

Chapter

01

# Processing-in-Memory 반도체 기술 동향

박성모\_한국전자통신연구원 책임연구원

최병건\_한국전자통신연구원 책임연구원

박필재\_한국전자통신연구원 책임연구원

박경환\_한국전자통신연구원 책임연구원

인공지능, 빅데이터 등 4차 산업혁명 도래에 따라 산업 및 사회 전반에서 반도체 의존도가 높아지고 있다. 뇌와 같은 저전력이면서 병렬처리가 가능한 컴퓨팅의 연구가 지난 수십 동안 연구가 계속되었다. 기존 폰노이만 구조는 연산과 메모리의 데이터 이동에 따른 문제점은 여전히 문제점을 가지고 있다. 이러한 문제점을 해결하기 위해서 메모리 중심의 컴퓨팅 기술인 PIM(Processing-In-Memory) 기술이 최근 관심을 받고 연구가 진행되고 있다. 특히, 인공지능반도체는 의료, 교육, 제조업, 국방 분야 등 사회 전반의 영역으로 확장되고 있다. 신개념의 PIM 반도체는 메모리 중심의 반도체 기술로 세계 최고의 메모리 기술력을 보유한 국내 산업이 그동안 CPU, GPU 중심의 인텔, AMD, NVIDIA 등이 주도하는 반도체 시장의 패러다임을 바꿀 기회이다. 본 고에서는 PIM 기술에 대한 최근 기술 동향에 대해서 알아보려고 한다. 인공지능반도체 기술에 대한 역할을 살펴보고, 차세대 반도체 구조인 PIM 방식에 대한 개발 동향에 대해 살펴본다.

## I. PIM 시장 현황

인공지능을 탑재한 자율주행차, 드론, 비행택시, 가전 AR/VR 등의 서비스는 날로 진화하여 대규모의 데이터를 고속으로 처리하기 위한 고성능 및 저전력의 컴퓨팅이 필요하게 되었다. 현재 프로세서의 폰노이만 방식의 컴퓨팅 구조는 연산과 저장 기능이 분리된 근본적인 구조로 인해서 성능과 전력 효율 향상에 문제가 있다. 반도체 내부의 대량의 데이터 이동으로 인한 데이터 병목현상과 집적도 향상이 한계에 도달하였다[2],[3]. 이런 문제점을 해결하

\* 본 내용은 박성모 책임연구원(☎ 042-860-5203, smpark@etri.re.kr)에게 문의하시기 바랍니다.

\*\* 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.

\*\*\*이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임  
(No. 2022-0-00181, 전파분야 소재부품 국산화(차량용 ROA 센서 칩/모듈 양산 시제품 개발))

기 위해 CPU 중심 컴퓨팅을 뇌 구조와 유사한 메모리 중심 컴퓨팅으로 전환하여 연산과 저장 기능을 통합한 PIM(Processing-In-Memory) 기술이 새로운 컴퓨팅으로 주목을 받고 있으며, 프로세서와 메모리를 통한 반도체 칩 상에서 최단 거리로 배치하거나 신소자로 연산과 메모리 기능을 동시 구현하여 성능 및 전력 효율 개선을 통해 현재의 컴퓨팅 구조의 근본적 문제를 해결하였다.

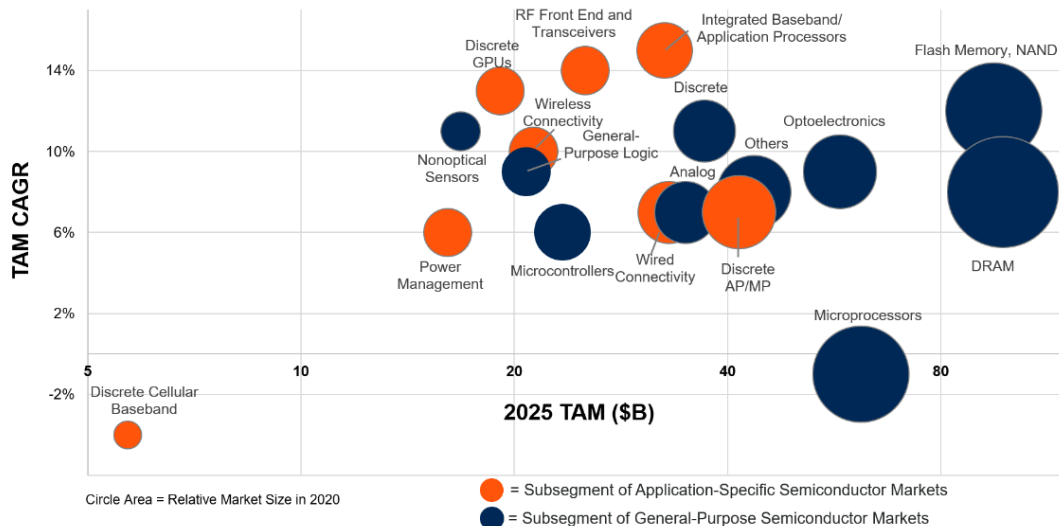
신개념의 PIM 반도체는 메모리 중심의 반도체 기술로 세계 최고의 메모리 기술력을 보유한 국내 산업이 그 동안 CPU, GPU 중심의 인텔, AMD, NVIDIA 등이 주도하는 반도체 시장의 패러다임을 바꿀 기회가 되었다. 또한, 기계학습 및 인공지능과 같은 새로운 응용 프로그램은 신경망에서 병렬 데이터를 처리하기 위한 저전력 하드웨어 가속기에 개발을 필요하게 되었다. 다중 및 누계(MAC) 연산은 신경망의 주요 산술 함수이며, 폰노이만 아키텍처에서는 연산부와 메모리가 분리된 구조로써 다중 및 누산 연산을 실행하기 위해서는 많은 양의 데이터가 연결 채널을 통해 전송되어야 한다. 이처럼 매우 빈번한 데이터 통신이 매우 높은 전력을 소비하는 문제점이 있다. 이는 에너지 효율적인 엣지 컴퓨팅 시스템에서는 큰 문제점으로 대두되고 있다.

PIM 아키텍처에서는 이러한 문제점을 해결하기 위해 병목현상을 극복하는 것으로 알려졌다[2]-[13]. PIM 아키텍처에서 각 처리 요소에는 컴퓨팅 회로와 메모리가 있어 외부 메모리와의 데이터 전송 빈도를 줄인다. 전력 소모가 많은 데이터 전송이 최소화되기 때문에 PIM 아키텍처는 에너지 효율성을 몇 배나 향상할 수 있으며, 이를 위해 SRAM 기반 PIM, DRAM 기반 PIM 및 ReRAM 기반 PIM과 같은 메모리 유형에 따라 최신 PIM 연구가 진행되고 있다. SRAM은 간단한 작동 모드와 성숙한 기술 덕분에 PIM 매크로를 구현하기 위한 다른 후보와 비교하여 가장 많이 연구가 진행되고 있다[6],[7]. 그러나 SRAM 셀의 면적은 DRAM 및 ReRAM과 같은 다른 후보보다 면적 면에서는 메모리 밀도가 낮다. DRAM 기반 PIM은 대용량 메모리로 대형 기계학습 모델을 가속하는 매력적인 솔루션이지만 DRAM 셀의 고밀도는 DRAM 기반 PIM의 구현에는 문제점이 있다. 그러나 3차원 적층 DRAM에 처음으로 제작된 PIM 칩으로 실현 가능성이 입증되었다.

세계 반도체 시장의 국가별 비중은 미국 47%, 한국 19%, 일본 10%, 유럽 10%, 대만 6%, 중국 5%이며, 미국은 시스템 반도체와 아날로그 반도체에서 시장을 주도하고 있으며, IDM(Integrated Device Manufacturer) 시장의 51%, 팹리스 시장의 65%를 차지한다.

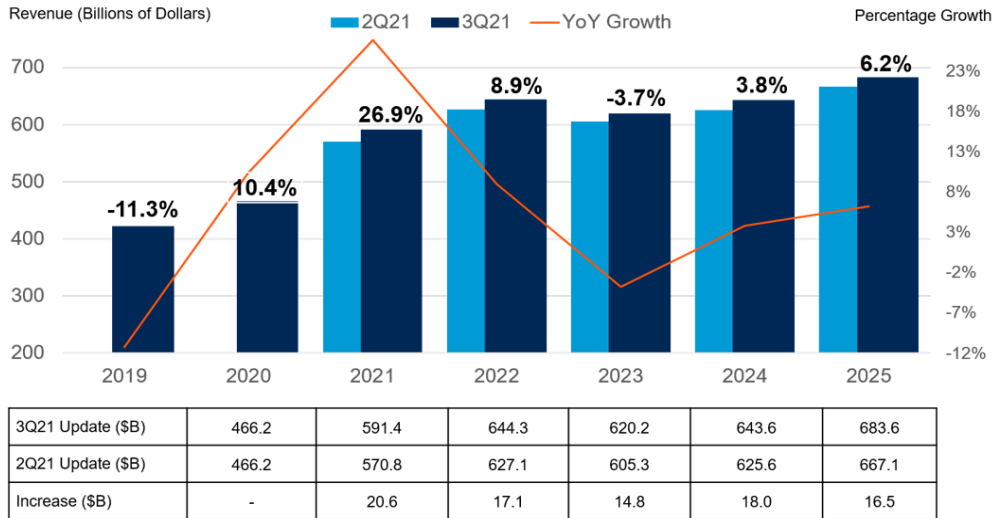
또한, 한국은 메모리 반도체, 유럽과 일본은 광 개별 소자 시장을 주도하고 있다. PIM 반도체의 주 수요처는 파운드리 시장 및 업체와 프로세서 업체로서, 글로벌 파운드리 시장은 미국이 AMD, 애플, 브로드컴, 퀄컴, NVIDIA 등과 같은 대형 팹리스 기업이 있어서 가장 수요가 많다. 글로벌 파운드리 시장에서는 상위 8개 기업이 88.4%를 점유하고 있으며, 그 중에서 대만 기업이 약 60%를 차지한다. 최근 5년 간 중국 파운드리 시장이 빠른 속도로 성장 중이며 국내 시장은 글로벌 메모리 반도체 강자인 삼성전자와 SK하이닉스가 파운드리 사업 부문을 분리하면서 파운드리 사업 강화와 함께 파운드리 시장 수요가 증가한다. 또한, 파운드리 분야 경쟁력을 강화하기 위해 메모리 반도체가 연산 기능을 수행하는 PIM 시장이 확대를 위한 PIM 사업을 통해서 인공지능 보편화와 빅데이터 처리 속도 향상에 이바지할 기술로 평가받고 있다.

PIM 시장은 앞으로 반도체 시장의 성장과 같이 성장할 것으로 예상된다. 반도체 시장규모를 예측하는 보고서들은 대부분 2025년까지는 6,000억~7,000억 달러 정도의 시장규모를 제시하고 있고 2021년을 기준으로 2025년에는 매년 140억 달러에서 200억 달러로 성장하고 있으며, 연평균성장률(Compound Annual Growth Rate: CAGR)은 2021년에 26.9%로 최고를 기록하다가 2025년에는 6.2% 이상을 예측된다[1](그림 [1-2] 참조).



〈자료〉 Gartner, "Semiconductor Revenue Forecast by Device Type Through 2025," 2021.

[그림 1] 반도체 디바이스별 시장 전망



〈자료〉 Gartner, "Semiconductor Revenue Forecast by Device Type Through 2025," 2021.

[그림 2] 반도체 시장 전망

본 고에서는 PIM 기술이 인공지능반도체 분야에 어떻게 적용되고 있는지를 설명하고자 한다. II 장에서는 SRAM 기반 PIM 기술에 관한 내용을 다루고, III 장에서는 DRAM 기반 PIM 기술에 대해 다룬다. IV장에서는 ReRAM 기반의 PIM 기술에 대해 언급을 하고, V장에서는 본 고의 결론을 제시한다.

## II. SRAM 기반 PIM 기술 동향

최근의 SRAM 기반의 PIM 기술은 인공지능반도체인 뉴로모픽은 MAC 연산을 위한 가중치의 값을 외부 메모리에서 읽고 쓰기가 빈번하여 대역폭 제한의 문제점이 있다. 이 문제를 개선하기 위해 다양한 구조연구가 진행되어 왔다. 기존 SRAM은 가장 직관적인 동작을 통한 이론 및 성숙한 제조 기술을 가지고 있어 PIM 매크로 구축의 어려움을 줄이고 인공 신경망 가속기 구축을 위한 가장 인기 있는 후보이다. 순수 메모리 어레이에서 PIM 매크로로의 전환에 따른 문제를 해결하기 위한 트레이드 오프 분석과 함께 여러 SRAM 기반 작업이 진행되었으며, 아날로그 및 디지털 기반의 솔루션을 포함하고 있다. 아날로그 PIM 매크로는 인공 신경망의 유연성과 분류 정확도의 한계를 표현하면서 높은 에너지/면적 효율 성능을

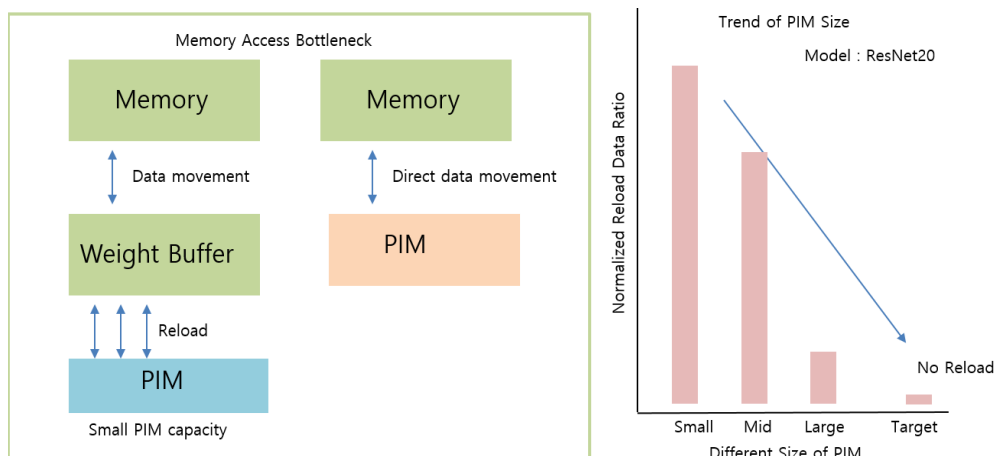
[표 1] SRAM 기반의 PIM

| 논문               | 연도      | 공정    | 비트 셀        | 기능   | 핵심 내용               |
|------------------|---------|-------|-------------|------|---------------------|
| 6T ML Classifier | JSSCC17 | 130nm | Standard 6T | MUL  | Compact Bitcell     |
| Split-6T PIM BNN | SOVC19  | 28nm  | Split WL 6T | XNOR | Compact Bitcell     |
| Charge-based PIM | JSSCC19 | 65nm  | 8T+1C       | XNOR | Less Variation      |
| CONV-SRAM        | JSSCC19 | 65nm  | 10T         | MUL  | Wide Dynamic Range  |
| C3SRAM PIM       | JSSCC20 | 65nm  | 8T+1C       | XNOR | Less Variation      |
| XNOR-SRAM        | JSSCC20 | 65nm  | 12T         | XNOR | Wide Dynamic Range  |
| 8T SRAM-PIM      | CICC20  | 65nm  | 8T          | MUL  | Diff. Read Bitline  |
| T8T SRAM PIM     | JSSC20  | 55nm  | T8T         | MUL  | Large Signal Margin |
| 7nm SRAM PIM     | JSSC21  | 7nm   | Foundry 8T  | AND  | Foundry Cel         |
| 6T SRAM 8b PIM   | ISSCC21 | 28nm  | 6T+SILMC    | MUL  | High Precision      |

〈자료〉 KIM et al., "OVERVIEW OF PROCESSING-IN-MEMORY CIRCUITS FOR ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING," IEEE Journal on emerging and selected topics in circuits and systems, Vol.12, No.2, June 2022, p.339.

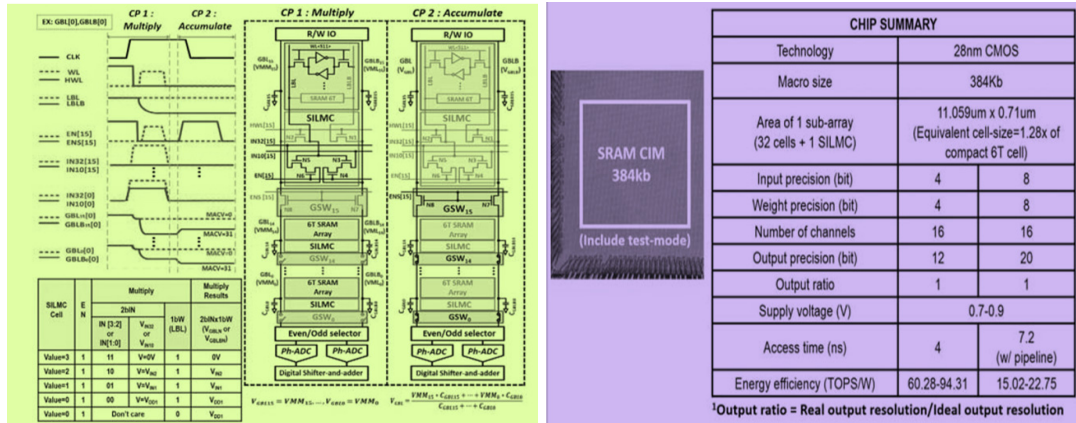
가지는 특징이 있다. 반면에 디지털 PIM 매크로는 효율성과 처리량이 적지만 물리적 변화를 피하는 데 장점이 가진다. 최신 SRAM 기반 PIM의 요약은 [표 1]과 같다[15].

[그림 3]은 SRAM 기반의 PIM 구조를 보여준다. 기존의 메모리 액세스 병목 문제를 해결



〈자료〉 Jian-Wei Su, et. al., "A 28nm 384kb 6T-SRAM Computation-in-Memory Macro with 8b Precision for AI Edge Chips," ISSCC 2021. p.251.

[그림 3] SRAM 기반 PIM 구조



Segment-bitline charge-sharing(SBCS) operation and sourceinjection local-multiplication cell(SILMC)

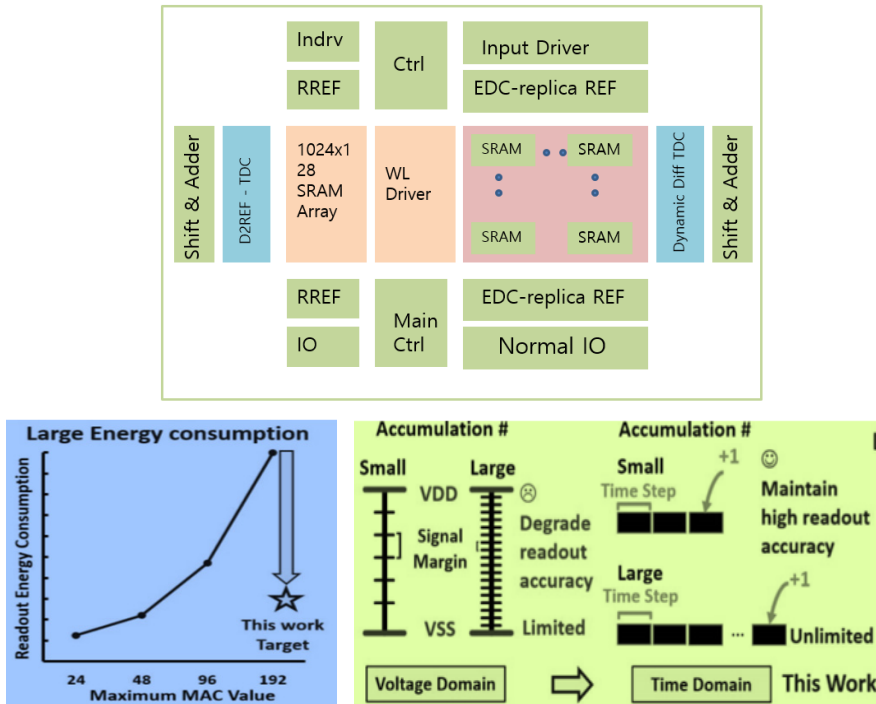
Die photo and summary table

〈자료〉 Jian-Wei Su, et. al, "A 28nm 384kb 6T-SRAM Computation-in-Memory Macro with 8b Precision for AI Edge Chips," ISSCC 2021. pp.251-252.

[그림 4] SILMC 기반 PIM 칩 사진

하기 위해서 새로운 구조를 제시하고 있다. 제안된 구조는 데이터의 이동을 최소화하기 위한 구조로써 기존에는 메모리와 가중치 버퍼의 데이터 이용이 빈번하여 데이터 처리에 제한적인 문제를 직접적으로 데이터 이동이 가능한 구조로 변경함으로써 해결하였다. 최종적인 목표는 다른 크기의 PIM을 데이터의 이동을 보장하는 것을 목표로 하고 있다. [그림 4]는 Segment Bitline Charge Sharing(SBCS) 동작이다. 초기에 PIM 모드에서 HWL=0은 N1/N2를 끄고 GBL/GBLB를 통한 LBL/LBLB의 사전 충전을 중지하며, GBL/GBLB 쌍은 VDD1로 사전 충전을 한다. 각 서브 어레이에서 선택된 WL이 활성화되어 액세스된 SRAM 셀에 저장된 1b 가중치 데이터를 자신의 LBL에 전달하여 N3/N5를 제어하며, 인접 GBL/GBLB 세그먼트에서 분리한다. W=1 및 EN=1일 때 주어진 4b 입력의 MSB 부분 아날로그 전압 절연 기생 커패시터로 전달한다. 전하 공유 작업을 통해 전압을 평균화 한다. 따라서 GBL 전압은 입력의 MSB 부분을 사용하여 16채널 누적으로 2bIN×1bW의 pMACV를 나타낸다.

제작된 칩은 최대 8bIN-8bW-20bOUT MAC 작업을 지원하는 파운드리 콤팩트-6T SRAM 셀을 사용하여 제작된 28nm 384kb SRAM-PIM이다. 셀 어레이는 일반 콤팩트 규칙 6T SRAM 어레이보다 1.28배 크고 8T 셀 어레이보다 작다. 8bIN-8bW-16채널 MAC 작업에서 tAC=7.2ns 및 22.75TOPS/W를 나타내고 있다[16].

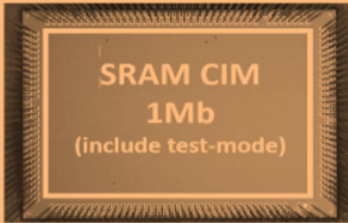


〈자료〉 Ping-Chun Wu, et. al, "A 28nm 1Mb Time-Domain Computing-in-Memory 6T-SRAM Macro with a 6.6ns Latency, 1241GOPS and 37.01TOPS/W for 8b-MAC Operations for Edge-AI Devices," ISSCC2022. p.191.

[그림 5] SRAM 기반 PIM

[그림 5]는 SRAM 기반의 PIM 구조에서 연산량이 많은 MAC에서 에너지 소비를 낮출 수 있는 구조를 새롭게 제안하였다. 기존의 전압 영역에서 MAC 연산의 전력 소모를 최소화한 구조였다면 타임 영역에서 에너지 소모를 최소화 하는 구조를 제안하였다. 구조는 16개의 6T-SRAM 셀, 4개의 EDC(Edge-Delay Cell) 및 1개의 EDC 멀티플렉서(EDC-MUX)로 구성된 2개의 열로 구성되며, SRAM 모드에서 읽기 및 쓰기 작업용으로 8b-IN MAC 작업을 위한 DCU는 30개의 트랜지스터와 32개의 SRAM 셀을 사용한다. 32개의 소형 영역 6T SRAM 셀에 대한 1.6배 영역 오버헤드가 있다. 선택된 WL은 4에서 N9/N10을 제어하기 위해 선택된 LBL/LBLB 및 MUX-트랜지스터를 통과하는 큰 전압 스윙으로 DCU의 액세스된 셀에 저장된 1b 가중치 데이터를 읽기 위해 활성화되며, VIN 레벨이 다르다면 단일 EDC 내에서 2b IN×1b W의 곱셈 결과의 차이로 인해 상승 에지 출력(EOUT10)에서 다른 지연(tEDC)이 생성된다. W=0 또는 IN=0일 때 MID=0 및 EDC는 EIN과 EOUT 단자 사이에



|  <p>SRAM CIM<br/>1Mb<br/>(include test-mode)</p> <p><sup>1</sup>Performance in 90% input sparsity<br/> <sup>2</sup>Using ResNet-20 model and the software baseline was 92.2%<br/> <sup>3</sup>Using ResNet-20 model and the software baseline was 68.03%</p> | CHIP SUMMARY                                                   |                                                                   |             |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|-------------------------------------------------------------------|-------------|
|                                                                                                                                                                                                                                                                                                                                               | Technology                                                     | 28nm CMOS                                                         |             |
|                                                                                                                                                                                                                                                                                                                                               | Macro size                                                     | 1Mb                                                               |             |
|                                                                                                                                                                                                                                                                                                                                               | Area of 1 sub-array<br>(32 cells + 4 EDCs)                     | 8.55um x 1.42um<br>(Equivalent cell-size=1.6x of compact 6T cell) |             |
|                                                                                                                                                                                                                                                                                                                                               | Input precision (bit)                                          | 4                                                                 | 8           |
|                                                                                                                                                                                                                                                                                                                                               | Weight precision (bit)                                         | 4                                                                 | 8           |
|                                                                                                                                                                                                                                                                                                                                               | Number of input channels                                       | 64                                                                | 64          |
|                                                                                                                                                                                                                                                                                                                                               | Number of output channels                                      | 256                                                               | 64          |
|                                                                                                                                                                                                                                                                                                                                               | Output precision (bit)                                         | 14                                                                | 22          |
|                                                                                                                                                                                                                                                                                                                                               | Supply voltage (V)                                             | 0.65-0.9                                                          |             |
|                                                                                                                                                                                                                                                                                                                                               | Access time (ns)                                               | 6.5 (0.9V)                                                        | 6.6 (0.9V)  |
|                                                                                                                                                                                                                                                                                                                                               | Throughput (GOPS)                                              | 4256-4964.8                                                       | 1050-1241.2 |
|                                                                                                                                                                                                                                                                                                                                               | Energy efficiency (TOPS/W)<br>(Average performance)            | 84.45-112.6                                                       | 21.19-27.75 |
|                                                                                                                                                                                                                                                                                                                                               | Energy efficiency (TOPS/W)<br>( <sup>1</sup> Peak performance) | 118.8-148.1                                                       | 27.21-37.01 |
|                                                                                                                                                                                                                                                                                                                                               | <sup>2</sup> Inference Accuracy (CIFAR-10)                     | 91.72%                                                            | 92.08%      |
|                                                                                                                                                                                                                                                                                                                                               | <sup>3</sup> Inference Accuracy (CIFAR-100)                    | 67.33%                                                            | 67.81%      |

〈자료〉 Ping-Chun Wu, et. al, "A 28nm 1Mb Time-Domain Computing-in-Memory 6T-SRAM Macro with a 6.6ns Latency, 1241GOPS and 37.01TOPS/W for 8b-MAC Operations for Edge-AI Devices," ISSCC2022. p.192.

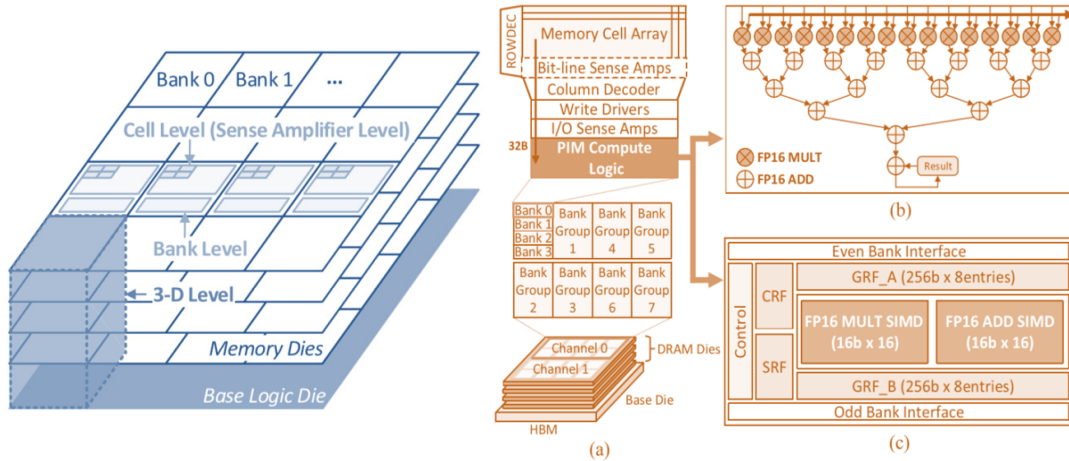
[그림 6] SRAM 기반 Time-Domain PIM 칩 사진 및 성능

고유 지연( $t_0$ )을 발생하며,  $W=1$  및 a 2b IN=01일 때  $V_{01}$ 은 P4의 게이트 지연을 변조하고 EIN과 EOUT 단자 사이에  $t_0 + \Delta t$ 의 지연을 생성한다. 제작한 공정은 28nm 1Mb SRAM-PIM 칩을 가지며, 6.6ns tAC 및 37.01TOPS/W 성능을 나타낸다. 또한, 이전 10배 이상 높은 FoM(IN 정밀도 $\times$ W 정밀도 $\times$ 출력 비율 $\times$ 처리량 $\times$ 용량)을 가지며 증가된 처리량, 더 높은 IN, W 및 OUT 정밀도, 더 큰 용량 및 더 빠른 tAC로 인해 시스템 수준 추론 정확도는 8b IN-8bW 정밀도의 ResNet20 모델을 사용하는 CIFAR-10 및 CIFAR-100 데이터 세트에 적용할 때 92.08% 및 67.81% 정확도를 나타내고 있다[17].

### III. DRAM 기반 PIM 기술

본 장에서는 DRAM의 기존 아키텍처와 작동뿐만 아니라 다양한 PIM 아키텍처와 그 구현을 소개한다. DRAM 아키텍처는 각 셀이 단일 트랜지스터와 커패시터의 단순한 구조를 갖는





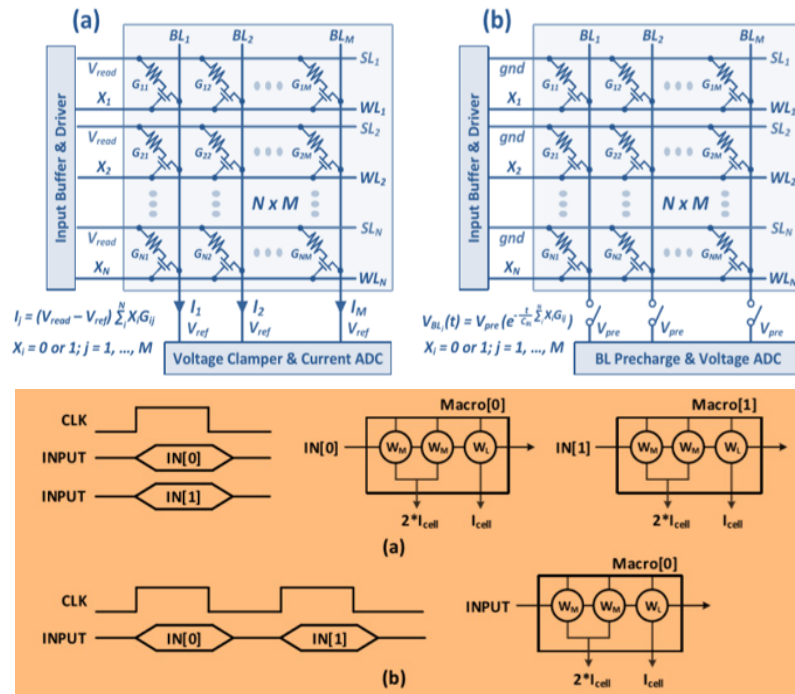
\* Different levels of DRAM PIN architectures (a) Bank-level PIM organization, (b) Newton compute logic, (c) HBM-PIM compute unit  
 <자료> KIM et al., "OVERVIEW OF PROCESSING-IN-MEMORY CIRCUITS FOR ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING", IEEE Journal on emerging and selected topics in circuits and systems, Vol.12, No.2, June 2022, p.339.

[그림 7] DRAM PIM 아키텍처 구조

셀 밀도에 초점을 맞춰 개발되었다. 메모리 셀의 구조적 단순성은 DRAM 기반 PIM을 위한 로직을 통합하는 데 흥미로운 아이디어를 불러 일으키지만 엄격한 물리적 제약이 있다. 많은 이전 연구에서 이 문제를 해결하기 위해 다양한 수준의 DRAM 아키텍처를 다룬다. [그림 7]과 같이 DRAM 기반 PIM을 논리 통합 수준에 따라 셀 수준, 뱅크 수준 및 3차원 수준의 세 가지 범주로 구분한다. DRAM 셀 레벨 PIM은 로우 레벨 트랜지스터를 통합한 구조, 메모리 뱅크의 전체 내부 대역폭을 활용하여 벌크 비트 연산을 수행하는 비트라인 감지 증폭기가 있는 논리 구조, 뱅크 수준 PIM은 각 뱅크의 열 디코더 뒤에 높은 수준의 처리 논리를 가지는 구조이다. 이 기술은 셀 수준 PIM에 비해 최대 내부 대역폭을 사용할 수 없지만 더 큰 논리 영역을 활용하여 DRAM 기반 PIM의 확장성을 가지며, 또한 3차원 레벨 PIM은 HMC(하이브리드 메모리 큐브)와 같은 기본 로직 다이와 함께 3차원 스택 메모리를 활용이 가능하다. TSV로 상호 연결된 스택형 메모리 다이에 컴퓨팅 로직 다이를 통합하여 이들 사이에 에너지 효율적인 고대역폭 통신을 가능하다. 3차원 적층 다이의 엄격한 물리적 및 타이밍 제약으로 인해 3차원 레벨 PIM의 실현은 여전히 어려운 문제가 있다[15].

## IV. ReRAM 기반 PIM 기술

본 장에서는 ReRAM의 배경과 ReRAM 기반 PIM 아키텍처에서 MAC 작업이 구현되는 방법을 소개한다. MAC 동작을 위한 ReRAM 기반 PIM에 대한 새로운 방향을 제시한다. ReRAM의 기초 ReRAM은 학계와 산업계에서 활발히 연구되고 있는 새로운 비휘발성 메모리의 일종이다. [그림 8]은 ReRAM 기반 PIM 로써 특히 읽기 및 쓰기 성능, 낮은 프로그래밍 전압, CMOS 제조 공정에 대한 확장성 및 호환성을 갖추고 있으며, 또한 이진 값을 저장하기 위해 고저항 상태(HRS) 또는 저저항 상태(LRS)로 프로그래밍할 수 있다. ReRAM의 아날로그 그적 특성 때문에 일부 작품에서는 메모리 용량을 향상시키기 위해 아날로그 저장을 위해 ReRAM을 사용했다[15].



\* ReRAM-based PIM architectures for MAC based on: (a) Currentmode sensing, (b) Voltage-mode sensing

\*\* Multiplication of 2-bit inputs and 2-bit weights in digital ReRAM based PIM: (a) One cycle and multiple macros, (b) Multiple cycles and macro

<자료> KIM et al., "OVERVIEW OF PROCESSING-IN-MEMORY CIRCUITS FOR ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING", IEEE Journal on emerging and selected topics in circuits and systems, Vol.12, No.2, June 2022, p.339.

[그림 8] ReRAM 기반 PIM

## V. 결론

인공지능 기술의 급격한 발전으로 인해서 데이터 사용이 폭발적으로 증가하였고 이에 따라서 기존의 폰노이만 방식의 컴퓨팅 구조에서는 데이터 처리에 대한 한계에 도달하게 되었다. 새로운 구조로 각광을 받고 있는 메모리 중심의 PIM 기술 구조가 새로운 대안으로 떠오르고 있다. 또한, 국내 메모리 업계에서 강점을 가진 기업인 SK하이닉스는 물론 삼성전자도 투자와 R&D를 통해 미래 시장을 위해서 PIM 기술 개발에 박차를 가하고 있다. 미래의 시장은 인공지능반도체를 중심으로 재편되면서 저전력과 메모리 대역폭 문제를 획기적으로 개선할 신기술의 필요성이 커지면서 솔루션으로 평가되고 있다. PIM 기반의 SRAM, DRAM, SRAM, ReRAM, eDRAM 등과 같은 다양한 솔루션에 대한 연구 개발을 통해서 인공지능반도체의 새로운 시장에 중심적인 역할이 기대된다. PIM 기반의 솔루션은 전송 과정에서 발생하는 지연시간과 전력 손실을 줄일 수 있다. 데이터를 보다 효율적으로 처리할 수 있고, 실시간으로 대용량 데이터를 처리하거나 데이터를 출력해야 하는 AI, 데이터센터, 고성능 컴퓨팅 등에 PIM 기술이 적용될 것으로 기대된다. 뉴로모픽 아키텍처는 초기에 연산과 저장이 분리되어 있고, 하나의 반도체가 모여서 연산을 수행하는 형태로 수행되었다. 기존 폰 노이만 아키텍처에서 메모리는 명령어와 연산자를 저장하고 프로세서는 메모리에서 명령어와 연산자를 가져와 연산을 수행한다. 그러나 PIM에서는 연산이 메모리에서 수행되는 아키텍처로, 신경망과 딥러닝이 이미지/음성 인식 등에 최적의 솔루션을 제공할 수 있으며 많은 연구가 진행되고 있다. 폰 노이만 구조의 한계를 넘어 프로세서와 메모리를 집적한 신개념 반도체 기술로 딥러닝에 최적화된 AI 반도체는 PIM 구조를 통해 반도체 성능과 전력 효율을 크게 향상시켜 저전력 소모로 고속 동작이 가능하다.

## ● 참고문헌

- [1] Gartner, "Semiconductor Revenue Forecast by Device Type Through 2025," 2021.
- [2] 박성모 외, "스파이킹 신경망 기반 뉴로모픽 기술 동향", TTA, 188호, 2020. 3, pp.28-33.
- [3] 박성모 외, "저전력 인공지능반도체 기술 동향", IITP, 2007호, 2021. 7, pp.6-10.
- [4] K. Ando et al., "BRein memory: A single-chip binary/ternary reconfigurable in-memory deep neural network accelerator achieving 1.4 TOPS at 0.6 W," IEEE J. Solid-State Circuits, Vol.53, No.4, 2018, pp.983-994,

- [5] M. Kang, S. K. Gonugondla, A. Patil and N. R. Shanbhag, "A multi-functional in-memory inference processor using a standard 6T SRAM array," IEEE J. Solid-State Circuits, Vol.53, No.2, 2018, pp.642-655.
- [6] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks," IEEE J. Solid-State Circuits, Vol.54, No.1, Jan. 2019, pp.217-230.
- [7] H. Valavi, P. J. Ramadge, E. Nestler and N. Verma, "A 64-tile 2.4-Mb in-memory-computing CNN accelerator employing charge-domain compute," IEEE J. Solid-State Circuits, Vol.54, No.6, 2019, pp.1789-1799.
- [8] S. Yin, Z. Jiang, J.-S. Seo and M. Seok, "XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks," IEEE J. Solid-State Circuits, Vol.55, No.6, 2020, pp.1733-1743.
- [9] S. K. Gonugondla, M. Kang and N. Shanbhag, "A 42 pJ/decision 3.12 TOPS/W robust in-memory machine learning classifier with on-chip training," in IEEE ISSCC Dig. Tech. Papers, 2018, pp.490-492.
- [10] J.-W. Su et al., "A 28 nm 64 Kb inference-training two-way transpose multibit 6T SRAM compute-in-memory macro for AI edge chips," in IEEE ISSCC Dig. Tech. Papers, Feb. 2020, pp.240-242.
- [11] X. Si et al., "A 28nm 64Kb 6T SRAM computing-in-memory macro with 8b MAC operation for AI edge chips," in IEEE ISSCC Dig. Tech. Papers, Feb. 2020, pp.246-248.
- [12] W.-S. Khwa et al., "A 65nm 4Kb algorithm-dependent computing-inmemory SRAM unit-macro with 2.3ns and 55.8 TOPS/W fully parallel product-sum operation for binary DNN edge processors," in IEEE ISSCC Dig. Tech. Papers, Feb. 2018, pp.496-498.
- [10] X. Si et al., "A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning," in IEEE ISSCC Dig Tech Papers, Feb. 2019, pp.396-398.
- [13] J.-M. Hung, C.-J. Jhang, P.-C. Wu, Y.-C. Chiu, and M.-F. Chang, "Challenges and trends of nonvolatile in-memory-computation circuits for AI edge devices," IEEE Open J. Solid-State Circuits Soc., Vol.1, 2021, pp.171-183.
- [14] P. Gu et al., "IPIM: Programmable in-memory image processing accelerator using near-bank architecture," in Proc. ACM/IEEE 47th Annu. Int. Symp. Comput. Archit(ISCA), May 2020, pp.804-817.
- [15] KIM et al., "OVERVIEW OF PROCESSING-IN-MEMORY CIRCUITS FOR ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING," IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS, Vol.12, No.2, June 2022, p.339.
- [16] Jian-Wei Su, et. al, "A 28nm 384kb 6T-SRAM Computation-in-Memory Macro with 8b Precision for AI Edge Chips," ISSCC2021, pp.251-252.

- [17] Ping-Chun Wu, et. al, "A 28nm 1Mb Time-Domain Computing-in-Memory 6T-SRAM Macro with a 6.6ns Latency, 1241GOPS and 37.01TOPS/W for 8b-MAC Operations for Edge-AI Devices," ISSCC2022. p.192.