주요 동향(2) : ICT

1 에이전트 시대 대응을 위한 '차세대 AI 메모리' 기술 경쟁 격화

에이전트 AI 급부상, 메모리가 핵심 인프라 병목으로

- AI 서비스, 응답 생성에서 자율 수행으로 진화
  - '25년 이후 AI 서비스의 작동 방식이 단발형 응답 생성에서 목표 기반 자율 수행 구조로 전환되는 흐름이 본격화, 주요 빅테크도 에이전트 AI 제품을 잇따라 출시하며 이 변화를 가속
  - IDC는 '29년까지 전 세계 운영 중인 AI 에이전트 수가 10억 개를 넘어설 것으로 전망했으며, 일일 수행 액션도 2,170억 건 이상으로 확대돼 추론 호출과 데이터 접근 수요 증가가 예상
  - 이에 따라 단발형 응답 생성 중심의 기존 추론 인프라만으로는 반복 호출·장기 맥락 유지·중간 결과 재사용에 따른 메모리 부하 대응이 어려워지며, AI 메모리를 포함한 인프라 전반의 기술 경쟁도 가속
- AI 메모리 수요 압력, 가격·공급·기술 포트폴리오 재편으로 가시화
  - HBM 가격 급등과 공급 부족이 지속되는 가운데, 에이전트 AI 확산이 AI 메모리 수요를 키우는 주요 요인으로 거론되며 메모리 시장의 가격·공급 압력이 확대
  - 이러한 수요 압력은 HBM 증설 경쟁에 그치지 않고, 고속 추론·메모리 확장·장기 기억·데이터 이동 최소화 등 사용 목적별 수요에 대응하는 기술 개발·투자 경쟁으로 확산
  - 이에 따라 HBM은 고속 추론, CXL은 메모리 확장·공유, 고성능 SSD·HBF는 장기 저장, PIM은 데이터 이동 최소화를 담당하는 기능 분담형 메모리 포트폴리오가 주요 기업을 중심으로 구체화

에이전트 AI가 AI 메모리 경쟁 기준을 바꾸는 방식

(1) AI 서비스 구조는 어떻게 달라지고 있는가?

- 계획·검색·실행·검증 반복하는 자율 수행 루프 구조로 전환
  - 에이전트 AI는 계획·검색·도구 실행·결과 검증을 반복하며 이전 대화·작업 이력·검색 결과·코드 실행 결과를 단계마다 참조하는 구조로, 단일 요청에서도 반복적인 모델 호출과 메모리 접근이 발생

- 에이전트 서비스는 사용자 맥락이 쌓일수록 정확도와 개인화 수준이 높아지는 자기강화 구조(컨텍스트 플라이휠)를 형성하며, 맥락 관리 역량이 핵심 경쟁 요소로 작동

〈 에이전트 AI 루프 단계별 메모리 접근 구조 〉

처리 단계	수행 내용	메모리 접근 유형	주요 메모리 계층
① 계획 수립	목표 분해·태스크 설계	컨텍스트 로드·KV 캐시 읽기	HBM·DRAM·CXL
② 정보 검색	벡터DB·외부 문서 호출	임베딩 벡터 검색	SSD
③ 도구 실행	코드·API 실행·외부 서비스 연동	중간 결과 저장·공유	CXL 메모리 풀
④ 결과 검증	출력 품질 평가·재실행 판단	KV 캐시 재참조	HBM·CXL
⑤ 이력 저장	작업 결과 장기 보존	대용량 영구 저장	SSD·외부 스토리지

자료 : SemiAnalysis; Vik's Newsletter('26.04); SK하이닉스('26.01) 자료 재구성

(2) AI 서비스 구조는 왜 메모리 수요를 바꾸는가?

● 반복 호출·맥락 누적이 AI 메모리 부하를 구조적으로 심화

- 장문맥 처리·멀티턴 대화·멀티에이전트 협업이 확산될수록, 하나의 요청에서도 이전 대화·작업 이력·검색 결과·중간 산출물을 함께 참조하며 데이터량과 메모리 재사용 부담이 동시 증가
- 특히 어텐션·KV 캐시 처리는 AI 연산 중 메모리 의존도가 가장 높은 영역으로, 에이전트 AI 확산에 따라 관련 연산 빈도와 메모리 접근 부하가 구조적으로 증가

● 단계별 데이터 재사용·다층 메모리 운용 구조로 메모리 운영 방식 전환

- 기존 추론에서는 메모리가 단일 요청 처리 시점에만 활용됐으나, 에이전트 AI는 작업 진행 전체에 걸쳐 단계별 데이터를 보관·호출하는 방식으로 메모리 운영 단위 확장
- 에이전트 AI는 계획·검색·실행·검증 과정에서 생성되는 중간 결과를 계속 이어 쓰기 때문에, 메모리도 단순 임시 저장 공간에서 작업 흐름을 유지하는 실행 기반으로 변화
- 결국 메모리는 하나의 응답을 만들기 위한 보조 자원을 넘어, 에이전트가 이전 맥락을 유지하며 다음 행동으로 이어가도록 뒷받침하는 핵심 인프라로 확대

(3) 차세대 메모리 경쟁의 평가 기준은 무엇인가?

● GPU 연산 중심 경쟁에서 대역폭·용량 등 목적별 역할 분담 체계 경쟁으로 전환

- 에이전트 AI 확산으로 AI 인프라 경쟁 기준이 GPU 연산 성능 중심에서 메모리 대역폭·용량·지연시간·전력효율·데이터 이동 비용 등 복합 평가 기준의 중요성 상승

- 기술 전문 미디어 SiliconANGLE은 AI 서버 BOM에서 메모리 비중이 65~70%에 달한다고 분석하며, AI 인프라의 무게중심이 연산 장치에서 메모리 체계로 이동하고 있다고 평가 ('26.2)
- 하나의 메모리 기술만으로 속도·용량·전력·비용 문제를 모두 해결하기 어려워지면서, 메모리 경쟁력도 개별 제품 성능뿐 아니라 다양한 메모리를 역할에 맞게 조합하는 설계 역량으로 확대

HBM부터 뉴로모픽까지, 에이전트 AI가 바꾸는 메모리 기술 경쟁

AI 서버 메모리, 4개 계층 배치 기반 분담 체계로 에이전트 AI 메모리 수요 대응

- AI 서버 메모리는 가속기와의 거리·속도·용량·휘발성을 기준으로 가속기 인접, 메모리 확장·공유, 장기 저장, 메모리 내 연산 등 4개 영역으로 구분

〈 AI 컴퓨팅 인프라의 메모리 계층 구조 〉

물리적 위치	계층	정의
GPU 내부	Register	AI 병렬 연산 장치
	SRAM	GPU 온칩 초고속 캐시
GPU 패키지 인접	HBM	패키지 적층형 고대역폭 DRAM
	HBF	NAND 기반 고대역폭·대용량 비휘발성 보조 메모리
서버 메인 메모리	DRAM	휘발성 시스템 메모리
서버 내·서버 간 메모리 풀	CXL 메모리 풀	CXL 기반 확장·공유 메모리 자원
AI 최적화 스토리지	SSD	NAND 기반 비휘발성 저장장치
데이터센터 전체	메모리·스토리지 풀	데이터센터 단위 메모리·스토리지 통합 인프라

자료 : 참고 자료 종합

  - (가속기 인접 계층) HBM·HBF가 GPU 인터포저 위에 초근접 배치되는 on-package 구조, 가속기 옆에서 직접 데이터를 공급하며 고속 추론 병목 완화 담당
  - (공유 계층) CPU 직접 연결 DRAM과 서버 간 자원을 통합하는 CXL 풀이 위치, 서버 간 데이터 정합성을 유지하는 고속 연결 기반으로, 메모리 자원을 필요에 따라 유연하게 배분하는 구조
  - (장기 저장 영역) SSD가 비휘발성 방식으로 데이터센터 전체 단위에 분산 배치, 전원 차단 후에도 데이터가 유지되며 대용량 맥락·이력 보관·호출 담당
  - (메모리 내 연산 영역) PIM·뉴로모픽이 메모리 내부에 연산 기능을 배치하는 메모리 내부 연산(in-memory compute) 구조, 데이터 이동 비용을 줄이는 미래형 분담 축
- 4개 영역은 상용화 단계가 서로 달라, HBM 중심의 가속기 인접 계층은 양산 경쟁, CXL·SSD 등은 초기 상용화, PIM·뉴로모픽은 실증·연구 단계로 구분

(1) HBM, AI 가속기 데이터 공급 병목을 완화하는 고속 추론 메모리로 부상

● HBM, AI 가속기 데이터 공급 병목 완화의 핵심기술

- AI 모델이 대형화되고 에이전트 AI 추론 호출량이 증가하면서, GPU 연산 대비 데이터 공급 대기 시간이 길어지며 메모리 대역폭이 시스템 성능을 좌우하는 핵심 병목으로 등장
- 다단계 추론 반복과 장문맥 처리가 결합되면서 GPU가 단위 시간당 처리해야 하는 데이터 요청량이 급증, 연산 속도보다 데이터 공급 속도가 먼저 한계에 도달하는 흐름으로 전환
- HBM은 AI 가속기와 인접한 초근접 적층 방식으로 일반 DRAM 대비 10~20배 높은 대역폭을 제공, 고속 추론 과정의 데이터 공급 지연을 줄이는 현실적 해법으로 자리매김
- AI 추론 수요 확대에 따라 HBM도 세대별로 고도화되며, 대형 LLM 추론(HBM3)에서 에이전트 AI 다단계 추론(HBM4) 단계로 처리 영역을 넓히며 AI 인프라 핵심 메모리로 정착

● HBM의 역할이 범용 부품 공급에서 가속기 공동 설계 파트너로 확장

- HBM은 단순한 빠른 메모리를 공급하는 역할을 넘어, 대역폭·발열 제어·패키징을 통합 최적화해 AI 가속기의 연산 처리량 상한을 좌우하는 핵심 기반으로 작동
- HBM4부터는 베이스 다이와 2.5D·3D 패키징이 가속기-메모리 연결 효율을 좌우, 표준 HBM으로는 최적 성능 구현이 어려워지며 고객사 가속기 구조에 맞춘 메모리 공동 설계 요구가 확대
- 이에 고객사 요구에 맞춰 대역폭·용량·전력·패키징 방식을 조정한 Custom HBM이 본격 등장, 빅테크가 자사 AI 가속기 설계 단계부터 메모리 기업과 사양을 공동 합의하는 방식으로 전환
- 결과적으로 메모리 기업의 역할도 범용 부품 공급자에서 가속기 공동 설계 파트너로 확장, HBM 경쟁의 무게중심도 가속기 최적화 메모리 체계 제공 능력으로 이동

● 韓·美·中 HBM 양산 경쟁, 세대별 격차 속 기업별 전략 분화

- '26년 상반기 한국·미국 메모리 기업이 HBM4 양산에 진입한 가운데, 중국은 정책 지원 기반으로 추격 중이나 HBM3 단계에 머물며 세대 격차가 유지
- (韓 SK하이닉스) HBM 핵심 공정 일부를 외부 파운드리에 맡기는 분업 방식으로 양산 효율과 고객 다변화 동시 추구, 가속기 고객사별 맞춤 사양 대응력 강화로 시장 선도 위치 유지

- (韓 삼성전자) Custom HBM 전담 조직과 자체 베이스 다이 공정을 앞세워 외부 위탁 없이 빅테크 맞춤 설계를 내부에서 완결하는 수직 통합 차별화
- (美 마이크론) 기존 DRAM 공정 활용 기조를 유지하며 검증된 기술 기반 안정성 확보에 주력하나, 베이스 다이·패키징 최적화 중심의 Custom HBM 경쟁에서는 상대적으로 후발 주자로 평가
- (中 CXMT) '24년 HBM2 양산 이후 HBM3 진입을 시도하나 소재·부품 발주 지연으로 '27년 이내 양산 목표로 조정, 자체 개발·정부 지원과 화웨이 협력 기반으로 추격 중

○ HBM 단독 한계 부각으로 다층 메모리 구조의 역할 분담 흐름

- HBM은 고속 추론 병목을 완화하는 핵심기술이나, 가격이 일반 DRAM 대비 약 5배에 달하고 공급도 제한되며 빅테크 중심의 선점 경쟁이 가속
- 또한, HBM은 GPU에 탑재할 수 있는 용량이 제한적이고 일반 DRAM 대비 전력 소모도 높아, 초장문맥 추론을 처리하거나 대규모 데이터센터를 운영할 때 비용·전력 부담을 높이는 요인
- 이에 HBM은 고속 추론을 담당하고, 메모리 확장·공유는 CXL이, 장기 문맥·대용량 데이터 처리는 HBF·SSD가, 데이터 이동 최소화는 PIM이 분담하는 다층 메모리 흐름 형성

(2) CXL, AI 데이터센터 메모리 불균형을 해소하는 확장·공유 인프라로 부상

○ AI 서버 메모리 비효율로 탄력적 배분 역량이 새 경쟁 변수로 등장

- AI 데이터센터는 GPU·CPU별 전용 메모리를 사전에 고정해 두는 방식으로 운영되나, 워크로드마다 필요한 메모리 용량과 사용 시점이 달라 서버 간 메모리 사용 편차 발생
- 이에 따라, 일부 워크로드는 메모리 부족으로 처리 지연이 발생하는 반면, 다른 워크로드는 할당된 메모리를 충분히 쓰지 못한 채 남겨 두는 자원 낭비 발생
- 특히 AI 서버에서는 대용량 메모리 수요가 특정 시점에 집중되는 특성으로 인해, 통상 운영 시 설치 메모리의 실제 활용률이 20~30% 수준에 그치는 사례도 보고
- 에이전트 AI 확산으로 부하 집중 시점의 변동성이 커지면서, 고정 할당 방식으로는 수요 변화에 실시간 대응하기 어려운 구조적 한계가 부각
- 이에 AI 데이터센터 효율성은 개별 서버의 연산 성능뿐 아니라 필요한 위치에 메모리를 적시에 배분하는 능력에 좌우되며, 탄력적 배분 능력도 HBM 대역폭과 구별되는 핵심 경쟁 변수로 확대

- CXL, HBM 용량 한계를 보완하는 메모리 확장·공유 인프라
  - CXL은 특정 서버의 유휴 메모리를 풀 형태로 통합하고, 여러 연산 장치가 워크로드에 필요한 메모리 용량을 동적으로 할당받을 수 있게 함으로써 메모리 재배분 가능
  - 그 결과 기존 고정 할당 구조에서 발생하던 메모리 부족과 유휴 메모리의 동시 발생 문제를 완화하고, 데이터센터 전체의 메모리 활용률 제고에 기여
  - 이에 CXL은 HBM의 탑재 용량 한계와 데이터센터 메모리 자원 비효율을 함께 보완하는 핵심기술로 부상, 표준 진화에 따른 성능 개선이 누적되며 상용 도입 기반을 확보
  - 현재 시점, CXL 메모리는 DRAM 기반 휘발성 모듈이 양산 주류이며, 휘발성·비휘발성을 함께 묶는 통합 모듈은 표준상 지원은 가능하나, 아직 상용화 초기 단계
- CXL 모듈 양산과 클라우드 채택에 따른 상용 운영 단계 진입
  - CXL은 CXL 2.0 기반 메모리 모듈의 제품화·고객 검증을 중심으로 초기 상용화 단계에 진입, CXL 3.0·3.1은 메모리 풀링·공유 기능을 고도화하는 차세대 적용 단계로 전개
  - (삼성전자) CMM-D 2.0 메모리 모듈을 '25년 주력 제품으로 고객 검증 단계에 공급하고, 차세대 표준 대응 CMM-D 3.1도 '25년 말 개발을 마쳐 도입 협의 단계 진입(삼성전자, '25.10.)
  - (SK하이닉스) OCP 2025에서 반복 사용 데이터를 외부 CXL 메모리에 임시 저장하는 시스템을 시연하며, GPU 메모리 부담을 줄이고 응답시간을 단축하는 효과 입증(SK하이닉스, '25.10.)
  - 수요 측면에서는 Microsoft Azure가 CXL 메모리 적용 클라우드 구성을 프리뷰 형태로 공개하는 등 초기 적용 사례가 등장, 주요 클라우드 사업자도 CXL 기반 메모리 확장 기술 검토 확대
- CXL 생태계 경쟁이 데이터센터 통합 설계 역량으로 확대
  - CXL이 데이터센터에서 실제로 작동하려면 표준·컨트롤러·스위치·메모리 모듈·운영 SW가 같은 방식으로 연동돼야 하며, 단일 부품 개발만으로는 상용 운영 구현에 한계
  - 이에 메모리와 CPU를 연결하는 컨트롤러, 메모리 풀을 구성하는 스위치, 풀 운영을 지원하는 Linux 기반 SW가 잇따라 등장하며 CXL 5개 계층 전반에서 전문 사업자 가시화

- 결국 CXL 경쟁의 핵심은 개별 부품보다 표준·컨트롤러·스위치·운영 SW를 연동하는 생태계 구축 역량으로 이동하며, 이 역량이 데이터센터 메모리 운영 효율을 좌우

(3) AI 스토리지·HBF, 에이전트의 장기 기억을 지탱하는 비휘발성 저장 인프라로 부상

● 스토리지, AI 장기 기억 인프라로 역할 확대

- 에이전트 AI는 사용자의 과거 작업, 외부 검색 결과, 도구 실행 이력을 지속적으로 참조하며 작동하므로, 이전 세션의 맥락을 다음 세션에서 다시 호출하는 흐름이 중요
- 누적 데이터를 HBM이나 DRAM 같은 휘발성 메모리에 모두 상주시키는 데는 한계가 있어, 대용량 데이터를 빠르게 불러오는 비휘발성 저장 계층이 추론 성능의 핵심 요소로 자리잡는 추세
- 특히 기업용 AI 에이전트는 문서·코드·업무 시스템·고객 데이터와 연결되면서, 단순 파일 보관을 넘어 필요한 지식을 즉시 호출할 수 있는 지속 기억 인프라를 필수 기반으로 요구
- NVIDIA는 '26년 3월 GTC 2026에서 HBM부터 외부 스토리지까지 데이터를 단계적으로 확장하는 구조를 발표하며, 스토리지를 추론 성능을 좌우하는 핵심 인프라로 제시

● 고성능 SSD 확산에 따른 HDD 대체 흐름 본격화

- 에이전트 AI 추론 워크로드는 무작위·고빈도 데이터 접근을 반복하면서 저장 장치의 응답 속도와 내구성 요구를 높이고 있으며, 이 과정에서 회전 디스크 기반 HDD의 한계가 부각
- 이에 따라 고성능 SSD는 모델 데이터, 벡터DB, 검색 인덱스 등 추론 핵심 데이터를 HDD보다 빠르고 효율적으로 저장·호출하는 대안 제품군으로 자리매김
- 주요 메모리 3사가 AI 추론에 특화된 SSD 라인업을 잇따라 출시하면서 AI 추론용 SSD 양산 경쟁이 본격화되고 있으며, 각사는 보유 역량과 제품 강점을 중심으로 차별화 전략을 전개
  - (마이크론) 245TB 초대용량 SSD를 통해 같은 랙 공간에 더 많은 데이터를 저장할 수 있도록 지원하고, HDD 대비 전력 사용과 운영 비용을 줄이는 데이터센터 효율화 전략에 주력
  - (삼성전자) 빠른 응답 속도가 필요한 추론 작업에는 Z-NAND를, 대용량 데이터 저장이 필요한 작업에는 PM1763을 활용하는 방식으로 제품군을 분화해 AI 워크로드별 대응력을 강화

• (Solidigm) SK하이닉스의 NAND 기술을 기반으로 122TB SSD를 출시해 대용량 SSD 라인업을 보강하고, 그룹 내 메모리 자산을 활용해 AI 추론용 SSD 시장 진입 확대

- 이러한 제품 전환은 단순한 저장 성능 개선을 넘어 데이터센터의 공간·전력·운영비 효율까지 변화시키고 있으며, AI 추론 인프라에서 SSD가 HDD를 대체하는 표준 선택지로 확산

○ HBF, DRAM과 SSD 사이를 잇는 신규 메모리 계층으로 부상

- DRAM은 접근 속도가 빠르지만, 용량 확장에 제약이 크고, SSD는 HDD보다 빠른 대용량 저장 장치이나 DRAM 수준의 지연시간과 대역폭을 제공하기 어려워 대규모 추론 데이터 호출에 한계
- 이러한 간극을 보완할 후보로 HBM 수준의 속도와 SSD 수준의 용량을 함께 지향하는 HBF가 등장하면서, DRAM과 SSD 사이를 잇는 새로운 메모리 계층으로 주목받기 시작
- HBF는 저장용 칩(NAND)을 HBM처럼 수직으로 쌓아 대역폭을 높이려는 고대역폭 플래시 계열 기술로, HBM보다 큰 용량과 SSD보다 높은 접근 성능을 지향하며 GPU 인접 배치 가능성도 제시
- SK하이닉스의 HBM·HBF 결합 아키텍처인 H^3 시뮬레이션 결과, 1천만 토큰 KV 캐시 처리량이 6.14배 개선되며 대규모 추론에 필요한 GPU 메모리 규모를 줄일 가능성 제시
- 이어 SanDisk와 SK하이닉스가 '26년 2월 OCP 산하 HBF 표준화 워크스트림을 발족하면서, HBF는 개별 기술 제안을 넘어 추론 인프라 비용 구조를 바꿀 가능성이 있는 산업 표준 후보로 확대

(4) PIM·뉴로모픽, 데이터 이동 병목을 줄이는 메모리 중심 컴퓨팅으로 진화

○ PIM, 메모리를 저장 공간에서 연산 보조 계층으로 확장

- PIM은 메모리 내부 또는 인접 영역에 연산 기능을 배치해 일부 연산을 직접 수행하는 차세대 컴퓨팅 기술로, 연산 장치와 메모리가 분리된 기존 구조를 보완하는 방식으로 설계
- 데이터를 CPU·GPU로 반복 이동시키지 않고 메모리 근처에서 단순·반복 연산을 처리함으로써, 데이터 이동 경로를 단축하고 전력 소모를 줄이는 효과를 기대
- 적용 영역은 HBM-PIM, LPDDR-PIM, NVM-PIM 등 메모리 종류별로 분화되고 있으며, 서버·온디바이스·엣지 등 처리 환경별 성능 및 전력 조건에 맞춰 적용 범위 확대

• (서버) HBM-PIM은 대규모 AI 학습·추론에서 데이터를 GPU로 반복 이동시키는 과정의 전력·지연 부담을 메모리 내부 연산으로 보완하는 서버용 고성능 PIM

• (온디바이스) LPDDR-PIM은 스마트폰·웨어러블 등 온디바이스 AI 기기의 배터리·발열 한계로 제약됐던 AI 연산 처리량을 끌어올리는 저전력 모바일 PIM

• (엣지) NVM-PIM은 상시 전원 공급이 어려운 센서·IoT 환경에서 전원 차단 후에도 학습 데이터를 유지해, 기존에 적용이 제한됐던 영역까지 AI 활용 범위를 확장하는 엣지용 비휘발성 PIM

○ 한국 PIM, 응용처별 효율 개선 실증으로 분담 전략 가속

- 한국 메모리사는 DRAM·HBM·모바일 메모리에서 축적한 경쟁력을 바탕으로, PIM을 차세대 메모리 주도권 확보를 위한 핵심기술 축으로 육성 중

• (SK하이닉스) DRAM 내부 연산형 메모리 AiM을 활용한 가속 카드 AiMX를 NVIDIA H100 GPU와 연동해 실제 추론 환경에서 구동하여, PIM의 데이터센터 적용 가능성 실증

• (삼성전자) 서버용 HBM-PIM 개발로 데이터센터 가속기 채택을 추진하고, 온디바이스용 LPDDR-PIM은 연산 전력 70% 이상 절감 기술로 제시하며 국제 표준화 주도

- 정부도 K-클라우드 기술개발('25~'30)을 통해 국산 NPU·PIM을 데이터센터 인프라·컴퓨팅SW· 클라우드 서비스와 연계하여, HW와 SW를 아우르는 풀스택 구조 구축 추진

- 기업 실증과 정부 K-클라우드 R&D가 결합되면서, 한국 PIM 전략은 실제 AI 서비스 환경에서 성능·표준·소프트웨어 호환성을 검증하는 상용화 준비 단계로 이동

○ 뉴로모픽, 연산·기억 통합을 지향하는 미래형 기술 후보로 주목

- 뉴로모픽은 인간 뇌의 신경세포 작동 방식을 반도체로 구현한 기술로, 전처리·인식·추론 등 AI 처리 과정의 전력 소모와 응답 지연을 줄일 대안으로 주목

- 글로벌 진영은 단순 칩 시연을 넘어 대규모 시스템 구축과 LLM 추론 효율 검증으로 실증 범위를 확대, 상용화보다는 시스템 확장성·저전력 추론·실시간 학습 가능성 검증에 우선 집중

• (Intel) Loihi 2 기반 대규모 뉴로모픽 시스템 Hala Point를 구축하고 1,152개 Loihi 2 프로세서와 11.5억 뉴런 규모를 집적해 이전 Pohoiki Springs 대비 성능 개선

• (IBM) 뇌 구조 모사형 AI 추론 칩 NorthPole을 통해 LLM 추론 효율 검증을 확대, 16개 NorthPole 프로세서를 탑재한 서버에서 저지연·저전력 추론 가능성 부각

- 국내에서는 초저전력 LLM 처리와 온디바이스 자가학습 칩 연구를 중심으로 뉴로모픽 원천 IP를 축적하고 있으며, 엣지·온디바이스 AI 적용 가능성 검증 단계로 전개

• 국내 반도체 기업 '엣지AI'는 'MDS인텔리전스'와 전략적 업무협약을 맺고 원격검침 등에 사용되는 스마트미터링에 탑재해 출시할 계획, 내년 초('27년 초) 대만의 TSMC를 통해 양산 예정

• KAIST는 멤리스터 기반 자가 학습 뉴로모픽 칩을 개발, 저장과 연산을 동시에 수행하는 소자 구조를 활용해 실시간 이미지 처리에서 오류를 스스로 학습·보정하는 기능을 검증

• 정부는 '25~'28년 SNN·DNN 혼합 가속기와 아날로그-디지털 혼성 초저전력 엣지 SoC 등에 대규모 투자를 진행하며, 뉴로모픽 기반 엣지 AI 반도체 원천기술 확보 본격화

- 다만 기존 AI 개발 환경과의 호환성 확보와 소프트웨어 최적화가 남아 있어 본격 상용화는 '30년 이후로 예상되며, 현시점에서는 원천기술 축적 중심의 미래 기술 트랙으로 전개 전망

출처 : Morgan Stanley 외(2026.5.)
www.morganstanley.com/insights/podcasts/thoughts-on-the-market/agentic-ai-supply-chain-market-shawn-kim
www.globalxetfs.com/articles/memory-is-the-new-bottleneck-in-ai-semiconductors
www.viksnewsletter.com/p/the-cpu-bottleneck-in-agentic-ai
www.idc.com/resource-center/blog/agent-adoption-the-it-industrys-next-great-inflection-point/
www.anthropic.com/engineering/multi-agent-research-system
www.vastdata.com/blog/kvcache-context-memory-storage
www.trendforce.com/insights/memory-wall
www.amd.com/en/blogs/2026/agentic-ai-changes-the-cpu-gpu-equation.html
www.newspim.com/news/view/20260414000262
www.patsnap.com/resources/blog/articles/in-memory-computing-architecture-landscape-2026/
www.patsnap.com/resources/blog/articles/edge-ai-inference-accelerators-2026-tech-landscape/