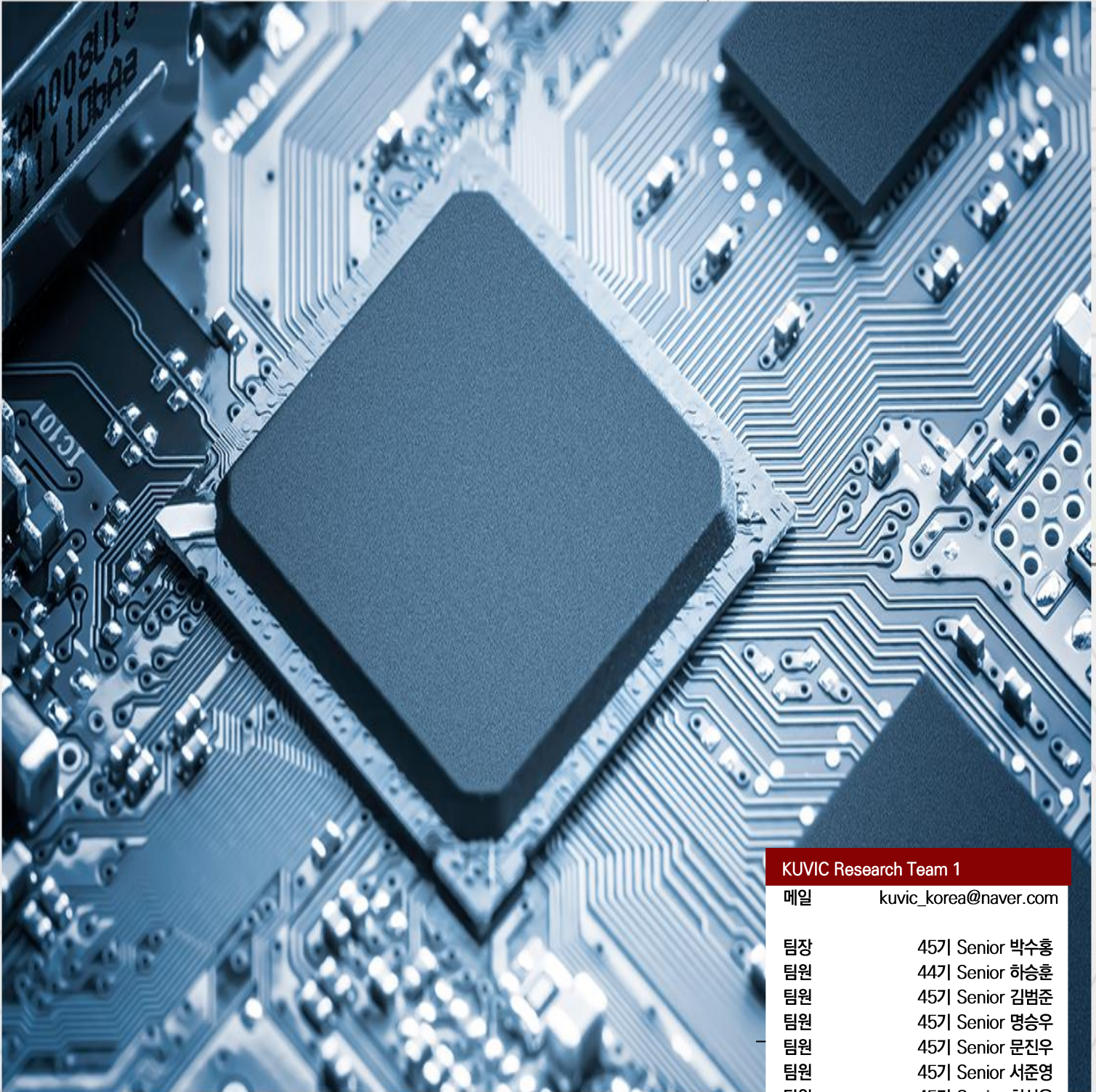


Industry Indepth 2026.07.30

[반도체와 반도체 장비] (비중확대)

간장공전공정정



KUVIC Research Team 1

메일 kuvic_korea@naver.com

팀장	45기 Senior 박수홍
팀원	44기 Senior 하승훈
팀원	45기 Senior 김범준
팀원	45기 Senior 명승우
팀원	45기 Senior 문진우
팀원	45기 Senior 서준영
팀원	45기 Senior 최서윤

CONTENTS

Summary	3
Key Chart	5
AI 수요	8
AI 컴퓨팅 수요	8
AI CAPEX 전망	11
AI 트렌드 분석	15
엔비디아와 CPU, LDPPR	15
HBM	18
NAND	19
메모리 가격추이	22
메모리 플레이어 분석	24
SK 하이닉스	24
삼성전자	25
마이크론	27
CXMT	28
DRAM, NAND 쇼티지 분석	30
DRAM 수요 및 공급 추정	30
NAND 수요 및 공급 추정	39
전공정	43
전공정이 핵심이다	43
5대 핵심 공정	45
TAM 추정	51
Company Analysis	62
ASML	

Summary

AI 수요

① AI의 무게중심이 학습에서 추론·에이전틱으로 이동하며 병목이 GPU 연산에서 메모리 계층 전반으로 확장, 이 구조적 전환이 HBM을 넘어 서버 DRAM과 NAND까지 아우르는 메모리 수요 팽창을 이끔

- 5대 하이퍼스케일러의 CapEx는 26년 \$808.3B(+83%YoY), 27년 \$1,031.8B(+28%YoY)까지 확대될 전망
- 하이퍼스케일러의 26년, 27년 CapEx/OCF 전망치는 99%로 각각의 연간 영업현금흐름 자체를 거의 다 소진하는 수준에 도달
- 단순히 투자 부담의 크기만 보기보다는 이미 확보한 클라우드 계약이 얼마나 빠르게 이익으로 전환되는지를 함께 볼 필요가 있음

AI 트렌드 변화

② 에이전틱 AI 전환으로 CPU가 핵심 프로세서로 부상, 서버당 DRAM 탑재량과 LPDDR 수요를 끌어올리고, HBM은 오프로딩에도 기본 요구 용량이 구조적으로 커지며, NAND는 단순 저장을 넘어 KV Cache를 담당하는 AI 연산 인프라로 승격되는 등 메모리 3대 축 전반에서 수요 저변이 넓어지고 있음.

- 에이전틱 AI 확산으로 CPU에 대한 중요도가 상승하며 CPU:GPU 비율이 최대 1:1까지 높아질 전망
- Vera CPU 1개당 LPDDR 탑재량이 Grace 대비 3배인 1.5TB로 늘어나 2026년 LPDDR 수요를 최대 60억GB까지 밀어올릴 것으로 전망
- 결론적으로 메모리 가격은 전 제품군에서 상승세를 이어가고 있으나, 상승폭은 점차 축소되는 국면
- 추론·에이전틱 AI 확산으로 HBM은 오프로딩에도 기본 요구 용량이 구조적으로 커지고, NAND는 단순 저장에서 KV Cache 오프로딩·HBF를 담당하는 AI 연산 인프라로 승격되며 AI 비중이 2028년 51.5%까지 확대될 것으로 전망
- 향후 실적과 산업 방향성을 결정하는 변수는 추가적인 가격 인상보다 비트 출하량 확대를 위한 생산능력 증설과 공급 확대가 될 가능성이 높음
- 결국 메모리 공급사들의 증설 규모와 생산성 개선 여부가 향후 업황을 좌우할 것이며, 이에 따라 전공정 투자와 장비 발주가 메모리 산업의 핵심 변수가 될 것으로 판단

메모리 시장 플레이어

③ 선단 노드 전환과 전방위 증설로 공급을 늘리려는 메모리 3사가 HBM·범용 DRAM·낸드 전반에서 주도권 경쟁을 벌이고 있고 CXMT는 범용 DRAM 내수를 잠식하는 잠재적 위협이 될 수 있으나 기술 격차로 3사에 영향을 미칠 수준은 아니라고 판단됨.

- SK하이닉스는 1Q26 HBM 점유율 58%·2026년 생산분 전량 완판으로 주도권을 쥔 가운데, 1b→1c HBM 캐파 확대와 범용 D램 잠식 둔화를 동시에 달성 목표

- 삼성전자는 2026년 2월 업계 최초 HBM4 양산·1Q26 D램 1위 탈환으로 반등에 시동을 걸었으며, 1c 캐파를 월 6~7만→20만 장으로 끌어올리는 공세적 전환 계획이 핵심
- 마이크론의 경우 1Q26 HBM 점유율 21%(3위)·엔비디아 HBM4 15% 내외로 규모는 열위이나, 업계 최초 6세대 1γ 출하와 미·일·대만 분산 거점이 강점. 그린필드 팹 기여가 2028년 이후로 밀려 단기는 1γ 전환 속도가 실적을 좌우.
- CXMT는 2025년 매출 617.99억위안(+155%)·첫 흑자로 글로벌 4위(Q4 점유율 7.67%)에 올랐으나, 웨이퍼 점유율 13→17% 대비 비트 점유율은 9%(26E)에 그칠 것으로 예상.

DRAM, NAND 쇼티지 분석

④ DRAM은 총족률이 2026~2028년 내내 -22.0%→-20.9%로 20%를 웃도는 쇼티지 발생. NAND 역시도 -2%대로 완화됐다가 2028년 비트 개선을 둔화로 -7.8%까지 공급부족 심화

- 2026~2030년 D램 총수요는 375bn Gb에서 869bn Gb로 연평균 약 23% 성장할 것으로 전망됨
- Server DRAM이 DRAM 시장의 성장을 주도할 전망이며, 이는 서버 출하량 증가보다 AI CPU 확산에 따른 서버 1대당 DRAM 탑재량 급증에 따른 결과
- HBM 수요는 2026년 5.88GB bn에서 2028년 10.74GB bn, ASIC 비중 33.9%로 확대 전망
- HBM 공급은 2026년 5.42 GB bn에서 2028년 10.46 GB bn으로 확대. 범용 DRAM 공급은 2026년 35.73 GB bn에서 2028년 53.55 GB bn으로 전망, HBM 우선 배분에 따른 웨이퍼 잠식으로 증가폭이 제한.
- 에이전틱 AI가 견인하는 eSSD 수요 폭발로 전체 NAND 수요는 26년 1,282EB에서 28년 1,870EB로 급성장(전체 수요 내 AI 비중 33.7% → 51.5%)할 전망이다.
- 반면, NAND 공급은 '클린룸 공간 부족'과 3D 적층 전환의 공정 난도 상승으로 인해 26년 1,254EB, 28년 1,725EB 수준까지 확대되는 데 그칠 것으로 추정됨.

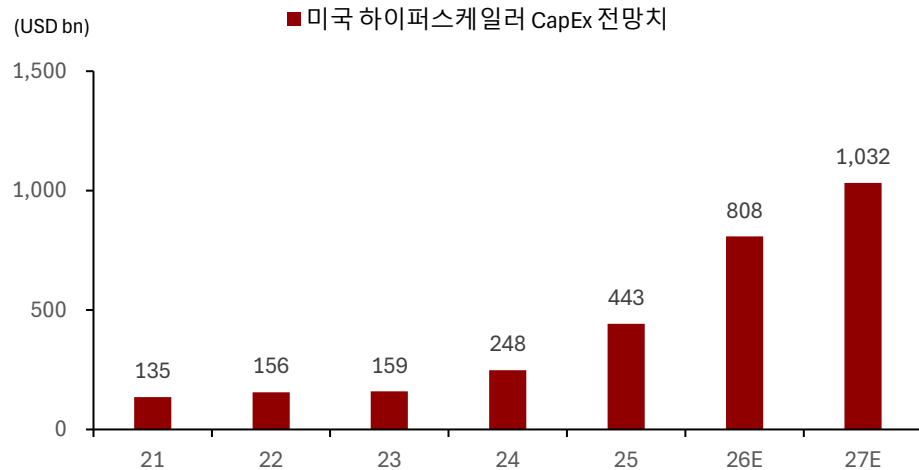
전공정

⑤ EUV 시장의 성장과 ASP 인상이 겹쳐 노광 시장이 CAGR 26.5%로 성장하고 전공정 대비 비중이 +2.5%p 상향되는 점, ASML 과점에 따른 실현 확실성·가격결정력을 종합하면 노광이 핵심 병목

- 이번 사이클의 공급 상한을 실제로 켜 것은 후공정이 아니라 전공정. TSMC가 2026년 CapEx의 70~80%를 첨단 전공정에 배정한 것도 이를 뒷받침.
- 노광은 미세화 수준에 따라 기술적 난도가 좌우되는 공정이자 미세화가 곧 진입 장벽 심화로 이어지는 구조. 표준 EUV 1.8억 유로·High-NA 3.8~4억 유로의 장비가·소재 과점까지 겹쳐 전공정 병목 형성
- 3D 전환과 EUV 레이어 증가로 전공정 장비 시장은 2026~2028년 전 공정에서 확대, 식각과 증착이 CAGR 26.6%·24.3%, 노광 역시 26.5%로 전공정 시장의 성장률을 주도.

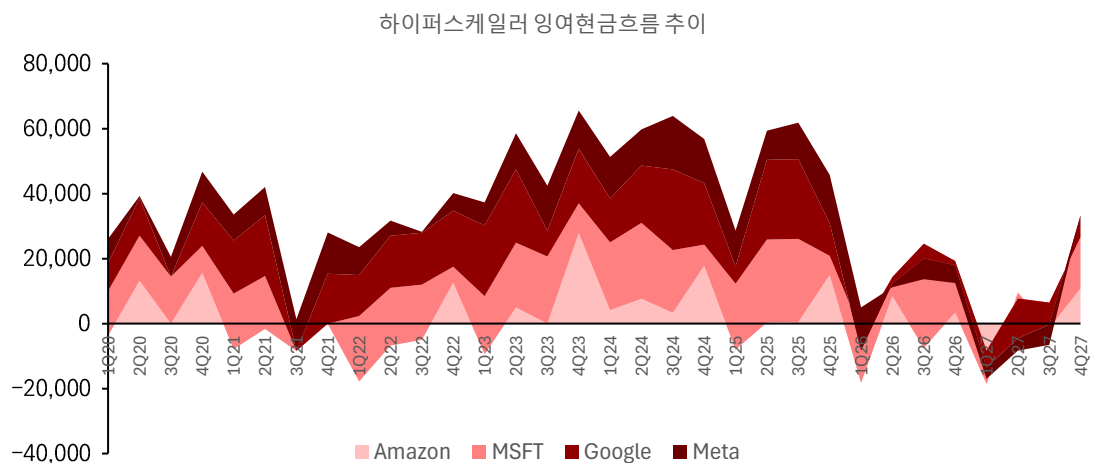
Key Chart

그림 1. 하이퍼스케일러 CapEx 전망치



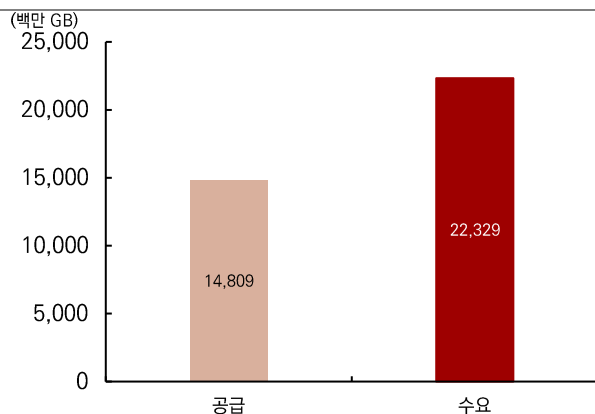
자료: Bloomberg, KUVIC 리서치 1팀

그림 2. 하이퍼스케일러 잉여현금흐름 추이



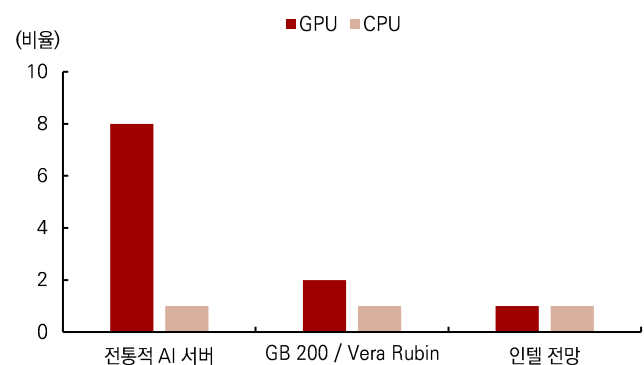
자료: Bloomberg, KUVIC 리서치 1팀

그림 3. SOCAMM으로 인해 가속화되는 LPDDR 쇼티지



자료: Trendforce, KUVIC 리서치 1팀

그림 4. CPU-GPU 비율 추이



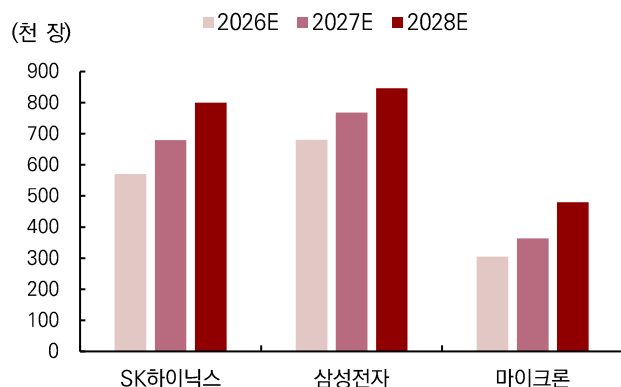
자료: 언론 종합, KUVIC 리서치 1팀

표 1. 메모리 3사 D램 증설 목표 타임라인 (~2029년, 예상 wpm)

기업	지역	공장	1H26	2H26	1H27	2H27	1H28	2H28	2029
삼성전자	평택	P4	1Q Ph3 45k	7월 Ph4 45k 11월 Ph2 45k					
		P5					팹1 오픈 약 300k		팹2 오픈 약 300k
	용인	1호 팹							10월 팹1 가동 목표
SK하이닉스	용인	Y1			2월 Ph1 60k	8월 Ph2 60k	2월 Ph3 60k	8월 Ph4 60k	2월 Ph5 60k
		Y2							오픈
	청주	M15X			4월 Ph3, 4 최대 80k				
마이크론	美 버지니아 마나사스	Fab 6	5월 1α 전환 20k						
	美 아이다호 보이시	ID1				3Q 오픈 80k			
		ID2						오픈 80k	
	美 뉴욕 클레이	-							오픈
	日 히로시마	Fab 15						신규장비 반입 45k	
	臺 통뤄	P5				4Q 오픈 50k			

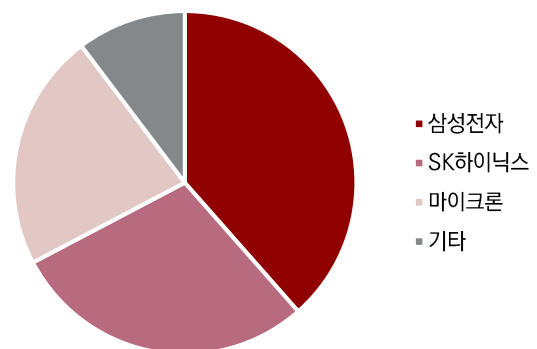
자료: 언론 종합, KUVIC 리서치 1팀

그림 5. 메모리 3사 월별 웨이퍼 투입 수 증가 추이



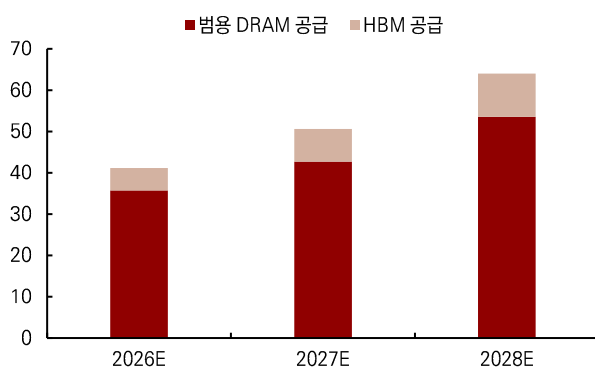
자료: KUVIC 리서치 1팀

그림 6. 1Q26 글로벌 D램 점유율



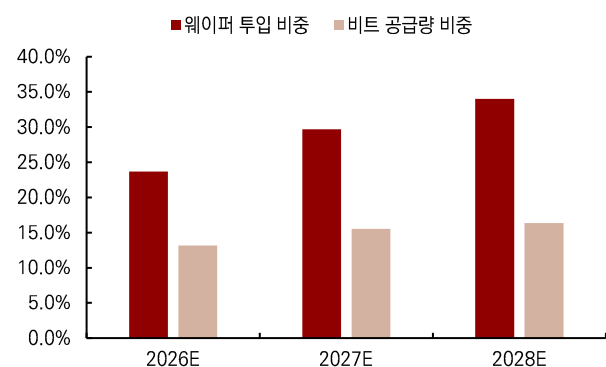
자료: TrendForce

그림 7. DRAM 공급량 (GB bn)



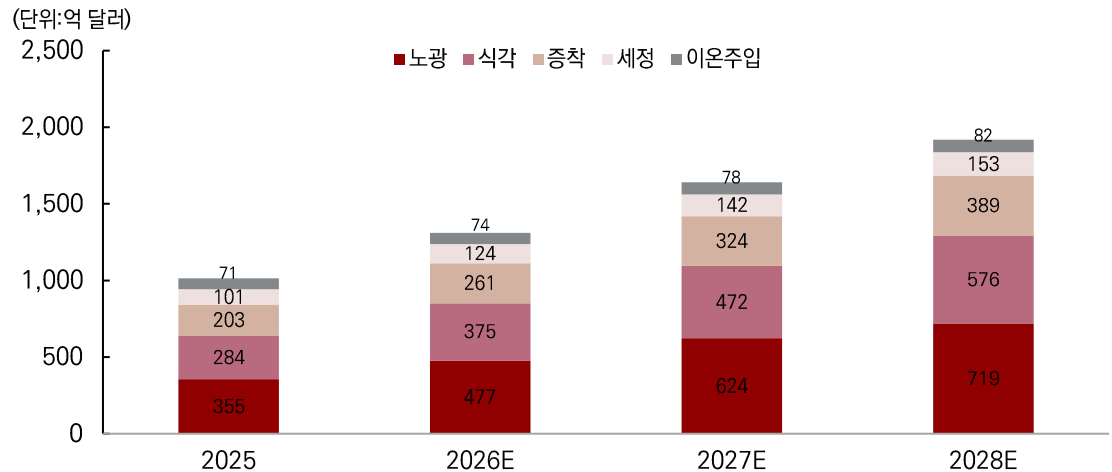
자료: KUVIC 리서치 1팀

그림 8. 웨이퍼 투입량 및 비트 공급량 내 HBM 비중



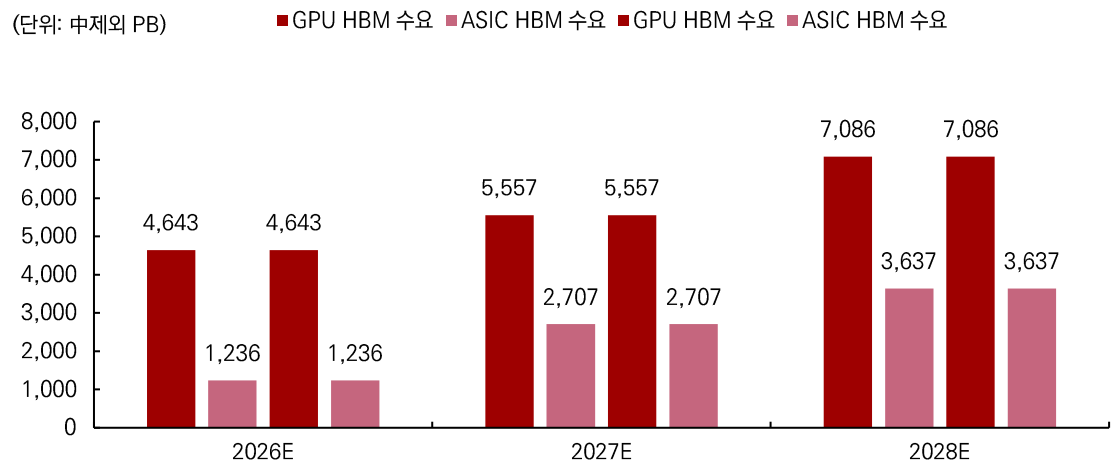
자료: KUVIC 리서치 1팀

그림 9. 전공정 공정별 시장 규모



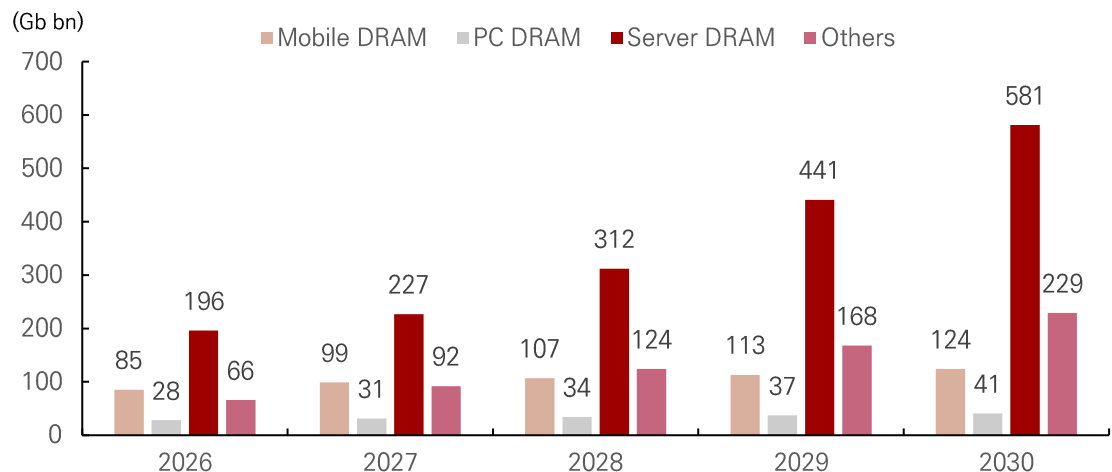
자료: SEMI, ASML, 도쿄일렉트론, KUVIC 산업분석 1팀

그림 10. GPU ASIC HBM 수요



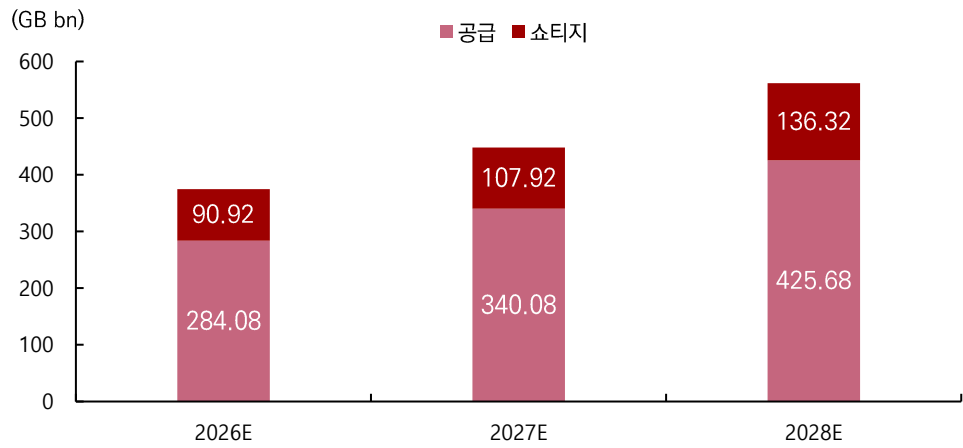
자료: Morgan Stanley, TrendForce, NVIDIA, Broadcom, Bloomberg, KUVIC 리서치 1팀

그림 11. 범용 D램 수요 전망



자료: KUVIC 리서치 1팀

그림 12. HBM 총수요 (공급 + 쇼티지)



자료: KUVIC 리서치 1팀

AI 수요

AI 컴퓨팅 수요

학습과 추론

AI 모델의 라이프사이클은 학습과 추론으로 구분

AI 모델의 라이프사이클은 **학습(Training)**과 **추론(Inference)** 단계로 구분된다. 학습은 대규모 데이터셋을 이용해 순전파(Forward Pass)와 역전파(Backward Pass)를 반복하며 모델의 파라미터를 최적화하는 과정으로, 막대한 연산량과 컴퓨팅 자원을 요구한다. 반면 추론은 학습이 완료된 모델의 파라미터를 고정한 상태에서 새로운 입력에 대해 순전파만 수행하여 결과를 생성하는 과정이다. **학습 단계는 모델의 성능을 결정하는 데 초점을 두는 반면, 추론 단계는 응답 지연시간(Latency), 처리량(Throughput), 비용 효율성이 핵심 성능 지표로 작용한다.**

표 2. 학습과 추론의 차이

	학습	추론
목적	정답을 바탕으로 가중치를 조정해 모델을 훈련	학습된 가중치로 새로운 입력의 결과를 예측
연산 과정	순전파 + 손실 계산 + 역전파	순전파
연산량	수천 개의 GPU 병렬 필요	엣지단 · 모바일에서도 가능
출력	모델 업데이트	예측값 출력
핵심 이슈	학습 속도, 에너지 효율, 데이터 품질	지연시간(Latency), 비용 절감

자료: KUVIC 리서치 1팀

무게중심은 학습에서 추론으로

최근 AI 산업의 무게중심은 학습 중심에서 추론 중심으로 이동하고 있다. 이러한 변화의 배경을 이해하기 위해서는 AI 모델의 성능 향상 방식이 시대별로 어떻게 발전해왔는지, 그리고 이에 따라 AI 워크로드와 요구되는 컴퓨팅 인프라가 어떻게 변화해왔는지를 살펴볼 필요가 있다.

AI 모델의 발전 과정

네 가지 스케일링 법칙: Pre-training Scaling, Post-training Scaling, Test-time Scaling, Agentic Scaling

그동안 AI 모델의 성능 향상은 **Pre-training Scaling, Post-training Scaling, Test-time Scaling의 세 가지 스케일링이 주도**해왔다. 각 단계는 모델의 성능뿐 아니라 AI 워크로드와 컴퓨팅 자원 수요를 함께 변화시키며 GPU, 메모리, 네트워크 등 AI 하드웨어 생태계의 발전을 이끌었다. 최근 엔비디아는 여기에 **Agentic Scaling**을 새로운 스케일링 법칙으로 제시하며 AI의 발전 방향을 한 단계 더 확장해 나가고 있다.

표 3. 네 가지 스케일링 법칙

	Pre-training Scaling	Post-training Scaling	Test-time Scaling	Agentic Scaling
핵심 목표	대규모 데이터 학습을 통해 기본 지식·추론 능력 확보	모델의 활용성 향상	문제 해결 능력 향상	복잡한 작업 수행
대표 모델	GPT-3	ChatGPT	OpenAI o1	Claude code, Codex
핵심 연산 구조	대규모 병렬 사전학습	사후학습	Chain-of-Thought (CoT)	추론 시점의 동적 연산, 다중 에이전트 협업(MAS), 반복적 피드백 루프
주요 병목	GPU, HBM, 네트워크	GPU, HBM, 네트워크	메모리 용량, 메모리 계층의 효율성	CPU, 메모리, 스토리지, 네트워크

자료: KUVIC 리서치 1팀

먼저 **Pre-training Scaling**은 모델 규모(파라미터 수), 학습 데이터, 컴퓨팅 자원을 확대해 AI 모델의 성능을 향상시키는 단계이다. 다양한 텍스트와 코드를 하나의 모델에 사전학습 하면서 특정 과제에만 사용되는 모델이 아닌 여러 영역에 재사용할 수 있는 **Foundation Model**이 본격적으로 발전했다. **모델 규모가 커질수록 단일 GPU의 성능보다 수천~수만 개 GPU를 연결한 대규모 AI 클러스터 구축이 핵심**

경쟁력으로 자리 잡았다. 이에 따라 GPU뿐만 아니라 HBM, 첨단 파운드리, 그리고 GPU 간 고속 통신을 담당하는 Scale-up · Scale-out 네트워크의 중요성이 함께 높아졌다. **결국 Pre-training Scaling은 AI 인프라 투자의 중심을 개별 반도체에서 데이터센터 단위의 통합 컴퓨팅 시스템으로 확장시키며** 현재 AI 하드웨어 생태계의 기반을 마련했다.

Post-training Scaling은 사전학습을 마친 Foundation Model을 사용자의 지시와 선호에 맞게 정렬해 실제 활용성을 높이는 후속학습 단계이다. 사전학습만으로는 지식과 언어 능력은 갖추더라도 사용자의 의도를 정확히 반영하는 데 한계가 있었지만, Instruction Tuning · RLHF · DPO 등의 기법이 적용되면서 답변의 품질과 신뢰성이 크게 향상됐다. GPT-3.5 기반 Chat GPT가 대표적인 사례로, 사후학습을 통해 범용 AI 서비스로 발전하며 생성형 AI의 대중화를 이끌었다. 생성형 AI의 대중화는 AI 서비스 이용자를 급격히 늘렸고, 이에 전 세계 사용자들의 요청을 동시에 처리하기 위해 동일한 모델을 다수의 GPU 서버에 배포하는 Serving Replication이 확대되면서 AI 데이터센터 증설과 GPU, HBM, 네트워크, 전력, 냉각 인프라 수요가 함께 증가했다.

Test-time Scaling은 추론 과정에서 더 많은 연산을 수행하고 더 긴 추론 과정을 거쳐 성능을 높이는 단계이다. OpenAI o1, DeepSeek R1 등이 대표적인 사례로, 사후학습된 모델에 추가 추론 연산을 배분해 더 높은 성능을 달성하였다. 이러한 접근은 **AI 성능 향상의 중심을 학습 단계에서 추론 단계까지 확장시켰으며, 사용자 수나 요청 횟수와 무관하게 요청당 추론 연산량 자체를 증가시켜 컴퓨팅 수요를 구조적으로 확대시켰다.** 또한 추론 과정이 길어질수록 토큰 생성량과 KV Cache 사용량이 크게 증가하면서 GPU 연산 성능보다 메모리 용량과 메모리 계층의 효율적인 활용이 추론 성능과 효율을 좌우하는 핵심 요소로 부상했다. 이러한 메모리 병목을 완화하기 위해 MoE(Mixture of Experts), Prefill-Decode 분리, Batching 및 KV Cache Management 등 동일한 비용과 지연시간에서 더 높은 추론 성능을 구현하기 위한 시스템 기술이 빠르게 확산되고 있다.

Agentic Scaling은 AI가 단순히 답변을 생성하는 수준을 넘어, 사용자의 목표를 이해하고 계획·실행·검증을 반복하며 다른 Agent와 협력해 실제 작업을 수행하도록 성능을 확장하는 단계이다. 이러한 변화로 단일 질의응답보다 훨씬 긴 실행 과정이 요구되면서 추론 연산량이 구조적으로 증가하고 있다. 하나의 작업이 여러 단계의 모델 호출과 외부 도구 및 시스템 연동으로 구성되고, Multi-Agent 환경에서는 여러 Agent가 동시에 동작해 추론 인프라의 활용도가 더욱 높아진다. 동시에, 장시간 작업 과정에서 작업 이력과 실행 결과를 지속적으로 유지해야 하므로 **Context와 KV Cache의 저장 및 재사용 중요성이 커지면서 메모리와 스토리지의 역할도 확대되고 있다.** 이와 함께 CPU 역시 GPU의 추론을 보조하는 역할을 넘어 에이전트 조율, Tool 호출, Workflow 제어 등 Agentic AI 시스템의 제어 계층을 담당하는 핵심 프로세서로 자리매김하면서 AI 시스템 내 전략적 중요성이 한층 높아지고 있다.

추론의 급격한 성장, 메모리의 부상

기존 하드웨어 핵심 병목은 GPU와 HBM

앞서 설명했듯이, 대규모 사전학습을 통해 모델 성능을 향상시켜온 시기에는 **GPU의 연산 성능과 HBM의 메모리 대역폭이 핵심 병목으로 부상했다.** 사전학습은 초거대 모델의 행렬곱과 역전파 연산을 수천 개 GPU에서 병렬 수행하는 과정으로, 연산량(FLOPs)뿐 아니라 가중치와 활성값을 실시간으로 공급하는 메모리 대역폭이 시스템 성능을 좌우한다. GPU는 Tensor Core 기반의 대규모 병렬 연산을 담당하고, HBM은 TSV 기반 3D 적층 구조와 초광폭 인터페이스를 통해 초고속 데이터 전송을 지원함으로써 메모리 병목을 최소화하고 GPU의 연산 활용률을 극대화하는 역할을 수행했다.

병목은 단순한 연산 성능에서 메모리 계층(Memory Hierarchy)의 효율성으로 확장

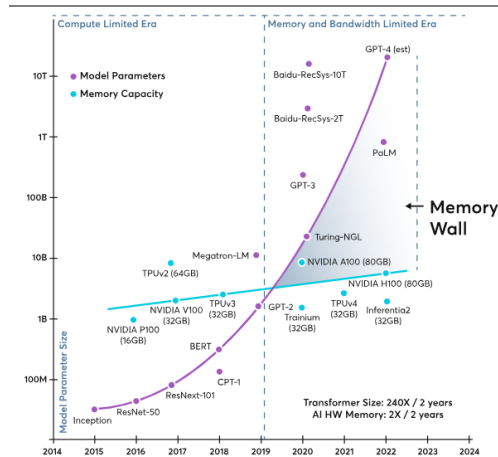
반면 AI 활용이 학습에서 추론 중심으로 이동하면서 병목은 **단순한 연산 성능에서 메모리 계층(Memory Hierarchy)의 효율성으로 확장되고 있다.** 추론에서는 단순한 연산량(FLOPs)보다 초당 생성 가능한 토큰 수와 토큰당 비용이 핵심 경쟁력이기 때문이다. 특히 긴 컨텍스트와 에이전틱 AI 확산으로 모델이 유지해야 하는 KV Cache가 급증하면서 모든 데이터를 HBM에 저장하는 것은 비용과 용량 측면에서 비효율적이게 되었다. 이에 연산이 집중되는 디코딩(Decode) 단계에서 자주 참조되는 활성 KV Cache는 HBM에 유지하고, 프리필 단계에서 생성된 KV Cache와 접근 빈도가 낮은 KV Cache는 시스템 DRAM과 SSD로 계층화하여 저장하는 구조가 확산되고 있다.

범용 DRAM과 NAND 역할의 확대

이러한 변화는 범용 DRAM과 NAND의 역할을 확대시키고 있다. 시스템 DRAM은 HBM과 SSD 사이의 중간 계층에서 대용량 KV Cache를 저장·관리하고, SSD는 장기 문맥과 RAG 데이터베이스 등 방대한 데이터를 저장하는 역할을 담당한다. 실제로 엔비디아는 올해 1월 ICMS(Inference Context Memory Storage) 아키텍처를 발표하며 루빈 플랫폼에서 HBM은 연산에 집중시키고, 대규모 문맥 데이터는 낸드 기반 SSD로 오프로딩하는 구조를 제시했다. 또한 16-Plane NAND를 적용해 병렬 입출력 성능을 높임으로써 대규모 추론 환경에서도 SSD를 효과적으로 활용할 수 있도록 설계했다.

결과적으로 AI 인프라의 핵심 메모리는 HBM에 국한되지 않는다. 추론 중심의 AI 시대에는 HBM이 연산 성능을 담당하는 반면, 범용 DRAM과 NAND는 KV Cache와 대용량 문맥 데이터를 저장·관리하는 메모리 계층으로 자리잡으며 구조적인 수요 확대가 기대된다.

그림 10. Memory Wall의 대안: HBM



자료: Aiarlabs

그림 11. 메모리 계층 구조



자료: SK HYNIX

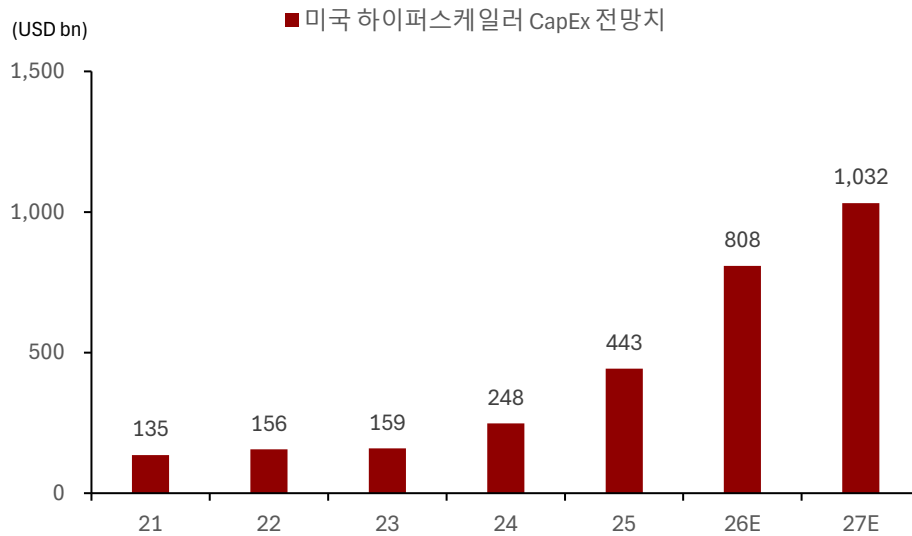
AI CapEx 전망

27년 CapEx가 1조 달러를 넘어서는 상황

하이퍼스케일러의 CapEx는 26년 \$808.3B, 27년 \$1,031.8B 전망

5대 하이퍼스케일러의 CapEx는 26년 \$808.3B(+83%YoY), 27년 \$1,031.8B(+28%YoY)까지 확대될 전망이다. 네오클라우드 4개사를 포함한 AI 기업 9개사 기준으로는 26년 \$876.5B, 27년 \$1,114.4B로 늘어날 것으로 예상되며, 이는 AI 인프라 확보가 현금흐름의 최우선 배분처로 자리 잡고 있음을 보여준다.

그림 12. 하이퍼스케일러 CapEx 전망치



자료: Bloomberg, KUVIC 리서치 1팀
주: 아마존, 구글, 마이크로소프트, 메타, 오라클의 연간 CapEx 전망치 컨센서스 합

표 4. AI 기업들 분기별 CapEx 전망

(USD Mn)	1Q26	2Q26E	3Q26E	4Q26E	1Q27E	2Q27E	3Q27E	4Q27E
Amazon	44,203	48,720	52,256	57,163	51,739	56,698	59,705	64,946
MSFT	30,876	35,208	44,969	49,321	45,729	47,554	51,236	53,106
Google	44,924	54,146	62,766	64,607	70,566	76,188	82,169	76,095
Meta	18,997	33,154	38,785	45,624	40,949	45,981	50,593	56,627
Oracle	16,493	19,641	21,961	24,483	26,592	23,123	25,106	27,098
CoreWeave	7,695	7,956	9,555	8,922	9,206	9,708	9,634	10,266
Nebius	2,473	4,883	6,778	8,690	6,994	7,469	7,626	7,642
IREN	949	1,452	2,184	2,288	2,155	2,188	2,244	2,192
Applied Digital	775	795	1,352	1,493	1,383	1,370	1,386	1,118
합계	167,385	205,954	240,607	262,592	255,313	270,278	289,699	299,091

자료: Bloomberg, KUVIC 리서치 1팀

하이퍼스케일러들의 자본지출 지속 가능성

알파벳의 CapEx 가이드선 상향과 잉여현금흐름 적자 전환

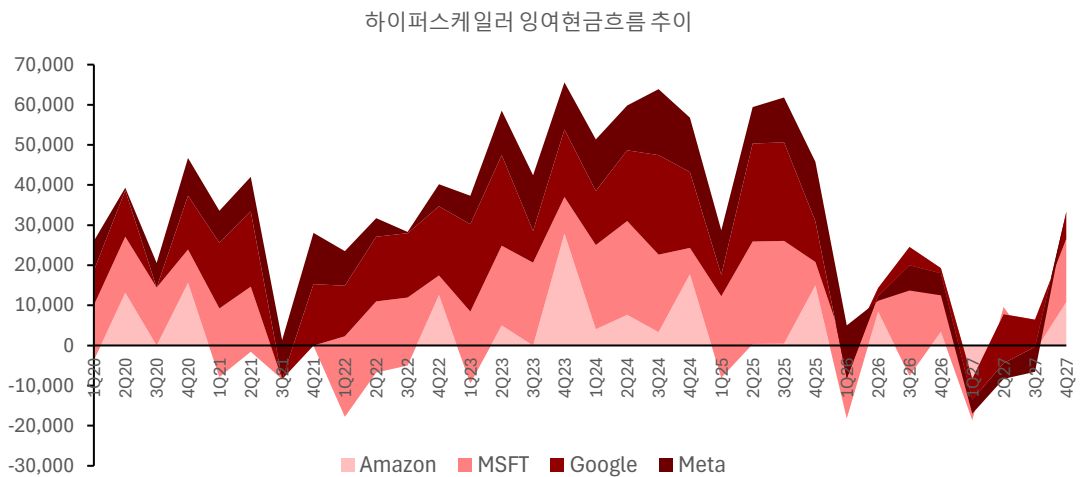
최근 알파벳은 2분기 실적 발표 이후 진행된 컨퍼런스 콜에서 27년 자본지출이 상당히 증가할 것으로 예상한다고 밝혔다. 해당 발표에서 시장이 주목한 부분은 **상향된 CapEx 가이드선**과 **잉여현금흐름의 변화**였다. 알파벳은 26년 연간 CapEx 전망치를 기존 1,800억~1,900억 달러에서 1,950억~2,050억 달러로 상향 조정했다. 상반기에 806억 달러를 집행하였으니 하반기에는 추가적으로 1,144~1,244억 달러를 투자하게 된다. 반면 잉여현금흐름은 1분기 101억 달러 흑자에서 2분기 59억 달러 적자로 전환됐다. CapEx 가이드선 상향과 잉여현금흐름의 적자 전환이 동시에 발표되자 자본지출 지속 가능성에 대한 투자자들의 우려가 커졌고, 발표 직후 알파벳의 시간 외 주가는 3.31% 하락했다.

AI CAPEX 확대는
향후 감가상각비
증가와 현금흐름
부담으로

하이퍼스케일러들이 자본지출을 지속적으로 확대하는 배경에는 AI 수요 성장성에 대한 기대와 선두 주도권 선점 경쟁이 있다. 주요 CSP에게 AI 투자는 검색·광고·클라우드·소프트웨어 등 본업 경쟁력과 직결되는 만큼 AI 경쟁력 확보를 위한 투자는 건조하게 유지될 가능성이 높다. 다만 하이퍼스케일러들의 대규모 AI CAPEX 확대는 향후 감가상각비 증가와 현금흐름 부담으로 이어질 수 있다.

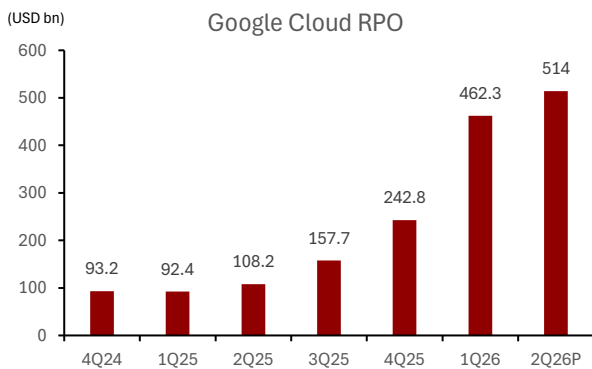
이와 별개로 시장에서는 토큰 비용 하락 속도가 AI 수요 증가 속도를 앞지르면서 결국 AI 인프라 투자가 둔화될 것이라는 우려도 제기되고 있다. 반면 AI 모델의 효율 개선은 인프라 수요를 위축시키기보다 AI 서비스의 가격 부담과 도입 장벽을 낮춰 사용자와 활용 영역을 확대하는 방향으로 작용할 것이라는 시각도 존재한다.

그림 13. 하이퍼스케일러 잉여현금흐름 추이



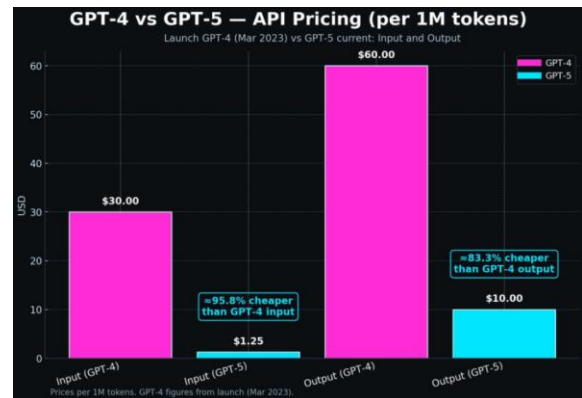
자료: Bloomberg, KUVIC 리서치 1팀

그림 14. Google Cloud RPO



자료: KUVIC 리서치 1팀
주: 1Q26부터 RPO 산정 기준 변경

그림 15. 토큰 비용 하락



자료: 언론 종합, KUVIC 리서치 1팀

하이퍼스케일러의
26년, 27년
CapEx/OCF 전망치는
99%

하이퍼스케일러의 26년, 27년 CapEx/OCF 전망치는 99%로 각각의 연간 영업현금흐름 자체를 거의 다 소진하는 수준에 도달했다. 긍정적인 부분은 빅테크들의 클라우드 매출이 지속적으로 증가 중이라는 것이다. 알파벳의 2Q26 클라우드 주주잔고(RPO)는 1분기 4,620억 달러에서 2분기 5,140억 달러로 11.2%QoQ 증가했다. RPO의 증가는 클라우드 컴퓨팅 수요가 강하다는 것을 의미하고, 잔고가 건조한 이상 추가적인 설비투자가 불가피하다. 단순히 투자 부담의 크기만 보기보다는 이미 확보한 클라우드 계약이 얼마나 빠르게 이익으로 전환되는지를 함께 볼 필요가 있다.

표 5. 하이퍼스케일러 CapEx/OCF

(mn USD)	2023	2024	2025	2026E	2027E
CapEx	153,171	233,527	394,528	725,718	929,881
Amazon	52,729	82,999	131,819	202,342	233,089
MSFT	35,202	55,552	83,094	160,374	197,625
Google	37,974	57,720	109,924	226,443	305,017
Meta	27,266	37,256	69,691	136,560	194,149
OCF	365,791	465,389	590,173	736,136	936,311
Amazon	84,946	115,877	139,514	183,064	236,663
MSFT	102,647	125,583	160,506	196,614	236,020
Google	107,085	132,601	174,353	212,455	282,411
Meta	71,113	91,328	115,800	144,003	181,216
CapEx/OCF	42%	50%	67%	99%	99%
Amazon	62%	72%	94%	111%	98%
MSFT	34%	44%	52%	82%	84%
Google	35%	44%	63%	107%	108%
Meta	38%	41%	60%	95%	107%

자료: Bloomberg, KUVIC 리서치 1팀

AI 투자 확대에 따른
재무 부담은 시장의
우려보다 제한적일 것

알파벳의 Cloud RPO 성장률이 어느 정도 수준을 유지해야 Cloud 사업부의 EBITDA가 27년에 추가되는 자본지출 금액인 786억 달러를 방어할 수 있는지 추정해보았다. Cloud RPO는 확보한 계약잔고가 향후 매출로 인식되는 구조인 만큼, RPO에 매출 인식률을 적용해 Cloud 매출과 EBITDA를 추정하였다. 매출 인식률은 2Q26 확정치인 5.36%를 시작으로 분기별로 소폭 상승해 4Q27에는 6.25%에 도달하는 것으로 가정했다. 이는 알파벳이 26년 1분기 실적발표에서 확보한 Cloud RPO의 50% 이상이 향후 24개월 내 매출로 인식될 것이라고 밝힌 점을 반영한 것으로, 시간이 지날수록 계약잔고의 매출 전환 속도가 점진적으로 높아지는 보수적인 시나리오를 적용했다. EBITDA 마진의 경우 Nebius AI Cloud 사업부의 1Q26 마진인 45%를 프록시로 사용하였다. 이후 27년 예상 Cloud EBITDA와 26년 대비 증가하는 CapEx를 비교해 투자 지속 가능성을 점검한 결과, **Cloud RPO가 약 10%QoQ 성장할 경우 27년 Cloud 사업부의 EBITDA가 추가되는 CapEx 부담을 대부분 흡수할 수 있는 것으로 나타났다.** 이는 2Q26 Cloud RPO 증가율인 11.2%QoQ를 유지할 수 있다면 Cloud 사업의 현금창출력만으로도 27년 추가되는 CapEx 부담을 대부분 흡수할 수 있음을 의미한다. 따라서 **현재 수준의 클라우드 수요가 이어진다면 AI 투자 확대에 따른 재무 부담은 시장의 우려보다 제한적일 것으로 판단한다.**

표 6. 알파벳 Cloud RPO 10%QoQ 성장 가정 시 Cloud 사업부의 분기별 EBITDA 추정치

(USD Bn)	1Q26	2Q26P	3Q26E	4Q26E	1Q27E	2Q27E	3Q27E	4Q27E
Cloud RPO	462.3	514.0	565.4	621.9	684.1	752.5	827.8	910.6
인식률		5.36%	5.51%	5.66%	5.81%	5.95%	6.10%	6.25%
Cloud Revenue	20.0	24.8	28.3	32.0	36.1	40.7	45.9	51.7
Cloud EBITDA	9.0	11.2	12.7	14.4	16.3	18.3	20.7	23.3

자료: KUVIC 리서치 1팀

AI 트렌드 분석

엔비디아와 CPU, LPDDR

에이전틱 AI 시대, CPU의 재조명

생성형 AI 에서
에이전틱 AI로의 전환,
핵심 구성요소로
부상하는 CPU

앞서 살펴본 바와 같이 생성형 AI 는 단순히 답변을 생성하는 단계를 넘어, 스스로 계획을 수립하고 다양한 도구를 활용해 작업을 수행하는 에이전틱 AI 시대로 진입하고 있다. 이 과정에서 CPU 의 중요성이 빠르게 부각되고 있다. 과거에는 GPU 가 핵심 프로세서이고 CPU 는 이를 보조하는 역할이라는 인식이 다수였다. 그러나 AI 의 활용 방식이 학습 중심에서 추론으로, 나아가 에이전틱 AI 중심으로 변화하면서 CPU 는 단순한 보조 프로세서를 넘어 핵심 구성요소로 부상하고 있다.

기존 챗봇형 AI 에서는 사용자가 질문을 입력하면 CPU 가 이를 전처리해 GPU 에 전달하고, GPU 가 대부분의 추론 연산을 수행한 뒤 CPU 가 결과를 후처리하는 구조였다. 즉 CPU 는 입력 정리와 결과 처리 중심의 역할을 담당했으며, AI 성능은 대부분 GPU 의 연산 능력에 의해 결정되었다.

반면 에이전틱 AI 는 사용자의 요청을 여러 단계로 분해한 뒤 필요한 도구를 선택하고, 웹 검색, 데이터베이스 조회, API 호출&코드 실행 등을 반복 수행하며 결과를 검증하고 필요 시 재실행하는 방식으로 동작한다. 이로 인해 대규모 행렬 연산을 병렬로 수행하는 GPU 보다, **다수의 작업을 순차적으로 스케줄링하고 시스템 전반의 흐름을 관리하는 CPU 의 역할이 더욱 중요해졌다.**

에이전틱 AI 시대,
4배이상 증가할 것으로
전망되는 CPU 코어 수요

이 같은 변화는 실제 업계에서도 확인되고 있다. Arm 의 르네 하스 CEO 는 기존 AI 워크로드에서는 데이터센터 용량 1GW 당 약 3,000 만 개의 CPU 코어가 필요했지만, 에이전틱 AI 시대에는 CPU 코어 수요가 약 4 배 증가할 것으로 전망했다.

CPU 수요
증가로 변화하는
CPU&GPU 비율

이에 AI 서버의 CPU 와 GPU 구성 비율 역시 빠르게 변화하고 있다. 전통적인 AI 서버에서는 GPU 대비 CPU 비율이 약 8:1~8:2 수준이었지만, GB200 과 베라-루빈(Vera-Rubin) 등 차세대 AI 가속 아키텍처에서는 2:1 수준으로 개선되었다. 인텔 역시 2026 년 1 분기 실적발표 컨퍼런스콜에서, **향후 데이터센터의 CPU 와 GPU 비율이 1:1 수준까지 높아질 수 있으며 장기적으로는 CPU 비중이 GPU 를 넘어설 가능성도 있다고 밝혔다.**

CPU 중심 서버 확산은
서버당 DRAM
탑재량까지 증가시킴

CPU 의 역할 확대는 메인 메모리(DDR5) 수요에도 직접적인 영향을 미친다. 에이전틱 AI 확산으로 CPU 중심의 AI 서버 구축이 증가하면서 서버당 메모리 탑재 용량 역시 빠르게 팽창하고 있다.

AI CPU 서버는 방대한 데이터를 병목 없이 처리하기 위해 듀얼 소켓(CPU 2 개) 구성이 표준으로 자리 잡았다. 중요한 것은 최신 AI 용 CPU 가 지원하는 메모리 채널 수가 크게 늘어났고, 대역폭을 극대화하기 위해 모든 슬롯에 64GB 이상의 고용량 DIMM 을 빈틈없이 채워 넣는다는 점이다. 이에 따라 비용 효율을 중시하여 슬롯을 일부 비워두거나 저용량 모듈을 쓰던 일반 서버 대비, AI 서버의 대당 DDR5 탑재량은 단숨에 2~4 배 이상 폭증하게 된다. 즉, 에이전틱 AI 의 확산은 CPU 자체의 수요 증가를 넘어 서버용 DRAM(DDR5)의 구조적 초과 수요를 이끄는 핵심 동력으로 작용하고 있다

그림 16. AI 추론 루프의 핵심 구성요소로 부상한 CPU

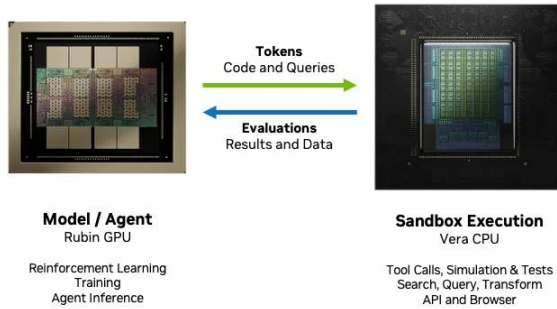
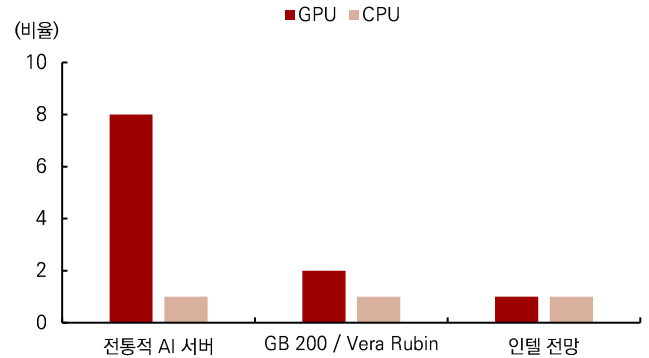


그림 17. CPU-GPU 비율 추이



자료: NVIDIA, KUVIC 리서치 1팀

자료: 언론 종합, KUVIC 리서치 1팀

기존 강자의 대응 - 인텔과 AMD

CPU의 역할이 이처럼 확대되자, 기존 서버 CPU 시장을 주도해 온 인텔과 AMD 역시 차세대 제품 경쟁을 본격화하고 있다. 양사는 모두 AI 데이터센터에 겨냥해 CPU 성능을 고도화하고 있지만, 제품 설계 방향에는 차이를 보이고 있다.

인텔은 워크로드별 전략으로 Xeon 6 라인업을 고도화함

인텔의 핵심 전략은 작업의 성격에 맞춰 코어를 철저히 분리하는 '이원화'다. AI 데이터센터의 워크로드가 무거운 추론 연산과 가벼운 스케줄링 작업으로 양극화됨에 따라, 인텔은 하나의 범용 칩으로 모든 것을 해결하려던 과거의 방식을 버렸다. 연산 성능이 중요한 영역에는 자체 AI 가속 블록(AMX)을 내재화하여 CPU 만으로도 AI 추론을 수행할 수 있는 'P-코어'를 투입한다. 반면, 에이전트 AI가 유발하는 수만 개의 자잘한 API 호출과 웹 검색 등 동시다발적인 스케줄링 작업에는 18A 선단 공정을 적용해 단일 칩에 무려 288 개의 코어를 집적한 'E-코어'를 투입한다. 즉, 전력 낭비 없이 작업 특성에 딱 맞는 맞춤형 연산 환경을 제공하는 것이 인텔의 무기다.

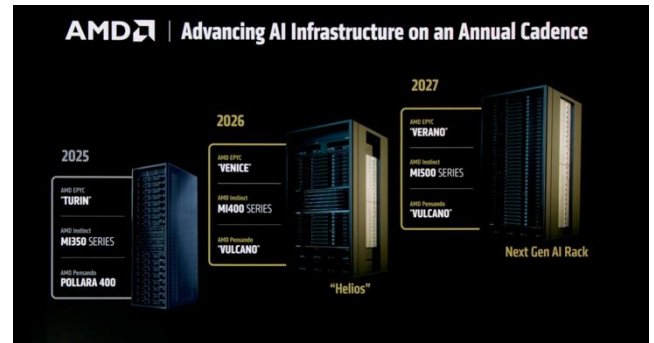
AMD 의 전략은 '칩렛(Chiplet)을 통한 극단적 확장성'과 'AI 통합 생태계 구축'으로 요약된다. AMD 는 인텔처럼 코어의 종류를 극단적으로 나누기보다는, 우수한 범용성을 가진 Zen 아키텍처를 블록 장난감(칩렛)처럼 이어 붙여 단일 칩에 192 개 이상의 코어를 집적하는 '규모의 경제'를 택했다. 더 나아가 AMD 의 진정한 목적은 'CPU 단품 판매'가 아니다. 압도적인 I/O(입출력) 처리 능력과 12 채널 메모리 지원 능력을 갖춘 EPYC CPU 를 중심에 두고, 자사의 AI 가속기(Instinct MI 시리즈)를 하나로 묶어 엔비디아의 독점에 대항하는 '통합 랙스케일 플랫폼'을 구축하는 데 사활을 걸고 있다.

이처럼 인텔은 '코어의 전문화'로, AMD 는 '확장과 플랫폼 통합'으로 각자의 해자를 지니고 있다. 하지만 꿰뚫어 보아야 할 가장 중요한 교집합은, 두 전략 모두 필연적으로 CPU 코어 수 확장(192~288 코어)과 메모리 채널 확대(8 채널 → 12 채널)로 '메모리 대역폭과 용량의 극대화'를 요구한다는 점이다.

인텔이 288 개의 E-코어를 가동해 방대한 데이터를 분산 처리하든, AMD 가 192 개의 코어와 GPU 를 묶어 거대한 데이터를 쏟아내든, 이 연산의 병목을 막아줄 유일한 탈출구는 CPU 와 맞닿아 있는 메인 메모리뿐이다. 이를 위해 양사 모두 차세대 플랫폼에서 12 채널 이상의 초고속 DDR5 지원을 표준화하고 있으며, 나아가 대역폭을 대폭 끌어올린 차세대 고용량 모듈(MCRDIMM/MRDIMM) 탑재를 강제하고 있다. 결론적으로 x86 진영의 치열한 아키텍처 경쟁은, 방식의 차이와 무관하게 서버용 범용 D 램 시장의 대당 탑재량을 구조적으로 격상시키는 가장 확실한 촉매제가 되고 있다.

그림 18. P-Core와 E-Core 이원화를 넘어서는 진화, Xeon 6+

그림 19. AMD 연간 AI 인프라 로드맵



자료: Intel, KUVIC 리서치 1팀

자료: AMD, KUVIC 리서치 1팀

새로운 경쟁자의 등장 – 엔비디아

Vera CPU를 통해 CPU 사업에 본격적으로 뛰어드는 엔비디아

이처럼 인텔과 AMD 가 기존 강자로서 점유율을 공고히 하고 있는 사이, 엔비디아는 GPU 기업에서 AI CPU 플랫폼 기업으로 영역을 넓히며 새로운 경쟁자로 부상하고 있다. 전 세대인 Grace CPU 는 엔비디아가 자체 설계한 Arm 기반 CPU 였으나, 주로 블랙웰 GPU 를 제어하는 용도로만 활용됐을 뿐 독립적인 CPU 제품으로는 공급되지 않았다. 반면 **차세대 제품인 Vera CPU 는 AI 데이터센터를 위한 독립적인 CPU 제품으로 공급될 예정이며, 이는 엔비디아의 CPU 사업이 본격화되는 전략적 전환점으로 평가된다.**

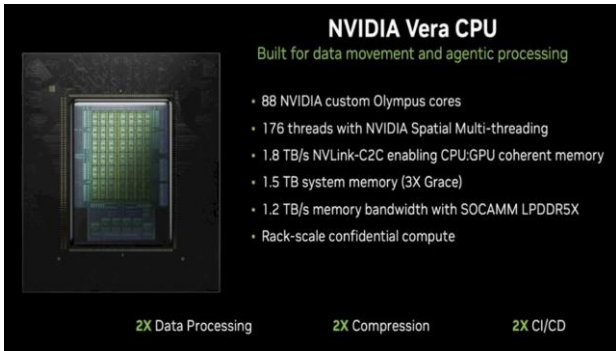
Vera CPU 는 Arm 기반 아키텍처 위에 엔비디아가 자체 설계한 88 개의 Olympus 코어를 탑재했으며, Rubin GPU 와 NVLink 인터페이스를 통해 연결된다. 기존 CPU 대비 최대 2 배의 효율성과 50% 향상된 성능을 제공하는 것이 특징이다. 현재 엔비디아는 Vera CPU 의 생산 단계에 진입했으며, OpenAI 와 SpaceX, Anthropic 등 주요 AI 기업을 대상으로 초기 시스템 공급을 시작했다고 밝혔다. 엔비디아가 CPU 시장에 본격적으로 뛰어 들었다는 점은, AI 시대의 핵심 인프라 기업으로 자리매김하기 위해서 GPU 뿐 아니라 CPU 시장에서도 중요한 위치를 선점해야 한다는 점을 시사한다.

또한 엔비디아의 2026 년 실적 발표를 통해서도 Vera CPU 의 성장성도 확인할 수 있었다. 엔비디아는 Vera CPU 독립 랙만으로도 올해 약 200 억 달러 규모의 매출 기회가 존재하며, 장기적으로는 관련 시장 규모가 약 2,000 억 달러까지 확대될 수 있다고 전망했다.

LPDDR 의 수요를 끌어올린 Vera CPU

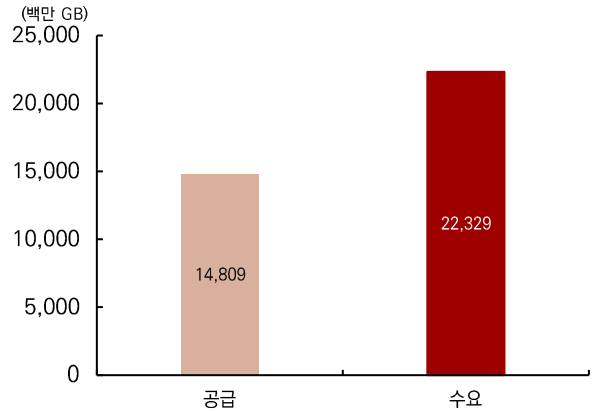
인텔과 AMD 가 전통적인 서버용 DDR5 의 수요를 견인하고 있다면, 새로운 경쟁자인 엔비디아는 차세대 Vera CPU 를 통해 모바일 전용이던 **LPDDR5X 를 데이터센터의 핵심 인프라로 편입시키며** 메모리 폼팩터의 거대한 지각변동을 일으키고 있다. Vera CPU 랙은 256 개의 Vera CPU 로 구성되며, CPU 1 개당 8 개의 SoCAMP2 메모리 모듈이 탑재된다. 이에 따라 단일 랙에는 총 2,048 개의 SoCAMP2 모듈이 필요하며, 이는 기존 NVL72 서버랙의 288 개 대비 약 7.1 배 증가한 수준이다. 또한 **Vera CPU 1 개에는 1.5TB 의 LPDDR(SoCAMP)이 탑재되는데, 이는 직전 세대 Grace CPU 의 480GB 대비 약 3 배 증가한 수치다.** 모건 스탠리에 따르면 엔비디아가 하이퍼스케일러들에게 VR200 NVL72 샴시에 담아 대량 판매할 경우 Vera CPU 는 개당 5 천 달러의 가격에 판매될 계획이라고 한다. 이러한 가격 전망과 200 억 달러 매출 전망을 통해 2026 년 Vera CPU 는 400 만개가 판매될 것으로 전망된다. 이에 따라 **Vera CPU 1 개에 LPDDR5X 1.5TB 가 사용됨으로 2026 년 증가하는 LPDDR 수요는 최대 60 억 GB 일 것으로 전망된다.** 과거 LPDDR 은 스마트폰 중심의 메모리로 인식되어 왔지만, Grace 에 이어 Vera CPU 까지 서버용 LPDDR5X 채택을 확대하면서 모바일 DRAM 공급이 AI 데이터센터로 넘어가는 상황이다. 이로 인해 모바일에 배분되는 공급이 상대적으로 줄어들며 쇼티지가 심화되고 있다. **결론적으로 Vera CPU 는 서버 CPU 향 LPDDR 채택 확대를 이끌어내며, LPDDR 메모리 수요 증가와 모바일 DRAM 의 쇼티지를 견인했다.**

그림 20. 엔비디아 Vera CPU



자료: NVIDIA, KUVIC 리서치 1팀

그림 21. SOCAMM으로 인해 가속화되는 LPDDR 쇼티지



자료: Trendforce, KUVIC 리서치 1팀

병렬 연산의 병목을 허무는 열쇠, HBM

초거대 AI 모델의 발전과 함께 수천 개의 GPU를 연결하여 병렬 연산을 수행하는 아키텍처가 데이터센터의 표준으로 자리 잡았다. 이 과정에서 마주한 가장 큰 기술적 난제는 연산 장치(GPU)의 막대한 처리 속도를 메모리의 데이터 전송 속도가 따라가지 못하는 **'메모리 병목(Memory Wall)'** 현상이다. HBM(High Bandwidth Memory)은 TSV(실리콘 관통 전극) 기술로 여러 개의 D램 다이(Die)를 수직 적층해 데이터 이동 경로를 기하급수적으로 늘림으로써, 이 병목을 해소하고 가속기의 연산 활용률을 극대화하는 필수재로 자리 잡았다.

최근 에이전틱 AI(Agentic AI)의 등장으로 과거의 문맥(Context) 데이터와 접근 빈도가 낮은 KV 캐시를 용량이 큰 범용 DRAM이나 NAND(eSSD)로 오프로딩(Offloading)하는 메모리 계층화가 가속화되고 있다. 그러나 이러한 하위 메모리로의 오프로딩 현상이 HBM의 수요 둔화를 의미하는 것은 결코 아니다. **가장 빠른 데이터 응답이 요구되는 생성 중인 '활성 KV 캐시(Active KV Cache)'는 낮은 지연 시간과 높은 대역폭이 필수적이므로 여전히 전적으로 HBM에 보관되어야 하기 때문이다.** 또한, 하위 스토리지로 데이터를 내리더라도 GPU가 연산 대기 상태에 빠지지 않도록 에이전트의 실행 흐름을 예측해 필요한 캐시를 HBM에 미리 적재(Pre-loading)해두어야 하므로, **결국 핵심 연산을 위한 HBM의 기본 요구 용량은 계속 커질 수밖에 없다.**

여기에 더해 AI 모델이 스스로 여러 경로를 탐색하고 검증하며 추론 단계의 연산량을 늘리는 '추론 모델(Reasoning Model, Test-time Scaling)'로 진화하면서 토큰 생성량과 KV 캐시 규모가 구조적으로 폭증하고 있다. 나아가 여러 에이전트가 각자의 독립된 컨텍스트를 바탕으로 병렬 추론을 수행하는 멀티 에이전트 시스템(Multi-Agent System)이 도입되며, **시스템 전체가 동시에 감당해야 할 KV 캐시의 총량은 더욱 기하급수적으로 늘어나 HBM 수요 팽창을 견인하고 있다.** 뿐만 아니라 추론 비용 효율화를 위해 도입되는 다양한 특화 ASIC 가속기들 역시 성능 확보를 위해 HBM 채택을 늘리고 있어 그 수요 저변은 더욱 넓어지고 있다

NAND

NAND가 뭔데?

NAND 플래시는 전원이 끊겨도 데이터가 유지되는 **비휘발성 메모리**다. 셀 내부의 절연막(차지 트랩)에 전자를 가둬 두는 방식으로 정보를 기록하기 때문에, 전하가 시간이 지나면 새어나가 주기적인 리프레이시가 필요한 DRAM과 근본적으로 다르다. 이 물리적 차이가 곧 용도의 차이로 이어진다. DRAM은 CPU·GPU가 나노초 단위로 즉시 접근하는 주기억장치이고, NAND는 마이크로초에서 밀리초 단위의 지연을 감수하는 **대신 비트당 원가를 압도적으로 낮춘 대용량 저장장치**다. BofA 추정 기준 2026년 DRAM의 8Gb 환산 평균판매가격(ASP)은 13.0달러, NAND의 256Gb 환산 ASP는 8.2달러로, 동일 비트로 환산하면 40배 안팎의 단가 격차가 존재한다.

DRAM 보다
비트당 원가가
압도적으로 낮은
NAND

선평의 DRAM vs
적층의 NAND

미세화의 방향도 다르다. DRAM이 평면 셀을 계속 줄여 나가며 1a·1b·1c 노드로 이행하고 EUV 도입을 확대해 온 반면, NAND는 2010년대 초반 평면 스케일링이 한계에 부딪히자 셀을 수직으로 쌓아 올리는 3D 구조로 전환했다. 즉 NAND의 원가 경쟁력은 선평이 아니라 적층 단수에서 나온다. 이 때문에 NAND는 EUV 의존도가 낮고 식각·증착 장비 비중이 높은, DRAM과는 다른 투자 프로파일을 갖는다.

표 7. DRAM과 NAND의 구조적 비교

구분	DRAM	NAND Flash
회발성	회발성 (주기적 리프레시 필요)	비회발성 (전원 차단 후에도 보존)
셀 구조	1T1C, 평면 셀 미세화 (1a→1b→1c)	차지 트랩 기반 3D 수직 적층
미세화 축	선평 축소, EUV 의존도 높음	적층 단수 증가, 식각·증착 중심
접근 단위	바이트 단위 랜덤 액세스	페이지 단위 읽기·블록 단위 쓰기
지연시간	수십 ns	수십 μ s ~ ms
내구성	사실상 제약 없음	셀당 쓰기 횟수(P/E cycle) 제한
주요 제품	HBM, DDR5, LPDDR, GDDR	eSSD, 클라이언트 SSD, UFS, 임베디드
2026E 시장	568.8B달러 (YoY +325%)	323.1B달러 (YoY +299%)

자료: BofA Global Research, KUVIC 리서치 1팀

NAND, 저장장치에서 AI 연산 인프라로

과거 NAND Flash는 데이터센터 내에서 사진, 영상 등 콜드 데이터(Cold Data)를 보관하는 단순한 저장 매체(Storage)로 여겨졌다. 그러나 에이전틱 AI(Agentic AI) 시대가 도래하며 모델이 수만~수십만 자의 긴 문맥(Long Context)을 이해하고 다단계 추론을 수행해야 하는 환경이 조성되면서, NAND 생태계의 위상이 'AI 연산 인프라'의 한 축으로 완전히 뒤바뀌고 있다.

에이전틱 AI의
KV캐시 폭증 →
NAND 오프로딩이
해법

이러한 지각변동의 핵심은 'KV 캐시 오프로딩(KV Cache Offloading)과 재사용'에 있다. 에이전트가 긴 시간 동안 여러 작업을 수행하다 보면 모델이 기억해야 할 KV 캐시 데이터가 기하급수적으로 쌓이게 된다. 하지만 물리적 공간과 비용의 한계가 뚜렷한 HBM이나 시스템 DRAM만으로는 이 방대한 데이터를 모두 품을 수 없다. 이에 따라 당장 연산에 쓰이지 않거나 접근 빈도가 낮은 과거의 문맥 및 캐시 데이터는 대용량 eSSD로 내려보내(Offloading) 저장해두고, 필요할 때마다 다시 불러오는 계층형 아키텍처가 필수적으로 자리 잡았다. 실제로 엔비디아 역시 차세대 인프라인 Vera Rubin POD 구성에 eSSD 기반의 'CMX(Context Memory Storage) 랙'을 공식 편입시켜, NAND를 GPU 연산을 보조하는 공용 캐시 계층으로 승격시켰다.

이러한 오프로딩 트렌드 속에서 HBM(연산)과 eSSD(저장)의 간극을 메우는 Tier-1.5 메모리로 'HBF(High Bandwidth Flash)'가 새롭게 부상하고 있다. HBF는 낸드 플래시를 TSV(실리콘 관통 전극) 기술로 수직 적층하여 대역폭과 용량을 동시에 끌어올린 차세대 솔루션이다. HBM이 D램을 적층해 '실시간 초고속 데이터 처리'에 집중한다면, HBF는 HBM 대비 속도는 다소 낮지만 10~100배 큰 저장 용량을 제공한다. 또한 비회발성 특성 덕분에 전력 공급 없이도 데이터를 유지할 수 있어 전력 효율성이 압도적으로 뛰어나다.

HBF=HBM·SSD
사이 신계층, HBM
대체 아닌 보완,

AI NAND 수요 비중
26년 32%→28년
51.5% 급증

'HBM(핵심 연산) - HBF(대용량 고속 저장 및 캐시) - eSSD(단순 저장 및 아카이빙)'의 3단계로 더욱 세밀하게 계층화될 전망이다. 에이전틱 AI가 견인하는 이러한 NAND 생태계 수요 급증에 힘입어, 전체 NAND 수요 내 AI가 차지하는 비중은 2026년 33.7%에서 2028년 51.5%까지 절반 이상으로 확대될 전망이다. 결론적으로 NAND는 더 이상 단순한 창고가 아니라 AI 추론 파이프라인이 멈추지 않게 돕는 핵심 확장 메모리로 진화했으며, 이에 따라 전례 없는 물량(Q) 중심의 구조적 수요 팽창 사이클에 진입했다.

메모리 제품별 가격 추이

범용 DRAM 가격 추이

범용 DRAM의 추가적인 가격 인상은 이전보다 제한적일 것

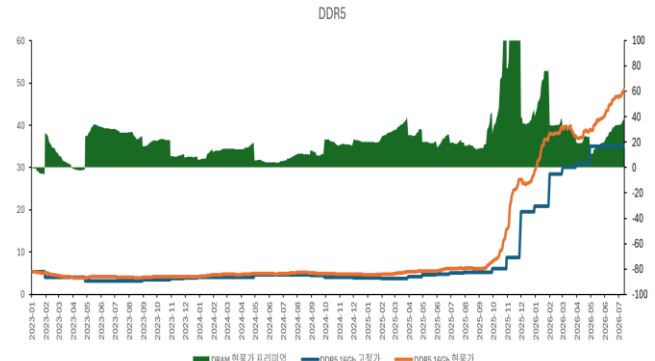
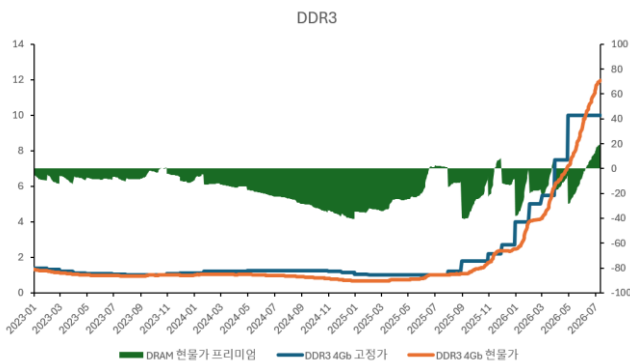
HBM의 공격적인 생산능력 확대와 높은 Trade Ratio는 범용 DRAM 생산능력을 잠식하며 공급 부족을 심화시켰다. 그 결과 **범용 DRAM 가격은 지난 1년간 5배 이상 상승했고, DDR5뿐 아니라 DDR4 등 레거시 제품까지 전 제품군에서 공급자 우위 시장이 형성됐다.** 다만 가격 상승 국면이 종료되는 것은 아니지만, 상승 속도는 점차 완만해질 가능성이 높다. 마이크론은 FY26 3분기 실적발표에서 4분기 Gross Margin 가이드스가 메모리 가격 상승률의 의미 있는 둔화(meaningful moderation in the rate of price increases)를 반영한다고 언급하며 가격 상승세 둔화를 시사했다. 또한 **최근 범용 D램의 마진이 HBM을 웃도는 수준까지 높아졌기에 추가적인 가격 인상 여력은 이전보다 제한적일 것으로 전망된다.**

3Q26 DRAM 계약가격은 전분기 대비 13~18% 상승 전망

TrendForce는 **3Q26 DRAM 계약가격이 전분기 대비 13~18% 상승할 것으로 전망했는데,** 이는 1Q26의 90~95%, 2Q26의 58~63% 대비 상승률이 크게 둔화된 수준이다. 현물가격은 공급 제약으로 강세를 이어가겠지만, 고정가격은 LTA를 기반으로 보다 완만한 상승 흐름을 보일 것으로 예상된다

그림 22. DDR3 4Gb 현물가 및 고정가 추이

그림 23. DDR5 16Gb 현물가 및 고정가 추이



자료: DRAmExchange, KUVIC 리서치 1팀 추정

자료: DRAmExchange, KUVIC 리서치 1팀 추정

HBM 가격 추이

HBM은 엔비디아 등 주요 고객사와의 장기공급계약(LTA)을 기반으로 가격이 결정되는 비중이 높아 단기간에 가격이 크게 변동하지 않는다. 시장 수급 변화가 현물가격과 계약가격에 비교적 빠르게 반영되는 범용 DRAM과는 다른 구조이다. 다만 DigiTimes는 **HBM4 가격이 올해 하반기 Gb당 약 2달러 수준에서 내년에는 4~5달러, 혹은 그 이상까지 상승할 수 있다고 전망했다.** 현재 Gb당 1.5~1.6달러 수준인 HBM3E 가격 역시 동반 상승이 예상된다. 삼성전자, SK하이닉스, 마이크론 등 글로벌 메모리 3사는 올해 4분기 중 내년 HBM 공급가격을 확정할 것으로 전망된다.

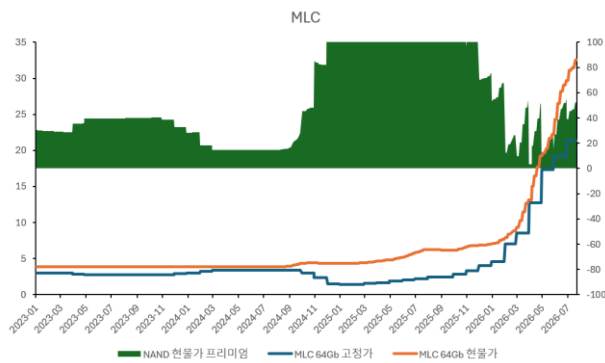
NAND 가격 추이

최근 일부 NAND 현물가격의 하락이 시장의 우려를 키우고 있다. 특히 512Gb TLC 현물가격은 지난 3월 중순을 정점으로 하락세를 보이고 있다. 그러나 극심한 공급 부족과 LTA 확대를 감안하면 현물시장은 전체 NAND 시장을 대표하기 어렵다. NAND 현물시장은 DRAM보다 제품별로 단절되어 있으며, 개별 품목의 거래 규모와 유동성이 제한적이기 때문이다. 따라서 최근 현물가격 조정은 일부 제품군의 단기 수급 완화를 반영한 것으로 해석하는 것이 적절하며, 기업용 SSD 중심의 계약가격 추세와 동일선상에서 해석하기에는 한계가 있다.

3Q26 NAND 계약가격은 전분기 대비 10~15% 상승 전망

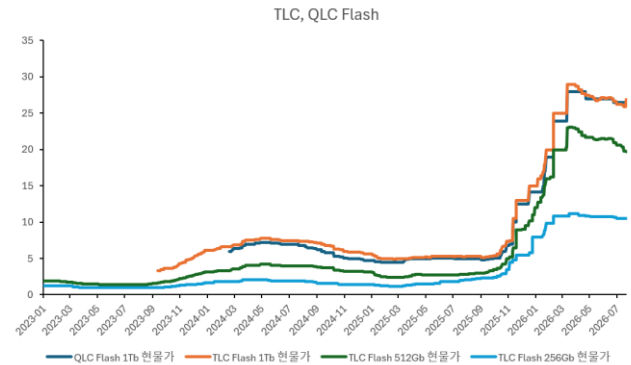
TrendForce는 3Q26 NAND 계약가격이 전분기 대비 10~15% 상승할 것으로 전망했는데, 이는 2Q26의 55~60% 대비 상승폭이 크게 둔화된 수준이다. 또한 3Q27부터는 TLC·QLC 웨이퍼 고정가격이 하락세로 전환돼 3분기 9.4%, 4분기 15.4% 하락할 것으로 전망했다. 기업용 SSD의 비트 출하량 증가세가 둔화되는 가운데 **삼성전자, SK하이닉스, 키옥시아의 신규 생산능력이 본격 가동되면 공급 증가 속도가 수요를 상회할 수 있다**는 분석이다.

그림 24. MLC 64Gb 현물가 및 고정가 추이



자료: DRAMeXchange, KUVIC 리서치 1팀

그림 25. TLC, QLC Flash 현물가 및 고정가 추이



자료: DRAMeXchange, KUVIC 리서치 1팀

가격 중심에서 물량 중심으로

전공정 투자와 장비 발주가 메모리 산업의 핵심 변수가 될 것

결론적으로 메모리 가격은 전 제품군에서 상승세를 이어가고 있으나, 상승폭은 점차 축소되는 국면에 진입하고 있다. 범용 DRAM 계약가격 상승률은 1Q26 90~95%에서 3Q26 13~18%로 둔화되고 있으며, NAND 역시 3Q27부터 가격 하락이 전망된다. 즉, 메모리 업황의 중심축이 **가격(P) 중심의 사이클에서 물량(Q) 중심의 사이클**로 이동하고 있는 것이다. 향후 실적과 산업 방향성을 결정하는 변수는 추가적인 가격 인상보다 비트 출하량 확대를 위한 생산능력 증설과 공급 확대가 될 가능성이 높다. 결국 메모리 공급사들의 증설 규모와 생산성 개선 여부가 향후 업황을 좌우할 것이며, **이에 따라 전공정 투자와 장비 발주가 메모리 산업의 핵심 변수가 될 것으로 판단한다.**

메모리 시장 플레이어 분석

표 9. 메모리 3사 D램 증설 목표 타임라인 (~2029년, 예상 wpm)

기업	지역	공장	1H26	2H26	1H27	2H27	1H28	2H28	2029
삼성전자	평택	P4	1Q Ph3 45k	7월 Ph4 45k 11월 Ph2 45k					
		P5					팹1 오픈 약 300k		팹2 오픈 약 300k
	용인	1호 팹							10월 팹1 가동 목표
SK하이닉스	용인	Y1			2월 Ph1 60k	8월 Ph2 60k	2월 Ph3 60k	8월 Ph4 60k	2월 Ph5 60k
		Y2							오픈
	청주	M15X			4월 Ph3, 4 최대 80k				
마이크론	美 버지니아 마나사스	Fab 6	5월 1α 전환 20k						
		ID1				3Q 오픈 80k			
	美 아이다호 보이시	ID2					오픈 80k		
	美 뉴욕 클레이	-							오픈
	日 히로시마	Fab 15					신규장비 반입 45k		
	臺 통류	P5				4Q 오픈 50k			

자료: 언론 종합, KUVIC 리서치 1팀

SK하이닉스: 지켜내야 할 HBM 주도권

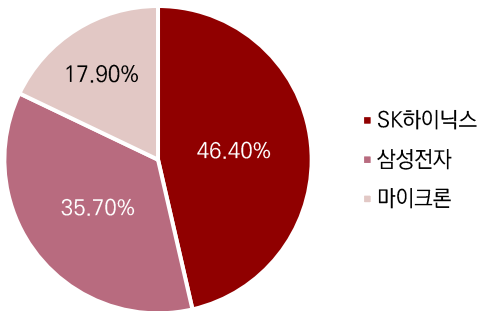
고객사 공급망,
기술력 기반으로
HBM 주도권 확보

HBM 시장에 일찍이 진입했던 SK하이닉스는 1Q26 기준 HBM 매출 점유율 약 58%를 차지하고 있으며, 이미 2026년 HBM 생산분은 전량 완판된 상태다. 이러한 지위는 시장 선점과 더불어 고객사 공급망 내 침투 깊이에서 비롯된다. 실제로 엔비디아는 2026년 HBM4 수요의 약 절반을 동사에 배정한 것으로 알려졌으며, 구글 역시 차세대 TPU v7p-v7e에 들어갈 HBM의 공급사로 동사를 선택할 것으로 예상되는 상황이다. 기술 측면에서도 선제적 대응이 이어지고 있다. 동사는 HBM4부터 고객 맞춤형 기능 탑재에 따른 수율 난이도 상승에 대응해 로직 다이를 자체 생산하는 대신 TSMC의 첨단 공정을 활용하고 패키징까지 협력하는 체계를 구축했으며, 2025년 9월 세계 최초로 HBM4 양산 체제를 확보하며 시장 개화에 선제적으로 대비했다. 결국 HBM 경쟁의 1차 분기점은 엔비디아 공급망 진입 시점과 그 안에서의 자리매김이었으며, 동사는 HBM3E 시대에서의 주도권을 HBM4 시대에까지 연장하려 하고 있다.

청주·용인 증설로
2030년까지 캐파
100만 장 목표

현재 동사의 D램 월 웨이퍼 투입 능력은 중국 우시 공장 생산량 약 20만 장을 포함해 월 55만 장 수준이나 2030년까지 캐파를 약 100만 장까지 끌어올린다는 계획인데, 청주 M15X 공장이 시발점이다. 20조 원 투자를 통해 기존 M15 공장을 확장한 M15X는 지난 4월부터 점진적 램프업을 거쳐 연말에는 월 3~5만 장, 내년에는 최대인 월 8만 장 가량의 생산 능력을 가지게 된다. 동사 증설의 핵심은 최대 600조 원을 투입해 총 4기의 팹을 구축하는 총 416만㎡ 규모의 용인 반도체 클러스터다. 팹 1기당 월 6만 장 생산 능력을 가진 클린룸 6개를 구축하는데, 2027년 2월 첫 클린룸(Ph1)을 오픈하는 것이 목표다. 이후 매 6개월마다 다음 클린룸을 오픈해 신규 투입량을 단계적으로 늘릴 예정인데, 계획대로라면 2030년 상반기엔 용인 1기 팹에서만 월 36만 장 규모의 신규 D램을 확보할 수 있다.

그림 26. 2026년 엔비디아향 HBM4 출하 비율 추정치



자료: TrendForce

그림 27. SK하이닉스 용인 1기 팹

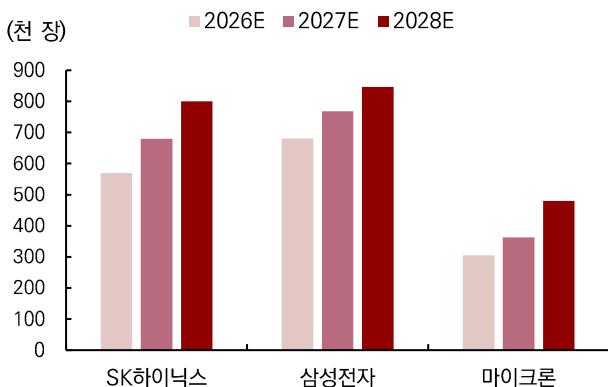


자료: SK하이닉스

1c 공정 전환으로 HBM 캐파 확대, 범용 D램 잠식 둔화 동시에 달성

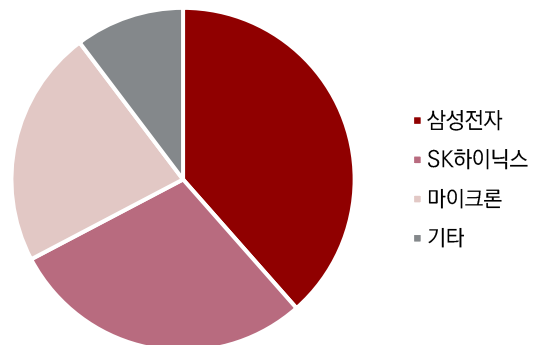
더불어 동사는 지난해 4분기부터 동사의 HBM3E 및 HBM4 제품의 코어 다이로 활용되는 1b D램용 설비투자를 M15X에 집행해 왔는데, 최근 들어 **1c D램용 설비를 발주하는 방향으로 선회**했다. 원래 동사는 올해 엔비디아향으로 60억 Gb 수준의 HBM4 출하량을 계획했으나, 베이스 다이 재설계 필요성 제기 및 데이터 처리 스펙 상향 요구에 따른 검증 지연 문제로 출하량 감소가 불가피해지면서 동사의 HBM4 목표 출하량도 이보다 20~30% 가량 낮아진 것으로 파악됐다. HBM4 하향분이 HBM3E, 서버용 D램, LPDDR 등으로 이관되며 **메모리 수요 총량을 훼손하지 않으면서도**, 1b D램의 생산능력을 1c D램으로 돌리면서 HBM4E 등 **향후 크게 늘어날 1c D램 수요를 사전에 받아내는 것**이다. 이로 인해 단기적으로 1b 노드 생산 공백이 발생하겠으나, 셀 면적 축소로 동일 웨이퍼에서 더 많은 다이를 확보할 수 있어 웨이퍼 추가 투입 없이 비트 공급을 확대할 수 있다. HBM이 범용D램 대비 웨이퍼를 3~4배 이상 더 소모하고, 코어 다이 미세화를 통해 다이 면적이 줄 동일 HBM 스택 생산에 필요한 웨이퍼가 감소하는 걸 고려하면 **1c 전환은 HBM 캐파 확대와 범용 D램 잠식의 속도 둔화를 동시에 달성**하는 방법이다.

그림 28. 메모리 3사 월별 웨이퍼 투입 수 증가 추이



자료: KUVIC 리서치 1팀

그림 29. 1Q26 글로벌 D램 점유율



자료: TrendForce

삼성전자: 1c 노드로 그리는 반등 시나리오

HBM 세대 전환기마다 후발 주자였던 삼성전자는 HBM4를 반등의 계기로 삼고 있다. HBM3E 세대에서 엔비디아 인증이 경쟁사보다 늦어 공급망 합류가 지연됐으나, 4Q25부터 점유율 회복에 나서 **1Q26 기준 전체 D램 시장 1위를 탈환**했고 HBM 점유율도 30% 수준까지 끌어올린 것으로 추산된다. 반등의 결정적 계기는 HBM4 선제 진입이다. 동사는 **2026년 2월 업계 최초로 HBM4 양산 출하를 시작**했으며, 베이스 다이에 4나노 선단 파운드리 공정을 적용해 **최대 13Gbps의 데이터 처리 속도와 이전 세대 대비 약 40% 개선된 전력 효율을 확보**했다. 이는 엔비디아가 HBM4에 요구한 11.7Gbps대 스펙을 상회하는 수준이다. 로직 다이 생산과 패키징을 TSMC에 의존하는 SK하이닉스와 달리, 메모리

설계·생산, 파운드리 로직 공정, 어드밴스드 패키징을 모두 내재화한 구조적 장점이 HBM4 세대에서도 드러났다.

표 10. 메모리 3사 HBM4 비교

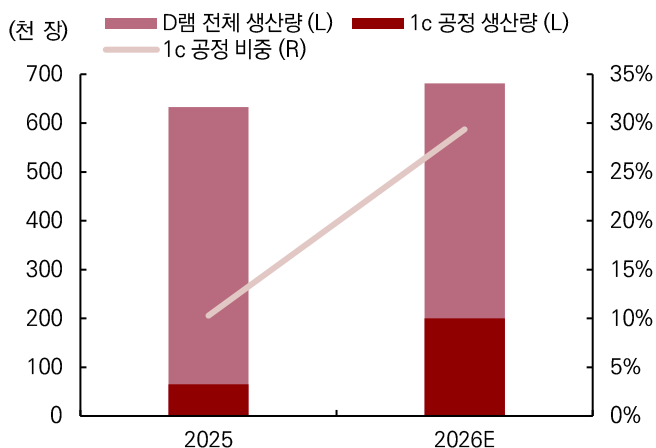
	삼성전자	SK하이닉스	마이크론
첫 양산/출하	2026년 2월 (업계 최초)	2026년 2월	2026년 1분기 (베라 루빈항)
코어 D램	1c (10나노 6세대)	1b (10나노 5세대)	1β (10나노 5세대)
베이스 다이 공정	자체 파운드리 4nm (SF4)	TSMC 12nm	자체 생산 (in-house)
동작 속도	최대 13 Gbps	11.7 Gbps	11+ Gbps
메모리 대역폭/스택	3.3 TB/s	2+ TB/s	2.8 TB/s
용량	24~36GB	36GB (12단) / 48GB (16단)	36GB (12단) / 48GB (16단 샘플)
엔비디아 탑재량	8개 (루빈, 총 288GB)	8개 (루빈, 총 288GB)	8개 (루빈, 총 288GB)
특징	HBM3E 대비 에너지 효율 40% 개선	MR-MUF, 웨이퍼 30μm 박막화	자체 베이스 다이로 성능 열위

자료: KUVIC 리서치 1팀

다만 조기 진입이 곧 주도권 회복을 의미하진 않는다. 2026년 엔비디아향 HBM4 조달 비중이 SK하이닉스에 크게 못 미친다는 분석이 팽배한데, 이는 동사가 고객사 내 선두 지위를 단기간에 가져오기 어렵다는 시장의 시각을 반영한다. 동사의 활로는 오히려 고객 다변화에 있다. 브로드컴·구글·아마존·마이크로소프트·메타 등 ASIC 업체를 아우르는 고객군을 확보한 만큼, 엔비디아 의존도가 낮다는 점은 커스텀 HBM 수요가 확대될수록 엔비디아향 열위를 상쇄하는 지렛대로 작용할 수 있다. 결국 동사의 반등은 시작됐으나, **HBM3E에서 벌어진 격차를 HBM4·HBM4E 세대에서 얼마나 회수하느냐가 남은 과제다.**

현재 D램 웨이퍼 투입 기준 월 66만 장 수준의 생산능력을 보유한 동사의 생산 전략은 지난 수년간 잃었던 기술 주도권을 되찾기 위한 공세적 전환에 가깝다. 전략은 크게 2가지로 정리된다. **10나노급 6세대 1c 노드로의 급격한 전환**과 HBM4 선제 투입, 마지막으로 **낸드 증설 투자를 통한 제품 믹스 재편**이다. 2가지 모두 2026년 현재 동시 진행되고 있으며, 각각 상당한 자본과 기술력을 수반한다는 점에서 동사의 올해 투자는 다소 공격적이라 평가된다.

그림 30. 삼성전자 1c D램 생산량 및 비중



자료: KUVIC 리서치 1팀

그림 31. 삼성전자 평택 공장 조감도



자료: 금융위원회

특히 동사는 속도감 있는 **1c 캐파 확대에 박차**를 가하고 있다. HBM4에 SK하이닉스는 이미 성숙 공정인 1b 공정을 적용했고 마이크론 역시 1b에 대응되는 1β 공정을 적용했으나, 동사는 11나노 수준인 1c 공정을 적용하기 때문에 1c 생산능력 및 수율이 매우 중요하다. 이에 동사는 평택 P4 페이지 3, 4, 2 증설을 통해 **2025년 말 웨이퍼 기준 월 6~7만 장 수준이었던 1c D램 생산능력을 2026년 말까지 월 20만 장으로 늘릴 방침**이다. 계획대로 라인이 갖춰지면 동사는 전체 D램 생산능력 중 1c D램으로만 약 30%를 채운다. 캐파 증가와 선단 공정 비중 확대를 동시에 취하는 구조다.

동사의 또다른 핵심 증설 계획인 평택 P5 공장은 다른 평택 공장들과 달리 3층 구조에 6개 클린룸으로

구성된 초대형 생산기지로, D램·낸드·파운드리 라인을 모두 수용하는 하이브리드 팹으로 조성된다. 당초 2026년 말께 클린룸 구축에 나설 예정이었으나 D램과 낸드에서의 수요 급증으로 인해 인프라 구축 일정을 앞당겼고, 팹1은 2028년, 팹2는 2029년 경 본격 가동될 것으로 예상된다. 규모 역시 상당한데, **팹1과 팹2에 각각 60조원 가량이 투입되어 각각 월 30만 장 규모의 웨이퍼 생산능력을** 갖출 예정이다.

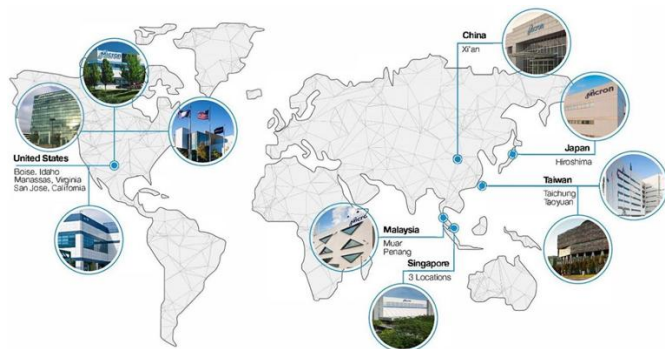
2번째 축인 **낸드 믹스 재편도 P5에서 구체화된다**. 동사는 2022년 이후 낸드 업황이 악화되자 신규 낸드 라인 구축 대신 D램 설비투자에 역량을 집중하고 낸드는 제한적 전환투자만 이어왔으나, 최근 기조를 바꿔 P5 낸드 증설을 결정했다. AI 수요가 학습에서 추론으로 확산되며 고용량 스토리지 수요가 늘어난 것이 배경이다. 관건은 이 재편이 D램 증설 여력을 잠식하는지 여부인데, P5가 신규 팹인 만큼 **기존 D램 캐파를 직접 깎아내기보다는 늘어난 클린룸 공간을 낸드와 D램에 배분**하는 성격에 가깝다. 다만 한정된 투자 재원과 장비를 두 제품이 나눠 갖는 구조인 이상, 낸드 수익성이 개선될수록 캐파 배분에서 낸드의 우선순위가 높아질 가능성은 D램 공급 측면에서 주시할 변수다.

마이크론: 규모는 열위, 강점은 분산

기존 팹 효율화 +
신규 공장 건설로
캐파 확장 중

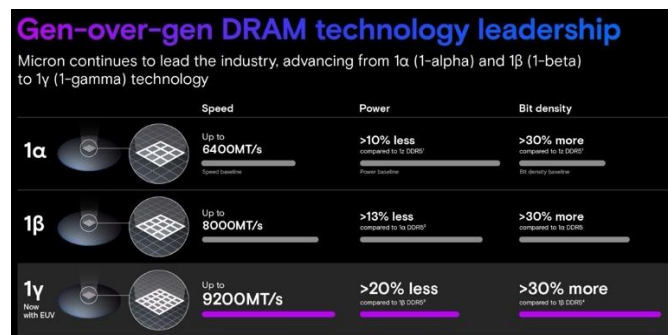
메모리 3사 중 후발 주자로 분류돼 온 마이크론은 1Q26 기준 HBM 점유율 약 21%로 3위에 머물러 있으나, **공격적 캐파 확장을 통해 판도 변화를 노리고 있다**. 동사는 뉴욕 팹 건설 발표 당시 향후 10년간 전 세계 최첨단 D램 생산량의 40%를 미국에서 생산하겠다고 공언한 바 있다. 이를 위한 **동사 증설 전략의 핵심은 기존 팹의 효율화와 신규 부지 확보를 병행**하는 것이다. 먼저 지난 5월 기존 미국 버지니아 마나사스 공장의 10나노급 1세대 1x, 2세대 1y 구형 제조 공정을 4세대 1α 공정으로 전환 완료하면서, DDR4나 LPDDR4 등 레거시 D램 공급 라인을 효율화했다. 또 미국 아이다호 보이시에 도합 월 16만 장의 생산량이 기대되는 팹 2기를 건설 중인데, 각각 2027년 중반과 2028년 말 D램 웨이퍼 생산 개시를 목표로 하고 있다. 이어 뉴욕주 클레이에도 2026년 1월 첫 번째 팹을 착공해 2029년부터 D램 생산을 개시할 전망이다. **미국 외 D램 생산 거점은 동아시아다**. 동사는 2026년 1월 대만 통뤄에 위치한 PSMC의 P5 팹을 18억 달러에 인수했다. 구형 DDR4를 생산하던 공장이지만 300mm 클린룸이 이미 있어 동사는 공장 신축 대비 빠르게 D램 생산 능력을 확장할 수 있었고, 2027년 하반기부터 본격적인 D램 생산 기여가 시작될 전망이다. 일본 히로시마에도 약 1조 5,000억 엔을 투입해 2028년 중반부터 공급을 더해줄 차세대 D램·HBM 생산시설을 구축하고 있다.

그림 32. 마이크론 생산 설비 현황



자료: Micron

그림 33. 마이크론 D램 세대별 성능 비교



자료: Micron

업계 최초 6세대
노드 출하

HBM 시장에서의 존재감도 확대되고 있다. 2026년 엔비디아 HBM4 시장에서 동사의 점유율은 약 15% 내외로 추정되며, 베라 루빈 플랫폼용 12단 HBM4 36GB 는 2026년 1분기 불륨 출하에 진입했다. 이는 자체 가이던스보다 한 분기 앞당겨진 것으로, 후발 주자로서의 실행 속도를 보여주는 대목이다. 다만 코어 다이는 아직 1β 기반으로, 1c를 선제 적용한 삼성전자나 1b에서 1c로 전환 중인 SK하이닉스 대비 HBM4 노드 측면에서는 후행하고 있다. 반면 **범용 D램에서는 노드 전환 속도가 두드러진다**. 동사는 **업계 최초로 6세대 1γ 노드를 출하**했으며, 1β D램의 핵심 생산 거점인 **히로시마 팹에 1γ에 EUV 노광을 처음 도입**했다. 1γ는 EUV와 High-K 메탈 게이트를 적용해 1β 대비 30% 이상의 비트 밀도 개선과 최대 9,200MT/s의 DDR5 속도를 구현한다. 동사는 1β에서 1γ로의 전환을 빠르게 진행해 **2026년 하반기 대부분의 비트를 1γ로 전환할 계획**이며, 이러한 효율화 조치만으로

2026년 약 20%의 추가 캐파 효과를 확보한 것으로 평가된다.

한편 동사의 다국적 공장 증설은 단순한 물량 증대를 넘어 지정학적 이점을 수반한다. 미국, 일본, 대만에 D램 생산 거점을 분산하면서 **미국 CHIPS Act 보조금과 일본 경제산업성의 지원을 동시에 확보했고, 히로시마 팹에는 최대 5,360억 엔의 정부 지원이** 배정됐다. 무엇보다 동사는 중국 내 D램 웨이퍼 생산 팹을 보유하지 않아, 우시 캐파가 장비 반입 규제로 동결된 SK하이닉스나 시안 팹 제약을 안고 있는 삼성전자에 비해 **중성 자유도에서 상대적 우위**를 갖는다. 미국 고객사 입장에서 공급망 다변화 니즈에 부합하는 선택지다.

증설 본격 반영은
2028년부터,
6세대 1γ 노드가
동사 전략의 핵심

다만 이 이점이 단기간에 점유율로 환산되기는 어렵다. 그린필드 팹의 실질 기여가 2027년 중반 이후로 밀려 있어 2026~2027년 구간은 1γ 노드 전환과 대만 P5 인수분배 의존하는 구조이며, 미국 내 최첨단 D램 40% 생산 목표 역시 10년 단위 과제다. 결국 **동사의 확장 전략은 2028년 이후를 겨냥한 포석**에 가까우며, **그 시기까지는 캐파 총량보다 1γ 전환 속도와 6세대 1γ 노드의 HBM4 코어 다이 적용 시점이 실적을 좌우할 변수로** 판단된다.

CXMT IPO, 글로벌 메모리 공급망 내 돌연변이?

CXMT 25년
매출+155%,
흑자전환, 26년 1Q
+719%, 다만 3사
점유율 90%+ 여전

CXMT 상장이 눈앞에 다가왔다. 더 이상 동사를 변방의 회사로 취급하기는 어려워졌다. 2025년 매출은 617.99억위안으로 전년 대비 약 155% 증가했고, 2023~2025년 매출 CAGR은 160.78%에 달하며, 2025년 처음으로 귀속순이익 18.75억위안의 흑자를 냈다. 2026년 1분기 매출은 508억위안(+719%)으로 튀었고, 회사는 상반기 매출 1,100~1,200억위안을 가이던스로 제시했다. Omdia 기준 2025년 4분기 매출 점유율은 7.67%로, 산정 기준상 중국 1위·글로벌 4위 DRAM 업체로 올라섰다. 삼성·SK하이닉스·마이크론 3사가 여전히 시장의 90% 이상을 쥐고 있으나, 새로운 메모리 공급사가 처음으로 유의미하게 끼어든 것은 분명하다.

가동률 95.7%로
풀가동,
DDR5/LPDDR
전환 가속

동사의 웨이퍼 캐파는 2025년 말 약 265k/월에서 2026년 말 350k/월로 확대될 예정이며, 글로벌 웨이퍼 점유율은 약 13%에서 17% 수준으로 올라간다. 가동률은 2024년 87.06%에서 2025년 95.73%로 사실상 풀가동에 근접했다. 제품 믹스의 고도화도 빠르다. **2024년 말 자체 DDR4 생산을 중단하고 DDR5·LPDDR5/5X로 전환하면서, DDR 계열 매출은 2024년 31.7억위안에서 2025년 195.3억위안으로 6.2배 뛰었고 비중은 31.87%까지 올라왔다. LPDDR 계열은 66.43%로 여전히 최대 라인이다.** 여기에 중국 DRAM 자급률이 아직 5% 안팎(목표 30%)이고 중국이 범용 DRAM 수요의 30~35%를 차지한다는 점을 겹치면, CXMT가 겨냥하는 '내수 대체'의 여지는 대단히 크다.

HBM 사실상 부재

그러나 이 서사를 곧바로 '3사 위협'으로 해석하기에는 시기상조다. **첫째, 이번 사이클의 엔진은 HBM인데 CXMT는 거기에 사실상 부재하다.** 이번 증권신고서 전문에는 'HBM'이라는 단어가 단 한 번도 등장하지 않으며, 90억위안 규모의 '전망기술' 프로젝트도 미래 선단 DRAM 적합 연구로만 기술되고 건설기간이 2026~2028년으로 잡혀 있을 뿐이다. HBM3 진입은 여전히 사업계획·언론 레벨의 목표이고, 재무적으로 인식되는 HBM 매출은 0이다. 전공정에서는 DUV 멀티패터닝으로 HBM3를 구현하려 시도 중이나 코어 다이 수율이 범용 대비 크게 낮고, 후공정에서는 다이 스테킹의 warpage·bonding·열응력 문제로 고전하고 있다. **사실상 HBM 공급에 수치적으로 의미있는 반영이 어렵다고** 판단된다. **지금 업황의 쇼티지를 촉발한 것은 HBM이나, CXMT는 이에 올라타지 못한 채 범용에서만 유의미하다.**

노드 2세대 격차,
웨이퍼점유율 대비
비트출하 점유율은
훨씬 낮음, 수율도
업계 하회

또한 노드와 수율의 격차가 실제 공급 기여를 제한한다. 주력 공정은 17nm급 제4세대 플랫폼이고, 해외 대형사 대비 두 세대 안팎의 격차가 유지된다. 업계가 이미 1b(10nm 5세대, HBM3e 적용)와 1γ를 달리는 동안 CXMT의 선단은 G4(16nm, 1z급)에 머물러 있다. 이 밀도 격차 때문에 웨이퍼 점유율(13~17%)과 비트 출하 점유율(2026년 약 9%, 2027년 약 12%) 추정치 사이에 뚜렷한 갭이 존재한다. 즉 350k/월이라는 웨이퍼 숫자는 마이크론에 근접해 보여도 비트로 환산하면 한참 못 미친다. 수율 역시 G3가 약 80%, 주력 G4는 업계 성숙 수준(85~90%)을 하회하는 것으로 파악되며, 신규 라인의 수율 램프가 캐파 목표를 실제 출하로 옮길지를 가르는 핵심 변수로 남아 있다.

흑자 전환을 더욱 구체적으로 뜯어보면 경쟁력을 확산하기가 어렵다. GPM은 2023년 -2.19%에서

2024년 5.00%, 2025년 41.02%로 급등했으나, 이는 주력 DRAM 판매단가가 2024년 55.08%, 2025년 33.69% 오른 가격 효과에 크게 기인한다. 규모의 경제는 아직 충분히 발현되지 않았고, 2025년 말 누적된 손실은 여전히 -366.5억위안이며, 고정자산 급증에 따라 감가상각비가 2023년 105.6억에서 2025년 246.8억위안으로 늘어났다. 이 구조는 다운사이클에서 그대로 역레버리지로 돌아올 수 있다. 소유구조도 함정이다. 2025년 전체 순이익 71.44억위안 중 귀속순이익은 18.75억위안으로 26%에 불과한데, 캐파 순증을 주도하는 신교·베이징 신편의 대주주가 국유펀드이기 때문이다. 상장사가 실제로 귀속받는 이익은 팍 경제성의 일부다.

현재 캐파 수준으로는 2026~2027년 예상 수요 증가분에 견줘 글로벌 수급에 미치는 영향이 제한적이며, 3사의 증설이 본격화되는 2028년 이후야 CXMT의 물량이 의미 있는 변수가 된다. 게다가 캐파가 2027년까지 사실상 예약이 있는데다 내수 초과수요 상태여서, 글로벌 시장으로의 침투 여력도 아직 크지 않다.

관전포인트 =
수율램프·HBM
첫매출시점,
수출비중 =>
아직 임계점 미도달

정리하면, CXMT는 '중국 내수 범용 DRAM을 잠식하는 실재하는 위협'이긴 하나 향후 메모리 공급을 주도할 핵심으로 보기엔 어렵다. 3사 관점에서 이 이름은 2028년 이후 범용 DRAM의 마진·점유율 리스크로 모델에 반영하는 것이 타당하며, 단기적으로는 3사의 시장에서 중국 내수 일부를 차지하는 변수로 보는 것이 정확하다. **결국 위협의 실현 여부는 세 가지 포인트로 압축된다. 신규 라인의 수율 램프, HBM3의 '첫 매출' 인식 시점, 그리고 내수를 넘어서는 수출 비중의 등장.** 이 셋 중 어느 것도 아직 임계점을 넘지 않았다는 점에서 결론은 명료하다. 위협의 가능성은 분명하나, 그 시점은 아직 도래하지 않았다.

표 11. CXMT 추정치

	2026E	2027E	2028E
총 DRAM CAPA(kwpm)	330	405	495
웨이퍼 점유율	16%	18%	19%
DRAM 공급량(GB/bn)	4.25	5.58	7.35
비트 점유율	11.8%	13.0%	13.7%

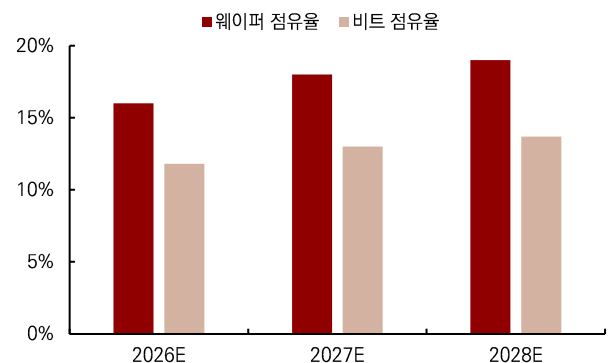
자료: SemiAnalysis, KUVIC 리서치1팀

그림 34. CXMT 허페이 Fab



자료: CXMT

그림 35. CXMT 웨이퍼 점유율 및 비트 점유율



자료: 언론 종합, KUVIC 리서치 1팀

Shortage 분석

D램 수요 전망

표 12. 범용 D램 수요 전망

구분	2026	2027	2028	2029	2030
Total DRAM Demand (bn Gb)	375	448	562	704	869
Mobile DRAM	85	99	107	113	124
PC DRAM	28	31	34	37	41
Server DRAM	196	227	312	441	581
Others	66	92	124	168	229
Set Units(백만개)					
Mobile	1,051	1,051	1,003	933	898
PC	245	250	255	258	263
Server	23	25	28	30	33
Contents/Box(GB)					
Mobile DRAM	10	12	13	15	17
PC DRAM	14	15	17	18	19
Server DRAM	1,063	1,134	1,418	1,795	2,145

자료: KUVIC 리서치 1팀

Mobile DRAM 수요 전망

메모리 시장이 호황기를 맞고 있음에도 Mobile DRAM 시장은 상대적으로 약한 상승세를 보이고 있다. 그 이유를 한 문장으로 요약하면 **DRAM 가격 급등으로 인한 스마트폰 출하량 감소**라고 할 수 있다. 엔비디아 AI 가속기와 데이터센터 서버 수요가 폭증하면서 메모리 업체들은 AI 데이터센터용 D램과 낸드 공급을 우선시했고, 이에 따라 메모리 가격은 분기 내내 큰 폭으로 상승했다. 최근 한 분기 동안 스마트폰용 DRAM과 NAND 가격은 80% 이상 상승하며 Mobile 메모리 시장의 원가 구조를 크게 변화시켰다.

실제로 **800달러 수준의 스마트폰을 기준으로 메모리가 차지하는 부품 원가 비중은 지난해 1분기 약 14%에서 올해 2분기 약 40%까지 상승**한 것으로 나타났다. 이에 스마트폰 제조사들은 높아진 메모리 원가를 제품 가격에 반영할 수밖에 없었다. 시장조사업체 Omdia에 따르면 전 세계 스마트폰 평균 판매가격(ASP)은 지난해 467달러(약 72만 원)에서 올해 565달러(약 87만 원)로 약 21% 상승할 것으로 전망했다.

글로벌 스마트폰 시장 점유율 1위 업체인 삼성전자의 가격 정책에서도 이러한 흐름을 확인할 수 있다. 삼성전자의 갤럭시 A 시리즈 신제품인 갤럭시 A27 5G의 미국 출시 가격은 전작 대비 17.1% 상승했으며, 국내에 출시된 갤럭시 A37 5G의 출고가 역시 전작 대비 19.8% 인상되었다. 이처럼 스마트폰 가격이 전반적으로 상승하면서 소비자 수요는 위축되었고, 결과적으로 스마트폰 출하량 감소로 이어졌다. TrendForce는 **2026년 글로벌 스마트폰 생산량이 전년 대비 16.2% 감소한 10억 5,100만 대에 그칠 것으로 전망**하였다.

다만 스마트폰 출하량 감소가 곧바로 Mobile DRAM 수요 감소로 이어진 것은 아니다. 2026년 1분기 기준 스마트폰의 평균 DRAM 및 NAND 탑재 용량은 전년 대비 각각 6.1%, 10.7% 증가했다. 메모리 가격 급등에 대응해 제조사들은 보급형과 중저가 모델에서는 메모리 용량을 동결하거나 고용량 모델을 축소한 반면, 프리미엄 제품에서는 소비자 기대와 가격 인상을 정당화하기 위해 메모리 용량을 지속적으로 확대하는 전략을 취했다. 그 결과 400달러 미만 스마트폰의 평균 DRAM 용량은 감소한 반면, 프리미엄 제품의 메모리 탑재량은 더욱 증가하는 양극화 현상이 심화되었다. 따라서 스마트폰 출하량은 주로 중저가 제품을 중심으로 감소했지만, 고가 제품의 높은 메모리 탑재량이 이를 상당 부분 상쇄하면서 **Mobile DRAM의 총수요는 큰 폭으로 감소하지 않고 2026년 85.3bn Gb, 2027년은 98.7bn Gb로 소폭 증가하는 모습을 보일 것으로 전망**한다.

표 13. Mobile Dram 수요 전망

구분	2026	2027	2028	2029	2030
Mobile Dram 총수요 (bn Gb)	85.3	98.7	107.3	113	124.2
출하량 (백만개)	1,051	1,051	1,003	933	898
Contents/Box(GB)	10.14	11.74	13.35	15.06	17.18

자료: KUVIC 리서치 1팀

PC DRAM 수요 전망

PC DRAM 시장은 Mobile DRAM 과 마찬가지로 높은 성장세를 보이긴 어려울 것이라 판단된다. 2026 년 1 분기 글로벌 PC 출하량은 전년 대비 3.2% 증가한 6,330 만 대를 기록하며 지속적인 상승세를 보이고 있는 듯 했다. **하지만 이는 메모리 가격이 본격적으로 상승하기 전에 미리 구매하려는 선구매 수요와 Windows 10 지원 종료에 따른 PC 교체 수요가 일시적으로 반영된 결과**라고 해석된다. 이러한 증가세는 지속적인 시장 성장이라기보다 수요가 단기적으로 앞당겨진 결과라고 판단된다.

PC용 메모리 가격도 앞서 언급한 것처럼 높은 상승세를 보였고 2026년 1 분기에는 전 분기 대비 약 두 배 가까이 상승했다. 이에 따라 기업들은 판매가격(ASP)을 인상해 비용 부담을 소비자에게 전가하는 전략을 취했다. 최근 시장에서 많은 주목을 받았던 애플의 맥북,아이패드 가격 인상 소식만 봐도 이러한 현상을 잘 파악할 수 있다. 애플은 최근 맥북과 아이패드 주요 제품 가격을 최대 300 달러 인상했다. 팀 쿡 최고경영자(CEO)는 "부품 가격 상승을 더 이상 자체적으로 흡수하기 어렵다"고 밝혔으며 향후 추가 가격 인상 가능성도 암시했다. 이러한 가격 인상 기조에 따라 출하량 또한 감소할 것으로 보인다. **글로벌 시장 조사 업체 Omdia 는 2026 년 전 세계 데스크톱, 노트북, 워크스테이션 출하량이 전년 대비 12% 감소할 것이라는 전망을 내비쳤다.**

다만 PC 시장은 교체 수요가 여전히 남아 있다는 점에서 감소폭은 제한적일 것으로 예상된다. **현재 설치된 PC 의 약 40%가 여전히 Windows 10 또는 그 이전 버전을 사용하고 있어 Windows 10 지원 종료에 따른 교체 수요가 지속될 것으로 보인다.** 또한 쉘컴, 인텔, AMD 가 AI PC 용 프로세서를 잇달아 출시하면서 **AI PC 보급 역시 확대되고 있다.** Microsoft 는 AI 서비스를 원활하게 활용하기 위한 AI PC 의 최소 메모리 사양을 16GB 로 제시하고 있어, 향후 PC 1 대당 평균 DRAM 탑재 용량은 지속적으로 증가할 것으로 예상된다. **보수적으로 중저가 노트북은 8GB, 고성능 노트북은 16GB 로 가정하고, 워크스테이션의 경우 기본 제품은 16GB, AI PC 는 32GB 라는 가정을 바탕으로 추정했다.** 이후 각각의 출하량 비중을 반영한 결과 PC 1 대당 평균 GB는 2026년 14.4GB, 2027년은 15.4GB로 전망된다. 따라서 PC 출하량은 둔화될 가능성이 높지만, AI PC 확산과 대당 메모리 탑재량 증가에 힘입어 PC DRAM 의 총수요는 2026 년 28.2 bn Gb, 2027 년은 30.8 bn Gb 에 달할 것으로 예상된다.

표 14. PC DRAM 수요 전망

구분	2026	2027	2028	2029	2030
PC DRAM 총수요 (bn Gb)	28.2	30.8	34.4	37.4	41
출하량 (백만개)	245	250	255	258	263
Contents/Box (GB)	14.4	15.4	16.9	18.1	19.5

자료: KUVIC 리서치 1팀

Server DRAM 수요 전망

Mobile DRAM과 PC DRAM 시장이 메모리 가격 상승에 따른 출하량 둔화의 영향을 받는 반면, Server DRAM 시장은 AI 데이터센터 투자 확대에 힘입어 가장 높은 성장세를 이어갈 것으로 전망된다. 즉, AI 인프라 투자가 본격화되면서 D램 수요의 중심축이 모바일과 PC에서 서버로 이동하고 있다.

Server 디램의 수요를 추정하기 위해서는 서버 CPU에 대해 분석해야 한다. 현재 서버 CPU 시장은 앞선 내용에서 언급했듯이 인텔과 AMD가 높은 점유율을 차지하고 있지만, 엔비디아 역시 Vera CPU를 통해 본격적으로 서버 CPU 시장에 진입한 상황이다. 각 기업별 서버 CPU 포트폴리오는 인텔 Xeon 시리즈, AMD EPYC 시리즈, 엔비디아 Vera CPU로 구성된다. Server 디램 수요 추정은 3사의 주요 서버 CPU 제품별 용량을 먼저 산출한 뒤, 소켓 개수를 반영해 서버 1대당 용량으로 환산하고, 여기에 제품별 가중치와 기업별 점유율을 부여해 서버 1대당 평균 GB를 도출하는 방식으로 진행했다. 그리고 여기에 서버 출하량을 곱해서 최종 Server 디램 총수요를 도출했다.

먼저 **제품별 용량 산출은 채널 수 × DIMM per Channel × DIMM 개수 × DIMM 용량(GB)을 통해 산출했다**. 주요 제품 기준으로 보면, 인텔의 5세대 Xeon 및 Xeon 6700/6500 시리즈는 8채널 구성에 DIMM 당 64GB를 적용해 CPU 1개당 512GB, 2소켓 서버 기준 대당 1,024GB 수준으로 추정된다. AMD의 EPYC 9004(Genoa) 및 9005(Turin) 시리즈는 12채널 구성으로 CPU 1개당 768GB, 서버 1대당 1,536GB로 산출된다. 엔비디아 Vera CPU는 서버 1대당 3TB 또는 1.5TB 모두 가능한 것으로 추정된다.

기업별 점유율의 경우, AMD와 인텔은 2025년 서버 CPU 시장 점유율 비율을 그대로 적용했다. 반면 엔비디아는 성장성을 반영해 2026년 Vera CPU 판매량이 400만 개에 달할 것이라는 회사 컨센서스를 바탕으로, 전체 서버 CPU 출하량 내 비중을 계산하여 이를 점유율에 반영했다. **이에 따라 서버 1대당 평균 용량은 2026년 1,063GB, 2027년 1,134GB로 추정된다**.

서버 출하량은 Trendforce의 서버 시장 성장률 전망치를 반영해 2026년 2,300만 대, 2027년 2,500만 대 수준으로 추정했다. 따라서 **출하량과 서버 1대당 GB를 반영한 Server DRAM 총수요는 2026년 195.6 bn Gb, 2027년 226.8 bn Gb에 이를 것으로 전망된다**.

표 15. Server DRAM 수요 전망

구분	2026	2027	2028	2029	2030
D램 총수요 (bn Gb)	195.6	226.8	312.1	441.2	580.9
Set Units (백만개)	23	25	27	30	33
Contents/Box (GB)	1,063	1,134	1,418	1,795	2,145

자료: KUVIC 리서치 1팀

HBM 수요 전망

표 16. 글로벌 HBM 수요 전망 (中 제외, PB)

구분	2026E	2027E	2028E
GPU HBM 수요	4,643	5,557	7,086
ASIC HBM 수요	1,236	2,707	3,637
총 HBM 수요 (PB)	5,879	8,264	10,724
YoY 성장률 (%)	-	40.6	29.8
ASIC 비중 (%)	21.0	32.8	33.9

자료: Morgan Stanley, TrendForce, NVIDIA, Broadcom, Bloomberg, KUVIC 리서치 1팀

HBM 수요는 AI 가속기 출하량과 가속기당 HBM 탑재량으로 결정되므로, 가속기를 GPU와 ASIC 두 경로로 나누어 각각의 물량에 세대별 HBM 용량을 곱하는 방식으로 추정했다. 두 경로 모두 앞서 정리한 AI 기업 9개사의 자본지출에서 출발하며, 여기에 하이퍼스케일러 밖의 소버린·엔터프라이즈 수요를 가산했다. 계산 방식은 두 경로가 다른데, GPU는 전력을 칩 수로 환산한 뒤 세대별 HBM을 곱하는 물리적 방식을, ASIC은 자체칩에 배분된 투자액을 칩 수로 환산하는 방식을 적용했다. 이렇게 추정한 글로벌 HBM 수요(중국 제외)는 2026년 5,879PB에서 2028년 10,724PB로, **연평균 35.1% 증가**하는 것으로 나타난다. 증가분의 상당 부분은 하이퍼스케일러의 자체칩 채택 확대에서 발생해, ASIC이 전체 수요에서 차지하는 비중은 2026년 21.0%에서 2028년 33.9%로 높아진다

분석 대상 9개사는 하이퍼스케일러 4개사(구글·아마존·마이크로소프트·메타)와 네오클라우드 5개사(오라클·코어위브·네비우스·IREN·Applied Digital)로 구성했다. 빅4는 **Morgan Stanley의 회사별 증설 GW 전망**을 그대로 쓰고, GW가 제시되지 않은 네오클라우드는 자본지출을 블렌디드 \$/GW로 나눠 GW로 환산했다. 이렇게 확보한 증설 GW에서 GPU 배분분을 분리한 뒤, 전력을 칩 수로 바꾸고 세대별 HBM을 곱해 수요를 산출했다. 전력을 칩 수로 환산하는 기준은 **GB300 NVL72 랙이 GPU 72개에 약 135kW를 소비**한다는 스펙이며, 칩당 HBM은 GB300·Vera Rubin 모두 288GB로 두었다. 산출 과정을 단계로 정리하면 다음 표와 같다.

표 17. GPU HBM 산출 방식 (단계별)

단계	산출식	가정
증설 GW	빅4: MS 전망 GW 네오클라우드: CapEx ÷ 블렌디드 \$/GW	GB300 \$39B·Rubin \$49B
GPU-GW	증설 GW × GPU 배분비중	회사별 GPU 비중
GPU 칩 수	GPU-GW × 칩/GW (533,333)	GB300 랙 135kW ÷ 72GPU
GPU HBM	GPU 칩 수 × HBM/GPU (288GB)	GB300·Vera Rubin 288GB

자료: Morgan Stanley (Exhibit 10 cost/GW), Lenovo·HPE, NVIDIA, KUVIC 리서치 1팀

이 방식으로 계산하면 **9개사 GPU HBM 수요는 2026년 2,321PB, 2028년 3,106PB로 산출**된다. 물량이 우상향하는 근본 동력은 CapEx 자체의 증가이며, 이는 5대 하이퍼스케일러 CapEx가 2026년 8,083억 달러에서 2027년 1조 318억 달러로 확대되는 흐름과 방향을 같이한다. 즉 투자 증가가 곧 가속기 물량과 HBM 수요 증가로 이어지는 구조다.

NVIDIA 데이터센터 매출이 FY27 1분기 기준 하이퍼스케일러 약 50%, 나머지 절반이 소버린·엔터프라이즈 등 비-하이퍼스케일로 구성된다는 점에 착안해, 9개사 GPU HBM에 이 비율을 적용해 하이퍼스케일러 밖 수요를 가산했다. 가산 비중은 소버린 매출이 전년 대비 80% 이상 성장하는 흐름을 반영해 연도가 갈수록 확대되도록 설정했다. 다만 이 부분은 개별 프로젝트 집계가 아니라 세그먼트 비율에 기반한 추정이므로, 규모의 정밀도보다 비-하이퍼스케일 수요가 구조적으로 커지는 방향성에 무게를 두었다. 그 결과 GPU HBM 총수요는 2026년 4,643PB에서 2028년 7,086PB로 추정된다.

표 18. GPU HBM 수요 상세

구분	2026E	2027E	2028E
데이터센터 증설 GW (GPU)	15.4	18.8	24.0
HBM per GPU (GB)	288	288	288
9사 GPU HBM (PB)	2,321	2,640	3,106
(+) 소버린·엔터·Non-HS (PB)	2,321	2,918	3,981
GPU HBM 총수요 (PB)	4,643	5,557	7,086

자료: Morgan Stanley, NVIDIA, KUVIC 리서치 1팀

ASIC은 GPU와 달리 랙 단위 전력 스펙이 정형화돼 있지 않아, 빅4의 증설 GW에서 ASIC 배분을 분리한 뒤 이를 칩 수로 환산하고 칩별 HBM을 곱하는 방식을 적용했다. ASIC 비중은 다세대에 걸쳐 자체칩을 운영 중인 구글·아마존은 높게, Maia·MTIA가 초기 단계인 마이크로소프트·메타는 낮게 두었다. 칩별 HBM 탑재량은 각 사가 공개한 스펙을 사용했으며, 세대 전환에 따라 용량이 상향되도록 반영했다.

이렇게 계산한 ASIC HBM 수요는 2026년 1,236PB에서 2028년 3,637PB로 연평균 71.5% 증가해 GPU보다 빠른 속도를 보인다. 회사별로는 자체칩이 가장 성숙한 구글이 전 구간에서 가장 큰 비중을 차지하고, 아마존이 뒤를 잇는다. 자체칩 비중이 상승하는 방향은 추론 확대에 따른 비용·전력 효율 우위에서 비롯되며, 아마존이 컨퍼런스콜에서 자사가 들여오는 AI 칩 중 가장 많은 수가 Trainium이라고 언급한 것과도 방향이 같다.

표 19. ASIC HBM 수요 상세 (PB)

구분	2026E	2027E	2028E
ASIC 증설 GW (4사)	8.2	12.5	18.2
구글 TPU	914	1,890	2,270
아마존 Trainium	269	533	782
마이크로소프트 Maia	54	116	297
메타 MTIA	0	169	289
ASIC HBM 총수요 (PB)	1,236	2,707	3,637

자료: Google·AWS·Microsoft·Meta (제품 스펙), Broadcom 1Q26 컨퍼런스콜, TrendForce, KUVIC 리서치 1팀

DRAM 공급추정

HBM bit 공급 전망

표 20. 메모리 3사 HBM 공급 추정

	2026E	2027E	2028E
웨이퍼 투입 (kwpm)			
SK하이닉스	190	250	310
삼성전자	180	250	300
마이크론	90	110	130
GB/웨이퍼			
SK하이닉스	994	1096	1203
삼성전자	895	1000	1095
마이크론	945	1039	1120
HBM bit 공급(GB bn)			
SK하이닉스	2.27	3.29	4.48
삼성전자	1.93	3.00	3.94
마이크론	1.02	1.37	1.75
3사 합계	5.22	7.66	10.16
산업 총계	5.42	7.86	10.46

자료: KUVIC 리서치 1팀

공급량 기본 계산식은 다음과 같다.

공급 (GB bn) = 웨이퍼 투입(kwpm) × 12 × 넷다이 × 수율 × 다이용량 / 10⁻⁶. GB 계산에 있어 적층 단수는 약분되어 사라지므로 고려하지 않았다.

SK하이닉스는 3사 중 최대 HBM 캐파를 유지할 것으로 보인다. **2025년 실적 기준 동사의 HBM 캐파는 153k wpm이었으며, M15X 램프업과 용인 1기 팹 가동을 반영할 경우 2028년까지 약 2배가 확대된다.** 이는 최태원 회장이 컴퓨텍스에서 5년 내 HBM 웨이퍼 캐파 2배를 언급한 점과 방향이 일치한다. 삼성전자는 지난 2024년 및 2025년, 투입량 월 100kwpm 부근의 수준인 2위 사업자였으나, HBM4 양산 시작과 투자 확대 기조에 따라 SK하이닉스에 근소하게 뒤지는 2위 수준으로 배치하였다. 마이크론은 후발주자로서 아이다호 신규 팹의 2027년 중반 가동과 히로시마 증설 일정을 반영해 상대적으로 가파른 증가율을 부여하였으나 28년 웨이퍼 투입량이 높게 인식됨에 따라 비중은 상대적으로 줄어드는 것으로 가정했다.

CXMT를 포함한 기타 회사 점유율도 존재하겠으나, 앞서 설명한 것처럼 고대역폭 메모리의 기술 장벽이 매우 크고 대중 수출 규제를 감안할 때 해당 물량이 실제 고사양 차세대 HBM으로 출하될 가능성은 제한적이고, 이는 의미있는 수치적 반영은 아니라 판단하였다. 따라서 2025년도 비트 공급량으로 flat하게 가정, 28년에만 공격적 증설을 반영 일부 증가한 수치를 가정하였다.

실제 메모리 3사 DRAM 캐파 추정치 대비 HBM 비중으로 검증해보아도, 25년 30% 수준인 SK하이닉스는 동사가 밝힌 27년 비중 40% 목표에 정합하고, 이는 HBM 시장 1위 사업자라는 점에서도 타당한 가정이라고 판단했다.

수율의 경우 회사 내부 데이터로서 정확한 공시가 되어있지 않고 발표되더라도 그 자체의 신뢰성 검증에도 어려움이 있다는 점이 큰 변수이다. **다만 출시 후 곧바로 신규 양산에 진입하는 HBM4, HBM4E 등의 경우 수율은 50% 수준이 일반적이고, 수년간 양산이 진행된 HBM3, HBM3E의 경우 70%대 수준까지도 안정화될 수 있다는 점을 고려했다.** 또한 SK하이닉스는 이미 성숙공정인 1b 공정을 HBM4 코어다이에 적용하여 수율을 높이는 한편, 삼성전자의 경우 선제적으로 1c 공정을 적용함에 따라 초기 수율 리스크를 감수하므로 이를 고려했다. 마이크론의 경우 HBM4E부터는 1γ공정을 적용하나, HBM4의 경우 성숙공정인 1β공정을 활용한다는 점을 고려하여 수율을 가정하였다.

웨이퍼 한 장에서 확보되는 넷다이의 개수의 경우 기하학적인 계산 결과이나, 세대별로 다이 면적이 상이하다. 기존 HBM3, HBM3E의 경우 $100\text{mm}^2 \sim 110\text{mm}^2$ 수준이었으나, 세대가 발전함에 따라 인터페이스 폭이 1024 bit에서 2048bit로 확대되면서 TSV 개수가 증가하고 코어다이 면적이 확대되기 때문에, 다이가 커지면 웨이퍼당 확보 가능 개수가 줄어든다. TSV 하나가 차지하는 순수 지름은 수 마이크로미터 수준이긴 하나, TSV 사이 간격을 유지해야하며, I/O 개수만 뚫는 것이 아닌 전력, 접지 등을 이유로 TSV를 추가적으로 뚫어줘야 하므로, 통상 10~15% 면적을 차지한다고 가정하여 HBM4의 다이 면적 증가율을 계산하였다. 이에 따른 세대별 넷다이 산업 표준 GDPW(Gross Die Per Wafer) 공식을 적용했다.

$$GDPW = \frac{\pi \cdot r^2}{A} - \frac{\pi \cdot d}{\sqrt{2A}}$$

이에 따른 HBM 총 공급량은 26년에 5.22GB bn, 27년 7.66GB bn, 28년 10.16GB bn와 같다.

범용 DRAM bit 공급 전망

표 21. 범용 DRAM 공급량

	2026E	2027E	2028E
범용 DRAM 공급(GB bn)	35.51	42.51	53.21
삼성전자	13.92	15.60	17.91
SK하이닉스	10.56	12.95	16.40
마이크론	5.81	7.41	10.68
CXMT	4.03	5.35	7.01
기타	1.2	1.2	1.2

자료: KUVIC 리서치 1팀

기본 계산식은 HBM과 동일한 Bottom-up 방식을 활용했고 투입 웨이퍼는 전체 DRAM 웨이퍼 캐파에서 HBM 라인으로 전환되는 웨이퍼를 차감하여 순수 범용 DRAM 캐파를 산출하였다.

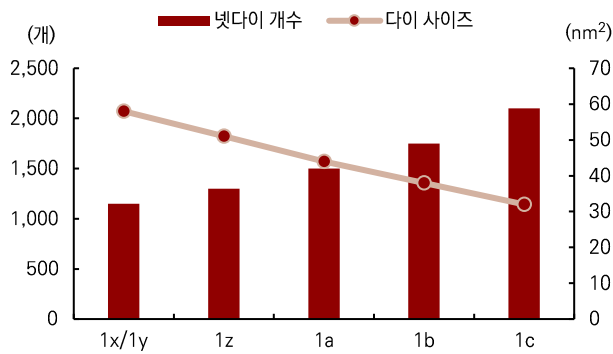
미세화 공정 진전에 따른 생산성 변화를 반영하기 위해 각 노드별 넷다이와 수율을 가정하였다. 구형 레거시(1x/1y) 노드는 웨이퍼당 넷다이 1,150개 및 다이 용량 1GB를 기준으로 설정했으며, 1z 노드는 넷다이 1,300개(2GB/die), 1a 노드는 1,500개(2GB/die), 1b 노드는 1,750개(2GB/die)로 산정하였다. 최선단 공정인 1c 노드의 경우 기존 대비 약 15~16%의 다이 사이즈 축소를 반영하여 웨이퍼당 넷다이를 2,100개(2GB/die)까지 확대 적용하였다. 해당 수치들은 모두 노드 별 다이 사이즈를 기준으로 역산하여 계산한 값이다.

통합 수율의 경우 기술 숙련도가 확보된 레거시부터 1b 노드까지는 연도별 90%의 안정적인 고수율을 유지하는 것으로 산정하였다. 반면 본격적인 램프업 단계에 진입하는 1c 노드는 초기 수율 80%(2026E)에서 시작하여 85%(2027E), 90%(2028E)로 단계적으로 개선되는 수율 곡선을 적용하였다. 마이크론의 경우 경쟁사 대비 공정 안정화 시차가 존재한다는 점을 고려하여 전체 수율 산출 시 오프셋을 적용했다.

메모리 3사의 노드 별 비중은 2022년~2025년 사이 노드 별 비중 변화에 대한 Trendforce 수치를 Historical 데이터로 인용하여 노드는 대략 2년 간격으로 도입되며, 각 노드 점유율은 pilot, 램프업, peak, 비중 축소의 s 커브 곡선을 이루는 비중 변화가 발생한다는 점을 기준으로 했다. 기존의 추이와, 이 중 HBM 생산에 투입되는 1b, 1c 공정 비중만큼은 제외하여 노드 별 비중을 산정하였다.

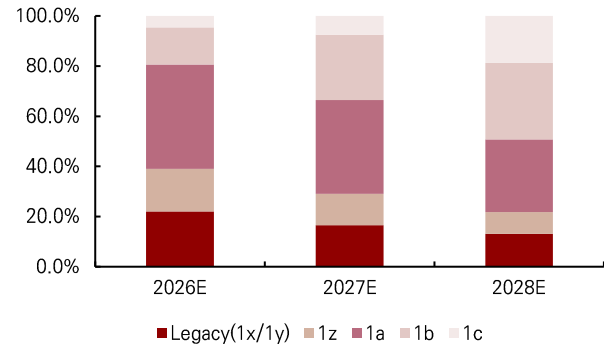
CXMT의 경우 기술적 격차가 아직 크게 존재한다는 점에서 그대로 활용하는 것이 아닌 SemiAnalysis의 조사에 따른 비트 밀도 수치와 별도의 노드 비중을 활용하여 비트 밀도, 수율, 웨이퍼당 가용 면적, 총 투입 웨이퍼를 곱하여 산출하였다.

그림 36. 넷다이 개수 및 다이 사이즈 추정치



자료: KUVIC 리서치 1팀

그림 37. 연도별 노드 믹스 추정



자료: KUVIC 리서치 1팀

이에 따른 범용 DRAM 및 전체 DRAM bit 공급량 추정치는 아래와 같다.

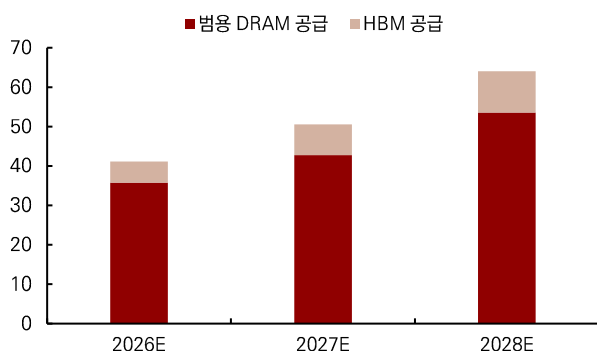
표 22. 메모리 3사 DRAM 공급량 내 HBM 비중

	2026E	2027E	2028E
범용 DRAM 공급(GB bn)	35.73	42.73	53.55
HBM 공급(GB bn)	5.42	7.86	10.46
전체 DRAM 공급(GB bn)	41.15	50.59	64.01
HBM 공급량 비중	13.2%	15.5%	16.3%

자료: KUVIC 리서치 1팀

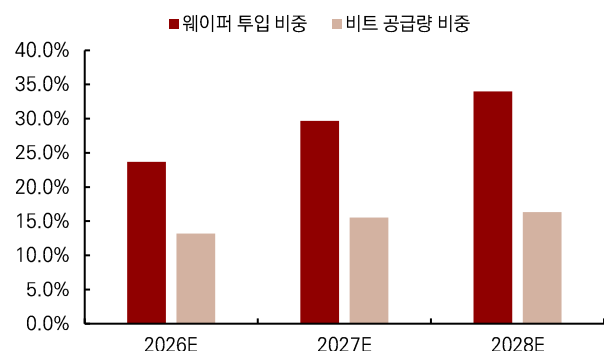
HBM은 웨이퍼 투입량 대비 실제 총 공급량(GB) 기여도가 현저히 낮다는 구조적 한계를 지닌다. 일반 범용 D램에 비해 다이 크기가 커서 웨이퍼 1장당 생산할 수 있는 넷다이 수가 급감하는 데다, TSV 등 고난도 패키징 공정으로 인해 수율 손실이 겹치기 때문이다. 특히 SK하이닉스의 경우 웨이퍼 캐파의 40% 이상을 HBM에 할당할 정도로 HBM 생산에 막대한 팹(Fab) 자원이 쏠리고 있음에도, **실제 시장에 출하되는 용량(GB) 기준 비중은 투입량에 크게 못 미치는 극심한 '웨이퍼 잠식 현상'이 발생한다.** 결과적으로 HBM의 거대한 웨이퍼 흡수력은 수익성 개선의 원동력이지만, 동시에 범용 D램의 생산 여력을 원천적으로 고갈시켜 향후 전체 메모리 시장의 구조적인 공급 부족을 촉발한다.

그림 38. DRAM 공급량 (GB bn)



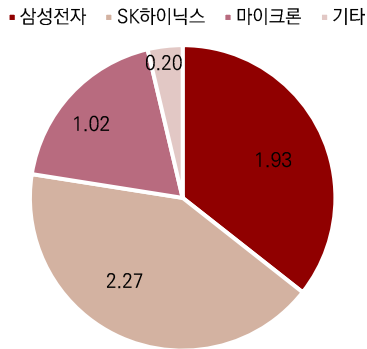
자료: KUVIC 리서치 1팀

그림 39. 웨이퍼 투입량 및 비트 공급량 내 HBM 비중



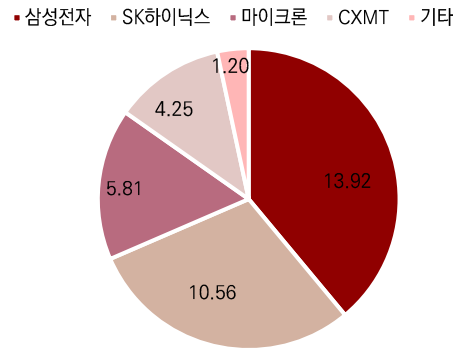
자료: KUVIC 리서치 1팀

그림 40. 회사별 HBM bit 공급량 (GB bn)



자료: KUVIC 리서치 1팀

그림 41. 회사 별 범용 DRAM bit 공급량(GB bn)



자료: KUVIC 리서치 1팀

NAND 시장 패러다임 변화 및 수급 분석

기술 및 제품 로드맵의 진화

SLC(1비트)~QLC(4비트): 비트 수↑ → 원가↓, 내구성↓

NAND는 하나의 셀에 몇 비트를 저장하느냐에 따라 SLC(1비트), MLC(2비트), TLC(3비트), QLC(4비트)로 나뉜다. 셀당 비트를 늘릴수록 같은 면적에서 뽑아내는 용량이 커져 GB당 원가는 내려가지만, 전압 구간을 더 잘게 쪼개야 하므로 쓰기 내구성과 속도, 신뢰성은 함께 저하된다. 결국 NAND 제품 전략은 원가와 내구성 사이의 어느 지점을 택할 것인가의 문제다.

데이터센터
주력=TLC, 아카이브
=QLC 침투 가속

현재 데이터센터의 주력은 TLC다. 삼성전자는 상대적으로 고수익이 기대되는 차별화 성능 TLC 제품군으로 서버 SSD 시장에 대응하겠다고 밝혔고, 마이크론 역시 데이터센터 믹스에서 TLC와 QLC를 고객 요구에 따라 병행 운영한다고 설명했다. 반면 읽기 비중이 압도적으로 높은 고용량 아카이브-컨텍스트 저장 영역에서는 QLC 침투가 빠르게 진행 중이다. SK하이닉스는 2026년 4월 세계 최초로 321단 QLC 개발과 고객 인증을 완료했고, 마이크론은 분기 기준 사상 최고 QLC 비트 믹스를 기록하며 업계 최초의 Gen5 QLC 클라이언트 SSD를 출시했다. 샌디스크의 QLC 기반 데이터센터 드라이브도 출하가 시작됐다.

한편 SLC·MLC는 산업용·네트워크 장비 등 니치 영역으로 남았지만, 2026년 들어 각국의 통신망과 데이터센터 구축이 본격화되며 수요가 되살아났다. 공급사들이 TLC·QLC 전환에 개파를 집중하며 SLC 생산을 줄여 온 탓에, 오히려 SLC의 가격 상승폭이 다른 제품군을 상회하는 현상이 나타나고 있다. DDR5 전환 과정에서 DDR4가 겪었던 구축 효과와 동일한 구조다.

표 8. 셀 타입별 NAND 비교와 2026년 동향

구분	SLC	MLC	TLC	QLC
셀당 비트	1비트	2비트	3비트	4비트
상대 내구성(P/E)	~10만회	~1만회	~3천회	~1천회
상대 원가(GB당)	최고	높음	중간	최저
주 용도	산업용·네트워크·부트	니치 산업용	서버 eSSD 주력, 모바일	고용량 아카이브·읽기 중심
2026년 동향	전환 여파로 공급 축소, 가격 상승폭 최대	비중 지속 축소	삼성 V9 기반 서버 SSD SKH 321단 QLC 세계 중심 대응	최초, MU 기록적 믹스

자료: 각 사 실적발표, 업계 자료, KUVIC 리서치 1팀

웨이퍼 증설 대신
단수 전환으로 비트
확보 → 웨이퍼
본딩이 면적효율 핵심
수단

공급 제약이 구조적인 만큼, 제조사들은 웨이퍼의 물리적 증설보다는 '단수 전환'을 통해 비트(Bit) 생산량을 확보하는 전략에 사활을 걸고 있다. SK하이닉스는 176단에서 321단으로 두 세대를 뛰어넘는 전환을 추진 중이며, 마이크론 역시 G9 노드 비중을 과반으로 올릴 전망이다. 다만, 고단화에 따른 식각 종횡비(Aspect Ratio) 난도를 극복하기 위해, 페리(Peripheral) 회로를 셀 어레이와 분리해 접합하는 '웨이퍼 본딩' 구조가 면적 효율과 수율 확보의 핵심 수단으로 자리 잡고 있다.

또한, 인터페이스와 폼팩터도 함께 진화하고 있다. 삼성전자는 하반기 신규 GPU 플랫폼 도입과 함께 PCIe 6세대

서버 시장이 급격히 확대될 것으로 보고 V9 기반 고성능 제품으로 초기 시장을 주도하겠다고 밝혔으며, 마이크론은 이미 G9 기반 PCIe Gen6 데이터센터 SSD를 양산 중이다. 용량 축에서는 마이크론의 122TB 고용량 SSD가 동일 용량의 HDD 구성 대비 와트당 순차 읽기 처리량이 16배에 달한다고 밝히며 HDD 대체 논리를 강화하고 있다..

떨리지 않는 공급 병목과 장기계약(SCA) 구조로의 전환

NAND 수요의 질적·양적 폭발에도 불구하고, 실제 시장에서 체감하는 공급 증가폭은 극도로 제한적이다. 이로 인해 가격 폭등과 더불어 NAND 시장의 거래 방식 자체가 근본적으로 재편되고 있다.

병목=클린룸,
증설해도 가동은
2028~29년 → 공급
확대 어려움

현재 NAND 공급 제약의 본질적인 원인은 웨이퍼 부족이 아니라 '클린룸(Cleanroom) 공간의 고갈'이다. 일부 공급사들이 NAND용 클린룸 공간을 D램으로 전환하고 있으며, 전체적인 클린룸 부족으로 인해 2026년 산업 전체의 **NAND 비트 출하 성장률은 20% 내외로 제약될** 전망이다. 마이크론의 싱가포르 신규 NAND 팹 착공에 따른 초기 웨이퍼 산출 시점과 삼성전자의 평택 P5 낸드 라인 가동이 모두 2028~2029년으로 예상되어, 당분간 물리적인 공급 확대 폭은 제한적일 수밖에 없다.

NAND 3Q 계약가
+35~40%, DRAM
을 상회하는 가격
상승

이러한 극심한 수급 불균형은 가격에 즉각 반영되고 있다. 2026년 3분기 계약가는 전분기 대비 35~40% 인상이 통보된 것으로 전해지는데, 이는 TrendForce가 앞서 제시한 10~15% 전망을 크게 상회하는 수치다. **UBS 역시 3분기 NAND 계약가 전망치를 기존 +17%에서 +30%로, 4분기는 +12%로 조정했다.** 샌디스크가 선제적으로 가격을 올린 뒤 나머지 공급사가 순차 동참하는 흐름이며, 기업용 SSD와 AI 스토리지 수요, 그리고 공급사의 의도적인 물량 통제가 배경으로 지목된다.

실적에도 그대로 반영됐다. 마이크론의 회계연도 3분기(캘린더 2026년 2분기) NAND 매출은 99억달러로 전년 동기 대비 361% 증가했고, 가격은 전분기 대비 80% 중반 상승했다. 데이터센터 SSD 매출만 50억달러를 넘기며 전분기 대비 두 배 이상 늘었다. SK하이닉스도 1분기 NAND ASP가 70% 중반 상승했으며, 출하량은 고부가 제품 믹스 전환과 생산 리드타임 증가로 오히려 10% 감소했다. BofA는 2026년 NAND 산업 매출을 3,231억달러(YoY +299%), 2027년을 4,204억달러로 추정하며, 같은 기간 마이크론과 키옥시아의 NAND 영업이익률을 각각 76%, 70% 수준으로 전망한다. 2026년 3분기 NAND 계약가는 전분기 대비 35~40% 인상이 통보되며 D램의 가격 상승폭을 크게 웃돌았다. 이는 마이크론(회계연도 3분기 NAND 매출 361% 증가)과 SK하이닉스(1분기 NAND ASP 70% 중반 상승) 등 주요 메이커들의 기록적인 실적 성장으로 직결되고 있다.

결과적으로 공급사의 의도적인 물량 통제와 불안정성에 직면한 고객사들의 니즈가 맞물려 **계약 구조의 패러다임이 변화하고 있다.** 마이크론은 해당 기간 NAND 물량의 약 3분의 1에 해당하는 16건의 전략적 고객계약(SCA)을 체결했으며, 샌디스크 역시 선제적 가격 인상과 함께 FY27 비트 공급의 3분의 1 이상을 장기계약으로 확정 지었다. 총이익률 80% 수준에서 형성되는 장기 계약 비중이 높아질수록 다운사이클에서의 이익 하방이 두꺼워지고, 셀 메이커는 증설에 나설 근거를 얻는다. 다만 모건스탠리가 DRAM과 레거시 메모리를 NAND보다 선호한다고 밝힌 것처럼, ASP 상승을 자체는 정점에 근접하고 있다는 시각도 공존한다. NAND 투자 논리의 무게중심이 가격에서 물량과 계약의 지속성으로 옮겨가는 국면이다

NAND 수요 추정

표 23. NAND 수요 전망 (AI / 비-AI)

구분	2025E	2026E	2027E	2028E
Total NAND 수요 (EB)	1,110	1,282	1,530	1,870
AI NAND	205	432	656	964
비-AI NAND	905	850	874	907
YoY 성장률 (%)	-	15.5	19.4	22.2
AI 비중 (%)	18.5	33.7	42.9	51.5
NAND 공급 (EB)	1,062	1,254	1,501	1,725
충족률 (공급/수요-1, %)	-4.3	-2.1	-1.9	-7.8

자료: Morgan Stanley, BofA Global Research, Bloomberg, KUVIC 리서치 1팀

NAND 수요는 DRAM과 달리 애플리케이션별 세트 출하량만으로는 설명되지 않는다. AI 데이터센터용 스토리지 수요가 기존 애플리케이션 수요와 전혀 다른 함수(가속기 출하량과 학습 데이터 규모)로 결정되기 때문이다. 이에 본 모델은 NAND 수요를 AI NAND와 비-AI NAND로 이원화하여, AI NAND는 AI CapEx에서 출발하는 bottom-up 방식으로, 비-AI NAND는 애플리케이션별 방향성 추정 방식으로 각각 산출한 뒤 합산하였다. 그 결과 전체 NAND 수요는 2026년 1,282EB, 2027년 1,530EB로 각각 전년 대비 15.5%, 19.4% 증가하며, 전체 수요 증분의 대부분이 AI에서 발생해 AI 비중은 2025년 18.5%에서 2027년 42.9%, 2028년 51.5%까지 확대될 전망이다.

AI NAND 수요 전망

AI NAND 수요는 다음의 논리 체인을 따라 추정하였다. ① Total CapEx → ② AI 인프라 CapEx → ③ AI 가속기 CapEx → ④ 가속기 출하량 → ⑤ 가속기당 NAND 탑재량 → ⑥ AI NAND 수요(EB). Total CapEx는 하이퍼스케일러·네오클라우드 9사와 OpenAI/Stargate를 대상으로 한 Bloomberg BEST_CAPEX 컨센서스를 기준으로 2025년 \$532B, 2026년 \$1,120B, 2027년 \$1,409B를 적용했다. 여기에 Goldman Sachs 및 SemiAnalysis가 제시한 AI 인프라 비중 75%를 곱해 AI 인프라 CapEx를 산출하고, 다시 Goldman Sachs의 AI 가속기 지출 추정치(\$180B/\$450B 40%)를 적용해 가속기 CapEx를 도출하였다. 이를 Bloomberg Intelligence(AI Accelerator Chips 2026 Outlook, 2026.1)의 blended ASP(2026년 \$15.7K, 2027년 \$17.6K)로 나누면 가속기 출하량은 2026년 21.4백만 개, 2027년 24.0백만 개로 추정된다.

핵심 가정은 가속기 1개당 NAND 탑재량이다. 이는 (i) 가속기 노드에 직결되는 로컬 NVMe와 (ii) 학습 데이터셋·체크포인트를 담는 공유 스토리지 tier로 구분된다. 로컬 NVMe는 NVIDIA DGX/HGX 시스템 스펙(8×3.84TB U.2 SSD ÷ 8 GPU)을 근거로 GPU당 3.84TB로 세대 불변 고정하였다. 반면 공유 tier는 학습 규모 확대에 따라 급증하는 구간으로, 병렬 파일시스템(체크포인트·데이터셋)을 로컬 대비 3.5배로 앵커링하고, 여기에 NVIDIA ICMS(Inference Context Memory Storage) 도입분을 더했다. ICMS는 BlueField-4 DPU 4개×150TB = 어플라이언스당 600TB, 트레이 16개 기준 9,600TB 구성이며, 8랙×72 = 576 GPU 기준 GPU당 16.7TB에 해당한다. Rubin 세대 침투율을 2026년 10%(2H26 램프), 2027년 40%, 2028년 65%로 가정해 반영했다. 그 결과 가속기당 총 NAND는 2026년 20.1TB, 2027년 27.3TB, 2028년 34.0TB로 확대된다.

표 24. AI NAND 수요 전망

구분	2025E	2026E	2027E	2028E
AI 가속기 CapEx (\$B)	160	336	423	529
Blended ASP (\$K)	-	15.7	17.6	18.7
가속기 출하량 (M units)	-	21.4	24.0	28.3
로컬 NVMe (TB/가속기)	3.8	3.8	3.8	3.8
공유 스토리지 (TB/가속기)	13.3	16.3	23.5	30.2
가속기당 NAND (TB)	17.1	20.1	27.3	34.0
AI NAND 수요 (EB)	205	432	656	964
YoY 성장률 (%)	-	110.7	52.0	46.8
(참고) MS AI NAND TAM (EB)	205	400	609	n/a

자료: Bloomberg, Goldman Sachs, Morgan Stanley, NVIDIA, KUVIC 리서치 팀

비-AI NAND 수요 전망

비-AI NAND는 Morgan Stanley의 NAND Sufficiency Estimates 내 애플리케이션별 분해를 기본 골격으로 삼고, BofA Global Research의 애플리케이션별 수요 전망으로 방향성을 교차검증하였다. 결론적으로 비-AI NAND 수요는 2026년 850EB로 전년 대비 6.1% 감소한 뒤 2027년 874EB, 2028년 907EB로 완만히 회복하는 데 그친다. 즉 비-AI 영역은 사실상 정체 국면이며, 전체 NAND 수요 성장은 전적으로 AI가 견인하는 구조다.

세부적으로는, PC SSD는 2025년 209EB에서 2026년 181EB로 감소한다. DRAM과 동일하게 메모리

가격 급등이 세트 ASP 인상으로 전이되면서 출하량이 위축되기 때문으로, Omdia는 2026년 데스크톱·노트북·워크스테이션 출하량이 전년 대비 12% 감소할 것으로 전망하고 있다. Windows 10 지원 종료에 따른 교체 수요가 하방을 지지하나, NAND의 경우 세트 업체들이 원가 부담을 흡수하는 과정에서 중저가 모델의 기본 SSD 용량을 축소하는 방향으로 대응하고 있어 대당 용량 증가 효과가 DRAM 대비 제한적이다. **Enterprise SSD(비-AI 서버)는 187EB→206EB→226EB로 유일하게 견조한 증가세를 보인다.** QLC 기반 대용량 SSD가 nearline HDD를 대체하는 구조적 흐름이 지속되기 때문이며, BofA 역시 서버용 SSD 수요를 가장 높은 성장 라인으로 제시하고 있다. **스마트폰은 2025년 337EB에서 2026년 291EB로 감소한다.** TrendForce가 전망한 2026년 스마트폰 생산량 16.2% 감소가 직접적 요인이거나, 대당 NAND 탑재량이 전년 대비 10.7% 증가하며 감소폭을 상당 부분 상쇄해 2027년에는 301EB로 반등한다. **Tablets(34EB), Flash Cards(15EB), USB(10EB), Other(113EB)는 이미 성숙 단계에 진입한 영역으로 별도의 성장 동인을 부여하지 않고 보합으로 가정하였다.** 2028년은 컨센서스 부재 구간이므로 라인별 성장률(PC +2%, Enterprise +9%, 스마트폰 +3%, 기타 0%)을 외삽하였으며, 이는 BofA의 애플리케이션별 추세와 방향성이 일치한다.

표 25. 비-AI NAND 수요 전망

구분	2025E	2026E	2027E	2028E
PC SSD (노트북+데스크톱)	209	181	175	179
Enterprise SSD (비-AI 서버)	187	206	226	246
Smartphones	337	291	301	310
Tablets	34	34	34	34
Flash Cards	15	15	15	15
USB Flash Drives	10	10	10	10
Other (임베디드·산업·기타)	113	113	113	113
비-AI NAND 총수요 (EB)	905	850	874	907
YoY 성장률 (%)	-	-6.1	2.8	3.8

자료: Morgan Stanley, BofA Global Research, TrendForce, Omdia, KUVIC 리서치 1팀

NAND 공급추정

NAND 공급은 웨이퍼 캐파 추정치를 통한 캐파 성장률에 적층단수 전환과 수율 개선을 고려한 비트 효율 개선을 추정치를 곱해 비트 성장률을 추정하였다. NAND 웨이퍼 캐파는 펄스 별 웨이퍼 캐파 추정치를 합산하였다. 삼성전자(시안, 평택 등), 키옥시아/샌디스크(윌카치, 기타카미), SK하이닉스/솔리다임(청주,대련), 마이크론(싱가포르), YMTC(우한) 등의 개별 웨이퍼 캐파 추정치를 합산하였다.

글로벌 웨이퍼 캐파는 2025년 1,651(k wpm)에서 2026E, 2027E, 2028E가 각각 1,726(k wpm), 1,796(k wpm), 1,878(k wpm)로 연간 +5% 수준으로 완만하게 증가할 것으로 예상했다.

NAND 업체들이 신규 웨이퍼 증설보다 적층단수 전환(공정 전환)을 통한 비트 성장에 초점을 맞추고 있고, 또한 일부 업체들은 낸드 라인 자체를 D램-HBM으로 전환하고 있기에 NAND에서의 웨이퍼 캐파 증가율은 미미할 것으로 예상된다.

웨이퍼당 비트출하 효율 개선율은 2023~2025년의 평균 개선률인 16.7%를 기준으로 2026E +13%, 2027E +15%, 2028E +10%를 가정하였다. 2026년의 경우 기술적으로는 삼성이 아직 V9(286단)이 아니라 V8(236단) 위주였고, SK하이닉스 321단(QLC)과 키옥시아 BiCS10(332단)은 막 램프업을 시작한 단계이다. 밀도 기준으로 232단→321단 전환 시 TLC는 +20%, QLC는 +44%까지 밀도가 오를 여력이 있는데, 아직 물량 비중이 크지 않은 채로 점진적으로 반영되는 국면이라 13%대의 완만한 개선을 가정하였다. 2027년의 경우 NAND 업체들의 신규 전환과 캐파 확대가 몰린다. 286단 QLC(V9)는 전세대 QLC(19.8) 대비 +44%의 가장 큰 밀도 점프를 기록하는데, 이 최대 폭의 전환이 신규 캐파 가동과 같은 해에 겹쳐 +15%의 개선을 가정하였다. 2028년의 경우 밀도로 볼 때 400단+ TLC가 사실상 286단 QLC와 사실상 비슷하다. 총수는 100단 넘게 늘어나도 셀당 저장 비트가 4비트(QLC)에서 3비트(TLC)로 줄면 밀도

이득이 거의 상쇄된다. 엔터프라이즈 SSD처럼 신뢰성이 중요한 수요가 늘면서 믹스가 QLC에서 다시 TLC 쪽으로 이동할 가능성이 있고, 여기에 마이크론은 2026~27년 내내 캐파가 고정돼 있어 이 회사의 공정전환 기여가 사실상 전무하다. 마이크론 싱가포르 신규 팹(Fab 10B)도 2H28에야 첫 가동해 그 해 기여는 제한적이다. 새로운 대형 레이어 점프(500단+)는 아직 로드맵상 확정되지 않아, 2028년은 "다음 점프 전 소강 국면"으로 설명할 수 있다. 따라서 +10%의 개선을 가정하였다.

따라서 예상되는 2026~2028년 NAND의 비트출하량은 매년 16~20%씩 늘어나 각 9.3→47.0→54.3(bn units)으로 추정된다. 이는 EB단위로 환산했을 때 1,254EB→1,501EB→1,725EB 수준이다.

메모리 수요 공급 쇼티지 분석

앞서 각각 추정한 수요와 공급을 하나의 표로 결합해 연도별 **수요충족률**을 산출해보면 다음 표와 같다.

표 26. 전체 DRAM 수급 및 충족률 전망 (GB bn)

구분	2026E	2027E	2028E
전체 DRAM 수요	52.8	64.3	81.0
범용 DRAM 수요	46.9	56.0	70.3
HBM 수요	5.9	8.3	10.7
전체 DRAM 공급	41.2	50.6	64.0
범용 DRAM 공급	35.7	42.7	53.6
HBM 공급	5.4	7.9	10.5
전체 DRAM 충족률 (%)	-22.0	-21.3	-20.9
범용 DRAM 충족률 (%)	-23.8	-23.7	-23.8
HBM 충족률 (%)	-7.8	-4.9	-2.5

자료: KUVIC 리서치 1팀

추정 결과 DRAM과 NAND 모두 2026~2028년 내내 수요충족률이 마이너스에 머무는 것으로 나타나며, 이는 이번 사이클의 공급 부족이 일시적 재고 조정이 아니라 캐파 제약에서 비롯된 현상임을 시사한다.

DRAM - 총량 부족과 HBM의 웨이퍼 잠식

전체 DRAM 수요는 범용 DRAM 수요(2026년 46.9 GB bn → 2028년 70.3 GB bn)와 HBM 수요(2026년 5.9 GB bn → 2028년 10.7 GB bn)를 합산해 산출하였다. **그 결과 전체 DRAM 수요는 2026년 52.8 GB bn에서 2028년 81.0 GB bn으로 연평균 24%가량 확대되는 반면, 공급은 41.2 GB bn에서 64.0 GB bn 증가에 그친다.** 전체 DRAM 충족률은 2026년 -22.0%, 2027년 -21.3%, 2028년 -20.9%로 3년 내내 20%를 상회하는 큰 폭의 부족을 이어간다. 부족 폭이 소폭씩 줄어드는 것은 메모리 3사의 증설과 공정 전환이 공급을 끌어올리기 때문이나, 출발선 자체가 -22%에 달한다는 점에서 초과수요 국면이 상당 기간 구조적으로 지속될 전망이다.

전체 지표 이면에는 범용 DRAM과 HBM이 서로 반대 방향으로 움직이는 이중 구조가 숨어 있다. 범용 DRAM 충족률은 2026년 -23.8%에서 3년 내내 거의 개선되지 않는 만성 부족에 머무는 반면, HBM 충족률은 2026년 -7.8%에서 2027년 -4.9%, 2028년 -2.5%로 부족 폭이 꾸준히 좁혀진다. **즉 HBM은 3사의 캐파 확대와 수율 개선이 겹치며 균형에 서서히 다가가는 반면, 범용 DRAM은 오히려 정체된 부족 상태를 벗어나지 못한다.** 그 이유는 HBM이 동일 용량을 생산하는 데 **범용 대비 웨이퍼를 3~4배 소모**하기 때문이다. 한정된 웨이퍼 캐파 안에서 HBM 캐파 확대가 최우선순위로 배정되면서, 그만큼 범용 DRAM에 돌아갈 웨이퍼가 줄어든다. 결국 이번 DRAM 부족의 실체는 HBM 자체의 공급난이 아니라, HBM 우선 배분이 범용 DRAM의 생산 여력을 지속적으로 잠식하는 웨이퍼 배분의

문제라고도 볼 수 있다. 수요 측면에서도 서버 DRAM이 2026년 이후 범용 수요 증분의 대부분을 견인하는 만큼, AI 인프라 투자가 이어지는 한 범용 DRAM의 만성 부족은 쉽게 해소되지 않을 전망이다.

NAND – AI 견인 수요와 비트 개선을 둔화

표 27. NAND 수급 및 충족률 전망 (GB bn)

구분	2025E	2026E	2027E	2028E
Total NAND 수요	1,110	1,282	1,530	1,870
AI NAND	205	432	656	964
비-AI NAND	905	850	874	907
YoY 성장률 (%)	-	15.5	19.4	22.2
AI 비중 (%)	18.5	33.7	42.9	51.5
NAND 공급	1,062	1,254	1,501	1,725
충족률 (공급/수요-1, %)	-4.3	-2.1	-1.9	-7.8

자료: Morgan Stanley, BofA Global Research, Bloomberg, KUVIC 리서치 1팀

NAND 총수요는 2025년 1,110 GB bn에서 2028년 1,870 GB bn으로 확대되며, 증분의 대부분을 AI가 견인한다. AI NAND 비중은 2025년 18.5%에서 2028년 51.5%로 절반을 넘어선다. 반면 공급은 1,062에서 1,725 GB bn 증가에 그쳐, 충족률은 2025년 -4.3%에서 2026~2027년 -2%대로 완화됐다가 2028년 -7.8%로 다시 확대되는 흐름을 보인다.

2028년 부족이 재차 깊어지는 원인은 공급 측에 있다. NAND 업체들이 신규 웨이퍼 증설보다 적층단수 전환을 통한 비트 성장에 집중하는 가운데, 2028년은 500단 이상의 대형 레이어 점프가 로드맵상 확정되지 않아 비트 개선율이 둔화되는 소강 국면에 들어선다. 여기에 엔터프라이즈 SSD 중심의 신뢰성 수요가 늘며 제품 믹스가 QLC에서 TLC로 회귀해 밀도 이득이 상쇄된다. 공급 탄력성이 떨어지는 국면에 AI 스토리지 수요가 계속 급증하면서, 완화됐던 부족이 다시 확대되는 것이다. NAND 역시 DRAM과 마찬가지로 공급의 열쇠가 웨이퍼 신규 증설과 공정 전환에 달려 있다는 점을 보여준다.

종합하면, DRAM과 NAND 모두 2026~2028년 내내 충족률이 마이너스에 머문다는 결과는 이번 공급 부족이 수요 급증만의 문제가 아니라 공급이 이를 따라잡지 못하는 **캐파 제약의 문제**임을 보여준다. 그리고 두 제품 공통으로 공급의 병목은 웨이퍼 신규 증설과 미세공정 전환, 즉 전공정 생산능력에 놓여 있다.

전공정

반도체 전공정 개요

정의와 구분 - 제조 흐름과 기술 방식은 총위가 다르다

8대 공정 분류는
단위공정과 작업
범주가 혼재

반도체 공정은 전공정과 후공정으로 나뉜다. 전공정은 웨이퍼 위에 회로를 만드는 단계이고, 후공정은 완성된 칩을 검사하고 패키징해 제품으로 만드는 단계다. 전공정 내부는 다시 FEOL과 BEOL로 갈린다. FEOL은 트랜지스터와 셀 등 소자를 형성하는 공정이고, BEOL은 소자 사이와 외부 접점을 잇는 금속 배선을 만든다. 둘 사이의 컨택 형성 구간을 MOL로 따로 두기도 한다. 칩 단면에서 차지하는 부피는 FEOL이 가장 작고 BEOL 배선이 대부분이지만, 공정 난이도와 수율 기여도는 반대다.

표 28. 반도체 공정 분류 체계

분류 축	구분	내용
1차 구분	전공정	웨이퍼 위에 회로 형성 (FEOL / MOL / BEOL)
	후공정	EDS 검사, 패키징, 최종 테스트
제조 흐름	8대 공정	웨이퍼 제조 → 산화 → 포토 → 식각 → 증착·이온주입 → 금속배선 → EDS → 패키징
기술 방식	단위공정	포토, 식각, 증착(CVD/PVD/ALD), 이온주입, CMP, 세정

자료: SK하이닉스 뉴스룸, KUVIC 산업분석 1팀

통용되는 8대 공정 분류는 단계가 섞여 있다. 웨이퍼 제조는 엄밀히 반도체 팹 내부의 공정이 아니고, 금속배선과 패키징, 테스트는 단위공정이 아니라 목적을 가진 작업 전체를 아우르는 범주다. 산화와 포토, 식각 같은 단위공정은 FEOL과 BEOL에서 사용되는 수단이며, 단계별 목적에 따라 사용 빈도와 횟수만 달라질 뿐 목적은 하나다. **미세한 패턴을 대량으로 만드는 것이다. 전공정의 실체는 증착과 포토, 식각, 세정이 반복되는 사이클**이고, 노드와 구조가 고도화될수록 이 사이클의 반복 횟수가 늘어난다. 공정 원가와 캐파 소요를 결정하는 것은 개별 단계의 난이도가 아니라 이 반복 횟수다.

지금은 전공정에 주목해야 할 시기

공급의 물리적
상한을 좌우하는
전공정

이번 AI 메모리 사이클에서 주목해야 할 곳은 후공정보다는 전공정이다. HBM 쇼티지, 범용 DRAM 가격 급등, 메모리 3사의 동시다발적 CAPEX 확대까지, 이번 사이클에서 관측되는 공급 타이트 현상은 그 원인을 거슬러 올라가면 결국 하나의 지점으로 수렴한다. 바로 웨이퍼 CAPA, 즉 전공정 생산능력의 구조적 경직성이다. 적층·본딩·테스트로 대표되는 후공정은 상대적으로 빠르게 늘릴 수 있는 반면, **웨이퍼 팹 증설은 클린룸 확보에서 장비 반입, 수율 안정화까지 최소 수 년이 걸린다**. 공급 곡선의 기울기를 결정하는 것도, 밸류체인 내 초과수익이 가장 오래 머무는 곳도 전공정이다. 언뜻 이번 HBM 병목은 TSV·본딩 같은 후공정 이슈처럼 보이지만, **공급의 물리적 상한을 실제로 쥐고 있는 것은 전공정**이다.

병목 근거 1)
기업의 자본 배분

가장 직접적인 근거는 **파운드리·메모리 기업 스스로의 자본 배분**이다. TSMC는 2026년 CAPEX 가이드를 기존 520~560억 달러에서 600~640억 달러로 상향했다. CFO 웬델 황(Wendell Huang)은 이 중 **70~80%**를 N2·A14 등 **첨단 전공정**에 배정하고, 첨단 패키징·테스트·마스크를 모두 합친 후공정에는 10~20%만 할당한다고 밝혔다. 생산을 책임지는 당사자가 자본의 우선순위를 전공정에 두고 있다. 황 CFO가 "N2 라인에서 월 1,000장 캐파를 짓는 데 드는 CAPEX가 N3보다 훨씬 크다"고 언급한 대목은, 미세화가 진행될수록 동일 캐파 확보에 필요한 전공정 투자 강도가 비선형적으로 커진다는 것을 보여준다.

메모리 3사도 구조는 동일하다. **삼성전자**의 팹택 P4·P5, **SK하이닉스**의 청주 M15X와 용인 클러스터, **마이크론**의 아이다호 신규 팹과 대만 팹 인수 등 이번 사이클을 대표하는 대형 CAPEX 이벤트는 예외 없이 웨이퍼 캐파, 즉 **전공정 확충**에 관한 것이다. HBM용 TSV·첨단 패키징 투자도 병행되지만, 절대 규모 면에서 전공정 팹 투자에 부속되는 항목에 가깝다.

병목 근거 2)
물리적 제한이 있는
전공정 캐파 증설
속도

또한 후공정은 늘리면 늘지만, **전공정은 늘린다고 곧바로 늘지 않는다**. 전공정이 병목인 본질적 이유는 증설의 비가역성이다. 전공정 캐파를 늘리려면 **클린룸 공간, EUV 등 핵심 장비의 리드타임, 팹 건설·인프라 셋업**이 동시에 충족돼야 한다. EUV 장비는 발주에서 제조·납품·설치·인증까지 통상 수개월에서 연 단위가 걸리고, 장비를 들여놓는 것으로 끝이 아니라 전력·용수·배기·진동 인프라까지 맞물려야 실가동으로 전환된다. 이 세 제약이 겹치면 CAPEX가 집행돼도 완공된 캐파로 연결되는 속도는 물리적으로 제한된다.

마이크론은 2026년 1월 대만 Tongluo(P5) 사이트를 현금 인수해 약 30만 제곱피트 규모의 300mm 클린룸을 확보했지만, 이 인수분의 DRAM 생산 기여 시점은 2027년 하반기다. 수요가 아무리 타이트해도 '공간 확보 → 설비 반입 → 가동'에 최소 연 단위 시차가 붙는다는 것을, 기존 팹을 통째로 사들이는 방식으로 겨우 앞당긴 것이다. TSMC 역시 CoWoS 병목 완화를 위해 남부 이노룩스(Innolux) LCD 공장을 인수·개조(AP8)하는 브라운필드 전략으로 준공 리드타임을 단축했다. **기업들이 신축 대신 인수·개조에 매달린다는 사실 자체가 전공정 캐파가 얼마나 경직적인지를 방증한다**.

반면 후공정은 상대적으로 장비 공급사가 다양하고 리드타임이 짧아, 급증하는 수요의 '급한 불'을 끄는데 훨씬 유연하다. TSMC 회장 C.C. 웨이(Wei)가 2026년 2분기 실적 콜에서 EMIB-T 등 경쟁 패키징 기술의 부상에 대해 "패키징 캐파가 워낙 타이트해 어떤 형태의 물량 부담이든 오히려 자사 전공정 웨이퍼 사업에 도움이 된다"는 취지로 환영한 발언은, 후공정 병목이 완화되면 그 부담이 결국 전공정으로 되돌아간다는 구조를 역설적으로 드러낸다.

병목 근거 3)
웨이퍼 증가가 아닌
미세공정 전환
중심의 증설 사이클

CAPEX가 늘어도 **비트 공급은 즉각적으로 늘지 않는다**. 전공정이 병목인 세 번째 이유는, 이번 사이클의 증설이 신규 웨이퍼 스타트 확대가 아니라 기존 라인의 미세공정 전환 중심으로 이뤄지기 때문이다. 전환은 장기적으로 웨이퍼당 비트 밀도를 올리는 방법이지만, **단기적으로는 동일 클린룸 안에서 장비를 재배치**하고 레시피를 재정렬하는 동안 가동률이 떨어지며 생산 공백을 만든다. 그 결과 CAPEX가 증가해도 상당 부분이 순수한 캐파 증가가 아니라 전환 비용과 난이도 상승에 흡수되고, 비트 공급은 자본 지출보다 훨씬 완만하게 상승한다.

결국, **전공정이 있어야 후공정도 있다**. 후공정은 전공정이 만들어낸 웨이퍼를 소화하는 다운스트림 공정이고, 진짜 공급 상한을 쥐고 있는 것은 전공정이다. TSV·본딩·적층 등 후공정의 병목도 결국 전공정에서 생산된 HBM 코어 다이를 전제로 한다. HBM 웨이퍼를 늘리는 만큼 같은 팹 안에서 범용 DRAM 웨이퍼 배분이 줄어드는 제로섬 구조이고, DRAM 3사의 HBM 웨이퍼 비중이 삼성 23%→34%, 하이닉스 32%→38%로 상향되는 흐름이 이를 그대로 보여준다. 아무리 적층·본딩 후공정 capa를 늘려도, 쌓을 코어 다이가 전공정에서 나오지 않으면 HBM은 늘지 않는다.

증설 자금의 블랙홀, 5대 핵심 전공정 장비

5대 핵심 공정:
노광, 식각, 증착,
세정, 이온주입

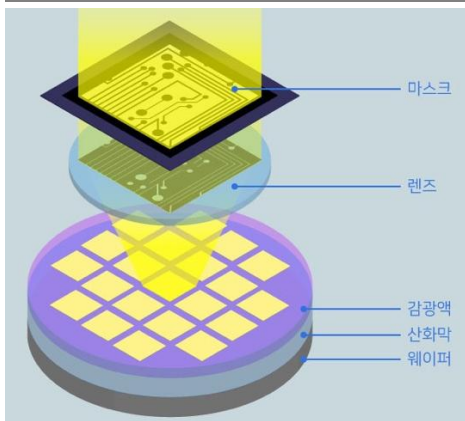
메모리 메이커들의 쏫티지 타개와 공격적인 증설 목표는 결국 '신규 팹(Fab) 건설'과 '장비 반입'으로 직결된다. 반도체 전공정은 웨이퍼 위에 미세한 회로를 만드는 과정으로, 공정이 고도화될수록 특정 단위 공정의 반복 횟수(Step 수)가 기하급수적으로 늘어난다. 공정 원가와 캐파(CAPA) 소요를 결정하는 것은 개별 단계의 난이도가 아니라 이 반복 횟수다. 결과적으로 팹 증설에 투입되는 막대한 자본(CAPEX)은 전공정 장비 발주로 향하게 된다. 산업 업계 자료에 따르면, **전체 전공정 장비 투자의 절대다수가 노광, 식각, 증착, 세정, 이온주입의 5대 핵심 공정에 집중되어 있다**. 증설 사이클의 진정한 수혜를 파악하기 위해서는 이 5대 공정의 역할과 병목 지점을 명확히 이해해야 한다.

5대 핵심 공정

1) 노광 공정

반도체 전공정 가운데 노광(포토)은 웨이퍼 위에 회로 패턴을 새기는 공정으로, 전공정 전체의 방향성을 결정하는 핵심 단계 중 하나에 해당한다. 노광은 설계된 회로도, 즉 마스크에 담긴 패턴을 웨이퍼에 실제로 옮기는 유일한 공정이며, 이후에 이어지는 식각과 증착을 비롯한 모든 후속 공정은 노광이 그려 놓은 패턴을 기준으로 진행된다. 결국 반도체의 미세화 수준을 최종적으로 규정하는 것이 노광 공정인 것이다. 회로 선폭이 곧 제품 경쟁력으로 직결되는 반도체 산업에서, 노광은 얼마나 미세하게 회로를 그려낼 수 있는가를 결정하는 기술적, 경제적 승부처가 된다.

그림 42. 노광 공정



자료: 삼성반도체이야기

그림 43. 노광 장비 사진



자료: ASML

노광 공정에서 비용과 병목이 발생하는 지점도 결국 이 구조에서 비롯된다. 더 미세한 회로를 그리기 위해서는 더 짧은 파장의 빛과 더 정교한 광학계가 필요한데, **이 조건을 충족하는 장비일수록 가격이 기하급수적으로 상승하고 공급 가능한 업체는 극소수로 좁혀진다.** 여기에 하나의 층을 완성하기 위해 여러 차례 노광을 반복해야 하는 경우가 많아, 공정이 미세해질수록 처리 시간과 원가 부담이 함께 커진다. 마스크와 포토레지스트 역시 미세 공정으로 갈수록 요구 품질이 높아지면서 소수의 선도 기업에 대한 의존이 심화된다. **광원부터 소재에 이르기까지 모든 요소가 고도화될수록 특정 업체에 대한 집중도가 함께 높아지는 이 구조가 바로 노광 공정을 전공정에서 가장 좁은 길목으로 만드는 근본 원인이다.**

노광 기술 역시 진화하고 있다. 회로 선폭이 10나노 이하로 진입하면서 기존 광원만으로는 한계에 부딪혔다. 전통적인 ArF 광원으로는 물리적 한계로 인해 같은 층을 여러 번 나누어 노광하는 멀티패터닝이 필요했는데, 이는 공정 횟수와 시간, 원가를 모두 늘리는 요인이었다. 이 문제를 근본적으로 해소한 것이 **13.5나노 파장의 극자외선을 사용하는 EUV다.** EUV는 훨씬 짧은 파장 덕분에 여러 번 나누어 그리던 패턴을 단일 노광으로 구현할 수 있다. 다만 높은 기술 장벽으로 EUV 장비를 특정 업체만 공급할 수 있는 구조로 귀결시켰고, 그 결과 **미세화가 곧 노광 장비 공급망의 집중으로 이어지는 패턴을 만들었다.**

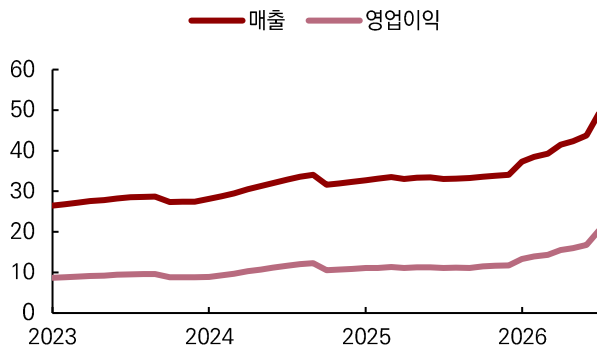
미세화 심화될수록
높아지는 진입 장벽

최근의 흐름은 EUV 안에서 다시 한 단계 진화하는 방향으로 전개되고 있다. 빛을 모으는 렌즈의 개구수를 기존 0.33에서 0.55로 높인 **High-NA EUV**를 통해 구현 가능한 최소 회로 크기가 한층 더 작아졌다. High-NA EUV는 2나노 이하 노드에 대응하기 위한 차세대 광원으로 자리매김하고 있으며, 2025년 SK하이닉스가 메모리 업체로는 처음으로 상용 High-NA EUV 장비를 팹에 도입하는 등 실제 양산 현장으로의 진입이 시작되었다. 다만 **High-NA 장비는 표준 EUV보다 두 배 이상 높은 가격과 늘어난 팹 인프라 요구를 동반해, 도입 가능한 기업이 최선단 대형 팹으로 좁혀진다.** 미세화의 진전이 곧 진입 장벽의 심화로 이어지는 노광 공정의 속성이 광원 기술이 세대를 거듭할수록 더욱 선명해지고 있는 상황이다.

최근 시장 트렌드와 노광 공정의 병목

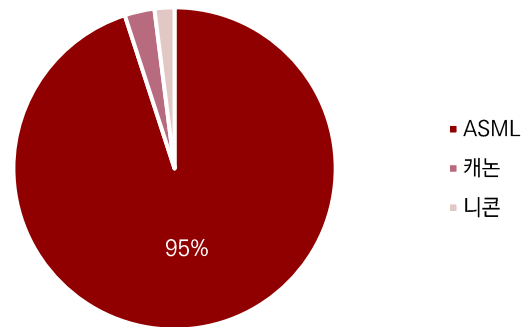
최근의 AI와 고성능 컴퓨팅, 데이터센터 워크로드는 모두 가장 앞선 미세 공정에서 제조된 칩 위에서 구동되며, 이러한 최선단 노드는 EUV 없이는 만들어질 수 없다. **AI 컴퓨팅 수요의 증가가 곧바로 EUV 장비 발주 증가로 직결되는 구조인 셈이다.** 여기에 메모리 업체들이 칩의 집적도와 성능을 끌어올리기 위해 EUV 적용을 넓히면서, 과거 파운드리 중심이던 EUV 수요가 메모리 영역으로까지 본격 확산되는 양상이다. 실제로 EUV 장비를 독점 공급하는 ASML의 수주 잔고와 매출 가이드스가 AI 수요를 배경으로 지속 상향되고 있다는 점은 이러한 흐름을 방증한다.

그림 44. ASML 12M Fwd 매출/영업이익



자료: Quantivise

그림 45. 노광 장비 글로벌 점유율



자료: JP모건

병목 1) 단일 공급자 의존

이처럼 노광 공정 수요는 구조적으로 확대되고 있지만, 정작 이 수요를 받는 공급 측면에서 뚜렷한 병목을 안고 있다. 첫 번째는 **ASML이라는 단일 공급자에 대한 의존**이다. EUV 장비는 전 세계에서 네덜란드의 ASML 한 곳만이 생산하며, 상대적으로 파장이 긴 DUV 장비까지 포함해도 ASML이 사실상 시장을 압도한다. **모든 최선단 칩이 한 기업의 장비에 의존한다는 것은 그 자체로 반도체 공급망 전체에 걸린 구조적 취약점**이다.

병목 2) 부담스러운 장비 가격과 인프라 투자

자연스럽게 **장비 가격과 인프라 투자 부담**이 두배다. 표준 EUV 장비 1대가 약 1.8억 유로, High-NA 장비는 3.8~4억 유로에 달하고, 장비 도입 외에도 진공·방진·전력 등 팍 인프라 전반의 보강이 뒤따른다. 이 막대한 비용을 감당할 수 있는 기업은 소수의 최상위 파운드리와 메모리 업체로 제한되며, 결국 자본력이 선단 공정의 진입을 결정하는 구조가 고착된다. 미세화가 심화될수록 이 격차는 더 벌어진다.

병목 3) 소재와 부품의 기술 장벽

마지막은 **소재와 부품의 기술 난도**다. 장비가 있더라도 그에 걸맞은 포토레지스트와 마스크, 펠리클이 없으면 공정은 완성되지 않는다. 특히 EUV용 포토레지스트는 높은 해상도와 감도, 안정성을 동시에 요구해 기술 장벽이 매우 높고, 오랫동안 일본의 소수 기업이 이를 과점해 왔다. 펠리클과 마스크 관련 소재 역시 상용화 수준의 품질을 확보한 공급처가 제한적이다. 장비의 병목이 소재·부품의 병목과 겹쳐 있다는 점에서, 노광 공정의 공급 제약은 단일 지점이 아니라 밸류체인 여러 층에 동시에 걸쳐 있다.

여기에 EUV 자체의 기술적 한계도 병목을 심화시키는 요인으로 작용한다. EUV 광원의 출력과 처리 속도에는 여전히 물리적 제약이 있어, 장비 한 대가 단위 시간당 처리할 수 있는 웨이퍼 수가 제한된다. **수요가 아무리 늘어도 장비의 생산성이 즉각 따라가기 어려운 구조인 것이다.** High-NA로의 전환 역시 해상도를 높여 주지만 앞서 언급한 가격과 인프라 부담을 동반하기에, 성능 향상과 도입 문턱 상승이 동시에 진행되고 있다.

2) 증착 공정

트랜지스터와 배선을 구성하는 다양한 박막을 형성하는 핵심 공정

ALD의 중요성이 빠르게 확대

증착 공정은 반도체 제조에서 **트랜지스터와 배선을 구성하는 다양한 박막을 형성하는 핵심 공정**이다. 공정 미세화가 진행될수록 증착막은 단순히 얇게 형성하는 것을 넘어, 원자 단위의 두께 오차까지 제어할 수 있는 높은 정밀도와 균일성이 요구된다. 특히 최근 반도체는 평면(Planar) 구조에서 3차원(3D) 구조로 빠르게 전환되고 있어, 깊고 좁은 홀(Hole)과 트렌치(Trench) 내부까지 동일한 두께의 박막을 형성하는 능력이 제품 성능과 수율을 좌우하는 핵심 경쟁력으로 자리잡고 있다. 이러한 변화의 배경에는 **공정 미세화와 적층 구조 확대**가 있다. DRAM은 1 α , 1 β , 1 γ 세대로 미세화가 진행되면서 커패시터와 비트라인 구조의 종횡비(Aspect Ratio)가 지속적으로 증가하고 있으며, 3D NAND 역시 200단을 넘어 400단 이상 적층을 목표로 개발이 진행되고 있다. 구조가 복잡해질수록 기존 화학기상증착(CVD) 방식만으로는 구조물 내부까지 균일한 막을 형성하기 어려워졌고, 이에 따라 **원자층 증착(ALD, Atomic Layer Deposition)의 중요성이 빠르게 확대**되고 있다.

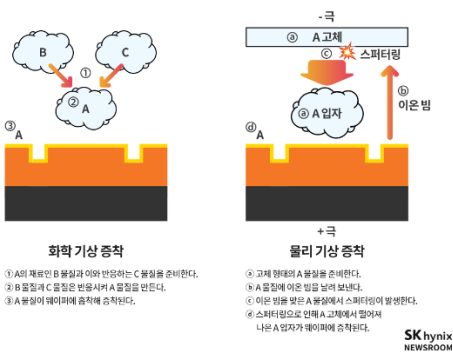
표 29. 증착 방식

방식	원리	강점	주 용도
CVD	전구체 가스의 화학반응	치밀한 막질 우수한 단차피복성	절연막(SiO ₂ , SiN _x)-금속막
PVD	증발-스퍼터링(물리적)	저온-빠른 증착 속도	금속 배선-전극막
ALD	원자층 단위로 반복 증착	초박막-초균일-높은 종횡비 대응	3D NAND 등 차세대 공정

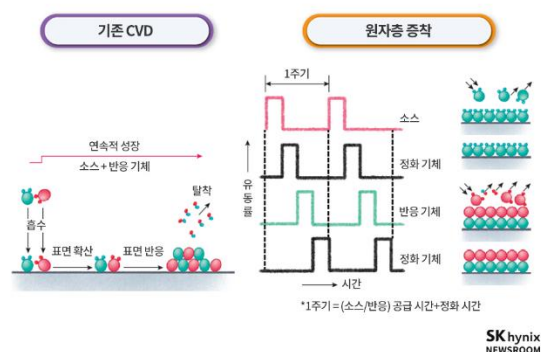
자료: SK하이닉스 뉴스룸, KUVIC 리서치 1팀

그림 46. CVD와 PVD의 개념

그림 47. CVD와 ALD의 개념



자료: SK hynix



자료: SK hynix

ALD는 **전구체(Precursor)를 순차적으로 공급하여 표면 반응만을 이용해 원자층 단위로 박막을 성장시키는 방식**이다. 증착 속도는 상대적으로 느리지만, 높은 단차 피복성(Step Coverage)과 두께 균일도를 구현할 수 있어 초미세 공정에서는 사실상 필수 기술로 자리잡았다. 특히 GAA(Gate-All-Around) 트랜지스터의 게이트 절연막과 금속 게이트, DRAM의 고종횡비 커패시터, 3D NAND의 채널 및 절연막 형성 등 첨단 소자 대부분에서 ALD 적용 범위가 확대되고 있다. 과거에는 일부 핵심 공정에 제한적으로 활용되던 기술이 이제는 첨단 로직과 메모리 공정 전반으로 빠르게 확산되는 추세다.

플라즈마 기반 저온 증착 기술 적용이 확대

저온 증착 기술 역시 중요한 기술 경쟁 요소로 부상하고 있다. 공정이 반복되면서 이미 형성된 금속 배선과 소자 구조 위에 새로운 박막을 증착해야 하는 경우가 많아졌는데, 고온 공정에서는 금속 확산이나 산화, 하부 구조 변형 등으로 인해 소자 특성과 수율이 저하될 수 있다. 이에 따라 300℃ 이하에서도 고품질 박막을 구현할 수 있는 **플라즈마 기반 저온 증착 기술의 적용이 확대**되고 있으며, 장비 업체들도 플라즈마 제어와 공정 균일성을 차별화 요소로 경쟁하고 있다.

시장 역시 이러한 기술 변화를 반영하고 있다. 일반 증착장비 시장은 공정 미세화와 반도체 설비투자 확대에 힘입어 안정적인 성장세를 이어가는 가운데, **ALD 장비 시장은 전체 증착 장비 시장을 크게 상회하는 성장률을 기록할 것으로 전망**된다. 이는 첨단 로직과 차세대 DRAM, 고적층 3D NAND 생산에서 ALD 적용 공정이 지속적으로 증가하고 있기 때문이다. 앞으로 공정 미세화와 적층 구조 확대가 계속될수록 증착장비는 단순한 박막 형성 장비를 넘어, 반도체 성능과 수율을 결정하는 핵심 공정 장비로서 중요성이 더욱 높아질 것으로 예상된다.

Applied Materials, Tokyo Electron (TEL), Lam Research의 과정 구조

증착 장비 시장은 Applied Materials, Tokyo Electron(TEL), Lam Research 등 종합 장비 업체가 시장을 주도하는 과점 구조다. 이들은 CVD, PVD, ALD 등 다양한 증착 장비 포트폴리오를 바탕으로 메모리와 파운드리 고객사에 폭넓게 공급하고 있다. 반면 ASM International은 ALD, Kokusai Electric은 LPCVD·Batch CVD 등 특정 증착 기술에 특화된 글로벌 강자로 자리 잡고 있다. 국내에서는 원익IPS와 주성엔지니어링이 CVD·ALD를 중심으로 삼성전자와 SK하이닉스 등에 장비를 공급하고 있으나, 글로벌 선도 업체 대비 제품 포트폴리오와 고객 기반은 제한적이다.

3) 이온주입 공정

순수한 실리콘 웨이퍼 내부에 붕소, 인 등의 불순물을 가속 주입하여 반도체로서의 전기적 특성(P형·N형)을 띠도록 만드는 공정이다. 선단 공정으로 갈수록 트랜지스터 구조가 복잡해지며 초정밀 도핑 기술의 중요성이 커지고 있으나, 노광이나 증착·식각에 비해 미세화에 따른 공정 횟수(Step 수) 증가 폭이 제한적이어서 전공정 투자 확대의 직접적인 수혜 크기는 상대적으로 작다. 이 시장은 AMAT와 2025년 비코(Veeco)와의 합병으로 포트폴리오를 확장하며 덩치를 키운 액셀리스(Axcelis)가 사실상 양강 구도를 형성하고 있다.

4) 식각 공정

이방성 확보 여부가 미세 패턴의 적용 한계를 결정

식각은 포토 공정으로 감광막에 형성된 패턴을 따라 하부 박막의 불필요한 영역을 선택적으로 제거해 회로를 입체적으로 구현하는 공정이다. 방식은 건식과 습식으로 나뉜다. 건식은 플라즈마와 반응성 가스를 사용해 수직 방향으로 깎아내는 이방성 식각이 가능해 미세 패턴 전 영역에 적용되며 시장의 89%를 차지한다. 습식은 화학 용액 기반의 등방성 방식으로, 저비용과 고처리량이 장점이어서 세정과 본딩 전 표면 처리 공정에 남아 있다. 패턴 크기가 줄어들수록 등방성 식각은 인접 패턴을 침범하므로, 미세화는 곧 건식 비중 확대를 의미한다.

그림 48. 식각 공정



그림 49. 식각 장비 사진



자료:삼성반도체이야기

자료: Lam Research

표 30. 건식 식각 방식별 특성

방식	원리	주 적용처
RIE	화학 반응과 이온 충격 병용	범용 이방성 식각
CCP	용량결합 플라즈마, 고에너지 이온	유전체 및 고종횡비 식각
ICP	유도결합 플라즈마, 고밀도	실리콘-메탈 식각
DRIE	식각과 보호막 형성 교대(보쉬 공정)	깊은 홀, TSV
ALE	원자층 단위 자기제한 반응	GAA 시트, 나노 구조
크라이오	극저온 유전체 식각	3D 낸드 채널홀

자료: Lam Research, Applied Materials, KUVIC 산업분석 1팀

3D 구조에서는 보잉과 로딩이 수율을 좌우

식각 품질은 다양한 기준으로 판단한다. 마스크 대비 하부막의 잔존율을 보는 선택비, 수직 절삭 여부를 보는 이방성과 프로파일, 깊이와 폭의 비율인 종횡비, 웨이퍼 전면의 균일도, 패턴 밀도에 따라 식각 속도가 달라지는 로딩 효과, 홀 중간부가 확장되는 보잉이 대표적이다. 평면 구조에서는 선택비와 균일도가 핵심이었으나, 3D 구조로 전환되면서 로딩과 보잉이 수율을 좌우하는 변수로 올라섰다. 깊은

홀을 균일하게 뚫는 능력이 곧 적층 한계를 정한다.

물리적 상한을 결정하는 식각

표 31. 메모리·로직별 식각 요구 사양

적용처	요구 사양	난이도 요인
3D 낸드 채널홀	중형비 60:1 초과, 200회 이상 식각 패스	측벽 보잉이 셀 효율 훼손
DRAM 커패시터	고중형비 홀 식각	노드 미세화에 따른 중형비 상승
HBM TSV	DRIE 기반 관통 전극	후공정 캐파의 실질적 제약
GAA 트랜지스터	ALE로 시트 두께 5~7nm 조절	구동 전류와 누설의 상충
백사이드 파워	중형비 100:1 비아	챔버 압력 ±2mTorr 제어 필요

자료: Lam Research, TSMC, Intel, Mordor Intelligence, KUVIC 리서치 1팀

3D 낸드에서 적층을 제약하는 것은 증착이 아니라 식각이다. 트렌치 중형비가 60:1을 넘어서면 측벽이 확장되는 보잉이 셀 효율을 훼손한다. 스택 하나를 완성하는 데 200회가 넘는 식각 패스가 필요하고, 텍을 추가하면 채널홀 식각과 정렬 공정이 그만큼 반복된다. 레이어 수가 같아도 텍 구성이 다르면 원가가 갈리는 이유다.

향후 기술적 방향은 극저온 식각이다. Lam Research의 Cryo 3.0은 기존 Flex와 Vantex 플랫폼과 호환되며, 주요 메모리 업체 전부가 3D 낸드 양산의 채널홀 식각에 사용하고 있다. 신규 장비를 전면 교체하지 않고 기존 설치 기반에 증설 형태로 적용할 수 있어 전환 비용이 낮다. 로직 쪽도 난이도가 함께 상승한다. TSMC의 백사이드 파워레일은 중형비 100:1 비아를 도입해 챔버 압력을 ±2mTorr로 제어해야 측벽 거칠기를 억제할 수 있고, 인텔은 ALE로 시트 두께를 5~7nm 범위에서 조절한다.

메모리 사이클에 직결되는 구조

표 32. 식각 장비 시장 개요

구분	내용	자료
시장 규모	2025년 \$20.0bn → 2033년 \$29.3bn (CAGR 5.7%)	Intel Market Research
과점 구조	Lam Research·TEL·Applied Materials 합산 약 90%	Intel Market Research
기술별 비중	건식 89%/습식 11%	Intel Market Research
응용별 비중	로직·메모리 약 80%	Intel Market Research

자료: Mordor Intelligence, Intel Market Research, KUVIC 리서치 1팀

Intel Market Research는 2025년 200억 달러에서 2033년 293억 달러로 연평균 5.7% 상승을 전망했다. Lam Research와 TEL, Applied Materials가 합산 90%를 점유하며, Lam은 플라스마 식각, TEL은 유전체 식각, Applied Materials는 증착과 식각을 묶은 통합 솔루션에 강점을 갖는다. 중국 NAURA와 AMEC는 자금 정책을 배경으로 내수 레거시 노드에서 점유를 확대하고 있다.

식각 장비는 메모리 사이클에 특히 민감하게 반응한다. 2024년 초 공급과잉 국면에서 SK하이닉스와 마이크론, 삼성전자가 식각 장비 구매를 약 25% 축소했다. 3D 낸드가 200회 넘는 식각 패스를 요구하기에, 메모리 업체의 지출 변동이 식각 장비 업체 매출로 직결된다.

비트 증가로 바로 이어지진 않는 레이어 증가

낸드 공급을 레이어 수로 환산하면 실제 비트 산출을 과대평가하게 된다. 레이어가 늘면 식각 스텝이 비선형으로 증가해 웨이퍼당 처리 시간과 장비 소요가 함께 늘고, 동일 클린룸에서 처리할 수 있는 웨이퍼 수는 줄어든다. 텍 구성이 다르면 같은 레이어에서도 식각 패스 수가 달라져 원가와 산출이 갈린다. 공급 추정에서 레이어 수 대신 단위 면적당 밀도와 텍 구성을 기준으로 삼아야 하는 이유다.

레이어 증가를 막는 것은 증착이 아니라 식각

크라이오 식각이 기존 설치 기반에 증설 형태로 적용

식각 패스 수가 많아 지출 변동이 증폭되는 구조

식각 스텝의 비선형 증가가 웨이퍼 처리량을 잠식

세대 전환에 선행하는 식각 CAPA

장비 수주가 램프
시점의 선행 지표

레이어 세대 전환은 식각 캐파 확보가 선행되어야 진행된다. 신규 세대의 채널홀 식각 조건을 확보하지 못하면 증착 장비가 준비되어도 양산에 들어갈 수 없다. 이 때문에 식각 장비 수주 흐름은 낸드 세대 전환의 타이밍을 양산 발표보다 먼저 드러낸다. 동시에 장비 리드타임은 HBM 증설과 낸드 전환 사이의 자원 배분을 제약한다. 클린룸 확보만으로 증설 속도가 결정되지 않는다는 점에서, D램 증설 계획을 평가할 때도 같은 기준이 적용된다.

5) 세정 및 검사 공정

건식 세정 & 초임계
CO₂ 세정 기술 도입
가속화 중

세정 공정은 단위 공정 후 웨이퍼 표면에 남아 있는 미세 입자, 유기 오염물, 금속 불순물 및 불필요한 자연 산화막을 제거하여 웨이퍼를 깨끗한 상태로 만드는 공정이다. 세정은 크게 습식 세정과 건식 세정으로 구분된다. 습식 세정은 황산과 과산화수소를 혼합한 SPM(피라나), 희석 불산(DHF), SC-1, SC-2 등 다양한 화학 약품과 초순수 증류수를 조합하여 수행한다. 과거에는 웨이퍼 여러 장을 수조에 담그는 배치 방식이 주를 이루었으나, 최신 선단 공정에서는 웨이퍼를 한 장씩 회전시키며 노즐로 약품을 분사하는 단엽식 방식이 표준으로 자리 잡았다. 나아가 미세 공정화에 따라 표면장력으로 패턴이 무너지는 현상을 방지하기 위해 **가스 및 플라스마를 사용하는 건식 세정이나 초임계 CO₂ 세정 기술의 도입이 가속화되고 있다.**

세정은 전공정 전체 단계 중 약 30% 이상을 차지할 정도로 가장 빈번하게 수행되는 공정이다. 선단 공정으로 갈수록 패턴의 종횡비가 급격히 커지며, **습식 세정 후 건조 과정에서 액체의 표면장력으로 인해 패턴이 서로 붙거나 무너지는 현상이 신규 병목으로 부상**하였다. 또한 EUV 노광 후 잔여물 제거 및 HBM 생산을 위한 TSV 딥 에칭 후 홀 내부 파티클 세정 등, 고도화된 패키징 및 선단 공정 영역에서 세정 공정의 난이도와 중요성은 한층 높아지고 있다.

반도체 제조 전공정에서 **검사 및 계측 공정**은 노광, 식각, 증착, CMP 등 단위 공정을 거친 웨이퍼의 물리적·전기적 상태를 정밀 분석하고 불량률 감지하는 대표적인 공정 제어 영역이다. 해당 영역은 **크게 검사와 계측으로 나뉜다**. **검사**는 웨이퍼 표면에 존재하는 미세 파티클, 패턴 단선 및 슛, 잔여물 등 결함의 위치와 개수를 검출하는 역할을 맡는다. 광각 광원을 활용해 고속으로 웨이퍼 전체를 스캔하는 광학 검사와, 광학으로 확인하기 어려운 극미세 결함을 전자빔으로 정밀 선별하는 전자빔 검사(E-Beam)가 대표적이다. 반면 **계측**은 웨이퍼 패턴의 수치적 정밀도를 계측하는 공정이다. 패턴의 선폭을 측정하는 CD-SEM, 하부층과 상부층 간의 정렬 오차를 측정하는 오버레이(Overlay), 증착된 박막의 두께 및 성분을 분석하는 Thin Film 계측 등으로 구체화된다.

공정 미세화에 따라
수율 관리 비용
역시 폭등 추세

3nm 이하 선단 파운드리 공정과 1b/1c nm D램, 고단화 D램 및 HBM 공정으로 진입함에 따라 수율 관리 비용은 폭증하는 추세이다. 단 한 개의 나노 파티클이나 1nm 이하의 오버레이 오차가 웨이퍼 전체 수율에 치명적인 영향을 미치기 때문이다. EUV 멀티패터닝, GAA 아키텍처, 300단 이상의 3D NAND 적용 등으로 전공정 단계는 1,000 Step 이상으로 대폭 늘어났다. 이에 따라 공정 중간 단계에서 결함을 적시에 선별하지 못할 경우, 후속 공정에 투입되는 비용이 고스란히 손실로 직결되는 병목이 발생한다. 단순 장비 납품을 넘어 검사 및 계측 과정에서 수집되는 광대한 웨이퍼 데이터와 AI 수율 관리 소프트웨어의 결합은 팹의 초반 수율 확보 기간을 결정짓는 핵심 요소이다.

전공정 TAM 추정

2025 전공정 시장 분석

2025년 전공정
장비 TAM 추정
: \$1013.5억

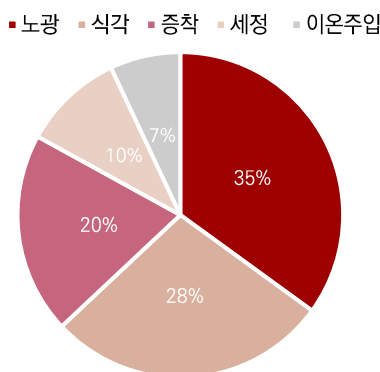
2025년 각 공정별 시장 규모를 측정하기 위해 전공정 장비 시장 규모에 각 공정 비율을 곱하여 추정하였다. 우선 전공정 장비 시장 규모를 구하기 위해 2025년 WFE 수치를 활용하였다. 시장에 나와있는 2025년 WFE(웨이퍼 팹 장비)는 1169억 달러다. 하지만 이 수치는 웨이퍼 가공, 마스크/레티클, 팹 설비 등이 포함되어 있는 값이다. 후공정 장비나 결정제조(잉곳·웨이퍼 생산 장비)는 포함하지 않았다.

2025년 반도체 CAPEX 비중에서 반도체 제조장비 75%, 기타 투자가 25%이고, 이 반도체 제조장비 안에서 전공정 48.7%, 후공정 15%, 공통설비 7.5%, 결정제조 3.8%로 나뉜다. 이에 따라 **전공정(48.7%)과 공통설비(7.5%, 계측·검사 장비) 두 항목을 실질적인 구성 요소라 파악하여 나머지 후공정과 결정제조는 계산에서 제외했다.** 이를 통해 계산한 전공정과 공통 설비의 비중은 각각 86.7%, 13.3%가 되며 이를 정규화하여 **전공정 장비 시장의 규모를 1013.5억 달러로 추정했다.**

2025년 전공정 장비 시장 = 1169 × (86.7/100) = **1013.5억 달러**

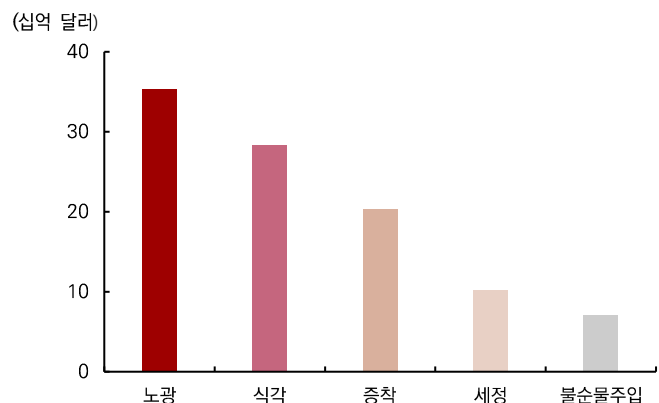
공정별 비율 노광 35%, 식각 28%, 박막형성(증착) 20%, 세정 10%, 불순물주입 7%를 앞서 구한 전공정 시장 규모 1013.5억 달러에 적용하여, **노광 장비 시장 354.7억 달러, 식각 장비 283.8억 달러, 증착 장비 202.7억 달러, 세정 장비 101.4억 달러, 이온 주입 장비 70.9억 달러로 각각 추정했다.**

그림 50. 2025년도 공정별 비중



자료: 업계 자료 종합, KUVIC 리서치 1팀

그림 51. 2025년도 전공정별 규모



자료: KUVIC 리서치 1팀

증착 & 식각

식각 TAM 전망
: 26년 \$374.6억
→ 27년 \$472억
→ 28년 \$576억

2026년은 반도체 업황 호황에 힘입어 투자 규모 자체가 상향되며 증착과 식각 공정 모두 큰 성장세를 보이는 중이다. 식각 공정에선 Applied Materials가 반도체 시스템 사업 가이드نس를 30% 이상으로 상향했고, Lam Research는 자사 사업(식각+증착)이 전체 반도체 장비 시장에서 차지하는 비중을 30%대 초반에서 30%대 중반으로 상향했다. Lam Research의 식각 시장에서의 높은 점유율을 고려, 식각 공정은 **시장 평균을 상회하는 +32% 정도의 성장률을 기록할 것으로 예상한다.** 이에 따라 **2026년 374.6억 달러 정도의 성장을 전망한다.**

증착 TAM 전망
: 25년 \$202.7억
→ 26년 \$261.5억
→ 27년 \$324.2억
→ 28년 \$389.1억

ASM International이 2026년 2분기 매출에서 가이드نس 중간값을 웃도는 수준으로 달성하며 견조한 수요를 확인시켜 준 점을 고려, 증착 역시 높은 성장세를 보일 것이라는 점을 확인했다. HBM 적층이나 GAA 게이트 형성 모두 재료를 쌓는 증착과 이를 깎아내는 식각이 한 세트 반복되는 공정이므로, 층수가 늘어날수록 두 공정은 함께 늘어나기에 증착 시장도 식각과 유사한 정도로 성장하여 **2025년 202.7억 달러에서 2026년 261.5억달러로 성장할 것을 전망한다.**

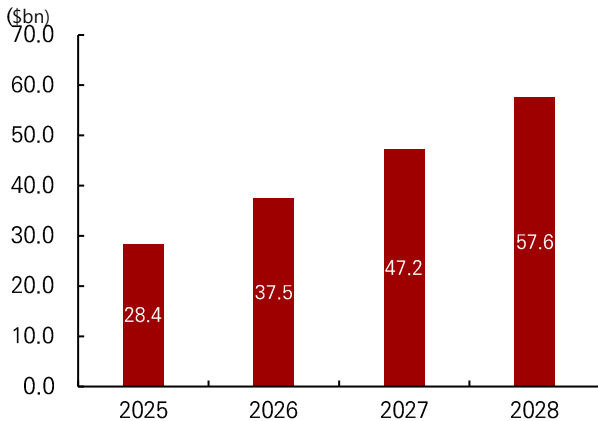
2026년의 폭발적인 성장이 전망되는 가운데 2027년과 2028년에도 식각과 증착 시장의 성장 가능성은 Lam Research 4분기 실적 발표를 참고했다. Lam은 **"산업이 여전히 의미있게 공급부족"이며, 올해의 공급부족이 내년으로 이월된다고** 밝혔다. 대형 고객 기준 팹 8~10곳이 지금부터 2027년 말 사이 신규 가동되고, 발주 논의가 리드타임을 넘어서는 구간에서 진행 중이라고도 언급했다. 즉 물량이 2026년에 당겨지는 것이 아니라,

병목으로 인해 2027년으로 이월되는 구조다.

CFO Doug Bettinger는 식각과 증착의 집약도 때문에 3년 연속 WFE 성장률을 상회하는 년도가 될 것이라 언급했다. 이에 따라 2027년과 2028년도도 식각 성장률이 WFE 성장률을 상회할 것이라 판단하여 **2027년도 472억 달러, 2028년 576억 달러로 시장 규모가 상승할 것으로 전망한다.**

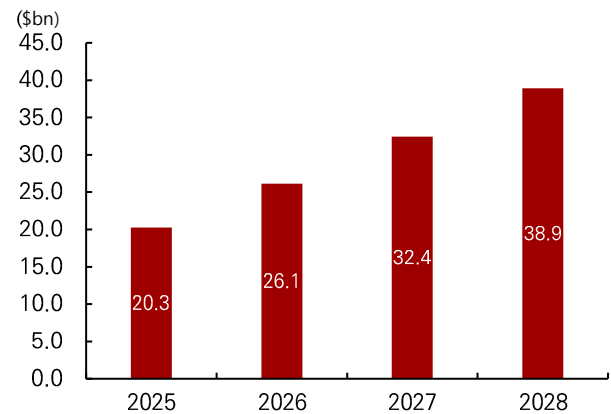
증착 또한 앞선 내용들과 Lam Research의 어드밴스드 패키징 성장 가이드가 3분기 연속 상향된 점을 고려하여 **27년도 324.2억 달러, 28년도 389.1억 달러의 시장 규모로 성장할 것이라 전망한다.**

그림 52. 식각 시장 규모



자료: KUVIC 리서치 1팀 추정

그림 53. 증착 시장 규모



자료: KUVIC 리서치 1팀 추정

노광

노광 TAM 전망
: 25년 \$15.4억
→ 28년 \$36.6억

노광 시장은 EUV, DUV, 트랙 장비의 분류를 통해 각 분야의 성장률을 종합적으로 고려하여 추정했다.

EUV 시장은 ASML 2Q26 컨퍼런스 콜의 "EUV net system sales growth of over 45%"를 직접 인용했다. AI 가속기와 메모리 선단 노드(DRAM 1c/1b)에서 EUV 레이어가 증가하는 것을 그 이유로 판단했다. 2027년은 ASML이 제시한 Low-NA EUV 출하 계획 85대(YoY +30%)와 "2027년 물량은 수주로 거의 충족"이라는 언급을 참고했다. 2028년은 ASML이 110대 캐파(YoY +30%)를 검토 중이라고 말한 부분에서 확정 발주가 아닌 점을 고려하여 YoY +20% 가량의 성장을 전망한다. **2025년 15.4억 달러에서 2028년 36.6억 달러 정도의 성장을 전망한다.**

DUV 시장은 ASML의 비EUV 가이드선스 +25%에서 비EUV 매출에 섞인 계측·검사 물량과 니콘·캐논의 상대적인 저성장을 고려하여 **2025년 15.9억 달러에서 2028년 24.7억 달러의 성장을 전망한다.**

트랙 시장은 TEL의 FY2027 코터/디벨로퍼 YoY+50% 이상 성장 전망을 활용했다. 트랙은 EUV·DUV 모든 노광 스텝에 1:1로 연결된다는 점을 고려, **2025년 4.2억 달러에서 2028년 7.8억 달러의 성장을 전망한다.**

노광 공정이 전공정
시장 확대의 핵심

이러한 전망에 기반한 **노광의 전공정 대비 비중이 +2.5%p 상향할 것으로 예측했으며**, 이는 동기간 타 공정들을 모두 앞서는 수치에 해당한다. 이를 통해 **노광 공정이 전공정 시장 확대의 핵심일 것이라 전망한다.**

세정

노광 시장은 EUV, DUV, 트랙 장비의 분류를 통해 각 분야의 성장률을 종합적으로 고려하여 추정했다면, 세정 시장은 2025년 WFE에 전공정 비중 86.7%와 세정 장비 비중 10%를 적용해 **2025년 101억 달러의 시장을 산출했다.**

세정 TAM 전망
 : 26년 \$124억
 → 27년 \$142억
 → 28년 \$153억

이온 주입 TAM
 : 25년 \$71억
 → 26년 \$74억
 → 27년 \$78억
 → 28년 \$82억

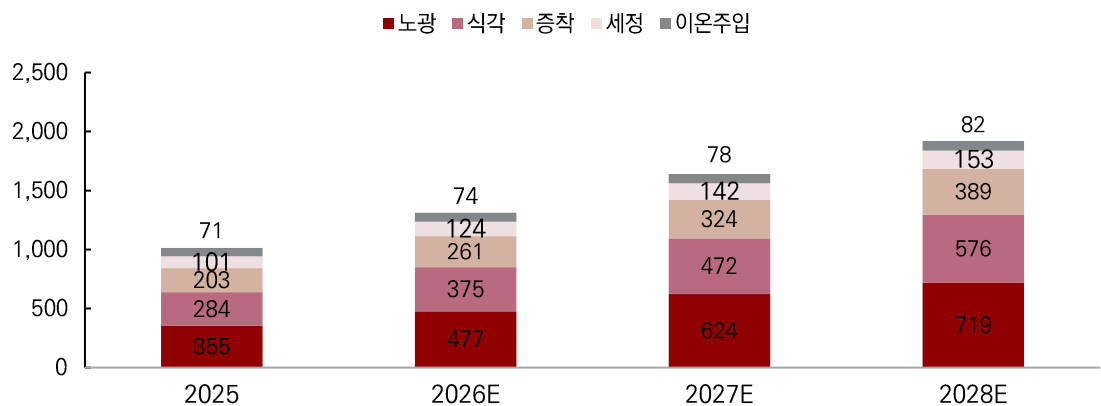
DRAM의 EUV 레이어 증가와 NAND의 고층형비 구조·웨이퍼 본딩 확대는 세정 난도와 장비 수요를 높인다. 다만 공정 복잡도 증가가 신규 장비 발주와 1:1로 연결되진 않다는 점을 고려했다. Lam Research의 시장 진입은 TAM 확대보다 기존 SCREEN·TEL과의 점유율 인 점을 고려했다. 또한, SCREEN의 FY2027 반도체 제조장비 가이던스를 참고했다. 다만 SCREEN 실적에는 세정 외 장비와 점유율 상승 효과가 포함된 점을 고려, **2026년 124억 달러, 2027년 142억 달러, 2028년 153억 달러의 성장을 전망한다.**

이온 주입

이온주입은 2025년 전공정 시장에 불순물주입 장비 비중 7%를 적용해 **2025년 71억 달러의 시장을 산출했다.** 다만 해당 추정이 순수 이온주입뿐 아니라 산화, 확산, 어닐링 장비가 포함되어 있다는 점을 고려, 이온 주입 장비 대표 기업인 Axcelis가 영위하는 이온주입 단독 시장이 아니라 관련 열처리 장비까지 포함한 시장임을 고려했다. 추정에는 Axcelis의 메모리, 로직 수요 전망을 반영하였고, 2026년 전사 매출은 SiC와 전력반도체 부진으로 정체되지만 DRAM·HBM 수요는 회복되고 있다는 점과 신규 메모리 캐파 확대를 고려했다. 또한, 이온주입은 신규 웨이퍼 캐파 증가에는 수혜를 받지만 식각, 증착처럼 3D 구조 전환에 따라 공정 횟수가 크게 늘어나지는 않는다는 점을 반영했다. 이를 통해 **2026년 74억 달러, 2027년 78억 달러, 2028년 82억 달러의 성장을 전망한다.**

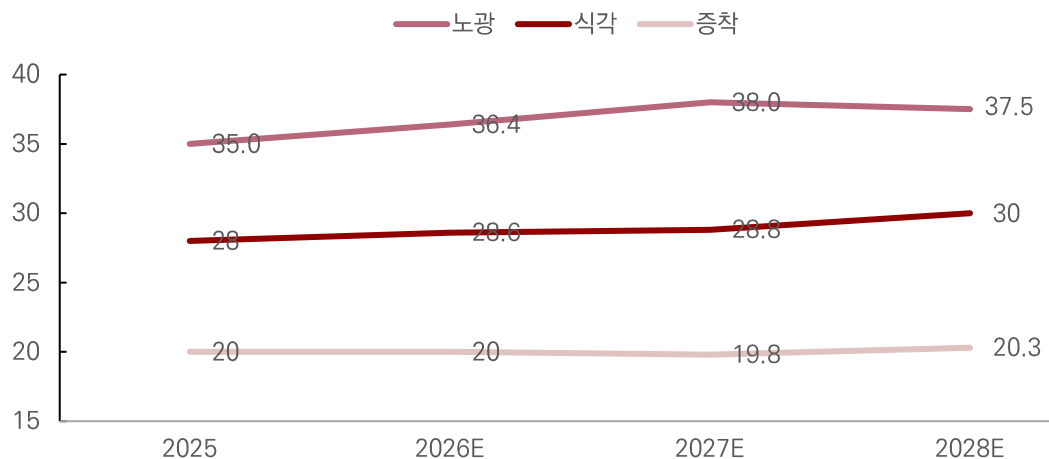
그림 54. 전공정 공정별 시장 규모

(단위:억 \$)



자료: SEMI, ASML, 도쿄일렉트론, KUVIC 리서치 1팀

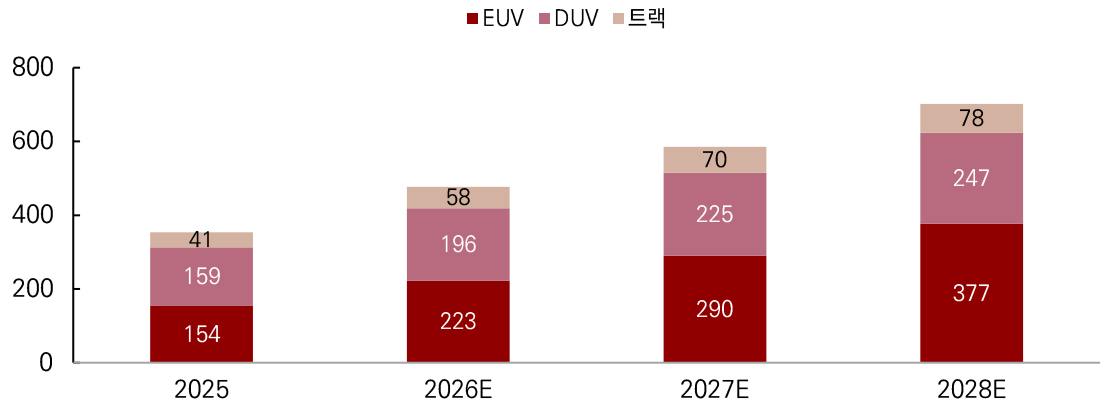
그림 55. 전공정 핵심 공정별 비중 변화



자료: KUVIC 리서치 1팀

그림 56. 노광 내 기술별 구성

(단위:억 \$)



자료: ASML, 도쿄일렉트론, KUVIC 리서치 1팀

3형제 중 1등, 노광.

전공정 주요 공정의 성장률을 보면 식각 CAGR 26.6%가 노광의 성장(23.8%)을 앞서고, 증착도 24.3%로 견조하다. Lam Research는 4Q26 컨콜에서 "식각-증착의 집약도 때문에 3년 연속 WFE 성장률을 앞서는 해를 바라본다"고 명시했다. 3D 구조 전환이 식각-증착 강도를 구조적으로 끌어올리는 국면이다.

다만 **본 리포트는 성장이 금액으로 실현하는 확실성과 추가 업사이드에서 노광이 앞선다는 점**에 주목했다. 실현 확실성, 가격결정력, 사이클에서의 지위 등을 종합적으로 고려했다.

식각-증착 시장은 Lam, AMAT, TEL 3사가 경쟁한다. 시장이 성장해도 개별 벤더의 점유율에는 변동 리스크가 있다. **반면 EUV는 ASML의 과점 시장이다.** EUV 시장이 성장하는 만큼 ASML 매출도 선형적으로 성장한다. 점유율을 뺏길 경쟁자가 상대적으로 강력하지 않다.

ASML은 value-based pricing으로 장비 대당 단가를 지속적으로 인상해 왔다. 2Q26 컨콜에서 "현재 환경의 가격 유연성이 과거보다 크다"고 반복했고, D→E→F로의 세대 전환과 High-NA 도입이 ASP를 끌어올린다. 본 리포트의 추정에선 해당 ASP 인상을 노광 시장 추정에 정량 반영했으며 이를 통해 노광 시장의 3개년 CAGR을 26.5%로 전망한다. 또한, ASP 인상은 매출뿐 아니라 마진까지 확장한다. SemiAnalysis에 따르면 장비사의 가격 인상분은 대부분이 총이익으로 연결된다. 부품 공급사는 가격결정력에서 열위에 놓여있기에 원가 부담은 부품사에 한정된다. 동일 사양 기준 10%만 가격을 올려도 장비사의 총이익률이 과거 최고치보다 1~5%p 높아질 수 있다. 물량 증가와 가격 인상이 겹치며 마진이 새로운 고점으로 향하는 상황이다.

식각과 증착 역시 이 가격 인상 흐름의 수혜를 받지만, 3사 경쟁으로 가격 프리미엄이 상대적으로 분산된다. 반면 노광의 EUV는 ASML 단독 공급이므로 인상분을 온전히 확보한다. **노광은 대수 성장에 ASP 상승이 더해지는 이중 수혜를 받으며, 그 인상분이 매출과 마진에 동시에 반영된다.**

핵심 장비사 6개 컨콜 모두 클린룸을 산업 성장의 최대 제약으로 지목했다. 병목 상황에서 핵심 질문은 "무엇이 전체 생산을 가로막는가"인데, **EUV가 바로 그 병목 자체다.** ASML이 캐파를 +30% 늘려도 수요가 공급을 상회하는 국면이다. 반면 식각과 증착은 상대적으로 공급 여유가 있다. 병목을 뚫은 공정이 사이클에서 가장 강한 가격결정력과 갖는다.

설령 노광 전체의 전공정 대비 비중이 정체하더라도, 노광 내부에서 저가 DUV → 고가 EUV로의 이동이 진행된다. DUV 1대보다 EUV 1대가 훨씬 비싸고 마진이 높으므로, 내부 수익성이 개선된다. 상기 사항등을 고려하여 본 리포트는 매출 개선과 마진 개선이 기대되는 상황에서 둘 중 하나만 이뤄져도 **그 수혜의 정도가 노광이 제일 큰 상황이라 판단한다.**

APPENDIX

노광 공정 상세 메커니즘

노광 공정은 크게 감광액을 바르는 **도포**, 빛을 쬌어 패턴을 옮기는 **노광**, 그리고 원하는 패턴만 남기는 **현상**의 세 단계로 이루어진다. 먼저 웨이퍼 표면에 ‘포토레지스트’라는 감광액을 균일하게 도포하는데, 이 감광액은 빛에 반응해 화학적 성질이 변하는 물질이다. 이어 **회로 패턴이 담긴 마스크에 빛을 통과시켜 웨이퍼 위에 축소해 투영하면, 빛을 받은 부분과 받지 않은 부분의 성질이 달라진다.** 마지막으로 현상 과정을 거치면 성질이 변한 영역이 선택적으로 제거되거나 남으면서 웨이퍼 위에 회로 패턴이 형성된다.

그림 57. 현상 공정

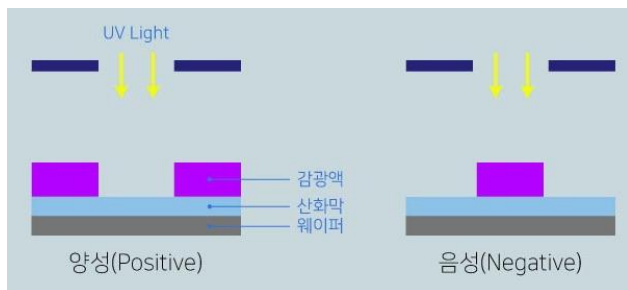
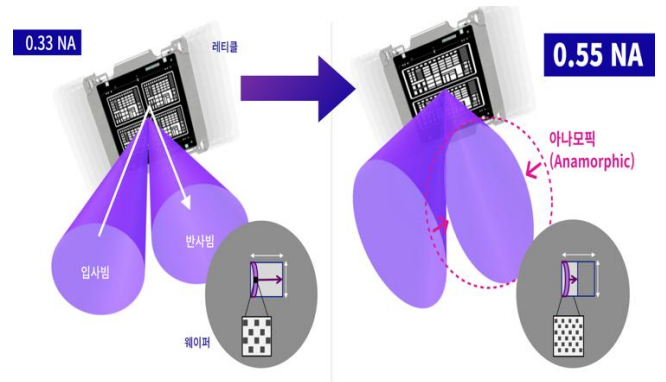


그림 58. High-NA EUV로의 기술 진보



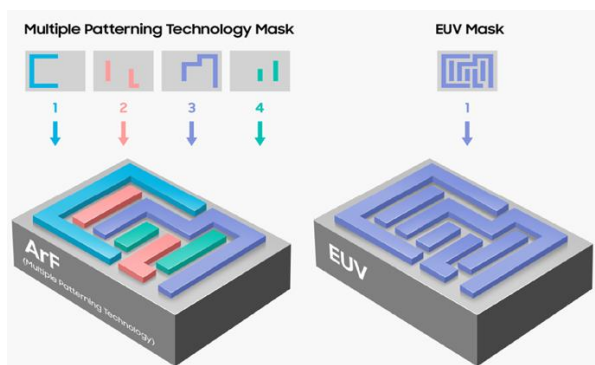
자료: 삼성반도체이야기

자료: ASML

광원, 마스크, 렌즈, 포토레지스트, 4가지 핵심 기술이 노광 경쟁력 결정

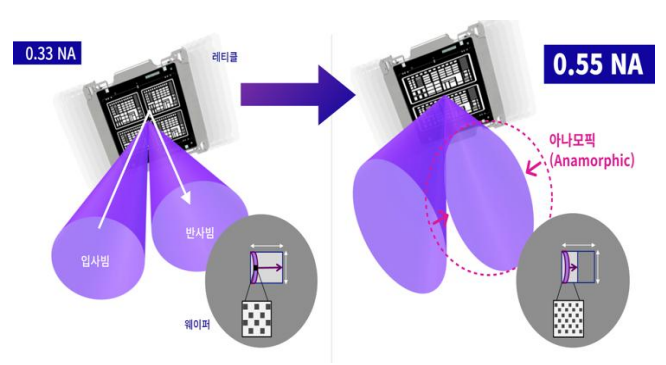
이 과정을 실제로 구현하는 데에는 4가지 요소가 핵심적으로 맞물린다. 회로를 새길 빛을 만들어내는 **광원**, 그려야 할 회로도가 담긴 **마스크**, 그 빛을 웨이퍼에 정밀하게 축소 투영하는 **렌즈**, 그리고 빛에 반응해 패턴을 형성하는 **포토레지스트**가 그것이다. 이 네 요소는 각각이 별개의 산업이자 독립된 기술 영역을 이루며, 하나라도 수준이 뒤처지면 전체 노광 성능이 그 낮은 수준에 맞춰 제약된다. 특히 **광원과 렌즈는 장비 자체의 성능을 규정하는 반면, 마스크와 포토레지스트는 소재·부품 영역에서 별도의 공급망을 형성한다.** 노광 공정의 경쟁력이 장비 한 대의 성능이 아니라 이 네 축을 아우르는 생태계 전반의 완성도에 달려 있다고 말하는 이유가 여기에 있다. 실제로 각 요소의 공급 구조를 들여다보면, 렌즈는 독일 자이스가, 광원 부품은 트럼프와 기가포톤 등이, 포토레지스트는 일본 소수 기업이 사실상 과점하는 형태로, 노광이라는 하나의 공정 안에 여러 독립된 과점 시장이 겹쳐 있는 구조가 드러난다.

그림 59. 멀티패터닝과 EUV



자료: 삼성반도체이야기

그림 60. High-NA EUV로의 기술 진보



자료: ASML

글로벌 플레이어와 점유율 구조

노광 시장을 독점한
ASML

노광 시장의 가장 두드러진 특징은 **ASML로의 극단적인 공급 집중**이다. 그나마 DUV 진영에서는 일본의 니콘과 캐논이 일정 부분 명맥을 유지하고 있으나, **선단 공정의 핵심인 EUV 시장에서는 사실상 완전 독점 구조가 형성**되어 있다. 모든 최선단 반도체가 단 하나의 기업이 만든 장비에 의존한다는 사실은 노광 시장을 이해하는 출발점이자, 이 시장이 왜 구조적 병목으로 지목되는지를 단적으로 보여준다.

다만 노광 장비 밸류체인을 더 깊게 파고 들어가면, ASML조차도 장비를 홀로 완성하는 것이 아닌 여러 전문 기업이 공급하는 핵심 요소를 통합하는 위치에 가깝다. 노광 장비의 성능을 좌우하는 정밀 광학계, 즉 렌즈와 거울은 독일의 자이스가 사실상 독점적으로 공급하며, EUV 광원을 만들어내는 데 필요한 레이저와 광원 부품은 독일의 트럼프, 일본의 기가포톤 등 소수 기업이 담당한다. 이들 광학·광원 협력사가 없이는 ASML의 장비 역시 성립하지 않는다는 점에서, **노광 장비 시장은 ASML을 정점에 둔 단일 독점이라기보다 그 아래에 다시 여러 겹의 과점이 겹쳐 있는 다층 구조로 이해하는 편이 정확하다.**

그림 61. 노광 장비 구성품별 시장 점유율

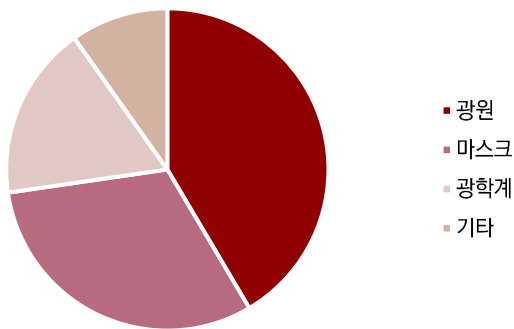
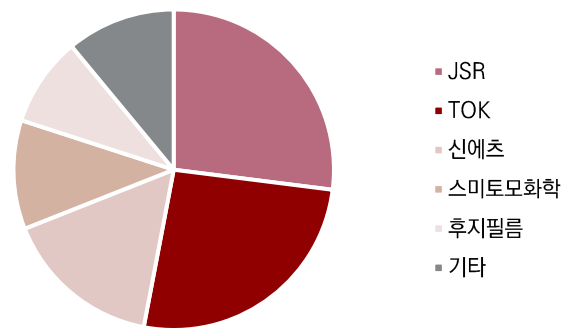


그림 62. 포토레지스트 점유율 구조



자료: Fortune Business Insights

자료: JP모건

소재 역시 장비 못지않게 집중도가 높다. 포토레지스트는 일본의 JSR, TOK, 신에츠 등 소수 기업과 미국 듀폰이 글로벌 시장을 선점해 왔으며, 특히 기술 난도가 가장 높은 EUV용 포토레지스트는 일본 기업의 과점이 오랫동안 유지되어 왔다. 마스크의 원판에 해당하는 블랭크마스크와 이를 보호하는 펠리클 역시 일본의 호야를 비롯한 소수 해외 기업이 사실상 독점해 온 영역이다. 결국 **노광 밸류체인은 장비에서 광학, 광원, 소재에 이르기까지 거의 모든 단계가 특정 소수 기업, 그것도 대다수가 일본과 유럽에 집중**되어 있다. 이러한 구조가 노광 공정의 공급 안정성을 근본적으로 취약하게 만들며, 동시에 후발 주자에게는 진입 장벽이자 국산화의 표적이 되는 지점이기도 하다.

증착(Deposition) 공정의 원리와 트렌드

증착 공정은 웨이퍼 표면에 절연막·금속막·보호막 등 얇은 막(박막)을 입히는 공정이다. 형성된 박막은 게이트 절연막, 배선, 스페이서, 하드마스크 등 다양한 소자 구조를 구성하며 이후 식각과 CMP 공정의 기준층 역할도 수행한다. 증착 방식은 화학 반응을 이용하는 CVD와 물리적으로 박막을 형성하는 PVD로 구분되며, 최근에는 원자층 단위로 막을 성장시키는 ALD의 활용이 빠르게 확대되고 있다. 특히 미세공정에서는 수십 나노미터 이하의 막 두께를 높은 균일도로 제어해야 하므로 증착 공정의 정밀도가 소자 성능과 수율을 결정하는 핵심 요소로 작용한다.

2026년 증착 공정의 핵심 기술 트렌드는 '저온 증착'과 '초미세 갭필'이다. 공정 미세화가 심화될수록 열에 민감한 구조가 증가하면서 낮은 온도에서도 고품질 박막을 형성하는 기술의 중요성이 커지고 있다. 동시에 3D NAND와

차세대 메모리의 고종횡비(HAR) 구조에서는 깊고 좁은 공간을 빈틈없이 채워야 하기 때문에, ALD의 우수한 균일도와 CVD의 높은 생산성을 결합한 하이브리드 증착 기술이 빠르게 확산되고 있다. 이러한 기술 변화는 고성능 증착 장비뿐 아니라 전구체(Precursor), 공정 소재 수요까지 함께 확대시키고 있다.

표 33. 증착 공정의 핵심 키워드

키워드	원리	강점	주 용도
저온 공정 확대	미세화로 열 손상 리스크 증가	PECVD·PEALD 채택 확대	저온 공정 확대
초미세 갭필	고종횡비 3D 구조 증가	ALD·하이브리드 방식 부상	초미세 갭필
막질 정밀 제어	누설 전류·신뢰성 요구 강화	막밀도 step coverage 관리 중요	막질 정밀 제어

자료: SK하이닉스 뉴스룸, KUVIC 리서치 1팀

이온주입(Ion Implantation) 공정의 원리와 트렌드

이온주입 공정은 붕소, 인, 비소 등의 불순물을 이온화한 뒤 가속시켜 웨이퍼 내부에 주입하는 공정이다. 이를 통해 반도체 내 P형·N형 영역을 형성하고 트랜지스터의 문턱전압(V_t), 소스·드레인 저항, 누설전류 등 전기적 특성을 제어한다. 불순물의 주입 깊이는 가속 에너지로, 농도는 주입량(Dose)으로 결정되며 이온주입 과정에서 발생하는 결정 손상을 복구하고 불순물을 활성화하기 위해 후속 어닐링(열처리) 공정이 함께 수행된다.

이온주입은 트랜지스터의 전기적 특성을 결정하는 핵심 공정으로, 미세한 오차만으로도 누설전류 증가와 성능 저하가 발생할 수 있다. 특히 GAA와 같은 차세대 트랜지스터 구조에서는 기존 평면 구조보다 도핑 균일성을 확보하기 어려워 공정 제어 난이도가 크게 높아졌다. 또한 소자의 접합 깊이가 수 nm 수준으로 얕아지면서 저에너지 이온주입과 초정밀 어닐링 기술의 중요성이 함께 확대되고 있다.

산화 공정

산화 공정이 뭘까?

산화 공정은 웨이퍼 표면의 실리콘(Si)을 고온에서 산소 또는 수증기와 반응시켜 이산화규소(SiO_2) 절연막을 형성하는 공정이다. 반응 온도는 통상 800~1,150°C 수준이며, 형성된 산화막은 회로를 불순물과 물리적 손상으로부터 보호하고 소자 간 전류 누설을 차단하는 절연체 역할을 수행한다.

산화 공정이 증착 공정(CVD·ALD)과 구분되는 지점은 막을 형성하는 방식에 있다. 증착이 웨이퍼 위에 별도의 물질을 쌓아 올리는 방식이라면, 산화는 **기판을 이루는 실리콘 자체를 소모하며 막을 성장시킨다**. 이 과정에서 성장한 산화막 두께의 약 46%는 기존 실리콘 표면보다 아래쪽에서 형성되므로, 하지 실리콘과의 계면 결함이 적고 막질이 균일하다. 이러한 계면 특성은 트랜지스터의 문턱전압 안정성과 누설 전류 특성을 좌우하며, 특히 **게이트 산화막**의 품질은 소자 신뢰성을 결정하는 출발점으로 평가된다.

공정 내 단계별 역할

산화는 반도체 8대 공정 중 두 번째 단계로 분류되지만, 실제 공정 흐름에서는 특정 시점에 한 번만 사용되는 것이 아니라 포토·식각·증착 공정 사이에 반복적으로 삽입된다. 산화막이 형성되는 대표적인 단계와 역할은 다음과 같다.

표 34. 산화막의 공정 내 적용 단계와 역할

적용 단계	형성 위치	역할
초기 산화막(때드·스크린)	웨이퍼 표면	이온주입·절취막 증착 전 표면을 보호하고 오염·손상을 방지
소자 격리막(STI 라인)	트렌치 내벽	소자 간 전기적 격리 및 트렌치 계면의 결함 안정화
게이트 산화막	게이트 하부	게이트와 채널 사이를 절연해 트랜지스터 스위칭 특성을 결정
터널 산화막(NAND)	채널-전하저장층 사이	전하의 주입·방출 경로를 형성해 데이터 저장·소거를 담당
희생 산화막	임시(형성 후 제거)	공정 중 표면을 정리하고 손상층·결함을 제거하기 위한 목적

자료: KUVIC 리서치 1팀

제품군에 따라 산화막의 비중은 달라진다. DRAM과 이를 수직 적층한 HBM에서는 셀 및 주변회로 트랜지스터의 게이트 유전막과 소자 격리막이 중심이 된다. 반면 **NAND**에서는 데이터 저장의 신뢰성을 좌우하는 터널 산화막과 전하를 가두는 **ONO(Oxide-Nitride-Oxide) 구조**가 핵심이며, 3D NAND에서 적층 단수가 높아질수록 관련 산화막의 두께 균일도와 열예산(Thermal Budget) 관리 난도가 함께 상승한다.

산화 방식별 특성 비교

산화는 반응에 사용하는 물질과 열처리 방식에 따라 막질과 성장 속도가 달라진다. 크게 반응 기체를 기준으로 건식·습식으로 구분되며, 장비·열처리 방식을 기준으로 배치형 퍼니스와 급속열처리, 라디칼 기반 저온 산화 등으로 나뉜다.

반응 기체를 기준으로 하면, **건식 산화(Dry)**는 산소를 사용해 성장 속도가 느린 대신 치밀하고 결함이 적은 고품질 막을 형성하므로 얇고 정밀한 게이트 산화막에 적합하다. **습식 산화(Wet)**는 수증기를 사용해 성장 속도가 빠른 대신 막질이 상대적으로 낮아, 두꺼운 필드 산화막이나 마스크·격리용 산화막에 주로 사용된다.

표 35. 건식 산화와 습식 산화 비교

구분	건식 산화 (Dry)	습식 산화 (Wet)
반응 물질	산소(O2)	수증기(H2O)
성장 속도	느림	빠름
막질·치밀도	우수(고품질·저결함)	상대적으로 낮음
주 용도	얇고 정밀한 게이트 산화막	두꺼운 필드·격리·마스크 산화막

자료: KUVIC 리서치 1팀

열처리 방식 측면에서도 차이가 존재한다. 전통적인 배치형 퍼니스는 다수의 웨이퍼를 한 번에 처리해 생산성이 높으나 고온·장시간 공정이 불순물 확산을 유발한다. 이 한계를 보완하기 위해 웨이퍼를 한 장씩 짧게 처리하는 급속열처리(RTP·RTO), 라디칼을 이용해 저온에서 고품질 박막을 얻는 ISSG, 계면 결함을 치유하는 고압수소어닐링 등 저열예산·고정밀 방식의 비중이 확대되고 있다.

표 36. 열처리 방식별 특징 비교

방식	처리 단위	특징	주 용도
배치형 퍼니스	다수(Batch)	고생산성·저비용, 고온·장시간 열처리	두꺼운 산화막, 대량 격리막
급속열처리(RTP·RTO) 단일 웨이퍼	단일 웨이퍼	단시간·정밀 온도제어, 열예산 절감	얇은 게이트 산화막
ISSG(라디칼 산화) 단일 웨이퍼	단일 웨이퍼	저온 라디칼 반응, 초박·고품질 박막	고신뢰 초박 산화막
고압수소어닐링(HPA) 단일 소량	단일 소량	계면 결함 치유, 수율·신뢰성 개선	선단 DRAM·NAND 신뢰성 보강

자료: KUVIC 리서치 1팀

금속배선 공정

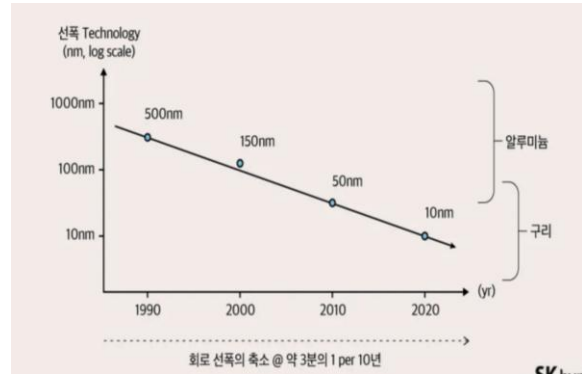
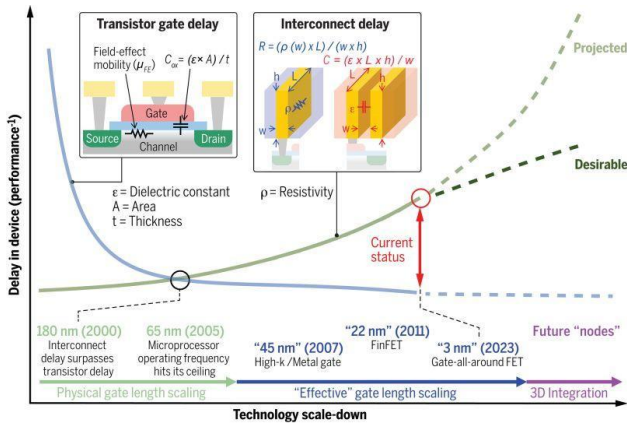
금속배선공정은 트랜지스터와 각 소자를 서로 연결해 전기 신호가 실제로 흐를 수 있도록 만드는 단계이다. 반도체는 개별 소자가 형성된 것만으로는 아무 의미가 없다. 따라서 금속배선을 통해 소자와 소자, 소자와 전원을 연결해주는 과정이 필요하며 이는 하나의 독립된 공정이 아니라 포토, 식각, 증착 등의 공정을 통해 금속배선이 형성된다.

배선 물질의 경우 소자의 구조가 고밀도로 발전하고 기능이 고속화됨에 따라 알루미늄(Al)에서 구리(Cu)로 전환되었고, 현재는 구리가 가장 보편적으로 사용되는 배선 물질이다. 알루미늄에서 구리로의 전환에 대해 조금 더 자세히 알아보면 과거에는 알루미늄이 금속 패턴을 내기 쉬운 재질이라는 점과 낮은 비저항 성분을 지니고 있다는 점으로 인해 많이 사용되었다. 그러나 기술이 발전하면서 금속선의 선폴이 좁아졌고, 이에 따라 알루미늄을 사용한 금속선은 선이 끊기거나 전류를 제대로 전달하지 못하는 등의 문제가 발생하였다. 이에 대한

대안으로 구리가 부상하였다. 구리는 알루미늄보다 약 30% 정도 낮은 전기 비저항 성질을 갖고 있어 전도도가 높기 때문에 알루미늄보다 더 작은 선폭을 구현할 수 있다. 또한 알루미늄의 단점이었던 전자 충돌에 의해 구리 입자가 움직이는 현상을 줄일 수 있었다. 이로 인해 배선이 끊어지는 간헐적 불량을 방지할 수 있게 되었고, 내구성과 신뢰성도 더 향상되었다.

그림 63. 반도체 미세화에 따른 배선 저항 및 RC 지연 증가

그림 64. 반도체 선폭 미세화에 따른 금속배선 소재의 변화



자료: Science

자료: SK 하이닉스 뉴스룸

반도체가 점점 더 미세화됨에 따라 금속배선 공정의 중요성은 커지고 있다. 왜 반도체 미세화가 금속 배선 공정의 중요성을 증가시킬까? 모든 금속은 저마다 다른 비저항 값을 가진다. 물체의 저항은 비저항이 클수록, 물체의 길이가 길수록, 단면적이 작을수록 커진다. 반도체 미세화로 인해 배선 선폭도 함께 좁아지기 때문에 전자가 부딪힐 확률이 높아지고 비저항 값도 급격히 상승하고 있다.

배선 저항의 증가는 RC 지연을 키우는 주요 원인이다. RC 지연은 배선의 저항과 정전용량에 의해 결정되는 신호 전달 지연으로, 고속 동작이 요구되는 반도체에서 중요한 성능 지표다. 공정이 미세화 될수록 트랜지스터 성능뿐 아니라 배선에서 발생하는 RC 지연이 칩 성능을 제한하는 요인으로 작용하게 되었다.

AI 가속기, HBM처럼 데이터 전송량이 큰 반도체는 매우 많은 데이터를 짧은 시간 안에 전달해야 하므로, 배선에서 발생하는 지연과 전력 손실이 전체 시스템 성능에 직접적인 영향을 미친다. 따라서 기업들은 금속 배선 층을 더 효율적으로 배치해 전력 소모를 줄이고, 초미세 공정용 금속 배선 소재를 개발하여 미세 공정으로 인한 한계점을 극복하고 있다.

CMP 공정

CMP공정은 반도체 제조 과정에서 화학적 반응과 기계적 힘을 동시에 활용하여 웨이퍼 표면을 평탄화하거나 불필요한 박막을 제거하는 공정이다. 이전 공정에서 형성된 구조물로 인해 웨이퍼 표면에 발생한 단차를 제거함으로써 후속 공정의 정밀도를 확보하는 역할을 한다. 반도체 소재가 미세화될수록 표면의 불균일성이 노광 공정의 초점 부정확성을 야기하는 문제가 심화되었으면 이를 해결하기 위한 공정으로 CMP가 자리매김했다.

핵심 원리는 화학적 작용과 물리적 연마의 유기적 결합이다. 먼저, 슬러리 내에 포함된 산화제 등의 화학 약품이 피연마 박막 표면 특성을 변화시켜 표면을 약화시킨다. 이후 슬러리 내 미세 연마 입자가 약화된 표면에 흡착되고, 웨이퍼 캐리어의 물리적 압력과 회전에 의해 해당 입자들이 표면으로부터 탈착되면서 박막이 제거되는 방식이다.

표 37.CMP 공정의 주요 변수

파라미터	의미
연마속도	단위 시간당 제거되는 박막의 두께
평탄도	연마 후 웨이퍼 표면의 평탄화 수준

파라미터	의미
균일도	웨이퍼 전면에 걸쳐 연마가 균일하게 이루어지는 정도
선택비	특정 막질을 다른 막질 대비 얼마나 선택적으로 연마할 수 있는지 나타내는 지표

CMP 공정은 반도체 소자의 미세화가 심화될수록 중요성이 더욱 부각되는 핵심 공정으로, 단순 표면 연마를 넘어 다층 배선 구조의 정밀한 형성과 후속 노광 공정의 수율 확보에 직접적인 영향을 미친다. 특히 3D NAND 및 첨단 로직 반도체의 적층 구조가 복잡해질수록 CMP 공정의 적용 횟수와 정밀도 요구 수준이 함께 증가하는 추세다. 공정 변수의 복잡성과 슬러리, 패드 등 소재의 고도화 필요성을 감안할 때, CMP 관련 정비 및 소재 분야는 반도체 공급망에서 지속적으로 높은 기술적 가치를 유지할 것으로 전망된다.

CMP 공정의 전공정 내 장비 투자 비중은 약 **3~4%** 수준으로, 주요 공정 대비 상대적으로 낮은 비중을 차지하고 있다. 비중은 낮지만, CMP 공정의 중요성과 투입 비중은 반도체 미세화 심화에 따라 구조적으로 증가하는 추세에 있다. 고집적화가 진행될수록 포토 공정이 허용하는 초점 심도(Depth of Focus)가 단차 이상으로 감소하는 문제가 발생하며, 이를 극복하기 위한 CMP 평탄화 공정의 필요성이 더욱 커지고 있다. 공정이 미세해질수록 표면 단차에 대한 허용 범위가 좁아지기 때문에, CMP 공정의 적용 횟수와 정밀도 요구 수준이 함께 상승하는 구조이다. 특히 D램 선단 공정 전환, HBM(고대역폭 메모리) 생산 확대, 로직 반도체의 고도화, 그리고 Advanced Packaging 확산이라는 반도체 산업의 방향성이 CMP 공정 수요를 구조적으로 끌어올리는 핵심 동인으로 작용하고 있다.

표 38. CMP 공정 밸류체인 내 수혜가능 기업

분야	기업	근거
CMP장비	케이씨텍	미세공정 CMP장비 국내 대표주자, 국산화 추진
슬러리 등 소재	솔브레인	공정 횟수가 늘어날수록 사용량 증가, 1Q26 슬러리 실적 가시화
계측 장비	파크시스템스	CMP 후 표면 계측 수요 증가

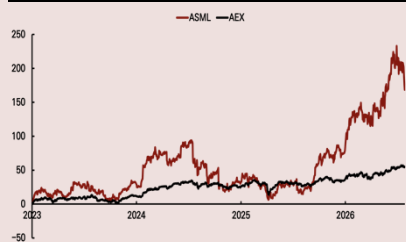
COMPANY ANALYSIS

Not Rated

Stock Information

시가총액	EUR 534.53B
발행 주식 수	384.05M
유동주식비율	99.97%
52주 최고가	EUR 1,741
52주 최저가	EUR 587.8
AEX	1,087.68

Price Trend



KUVIC Research Team 1

메일 kuvic_korea@naver.com

팀장	45기 Senior 박수홍
팀원	44기 Senior 하승훈
팀원	45기 Senior 김범준
팀원	45기 Senior 문진우
팀원	45기 Senior 명승우
팀원	45기 Senior 서준영
팀원	45기 Senior 최서운

Who We Are



ASML Holding NV (ASML)

A Solid MonopoLy

P와 Q의 동반 확대

2nm 이하 로직 공정에서는 미세화가 진행될수록 EUV 노광이 필요한 레이어 수 자체가 늘어나 웨이퍼 1장당 노광 공정 횟수가 증가하게 된다. HBM4로 넘어가며 코어 다이의 미세공정 전환과 로직 베이스 다이 도입이 맞물려, 기존에는 EUV 적용이 제한적이었던 메모리 공정에서도 EUV 레이어가 새롭게 늘어나는 추세다. 즉 로직과 메모리 양쪽에서 동시에 노광 강도가 높아지는 구조로, 두 축의 수요가 겹치면서 EUV 장비 대수 자체에 대한 요구가 가속화되고 있다. 동사는 2026년 EUV 출하량을 전년 대비 25% 늘리고 2027년엔 High-NA 비중까지 확대할 계획인데, 약 5,000억 원을 호가하는 High-NA 장비는 기존 EUV 대비 단가가 2~3배 높아 가격(P) 역시 Q와 함께 상승할 것으로 기대되는 상황이다.

전공정 최대 병목은 노광

반도체 회로가 점점 미세해지는 지금, 그 미세화 수준을 최종적으로 규정하는 기술적, 경제적 승부처는 바로 노광 공정이다. 전공정 핵심 장비 중 하나인 EUV 장비는 발주에서 제조·납품·설치·인증까지의 리드타임이 가장 길고, 단가 역시 약 2,000~5,000억 원으로 굉장히 높다. 파운드리·메모리 업체의 CapEx는 사상 최대 수준으로 확대되고 있지만, 실질적 캐파로 연결되는 속도는 노광 공정에서 물리적으로 제한된다. 즉 동사는 캐פק스 사이클에서도 가격결정력과 실적 방어력이 구조적으로 높다. 실제로 2025년 4분기 순수주가 132억 유로, 이 중 EUV가 74억 유로로 시장 예상을 크게 상회하며, 병목 공정에 대한 선제적 확보 수요가 여전히 견조함이 증명됐다.

노광 공정 내 사실상 독점 지위

니콘과 캐논이 EUV 사업에서 철수하면서 ASML은 EUV 노광장비 시장을 사실상 100% 독점하고 있다. 최근 중국의 자체 DUV 노광장비 국산화 시도가 리스크로 거론되지만, 이는 선단공정 필수 장비인 EUV와는 별개 영역의 이슈다. 중국 국영기업의 자국산 DUV 생산 계획도 2026년 5대, 2027년 약 20대 수준에 그쳐, ASML의 연간 시스템 출하 규모 대비 미미한 물량이다. 이러한 독점적 지위는 CapEx 사이클과 무관하게 ASML에 구조적인 가격 결정력과 높은 마진을 보장하는 핵심 해자로 작동한다.

Earnings and valuation metrics

결산기 (12월)	2021	2022	2023	2024	2025
매출액 (십억원)	18,611	21,173	27,559	28,263	32,674
YoY (%)	33.1%	13.8%	30.2%	2.6%	15.6%
영업이익 (십억원)	6,277	6,530	9,039	9,018	11,300
YoY (%)	68.6%	4.0%	38.4%	-0.2%	25.3%
영업이익률 (%)	33.7%	30.8%	32.8%	31.9%	34.6%
당기순이익 (십억원)	5,883	5,624	7,839	7,572	9,551
EPS (원)	14.36	14.14	19.91	19.25	24.62
P/E (배)	49.2	38.6	34.2	35.8	38.6

주: IFRS 연결 기준, 순이익은 당기순이익

자료: KUVIC 리서치 1팀

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.
- 본 보고서에 대한 지적재산권은 고려대학교 가치투자동아리 KUVIC에 있으며 어떠한 경우에도 법적 책임소재의 증빙자료로 사용될 수 없습니다.

