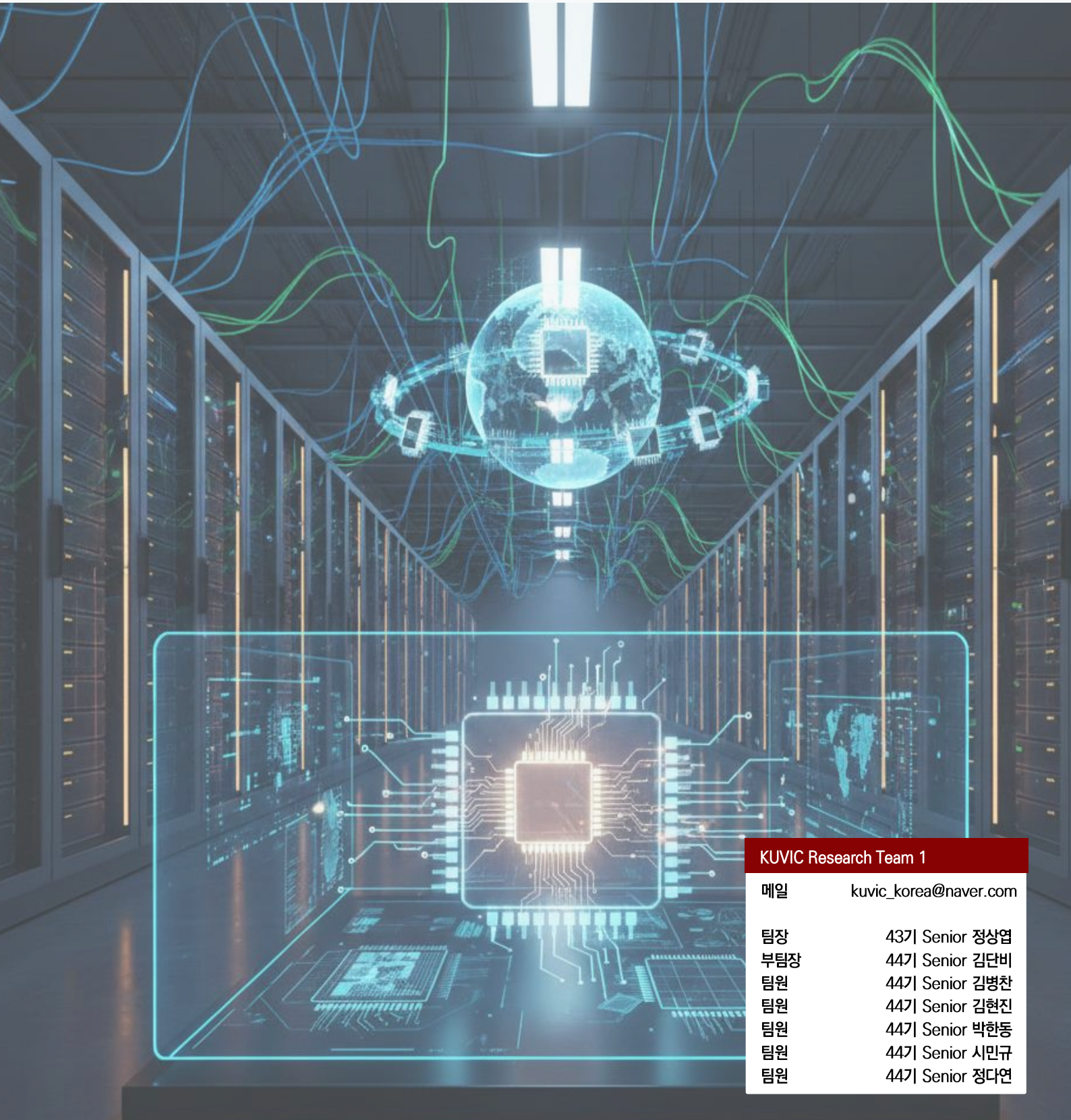
Industry Indepth| 2026.01.26

반도체 (비중확대)

이번 사이클 운동 많이 될 거야



KUVIC Research Team 1

메일 kuvic_korea@naver.com

팀장 43기 Senior 정상엽
부팀장 44기 Senior 김단비
팀원 44기 Senior 김병찬
팀원 44기 Senior 김현진
팀원 44기 Senior 박한동
팀원 44기 Senior 시민규
팀원 44기 Senior 정다연

CONTENTS

Summary

AI 수요

① 26년 AI 수요는 추론을 중심으로 급증, 8,480억 달러의 CAPEX 집행을 전망함.

- AI 서비스 이용량 및 모델 크기 증가로 지속적 컴퓨팅 비용이 발생하는 추론의 비중이 학습보다 커짐

- 26년은 에이전틱 AI를 위주로 토큰 소비량이 매년 10배 이상 증가 전망, AI 발전단계의 최종 목적지로 피지컬 AI가 제시되는 중

- CAPEX 투자 회수 가능성에 대한 의구심 존재하나, 수익 기반 다각화 및 추론 비용 급감으로 해결하리라는 반론 존재

- 글로벌 AI 데이터센터 CAPEX는 26년 기준 8,480억 달러, 27년 1조 240억 달러로 추정되며, 이는 블랙웰 GPU 26년 1,970만 장, 27년 2,386만 장에 달하는 규모로 TSMC CoWoS 규모와 유사

- TSMC의 CoWoS가 AI 사이클의 주요 병목으로, AI 가속기 시장에 초과 수요가 존재함을 시사.

- 현재 AI 매출은 막대한 CAPEX로부터 비롯된 단기 운영 비용조차 보전하기 어려운 상태로, 26년은 추론 비용에서의 급감과 글로벌 AI 매출 연 50% 이상 고성장이 확인돼야.

엔비디아와 ASIC

② 26 년 엔비디아는 블랙웰보다 3~4 배 연산 능력이 개선된 Rubin 및 추론에 특화된 첫 GPU 인 Rubin CPX 출시를 앞두고 있으며, Vera Rubin NVL144 로 첫 엑사플롭스 단위의 랙을 출시 예정.

- 엔비디아 매출에서 차지하는 비중은 마이크로소프트 15~17%, 메타 13~15%, 알파벳 및 아마존 각 6~8%로 빅테크가 대부분 점유율을 차지

- ASIC 은 주로 추론 영역에서 운영 비용 절감 및 엔비디아 의존도 헷징을 목적으로 GPU 대비 30~60% 수준의 전력 효율성을 가지도록 설계

- CUDA 생태계 지배력, 칩 성능 시간차 격차로 인해 ASIC 의 부상에도 GPU 지배력은 일정기간 유지될 것으로 전망

TSMC

③ AI 사이클로 인해 선단 공정과 CoWoS 모두 병목을 겪고 있으며, 이로 인해 파운드리는 삼성전자가, 첨단 후공정은 인텔 EMIB 의 수혜가 예상.

- TSMC CoWoS 캐파는 26 년 말 12.5 만 장, 27 년 말 14 만 장으로 이는 블랙웰 GPU 기준 26 년 1,755 만 장, 27 년 2,385 만 장에 달하는 규모

- 3nm 선단공정은 3Q25 월 15 만 장의 풀캐파 가동수준이며, 26 년 말 20~22 만장 수준까지 증설 예정, 2nm 의 경우 26 년 말까지 월 10 만 장 생산 목표이며, 절반 이상 이미 애플이 선점

- 선단 공정 병목 지속될 경우, 외부 HPC 칩 양산 경험 및 테슬라 AI5, AI6 수주 이력 있는 삼성 파운드리 수혜 전망, 엑시노스 2600 으로 경쟁력 파악 가능

- AI 칩의 경우 선단 공정보다 CoWoS 병목에 직접적 제한받으며, 구글 TPU v9 및 메타 MTIA 등에서 도입 고려되는 인텔 EMIB 가 첨단 후공정 병목 지속 시 수혜 전망

- TSMC 캐파로부터 추정한 HBM 수요 용량은 26 년 337 억 Gb, 27 년 458 억 Gb

메모리

④ 범용 DRAM 쇼티지는 공급 물량 대비 26 년 66%, 27 년 58%로, 27 년까지 지속될 것으로 전망.

- 범용 DRAM 쇼티지 26 년 1,754 억 Gb, 27 년 2,009 억 Gb, HBM 의 경우 Bit 용량보다 성능, 수율, 후공정 증설 여부가 더 중요

- 26 년은 웨이퍼 순증 25 년 +215K/월 대비 26 년 +40K/월로 제한적이며, 공정 전환을 통해 소규모 Bit Growth 확보

- 27 년은 용인 클러스터(SK 하이닉스) 등 신규 공간 증설로 하반기부터 물리적 증설 가능

- HBM 및 1c 등 미세 공정 전환 위주의 Bit Growth, 클린룸 부족 등의 현재 증설 병목의 주요 원인

- SK 하이닉스의 경우 HBM3E 에서의 압도적 경쟁력을 바탕으로 26 년도 HBM 점유율 일부 유지 전망

- HBM4 부터 로직 다이 결합 최적화, 엔비디아 품질 기준 상향으로 삼성전자의 빠른 격차 축소 전망(엔비디아 퀄테스트 통과)

- 범용 DRAM 수요 증가 원인: 1) 추론 워크로드 증가로 AI 데이터센터 메모리 구조 개편 2) 일반 서버가 AI 서버 따라 증가 3) 빅테크 조달 경쟁 심화 4) 16~18 년 일반 서버 사이클의 교체 주기 도래

- 하이브리드 본딩은 HBM4E(28 년 이후)부터 도입 전망, 엔그램이 도입된 딥시크 V4 출시로 CXL 의 메모리 풀링 기능 주목 및 DRAM, SSD 수요가 다시 자극 받을 가능성

Key Chart

표 1. 빅테크 ASIC 생산 중 제품 사양 비교

구분	적용공정	소모전력	다이 개수	AI연산(FP4)	AI연산(FP8)	메모리 종류	메모리 용량	가격(K$)	출시 일정
B200	4NP	1000w	2개	14 PFLOPS	7 PFLOPS	HBM3e	192GB	30~40	24년
TPU v6	3nm	400w~500w (E)	1개			HBM3	32GB		24년
Trainium v2	5nm	500w~600w (E)	2개	5.2 PFLOPS	1.3 PFLOPS	HBM3	96GB		26년 상반기
Maia v1	5nm	1,000w	2개			HBM2e	64GB		24년
MTIA v2	5nm	1,000w (E)	2개			LPDDR5	128GB		25년 4분기

자료: 각종 보도자료 종합, KUVIC 리서치 1팀

표 2. 빅테크 ASIC 최신 제품 사양 비교

구분	적용공정	소모전력	다이 개수	AI연산(FP4)	AI연산(FP8)	메모리 종류	메모리 용량	가격(K$)	출시 일정
R100	3nm	1,000w (E)	2개	50 PFLOPS (E)	25 PFLOPS	HBM4	288GB		26년 하반기
TPU v7	3nm	600w~800w (E)	2개			HBM3e	216GB		26년 하반기
Trainium v3	3nm	500w~600w (E)	2개 (E)			HBM3e	288GB		26년 1분기
Maia v3	3nm	1,000w (E)	2개			HBM4			27년
MTIA v3	3nm		2개			HBM3e			26년 상반기

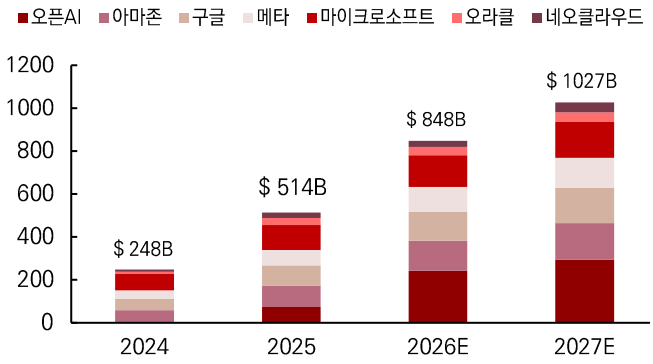
자료: 각종 보도자료 종합, KUVIC 리서치 1팀

그림 1. UBS: 사용 사례 당 추론 연산량 수요 추정

사용 사례	2024년 수요	2030년 전망
챗봇(예: ChatGPT)	10 exaFLOP/s	200 exaFLOP/s
기업용 AI (예: 사기 탐지)	15 exaFLOP/s	440 exaFLOP/s
에이전트형 AI	수백 exaFLOP/s	14 zetta FLOP/s
물리적 인공지능(예: 로봇공학)	X	? yottaFLOP/s

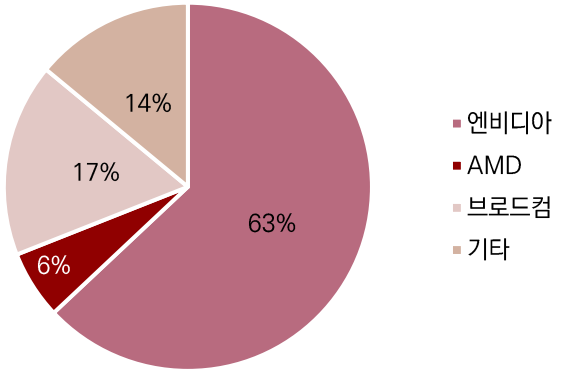
자료: UBS, KUVIC 리서치 1팀

그림 2. AI 데이터센터 CAPEX



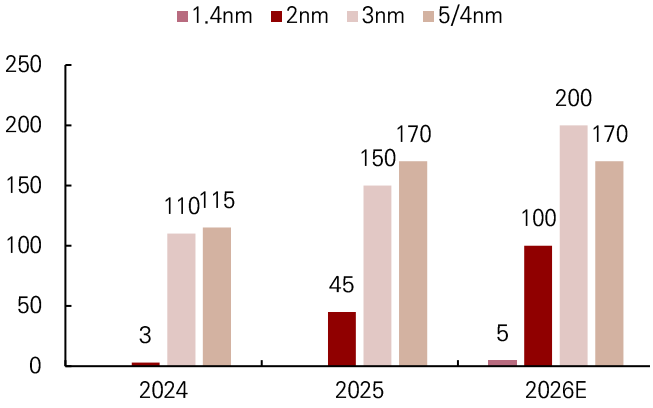
자료: Morgan Stanley, 각 사, KUVIC 리서치 1팀

그림 3. CoWoS 고객사별 비중



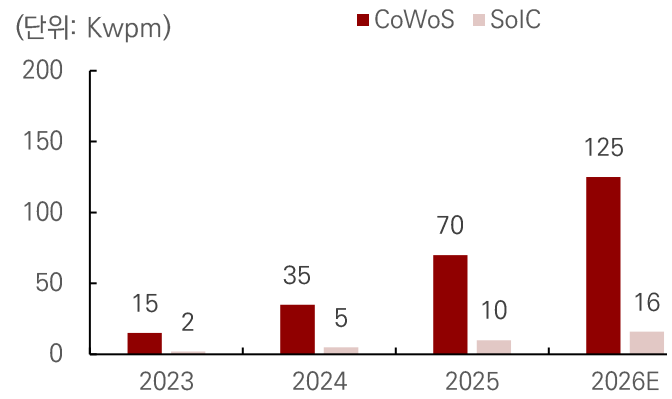
자료: Trendforce, KUVIC 리서치 1팀

그림 4. TSMC 선단 공정 생산량 전망



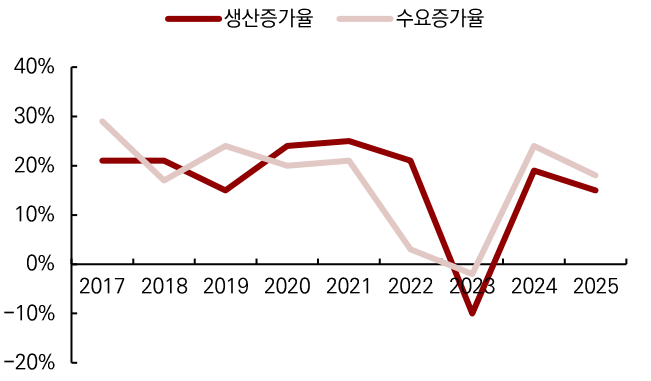
자료: KUVIC 리서치 1팀

그림 5. TSMC CoWoS/ SoIC capa 추이 및 전망



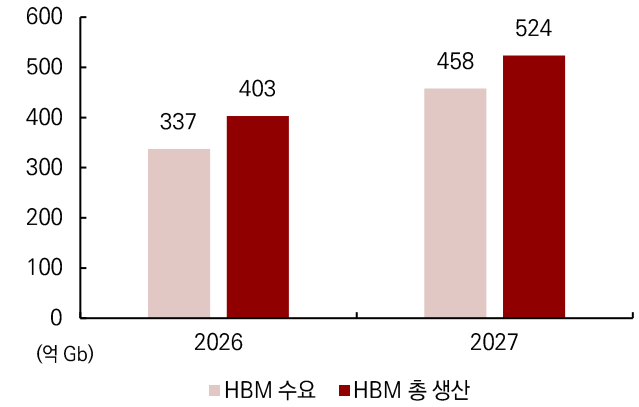
자료: TSMC, 언론종합

그림 6. DRAM 수요공급 증감률 추이



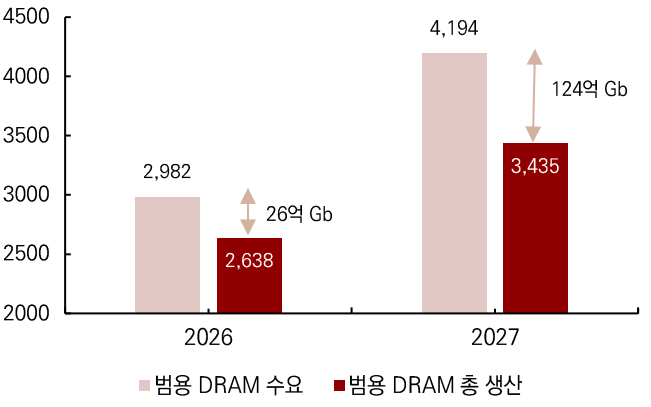
자료: OMDIA, WSTS, KUVIC 리서치 1팀

그림 7. 데이터센터 향 HBM 수요 및 공급



자료: KUVIC 리서치 1팀

그림 8. 데이터센터향 범용 DRAM 수요 및 공급



자료: KUVIC 리서치 1팀

표 3. 메모리 3사 증설 타임라인

		2025	2026F	2027F	2028F	2029F
삼성전자	P4(평택)	Ph1: DRAM 30k NAND15k	Ph3: DRAM 45k	Ph4(3Q26): DRAM 45k	Ph2(4Q26): DRAM 45k	
	P5 (평택)					9월 open
SK하이닉스	M15X (청주)		Ph3: DRAM 40k	Ph4: DRAM 40k		
	Y1 (용인)			2월 open		
마이크론	Fab16 (A3)	Ph2: DRAM 15k	Ph3: DRAM 15k			
	ID1			3Q open		
	P5			하반기 open		

자료: KUVIC 리서치 1팀

AI 수요

AI 컴퓨팅 수요 양상

추론의 급격한 성장

학습 vs 추론
학습은 대규모 일회성 비용, 추론은 지속적 비용

AI 모델의 라이프사이클은 두 개의 구별되는 단계로 나뉜다. **학습(Training)은 대규모 데이터셋으로부터 모델의 가중치를 최적화하는 일회성 고비용 작업**으로, 수일에서 수주가 소요되며 막대한 GPU 자원을 필요로 한다. 반면 **추론(Inference)은 학습 완료 후 고정된 모델을 사용해 새로운 입력값에 대해 실시간 예측**을 수행하는 단계로, ChatGPT에 질문을 입력했을 때 응답이 돌아오거나 YouTube가 사용자에게 맞춤 추천 영상을 제시하는 것이 모두 추론 단계에 해당한다. 두 단계 모두 컴퓨팅 자원을 소비하지만, 학습이 모델 개발 과정에 일회성으로 대규모의 연산량을 요구하는 반면 **추론은 배포 후 서비스 제공 기간 동안 지속적으로 발생**하는 특성을 가진다.

표 1. 비교: 학습 vs 추론

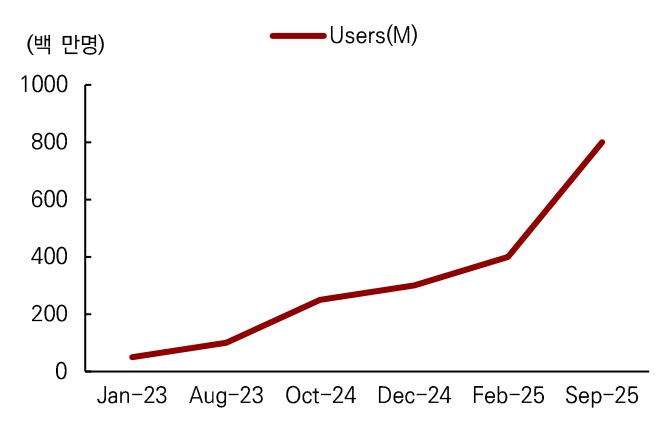
항목	학습 (Training)	추론 (Inference)
목적	과거 데이터를 알고리즘에 입력하여 패턴과 관계를 학습	학습된 모델이 패턴을 새로운 데이터에 적용하여 예측
데이터 흐름	대규모 레이블링된 데이터셋 입력 → 모델 파라미터 업데이트 → 손실 함수 최소화	학습된 고정 파라미터로 새로운 개별 데이터 처리 → 분류/확률/텍스트 생성
계산 요구 사항	고강도 연산	적은 연산량을 지속적 실행
지연 시간 및 성능	높은 지연 시간 허용, 사용자 영향 X	실시간 응답 요구, 사용자 경험 중요
비용 및 에너지 소비	일회성/대규모 투자, 주기적인 모델 업데이트	지속적 누적 비용
하드웨어 요구 사항	고성능 GPU/TPU 클러스터 필수	GPU/CPU/FPGA/NPU/엣지 디바이스 다양 선택 가능. 비용-성능 균형 최적화 필요.

자료: KUVIC 리서치 1팀

학습보다 추론을 더
26년에 AI 인프라 지출의 55%, 연산 요구량의 66% 점유 전망

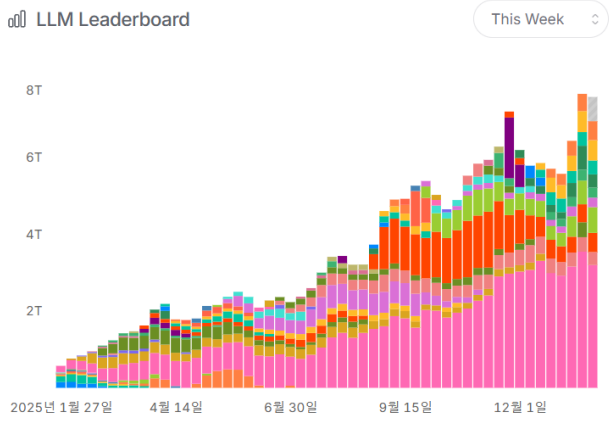
현재 AI 컴퓨팅 수요의 급증은 단연 **추론**으로 주도되고 있다. 다수의 업계 분석가들은 25년을 지출액과 컴퓨팅 수요의 측면에서 추론 수요가 학습 수요를 넘어서는 원년으로 해석한다. Gartner 분석에 따르면 AI 클라우드 인프라 지출에서 추론이 2025년 92억 달러(전체 183억 달러)에서 2026년 206억 달러로 2배 이상 증가하여 **전체 375억 달러의 55%**를 차지할 것으로 추정했다. Deloitte는 전체 연산량 중 추론 워크로드의 비중이 23년 3분의 1, 25년 5분의 1에서 **26년 3분의 2**가 되며 가장 주목받는 분야가 될 것으로 예측했다.

그림 1. ChatGPT 이용자 수



자료: OpenAI, KUVIC 리서치 1팀

그림 2. Openrouter API 주간 토큰 사용량 (1년 추이)



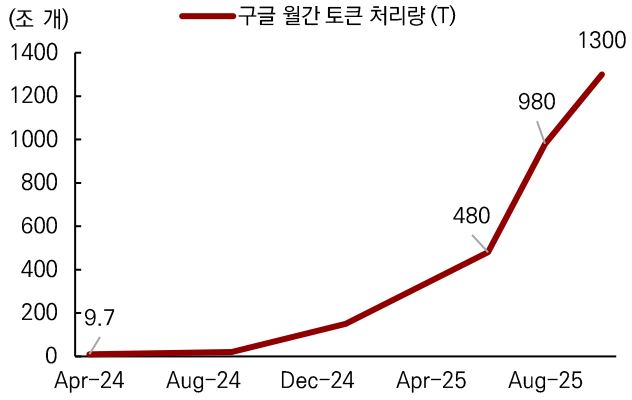
자료: Openrouter

추론 수요 폭증:
1) 이용량 증가
2) 모델 크기 증가

지속성이라는 추론의 구조적 특성 아래 **이용량과 모델 크기가 동시에 상승하는 것이 추론 수요 폭증의 원인**이다. 첫째, **AI 서비스 보편화로 이용량이 급증**한다. ChatGPT와 Gemini는 25년 초와 25년 말을 비교하여 대략 2배씩 이용자 수가 증가했다. (ChatGPT WAU 2월 4억 명 → 9월 8억 명, Gemini MAU 3월 3.5억 명 → 10월 6.5억 명) AI 서비스 종류도 다양화·전문화되며 기하급수적으로 증가하는 중이다. 또한 미국 기업 40%가 이미 AI 도구에 비용을 지불 중이며 2028년 80%, 2030년 거의 100%에 달할 전망이다. 둘째, **소프트웨어 혁신으로 더 복잡한 작업이 가능**해지면서 모델 크기 증가는 필연적인 흐름이다. ChatGPT는 이용자가 4.5배 증가할 동안 총 토큰 사용량이 50배 늘었는데, 이는 유료 사용자들의 1인당 사용량 증가와 복잡 작업 위임을 반영한다. 더불어 Google의 Veo, OpenAI의 Sora 등 영상 생성 월드 모델이 출시 및 업그레이드되며 이전 대비 훨씬 큰 연산량을 요구하고 있다.

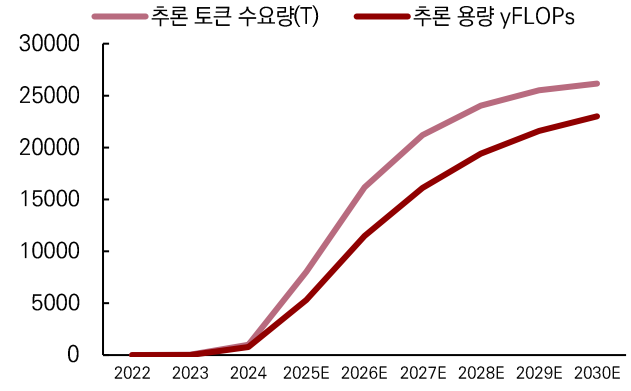
이렇게 이용량과 모델 크기 동반 상승으로 볼륨이 커지는 AI 서비스를 영구적으로 지원하며 추론 워크로드가 폭발적으로 가중되는 와중, 학습 워크로드는 학습할 데이터 자체가 부족해지거나 피지컬 AI와 같이 데이터 생성 비용 자체가 증가하며 점차 둔화될 것이라는 예측이 지배적이다.

그림 3. 구글 월간 토큰 처리량



자료: 구글, KUVIC 리서치 1팀

그림 4. Eric Ding: 추론 컴퓨팅 수요 추정 (토큰/용량)



자료: Eric Ding, KUVIC 리서치 1팀

추론이 촉발한 메모리 시대

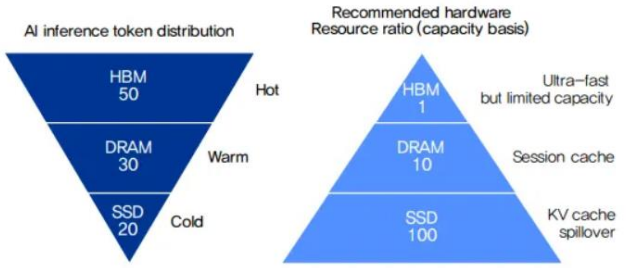
기존 병목은 HBM
대역폭이 높지 않으면 GPU 높은 성능 사용 불가

AI 사이클의 **기존 메모리 병목은 대역폭(Memory Bandwidth)** 이슈였다. GPU의 컴퓨팅 성능이 메모리 전송 속도의 수십~수백배 빨라서, 결국 메모리 대역폭이 높지 않으면 GPU의 높은 컴퓨팅 성능을 이용할 수 없다. 따라서 3D 수직 스택 구조로 테라바이트/초 대역폭을 제공하는 **고대역폭 메모리(HBM: High-Bandwidth Memory)**이 GPU의 필수 요소가 되었고, DRAM을 수직으로 쌓는 TSV 공정과, HBM을 인터포저를 통해 GPU die와 2.5D로 연결하는 CoWoS 등의 **첨단 후공정 기술이 새로운 병목**으로 부상하게 되었다.

롱컨텍스트 추론
범용 DRAM 및 SSD 수요 폭발

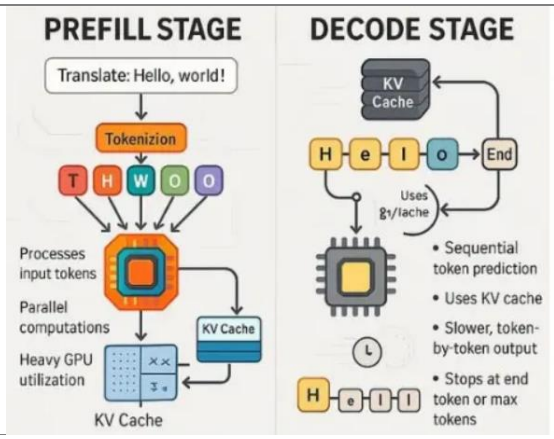
그러나 추론 과정이 복잡해지는 **롱컨텍스트 추론(100K~1M 토큰) 국면**에서는 새로운 병목이 나타났다. **KV 캐시** 크기가 컨텍스트 길이에 선형으로 증가해 **GPU의 HBM 용량을 초과**하는 경우가 빈번해졌으며, CSP들은 이를 해결하기 위해 **Prefill(프롬프트 처리, 컴퓨팅 집약)**과 **Decode(토큰 생성, 메모리 집약)** 단계에 **서로 다른 메모리 계층을 할당**하기 시작했다. 구체적으로 Decode 단계에만 HBM을 집중 배치하고, Prefill과 KV 캐시 중 비활성 부분은 범용 DRAM이나 NVMe SSD로 오프로드하여 **HBM:DRAM:SSD = 1:10:100** 구조로 재편(기존 HBM:DRAM = 1:1 구조)되면서 **특히 범용 DRAM과 SSD 수요가 폭증**했다. 이렇게 AI 사이클의 추론 국면이 데이터센터의 메모리 구성을 근본적으로 변화시키며 범용 DRAM을 AI 밸류체인에서 가장 중요한 위치로 새롭게 올려놓게 되었다.

그림 5. 추론 AI 메모리 계층 구조



자료: Omdia

그림 6. 프리필 & 디코드 연산 과정



자료: Medium

생성형 AI: LLM부터 에이전틱AI까지

계속 커질 GenAI
생성형 AI의 작업 차원은 기하급수적으로 확대

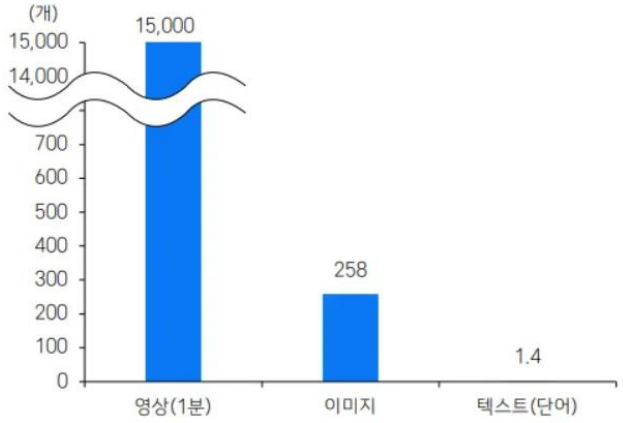
생성형 AI의 크기는 점차 증가할 예정이다. GPT-4는 약 500B, GPT-4o는 1.8T, GPT-4.5는 5~7T, Gemini 2.5 Pro는 5T 수준으로, 시간의 흐름에 따라 LLM의 파라미터 수는 몇 배씩 증가하는 양상을 보인다. MoE 아키텍처가 도입되며 추론 연산량 감소 혁신이 있었지만, 작업 난이도가 높아지며 실제 활성 파라미터수 또한 증가할 수밖에 없는 흐름이다. 한 예로, 구글은 영상 생성을 넘어 3D 가상 공간을 구현하는 월드 모델 연구를 가속화 중이다. 텍스트→이미지→영상→가상 공간으로 생성형 AI의 작업 차원이 확대될수록 미래 컴퓨팅 수요는 현재 수준을 압도하는 기하급수적 증가를 보일 전망이다.

그림 7. 구글 생성형 AI 연구개발 로드맵

구분	Veo	GameNGen	Genie 3
현재 상태	대중 공개	연구용	연구용이나 향후 공개될 듯
도메인	비디오 및 음성 생성	특정 게임을 생성한 후 플레이	월드 모델을 동시에 생성/플레이
해상도	720p ~ 4K	320p	720p
생성 길이	8-10초	수 초	수 분
제어 방식	텍스트와 이미지, 비디오로 생성 명령	키보드·마우스로 제어	텍스트와 이미지로 명령 내리고 키보드·마우스로 제어
상호작용	실시간 상호작용 불가능	이미 학습된 특정 게임 내 실시간 플레이 가능	사용자 입력(키보드, 마우스)에 따라 세상 자체의 환경을 실시간으로 한 프레임씩 생성

자료: 구글

그림 8. 영상, 이미지 텍스트 생성 시 필요한 토큰 수 비교



자료: 구글

에이전틱 AI
스스로 계획하여 과업을 완수하는 진화된 LLM

LLM의 지능 향상과 프레임워크 및 프로토콜의 확산이 자율성과 행동력을 갖춘 **에이전틱 AI 시대**를 열고 있다. JP Morgan에 따르면, 22년 Chatbot에서 2024년 Reasoning, 그리고 2026년 Agentic AI로 2년 주기로 새로운 AI 워크로드 진화가 이뤄지며 **토큰 소비량은 매년 10배 이상 증가**할 것으로 분석했다. 에이전틱 AI는 최소한의 인간 개입으로 명령을 수행하고, 계획 수립부터 실행까지 다단계 추론을 통해 과업을 완수한다. 이러한 구조적 특성상 **토큰 처리량의 급증은 불가피**하며, 모델의 고도화는 범용 에이전트의 확장과 산업별 특화 에이전트의 두 방향으로 이루어질 것으로 전망된다.

에이전틱 AI
26년 컴퓨팅 수요 폭증의 주인공 예약

여러 에이전트가 문제를 분할·병렬적으로 수행하는 '다중 에이전트 시스템(Multi-Agent System)'은 단일 에이전트 대비 약 90% 이상 높은 성능을 보였으나, 일반적인 채팅 대비 **단일 에이전트의 토큰 소모량은 4배**, **다중 에이전트 시스템은 15배 이상**에 달했다. 이는 에이전트 시대의 컴퓨팅 요구가 과거 챗봇

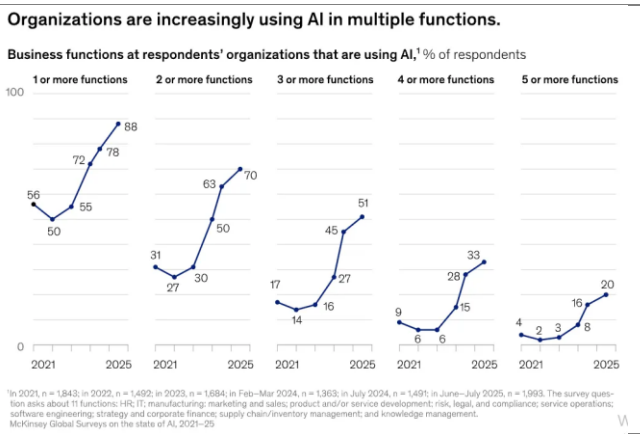
중심 시대와는 근본적으로 다른 차원임을 시사한다. UBS는 에이전틱 AI 관련 컴퓨팅 수요가 2024년 수백 exaFLOP/s에서 2030년 14 zettaFLOP/s로 확대될 것으로 전망하며, 가장 높은 성장세가 예상되는 분야로 제시하고 있다.

그림 9. UBS: 사용 사례 당 추론 연산량 수요 추정

사용 사례	2024년 수요	2030년 전망
챗봇(예: ChatGPT)	10 exaFLOP/s	200 exaFLOP/s
기업용 AI (예: 사기 탐지)	15 exaFLOP/s	440 exaFLOP/s
에이전트형 AI	수백 exaFLOP/s	14 zetta FLOP/s
물리적 인공지능(예: 로봇공학)	X	? yottaFLOP/s

자료: UBS, KUVIC 리서치 1팀

그림 10. McKinsey: 다중 에이전트 사용 기업 설문조사



자료: McKinsey

생성형 AI: LLM부터 에이전틱AI까지

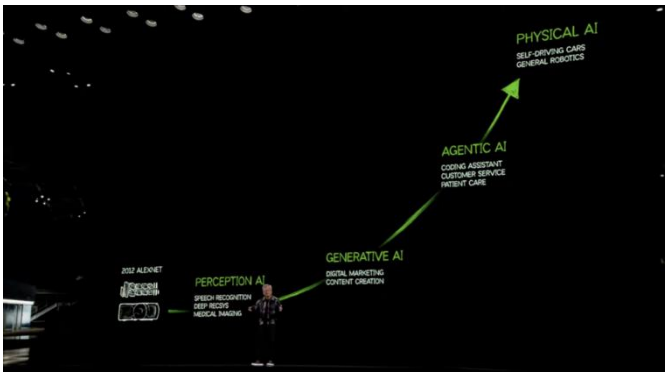
피지컬 AI
엔비디아의 궁극적 지향점

엔비디아는 AI의 궁극적 지향점으로 피지컬 AI를 제시하며, 세계 최초로 **"월드 파운데이션 모델: Cosmos"**를 출시했다. Cosmos는 물리 법칙에 정렬된 가상 세계를 2D 비디오로 생성하는 모델로, 2천만 시간 분량의 물리적 속성이 강하게 드러나는 비디오와 Omniverse 3D 에셋을 통합 학습한 결과물이다. 이는 로보틱스 및 자율주행 AI 학습의 근간으로 활용되며, **AI for AI traning**으로서의 역할을 수행한다.

엔비디아 vs 테슬라
Cosmos로 물리 데이터 무한 복제
vs
대규모 현실 데이터 때려넣기

로보틱스 발전의 가장 큰 장벽은 **훈련 데이터의 절대적 부족**이다. 인터넷 수집이 불가능하고 물리 세계에서 직접 생성해야 하는 구조적 한계 때문이다. **엔비디아**는 **"소량의 실제 데이터로 무한한 양의 고품질 합성 데이터를 생성"하는 Cosmos로 이를 해결**한다. 구글의 월드 파운데이션 모델 Genie 3도 동일한 목적성을 띤다. 이에 반해 **테슬라**는 독보적인 자율주행차 판매량을 기반으로 확보한 **대규모 현실 비전 데이터를 학습**하고, 휴머노이드 학습에도 센서 데이터나 유튜브 영상 등 유사 접근을 병행한다.

그림 11. 젠슨 황의 AI 4단계 진화론



자료: 엔비디아

그림 12. 엔비디아와 테슬라 피지컬 AI 훈련법 비교대조

구분	테슬라	엔비디아
기본 철학	양이 질을 이긴다	"소량 정예 + AI 생성"
데이터 원천	자율주행, 유튜브 영상 + 실제 인간 작업	소량 고품질 텔레오퍼레이션 데이터
필요 데이터 규모	수십억 시간(무제한)	수십~수천 개 현실 데이터
학습 방식	End-to-End 직접 학습	4단계 파이프라인(추론 → 학습 → 추론)
비용 구조	예측 불가능	비교적 예측 가능(샘플당 $0.34)

자료: 언론 보도 종합, KUVIC 리서치 1팀

누가 이기든
피지컬 AI 훈련 컴퓨트 수요는 대규모

훈련 데이터가 시뮬레이션 기반이든 현실 기반이든, **피지컬 AI 훈련 컴퓨트 수요는 지속**될 전망이다. 엔비디아는 Cosmos 데이터 생성(추론)과 이를 학습하는 과정에서 폭발적인 연산량 수요가 발생한다. 테슬라도 자율주행 상용화 및 휴머노이드 양산 직전임에도 **막대한 훈련 연산량을 확보** 중이며, 이는 간접적으로 수요를 확인시켜 준다. xAI의 일론 머스크는 5년 내 H100 동급 5,000만 개 GPU 확보를 목표로 하며, 예상 전력 소모량은 약 35GW(인류 전체 전력 소비의 2%에 달한다.

GPU에 미친 xAI
세계에서 가장 높은 연산량 보유가 목표

xAI는 현재 모델 학습을 위해 23만 개 GPU(이중 GB200 3만 개)로 구성된 Colossus 1단계를 운영 중이다. 차세대는 테네시주 **Colossus 2단계**로, 11만 개 GB200으로 시작해 55만 개 GB200·GB300, **최종 100만 개 이상 Blackwell GPU 집적을 목표**한다. 완공 시 20,000 ExaFLOPS(FP8) 연산 능력을 확보하며, 메타·OpenAI 클러스터 대비 50배 이상, 현존 9개 초대형 클러스터 합산치보다 11배 큰 압도적 선두를 점하게 된다.

표 2. Colossus 2단계 관련 주요 내용

항목	내용
부지	Tulane industrial park, Shelby County, 테네시주
GPU 1차 배치	110k GB200 (2월 중순 설치 및 시범 가동 중) (Colossus 1단계 급 규모 연산량)
우선 목표(01/23 달성)	550k GB200 & GB300 01/23 가동 개시 (1GW 급)
중간 목표	4월 중 1.5GW 급 업그레이드 (Blackwell 750K 급)
최종 목표	Blackwell 급 GPU 1000K 이상 집적 (2GW 급)
냉각	전량액체냉각(콜드-플레이트)으로 이동식냉각장치사용
백업 전력	테슬라Megapack (수백MWh 규모) + 다중 전원 루프

자료: xAI, 언론 보도 종합, KUVIC 리서치 1팀

왜 이렇게 많이?
1) 안전 문제는 아무리 해도 부족
2) 강화학습 스케일링으로 추론 수요 폭발

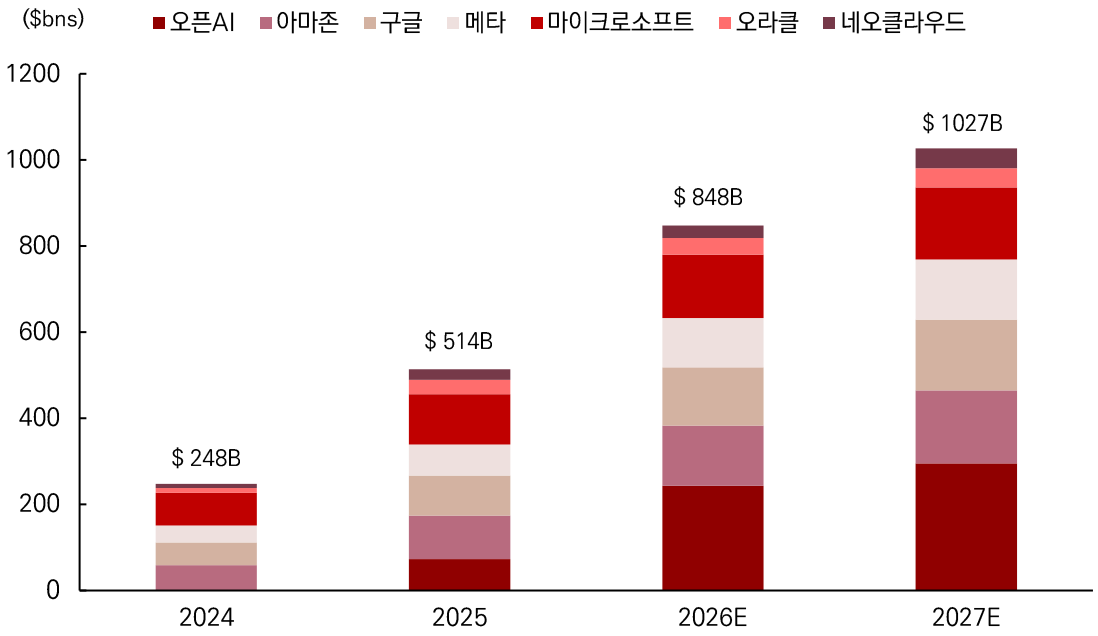
테슬라의 연산력 확보 배경은 다음과 같다. **물리 세계에서 99.99%와 99.9999% 신뢰도는 인간 안전 기준으로 완전히 다른 차원의 의미**를 가지기에 지능 향상 필요성이 지속된다. FSD가 오스틴에서 상용화에 성공하더라도 서비스 지역 확장을 위해 모델 크기를 몇 배 이상 키워야 한다. 또한 AI 업계는 단순 모델 크기 확대를 넘어 **"강화학습 스케일링"** 패러다임으로 전환하며 추론 측면에서도 연산 요구량 또한 폭발할 것이다. 이는 훈련 FLOPs 경쟁이 아닌, 추론 단계에서 수많은 롤아웃(답변 생성)을 통해 최적 선택으로 모델을 다듬는 방식으로, 극도로 추론 집약적인 과정이다.

AI 데이터센터 CAPEX 전망

OpenAI와 빅테크 주도의 전례 없는 AI 사이클 진입

2026년 글로벌 AI 데이터센터 CAPEX는 OpenAI, 빅테크 4사, 오라클 및 네오클라우드 기업들의 합산 기준 **8,480억 달러**에 달할 것으로 전망된다. 이는 2024년 기록한 2,480억 달러 대비 **약 242% 급증한 수치**이며, 전년도인 2025년의 5,140억 달러와 비교해도 **약 65%의 가파른 성장세**를 유지하는 수준이다. 인프라 확장이 본격화됨에 따라 차기 연도인 **2027년에는 전체 투자 규모가 1조 270억 달러**에 이를 것으로 예측되며, AI 산업의 자본 집약적 성장 기조는 더욱 심화될 것으로 보인다.

그림 13. AI 데이터센터 CAPEX

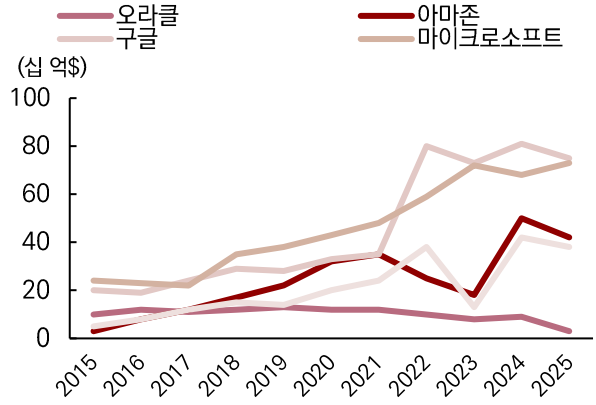


자료: Morgan Stanley, 각 사, KUVIC 리서치 1팀

Base -> 70%
Bull -> 100%

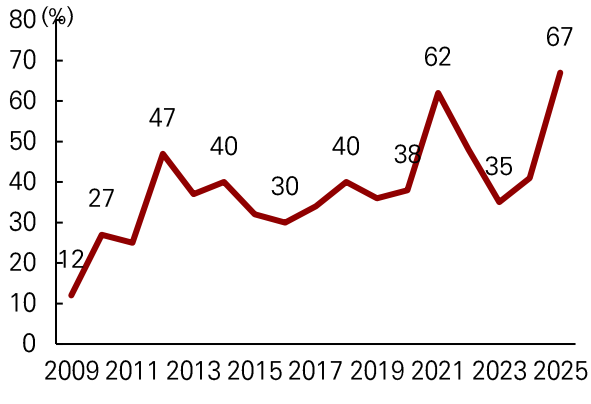
오라클과 빅테크 4사의 현금 흐름을 분석한 결과, **잉여현금흐름(FCF)의 우상향 추세에도 불구하고 AI 사이클에 따른 자본지출 강도는 전례 없는 고점**을 향하고 있다. 특히 **영업현금흐름(OCF) 대비 CAPEX 비중이 70%를 상회**하며 지속적으로 상승하고 있다는 점에 주목해야 한다. 본 분석에서는 이러한 추세를 반영하여 **현재 비중인 70%를 Base 시나리오**로 설정하였으며, **영업현금흐름 전액을 설비 투자에 투입하는 공격적인 상황을 Bull 시나리오(100%)**로 가정하여 수요 성장의 상단 가능성을 검토하였다.

그림 14. 하이퍼스케일러 잉여현금흐름*(FCF) 추이*



자료: Bloomberg, KUVIC 리서치 1팀

그림 15. 하이퍼스케일러 CAPEX/OCF 추이 (%)



자료: Bloomberg, KUVIC 리서치 1팀

빅테크 4사

2022년 ChatGPT 출시 이후 주요 빅테크의 CAPEX는 약 2.5배 급증했으며, 2025년부터 2027년까지 4사의 누적 지출액은 **약 1.5조 달러**에 달할 전망이다. **아마존**은 자체 개발 칩인 Trainium 도입을 통해 원가의 30~40% 절감 및 연산 병목 현상 해소를 꾀하고 있으며, 2027년 CAPEX는 1,700억 달러로 4사 중 **단일 기업 기준 최대치**를 기록할 것으로 전망된다. **구글** 또한 자체 개발 TPU v6/v7의 대량 배포에 힘입어 2025년 75%의 높은 성장률을 기록했으며, 순다르 피차이 CEO가 Bloomberg Tech conference에서 언급한 **'의미 있는 증가(significant increase)' 기조**에 따라 2026년에도 공격적인 인프라 확장을 지속할 것으로 예측된다.

올해 초 가동을 목표로 하는 초대형 AI 데이터센터 Fairwater 프로젝트

메타는 엔비디아 H100 의존도를 낮추기 위해 **자체 칩 MTIA v2/v3 비중을 확대**하고 있으며, 2025년 기준 85%의 CAPEX 성장률을 기록하며 인프라 확보에 사활을 걸고 있다. 특히 마크 저커버그 CEO가 밝힌 향후 수년간 **6,000억 달러 규모의 데이터센터 클러스터 구축 계획**은 메타의 장기적인 AI 리더십 확보 의지를 반영한다. **마이크로소프트**는 **Fairwater 프로젝트**를 필두로 초대형 인프라 구축을 주도하고 있으며, 2025~2027년 누적 기준 **빅테크 중 가장 압도적인 투자 규모**를 유지하며 Azure OpenAI를 통한 수익화 모델을 공고히 할 것으로 판단된다.

표 3. 빅테크 4사 CAPEX

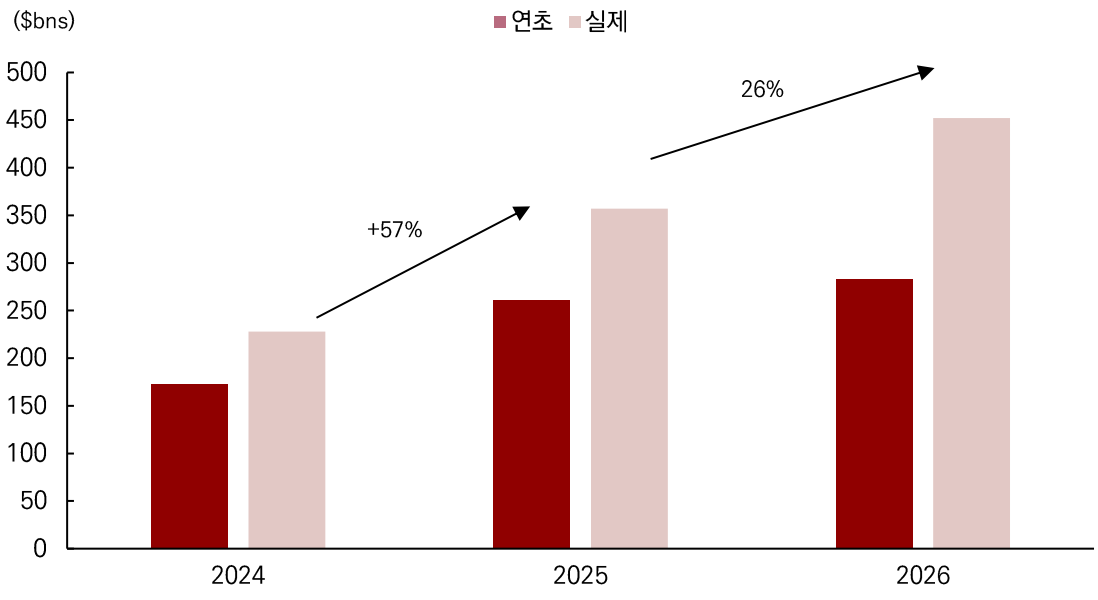
(단위: $B)	2024	2025	2026E	2027E
아마존	59	101	140	170
구글	53	93	135	164
메타	39	72	115	140
마이크로소프트	76	117	147	167

자료: Morgan Stanley, KUVIC 리서치 1팀

연초 예상치를 상회하던 실질 CAPEX

아래 그래프를 보면 **연초 대비 실제 CAPEX 예상치가 상당히 상이**한 것을 볼 수 있다. 24년에는 32%의 상향이 이루어졌으며 25년에는 37%의 상향이 이루어졌다. 이와 같은 추세를 감안하였을 때, **올해 역시 연초의 예상치와 실질적인 실제 CAPEX가 다를 가능성이 충분**하며, 위 표의 예측치가 보수적인 수치일 수 있음을 감안해야 한다. Bloomberg의 컨센서스에 따르면 26년 빅테크 4사의 CAPEX는 YoY 26% 성장이 예측되는데, 고질적인 AI CAPEX 과소평가 현상을 고려해야 한다.

그림 16. 빅테크 4사 CAPEX 연초 예상치 vs 현재 예상치



자료: Bloomberg, KUVIC 리서치 1팀

오픈 AI의 막대한 CAPEX 투자

기존 빅테크 CAPEX를 뛰어넘는 오픈AI의 대규모 투자

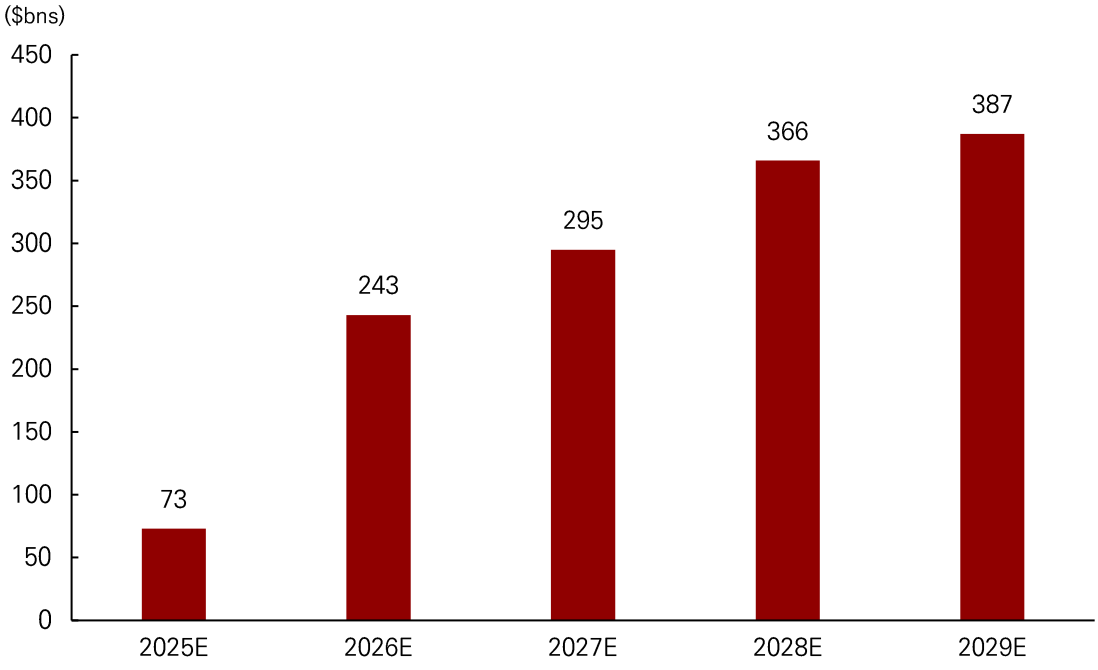
오픈AI의 **스타게이트(Stargate) 프로젝트**는 인공지능 인프라 역사상 최대 규모인 **5,000억 달러**를 4년간 투입하는 초대형 자본 집약화 사업이다. 이 프로젝트는 오픈AI와 소프트뱅크가 공동 주도하며, 오라클과 UAE 기반 투자사 MGX가 주요 자본 파트너로 참여해 설립된 **독립 법인인 스타게이트 LLC**를 통해 추진된다. 소프트뱅크가 재무적 책임을, 오픈AI가 설계 및 운영 책임을 맡는 구조를 통해, 개별 기업의 재무적 한계를 넘어선 천문학적인 자본을 AI 가동 능력 확보에 집중시키고 있으며, 이는 단순한 서버 임대를 넘어선 독자적인 인프라 구축 모델을 지향한다.

이에 따라 **26년부터 오픈 AI의 스타게이트와 반도체 기업 파트너십을 통한 CAPEX 확보가 본격적으로 반영**된다. 해당 프로젝트는 기존 CSP와의 파트너십에 기반하고 있지만, 그 자본의 원천과 기술적 설계, 에너지 전략 면에서 CSP의 유기적 성장과는 구분되는 독립성을 지닌다. 외부 재무적 투자자와 부채 시장에서 유입되는 수천억 달러의 자금은 빅테크를 포함한 기존 테크 기업의 CAPEX 예산 밖에서 형성되는 신규 투자이며, 이를 추산하여 향후 AI 데이터센터 CAPEX에 추가되는 자금 흐름을 파악할 수 있다.

5,000억 달러가 26년부터 29년까지 4년간 CAPEX에 반영됨

그림 2에서 2026년 투자가 2,430억 달러로 급격히 가속화되는 이유는 **엔비디아의 차세대 '루빈(Rubin)' 플랫폼 출시와 브로드컴 기반의 10GW급 맞춤형 가속기 대량 도입** 시점(29년 완료)이 맞물리기 때문이다. 5,000억 달러의 CAPEX 투자가 2029년까지 4년간 반영되어 적용되며 **2033년까지 1조 4,000억 달러 규모의 장기 인프라 확보**를 약속했기에 올해부터 매년 최소 2000억 달러 이상의 자본을 투자할 것으로 전망된다.

그림 17. 오픈AI 프로젝트 CAPEX 금액 전망



자료: 언론 보도 종합, KUVIC 리서치 1팀

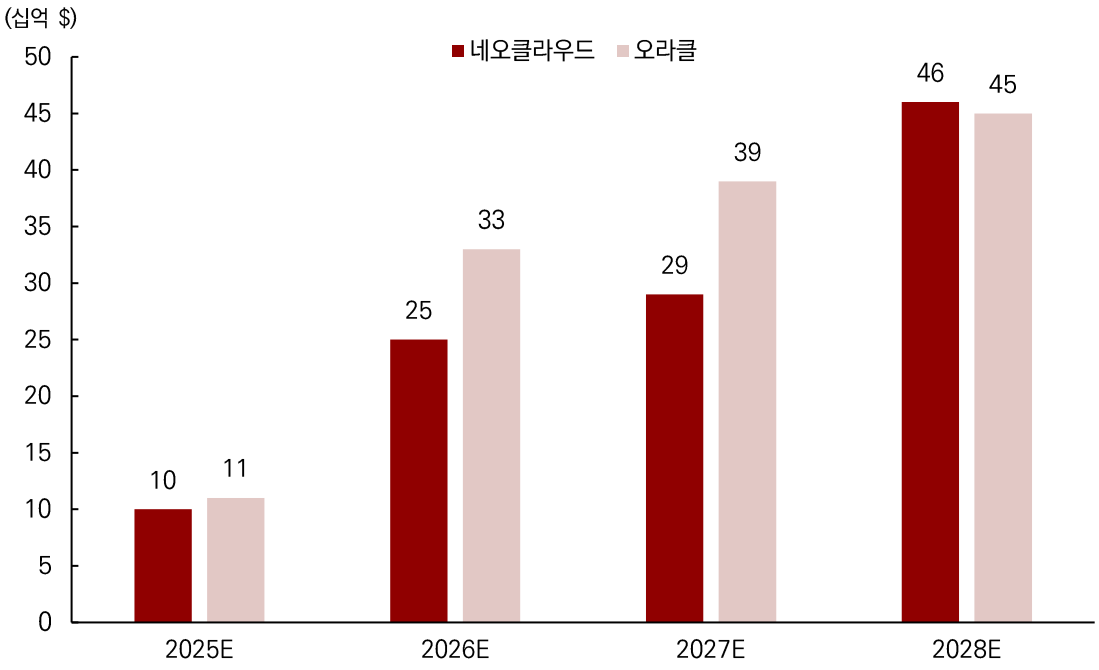
네오클라우드와 오라클

네오클라우드 기업인 **코어위브(CoreWeave), 네비우스(Nebius), 아이렌(IREN)** 등은 빅테크 하이퍼스케일러의 캐파 쇼티지(Capacity Shortage) 이슈를 해결할 실질적인 대안으로 급부상하고 있다. 현재 인공지능 인프라 시장은 견조한 고객 수요를 바탕으로 공급이 수요를 따라가지 못하는 국면이 지속되고 있으며, 이에 대응하기 위한 네오클라우드 업체들의 **공격적인 CAPEX 확장 의지**가 확인된다. 네오클라우드 기업들의 전체 CAPEX 규모는 **2025년 250억 달러에서 2030년 890억 달러까지 연평균(CAGR) 29%의 고성장을 기록**할 것으로 전망된다.

AI 인프라 시장의 핵심 공급자이자 공격적인 CAPEX 투자를 보이는 네오클라우드 기업과 오라클

오라클은 기존 하이퍼스케일러 중 네오클라우드 생태계와 가장 밀접하게 결합하여 독보적인 성장 동력을 확보하고 있다. 특히 2025년 9월 **오픈AI와 체결한 5년간 3,000억 달러 규모의 초대형 클라우드 컴퓨팅 계약(Stargate 프로젝트)**은 오라클이 AI 인프라 시장의 핵심 공급자임을 입증하는 사례이다. 이러한 대규모 수주는 오라클의 인프라가 오픈AI와 같은 최상위 AI 기업의 요구사항을 충족할 수 있는 기술적, 자본적 역량을 갖추었음을 의미하며, 향후 오라클의 CAPEX 집행이 네오클라우드 시장 전체의 흐름을 주도하는 지표가 될 것임을 시사한다.

그림 18. 네오클라우드 상장사와 오라클 CAPEX 전망



자료: Morgan Stanley, 각 사, KUVIC 리서치 1팀

규모 분석

표 4. 2026년 및 2027년 AI 가속기 칩 수요

계산식	2026년	2027년	단위
AI 데이터센터 CAPEX 종합	848	1027	$B
AI 가속기 칩 구매 비중	40	40	%
엔비디아 GPU 비중	60	60	%
엔비디아 GPU 단가	22,000	22,000	$
ASIC 단가	10,000	10,000	$
전체 AI 가속기 칩 평균 단가	17,200	17,200	$
2026년 예상 수요(Base / Bull)	**1,970 / 2,814**	**2,386 / 3,408**	**만대**

자료 : KUVIC 리서치 1팀

CAPEX를 기반으로 한 AI 가속기 칩 수요 추정

종합적으로, **오픈AI를 포함한 빅테크와 네오클라우드 기업들의 26년 ai 데이터센터 CAPEX는 8,480억 달러**이며, 이는 블랙웰 **GPU 1,970만 장**에 달하는 규모이다. **27년 CAPEX는 1조 270억 달러**로 블랙웰 **GPU 2,386만 장**에 달한다. 후술한 본 보고서의 글로벌 CoWoS Capa 추정치가 블랙웰 GPU 환산 시 26년 1,755만 장, 27년 2,385만 장 규모임을 고려하면, **빅테크가 각자 TSMC로부터 확보 가능한 물량을 기준으로 CAPEX를 책정**한 것을 알 수 있다. 이는 생산 가능 물량이 모두 판매된다는 것을 의미하므로 **AI 가속기 시장에 초과 수요**가 존재하며, **TSMC의 CoWoS가 AI 사이클 주요 병목**임을 단적으로 시사한다. 아래는 추정 논리에 대한 설명이다.

과거 데이터센터 투자에서는 건물과 전력 설비 비중이 높았으나 AI 데이터센터는 역전된 비용 구조를 띈다. 모건스탠리의 리서치에 따르면 **GPU 구매 비용이 CAPEX의 41%**에 이르며 TrendForce는 AI 데이터센터에서 서버가 전체 CAPEX의 60%를 차지하고 이중 대부분이 GPU 비용에 포함된다. 따라서 전체 CAPEX의 40%를 AI 가속기 칩 구매 비용이라 추정하였다.

시장은 엔비디아의 H100/B200의 가격($25K~$30K)에 주목하지만, 실제 구매 수량은 바벨 전략에 따라 결정된다. 모건 스탠리 및 각종 언론 보도에 따르면, **엔비디아가 2026년 CoWoS 웨이퍼 총 수요의 60%를 차지할 것으로 전망**하며, 나머지 **40%는 구글의 TPU, 메타의 MTIA, 아마존의 Tranium과 같은 북미 CSP들의 자체 칩, 즉 ASIC이 차지**하고 있다. 위 비율은 CUDA 생태계 및 성능 우위 등의 AI 생태계 내 엔비디아의 독보적 지위를 고려하여, ASIC의 약진에도 불구하고 계속해서 유지될 것으로 전망한다. 위 ASIC 칩들의 가격은 NVIDIA의 MSRP $30,000의 3분의 1 수준인 $10,000로 예측된다. 이처럼 올해는 하이퍼스케일러들이 자체 칩으로 대거 전환하는 임계점이며, Citi Research에 따르면 ASIC 가격 경쟁으로 NVIDIA는 점유율 이탈을 막기 위해 **공격적인 번들링 할인**을 확대할 전망이다.

22,000달러의 GPU가 60%, 10,000달러의 ASIC이 40%

NVIDIA는 직접 판매보다는 DELL, HPE, Supermicro와 같은 공인 OEM 파트너를 통해 시스템 형태로 제품을 공급하는 경우가 많고 H100의 사례를 보면 8개 이상의 GPU 주문 시 유닛당 가격이 **22,000달러에서 26,000달러 사이**로 형성된다. 더 나아가 100개 이상의 유닛을 주문하는 기업용 대규모 계약의 경우 **20,000달러에서 24,000달러까지 하락**하게 된다. 특히 NVIDIA의 Blackwell 기반 NVL72시스템은 72개의 GB200 GPU와 36개의 Grace CPU를 단일 랙에 통합한 액체 냉각 시스템이다. 가격은 300만 달러로 형성되어 있는데, 절반 가량을 차지하는 냉각 인프라와 케이싱, PSU의 비중을 제외하면 **개별 칩 가격은 22,000달러 수준**까지 내려오게 된다. 이러한 Blackwell 칩과 Hopper 칩의 할인율, 그리고 ASIC의 비중 확대 추세를 반영하여 22,000달러의 단가를 NVIDIA의 GPU가 형성한다 가정하였다.

Base -> 1,970만
Bull -> 2,814만

즉, 이를 기반으로 ASIC 칩을 포함한 전체 AI 가속기 칩의 ASP를 계산해보면 **17,200달러**라는 단가가 완성된다. 따라서 예상 CAPEX 기반으로 수요를 구해보면, **Base 케이스에서는 26년 1,970만개**가 나오게 된다. 앞서 CAPEX 전망 부분에서 기술했듯, **Bull 케이스, 즉 CAPEX/OCF 비율을 수요 성장의 상단에서 100%까지 적용한다면 2,814만개**의 수요가 요구된다. 다만 Bull Case 물량은 TSMC CoWoS Capa 제한으로 현실적으로 달성 불가능하며, AI CAPEX의 추가 집행 여력을 GPU 규모로 환산하여 체감하기 위한 참고자료로서 기술하였다.

CAPEX 리스크

CAPEX 투자와 매출이 직결되지 않는다

AI BM:
1) 광고
2) 클라우드

빅테크의 막대한 AI CAPEX 집행은 현재 **광고 매출**과 **클라우드 매출**로 정당화된다. 첫째, **머신러닝 및 생성형 AI 기반 광고 자동화·개인화 기술**이 대표적이며, 구글 검색·유튜브, 메타 앱 패밀리(페이스북·인스타그램·왓츠앱·쓰레드 등)에서 개인정보 처리 확대를 위한 AI 하드웨어 투자가 매출 증대를 뒷받침한다. 둘째, OpenAI(GPT), Anthropic(Claude), Oracle 등 대형 고객을 대상으로 **AI 학습·추론을 지원하는 클라우드 서비스**로, 아마존 AWS, 마이크로소프트 Azure, 구글 클라우드가 대표적이다.

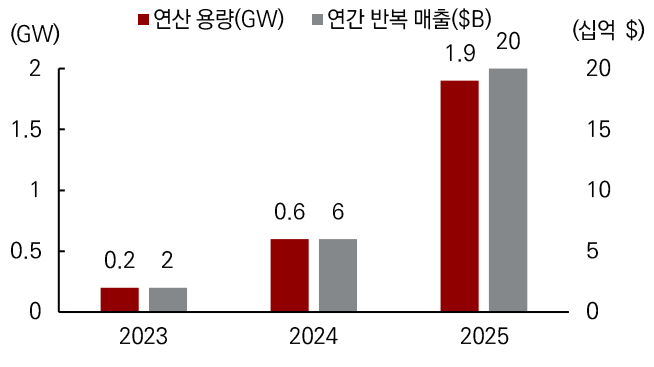
투자 = 매출 (x)
AI 인프라 확대가 단기 안에 매출로 직결되지 않는 구조

그러나 광고 고도화의 연산량 점유율은 공개되지 않지만 빅테크 CEO들의 강조 대상(월드 모델·에이전틱 AI·피지컬 AI 등 미래 영역)이 아니라는 점에서 대규모 CAPEX를 설명하는 수준이 아니라고 유추할 수 있다. 클라우드의 경우, "기업용 LLM → 에이전틱 AI"가 최종 소비자 생산성 혁신으로 CAPEX를 상쇄할 만큼의 성과를 내야 하나 아직 확인되지 않았다. 결국 현재의 CAPEX 증가세를 주도하는 동력은 '과소 투자에 따른 경쟁 도태'에 대한 공포(FOMO)다. 알파벳과 메타의 CEO들은 공통적으로 인공지능 분야에서 과잉 투자의 위험보다 과소 투자의 위험이 훨씬 크다는 점을 강조하며 공격적인 지출을 정당화해 왔다.

OpenAI:
컴퓨팅 용량이 매출과 직결, 그러나 AI 서비스 만으로 수익성 보장 X

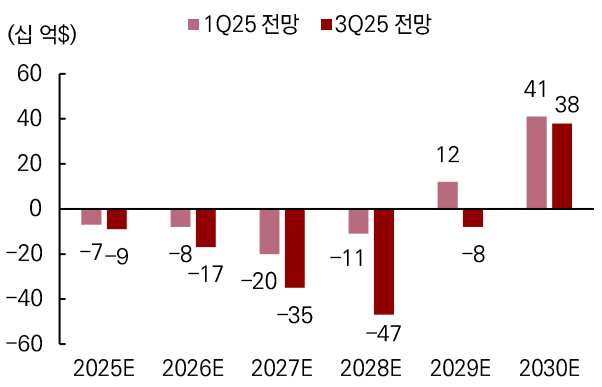
빅테크 4사는 기존 광고·CSP 사업과 연계해 투자 회수를 정당화하나, **OpenAI는 프론티어 AI 소프트웨어 직접 제공으로 컴퓨팅 용량과 매출이 직결**된다. 구독(개인+기업) 70%, API(기업 사용량 직결) 30% 구조에도, 추론 성능 변화에 유료 구독자가 서비스를 기민하게 전환함에 따라 전체 매출이 컴퓨팅 용량과 거의 비례함을 알 수 있다. 그러나 이렇게 직접 AI로 돈을 버는 OpenAI의 수익성 전망은 밝지 않다. The Information에 따르면 OpenAI는 2026년 140억 달러 적자, **2028년 말 누적 440억 달러 손실** 후 2029년 140억 달러 흑자 전환을 예상했다. 도이치은행은 OpenAI의 1.4조 데이터센터 투자 약속을 제외하고도 **2029년까지 누적 손실을 1,430억 달러**까지 전망했다.

그림 19. OpenAI 연산 용량-매출 비교대조



자료: OpenAI, KUVIC 리서치 1팀

그림 20. OpenAI 지연된 FCF 기준 흑자 전환 시점



자료: 언론 보도 종합, KUVIC 리서치 1팀

수익성 리스크 분석

26년은 수익성을 입증해야 할 원년

따라서 수익성에 의한 CAPEX 축소 리스크를 분석한다. **2026년은 집행된 자본 투자의 수익성을 입증해야 할 원년**이 될 것으로 전망한다. 이미 주식 시장에서는 25년 10월 메타의 대규모 부채 발행과 26년 1월 마이크로소프트의 자본 지출 급증에서 과잉 투자를 우려했다. 이는 19세기 철도 버블이나 20세기 말 IT 버블 당시, 시장의 관심사가 '확장성'에서 '수익성'으로 이동하며 인프라 구축의 고점을 달성했던 역사적 사례와 궤를 같이한다. 결국 확실한 수익성에 대한 근거가 부족하다면 2026년을 기점으로 CAPEX

축소 리스크를 부각될 가능성이 크다.

운영비용부터 회수할 수 있어야

다만, 현실적인 관점에서 시장은 투입된 누적 CAPEX 전체의 단기 회수 가능성보다는, **AI 서비스를 통해 발생하는 매출이 매년 발생하는 단기 가변 비용(OPEX)을 우선적으로 보전할 수 있는지** 여부에 집중할 것으로 예상한다. 이에 따라 리서치 1팀은 GPU 구매 및 데이터 센터 건축 등 일회성 선투자 비용(장기 회수 가능한 고정 비용)과 GPU 감가상각, 전력 공급 및 냉각 시스템 유지비 등 운영 비용(단기 비용)을 분리하여, **운영 비용을 위주로 CAPEX가 정당화되는 AI 매출 목표를 산출**했다. 또한, 토큰 당 추론 비용 감소, ASIC 확대 등 **수익성 개선 변수가 이 간극을 메우기 위해 어느 정도로 개선되어야 하는지**를 확인했다.

25년 CAPEX 정당화를 위해 26년에는 1,665억 달러는 벌어야

결론적으로, **CAPEX 정당화를 위한 단기 AI 매출 목표는 25년 1,665억 달러, 26년 2,748억 달러**로 전망된다. 세쿼이아 캐피털(Sequoia Capital)이 25년 초 추정한 AI 매출 규모는 1,000억 달러로 **현재 AI 관련 매출은 단기 운영 비용조차 보전하기 어려운 상황**임이 확인된다. 따라서 26년 내 AI 매출 규모나 비용 측면에서 획기적인 개선이 이루어지지 않는다면 현재의 CAPEX 규모가 지속되기 어려울 것으로 전망한다.

표 5. CAPEX 정당화 매출 목표 추정

계산식	2025년	2026년	단위 및 추정 논리
일회성 비용 총합	200%	200%	
- GPU 구매 비용	100%	100%	GPU 구매 비용을 기준으로
- 건설 및 냉각 등 인프라 비용	100%	100%	인프라 비용은 GPU 구매 비용만큼 발생한다고 가정
운영 비용 총합	40.5%	40.5%	
- GPU 감가상각	33%	33%	실질적 GPU 교체 주기 3년 가정, 매년 GPU 구매 비용의 33%
- 전력 비용	5%		블랙웰 GPU 연간 소요 전력(가동률 61% 가정) 및 미국 상업용 전기료 바탕 추정
- 냉각 비용	2.5%		GPU 소요 전력 비용의 50% 가정
AI 데이터센터 CAPEX 종합	5,140	8,480	단위: 억 달러
GPU 구매 금액	2,056	3,392	전체 CAPEX의 40% 가정
연도 별 발생 일회성 비용	4,112	6,784	단위: 억 달러
연도 별 발생 운영비용	833	1,374	단위: 억 달러
합리적 매출총이익률 가정	50%	50%	AI 최종 소비자(기업) 입장
운영 비용 정당화 매출 목표	**1,665**	**2,748**	단위: 억 달러
일회성 포함 정당 매출	5,777	9,532	단위: 억 달러

자료 : KUVIC 리서치 1팀

추론 비용 절감과 ASIC 확산이 아주 획기적이어야

비용 감소만으로 수익성을 입증하기 위해서는 **GPU 구매 비용의 33% 수준인 GPU 감가상각 비용을 16%로 절반 가량 축소**해야 한다. 이는 **토큰당 추론 비용의 급감과 ASIC(주문형 반도체) 확대**를 통한 실질 칩 단가 인하를 통해 달성 가능할 전망이다. DRAM과 eSSD 설계 개편을 통한 데이터 센터 가동률 상승으로 토큰 당 추론 비용이 추세적으로 연 2배씩 감소하는 와중, 2026년 본격 양산될 엔비디아 루빈(Rubin) 플랫폼이 추론 비용 10배 절감을 목표로 하고 있다. 또한 ASIC의 경우 통상 GPU 대비 3분의 1 가격을 가진다. **26년에 수익성 우려를 판단할 때 추론 효율화를 나타내는 지표들에서 실제로 획기적인 원가 절감 추세가 나타나는지를 확인해야**할 것이다.

근본적으로 AI 서비스 매출 연 50% 성장해야

그러나 근본적으로는 최종 사용자 단계에서의 **AI 서비스 매출이 연간 50% 이상 성장해야만 하는 상황**이다. GPU 감가상각 비용의 감소는 곧 GPU 신규 구매 수요의 둔화이며, 이는 전체 CAPEX 총량을 압축하는 핵심 요인이다. 특히 기술 혁신으로 인한 토큰당 추론 비용의 하락은 개별 작업에 투입되는 연산 자원량(Q)을 감소시키고, ASIC 도입 확대는 칩 단가(P)의 하락을 유도하는 하방 압력으로 작용한다. 이러한 공급측면의 효율화 수혜를 누리면서도 현재의 CAPEX 투자 규모를 유지하기 위해서는 '제번스의 역설(Jevons Paradox)'이 전제하는 것과 같이, **26년에는 낮아진 단가가 거의 무한한 신규 수요를 창출**

한다는 사실이 명확히 확인되어야 한다.

에이전틱 AI가 입증할 차례 그러나 쉽지 않은 현실

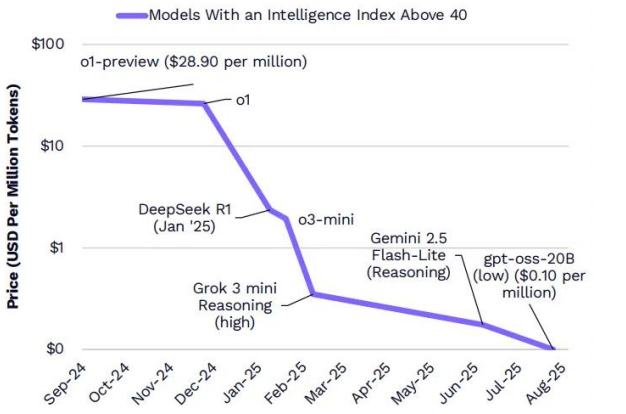
따라서 2026년부터는 단순한 인프라 구축을 넘어, **에이전틱 AI(Agentic AI)가 실무 프로세스에서 유의미한 인건비 대체 및 생산성 향상을 이끌어내며 서비스 매출의 가파른 우상향을 증명**해내야 한다. 시장은 특히 2026년을 인프라가 실질적인 기업 이익으로 치환되는 '수익화의 원년'으로 보며 에이전틱 AI의 실질적인 확산 여부에 주목할 것으로 보인다. 그러나 최근 MIT의 보고서에 따르면 기업들이 구축한 커스텀 AI 툴의 단 5%만이 실제 생산 단계에 진입했을 뿐이며, 도입 기업의 95%가 실질적인 비즈니스 가치를 창출하지 못한 채 '실험 단계'에 머물러 있는 것으로 나타났다. 또한 맥킨지(McKinsey)의 2025년 조사에 따르면 전 세계 기업의 88%가 업무에 AI를 활용하고 있으나 전사적 차원에서 이를 확장한 곳은 33%에 불과하며, 도입 기업의 39%만이 영업이익(EBIT)에 기여했다고 답했을 뿐 그마저도 기여도는 5% 미만에 그치고 있다. 이는 **최종 사용자 단에서의 AI 서비스 수요가 여전히 불투명하며, 인프라 투자 규모와 실질 수익 사이의 간극이 단기간에 해소되기 어려울 수 있음**을 시사한다.

그림 21. 26년 AI 서비스 매출 추정

기업	2026년 예상 AI 서비스 매출
마이크로소프트	235 ~ 260억 달러
구글	180억 달러
OpenAI	200억 달러
앤스로픽	180억 달러
xAI	50억 달러
그 외	130 ~ 150억 달러
합산	975 ~ 1,020억 달러

자료: 언론 보도 종합, KUVIC 리서치 1팀

그림 22. 동일 작업의 토큰 당 추론 비용 하락



자료: Ark Invest

엔비디아와 ASIC

NVIDIA: AI사이클 설계자 그리고 그의 진화

AI 사이클의 설계자, 그리고 GPU의 진화

지난 30년 간 그래픽 처리 장치(GPU) 제조사였던 엔비디아(NVIDIA)는 이제 인공지능(AI) 시대의 가장 강력한 '플랫폼 기업'으로 진화했다. 엔비디아의 역사는 단순한 하드웨어 성능 개선의 역사가 아니라, **컴퓨팅의 패러다임 전환(Paradigm Shift)** 그 자체로 봐도 무방하다.

엔비디아의 역사는 단순한 하드웨어 성능 개선의 과정이 아니라, 컴퓨팅의 단위와 정의를 바꾸는 패러다임 전환의 역사다. 2006년 CUDA를 통해 GPU를 범용 연산 장치로 해방시키며 생태계의 해자(Moat)를 구축한 엔비디아는, AI 연산 수요가 폭발할 때마다 준비된 하드웨어를 제시하며 시장의 표준을 정의해왔다.

AI의 시대를 연 엔비디아의 제품들

특히 최근의 생성형 AI(Generative AI) 사이클은 엔비디아의 아키텍처 로드맵인 **Hopper, Blackwell, Rubin**으로 이어지는 기술적 진보와 궤를 같이한다. 'AI의 아이폰 모멘트'를 연 Hopper(H100) 아키텍처는 트랜스포머 엔진을 탑재하여 거대언어모델(LLM) 학습의 물리적 토대를 마련했다. 이어 등장한 Blackwell은 단일 칩의 물리적 한계를 **칩렛(Chiplet) 구조**로 돌파하고, 데이터센터를 하나의 거대한 GPU처럼 작동하게 만드는 **'랙 스케일(Rack-Scale)' 개념**을 현실화했다. 향후 2026년 도래할 Rubin 아키텍처는 HBM4 메모리와 3nm 공정을 결합하여, 인간의 개입 없이 스스로 사고하는 '에이전틱 AI(Agentic AI)' 시대를 위한 엑사스케일(Exascale) 컴퓨팅 환경을 제공할 것으로 전망된다.

무어의 법칙이 둔화됨에 따른 새로운 기술들의 도입

무어의 법칙이 둔화됨에 따라, 엔비디아는 단일 칩의 성능을 높이기 위해 미세 공정 전환뿐만 아니라 **아키텍처 혁신, 패키징 기술(CoWoS), 그리고 정밀도 최적화(FP4/FP8)**에 집중하고 있다. 특히 Hopper에서 Blackwell로 넘어가면서 도입된 칩렛(Chiplet) 구조는 반도체 노광 장비가 찍어낼 수 있는 **면적의 한계(Reticle Limit)를 극복하기 위한 필연적인 선택**이다. 또한, 추론 시장의 효율성을 극대화하기 위해 데이터 크기를 4비트로 줄이면서도 정확도를 유지하는 **FP4 정밀도의 도입**은 하드웨어 성능을 소프트웨어적으로 배가시키는 핵심 기술이다.

표 6. 엔비디아 제품 정리

구분	적용공정	소모전력	다이 개수	AI연산(FP4)	AI연산(FP8)	메모리 대역폭	메모리 종류	메모리 용량	가격(K$)
H100	4nm	700w	1개		4 PFLOPS	3.35TB/s	HBM3	80GB	25~30
H200	4nm	700w	1개		4 PFLOPS	4.80TB/s	HBM3e	141GB	27~28 (E)
B100	4NP	700w	2개	14 PFLOPS	7 PFLOPS	8.00TB/s	HBM3e	192GB	
B200	4NP	1,000w	2개	20 PFLOPS	10 PFLOPS	8.00TB/s	HBM3e	192GB	30~40
R100	3nm	1,000w (E)	2개	50 PFLOPS (E)	25 PFLOPS	22.00TB/s	HBM4	288GB	
Rubin CPX	3nm	350~450w (E)					GDDR7	128GB	

자료: 각종 보도자료 종합, KUVIC 리서치 1팀

랙 스케일(Rack-Scale) 아키텍처: AI 인프라 회사로의 변화

AI 연산 수요가 폭증하면서 데이터센터의 **기본 구매 단위는 더 이상 개별 GPU 서버가 아니라, 수십 개의 GPU가 하나의 거대한 연산 시스템으로 통합된 랙(Rack) 단위로 빠르게 이동**하고 있다. 이는 단순한 하드웨어 확장이 아니라, AI 컴퓨팅의 병목이 "칩 성능"에서 "시스템 아키텍처 전체"로 이동했음을 의미한다.

–

호퍼(Hopper) 세대까지는 공랭 기반의 개별 GPU 서버를 네트워크로 연결하는 구조가 주류였다. 그러나 블랙웰(Blackwell) 세대의 **GB200 NVL72**부터는 설계 철학이 근본적으로 달라졌다. 엔비디아는 72개의 GPU를 하나의 NVLink 도메인으로 통합하기 위해 랙 후면에 5,000가닥 이상의 구리 백플레인(Copper Backplane)을 배치했다. **이는 광케이블 사용 시 필요한 광 트랜시버를 제거함으로써 랙당 약 20kW 수준의 전력 소모를 줄이고, 동시에 신호 지연(latency)을 최소화하기 위한 구조적 선택이다.**

이와 함께, 랙당 100kW를 상회하는 발열을 제어하기 위해 액체 냉각(Liquid Cooling)이 필수적으로 도입되었다. 결과적으로 GB200 NVL72는 단순히 GPU를 다수 장착한 서버 묶음이 아니라, **72개의 GPU가 하나의 거대한 가상 '슈퍼 칩'처럼 동작하는 단일 연산 도메인**에 가깝다. 이는 초거대 모델 학습 시 GPU 간 통신 병목을 대폭 줄여 실효 연산 성능(utilization)을 크게 향상시키는 구조다.

패러다임이 바뀐 AI시대의 경쟁력

이러한 변화는 데이터센터 인프라 전반의 재설계를 요구한다. 전력 밀도, 냉각 설비, 랙 배치 구조까지 모두 GPU 중심으로 다시 설계되어야 한다. 즉, AI 시대의 경쟁력은 개별 칩 성능이 아니라 **"GPU를 얼마나 효율적으로 대규모로 묶어 하나의 시스템처럼 동작시키느냐"**에 의해 결정된다.

이 흐름은 차세대 **루빈(Rubin)** 플랫폼에서 더욱 극대화될 전망이다. 2026년 출시가 예상되는 루빈 아키텍처에서는 집적도가 더욱 높아져, 최대 144개의 GPU가 하나의 랙(NVL144)에서 동작하는 구조가 제시되고 있다. 단일 랙 단위에서 엑사플롭스(ExaFLOPS)급 성능을 구현하는 것이 목표이며, 이는 데이터센터의 최소 연산 단위가 사실상 "슈퍼컴퓨터급 모듈"로 재정의되고 있음을 의미한다.

변화하는 엔비디아의 BM

결과적으로 엔비디아의 사업 모델도 구조적으로 변화하고 있다. 과거에는 GPU 칩 판매가 중심이었으나, **현재는 DGX 서버, HGX 플랫폼, NVL 랙, 나아가 데이터센터 단위 설계까지 포함하는 AI 인프라 스택 전반을 공급하고 있다.** GPU, NVLink, 네트워크, CPU(Grace), 소프트웨어(CUDA)까지 통합된 구조는 고객을 엔비디아 생태계에 깊게 결속시키는 효과도 가진다.

표 7. 랙 스케일 아키텍처 사양 정리

구분	적용공정	소모전력	다이 개수	AI연산(FP4)	AI연산(FP8)	메모리 대역폭	메모리 종류	메모리 용량	가격(10K$)
DGX H100	4nm	40kW	1개		~128 PFLOPS	107TB/s	HBM3	2.5TB	150~200
GB200 NVL72	4NP	130~150kW	2개	1.44 EFLPOS	~720 PFLOPS	576TB/s	HBM3e	13.5TB	300~400
Vera Rubin NVL144	3nm	215kW	2개	~3.60 EFLPOS	~1.2 EFLOPS	4.6PB/s	HBM4	~100TB	500 (E)

자료: 각종 보도자료 종합, KUVIC 리서치 1팀

빅테크의 엔비디아 의존도와 그를 벗어나기 위한 시도

빅테크의 엔비디아 매출 비중

엔비디아의 데이터센터 매출 절반 가까이 차지하는 빅테크 기업들

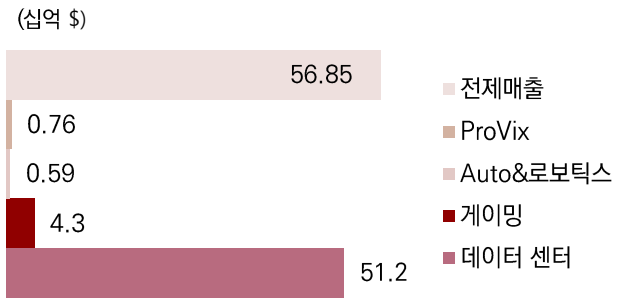
2025년 3분기 기준 엔비디아의 **데이터센터 매출은 512억 달러로 총 매출액인 568억 달러의 90%에 해당**하며, 이 매출액에서도 빅테크들의 비중이 압도적으로 높은 구조를 보인다.

우선 엔비디아의 가장 큰 단일 고객으로 평가되는 **마이크로소프트는 전체 매출의 약 15~17%를 차지하는 것**으로 추정되는데, 이러한 대규모 지출은 **Azure AI 인프라의 공격적인 확장과 오픈AI에 대한 초대형 컴퓨팅 파워 지원에 기인**한다. 마이크로소프트는 이글(Eagle) 슈퍼컴퓨터를 포함한 대규모 AI 클러스터를 구축하였으며, 이는 엔비디아의 GPU 및 인피니밴드(InfiniBand)와 같은 인터커넥트 기술에 거의 전적으로 의존하는 구조이다. 2024년 한 해 동안 48만 5천 개 이상의 GPU를 확보한 것으로 추정되며 2025년 말부터는 블랙웰(B200) 기반의 GB200 NVL72 랙 시스템 도입을 주도하고 있다. 비록 마이크로소프트가 자체 AI 가속기인 마이아(Maia)를 개발하고 있으나, GPT-5 이후의 최전선 모델 학습 로드맵은 여전히 CUDA 생태계와 블랙웰 아키텍처가 제공하는 압도적인 메모리 대역폭 및 확장성에 강하게 종속되어 있어, 자체 칩은 주로 추론이나 제한적 워크로드에 활용될 뿐 핵심 대규모 학습 영역에서는 엔비디아 GPU가 사실상 대체 불가능한 위치를 유지하고 있다.

다음으로 엔비디아 매출의 두 번째 축을 형성하는 **메타 플랫폼은 전체 매출의 약 13~15%를 기여하는 것**으로 분석되는데, 마이크로소프트와 달리 메타의 GPU 수요는 클라우드 재판매가 아닌 내부 AI 모델 학습과 추천 시스템 고도화에 온전히 집중되어 있다는 점이 특징이다. 라마(Llama) 시리즈의 오픈 웨이트 모델 전략은 메타로 하여금 경쟁사들을 압도하는 초대형 GPU 클러스터 구축을 불가피하게 만들었으며, 2025년 말 기준 메타는 60만 개 이상의 H100 환산 GPU를 확보한 것으로 보고되는데 이는 단일 기업 기준 세계 최대 수준이다.

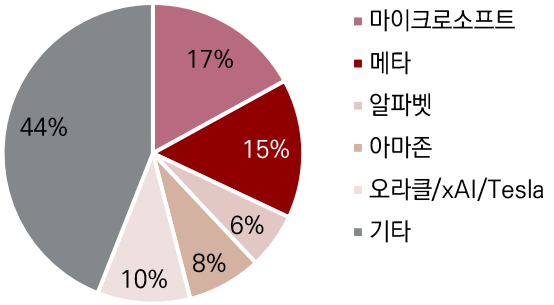
마지막으로 **각각 엔비디아 전체 매출의 약 6~8%를 차지하는 핵심 고객군인 알파벳과 아마존**은 자체 실리콘과 엔비디아 의존성이 공존하는 독특한 양상을 보인다. 이들 기업은 TPU와 Trainium·Inferentia 등 이미 성숙한 단계의 자체 AI 가속기 생태계를 보유하고 있음에도 불구하고 엔비디아 GPU에 대한 대규모 구매를 지속하고 있는데, 이는 자체 실리콘이 특정 내부 워크로드에서는 비용 효율성을 제공하나 범용성이나 소프트웨어 생태계 그리고 초대형 모델 학습 성능 측면에서는 여전히 한계를 노출하고 있기 때문이다. 특히 CUDA 기반 환경에 대한 개발자들의 선호도는 여전히 절대적이며 이는 사실상 산업의 표준으로 기능하고 있어, 클라우드 사업자의 핵심 경쟁력이 단일 아키텍처의 우월성이 아닌 고객이 요구하는 모든 인스턴스를 제공할 수 있는 선택지의 폭에 있다는 점을 고려할 때 AWS와 Google Cloud는 엔비디아 GPU 인스턴스를 제공하지 않을 경우 핵심 고객 이탈이라는 구조적 리스크에 직면하게 된다. 결과적으로 알파벳과 아마존은 자체 실리콘을 비용 최적화 및 워크로드 분산 용도로 활용하면서도 최신 엔비디아 GPU 및 GB200 NVL72와 같은 랙 스케일 솔루션을 병행 도입하는 이중 전략을 채택하고 있으며, 이는 엔비디아와의 단순한 경쟁 관계라기보다 상호 의존적인 공존 관계에 가까운 구조로 평가된다.

그림 23. 25년 3분기 엔비디아 매출 비중



자료: 엔비디아, KUVIC 리서치 1팀

그림 24. 빅테크의 엔비디아 데이터 센터 매출 차지 비중



자료: 엔비디아, KUVIC 리서치 1팀

빅테크의 반란: ASIC

엔비디아에 맞서 자체 AI 가속기를 개발하는 빅테크 기업들

글로벌 빅테크 기업들은 NVIDIA GPU에 대한 높은 의존도를 낮추기 위해 대규모 투자를 바탕으로 **자체 AI 가속기(ASIC) 개발을 가속화**하고 있다. Google의 TPU는 이미 다수 세대에 걸쳐 고도화된 성숙 단계에 진입한 상태이며, Amazon의 Trainium2와 Microsoft의 Maia 100 역시 본격적인 도입 구간에 진입하며 ASIC 시장의 존재감이 빠르게 확대되고 있다. 이는 단순한 기술 다변화가 아니라, **AI 인프라의 구조적 비용 문제에 대응하기 위한 전략적 선택으로 해석**된다.

ASIC 확산의 배경 전력비용 상승과 이에 따른 TCO 부담 증가

ASIC 확산의 가장 핵심적인 **배경은 전력 비용 상승과 이에 따른 TCO(Total Cost of Ownership) 부담 증가**에 있다. **AI 데이터센터에서 연산 비용의 상당 부분은 전력과 냉각에서 발생**하며, GPU 중심 인프라 구조는 고성능과 동시에 높은 전력 밀도를 수반하는 특성이 있다. 반면 빅테크가 설계한 ASIC은 특정 워크로드에 최적화된 구조를 통해 **전력 대비 연산 효율을 높이는 방향으로 설계**되는 특징이 있다. 실제로 전력 효율 측면에서 NVIDIA H100이 5~10 수준으로 평가되는 반면, Google TPU Trillium은 15~20, AWS Trainium2는 10~15, Intel Gaudi3는 12~18, Groq LPU는 20 이상 수준으로 제시되며,

동일 전력 조건에서 더 많은 연산을 처리할 수 있는 구조적 가능성이 부각되고 있다.

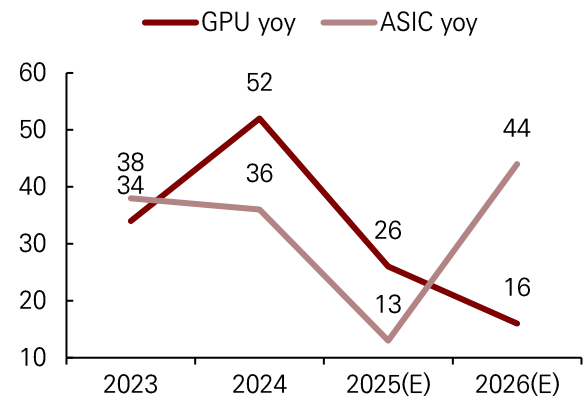
또한 ASIC은 에너지 절감 측면에서도 데이터센터 운영 부담을 낮추는 대안으로 평가된다. 대규모 AI 클러스터 환경에서는 단순 전기요금뿐 아니라 전력 설비 증설, 냉각 인프라 투자, 랙 전력 밀도 제한 등이 중요한 제약 요인으로 작용한다. 전력 효율이 개선될 경우 이러한 인프라 투자 부담이 동반 축소되는 효과가 발생하며, 이는 AI 서비스의 장기적인 수익성 확보 측면에서 중요한 요소로 작용한다.

ASIC의 강점

여기에 모델 및 서비스 맞춤형 최적화가 추가적인 동력으로 작용한다. GPU가 범용 연산에 강점을 가지는 반면, 빅테크의 ASIC은 자사 모델 구조와 서비스 패턴에 맞춰 불필요한 연산 경로를 제거하고 데이터 흐름을 단순화하는 구조로 설계되는 특징이 있다. 이는 동일한 연산 결과를 더 낮은 전력과 자원으로 처리할 수 있도록 하여 비용 구조를 근본적으로 개선하는 효과를 가져오게 된다.

결과적으로 전력 효율 개선과 워크로드 맞춤 최적화라는 두 축이 결합되면서 AI 인프라의 중심 구조는 GPU 단독 체계에서 GPU와 ASIC이 병행되는 하이브리드 체계로 이동하는 흐름이 점차 뚜렷해지고 있다. 이는 NVIDIA 중심의 AI 반도체 생태계에 대한 견제이자, 빅테크가 인프라 주도권을 다시 확보하려는 구조적 전략 변화로 해석할 수 있다.

그림 25. AI서버 성장 추이: GPU vs ASIC



자료: TrendForce, KUVIC 리서치 1팀

그림 26. ASIC 전력 효율성 비교

가속기	전력효율	GPU대비 에너지 절감율 (%)
NVDIA H100	5~10	기준값(Baseline)
Google TPU Trillium	15~20 (전력효율 67%개선)	30~60
AWS Trainium2	10~15	40
Groq LPU	20+	50~70 (약 1/10 수준 에너지)
Intel Gaudi3	12~18	40
Cerebras WSE-3	15~25	30~40

자료:Bestgpusforai, KUVIC 리서치 1팀

표 8. 빅테크 ASIC 생산 중 제품 사양 비교

구분	적용공정	소모전력	다이 개수	AI연산(FP4)	AI연산(FP8)	메모리 종류	메모리 용량	가격(K$)	출시 일정
B200	4NP	1000w	2개	14 PFLOPS	7 PFLOPS	HBM3e	192GB	30~40	24년
TPU v6	3nm	400w~500w (E)	1개			HBM3	32GB		24년
Trainium v2	5nm	500w~600w (E)	2개	5.2 PFLOPS	1.3 PFLOPS	HBM3	96GB		24년~26년 상반기
Maia v1	5nm	1,000w	2개			HBM2e	64GB		24년
MTIA v2	5nm	1,000w (E)	2개			LPDDR5	128GB		25년 4분기 이후

자료: 각종 보도자료 종합, KUVIC 리서치 1팀

표 9. 빅테크 ASIC 최신 제품 사양 비교

구분	적용공정	소모전력	다이 개수	AI연산(FP4)	AI연산(FP8)	메모리 종류	메모리 용량	가격(K$)	출시 일정
R100	3nm	1,000w (E)	2개	50 PFLOPS (E)	25 PFLOPS	HBM4	288GB		26년 하반기
TPU v7	3nm	600w~800w (E)	2개			HBM3e	216GB		26년 하반기 이후
Trainium v3	3nm	500w~600w (E)	2개 (E)			HBM3e	288GB		26년 1분기
Maia v3	3nm	1,000w (E)	2개			HBM4			27년
MTIA v3	3nm		2개			HBM3e			26년 상반기

자료: 각종 보도자료 종합, KUVIC 리서치 1팀

그럼에도 불구하고 다시 엔비디아

엔비디아의 강력한 해자에서 벗어날 수 없다.

빅테크 기업들이 자체 칩(ASIC) 개발을 통해 엔비디아 의존도를 낮추려는 시도는 분명 존재하지만, 이는 어디까지나 내부 비용 통제를 위한 헷징(Hedging) 전략일 뿐 엔비디아의 지배력을 근본적으로 대체하기에는 역부족이다. 엔비디아는 단순한 반도체 칩을 넘어 **소프트웨어(CUDA), 네트워크(NVLink), 그리고 시스템 아키텍처(AI Factory)**를 아우르는 **'풀 스택(Full-stack)'**을 통해 경쟁사가 모방할 수 없는 구조적 해자(Moat)를 구축했기 때문이다. 이러한 구조적 해자는 **소프트웨어 생태계에 기반한 개발자 락인, 범용성과 경제성의 균형에서 발생하는 ASIC의 구조적 한계, 하드웨어와 AI 알고리즘이 함께 진화하며 벌어지는 세대 간 시간차 격차**라는 세 가지 축을 통해 형성된다.

CUDA 생태계에서 벗어날 수 없는 개발자들_

첫째, CUDA 생태계가 만든 '개발자 락인(Lock-in)'은 기술적 우위를 넘어선 산업 표준이다. 경쟁사들이 하드웨어 스펙을 따라잡더라도 소프트웨어 생태계의 격차는 메우지 못하고 있다. 현재 **전 세계 CUDA 개발자는 약 400만 명**에 육박하는 반면, 경쟁사인 AMD의 소프트웨어 스택(ROCm)을 실무적으로 다룰 수 있는 전문 인력은 2,000명 수준에 불과하다. 기업 입장에서 엔비디아 외의 대안을 선택하는 것은 인력 수급의 난항과 막대한 코드 변환 비용을 감수해야 하는 모험이다. 실제로 독자 노선을 고집하던 애플조차 자사의 머신러닝 프레임워크(MLX)가 CUDA를 지원하도록 수정했을 만큼, 고성능 AI 개발 환경에서 엔비디아는 거스를 수 없는 표준으로 자리 잡았다. 개발자나 기업은 복잡한 CUDA 함수를 처음부터 짤 필요 없이, 라이브러리를 가져다 쓰기만 하면 즉시 세계 최고 수준의 GPU 가속 성능을 얻게 된다. CUDA는 단순한 프로그래밍 모델이 아니라, 특정 산업의 문제를 해결하는 구체적이고 강력한 솔루션의 집합체로 엔비디아의 해자를 구축하고 있다.

표 10. CUDA 기반 핵심 라이브러리 및 개발 도구 생태계

라이브러리	분야	설명/기능
CuPy	수치 연산	가장 널리 사용되는 수치 라이브러리 Numpy 문법을 CUDA로 호환/구현한 것
Aerial & Shona	통신	세계 최초 GPU 가속 5G/6G 무선 신호 처리. 소프트웨어 정의 방식으로 AI를 5G/6G에 도입 가능
Parabricks	유전체학	유전체 분석
MONAI	의료 영상	의료 영상 처리
Earth-2	기상 예측	기상 예측
cuQuantum	양자 컴퓨팅	양자-고전 컴퓨터 아키텍처 및 시스템
cuEquivariance, cuTensorNet	텐서 수학	텐서 수학 라이브러리
Megatron, TensorRT-LLM, NeMo, Dynamo	딥러닝	딥러닝 학습 및 추론 라이브러리. 최근 대규모 AI 팩토리를 위한 새로운 운영체제(Dynamo)도 포함
cuDF	데이터 프레임	Spark, SQL과 같은 구조화된 데이터 가속
cuML	머신러닝	고전적 머신러닝
Warp	시뮬레이션	CUDA 커널을 기술하기 위한 Pythonic 프레임워크. 매우 성공적
cuOpt	최적화	외판원 문제, 공급망 최적화 등 제약이 많은 대규모 변수 최적화 문제 해결
cuDSS, cuSparse	시뮬레이터	희소 구조 시뮬레이터 (CAE, CAD, 유체 역학, 유한 요소 분석 등 EDA/CAE 산업에 중요)

자료: 엔비디아 , KUVIC 리서치 1팀

자체 칩 개발에 어려움을 겪는 빅테크들

둘째, 빅테크들의 자체 칩(ASIC) 개발은 기술적 난이도와 범용성의 한계에 봉착해 있다. 구글(TPU), 아마존(Trainium), 마이크로소프트(Maia)가 자체 칩을 개발하고 있으나, 이는 엔비디아 GPU 비용을 줄이기 위한 '내부용(Internal)' 성격이 강하다. 클라우드 비즈니스의 핵심인 외부 개발자들은 여전히 범용성과 호환성이 뛰어난 엔비디아 GPU를 요구하기 때문에, 빅테크들은 경쟁력을 위해 울며 겨자 먹기로 엔비디아 칩을 구매해야 한다. 또한 칩 개발의 난이도는 예상보다 훨씬 높다. 마이크로소프트는 차세대 칩인 Maia 200(코드명: Braga)의 출시를 2026년으로 연기했으며, 출시되더라도 엔비디아의 차세대 칩인 Blackwell 성능에 미치지 못할 것으로 전망된다. 빅테크가 3년 걸려 특정 워크로드에 최적화된 칩을 개발해 내놓으면, 엔비디아는 이미 1년 주기로 범용 신제품을 출시하며 세대 격차를 벌려버리는 구조다.

젠슨 황이 "**구매 가능한 GPU보다 성능이 떨어진다면 ASIC을 개발할 이유가 무엇인가?**"라고 반문한 것은 이러한 자신감을 보여준다.

셋째, 하드웨어와 소프트웨어의 공진화(Co-evolution)를 통한 '시간차 격차'이다. 엔비디아는 AI 알고리즘의 발전 방향을 미리 파악하고 이에 최적화된 하드웨어를 내놓는 선순환 구조를 완성했다. 최신 AI 모델인 트랜스포머(Transformer) 아키텍처가 엔비디아 GPU 구조에 맞춰 진화했듯, 엔비디아는 새로운 아키텍처(Blackwell 등)를 내놓을 때마다 저정밀도 연산(FP4)과 같은 전용 기능을 탑재하여 성능을 비약적으로 향상시킨다. 스타트업이나 경쟁사가 특정 모델에 최적화된 칩을 개발해 내놓을 시점이면, 엔비디아는 이미 다음 세대 기술로 시장의 표준을 바꿔버려 후발 주자의 칩을 구식으로 만들어버린다.

결론적으로 엔비디아의 경쟁력은 단일 칩의 성능이 아니라, 칩·네트워크·소프트웨어가 결합된 **시스템 전체의 우위**에서 나온다. 경쟁사들이 부품을 만들 때 엔비디아는 'AI 팩토리'라는 공장 전체를 설계하고 있어, 당분간 그 격차를 좁히기는 어려울 것이다.

표 11. 엔비디아의 시스템적 우위

핵심 경쟁력	주요 전략 및 기술	이를 통한 우위	경쟁사에 미치는 영향
시스템 전체의 격차	단일 기술이 아닌 공급망, 기술, 속도 전체를 결합	- 견고한 시장 지배력 구축 - 작은 성능 우위는 쉽게 무력화 가능	- 칩 스펙 하나만으로는 공략 불가 - 생존을 위해선 몇 배의 압도적 효율격차 필요
네트워킹 장악	- 스케일업 독자 규격 “NVLink” - 스케일아웃 독자 규격 “InfiniBand”	- 수만 개 GPU 클러스터의 병목 구간 완벽 통제 - 개별 칩이 아닌 클러스터 전체의 효율 극대화	- 단순 GPU 칩만으로는 흉내 낼 수 없는 시스템 레벨의 격차 발생
메모리 및 공정 선점	- SK하이닉스와의 파트너십으로 최첨단 HBM 선제 확보 - TSMC 최첨단 공정 우선 활용	- 최신 HBM 메모리 최초, 최대 물량 확보 - 근본적인 물리적, 기술적 우위 선점	- 경쟁사는 항상 이전 세대 기술 또는 부족한 물량으로 경쟁해야 함
속도의 지배	기획 → 출시 → 대량 생산까지의 압도적 사이클 속도	- 시장 변화에 가장 빠르게 대응 - 차세대 제품을 통한 시장 연속 지배	- 어렵게 제품을 출시해도, 엔비디아가 이미 다음 세대 제품을 발표/공급하는 사이클

자료: Dylan Pate, KUVIC 리서치 1팀

TSMC

AI 사이클이 만든 TSMC의 독점적 지위

선단공정 수율이 만든 AI 반도체 독점 구조

공정 격차와 엔비디아의 절대 의존: 'AI 필수 인프라'가 된 TSMC

TSMC는 글로벌 파운드리 산업 내에서 선단공정(7nm 이하) 기준 수율과 생산 안정성 측면에서 구조적인 경쟁우위를 확보한 유일한 사업자다. 특히 AI 가속기와 같이 칩 면적이 크고 설계 복잡도가 높은 제품군일수록 초기 수율 격차가 원가 구조, 성능 구현, 출하 안정성에 미치는 영향은 비선형적으로 확대된다. 이 영역에서 TSMC와 경쟁사 간 기술 격차는 단기간 내 해소되기 어려운 수준에 머물러 있다.

이러한 환경 속에서 엔비디아는 공시상 복수 파운드리(TSMC·Samsung)를 활용하고 있으나, A100·H100·Blackwell로 이어지는 **데이터센터용 AI GPU는 TSMC 생태계에 대한 의존도가 사실상 절대적인 구조를 형성하고 있다.** 데이터센터용 GPU 수요가 급증하는 과정에서 엔비디아는 2025년 중 TSMC의 최대 고객으로 부상하며, 기존 최대 고객이었던 애플을 추월했다.

그 결과, 엔비디아 GPU의 시장 지배력 확대 → TSMC 선단공정 가동률 상승 → TSMC의 기술·가격 결정력 강화로 이어지는 구조가 형성되었다. 이는 AI 반도체 생태계 전반에서 TSMC가 단순한 제조 파트너를 넘어 사실상 필수 인프라로 기능하게 된 핵심 배경으로 판단된다.

AI성능 진화가 만든 CoWoS 필수화

AI성능 고도화의 병목, HBM과 CoWoS의 필수적 결합

본 보고서 추정에 따르면 **26, 27년 말 기준 CoWoS 웨이퍼 캐파는 12.5만 장, 14만 장이며, 이를 GPU 규모로 환산하면 26년 연간 1755만 장, 27년 2385만 장 규모에 달한다**. 이는 앞서 언급했듯, 빅테크 CAPEX로 환산한 GPU 규모(26년 1,755만 장, 27년 2,386만 장)와 비슷한 수준으로, **TSMC CoWoS 캐파가 글로벌 AI 가속기 물량을 결정하는 초과 수요의 양상을 여실히 보여준다. 이러한 첨단 후공정 병목 상황은 ASIC칩 주도로 인텔 첨단 후공정인 EMIB까지 기회를 줄 것으로 전망한다.**

AI 모델의 성능 고도화는 연산 유닛의 개선에 그치지 않고, 메모리 대역폭의 구조적 확장을 전제 조건으로 요구하고 있다. 파라미터 수 증가와 학습·추론 과정에서의 토큰 처리량 급증으로 인해, 단일 DRAM 구성이나 기존 패키징 구조만으로는 GPU 연산 성능을 충분히 활용하기 어려운 국면에 진입했다. 이에 따라 HBM의 다층 적층과 GPU 간 초고대역폭 연결을 구현하는 첨단 패키징 기술이 AI 성능 향상의 핵심 요소로 부상했으며, 이 과정에서 TSMC의 CoWoS(Chip-on-Wafer-on-Substrate)는 사실상 표준 공정으로 자리매김하고 있다.

CoWoS는 단순한 후공정 기술을 넘어, 선단 로직 공정과 HBM 인터페이스의 정합성, 대면적 인터포저 기반 수율 관리, 대량 양산 경험이 동시에 요구되는 고난도 통합 공정이다. 이로 인해 OSAT이나 타 파운드리 업체가 단기간 내 동일한 수준의 대체 역량을 확보하기는 제한적인 상황이다. 결과적으로 AI GPU 수요 확대는 곧바로 CoWoS 캐파 부족으로 연결되고 있으며, 이는 TSMC 첨단 패키징의 전략적 중요성을 한층 부각시키고 있다.

종합하면, AI 성능 고도화 → HBM 다층 적층의 필수화 → CoWoS 수요 급증 → TSMC의 전략적 지위 강화라는 구조가 고착화되고 있다. 이러한 흐름은 AI 사이클이 지속되는 한, TSMC가 AI 반도체 밸류체인 내에서 차지하는 위상이 구조적으로 훼손되기 어렵다.

선단 공정에서의 구조적 병목

선단 공정 수요 증가

HPC중심의 매출 믹스 전환: 3nm 풀캐파 도달

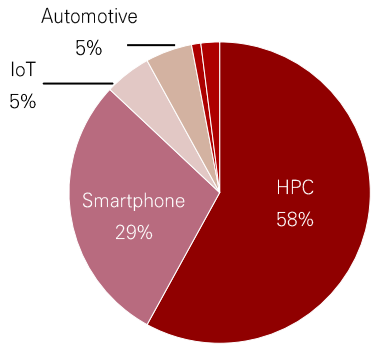
TSMC의 2025년 매출에서 HPC 부문은 약 58%를 차지하며, 스마트폰·IoT 등 기타 플랫폼 대비 가장 높은 비중을 기록했다. 특히 HPC 매출은 YoY 48% 성장하며, 전체 매출 성장을 견인하는 핵심 동력이다. 이러한 매출 믹스 변화는 최근 수년간 HPC 비중 확대가 곧바로 5nm·3nm 등 선단공정 수요 증가로 연결되는 구조가 고착화되었음을 시사한다.

AI 가속기, 데이터센터용 CPU 및 ASIC 등 HPC용 칩은 대면적 설계와 높은 공정 난이도를 특징으로 하며, 이로 인해 초기 수율과 생산 안정성이 핵심 경쟁 요소로 작용한다. 해당 영역에서 TSMC는 경쟁사 대비 우수한 수율과 축적된 양산 경험을 기반으로 사실상 독보적인 위치를 확보하고 있으며, 이는 주요 AI 고객사의 주문이 선단공정에 집중되는 구조를 강화하고 있다.

이러한 수요 환경 속에서 선단공정 캐파는 AI 수요 급증 국면에서 구조적인 병목 요인으로 작용하고 있다. 특히 3nm 공정은 2025년 들어 가동률이 빠르게 상승하며, 3분기 기준 사실상 풀캐파 가동 수준에 도달한 것으로 파악된다. TSMC는 2026년 말까지 3nm 월 생산능력을 약 20만 장 수준으로 확대할 계획이나, 증설이 진행되더라도 AI 고객사의 수요 증가 속도를 감안할 경우 공급 타이트 현상은 상당 기간 지속될 가능성이 높다.

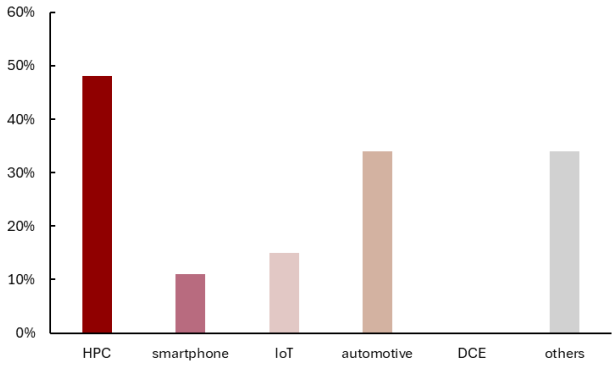
이는 선단공정 증설이 대규모 자본투자와 장기간의 수율 안정화 과정을 수반하는 특성에 기인한다. 결과적으로 **TSMC 선단공정 캐파 부족은 AI 반도체 공급 사이클 전반을 제약하는 구조적 병목으로 작용하고 있다**.

그림 27. TSMC 2025 플랫폼별 매출



자료: TSMC

그림 28. TSMC 2025 플랫폼별 성장률(YoY)



자료: TSMC

TSMC의 선단공정 증설에도 초과하는 수요

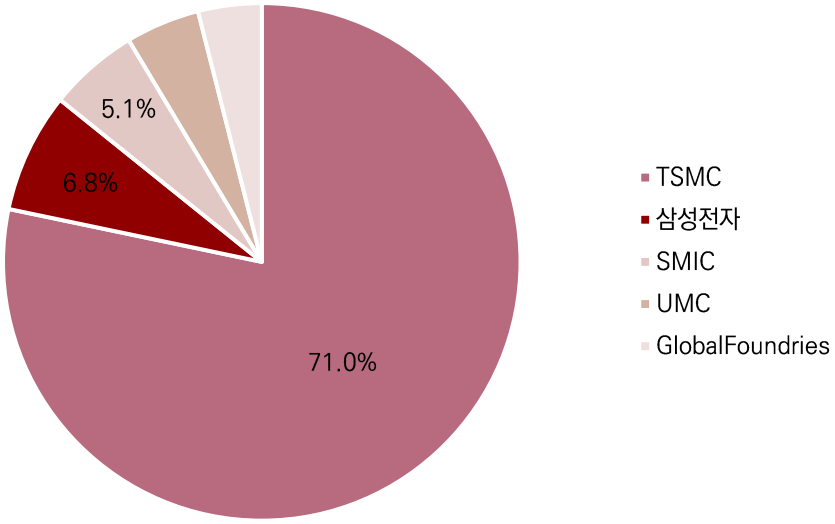
TSMC의 증설 한계와 삼성 파운드리의 낙수효과 수혜 전망

하지만 이러한 **선단공정에서의 초과 수요분이 삼성 파운드리로 넘어오면서 수혜를 입을 것으로 예상된다**. TSMC는 공격적인 증설 계획을 통해 파운드리에서의 우위를 더욱 강화하고 있다. 2nm 공정의 경우 2026년 말까지 월 10만 장 생산 목표를 설정했으며, 이 중 절반 이상은 이미 애플이 선점한 상태다. 2027년에는 월 15만 장까지 확대할 것으로 전망된다. 3nm 공정에서는 2025년 말 기준 월 15만 장을 돌파했으며, 2026년 말에는 월 20~22만 장 수준까지 증설될 것으로 예상된다

그러나 이러한 **대규모 증설에도 불구하고 선단공정 수요는 공급을 지속적으로 초과할 것으로 보인다**. AI 가속기, 스마트폰 AP, HPC 칩 등 선단공정을 요구하는 프로세서의 수요가 급증하고 있으며, TSMC의 증설 속도로는 모든 잠재 고객의 수요를 충족시키기 어려운 구조적 병목이 존재한다. TSMC가 애플, 엔

비디아, AMD, 퀄컴 등 기존 주요 고객에게 우선적으로 캐파시티를 배정하면서, 신규 진입 고객이나 중소형 팹리스 업체들은 필요한 물량을 확보하지 못하는 상황이 지속되고 있다. **삼성 파운드리는 인텔과 달리 AI4 등 고성능 칩 생산 경험과 빅테크 수주 레퍼런스가 확보되어 있는 상태이기에 위탁 생산을 맡기는 기업의 입장에서도 인텔보다 삼성 파운드리를 선택할 유인이 더 크다.**

그림 29. 파운드리 시장 점유율



자료: Trendforce, KUVIC 리서치 1팀

인텔 파운드리(18A)

펜서레이크를 통한 수율 검증 성공, 그러나 고객사 확보는 미진

인텔은 18A 공정을 통해 의미 있는 기술적 성과를 달성했다. CES 2026에서 공식 출시된 팬서레이크(Core Ultra Series 3)는 해당 공정으로 양산 중인 대표 제품이다. 최고 성능은 전작 대비 약 10% 향상에 그쳤지만, 전력 효율 및 그래픽 성능이 60% 이상 상승하며 노트북 CPU 시장에서 경쟁력을 확보했다. 18A 공정의 수율은 60% 이상으로 알려져 있어 안정화에 성공했다고 볼 수 있으며, 팬서레이크 출시와 함께 대량 양산 단계로 진입 중이다. 미국 애리조나 오코틸로 캠퍼스를 중심으로 18A 공정 파운드리 서비스를 제공하고 있고, 양산 능력은 확보했지만 외부 고객 유치에는 어려움을 겪고 있다. 따라서 주로 자사 제품 생산에 대부분의 시설을 활용하고 있는 상황이다.

표 12. 파운드리별 선단공정 양산 로드맵

	2025	2026	2027	2028	2029
TSMC	N2(2nm)	A16(1.6nm)	N2P	A14(1.4nm)	
삼성 파운드리	SF2(2nm)	SF2P(AI6적용)			SF1.4(1.4nm)
인텔 파운드리	18A(1.8nm)			14A(1.4nm)	

자료: 언론 보도 종합, KUVIC 리서치 1팀

인텔은 2028년 양산을 목표로 14A 공정을 준비 중이다. 이는 TSMC와 삼성전자보다 한 단계 앞선 로드맵이지만, 과거 인텔의 공정 개발 이력을 고려할 때 공정 노드의 우위가 곧 기술력 우위를 보장하지는 않는다는 점을 유의해야 한다.

인텔 파운드리의 한계

빅테크 기업들이 인텔 파운드리에 외주를 주저하는 가장 큰 이유는 신뢰성 부족이다. 특히 AI용 GPU 생산 경험 부재로 인한 낮은 수율이 문제로 지적된다. **인텔은 18A 양산에 성공했지만, 엔비디아 GPU와 같은 대형 고성능 AI 칩 생산 경험이 전무하다.**

GPU 생산의 높은 진입장벽: 엔비디아 테스트 중단이 시사하는 인텔의 한계

AI용 GPU는 CPU보다 결함 발생 확률이 높아, 인텔이 달성했다는 60% 수율이 GPU 제조에서는 훨씬 낮을 수밖에 없다. 실제로 **엔비디아는 2025년 말 인텔 18A를 활용한 차세대 칩 생산 테스트를 진행했으나 결국 중단**했다. NPU급(AI 노트북, 서버용 CPU)의 칩까지 양산이 가능한 상황이다.

반면 삼성전자는 2020년 8nm 공정으로 엔비디아 GPU '지포스 RTX 30 시리즈'를 생산한 노하우를 보유하고 있으며, 최근에는 닌텐도 스위치 2에 탑재된 엔비디아 GPU T239을 생산했다. 또한 AI 가속기 생산 경험도 축적하고 있다. 이러한 이유로 신뢰성 측면에서 인텔은 삼성전자에 크게 뒤처진다.

삼성 파운드리(SF2)

삼성 2nm수율의 반등과 '엑시노스 2600'양산 본격화

삼성전자의 2nm 공정 수율은 2026년 1월 기준 50~60% 수준에 도달한 것으로 보고되었다. 이는 초기 10~20%대에서 크게 개선된 수치이지만, 일부 업계에서는 40% 미만일 것으로 추정하기도 했다. TSMC가 2nm 공정에서 이미 70% 이상의 수율을 기록하고 있는 것과 비교하면 여전히 열위에 있지만, 3nm 공정 초기의 참담한 실패에 비하면 의미 있는 진전이다.

삼성전자는 2025년 초 30%였던 수율을 2025년 중반 50%까지 끌어올렸다. 통상적으로 파운드리 고객사가 양산을 맡기기 위해서는 최소 60% 이상의 수율이 보장되어야 하므로, 삼성전자로서는 2026년 상반기 중 수율 안정화가 매우 중요한 과제다.

수율 안정화를 증명할 수 있는 제품이 바로 삼성전자가 현재 양산중인 '엑시노스 2600'이다. 삼성전자는 2025년 12월 19일 2nm 공정으로 제조한 모바일 AP 엑시노스 2600의 양산에 돌입했다고 발표했다.

엑시노스 2600: 2nm 공정의 시험대

모바일에서 HPC까지: 2nm 레퍼런스 확보를 통한 1nm공정 반격 준비

엑시노스 2600은 2026년 2월 예정된 갤럭시 S26 언팩에서 공식 성능이 공개될 예정이며, 갤럭시 S26 일반 모델과 플러스 모델에 탑재되어 전체 물량의 25~30%를 차지할 것으로 전망된다. 더 중요한 것은 제품 출시 이후 실제 사용자들의 발열, 성능 등에 대한 평가다. 과거 엑시노스 2500이 3nm 공정 수율 목표 미달로 갤럭시 S25 시리즈 탑재에 실패했던 전례가 있어, **업계에서는 엑시노스 2600의 성공 여부를 주시하고 있다.**

삼성전자 파운드리가 TSMC에 압도적으로 밀린 결정적 이유는 3nm 공정 도입 초기 수율 부진이었다. 3nm 공정의 수율이 3년째 50% 수준에 머물러 있는 반면, TSMC는 90% 이상의 수율로 애플, 퀄컴, 엔비디아, 구글 등 주요 고객을 독점하고 있다.

흥미롭게도 삼성전자는 선단공정보다 8nm, 4nm 공정에서 더 많은 주문을 받고 있다. **파운드리 사업부의 과도한 '속도전' 전략은 사실상 폐기됐으며 기초를 잘 쌓아 시장의 신뢰를 되찾는 것이 목표다**. 모바일 AP를 시작으로 서버, HPC용 빅칩에서도 수주 실적을 쌓아올려 2나노 이후 1나노대 공정에서 반전을 노린다는 방침이다.

테슬라 수주의 전략적 의미 및 파운드리 시장 전망

삼성전자는 2025년 7월 테슬라와 약 23~24조 원(약 165억 달러) 규모의 AI6 반도체 공급 계약을 체결했다. 또한 지난해 10월에도 테슬라 실적을 발표하며 "삼성전자와 TSMC 모두가 AI5 생산에 참여할 것"이라고 밝히기도 하였다.

테슬라 AI5 칩은 단일 SoC 구성에서 엔비디아의 호퍼(Hopper) 아키텍처급 성능을, 듀얼 SoC 구성에서는 블랙웰(Blackwell)급 성능을 구현하는 것으로 알려졌다. 일론 머스크는 AI5가 일부 지표에서 AI4 대비 40배 향상되었으며, 엔비디아 칩 대비 10배 저렴한 비용으로 추론을 실행할 수 있다고 밝혔다. 특히 AI5는 150W의 전력만 소비하면서 700W를 요구하는 엔비디아 H100과 경쟁할 수 있는 에너지 효율을 자랑한다.

테슬라 AI16수주: 삼성 GAA 2nm공정의 기술적 신뢰성 검증

AI4는 삼성전자의 7nm 공정으로 생산되었으며, INT8 기준 100~150 TOPS의 성능과 약 384 GB/s의 메모리 대역폭을 제공한다. AI5는 TSMC와 삼성전자의 2~3nm 공정에서 양산될 예정이고, 머스크는 AI5가 AI4 대비 메모리 대역폭이 5배 증가할 것이라고 언급했다. AI6는 삼성전자가 GAA(Gate-All-Around) 기반의 2나노 양산 능력을 테슬라라는 빅테크 고객을 통해 입증했다는 데에 큰 의의가 있다. 이는 **모바일(Exynos)에 편중되었던 포트폴리오를 고성능 컴퓨팅(HPC) 및 전장(Automotive) 분야로 확장하는 계기가 되며, 특히 2027년경으로 예상되는 파운드리 흑자 전환과 TSMC 추격을 위한 기술적 신뢰도를 확보했다는 점에서 삼성전자에게 핵심적인 기회 요인으로 작용할 것이다.**

테슬라가 삼성전자를 선택한 핵심 이유는 TSMC의 생산 능력 한계와 물량 확보 문제다. TSMC의 선단 공정 라인, 특히 2nm와 3nm는 생산능력 포화 상태이며, 테슬라는 TSMC의 우선순위 고객이 아니다. 테슬라는 AI5 칩이 자동차뿐만 아니라 옵티머스 휴머노이드 로봇에도 탑재되어야 하므로, TSMC 단독 생산으로는 필요한 물량을 맞출 수 없다고 판단했다.

또한 삼성전자가 메모리와 파운드리를 동시에 운영하는 유일한 기업이라는 점도 중요한 차별화 요소다. HBM 개발 경험과 노하우를 파운드리에 응용할 수 있는 시너지 효과가 기대된다. 테슬라 AI6에 HBM 탑재 가능성이 언급되면서 삼성전자의 이러한 턴키(Turn-Key) 전략이 더욱 주목받고 있다.

파운드리 시장 전망: 삼성전자의 2위 입지 확보 기대

결과적으로 선단 공정 내 대형 AI 칩 생산의 신뢰성 측면에서 인텔은 삼성전자에 비해 열위에 있다. 자사 CPU 생산에 주력해 온 인텔과 달리, **삼성전자는 8nm GPU부터 AI 반도체 양산에 이르기까지 폭넓은 외부 고객사 레퍼런스를 축적**해 왔기 때문이다. TSMC의 선단 공정 점유율이 한계치에 도달하며 기존 고객사에 집중된 현 상황에서, GPU 생산을 통해 수율과 신뢰성을 선제적으로 확보한 삼성전자가 파운드리 시장 내 '2위 입지'를 공고히 할 것으로 기대된다

AI 칩에 의한 CoWoS 패키징 병목

CoWoS 패키징의 지위

메모리 대역폭 한계와 첨단 패키징 필수화: 수요를 하회하는 CoWoS 캐파

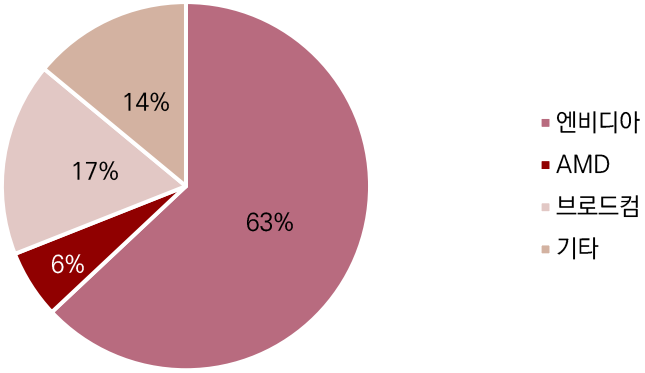
AI 모델이 고도화될수록 연산 성능 향상만으로는 성능 개선에 한계가 나타나고 있으며, 메모리 대역폭 확대가 사실상 전제 조건으로 작용하고 있다. 파라미터 규모 확대와 학습·추론 과정에서의 데이터 이동량 증가로 인해, GPU와 HBM을 고밀도로 연결할 수 있는 첨단 패키징에 대한 수요가 구조적으로 증가하고 있다. 이러한 환경에서 TSMC의 CoWoS(Chip-on-Wafer-on-Substrate)는 실리콘 인터포저 위에 로직 칩과 HBM을 배치하는 2.5D 패키징 방식으로, AI GPU와 AI ASIC 생산에 필수적인 기술이다.

CoWoS는 장비 도입부터 공정 안정화까지의 리드타임이 길고, 단기간 내 생산능력을 확대하는 데 구조적인 제약이 존재한다. 그 결과 **현재 AI GPU 공급은 설계나 전공정 능력보다도 CoWoS 패키징 캐파에 의해 직접적으로 제한받는 국면이 이어지고 있다**. 삼성전자의 I-Cube, 인텔의 EMIB·Foveros 등 대안적 패키징 기술이 존재하나, 주요 팹리스 고객사의 주력 AI 칩 양산은 여전히 TSMC CoWoS에 집중되어 있다.

TSMC는 지난 10여 년간 CoWoS 기술을 지속적으로 고도화하며 글로벌 고대역폭 패키징 분야에서 사실상의 표준적 지위를 구축해왔다. 로직 칩, 메모리 칩, 아날로그 칩을 하나의 대형 인터포저 위에 통합하는 이 구조는 엔비디아의 H100·H200·GB200(CoWoS-L 적용)을 비롯해 AMD의 MI300 시리즈 등에 적용되고 있다.

문제는 이러한 기술적 성숙도가 공급 측면에서는 오히려 병목을 심화시키고 있다는 점이다. 현재 CoWoS 생산능력은 구조적인 공급 부족 상태에 놓여 있으며, **엔비디아 단일 고객이 TSMC CoWoS 캐파의 절반 이상을 차지하고 있다**. UBS는 엔비디아의 Blackwell, Blackwell Ultra, Rubin으로 이어지는 제품 로드맵을 반영할 경우, 2026년 엔비디아의 CoWoS 웨이퍼 수요가 전년 대비 약 40% 증가한 67만 장 수준에 이를 것으로 전망하고 있다.

그림 30. CoWoS 고객사별 비중



자료: Trendforce, KUVIC 리서치 1팀

인텔 EMIB: 비용 효율적인 대안

EMIB: 고비용 CoWoS를 대체할 '가성비' 패키징 솔루션

인텔의 EMIB(Embedded Multi-die Interconnect Bridge)는 구조적으로 TSMC의 CoWoS에 비해 대역폭은 낮고 지연 시간은 다소 길지만, 비용과 수율 측면에서 가성비를 앞세운 패키징 솔루션이다. CoWoS가 대면적 실리콘 인터포저를 사용해 높은 대역폭과 낮은 지연 시간을 확보하는 대신 매우 높은 공정 비용과 열·기계적 복잡성을 감수하는 구조라면, **EMIB는 필요한 구간에만 실리콘 브리지를 삽입해 인터커넥트를 구현함으로써 설계·제조 난이도를 줄인다.**

현재 고성능 AI GPU 및 HBM 기반 가속기의 대다수는 CoWoS 계열 패키징을 채택하고 있으며, 실제로 엔비디아·AMD·브로드컴 등 주요 AI/네트워크 고객은 TSMC CoWoS에 강하게 락인(lock-in)되어 있다. 반면 EMIB는 아직까지 주로 인텔 자체 AI 칩 및 서버 CPU, 그리고 제한적인 외부 고객 설계에 쓰이는 단계로, 시장 점유율은 CoWoS 대비 한참 낮은 편이다. 즉, 패키징 스펙 기준으로는 CoWoS가 플래그십, EMIB가 가성비 옵션에 가까운 포지션이다.

ASIC전환과 빅테크의 탈 CoWoS흐름: EMIB의 실질적 수혜 전망

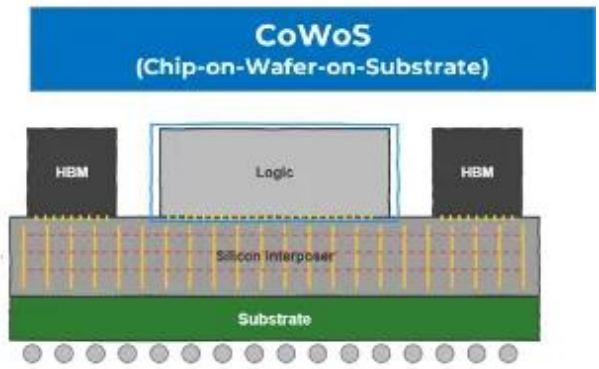
다만 AI 워크로드 구조가 변화하면서 EMIB의 전략적 가치는 점차 커질 가능성이 크다. 현재 GPU 중심의 범용 가속 환경에서는 극단적인 메모리 대역폭과 최소 지연 시간이 중요해 CoWoS가 최적화된 솔루션이지만, **향후 추론 비중 확대와 함께 ASIC(주문형 반도체) 기반 가속기의 비중이 높아질수록 '적당한 수준의 대역폭'과 '더 낮은 비용'의 조합을 원하는 수요가 늘어날 수밖에 없다. 이 지점에서 EMIB는 CoWoS 대비 성능에서 한 단계 양보하는 대신, 비용·수율·패키지 스케일 면에서 매력적인 대안으로 자리 잡을 수 있다.**

실제 빅테크들도 CoWoS 의존도를 낮추기 위한 전략적 옵션으로 EMIB를 검토하고 있다. CoWoS 캐파가 엔비디아와 일부 대형 고객에 우선 배정되면서, 구글 TPU, 아마존 Trainium 등 ASIC 계열 칩들의 패키징 수요는 구조적으로 후순위로 밀리는 상황이다. 이 때문에 **ASIC 진영은 CoWoS에만 의존하지 않는 대체 기술을 찾고 있으며, 인텔 EMIB와 유사한 브리지 기반 2.5D 구조, 혹은 OSAT의 CoWoS 유사 공정(Amkor, SPIL 등) CoWoS 물량 일부가 비(非)TSMC 라인으로 분산되는 흐름이 나타나고 있다.** 구글은 2027년 출시 예정인 TPU v9에 EMIB 적용을 검토하고 있고, 메타도 자사 훈련·추론 가속기 MTIA에 EMIB 도입을 고려하는 것으로 알려졌다.

기술적 관점에서 EMIB의 장단점은 다음과 같이 정리할 수 있다. 장점 측면에서는 첫째, 패키지 일부에만 실리콘 브리지를 삽입하는 구조 덕분에 전체 패키지의 열팽창계수(CTE) 불일치를 줄여 칩 휨(warpage)과 장기 신뢰성 문제를 완화한다. 둘째, 대면적 실리콘 인터포저를 사용하지 않기 때문에 제조 수율이 상대적으로 높고, 인터포저 사이즈 한계를 크게 받지 않아 대형 패키지 설계에 유리하다. 실제로 CoWoS-L의 레티클 크기 확장이 약 3.5배 수준으로 제한되는 반면, EMIB-M은 그보다 훨씬 큰 유효 패키지 면적을 지원하는 것으로 알려져 있어, 차세대 초대형 AI 패키지 설계에서 확장성이 높다. 셋째, 실리콘이 필요한 구간에만 브리지 형태로 들어가기 때문에, 동일 면적 기준 재료비 측면에서 CoWoS 대비 비용 효율이 높다.

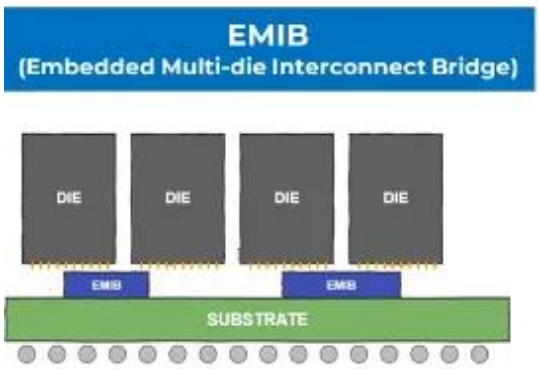
반면 **단점도 분명하다. EMIB는 구조적으로 데이터가 브리지를 통해 더 긴 경로를 이동해야 하므로, CoWoS 대비 지연 시간이 길고 인터커넥트 대역폭에도 현실적인 제한이 존재한다**. 이는 극단적인 메모리 대역폭과 최소 레이턴시가 곧바로 모델 학습 성능으로 직결되는 GPU·HPC용 플래그십 제품에는 불리한 조건이다. 따라서 EMIB는 엔비디아급 GPU 메인스트림보다는, 상대적으로 요구 사양이 낮고 TCO(총소유비용)와 전력 효율·가격을 더 중시하는 TPU, MTIA, Trainium과 같은 ASIC 계열에 더 매력적인 선택지가 된다.

그림 31.CoWoS 구조도



자료: TSMC

그림 32. EMIB 구조도



자료: 인텔

표 13. Intel EMIB vs. TSMC CoWoS

	Intel		TSMC		
	EMIB-M	EMIB-T	CoWoS-S	CoWoS-R	CoWoS-L
인터포저 (Interposer)	X	X	실리콘 (Silicon)	유기 폴리머 (Organic polymer)	유기 폴리머 (Organic polymer)
실리콘 브릿지 (Silicon Bridge)	MiM 커패시터 통합 및 기판 내장	TSV 통합 및 기판 내장	X	X	RDL 인터포저 내에 LSI 내장
레티클 사이즈 (Reticle Size)	6x	12x	3.3x	9x	9x
주요 적용 제품	Sapphire Rapids, Emerald Rapids, Granite Rapids	-	Hopper, TPU, MTIA, Maia…	Trainium	Blackwell (3.5x), Rubin (5.5x)
차이점	- 비용 효율적 - 더 큰 레티클 크기 - 더 적은 CTE(열팽창) 불일치 - 상대적으로 낮은 대역폭		- 높은 비용 - 상대적으로 작은 레티클 크기 - 높은 대역폭		

자료: TrendForce, KUVIC 리서치 1팀

삼성전자 I-Cube: 관심 부재

삼성전자의 I-Cube는 기술적으로 CoWoS와 유사하게 실리콘 인터포저 위에 칩을 배치하는 2.5D 패키징 기술이다. 삼성전자는 I-Cube(2.5D)와 H-Cube(3D) 플랫폼을 통해 로직 칩과 차세대 HBM을 통합할 수 있는 역량을 갖추고 있으며, **메모리와 파운드리를 동시에 운영하는 유일한 기업이라는 점이 차별화 요소다.**

대형 수주 부재

그러나 **수주 실적은 매우 제한적이다.** 2024년 국내 AI 스타트업 리벨리온의 'Rebel' 칩 생산 이후 눈에 띄는 대규모 수주 사례가 보고되지 않았다. 리벨리온의 Rebel 칩은 삼성전자 4nm 공정과 I-Cube-S 인터포저 패키징 기술을 활용하며, HBM3E 메모리 4스택을 통합하여 총 4.8TB/s의 메모리 대역폭을 제공한다.

표 14. 2.5D vs 3D

	2.5D IC					3D IC	
TSMC	InFO		CoWoS			SoIC	
	InFOPoP	InFOos	CoWoS-S	CoWoS-R	CoWoS-L	CoW	WoW
Intel			EMIB			Foveros	
			EMIB	Co-EMIB		Foveros	Foveros Direct
SAMSUNG			I-Cube			X-Cube	
			I-Cube-S	I-Cube-E	H-Cube	X-Cube bump	X-cube Hybrid bonding

자료: KUVIC 리서치 1팀

패키징 시장 전망

첨단 패키징 시장의 가장 큰 진입장벽은 설계 교체의 어려움이다. 칩 설계 단계에서부터 특정 패키징 기술과 긴밀하게 통합되어 개발되므로, 이미 CoWoS로 설계된 제품을 I-Cube나 EMIB로 전환하는 것은 상당한 재설계 비용과 시간이 소요된다. 엔비디아, AMD, 브로드컴 등 주요 고객사들이 이미 CoWoS 생태계에 깊이 뿌리내린 상황에서, 삼성전자나 인텔이 단기간에 시장 점유율을 확보하기는 어렵다.

그럼에도 불구하고 CoWoS 캐파 부족은 삼성전자와 인텔에게 기회가 된다. TSMC가 2026년 말까지 공격적으로 증설해도 월 120,000~130,000장 수준으로, AI 칩 수요 증가 속도를 따라가기에 충분하지 않다. 고객사들은 공급망 리스크 분산을 위해 멀티 벤더 전략을 추구하고 있으며, 이는 TSMC 외 파운드리의 패키징 시장 점유율 확대 가능성을 시사한다.

CoWoS 공급 부족과 ASIC 시장 확대로 부상하는 인텔 EMIB

TSMC CoWoS 캐파의 증설 속도가 폭발적인 수요를 하회하고, AI 산업의 중심축이 학습에서 추론으로 이동함에 따라 인텔 EMIB가 패키징 시장의 강력한 차선책으로 부상하고 있다. 특히 **AI 가속기 시장이 GPU에서 효율 중심의 ASIC으로 재편되는 과정에서 EMIB의 비용 경쟁력은 핵심적인 선택 기준이 될 것이다.** 인텔은 CoWoS 대비 구조적 단순성과 경제적 우위를 바탕으로, 오버스펙을 지양하는 AI 추론 칩 및 네트워크 ASIC 영역에서 점유율을 빠르게 확대하며 패키징 시장 내 '실질적 2위' 입지를 굳힐 것으로 전망한다.

TSMC 차세대 기술

차세대 공정: N2, N2P

AI 워크로드의 확산은 반도체 공정 전반에 구조적인 변화를 요구하고 있다. 연산 성능과 전력 효율에 대한 요구 수준이 동시에 높아지면서, 선단 공정의 중요성은 과거 대비 더욱 확대되고 있다. 이러한 환경에서 TSMC는 차세대 공정 전환을 계획대로 진행하며, 기술 로드맵 측면에서 경쟁사 대비 한발 앞선 위치를 유지하고 있다.

TSMC의 2nm(N2) 공정은 2025년 4분기 양산에 돌입했으며, 이는 선단 공정 경쟁에서 중요한 분기점으로 평가된다. N2는 기존 세대 대비 성능과 전력 효율 측면에서 의미 있는 개선을 제공하는 공정으로, 고성능·고집적 칩에 대한 수요 증가에 대응하기 위한 핵심 노드로 자리 잡을 전망이다. 특히 AI, 데이터센터, 고성능 컴퓨팅(HPC) 등 차세대 수요처에서의 채택 가능성이 높게 점쳐지고 있다.

TSMC는 단일 노드에 그치지 않고, 2nm 공정의 확장 버전인 **N2P**를 통해 선단 공정의 활용 범위를 넓힐 계획이다. N2P는 N2 대비 성능과 전력 효율을 추가로 개선한 파생 공정으로, 2026년 하반기 양산이 예정돼 있다. 이를 통해 TSMC는 초기 양산 이후에도 선단 공정의 경쟁력을 단계적으로 끌어올리며, 고객사의 고성능 제품 로드맵에 유연하게 대응할 수 있는 구조를 구축하고 있다.

N2양산 돌입과 N2P · A16으로 이어지는 로드맵을 통한 TSMC의 기술 리더십 공고화

종합하면, N2 및 N2P로 이어지는 2nm 공정 로드맵은 TSMC가 선단 공정에서의 기술 리더십을 중장기적으로 유지하기 위한 핵심 축으로 판단된다. 첨단 패키징(CoWoS)과 선단 공정이 동시에 확장되는 구조 속에서, TSMC는 AI 반도체 수요 확대 국면에서도 안정적인 공급 능력과 높은 고객 의존도를 유지할 가능성이 높다.

차세대 패키징: SoIC 및 SoW시리즈가 만드는 시스템 수준 통합

차세대 공정 경쟁이 심화되는 가운데, 패키징 기술은 더 이상 전공정을 보완하는 후단 공정이 아니라 성능·전력·면적 효율을 좌우하는 핵심 요소로 부상하고 있다. 특히 AI와 HPC 환경에서는 칩 자체의 성능뿐 아니라 칩 간 연결 구조가 시스템 전체 성능을 결정하는 요인으로 작용하면서, 패키징 기술의 전략적 중요성이 빠르게 확대되고 있다.

이러한 흐름 속에서 **TSMC는 2022년부터 3D 적층 패키징 기술인 SoIC(System on Integrated Chips)를 양산 단계에 도입하며 시스템 수준 통합 역량을 강화하고 있다**. SoIC는 μ-bump 기반의 SoIC-P와 하이브리드 본딩 기반의 SoIC-X로 구분되며, **칩 간 연결 거리를 최소화함으로써 신호 지연과 전력 손실을 억제할 수 있다는 점에서 고성능 컴퓨팅 및 AI SoC 구현에 적합한 솔루션으로 평가된다**. 현재 SoIC-X는 AMD의 MI300 시리즈와 3D V-Cache CPU 등 고성능 제품군에 적용되며 실질적인 양산 경험이 축적되고 있고, 향후 애플의 M5 시리즈, 엔비디아의 Rubin 등 차세대 HPC·AI 플랫폼으로 적용 범위가 확대될 가능성이 높다.

TSMC는 여기서 한 단계 더 나아가, **패키징의 다음 세대로 SoW(System on Wafer) 시리즈를 준비하고 있다. 이 중 SoW-X는 웨이퍼 단위에서 다수의 로직 칩과 HBM 스택을 통합하는 구조로, 기존 CoWoS 대비 훨씬 높은 수준의 시스템 집적을 목표로 하고 있다**. 최대 수십 개의 HBM 스택과 대형 로직을 하나의 웨이퍼 상에서 통합할 수 있도록 설계된 **SoW-X**는, 전력 효율과 시스템 확장성이 중요한 차세대 AI 컴퓨팅 환경에서 차별화된 대안으로 평가된다. TSMC는 2027년 양산을 목표로 SoW-X를 준비 중이며, AI 스타트업과 대형 컴퓨팅 플랫폼 고객을 중심으로 수요를 창출할 수 있을 것으로 전망된다.

SoIC와 SoW로 이어지는 TSMC의 로드맵

한편 **SoW-P**는 모바일 및 엣지 AI 등 보다 폭넓은 응용처를 겨냥한 구조로 개발되고 있다. SoIC, CoWoS, SoW로 이어지는 TSMC의 패키징 로드맵은 단일 칩 제조를 넘어 시스템 수준에서의 통합(System-level integration)을 강화하려는 전략적 방향성을 명확히 보여준다. 이는 중장기적으로

TSMC가 AI 반도체 밸류체인 전반에서 기술적·구조적 차별화를 확대하는 핵심 요인으로 작용할 것으로 판단된다.

TSMC 증설 타임라인

2nm 및 1.4nm 선단 공정 캐파 확충

TSMC의 선단공정 증설은 Fab 25(타이중)가 2027년 말 리스크 프로덕션 후 2028년 하반기 1.4nm(A14) 양산을 개시하며, 애리조나 P2(Fab 21 Phase 2)는 2026년 3분기 장비 반입 후 2027년 하반기 3nm 양산을 시작하고 동시에 2nm 생산도 병행하여 초기 월 2만 장(20k wpm) 규모로 운영될 예정이다. 2nm 전체 생산능력은 2025년 말 월 4만 장에서 2026년 말 10만 장, 2028년 20만 장으로 확대되며, 이 중 30%가 미국 거점에 배치될 계획이다. 첨단패키징 증설은 AP8(타이난)이 2025년 하반기 가동 시작 후 2026년 현재 램프업 중이며, AP7(자이)은 Phase 2가 2026년 양산, Phase 1이 2027년 양산을 개시해 CoWoS 월간 생산능력을 2024년 3.5만 장에서 2026년 말 12~13만 장으로 약 3.5배 증가시킬 전망이다. AP7은 SoIC 용량을 2024년 4~5천 장에서 2026년 2만 장으로 4배 확대하며, CoPoS·WMCM 등 차세대 패키징 기술도 2028년까지 순차 투입할 계획이다

TSMC의 선단 공정 증설 전략은 기본적으로 대만에서 최선단 공정을 먼저 양산한 뒤, 해외 생산기지는 일정 시차를 두고 추격하는 구조로 설계돼 있다. 이는 선단 공정 초기 단계에서 요구되는 수율 안정화, 공정 학습 효과, 협력 생태계 밀집도를 고려한 전략으로, AI·HPC 수요가 집중되는 핵심 캐파는 대만에 우선 배치하는 방식이다.

대만에서는 3nm(N3) 공정이 이미 안정적인 양산 단계에 진입한 가운데, 2nm(N2) 공정이 2025년 4분기부터 HVM(대량 양산)에 진입하며 선단 로드맵이 계획대로 이행되고 있다. TSMC는 N2 공정을 신주(Hsinchu)와 가오슝(Kaohsiung) 양쪽에서 병행 램프업하고 있으며, 초기 수율 또한 양호한 수준으로 평가된다. 이는 AI 가속기(GPU·ASIC) 및 차세대 HPC 칩 수요에 대응하기 위한 핵심 캐파가 여전히 대만 본토에 집중되고 있음을 의미한다.

한편 미국 애리조나에서는 N4(4nm) 공정을 시작으로 선단 공정 캐파를 단계적으로 확대하는 전략이 추진되고 있다. 1공장은 N4 기준으로 양산에 돌입했으며, 2공장은 기존 계획보다 앞당겨진 일정으로 3nm급 공정 도입이 논의되고 있다. 최근 보도에 따르면 TSMC는 미국 정부의 정책적 압박과 보조금 협상 과정에서 최신 공정 도입 시점을 일부 앞당기는 방향으로 조정하고 있는 것으로 알려졌다. 다만 회사 측은 최신 공정의 해외 이전에는 최소 1년 이상의 안정화 기간이 필요하다는 점을 분명히 하며, 대만과 동일 시점의 최선단 양산 가능성에는 선을 긋고 있다.

특히 애리조나 2공장의 양산 시점이 2027년 하반기 수준으로 당겨질 가능성이 거론되지만, 이는 대만의 N2 양산 이후 일정 시차를 둔 도입에 해당한다. 더 나아가 3공장에서는 2nm 또는 A16급 공정 도입이 2030년 전후로 계획되고 있어, 중장기적으로 미국 내 선단 캐파 비중은 확대되겠으나 단기적인 공급 결정력은 제한적일 전망이다.

AP8 가동을 통한 CoWoS 병목 해소 및 AI 가속기 시장 독점력 강화

TSMC는 AI 가속기용 CoWoS를 중심으로 한 첨단 패키징 수요 급증에 대응하기 위해 대만 남부 이노룩스 LCD 공장을 인수·개조한 AP8 팹의 증설 및 램프업을 본격화하고 있으며, 이를 통해 2024년 대비 2026년 CoWoS 생산능력을 약 3배 수준으로 확대해 기존 AP6 중심의 병목을 완화하고 리드타임을 단축할 것으로 전망된다. AP8은 Brownfield 전략을 통한 준공 리드타임 단축, CoWoS와 향후 SoIC 등 3D 패키징까지 수용 가능한 인프라 확보를 강점으로 하며, NVIDIA·AMD 등 주요 하이엔드 고객의 AI 관련 CAPEX 집행을 TSMC로 견인함으로써 패키징까지 포함한 턴키(value chain 통합) 경쟁력을 강화할 것으로 판단된다. 초기에는 LCD 공장 개조에 따른 수율 안정화 및 전력·용수 등 인프라 리스크가 존재하나, 2025~2026년 가동률 상승과 함께 첨단 패키징 ASP 및 믹스 개선이 수익성에 우호적으로 작용할 것으로 보이며, 관련 후공정 장비·소재 공급망 전반에도 동반 수혜가 예상된다.

SoIC · WMCM등 차세대 기술 다변화와 애플 · 엔비디아 락인 효과 극대화

대만 자이(Chiayi)에 위치한 AP7 팹은 8단계(8 Phase)로 구성된 차세대 첨단 패키징 거점으로 개발하고 있으며, 2026년 1월 22일 현지 언론에 최초로 공개하는 등 가시성을 높이고 있다. AP7은 Phase 2가 2025년 하반기 장비 반입을 시작해 2026년 양산 개시를 목표로 하며, Phase 1은 고고학적 유적 발굴로 인한 지연으로 2026년 장비 반입 후 2027년 본격 양산에 돌입할 계획이다. 기술 로드맵 측면에서 **AP7은 Phase 1·3에서 SoIC(System on Integrated Chips) 3D 패키징 확대, Phase 2는 Apple 전용 WMCM(Wafer-Level Multi-Chip Module) 생산 거점, Phase 4 이후로는 차세대 CoPoS(Chip-on-Panel-on-Substrate) 대량생산을 2028년 말까지 목표로 하여 CoWoS를 넘어서는 이종집적(Heterogeneous Integration) 플랫폼을 구축**한다. AP7은 AP8(타이난)의 CoWoS 물량 집중 전략과 차별화하여 SoIC·WMCM·CoPoS 등 기술 다변화 및 Apple·NVIDIA 등 전략 고객 전용 라인 운영을 통해 ASP 프리미엄 및 공급망 락인(Lock-in) 효과를 극대화할 것으로 판단되며, 2026년 이후 TSMC 첨단 패키징 매출 믹스 개선의 핵심 동력으로 작용할 전망이다.

표 15. TSMC fab 현황

대만					
신주	Fab12A, B	신형	타오위안	AP3	후공정
	Fab3	구형	타이난	Fab14	신형
	Fab5	구형		Fab18	신형
	Fab8	구형		Fab6	레거시
	Fab2	구형		AP2	첨단 패키징
	AP1	후공정		AP8 (2026E)	첨단 패키징
타오중	Fab20 (2025)	2nm	가오슝	Fab22	2nm
	Fab15	신형	주난	AP6	첨단 패키징
	AP5	후공정	자이	AP7 (2027E)	첨단 패키징
	Fab25 (2028E)	2nm			
미국			**일본**		
워싱턴주 카마스	Wafer Tech	레거시	규슈 구마모토	JASM1	12/16nm
애리조나주 피닉스	P1 (2025 양산)	4nm			22/28nm
	P2 (2028E 양산)	2/3nm		JASM2 (2027E)	6-7nm
	P3 (2030E 양산)	〈=2nm			

자료: TSMC, KUVIC 리서치 1팀

TSMC 생산량 추정

전공정

TSMC의 3nm 공정 생산능력 확장은 시장의 기존 가정을 상회하는 속도로 진행되고 있다. 2025년 말 기준 3nm 월 생산능력은 이미 15만 장을 상회하며 당초 목표치를 조기 달성했으며, 2026년에도 증설 속도는 둔화되지 않고 있다. 이는 주로 18B 팹 내 P7~P8 단계의 본격 가동에 기인한다. TSMC의 표준 단계당 생산능력(월 2.5만 장)을 감안할 경우, 18B 팹의 P1~P8 단계가 모두 가동될 시 이론상 월 20만 장 수준의 캐파가 형성된다. 현재 장비 반입 및 셋업 진행 상황을 고려하면, **2026년 말 3nm 월 생산능력은 월 20만~22만 장 수준에 도달할 전망이다.**

Fab 15B의 생산 구조에서도 변화가 관측된다. 일부 5nm·3nm 후공정 장비가 공유되면서 타 공장에서 생산되던 물량이 Fab 15B로 이전되고 있으며, 동시에 6/7nm 공정 수요도 재차 증가하고 있다. 구체적인 최종 수요처는 다변화되어 있으나, 이 같은 수요 회복은 Fab 15B의 가동률을 끌어올리며 3nm 전환 여력을 확보하는 데 기여하고 있다.18B 팹의 후속 단계인 P9는 건설을 완료했으며, 2027년 1분기 장비 반입을 거쳐 2분기 말 양산 개시가 예상된다. 통상적인 램프업 속도를 감안할 경우, 해당 단계 역시 1년 내 월 2.5만 장 수준의 정상 캐파에 도달할 전망이다. 여기에 Fab 15B 전환 물량을 포함하면, **TSMC의 3nm 월 생산능력은 2026년 약 20만 장에서 2027년 25만 장 수준으로 확대될 것으로 예상된다.**

기존 팹의 공정 전환과 18B신규 단계 가동을 통한 3nm 생산 능력 확충
유연한 팹 설계로 선단 공정 수요 변동에 최적화된 생산 체계 구축

한편, P10~P12 단계 증설 계획도 유지되고 있으나, 본격 양산 시점은 2028년 이후로 지연될 가능성이 높다. 이는 인력 제약과 함께 2nm(N2) 공정 확장에 대한 우선순위가 반영된 결과로 판단된다. 다만 주목할 점은 해당 단계들이 3nm와 2nm 간 전환이 가능한 구조로 설계되고 있다는 점이다. 향후 2nm 수요가 예상보다 빠르게 확대될 경우, 일부 단계는 2nm 라인으로 전환될 수 있으며, 이는 TSMC 선단 공정 캐파 운영의 유연성을 높이는 옵션으로 작용할 전망이다.

2027년 **TSMC 7/6nm 생산능력**은 선단 공정 전환 압력에도 불구하고 **월 15만 장 수준에서 안정적인 보합세를 유지**할 것으로 전망된다. 3nm/2nm 등 초미세 공정으로의 투자 집중 기조 속에서도, 전장용 반도체 및 가전, Edge AI 등 비용 효율성을 중시하는 수요처들의 발주가 견조하게 이어지며 물리적인 Capa 감축 요인을 상쇄할 것으로 판단된다. 이는 7nm 공정이 성숙기에 진입함에 따라, 무리한 라인 전환보다는 기존 유휴 설비의 효율적 재배치를 통해 꾸준한 가동률을 확보하고 현금 흐름을 창출하는 역할을 유지할 것으로 예상한다.

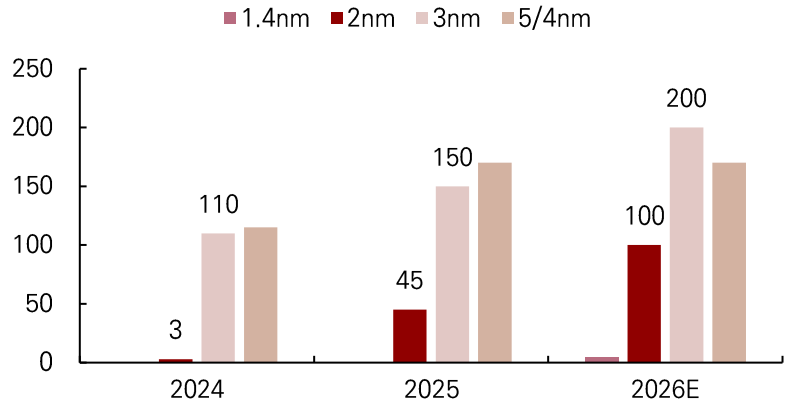
2027년 TSMC의 2nm(N2) 생산능력은 Fab 20와 Fab 22의 동시 다발적 램프업에 힘입어 월 10만 장에서 15만 장 수준으로 폭발적인 성장세를 보일 전망이다. Fab 20의 P1·P2 라인이 2026년 안정적 수율을 확보한 데 이어 2027년 P3·P4 라인이 가세하고, Fab 22 역시 P1~P3 라인이 풀가동 체제에 돌입하며 Apple, AMD, Intel 등 초기 고객사의 대규모 주문에 대응할 것으로 예상된다. 이러한 공격적인 증설은 HPC 시장의 패러다임이 GAA(Gate-All-Around) 기반의 2nm로 급격히 이동하고 있음을 시사하며, 2027년은 TSMC가 3nm에 이어 2nm에서도 압도적인 파운드리 주도권을 공고히 하는 해가 될 것이다.

그림 33. TSMC 라인별 생산량

(단위: kwpm)	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26	2Q26	3Q26	4Q26	2027E
1.4nm	0	0	0	0	0	0	0	0	0	0	0	5	5
2nm	0	0	3	3	5	25	40	45	60	70	75	100	150
3nm	70	80	100	110	110	115	120	150	160	170	180	200	250
5/4nm	135	135	125	115	125	140	150	170	170	170	170	170	170
7/6nm	150	150	150	150	150	150	15	150	150	150	150	150	150
16/12nm	160	160	160	160	160	160	160	160	165	165	165	165	165
20nm	5	5	5	5	5	5	5	5	5	5	5	5	5
28/22nm	240	240	245	245	240	240	245	250	260	260	260	260	260
45/40nm	175	175	175	175	175	175	175	175	175	175	175	175	175
65/55nm	125	125	125	125	125	125	125	125	125	125	125	125	125
90/80nm	73	73	73	73	73	73	73	73	73	73	73	73	73
0.1Xum	237	237	237	237	238	226	222	217	211	211	202	197	197
0.25um and above	60	60	60	60	60	60	60	60	60	60	60	60	60
Total	1,430	1,440	1,458	1,458	1,467	1,495	1,390	1,580	1,614	1,634	1,640	1,685	1,785

자료: 트렌드포스, 언론종합

그림 34. TSMC 선단 공정 생산량 전망



자료: KUVIC 리서치 1팀

후공정

TSMC의 첨단 패키징(CoWoS·SoIC) 생산능력은 AI 반도체 수요 확대에 대응해 중기적으로 가시적인 확장 국면에 진입했으며, 후공정 캐파는 2026년 이후 실적과 고객 구조를 설명하는 핵심 변수로 자리 잡고 있다.

공격적인 CoWoS Capex확장

2026년 초 기준, TSMC는 CoWoS를 중심으로 첨단 패키징 생산능력을 지속적으로 확대하고 있다. 글로벌 IB 추정에 따르면 TSMC의 **CoWoS 월 생산능력은 기존 시장 기대치였던 10만 장 수준을 상회해, 2026년 말 기준 12만~13만 장 수준까지 확대될 예정이다.** (26년 평균 kwpm은 25년 말 7만 장과 26년 말 kwpm 12.5만 장의 중간값인 9.75만 장으로 추정)

첨단 패키징 부문에 대한 자본 투입 역시 구조적으로 확대되고 있다. CoWoS, SoIC 등 후공정 관련 자본지출은 2025~2027년 기간 동안 두 자릿수 중반대의 CAGR이 예상된다. 이를 통해 첨단 패키징 생산능력이 단순한 보조 공정을 넘어, AI 반도체 출하 확대와 제품 믹스 고도화를 직접적으로 뒷받침하는 핵심 생산 요소임을 알 수 있다.

SoIC 적용 확대 및 AI · HPC 시장 내 독보적 지배력 강화

SoIC 역시 점진적인 캐파 확대 단계에 진입하고 있다. 기존에는 일부 고성능 제품을 중심으로 제한적으로 활용되던 SoIC가, 2026년 이후 차세대 AI 및 HPC 플랫폼으로 적용 범위가 확대될 가능성이 높아지면서 중기적으로 의미 있는 생산능력 확장이 예상된다. CoWoS 단일 축에 의존하던 후공정 구조는 다층화되고 있으며, TSMC의 첨단 패키징 포트폴리오가 양적·질적으로 확장되고 있다.

한편 TSMC의 전사 자본지출 확대는 이러한 후공정 캐파 증설을 구조적으로 뒷받침하고 있다. 2025년 자본지출은 사상 최대 수준을 기록한 데 이어, 2026년에도 높은 투자 기조가 유지될 가능성이 크다. 투자금의 상당 부분은 선단 공정과 함께 CoWoS·SoIC 등 첨단 패키징에 집중될 것이며, 주요 AI 고객사의 중장기 수요 가시성이 확보된 가운데 후공정 생산능력이 실적 레버리지로 작용할 수 있다.

2027년 TSMC의 CoWoS 생산능력(Capa)은 월 140,000장의 추정치가 제시된다. 이는 기존 Zhunan(AP6)의 안정적 가동 기반 위에, Chiayi(AP7) 팹의 신규 Phase 램프업이 본궤도에 오르고 Innolux 유휴 공장을 인수한 Tainan(AP8) 라인이 2026년 하반기 셋업을 거쳐 2027년 본격 양산 체제로 전환되는 데에 기인한다. (27년 평균 kwpm은 26년 말 12.5만 장과 27년 말 kwpm 14만 장의 중간값인 13.25만 장으로 추정)

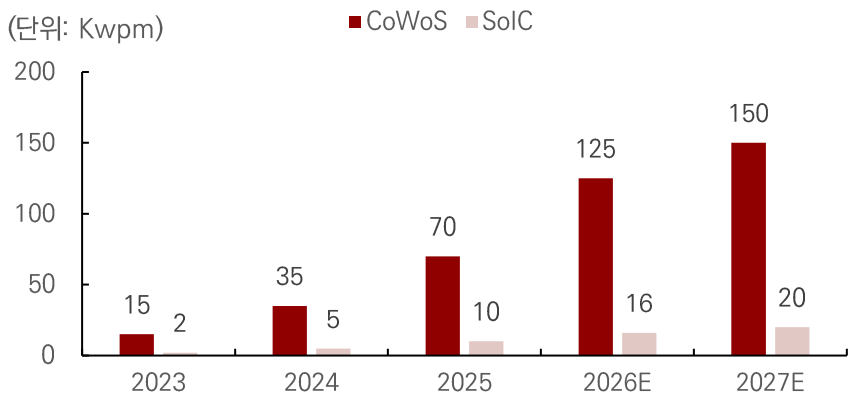
2027년 TSMC의 SoIC(System-on-Integrated-Chips) 생산능력(Capa)은 Chiayi(AP7) 팹의 Phase 1 라인 가동이 본격화됨에 따라 월 2만 장 수준까지 확대되며 틈새(Niche) 공정에서 주력 공정으로 확대될 예정이다. **2026년 말까지 AP6(NanKe) 증설을 통해 월 1.6만 장 수준을 확보**하고, **2027년 AP7 팹의 본격 가세를 통해 Capa를 두 배 수준인 2만 장**으로 증설하며 3D 패키징 시장에서의 기술 및 양산 주도권을 더욱 공고히 할 것으로 판단된다. TSMC의 첨단 패키징 Q는 2026년을 기점으로 구조적인 확장 국면에 진입했으며, CoWoS를 중심으로 SoIC까지 확대되는 후공정 캐파는 AI 반도체 공급 확대와 고객사 락인을 동시에 강화하는 요인으로 작용할 것이다.

본 보고서 추정에 따르면 26, 27년 말 기준 CoWoS 웨이퍼 캐파는 12.5만 장, 14만 장이며, 이를 GPU 규모로 환산하면 26년 1755만 장, 27년 2385만 장 규모로 추정된다.

앞서 산출한 GPU(AI 가속기) 생산 가능 물량을 기반으로 실제 필요한 HBM의 총 소요량을 추정하면 다음과 같다. 본 추정에서는 AI 가속기 1개당 탑재되는 HBM 슬롯(Slot) 수를 8개로 고정하였으며, 세대 교체(HBM3E → HBM4) 및 적층 단수 상향(8Hi → 12Hi)에 따른 믹스(Mix) 변화를 고려하여 스택(Stack)당 평균 메모리 용량을 30GB로 가정하였다. 이러한 전제를 대입할 때, CoWoS 캐파가 전량 가동된다고 가정할 경우 대응 가능한 연간 메모리 총 용량은 2026년 기준 42.1억 GB, 2027년 기준 57.2억 ~ 63.7억 GB 수준으로 추산된다. 이를 메모리 반도체 산업의 표준 단위인 기가비트(Gb)로 환산할 경우, 26년 약 337억 Gb, 27년 약 458억 Gb의 거대한 비트 그로스(Bit Growth)가 요구됨을 의미

한다. 결과적으로 파운드리단의 CoWoS 병목 해소는 곧 HBM의 구조적 수요 증가로 직결되며, 27년까지 연평균 40% 가량의 가파른 수요 확장이 지속될 것으로 판단된다.

그림 35. TSMC CoWoS/ SoIC capa 추이 및 전망



자료: TSMC, 언론종합

표 16. CoWoS 연간 생산량 및 요구 메모리 크기

	2026	2027	단위
월간 웨이퍼 생산량	9.75	13.25	만 장
연간 웨이퍼 생산량	117	159	만 장
웨이퍼당 생산 가능 칩 개수	15	15	개
생산 가능 칩 개수	1,755	2,385	만 대
HBM 스택 당 메모리 용량 평균	30	30	GB
HBM 슬롯 개수	8	8	개
대응 가능 메모리 용량	4,212	5,724	백 만 GB
대응 가능 메모리 용량	33,696	45,792	백 만 Gb

자료: KUVIC 리서치 1팀

메모리

HBM 및 범용 DRAM 공급 추정

연도별 생산량

표 17. 메모리 3사 DRAM 생산 구조 (단위: 백 만 Gb)

구분	전체 DRAM 생산 Gb			HBM 생산 Gb			HBM Gb 비중		
	2025	2026E	2027E	2025	2026E	2027E	2025	2026E	2027E
삼성전자	111,890	129,287	152,393	11,765	14,602	20,059	11%	11%	13%
SK하이닉스	90,618	102,003	143,283	14,803	17,626	22,421	16%	17%	16%
Micron	66,304	72,876	100,205	5,904	8,093	9,871	9%	11%	10%
합계/평균	268,812	304,166	395,880	32,472	40,320	52,351	12%	13%	13%

자료: OMDIA, KUVIC 리서치 1팀

26년에도, 27년에도 범용 DRAM은 강력한 Shortage

표 18. HBM 및 범용 DRAM 수요&공급 (단위: 백 만 Gb)

	2026E	2027E
수요	472,994	590,271
HBM	33,696	45,792
범용 DRAM	439,298	544,479
(데이터센터향)	266,474	356,013
(기타)	172,824	188,466
공급	304,166	395,880
HBM	40,320	52,351
범용 DRAM	263,846	343,529
Shortage		
HBM	6,624	6,559
범용 DRAM	-175,452	-200,950

자료: KUVIC 리서치 1팀

본 리서치 팀은 TSMC의 CoWoS 연간 생산량으로 통한 HBM 수요 추정과 메모리 3사 DRAM 생산량 추정을 토대로 **범용 DRAM에서 2026E 1,754억Gb, 2027E 2,009억Gb의 쇼티지**가 발생할 것으로 예측하며, 이는 **범용 DRAM 공급 물량의 2026E 66%, 2027E 58%**에 달하는 규모이다.

이를 토대로 **HBM에서 범용 DRAM으로 병목의 중심이 이동**하는 흐름을 확인할 수 있다. HBM의 경우 숫자 상으로는 66억 Gb 정도 초과 공급이 발생하나 이는 과잉 공급보다 균형 지속으로 해석해야 한다. 데이터센터향 HBM은 TSMC의 CoWoS 캐파가 정확한 수요 상한을 결정하기 때문이다. 성능 및 수율이 엔비디아 등 전방 고객 인증에 중요한 변수이기에, 용량 자체의 중요성은 크지 않다. 반면 범용 DRAM은 AI 추론 확산으로 서버 내 메모리 구성이 범용 DRAM 중심으로 두꺼워지고, 일반 서버의 대당 탑재량 상향까지 겹치며 수요가 구조적으로 상승한다는 것을 위 분석을 통해 여실히 확인할 수 있다. 또한 27년까지 쇼티지의 공급 물량 대비 비율도 크게 줄어들지 않아, **범용 DRAM에서의 쇼티지와 조달 불안은 27년까지 지속될 것으로 전망**한다.

수요 추정 논리

HBM 수요는 CoWoS 처리량을 출하 상한으로 두고 역산했다. 2026년 CoWoS 연간 웨이퍼 117만장, 웨이퍼당 칩 15개, 칩당 HBM 슬롯 8개, 스택당 평균 30GB(8단 24GB와 12단 36GB를 1:1 평균) 가정을 적용하면 데이터센터향 HBM 수요는 약 337억Gb로 계산된다. 2027년은 연간 웨이퍼 159만장으로 상향하고 동일 가정을 적용해 약 458억Gb로 산출된다.

범용 DRAM 수요는 데이터센터 내부 구성 변화가 상방 요인이다. AI 서버의 메모리 구성은 **2026년 HBM:범용 DRAM 1:1에서 2027년 1:2로 이동**하는 것으로 둔다. **추론 비중 확대와 롱컨텍스트 운용 확산으로 KV 캐시 부담이 커지며 범용 메모리의 필요 용량이 상대적으로 더 빠르게 늘어난다**는 판단이다. 일반 서버도 차세대 CPU의 메모리 채널 확장 효과로 대당 탑재량이 2.2TB에서 10% 성장률을 곱한 2.42TB로 상향 평준화되는 것으로 둔다. 반면 스마트폰, PC 등 비데이터센터 수요는 플랫을 가정한다. 이러한 가정으로 2026년 4729억Gb, 2027년 5902억Gb로 산출된다.

공급 추정 논리

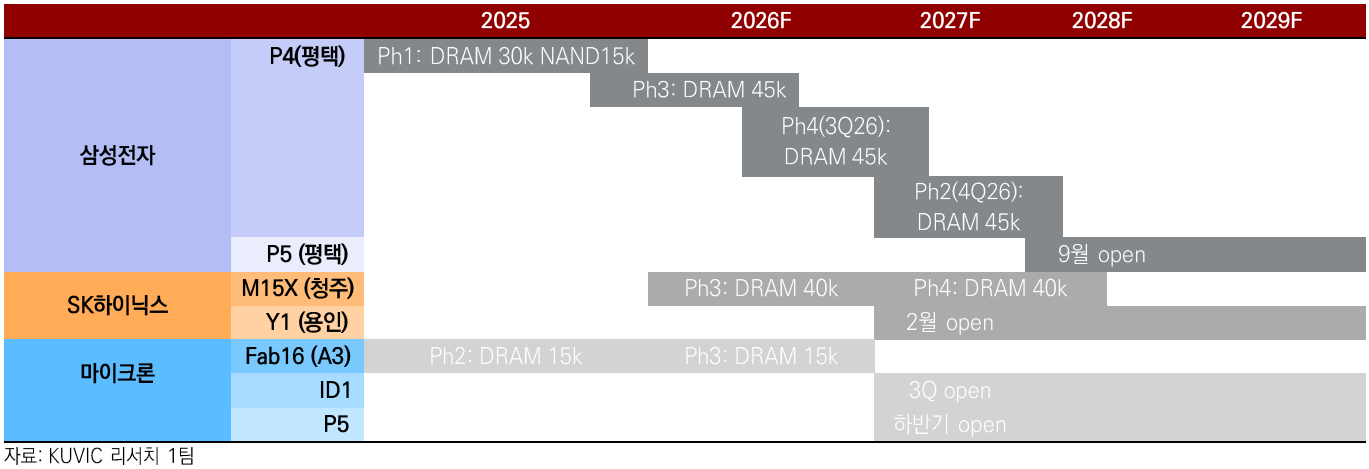
메모리 공급은 **2026년까지 OMDIA 데이터로 기준 수준을 고정**하고, **2027년 이후는 증설 일정 반영과 연평균 성장률 적용**을 바탕으로 믹스·수율·가격·환율 가정을 결합해 확장했다. HBM은 2025년 HBM3·HBM3E·HBM4 병존 이후 2026년부터 HBM3가 종료되고 HBM3E·HBM4 중심으로 전환되며, 2027년에는 HBM4E가 추가되며 세대 구성이 재편되는 흐름을 반영했다.

추정 결과 HBM 공급은 2026년 403억Gb, 2027년 523억Gb, 범용 DRAM 공급은 2026년 2,638억Gb, 2027년 3,435억Gb로 산출된다.

DRAM 증설 타임라인과 공급 병목 구조

메모리 3사 증설 타임라인

표 19. 메모리 3사 증설 타임라인



자료: KUVIC 리서치 1팀

2026년: 제한적 순증, 전환·램프업 중심

2026년 DRAM 공급은 전환과 램프업 속도가 기여 시점을 좌우

2026년 삼성전자·SK하이닉스·마이크론 DRAM 공급의 타임라인은 **"대규모 웨이퍼 스타트 확대"보다는 전환과 램프업의 진행 속도에 따라 공급 기여가 단계적으로 반영**되는 구조로 볼 수 있다. DRAM 기준 2026년 웨이퍼 출하는 2025년 대비 +40K/월 증가로 제시되는데, 2025년에 이미 +219K/월 확대가 반영된 이후라는 점을 감안하면 **2026년은 웨이퍼 순증 폭 자체가 제한적**이며, 대신 **공정 미세화(1c) 확**

대를 통해 비트 성장률을 확보하는 방향이 중심으로 된다.

삼성은 2026년에 평택(P4) 증량 일정이 잡혀 있으나, 동시에 **일부 라인에서 공정 전환에 따른 웨이퍼 감소가 동반되어 순증이 상쇄**되는 형태로 보인다(평택 P4의 증량이 존재해도 타 라인의 감소가 함께 나타나는 구조). 이 과정에서 감소는 수요 둔화가 아니라 공정 전환 영향으로 설명되며, 2026년의 공급 변화는 웨이퍼 절대량보다 1c 웨이퍼 확대 등 전환 진척에 의해 설명되는 비중이 커진다. 또한 평택 P4 증량은 2026년 하반기에 걸쳐 분기 단위로 반영되는 형태로 제시되어, 연중 공급 기여가 특정 시점에 집중되기보다 분산되는 스케줄로 이해하는 것이 자연스럽다.

SK하이닉스는 2026년 웨이퍼 순증이 +10K/월로 예측되며, 내부적으로는 M15X 증가(+37K/월)가 존재해도 **다른 거점 감소가 동반되어 순증이 제한**되는 형태로 보인다. 최근에는 청주 M15X가 2026년 2월부터 웨이퍼 투입을 시작하는 일정이 확인되며, 2026년 공급 기여는 이 웨이퍼 투입 개시 이후 램프업 속도에 의해 결정될 것으로 확인된다.

마이크론은 2026년 웨이퍼 물량보다 1c 전환을 통해 웨이퍼당 산출 비트를 높이는 전략

마이크론은 2026년 DRAM 웨이퍼가 2025년과 동일한 수준(300K/월 내외)으로 제시된다. 그 이유로는 공급 확대의 초점은 **웨이퍼 증설이 아니라 공정 전환을 통한 비트 성장**에 맞춰져 있기 때문이다. 구체적으로 **1c 비중을 2025년 말 8%에서 2026년 말 38%로 확대**하는 경로가 강조되는데, 이는 신규 라인 증설로 웨이퍼 스타트를 늘리는 방식보다 기존 라인의 미세화 전환을 통해 단위 웨이퍼당 산출 비트를 끌어올리는 전략에 가깝다. 따라서 2026년 공급 변화는 웨이퍼 절대량 증가보다, 1c 전환 진척에 따른 비트 기준 공급 증가로 먼저 반영되는 구간으로 정리된다.

2027년: 신규거점 가동, 공간·후공정 반영

2027년은 전공정, 후공정, 공간 확보 일정이 실제 생산에 연결되며 공급 기여 시점이 구체화

2027년은 **전공정·후공정 및 공간 확보 관련 일정이 공급 기여로 연결되는 시점들이 구체화**되는 해다. SK하이닉스는 용인 클러스터 첫 공장의 가동 시점을 2027년 2월로 제시하고 있으며, 이는 2027년 상반기에 전공정 측면의 물리적 캐파가 추가되는 일정으로 보인다. 다만 HBM은 전공정 웨이퍼뿐 아니라 후공정 수용력이 함께 제약이 될 수 있으므로, 하이닉스가 발표한 첨단 패키징 공장의 2026년 4월 착공–2027년 말 완공 일정은 **2027년 하반기(연말) 시점이 되어야 후공정 병목 완화**가 반영될 수 있음을 시사한다.

마이크론은 2026년 1월에 대만 Tongluo P5 사이트 인수를 통해 300,000제곱 ft 규모의 300mm 클린룸을 확보하는 계획을 공개했고, 이 인수의 DRAM 생산 기여 시점을 2027년 하반기로 설명했다. 또한 아이다호 신규 팹(ID1)에서 DRAM 출력이 2027년 하반기에 시작될 가능성이 높아 미국 신규 거점도 2027년 DRAM 생산에 기여할 것이다.

삼성은 2026년 하반기에 평택 P4 증량이 분기 단위로 반영되는 일정이 제시되고, 더 큰 물리적 증설 이벤트는 평택 P5 2028년 9월 오픈으로 명시되어 있어, 2027년은 **전환·램프업 기조의 연장선**으로 해석하는 쪽이 타당할 것으로 보인다.

종합하면, **2026년은 웨이퍼 순증이 제한된 가운데 전환·램프업이 공급 변화를 설명**하는 비중이 커지고, **2027년은 상반기(전공정 신규 거점 가동)와 하반기(공간 확보 자산의 생산 기여, 후공정 완공 등)로 나뉘어 공급 기여 시점이 분기**되는 형태로 구성할 수 있다.

"증설=공급 증가"가 아닌 구조: 공정 전환이 만드는 비트 공급 공백

CAPEX가 늘어도 미세공정 전환은 단기 생산·수율 공백을 키우기 쉬움

메모리 산업에서 CAPEX가 늘었다고 해서 곧바로 비트 공급이 같은 속도로 늘어나지는 않는다. 특히 최근 구간은 신규 팹을 크게 짓는 방식보다, **기존 라인을 더 미세한 공정으로 전환하는 방식이 중심**이기 때문이다. 전환은 장기적으로 비트밀도를 올리는 방법이지만, 단기에는 생산능력을 오히려 갉아먹는다. 동일한 클린룸 면적 안에서 장비 배치가 바뀌고 공정 레시피가 재정렬되며, 전환 기간 동안은 가동률이

내려가거나 수율 램프업 구간이 길어져 **실생산이 줄어드는 공백**이 발생한다. 실제로 미세화가 진행될수록 공정 난이도가 올라가 수율을 안정화시키는 데 시간이 더 걸리고, 이 때문에 **주력 노드가 빠르게 새 노드로 완전히 대체되지 못한 채 여러 세대가 병존하는 시간이 길어진다.**

HBM 중심 믹스와 제조·검증·패키징 제약으로 유효 비트가 투자만큼 늘기 어려움

여기에 최근 증설 믹스가 범용 DRAM의 단순 증설이 아니라, **HBM 중심으로 재편**되면서 비트 공급의 체감 증가폭을 더 낮춘다. 같은 웨이퍼라도 어떤 제품을 뽑느냐에 따라 시장에 풀리는 '유효 비트'가 달라지는데, 고부가 제품은 **제조·검증·패키징 제약** 때문에 라인 투자 규모 대비 출하 비트가 기대만큼 따라오지 못하는 구간이 생긴다. 결과적으로 CAPEX가 증가하는 것처럼 보여도, 그 CAPEX의 상당 부분이 **순수한 웨이퍼 스타트 증가가 아니라 전환 비용·난이도 상승에 따른 장비/공정 투자로 흡수**되면, 공급 곡선은 생각보다 완만해진다.

또 하나 중요한 점은, 기업이 전환 속도를 마음대로 끝까지 올리기 어렵다는 것이다. 시장에는 여전히 **긴 수명주기 제품(레거시/특정 고객용) 수요**가 남아 있고, 특정 노드는 단기간에 완전히 접기 어렵다. 결국 가장 앞선 최신 노드로 전부 갈아타면 끝이 아니라, **여러 노드를 병행 운영하면서 전환을 끌고 가야 하는 구조**가 되고, 이는 전환기에 공급이 공격적으로 늘기 어렵게 만든다.

다운사이클 경험으로 선제 증설보다 수익성·현금흐름 우선해 공급 램프가 완만해짐

여기에 더해, 설령 기술적으로 증설이 가능하더라도 기업 입장에서 CAPEX를 공격적으로 당기기 어려운 행동 요인이 존재한다. 직전 다운사이클에서 **과잉 공급과 재고 조정이 수익성에 큰 타격**을 줬던 경험이 남아 있고, 그 이후로는 **수요가 확인되기 전에 먼저 깔아두는 선제 증설보다 수익성·현금흐름을 우선하는 투자 규율**이 강화되는 경향이 있다. 결국 2025~2027년 구간은 "필요하면 언제든 빨리 늘린다"는 구간이 아니라, 수요가 강해도 전환·검증·공간 제약을 감안해 단계적으로 램프를 올리고, 동시에 **과잉투자 리스크를 피하려는 태도**가 겹치면서 공급이 더 완만하게 늘어날 가능성이 커진다.

물리적 제약의 삼중 병목: 공정 난이도·장비 리드타임·클린룸 부족

공정 전환이 가능하려면 **장비 수급, 클린룸 공간, 공정 안정화**가 동시에 충족돼야 한다. 그런데 최근 메모리는 **이 세 가지가 동시에 제약**이 되는 구간이다.

첫째, 공정 난이도 자체가 올라가면서 **장비 집약도가 커진다**. 미세화는 동일 면적에서 더 많은 비트를 생산하게 해주지만, 그 대가로 **공정 스텝이 복잡해지고 공정 창(window)이 좁아진다**. 결과적으로 수율 안정화가 어려워지고, 장비를 투입해도 곧바로 비트가 늘지 않는 구간이 길어진다.

둘째, 장비 리드타임과 설치 속도는 생각보다 느리다. EUV 같은 핵심 장비는 **발주→제조→납품→설치→인증**까지 시간이 길고, 업계에서도 **장비 납기 자체가 수개월 단위**로 잡힌다는 언급이 반복된다. 팹은 장비만 놓는다고 끝이 아니라 유틸리티/전력/배기/케미컬/진동 등 인프라를 맞춰야 하고, 이 과정이 병목이 되면 CAPEX가 집행돼도 완공된 캐파로 전환되는 속도가 제한된다. 실제로 장비 납기와 공장 건설 자체가 오래 걸린다는 점은 글로벌 메이저 장비사/언론 보도에서도 일관되게 확인된다.

클린룸 유휴공간이 부족하면 장비를 사도 못 깔아 신규 팹보다 기존 라인 전환이 중심

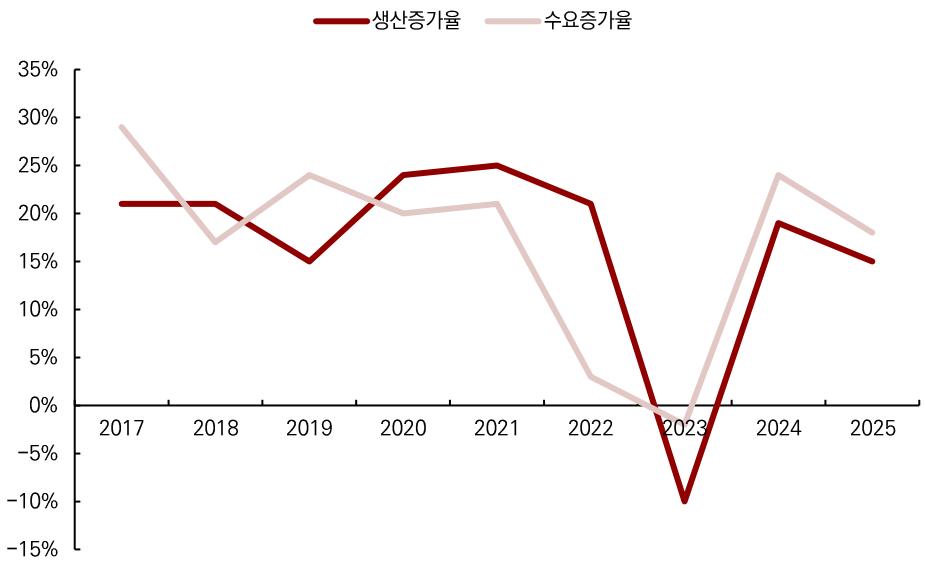
셋째, 가장 단단한 제약은 클린룸 면적(유휴공간)이다. 클린룸이 부족하면 장비를 더 사서 깔아 증설하는 방식 자체가 불가능해지고, 전환도 결국 기존 공간을 비워가며 진행해야 한다. 이때 기업의 선택지는 **(1) 신규 팹을 짓거나, (2) 기존 팹을 리모델링/전환하거나**인데, 전자는 시간과 비용이 너무 크고 수요 불확실성도 있어 쉽게 선택할 수 없다. 그래서 현실적으로는 **기존 라인의 전환이 중심**이 되며, 이 구조가 공급 증가 속도를 더 저하시킨다. 실제로 클린룸 쇼티지가 중장기까지 이어진다는 판단 아래, 신규 팹보다는 **1c/1d 중심의 전환**이 강조되는 흐름이 진행중이다.

이 **클린룸 병목**은 개별 사례에서도 확인된다. 예컨대 마이크론은 대만에서 DRAM 생산능력을 늘리기 위해 기존 팹을 현금으로 인수해 클린룸 면적을 확보했고, 그 생산 가동 시점을 **2027년 하반기**로 잡았다. 즉 수요가 타이트하다고 해서 공급이 즉시 따라오는 게 아니라, **공간 확보→설비 반입→가동**이 최소 연 단위로 지연되는 구조가 실제로 드러난다.

그림 36. 삼성전자 주가추이 (2018~2025)

자료: KUVIC 리서치 1팀

그림 37. DRAM 수요공급 증감률 추이



자료: OMDIA, WSTS, KUVIC 리서치 1팀

25~27년 공급은 전환·수율·공간·장비 제약 속 완만 증가 구조라 타이트 수급 장기화 가능성 높음

공급 측면에서 보면, 2025~2027 구간의 메모리 공급은 전환(시간)·난이도(수율)·공간(클린룸)·장비(리드타임) 제약이 동시에 걸려 **증가 속도가 구조적으로 제한**될 가능성이 크다. 이때 메모리 사이클이 반복되는 핵심은 수요의 변동 자체보다 **수요 증가율과 공급 증가율의 상대속도가 엇갈리며 교차하는 지점에서 변곡점이 형성**된다는 데 있다. 통상 업사이클은 수요 증가율이 공급 증가율을 상회해 재고가 빠르게 줄어드는 구간에서 시작되고, 이후 CAPEX·라인 전환·증설이 시차를 두고 반영되며 공급 증가율이 수요를 상회하면 재고가 누적되면서 다운사이클로 전환된다. 따라서 이번 국면도 사이클의 매커니즘을 따르지만 **공급 반응의 기울기 자체가 완만**해질 수 있어, 수요가 유지되는 한 수요>공급 구간이 과거보다 길어지며 타이트 수급이 장기화될 소지가 크다.

HBM: 병목의 중심, 수급의 중심

HBM은 무엇이고 왜 필요한가

HBM은 가속기 옆에서 대역폭 병목을 풀어 실제 처리량과 시스템 효율을 끌어올림

HBM(High Bandwidth Memory)은 AI 가속기(GPU/ASIC) 주변에서 후술할 **TSV(Through Silicon Via) 기술을 적용시켜 초고대역폭을 제공하도록 설계된 DRAM**이다. 범용 DRAM이 오랫동안 용량 확대와 단가 하락을 중심으로 발전해 왔다면, HBM은 출발점부터 **대역폭 병목을 해소**해 가속기의 처리 성능을 안정적으로 끌어올리는 데 초점이 맞춰져 있다. 그래서 평가 기준도 다르게 잡힌다. 범용 DRAM은 용량(GB), 단가, 전력 효율, 범용 호환성이 중심인 반면, HBM은 대역폭(GB/s), 데이터 공급 효율, 가속기 성능 기여도, 시스템 레벨 효율이 핵심이다. 같은 DRAM이라도 용량보다 **연산 유닛에 대한 데이터 공급 능력**이 먼저 평가되는 구조다.

병목이 연산에서 데이터 공급으로 옮겨가며 HBM이 처리량을 좌우

HBM을 단순히 "빠른 메모리"로 정의하면 핵심을 놓친다. HBM의 포지션은 **가속기 근접(near-memory)** 계층을 구성하는 사실상의 표준 부품이다. AI 워크로드는 가중치·활성값·중간 텐서·캐시 데이터가 계속 오가며, **연산 유닛이 데이터를 기다리는 시간이 성능을 직접 잠식**한다. 가속기 연산 성능이 높아질수록 병목은 계산이 아니라 공급에서 먼저 드러나고, 이때 근접 대역폭이 부족하면 가속기 활용률이 떨어지며 성능이 스펙 대비 크게 후퇴한다. HBM은 가속기 옆에서 대역폭을 크게 확보해 이런 대기 시간을 줄이고, 실제 처리량을 끌어올리는 역할을 한다. 결과적으로 HBM은 메모리 시장에서 하나의 제품군을 넘어, AI 하드웨어 스택의 성능과 효율을 좌우하는 핵심 축으로 자리 잡는다.

HBM의 부상은 AI 데이터센터에서 병목이 연산 능력 자체보다 **메모리 대역폭과 데이터 이동 비용**으로 자주 나타나는 현실과 맞물린다. 대역폭이 부족하면 처리량이 줄어드는 데서 끝나지 않고, 동일 목표 처리량을 맞추기 위해 서버 대수가 늘면서 **전력·냉각·공간·네트워크 비용이 함께 증가**한다. 반대로 근접 대역폭이 충분하면 가속기는 더 높은 활용률로 돌아가고, 같은 작업을 더 짧은 시간에 끝내며, 데이터센터 관점에서 성능/와트와 성능/랙이 개선된다. 특히 학습에서 추론으로 무게중심이 이동할수록, 동시 요청 처리와 긴 컨텍스트로 인해 메모리 트래픽과 상주 부담이 커지면서 대역폭 부족이 지연·처리량 저하·서버 증설로 직결된다. 이 구조가 HBM을 선택 옵션이 아니라 **시스템 효율을 규정하는 필수 요소**로 만든다.

HBM을 둘러싼 구조와 수급

AI 서버는 HBM–DDR–스토리지 계층으로 최적화

AI 서버 메모리 설계와 HBM 수급은 결국 **계층 구조에서의 역할**과 **공급 제약이 만드는 시장 메커니즘**이 맞물려 결정된다. 즉 HBM은 DDR·스토리지와 결합된 설계 안에서 성능·지연 특성을 규정하는 한편, 후공정·플랫폼 검증 리드타임 때문에 공급이 탄력적으로 늘기 어려워 조달 방식과 가격 구조까지 함께 흔든다. 이러한 연결고리를 기준으로, AI 서버 메모리 설계와 HBM 수급을 관통하는 핵심 논점을 네 가지로 정리해 설명할 수 있다.

첫째, AI 서버의 메모리 설계는 HBM만으로 완결되지 않는다. 현실적으로는 HBM(가속기 근접 고대역폭)–DDR(호스트 메인메모리)–스토리지(데이터 보관/캐시)로 이어지는 **메모리 계층 구조**를 전제로 최적화된다. HBM은 가속기 바로 옆에서 대역폭을 제공해 연산 유닛이 데이터 부족으로 멈추지 않게 하고, DDR은 CPU 및 시스템 영역에서 큰 용량 기반을 담당하며, 스토리지는 데이터셋·체크포인트·로그·캐시를 뒷받침한다. 즉 서버는 대역폭(근접)과 용량(호스트)과 저장(스토리지)을 분업시키고, 워크로드 성격에 맞춰 이 조합을 조정하면서 목표 성능과 비용을 맞춘다.

추론에선 지연·동시성 때문에 HBM 제약이 설계 제약이 됨

둘째, 이 계층 구조에서 HBM의 의미는 "최고 성능"만이 아니다. 실제 데이터센터 운영에서는 평균 처리량뿐 아니라 지연 분포, 동시성 하 성능 안정성, 워크로드 변동성 대응이 중요하다. 근접 대역폭이 부족하면 특정 구간에서 지연이 급증하고, 이는 서비스 품질 저하로 이어진다. 반대로 HBM이 충분하면 가속기 활용률이 안정화되고, 요청 처리의 변동성이 완화되어 예측 가능한 성능을 만들 수 있다. 또한 추론

국면에서 KV cache 등 구조적 메모리 수요가 커지면, HBM을 정점에 두되 DDR/스토리지와의 조합으로 상주·이동을 조정하는 설계가 강화된다. 중요한 것은 이 과정에서도 HBM이 계층의 최상단에서 **가속기 성능을 규정한다**는 점이다. HBM의 대역폭/용량 제약은 곧바로 배치 전략, 캐시 정책, 병렬화 방식 같은 시스템 소프트웨어의 설계 제약으로 내려온다.

셋째, HBM 시장의 또 다른 핵심은 구조적 쇼티지(공급 타이트)다. HBM은 수요가 늘어도 단기간에 공급이 탄력적으로 따라붙기 어렵다. 고성능 제품 특성상 공정 난이도와 품질 관리 부담이 크고, 출하 관점에서 명목 생산량보다 스펙을 만족하는 **유효 물량**이 중요해지기 때문이다. 특히 데이터센터급 제품은 신뢰성 요구가 높고 장시간 안정 동작을 전제로 하기 때문에 품질 기준이 더 엄격해진다. 게다가 공급 확대는 장비 투입만으로 해결되지 않고 양산 안정화와 고객 플랫폼 검증(qualification)을 동반해 리드타임이 길어진다. 단순히 만드는 것이 아니라 플랫폼에서 문제없이 돌아가는 조합으로 확정되어야 하므로 전환과 램프업이 느려지기 쉽다.

타이트 수급은 선행 주문을 부르고, 프리미엄이 길어짐

넷째, 수요 측면에서도 타이트 수급은 강화될 수 있다. 가속기 수요가 급증하는 구간에서는 "필요할 때 못 구할 위험"을 줄이기 위해 조달을 앞당기거나, 안전 재고 성격의 주문이 늘어나면서 실수요 대비 주문이 선행되는 현상이 나타날 수 있다. 이때 수급은 더 타이트해지고, 타이트 수급 인식이 다시 조달 행태를 강화하는 순환이 생긴다. 이런 구조는 가격에도 반영되어, HBM은 수요 강세 국면에서 단기 급등보다 **프리미엄이 오래 유지**되는 형태로 나타날 가능성이 크다. 더 나아가 HBM 타이트 수급은 시스템의 대역폭–용량–전력–폼팩터 균형 재설계를 촉발해 DDR5·GDDR·LPDDR 기반 구성 등 **범용 메모리 채택 논리로 전이**되며, 메모리 사이클을 "수요 변동"뿐 아니라 **"공급 제약"** 중심으로 재해석하게 만든다.

HBM의 현 단계: SK하이닉스 우위 국면

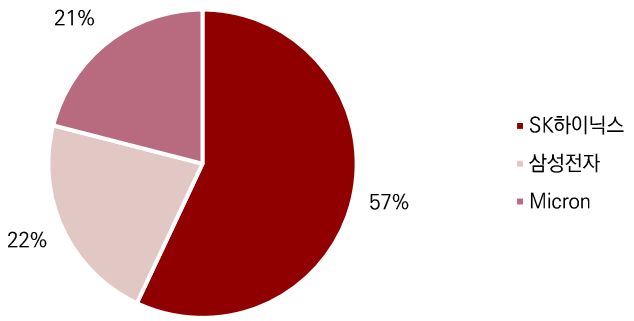
HBM 구조적 타이트 수급·우위 지속성

앞에서 언급했듯이 HBM은 가속기 옆에서 대역폭 병목을 풀어 실제 처리량과 시스템 효율을 끌어올리는 핵심 부품이고, 동시에 **플랫폼 인증과 후공정 제약 때문에 수급이 구조적으로 타이트하게 형성**된다. 이 시장에서 공급사 우위는 신제품 발표나 단기 가격 경쟁으로 쉽게 바뀌지 않는다. 고객 플랫폼에서 장시간 안정 동작을 전제로 검증을 통과해야 하고, 검증이 끝난 조합이 곧바로 조달과 양산, 출하로 연결되기 때문이다. 따라서 **이미 형성된 공급 포지션이 그대로 연장**되는 성격이 강하다.

2025년 점유율 기준 SK하이닉스 우위

HBM에서의 우위는 점유율에서 먼저 확인된다. **2025년 3분기 기준으로 제시된 HBM 시장점유율은 SK하이닉스 57%, 삼성 22%, 마이크론 21% 수준**으로, SK하이닉스가 우위를 점하고 있다.

그림 38. 글로벌 HBM 시장 점유율



자료: Counterpoint Research, KUVIC 리서치 1팀

SK하이닉스의 1위는 HBM3E에서 가장 먼저 양산·출하 레퍼런스를 구축하며 **고객 조달의 기준점을 선점**한 데서 출발한다. 이어 HBM의 승부처인 후공정(적층·본딩·TSV) 수율과 장시간 안정 동작 신뢰성에

서 삼성전자와 마이크론 대비 우위를 확보해, 실제 출하 가능한 물량을 가장 크게 만들었다. 그 결과 고객은 검증된 조합을 중심으로 물량을 장기 배정하게 되었고, 이 배정 구조가 **2025년 점유율 우위를 그대로 고착화**시켰다.

이 점유율 우위가 2026년에도 직접적인 영향을 미칠 가능성이 높다. 이유는 조달 방식 때문이다. **HBM은 단년도 스팟 조달보다 다년 공급 계약과 장기 배정이 선호**되는 방향으로 움직였다. 고객은 가속기 출하 일정과 데이터센터 증설 계획에 맞춰 메모리 조달을 선제적으로 고정하려 하고, 공급사는 검증된 조합을 중심으로 물량을 배분한다. 그 결과 기존에 구축된 공급 이력과 인증 레퍼런스가 다음 해 물량 배분의 출발점이 된다. 따라서 이러한 과거의 흐름 상 **2026년 경쟁 구도는 단기간에는 급변하지 않을 가능성**이 크다.

HBM4·HBM4E에서 3사 격차 축소

HBM4·HBM4E 전환으로 격차 축소 가능성

그럼에도, **HBM4·HBM4E 전환은 격차가 줄어들 수 있는 이벤트**이다. 세대 전환기에는 고객 요구 사양이 상향되면서 설계와 검증이 사실상 리셋되고, 기존 레퍼런스의 영향력이 상대적으로 약해진다. 특히 HBM4부터는 **단순 대역폭 경쟁을 넘어 전력·열·신호 무결성, 커스텀 로직/베이스 다이 최적화, 패키징 병목 해소 능력이 동시에 성패를 가르는 구조**로 이동한다. 이 구조에서는 선두가 단독으로 치고 나가기보다, 후발주자가 동일 조건의 재검증 레이스에서 격차를 좁힐 여지가 커진다. 결과적으로 **HBM4·HBM4E 전환은 SK하이닉스의 독주를 제어하고 3사 경쟁 구도를 재정렬할 변곡점**으로 작동할 수 있다.

HBM4 전환은 동시 재검증이 먼저 작동하는 국면이다. Rubin 플랫폼 기준으로 HBM4 요구 사양이 상향되면서 per-pin 속도 요구치가 11Gbps를 상회하는 수준으로 올라갔고, 그 결과 일부는 설계조정이 불가피해졌다. 이 이벤트는 특정 1개사의 단독 질주를 막고, 동일 조건에서 샘플 재제출과 재검증을 반복하는 구조를 만든다. 동시에 **HBM4 양산 램프업이 2026년 1분기 말 이후로 후행하는 전망**이 제시된다. 일정이 밀릴수록 승부처는 발표 시점이 아니라 인증 통과 속도, 수율 안정화 속도, 후공정 병목 해소로 이동하며, 이 이동이 격차 축소를 이끈다.

삼성전자: 커스텀/로직다이·파운드리 결합으로 추격 경로 열림

이 부분에서 삼성전자는 격차 축소 변수로 들어온다. 삼성전자는 **엔비디아의 HBM4 최종 품질 검증을 통과**했고, 2월부터 엔비디아·AMD에 양산 제품을 **업계 최초로 정식 납품**한다. 고객 요구치(초당 10Gb)를 상회하는 초당 11.7Gb 동작속도를 구현했고, 대역폭은 초당 2.8TB 수준으로 올라섰다. 또한 전력효율도 HBM3E 대비 40% 이상 개선되며, 고속·고대역폭·저전력의 3요소를 동시에 끌어올렸다. 이런 성능을 낼 수 있었던 배경은 명확하다.

첫째, 삼성은 HBM4의 기본 재료인 D램을 1c로 끌어올렸다. 둘째, 로직/베이스 다이에 4nm 파운드리 공정을 선제 적용해 성능 헤드룸을 확보해 둔 상태였다. 그 결과 재설계 없이 검증을 통과했다. HBM4 이후 커스텀 HBM의 비중이 커지며 로직/베이스 다이의 역할이 확대되는 전환점에서, 삼성은 로직 다이 공정을 4nm에서 2nm급까지 확장하는 방향 또한 검토하고 있다. 메모리 단독 성능 경쟁에서 **메모리+로직+공정(파운드리) 결합 최적화**로 이동하고, 이 전환이 본격화되는 순간 삼성의 추격 경로가 구조적으로 열린다.

HBM4E에서는 **타임라인 수렴이 더 직접적**이다. 커스텀 HBM4E 설계 완료 목표가 2026년 5~6월로 제시되고, SK하이닉스와 마이크론도 유사한 타임라인일 것으로 예측된다. 설계 시기가 수렴하면 2026년 말~2027년 경쟁은 **인증, 수율, 패키징, 고객별 커스텀 대응**으로 정리된다. 이 네 가지는 한 기업의 독주보다 수렴을 강화하는 항목들이다. 다만 최근 삼성전자가 엔비디아 품질 검증을 가장 빠르게 통과하며 '검증 레이스'에서 앞선 이력이 확인된 만큼, 이 네 가지 경쟁축이 본격화될수록 삼성전자 쪽에 우세가 실릴 가능성이 높다.

HBM만이 아니다 범용DRAM과 NAND가 같이 간다

DDR4/DDR5: 서버·PC 메인메모리의 기초 수요

DDR은 서버·PC의 기본 메인메모리라 일반 서버 증설로 수요가 깔림

범용 DRAM을 설명할 때 가장 중심축은 **서버와 PC의 메인메모리**를 담당하는 DDR4/DDR5다. 데이터센터에서 HBM이 가속기 성능을 좌우하는 '상단 메모리'라면, DDR은 OS·가상화·데이터 처리·네트워크 스택을 운영하기 위한 **시스템 기본 용량**을 맡는다. 그래서 일반 서버 증설이 지속되는 한, 가속기 유무와 무관하게 DDR 기반 비트 수요는 구조적으로 발생한다. 특히 기업 데이터센터나 클라우드의 일반 컴퓨트 노드는 AI 서버만큼 화려하게 보이지 않지만, 실제로는 서비스 트래픽, 데이터 처리량, 스토리지 계층과 캐시 운영을 받치기 위해 꾸준히 증설된다. 이때 메모리는 성능 옵션이 아니라 **운영을 위한 전제조건**이기 때문에, DDR 수요는 경기 사이클을 타더라도 완전히 꺼지기 어렵다.

LPDDR: 모바일·엣지에서 전력 아끼면서 더 빠르게

LPDDR은 전력 제약 환경의 상주·트래픽 수요를 받침

모바일과 초경량 기기, 전력 예산이 타이트한 시스템에서는 LPDDR(Low Power Double Data Rate)이 표준이다. LPDDR은 **저전력·저발열·소형 실장**을 최우선 목표로 설계된 DRAM 규격으로, 스마트폰·태블릿·초경량 노트북 등 **배터리와 열 예산이 타이트한 시스템의 메인메모리**로 사용된다 LPDDR은 고성능 자체보다 전력 효율과 발열, 폼팩터 제약(기판 면적·실장 구조)을 우선순위로 발전해 왔고, "더 적은 전력으로 더 많은 트래픽을 감당"하는 방향으로 진화한다. 최근에는 온디바이스 AI, 고해상도 멀티미디어, 고성능 모바일 게임, 얇은 노트북에서의 상시 대기 같은 사용 패턴이 확대되면서 **메모리 트래픽**과 **상주 부담**이 커지고 있다. 즉 메모리는 단순 저장 공간이 아니라 체감 성능과 배터리 지속시간을 동시에 좌우하는 요소가 된다.

로컬 처리 비중이 커질수록 LPDDR은 '연산을 위한 상주 공간'이 되어 탑재량이 늘기 쉬움

LPDDR 수요는 소비 경기 영향을 받지만, 장기적으로는 기기당 평균 탑재 용량이 늘어나는 흐름이 강해 전체 비트의 최소한의 수요를 유지하게 한다. 특히 엣지에서 로컬 처리 비중이 커질수록, 데이터를 매번 네트워크로 보내기보다 기기 내부에서 처리·압축·요약하려는 요구가 커지고, 그 과정에서 메모리는 **"연산을 돌리기 위한 상주 공간"**이 된다. 전력·열 제약이 강한 환경에서 이런 상주 수요를 감당하려면, 단순히 빠르기만 한 메모리보다 효율적으로 트래픽을 처리하는 저전력 메모리의 중요도가 올라간다. 그래서 LPDDR은 범용 DRAM 내부에서 "모바일 전용 규격"이 아니라, 전력 제약 기반의 컴퓨팅이 확산될수록 **적용 범위가 넓어질 수 있다**.

GDDR·SOCAMM: 대역폭은 GDDR, 실장은 SOCAMM

GDDR은 HBM의 대체가 아니라 비용·공급 제약 속에서 일정 대역폭을 맞추는 대안

그래픽 및 고대역폭이 요구되는 영역에서는 GDDR(Graphics Double Data Rate)이 중요하다. GDDR은 GPU 중심 워크로드를 위해 설계된 **고대역폭 DRAM 규격**으로, 그래픽 처리·게이밍·전문 GPU에서 주로 사용되며, 일부 AI 가속 구성에서는 **HBM을 보완하거나 대체하는 비용 효율형 메모리**로 활용될 수 있다. 전통적으로는 GPU의 그래픽 렌더링과 게이밍을 위해 발전했지만, 최근에는 AI 워크로드가 확대되면서 일부 구성에서 가속기 메모리로도 활용된다. 다만 GDDR은 HBM을 직접 대체한다기보다, 비용·공급·설계 난이도 제약 속에서 특정 성능 목표를 맞추기 위한 **현실적 대안**이다. 하이엔드에서는 HBM이 성능을 규정하지만, 더 넓은 시장에서는 **HBM 없이도 일정 수준의 가속 성능을 구성해야 하는 세그먼트**가 존재하고, 이 구간에서 GDDR은 접근성과 대역폭 측면의 균형점을 제공한다. AI 수요가 확산될수록 이런 중간 지대가 커질 수 있고, 그만큼 GDDR의 활용 범위도 넓어질 여지가 있다.

SOCAMM은 서버가 고집적/고열로 가며 생기는 실장·신호·냉각·업그레이드 한계를 풀려는 폼팩터

SOCAMM은 범용 DRAM이 AI/서버 환경의 새로운 제약에 적응하는 흐름으로 이해할 수 있다. SOCAMM은 서버·워크스테이션에서 **고집적·고열·신호 무결성·업그레이드** 요구를 동시에 충족시키기 위해 설계된 **모듈형 메모리 폼팩터**로, 실무적으로는 **LPDDR 계열을 모듈화**해 시스템 통합과 냉각·실장 효율을 높이려는 접근으로 볼 수 있다. 서버·가속기 플랫폼이 고집적·고전력·고열 구조로 이동하면서, 전통적인 DIMM 중심 구성만으로는 실장 밀도, 전력 전달, 신호 무결성, 냉각, 업그레이드 유연성을 함께 만

족시키기 어려운 구간이 생긴다. SOCAMM은 이런 제약을 완화하면서 **용량 확장성과 시스템 통합 효율**을 개선하는 선택지로 의미가 있다. 핵심은 SOCAMM이 완전히 새로운 메모리라기보다는, **LPDDR의 전력 효율·집적도·실장 유연성**을 서버 쪽 폼팩터로 끌어와 **더 효율적인 메모리 구성을 가능하게 하는 방식**이라는 점이다.

즉, LPDDR이 모바일/엣지에서 전력 효율을 우선하는 규격이라면, SOCAMM은 그 LPDDR의 장점을 살리면서도 모듈형 폼팩터를 통해 서버·고집적 시스템에서 필요한 실장·냉각·업그레이드 요구를 맞추려는 흐름에 가깝다. 그래서 SOCAMM을 설명할 때 LPDDR이 빠질 수 없다. SOCAMM은 HBM처럼 초고대역폭을 제공하는 제품이 아니라, 시스템 설계 관점에서 더 현실적인 방식으로 메모리 용량과 폼팩터를 맞추는 선택지이고, 그 기반에 LPDDR의 효율 특성이 깔려 있다.

NAND SSD: 데이터센터 저장·캐시 계층이 함께 두꺼워진다

범용 DRAM 재편과 함께 NAND 스토리지 계층 수요도 동반 확대되는 흐름

범용 DRAM이 AI/서버 환경의 제약에 맞춰 규격과 폼팩터를 재배치하는 흐름이라면, NAND는 **그 아래에서 데이터의 저장·공급·재사용을 담당하는 스토리지 계층**으로 같이 커진다. 학습에서는 대용량 데이터셋과 반복적인 체크포인트 저장·복원이 필수이고, 추론으로 무게중심이 이동할수록 모델 버전 관리, 로그·리플레이, 벡터DB(RAG)와 캐시 운영처럼 읽기·쓰기 빈도가 높은 워크로드가 늘어난다. 이때 NVMe 기반 엔터프라이즈 SSD는 단순히 용량을 쌓는 부품이 아니라, **데이터가 위로(DDR/HBM) 올라가기 전 병목을 줄여 전체 클러스터 효율을 좌우**하는 인프라로 의미가 커진다. 결국 상단의 HBM과 중단의 DDR/LPDDR이 빨라질수록, 데이터를 얼마나 안정적으로 쌓아두고 빠르게 흘려보내는지가 더 중요해지며, AI 시대의 메모리 수요 변화는 **범용 DRAM의 재편과 함께 NAND 스토리지 계층의 동반 확대**로 이어진다.

결국 **범용 DRAM은 AI 시대에 구조적으로 필요성이 강화**된다. DDR4/DDR5가 데이터센터와 PC의 용량 기반을 만들고, LPDDR이 전력 제약 컴퓨팅의 상주 수요를 지지하며, GDDR이 비용·성능 균형 구간에서 대역폭을 보완하고, SOCAMM이 LPDDR의 장점을 모듈 형태로 끌어와 시스템 통합 효율을 높이는 방식으로 **메모리 믹스가 재편**된다. 이 변화는 **범용 DRAM 수요를 더욱 두껍게 만드는 방향**이다. 여기에 NAND SSD는 학습과 추론이 모두 커질수록 데이터 파이프라인의 병목을 줄이며 인프라 효율을 함께 끌어올리는 축이 된다. 즉, AI로 인한 메모리 수요의 변화는 HBM만 늘어나는 것이 아닌, 전체 시스템이 재구성되면서 범용 DRAM의 역할과 형태, 그리고 이를 받치는 저장 계층까지 함께 확장되는 과정으로 볼 수 있다.

그림 39. 범용 DRAM 유형별 핵심 비교

구분	정의	주 사용 위치	핵심 강점	핵심 제약
DDR4/DDR5	서버의 표준 호스트 메인 메모리	CPU 옆	대용량·확장성·범용성	가속기 근접 대역폭 용도엔 한계
LPDDR	가속기용 고대역폭 DRAM	GPU/가속기 보드	대역폭↑, HBM보다 구성 유연	전력/발열·지연 측면에서 부담
GDDR	저전력 DRAM	SoC 근접(온보드/온패키지)	전력 효율(대역폭/와트)	증설/모듈 교체 제약
SOCAMM	서버용 모듈형 LPDDR	특정 CPU/플랫폼	LPDDR 효율을 '서버 모듈'로 제공	표준 DIMM처럼 범용 아님

자료: KUVIC 리서치 1팀

범용 메모리 수요 성장의 배경

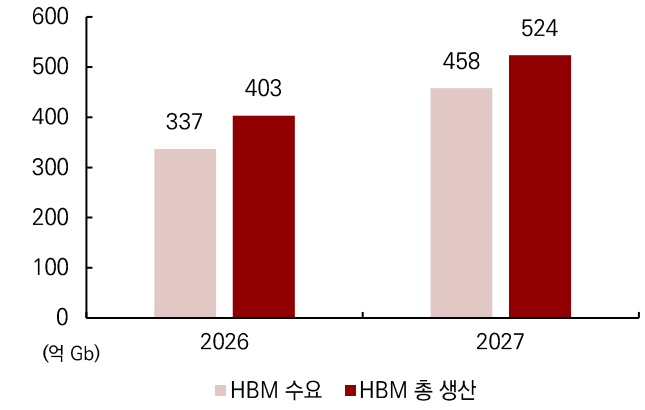
범용 데이터센터 증설이 만드는 기초 수요

AI 트래픽이 커지면 인프라는 AI 서버뿐 아니라 일반 DC로 파급

범용 메모리 수요가 커지는 첫 번째 축은, 업황이 단순히 AI 측면에서 설명되는 게 아니라 **AI가 촉발한 인프라 투자와 트래픽 증가가 범용(일반) 데이터센터로 파급·확산**되며, 결과적으로 **범용 서버/스토리지까지 증설 압력이 커지는 국면**으로 넘어가고 있다는 점이다. 올해 봄부터 이어진 업사이클은 겉으로는 AI가 끌어온 듯 보이지만, 중간까지의 흐름과 최근의 흐름은 성격이 다르다. 한동안은 PC·모바일 같은 B2C 수요가 회복되지 않은 상태에서 AI 수요만 강했고, 범용 수요 중에서도 일반 서버는 AI 서비스가 커지면서 같이 늘어나는 정도에 그쳤다. 그래서 전체 사이클도 **AI 수요가 앞에서 끌고 가는 형태**로 보였다.

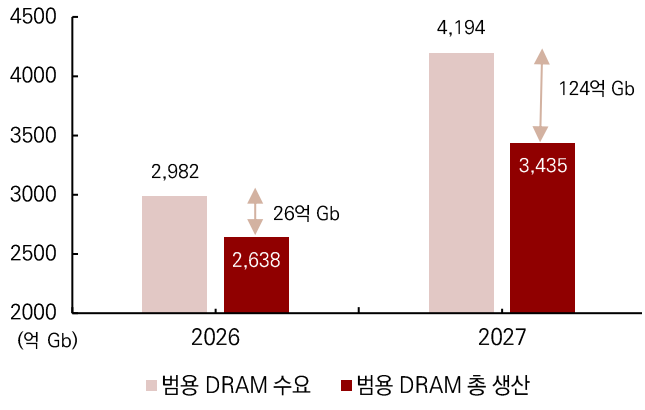
그런데 일정 시점 이후 분위기가 달라진 핵심은, **일반 서버 수요가 단순 동행이 아니라 '확대'로 전환**됐다는 것이다. 범용 수요가 강해지기 시작하면, 시장은 특정 제품(HBM)만 타이트한 업황이 아니라 **범용 DRAM/낸드까지 포함한 전반적 타이트함**으로 체감이 바뀌고, 가격과 조달 행동도 더 공격적으로 움직이기 쉽다. 실제로 이 구간에서 나타난 변화는 수요가 조금 좋아지는 정도가 아니라, 구매자 입장에서 **조달 불안이 급격히 커지는 형태**로 나타난다.

그림 40. 데이터센터 향 HBM 수요 및 공급



자료: KUVIC 리서치 1팀

그림 41. 데이터센터향 범용 DRAM 수요 및 공급



자료: KUVIC 리서치 1팀

HBM: 수급 균형 유지

범용 DRAM: 쇼티지 심화

그에 따라 AI 및 범용 데이터센터향 HBM과 범용 DRAM 공급을 추정한 자료는 위와 같다. **데이터센터향 HBM 수요는 전체 공급과 엇비슷한 수준을 유지**할 것으로 보이는 반면, **범용 DRAM의 경우 26년 26억Gb에서 27년 124억Gb로 1년 만에 375%에 달하는 가파른 상승세로 쇼티지가 발생**할 것으로 보인다.

AI 서버의 경우, 본 리서치 팀이 추정한 데이터센터향 HBM의 수요를 바탕으로 **HBM과 범용 DRAM의 용량 구성 비율이 2026년에는 1:1 비율에서 2027년 1:2 비율로 증가할 것으로 가정하였다.** 일반 서버의 경우, 차세대 CPU의 메모리 채널 확장 효과로 대당 용량이 2.2TB에서 2.42TB(yoy +10%)로 상향 평준화될 것으로 보고 추정을 진행하였다.

이렇듯 일반 서버 수요의 증폭이 발생하게 되는 데에는 크게 세 가지 요인이 겹친다.

첫째, **AI 서비스 사용량 증가가 'AI 서버'뿐 아니라 '일반 서버' 투자를 직접 자극**한다. AI 서비스로 단순히 가속기 추론 연산만 늘어나진 않는다. 서비스 앞뒤로 붙는 데이터 처리(입력 정리, 라우팅, 검색·조회, 후처리, 응답 전송)가 같이 커지고, 이 역할을 맡는 쪽이 일반 서버다. 결과적으로 AI 트래픽이 늘수록 "가속기만 늘리면 된다"가 아니라 **일반 서버까지 같이 증설해야 처리량이 유지되는 구조**가 된다. 이때 일반 서버는 기본적으로 DDR 중심의 용량을 필요로 하므로 범용 DRAM 비트 수요가 두껍게 붙는다.

경쟁 국면에선 '필요'보다 선점·리스크 헤지 목적의 조달

둘째, **빅테크/클라우드의 투자 경쟁이 심화되면서 조달이 '선행'**한다. AI 경쟁 구도가 격화되면, 인프라 투자는 필요해서 사는 것에서 뒤처지면 안 되기에 먼저 확보하는 것으로 성격이 바뀐다. 특히 전력·데이터센터 용량(수 GW 단위)·컴퓨팅 계약이 연쇄적으로 발표되는 구간에서는, 밸류체인 전반이 부품을 제때 못 구할 수 있다는 리스크를 더 크게 가격에 반영한다. 이때 범용 메모리는 AI 서버에 직접 붙는 HBM과 달리 **일반 서버 증설의 즉시성**에 더 민감하게 반응할 수 있다.

셋째, **일반 서버 교체주기(리프레시) 도래가 DDR 수요를 아래에서 떠받쳐 준다.** 2016~2018년 '슈퍼사이클' 때처럼 서버용 DRAM은 HBM과 달리 3사 제품 간 성능 격차가 크지 않아, 결국 누가 먼저 물량을 확보하느냐가 가격과 업황을 좌우했다. 이후 공급과잉으로 사이클이 꺾였지만, 구조적으로 서버 DRAM의 교체주기는 통상 5~7년(보통 5년)이고, 클라우드 업체들이 소프트웨어 최적화·업그레이드로 교체를 버티면서 실제로는 **7~8년까지 늘어진 구간**이 누적됐다. 그 결과 **2023년** 이후부터는 **더 미룰 수 없는 리프레시 수요**가 본격화될 수밖에 없는 시점이 된 것이다. 더 중요한 건 이번 교체가 단순 "고장난 장비 대체"가 아니라 **플랫폼 업그레이드 + DDR5 전환**과 함께 진행되기 쉽다는 점이다. 이 경우 메모리 채널 구성 변화와 평균 탑재 용량 증가로 비트 수요가 점프하며, 메모리 구매는 선택적 지출이 아니라 서비스 운영을 위한 필수 지출로 성격이 바뀐다. 2023년 이후 AI 투자로 범용 서버 교체가 일부 지연됐더라도, **8년 누적 교체분이 한꺼번에 돌아오는 타이밍**에 AI 트래픽 확산과 선행 조달까지 겹치면, 범용 DRAM 수요는 약해 보이던 국면에서도 단기간에 **'확대 국면'으로 전환**되기 쉽다.

그림 42. AI 데이터센터 vs 범용 데이터센터

구분	AI 데이터센터	범용 데이터센터
목적	AI 학습·추론 처리량 극대화	범용 서비스 처리(웹·DB·스토리지·플랫폼)
주 메모리	HBM 중심	범용 DRAM 중심
비중/구성	가속기(HBM 포함) 비중 높음, 클러스터 단위 증설	CPU·일반 서버 비중 높음, 점진적 증설
핵심 병목	메모리 대역폭·지연·동시성	용량·IO·비용 효율

자료: KUVIC 리서치 1팀

그림 43. 서버 교체주기 조정과 범용 DRAM 리프레시 시점

구분	Microsoft	Alphabet	Amazon
서버/네트워크 교체주기 정책	4년 → 6년	4,5년 → 6년	6년 → 5년
정책 변경의 트리거	감가상각·비용 최적화	유효수명 재평가	AI/ML 중심 발전 속도 반영
리프레시 수요 구간	2025 ~ 2026년	2025 ~ 2027년	2025 ~ 2026년

자료: KUVIC 리서치 1팀

표 20. 전세계 Server 공급 전망

(단위: 천 대)

공급업체	2024	1Q25	2Q25	3Q25	4Q25F	2025F	1Q26F	2Q26F	3Q26F	4Q26F	2026F
HPE	1,231	288	300	280	315	1,183	260	262	263	280	1,065
Dell	1,433	355	340	330	370	1,395	346	350	361	358	1,415
Lenovo	810	200	210	250	240	900	229	231	232	230	922
Oracle	68	36	64	75	70	245	77	91	91	92	351
IEIT (RithMax)	1,055	260	280	270	260	1,070	276	274	272	270	1,092
xFusion	390	90	100	110	110	410	105	106	103	101	415
H3C	323	80	80	60	50	270	81	83	62	52	278
ZTE	192	50	52	60	43	205	51	52	60	43	206
Nettrix (Smooth Compute)	242	60	68	60	50	238	60	68	61	50	239
ODM Direct/ White Brand	5,067	1,380	1,410	1,474	1,592	5,856	1,915	1,967	1,771	1,769	7,422
OEM Others/ Self Build (SI)	2,852	704	706	739	780	2,929	788	819	600	500	2,707
전체 Server	13,663	3,503	3,610	3,708	3,880	14,701	4,188	4,303	3,876	3,745	16,112
AI Server	1,652	490	500	560	615	2,165	691	710	736	749	2,886
General Server	12,240	3,013	3,110	3,148	3,265	12,536	3,497	3,593	3,140	2,996	13,226
AI Server 비중 (%)	12.1	14.0	13.9	15.1	15.9	14.7	16.5	16.5	19.0	20.0	17.9

자료: TrendForce, KUVIC 리서치 1팀

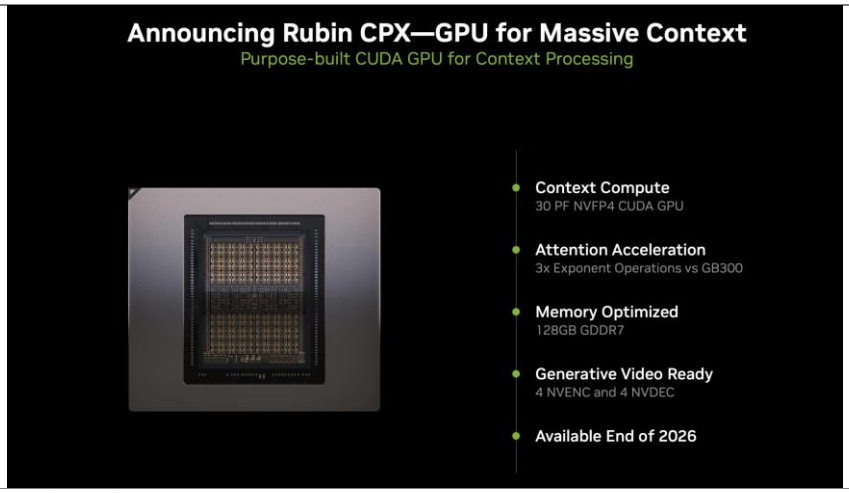
학습 → 추론 전환과 메모리 수요의 '옆으로 확산'

추론은 프리필/디코드로 성격이 갈려 단계별 자원 최적화가 필요해짐

두 번째 축은 AI 워크로드의 중심이 학습에서 추론으로 이동하면서, 추론 인프라가 HBM 달린 가속기만 늘리는 구조에서 벗어나 **단계별 역할 분화**로 확장된다는 점이다. 추론은 크게 프리필(prefill)과 디코드(decode)로 성격이 갈린다. 프리필은 긴 입력(프롬프트/컨텍스트)을 한 번에 처리해 토큰 상태를 준비하는 구간이라 순간적인 연산량이 크고, 디코드는 토큰을 순차 생성하면서 **메모리 트래픽과 캐시(KV) 유지 부담**이 병목이 되기 쉽다. 추론이 커질수록 고가의 HBM 자원을 모든 단계에 풀로 적용하는 방식은 비용 효율이 나빠지고, 그래서 단계별로 최적 자원을 분리해 쓰는 설계가 힘을 받는다.

이 흐름을 상징적으로 보여주는 게 **Vera Rubin CPX**다. CPX는 한마디로 "추론 중에서도 프리필 비중이 큰 구간을 전용으로 처리해, **전체 처리량(throughput)을 가장 저렴하게 최적화**하려는 계층"이다. 프리필은 compute 성격이 강해서, 무조건 HBM을 최대로 얹는 구성보다 더 경제적인 메모리 조합(GDDR 계열 같은)으로 효율을 낼 수 있는 영역이 생긴다. 반대로 디코드는 긴 컨텍스트·동시 요청에서 KV 캐시가 커지며 **대역폭/지연/상주 공간**이 성능을 좌우하니, HBM 중심의 고대역폭 자원이 계속 핵심이 된다. 결국 CPX는 HBM을 줄인다기보다는, **HBM을 정말 필요한 구간(디코드)에 집중시키고**, 프리필은 **상대적으로 범용·저비용 메모리로 확장**해 랙 단위 처리량을 키우는 방향이다.

그림 44. Vera Rubin CPX 사양



자료: NVIDIA, KUVIC 리서치 1팀

계층화가 진행될수록 추론 인프라는 DDR/GDDR/LPDDR로 메모리 믹스가 넓어짐

이 구조가 범용 메모리 수요를 자극하는 포인트는 명확하다.

첫째, 프리필 계층이 커질수록 추론 인프라에 들어가는 메모리 믹스가 HBM 일변도에서 벗어나 **GDDR·LPDDR(모듈형 포함)·DDR 같은 범용 메모리 비중이 커지는 방향**으로 재편된다. 즉 추론이 늘면 HBM만 더 필요한 것이 아니라, **추론 파이프라인을 더 싸고 넓게 깔기 위해** 범용 메모리를 더 쓰는 구조가 만들어진다.

둘째, CPX 같은 계층화가 등장할수록 데이터센터는 특정한 종류의 서버로 밀어붙이기보다, 워크로드를 쪼개서 **각 단계에 최적화된 서버/가속기 풀**을 운영하려고 한다. 이때 풀 자체가 커지고 다양해지니, 그 주변을 받치는 **범용 서버(호스트) 메모리**도 같이 증가한다.

셋째, 범용 메모리 수요를 특히 강하게 자극하는 건 "추론은 상시 운영"이라는 특성이다. 상시 운영에서는 평균 성능보다 **피크 트래픽을 견딜 여력**과 **지연이 무너지지 않는 안정성**이 중요하다. 이를 맞추려면 CPU 서버 수를 늘리거나 서버당 메모리를 늘려 캐시·세션·중간 데이터 처리 여력을 확보하는 쪽으로 설계가 가기 쉽다. 컨텍스트가 길어지고 동시성이 늘면 서비스는 더 많은 데이터를 더 오래 붙잡고 처리하

게 되고, 이때 범용 메모리는 “있으면 좋은 옵션”이 아니라 **서비스 처리량과 단가를 받치는 운영 자원**으로 바뀐다.

추론 서비스를 흐름으로 보면 더 직관적이다. 사용자가 요청을 던지면, 먼저 일반 서버가 입력을 받아 전처리하고, 필요하면 RAG을 위한 벡터 DB 조회 같은 검색/조회 작업을 수행한다. 다음으로 가속기/추론 계층(CPX 포함)이 프리필·디코드 연산을 수행하고, 다시 일반 서버가 결과를 받아 후처리(형식화, 안전 필터링, 로그/저장, 라우팅)를 거쳐 응답을 전달한다. 즉 추론이 커진다는 건 가속기 옆에 HBM만 늘리는 문제가 아니라, 앞·뒤 단계의 **일반 서버 트래픽과 상주 데이터**가 같이 커지는 문제다. 그래서 추론 트래픽이 늘수록 범용 서버(그리고 그 안의 DDR5 중심 용량)가 함께 증가한다.

학습→추론 전환은 범용 메모리 수요를 옆으로 확장

정리하면, 학습 → 추론 전환은 HBM 쇼티지의 이야기로만 끝나지 않는다. **Vera Rubin CPX처럼 프리필 계층이 분리·확대**되면서, 추론 인프라는 단계별 최적화를 위해 메모리 믹스를 바꾸고, 그 결과 범용 메모리(DDR/LPDDR/GDDR)가 **‘옆으로’ 크게 확산**된다. 추론이 커질수록 메모리는 단순 부품이 아니라 **처리량과 TCO를 동시에 결정하는 핵심 변수**가 되고, 범용 메모리 수요는 “AI와 무관한 꾸준함”을 넘어 “AI 추론 확대가 직접 자극하는 성장”으로 성격이 변화한다.

패키징: HBM 경쟁력의 핵심

병목 해결의 key, 패키징 기술

전공정 미세화의 기술적·비용적 한계로 인해 첨단 패키징 기술이 주목받음

그동안 반도체의 성능 향상은 회로 선폭을 좁히는 전공정 미세화가 주도해왔다. 그러나 선단 노드 개발 비용의 급격한 상승과 물리적 한계로 인한 누설 전류 및 발열 문제가 심화되면서 **전공정 위주의 발전은 한계에 도달했다.** 특히 전공정의 속도를 후공정의 데이터 통로(I/O)가 따라가지 못하는 병목 현상이 발생함에 따라, 업계는 **후공정 패키징 기술을 통해 성능을 개선하는 첨단 패키징(Advanced Packaging)에 주목하기 시작했다.** 다시 말해, **고도의 적층 기술**을 도입하여 수율을 높이고 비용을 절감하는 방식이 현대 반도체 설계의 핵심 전략으로 자리 잡은 것이다.

TSV 기술로 적층

HBM의 기본 원리가 바로 TSV 패키징

전통적인 그래픽 메모리인 GDDR은 핀(I/O) 수의 물리적 한계로 인해 대량의 데이터를 처리할 때 병목 현상을 겪어왔다. 이를 해결하기 위해 **TSV(Through Silicon Via) 패키징 기술을 적용시킨 것이 바로 2013년 SK하이닉스가 개발한 HBM(High Bandwidth Memory)이다.** TSV는 **칩 내부에 수직 구멍을 뚫어 본딩을 하는 방식**이기에 와이어 본딩과 달리 연결 공간이 필요 없어 **패키지를 줄일 수 있으며, 핀의 개수 또한 늘릴 수 있다**는 장점이 있다. 따라서 DRAM은 수직 적층을 통해 좁은 공간 내에서 한 층당 1,024개의 데이터 통로를 확보함으로써 GDDR과는 비교할 수 없는 압도적인 대역폭을 형성할 수 있다. 이러한 고속 데이터 전송 능력과 집적도는 막대한 양의 데이터를 처리해야 하는 **AI 학습(Training) 분야에서 HBM을 필수적인 존재로 만들었다.**

하이브리드 본딩을 향해

HBM의 적층 경쟁은 단순히 칩을 쌓는 단계를 넘어, **칩 사이를 어떻게 연결하고 채우느냐는 본딩 기술**과 **반도체 칩을 배치하고 연결하는 구조인 패키지 플랫폼** 기술이 경쟁력을 결정하고 있다.

표 21. HBM 본딩 기술

	주요 특징 및 메커니즘	장점 및 핵심 경쟁력	한계 및 향후 과제	주요 기업
TC-NCF	칩 사이에 절연 필름 넣고 열·압력으로 압착	고단 적층 시 전체 두께 제어에 유리	균일 압력 어려움, 생산 효율 낮음	삼성전자, 마이크론
MR-MUF	칩 쌓은 후 액체 보호재 주입해 일괄 경화	방열 성능 우수, 생산 속도 매우 빠름	소재 배합 노하우 등 진입장벽 높음	SK하이닉스
Hybrid	범프 없이 구리(Cu) 패드끼리 직접 접합	16단 이상 필수 기술	극도의 공정 정밀도 및 비용 발생	TSMC

자료: KUVIC 리서치 1팀

TC 본딩은 압력 균일성에 한계가 있으나 두께 균일성이 탁월

칩을 어떤 방식으로 연결하는지에 관한 본딩(Bonding) 기술의 경우 크게 세 가지로 방식으로 발전해왔다. **삼성전자와 마이크론이 주력하는 TC-NCF 방식**은 칩 사이에 비전도성 필름을 넣고 열과 압력으로 누르는 기술이다. **압력 균일성을 확보하기 어려워 대량 생산에 한계가 있다는 지적도 있으나, 삼성전자는 축적된 노하우와 어드밴스드 NCF 기술을 통해 칩 휘어짐을 보완하고 있다.** 특히 필름 형태의 특성상 두께 균일성 제어에 강점이 있어, 적층 단수가 높아질수록 고단 적층에서 유리한 고지를 점할 수 있다는 전망이다.

MR 본딩은 고사양 AI 서버 시장의 중심

반면 **SK하이닉스**는 액체 보호재를 주입해 한 번에 굳히는 **MR-MUF 공정**을 통해 시장을 주도하고 있다. 이 방식은 **생산 효율성이 높고 방열 성능이 뛰어나 고사양 AI 서버 시장에서 강력한 차별점**을 갖는다. 현재 일본 나믹스(Namics)와의 독점적 공급망을 바탕으로 HBM3 시리즈까지 기술적 우위를 이어가

고 있다.

하이브리드 본딩은 2028년부터

향후 12단 이상의 초고단 적층 시대에는 **하이브리드 본딩**이 차세대 솔루션으로 부상할 것으로 전망되는데, 이는 **납땜용 범프 없이 칩 단면을 직접 접합하는 구조를 실현하여 칩 사이의 거리를 획기적으로 줄여준다.** 그렇기에 하이브리드 본딩은 미래 HBM 제조의 필수 공정이 될 것으로 보이며, 성능 고도화의 결정적 열쇠가 될 전망이다. 최근 JEDEC의 HBM 높이 규제 완화로 기존 본딩 방식의 기술 수명이 연장됨에 따라 하이브리드 본딩의 조기 도입은 당분간 지연될 것으로 보이나, 16단 이상의 초고적층이 필수적인 HBM4E 세대부터는 기술적 임계점 도달로 인해 하이브리드 본딩 채택이 불가피할 전망이다. 이미 **삼성전자와 SK하이닉스 모두 HBM4E를 기점으로 해당 기술 도입을 공식화**했으며, 엔비디아 등 핵심 고객사가 HBM5 및 하이브리드 본딩 패키징을 차세대 GPU 아키텍처에 적용할 것으로 예상되는 **2028~2029년경이 업계 전반의 본격적 기술 보급 시점이 될 것으로 보인다.**

적층에 적층을 더해, 3D 패키징

표 22. HBM 패키징 기술

	주요 특징 및 메커니즘	장점 및 핵심 경쟁력	한계 및 향후 과제	주요 기업
2.5D	인터포저 위에 로직 칩-HBM 수평 배치	현재 AI 가속기(GPU)의 표준 구조	인터포저 비용 및 공간 점유 문제	TSMC, SK·엔비디아 연합
3D	인터포저 없이 수직으로 완전 적층	데이터 경로 최단화, 성능 극대화	열 방출 문제 및 하이브리드 본딩 필수	삼성전자, TSMC

자료: KUVIC 리서치 1팀

반도체 칩을 배치하고 연결하는 구조인 패키지 플랫폼 역시 성능 극대화를 위해 진화하고 있다. 가장 기본적인 **2D 패키징**이 두 개 이상의 칩을 수평으로 나열하는 방식이라면**, 3D 패키징은 칩을 수직으로 쌓아 올려 TSV로 연결하는 입체적인 구조**를 말한다. 최근에는 구조적으로는 수평 배치를 따르되 성능은 3D에 준하도록 설계된 **2.5D 패키징이 AI 반도체 시장의 주류로 떠오르고 있다.**

2.5D 패키징은 실리콘 인터포저로 로직 칩과 HBM을 수평으로 연결

2.5D 패키징의 핵심은 반도체 칩과 기판 사이에 위치하는 **실리콘 인터포저**라는 추가 계층이다. 인터포저는 수평으로 배치된 서로 다른 칩들 사이에서 전기적 연결을 중계하는 역할을 수행한다. 특히 HBM은 그 자체로 이미 적층된 구조를 가진 반(半) 패키지 제품인데, 이를 **로직 칩(CPU/GPU) 옆에 평행하게 배치하고 인터포저로 통합**함으로써 하나의 고성능 시스템을 완성한다.

3D 패키징은 칩을 수직으로 완전 적층

최종적인 지향점인 **3D 패키징은 칩을 완전히 위아래로 겹쳐 쌓기 때문에 실리콘 인터포저가 필요하지 않으며,** 평면적인 공간 점유를 최소화할 수 있다. 이는 데이터 처리 속도를 비약적으로 높여주지만, 이를 실현하기 위해서는 전기 신호의 밀도를 극도로 높여야 하는 과제가 뒤따른다. 결국 3D 패키징의 완성은 **앞서 언급한 하이브리드 본딩 기술과 직결된다.** 범프를 완전히 제거하고 구리 패드를 직접 연결하는 상태에 도달함으로써 비로소 진정한 의미의 수직 통합이 가능해지기 때문이다. 현재 TSMC는 'SoIC'라는 명칭으로 이러한 3D 패키징 서비스를 선도하고 있으며, 삼성전자 역시 기술 격차를 좁히기 위해 차세대 3D 솔루션 개발에 역량을 집중하고 있다.

HBM을 이을 차세대 메모리 기술

메모리 기술, 더 높아진다

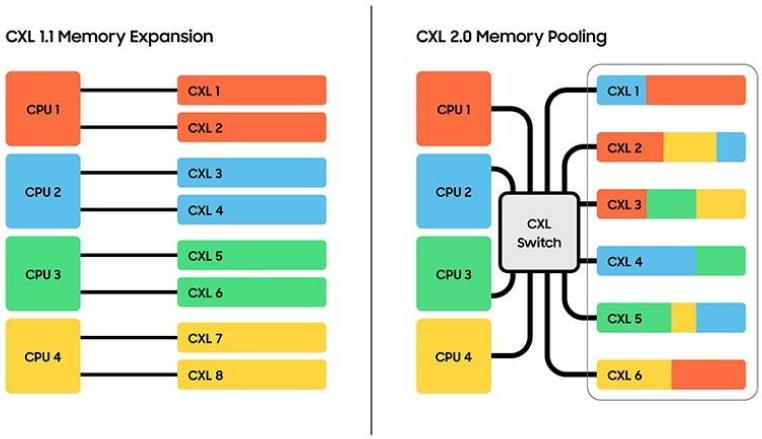
기존 메모리 용량·전력 한계 극복하기 위해 차세대 기술(CXL, LPCAMM, PIM)이 필수적

HBM과 범용 DRAM의 시너지로 대역폭 병목이 완화되고 있으나, **폭발적으로 늘어나는 AI 추론 데이터를 기존의 DRAM 구조만으로 감당하기에는 여전히 용량의 물리적 한계가 존재한다.** 현재의 서버 구조는 CPU당 결합할 수 있는 메모리 모듈 수가 제한적이며, 데이터가 프로세서로 이동하는 과정에서 막대한 전력이 소모되고 있다. 따라서 범용 메모리의 확산은 단순한 판매량 증대를 넘어, 시스템의 구조적 비효율을 해결할 새로운 폼팩터와 아키텍처의 등장을 필연적으로 요구한다. 이에 따라 **확장성(CXL), 전력 효율성(LPCAMM), 연산 융합(PIM)을 골자로 하는 차세대 기술들이 시장의 주류로 부상하고 있다.**

CXL (Compute Express Link)은 메모리 신대륙

현재 데이터센터는 유휴 자원을 효율적으로 해결하지 못하는 문제를 당면하였다. 서버 A는 메모리가 부족해 연산이 멈추는데, 옆에 있는 서버 B는 **메모리가 남아도는 유휴 메모리(Stranded Memory)가 발생하는 이유는 기존 구조에서는 CPU별로 메모리가 물리적으로 종속되어 있어, 남는 자원을 다른 서버가 끌어다 쓸 수 없기 때문이다.**

그림 45. CXL



자료: 삼성전자, KUVIC 리서치 1팀

메모리 풀링 기술로 유휴 메모리 낭비 해결하는 CXL

이에 대한 대안인 CXL은 **PCIe 인터페이스를 기반으로 CPU, GPU, 메모리 등을 연결하는 통합 규격**이다. 이를 통해 DRAM 용량을 무한대로 확장할 수 있을 뿐 아니라, **여러 프로세서가 거대한 메모리 풀(Pool)을 공유하는 '메모리 풀링(Pooling)'**이 가능해진다. 그에 따라 **기존 서버 구조에서는 슬롯 수 제한으로 DRAM 증설이 물리적으로 불가능했다면, CXL 도입은 이를 무한대로 확장하게 해주어 서버당 장착 가능한 메모리 용량의 하방을 지지하고 상단을 여는 효과를** 가져올 것이다.

현재 메모리 3사의 CXL 개발 현황에 있어 **삼성전자가 선두**에 있고 SK 하이닉스는 상당수 리소스가 HBM 개발에 집중되고 있기에 개발 속도는 느리나, 자체 소프트웨어 개발에 힘써 CXL 생태계를 구축하는 록인(Lock-in) 전략을 추구하고 있다.

표 23. CXL 개발 현황

	삼성전자	SK하이닉스	마이크론
CXL 2.0	24년 하반기부터 양산 시작	96GB 제품 고객 인증 완료 **128GB 제품 고객 인증 진행 중**	메모리 모듈 공개 및 고객 평가 단계
CXL 3.0	**26년 상반기 공급 목표**	공식 양산 시점 미공개	CXL 3.0 지원 계획은 공개 수준
전략적 차별성	D램 + 낸드 결합형 CXL 메모리(CMM-H) 27년 출시 예정	HMSDK 소프트웨어 개발해 리눅스에 적용	상용화 드라이브 약한 상태

자료: KUVIC 리서치 1팀

현재 상용화 된 제품은 CXL 2.0을 지원하는 삼성전자의 CMM-D인데, 24년 하반기부터 양산을 시작했으나 클라우드 서비스 공급자(CSP) 등 고객들의 수요는 아직 미미하다. CXL 3.0에 탑재된 패브릭 기능은 연결 가능 노드 수를 기존 16개에서 4,096개로 대폭 확장하는데, 이러한 압도적인 확장성을 바탕으로 CXL 시장은 도입기를 지나 본격적인 개화기에 진입할 전망이다.

CXL 3.0 상용화는 27년부터

종합하여, CXL 기술의 선두주자 **삼성전자가 CXL 3.0을 지원하는 CMM-D는 26년 상반기 공급을 목표로** 하고 있다는 점, CXL 3.0을 지원하는 프로세서인 **인텔의 다이아몬드 래피즈가 26년 하반기 출시 예정**인 점을 고려했을 때, 데이터센터에 CXL이 본격적으로 도입됨에 따라 **2027년부터 시장이 본격 개화할 것으로 보인다.**

다만 CXL 도입을 둘러싼 구조적 한계와 이에 대한 반론 역시 함께 검토할 필요가 있다. **CXL은 서버 및 데이터센터의 메모리 구조 전반을 재설계해야 하는 기술**로, 기존 CPU-GPU-메모리 결합 구조에 최적화된 인프라를 운영해 온 기업들 입장에서는 **초기 도입에 보수적인 태도를 보일 가능성이 높다.**

DC의 소극적 도입과 메모리 수요 잠식에 대한 우려 존재

또한 메모리 풀링을 통해 시스템 전반의 메모리 활용 효율이 개선될 경우, 이론적으로는 서버당 필요 메모리 용량이 감소할 수 있다는 우려도 존재한다. 메모리 중복 할당이 줄고 유휴 메모리 활용도가 높아질 경우, 단위 서버 기준 메모리 탑재량이 축소될 수 있다는 해석이다. 이에 따라 **CXL이 오히려 메모리 수요를 잠식할 수 있다는 시각도 일부 제기된다.**

지금은 비용 절감보다 성능에 신경써야 할 때, 메모리를 축소시키지 않을 것

그러나 이러한 해석은 실제 데이터센터의 확장 논리와는 다소 괴리가 있다. **메모리 효율성 개선은 비용 절감보다는 처리량 확대와 서비스 품질 개선으로 전환되는 경우가 일반적**이다. 특히 AI 추론 및 멀티테넌트 클라우드 환경에서는 메모리 효율이 확보될수록 더 많은 동시 세션과 더 긴 컨텍스트를 수용하려는 수요가 동반 확대된다. 결과적으로 **메모리 풀링은 서버당 메모리 축소로 이어지기보다는, 시스템 전체의 메모리 사용 범위를 확장시키는 방향으로 작동할 가능성이 높다.**

가속기 기업이 채택하지 않는 것에 대한 우려 존재

한편 엔비디아와 AMD 등 주요 가속기 업체들이 CXL 아키텍처를 전면적으로 채택하지 않고 있다는 점 역시 한계 요인으로 지적된다. 엔비디아는 NVLink라는 독자적인 고속 인터커넥트를 통해 GPU 간 통신과 메모리 접근 문제를 자체적으로 해결하고 있으며, 개별 GPU 또는 GPU 클러스터 단위에서 대규모 연산을 처리하는 구조를 채택하고 있다. 이로 인해 추론 과정에서의 지연 시간 문제는 상대적으로 덜 부각되며, CXL을 통한 외부 메모리 확장의 필요성도 제한적으로 인식되고 있다.

적용 영역에 대한 오해에 불과하다

다만 이는 CXL의 유효성이 낮다는 의미라기보다는 적용 영역의 차이에 가깝다. NVLink가 GPU 내부 및 GPU 간 통신에 최적화된 폐쇄형 구조인 반면, CXL은 CPU-GPU-메모리를 아우르는 범용 인터커넥트로서 데이터센터 전반의 자원 유연성을 높이는 데 목적이 있다. 다수의 워크로드가 혼재된 범용 클라우드 환경이나, GPU 활용률 편차가 큰 추론 중심 인프라에서는 NVLink만으로 해결하기 어려운 메모리 비효율 문제가 구조적으로 발생할 수 있다.

추론은 HBM에, 지식은 CXL에

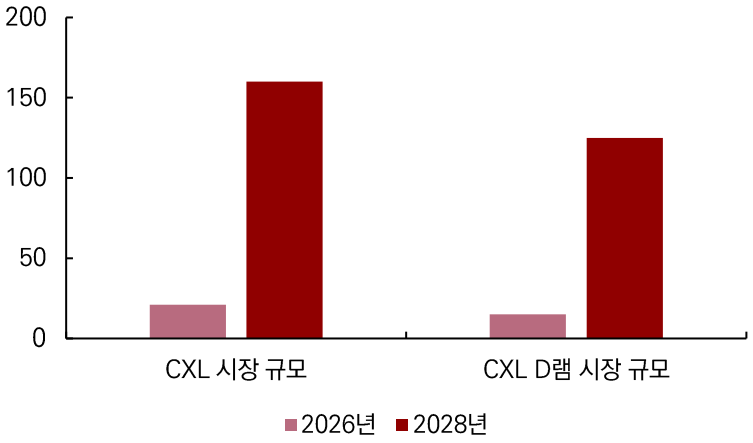
CXL은 엔그램의 직접적 수혜 기술

이러한 맥락에서 **CXL은** 2026년 2월 중순 출시되는 **딥시크 V4에 탑재될 '엔그램(Engram)' 메모리 아키텍처의 직접적 수혜 기술로 평가된다.** Engram은 LLM 추론 과정에서 KV cache를 보다 효율적으로 관리함으로써 GPU당 HBM 사용 효율을 개선하는 구조를 제시한다. 추론(Reasoning)과 지식(Knowledge)을 분리해, **지연 민감도가 높은 추론 영역은 HBM에 유지하고, 상대적으로 접근 빈도가 낮은 지식 영역은 별도의 메모리 계층으로 오프로딩**하는 방식이다. 이 과정에서 **CXL 메모리는 대용량 지식 메모리를 담당하는 핵심 인터페이스로 활용될 가능성이 높다.**

CXL은 HBM의 대체재가 아닌 보완재

다만 **Engram 역시 HBM의 역할을 축소시키는 기술은 아니다.** Engram은 HBM을 덜 사용하는 구조가 아니라, HBM을 지연 민감 연산에 집중시키는 방향으로 재정의한다. **첫째,** 추론 효율 개선은 동시 처리 가능한 세션 수 증가로 이어지며, **서비스 사업자는 이를 비용 절감보다는 전체 추론 처리량 확대로 전환하는 경향**이 강하기에 GPU 및 가속기 증설 수요를 자극해 전체 HBM 탑재량이 증가할 것이다. **둘째,** LLM 서비스의 진화 방향은 멀티모달, 에이전트 기반 추론, 장기 대화형 서비스 등 **컨텍스트 길이 확대**에 있으며, **이 영역은 반드시 GPU 인접 초고대역폭 메모리인 HBM이 담당**해야 한다. **셋째, 활성 데이터 및 지연 민감 데이터의 비중이 높아지면서** HBM의 전략적 중요성은 오히려 강화된다.

그림 46. CXL 시장 규모



자료: Yole Intelligence, KUVIC 리서치 1팀

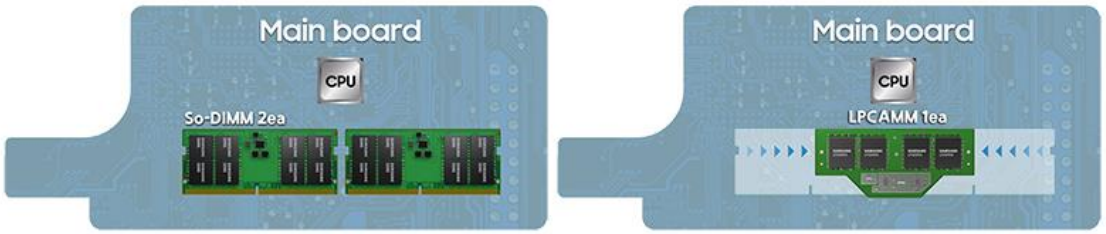
CXL은 HBM과 함께 성장해나갈 것

결론적으로 CXL은 단기적으로는 인프라 전환 부담과 기존 가속기 생태계의 관성으로 인해 도입 속도가 제한될 수 있다. 그러나 AI 추론 워크로드 확대와 메모리 중심 병목 구조의 심화, 자원 활용 효율에 대한 요구가 높아질수록 CXL의 필요성은 점진적으로 부각될 가능성이 높다. 특히 Engram과 같은 차세대 추론 아키텍처가 확산될 경우, 메모리는 '추론은 HBM, 지식은 CXL'로 역할이 분화되는 계층 구조로 진화할 가능성이 크다. 그 결과, **CXL은 HBM을 중심으로 한 메모리 생태계를 확장하는 필수적 기술로서 성장해나갈 것임을 예상할 수 있다.**

LPCAMM (Low Power Compression Attached Memory Module)

기존 노트북 메모리는 성능과 교체 편의성 사이의 딜레마에 빠져 있었다. 탈부착이 가능한 SO-DIMM은 두껍고 전송 속도가 느려 AI 연산에 부적합했고, 속도가 빠르고 전력이 낮은 LPDDR은 메인보드에 납땜(Soldering)되어야 해 용량 확장이 불가능했다. 온디바이스 AI 구동을 위해서는 고속·고용량이 필수적이나, **기존 폼팩터로는 이를 동시에 만족시킬 수 없었다.**

그림 47. LPCAMM



자료: 삼성전자, KUVIC 리서치 1팀

부피와 전력을 줄이면서도 탈부착이 가능한 LPCAMM

이를 해결하기 위해 등장한 **LPCAMM은 LPDDR 패키지를 기판에 탑재해 탈부착이 가능하게 만든 모듈**이다. 기존 SO-DIMM 대비 부피를 60% 줄이면서도 전력 효율은 70% 개선하여, 얇은 노트북에서도 고성능 AI 모델을 구동하고 필요시 용량을 늘릴 수 있는 유연성을 제공한다. LPCAMM의 등장으로 메모리 업체들이 그동안 **부품으로만 팔던 LPDDR을 모듈 형태의 완제품 솔루션으로 판매하며 ASP를 높일 수 있을 것**이다. 또한, 얇은 폼팩터에서도 고용량 구현이 가능해짐에 따라 **AI PC 교체 수요와 맞물려 프리미엄 메모리 시장의 새로운 캐시카우가 될 전망**이다.

LPCAMM 상용화에 있어 가장 큰 장벽은 **하드웨어 생태계의 변화**다. LPCAMM을 쓰려면 노트북 메인보드의 커넥터 설계 자체를 완전히 바꿔야 하기에 제조사(OEM)들의 설계 변경 비용이 발생한다. 하지만 2024년 JEDEC 표준화(LPCAMM2)가 완료되었고, 삼성전자와 마이크론이 양산을 시작하였기에 **2025~2026년 출시되는 고성능 AI 노트북(워크스테이션급)을 시작으로 2027년 이후에는 일반 소비자용 노트북까지 침투율이 확대될 것으로 예상된다.**

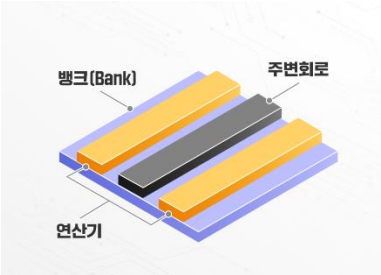
PIM (Processing in Memory)

현대 컴퓨팅의 가장 큰 비효율은 '폰 노이만 병목(Von Neumann Bottleneck)'에서 온다. **데이터가 메모리(저장)와 프로세서(연산) 사이를 끊임없이 오가는 과정에서 전체 에너지의 80% 이상이 소모되고 성능 저하가 발생한다.** 아무리 빠른 HBM을 써도 데이터 이동 거리 자체가 줄어들지 않으면 근본적 해결이 어렵다.

메모리 내부에서 직접 연산을 수행해 데이터 이동을 최소화하는 PIM

PIM은 메모리 내부에 연산 기능을 통합하여 비메모리 반도체의 역할을 일부 수행하게 하는 기술로, PIM 기술이 적용되는 메모리 종류에 따라 HBM-PIM, GDDR-PIM 등 다양한 형태로 구현 가능하다. PIM은 **데이터 이동을 최소화하여 처리 속도를 높이고 시스템 전력 소모를 획기적으로(최대 70% 이상) 줄일 수 있게 해준다. 연산 비용의 절감은 고성능 AI 구현을 위한 메모리 탑재량 확대의 기폭제**가 될 것이며, 이는 메모리 수요의 양적·질적 성장으로 이어질 것을 기대해볼 수 있다.

그림 48. PIM



자료: SK하이닉스, KUVIC 리서치 1팀

PIM이 아직 상용화되지 못한 결정적 이유는 **소프트웨어 스택(Software Stack)의 부재**다. 기존 프로그램 코드는 연산은 CPU/GPU가 한다고 가정하고 짜여 있어, PIM을 활용하려면 코드를 수정하거나 별도의 컴파일러가 필요하다. 즉, 프로그래머가 쓰기 어렵다는 점이 확산을 막고 있다. 현재는 특정 AI 연산 전용 가속기 시장부터 진입 중이며, **표준화(JEDEC)와 소프트웨어 호환성이 확보되는 2027~2028년 이후에야 범용적인 상용화 단계에 진입할 것으로 관측된다.** 이 과정에서 **로직과 메모리 제조 기술을 동시에 보유한 기업들이 시장 선점에 유리할 것**으로 판단되며, **칩 간 접합(Hybrid Bonding) 장비와 미세 공정용 EUV/ALD 소재 기업의 수혜가 예상된다.**

메모리 패러다임의 전환

차세대 기술은 메모리 수요의 양적·질적 성장 기폭제

결국 이 모든 기술적 진보가 가리키는 방향은 단 하나다. **메모리는 이제 '싸게 많이 찍어내는 범용품'에서 벗어나 '시스템의 문제를 해결하는 고부가가치 솔루션'으로 진화하고 있다**는 점이다. 일각에서는 최적화 기술이 나오면 메모리를 덜 쓰게 되어 수요에 부정적일 것이라 우려하지만, 이는 과거 하드디스크 용량이 커지면 데이터 저장 수요가 줄어들 것이라 예상했던 것과 같은 오류다. **기술 혁신으로 데이터 처리 비용이 낮아지면, 시장은 그 절감된 비용만큼 더 많은 데이터를 처리하고 더 거대한 AI를 돌리는 방향으로 움직인다.** 따라서 차세대 기술(CXL, LPCAMM, PIM)은 메모리 기업들에게 업황의 사이클 변동성을 줄이고 지속적인 이익 성장을 위한 가능하게 하는 촉매가 될 전망이다.

DeepSeek-V4: mHC, 엔그램, 추론 혁신

3가지 논문

3가지 논문: mHC, 엔그램, R1 기술분석 보고서

25년 초 R1 모델을 발표하면서 엔비디아 주가를 폭락시켰던 DeepSeek가 2월 중순 V4 모델 출시를 앞두고 있다. R1 모델 출시 당시 아키텍쳐 개선으로 학습 연산량을 혁신적으로 줄일 수 있다는 사실을 증명하며, GPU 구매 수요 감소에 대한 공포를 시장에 불어넣었다. 그리고 26년 1월 초부터 지금까지 한달 간 발표된 논문들에 의하면 V4에 담길 두 번째 혁신은 '메모리'와 '추론'을 향할 것이라는 점을 알 수 있다. 해당 논문들은 **'mHC(Manifold-Constrained Hyper-Connections)', '엔그램(Engram) 조건부 메모리', 그리고 'DeepSeek-R1'**의 업데이트된 기술 보고서이다.

mHC(Manifold-Constrained Hyper-Connections)

mHC: 학습 요율 및 성능 혁신

1월 1일 공개된 **mHC** 논문은 현대 딥러닝 모델의 구조적인 문제인 '심층 신경망의 학습 불안정성'을 해결하기 위한 **학습 효율과 성능에서의 혁신**을 담고 있다. 기존의 HC(Hyper-Connection) 연결 구조에서는 모델의 깊이가 깊어질 수록 잔차 연결(Residual Connection) 방식으로 인해 각 층 별로 신호가 누적되어 10의 제곱 단위로 신호가 증폭되는 신호 폭발 문제가 구조적으로 존재했다. 그러나 딥시크는 이중 확률 행렬을 통해, 데이터셋의 길이(norm)를 1로 제한하면서 신호 증폭률을 획기적으로 낮췄다. 기존 HC 구조가 깊이 64층에서 신호 증폭율이 10의 16승 이었다면, mHC 구조에서 1.6배까지 획기적으로

낮춘 것이다. 이는 **1,000조 배 이상의 신경망 안정성을 확보**한 것으로 학습 효율을 비약적으로 향상시켰다.

그러나 더 중요한 것은 모델 성능까지 상승했다는 점이다. 안정성 및 학습 효율 확보 시 일부 성능 희생이 일반적이나, mHC 모델은 신호를 왜곡 없이 깊은 층까지 전달하며 모델 성능까지 향상시켰다. 특히 모델 파라미터수가 크고 학습 시간이 길수록 이러한 효과가 뚜렷했다. 따라서 mHC는 학습 속도와 비용을 동시에 절감하는 핵심 기술로 자리잡을 전망이다. R1의 MoE로 연산량을 수십 배 줄인 데 이어 AI 학습 측면에서 두 번째 혁신이 이루어지는 셈이다.

엔그램(Engram); 조건부 메모리(Conditional Memory)

엔그램: HBM을 DRAM, SSD로 대체

1월 13일 공개된 **엔그램 논문은 현재의 메모리 병목 상황을 겨냥**했다. 기존 모델은 지식을 신경망 전체 가중치에 녹여서 저장하기에, 간단한 책 한 권도 온 도서관을 뒤지는 식으로 비효율적이었다. 그러나 딥시크는 논문에서 뇌의 기억 저장 방식(N-gram)에 착안하여, 추론 기능과 지식 저장 기능을 물리적으로 분리했다. 거대한 저장소를 '룩업 테이블'로 두고, 추론 기능에는 해당 테이블의 위치를 연상하는 기본 정보만 저장하는 구조이다. 즉, 단서를 기반으로 연상하여 지식을 검색하는 인간의 기억 방식과 유사하다.

연구진은 **파라미터수의 20~25%라는 엔그램 메모리 할당 비율 최적치**를 찾아냈다. 이보다 적으면 적절한 지식 찾기 실패로 환각이 발생하고, 많으면 느려지면서 추론 능력이 저하된다. 더욱 중요한 것은 엔그램 메모리는 연산이 필요 없기 때문에 HBM이 아닌 RAM, SSD, CXL 확장 메모리에 저장이 가능하다. 100B 파라미터 규모의 룩업 테이블을 외부화해도 추론 속도 저하는 3%에 불과했다. 즉, **HBM을 LPDDR, SSD로 대체할 수 있는 기술적 토대**가 마련된 것이다.

또한 엔그램은 **'무한 문맥'**이라는 파괴적 혁신을 가능케 할 수 있다. 기존 모델은 문맥의 길이가 길어지면 연산량이 폭증하고 중간 정보를 망각하는 문제가 발생하는데, 엔그램은 해시 기반 조회로 문맥 길이와 무관한 상수 수준의 복잡도로 정보 검색이 가능하다. 이는 수백 페이지 PDF에서 특정 정보를 정확히 추출하는 등의 작업이 실현 가능함을 의미한다.

추론 혁신; R1 기술 분석 보고서

1) 연산 비용을 아낄 수 있었던 이유
2) 엣지 AI 가능성

1월 20일 경 업데이트 된 DeepSeek-R1 기술 분석 보고서(20pg → 86pg)에서는 순수 강화학습을 통한 추론 영역의 혁신을 기술한다. 기존 PPO(Proximal Policy Optimization) 방식은 답변에 대한 보상 함수를 추정하는 Critic 모델을 사용하여 선호 답변을 모방했지만 그 대가로 메모리를 2배로 사용했다. 반면 딥시크의 **GRPO(Group Relative Policy Optimization) 모델**은 답변 그룹 내에서 상대적 우열만을 평가하여 메모리 사용량을 획기적으로 절감했다. 따라서 훨씬 적은 비용으로 OpenAI 모델 등에 비견된 고성능 모델을 학습시킬 수 있었던 것이다.

또한 해당 보고서는 **증류를 통한 엣지 AI 가능성**을 입증했다. 딥시크는 R1 모델이 생성한 고품질의 사고 데이터(Reasoning Trajectories)를 사용하여 1.5B, 7B, 14B 등 작은 크기의 모델들을 미세 조정했는데, 이들은 R1의 사고 패턴을 모방하며 추론 점수가 비약적으로 상승했다. 즉, 향후 온디바이스(On-device) AI나 엣지 컴퓨팅 환경에서도 고도화된 추론 기능을 사용할 수 있는 길을 열었다. 최상위 모델 하나만 강력하게 학습시키면 경량화 모델을 효율적으로 교육시키는 지식의 낙수효과가 가능해진 것이다.

총평

mHC는 학습 안정성을, 엔그램은 메모리 효율성을, R1 업데이트는 추론 효율성을 극대화하며 딥시크 AI의 경쟁력은 V4를 필두로 재강화될 전망이다. 또한 Engram과 R1 기술 분석 보고서는 AI 하드웨어 시장에 중대한 영향을 미칠 가능성이 있다.

엔그램은 결국 **고가의 HBM 수요가 DRAM이나 낸드로 일부 대체될 수 있음을 시사**하기 때문에 전력 및 비용 효율성이 높은 **LPDDR**과, 속도가 느려도 가격이 저렴한 **eSSD**의 수요가 한 층 더 높아질 트리거가 될 수 있다. 즉, 지금의 범용 DRAM 및 낸드 병목 상황이 더욱 심화될 수 있다는 것이다.

R1 기술 분석 보고서는 고성능 모델도 상대적으로 적은 연산량과 메모리로 학습시킬 수 있다는 점을 시사한다. 그러나 2025년 내내 **제번스의 역설**에 따라 컴퓨팅 비용 하락이 오히려 수요를 창출해온 점을 감안할 때, 지식 증류를 통한 엣지 AI 확산 가능성이 더 중요한 의미를 가진다. 모든 AI 디바이스, 자율주행차, 휴머노이드, IoT 기기에 고가의 최상위 성능 칩이 불필요해지며 저비용 고성능 AI 기기 보급이 가능해졌기 때문이다. 따라서 **중간 성능 AI 가속기 및 범용 DRAM 수요 확대의 기술적 기반**으로 해석된다.

Appendix

AI CAPEX

표 24. 2025년 이후 예정된 데이터센터 건설 프로젝트

프로젝트명	Country	Owner	전력용량(MW)	Operational Date
Telangana Yotta H1 Hyderbad AI City Cluster Phase 1	India	Government of Telangana	6	2026.12
S. Korea 6th national supercomputer	South Korea	Ministry of Science and ICT	12	2026.12
Foxconn Big Innovation Cloud AI factory	Taiwan	Foxconn	30	2026.12
Nscale Loughton	UK	-	90	2026.12
Sesterce Valence	France	Sesterce	96	2026.12
CoreWeave Muskogee	USA	CoreWeave	100	2026.12
EU AI Gigafactory #1	미정	EU	100	2026.12
EU AI Gigafactory #2	미정	EU	100	2026.12
EU AI Gigafactory #3	미정	EU	100	2026.12
Applied Digital CoreWeave Ellendale Phase 2	USA	Applied Digital	150	2026.12
xAI Colossus 2 Memphis Phase 2	USA	xAI	170	2026.12
CyrusOne × Calpine	USA	ECP, KKR, CyrusOne	190	2026.12
Stargate UAE Phase 1	UAE	Stargate (OpenAI)	200	2026.12
Stargate OpenAI Norway	Norway	Nscale	230	2026.12
Nebius New Jersey	USA	Nebius AI	300	2026.12
OpenAI/Microsoft Mt Pleasant, Wisconsin Phase 1	USA	OpenAI, Microsoft	300	2026.12
Tesla Cortex Phase 2	USA	Tesla	370	2026.12
together.ai project Phase 2	Europe	TogetherAI, Hypertec	600	2026.12
Stargate Abilene Supercluster Phase I-2	USA	Oracle	1,000	2026.12
Meta Prometheus	USA	Meta AI	1,000	2026.12
Coreweave 펜실베니아	USA	CoreWeave	100	2026.12
Pennsylvania Digital I (PAX) project Phase 1	USA	PowerHouse Data Centers	450	2027.06
TACC Horizon phase 1	USA	University of Texas at Austin	18	2027.06
OpenAI/Microsoft Atlanta Phase 1	USA	OpenAI, Microsoft	105	2027.06
울산 데이터센터 Phase 1	South Korea	SK and Amazon	41	2027.06
Brazil Scala AI City Phase 1	Brazil	Scala Data Centers	54	2027.06
EU AI Gigafactory #4	미정	EU	100	2027.06
EU AI Gigafactory #5	미정	EU	100	2027.06
Applied Digital Ellendale Possible Phase 3	USA	Applied Digital	150	2027.06
Coreweave EcoDataCenter Phase 1	Sweden	CoreWeave	240	2027.06
CoreWeave Denton GB200s OpenAI/Microsoft	USA	CoreWeave	260	2027.06
Cinco data center campus	USA	Rowan Digital Infrastructure	300	2027.12
Reliance Industries Supercomputer	India	Reliance Industries	1,000	2027.12
Wonder Valley Datacenter Phase 1	Canada	-	1,400	2027.12
Meta Hyperion Phase 1	USA	Meta AI	1,500	2027.12
Microsoft Sweden Gävle	Sweden	Microsoft	미공개	2027.12

Microsoft Sweden Sandviken	Sweden	Microsoft	미공개	2027.12
Microsoft Sweden Staffanstorp	Sweden	Microsoft	미공개	2027.12
Tata Group GH200 supercomputer	India	Tata Group	미공개	2027.12
Meta Temple Texas	USA	Meta AI	미공개	2027.12
Sakura's B200s Phase 2	Japan	Sakura Internet	16	2028.03
Pennsylvania Digital I (PAX) project Phase 2	USA	PowerHouse Data Centers	450	2028.03
OpenAI/Microsoft Atlanta Phase 2	USA	OpenAI, Microsoft	110	2028.06
Sesterce Southern France 250MW	France	Sesterce	250	2028.09
DataVolt Neom 1.5 GW Phase 1	Saudi Arabia	DataVolt	300	2028.12
Sesterce Grand Est France A	France	Sesterce	300	2028.12
Sesterce Grand Est France B	France	Sesterce	300	2028.12
together.ai project Phase 3	Europe	TogetherAI, Hypertec	800	2028.12
Fluidstack France Gigawatt Campus	France	–	1,000	2028.12
South Korea Planned 3GW Cluster	South Korea	SFR	3,000	2028.12
Pennsylvania Digital I (PAX) project Phase 3	USA	PowerHouse Data Centers	450	2029.06
OpenAI/Microsoft Atlanta Phase 3	USA	OpenAI, Microsoft	110	2029.06
울산 데이터센터 Phase 2	South Korea	SK and Amazon	62	2029.12
Stargate UAE Phase 2	UAE	Stargate (OpenAI)	800	2029.12
Stargate OpenAI Oracle OCI Supercluster Phase	USA	Oracle	4,500	2029.12
Stargate Project 종합	USA	–	5,000	2029.12

자료: KUVIC 리서치 1팀

표 25. 2025년 이후 예정된 데이터센터 건설 프로젝트

프로젝트명	Country	Owner	전력용량(MW)	Operational Date
Viettel Project	Vietnam	Viettel	140	2030.12
HUMAIN Saudi Arabia Phase 2	Saudi Arabia	Humain	500	2030.12
Abu Dhabi UAE/USA 5GW Campus Phase 2	UAE	–	3,940	2030.12
DataVolt Neom 1.5 GW Phase 2	Saudi Arabia	DataVolt	1,200	2031.12
Brazil Scala AI City	Brazil	Scala Data Centers	4,696	2033.12
Tesla Dojo 1 Phase 1	USA	Tesla	2	Planned
Meta Hyperion Phase 2	USA	Meta AI	3,500	Planned
Project Rainier	USA	Amazon	2,250	Planned
Google Frech Data Center	France	Google	2,000	Planned
Vantage Data Centers – frontier project	USA	Vantage	1,400	Planned
Stargate OpenAI India Project	India	OpenAI, Microsoft	1,000	Planned
Google India Data center	India	Google	1,000	Planned
울산 데이터센터 Phase 3	South Korea	SK and Amazon	897	Planned
Gigapop	USA	–	540	Planned
Google Papillion Nebraska	USA	Google	250	Planned
Coreweave EcoDataCenter Phase 2	Sweden	CoreWeave	120	Planned
Telangana Yotta Hyderbad AI City Cluster Phase 2	India	Government of Telangana	50	Planned
TACC Horizon phase 2	USA	University of Texas at Austin	36	Planned
CoreWeave Hillsboro Oregon	USA	CoreWeave, Digital Realty	36	Planned
Inflection AI Cluster	USA	Inflection AI	31	Planned
NRT14 Campus	Japan	–	31	Planned
G42 Microsoft 30MW UAE Cluster A	UAE	G42, Microsoft	30	Planned
G42 Microsoft 30MW UAE Cluster B	UAE	G42, Microsoft	30	Planned
NVIDIA Israel Blackwell Supercomputer	Israel	NVIDIA	30	Planned

Sesterce Pegasus	–	Sesterce	25	Planned
Project Ceiba Phase 1	USA	Amazon, NVIDIA	23	Planned
Poolside 10k Cluster	–	Poolside	14	Planned
Saudi Data & AI Authority Sovereign AI factory	Saudi Arabia	Saudi Data and AI Authority	12	Planned
Foxconn Hon Hai Kaohsiung Supercomputer	Taiwan	Foxconn	11	Planned
SoftBank Planned B200 Superpod	Japan	Softbank	8	Planned
Voltage Park Location 5	USA	Voltage Park	6	Planned
Voltage Park Location 6	USA	Voltage Park	6	Planned
Voltage Park Texas Phase 2	USA	Voltage Park	6	Planned
Voltage Park Utah	USA	Voltage Park	6	Planned
Voltage Park Virginia	USA	Voltage Park	6	Planned
Voltage Park Washington	USA	Voltage Park	6	Planned
OneAsia OBON Clusters	Thailand	OneAsia	6	Planned
KAUST Shaheen-III	Saudi Arabia	KAUST	5	Planned
Google Council Bluff Iowa	USA	Google	미공개	Planned
Google Lancaster Ohio	USA	Google	미공개	Planned
Google Omaha Nebraska	USA	Google	미공개	Planned
Google south Columbus Ohio	USA	Google	미공개	Planned
Amazon Mexico Central Region	Mexico	Amazon	미공개	Planned
Amazon Thai Asia Central Region	Thailand	Amazon	미공개	Planned
Oracle $14B Saudi Arabia Investment	Saudi Arabia	Oracle	미공개	Planned
ParTec ELBJUWEL	Germany	HZDR	미공개	Planned
IndiaAI Mission Supercomputer	India	India	미공개	Planned
Neevcloud planned GPUs	India	NeevCloud	미공개	Planned
NHRI Taiwan Supercomputer (NHRI-1)	Taiwan	NHRI Taiwan	미공개	Planned
OTP SambaNova Supercomputer	Hungary	Hungary	미공개	Planned
S. Korea 7th national supercomputer	South Korea	Ministry of Science and ICT	미공개	Planned
SoftBank Planned GB200	Japan	Softbank	미공개	Planned
Tesla Buffalo Dojo	USA	Tesla	미공개	Planned

자료: 다올투자증권

DRAM 부문 매출추정

메모리 3사 DRAM 부문 매출추정

표 26. 삼성전자 DRAM 매출 구성 (단위: 백만원)

구분	2025	2026F	2027F
HBM	26,488,788	49,836,384	59,733,504
범용 DRAM	49,832,312	99,776,298	133,134,508
전체 DRAM	76,321,101	149,612,682	202,019,716

자료: OMDIA, KUVIC 리서치 1팀

표 27. SK하이닉스 DRAM 매출 (단위: 백만원)

구분	2025	2026F	2027F
HBM	37,255,490	57,428,352	75,857,040
범용 DRAM	37,733,026	73,408,338	120,219,734
전체 DRAM	74,988,516	130,836,690	198,106,310

자료: OMDIA, KUVIC 리서치 1팀

표 28. Micron DRAM 매출 구성 (단위: 백만원)

구분	2025	2026F	2027F
HBM	15,193,786	26,208,576	35,172,360
범용 DRAM	30,061,080	56,361,384	90,891,728
전체 DRAM	45,254,866	82,569,960	126,857,528

자료: OMDIA, KUVIC 리서치 1팀

본 추정은 2026년까지 OMDIA 데이터로 기준 수준을 고정하고, 2027년 이후는 증설 반영과 연평균 성장률, 믹스·수율·가격·환율 가정을 결합해 확장하였다. HBM은 2025년 HBM3·HBM3E·HBM4 병존 이후 2026년부터 HBM3가 종료되며 HBM3E·HBM4 중심, 2027년부턴 HBM4E가 추가되며 재편되는 흐름을 반영했다

HBM 물량(웨이퍼) 산정: OMDIA 기준치 + 2027년 외삽

HBM 웨이퍼 물량은 3사(삼성전자·SK하이닉스·마이크론) 모두 2025~2026년 구간은 OMDIA 자료를 기준치로 반영하였다. 세대 구성은 2025년에 HBM3·HBM3E·HBM4가 병존하는 것으로 처리하되, 2026년부터는 HBM3가 실제로 생산·공급에서 제외되는 구간으로 HBM3 물량을 0으로 반영하고, 2026년 물량은 HBM3E와 HBM4로만 구성하였다.

2027년 HBM 웨이퍼 물량은 2025~2026년의 OMDIA 기준치에서 출발하여, 중장기 확대 흐름을 감안하여 HBM의 2030년까지의 연평균 성장률(CAGR) 추정치인 30%를 적용한 후 산정하였다. 또한 2027년에는 세대 전환이 본격화된다는 점을 반영해 HBM4E를 추가하여 전체 DRAM 생산의 20%를 차지한다고 가정하였고, HBM3E가 줄어드는 믹스 변화를 적용하여 HBM4 : HBM3E = 2 : 1 수준으로 HBM4 비중이 우세하도록 설정하였다.

HBM 수율 가정: 선도업체 동률 + 삼성 격차 축소

HBM 최종수율은 SK하이닉스와 마이크론이 동일한 수준이라고 가정하였다. 반면 삼성전자는 선도업체 대비 수율이 낮은 구간이 존재한다고 보아 2025년까지는 두 회사 대비 20% 이상 뒤처진 상태로 설정하였고, 2026년부터는 격차가 10% 미만으로 축소되는 것으로 반영하였다. 이는 2026년 이후 삼성전자의 수율 개선이 유의미하게 진행될 것을 반영한다. 2027년의 HBM4E 수율은 세 회사 모두 동일한 것

으로 가정하였다.

전체 DRAM 비트 생산: OMDIA('26) + 증설 반영('27) + 평균 수율 개선

전체 DRAM 기가비트(Gb) 생산량은 2026년까지 OMDIA 자료를 그대로 반영하였다. 2027년부터는 OMDIA의 연도별 총량 대신, 각 사의 공장 증설분을 반영하여 전체 생산 웨이퍼(wafer start) 규모를 재구성한 뒤 이를 기반으로 총 생산능력을 산정하였다. 또한 2027년에는 공정 안정화 및 운영 효율 개선을 반영하여 평균 수율이 전년 대비 10% 개선되는 것으로 가정함으로써, 웨이퍼 기준 생산능력이 실제 비트 생산으로 연결되는 정도를 상향 조정하였다.

가격(ASP) 가정: 세대별 레벨 설정 + 2027년 정상화/타이트 시나리오

가격은 기가비트 기준 ASP(US$/Gb)로 세대별 단가 레벨을 구분하여 적용하였다.

- HBM3는 2025년 가격을 $1.3/Gb로 추정하였으며, 앞서 세대 구성 처리와 동일하게 2026년 이후에는 생산이 없으므로 물량 및 매출 반영 대상에서 제외하였다.
- HBM3E는 2025년 가격 추정치 $1.8/Gb 대비 2026년 15~20% 상승한 $2.1/Gb를 적용하였다. 2027년에는 세대 교체 진전 및 수급 완화를 반영해 $1.8/Gb로 재하락하는 정상화 시나리오를 반영하였다.
- HBM4는 2025년 가격 추정치 $2.0/Gb 대비 2026년 20% 상승한 $2.4/Gb로 가정하였다. 2027년에는 쇼티지 해소를 전제로 가격이 $2.4/Gb 수준에서 유지되는 것으로 설정하였다.
- HBM4E는 2026년 HBM4 가격 추정치 $2.4/Gb 대비 약 30% 상승한 $3.2/Gb로 가정하였다.
- 범용 DRAM은 2025년 $0.35/Gb 추정치를 기준으로, 2026년에는 범용 DRAM 쇼티지를 반영해 약 70% 상승한 $0.60/Gb를 적용하였다. 2027년에는 쇼티지가 완화되더라도 잔존 타이트함이 지속된다는 전제하에 $0.70/Gb로 추가 상향하였다.

환율 가정: 2025 실적 환율 + 2026~2027 고환율 기조

환율은 원화 환산 매출 추정의 일관성을 위해 각 년도마다 단일 평균환율을 사용하였다. 2025년은 연평균 환율 1,422원/$를 적용하였고, 2026~2027년은 고환율 기조를 반영해 1,450원/$로 상향하여 적용하였다.

DRAM 3사 생산구조: 라인·공정·HBM 비중

라인별 현황

표 29. 삼성전자 DRAM 라인별 현황 (단위: K wafers quarterly and annualy)

구분	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26F	2Q26F	3Q26F	4Q26F
Line 13	120	105	90	30	-	-	-	-	-	-	-	-
Line 15	540	525	510	495	450	465	480	480	480	480	465	450
Line 16	120	180	240	240	210	210	225	225	225	210	210	195
Line 17	210	240	270	285	285	285	285	285	285	285	285	285
Pyeongtaek	645	825	885	915	885	930	960	1,035	1,050	1,080	1,065	1,125
합계	1,635	1,875	1,995	1,965	1,830	1,890	1,950	2,025	2,040	2,055	2,025	2,055

(단위: 1Gb million units)

구분	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26F	2Q26F	3Q26F	4Q26F
12" 합계	20,522	25,588	28,510	28,407	25,548	27,214	28,470	30,658	31,676	32,341	32,217	33,053

자료: OMDIA, KUVIC 리서치 1팀

표 30. SK하이닉스 DRAM 라인별 현황 (단위: K wafers quarterly and annualy)

구분	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26F	2Q26F	3Q26F	4Q26F
M-14	480	480	480	495	495	480	465	450	450	435	420	414
WUXI-China	480	510	540	570	570	570	570	570	570	555	540	525
M-16	210	240	300	330	330	450	540	570	570	570	570	570
M-15X	-	-	-	-	-	-	-	-	-	15	75	111
합계	1,170	1,230	1,320	1,395	1,395	1,500	1,575	1,590	1,590	1,575	1,605	1,620

(단위: 1Gb million units)

구분	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26F	2Q26F	3Q26F	4Q26F
12" 합계	16,585	17,771	19,386	21,055	20,212	22,185	23,731	24,490	24,389	25,052	25,866	26,696

자료: OMDIA, KUVIC 리서치 1팀

표 31. Micron DRAM 라인별 현황 (단위: K wafers quarterly and annualy)

구분	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26F	2Q26F	3Q26F	4Q26F
Fab 4	81	81	81	81	81	81	81	81	81	81	81	81
Fab 11	249	264	264	264	264	264	264	264	264	264	264	264
Fab 15	255	264	264	264	264	264	264	264	264	264	264	264
Fab 16	291	291	291	291	291	291	291	291	291	291	291	291
합계	876	900	900	900	900	900	900	900	900	900	900	900

(단위: 1Gb million units)

구분	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26F	2Q26F	3Q26F	4Q26F
12" 합계	13,447	14,804	15,381	15,743	15,928	16,330	16,789	17,257	17,296	17,903	18,559	19,118

자료: OMDIA, KUVIC 리서치 1팀

공정별 비중 추이

표 32. 삼성전자 DRAM 공정별 비중 추이

구분	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26F	2Q26F	3Q26F	4Q26F
2znm	4%	4%	3%	2%	2%	1%	2%	1%	-	-	-	-
1xnm	7%	6%	6%	6%	6%	5%	4%	2%	1%	-	-	-
1ynm	19%	16%	13%	11%	8%	5%	2%	1%	-	-	-	-
1znm	26%	27%	28%	30%	28%	23%	21%	17%	12%	7%	4%	3%
1anm	37%	35%	33%	32%	28%	36%	32%	33%	32%	31%	26%	21%
1bnm	6%	12%	18%	20%	28%	30%	36%	41%	43%	46%	44%	43%
1cnm	-	-	-	-	-	-	3%	5%	11%	16%	25%	32%
1dnm	-	-	-	-	-	-	-	-	-	-	-	1%

자료: OMDIA, KUVIC 리서치 1팀

표 33. SK하이닉스 DRAM 공정별 비중 추이

구분	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26F	2Q26F	3Q26F	4Q26F
1y	23%	20%	19%	15%	14%	8%	5%	2%	1%	-	-	-
1z	47%	40%	28%	18%	15%	12%	9%	6%	6%	5%	4%	3%
1a	26%	31%	37%	44%	44%	49%	51%	50%	48%	43%	39%	32%
1b	4%	9%	16%	24%	27%	30%	34%	41%	41%	42%	38%	40%
1c	-	-	-	-	-	-	-	1%	4%	10%	18%	24%
1d	-	-	-	-	-	-	-	-	-	-	1%	1%
1d	-	-	-	-	-	-	-	-	-	-	-	1%

자료: OMDIA, KUVIC 리서치 1팀

표 34. Micron DRAM 공정별 비중 추이

구분	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26F	2Q26F	3Q26F	4Q26F
2znm	2%	1%	1%	1%	1%	1%	1%	1%	1%	1%	-	-
1xnm	2%	2%	2%	2%	2%	2%	2%	1%	1%	1%	1%	1%
1ynm	-	-	-	-	-	-	-	-	-	-	-	-
1znm	12%	7%	3%	2%	2%	1%	-	-	-	-	-	-
1αnm	62%	65%	64%	62%	59%	56%	48%	40%	34%	27%	20%	17%
1βnm	21%	24%	27%	31%	35%	39%	45%	50%	50%	51%	48%	43%
1γnm	2%	2%	2%	2%	2%	2%	5%	8%	14%	20%	31%	38%
1δnm	-	-	-	-	-	-	-	-	-	-	-	1%

자료: OMDIA, KUVIC 리서치 1팀

HBM 생산능력 추이

표 35. 메이커별 HBM 세대별 Wafer 생산능력 추이 (단위: K/월)

구분	2025			2026F			2025	2026F
	HBM3	HBM3e	HBM4	HBM3	HBM3e	HBM4		
삼성전자	60	85	5	-	60	130	150	190
SK하이닉스	10	140	10	-	80	120	160	200
Micron		60	5	-	40	50	65	90
합계	70	285	20	-	180	300	375	480

자료: OMDIA, KUVIC 리서치 1팀

연도별 생산량

표 36. DRAM 3사 웨이퍼 구성: 전체 대비 HBM 웨이퍼 비중 (단위: K wafers)

구분	전체 웨이퍼 장수			HBM 웨이퍼 장수			HBM 웨이퍼 비중		
	2025	2026F	2027F	2025	2026F	2027F	2025	2026F	2027F
삼성전자	7,695	8,175	8,760	1,800	2,280	3,000	23%	28%	34%
SK하이닉스	6,060	6,390	8,160	1,920	2,400	3,120	32%	38%	38%
Micron	3,600	3,600	4,500	780	1,080	1,440	22%	30%	32%
합계/평균	17,355	18,165	21,420	4,500	5,760	7,560	26%	32%	35%

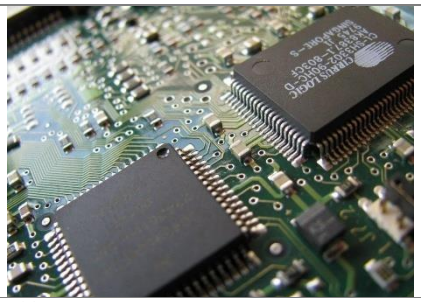
자료: OMDIA, KUVIC 리서치 1팀

표 37. 반도체 제품의 종류

사스템 반도체 (비메모리)	마이크로컴포넌츠	MPU
	아날로그IC	PMIC
	로직IC	DDI
메모리	RAM	SRAM
		DRAM
	ROM	NAND FLASH
		NOR FLASH

자료: KUVIC 리서치 1팀

그림 49. 집적회로(IC)



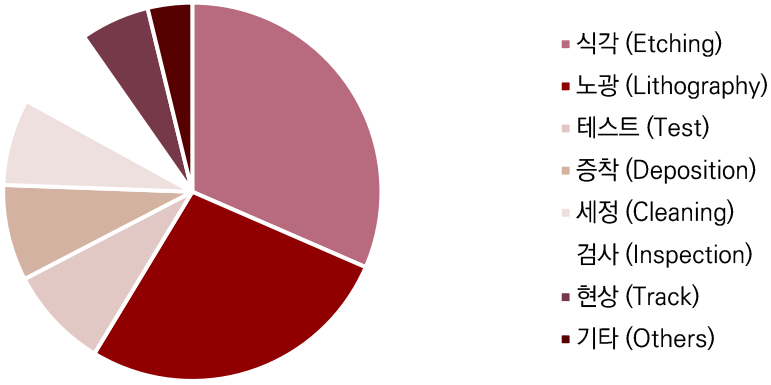
자료: 한국 전자 기술

표 38. 반도체 8공정

1. 웨이퍼 제조	실리콘을 녹여 고순도 잉곳(Ingot)을 만든 후 얇게 절단하여 원판 형태의 웨이퍼를 생산 회로 정밀도를 높이기 위해 표면을 매끄럽게 연마(CMP)하며, 현재는 12인치 웨이퍼가 주류
2. 산화 공정	웨이퍼 표면에 산화막을 형성하여 불순물로부터 회로를 보호하고 전류 누설을 막는 절연체 역할을 수행 건식 산화는 품질이 우수하고, 습식 산화는 막을 빠르게 형성함
3. 포토 공정	빛을 이용해 웨이퍼 위에 미세한 회로 패턴을 그려 넣는 가장 핵심적이고 난이도가 높은 공정 감광액 도포, 노광, 현상 단계를 거치며, 초미세 회로 구현을 위해 EUV(극자외선) 장비가 필수적
4. 식각 공정	포토 공정 후 불필요한 박막 부분을 선택적으로 제거하여 실질적인 회로 패턴을 완성 미세 공정에는 방향성이 뛰어난 건식 식각이 유리하며, 최근에는 물리·화학적 장점을 결합한 RIE 방식이 주로 쓰임
5. 증착 및 이온 주입	증착은 얇은 박막을 입혀 전기적 특성과 절연층을 형성하는 과정으로, 원자 단위 조절이 가능한 ALD 기술이 각광받음 이온 주입은 부도체인 웨이퍼에 불순물을 넣어 전기가 통하는 반도체 성질을 부여함
6. 금속 배선	형성된 소자들을 금속 선으로 연결하여 전원과 신호가 흐를 수 있는 통로를 구축하는 단계 알루미늄이나 구리 등의 소재를 사용하며, 배선 설계 방식에 따라 CPU나 GPU 등 반도체의 용도가 결정됨
7. 테스트 공정	웨이퍼 상태에서 개별 칩의 품질을 확인하는 EDS 테스트와 패키징 후의 최종 검사를 포함 프로브 카드와 테스트 소켓을 사용하여 전기적 동작 여부를 확인하고 불량품을 선별
8. 패키징 공정	칩을 보호하고 외부 기기와 신호를 주고받을 수 있도록 규격화된 형태로 조립하는 과정 웨이퍼를 얇게 갈고 낱개로 잘라 기판과 연결하며, 최근에는 고성능 구현을 위한 FC-BGA 기술 등이 중요해지고 있음

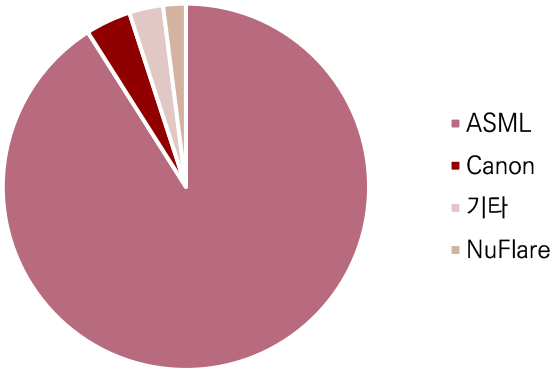
공정별 점유율

그림 50. 글로벌 장비 시장 (공정별) 점유율



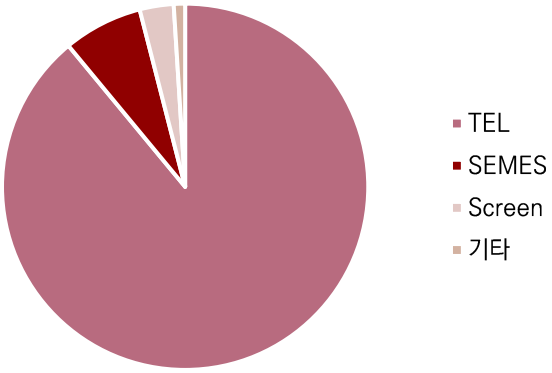
자료: KUVIC 리서치 1팀

그림 51. 노광 장비 점유율



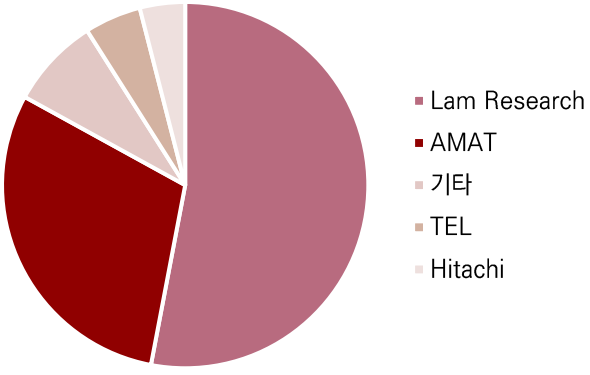
자료: KUVIC 리서치 1팀

그림 52. 현상 장비 점유율



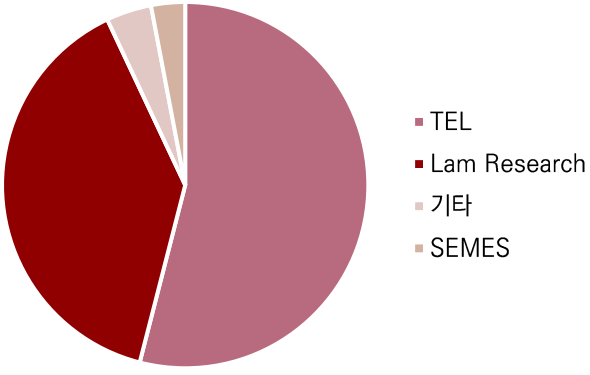
자료: KUVIC 리서치 1팀

그림 53. Conductor(전도체) 식각 장비 점유율



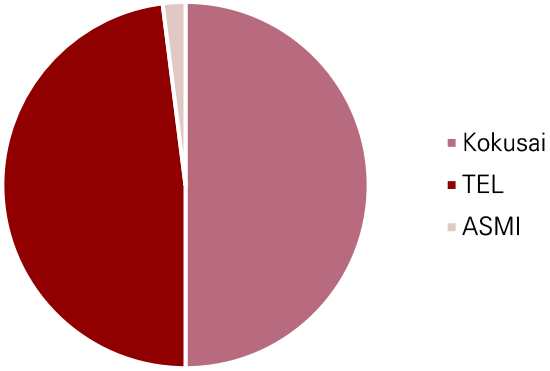
자료: KUVIC 리서치 1팀

그림 54. Dielectric(유전체) 식각 장비 점유율



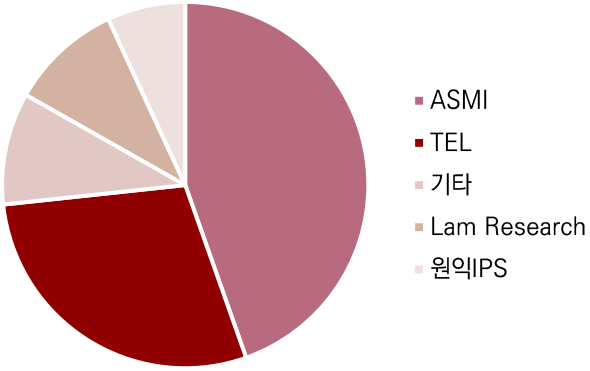
자료: KUVIC 리서치 1팀

그림 55. CVD 증착 장비 점유율



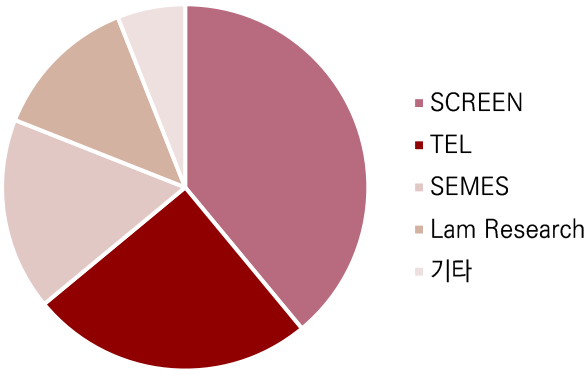
자료: KUVIC 리서치 1팀

그림 56. ALD 증착 장비 점유율



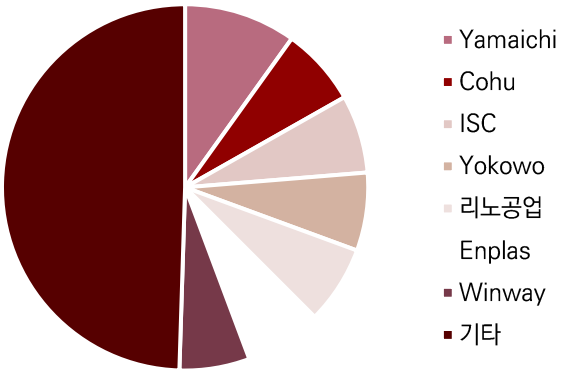
자료: KUVIC 리서치 1팀

그림 57. 세정 장비 점유율



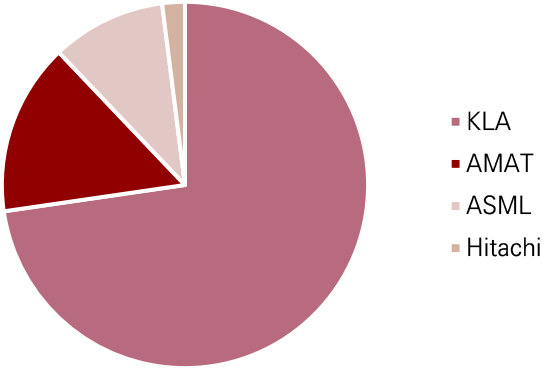
자료: KUVIC 리서치 1팀

그림 58. 소켓 점유율



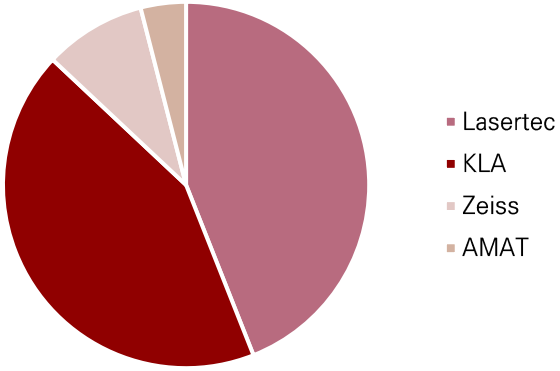
자료: KUVIC 리서치 1팀

그림 59. 검사 장비 점유율



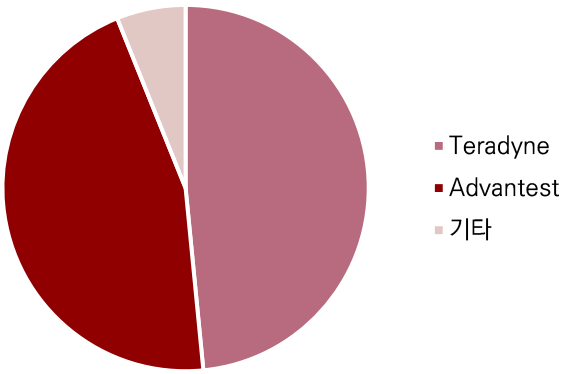
자료: KUVIC 리서치 1팀

그림 60. MASK 검사 장비 점유율



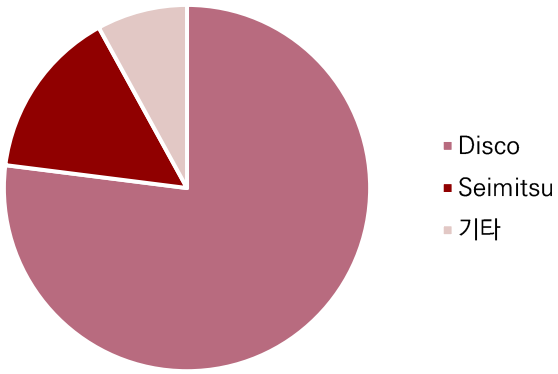
자료: KUVIC 리서치 1팀

그림 61. 테스트 장비 점유율



자료: KUVIC 리서치 1팀

그림 62. 절단 및 그라인더 장비 점유율



자료: KUVIC 리서치 1팀

표 39. 한국 반도체 밸류체인

설계 및 디자인 (Design & IP)	
팹리스 (Fabless)	어보브반도체, 제주반도체, 텔레칩스
디자인하우스 (DSP)	에이직랜드, 가온칩스, 에이디테크놀로지
IP (지식재산권)	오픈엣지테크놀로지, 칩스앤미디어, 퀄리타스반도체
웨이퍼 및 기판 (Wafer & Substrate)	
웨이퍼 제조	SK실트론(웨이퍼), 유니온머티리얼(SiC웨이퍼), 티씨케이(화합물웨이퍼), 하나머티리얼즈(잉곳), OCI(폴리실리콘)
PCB (인쇄회로기판)	비에이치, 대덕전자, 심텍, 해성디에스, 코리아써키트, 인터플렉스, 이수페타시스, 삼성전기, LG이노텍
기판 검사	인텍플러스, 고영, 기가비스, 펨트론
산화 (Oxidation)	
장비	피에스케이(Dry Strip), AP시스템, 원익IPS(RTP)
포토 (Photo)	
PR/원재료	동진쎄미켐, 이엔에프테크놀로지, 와이씨켐, SK머티리얼즈, 경인양행, 미원상사, 켐트로스, 송원산업, 엘티씨, 램테크놀러지
마스크/펠리클	에스앤에스텍(블랭크마스크/펠리클), SK엔펄스, 에프에스티(펠리클)
장비/기타	세메스(Coater), 파크시스템스(EUV 리페어/원자현미경), 오로스테크놀로지(오버레이), 엘티씨, 램테크놀러지(박리액)
식각 (Etch)	
장비 (Etcher)	테스, 피에스케이, 에이피티씨, 세메스
소재 (가스/액)	후성, 원익머트리얼즈, SK머티리얼즈(특수가스), 솔브레인, 이엔에프테크놀로지, 동진쎄미켐, 램테크놀러지
부품 (Ring/Quartz)	티씨케이, 하나머티리얼즈, 월덱스, 케이엔제이, 원익QnC, 비씨엔씨
세정 (Cleaning)	
장비/부품	제우스, 케이씨텍, 세메스, 피에스케이(Dry Cleaning)
소재/코팅	한솔케미칼, SK머티리얼즈, 원익머트리얼즈, OCI(과산화수소), 미코(코미코), 아이원스, 원익QnC, 포인트엔지니어링, 엘티씨
박막/증착 (Thin Film/Deposition)	
장비 (CVD/ALD)	유진테크, 주성엔지니어링, 원익IPS, 테스
전구체 (Precursor)	솔브레인, 한솔케미칼, 원익머트리얼즈, 레이크머트리얼즈, 원익QnC, TEMC CNS, 동진쎄미켐, 와이씨켐, 제이아이테크, 메카로, 디엔에프, 덕산테코피아, SK머티리얼즈, 후성
부품/열처리	원익QnC(Quartz), 이오테크닉스(Annealing), AP시스템, HPSP(고압수소어닐링), 미코(Heater/ESC), SK엔펄스, 월덱스
연마(CMP)	케이씨텍(장비/슬러리), 솔브레인, 와이씨켐, 이엔에프테크놀로지, 동진쎄미켐, SKC, 나노신소재(슬러리), 에프엔에스테크, 시노렉스(패드/필터)
금속배선 (Metalization)	
특수가스	SK머티리얼즈, 원익머트리얼즈(WF6)
테스트 (Testing)	
소켓/프로브카드	티에스이, 리노공업, ISC, 티에프이, 오킨스전자, 마이크로컨텍솔, 샘씨엔에스, 프로텍, 코리아인스트루먼트
테스터/핸들러	테크윙, 제너셈, 제이티, 미래산업, 유니테스트, 엑시콘, YC, 디아이
패키징 (Packaging)	
장비 (Bonder/Dicing)	이오테크닉스(Dicing/Marking), 한미반도체(Bonder/Vision), 프로텍, 탑엔지니어링, 에스티(가압장비), 제너셈(마킹), 레이저쎌(Bonding)
리플로우/기타	피에스케이홀딩스, 에스티아이, 레이저쎌, 프로텍
소재 (Ball/Wire)	덕산하이메탈, 엠케이전자, 휘닉스소재(솔더볼), 해성디에스(리드프레임), 엠케이전자(본딩와이어)
검사 (Inspection)	
Vision/TSV/Pkg	한미반도체(Vision Placement), 오로스테크놀로지(TSV Bump), 인텍플러스, 펨트론, 고영, 파크시스템스, 넥스틴, HB솔루션, 에프에스티(오로스테크놀로지), 유니테스트, 엑시콘, 네오셈
OSAT	
패키징 및 테스트	하나마이크론, SFA반도체, 네패스(네패스아크), 엘비세미콘(엘비루셈), 윈팩, 영풍(시그네틱스), 두산테스나, 에이팩트
인프라 및 설비	
진공 (Vacuum)	엘오티베큠
이송/자동화	싸이맥스(웨이퍼이송), 에스에프에이, 로체시스템즈, 라온테크
CCSS (중앙공급장치)	에스티아이, TEMC CNS, 한양이엔지, 씨엔지하이테크, 원익홀딩스(가스배관)
칠러/스크러버/필터	유니셈, GST, 에프에스티, 지앤비에스에코, 젬백스앤카엘(필터)
클린룸	신성이엔지, 한양이엔지, 성도이엔지
기타	
소재/장비	한미반도체(EMI Shield), 서플러스글로벌, 러셀, 기가레인(중고장비)

자료: KUVIC 리서치 1팀

COMPANY ANALYSIS

–

Not Rated

삼성전자 (005930)

왕의 귀환

투자포인트 1. 숨은 주역 범용 DRAM의 점유율 깡패

2026~2027년 메모리 수급의 핵심은 HBM이 아닌 **범용 DRAM**이며, AI 추론 확산과 일반 서버 수요 회복이 맞물리며 범용 DRAM 수요는 구조적으로 증가하는 반면 공급은 공정 전환과 증설 제약으로 빠르게 확대되기 어렵다. 이로 인해 대규모 쇼티지가 장기화될 가능성이 높은 가운데, **글로벌 DRAM 생산 비중 1위**인 동사는 압도적인 **생산능력과 선단 공정 양산 역량, 폭넓은 고객 커버리지**를 바탕으로 물량 대응력과 가격 협상력에서 독보적인 우위를 확보하고 있으며, **범용 DRAM 수급 불안이 심화될수록 가장 큰 구조적 수혜가 기대된다.**

투자포인트 2. 삼성 파운드리까지 기회

TSMC의 HPC 중심 수요 급증으로 3nm 공정이 풀캐파 도달, 2026년 말까지 3nm 월 20만 장·2nm 월 10만 장 증설에도 애플·엔비디아 등 주요 고객 우선 배정으로 신규 팹리스의 물량 확보가 구조적으로 제약되며 **초과 수요가 삼성 파운드리로 이전되는 낙수효과가 본격화될 전망**이다.

삼성전자는 2nm 공정 수율을 50~60% 수준으로 개선해 2025년 12월 엑시노스 2600 양산에 돌입했으며, 테슬라 AI6 칩 수주(약 23조 원)를 통해 GAA 기반 2nm 공정의 빅테크 신뢰성을 검증받았다. 경쟁사 인텔은 18A 공정(수율 60%)에서 CPU 생산에 국한되어 엔비디아 GPU 테스트 중단 등 대형 AI 칩 양산 경험 부재로 외부 고객 유치에 난항을 겪는 반면, **삼성전자는 과거 8nm 엔비디아 GPU 및 AI 가속기 양산 노하우를 보유해 신뢰성 우위를 확보**하고 있다.

투자포인트 3. HBM4 및 CXL 경쟁력

HBM4 진입은 설계·공정·패키징을 포함한 종합 기술력이 동일 선상에서 재검증되는 국면이다. 이 과정에서 동사는 **1c DRAM 공정과 미세 공정 기반의 기술적 준비도가 높으며 메모리-로직-공정을 함께 가져갈 수 있다는 점**에서 HBM 시장에서 왕위를 탈환할 기회로 작용한다. 동사는 또한 **딥시크-엔그램으로 수혜가 예상되는 CXL 기술을 선도**하는 등 기술적 선도가 높아 HBM4가 도입되는 시기와 맞물려 집중적 수혜를 받을 것으로 전망된다.

Earnings and valuation metrics

결산기 (12월)	2020	2021	2022	2023	2024
매출액 (십억원)	236,807	279,605	302,231	258,936	300,871
YoY (%)		18%	8%	-14%	16%
영업이익 (십억원)	35,994	51,634	43,377	6,567	32,726
YoY (%)		43%	-16%	-85%	398%
영업이익률 (%)		18%	14%	3%	11%
당기순이익 (십억원)	26,408	39,908	55,654	15,487	34,451
EPS (원)	3,841	5,777	8,057	2,131	4,950
P/E (배)	21.09	13.55	6.86	36.84	10.75

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

Stock Information

시가총액	9,911,949억원
발행 주식 수	591,964만주
유동주식비율	75.32%
52주 최고가	152,300원
52주 최저가	51,000원
외국인 지분율	51.89%
KOSPI	4949.59
KOSDAQ	1064.41

Price Trend

KUVIC Research Team N

메일 kuvic_korea@naver.com

팀장 43기 Senior 정상엽
부팀장 44기 Senior 김단비
팀원 44기 Senior 김병찬
팀원 44기 Senior 김현진
팀원 44기 Senior 박한동
팀원 44기 Senior 시민규
팀원 44기 Senior 정다연

Who We Are

Compliance Notice

- 본 보고서는 고려대학교 가치투자동아리 KUVIC의 리서치 결과를 토대로 한 분석 보고서입니다.
- 본 보고서에 사용된 자료들은 고려대학교 가치투자동아리 KUVIC이 신뢰할 수 있는 출처 및 정보로부터 얻어진 것이나 그 정확성이나 완전성을 보장하지 못합니다.
- 본 보고서는 투자 권유 목적으로 작성된 것이 아닌 고려대학교 가치투자동아리 KUVIC의 스터디 목적으로 작성되었습니다.
- 따라서 투자자 자신의 판단과 책임 하에 종목선택이나 투자시기에 대한 최종 결정을 하시기 바랍니다.