AI 추론 효율화를 위한 HBF의 등장

산업기술리서치센터 디지털·반도체팀
장 은 현 (ehj@kdb.co.kr)

◆ 최근 AI 산업이 '학습'에서 '추론' 중심으로 이동하며 대용량 맥락 데이터의 장기 저장 수요가 증가하고 있으나, HBM은 용량 제한 및 데이터 휘발성 등 한계 보유

◆ AI 추론에서 HBM의 한계를 보완할 메모리로 낸드플래시를 수직 적층한 'HBF'가 등장하였으며, 업계는 HBF 표준화 및 구조 혁신 등을 통해 신시장 개화에 대응

□ 맥락 기반 AI 추론 효율화를 위해 대용량 데이터의 장기 저장 수요 증대

○ 최근 AI 산업은 거대언어모델(LLM)을 구축하는 '학습' 단계에서 장기 기억 기반의 개인화 서비스를 제공하는 '추론' 중심으로 이동

- AI가 사용자의 요구에 단순히 응답하는 단계를 넘어, 대화로 축적된 맥락(Context)*을 기억하여 사용자 맞춤형 서비스를 제공하는 추론 서비스가 확산

* AI가 답변 생성 시 참고하는 대화 이력, 지시 사항, 외부 문서 등을 의미

○ 방대한 맥락 데이터 기반의 추론 작업이 요구됨에 따라 효율적인 추론을 돕는 'KV 캐시*'가 주목받고 있으며, 이를 위해서는 추가 데이터 저장 매체가 필요

* Key Value Cache : 이전 대화로 축적된 맥락 데이터를 저장해두어 반복 계산 없이 추론 속도를 높이는 기술

- 엔비디아는 CES 2026에서 eSSD*를 활용한 KV 캐시 저장용 新 메모리 플랫폼 'ICMSP(Inference Context Memory Storage Platform)'을 제시

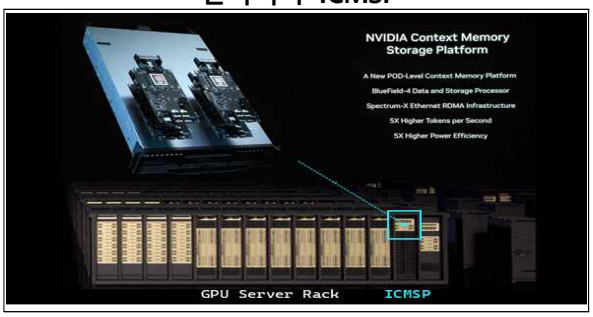
* Enterprise Solid State Drive : 데이터센터, 서버 및 고부하 AI 연산 환경 등 기업용에 최적화된 고성능·대용량 낸드플래시 기반 저장장치

엔비디아 스토리지 계층

구 분	스토리지 계층
G1	GPU HBM
G2	CPU/시스템 메모리 (LPDDR 등)
G3	서버 내 SSD
G3.5	**KV 캐시 전용 외부 SSD (ICMSP)**
G4	공유 스토리지

주 : 엔비디아의 스토리지 계층은 AI 데이터 병목 해소를 위해 세분화되었으며, ICMPS는 기존 계층에 G3.5 레이어로 추가되어 외부 KV 캐시 저장소 역할
자료 : 유진투자증권('26.1), "Memory Watch"

엔비디아 ICMSP



자료 : 엔비디아, 산업은행 재구성

□ 높은 대역폭을 보유한 HBM은 AI 학습 등의 필수 메모리로 수요가 급증하고 있으나, 맥락 기반 추론 시 용량 제한 및 데이터 휘발성 등 한계점 보유

○ 'HBM(High Bandwidth Memory)'은 메모리 병목현상*을 개선하기 위해 기존 D램보다 대역폭을 높여 데이터 이동량을 증가시킨 메모리로, AI 모델 학습 등을 위한 핵심 부품으로 주목받으며 수요 급증

* 프로세서(CPU, GPU 등)의 성능 향상 속도 대비 메모리 성능(용량, 대역폭 등) 개선 속도가 느려, 시스템 전체 연산 성능이 저해되는 현상

- HBM은 D램을 적층한 후 TSV*로 칩을 연결하여 대역폭**을 확장한 메모리로, GPU 근처에서 데이터를 주고받으며 AI 연산을 지원

* Through Silicon Via : 적층된 칩 간 전기적 연결을 위해 형성한 수직형 구멍

** 프로세서와 메모리 간 단위 시간 동안 전송할 수 있는 데이터양을 의미하며, 대역폭이 높을수록 데이터를 더 빠르게 공급 가능

○ 그러나, HBM은 대규모 맥락 데이터의 장기 보관이 요구되는 추론 영역에서 용량 제한 및 데이터 휘발성 등 한계점 보유

- 기술 난이도가 높아 적층 단수를 높이기 어렵고, GPU 연산을 지원하기 위해 근처에 위치하므로 용량 증가를 위한 물리적 공간 제약
- 또한, 전원이 공급될 때만 데이터를 저장하는 휘발성 메모리로 이전 대화의 맥락 저장을 위해서는 비휘발성 메모리 대비 전력 소모 증가

□ AI 추론에서 HBM 한계를 보완할 메모리 중 하나로 낸드플래시 기반 'HBF'가 등장함에 따라, 업계는 HBF 표준화 추진 및 차세대 메모리 계층 구조를 제시

○ HBF(High Bandwidth Flash)는 D램 기반의 HBM과 유사하게 낸드플래시*를 수직 적층하여 대역폭 및 용량을 확장한 메모리로, HBM의 한계점을 보완할 메모리 중 하나로 등장

* D램 대비 속도가 느리나 대용량 구현이 용이한 비휘발성 메모리

HBM 및 HBF 비교

구 분	HBM4[주]	1세대 HBF(예상)
기반	D램	낸드플래시
용량(GB)	64	512
대역폭	~2 TB/s	1,638 GB/s
데이터 유지	휘발성(전원 차단 시 소멸)	비휘발성(전원 차단에도 유지)
주요 역할	AI 학습, 고성능 연산	AI 추론, 저장·연산 보조 메모리
전력 효율	낮음(대기전력·발열 부담)	높음

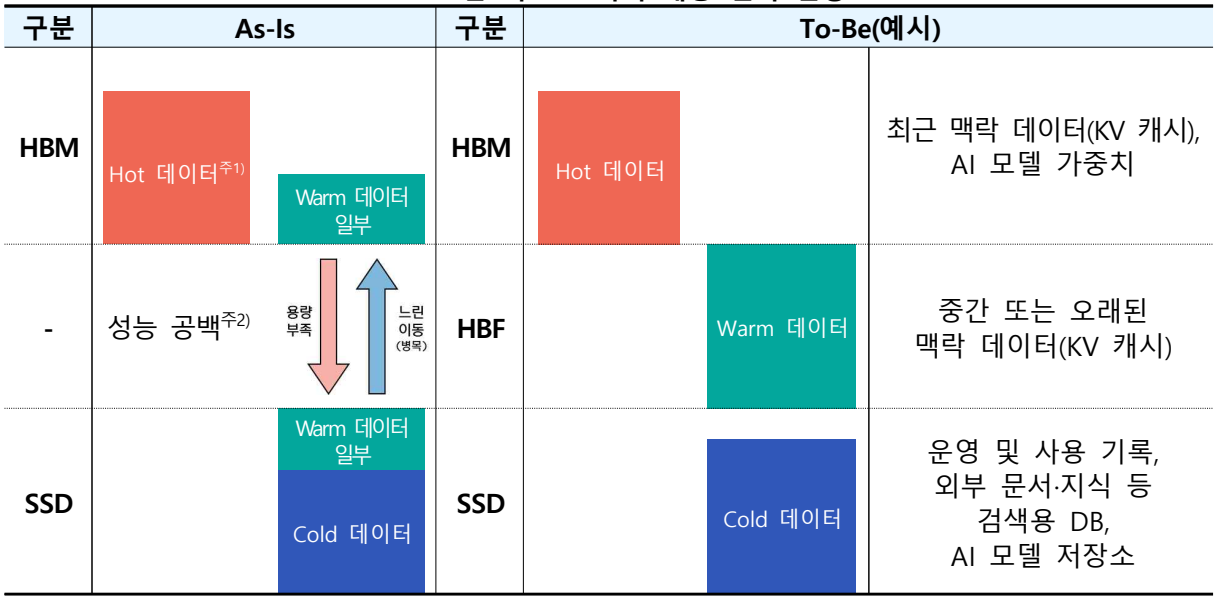
주 : 국제반도체표준협의기구(JEDEC) 기준

자료 : 샌디스크('25.7), 뉴스웨이('25.12.17.) "HBM만으론 부족...HBF 부상에 한미반도체 재조명" 산업은행 재구성

○ AI 모델 학습·추론 등에서 작업별 메모리 계층 최적화가 연산 효율 향상의 핵심으로, 이를 위해 AI 메모리 계층은 HBM의 고대역폭, HBF의 대용량 등 각 장점을 결합한 형태로 발전할 전망

- AI 추론에 활용되는 맥락 데이터(KV 캐시) 등은 HBM에 저장되고, 대화 길이 증가 등으로 HBM 용량이 부족할 경우 맥락 데이터 일부가 SSD로 이동
- 그러나, 데이터 읽는 속도가 느린 SSD로부터 AI 추론을 위해 데이터를 다시 호출할 때 추론 지연 유발
- HBF가 추가된 스토리지 계층에서는 일부 맥락 데이터를 SSD 대신 HBF에 저장함으로써 HBM 용량 한계를 해결하고 SSD의 느린 속도 개선 효과 기대

HBF 도입 시 스토리지 계층 변화 전망



주1 : 데이터는 활용 빈도에 따라 Hot/Warm/Cold 데이터로 구분되며, ①Hot 데이터는 최근 기록되어 추론에 곧바로 사용되는 맥락 데이터, ②Warm 데이터는 대화로 누적된 맥락 데이터, ③Cold 데이터는 운영·사용 기록, 검색용 지식 DB 및 모델 등 장기 기억 데이터 등을 의미

주2 : HBM 용량을 초과하는 Warm 데이터는 SSD로 저장되며, 추론 시 필요 데이터가 SSD에서 HBM으로 이동하나, SSD는 낸드플래시 기반으로 데이터 이동 속도가 느려 병목 발생

○ SK하이닉스는 미국 샌디스크와 협력하여 HBF 표준화 및 생태계 구축을 추진 중이며, 최근 HBF를 활용한 차세대 메모리 아키텍처 'H^3(Hybird architecture using HBM and HBF)' 구조를 제시하며 기술 기반 강화 중

- 낸드플래시 제조사인 샌디스크는 HBF 시제품 공개('26년 하반기) 및 HBF를 탑재한 AI 가속기 샘플 출시('27년 초) 계획 발표
- 'H^3' 구조는 HBF에 읽기 전용 데이터를 저장하고, HBM은 업데이트가 잦은 동적 데이터를 저장하여 용도별 메모리 계층을 세분화한 모델