인공지능 반도체 메모리 기술 동향

Trends in Artificial Intelligence Semiconductor Memory Technology

황규동 (K.D. Hwang, kyuba01@etri.re.kr) 지능형엣지반도체연구실 선임연구원
오광일 (K.I. Oh, kioh@etri.re.kr) 지능형엣지반도체연구실 책임연구원
이재진 (J.J. Lee, ceicarus@etri.re.kr) 지능형엣지반도체연구실 책임연구원/실장
구본태 (B.T. Koo, koobt@etri.re.kr) 지능형반도체연구본부 책임연구원/본부장

ABSTRACT

Memory can refer to a storage device that collects data, and it has evolved to increase the reading/writing speed and reduce the power consumption. As large amounts of data are processed by artificial intelligence services, the memory data capacity requires expansion. Dynamic random-access memory (DRAM) is the most widely used type of memory. In particular, graphics double date rate and high-bandwidth memory allow to quickly transfer large amounts of data and are used as memory solutions for artificial intelligence semiconductors. We analyze development trends in DRAM from the perspectives of processing speed and power consumption. We summarize the characteristics required for next-generation memory by comparing DRAM and other types of memory implementations. Moreover, we examine the shortcomings of DRAM and infer a next-generation memory for their compensation. We also describe the operating principles of spin-torque transfer magnetic random access memory, which may replace DRAM in next-generation devices, and explain its characteristics and advantages.

KEYWORDS DRAM, STT-MRAM, 메모리, 반도체, 인공지능

Ⅰ. 서론

최근 많은 AI(Artificial Intelligence) 서비스가 등장하여 뛰어난 성능으로 사람들을 놀라게 하고 있다. GPT(@Open-AI)를 필두로 해서, LLaMA(@Meta), Claude(@Anthropic), Gemini(@Google) 등과 같은 LLM (Large Language Model)들이 치열하게 경쟁하고 있다. 이와 같은 인공지능 서비스들이 지속적으로 발전하려면, 소프트웨어와 함께 하드웨어의 발전도 필수적이다. 실제로 인공지능 반도체의 시장 규모는 수백억 달러 정도로 해마다 급격하게 증가하고 있다.

많은 Data를 다루는 것이 특징인 인공지능 반도

* DOI: https://doi.org/10.22648/ETRI.2024.J.390503
* This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korea government[24ZS1230, Memory-computation convergence neuromorphic computing technology].

체는 Data를 처리하는 비메모리와 Data를 저장하는 메모리로 구분된다.

비메모리 중 인공지능에 많이 사용되는 반도체는 GPU(Graphic Processing Unit)와 NPU(Neural Processing Unit)가 있다.

GPU는 1990년대 후반에 CPU(Central Processing Unit)만으로 처리하기 어려운 그래픽 작업을 보조하기 위해 개발되었다. 개발 초기에는 게임과 CAD(Computer-Aided Design) 등에 사용되었다. 2000년대 초반에 GPU를 그래픽 처리뿐만 아니라 범용적인 계산에 사용하는 GPGPU(General-Purpose GPU) 기술이 등장했다. 이 기반으로 2007년 NVIDIA의 CUDA(Compute Unified Device Architecture)가 개발되면서, GPU가 본격적으로 인공지능 반도체로 사용되기 시작하였다. 이후, GPU는 인공지능 학습을 위해 없어서는 안 되는 중요한 도구가 되었으며, 품귀현상이 벌어지기도 하였다. 그 결과, 2024년 현재 NVIDIA는 인공지능 반도체 시장의 80% 이상을 점유하고 있다.

하지만, GPU의 가장 큰 단점은 전력 사용량이 너무 크고, 가격이 비싸다는 것이다. 모바일 기기나 자율주행 차량에서 On-device AI 반도체를 사용하기 위해서는 전력 사용량을 줄이고, 가격을 낮추는 것이 필요하다. 이를 위해 2015년부터 NPU가 개발되기 시작하였다. 기존 GPU에서 CUDA를 통해 소프트웨어로 처리했던 인공지능 연산을, NPU는 하드웨어로 고정해 설계한다. 특정 Neural Network에 특화되어 설계되는 NPU는 다양한 용도로 활용할 수 없다는 단점이 있지만, GPU 대비 저전력 특성을 갖는다. 2024년 현재 NPU는 Qualcomm, Apple, Tesla, Google, NVIDIA, Samsung 등의 글로벌 기업에서 모두 자체적으로 제작하고 있으며, On-device AI 반도체로 널리 사용되고 있다.

이처럼 인공지능 반도체의 한 축인 비메모리는 인공지능의 발전과 함께 꾸준히 발전해 왔다. 그렇다면, 인공지능 반도체의 다른 축인 메모리는 어떻게 발전해 왔을까?

본고에서는 메모리에 관해 좀 더 집중적으로 탐구해 보고자 한다. 메모리가 어떻게 발전해 왔으며, 이 중 인공지능 반도체에 사용되는 메모리들의 특징은 무엇인지, 미래에 이를 대체할 가능성이 있는 새로운 메모리는 무엇이 있는지에 대해 알아보고자 한다.

II. 메모리의 발전 흐름

1. 메모리 속도의 발전 방향

현재 전 세계적으로 가장 많이 사용되는 메모리는 DRAM(Dynamic Random Access Memory)이다. 주변의 모든 Computer 및 전자 장비에 DRAM이 들어가지 않는 것을 찾기 어렵다. DRAM이 이렇게 널리 사용되는 이유는 빠른 속도와 싼 가격에 있다. DRAM 속도의 중요성에 맞춰, DRAM의 pin당 Data-rate는 개발 초기부터 지금까지 꾸준히 증가해 오고 있다.

DRAM은 1968년에 Robert H. Dennard에 의해 최초로 개발되었다[1]. 처음에는 SDR(Single Data-Rate)이 사용되었지만, 2000년부터 DDR(Double Data-Rate)이 상용화되었다[2]. 그림 1은 DDR의 등장 이후부터 현재까지의 DDR DRAM의 속도 발전을 시간순으로 나타낸다. x축은 연도이고, y축은 DRAM의 하나의 pin당 Data-rate를 나타낸다. 시간의 흐름에 따른 전체 그래프의 방향성은 “속도가 증가한다”이다.

DRAM은 사용처와 기능에 따라 4종류로 분류한다. DDR은 데스크탑과 서버에 사용되며, 메인보드의 Channel Loss가 크기 때문에, Channel Loss를 보상하는 것을 목표로 한다. LPDDR(Low-Power

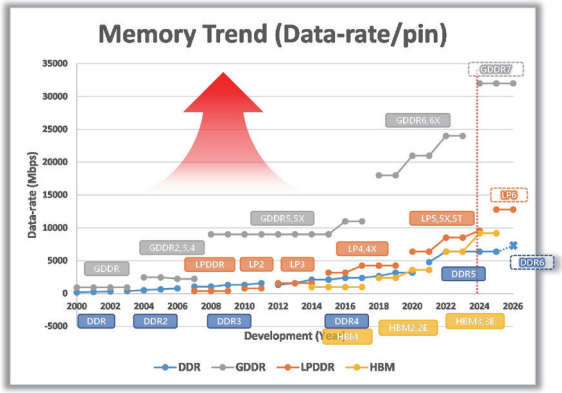


그림 1 DRAM 속도의 발전 방향

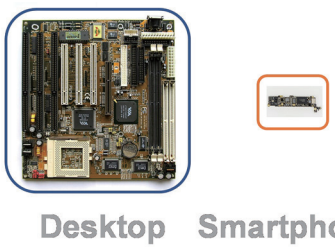
DDR)은 모바일 제품들에 사용되며, 전력 소모를 최소화하는 것이 목표이다. GDDR(Graphic DDR)은 Graphic 제품들에 사용되며, 가장 높은 속도를 목표로 한다. HBM(High-Bandwidth Memory)은 인공지능 Computing을 위해 사용되며, Total Bandwidth를 최대로 키우는 것이 목표이다.

DDR과 GDDR은 2000년에 상용화가 시작되었다. DDR은 CPU를 보조하면서, 합리적인 성능을 얻고, 합리적인 전력을 소모하도록 설계되었다[3]. 반면, GDDR은 GPU를 보조하면서, 최고의 성능을 얻기 위해 설계되었고, 제품의 세대가 바뀔 때마다 속도 증가가 DDR 대비 매우 큰 것이 특징이다[4].

2007년 새로운 메모리가 새로운 제품과 함께 등장하였다. 바로 Apple사의 아이폰이 2007년 LPDDR 제품을 탑재하고 출시되었다. 이후 스마트폰 시장과 함께 LPDDR 시장도 빠르게 성장하였다. 현재는 스마트폰뿐만 아니라 노트북, 태블릿 등의 많은 휴대용 기기에서 LPDDR이 사용되고 있다. 저전력의 장점 덕분에 최근에는 데스크탑이나 서버와 같은 다른 Application에서도 사용하려는 움직임이 있다.

그림 1의 Speed Trend에서 LPDDR 발전 과정의 독특한 부분을 확인할 수 있다. 2007년 첫 등장 시에 Low-power를 표방하고 등장하였기 때문에, 초기의 LPDDR은 DDR보다 전력을 감소시키는 방향으로 개발되었다. 전력을 절약하다 보니 초기에는 LPDDR이 DDR보다 속도가 낮게 개발되었다. 하지만, 2015년 LPDDR4가 개발되면서부터 LPDDR은 저전력이면서, 속도에서도 DDR을 넘어서게 된다[5]. 현재까지 그 추세는 계속 이어져 오고 있으며, DDR과 LPDDR의 성능 차이는 점점 벌어지고 있다.

그렇다면 LPDDR은 왜 DDR보다 전력을 적게 쓰면서도, 더 높은 속도를 구현할 수 있는 것일까? 그 이유는 LPDDR의 Channel 환경이 DDR보다 Signal 전송에 유리하기 때문이다. LPDDR Channel의 첫 번째 장점은 크기이다. 그림 2는 일반적인 데스크탑과 스마트폰의 메인보드 크기를 상대적으로 보여준다. 데스크탑에서 CPU와 Memory 사이의 거리가 스마트폰에서 AP와 Memory 사이의 거리보다 훨씬 길다는 것을 쉽게 추측할 수 있다. 두 번째 장점은 Channel 환경의 불변성이다. 모바일 제품들은 내구성을 증가시키기 위해 폐쇄적으로 설계되고, 소비자가 Memory 변경을 하는 것이 어렵게 되어있다. 반면에 데스크탑은 사용자가 DIMM(Dual In-line Memory Module)을 추가하는 것으로 쉽게 메모리 Customizing을 할 수 있다. Customizing 방식마다 Channel 특성이 달라지고, 모든 Channel 특성에서 안정적인 성능을 갖도록 설계되어야 하기 때문에, DDR은



출처 Reproduced from Anabase, CC BY-SA 3.0 〈https://creativecommons.org/licenses/by-sa/3.0〉, via Wikimedia Commons and © Raimond Spekking, CC BY-SA 4.0 (via Wikimedia Commons).

그림 2 DDR과 LPDDR의 Mainboard 크기 비교

LPDDR보다 속도를 향상시키는 것이 어렵다.

고성능 Memory의 영역을 GDDR이 담당하고 있었지만, 단일 Channel에서 속도를 증가시키는 것에는 언젠가 한계가 온다는 것을 모두 알고 있었다. 이에 따라 다른 관점에서 Total Bandwidth를 개선하려는 움직임들이 나타나기 시작했고, 2014년에 HBM이 개발되었다. HBM은 이름에서도 알 수 있듯이 Total Bandwidth를 증가시키는 것이 목적이다. 이를 위해서 속도를 증가시키는 것이 아니라 pin 개수를 늘려서 Bandwidth를 증가시킨다. 초기의 HBM은 가격이 비싸고, GDDR 대비 큰 장점이 없고, 수율이 떨어져서 수익성이 없었다. 하지만, 2020년의 HBM2E부터 성능이 개선되고, 전 세계적으로 인공지능 개발의 광풍이 불면서 인기가 폭발적으로 급상승하였다[6]. 다음 절에서는 인공지능 컴퓨팅을 위해 가장 많이 사용되는 GDDR과 HBM의 구조적 차이 및 Total Bandwidth 차이 등에 대한 이야기를 좀 더 자세하게 해보려고 한다.

2. GDDR과 HBM

GDDR은 세대를 거듭하며, DDR, LPDDR의 다른 DRAM들과의 속도 차이가 매우 크게 벌어져 왔다. 하지만, GDDR의 속도가 점점 빨라질수록 DRAM과 GPU 간에 Data를 전송할 때 Signal 특성을 확보하는 것이 점점 힘들어지고 있었다. GDDR의 속도 한계는 명확했기 때문에 많은 Data를 전달하기 위한 새로운 방법이 필요했다.

HBM은 속도를 늘리는 것보다 개수를 늘리는 방식을 선택했다. 1-pin의 속도를 올리는 것이 어렵기 때문에 pin의 개수를 늘려서 많은 Data를 Parallel 전송하는 것이다. GDDR이 1명이 1시간 동안 10박스를 옮긴다면, HBM은 100명이 1시간 동안 각자 1박스씩을 옮긴다. 1명의 속도는 GDDR이 빠르지만,

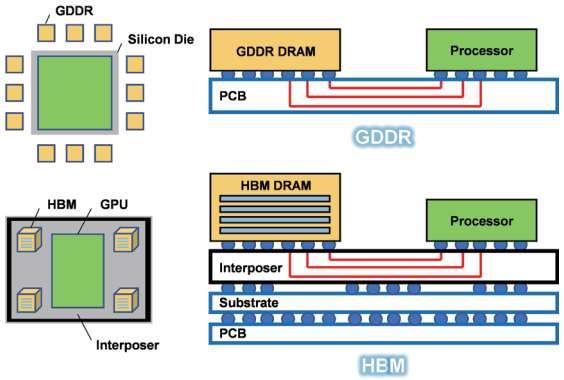


그림 3 GDDR과 HBM의 구조 비교

전체 옮겨진 박스는 GDDR이 10개, HBM이 100개가 된다. 실제로 Data 전송을 위해서 GDDR은 32개 pin을 사용하지만, HBM은 32배 많은 1,024개 pin을 사용한다.

Pin 수 증가와 함께 HBM 개발을 위해 반드시 선행되어야 하는 것이 DRAM 칩을 수직으로 쌓고 연결할 수 있어야 한다. 그림 3과 같이 GDDR은 단층의 메모리로 GPU와 연결되어 있지만, HBM은 복층의 메모리들이 연결되어 있다. 이렇게 메모리를 아파트처럼 수직으로 쌓는 연결 기술을 TSV(Through-Silicon Via)라고 하며, 각 메모리 회사들의 HBM 수율을 결정짓는 주요 기술이다. 여러 개의 Chip에 구멍을 뚫어서 수많은 Pin이 틀어짐 없이, 모든 층이 완벽하게 붙어야만 1개의 HBM이 완성되는 것이다. 만약 중간에 연결이 잘못되거나 불순물이 끼게 된다면, 적층한 모든 칩은 한꺼번에 폐기해야 한다. HBM의 가격이 다른 메모리보다 월등하게 비싼 이유는 이런 공정상의 어려움이 존재하기 때문이다. 최근에는 메모리 회사들이 회로적인 방법과 물성적인 방법들로 많은 노력을 하여, HBM에서도 높은 수율을 확보하고 있다.

그림 4는 GDDR과 HBM의 Total Bandwidth 발전을 시간순으로 나타낸다. 그림 1과 비교하여 HBM이 GDDR 대비 매우 높은 Bandwidth를 갖는 것을

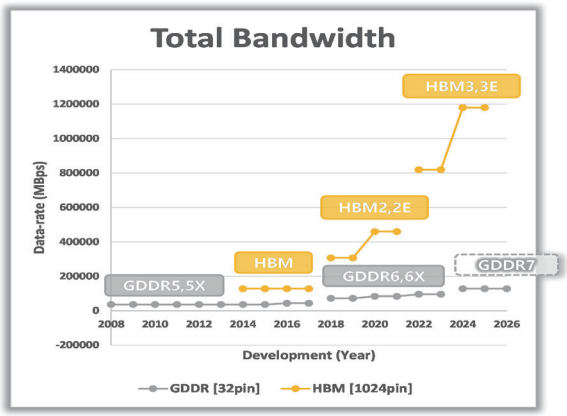


그림 4 GDDR과 HBM의 Total Bandwidth 비교

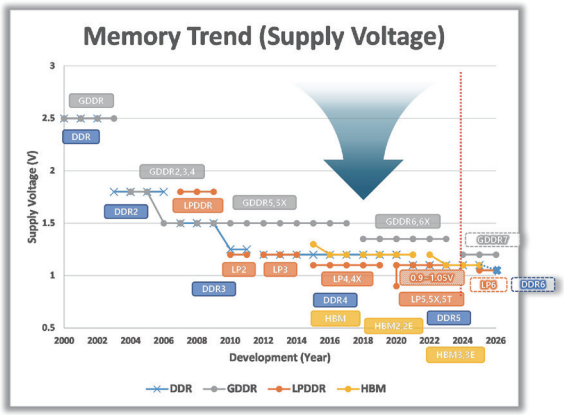


그림 5 DRAM 전압의 발전 방향

확인할 수 있다.

그러나 2014년의 HBM 개발 초기에는 높은 가격과 공정 안정성의 문제로 크게 주목받지 못했다. 최근 ChatGPT와 같은 인공지능 학습을 위해 NVIDIA의 AI GPU가 대량으로 사용되고, 여기에 HBM이 적극적으로 채용되면서, 가치와 수요가 급상승하고 있다.

Bandwidth 이외에도 HBM은 GDDR 대비 전력 소모 효율이 좋은 장점을 갖는다. 칩 1개가 소모하는 전력량은 HBM이 많을 수 있지만, 전송되는 Data당 사용되는 전력은 HBM이 더 적다.

반면, HBM은 GDDR 대비 발열 문제가 더 큰 단점이 있다. GDDR도 빠른 속도로 인해 발열을 무시할 수 없는데, HBM은 높은 Bandwidth와 함께 적층된 여러 개의 칩이 열의 방출을 방해하기 때문에, 발열 문제가 더 심각해지는 것이다.

3. 메모리 전력 소모의 발전 방향

그림 5는 DDR의 등장 이후부터 현재까지의 DDR DRAM의 전압 변화를 시간순으로 나타낸다. x축은 연도이고, y축은 전압을 나타낸다. 시간의 흐름에 따른 전체 그래프의 방향성은 "전압이 감소한다"이다.

전압이 계속 감소하는 이유는 전력 소모를 줄이기 위한 가장 효과적인 방법이기 때문이($P=VI=V^2/R$). 즉, 메모리는 전력 소모를 줄이는 방향으로 지속적인 발전을 하고 있다.

하지만, 전압을 감소하는 것에는 한계가 존재한다. 그림 5에서 각 제품마다 초기의 전압은 다르지만, 최근에는 1.1V 근방으로 수렴하는 것을 확인할 수 있다. Transistor가 기본적으로 동작할 수 있는 전압이 확보되어야지만 고속으로 Data를 전송할 수 있는데, 현재 DRAM 공정으로는 그 한계가 1V 정도로 나타나고 있다. 최신의 LPDDR 제품에서는 이런 상황에서도 최대한 전력을 줄이기 위해서 칩이 저속으로 동작할 때는 0.9V 정도로 전압을 낮추는 가변 전압 기법을 도입하였다[7]. 하지만, 이런 방식은 근본적인 해결책은 될 수 없고, 새로운 소자나 공정의 개발이 필요하다.

4. 메모리의 Signaling 동향

전통적으로 메모리의 Signaling은 그림 6(a)와 같은 NRZ(Non Return to Zero)를 사용하고 있다. NRZ는 1개의 Data를 표현하기 위해 High 전압과 Low

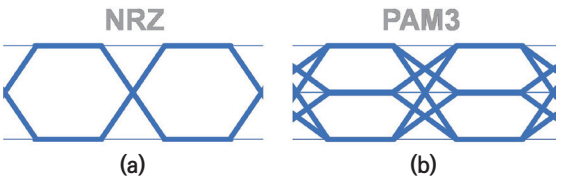


그림 6 NRZ와 PAM3의 파형 비교

전압의 2개 전압 Level만 사용하는 방식이다. 가장 간단한 구조의 전송 방식이기에 메모리에서 적극적으로 사용되어왔다. 하지만 1개의 Channel에서 NRZ로 전송할 수 있는 Speed가 한계에 다다르고 있다. DRAM에서 가장 빠른 GDDR6X에서는 NRZ 방식으로 24Gbps에 근접한 Signaling을 한다. 이 경우에 1개의 Data를 나타내는데 주어진 시간은 41ps 정도밖에 되지 않고, Channel에서 정확한 Signaling을 하는 것이 점점 어려워진다.

이런 문제를 해결하기 위해 GDDR7부터는 NRZ를 대신하여, Multi-level Signaling이 사용된다[8]. 그림 6(b)에 보이는 것과 같이 PAM3(Pulse-Amplitude Modulation 3)는 Data를 표현하기 위해 Middle 전압을 추가하여 3개의 Level을 사용한다. PAM3는 같은 시간 동안 NRZ 대비 약 1.5배의 Data를 전송할 수 있다.

PAM3의 단점으로는 Signal의 Height가 절반으로 감소하기 때문에, Noise에 취약하고, 회로가 복잡해진다. 하지만, NRZ의 속도 한계는 명확하기 때문에, GDDR을 시작으로 향후 DDR, LPDDR 등도 PAM3와 같은 Multi-level Signaling으로 발전할 가능성이 있다.

Ⅲ. 차세대 메모리 개발 동향

1. 기존 메모리 간의 비교

지금까지 메모리 중 DRAM의 변천사에 대해서 알아봤다. Data를 저장할 수 있는 메모리에는 DRAM 외에도, SRAM과 Flash가 있다[9,10]. DRAM이 빠르고 싼 가격으로 메모리 시장에서 가장 널리 사용되고 있지만, SRAM과 Flash도 각자의 특화된 영역에서 사용되고 있다. 인공지능 반도체를 위한 차세대 메모리에 필요한 특성들을 알기 위해 먼저 세 종류의 메모리들을 비교하려고 한다.

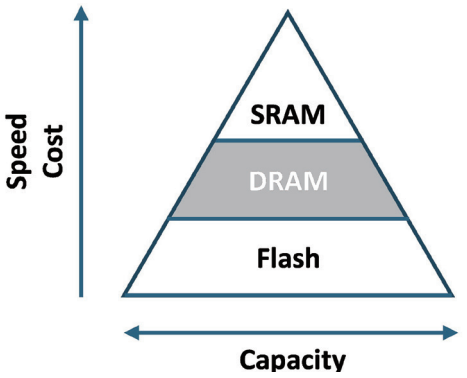


그림 7 기존 메모리 간의 피라미드 비교

그림 7은 메모리 간의 대표적인 특성 차이를 피라미드 형태로 보여준다. 중앙에 위치한 DRAM은 Speed, 가격, 대용량의 측면에서 어느 한 쪽으로 치우쳐 있지 않기 때문에 현재까지 가장 인기 있는 메모리로 사용될 수 있었다. SRAM은 DRAM보다 속도가 더 빠르지만, 비싸고, 대용량이 불가능하다. 반면, Flash는 싸고, 대용량이 가능하지만, 속도가 느리다.

하지만, DRAM도 몇몇 단점을 갖고 있고, 차세대 메모리의 개발을 위해서는 이런 특성들도 함께 비교되어야 한다. 가장 큰 단점은 Volatile 특성이다. 향후 인공지능 반도체는 메모리도 일부 연산에 관여하는 것이 필요한데, 전원이 꺼지면 모든 Data가 사라지는 Volatile 특성은 인공지능 반도체를 위한 차세대 메모리에 어울리지 않는다. DRAM에서 Data를 저장하는 Cell 구조는 Transistor 1개와 Capacitor 1개(1T1C: 1 Transistor 1 Capacitor)로 매우

단순하게 구성되어 있다. Capacitor에 전하를 저장하는데 전원이 인가되지 않으면 전하들이 사라지고, 결국 Data가 사라지는 것이다. 또한, Cell 구조상 Leakage로 인해 자연적으로 발생하는 Data 오염을 막기 위해 주기적으로 Refresh를 해주는 것으로 전력을 소모한다. Data에 접근하지 않을 때도 전력을 소모하는 특성은 저전력 반도체와는 어울리지 않는다. Cell에 사용되는 Capacitor는 Low-voltage로 구동되기 어렵고, 미세화하는 것도 어렵다.

그림 8은 다양한 특성 지표에서 DRAM과 SRAM, Flash 간의 비교를 레이더 차트로 보여준다. 차세대 메모리를 위한 항목들에서 SRAM과 Flash가 DRAM보다 나은 특성을 보이는 항목들도 있지만, DRAM

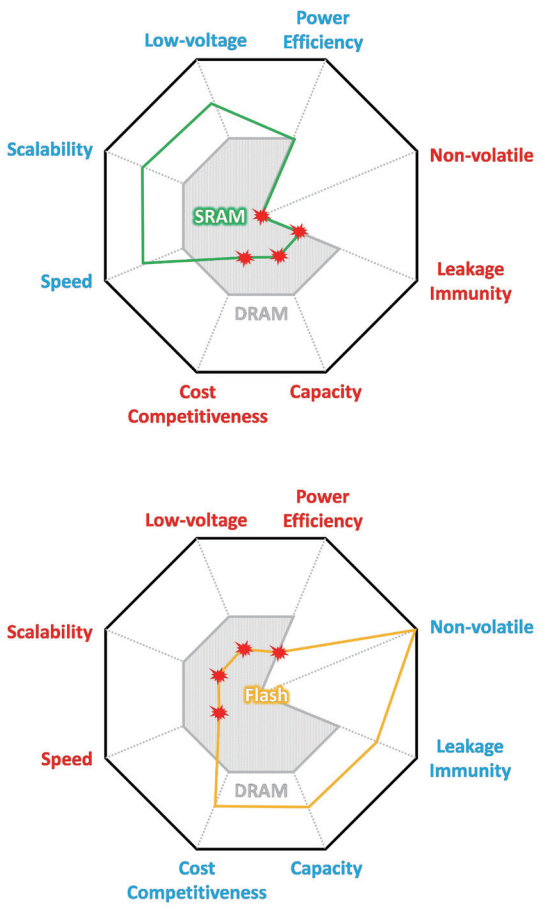


그림 8 기존 메모리 간의 레이더 차트 비교

의 단점을 극복하지 못했거나, 더 안 좋은 특성을 보이는 것을 확인할 수 있다. 결론적으로, SRAM과 Flash가 인공지능 반도체를 위한 차세대 메모리로 DRAM을 대체하여 사용되기는 어렵다는 것을 쉽게 알 수 있다.

2. 차세대 메모리의 후보

DRAM은 많은 장점이 있는 메모리지만, 앞에서 살펴본 것처럼 단점 또한 존재한다. 현재 DRAM의 단점은 1T1C 구조에서 발생하는 특성들이 많다. 그러므로, 차세대 메모리를 위해서는 새로운 Cell을 사용하는 메모리들이 고려되는 것이 필요하다.

새로운 Cell을 사용하는 메모리는 다양하게 연구되고 있다. 기존의 DRAM이 Capacitor에 저장된 Charge를 이용했다면, 새로운 메모리들은 전자의 Spin을 이용하거나, 저항성을 이용하거나, 위상을 이용하거나, 화학적 방법을 사용하는 방법 등이 있다.

여러 가지 메모리 중 현재 양산성이 가장 높은 것은 Spin 특성을 사용하는 STT-MRAM(Spin-Transfer Torque Magnetic RAM)이다[11]. STT-MRAM은 DRAM과 많은 부분에서 비슷하거나 좋은 특성을 갖는다. 다음 장에서는 STT-MRAM이 차세대 메모리로서 가능성이 있는지 확인하기 위해 DRAM의 특성들과 비교해 보고자 한다.

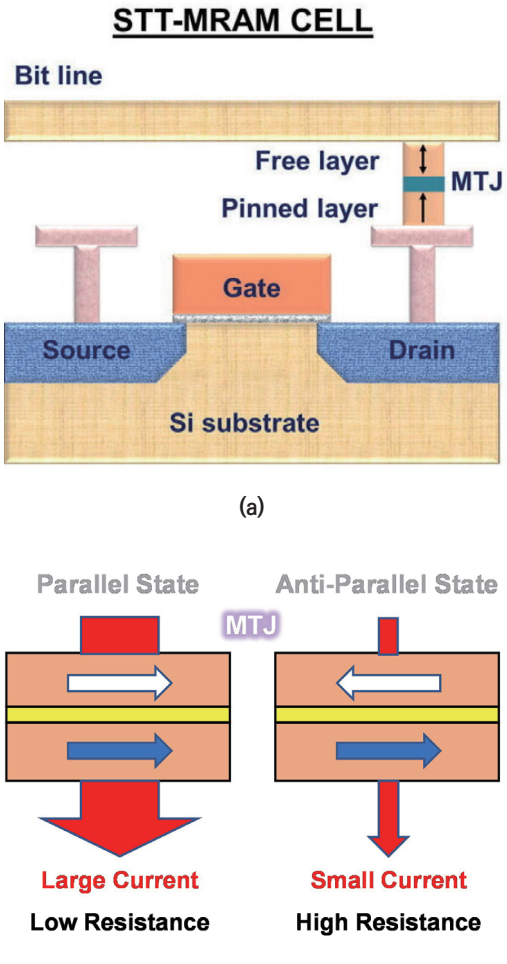
Ⅳ. STT-MRAM의 가능성

1. STT-MRAM의 기초

STT-MRAM의 특성을 비교하기 전에 먼저 동작 원리에 대해 알아보자. STT-MRAM은 자성의 근원인 전자의 Spin 특성을 이용한다. 그림 9(a)[12]에서 표시된 MTJ(Magnetic Tunnel Junction) 소자의 Spin 방향을 조절해서 Data를 저장하고 읽어낸다.

Data를 저장하기 위해 STT-MRAM은 전자가 이동하며, 주변의 Spin 방향에 영향을 주는 Torque로 MTJ를 조절한다. MTJ에 흐르는 전류의 방향을 바꾸면, MTJ의 Spin 방향이 바뀌고, MTJ에 저장되는 Data가 변경되는 것이다.

Data를 읽기 위해서는 그림 9(b)와 같은 GMR (Giant Magneto Resistance) 효과를 이용한다. GMR 효과는 MTJ의 2개 Layer에서 Spin 방향이 같으면, 전류가 많이 흐르고, 반대일 경우에는 전류가 거의 흐르지 않는 현상이다. 즉, MTJ에 작은 전류를 흘리고, 저항을 측정하는 것으로 저장된 Data가 0인지 1인지 읽어낼 수 있다.



출처 (a) Reprined from S. Bhatti et al., "Spintronics based random access memory: a review," Materials Today, vol. 20, no. 9, 2017, CC BY NC-ND.

그림 9 STT-MRAM의 구조 및 동작 원리

2. DRAM과 STT-MRAM

MTJ 사용으로 얻어지는 STT-MRAM의 가장 큰 장점은 Non-volatile이다. DRAM은 Capacitor에 전하 형태로 Data를 저장하는 방식이기 때문에 컴퓨터의 전원을 끄면 저장된 모든 정보가 소실된다. 또한 Cell의 값이 Leakage에 의해 변하고, 이를 방지하기 위해 주기적으로 Data를 다시 써주는 Refresh 동작이 필요하다. Refresh 동작은 DRAM의 성능을 제한하는 것뿐만 아니라 전류 소모도 증가시킨다. 반면, MTJ의 Spin 방향은 시간이 지남에 따라 변하지도 않고, Leakage가 존재하지도 않으며, 전원의 여부와 상관없이 유지된다. 그러므로, STT-MRAM에 저장된 Data는 DRAM과 다르게 항상 유지되고, Refresh 동작도 필요하지 않다.

그림 10은 STT-MRAM의 특징들을 DRAM과 비교하여 레이더 차트로 나타낸 것이다. 먼저, STT-MRAM이 DRAM을 대체할 가능성이 있는 이유는

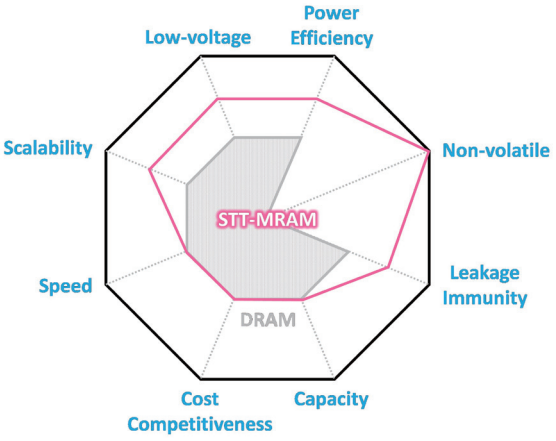


그림 10 DRAM과 STT-MRAM의 비교

Speed와 가격, 대용량 특성이 DRAM과 비슷하기 때문이다. 기본적인 특성이 비슷하면서도, DRAM의 단점들을 보완하고 있기 때문에 차세대 반도체로서의 가능성이 있다.

STT-MRAM은 저전력을 위한 Low-voltage에 더 유리하다. DRAM은 Cell에 고전압을 사용해야 하기 때문에 Low-voltage로의 전환이 쉽지 않다. STT-MRAM은 Leakage도 없고, Refresh도 필요 없다. 이런 점들은 종합적으로 Power Efficiency를 개선한다. 공정이 미세화되면서 칩의 크기도 줄어들 것인가를 나타내는 Scalability도 DRAM Cell의 Capacitor의 크기 감소가 제한적이기 때문에 STT-MRAM이 더욱 유리하다.

이와 같은 특징들로 인해 STT-MRAM은 새로운 인공지능 반도체를 위한 차세대 메모리 중에서 가장 선두에 서 있다고 할 수 있다. 뿐만 아니라 차세대 메모리들 중 양산성이 가장 빠르게 확보되고 있기 때문에 향후 STT-MRAM의 발전이 더욱 기대된다.

V. 결론

현재 전 세계는 인공지능의 거대한 바람을 맞이하고 있다. LLM을 필두로, 생성형 AI들이 빠르게 발전하고 있으며, 실생활에도 많이 사용되고 있다. 지금까지는 소프트웨어 위주의 발전이 주를 이루고 있지만, 새로운 하드웨어의 필요성도 주목받고 있다. 메모리 중에서 널리 사용되고 있는 DRAM은 속도와 전력에서 꾸준히 많은 발전을 이뤄왔고, GDDR과 HBM이 인공지능 반도체를 위해서 사용되고 있다. 하지만, DRAM Cell의 Volatile 특성이 인공지능 반도체의 구현을 위해서는 적합하지 않다. 속도와 전력이 DRAM과 비슷하고, Non-volatile 특성을 갖는 STT-MRAM은 차세대 인공지능 메모리로서의 가능성을 충분히 갖고 있다고 전망된다.

약어 정리

1T1C	1 Transistor 1 Capacitor
AI	Artificial Intelligence
CAD	Computer-Aided Design
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
DDR	Double Data-Rate
DIMM	Dual In-line Memory Module
DRAM	Dynamic Random Access Memory
GDDR	Graphic DDR
GMR	Giant Magneto Resistance
GPGPU	General-Purpose GPU
GPU	Graphic Processing Unit
HBM	High-Bandwidth Memory
LLM	Large Language Model
LPDDR	Low-Power DDR
MTJ	Magnetic Tunnel Junction
NPU	Neural Processing Unit
NRZ	Non Return to Zero
PAM3	Pulse-Amplitude Modulation 3
SDR	Single Data-Rate
STT-MRAM	Spin-Transfer Torque Magnetic RAM
TSV	Through-Silicon Via

참고문헌

[1] R.H. Dennard, "Field-effect transistor memory," US3387286A, Jun. 4, 1968.

[2] https://www.jedec.org/about-jedec/jedec-history/2000s/

[3] H.Y. Yoon et al., "A 2.5-V, 333-Mb/s/pin, 1-Gbit, double-date-rate synchronous DRAM," IEEE J. Solid-State Circuits, vol. 34, no. 11, Nov. 1999, pp. 1589-1599.

[4] K.-D. Hwang et al., "A 16 Gb/s/pin 8 Gb GDDR6 DRAM with bandwidth extension techniques for high-speed applications," in IEEE Int. Solid-State Circuits Conf., (San Francisco, CA, USA), Feb. 2018, pp. 210-212.

[5] T.-Y. Oh, "A 3.2 Gbps/pin 8 Gbit 1.0 V LPDDR4 SDRAM with integrated ECC engine for sub-1 V DRAM core operation," IEEE J. Solid-State Circuits, vol. 50, no. 1, Jan. 2015, pp. 178-190.

[6] D.U. Lee et al., "A 128 Gb 8-high 512 GB/s HBM2E DRAM with a pseudo quarter bank structure, power

dispersion and an instruction-based At-speed PMBIST," in IEEE Int. Solid-State Circuits Conf. (San Francisco, CA, USA), Feb. 2020, pp. 334-336.

[7] D.-H. Kim et al., "A 16 Gb 9.5 Gb/S/pin LPDDR5X SDRAM with low-power schemes exploiting dynamic voltage-frequency scaling and offset-calibrated readout sense amplifiers in a fourth generation 10nm DRAM process," in IEEE Int. Solid-State Circuits Conf. (San Francisco, CA, USA), Feb. 2022, pp. 448-449.

[8] J.-H. Yang et al., "13.1 A 35.4Gb/s/pin 16Gb GDDR7 with a Low-Power Clocking Architecture and PAM3 IO Circuitry," in IEEE Int. Solid-State Circuits Conf. (San Francisco, CA, USA), Feb. 2024, pp. 232-233.

[9] K. Ishibashi et al., "Low Power and Reliable SRAM Memory Cell and Array Design," Berlin, Germany, Springer, 2011.

[10] H.-J. Kim et al., "1 GB/s 2Tb NAND flash multi-chip package with frequency-boosting interface chip," in IEEE Int. Solid-State Circuits Conf. (San Francisco, CA, USA), Feb. 2015, pp. 1-3.

[11] S. Jung et al., "A crossbar array of magnetoresistive memory devices for in-memory computing," Nature, vol. 601, 2022, pp. 211-216.

[12] S. Bhatti et al., "Spintronics based random access memory: a review," Materials Today, vol. 20, no. 9, 2017, pp. 530-548.