Chapter 02

AI 반도체의 기술 장벽과 전망

- 에너지, 메모리, 패키징 -

문제현_한국전자통신연구원 책임연구원

거대언어모델(Large Language Model: LLM) 및 심층신경망(Deep Neural Network: DNN)을 필두로 한 인공지능(Artificial Intelligence: AI) 워크로드의 확산은 현대 반도체 시스템 설계의 패러다임을 전환시키고 있다. 본 고에서는 AI 반도체의 3대 기술적 과제인 메모리 벽(Memory Wall), 전력 및 열 제약, 패키징을 이슈를 살펴본다. 프로세서의 연산 성능과 메모리 대역폭 간의 불균형으로 인해 발생하는 데이터 이동 오버헤드 문제의 등에 관련된 메모리 벽을 진단하고, 이를 해결하기 위한 고대역폭 메모리(High Bandwidth Memory: HBM), 지능형 메모리 반도체(Processing-In-Memory: PIM), 모놀리식 3D(M3D) 집적 기술의 효용성을 검토한다. AI 가속기의 전력 소비 급증에 따른 데이터센터 차원의 열관리 설계 제약 사항을 고찰하며 이와 직접 연계되는 패키징 기술에 대해 고찰한다. 2.5D 인터포저 기반 이종 집적 기술을 통한 고밀도 구현 가능성과 더불어, 이로 인해 발생하는 열적 결합(Thermal Coupling) 및 제조 원가 상승 문제를 분석한다. 마지막으로 폰 노이만 아키텍처의 구조적 한계를 근본적으로 극복하기 위한 장기적 대안으로서 뉴로모픽 컴퓨팅 및 실리콘 포토닉스 인터커넥트 기술의 발전 방향을 제시한다.

I. 서론

인공지능(Artificial Intelligence: AI) 반도체의 기술적 과제와 설계 패러다임의 변화, AI 워크로드, 특히 거대언어모델(Large Language Model: LLM)과 심층신경망(Deep Neural Network: DNN)은 반도체 설계의 우선순위를 근본적으로 재편하고 있다. 전통적인 연산 중심 아키텍처와 달리, AI 가속기는 단순히 트랜지스터의 성능뿐만 아니라 메모리 대역폭(bandwidth), 전력 밀도, 집적 복잡도와 관련된 시스템 수준의 병목현상에 의해 제약을 받

* 본 내용은 문제현 책임연구원(☎ 042-860-5639, jmoon@etri.re.kr)에게 문의하시기 바랍니다.
** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.
***본 고는 정부(산업통상부)의 재원으로 한국산업기술기획평가원의 지원을 받아 수행되었음[연구과제명: 메타버스 구현을 위한 웨어러블 기기용 반도체 및 디스플레이의 특성평가 국제표준 개발(국가표준기술력향상-20025884)]

는다. 메모리 벽, 전력 및 열관리 그리고 첨단 패키징의 3가지 기술적 과제가 AI 반도체 로드맵의 핵심 요소로 부상하고 있다. 첫째, 메모리 벽은 AI 시스템에서 가장 지배적인 아키텍처적 제약 요인으로 나타나고 있다. 본래 윌리엄 울프(William A. Wulf)와 샐리 맥키(Sally A. McKee)가 정의한 이 개념은 프로세서 성능과 메모리 대역폭 및 레이턴시(Latency, 지연) 간의 격차 확대를 설명하기 위해 도입되었다[1]. GPU(Graphics Processing Unit) 및 영역 특화 프로세서(domain-specific processors)와 같은 현대의 AI 가속기에서 데이터 이동에 소모되는 에너지와 레이턴시는 산술 연산 비용보다 커지는 경우도 발생한다. 최근의 HBM (High Bandwidth Memory) 3급은 초당 테라바이트급 대역폭을 제공함으로써 이러한 제약을 일부 완화하고 있다. 그러나 대역폭의 확장 속도는 여전히 연산 성능의 확장 속도를 밑돌고 있으며, 이에 따라 메모리 아키텍처, 데이터 지역성 최적화 그리고 니어 메모리(Near-memory) 통합이 주요 연구 방향으로 자리 잡고 있다. 둘째, 전력 공급 및 열관리는 설계 시 최우선으로 고려해야 할 제약 조건이 되었다. 최첨단 AI 가속기는 장치 당 수백 와트의 전력을 소모하며, 시스템 수준의 랙(Rack) 단위에서는 수십 킬로와트를 초과하기도 한다. 트랜지스터 집적도가 높아짐에 따라 전력 밀도(W/cm^2) 또한 지속적으로 상승하고 있으며, 이는 온칩 전력 분배 네트워크와 패키지 수준의 전압 조정기에 상당한 부담을 가하고 있다. 동시에 2.5D 및 3D 통합 기술은 열방산 경로의 감소와 국소적 핫스팟(hot spot) 발생으로 인해 열관리 문제를 더욱 어렵게 하고 있다. 이에 따라 수랭식 냉각, 개선된 열 인터페이스 물질(Thermal Interface Material: TIM), 전력-열 공동 설계 아키텍처와 같은 고급 냉각 솔루션은 필수 고려 사항이 되었다. 이와 관련된 에너지 효율(Tera Operations Per Second per Watt: TOPS/W)은 자체 AI 가속기 성능만큼이나 중요한 지표로 자리 잡고 있다. 셋째, 소자 구성요소 미세화보다는 첨단 패키징 기술이 매우 중요한 요소로 자리잡고 있다. 이에 대응하는 기술로 이종 집적(heterogeneous integration)이 제시되고 있다. 2.5D 실리콘 인터포저, 실리콘 관통 전극((Through Silicon Via: TSV)을 이용한 3D 다이 적층 그리고 칩렛(chiplet) 기반의 모듈형 아키텍처는 로직과 메모리 간의 고밀도 인터커넥트를 가능하게 한다. TSMC의 CoWoS(Chip-on-Wafer-on-Substrate) 플랫폼과 UCIe 컨소시엄의 UCIe(Universal Chiplet Interconnect Express)와 같은 신규 다이 간 연결 표준은 시스템 수준의 스케일링으로의 전환을 잘 보여준다. 그러나 이러한 접근 방식은 수율 관리, 열 결합, 신호 무오류 및 비용 구조 측면에서 새로운 도전 과제를 제기한다. 요컨대 AI 반도체의 혁신은 더 이상

트랜지스터 밀도만으로 정의되지 않는다. 대신 성능 확장은 메모리 대역폭의 한계 극복, 극심한 전력 및 열 밀도 관리 그리고 이종 집적을 위한 첨단 패키징 활용 여부에 달려 있다. 이 세 가지 영역에서의 기술적 진보가 차세대 AI 컴퓨팅 시스템의 효율성, 확장성 및 경제적 타당성을 결정짓는 핵심 지표가 될 것이다.

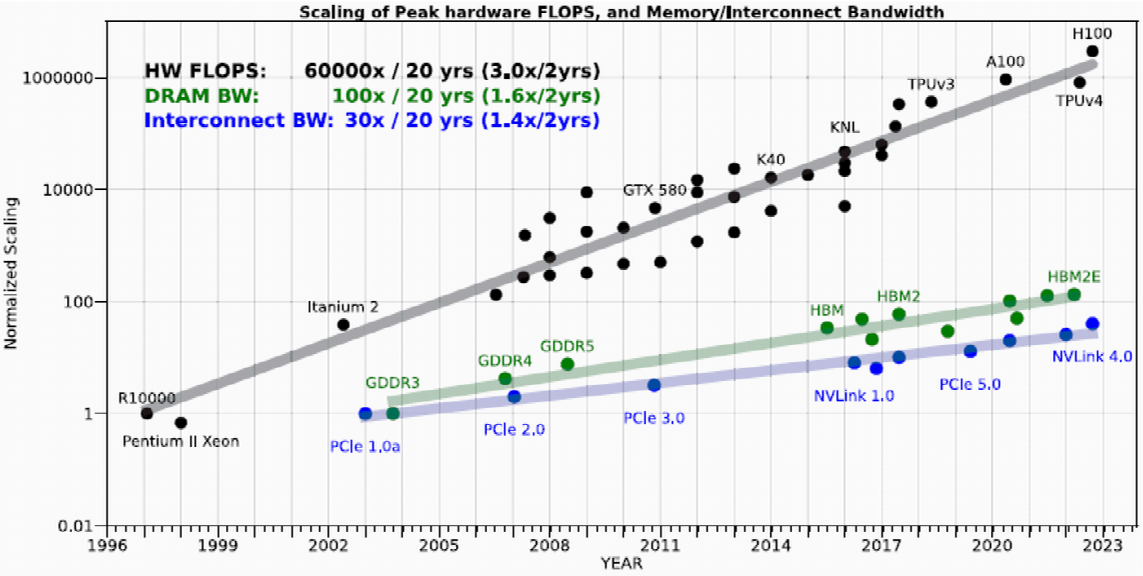
II. 메모리 병목현상

메모리 벽이란 프로세서의 연산 속도 증가율이 메모리 시스템의 데이터 공급 속도를 상회함에 따라 발생하는 성능 병목현상을 의미한다[2]. 이로 인해 연산부(로직)는 데이터 전송을 기다리는 동안 유휴 상태(Idle)에 머물게 된다. 메모리 벽은 본래 CPU와 메모리 간의 지연을 설명하기 위해 등장한 개념이었으나, AI 분야에서는 메모리 대역폭및 용량의 한계가 시스템 성능 확장성을 저해하는 포괄적인 제약 및 목적 성능 미달성 요인을 지칭한다.

LLM이나 DNN과 같은 AI 워크로드는 프로세서 코어와 메모리 간의 막대한 데이터 이동을 수반한다. 연산 능력과 메모리 처리량 사이의 불균형으로 인해 실제 실행 시간은 산술 연산이 아닌 메모리 액세스에 의해 결정되는 경우가 많다. AI 학습 및 추론 과정에서 요구되는 메모리 대역폭은 기존 DRAM(Dynamic Random Access Memory)의 성능 범위를 크게 벗어나 있으며, 이는 연산장치(예; GPU, NPU, TPU)가 데이터를 아주 빠르게 확보하지 못해 목적 성능이 충분히 발휘되지 못하는 병목현상으로 이어진다. 참고로 엔비디아 H100 AI 가속기가 LLM 워크로드를 효율적으로 수행하기 위해서는 2TB/s 수준의 매우 높은 메모리 대역폭이 필요하다.

메모리 병목현상은 연산부와 메모리부가 물리적 분리 및 인터커텍트의 폰 노이만(von Neumann) 구조에 기인하는 현상이다. 연산장치는 트랜지스터(예; MOSFET)의 스위칭 속도를 높이는 데 최적화되어 있지만 메모리(예; HBM)는 데이터 저장에 관련된 정전 용량부에 최적화되어 있다. 이 때문에 두 요소의 대역폭은 차이가 날 수밖에 없는 근본적 한계가 있다. 또한, 연산부와 메모리를 연결하는 인터커넥트에서도 지연이 발생한다.

이 병목현상은 메모리 기술의 진보에도 불구하고 오늘날까지 단점으로 작용하고 있다. [그림 1]을 보면 FLOPS(Floating Point Operations Per Second, 연산 속도)의 증대보다 대역폭의 증대가 완만함을 볼 수 있다. 구체적으로 지난 20년 간 서버급 AI 연산장치의 최대

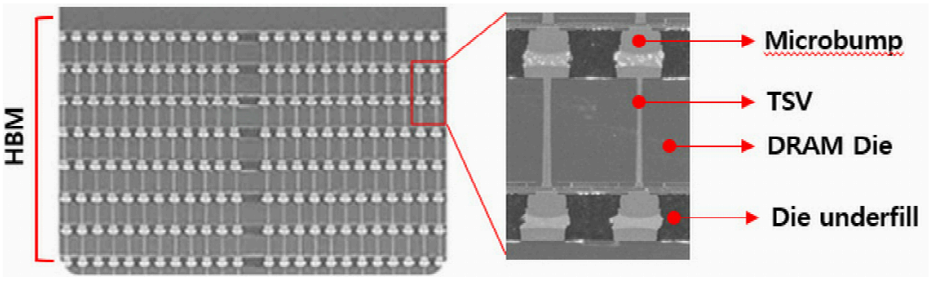


〈자료〉 Gholami, "AI and Memory Wall", IEEE Micro, 44, 2024.

[그림 1] 연도별 FLOPS(연산 속도) 및 메모리/인터커넥트의 대역폭의 발전 추이

연산 성능(peak compute)은 약 6만 배 증가한 반면, DRAM 대역폭은 100배, 인터커넥트 대역폭은 30배 증가하는 데 그쳤다. 메모리 및 인터커넥트 대역폭을 확장하는 데 수반되는 근본적인 기술적 난제들은 폰노이만 구조와 스위치/메모리의 상이점이 해결되지 않은 이상 어려운 기술적 장벽으로 남아 있을 수밖에 없다.

지연을 개선하기 위해서는 여러 가지 기술들이 활용 및 제안되고 있다. 첫째는 DRAM을 적층한 HBM(고대역폭 메모리) 있다. 이 방식은 이미 상용화가 되었고 계속해서 기술적으로 진화된 제품들이 발표되고 있다. 현재 상용화된 HBM3E는 12단 적층으로 최대 전송속도 9.6Gbps 및 메모리 용량 32GB를 제공하고 있다. 삼성전자는 2026년 2월 초에 11.7Gbps 전송속도 및 최대 대역폭 3.3TB/s를 갖는 HBM4를 성공적으로 개발하여 양산에 개시하였다[2]. 최근의 HBM 사양 개선은 적층 메모리를 연결해 주는 TSV(Through Silicon Via, 실리콘 관통 전극) 고종횡비 보쉬(bosch) 식각기술 및 1,024비트 이상의 인터페이스 기술에 힘입은 바가 크다. [그림 2]는 DRAM이 TSV로 연결된 HBM의 예를 보여준다. DRAM와 DRAM 사이는 일종의 접착제인 언더필(underfill)로 충전되어 있다. HBM의 제작 핵심은 TSV 공정과 DRAM 접합 공정인 언더필 공정이다. 최근에는 열 방출에 유리한 매스 리플로우(mass reflow) 공정이 언더필에 적용되고 있다.



〈자료〉 Choe, "Memory Technology Update: Roadmap and Insights on HBM2E and 3D DRAM", Tech Insights, 04, 2020.

[그림 2] HBM의 적층구조

병목현상을 해결하기 위한 또 다른 방법으로는 프로세싱 인 메모리(Processing-in-Memory: PIM) 방식이 있다[4][5]. 이 방식에서는 연산기능과 메모리 작용이 하나에 다이에 집적화되어 있다. 이 방식에서는 폰 노이만 구조에서 야기되는 지연을 해결할 수 있을 것으로 기대된다. 하지만 PIM은 로직 소자와 메모리 소자의 제작 공정에 수반에 되는 열에 큰 차이가 있어 제작 공정이 쉽지 않다. 이를 개선하기 위해 연산부와 메모리부가 3D로 적층하는 구조가 제안되고 있다. 이 구조에서는 고온 공정이 수반되는 로직이 먼저 제작된다. 저온 BEOL 공정이 확립되는 모노리식 3D 집적화는 기술적으로 가능하기 때문에 IEEE IRDS에서는 이를 중요한 기술 이정표로 취급하고 있다[6]. 3D 집적화는 인터커넥트에서 발생하는 지연을 줄이는 기술로 이해될 수도 있다[7]-[9]. 인터커넥트의 금속 배선은 열과 신호교란의 단점이 있다. 이 때문에 이를 광 방식 실리콘 포토닉스로 대체하려는 기술적 시도가 있다. 금속 배선에 비해 광 전송은 열 장애가 없으며 거리에 제한이 없다는 장점이 있다. 현재는 충분한 데이터 전송률을 확보하기 위해 폰 노이만 구조에서 메모리와 연산장치와 물리적으로 최대한 근접 배치해 대역폭을 개선하여 지연을 줄이는 접근법이 활용되고 있다[표 1].

[표 1] 메모리 병목을 해결하기 위한 기술 예

기술명	병목 해결 요소	기술적 도전
HBM (High Bandwidth Memory)	고대역(TB/s) 제공	DRAM 자체의 한계, 고종횡비 TSV 확보를 위한 식각기술
PIM(Processing-in-Memory)	연산과 메모리 동일 구역 작업 수행	로직과 메모리의 열적 공정호환성 부조화
Monolithic 3D Integration	연산부과 메모리부의 인터커넥트 물리적 거리 최소화	상부(메모리부)의 저온 공정 확립

〈자료〉 Gholami, "AI and Memory Wall", IEEE Micro, 44, 2024.

[표 2] 메모리 병목

기술명	HBM	PIM	3D Integration
대역폭	1.5~2TB/s(HBM4)	5TB/s	〉 10TB/s(추정치)
메모리 지연	~200ns	~50ns	〈 10ns(추정치)
인터커넥트길이 및 밀도	~5mm, ~10^3/mm^2	~50μm, ~10^4/mm^2	~50nm, ~10^6/mm^2
기술성숙도	상용화	프로토타입 시제품	연구개발 단계

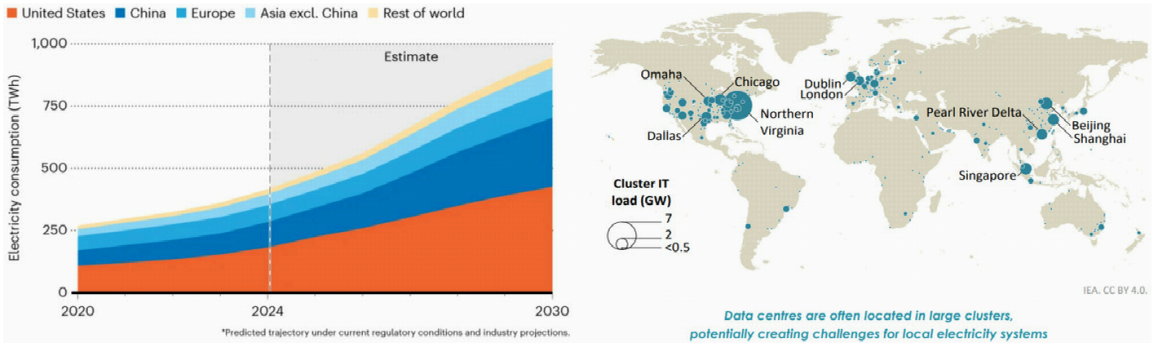
〈자료〉 Asifuzzaman, "A survey on processing-in-memory techniques: Advances and challenges", Memories - Materials, Devices, Circuits and Systems, 4, 2023.

[표 2]에는 [표 1]에 관련된 정량적 수치를 정리하였다. 수치들에서 볼 수 있듯이 인터커넥트 길이를 줄이고 밀도를 높이면 대역폭 증대와 지연시간 저감을 달성할 수 있다. 그렇지만 여기에는 반드시 열관리 기술이 동반되어야 하므로 공정 개선만으로는 쉽게 메모리 병목을 쉽게 해소할 수 없다.

III. 전력 소모 및 발열관리

AI 칩은 역사상 가장 높은 전력을 소모하는 컴퓨팅 장치 중 하나이다. GPU 및 가속기의 높은 연산 밀도는 데이터센터의 전력 인프라(PDU, UPS)뿐만 아니라 칩 수준의 전력 분배 네트워크(Packet Data Network: PDN)에도 상당한 부담을 준다. 수십 테라플롭스(TFLOPs) 이상의 성능을 내는 AI 반도체들은 장치당 수백 와트의 전력을 요구하며, 이는 시스템 에너지 비용 상승과 확장성 저해의 원인이 된다.

[그림 3]은 전세계 AI데이터센터의 지역별 전력 소모와 위치를 보여준다. AI데이터센서국제에너지기구(IEA)의 모델 분석에 따르면, AI데이터센터의 전력 소비량은 2030년에 945 TWh에 달할 것으로 전망되며, 이는 현재 일본의 연간 총 전력 소비량과 맞먹는 엄청난 수준이다[10]. 참고로 2024년 기준 데이터센터 소비 전력은 전 세계 총 전력 소비량의 약 1.5%에 해당하는 415 TWh를 기록한 바 있다. 즉, 2030년의 전력 소모량은 2024년의 소모량의 2배 이상이 되리라 예측된다. 전세계 AI데이터센터의 전력량의 약 85%는 미국, 중국, 유럽 등에서 소모되고 있다. 이는 AI데이터센터의 지역적 불균형을 반영하는 수치로 해석될 수 있으며 장래 AI 관련 분야의 중요성을 고려하면 국력에 관련되는 큰 함의가 있다. 막대한 전력 소모는 기후변화와도 연동이 되기 때문에 AI 지향 경제 발전과 환경보존 사이에 갈등의



〈자료〉 International Energy Agency, “Energy and AI”, IEA 04, 2025.

[그림 3] AI데이터센터의 지역별 전력 소모량과 위치

요소가 있다.

AI데이터센터의 전력 소모는 크게 가속기 운용, 정보 저장, 네트워킹 등과 같이 직접 AI 컴퓨팅에 관련된 부분과 냉각 및 시설 유지에 관련된 인프라로 크게 구분할 수 있다[11]. [표 3]은 전력 소모에 관련된 정량적 수치를 정리하였다. 전력 소모 비율에서 볼 수 있듯이 냉각에 소요되는 전력 소모가 상당히 높음을 알 수 있다. AI 컴퓨팅의 성능 개선을 위한 기술적 요구가 지속적으로 발생하고 있는 상황에 대응하여 반도체 제조업체들은 FLOPs, 대역폭 및 저장용량이 개선된 제품을 꾸준히 발표하고 있으며 운용을 위한 소모 전력도 증가하고 있다. 반도체 소자 복잡도 증가에 따라 열 발생이 비례적으로 증가하기 때문에 AI 데이터센터 냉각에 필요한 소모 전력 증대도 동반된다.

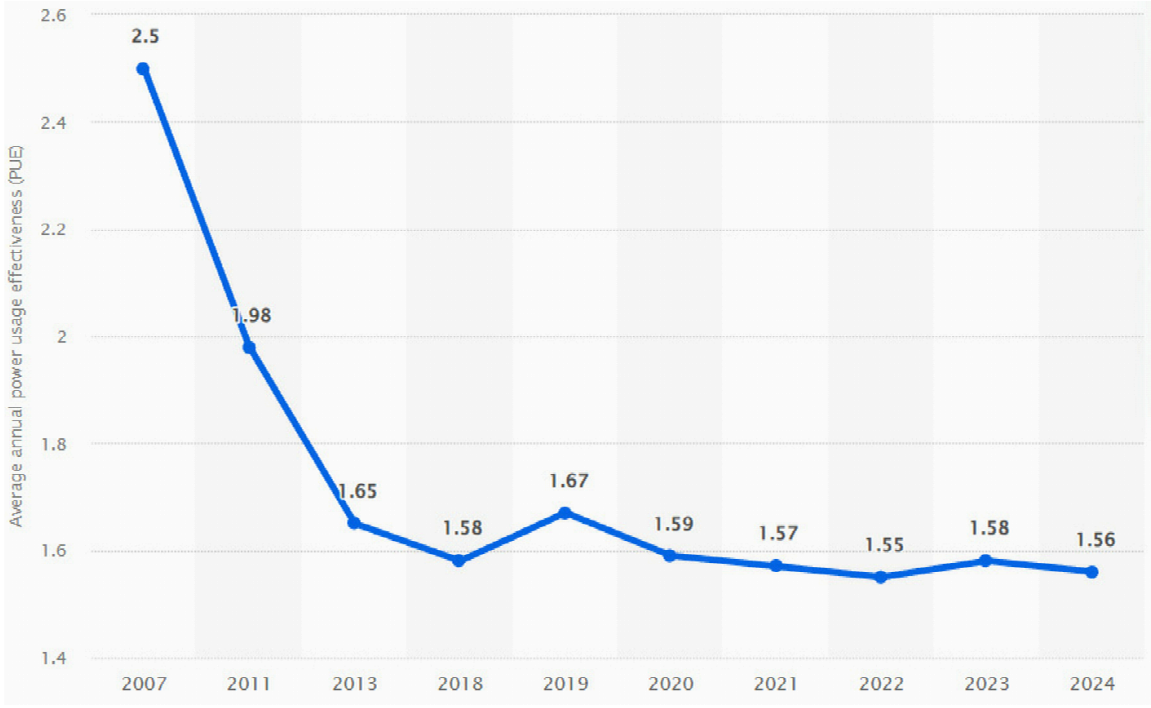
AI데이터센터의 전력 소모 중에서 냉각시스템을 평가하기 위한 지표로 전력효율지수(Power Usage Effectiveness: PUE)가 널리 사용되고 있다[12]. PUE는(시설의 전체 사용 전력)/(컴퓨팅부 사용 전력)으로 정의된다. PUE가 1.0에 근접할수록 컴퓨팅에 소요되는

[표 3] AI데이터센터 운용/유지에 소요되는 작업별 전력 소모 비율

작업 영역	업무	대표 예	전력 소모 비율
컴퓨팅	연산, 학습	GPU, AI 가속기	~50%
	정보저장	HBM, SSD 저장	~13%
	네트워킹	인트라 라우팅, 인터넷 전송	~5%
인프라	냉각	액상냉각, 공냉, 칠러 유지	~30%
	기타	외부파워서플라이 조명, 보안,	〈 2%

〈자료〉 International Energy Agency, “Energy and AI”, IEA 04, 2025.

전력이 높다. PUE가 2.0 보다 큰 경우 컴퓨팅이 아닌 인프라 부분에 전력 소모가 높은 경우로 비효율적인 경우이다. AI데이터센터의 PUE를 개선하는 가장 효과적인 방법은 냉각 소요 전력량 저감이다. 냉각으로 크게 액상냉각법을 활용하는 외적인 방법과 AI 반도체의 소자, 소재 및 구조를 변경하는 내적인 방법으로 구분할 수 있다[13]. 구글, 마이크로소프트(Azure), 메타, 아마존(AWS) 등에서는 최고 성능으로 PUE ~1.1 수준의 값을 발표하고 있다. PUE 1.1 수준은 현재 기술적인 한계로 인식되고 있다. [그림 4]는 연도별 AI데이터센터 PUE의 변화 양상이다. 2020년 이후로는 PUE(~1.5)에 큰 변화가 없다는 사실을 알 수 있다. 이는 AI 반도체/시스템의 성능을 희생시키면서 전력 소모를 낮추기 어려운 이유와 냉각 방식의 한계 때문이다. 고밀도 집적은 국소적인 열 발생을 심화시키며, 이를 적절히 방출하지 못할 때 성능 저하나 소자 손상을 초래할 수 있다. 기존의 공랭식이나 기본적인 수랭식 냉각 방식으로는 3D 적층 구조 및 고전력 칩의 발열을 제어하기에 역부족이다. 열관리의 핵심은 반도체 소자 분야와 인프라 분야가 협업으로 온도 발생 및 방열 구조 개발이 더욱 긴급한 시점에 와 있다.



〈자료〉 Taylor, "Global data center average annual power usage effectiveness(PUE)", Statista. 11, 2026.

[그림 4] 연도별 AI데이터센터의 PUE 변화 양상

IV. AI 반도체 패키징

거대규모의 AI데이터센터가 등장하기 이전에는 반도체 공정은 주로 극미세 리소그래피(5nm) 패터닝, 고집적화 및 성능 개선에 초점이 맞춰져 있었다. 소자 크기가 작아질수록 소자 특성 변동성, 누설전류를 조절하기 위한 복잡도가 높은 공정들이 제안되고 시도되었다. AI데이터센터의 본격적 등장 이후(2020년 이후)에는 미세공정보다는 열관리, 전력 소모 및 소자 재현성 및 수율에 관여된 기술적 장벽이 중요하게 취급되었다. 집적화 단계인 패키징에서는 연산부와 메모리부를 하나의 시스템으로 통합함과 동시에 열관리, 전력 공급 등이 원활히 해야 하는 통합적인 핵심 기술의 개발이 필요하게 되었다.

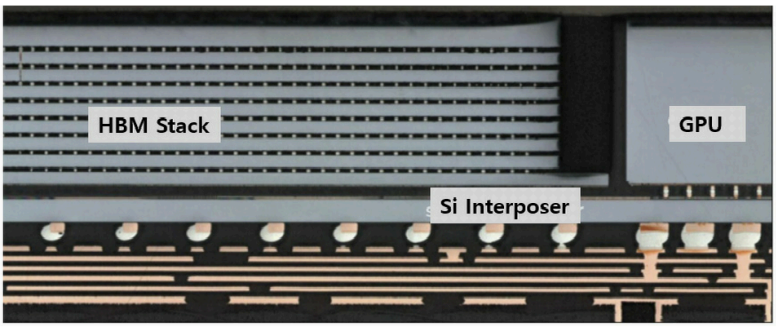
패키징 이슈에서 가장 중요한 이슈는 열관리이다. 전 절에서는 다소 거시적인 관점에서 전력 소모 및 열을 냉각 관점에서 살펴보았다. 여기서는 발열처는 칩 본체에 대해 살펴본다. 칩 수준에서 발열처는 AI 반도체 구성 요소인 GPU와 HBM, 인터커넥트이다. 전기적인 관점에서 구성 요소 발열은 GPU의 동적 스위칭, DRAM의 활성화 및 리프레쉬(Refresh), 인터커넥스와 TSV 배선 내 저항열, 정전용량 성분 발열에 기인한다. 연산부 GPU에서 발생하는 열이 ~85% 이상이며, HBM 부에서는 ~10% 인터커넥트에서는 ~5%의 등으로 열 발생이 AI 반도체 구성된다[표 4].

GPU, HBM 및 인터커넥트 구조 내부에서 발생한 열은 마이크로범프, 인터포저, 열 인터페이스 물질, 히트 스프레더 및 냉각 시스템을 포함하는 복잡한 열적 적층 구조를 통해 전도된다. 이러한 계층 중 열 인터페이스 물질은 접합부에서 냉각수에 이르는 전체 열 저항에서 가장 지배적인 비중을 차지하는 경우가 많다. 따라서 차세대 AI 반도체 시스템 설계에 있어 첨단 냉각 기술, 개선된 열 인터페이스 물질 그리고 최적화된 패키지 아키텍처를 통한 열 발산의 설계/제조 기술은 패키징의 핵심으로 대두하고 있다. 이를 고려한 구조로 2.5D 아키

[표 4] AI 반도체 내 발열부 및 발열 원인 비율

요소	열 발생(W/cm^2)	피크온도(℃)	열원
GPU 로직다이	80~150	〉 100	스위칭 및 누설
HBM	10~20(적층 단위당)	~80	DRAM 활성화 및 리프레쉬
인터커넥트	5~15	~60	저항열 및 스위칭

〈자료〉 Chen, "Breaking Thermal Bottleneck in 3D HBM-on-GPU Integration via System-Technology Co-Optimization", IEEE IEDM, 1, 2025.



〈자료〉 Yole Group, “High-End Performance Packaging”, SystemPlus Consulting, 05, 2023.

[그림 5] NVIDIA 2.5D 아키텍처의 단면도

텍처가 제안되고 있다[13].

[그림 5]는 실제 엔비디아의 2.5D 아키텍처의 단면도 사진이다. 이는 기존의 2D보다는 집적도가 높고 3D 집적화보다는 공정이 용이한 장점이 있다. 이 아키텍처에서는 실리콘 인터포저 상에 GPU와 HBM가 동일 평면상에 고밀도로 배치되어 전력 소모와 열 발생이 줄어든다. 이의 2.5D를 구현하는 공정은 통상적으로 CoWoS이라 일컫는다.

최근 들어, AI 반도체의 패키징은 전 공정으로 자리를 잡고 있을 정도로 중요도가 높다. 패키징은 3D 이종 집적(heterogeneous integration), 칩렛 기반 아키텍처 및 첨단 냉각 기술이 융합되어야 열 관리에 용이하며, 여기에 상호 연결된 지연 등의 장벽을 극복할 수 있다.

V. 대안 기술

연산(GPU 로직 다이)과 메모리(HBM)를 물리적으로 분리된 폰 노이만 구조에서는 지연, 대역폭 제한 및 열 발생 등의 성능저하 요인들이 자리잡고 있다. 이를 위한 해결책으로 생물모사의 뉴로모르픽 아키텍처가 활발히 연구되고 있다[15]. 이 아키텍처에서는 연산과 저장이 물리적으로 같은 곳에 있기 때문에 지연 및 대역폭의 장애를 없앨 수 있다. 또한, 작동 방식이 이벤트 기반(event driven) 기반한 스파이크형이므로 전력 소모 또한 현저히 줄일 수 있으며 병렬컴퓨팅 수행이 가능한 큰 장점도 있다. 아직 뉴로모르픽의 소자/소재는 확정되지 않았고 페리퍼럴 회로 필요성 및 학습 훈련이 용이하지 않다는 부분들이 뉴로모르픽 기반 AI의 기술적

도전으로 남아 있다.

실리콘 인터포저를 활용하는 2.5D 아키텍처에서는 배선을 통해 연산과 메모리가 연결되어 있다. 이 때문에 폰 노이만의 단점들은 여전히 상존한다. 전기적 인터커넥트를 광학적 방식으로 대체하는 방식이 실리콘 포토닉스이다[15]. 실리콘 포토닉스에서는 저항 및 정전용량 요소가 부재하므로 비트 당 소모되는 전력을 현저히 줄일 수 있으며, 다이 수준 물리적 크기에서는 신호 저하(signal degradation)의 무시할 만한 수준이다. 또한, 이론적으로 링크 당 ~200Gb/s의 전송이 가능하기 때문에 데이터 전송이 배선보다 훨씬 용이하다. CMOS/MOSFET 공정에 실리콘 포토닉스의 요소 및 레이저 광원의 집적화는 여전히 기술적 도전으로 남아 있기에, 이 분야는 추가적인 연구 개발이 필요하지만 유망한 기술임은 확실하다.

Ⅵ. 맺음말

차세대 AI 반도체 시스템이 직면한 핵심 기술적 과제를 고찰하고, 이러한 한계를 극복하기 위한 잠재적 기술 경로를 분석하였다. 첫째, 메모리 벽은 AI 컴퓨팅 시스템에서 어려운 아키텍처 제약으로 남아 있다. 지난 20년 간 현대 가속기의 연산 처리량은 비약적으로 증가한 반면, 메모리 대역폭 및 인터커넥트 성능의 개선 속도는 이에 크게 미치지 못했다. 그 결과, AI 워크로드에서는 산술 연산보다 데이터 이동이 실행 시간을 지배하는 현상이 빈번하게 발생하고 있다. 이러한 병목현상을 완화하기 위해 메모리 대역폭 확장, 지연시간 단축 그리고 연산-메모리 소자 간의 물리적 거리 축소를 목표로 하는 HBM, PIM 아키텍처, 모놀리식 3D 집적 기술 등이 제안되었다. 둘째, 전력 소모 및 열관리는 AI 반도체 설계의 핵심 현안이 되었다. 현대 AI 가속기는 수백 와트 이상의 전력 수준에서 구동되는 경우가 많으며, 대규모 AI데이터센터는 향후 10년 내에 전 세계 전력 소비의 상당 부분을 차지할 것으로 전망된다. 데이터센터 에너지 소비의 상당 부분이 냉각 시스템에 투입됨에 따라, 와트 당 연산 성능(TOPS/W) 및 PUE와 같은 지표가 핵심 설계 고려 사항으로 부상하였다. 이에 따라 액체 냉각 및 개선된 열 인터페이스 물질을 포함한 첨단 냉각 기술이 고성능 AI 시스템의 필수 요소가 되고 있다. 셋째, 첨단 패키징 기술은 로직과 메모리 구성 요소의 이종 집적(heterogeneous integration)을 실현하는 데 중추적인 역할을 수행한다. 2.5D 실리콘 인터포저 플랫폼과 같은 아키텍처는 GPU와 HBM 스택을 근접 배치함으로써 기존 시스템 아키텍처 대비 대역폭

향상과 지연시간 단축을 가능하게 한다. 그러나 이러한 솔루션은 열적 결합(thermal coupling), 제조 수율, 비용 효율성 측면에서 새로운 도전 과제를 야기한다.

향후, 새로운 컴퓨팅 패러다임이 기존 아키텍처의 한계에 대한 근본적인 해결책을 제시할 수 있을 것으로 기대된다. 뉴로모픽 컴퓨팅은 아키텍처 내에 메모리와 연산을 통합함으로써 에너지 효율과 병렬 처리 능력을 획기적으로 개선할 가능성이 있다. 실리콘 포토닉스 인터커넥트는 칩 내부 및 칩 간의 초고속 광학 데이터 전송을 가능하게 함으로써 전기적 인터커넥트의 대역폭 및 전력 한계를 극복할 수 있는 기술력 제시되고 있다.

결론적으로, AI 반도체 기술의 지속적인 발전은 소자 물리, 메모리 아키텍처, 시스템 설계, 패키징 기술 그리고 에너지 관리를 아우르는 다학제적 접근이 필요하다. 이러한 기술적 과제들을 해결하는 것은 미래 AI 컴퓨팅 인프라의 성장과 확장성을 유지하는 데 필수적인 요건이 될 것이다.

참고문헌

[1] W. Wulf and S. McKee. “Hitting the memory wall: Implications of the obvious”, ACM SIGARCH Computer Architecture News 23, 1995, pp.20-24.

[2] A. Gholami, Z. Yao, S. Kim, C. Hooper, M. W. Mahoney and K. Keutzer, “AI and Memory Wall”, IEEE Micro, 44, 2024, pp.33-39.

[3] Samsung Electronics, “Samsung Ships Industry-First Commercial HBM4 With Ultimate Performance for AI Computing”, Samsung Global Newsroom, Feb. 2026

[4] S. Lee et al., “Hardware Architecture and Software Stack for PIM Based on Commercial DRAM Technology: Industrial Product”, 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture(ISCA), Valencia, Spain, 2021, pp.43-56

[5] K. Asifuzzaman et al., “A Survey on the Expanding Scope and Interdisciplinary Opportunities for Processing-in-Memory Techniques”, in IEEE Access, Vol.14, 2026, pp.18408-18430.

[6] THE INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS: 2024 IEEE.

[7] M. M. Shulaker et al., “Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs”, 2014 IEEE International Electron Devices Meeting, San Francisco, CA, USA, 2014, pp.27.4.1-27.4.4,

[8] H. Amrouch, N. Du, A. Gebregiorgis, S. Hamdioui and I. Polian, “Towards Reliable In-Memory Computing:From Emerging Devices to Post-von-Neumann Architectures”, 2021 IFIP/IEEE 29th International Conference on Very Large Scale Integration(VLSI-SoC), Singapore, Singapore, 2021, pp.1-6

[9] J. Wu, F. Mo, T. Saraya, T. Hiramoto and M. Kobayashi, “A Monolithic 3-D Integration of RRAM Array and Oxide Semiconductor FET for In-Memory Computing in 3-D Neural Network”, in

IEEE Transactions on Electron Devices, Vol.67, No.12, Dec. 2020, pp.5322-5328.
[10] International Energy Agency(IEA), Energy and AI, IEA Report, Paris, 2025.
[11] Y. Jia, "Analysis of the Impact of Artificial Intelligence on Electricity Consumption", 2024 3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology(AIoTC), Wuhan, China, 2024, pp.57-60.
[12] R. C. Zoie, R. Delia Mihaela and S. Alexandru, "An analysis of the power usage effectiveness metric in data centers", 2017 5th International Symposium on Electrical and Electronics Engineering(ISEEE), Galati, Romania, 2017, pp.1-6.
[13] R. Mahajan, Chia-pin Chiu and G. Chrysler, "Cooling a Microprocessor Chip", in Proceedings of the IEEE, Vol.94, No.8, Aug. 2006, pp.1476-1486.
[14] J. Kim et al., "Architecture, Chip, and Package Co-design Flow for 2.5D IC Design Enabling Heterogeneous IP Reuse", 2019 56th ACM/IEEE Design Automation Conference(DAC), Las Vegas, NV, USA, 2019, pp.1-6.
[15] J. Moon, K. Kim, J.Kim, S. Park, Y. Yi, J. Woo, S.-Y. Kang, "Nonplanar Atomic Layer Deposition (ALD)-Niobium Oxide(NbOx) Neurons for Oscillatory Neural Network Applications", Adv. Intell. Syst., 2026, e202501015.
[16] H. Hsia et al., "Integrated optical interconnect systems(iOIS) for silicon photonics applications in HPC", in Proc. Electron. Components Technol. Conf.,(Orlando, FL, USA), May. 2023, pp.612-616.