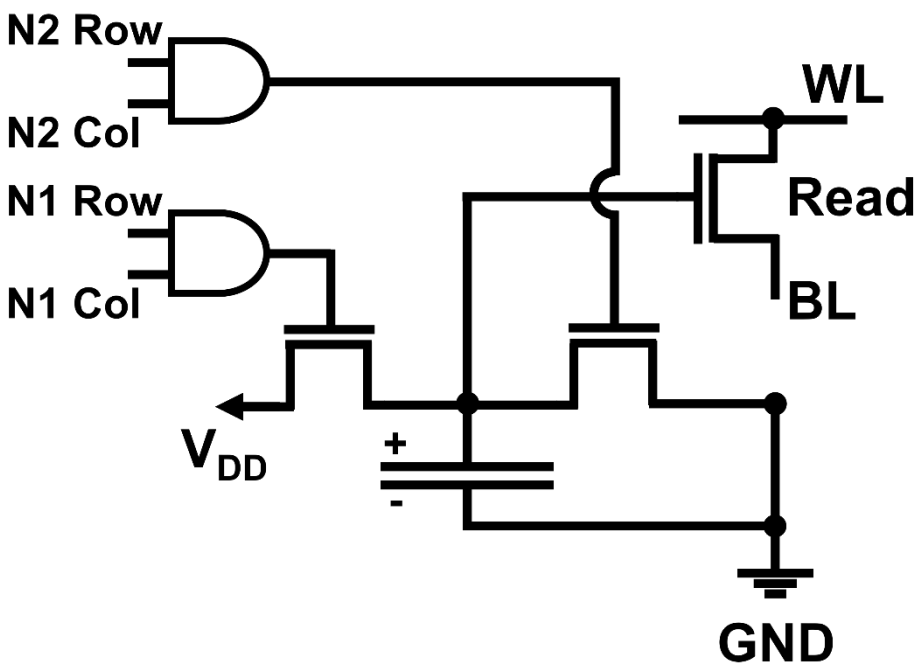

Figure 3.2.3 어레이에서의 3T1C 회로도. 선택 소자로 AND gate를 사용한다.

3.3 심층신경망 가속을 위한 시냅스 소자의 필요성

3T1C 시냅스 소자 및 array 측정은 Arduino DUE micro controller unit (MCU)로 구동되는 PCB 주변 회로로 진행하였다. 측정 환경은 Figure 3.3.1과 같다. 3T1C array 측정을 위해서는 많은 수의 pad가 연결되어야 하므로 45개의 probe가 일렬로 나열된 probe card를 사용하여 측정하였다.

MCU에 미리 시냅스 가중치 갱신, 가중치 retention, cycling endurance test 등의 모든 동작 시나리오를 프로그래밍한 뒤, 실험 조건을 제어 컴퓨터로 MCU와 주변 장비들에 명령을 내리면 조건에 맞추어 PCB 회로와 3T1C 소자에 적절한 전압 신호가 가해지고, 전류 신호를 읽어오는 방식이다. Update 트랜지스터에는 전압 신호가 가해지며, 읽기 과정에서는 read 트랜지스터의 drain에 전압을 가해준 뒤 흐르는 전류를 PCB의 적분기 회로와 MCU의 ADC로 10-bit 정확도를 가지는 정수로 출력한다(Figure 3.3.1). PCB 주변 회로에는 소자에 가해질 전압을 출력하는 power supply가 연결되어 있으며, MCU가 가하는 전압 신호에 따라 3T1C의 각 트랜지스터에 OFF 또는 ON 전압을 가한다. MCU가 가할 수 있는 전압 신호의 최소 길이는 약 100 ns이기 때문에 이보다 더 짧은 전압 신호를 가할 때는 81110A, 81115A 신호 발생기를 사용하였다. PCB 주변 회로에서 나오는 신호를 트리거로 사용하여 짧은 전압 신호를 3T1C 트랜지스터에 가할 수 있다.

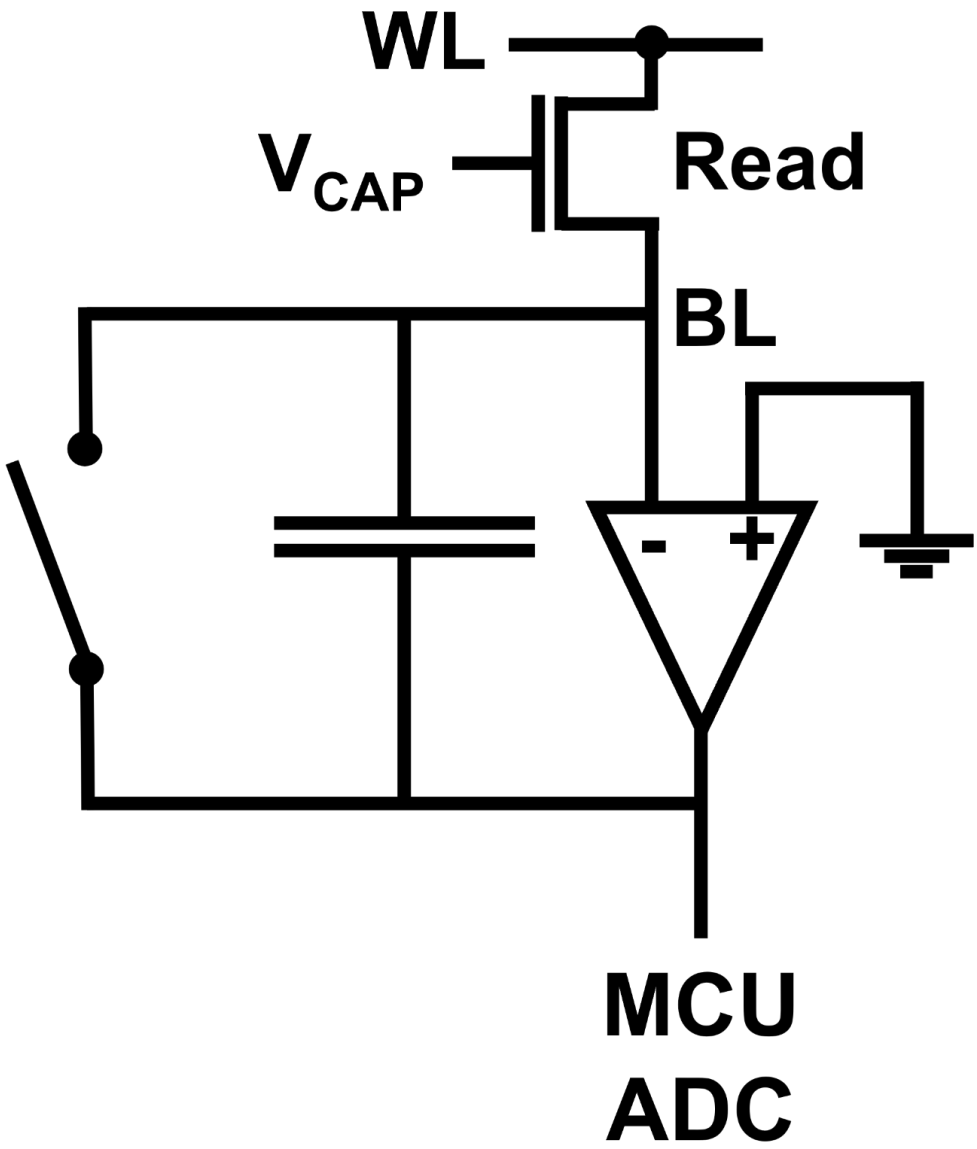

Figure 3.3.1 3T1C read 회로. OP amplifier가 BL을 GND로 잡아주는 동시에 적분 커패시터와 함께 read 동작 또한 수행한다.

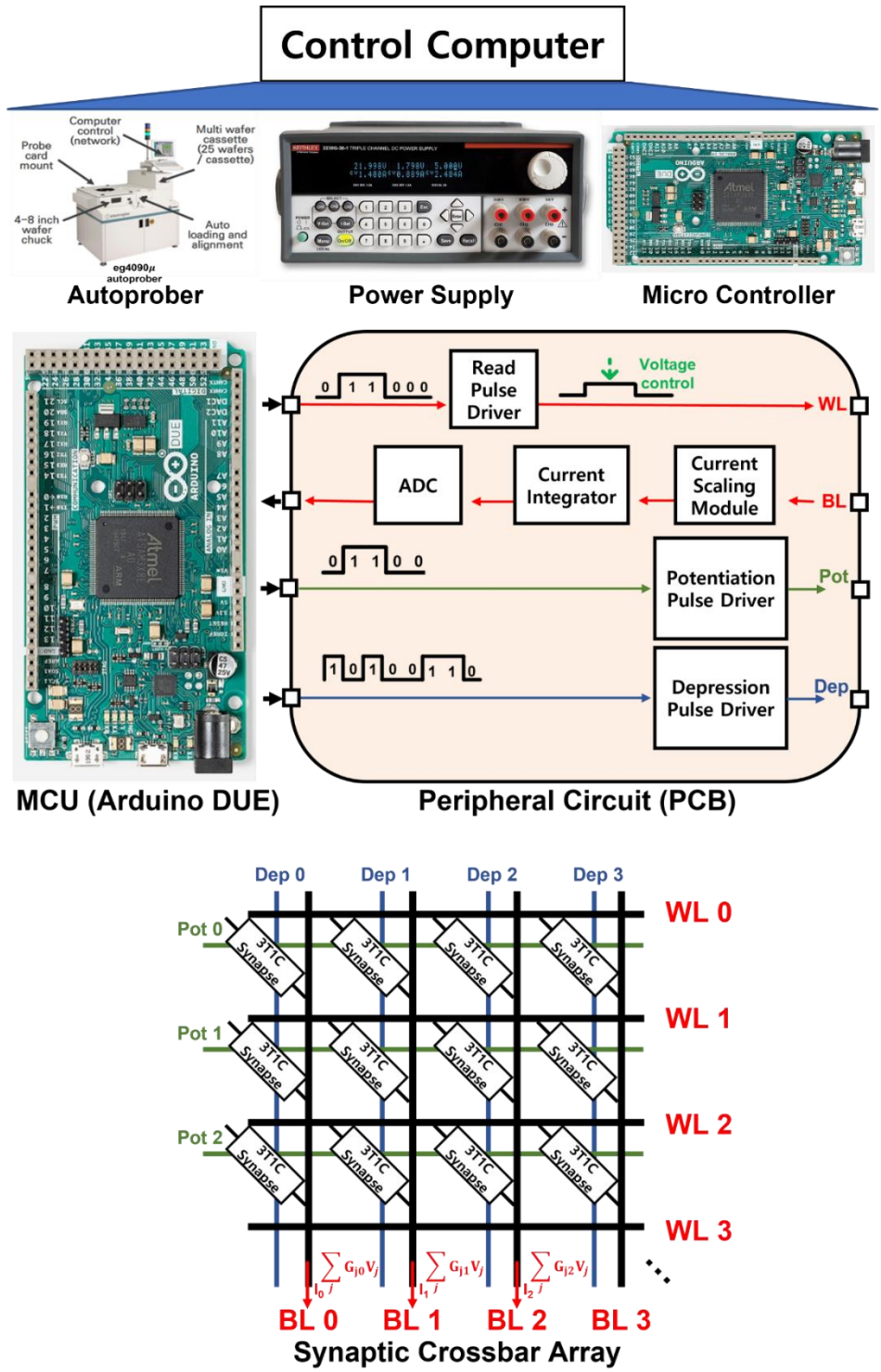

Figure 3.3.2 MCU, PCB를 이용한 주변회로와 측정 환경.

4. 결과 및 논의

4.1 3T1C 가중치 갱신 모델링

4.1.1 모델링 방식

선호되는 가중치 갱신 특성을 가질 수 있는 IGZO TFT로 제작된 3T1C 시냅스 소자이지만, 비이상적인 특성이 존재할 수 있다. 우선 커패시터에 전류를 주입하는 과정인 potentiation은 NMOS만으로 이루어진 소자에서는 완벽히 선형적일 수 없다. N1 트랜지스터의 source 전압은 커패시터의 전압이고, 커패시터에 저장된 전압에 따라 N1 트랜지스터의 V_{DS}와 V_{GS}가 변한다. 이처럼 potentiation에 따라서 N1 트랜지스터가 켜지는 정도가 변하고, 가중치 갱신량이 달라져 비선형적인 가중치 갱신이 일어날 수 있다. 또한, 반대 방향의 갱신 과정인 depression도 일정 구간에서는 비선형적인 갱신이 일어날 수 있다. 트랜지스터가 saturation 영역에서 동작하면 저장된 커패시터 전압에 무관하게 N2 트랜지스터에 일정 전류가 흘러 선형적으로 갱신되지만, 가중치가 충분히 작아져 N2 트랜지스터의 V_{DS}가 V_{GS}보다 작아진다면 트랜지스터가 linear 영역에서 동작해 비선형적인 갱신이 일어날 수 있다. 이런 이상적이지 않은 특성들을 파악, 개선하고, 최종적으로는 NeuroSim[7]이나 PyTorch와 같은 인공 신경망 시뮬레이터로 학습 정확도를 파악하기 위해서는 모델링이 필요하다. 따라서 본 연구에서는 유한요소법(Finite Element Method, FEM)을 이용한 근사 모델과 수식적으로 풀어낸 모델 두 가지로 3T1C의 가중치 갱신을 평가하였다.

모델링은 흔히 알려진 NMOS drain 전류 식인 수식 (3), (4)를 이용하였다. μ_n는 전자의 mobility, C_{ox}는 gate insulator의 단위 면적 당 커패시턴스, λ는 saturation region에서의 비이상적 전류를 표현하기 위한

파라미터이다.

$$I_{DS} = \mu_n C_{ox} \frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right](1 + \lambda V_{DS})$$

$$(V_{GS} > V_T, V_{DS} \leq V_{GS} - V_T) \quad (3)$$

$$I_{DS} = \frac{1}{2}\mu_n C_{ox} \frac{W}{L}(V_{GS} - V_T)^2(1 + \lambda V_{DS})$$

$$(V_{GS} > V_T, V_{DS} > V_{GS} - V_T) \quad (4)$$

FEM 모델링의 결과는 Figure 4.1.1.1과 같았다. 3T1C는 동작마다 하나의 트랜지스터만 동작하기 때문에 간단한 in-house MATLAB 모델을 제작하였다. 외부 전압 조건과 커패시터 전압에 따라 수식 (3), (4)를 이용하여 update 트랜지스터에 흐르는 전류를 계산한 뒤, 흐르는 전하량만큼 커패시터의 전압을 갱신하는 과정을 반복하였다. 전류를 계산할 때 필요한 파라미터들은 Figure 3.1.6에서 측정한 값을 사용하였다. 예측하였던 것과 같이 potentiation은 전반적으로 비선형성이 드러났고, depression은 낮은 커패시터 전압 영역에서만 비선형적인 가중치 갱신이 일어남을 확인하였다.

수식적으로 해를 구하는 방식 또한 시도하였다. Read 트랜지스터가 완벽하게 선형적인 V_G-I_{DS} 관계를 보인다고 가정하였을 때 가중치의 선형성은 커패시터 전압이 시간에 따라 얼마나 선형적으로 변화하는지를 분석하면 된다. Potentiation은 특이하게 수식 (5)에 의해 외부 조건인 V_{DD}, $V_{N1,ON}$ 조건에 의해서 saturation 또는 linear 영역에서 동작할지가 결정된다. 따라서 potentiation 전 구간을 하나의 수식으로 나타낼 수 있었고, 이는 수식 (6), (7)을 풀어서 얻을 수 있었다. 수식 (9)에서는 가중치 갱신량이 가중치에 대한 이차식 형태를 가진다는 것을 확인할 수 있었으며, 수식 (10)을 통해서는 saturation region에서의 potentiation

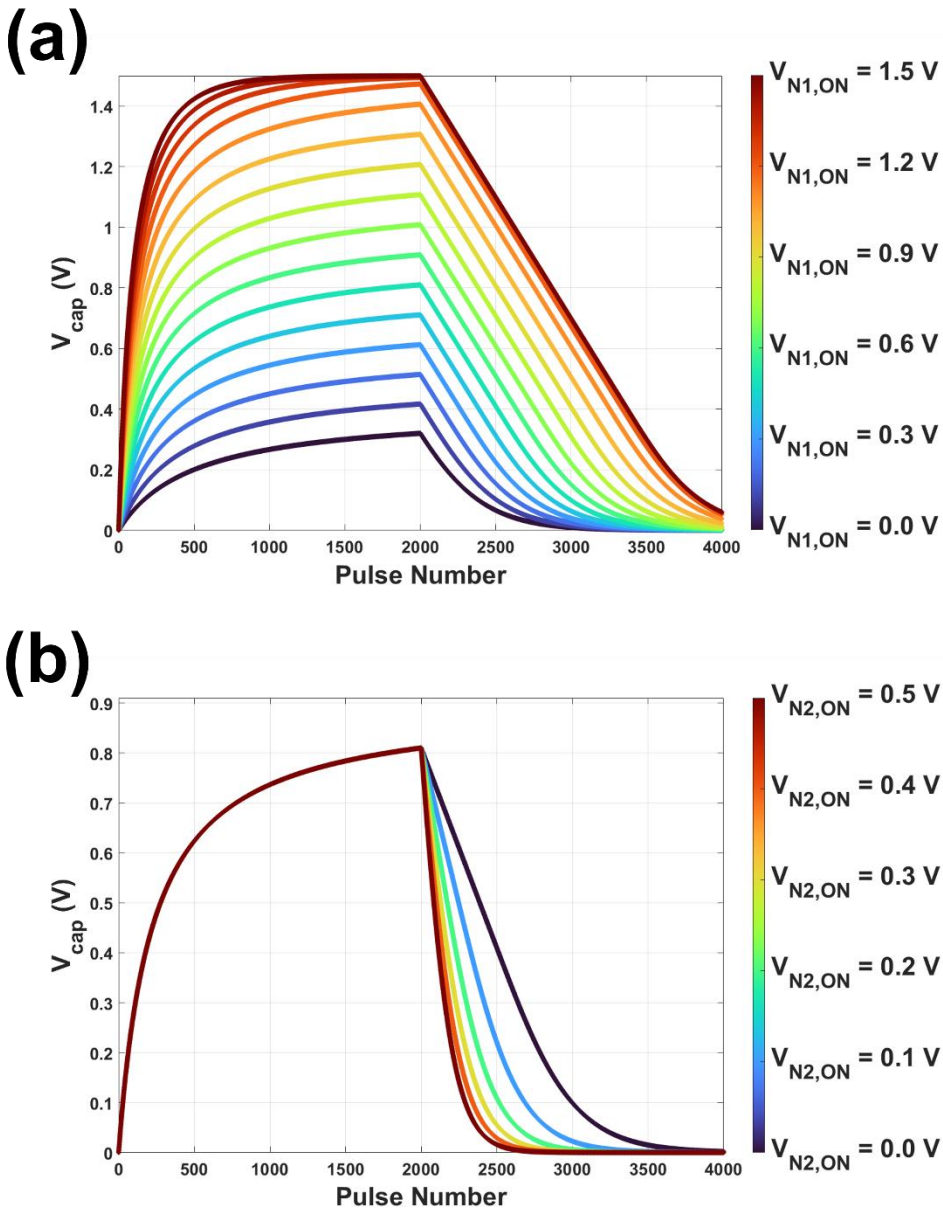

Figure 4.1.1.1 FEM 모델에서의 전압에 따른 가중치 갱신. (a)는 N1 트랜지스터의 gate에 가해지는 전압, (b)는 N2 트랜지스터의 gate에 가해지는 전압에 따른 potentiation-depression 경향성 변화 결과이다.

가중치 변화 양상을 예측할 수 있었다. 매우 작은 값인 λ는 무시하였다.

$$\begin{aligned} V_{DS} - V_{GS} + V_T &= (V_{DD} - V_{cap}) - (V_{N1,ON} - V_{cap}) + V_T \\ &= V_{DD} - V_{N1,ON} - V_T \end{aligned} \quad (5)$$

$$\frac{dI_{DS}\left(V_{cap}(t)\right)}{dV_{cap}(t)} \approx K\left(V_{cap}(t) - (V_G - V_T)\right) \quad (6)$$

$$V_{cap}(t) = \left(\frac{1}{C_{str}}\right)\int_0^t I_{DS}\left(V_{cap}(t')\right)dt' \quad (7)$$

$$C_{str}V''_{cap}(t) = KV'_{cap}(t)\left(V_{cap}(t) - (V_G - V_T)\right) \quad (8)$$

$$\frac{dG}{dn} \propto \frac{dV_{cap}(t)}{dt} = \frac{K}{2C_{str}}\left(V_G - V_T - V_{cap}(t)\right)^2 + D \quad (9)$$

$$V_{cap}(t) = (V_G - V_T) - \frac{1}{\frac{K}{2C_{str}}t + \frac{1}{V_G - V_T}} \quad (10)$$

Depression도 앞선 방법과 동일한 방식으로 수식적인 해를 구할 수 있었다. 다만 depression 과정은 가중치에 따라 N2 트랜지스터가 동작하는 영역이 달라지기 때문에 두 구간으로 나누어 해를 구했다. Saturation 영역에서 동작할 때는 항상 동일한 전류가 흐르기 때문에 간단하게 수식 (11)과 같은 해를 보이며, 커패시터 전압이 $V_G - V_T$보다 작아지는 구간부터는 수식 (12)-(16)과 같은 방식으로 해를 구할 수 있었다.

$$V_{cap}(t) = V_{max} - \frac{K}{2C_{str}}(V_G - V_T)^2 t \quad (11)$$

$$\frac{dI_{DS}\left(V_{cap}(t)\right)}{dV_{cap}(t)} \approx K\left((V_G - V_T) - V_{cap}(t)\right) \quad (12)$$

$$V_{cap}(t) = (V_G - V_T) - \left(\frac{1}{C_{str}}\right)\int_0^t I_{DS}\left(V_{cap}(t')\right)dt' \quad (13)$$

$$C_{str}V''_{cap}(t) = -KV'_{cap}(t)\left(V_{cap}(t) - (V_G - V_T)\right) \quad (14)$$

$$\frac{dV_{cap}(t)}{dt} = \frac{K}{2C_{str}}\left(V_G - V_T - V_{cap}(t)\right)^2 - \frac{K}{2C_{str}}(V_G - V_T)^2 \quad (15)$$

$$V_{cap}(t) = \frac{2(V_G - V_T)exp(-\frac{K}{C_{str}}(V_G - V_T)t)}{exp\left(-\frac{K}{C_{str}}(V_G - V_T)t\right) + 1} \quad (16)$$

Depression 또한 potentiation과 마찬가지로 이차식의 갱신량-가중치 관계식을 가졌으며, 수식 (10), (11), (16)을 이용해 potentiation-depression 과정을 예측한 결과는 Figure 4.1.1.2와 같다. FEM 모델과 수식적인 해가 동일한 가중치 갱신 양상을 가졌고, 모델링이 문제없이 되었다는 것을 확인하였다.

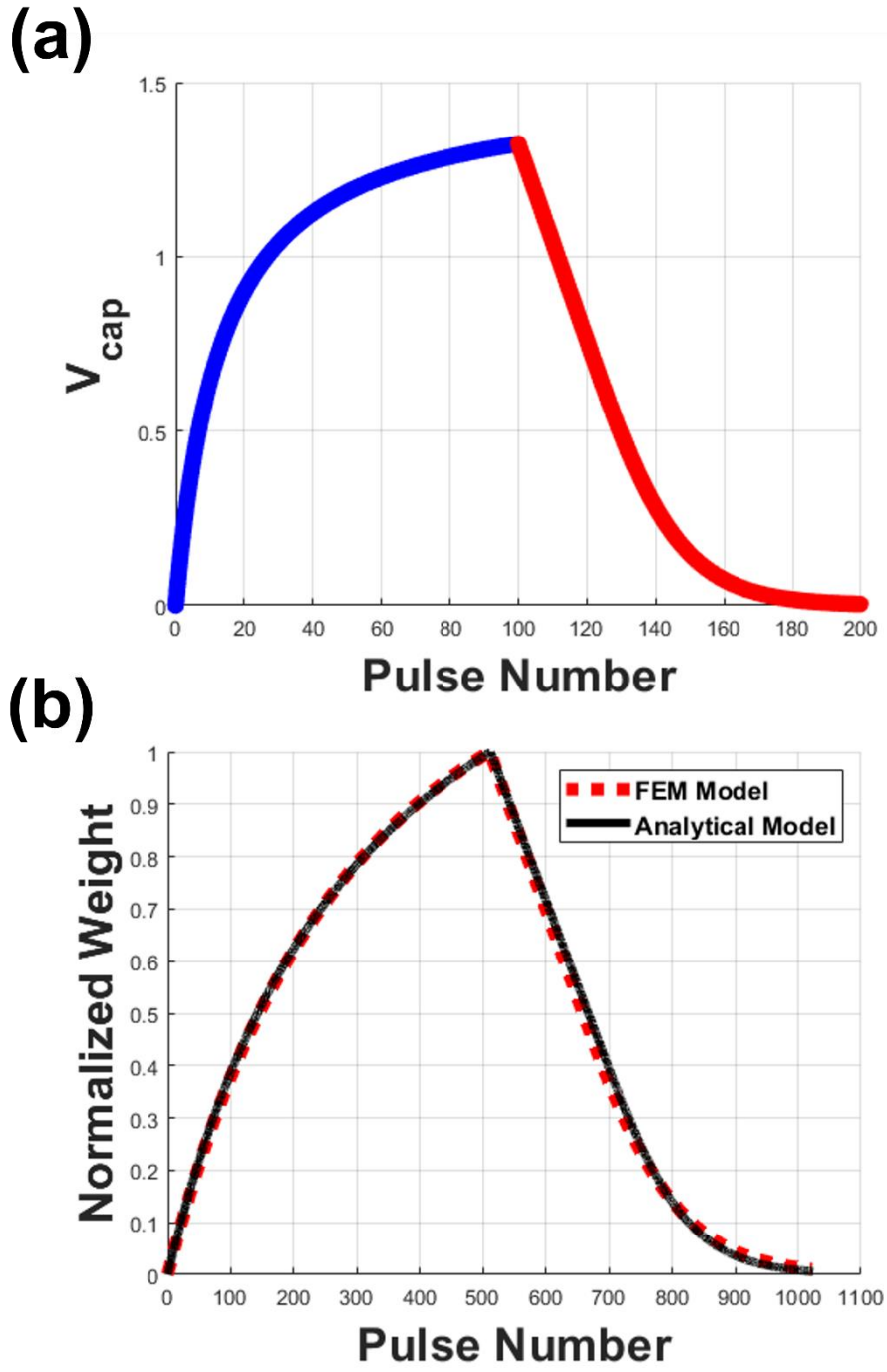

Figure 4.1.1.2 수식적인 모델의 결과. (a)는 수식적인 해를 통해 예측한 potentiation-depression 결과이며, (b)에서 수식적인 해와 FEM 모델이 일치하는 것을 확인하였다.

4.1.2 가중치 갱신의 선형 대칭성 평가 방법

최적의 3T1C 동작 전압 조건 탐색을 위해서는 가중치 갱신의 비선형성, 비대칭성의 정량적인 평가 방법이 필요하다. [6, 7, 32]과 같이 연구나 소자에 따라서 평가 방법이 다르지만, 대체로 지수함수 기반의 모델을 이용한다. 다만, 시냅스 소자에 따라 비선형성이 나타나는 물리적인 이유가 다르기 때문에 3T1C에 한해서는 새로운 선형 대칭성 평가 방법이 필요하다.

앞서 수식적으로 구한 해인 수식 (9)와 수식 (15)는 Brivio *et al.*의 연구와[4] 비슷한 형태이지만, 두 식은 시냅스 소자의 동작 영역에서의 비선형성, 비대칭성을 평가하기에는 알맞지 않다. 3T1C 가중치 갱신의 평가 지표가 될 수 있는 수식이 필요하며, 본 연구에서는 표준편차 기반의 수식 (17)과 수식 (18)을 통해 가중치 갱신의 선형성과 갱신량을 평가하였다.

$$\boldsymbol{NL = \frac{1}{< \Delta ADC >}\sqrt{\frac{1}{N}\Sigma(\Delta ADC - < \Delta ADC >)^2} \times 100\%} \qquad \textbf{(17)}$$

$$\boldsymbol{\alpha = \frac{< \Delta ADC >}{n}} \qquad \textbf{(18)}$$

NL은 가중치 갱신의 선형성을 평가하는 지표이며, α 는 가한 programming pulse 당 가중치 갱신의 양을 나타낸다. NL이 0에 가까울수록 평균에서 벗어나지 않는 선형적인 갱신을 의미하며 α가 클수록 한 programming pulse가 일으키는 가중치 갱신이 크다. $< \Delta ADC >$는 사용하는 커패시터 전압 구간에서의 평균적인 가중치 갱신을 의미하며, N은 읽은 지점들 수, n은 읽은 지점 사이에 가한 pulse 수를 의미한다. 일반적으로 통계 기반의 선형성 모델은 가중치 갱신의 경향을 정확하게 파악

하기에 불리하지만, IGZO 3T1C의 경우 potentiation, depression 모두 가중치 갱신이 정확하게 모델링 되기 때문에 이러한 평가법을 사용해도 문제가 없다.

Potentiation과 depression의 비대칭성은 수식 (19), (20)으로 선형성의 대칭성과 갱신량의 대칭성을 평가하였다.

$$\boldsymbol{max}\left(\frac{\boldsymbol{NL_{pot}}}{\boldsymbol{NL_{dep}}}, \frac{\boldsymbol{NL_{dep}}}{\boldsymbol{NL_{pot}}}\right) \quad (\mathbf{19})$$

$$\boldsymbol{max}\left(\frac{\boldsymbol{\alpha_{pot}}}{\boldsymbol{\alpha_{dep}}}, \frac{\boldsymbol{\alpha_{dep}}}{\boldsymbol{\alpha_{pot}}}\right) \quad (\mathbf{20})$$

4.2 3T1C 가중치 갱신

4.2.1 측정 결과와 모델링 비교

가중치 갱신 실험은 우선 N2 트랜지스터를 충분히 긴 시간 동안 켜서 커패시터 전압을 0 V로 만든 뒤, potentiation과 depression 순서대로 진행하였고, 시냅스 가중치는 read 트랜지스터에 흐르는 전류를 analog to digital converter (ADC)를 통해 디지털화하여 측정하였다. 커패시터에 저장된 전압은 ADC 측정값으로부터 간접적으로 계산하였다. Figure 4.2.1.1과 같이 읽기 조건과 동일한 V_D, V_S에서의 read 트랜지스터의 transfer curve를 측정한 뒤, transfer curve에서 변환된 I_{DS}에 대응되는 전압으로 커패시터에 저장된 전압을 추측하였다. I_{DS} 값은 ADC 측정값을 수식 (21)를 통해 변환하였다. 측정한 transfer curve 데이터에 해당하는 전류 값이 없는 경우는 선형 보간법을 통해 커패시터 전압을 산출하였다. C_{int}와 V_{int}는 적분기 커패시터의 커패시턴스와 전압이며, ADC가 0–3.3 V를 10–bit precision으로 변환하기 때문에 수식 (21)와 같이 계산하였다.

$$\boldsymbol{I_{DS}} = \frac{\boldsymbol{Q}}{\boldsymbol{t}} = \frac{\boldsymbol{Q_{int}}}{(\boldsymbol{read\ time})} = \boldsymbol{C_{int}} \times \frac{\boldsymbol{V_{int}}}{(\boldsymbol{read\ time})}$$
$$= \boldsymbol{C_{int}} \times (\boldsymbol{ADC\ value}) \times \frac{\frac{\boldsymbol{3.3}}{\boldsymbol{1023}}}{(\boldsymbol{read\ time})} \qquad (\boldsymbol{21})$$

측정 결과는 Figure 4.2.1.2와 같다. FEM 모델이 예측한 가중치 갱신의 경향과 실제 측정된 potentiation–depression 경향이 상당 부분 일치하는 것을 확인하였다. 모델과 측정 사이의 차이는 update 트랜지스터들의 산포, read 트랜지스터의 비선형성 등의 효과에 의해 나타났다고 추정하였다.

IGZO TFT는 n-type만 존재하기 때문에 예상하였던 것처럼 potentiation 과정에서 약간의 비선형성을 확인할 수 있었고, depression도 높은 커패시터 전압 구간에서는 선형적인 갱신이 일어나지만 낮은 전압에서는 비선형적인 갱신이 일어나는 것을 확인하였다.

Figure 4.2.1.3에서는 potentiation의 $V_{cap}-dV_{cap}$을 나타냈다. 수식 (9)이 예측한 이차식의 경향이 나타났으며, 이차식을 fitting 했을 때 V_G-V_T의 값 또한 실험에 사용한 소자와 동작 조건에 일치하였다.

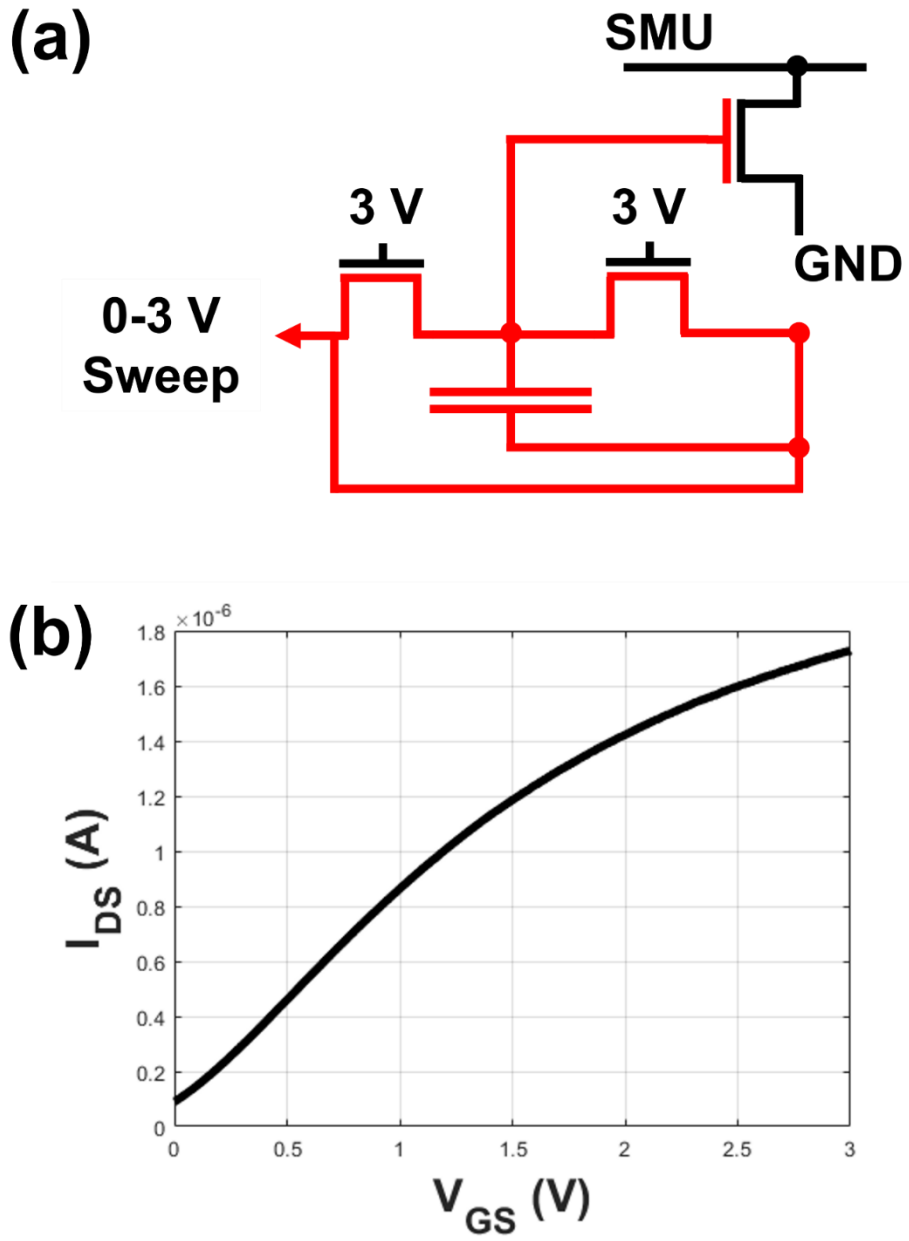

Figure 4.2.1.1 Read 트랜지스터 (a)측정 방법과 (b)결과. N1과 N2를 충분히 키면 V_{DD}/GND 전압이 온전히 read 트랜지스터의 gate에 전달될 수 있다.

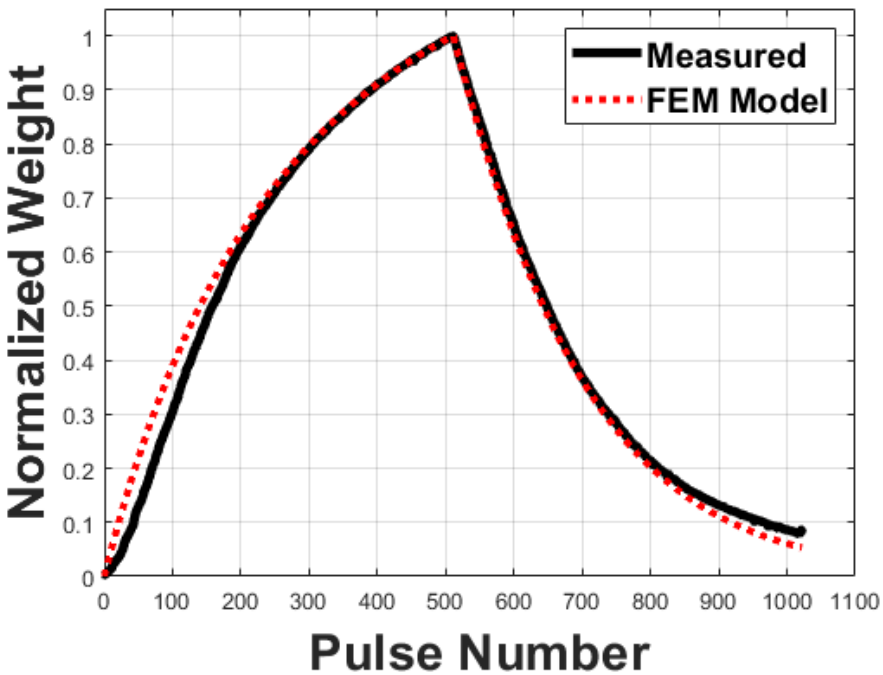

Figure 4.2.1.2 가중치 갱신의 측정값과 FEM 모델의 비교

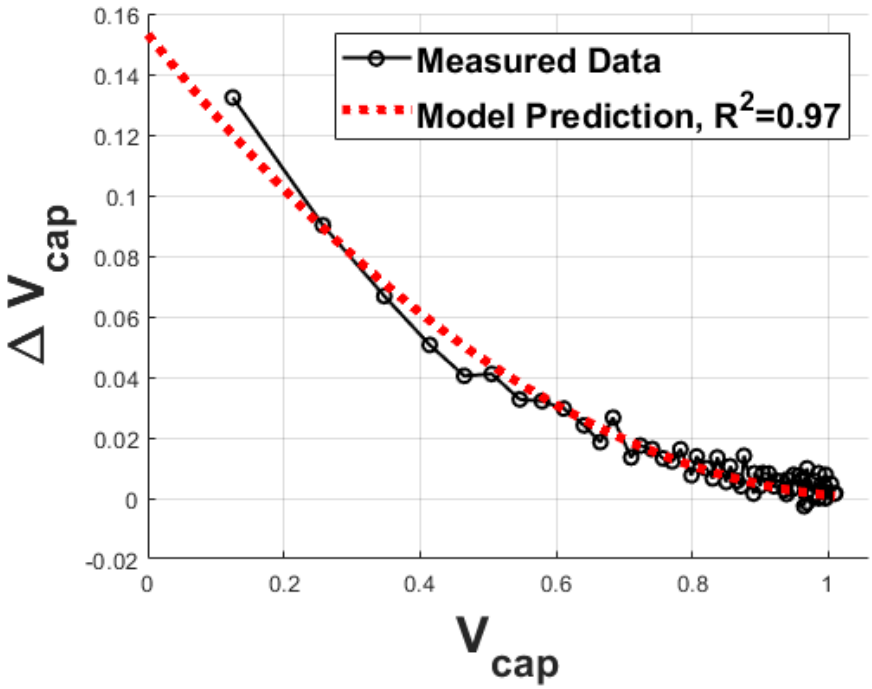

Figure 4.2.1.3 가중치 갱신 변화의 측정값과 FEM 모델의 비교

4.2.2 시냅스 소자의 고속 동작

비휘발성 메모리가 아닌 트랜지스터를 이용하여 시냅스를 제작했을 때의 장점 중 하나는 가중치 갱신 동작이 고속으로 가능하다는 것이다. IGZO TFT는 수 ns 수준으로 동작할 수 있다는 사실이 보고된 바 있으며[50], 최적화를 통해 더 빠른 동작 또한 가능할 것이다. 또한, 보다 작은 면적의 커패시터를 사용하기 위해서는 한 번의 갱신 때 N1, N2 트랜지스터에 흐르는 전류량을 자유롭게 조절할 수 있어야 한다. 신호발생기로 ns 수준의 가중치 갱신 실험 결과는 Figure 4.2.2.1과 같다. 8 ns 갱신 펄스까지 시냅스가 정상 동작하였고, 이를 통해 3T1C 시냅스의 고속 동작 가능성뿐만 아니라 scalability, 저전력 동작 또한 확인할 수 있었다.

표 2 3T1C 소자 동작 속도 및 전력 소비 예상

	MCU Peripheral	FPGA Peripheral*
Program Speed	1 μs	< 5 ns
Read Speed	10 - 20 ms	< 8 μs
Program Energy	< 1 pJ	< 1 fJ
Read Energy	< 10 nJ	< 3 pJ

*FPGA와 tape-out 주변회로 chip을 사용한 측정

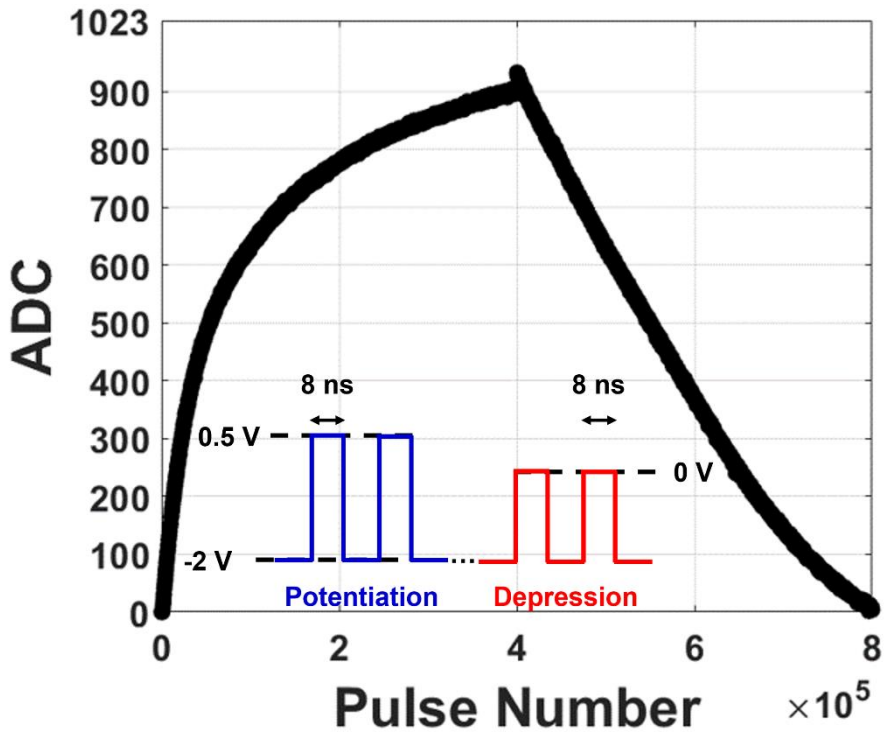

Figure 4.2.2.1 ns 수준의 시냅스 동작

4.2.3 전압 조건에 따른 가중치 갱신

수식 (9)에서는 항상 커패시터 전압 변화가 전압의 이차식에 비례하는 것으로 계산이 되지만, 실제 시냅스 동작에서는 비교적 선형적인 가중치 갱신이 측정되었다. 특히 Figure 4.1.1.1의 FEM 시뮬레이션 결과에서도 전압 조건에 따라서 시냅스의 거동이 달라지는 것을 확인할 수 있었다. 높은 학습 정확도를 가지는 3T1C array 구동을 위해서는 전압 조건에 따른 시냅스 가중치 갱신의 경향을 파악, 최적화하는 과정이 필요하다.

Figure 4.2.3.1은 update 트랜지스터 동작 전압 조건에 따른 가중치 갱신의 선형성과 대칭성을 나타냈다. Potentiation은 N1 gate 전압이 증가할수록 선형성은 향상되었으며, α 는 증가하는 경향을 보였다. Depression도 N2 gate 전압 증가에 따라 α 가 증가하였지만, potentiation과 반대로 비선형성을 나타내는 NL은 높은 gate 전압이 가해질수록 증가하였다.

동작 전압 상승에 따른 α 증가는 수식 (3), (4)으로 설명 가능하였다. 당연하지만 트랜지스터의 gate 전압이 클수록 더 큰 I_{DS}가 흐르기 때문에 한 갱신 신호 당 변하는 ADC 값이 커진 것이다. Potentiation에서 선형성 향상은 Figure 4.2.3.2로 설명할 수 있다. 전압 갱신량은 수식 (9)에서 증명하였듯이 커패시터 전압에 대한 이차식의 관계를 가지는데, N1 트랜지스터의 동작 전압이 상승하면 V_G-V_T가 증가해 동일한 전압 범위에서는 선형성이 향상된다. Depression의 선형성이 potentiation과 반대의 경향을 가지는 이유는 N2 트랜지스터가 linear mode로 동작하는 전압 구간이 달라지기 때문으로 설명하였다. N2 트랜지스터의 동작이 saturation에서 linear 영역으로 바뀌는 경계는 수식 (4)에서 V_G-V_T이다. 동작 전압이 증가하면 V_G-V_T가 증가하여 더 높은 전압 범위부터 linear 영역에서 동작하고, 선형적인 갱신이 일어나는 saturation 영역에

서 동작하는 범위가 줄어 선형성이 나빠지는 것이다. 다만 depression 과정이 대체로 linear 하기 때문에 potentiation만큼의 변화는 없는 것으로 추정되었다.

실험 결과를 통해 선형적인 갱신을 위해서는 potentiation은 높은 gate 전압, depression은 낮은 gate 전압에서 동작해야 한다는 결론을 내릴 수 있었다. 다만 이 경우 선형성과 선형성의 대칭성(수식 (19))은 개선할 수 있지만, 갱신량의 대칭성(수식 (20))은 악화되는 trade-off가 있다. 이는 Figure 4.2.3.3과 같이 커패시터의 하단 전극, V_{DD}, potentiation 전압을 향상시키는 것으로 해결할 수 있다. Potentiation 과정의 전압은 모두 동일한 양이 상승하였기 때문에 갱신에 변화가 없지만, N2 트랜지스터에서는 V_D가 크게 증가한 효과이기 때문에 커패시터에 저장된 전압에 무관하게 항상 saturation 영역에서 동작하게 할 수 있다. Depression 트랜지스터에서 더 이상 선형성을 위해 낮은 전압을 사용할 필요가 없어지기 때문에 자유롭게 갱신량의 대칭성도 해결된다. 커패시터 하단 전극 전압의 조절로 read 트랜지스터의 가장 선형적인 구간을 사용할 수 있다는 부가 효과도 얻을 수 있다.

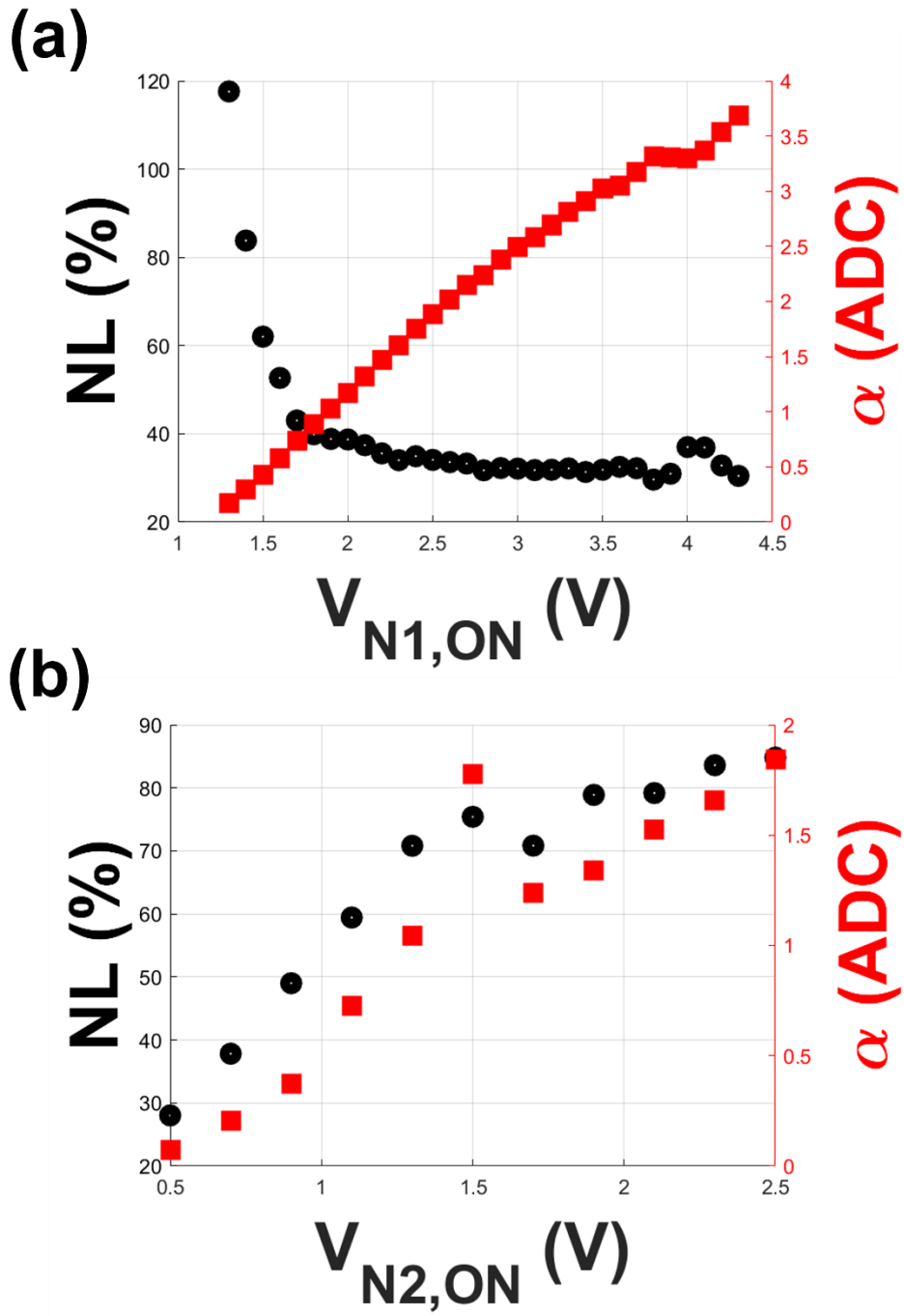

Figure 4.2.3.1 전압에 따른 가중치 갱신의 선형성. (a)는 N1 gate 전압에 따른 potentiation 변화, (b)는 N2 gate 전압에 따른 depression 변화이다.

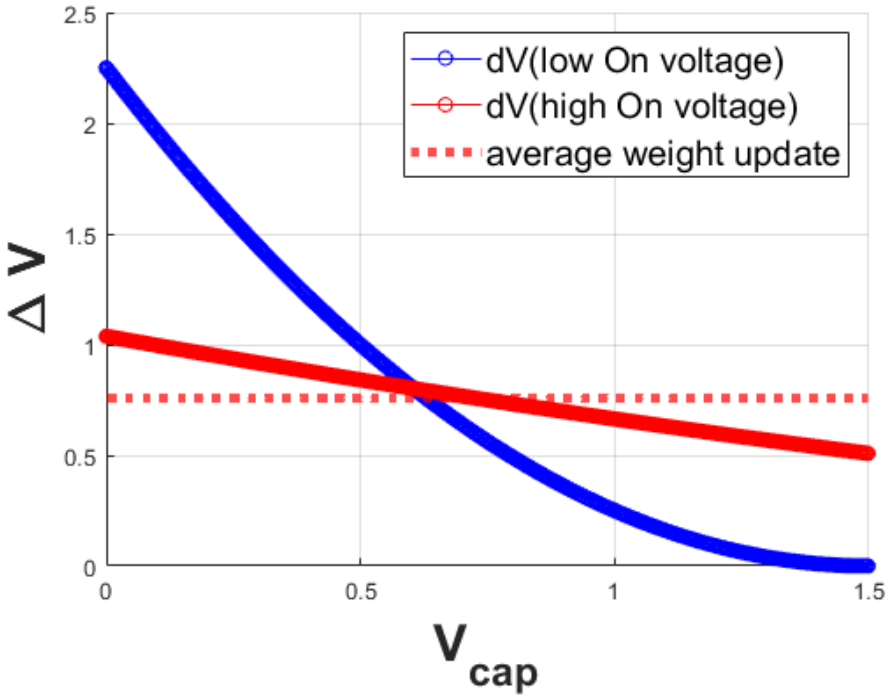

Figure 4.2.3.2 높은 전압에서의 선형적 갱신 해석. 낮은 N1 gate 전압(blue)에 비해 높은 N1 gate 전압(red)이 평균 가중치 갱신량이 같을 때 더 선형적인 가중치 갱신이 일어난다.

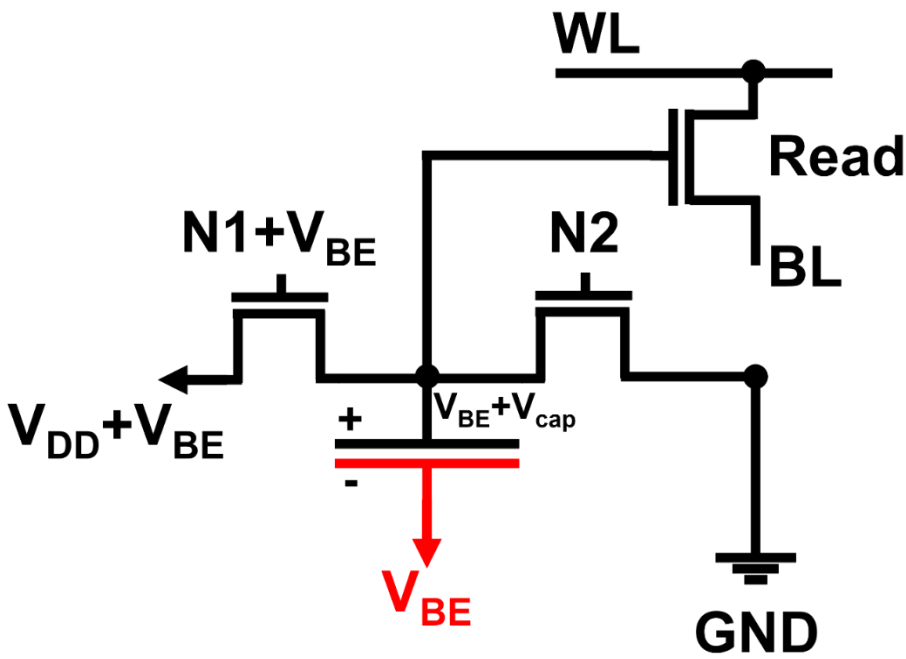

Figure 4.2.3.3 커패시터 하단 전극 boosting. 하단 전극에 V_{BE}가 가해지면 상단 전극에는 $V_{BE}+V_{cap}$이 가해져 N2 트랜지스터의 source-drain 사이 더 큰 전압이 걸리고 saturation region에서 동작하게 된다.

4.2.4 시냅스 산포 평가

3T1C 시냅스 간 산포는 device-to-device (D2D) 산포와 cycle-to-cycle (C2C) 산포 두 종류로 분류할 수 있다. 인공 신경망에서의 아날로그 연산은 fault-tolerant한 것으로 알려졌지만, 산포가 일정 수준을 넘는 경우 on-chip learning의 최종 정확도에 영향을 준다. 특히 D2D는 산포가 보정되는 방향으로 학습이 되며 최종 정확도에 큰 영향을 주지 않지만, C2C 산포 증가는 정확도 하락의 원인이 되는 것으로 알려져 있다[15].

D2D, C2C 산포는 동일 wafer 위의 소자들을 비교하였다. C2C 산포는 한 소자에서 10회 반복 측정해 평가하였고 D2D 산포는 한 die 위의 25개 소자를 비교하였다. Figure 4.2.4.1에서 시냅스가 일관성 있게 동작하는 것을 확인하였다. 낮은 ADC 값에서는 산포가 상대 표준편차가 큰 것으로 나타나는데, 이는 ADC 회로의 noise에 의한 것으로 판단된다. 시냅스 상의 C2C 산포는 무시 가능하며, 커패시터 전압에 무관하게 존재하는 주변 회로의 noise에 의해 발생하였기 때문에 Figure 4.2.4.1(b)와 같은 경향성이 나타났다. 이와 같은 부분은 실제 tape-out 주변 회로를 사용하면 개선이 있을 것으로 기대된다. Figure 4.2.4.2에서 소자 간 산포는 C2C에 비해 큰 것을 확인하였다. 그러나 D2D 산포 또한 낮은 ADC 영역을 제외한다면 상대 표준편차 20% 이하의 일관성 있는 potentiation-depression이 일어났음을 Figure 4.2.4.2(b)에서 볼 수 있다. 소자가 모두 동작하며, D2D 산포가 심각하지 않기 때문에 학습 능력에는 영향이 없을 것으로 판단하였다. D2D 산포는 공정 최적화를 통해 추가 개선이 가능하다.

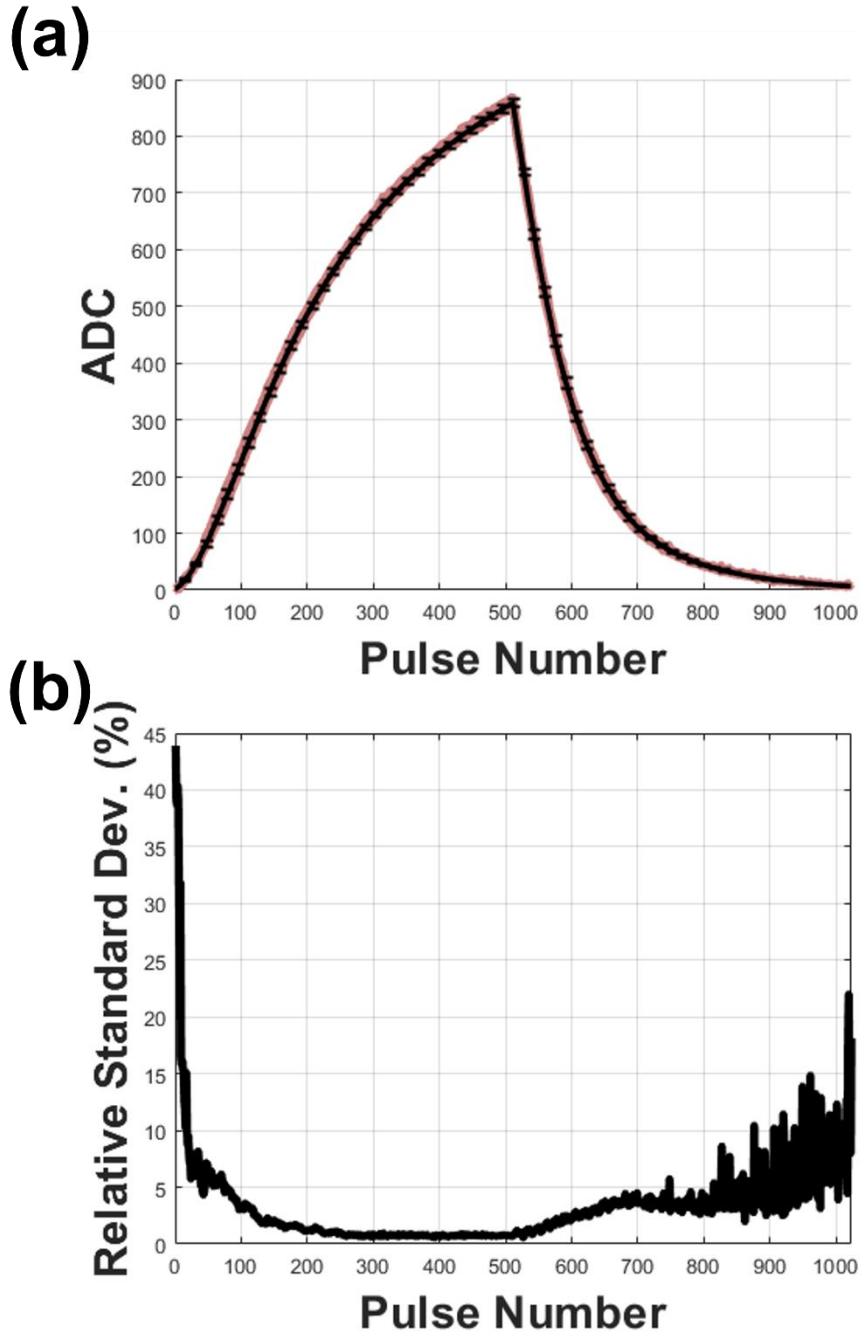

Figure 4.2.4.1 시냅스 소자 가중치 갱신의 cycle-to-cycle 산포. (a)는 10회의 potentiation-depression 과정을 나타낸다. (b)는 pulse number에 따른 상대 표준편차를 나타낸다.

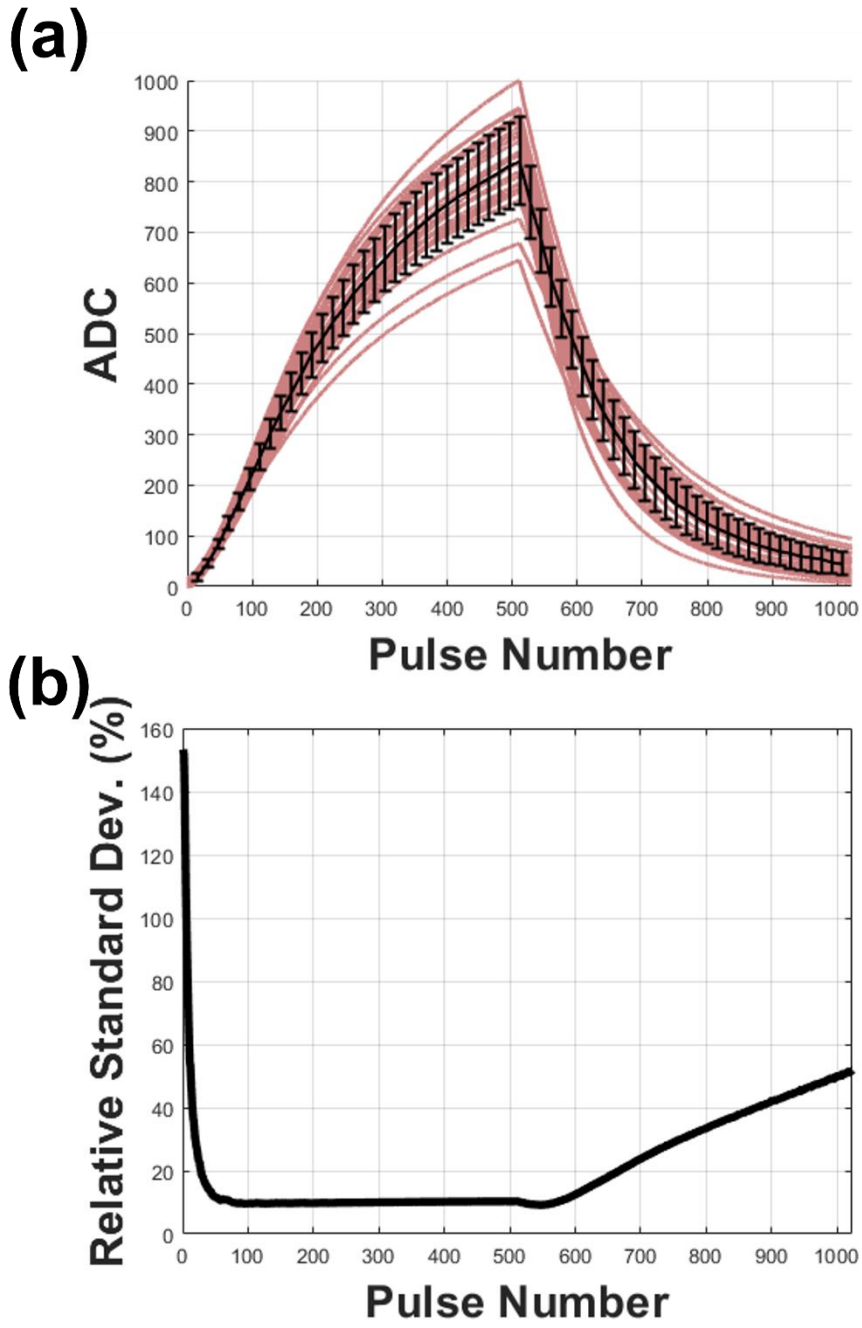

Figure 4.2.4.2 시냅스 소자 가중치 갱신의 device-to-device 산포. (a)는 25개의 device 간 산포를 나타낸다. (b)는 pulse number에 따른 상대 표준편차를 나타낸다.

4.2.5 목표 가중치 도달 능력 확인

시냅스가 인공 신경망 내에서 의도한 대로 동작할 수 있는지 검증하기 위해 시냅스의 가중치가 목표하는 값으로 수렴하는지를 확인하였다. 목표하는 ADC 값을 설정한 뒤, 현재 ADC 값과 목표를 비교하여 가중치를 갱신하는 과정을 반복하여 목표에 수렴할 수 있는지를 확인하였다. 시냅스를 동작한 방식은 Figure 4.2.5.1와 같다. 확률적인 update는 Gokmen *et al.*의 연구에서 입력값과 역전파된 오차 값의 곱을 별도의 연산 장치 없이 계산하고자 사용되었던 방법이며[15], 본래라면 N1, N2의 gate에 연결된 AND gate의 두 입력 값이 모두 도달할 때 update가 되는 방식이지만, 본 연구에서는 AND gate를 구현할 수 없었기 때문에 N1, N2의 gate에 직접 MCU 소프트웨어 상에서 계산된 확률적 update 신호를 인가하였다.

실험 결과는 Figure 4.2.5.2와 같았으며, 가중치의 시작점과 목표에 무관하게 수렴하였다. Potentiation의 경우 Figure 4.2.1.2에서 확인하였던 바와 같이 가중치가 높을수록 포화되는 현상이 있다. Potentiation과 depression으로 동일한 ADC 차이를 갱신하는데 필요한 update cycle 수를 비교했을 때, 적은 ADC 갱신은 potentiation과 depression 간 차이가 없었던 반면, 더 큰 간격의 갱신에서는 potentiation이 depression 보다 더 느리게 수렴하였다. 그러나 수렴 속도에 큰 차이는 없었으며, 모든 실험에서 목표에 수렴하였기 때문에 3T1C의 시냅스 소자로의 가능성을 검증할 수 있었다.

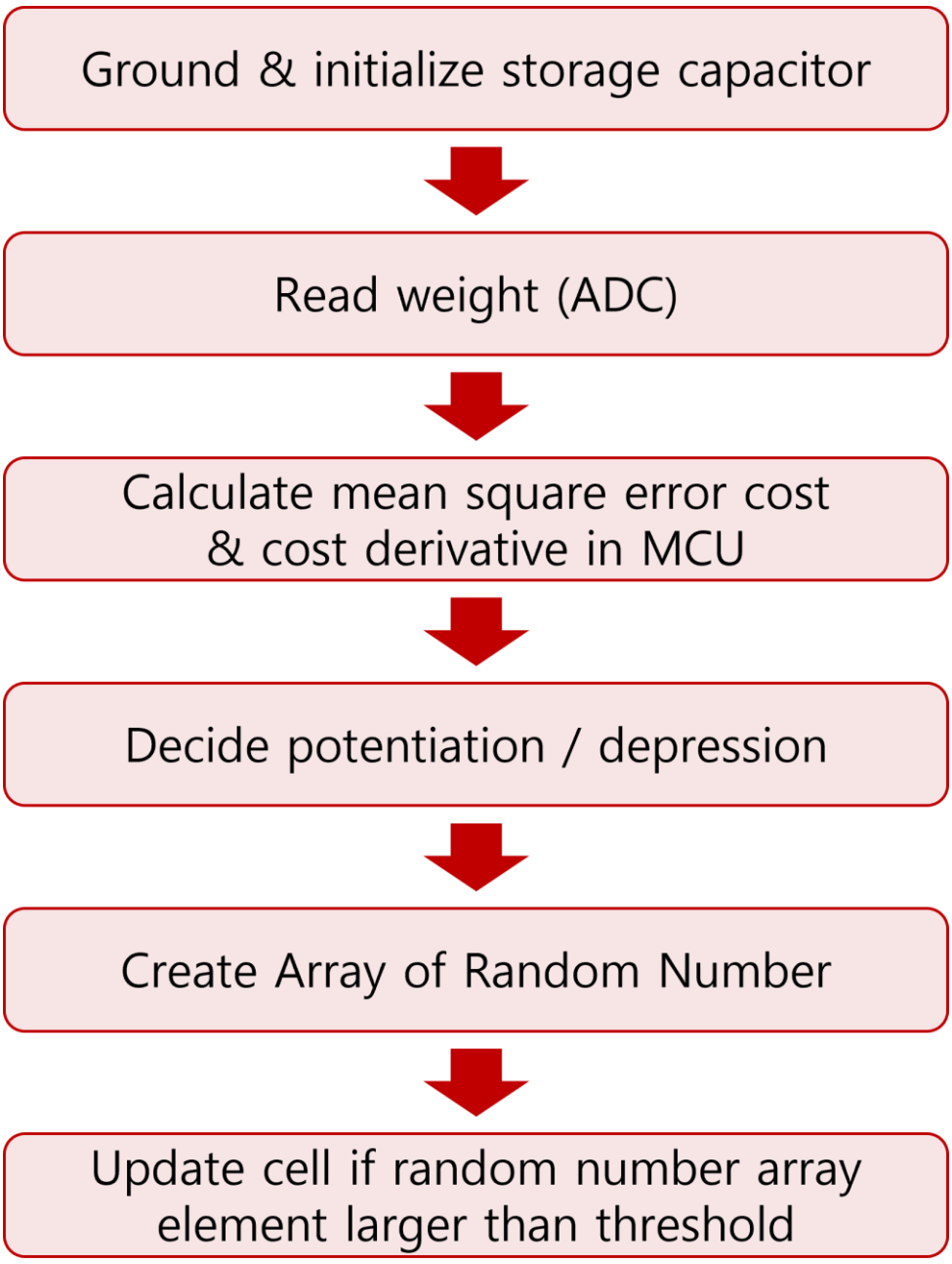

Figure 4.2.5.1 목표 가중치 도달 실험 방법.

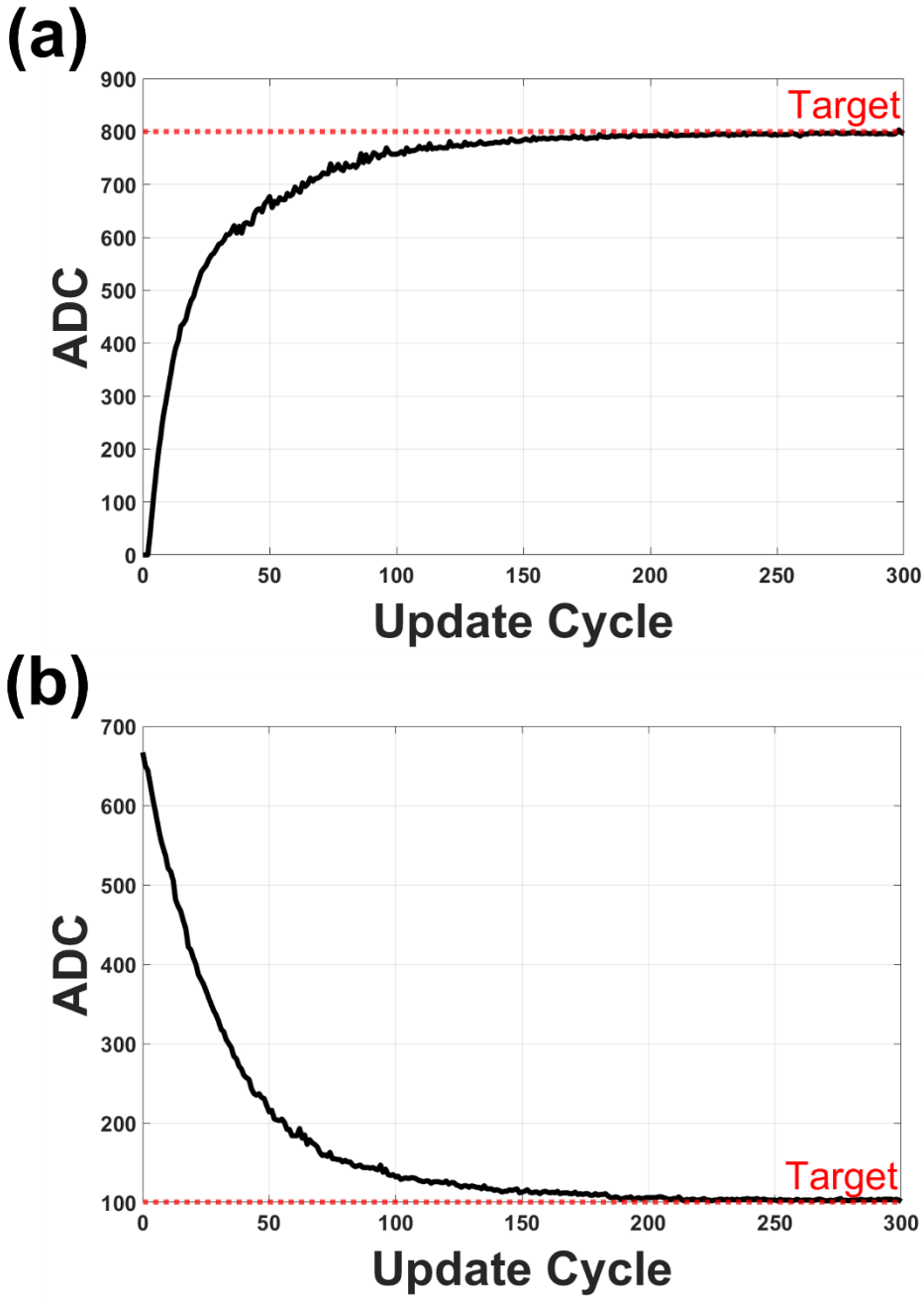

Figure 4.2.5.2 목표 가중치 도달 실험 결과. (a)는 potentiation을 사용한 가중치 갱신, (b)는 depression을 사용한 가중치 갱신이다.

4.3 가중치 retention

4.3.1 가중치 retention 실험 결과

거대한 신경망을 학습시킬 때나 장시간 추론을 하는 경우 시냅스 소자의 retention이 중요하다. Retention 실험은 시냅스 가중치를 특정 수준으로 update 한 뒤, 일정 간격으로 가중치 값을 읽어 ADC 값의 변화를 분석하였다. 가중치 감소는 간단한 RC 회로에서의 방전 모델인 exponential decay 모델을 사용할 수도 있지만, 실험 결과는 수식 (22)이 더 잘 표현하였기 때문에 수식 (22)를 분석에 이용하였다[46]. β는 1 이하의 상수로 exponential을 extended exponential로 만드는 역할을 하며[34] retention 성능은 시간 상수 τ의 크기로 평가할 수 있다.

$$\boldsymbol{ADC(t) = A \times exp\left(-\left(\frac{t}{\tau}\right)^{\beta}\right)} \quad \textbf{(22)}$$

Retention 실험의 결과는 Figure 4.3.1.1과 같다. 수식 (22) 모델이 누설 전류에 의한 가중치 변화를 잘 모델링 하는 것을 확인할 수 있었다. β 값 또한 [46]의 연구와 비슷한 값을 가졌다. 누설 전류 수준이 작은 a-IGZO TFT를 사용하였기 때문에 retention 시간 상수가 약 10,220분 수준으로 Si-CMOS 기반 3T1C에 비해 뛰어났다. ADC 출력값을 커패시터 전압으로 환산한 결과, 3T1C에서의 총 누설 전류 수준은 $<10^{-15}\mathrm{A}/\mu m$ 정도이다. Si MOSFET에 비해서는 훌륭한 수치이지만, 이전 보고된 값들에 비해서는 떨어지는데, 이는 3T1C에 사용되는 트랜지스터가 여러 개이며, 커패시터 등 추가로 누설 전류가 발생할 수 있는 통로가 있기 때문으로 추측하였다.

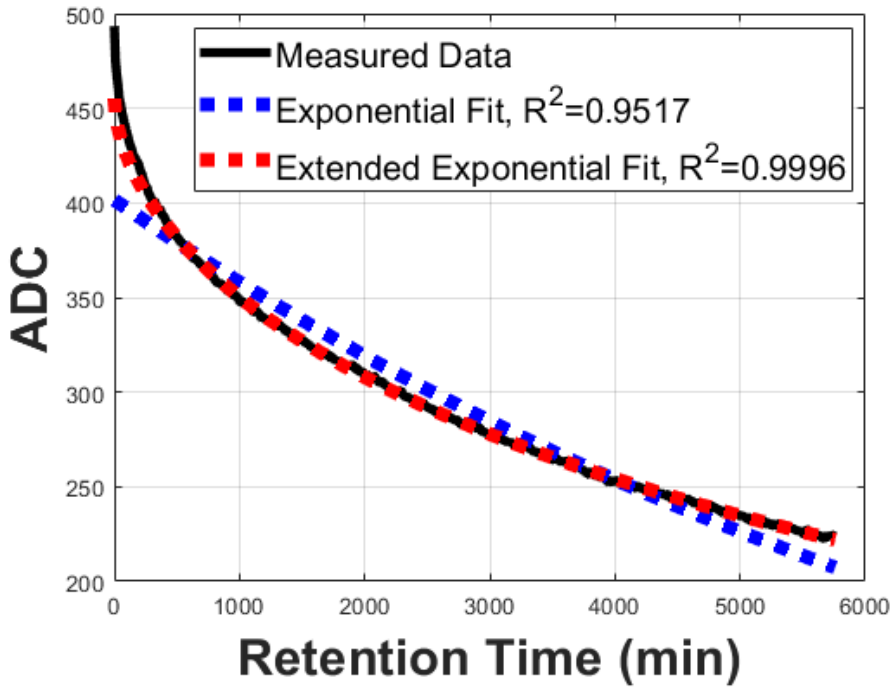

Figure 4.3.1.1 가중치 retention 실험 결과. 파란 선은 일반 exponential decay 모델, 빨간 선은 extended exponential decay 모델을 사용하여 평가한 결과이다.

4.3.2 5T1C와 3T1C의 retention 차이

5T1C(6T1C) 구조는 a-IGZO TFT는 NMOS만 존재한다는 단점을 보완하기 위해 제시된 시냅스 회로이다[51]. NMOS는 전류 주입에 부적합해 potentiation이 비선형적이게 되고, 가중치 갱신이 비대칭적으로 일어나기 때문에 potentiation 과정을 전류를 방전하는 구조로 변경하였다. Figure 4.3.2.1에서와 같이 potentiation에서는 N1, N2 트랜지스터를, depression에서는 N3, N4 트랜지스터를 사용하며 실제 전류가 흐르는 N2, N4 트랜지스터에는 V_{GS}가 일정하게 유지되도록 하였다. 이런 특징에 의해 3T1C에서의 depression처럼 potentiation도 선형적일 수 있으며, 대칭적인 가중치 갱신이 가능하기 때문에 높은 신경망 학습 정확도를 기대할 수 있다. 또한, N1과 N2 (N3와 N4)가 동시에 켜져야 갱신이 일어나기 때문에 별도의 AND gate 회로가 없어도 array에서 사용할 수 있다는 장점이 있다. 다만 커패시터가 floating 상태이기 때문에 커패시터 전압을 읽기 위해서는 N3 트랜지스터를 켜서 커패시터의 하단 전극을 $V_{DD}/2$로 정의하는 과정이 필요하다.

많은 장점을 가지는 5T1C 시냅스 소자이지만, 가중치 읽기 과정의 N3를 키는 과정에서 기생 커패시턴스에 의해 가중치가 저장된 커패시터의 전압이 변하는 문제가 있다. Figure 4.3.2.2에서 retention 실험 중 읽기 빈도에 따라 달라지는 retention 시간 상수를 확인할 수 있었다. a-IGZO TFT로 누설 전류가 흐르는 경우는 가중치 감소가 수식 (22)을 따라야 하지만 기생 커패시턴스에 의해 가중치가 손상되기 때문에 exponential decay 식을 따랐다. 신경망의 학습, 추론 과정에서 읽기 과정은 빈번하게 일어나는데, 과정마다 가중치가 변화하면 정확도가 떨어질 수밖에 없기 때문에 치명적인 문제이다. 또한, 읽기 과정뿐만 아니라 array 내에서 가중치 갱신을 하는 경우 선택되지 않은 소자들은 update

트랜지스터 중 하나만 켜지게 되는 half-select 상태에 있게 되는데, 이 과정 또한 동일하게 커패시터 전압에 간섭을 일으켜 문제가 될 수 있다. 반면 3T1C는 가중치가 저장되는 커패시터의 하단 전극이 항상 GND로 정의되기 때문에 읽기 과정에서 update 트랜지스터가 켜질 이유가 없으며, array 구동 또한 N1, N2 트랜지스터에 연결된 AND gate 바탕으로 이루어지기 때문에 5T1C에서와 같은 문제가 발생할 수 없다. 5T1C에서는 읽기 과정의 빈도와 retention 시간 상수 사이의 관계가 읽기 과정에 의해 병렬적인 누설전류 path가 있는 것처럼 모델링이 되지만, 읽기 빈도를 바꾸어 가며 retention 실험을 한 결과 3T1C에서는 시간 상수와의 경향성을 찾을 수 없었다.

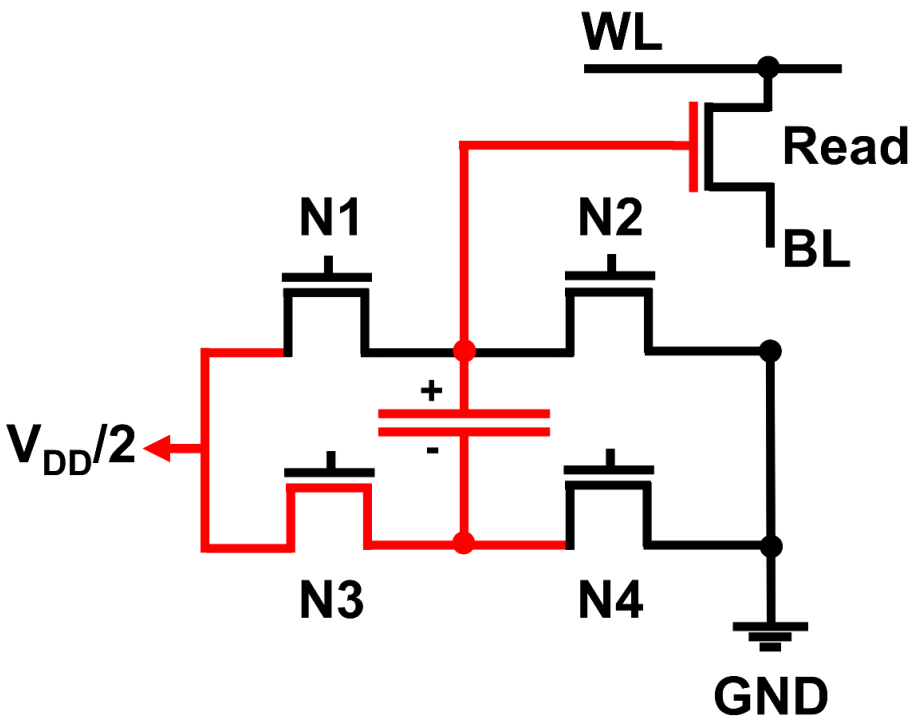

Figure 4.3.1.2 5T1C 회로도와 읽기 방식. Read 과정에는 빨간 선을 따라 read 트랜지스터의 gate에 $V_{DD}/2+V_{cap}$의 전압이 가해진다.

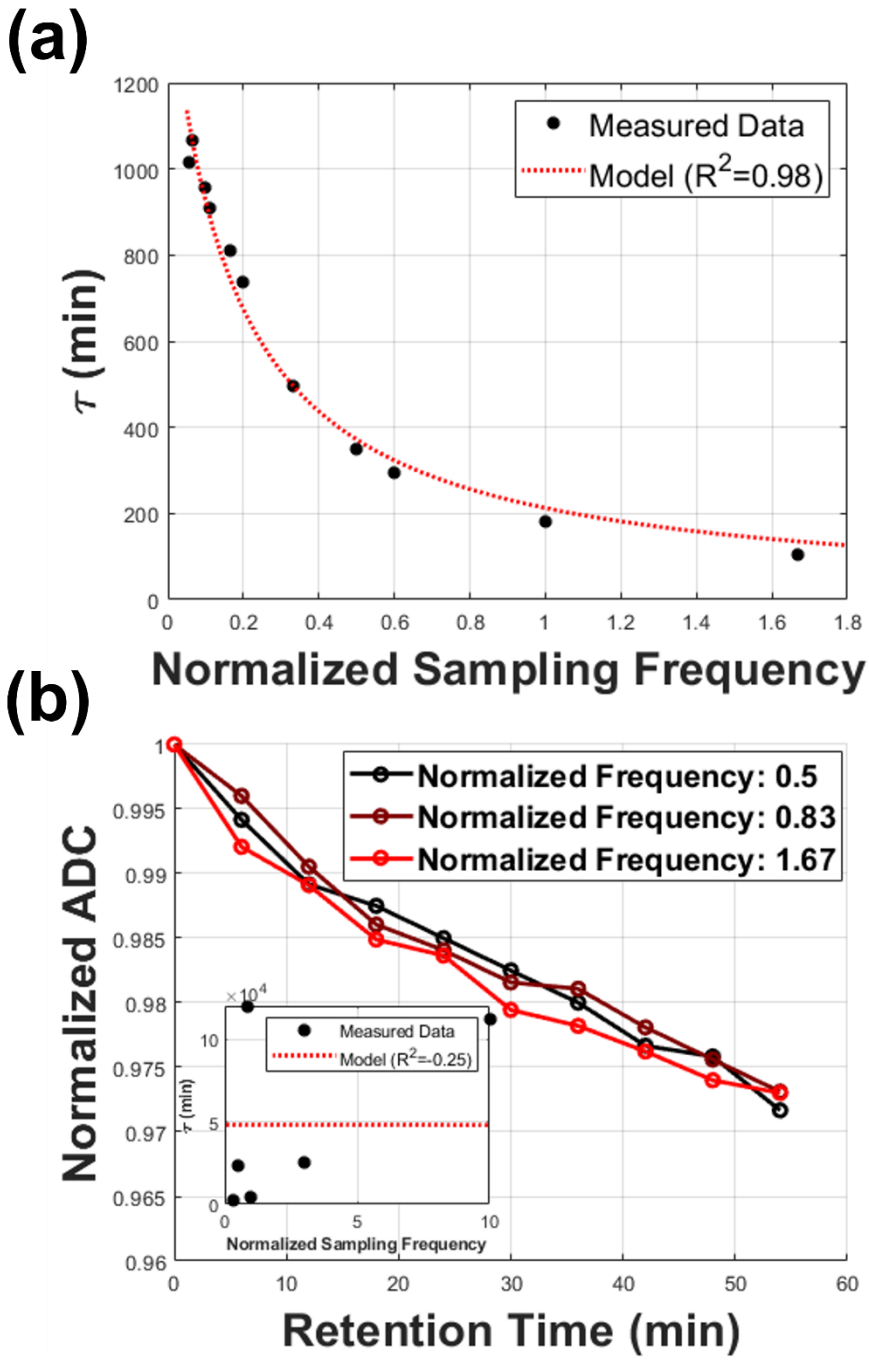

Figure 4.3.2.2 5T1C와 3T1C의 read 빈도에 따른 retention 성능 차이. (a)는 5T1C의 결과, (b)는 3T1C의 결과이다. (a)는 read 빈도에 일정한 반비례 관계를 보였다.

4.3.3 Retention 실험 전후 시냅스 성능 평가

비정질 산화물 기반 TFT는 bias stress에 큰 영향을 받는 것으로 알려져 있다. 3T1C 내 개별 트랜지스터들은 retention 실험 동안 bias stress를 받게 된다. N1과 N2는 항상 음의 전압으로 꺼진 상태이기 때문에 NBS가 가해지며, 커패시터에는 양의 전압이 저장되어 있기 때문에 read 트랜지스터는 PBS가 가해진다. Bias stress에 의해서 트랜지스터의 전기적 성질이 변화하고, 3T1C의 가중치 갱신에 변화가 생길 수 있기 때문에 retention 실험 전후로 각 트랜지스터의 transfer curve를 비교하였다.

Read 트랜지스터는 Figure 4.2.1.1와 동일한 방식으로 측정하였으며, N1과 N2 트랜지스터는 Figure 4.3.3.1과 같이 측정하였다. 각 트랜지스터에 직접적으로 전압을 가할 수 있는 방식이 없기 때문에 이처럼 간접적으로 실험하였다. 측정하지 않는 트랜지스터는 양의 gate 전압을 크게 걸어주고, 측정하고자 하는 트랜지스터의 gate 전압을 sweep해서 transfer curve를 측정하였다. Sweep하는 전압이 V_{th}와 비슷한 정도일 때는 직렬로 연결된 트랜지스터들에 흐르는 전류가 작기 때문에 V_{DD} node에 가하는 전압이 모두 측정하고자 하는 트랜지스터에 전달되지만, 더 큰 전류에서는 V_{DD} node에 가해준 전압이 두 트랜지스터에 나뉘어 가해지기 때문에 V_{th}는 측정이 가능하지만 on current는 정확한 분석이 어렵다.

실험 결과는 Figure 4.3.3.2와 같다. 총 세 가지 경우: 1) retention 실험 이전, 2) 빛을 쪼이며 retention 실험한 이후, 3) 빛을 쪼이지 않으며 retention 실험한 이후에 대해서 측정한 결과이다. NBS가 가해지는 N1, N2 트랜지스터의 경우 빛이 쪼여지는 경우는 음의 방향으로 V_{th}가 이동하는 반면 빛이 쪼이지 않는 경우 retention 실험 전후 차이가 미미하였

다. 이는 a-IGZO 내 hole carrier 농도가 작아 hole trapping이 일어나기 어렵지만, 빛이 쪼여지는 경우 electron-hole pair가 생성되어 V_{th}가 변한 것으로 설명하였다[29]. Update 트랜지스터의 V_{th}가 작아지면 동일 동작 전압에서 더 많은 전류가 흐를 수 있고, 세밀한 가중치 갱신이 불가할 수 있지만, 빛만 막는다면 문제를 방지할 수 있다. Read 트랜지스터의 PBS에 의한 전기적 성질 변화는 electron trapping에 일어나는 현상이기 때문에 빛의 유무와 무관하게 일어났다. 이전 연구에서 보고된 바와 동일하게[10] V_{th}가 증가하였으며 on current가 감소하였다. Read 트랜지스터의 열화는 동일한 가중치가 다르게 읽히게 하므로 추가 연구가 필요할 수 있다.

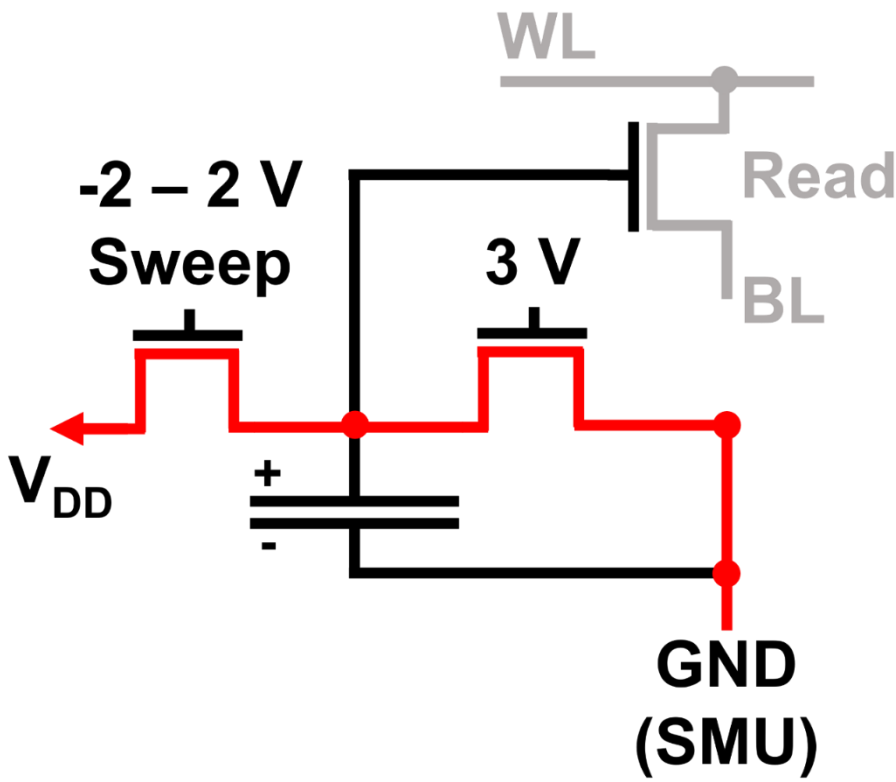

Figure 4.3.3.1 3T1C 소자 내 N1, N2 트랜지스터 transfer curve 측정 방법. N2를 측정할 때는 그림에서 N1과 N2에 가하는 전압을 바꾸어 가하였다.

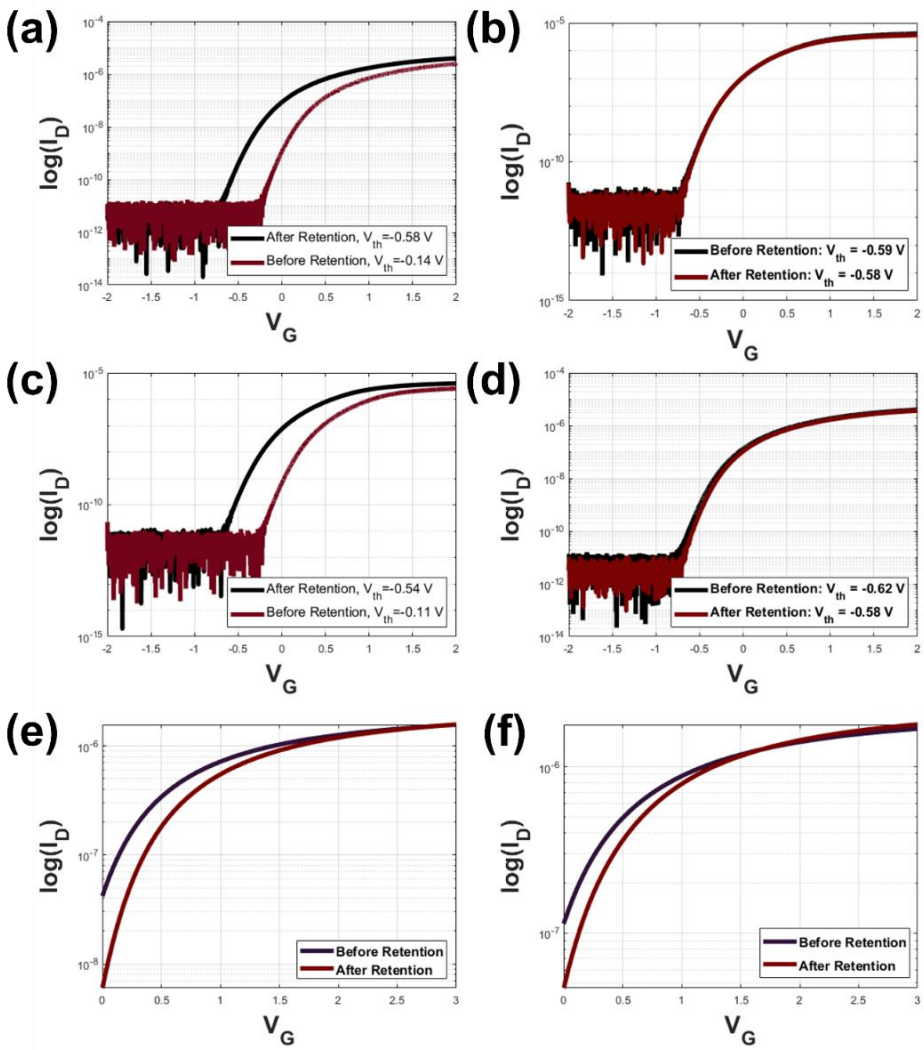

Figure 4.3.3.2 N1, N2 트랜지스터의 bias stress에 대한 안정성. (a)와 (b)는 N1, (c)와 (d)는 N2, (e)와 (f)는 read 트랜지스터의 결과이다. (a), (c), (e)는 빛이 있는 상태에서의 bias stress에 대한 변화, (b), (d), (f)는 어두운 상태에서의 bias stress에 대한 변화이다.

4.4 Cycling endurance

4.4.1 Endurance 실험 방법 및 결과

추론 array와 달리 on chip learning을 위해서는 많은 epoch 동안 시냅스 소자 갱신이 가능해야 하므로 좋은 내구성이 필요하다. Cycling endurance 실험은 4.2.2에서의 potentiation-depression 과정을 반복하였다. 실험에 사용하는 PCB 주변 회로의 한계상 read 과정이 ms 수준으로 오래 걸리기 때문에 endurance 실험에서는 매 potentiation-depression cycle마다 읽는 것이 아닌, 10의 거듭제곱과 같은 특정 시점에서만 읽고, 이외에는 update만 하였다. 이미 4.3.3에서 빛을 쪼이지 않았을 때 N1과 N2의 negative bias stress에 대한 안정성이 뛰어나다는 사실을 확인하였기 때문에 endurance 실험에서는 빛을 차단하였다.

측정 결과는 Figure 4.4.1.1과 같다. 5×10^7 cycle (cycle 당 potentiation 2,000번, depression 1,500번) 동안 endurance 실험을 하였으며, 시냅스 가중치 갱신 경향에 큰 변화 없이 잘 동작하는 것을 확인하였다. PRAM과 같이 재료를 물리적으로 변경하는 것이 아닌[56], 트랜지스터를 통해 전류를 흘리는 방식이기 때문에 DRAM과 같이 훌륭한 endurance 성능을 가지는 것을 확인할 수 있었다. Figure 4.4.1.2에서는 가중치 갱신의 선형성, 갱신량 등에 크지는 않지만 potentiation-depression cycle을 반복할수록 선형성, 대칭성이 개선되는 변화를 보였다. Cycling 상황에서 가장 취약한 트랜지스터 파악과 선형성, 대칭성 개선의 원인 분석을 위해 4.3.2와 동일한 방식으로 개별 트랜지스터들의 transfer curve를 측정하였다.

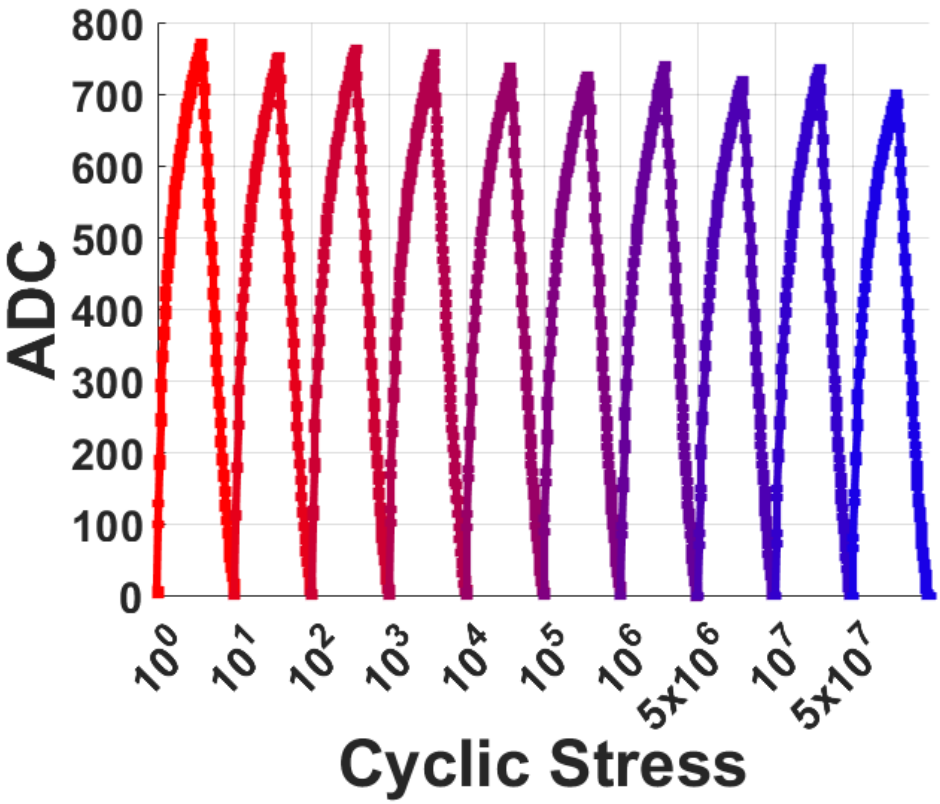

Figure 4.4.1.1 Cycling endurance 실험 결과. $5x10^7$ cycle의 potentiation-depression 동안 큰 변화 없이 동작하였다.

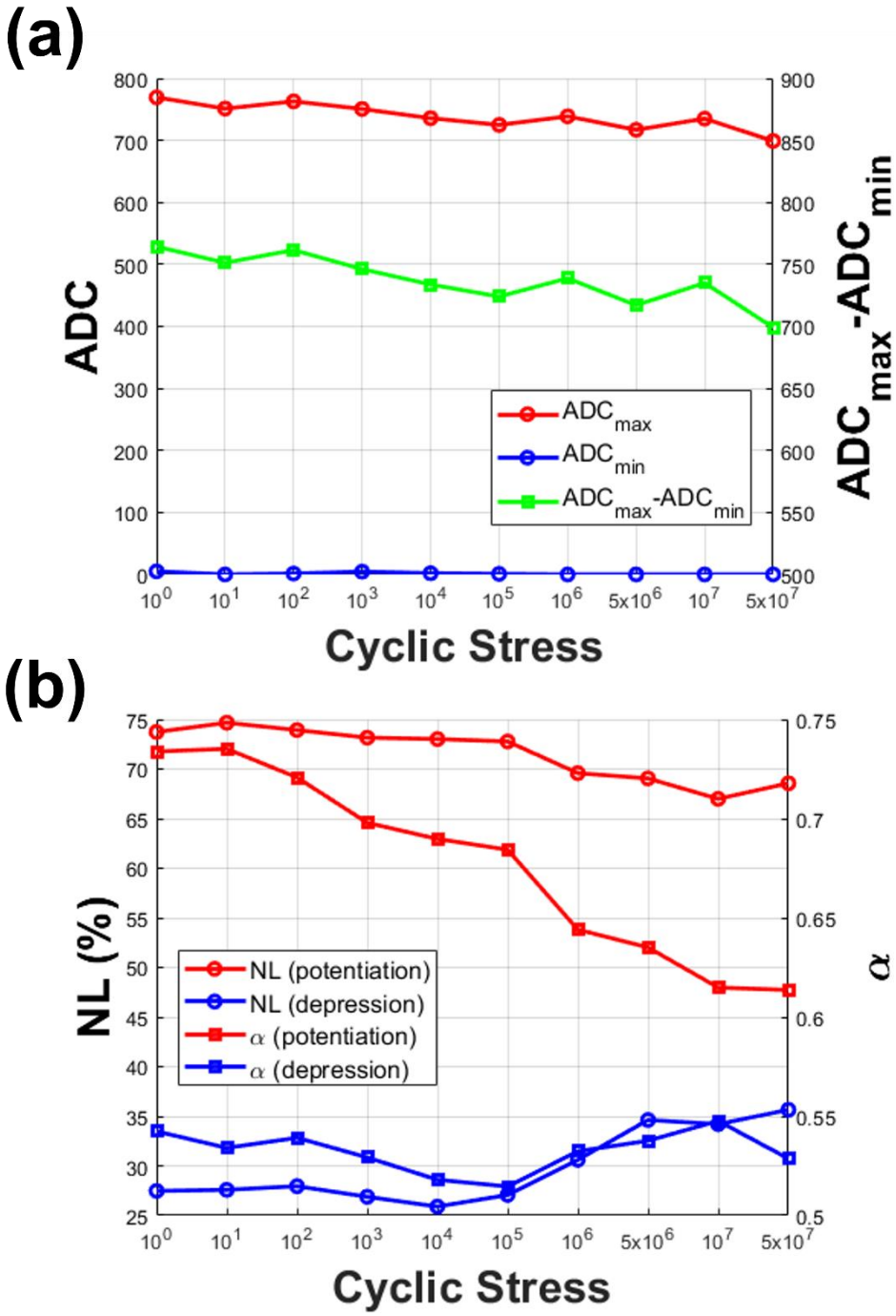

Figure 4.4.1.2 Cyclic stress가 시냅스에 미치는 영향. (a)는 ADC range, (b)는 가중치 갱신의 양과 선형성을 정리한 결과이다.

4.4.2 Cycling 부하가 시냅스 성능에 미치는 영향

일정한 bias stress가 가해지는 retention 실험과는 달리 cycling endurance 실험에서는 AC bias stress가 가해진다. N1, N2 트랜지스터는 on/off가 반복되며 read 트랜지스터에는 커패시터 충·방전에 의해 양의 AC bias stress가 가해진다. Figure 4.4.2.1은 Figure 4.4.1.2에서의 소자의 cycling endurance 실험 전후 개별 트랜지스터 측정 결과이다. AC stress가 가해졌지만, retention 실험 때와 동일하게 N1과 N2는 빛이 쪼여지지 않았기 때문에 V_{th} 변화가 미미하였다. 그러나 read 트랜지스터는 AC PBS의 영향에 의해 retention과 동일한 경향으로 V_{th}가 변화하였고, on-current의 열화 또한 확인하였다.

개별 트랜지스터 성능 측정을 통해 cyclic stress에 따른 가중치 갱신의 선형성, 대칭성 개선은 read 트랜지스터에 의한 효과라고 결론지을 수 있었다. Read 트랜지스터가 PBS에 노출되었을 때 전류 수준이 낮아지는 현상 이외에도 V_{th} 변화에 의해 낮은 커패시터 전압 영역이 subthreshold region에 가까워지며 transfer curve가 아래로 볼록해지는 변화가 일어나는데, 이것이 potentiation의 비선형성을 상쇄시킨 것으로 추측하였다. 낮은 커패시터 전압에서 N1 트랜지스터에 많은 전류가 흐르는 것이 potentiation의 비선형성에 큰 영향을 준다. Read 트랜지스터가 변하면 실제로 커패시터의 전압은 급격하게 갱신이 일어나지만 읽기 과정에서 상쇄되어 BL에 흐르는 전류는 기존보다 선형적으로 증가하는 것으로 결론지었다. 실제로 potentiation을 구간별로 나누어 비교하였을 때, cyclic stress에 따라 potentiation 초기에 일어나는 갱신량에 급격한 감소가 있음을 확인하였다. 따라서 적절한 read 트랜지스터 설계를 통해서도 더 선형적이고 대칭적인 가중치 갱신이 가능할 것이다.

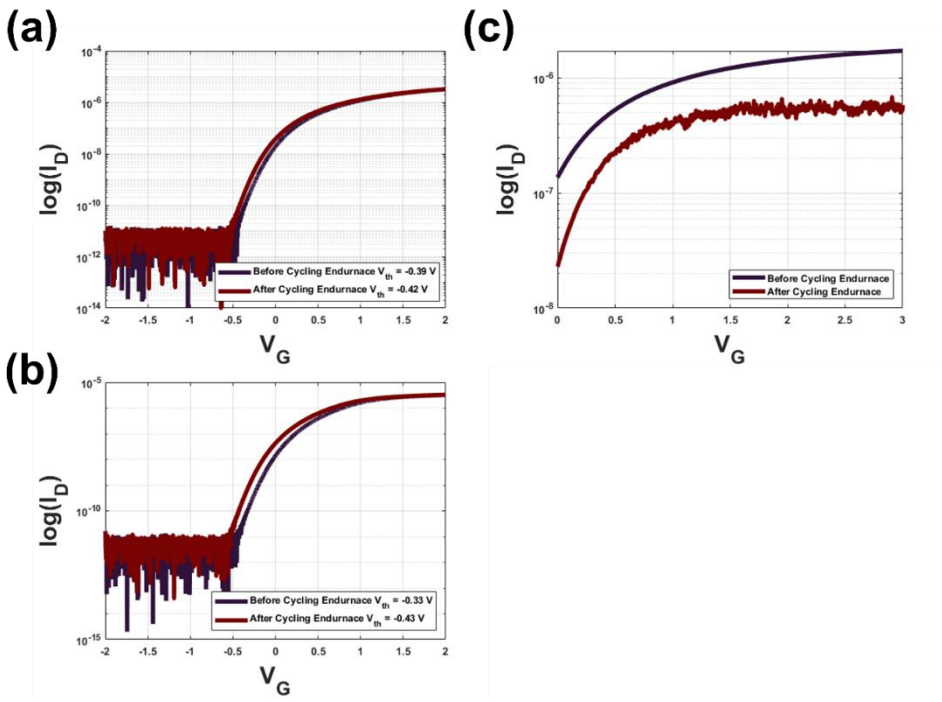

Figure 4.4.2.1 Cyclic stress 전후의 트랜지스터 성능 변화. (a)는 N1, (b)는 N2, (c)는 read 트랜지스터의 성능 변화를 나타낸 결과이다.

5. 결 론

본 연구에서는 현재 컴퓨팅이 심층 신경망 학습 시에 가지는 전력 소모, 연산 속도 측면에서의 한계를 극복할 수 있는 crossbar 형 심층 신경망 가속기에 적합한 가중치 소자로 a-IGZO TFT와 커패시터를 이용한 3T1C 시냅스를 제시하였다. PRAM, RRAM 등의 차세대 비휘발성 메모리에 비해 선형적이고 대칭적인 가중치 갱신을 제공하는 동시에, Si CMOS 기반의 시냅스에 비해 상대적으로 휘발성이 적은 장점을 가지기 때문에 대규모 신경망의 on-chip learning에 적합하다.

실험을 통해 고속의 가중치 갱신이 가능함을 확인하였으며, 가중치 retention과 endurance 성능 또한 뛰어남을 확인하였다. N-type만 존재하는 a-IGZO TFT의 특성상 potentiation 과정이 비선형적이었지만 동작 전압 조건의 최적화를 통해 이를 최적화할 수 있었으며, read 트랜지스터의 설계로 비선형성과 비대칭성을 상쇄할 수 있는 가능성을 보였다. Retention 또한 시간 상수가 10,000분 이상으로 뛰어났으며, 추후 공정 최적화로 성능 향상을 기대할 수 있었다. Cycling endurance도 트랜지스터 기반 시냅스 소자이기 때문에 훌륭하였다.

CMOS BEOL 공정에 호환되는 공정이기 때문에 Si-CMOS 공정과 수직 적층이 가능하다는 장점이 있지만, 3T1C는 선택 소자로 면적이 큰 AND gate를 사용해야 한다는 단점이 있다. 면적 감소를 위해 빠른 동작 속도를 포기하고 potentiation과 depression에 동일한 update 트랜지스터를 사용하는 2T1C 구조가 도움이 될 수 있다. 또는 AND gate 대신 dual gate update 트랜지스터를 통해 update 트랜지스터에서 선택 과정까지 일어날 수 있게 제작할 수 있다. 시냅스 동작에 따라 read 트랜지스터가 받는 bias stress에 대한 안정성도 반드시 필요하다. 이는 공정 최적화를 통한 막질 개선 등을 통해 이룰 수 있을 것으로 기대된다. 마

지막으로 potentiation과 depression의 비선형성, 비대칭성의 경우 zero-shifting[28]과 같은 3T1C에 최적화된 학습 알고리즘을 이용하면 학습 정화도 열화를 막을 수 있을 것으로 예상된다.

요약하자면, 본 연구에서는 트랜지스터를 사용하지만 속도, 전력 소모, 데이터 유지 능력 측면에서 합리적인 성능을 가지는 시냅스 소자를 제안하였다. 공정 최적화와 적절한 알고리즘 개발을 통해 3T1C의 on-chip learning array로의 성능을 향상시킨다면 현재 폰 노이만 기반 컴퓨팅의 한계를 극복하며, on-device AI 등 저전력, 고성능 AI 프로세서가 필요한 분야에서 사용될 수 있을 것으로 기대된다.

참고 문헌

[1] S. Ambrogio et al., Equivalent-accuracy accelerated neural-network training using analogue memory. Nature. 558, 60–67 (2018).

[2] H. Baba et al., in Technical Digest - International Electron Devices Meeting, IEDM (Institute of Electrical and Electronics Engineers Inc., 2021), vols. 2021-December, pp. 21.2.1-21.2.4.

[3] A. Belmonte et al., in Technical Digest - International Electron Devices Meeting, IEDM (Institute of Electrical and Electronics Engineers Inc., 2020), vols. 2020-December, pp. 28.2.1-28.2.4.

[4] S. Brivio, D. R. B. Ly, E. Vianello, S. Spiga, Non-linear Memristive Synaptic Dynamics for Efficient Unsupervised Learning in Spiking Neural Networks. Frontiers in Neuroscience. 15 (2021), doi:10.3389/fnins.2021.580909.

[5] G. W. Burr et al., Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element. IEEE Transactions on Electron Devices. 62, 3498–3507 (2015).

[6] C. C. Chang et al., Mitigating Asymmetric Nonlinear Weight Update Effects in Hardware Neural Network Based on Analog Resistive Synapse. IEEE Journal on Emerging and Selected Topics in Circuits and Systems. 8, 116–124 (2018).

[7] P. Y. Chen, X. Peng, S. Yu, in Technical Digest - International

Electron Devices Meeting, IEDM (Institute of Electrical and Electronics Engineers Inc., 2018), pp. 6.1.1–6.1.4.

[8] Y. Chen et al., An 18.6–μm–Pitch Gate Driver Using a–IGZO TFTs for Ultrahigh–Definition AR/VR Displays. IEEE Transactions on Electron Devices. 67, 4929–4933 (2020).

[9] Y. T. Chien et al., Performance Enhancement of InGaZnO Top–Gate Thin Film Transistor with Low–Temperature High–Pressure Fluorine Treatment. IEEE Electron Device Letters. 42, 1611–1614 (2021).

[10] M. H. Cho et al., Comparative Study on Performance of IGZO Transistors With Sputtered and Atomic Layer Deposited Channel Layer. IEEE Transactions on Electron Devices. 66, 1783–1788 (2019).

[11] S. Choi et al., Positive Bias Stress Instability of InGaZnO TFTs with Self–Aligned Top–Gate Structure in the Threshold–Voltage Compensated Pixel. IEEE Electron Device Letters. 41, 50–53 (2020).

[12] X. Duan et al., Novel Vertical Channel–All–Around (CAA) In–Ga–Zn–O FET for 2T0C–DRAM With High Density Beyond 4F2by Monolithic Stacking. IEEE Transactions on Electron Devices. 69, 2196–2202 (2022).

[13] R. Garcia et al., A compact drain current model for thin–film transistor under bias stress condition. IEEE Transactions on Electron Devices. 65, 1803–1809 (2018).

[14] T. Gokmen, W. Haensch, Algorithm for Training Neural Networks on Resistive Device Arrays. Frontiers in

Neuroscience. 14 (2020), doi:10.3389/fnins.2020.00103.

[15] T. Gokmen, Y. Vlasov, Acceleration of deep neural network training with resistive cross-point devices: Design considerations. Frontiers in Neuroscience. 10 (2016), doi:10.3389/fnins.2016.00333.

[16] K. He, X. Zhang, S. Ren, J. Sun, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (IEEE Computer Society, 2016), vols. 2016-December, pp. 770–778.

[17] T. Hirofuchi, R. Takano, A prompt report on the performance of intel optane DC persistent memory module. IEICE Transactions on Information and Systems. E103D, 1168–1172 (2020).

[18] Y. Hu, Y. Wang, T. Lei, F. Wang, M. Wong, Neuromorphic Implementation of Logic Functions Based on Parallel Dual-Gate Thin-Film Transistors. IEEE Electron Device Letters. 43, 741–744 (2022).

[19] Y. Hu, T. Lei, Y. Wang, F. Wang, M. Wong, An Artificial Neural Network Implemented Using Parallel Dual-Gate Thin-Film Transistors. IEEE Transactions on Electron Devices (2022), doi:10.1109/TED.2022.3201836.

[20] S. Huang, X. Sun, X. Peng, H. Jiang, S. Yu, in Proceedings of the 2020 Design, Automation and Test in Europe Conference and Exhibition, DATE 2020 (Institute of Electrical and Electronics Engineers Inc., 2020), pp. 1025–1030.

[21] R. Islam et al., Device and materials requirements for neuromorphic computing. Journal of Physics D: Applied Physics.

52 (2019), , doi:10.1088/1361-6463/aaf784.

[22] M. Jerry et al., in Technical Digest - International Electron Devices Meeting, IEDM (Institute of Electrical and Electronics Engineers Inc., 2018), pp. 6.2.1-6.2.4.

[23] S. Jung et al., A crossbar array of magnetoresistive memory devices for in-memory computing. Nature. 601, 211–216 (2022).

[24] T. Kamiya, K. Nomura, H. Hosono, Present status of amorphous In-Ga-Zn-O thin-film transistors. Science and Technology of Advanced Materials. 11 (2010), doi:10.1088/1468-6996/11/4/044305.

[25] F. Kiani, J. Yin, Z. Wang, J. Joshua Yang, Q. Xia, A fully hardware-based memristive multilayer neural network. Science Advances. 7 (2021), doi:10.1126/sciadv.abj4801.

[26] S. Kim, T. Gokmen, H. M. Lee, W. E. Haensch, in Midwest Symposium on Circuits and Systems (Institute of Electrical and Electronics Engineers Inc., 2017), vols. 2017-August, pp. 422–425.

[27] W. Kim et al., in Digest of Technical Papers - Symposium on VLSI Technology (Institute of Electrical and Electronics Engineers Inc., 2019), vols. 2019-June, pp. T66–T67.

[28] H. Kim et al., Zero-shifting technique for deep neural network training on resistive cross-point arrays. arXiv preprint arXiv:1907.10228.

[29] C. W. Kuo et al., On the Optimization of Performance and Reliability in a-InGaZnO Thin-Film Transistors by Versatile

Light Shielding Design. IEEE Transactions on Electron Devices. 68, 1654–1658 (2021).

[30] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. Proceedings of the IEEE. 86, 2278–2323 (1998).

[31] C. Lee, K. Noh, W. Ji, T. Gokmen, S. Kim, Impact of Asymmetric Weight Update on Neural Network Training With Tiki-Taka Algorithm. Frontiers in Neuroscience. 15 (2022), doi:10.3389/fnins.2021.767953.

[32] G. H. Lee et al., Effect of weight overlap region on neuromorphic system with memristive synaptic devices. Chaos, Solitons and Fractals. 157 (2022), doi:10.1016/j.chaos.2022.111999.

[33] Y. Li et al., in Digest of Technical Papers - Symposium on VLSI Technology (Institute of Electrical and Electronics Engineers Inc., 2018), vols. 2018-June, pp. 25–26.

[34] A. Lukichev, Physical meaning of the stretched exponential Kohlrausch function. Physics Letters, Section A: General, Atomic and Solid State Physics. 383, 2983–2987 (2019).

[35] Y. Luo, S. Yu, Accelerating Deep Neural Network In-Situ Training with Non-Volatile and Volatile Memory Based Hybrid Precision Synapses. IEEE Transactions on Computers. 69, 1113–1127 (2020).

[36] N. Lv et al., Suppression of the Short-Channel Effect in Dehydrogenated Elevated-Metal Metal- Oxide (EMMO) Thin-Film Transistors. IEEE Transactions on Electron Devices. 67, 3001–3004 (2020).

[37] A. Mehonic and A. J. Kenyon., Brain-inspired computing needs a master plan. Nature. 604.7905, 255-260 (2022).

[38] M. Oota et al., in Technical Digest - International Electron Devices Meeting, IEDM (Institute of Electrical and Electronics Engineers Inc., 2019), vols. 2019-December.

[39] K. Roy, S. Mukhopadhyay, H. Mahmoodi-Meimand, Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits. Proceedings of the IEEE. 91, 305–327 (2003).

[40] D. Saito et al., IGZO-Based Compute Cell for Analog In-Memory Computing - DTCO Analysis to Enable Ultralow-Power AI at Edge. IEEE Transactions on Electron Devices. 67, 4616–4620 (2020).

[41] Y. Sekine et al., (Invited) Success in Measurement the Lowest Off-state Current of Transistor in the World. ECS Transactions. 37, 77–88 (2019).

[42] M. Shoeybi et al., Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. (2019), arXiv:1909.08053v4.

[43] X. Sun et al., PCM-Based Analog Compute-In-Memory: Impact of Device Non-Idealities on Inference Accuracy. IEEE Transactions on Electron Devices. 68, 5585–5591 (2021).

[44] M. Suri et al., in Technical Digest - International Electron Devices Meeting, IEDM (2011).

[45] M. Suri et al., Physical aspects of low power synapses based on phase change memory devices. Journal of Applied

Physics (2012), vol. 112.

[46] M. Tsubuku et al., Analysis for Extremely Low Off-State Current in CAAC-IGZO FETs. ECS Transactions. 67, 17–22 (2015).

[47] H. Wang, M. Wang, D. Zhang, Q. Shan, Degradation of a-InGaZnO TFTs under Synchronized Gate and Drain Voltage Pulses. IEEE Transactions on Electron Devices. 65, 995–1001 (2018).

[48] P. Wang, S. Yu, Ferroelectric devices and circuits for neuro-inspired computing. MRS Communications. 10, 538–548 (2020).

[49] S. Wang et al., Resilience of Fluorinated Indium-Gallium-Zinc Oxide Thin-Film Transistor against Hydrogen-Induced Degradation. IEEE Electron Device Letters. 41, 729–732 (2020).

[50] Y. Wang et al., Amorphous-InGaZnO Thin-Film Transistors Operating beyond 1 GHz Achieved by Optimizing the Channel and Gate Dimensions. IEEE Transactions on Electron Devices. 65, 1377–1382 (2018).

[51] J. Won et al., Device-algorithm co-optimization for an on-chip trainable capacitor-based synaptic device with IGZO access transistor and retention-centric Tiki-Taka algorithm [Unpublished manuscript]. Department of Materials Science and Engineering, Seoul National University (2023)

[52] H. S. P. Wong et al., in Proceedings of the IEEE (Institute of Electrical and Electronics Engineers Inc., 2012), vol. 100, pp. 1951–1970.

[53] W. Wu et al., Improving Analog Switching in HfOx-Based Resistive Memory with a Thermal Enhanced Layer. IEEE Electron Device Letters. 38, 1019–1022 (2017).

[54] Q. Xia, J. J. Yang, Memristive crossbar arrays for brain-inspired computing. Nature Materials. 18 (2019), pp. 309–323.

[55] Y. Xiang et al., Impacts of State Instability and Retention Failure of Filamentary Analog RRAM on the Performance of Deep Neural Network. IEEE Transactions on Electron Devices. 66, 4517–4522 (2019).

[56] Y. Xie et al., Self-Healing of a Confined Phase Change Memory Device with a Metallic Surfactant Layer. Advanced Materials. 30 (2018), doi:10.1002/adma.201705587.

[57] X. Xu et al., Scaling for edge inference of deep neural networks. Nature Electronics. 1, 216–222 (2018).

[58] C. X. Xue et al., in Digest of Technical Papers - IEEE International Solid-State Circuits Conference (Institute of Electrical and Electronics Engineers Inc., 2020), vols. 2020-February, pp. 244–246.

[59] S. Yu, Neuro-Inspired Computing with Emerging Nonvolatile Memorys. Proceedings of the IEEE. 106, 260–285 (2018).

Abstract

3T1C Charge-Storage Type Synapse Using InGaZnO Thin-Film-Transistors for Deep Neural Network Acceleration

Minseung Kang
Materials Science and Engineering
The Graduate School
Seoul National University

Artificial intelligence (AI) has achieved remarkable progress in various fields such as image recognition and natural language processing. However, the complexity of emerging AI algorithms results in high power consumption and long training periods in conventional von-Neumann computing. While accelerating matrix-vector multiplication in crossbar arrays of nonvolatile memories has been suggested as a remedy for von-Neumann bottleneck issue, inherent nonidealities of nonvolatile memories, especially nonlinear and asymmetric weight updates, prevent its application. Si-CMOS and capacitor-based synapse may have linear, symmetric weight updates, but is volatile in nature. Amorphous InGaZnO thin film transistor 3T1C synapse circuit as a training accelerator is suggested in this work. Nonlinearity and asymmetry can be expected due to only n-type transistors existing for a-IGZO TFTs, but no issues were

found in weight updating in simulated learning schemes. Mitigation methods have also been suggested founded on weight update models and experiments. Outstanding retention performance of more than 10,000 min was measured as expected of low off current of a-IGZO TFTs. Synaptic operation did not experience significant changes after 5×10^7 weight update cycles. Combined with optimized learning algorithms, a-IGZO 3T1C synapse can be a candidate for low power, high-speed AI accelerator.

Keywords : InGaZnO TFTs, DNN accelerator, charge-storage type synapse, low off-current, linear and symmetric weight update
Student Number : 2021-28307