Industry Indepth| 2026.6.8

[반도체/반도체와 장비] (비중확대)

NAND, LOVE LOVE LOVE

KUVIC Research Team 2

메일 kuvic_korea@naver.com

팀장 44기 Senior 김서정
팀원 44기 Senior 김현진
팀원 43기 Senior 정상엽

CONTENTS

Summary

AI 추론 시대, 토큰 폭발로 인한 메모리의 위기

'1조 파라미터 LLM 단일 사용자 1명에게 필요한 **KV 캐시(Key-Value Cache) 용량 310GB, GPU 텐서 코어의 99% 유휴 시간**', AI 추론 인프라가 직면한 현실 수치이다. 에이전틱 AI의 부상으로 기존 챗봇 대비 **10~100배 이상의 토큰을 생성**하면서 KV 캐시가 폭증하고 있다. 그러나 차세대 루빈 GPU에 탑재되는 **HBM4의 용량은 288GB에 불과**해, 1조개를 넘어 10조개 규모의 파라미터 모델 시대에는 **단일 가속기 메모리만으로 추론이 사실상 불가능한 구조적 병목**에 직면해 있다. 이러한 메모리 월(Memory Wall) 현상 해결을 위해, 엔비디아는 **CMX라는 3계층 메모리 아키텍처**를 도입하여, '**HBM4 → SOCAMM2 → NVMe SSD**'로 이어지는 메모리 계층 구조가 표준화되고 있다. 즉, **NAND 플래시가 단순 보조 스토리지에서 추론 인프라의 핵심 메모리 계층으로 격상**되면서, 앞으로 NAND 플래시에 주목해야 하는 이유이다.

변화하는 NAND 산업

이렇듯 AI 추론 시대가 본격화되고 NAND가 메모리 계층의 중추로 진입함에 따라, 그 수혜는 **NAND 산업 전반과 차세대 메모리 HBF**에 고스란히 전달될 것이다. 본 리서치 팀은 **NAND 산업의 3대 변화(엔비디아 CMX, 적층 단수 증가, TLC→QLC 전환)**가 HBF의 기술적 기반을 형성하는 흐름을 분석했으며, **HBM 대비 8~16배 용량을 1/5~1/10의 가격에 제공하는 HBF**가 AI 추론 시대의 중요한 역할이 될 것으로 판단한다.

NAND 적층 단수의 경우, 2025~2026년 양산 주력이 **200~230단대에서 2026~2027년 300단대로 본격 전환**될 전망이다. 특히 단수 경쟁이 과거와 달리 2025년 이후 **한 세대당 100단 이상씩 점프**하는 양상으로 변화하고 있으며, **삼성과 키옥시아는 각각 2030년과 2027년 1,000단 양산을 선언**했다. 이러한 단수 경쟁의 핵심 기술인 **CBA(CMOS Bonding Array)와 하이브리드 본딩이 HBF의 적층 공정과 기술적 뿌리를 공유**한다는 점에서, **적층 단수 경쟁은 HBF 양산 가능성을 이끌 주요 요인**이 될 것이다.

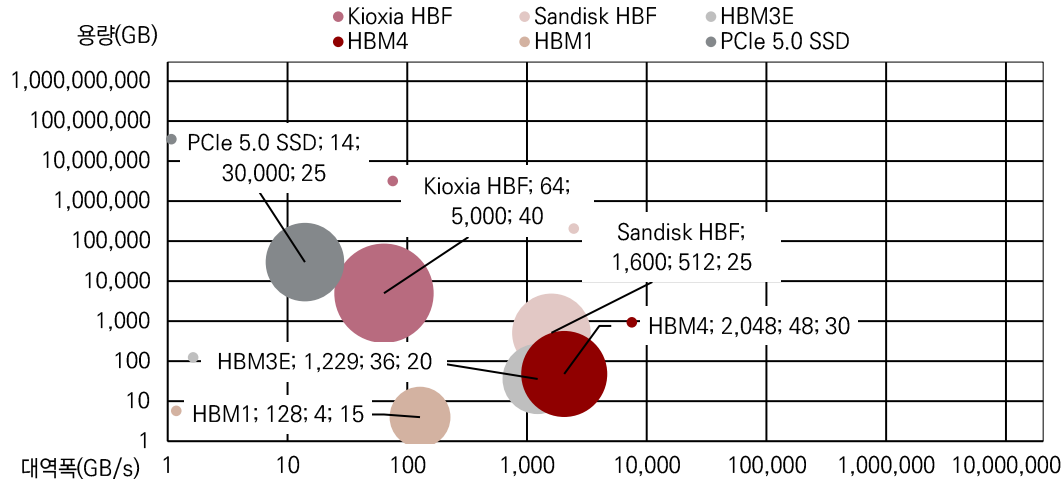
TLC에서 QLC로의 전환 역시 AI 추론 워크로드의 **'한 번 쓰고 수백만 번 읽는' 특성**이 **QLC의 짧은 쓰기 수명(P/E 사이클 ~1,000회)을 사실상 무력화**시키면서 가속화되고 있다. HDD를 QLC SSD로 대체 시 **"랙 공간 1/3 축소 +·전력비 20% 절감 + 총 저장 비용 31% 절감 효과"**가 입증되었으며, **2027년에는 QLC 출하량이 TLC를 추월할 전망**으로, 이제는 QLC가 주류가 되는 것이다.

HBF의 엣지

HBF의 경우, **HBM4와 동일한 폼팩터·전력 프로파일을 유지하면서 스택당 용량을 36~48GB에서 512GB로 약 10배 이상 확장**하는 **드롭인 호환 솔루션**이다. 본 리서치팀의 추정에 따르면 **HBF Gen1의 GB당 단가는 약 $3.92/GB(Base Case 기준)**로 산출되었으며, HBM4 스택과 거의 동일한 비용에 10배의 용량을 제공한다는 점에서, **NVIDIA·AMD가 추론용 GPU에 HBF를 채택할 경제적 해자가 충분**한 것으로 보인다.
HBF 시장 규모는 2027년 양산 본격화 시점 기준 약 $401.4억 달러 규모로 추정되며, 글로벌 NAND TAM의 25% 수준이 될 것으로 전망한다. 또한 이는 **메모리 3사 및 샌디스크의 매출 성장에 직접적으로 기여**할 것으로 전망된다.

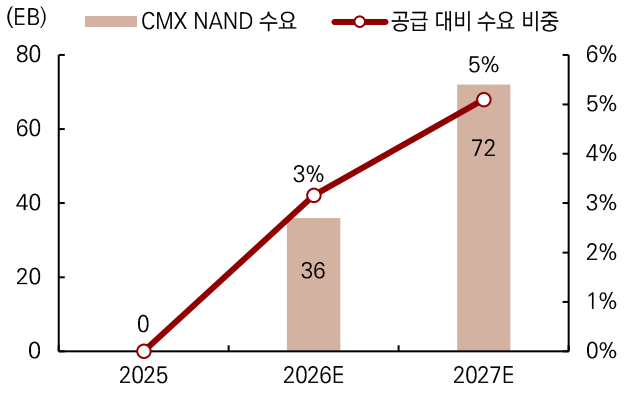
Key Chart

그림 1. 기업별 HBF, HBM 스펙 포지션



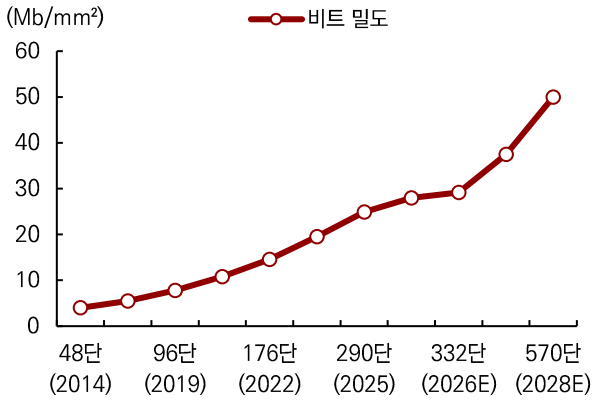
자료: 키옥시아, 샌디스크, TrendForce, KUVIC 리서치 2팀

그림 2. NVIDIA CMX eSSD 수요 추정



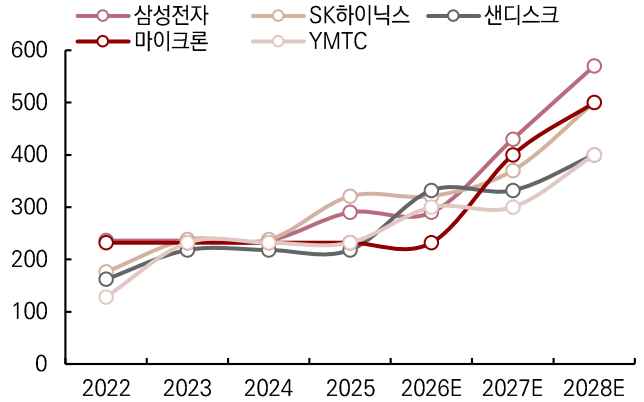
자료: KUVIC 리서치 2팀

그림 3. 단수 증가에 따른 비트 밀도 향상 추이



자료: AnandTech, TechInsights, TrendForce, KUVIC 리서치 2팀

그림 4. 기업별 적층 단수 변화



자료: 각 사, KUVIC 리서치 2팀

그림 5. HBF Q 추정 주요 가정

1) HBF Q 추정 주요 가정	
27 년 이후 GPU 판매량 가정	1,000 만개
HBF 1 스택당 용량	512GB
GPU 당 HBF 스택수 가정	8 개
프리필용 GPU 점유율 가정	25%
HBF 수요 추정	**102.4EB**

자료: KUVIC 리서치 2팀 추정

엔비디아 CMX

CMX의 등장을 낳은 KV 캐시 폭발

에이전틱 AI의 부상과 KV 캐시 용량의 폭발적 증가

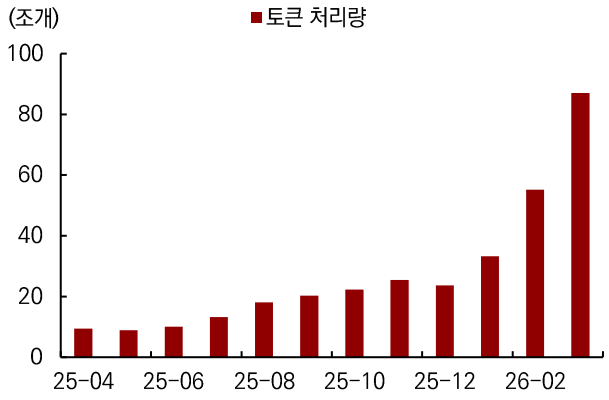
에이전틱 AI와 RAG로 캐시 폭발, 대역폭 병목

생성형 AI 서비스가 단발성 질의응답을 처리하던 초기 단계에서 벗어나 사용자 개입 없이 복잡한 목표를 스스로 계획하고 실행하는 **'에이전틱 AI(Agentic AI)'**로 진화함에 따라, 처리해야 할 데이터 규모가 급증하는 '토큰 폭발(Token Explosion)' 현상이 나타나고 있다. 에이전트 기반 작업은 수 시간 이상 세션을 유지하며 방대한 대화 문맥을 쌓아 나가고, 최신 추론 모델들은 정답 도출을 위해 확장된 사고 과정을 거치며 기존 챗봇 대비 10배에서 100배 이상의 토큰을 생성한다. 이처럼 늘어난 토큰은 에이전틱 AI가 맥락을 기억하기 위해 사용하는 임시 메모리 공간인 **'KV 캐시(Key-Value Cache)'의 폭증**으로 이어진다. 대화나 문서가 길어질수록 이 캐시 크기는 정비례하여 커지기 때문에, 추론 가속기 내부의 제한된 메모리 공간은 순식간에 불어나는 캐시 데이터로 인해 극심한 용량 압박을 받게 된다. 또한 외부 지식 문서를 실시간으로 참조하는 **검색 증강 생성(RAG) 기술**이 보편화되면서, 하위 스토리지에 보관된 방대한 데이터를 밀리초(ms) 단위로 HBM까지 빠르게 퍼 올려야 하는 대역폭 병목 현상이 심화되고 있다.

모델 파라미터 개수 늘어나며 KV 캐시가 HBM 용량을 넘어

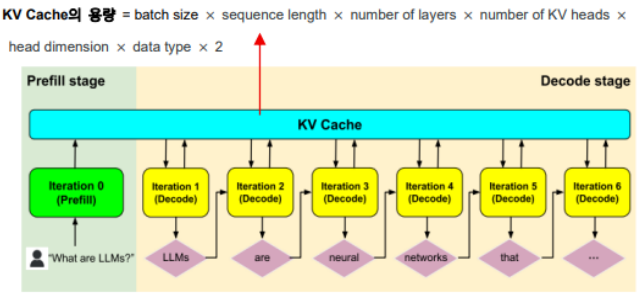
구체적인 수치로 가이드라인을 분석하면, 매개변수 70B(700억개) 규모의 고성능 LLM을 구동할 때 동시 가동되는 사용자 단 1명에게 할당해야 하는 KV 캐시 용량은 인풋 문맥의 길이에 따라 100GB 이상, 최대 310GB에 육박하게 된다. 이는 단일 GPU 하드웨어에 탑재된 초고속 고대역폭메모리(HBM)의 물리적 탑재 용량을 초과하는 수치(HBM4의 용량은 288GB)이다. 또한 1조개를 넘어 10조개 규모의 파라미터를 가진 모델이 등장할 것으로 예상되는 바, **KV 캐시 용량은 수천 GB까지** 커질 것으로 전망된다. 따라서 고용량 모델의 다중 사용자 추론 국면에서는 가속기 자체의 메모리 한계로 인해 연산 전체가 중단되거나 성능이 급락하는 구조적 병목 현상이 강제적으로 발생한다.

그림 6. 월별 토큰 처리량 추이(25.04 – 26.03)



자료: OpenRouter, KUVIC 리서치 2팀

그림 7. Transformer 모델의 KV 캐시 용량



자료: Google, KUVIC 리서치 2팀

추론 단계의 메모리 대역폭 한계와 계층적 오프로드의 필연성

추론 = 프리필 + 디코딩

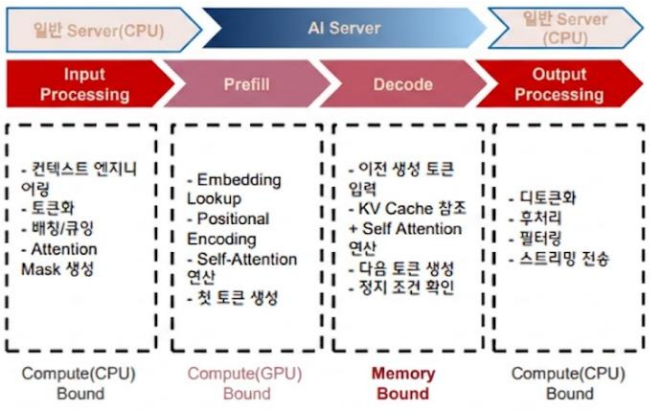
AI 추론 연산 프로세스는 크게 입력된 프롬프트 전체를 한 번에 처리하는 **프리필(Prefill)** 단계와 이후 결과 토큰을 하나씩 순차적으로 생성해 나가는 **디코드(Decode)** 단계로 구분된다. 프리필 단계는 대량의 데이터를 동시에 처리하므로 가속기의 연산 성능이 중심이 되는 **연산 중심(Compute-bound)** 특성을 나타낸다. 반면 디코드 단계는 새로운 토큰을 생성할 때마다 직전 단계까지 누적된 KV 캐시 메모리 전체를 매번 새로 읽어와야 하므로, 연산 장치의 처리 속도보다 메모리 인터페이스의 전송 속도가 효율을 지배하는 **메모리 중심(Memory-bound)** 특성을 띤다.

추론이 메모리 병목 현상을 야기

이러한 디코드 작업의 반복은 추론 가속기의 연산 대역폭에 비해 데이터 전송 속도가 턱없이 부족하여 연산 장치가 병목을 겪는 일명 **'메모리 월(Memory Wall)' 현상**을 심화시킨다. 데이터가 가속기 내부로 공급되는 속도가 연산 속도를 받쳐주지 못함에 따라, 고성능 GPU의 핵심 연산 유닛인 텐서 코어는 약 99%에 달하는 시간 동안 데이터 입력을 기다리는 유휴(Idle) 상태에 머물게 된다. 결과적으로 전체 시스템의 바이트당 연산 효율이 2 FLOPS 수준으로 급락(H100 최적 연산량은 295 FLOPS)하며 인프라 자원의 막대한 낭비가 발생한다.

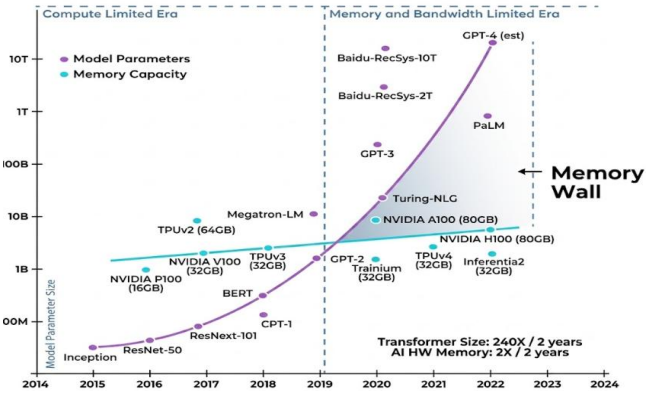
가속기 내부 HBM의 물리적 한계를 넘어서는 방대한 KV 캐시를 제어하고 텐서 코어의 가동률을 복구하기 위해서는 자원의 효율적 재배치를 위한 **'계층적 오프로드(Offload)' 아키텍처**가 필연적으로 요구된다. 이는 **연산의 시급성이 낮은 대기열 상태의 컨텍스트 데이터를 HBM 외부의 물리적 메모리 계층으로 밀어내어 저장 공간을 확보**한 뒤, 해당 사용자의 차례가 돌아오는 디코딩 시점에 맞춰 데이터를 다시 고속으로 끌어오는 제어 방식이다. 이에 따라 전체 AI 추론 인프라는 **HBM**에서 시스템 메인 **DRAM**을 거쳐 최종적으로 고성능 **NVMe SSD**로 이어지는 다단계 메모리 계층 구조를 형성하게 된다.

그림 8. AI 추론의 4단계 과정



자료: KUVIC 리서치 2팀

그림 9. AI Memory Wall



자료: Aya Labs, KUVIC 리서치 2팀

엔비디아 CMX와 Vera Rubin Pod의 등장

엔비디아의 통합 컨텍스트 메모리 시스템(ICMS/CMX) 3계층 분산 아키텍처

KV캐시를 순차적으로 내리는 CMX

엔비디아는 단일 가속기의 사양 업그레이드만으로는 폭발하는 KV 캐시와 메모리 월 한계를 극복할 수 없다고 판단하여 데이터센터 하드웨어 전반을 컨텍스트 메모리용 시스템으로 통합 재설계하는 **CMX(Context Memory eXtension) 아키텍처**를 도입했다. CES 2026에서 ICMS(Inference Context Memory Storage)라는 이름으로 처음 등장한 개념은 GTC 2026에서 CMX로 재정의 되었다. 이는 기존의 독립된 연산 장치 중심 패러다임에서 벗어나 **시스템 전체의 메모리 풀을 연산 가속을 위해 유기적으로 결합**하는 인프라 시프트의 결과물이다. 젠슨 황 엔비디아 CEO가 공식 석상에서 컨텍스트 메모리 시스템의 전면적 재설계를 선언한 것도 이와 같은 맥락이다.

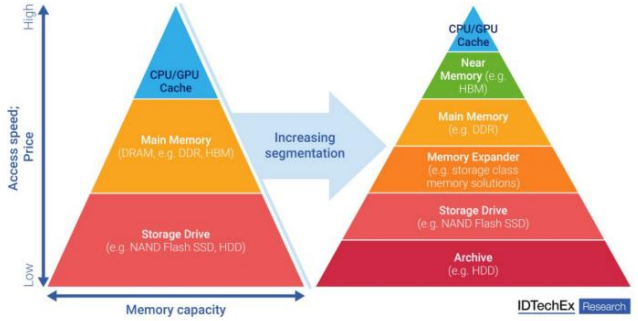
CMX의 운영체제 “Dynamo”

이 거대한 하드웨어 자원을 제어하는 소프트웨어 핵심은 **'다이나모(Dynamo)' 분산 운영체제**이다. 다이나모 OS는 데이터센터 내부에서 물리적으로 분리되어 있는 복수의 GPU, 호스트 CPU, 그리고 고속 스토리지 서버들을 단일 가상 메모리 주소 공간으로 묶어주는 역할을 수행한다. 이를 통해 물리적 거리에 따른 전송 제약을 소프트웨어 단에서 가상화하여 관리함으로써 가속기가 필요한 컨텍스트 데이터의 위치를 실시간으로 추적하고 최적의 경로로 매핑한다.

G1 ~ G4, CMX는 G3.5

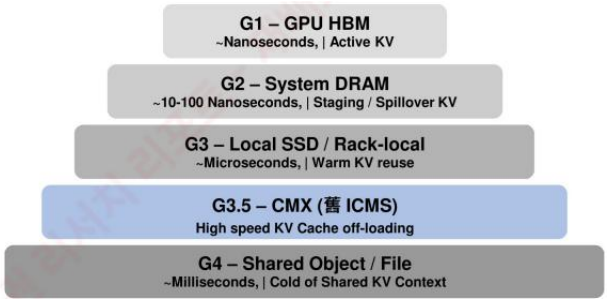
CMX 아키텍처는 컨텍스트 데이터의 접근 빈도와 신속성에 따라 메모리 자원을 총 **3개의 물리적 계층**으로 배분하여 구동 효율을 극대화한다. 최상위 1계층은 **루빈 GPU에 내장된 초고속 HBM4**로, 실시간 디코딩이 진행 중인 핵심 컨텍스트를 상주시킨다. 2계층은 메인보드 단에 위치한 **베라 CPU 기반의 고용량 SOCAMM2 메인 메모리**로, 랙당 1TB 수준의 준고속 버퍼 역할을 수행한다. 마지막 3계층은 스토리지 풀에 배치된 **고성능 NVMe SSD**로 구성되며, 대기 상태의 방대한 장기 컨텍스트 데이터를 저장하는 최종 오프로드 레이어로 기능한다.

그림 10. AI 메모리 계층 구조



자료: IDTechEX, KUVIC 리서치 2팀

그림 11. CMX: NAND를 KV 캐시 오프로딩에 활용



자료: NVIDIA, KUVIC 리서치 2팀

Vera Rubin Pod: 40개의 특수 랙으로 에이전틱 AI 하나 뚝딱

40개 특수 랙으로 이루어진 Vera Rubin Pod

엔비디아가 제시하는 **'베라 루빈 포드(Vera Rubin Pod)'**는 단순한 서버 랙의 물리적 결합을 넘어, 에이전틱 AI 워크로드를 단일 인프라 내에서 완벽하게 종결짓기 위해 표준화된 엔비디아 최상위 규격의 슈퍼컴퓨팅 단위이다. 기존의 생성형 AI가 단발성 응답 위주여서 GPU 연산 성능에만 병목이 걸렸다면, 에이전틱 AI 시대에는 데이터베이스 접근, 외부 툴 호출, 가상 환경 내 재추론 등 **복잡한 오케스트레이션 작업이 급증**하게 된다. 엔비디아는 이러한 Non-GPU 워크로드의 폭발에 대응하기 위해 프로세서 구성비 자체를 근본적으로 혁신하고, **CMX 메모리 계층 구조**를 담았다. 과거 NVL72 랙 기준 2:1 수준이었던 **GPU 대 CPU 개수 비율**을 VR Pod 시스템 전체 기준으로 약 **1:1 수준**(루빈 GPU 1,152개 대 베라 CPU 1,088개)까지 끌어올렸다. 이 거대한 인프라는 차세대 NVLink 6 인터커넥트와 광학 스위치(OCS) 네트워킹을 통해 결합되어, 외형상 수십 개의 랙임에도 소프트웨어와 연산 단에서는 마치 하나의 거대한 초고속 단일 GPU처럼 유기적으로 구동된다.

NVL72 16개, Vera CPU 랙 2개

전체 시스템은 5개 종류, 총 **40개의 특수 목적으로 고안된 랙**들이 유기적인 다단계 오케스트레이션 구조를 형성한다. 먼저 대규모 연산과 문맥 입력(Prefill) 단계를 전담하는 핵심 엔진으로서 **'Vera Rubin NVL72 랙' 16개**가 배치된다. 여기에 탑재되는 **루빈 GPU**는 HBM4를 통해 **단일 칩당 288GB의 용량과 초당 22TB(22TB/s)의 대역폭**을 확보하며, **베라 CPU**는 개당 8개의 차세대 SOCAMM2 모듈을 결합하여 **랙당 1TB**에 달하는 고용량 메인 메모리와 **초당 1.2TB**(1.2TB/s)의 전송 속도를 구현한다. GPU 없이 오직 **256개의 베라 CPU**만으로 빽빽하게 채워진 **'Vera CPU 전용 독립 랙' 2개**가 별도로 탑재되어 시뮬레이션 전용 데이터센터의 역할을 수행한다.

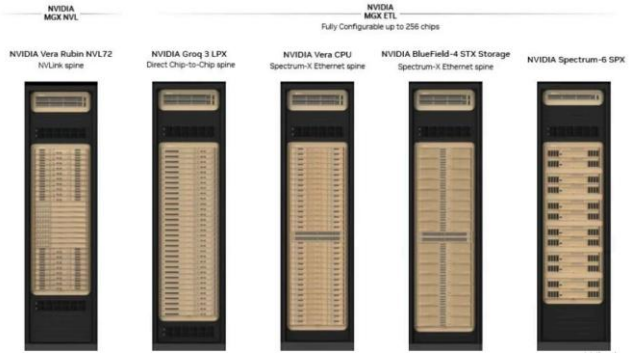
Groq LPX 랙 10개, STX 랙 2개, SPX 랙 10개

실시간 대화 및 추론 과정에서 발생하는 지연 시간(Latency)을 극단적으로 제어하고 디코드 단계를 가속하기 위해서는 **'Groq 3 LPX 랙' 10개**가 투입된다 이 초고속 추론 가속 랙에는 랙당 **256개의 LPX 칩**이 탑재된다. LPX 랙의 핵심인 LPU는 초당 무려 **150TB(150TB/s)**라는 압도적인 대역폭을 자랑하는 **500MB 용량의 SRAM**을 탑재하여, 디코딩 시 필수적인 초고속 데이터 레이턴시를 보장하고 시스템 전체의 추론 효율을 극대화한다.

이와 동시에, 에이전트가 사고하는 과정에서 발생하는 방대한 KV 캐시 데이터를 실시간으로 오프로드하고 임시 저장하기 위한 AI 전용 스토리지로서 **'BlueField-4 STX 랙' 2개**가 결합된다. 이 스토리지 랙은 **4개의 DPU 및 고성능 NIC**가 통합된 하드웨어 구조를 지니며, **총 576TB(18TB SSD 32개)**의 무지막지한 초고용량 NVMe 풀을 제공한다. STX 랙은 DPU 당 800Gbps의 속도로 랙 당 400GB/s의 스토리지 전송 속도를 통해 가속기 외부로 밀려나는 대규모 캐시 데이터를 실시간으로 수용한다. 마지막으로 이 모든 컴퓨팅, 추론 가속, CPU 전용 시뮬레이터, 캐시 스토리지 랙들을 하나로 묶어 거대한 단일 가상 메모리 공간으로 기능하게 만드는 통신 센터로서 **'Spectrum-6 SPX 네트워킹 스위치 랙' 10개**가 허브 역할을 수행한다.

결과적으로 베라 루빈 포드에서 이 40개의 랙 스펙트럼은 분산 운영체제(OS)인 **다이나모의 통제**에 따라 지연 시간을 최소화하는 최적의 경로로 매핑되며, 단일 클러스터 내에서 에이전틱 AI의 추론부터 스토리지 오프로드까지 모든 인프라 병목을 완전히 해소하는 종단간(End-to-End) 아키텍처의 실체이다.

그림 12. 엔비디아 5가지 특수 목적 랙



자료: NVIDIA, KUVIC 리서치 2팀

그림 13. 엔비디아 Vera Rubin Pod 구성

랙 종류	랙 개수	칩	메모리	대역폭
NVL72	16	GPU	HBM4	22TB/s
Vera CPU	2	CPU	LPDDR5X	1.2TB/s
Groq 3 LPX	10	LPU	SRAM	150TB/s
Bluefield-4 STX	2	DPU	NAND	400GB/s
Spectrum-6 SPX	10			

자료: KUVIC 리서치 2팀

DPU 기반 SSD 스토리지 풀링 기술 설명

DPU 중심의 초고속 데이터 전송 및 독점 생태계 구축

STX랙의 Bluefield-4가 eSSD 오프로드 핵심 기술

CMX 아키텍처 환경에서 연산 가속기와 외부 스토리지 풀 간의 데이터 이주를 전담하는 핵심 하드웨어 장치는 **블루필드(BlueField) DPU**이다. 이 장치는 가속기 노드와 스토리지 노드의 최전선에 배치되어 호스트 CPU의 개입을 완전히 차단한 채 독립적으로 네트워크 패킷을 제어하고 연산을 보조한다. 이로 인해 대규모 오프로딩 과정에서 발생하던 호스트 CPU의 연산 오버헤드와 자원 간섭 현상이 원천적으로 제거되며, 전체 인프라의 데이터 전송 효율이 극대화된다.

가속기가 호스트 메인 메모리를 거치지 않고 스토리지 파일에 직접 액세스하도록 지원하는 **GPU Direct Storage(GDS) 기술**은 전송 레이턴시를 극한으로 단축시키는 핵심 기제이다. 이 기술은 이더넷(Ethernet) 네트워크 위에서 구현되는 RoCE 기반의 **RDMA(Remote Direct Memory Access)** 규격과 결합하여 물리적 거리감을 무력화한다. 결과적으로 NVMe SSD 스토리지 풀에 저장된 대용량 KV 캐시 데이터는 프로토콜 변환 손실 없이 가속기 메모리로 초고속 직결 수송된다.

네트워크 카드로 묶인 페타바이트급 가상 단일 스토리지 풀 구현

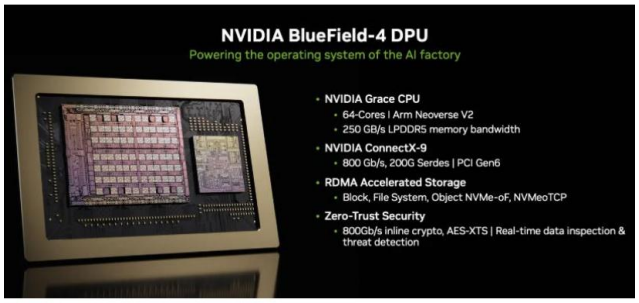
랙 하나 당 9,600TB 라는 압도적인 용량

STX 랙은 이렇게 Bluefield-4 DPU를 통해 거대한 스토리지 자원을 네트워크로 결합해 유연하게 분배하는 가상화 환경을 지향한다. STX 랙은 총 **16개의 독립적인 SSD 트레이(Tray)**로 구성되었다. 각 트레이마다 **4개의 DPU**가 독립적으로 배치되며, 이 **DPU 하나당 150TB**의 초고용량 저장 공간을 관리하는 구조를 취한다. 결과적으로 16개의 트레이와 총 64개의 DPU가 결합함으로써 하나의 STX랙은 무려 **9,600TB(9.6페타바이트)**에 달하는 거대한 물리 컨텍스트 저장 공간을 확보하게 된다.

이 방대한 스토리지 유닛들은 DPU 당 800Gbps 대역폭을 지원하는 최신 스마트닉(SmartNIC) 인터커넥트 백본을 매개로 상호 직결된다. 트레이당 4개의 DPU 채널이 동시 가동됨에 따라, 단일 스토리지 풀에서 뿜어지는 데이터 전송 속도는 **초당 400GB(400GB/s)**라는 속도를 달성한다. 이처럼 페타바이트(PB)급으로 묶인 가상 단일 스토리지 풀을 통해 에이전틱 AI 구동 시 발생하는 초고용량 KV 캐시를 소화한다.

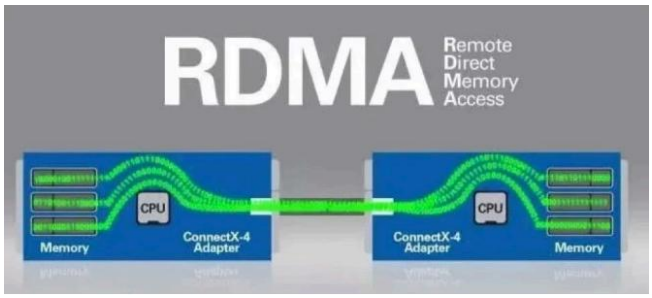
최근 Google 또한 TPU v8을 공개하며 **TPU Direct Storage**를 공개한 바 있다. NVIDIA CMX 와 유사하게 **TPU가 스토리지에 직접 접근 가능**하게 함으로써 CPU 처리의 병목 현상을 우회하고 학습 지연을 방지하는 방법론이다. CMX와 마찬가지로 고성능 eSSD에 대한 수요 증가를 이끌 것으로 보인다.

그림 14. Bluefield-4



자료: NVIDIA, KUVIC 리서치 2팀

그림 15. RDMA 시각화



자료: Fibermall, KUVIC 리서치 2팀

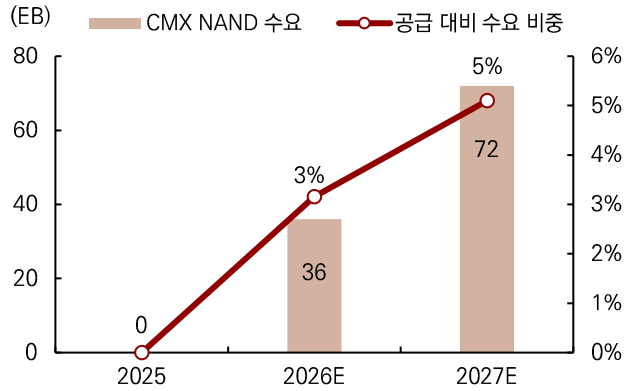
CMX로 인한 NAND 신규 수요 추정

2026E 36EB, 2027E 72EB

엔비디아의 CMX 아키텍처로 인한 NAND 신규 수요 크기는 2026년 36EB(엑사바이트, 10억 GB), 2027년 72EB로 전망되며, 이는 각 연도 별 NAND 공급 용량의 3%, 5% 수준이다.

GPU 당 SSD 용량의 경우 젠슨황이 CES 2026에서 언급한 **GPU 당 16TB**를 사용하였으며, 공개된 **STX랙 구성과 Vera Rubin Pod 랙별 구성비**를 통해 이를 검증하였다.

그림 16. NVIDIA CMX eSSD 수요 추정



자료: KUVIC 리서치 2팀

그림 17. GPU 당 SSD 용량 검증

1) CMX NAND 수요	
Bluefield-4 STX 기반 추정	
SSD 트레이	16 개
트레이당 DPU 개수	4 개
DPU 당 관리 용량	150TB
STX 랙당 SSD 용량	9,600TB

2) Vera Rubin Pod 구성비	
NVL72 : STX	8 배
NVL72 GPU 개수	72 개
GPU 당 SSD 용량	**16.7TB**
NVL72 당 SSD 용량	1,200TB

자료: KUVIC 리서치 2팀

26년의 경우 GB NVL72가 25년 하반기 출시되어 3만대가량 판매된 것을 고려, VR NVL72도 26년 하반기 출시되어 반기 간 3만대 판매를 가정하였고, 27년은 온기 반영하여 6만대 판매를 가정하였다. 27년의 북미 CSP사의 CAPEX 규모가 26년 대비 30 ~ 40%가량 증가하여 GPU 및 랙 스케일 구매량이 증가할 것으로 전망되기에 해당 가정은 보수적인 수치로 판단된다. NAND bit 공급량의 경우 Trendforce의 추정치를 사용하였다. 샌디스크 CEO는 **2027년 75 ~ 100EB**가량의 글로벌 CMX 신규 수요를 전망하였는데, 본 리서치 팀의 추정 결과 해당 전망치는 어느정도 달성 가능한 수치로 판단된다.

그림 18. CMX 수요 및 공급 세부

3) CMX NAND 수요 추정 : Vera Rubin NVL72 랙 판매량 가정	2025	2026E	2027E
(단위: 개)	0	30,000	60,000
(단위: EB)	**0**	**36**	**72**

4) NAND 공급 vs CMX 수요	2025	2026E	2027E
(단위: EB)	943	1,140	1,413
공급 대비 비중	**0%**	**3%**	**5%**

자료: Trendforce, KUVIC 리서치 2팀

그림 19. NAND 수요 구분

NAND 수요 구분 (단위: EB)	2025	2026E
Others	170	210
Mobile	324	295
Client SSD	209	177
Enterprise SSD	297	530
총합	**1,000**	**1,212**

자료: Trendforce, KUVIC 리서치 2팀

NAND 산업의 변화

NAND가 왜 여기서 나와?

엄중해진 NAND

메모리 월 병목 해소를 위해 다시 대두된 NAND

메모리와 프로세서의 속도 사이의 불균형으로 인해 **메모리 월이 병목**으로 드러나게 되면서 속도 격차를 메우기 위한 대용량의 NAND가 메모리 계층에서 주목을 받게 되었다. NAND가 용량이 큰 이유는 물리적 구조의 단순함과 높은 공간 효율성 때문이다. 구조적으로 NAND는 트랜지스터와 커패시터가 모두 필요한 DRAM과 달리, **오직 트랜지스터로만 이루어져 있어 물리적 크기를 극단적으로 줄일 수 있다**. 여기에 수십 개의 셀을 직렬로 연결하는 아키텍처를 채택해 데이터 이동 통로를 최소화했으며, DRAM처럼 칩 단위가 아닌 **셀 단위의 3차원 적층이 가능**해 제한된 공간 안에 엄청난 수의 셀을 쌓아 올릴 수 있다. 더욱이 셀 하나에 0과 1의 두 단계 신호만 구분하는 DRAM과 달리, NAND는 TLC나 QLC 방식을 통해 하나의 셀에서 **8단계에서 16단계까지 신호를 미세하게 분할 인식**하므로 **데이터 집적도 측면에서 비교가 불가능한 우위**를 점한다.

그럼에도 불구하고 **NAND 플래시를 컴퓨터의 메인 작업 메모리로 활용할 수 없는 명확한 한계**가 존재한다. DRAM이 각 셀마다 단독 통로를 배정받아 바이트 단위로 신속하게 소통하는 반면, NAND는 전자가 산화막을 강제로 뚫고 들어가야 하는 물리적 메커니즘을 사용하며 직렬 구조상 특정 셀을 읽기 위해 주변 셀을 모두 켜야만 한다. 이로 인해 **읽기와 쓰기 지연 시간이 DRAM보다 수천에서 수만 배 이상 느리고**, 덮어쓰기가 불가능해 **지우기 단계에서는 수십만 배의 시간이 소요**된다. 또한 데이터에 접근하는 기본 단위가 바이트가 아닌 수 킬로바이트(KB) 단위의 페이지나 블록으로 묶여 있어 세밀한 연산 작업에 부적합하다. 가장 치명적인 문제는 내구성에 있다. 이론상 수명이 무한한 DRAM에 비해 NAND는 **셀당 쓰기 및 지우기 가능 횟수가 수천 회 수준에 불과**하므로, 끊임없이 데이터가 교체되는 메인 메모리로 사용할 경우 시스템이 순식간에 파괴된다.

표 1. DRAM, NAND 특성 비교

비교 축	DRAM	NAND 플래시
셀 구조	트랜지스터 1 + 커패시터 1 (1T1C)	트랜지스터 1 (커패시터 불필요)
집적 방식	칩 단위 적층	셀 단위 3D 수직 적층
셀당 저장 비트	1비트(2단계)	TLC 3비트(8단계)~QLC 4비트(16단계)
접근 단위	바이트(Byte)	페이지·블록(수~수십 KB)
읽기 레이턴시	10~50 ns	25~125 µs
쓰기(프로그램) 레이턴시	10~50 ns	수십~수백 µs
지우기(Erase)	덮어쓰기로 불필요	블록 단위, 1~5 ms
내구성(P/E 사이클)	~10^{15}회 (사실상 무한)	SLC 10^5 / TLC 1k~3k / QLC 100~1k회

자료: KUVIC 리서치 2팀

이러한 특성 때문에 **NAND를 메인 연산 메모리가 아닌 KV 캐시 전용 저장소로 활용하는 대안**이 부상하게 된 것이다. AI 추론 과정에서 이전 대화 맥락을 담은 대규모 데이터는 한 번 생성되면 거의 영구적으로 유지되며, 사용자가 해당 세션으로 돌아와 다시 질문할 때 복기하는 용도로 쓰인다. 즉, 수백 번의 질문이 오가는 동안 NAND에서 데이터를 읽어오는 과정은 빈번하게 일어나지만, **기존 데이터를 지우거나 새로 고쳐 쓰는 일은 거의 발생하지 않는다**. 따라서 지우기 횟수가 극도로 적고 대용량 데이터를 장기 보존해야 하는 KV 캐시의 작동 환경은 **NAND 플래시의 치명적인 단점인 짧은 수명과 느린 쓰기 속도를 회피하면서, 가성비 좋은 대용량 저장 능력을 완벽하게 활용할 수 있는 최적의 무기**가 된다.

적층 단수 증가와 집적도 향상

현재 NAND 산업 내의 가장 큰 트렌드 두 가지는 **적층 단수 증가와 집적도 향상**이다. **비트당 원가를 낮추면서 압도적 용량 확장**을 이뤄내기 위한 증분의 국면으로 들어서게 되며 필연적으로 이와 같은 변화가 생겨나고 있는 것이다.

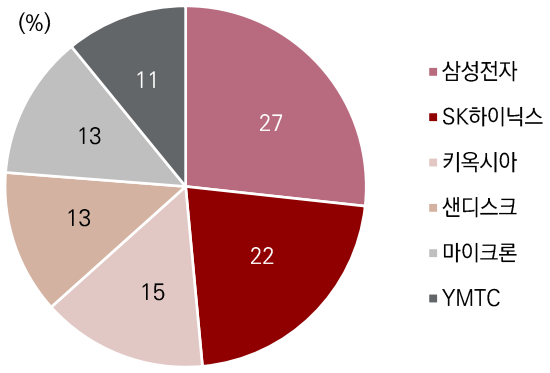
적층 단수가 뭐예요?

NAND 플래시 메모리는 메모리 셀을 평면에 배치하는 2D 구조에서 출발하였으나, 미세화 한계에 직면하며 **2013년경부터 셀을 수직으로 쌓아 올리는 3D NAND 구조로 전환**되었다. 적층 단수(layer count)란 수직으로 쌓아 올린 메모리 셀의 층수를 의미하며, **단수가 늘어날수록 동일 웨이퍼 면적당 저장 가능 비트가 비례적으로 증가**하여 GB당 원가가 하락하는 구조다. 이 때문에 **NAND 산업의 기술 경쟁은 본질적으로 적층 단수의 경쟁**이라 할 수 있다.

3D NAND는 **2025년 기준 글로벌 NAND 플래시 시장의 86.85%**를 차지하며, 사실상 2D NAND를 완전히 대체한 상태다. 일부 항공·국방용 모듈에서만 초저지연 특성을 위해 2D NAND가 잔존하고 있을 뿐이다.

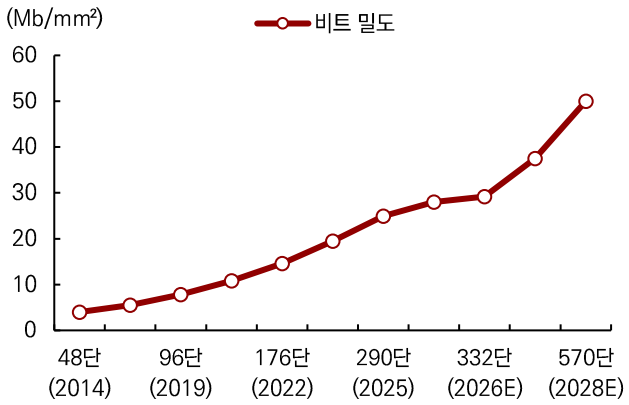
적층 단수 증가는 단순히 용량을 늘리는 것 이상의 의미가 있다. 동일 GB를 생산하는 데 필요한 웨이퍼 면적이 줄어들면서, **'단수 증가 → 비트 밀도 향상 → GB당 원가 하락'**의 원리가 작동한다. 실제로 **200단을 넘어선 3D NAND는 밀도를 45% 향상시키면서 비트당 원가를 28% 절감**하는 효과를 보였다고 한다. **적층 단수가 곧 비용 경쟁력의 핵심 지표**인 셈이다.

그림 20. NAND 시장 점유율 (4Q25 기준)



자료: Counterpoint Research, KUVIC 리서치 2팀

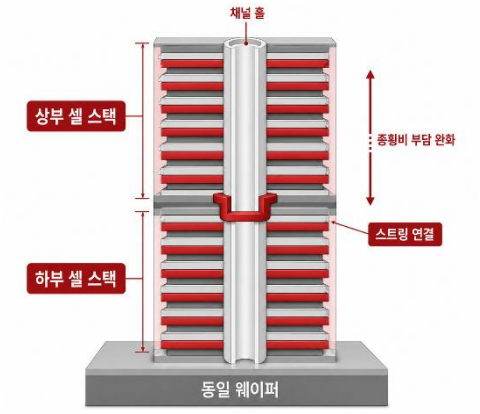
그림 21. 단수 증가에 따른 비트 밀도 향상 추이



자료: AnandTech, TechInsights, TrendForce, KUVIC 리서치 2팀

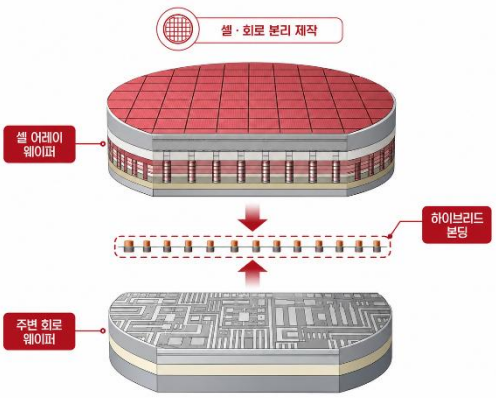
다만 단수가 일정 수준을 넘어서면 **단일 적층 방식의 물리적 한계**에 부딪힌다. 셀을 수직으로 쌓을수록 식각해야 하는 채널 홀(channel hole)의 종횡비가 가팔라지고, 식각 도중 웨이퍼가 휘어지거나 셀 간 간섭이 발생하기 때문이다. 이를 우회하기 위한 핵심 기술이 **스트링 적층과 웨이퍼 본딩**이다. 전자는 두 개의 셀 스택을 별도로 만든 뒤 동일 웨이퍼 위에 이어 붙이는 방식이고, 후자는 셀 어레이와 주변 회로를 별도 웨이퍼에서 제작한 뒤, **하이브리드 본딩으로 결합**하는 방식이다.

그림 22. 스트링 적층 방식



자료: KUVIC 리서치 2팀

그림 23. 웨이퍼 본딩 방식



자료: KUVIC 리서치 2팀

표 2. NAND 주요 기술

	적층 단수	대표 채택 기업	장점
Single Stack (모놀리식)	~128단	초기 3D NAND	단순한 공정
String Stacking (다중 데크)	128~321단	삼성, SK, 키옥시아 등	단수 증가 용이
웨이퍼 본딩 / CBA	232단~	YMTC, 키옥시아 BiCS9	셀/주변회로 분리
하이브리드 본딩 (W2W)	370단~	삼성 V10+, SK V10+	400단 이상 가능

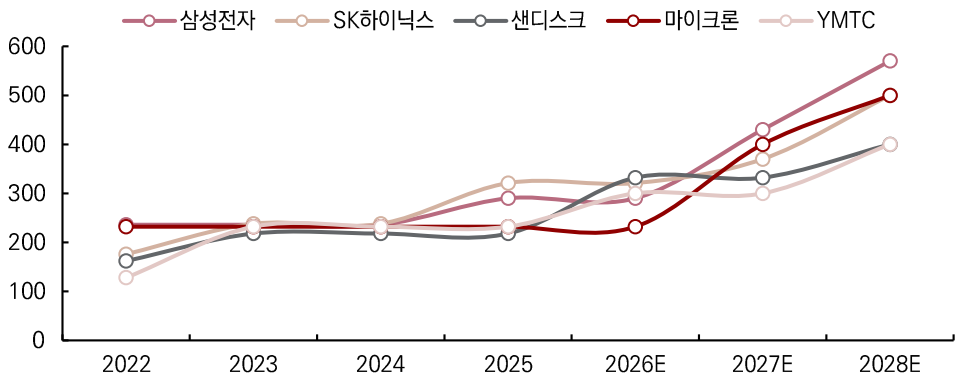
자료: 샌디스크, KUVIC 리서치 2팀

지금은 300단이 대세

2025부터 2026년 현재까지의 **주로 양산되는 단수는 200~230단대**다. 삼성전자의 V8(236단), SK하이닉스 V8 4D PUC(238단), 마이크론 Gen6(232단), 샌디스크 BiCS8(218단), YMTC Xtacking 3.0(232단)이 각 사의 주력 양산 제품군을 형성하고 있다. 단, **2026~2027년에는 300단대로의 본격 전환**이 예상되고 있다. **SK하이닉스는 321단 TLC 4D NAND를 2024년 11월 양산 개시**했으며 2025년 상반기부터 출하했다. 321단 QLC NAND는 2025년 8월 양산을 시작해 2026년 상반기 출시 예정이며, **삼성전자의 V10은 약 430단**으로 알려져 있다. 이처럼 300단대에서 400단 대 혹은 그 이상으로 단수가 늘어날 전망이다. 특히 주목할 점은 단수 경쟁이 1년에 평균 50~70단씩 증가하던 과거와 달리, 2025년 이후 **한 세대당 100단 이상씩 큰 단위로 점프하는 양상**으로 변화하고 있다는 것이다.

2028년 이후에는 **500단~1,000단대로 경쟁이 확장**될 수 있다. **삼성전자는 2030년까지 1,000단 NAND 양산을 목표**로 하고 있으며, **키옥시아는 이보다 더 공격적으로 2027년 1,000단 3D NAND 양산을 선언**했다. 삼성전자는 이미 **2026년 5월 900단 시제품을 공개**하였으며, 이는 CMB(Cell Multi-Bonding) 기법을 통해 별도의 NAND 셀 웨이퍼 두 장을 정밀 본딩하여 구현한 것이다.

그림 24. 기업별 적층 단수 변화



자료: 각 사, KUVIC 리서치 2팀

적층 단수 증가의 영향

이와 같은 NAND의 적층 단수 변화 흐름은 향후 메모리 산업에 미치는 영향은 **'CAPEX 구성', 'HBF'** 면에서 있을 것으로 보인다. 먼저 **CAPEX 구성의 변화**는, 단순 CAPA 증설이 줄고 **하이브리드 장비·고종횡비 식각 장비·정밀 계측 장비에 대한 투자 비중이 커지면서, 후공정 장비 업체와 정밀 공정 장비 업체의 수혜가 집중**될 가능성이 있다.

또한 앞서 언급한 기업 간의 경쟁 구도도 주목해야한다. **단수가 곧 비트당 원가를 결정**하기 때문에, 단수 경쟁에서 뒤처지는 업체는 가격 경쟁력 측면에서 빠르게 도태된다. **SK하이닉스가 적층 단수 1위 자리를 유지**하고, 키옥시아와 샌디스크가 BiCS10로 추격에 나서며, **삼성전자가 V10·V11에서 한 번에 점프하려는 구도**가 향후 1~2년의 핵심 관전 포인트다.

적층 단수가 증가하는 것은 HBF로의 수요로도 이어질 수 있는데, **BiCS9가 HBF 전용 트랙으로 분기된 사례**에서 보듯, 단수 경쟁의 한 축은 일반 NAND(SSD향), 다른 축은 HBF향 고대역폭 NAND로 분화되고 있다. 이는 **NAND 산업이 단순 용량 경쟁에서 용량과 대역폭 두 축의 경쟁으로 재구성**됨을 의미한다.

워크로드에 따른 TLC, QLC의 분업화

NAND플래시 메모리는 셀 하나에 저장하는 비트 수에 따라 세대가 구분된다. 셀당 3비트를 저장하는 **TLC(Triple Level Cell)와 4비트를 저장하는 QLC(Quad Level Cell)**가 대표적이다. 셀에 담는 비트 수가 늘어날수록 동일 면적에 더 많은 데이터를 저장할 수 있어 원가가 절감된다. 반면 **한정된 전압 범위를 더 잘게 쪼개어 제어해야 하므로 읽기·쓰기 속도와 수명이 저하**되는 특성을 지닌다. 최근 AI 추론 워크로드가 데이터 성격에 따라 다변화되면서, 하드웨어 스펙에 맞춰 **TLC와 QLC가 각 계층을 분담하는 최적화 구조가 정착**되고 있다.

고속 응답성과 강력한 수명이 동시에 요구되는 실시간 오프로딩 영역에서는 **고성능 TLC eSSD가 독점적인 지위**를 확보하고 있다. AI 추론 과정 중 디코딩 단계는 기존의 KV 캐시를 지속해서 읽는 동시에, 매 연산 주기마다 새로 생성되는 토큰 데이터를 고속으로 기록해야 하는 읽기·쓰기 혼합 워크로드다. 따라서 추론 인프라의 스토리지는 끊임없이 발생하는 고속 쓰기 압박을 실시간으로 견뎌야 한다. TLC는 QLC와 비교해 **전압 통제 레벨이 절반 수준이어서 반복적인 고속 쓰기 시 절연막 마모 속도가 느리기에 강력한 쓰기 내구성**의 기반이 된다. 또한 TLC는 QLC 특유의 미세 전압 제어 단계인 ISPP(Incremental Step Pulse Programming) 루프를 간소화할 수 있어 응답 속도가 빠르며 이 레이턴시 단축 효과는 서비스 사용자 경험을 결정짓는 핵심 지표인 **TTFT(Time to First Token, 첫 토큰 소요 시간)를 최소화**한다.

표 3. 추론 워크로드 기반 NAND 계층 세분화 맵

계층	대표 데이터	읽기·쓰기 성격	지연 민감도	NAND 유형	대표 제품
Cold Data	모델 가중치, RAG·벡터 DB, 콜드(Cold) KV	읽기 집약·쓰기 희소 (WORM)	낮음	QLC	키옥시아 LC9(245TB), 마이크론 6600 ION(122TB)
Hot/Warm Data	핫(Hot) 동적 KV 캐시, 실시간 KV 갱신·오프로드	읽기·쓰기 혼합	높음	TLC (1~3 DWPD)	키옥시아 CM9(25.6TB), 삼성 PM1753

자료: KUVIC 리서치 2팀

TLC는 Hot Data, QLC는 Cold Data 워크로드 따른 분업

반면 QLC는 데이터 기록 속도가 느리고 쓰기 수명이 상대적으로 짧아 실시간 연산용으로는 부적합하다는 평가를 받았다. 그러나 데이터를 한 번 기록한 뒤 조회와 참조만 반복하는 **WORM(Write Once Read Many) 워크로드에서는 수명 제약이 걸림돌이 되지 않기**에 반복적 쓰기로 인한 셀 열화 압박이 사라지므로 소자의 장기 안정성을 유지할 수 있는 최적의 구동 환경이 조성된다. 에이전틱 AI 아키텍처에서 장기 보존해야 하는 방대한 정적 지식 베이스, 모델 가중치, RAG(검색 증강 생성) 참조용 벡터 데이터 등은 업데이트 주기가 수 주에서 수 개월로 길고 참조 빈도만 극단적으로 높은 대표적 장기 컨텍스트 영역이다. 이러한 고정성 데이터 계층에 **QLC eSSD를 배치하면 쓰기 동작이 최소화**되면서 QLC의 물리적 수명 한계가 소프트웨어적 워크로드 특성 매칭을 통해 자연스럽게 극복된다. 이에 따라 QLC 고유의 강점인 저장 밀도와 비트당 원가 절감 효과가 극대화되며 솔리다임의 TCO 시뮬레이션 결과에 따르면, 기존 HDD를 QLC SSD로 대체할 경우 데이터센터의 **랙 공간을 3분의 1로 줄이고 전력비를 20% 절감하며 총 저장 비용을 31% 낮추는 효과**가 존재한다.

표 4. QLC 도입 및 엔터프라이즈 SSD 현황

기업	QLC 제품 및 양산 현황	주요 데이터센터 및 기업용 솔루션 애플리케이션
SK하이닉스	321단 2Tb QLC 양산 성공	초고용량 엔터프라이즈 SSD(244TB급) 및 UFS 적용 확충
마이크론	G9 2Tb QLC (6-plane, 3.6GB/s)	9650·6600 ION SSD 기반 최대 122TB 라인업으로 HDD 대체 겨냥
삼성전자	V9 QLC 기반 고집적 포트폴리오	BM1743 등 초고용량 엔터프라이즈 SSD 시장 주도
키옥시아	BiCS 아키텍처 기반 QLC 개발	CD9P 등 고효율 61.44TB 엔터프라이즈 SSD 공급
YMTC	multi-Xtacking 기반 QLC 라인업	저비용·고밀도 스토리지 시장 집중 공략

자료: Tom's Hardware, Mordor Intelligence, KUVIC 리서치 2팀

공급 측면에서는 삼성전자·SK하이닉스 양사가 2026년 NAND 웨이퍼 투입량을 오히려 축소(**삼성전자 490만→468만장, SK하이닉스 190만→170만장**)하면서도, **QLC 전환을 통해 실질 비트 출하는 늘리는 전략**을 취하고 있다. 특히 SK하이닉스는 제품 믹스 내 QLC 비중이 경쟁사보다 높은 것으로 알려져 있으며 **321단 QLC를 2026년 상반기부터 글로벌 고객사에 본격 공급**하기 시작했다. 삼성은 QLC 후발주자였으나 286단 V9 QLC의 선단 공정 전환 투자를 다각화하며 추격을 가속하고 있다. 이처럼 기존에 널리 쓰이던 **TLC는 추론의 Warm Data 위주로 지속 성장하고, QLC 또한 Cold Data 계층에 안착하면서 동반 성장**하는 구도를 형성하고 있다.

HBF

HBF는 뭔데?

HBF 기술의 본질

NAND를 HBM처럼 GPU 패키지 내부로 직결시키는 HBF

NAND의 미세화 및 고집적화 흐름은 단순히 가성비 좋은 SSD를 만드는 데 그치지 않고, 새로운 메모리 카테고리의 탄생으로 이어진다. 그것이 바로 최근 업계의 가장 뜨거운 화두인 **HBF(High Bandwidth Flash)**다. HBF는 단순히 셀을 높이 쌓거나 단수를 늘려 원가를 낮추는 기존의 생산성 경쟁과 결이 다르다. **이는 NAND 플래시를 HBM처럼 GPU 가속기 패키지 내부로 직접 끌어들여, 컴퓨터의 메모리 계층 구조 자체를 전면 재편하려는 시도다.**

구조적으로 HBF는 HBM의 제조 방식과 매우 닮아 있다. HBM이 기판 위에 DRAM을 수직으로 쌓아 올리듯, HBF는 **3D NAND 코어 다이를 수직으로 적층하고 최하단에 이를 제어할 로직 다이(컨트롤러)를 배치**한 형태다. 이렇게 완성된 NAND 스택을 **HBM과 나란히 메인 기판(인터포저) 위, 즉 GPU 바로 옆에 배치**한다.

여기서 앞서 언급한 NAND 업계의 300단 돌파 변곡점인 **하이브리드 본딩(CBA) 기술이 빛을 발한다**. **샌디스크 기준 16단으로 구성되는 HBF의 적층 공정**은 NAND 제조사들이 400단 너머의 초고적층 NAND를 위해 선제 투자해 온 CBA 공정과 기술적 뿌리가 같다. 즉, HBF는 아예 새로운 설비를 까는 것이 아니라 **기존에 확보한 첨단 NAND 공정 기술을 응용해 만들어내는 신제품**이다.

HBF 성능의 핵심은 스택 최하단에 위치한 로직 다이에 있다. **일반적으로 8~16개의 채널을 다루는 일반 SSD 컨트롤러와 달리, HBF 컨트롤러는 수천 개의 NAND 채널을 동시에 병렬로 제어한다.** 여기에 AI의 데이터 접근 패턴을 미리 예측해 데이터를 알아서 읽어오는 **프리페치(Prefetch) 알고리즘**까지 탑재한다. 이를 통해 HBM 수준의 막강한 데이터 전송 속도(대역폭)를 유지하면서도, **HBM 대비 8~16배에 달하는 대용량을 훨씬 저렴한 비용으로 구현**하는 것이 HBF의 궁극적 목표다.

최근 샌디스크와 SK하이닉스가 OCP(Open Computing Project) 표준화 킥오프를 통해 공개한 공식 로드맵과 스펙을 살펴보면 성능의 진화 방향이 뚜렷하게 나타난다.

표 5. HBF 세대별 기술 로드맵

구분	GEN 1	GEN 2	GEN 3
스택당 용량	512 GB	최대 1 TB	최대 1.5 TB
읽기 대역폭	1.6 TB/s	2 TB/s 이상	3.2 TB/s 이상
전력 소모 (Gen1 대비)	1.0x	0.8x	0.64x

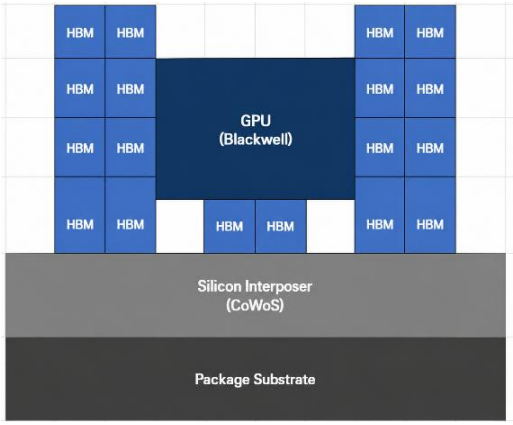
자료: 샌디스크, KUVIC 리서치 2팀

HBM vs. HBF

앞서 살펴보았듯이, HBF는 HBM의 대체재까지는 아니다. 다만, AI 추론형 수요를 감당할 수 있는 대용량의 스펙을 가졌다는 것이 큰 강점이다. 특히 아래의 표를 보면, **HBF 1세대의 스펙은 HBM4 용량 대비 최소 10배는 큰 용량을 보유**하여, AI 메모리 병목을 해결해줄 하나의 솔루션이 될 수 있다. 또한 HBF와 기존 HBM의 큰 차이점 중 하나는 **HBM은 AI 학습용 혹은 추론용 워크로드에 적합한, HBF의 경우 추론형 워크로드에 적합한 메모리**이다. 이는 단순 학습형보다 추론형 LLM 수요가 늘면서, HBF의 역할이 부각되는 이유이기도 하다.

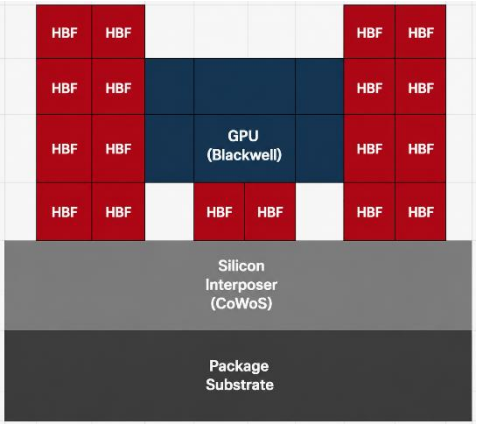
추가로 HBF 시스템의 구조는 HBM4에서 **HBM을 HBF로만 드랍인(Drop-in) 방식으로 교체**하면 되는 구조이다. 특히 두 시스템의 스택당 용량, 총 메모리와 스택당 가격은 다르더라도 **폼팩터와 전력 스펙은 동일하기 때문에 교체하는 식으로 호환이 가능**한 것이다.

그림 25. HBM4 시스템 구조



자료: KUVIC 리서치 2팀

그림 26. HBF 시스템 구조



자료: KUVIC 리서치 2팀

표 6. HBF 버전별 스펙 비교

	HBF Gen1	HBF Gen2	HBF Gen3
기반 소자	NAND (BiCS9)	NAND	NAND
적층 수	16-Hi		
스택당 용량	512GB	1TB	1.5TB
읽기 대역폭	1.6 TB/s	〉2 TB/s	〉3.2 TB/s
레이턴시	~μs (마이크로초)	~μs	~μs
쓰기 내구성	~100K P/E		
전력 (정적)	비휘발성, 낮음	Gen1×0.8	Gen1×0.64
I/O 핀 수	별도 프로토콜		

자료: 샌디스크, Videocardz, KUVIC 리서치 2팀

표 7. HBM, HBF 스펙 비교

	HBM3E (12-Hi)	HBM4 (16-Hi)	HBF Gen1
기반 소자	DRAM	DRAM	3D NAND (BiCS9)
스택당 용량	36 GB	48-64 GB	512 GB
읽기 대역폭	1.2 TB/s	~2 TB/s	1.6 TB/s
레이턴시	~ns (나노초)	~ns	~10 μs
쓰기 내구성	사실상 무제한	무제한	~100K P/E 사이클
전력 (정적)	높음 (리프레시 필요)	높음	낮음 (비휘발성)
적합 워크로드	학습+추론 전반	학습+추론 전반	추론 전용 (읽기 집약)
폼팩터	표준 HBM	HBM4 규격	HBM4와 동일
상용화 시점	양산 중 (2024~)	2026 양산 시작	2026H2 샘플, 2027 시스템

자료: 샌디스크, TrendForce, KUVIC 리서치 2팀

HBF 양산까지 남은 고비

수율과 컨트롤러가 관건

물론 상용화를 위해 넘어야 할 기술적 난제도 적지 않다. 하나는 **적층·패키징 수율**이고, 다른 하나는 **컨트롤러 설계의 진입장벽**이다.

우선 미세한 NAND 칩을 HBM 방식으로 두껍게 쌓다 보니 수율 확보의 문제가 발생한다. HBM의 사례만 보더라도 동일 공정 세대 기준 **일반 DDR 대비 다이 크기가 35~45%가량 커지는데, TSV 배선을 위한 주변부 영역 확장** 때문이다.

이를 극복하기 위해 **SK하이닉스는 VFO 방식을 택했다. TSV를 칩 내부로 관통시키는 대신 칩 외곽을 따라 수직 배선을 연결**하는 패키징 구조를 도입한 것이다. 고난도의 TSV 공정을 우회해 수율 저하를 막고, 그동안 축적한 HBM 패키징 노하우를 TSV 없는 고대역폭 패키지 형태로 확장하려는 전략이다.

삼성전자는 FinFET NAND 기반 접근을 취한다. 파운드리의 첨단 **FinFET(14nm급) 공정을 HBF 최하단의 로직 다이에 적용**해 연산 성능과 전력 효율, 수천 개 채널의 병렬 제어 능력을 극대화하려는 방향이다. 패키징 공정보다 로직 다이 자체의 논리 처리 성능에서 차별화를 두겠다는 구상이다.

두번째 문제인 수천 개의 채널을 칩 하나로 통제하는 컨트롤러 설계 역시 고난도 작업이다. **키옥시아가 선보인 초기 프로토타입**의 경우, **단일 패키지로 5TB라는 압도적 용량은 구현했으나 대역폭이 64GB/s 수준에 그쳐** 초기 HBM보다 느렸고, 전력 효율성(pJ/bit) 역시 **HBM3E 대비 25~40배 격차**가 벌어지는 한계를 보이기도 했다. 업계는 이러한 단점을 보완하여 **2026년 하반기 샘플 공급, 2027년 본격적인 양산 및 시스템 탑재**를 목표로 개발 속도를 올리고 있다.

표 8. HBF 기업별 로드맵

시점	마일스톤
2025.02	샌디스크, HBF 컨셉 최초 공개
2025.08	샌디스크-SK하이닉스 HBF 표준화 MoU 체결 (2025.08.06, 업계 최초 규격 논의 공식화)
2H25	키옥시아 프로토타입 공개 (5TB / 64GB/s)
2026	샌디스크-SK하이닉스 1차 HBF 샘플 출하 (BiCS9 기반)
2027	양산(Mass Production) 목표 / HBF 탑재 1세대 AI 추론 디바이스 등장

자료: 샌디스크, KUVIC 리서치 2팀

HBF의 포지션: 특수목적 고용량 계층

고용량 하이브리드 계층
대체재 아닌 모델 가중치 적재 전용 Read-heavy 계층

이러한 한계를 종합하면, HBF는 **HBM의 대체재라기보다 고용량 하이브리드 계층**으로 정의하는 편이 타당하다. 이 기술은 지연 시간보다 절대적인 메모리 용량이 병목으로 작용하는 **모델 가중치 적재 계층에서 가장 강력한 효용**을 발휘한다. 가중치는 프리필과 디코딩 전 구간에 걸쳐 반복적으로 읽히지만 한 번 적재된 뒤에는 거의 갱신되지 않으므로, 나노초(ns) 단위의 지연 최적화보다 **수 TB에 달하는 파라미터를 단일 패키지 안에 통째로 올려둘 수 있는 용량**이 훨씬 중요하다.

엔비디아의 차세대 Rubin GPU에 탑재될 **HBM4 용량(288GB)만으로는 1조개 이상의 파라미터(1TB 이상)를 가진 초거대 모델을 단일 칩에서 구동할 수 없어** 복잡한 파이프라인 병렬 처리가 강제된다. 반면 **HBF가 제공하는 수 TB급 용량**을 활용하면 이러한 시스템 복잡도와 병렬화 오버헤드를 획기적으로 줄일 수 있다.

결국 HBF는 고속 연산을 담당하는 HBM과 병행되어, **HBM이 대역폭(속도)을, HBF가 용량을 분담**하는 형태로 도입될 가능성이 높다. 다만 2032년 출시 예정인 HBM6 세대부터는 핀당 데이터 전송 속도가 16Gbps로 상향될 예정이어서, HBM과 HBF 간 절대적인 대역폭 격차는 앞으로 더 벌어질 수 있다. 이는 HBF의 장기 포지션이 속도 경쟁보다는 쓰기 빈도가 낮고 읽기 요청이 압도적인 **가중치 적재 전용 대용량 계층(Read-heavy layer)**으로 굳어질 것임을 시사한다.

SSD 풀이 있는데 왜 HBF가 또 필요한가?

여기서 한 가지 의문이 생길 수 있다. **초고속 네트워크 기반의 SSD 풀(ICMS)과 HBF 모두 결국 NAND 플래시**이다. 같은 기술에서 출발했음에도 불구하고 왜 HBF의 필요성이 대두된 것일까?

물리적 거리가 차이
SSD 풀은 랙 단위 네트워크 공유, HBF는 패키지 내부 전용 통로

두 기술은 GPU 가속기로부터의 물리적 거리에서 출발점이 갈린다. **SSD 풀은 가속기 본체와 분리된 랙 단위의 독립 공간**에 위치하며, BlueField-4 DPU를 통한 네트워크(RDMA) 링크로 GPU와 연결된다. 대규모 클러스터 전체를 아우르기에 총용량은 거대하지만, **여러 GPU가 통로를 공유해야 하므로 개별 GPU가 체감하는 실효 대역폭은 제한적**이다.

반면 **HBF는 GPU와 한 패키지 내부(인터포저 위)에 전용 통로로 직결**된다. 이 패키지 근접성 덕분에 **개별 GPU가 독점하는 실효 속도는 SSD 풀보다 한 자릿수 이상 압도적으로 높다**. 이로 인해 HBF는 용량에 제한이 있지만 가속기와 가까워 대역폭이 높다는 고유한 특성을 가지게 된다.

표 9. HBF와 SSD 풀의 계층적 특성 비교

구분	SSD 풀 (ICMS)	HBF
물리적 위치 및 접근	**랙 단위 인클로저, 네트워크(RDMA) 공유**	**패키지 내부(인터포저), 전용 통로 직결**
GPU당 실효 대역폭 (속도)	약 0.1 TB/s (공유 네트워크 링크)	스택당 1.6 TB/s (독점 전용 통로)
총용량 규모	슈퍼포드 기준 최대 ~9,600 TB	스택당 512 GB (8스택 빌드 시 ~4 TB)
최적화 강점 구간	장문맥 적재 (순차·예측 가능)	프리필(Decoding) 중 KV 재사용 (지연 민감)
주요 저장 데이터	클러스터 공유 컨텍스트, 사용 빈도 낮은 콜드(Cold) KV	AI 모델 가중치(Weights), 자주 읽는 핫(Hot)한 정적 KV
기술적 한계	모델이 커질수록 프리필로 가져와야 할 파라미터 정보 접근 시 속도 한계로 GPU 유휴 유발	SSD 풀 대비 시스템 전체 총용량 규모가 작음

자료: NVIDIA, 샌디스크, KUVIC 리서치 2팀

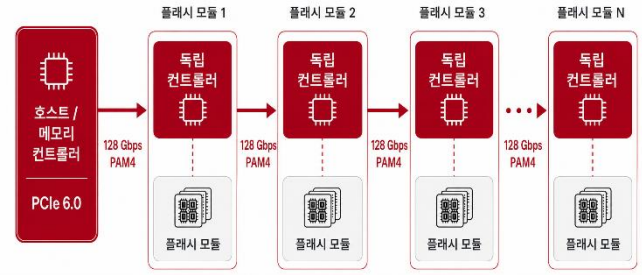
키옥시아 vs. 샌디스크

키옥시아와 샌디스크 모두 HBF를 개발 중인 단계에 있다. 두 HBF 제조 회사의 HBF 프로토타입과 기술, 아키텍처를 비교해 HBF 생산의 우위를 판단해보고자 한다.

키옥시아의 설계는 기존 HBM처럼 TSV로 수직 적층하는 방식이 아닌, [그림 27]처럼 **각 플래시 모듈마다 독립적인 컨트롤러를 배치하고, 모듈 간을 데이지체인 방식으로 직렬 연결하는 분산형 아키텍처**를 채택했다. **128Gbps PAM4 시그널링**을 사용하며, 모듈 수를 늘려도 대역폭이 저하되지 않는 구조다. PCIe 6.0을 호스트 인터페이스로 사용하는 메모리 컨트롤러와 메모리 모듈을 프로토타이핑했으며, **5TB 용량과 64GB/s 대역폭을 40W 미만의 전력으로 달성**함을 검증했다. 또한 키옥시아의 HBF 프로토타입은 **엣지 서버를 타겟으로 하며**, 5G, 6G 셀룰러 네트워크를 통해 IoT 디바이스와 연결되는 모바일 엣지 유닛에서 AI 모델 구동을 위해 DRAM을 보완하는 용도로 설계됐다.

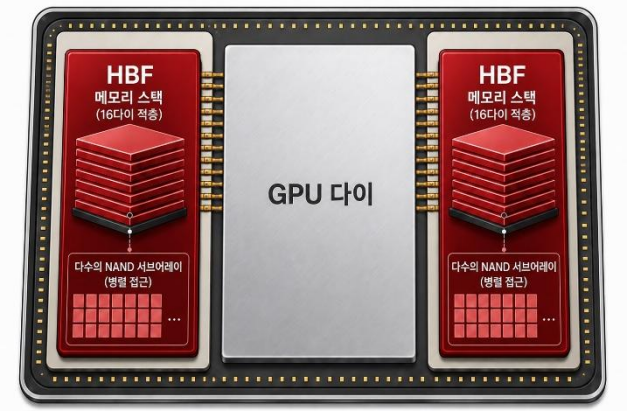
반면 **샌디스크의 HBF는 고대역폭에 특화되고, GPU 패키지 내 탑재**되는 것이 특징이다. 샌디스크 1세대 HBF의 경우, **읽기 대역폭 최대 1.6TB/s, 다이당 256GB, 16다이 스택 기준 총 512GB 용량**의 스펙을 목표로 하고 있으며, 스택은 **HBM4의 물리적 폼팩터, 전력 프로파일, 스택 높이와 일치**하도록 설계되었다.

그림 27. 키옥시아 HBF 아키텍처



자료: 키옥시아, KUVIC 리서치 2팀

그림 28. 샌디스크 HBF 아키텍처



자료: 샌디스크, KUVIC 리서치 2팀

표 10. 샌디스크 HBF 세대별 스펙 비교

	HBF Gen1	HBF Gen2	HBF Gen3
용량 (스택당)	512 GB	1 TB	1.5 TB
읽기 대역폭	1.6 TB/s	〉2 TB/s	〉3.2 TB/s
확장 구성 (8스택)	4 TB	8 TB	12 TB
전력 프로파일	HBM4와 동일	Gen1 × 0.8	Gen1 × 0.64
적층 방식	TSV + CBA 수직 적층	동일	동일
폼팩터	HBM4와 동일	동일	동일
타겟 용도	AI 추론 (데이터센터)	동일	동일

자료: 샌디스크, KUVIC 리서치 2팀

샌디스크 HBF의 아키텍처를 보면, **거대한 NAND 어레이를 수많은 서브어레이로 분할**해 각 서브어레이에 병렬 접근하는 구조다. **SK하이닉스는 TSV 및 수직 적층 패키징 기술을 제공**하고, **샌디스크는 BiCS NAND와 CBA(CMOS Bonding Array) 아키텍처를 제공**한다. HBF는 **HBM4와 동일한 핀 레이아웃, 치수, 전력 프로파일**을 사용하므로 하드웨어 제조사가 대규모 시스템 재설계 없이 채택할 수 있도록 설계됐다. 또한 샌디스크는 BiCS9 기술을 HBF에 적용할 예정이며, **BiCS9는 기존 8세대 대비 NAND 인터페이스 속도를 33% 개선**하여 4.8Gb/s를 달성한다. 참고로 현재 양산 중인 HBM3E의 핀당 인터페이스 속도는 9.2Gb/s다.

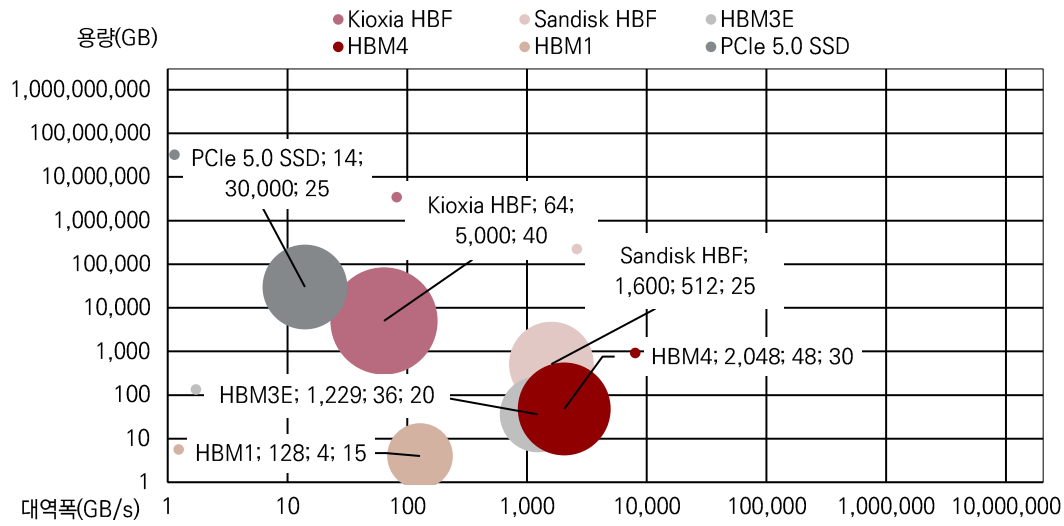
표 11. 샌디스크 HBF 세대별 스펙 비교

	키옥시아 HBF	샌디스크 HBF
대역폭 (단위당)	64 GB/s (모듈 1개)	1,600 GB/s (스택 1개)
용량 (단위당)	5 TB (모듈 1개)	512 GB (스택 1개)
전력	〈40W (모듈)	HBM4 수준 (~20~30W 추정)
인터페이스	PCIe 6.0 (외부 버스)	TSV 직접 연결 (GPU 내부)
GPU 패키지 내 탑재	불가 (외부 모듈)	가능 (HBM4 폼팩터)
아키텍처	데이지체인 직렬	TSV 병렬 서브어레이
타겟 시장	엣지 AI, IoT	데이터센터 AI 추론

자료: 샌디스크, KUVIC 리서치 2팀

그래프는 HBM 세대별 제품군(HBM1·HBM3E·HBM4)과 두 HBF 공급사(키옥시아·샌디스크)의 1세대 HBF, 그리고 비교군인 PCIe 5.0 SSD의 스펙을 대역폭-용량 평면 위에 동시 매핑한 결과이다. **X축은 제품 1단위(스택 또는 모듈)당 읽기 대역폭**(GB/s, 로그 스케일), **Y축은 제품 1단위당 저장 용량(GB, 로그 스케일)**, 그리고 **점의 크기는 단위당 전력 소비량(W)**을 의미한다. 양 축에 로그 변환을 적용한 것은 HBM(수십 GB)과 SSD(수만 GB) 사이의 용량 격차, 그리고 PCIe SSD(수십 GB/s)와 HBM(수천 GB/s) 사이의 대역폭 격차가 3~4자릿수에 이르기 때문에, 모든 제품을 한 평면에 동등하게 비교하기 위함이다.

그림 29. 기업별 HBF, HBM 스펙 포지션



자료: 키옥시아, 샌디스크, TrendForce, KUVIC 리서치 2팀

키옥시아와 샌디스크 HBF의 스펙을 비교하면, **키옥시아는 대역폭이 64 GB/s로 매우 낮고, 샌디스크의 HBF 1세대는 HBM3E보다는 높고, HBM4보다는 낮은 대역폭**을 지닐 전망이다. 특히 키옥시아와 샌디스크 모두 HBF의 기술을 좌우할 **CBA 본딩 기술을 보유하고 있다는 것이 공통점**이며, 현재 상태로는 키옥시아는 샘플 제시 단계, 샌디스크는 로드맵을 제안할 단계임을 고려해야한다. 요약을 하면, 키옥시아의 HBF와 샌디스크의 HBF는 **아키텍처, 목표 시장, 성능, 포지셔닝 면에서 아예 다른 구조**를 보인다.

HBF는 도입 우선순위인가?

CPU(가동률)
CPO(통신)
HBF(용량)
서로 다른 병목
동시 공략

GPU 성능은 칩 하나로 결정되지 않고 HBM, 첨단 패키징(CoWoS) 등 밸류체인 전반의 고도화에 좌우되며, 그 과정에 막대한 비용이 따른다. 그래서 업계는 GPU 자체를 키우는 대신 시스템 인프라를 혁신해 가속기 성능을 끌어올리는 길을 택해 왔고, **CPU 비중 확대 → CPO → HBF**가 그 흐름으로 거론된다.

순서대로 등장하다 보니 HBF를 CPU 병목과 CPO가 모두 해결된 뒤에야 차례가 오는 최후순위 기술로 보는 시각이 있다. 그러나 세 기술은 한 줄로 차례를 기다리지 않는다. **각각 가동률·통신·용량이라는 서로 다른 병목을 동시에 겨냥하는 병렬 트랙**이다.

표 12. CPU, CPO, HBF의 도입 범위

병목	해결책	핵심 질문
GPU 가동률 저하	CPU 비중 확대 (CPU:GPU 1:4 → 1:2)	GPU를 쉬지 않게 쓰는가
GPU 간 통신 지연 (Scale-out 보틀넥)	CPO (광학 인터커넥트)	묶인 GPU를 빠르게 잇는가
GPU당 용량 부족 (조 단위 모델·KV 캐시)	HBF	GPU 하나가 많이 담는가

자료: KUVIC 리서치 2팀

HBF가 독립된 트랙인 근거는 세 가지로 정리된다.

① HBF는 다른 종류의 문제를 푼다

HBF만이 절대 용량을 해결

CPU 비중 확대와 CPO는 이미 가진 자원을 더 알뜰하게 쓰는 효율 최적화다. CPU를 늘려 노는 GPU를 깨우고, CPO로 멀리 떨어진 GPU 간 통신 지연을 줄인다. 반면 **HBF는 GPU 한 개가 담는 절대 용량의 천장을 끌어올린다**.

문제의 성격이 다르므로 HBF는 앞선 두 기술의 완성을 기다릴 이유가 없다. 가동률이 100%에 이르고 수백 개의 GPU가 단일 칩처럼 통신하더라도, 모델이 메모리에 적재되지 못하면 소용이 없다. **차세대 초거대 모델의 추론 용량이 36TB에 이른다면, NVL72 랙이 공유하는 HBM 풀 13.5TB로는 모델 자체가 올라가지 못한다. 절대 용량 부족은 효율 최적화가 건드리지 못하는 별개의 문제**이며, HBF만 이를 직접 겨냥한다.

② HBF가 올라탈 데이터 경로는 이미 깔리고 있다

GIDS가 길을 먼저 깐다

HBF는 NAND 플래시를 패키지 안에 통합해 대용량 메모리로 쓰는 기술이다. 다만 GPU가 데이터를 읽을 때마다 멀리 있는 CPU의 승인을 기다린다면 NAND 고유의 접근 지연이 더 커진다. 따라서 HBF가 제 성능을 내려면 **GPU가 CPU를 거치지 않고 스토리지에 직접 접근하는 경로가 전제**된다.

이 경로는 이미 만들어지고 있다. 엔비디아는 차세대 플랫폼 **Vera Rubin(2026년 하반기)부터 GIDS(GPU-Initiated Direct Storage Access)를 도입**하는 것으로 알려져 있다. 기존 GDS가 CPU를 통해 스토리지에 데이터를 요청했다면, **GIDS는 GPU가 직접 읽기 명령을 내려 CPU와 DRAM을 우회**한다. **데이터 경로(2026년)가 HBF 상용화(2027년)보다 한발 앞서 깔리는 셈**이며, HBF는 이렇게 먼저 놓인 길 위에 올라타 성능을 낸다.

③ 스케일업·CPO와 HBF는 서로를 보완한다

랙 단위 스케일업이 고도화되고 CPO로 GPU 간 통신이 빨라지면 HBF가 필요 없어진다는 우려가 있다. 그러나 두 흐름은 보완 관계에 가깝다.

H³ 아키텍처의 임팩트

첫째, 인터커넥트가 아무리 발전해도 **HBM 자체의 용량 부족은 풀리지 않는다**. NVL72는 13.5TB, 차세대 Rubin NVL144도 20.7TB에 그쳐 36TB 수준의 추론 용량 요구에 못 미친다. 통신을 최적화해도 용량을 채우려면 결국 GPU를 더 사야 하고, 이는 필요 없는 연산 자원까지 떠안는 가장 비싼 방식이다. HBF는 **GPU 추가 없이 GB당 HBM의 약 10분의 1 비용으로 용량을 확보**한다.

둘째, HBF는 **묶어야 할 가속기 수 자체를 줄여 스케일업·CPO의 부담을 덜어**준다. SK하이닉스가 IEEE 논문에서 공개한 **H³ 아키텍처(B200에 HBM3E 8스택과 HBF 8스택 혼합)**가 이를 정량으로 보여준다.

표 13. H³ 아키텍처

지표 (1,000만 토큰 KV 캐시 기준)	HBM 단독 대비
토큰 처리량	6.14배
동시 쿼리 처리(배치)	18.8배
와트당 성능	2.69배
필요 GPU 수	32개 → 2개

자료: IEEE, KUVIC 리서치 2팀

토큰 처리량 개선폭은 100만 토큰에서 1.25배였다가 1,000만 토큰에서 6.14배로 커져, 데이터 규모가 클수록 효과가 두드러진다. **과거 GPU 32개와 그에 딸린 HBM 풀이 필요했던 워크로드를 단 2개로 소화**하는 셈이다. 묶을 노드가 줄면 스케일업 규모와 CPO가 감당할 통신 부하도 함께 낮아진다.

④ HBF와 CXL은 데이터 성격에 따라 일을 나눈다

모델 가중치는 HBF(불변), KV 캐시는 CXL(재기록)

용량 부족의 또 다른 대안으로 CXL(Compute Express Link) 기반 D램 확장이 함께 거론된다. 둘은 같은 자리를 두고 경쟁하기보다 **데이터의 재기록 빈도에 따라 역할을 나눈다**. **NAND는 약 10만 회의 쓰기·소거면 수명이 다하는 반면 D램은 사실상 무제한 재기록이 가능**하기 때문이다.

표 14. 데이터에 따른 메모리 계층

데이터	특성	메모리 계층
모델 가중치	학습 후 고정된 읽기 중심 대용량 데이터	HBF
KV 캐시	토큰마다 갱신·삭제되는 재기록 집약 데이터	CXL / DRAM
활성 데이터 (Activation)	프로세서와 가장 빈번히 주고받는 데이터	HBM

자료: KUVIC 리서치 2팀

한번 적재되면 변하지 않는 모델 가중치는 NAND의 내구성 한계를 우회할 수 있어 HBF에 두는 편이 효율적이다. 끊임없이 생성·소멸하는 **KV 캐시는 CXL 메모리 풀이나 D램이 맡는 편이 합리적**이다. 실제로 KV 캐시를 CXL로 오프로드한 연구에서는 프리필 단계 GPU 가동률이 큰 폭으로 높아지고 처리량이 배 단위로 늘었다. **SK하이닉스 H³가 NAND 지연을 SRAM 기반 버퍼(Latency Hiding Buffer)로 가리면서**, 읽기 요청이 압도적인 정적 KV 캐시만 골라 HBF에 배치한 것도 같은 분업 원칙이다.

요약하면 **HBF는 세 좌표를 갖는다**. GPU가 스토리지를 직접 제어하는 GIDS라는 데이터 경로 위에서 작동하고, **스케일업·CPO가 못 채우는 절대 용량을 저비용으로 메우며**, CXL과는 데이터 성격에 따라 역할을 나눈다. 결국 HBF는 다른 기술에 밀려 도입이 미뤄지는 후순위 대안에 머물지 않는다. GPU 중심으로 재편되는 차세대 AI 인프라에서 **절대 용량과 구축 비용이라는 병목을 풀기 위해 반드시 자리 잡을 독립 계층**이다.

HBF 시장규모 분석

본 리서치 팀 추정 결과, HBF TAM은 추론 워크로드에서의 중요도에 따른 가격 협상력에 따라 **Base Case 401.4억 달러**에서 **Bull Case 783.4억 달러** 수준으로 결정될 것으로 전망한다. Base Case의 경우에도 **26년 글로벌 NAND TAM의 25% 수준**으로, HBF 상용화 시 NAND 시장에서 매우 큰 비중을 차지할 것으로 판단된다.

HBF Bit 수요 추정

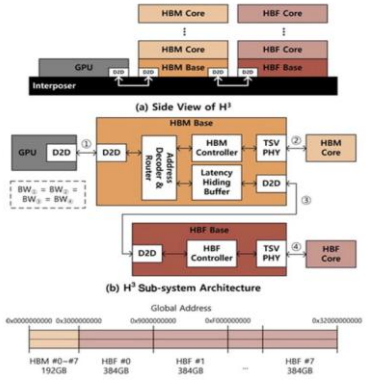
HBF의 Bit 수요 TAM은 102.4EB일 것으로 추정된다. 2027년 이후 GPU 판매량은 컨센서스인 **1,000만대**로 가정하였으며, 샌디스크가 제안한 설계 구조를 바탕으로 32GB 용량의 SLC NAND die 16개를 적층한 **512GB 용량의 HBF 1스택**을 기본 단위로 설정하였다. 현재 단일 GPU에 패키징될 HBF 스택 수에 대해서는 표준이 정해지지 않은 상태이다. 대표적으로 2개의 GPU가 48GB GDDR(4GB x 12) 및 4TB HBF(512GB x 8)와 연결되는 구조나, 96GB HBM(24GB x 4)과 가 결합하는 아키텍처에서는 GPU당 4개의 HBF 탑재가 도출된다. 반면 SK하이닉스가 IEEE에서 발표한 H3 아키텍처의 경우 GPU당 8개의 HBM과 8개의 HBF를 동수 탑재하는 시뮬레이션을 제시하였으며, 이를 통해 HBM 단독 탑재 대비 1,000만 토큰 처리 시 TPS 6.14배, 동시 쿼리 처리 용량 18.8배, 전성비 2.69배라는 압도적인 성능 개선을 선보인 바 있다.

그림 30. HBF Q 추정 주요 가정

1) HBF Q 추정 주요 가정	
27 년 이후 GPU 판매량 가정	1,000 만개
HBF 1 스택당 용량	512GB
GPU 당 HBF 스택수 가정	8 개
프리필용 GPU 점유율 가정	25%
HBF 수요 추정	**102.4EB**

자료: KUVIC 리서치 2팀

그림 31. SK하이닉스가 공개한 H3 구조



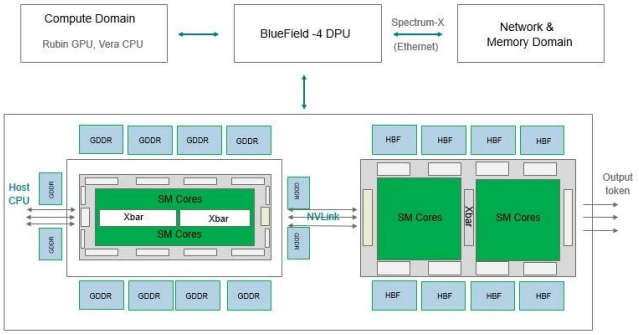
자료: SK하이닉스, KUVIC 리서치 2팀

시장 일각에서는 가성비를 고려해 GPU당 4개 탑재안을 무게 있게 다루고 있으나, 2팀은 **HBM과 HBF가 1:1 대칭 구조**로 탑재되는 메커니즘과 엔비디아의 하이엔드 전략에 주목하여 최종적으로 **GPU당 8개 탑재**가 주류를 이룰 것으로 판단한다. 대규모 언어 모델(LLM)의 파라미터 수가 기하급수적으로 증가함에 따라 고속 연산을 수행하는 HBM과 초고용량을 지원하는 HBF의 탑재량은 모델 스케일에 비례하여 동반 상승해야 하기 때문이다. 비록 HBF가 HBM에 담기지 못한 캐시 메모리를 오프로드하는 역할을 수행하지만, 실시간 데이터 처리를 담당하는 작업 메모리 성격의 HBM과 상대적으로 콜드 데이터에 가까운 가중치를 저장하는 HBF는 메모리 자원의 성격이 근본적으로 다르다. 따라서 가성비를 이유로 기존 하이엔드 GPU의 HBM 탑재 개수를 줄이고 이를 HBF로 대체하는 아키텍처는 심각한 병목 현상을 야기할 위험이 크다. 결국 **기존 HBM의 물리적 채널을 유지한 채 동수의 HBF를 추가 탑재하는 방향**이 기술적으로 타당하며, 비용 절감보다 경쟁사 대비 **압도적인 성능 우위를 지향해 온 엔비디아의 차세대 제품 전략**을 고려할 때도 추론 워크로드 강화를 위한 8개 탑재가 지배적인 규격이 될 가능성이 높다.

한편 2026년을 기점으로 전체 AI 워크로드에서 추론이 차지하는 비중이 학습을 추월할 것으로 예상되는 가운데, 당사는 HBF의 탑재 대상을 추론 워크로드 중에서도 **프리필(Prefill) 단계**를 수행하는 GPU로 한정하여 수요를 추정하였다. 추론의 두 번째 단계인 디코딩(Decoding) 워크로드의 경우 지연 시간 극단화에 특화된 Groq의 LPU 같은 전용 랙 솔루션이 이미 제안되어, **HBF는 프리필 최적화 가속기에 집중 탑재**될 것으로 판단된다. 프리필은 모델이 사용자의 프롬프트 전체를 일괄적으로 읽어 들이는 과정으로, 나노초 단위의 미시적 지연 시간 최적화보다는 방대한 모델 가중치(Weight)를 단일 칩 메모리 풀에 온전히 올려둘 수 있는 절대적인 저장 용량이 핵심이다.

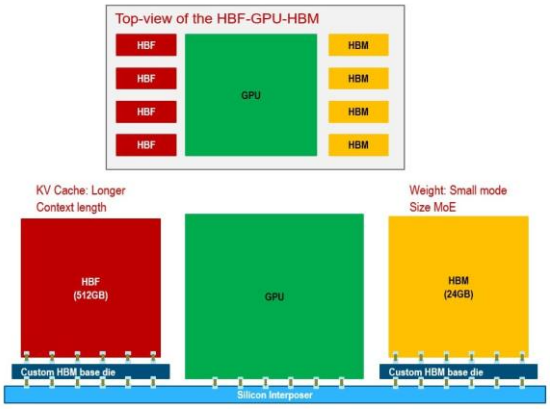
HBF가 제공하는 수 테라바이트 수준의 압도적인 용량은 대규모 네트워크 트래픽을 유발하는 복잡한 파이프라인 병렬 처리의 필요성을 획기적으로 줄여줄 수 있다. 즉, HBF는 HBM을 전면 대체하는 것이 아니라 초거대 모델의 가중치를 단일 칩 수준에서 수용해 프리필 단계의 물리적 메모리 병목을 해소하는 특수 목적의 고용량 메모리로 도입될 전망이다. 이에 따라 전체 추론 워크로드 비중을 보수적으로 50%로 가정하고, 이를 다시 프리필과 디코딩 단계로 양분하여 최종적으로 **프리필용 GPU 점유율을 25%**로 산출하였다. 결과적으로 2027년 이후 연간 1,000만대의 GPU 수요 중 25%에 해당하는 250만대의 프리필용 GPU에 스택당 512GB의 HBF가 각각 8개씩 탑재됨으로써, 최종적인 HBF Bit 수요는 102.4EB 수준의 거대한 신규 시장을 형성할 것으로 전망된다.

그림 32. GDDR-HBM 구조



자료: JP모건, KUVIC 리서치 2팀

그림 33. HBM-GPU-HBF 구조



자료: JP모건, KUVIC 리서치 2팀

HBF GB 당 ASP 추정

표 15. HBF ASP 분석 세부 가정

2) HBF ASP 분석 세부 (HBM 기반)					
서버용 DRAM 1GB BOM	$22.32	Trendforce 26년 5월	서버용 NAND 1GB BOM	$0.37 $	Trendforce 26년 5월
마진	$17.86	2Q26 DRAM 영익률 80% 전망	마진	$0.24 $	26E NAND 영익률 65% 전망
원가	$4.46		원가	$0.13 $	
HBM4 1GB BOM	$39.20	실리콘 애널리스트 추정	HBF 1GB BOM 추정	3.92 ~$7.84	
마진	$23.52	2Q26 HBM4 영익률 60% 전망	마진	$1.24~$5.16	
DRAM 원가	$13.39	범용 DRAM 대비 캐파 3배 소모	SLC NAND 원가	$0.39	TLC → SLC 전환 기회비용 3배
후공정	$2.29		후공정	$2.29	HBM과 동일

자료: KUVIC 리서치 2팀 추정

샌디스크 및 SK하이닉스가 제시한 가이드라인과 정밀한 원가 구조 분석을 결론으로 결합하여 **HBF의 1GB당 예상 판매단가(ASP) 범위를 도출**하였다. 두 기업의 전망에 따르면 **HBF의 가격은 HBM의 5분의 1에서 10분의 1 수준에서 형성될 것으로 예측**된다. 시장에서 예상하는 차세대 HBM4의 1GB당 가격이 39.20달러 수준임을 감안할 때, HBF의 1GB당 단가는 단순 계산상으로 **3.92달러에서 7.84달러** 사이에서 결정될 가능성이 높다. 그러나 2팀은 이 단순 비교 방식에서 나아가, 시장 공급망 데이터와 공정 특성을 반영한 원가 역산 모델을 구축하여 논리적이고 타당한 가격 밴드를 검증하였다.

이를 위해 시장조사기관 트렌드포스(Trendforce)의 최신 가격 데이터와 메모리 업계의 영업이익률 컨센서스를 적용하여 **서버용 DRAM 및 NAND의 제조원가를 역산**하였다. HBF는 성능 최적화를 위해 eSSD의 주력인 TLC(Three-Level Cell)나 QLC(Quad-Level Cell) 대신 32GB 용량의 SLC(Single-Level Cell) NAND die를 채택하지만, **원가 분석의 기준이 되는 1GB당 단가는 TLC 및 QLC의 혼합(Blended) ASP를 기준으로 산정**하는 것이 타당하다. 현재 시장에 형성된 순수 SLC 가격은 구형 공정과 소규모 특수 수요로 인해 기형적으로 높게 왜곡되어 있기 때문이다. SLC die는 물리적 공정 전환 없이 소프트웨어적인 셀 제어 설계만 달리하면 기존 고성능 TLC·QLC 생산 라인을 그대로 활용해 대량 생산할 수 있으므로 범용 NAND 가격을 모형의 출발점으로 삼았다.

HBF의 원가 구조를 정교화하기 위해 하이엔드 메모리인 **HBM4의 원가 산정 방식을 벤치마크 모델로 차용**하였다. HBM4의 원가는 전체 가격에서 마진을 제외한 금액이 DRAM die 원가와 TSV(관통전극) 본딩 및 테스트를 포함한 후공정(Advanced Packaging) 비용으로 구성된다고 가정하였다. 이때 HBM 내부 DRAM die 원가는 생산 수율(Net Die Yield)이 범용 DRAM의 30% 수준에 불과해 생산 캐파(CapEx)를 약 3배 소모한다는 점을 반영하여 **일반 서버용 DRAM 원가의 3배**인 13.39달러를 적용했으며, 여기에 2.29달러의 후공정 비용이 가산된다. HBF에 탑재될 SLC NAND die 원가는 기회비용 논리를 적용했다. 기존 TLC 라인을 활용해 저비용 대량생산이 가능하더라도, 셀당 3비트를 저장하던 TLC에서 1비트만 저장하는 SLC로 전환되면 **셀당 저장 용량이 3배 감소**하게 된다. 비록 SLC가 TLC 대비 읽기·쓰기 지연 시간과 쓰기 내구성(TBW) 측면에서 압도적인 기술적 우위를 지니지만, 웨이퍼 면적당 용량 손실이라는 물리적 기회비용을 상쇄하기 위해 **SLC die 원가를 서버용 NAND 원가(0.13달러)의 3배**인 0.39달러로 산정하였다. 아울러 HBF가 HBM과 유사한 후공정 구조를 공유한다고 판단하여 후공정 비용은 HBM과 동등한 2.29달러를 반영하였다.

이상의 원가 구조를 종합하면, HBF가 제조 기업 관점에서 적자를 면하고 흑자 마진을 확보하기 위한 최소한의 **1GB당 하한선 가격은 die 원가와 후공정 비용의 합산인 2.68달러**로 계산된다. 이를 기준으로 볼 때, 샌디스크와 SK하이닉스가 제안한 가격 밴드의 최하단이자 HBM4 가격의 **10분의 1 수준인 3.92달러는 제조사에 1.24달러의 안정적인 이윤을 보장**하므로 충분히 시장에 안착 가능한 가격대로 판단된다. 나아가 글로벌 공급 부족(Shortage) 국면 속에서 2026년 서버용 NAND 시장의 전체 **영업이익률이 65% 수준**까지 치솟을 것으로 전망됨에 따라, HBF 공급사들이 고부가 신제품에 대해 기존 NAND 사업부와 동등한 수준의 수익성을 요구할 가능성도 상존한다. 원가 2.68달러에 65%의 영업이익률을 역산하여 반영할 경우 HBF의 가격은 **1GB당 최대 7.65달러까지 상승**할 수 있다. 다만 현재의 극심한 NAND 쇼티지와 그로 인한 65%의 초고이익률은 역사적 평균에서 다소 과도하게 벗어난 단기적 현상일 수 있음을 고려해야 한다. 따라서 당사는 원가 모델 상 보수적이면서도 제조사 참여가 합리화되는 **3.92달러를 Base Case ASP로 설정**하고, NAND 업황의 강세와 초기 독점적 지위를 가정한 **7.65달러를 Bull Case ASP로 최종 결정**하여 전체 TAM 전망의 신뢰성을 확보하였다.

COMPANY ANALYSIS

—

Not Rated

테스 (095610)

테스형, NAND가 왜 이래

NAND 고단화 트렌드, ACL 장비에 수혜

테스(TES)는 3D NAND 고단화 공정의 최대 수혜를 입는 **ACL 장비** 전문 기업이다. 증착 공정의 **PECVD 장비의 일종**으로, **NAND 적층수가 300단 이상**으로 올라갈수록 식각 공정에서의 포토레지스트 마모로 인한 패턴 붕괴를 방지하는 강고한 **탄소 하드마스크(ACL)의 두께와 공정 수가 필수적으로 늘어난다.** 이에 따라 테스가 핵심 NAND 설비 투자에서 확보하는 단위당 수주 단가는 **일반 DRAM 투자의 약 1.5배 높은 수준**에 형성된다. 전사 매출액 중 무려 70%에서 80% 상당을 안정적으로 점유하는 PECVD 핵심 장비 믹스는 초고단화 NAND와 HBF 아키텍처 도입 흐름 속에서 테스의 압도적인 장기 실적 성장을 견인하는 강력한 지렛대 역할을 완성할 것이다.

고객사 설비 투자에 따른 장비 수주 모멘텀

핵심 고객사의 선단 공정 증설 일정이 선명하게 구체화되고 있다. 먼저 **삼성전자**는 **2026년 하반기부터 시안 2팹에 약 30K에서 45K 규모의 V9 V-NAND 전환 설비 투자**를 본격 개시할 예정이며, 국내 팹 역시 **V10 신규 전환**을 긴밀하게 추진 중이다. **SK하이닉스** 또한 핵심 **M15 팹**을 중심으로 차세대 300단대 NAND 공정 전환 수주를 가쁘게 이어가고 있다. 나아가 **2027년 신규 가동될 삼성전자의 P5 팹향 하이브리드 증설 투자**가 대형 수주 성과로 순차 결실을 맺을 것으로 기대되어 테스의 중장기 실적 우상향 구도는 확실하다.

하이브리드 본딩 협력과 BSD 장비 모멘텀

HBF 후공정의 구현의 필수 전제인 구리-구리 직접 연결 방식의 **하이브리드 본딩** 공정에서도 테스의 지위가 독보적이다. 테스는 박막 증착 및 플라즈마 클리닝 기술력을 결합해 글로벌 1위 본딩 기업인 **한미반도체와 하이브리드 본더 장비를 공동 개발하여 2027년 출시**를 정조준한다. 또한, 초적층 NAND에서 die 두께 감소로 인해 필연적인 기판 휘어짐을 상쇄 제어하는 특수 **BSD(Backside Deposition) 장비**를 자체 설계 완료하여 **주요 고객사 NAND 라인에 이미 2025년부터 양산 공급 중**이며 **HBM 및 파운드리 확장**도 함께 타진하여 기술 독점 영역을 크게 공고히 다져가고 있다.

Earnings and valuation metrics

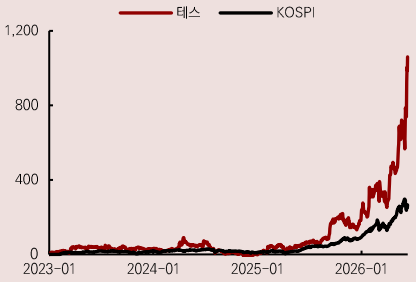
결산기 (12월)	2021	2022	2023	2024	2025
매출액 (십억원)	375	358	147	240	351
YoY	-	*-5%*	*-59%*	*63%*	*46%*
영업이익 (십억원)	62	56	-6	39	58
YoY	-	*-10%*	*-111%*	*-753%*	*50%*
영업이익률 (%)	32.8	34.2	21.7	45.7	43.6
당기순이익 (십억원)	74	47	2	43	57
EPS (원)	3,743	2,366	79	2,158	2,885
P/E (배)	8.0	6.6	253.7	7.2	15.4

주: K-IFRS 연결 기준, 순이익은 당기순이익
자료: KUVIC Research 2팀

Stock Information

시가총액	3조 5,274억원
발행 주식 수	1,936만주
유동주식비율	62.1%
52주 최고가	182,200원
52주 최저가	23,900원
외국인 지분율	9.6%
KOSPI	8,123.0
KOSDAQ	1,029.1

Price Trend



KUVIC Research Team 2

메일 kuvic_korea@naver.com

팀장 44기 Senior 김서정
팀원 44기 Senior 김현진
팀원 43기 Senior 정상엽

Who We Are

KUVIC

Compliance Notice