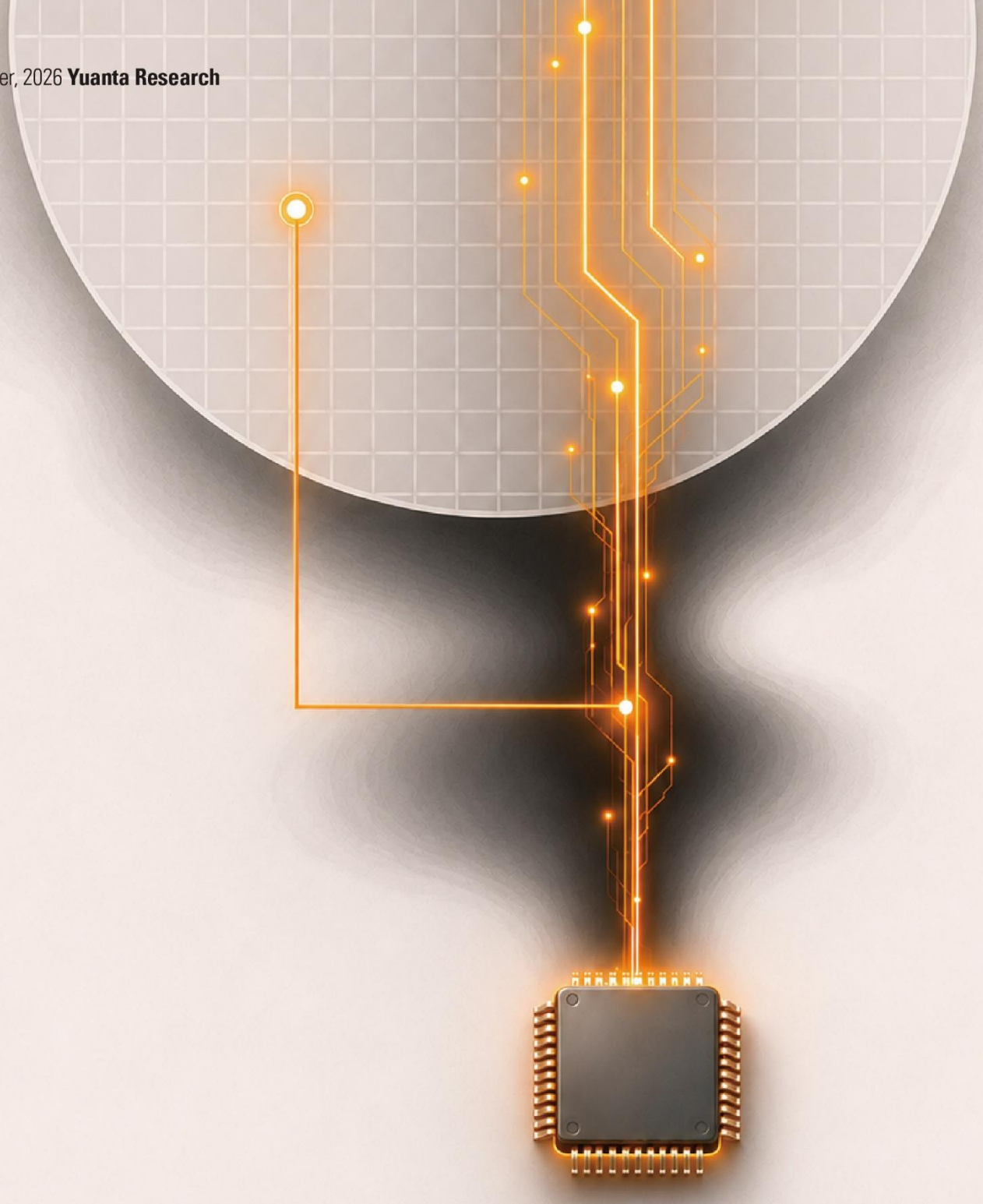


## 반도체

2027년 연간전망

# 수요의 재설계

## AI 수익화와 메모리 접근 구조의 변화

**백길현** USCPA, 반도체
02 3770 5635
gilhyun.baik@yuantakorea.com

**임석민** Research Assistant
02 3770 3648
seokmin.lim@yuantakorea.com

유안타증권
Yuanta Securities (Korea)

# 반도체/장비



백길현, USCPA 반도체
gilhyun.baik@yuantakorea.com
임석민 Research Assistant
seokmin.lim@yuantakorea.com

## 2027년 연간전망 (산업편): 수요의 재설계

### 2027년 전망: 토큰이 다시 그리는 반도체 지도

2027년 AI 반도체를 보는 가장 중요한 변수는 메모리를 얼마나 빨리, 어떤 구조로 읽느냐다. 과거 사이클의 수요 둔화가 재고 조정이었다면, 지금의 변화는 '아키텍처 조정'이기 때문이다. Compute와 AI 모델이 모두 메모리 접근의 효율화를 요구하고 있다. Compute에서는 연산 성능이 메모리 대역폭보다 빠르게 늘어나면서, 가속기 설계의 초점이 연산기의 숫자를 늘리는 것에서 필요한 데이터를 연산 유닛 가까이에 두는 것으로 옮겨가고 있다. AI 모델에서는 추론이 길어질수록 연산기가 데이터를 기다리는 Decoding 비중이 커져, 메모리 대역폭과 데이터 이동 효율이 추론의 속도와 비용을 좌우한다. 이 요구에 맞춰 새로운 메모리가 존재감을 드러내기 시작한다. 베이스다이가 로직 공정으로 넘어가 데이터 이동 경로의 일부가 되는 Custom HBM, 대역폭 한계를 푸는 MRDIMM, 연산 일부를 메모리 안에서 처리하는 PIM이 대표적이다.

### 2028~2030년: 추론 다음의 수요, 미세화 다음의 기술

2028년 이후에 대한 질문은 수요가 이어지는가, 그리고 기술이 그 수요를 따라갈 수 있는가다. 수요에서는 추론 다음 단계인 '행동'에 주목한다. 에이전트가 거래와 업무를 직접 실행하고 로봇이 움직이는 단계로, 수요를 설명하는 변수는 토큰에서 실행 건수와 기기 대수, 가동 시간으로 넓어진다. 기술의 방향은 '줄이는 것'에서 '세우고, 뒤집고, 붙이는 것'으로 바뀐다. 로직은 GAA와 후면 전력 공급, High-NA EUV로, DRAM은 $4F^2$ VCT로, NAND는 웨이퍼 본딩으로 간다. 연결은 구리에서 빛으로 넘어가고, 마지막 병목은 반도체가 스스로 풀 수 없는 전력이다.

### 반도체 업종 투자 전략

2027년은 피크아웃이 아니라, 시장이 불편해하는 우려가 하나씩 걷히면서 업종이 균형을 맞춰 우상향하는 해가 될 전망이다. 시장이 가장 불편해하는 것은 AI 수익화이며, 그 경로는 인터넷 플랫폼이 걸어온 길과 유사할 것으로 예상한다. 플랫폼은 무료로 이용자를 확보한 뒤 수익화에 나섰고, 그 사이 구간에서는 의심이 주가의 변동성을 키웠다. AI 수익화에 대한 시장의 현재 위치는 이 의심 구간의 후반부로 판단한다. 토큰이 매출이 되는 여섯 가지 경로가 열리고 있고, 수익화가 숫자로 확인된다면 빅테크와 반도체 주가는 정상화를 넘어 재평가 국면에 들어설 수 있을 것이다. 그 밖의 피크아웃 우려에 대한 당사 판단은 본문의 피크아웃 우려 점검표(표22)에 정리했다.

수익화 다음으로 봐야 할 시장은 메모리다. 메모리는 AI 반도체 안에서 입지와 중요도가 구조적으로 높아지고 있는 전략 자산이며, 비중 확대 의견을 유지한다. HBM의 웨이퍼 잠식으로 범용 DRAM의 공급 제약이 길어지는 가운데 가격은 상승 폭이 줄어도 우상향하고, 메모리는 계약의 산업으로 바뀌며, 커진 이익은 주주에게 돌아오고 있다. 메모리를 넘어선 기회는 병목 대응 전략으로 정리한다. 병목은 메모리 생산능력과 첨단 패키징(2027년), 광연결(2028년), 전력(2030년)으로 정점을 옮겨가고, 고객은 높은 가격을 감수하거나 대안을 채택한다. 대안이 채택되면서 대체 수요가 생기는 영역 가운데서는 클린룸 부족에 따른 메모리 후공정 외주의 강도가 가장 크고, LPDDR·PIM, OSAT, 삼성전자 2nm가 뒤를 잇는다. 대안이 없어 공급자가 가격 결정력을 갖는 영역은 최상단 첨단 패키징, 테스트다.

| 종목 | 투자의견 | 목표주가 (원) |
|---|---|---|
| 삼성전자 | BUY (M) | 690,000 (M) |
| SK 하이닉스 | BUY (M) | 3,700,000 (M) |
| Micron | Not Rated (M) | - (M) |
| SanDisk | Not Rated (I) | - (I) |
| Kioxia | Not Rated (I) | - (I) |
| Nvidia | Not Rated (M) | - (M) |
| AMD | Not Rated (M) | - (M) |
| Intel | Not Rated (M) | - (M) |
| Broadcom | Not Rated (M) | - (M) |
| MediaTek | Not Rated (M) | - (M) |
| Arm | Not Rated (M) | - (M) |
| TSMC | Not Rated (M) | - (M) |
| ASE | Not Rated (I) | - (I) |
| Amkor | Not Rated (I) | - (I) |
| Advantest | Not Rated (I) | - (I) |
| Keysight | Not Rated (I) | - (I) |
| Celestica | Not Rated (I) | - (I) |
| Onsemi | Not Rated (I) | - (I) |
| Texas Instruments | Not Rated (I) | - (I) |
| Renesas | Not Rated (I) | - (I) |
| ROHM | Not Rated (I) | - (I) |
| Monolithic Power | Not Rated (I) | - (I) |
| Navitas | Not Rated (I) | - (I) |
| Delta Electronics | Not Rated (I) | - (I) |
| Lite-On | Not Rated (I) | - (I) |
| Alphabet | Not Rated (M) | - (M) |
| Meta | Not Rated (M) | - (M) |
| Microsoft | Not Rated (M) | - (M) |
| Amazon | Not Rated (M) | - (M) |
| Oracle | Not Rated (M) | - (M) |
| CoreWeave | Not Rated (I) | - (I) |
| Tesla | Not Rated (M) | - (M) |
| Walmart | Not Rated (M) | - (M) |
| Shopify | Not Rated (M) | - (M) |
| Visa | Not Rated (I) | - (I) |
| Mastercard | Not Rated (I) | - (I) |
| Disney | Not Rated (I) | - (I) |

# 목차


## 01 2026년 리뷰: 병목의 이동 개시 3

AI 클러스터 대형화와 Memory Super Cycle ....3
Compute Architecture의 다변화 ....5
2026년이 남긴 질문: 재고 조정이 아닌 아키텍처 조정 ....7

## 02 2027년 전망: 토큰이 다시 그리는 반도체 지도 8

1. Compute: Domain-Specific 시대와 Rack 단위 경쟁 ....10
2. 추론의 경제학: 토큰 한 개의 손익계산서 ....12
3. 메모리의 파운드리화: 존재감을 드러내는 새로운 메모리 ....20
4. 패키징·테스트: 가격으로 번지는 병목 ....26
5. 엣지: 메모리 원가가 시험하는 소비자 기기 ....29

## 03 2028~2030년: 추론 다음의 수요, 미세화 다음의 기술 32

1. 공정: 줄이는 시대에서 세우고, 뒤집고, 붙이는 시대로 ....32
2. 패키징: 반도체의 공통 언어, 본딩 ....40
3. 연결: 빨라진 신호가 구리선을 떠나는 곳 ....44
4. 전력: 마지막 병목이자 새로운 반도체 수요처 ....48

## 04 반도체 업종 투자전략 50

1. 리스크 포인트 점검: 하나씩 걷히는 피크아웃 우려 ....51
2. 메모리: 비중 확대 유지, 늘어나는 토큰과 모자란 웨이퍼 ....56
3. 병목 대응 전략: 대체 수요와 가격 결정력 ....60

## 05 Appendix 64

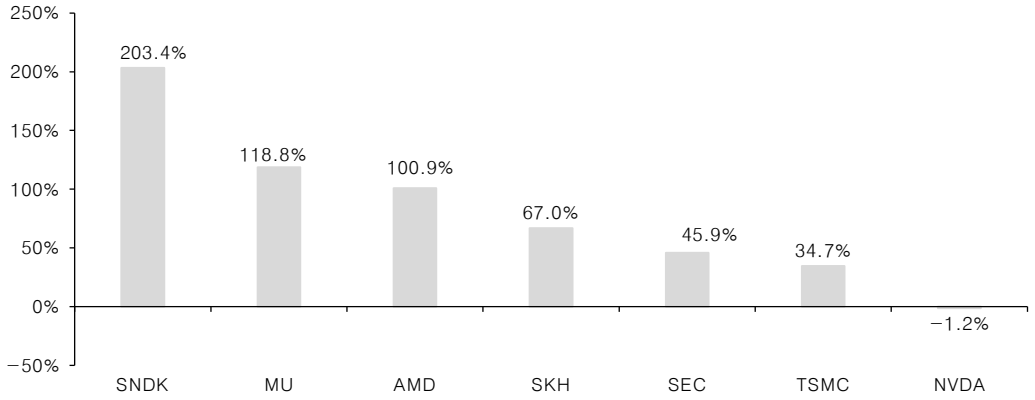
# 2026년 리뷰: 병목의 이동 개시

## AI 클러스터 대형화와 Memory Super Cycle

2026년 반도체 시장의 주인공은 단연 Memory였다. 2024~2025년 AI 인프라 투자가 GPU 확보와 연산 성능 확대에 집중됐다면, 2026년에는 AI Cluster의 대형화와 함께 병목의 중심이 Compute에서 Memory로 본격적으로 이동하기 시작했다. HBM 수요의 고성장이 지속되는 가운데 Server DRAM과 Enterprise SSD 수요까지 동시에 확대되면서, 메모리 업황은 일부 제품의 Shortage를 넘어 HBM과 Non-HBM 전반의 공급 부족으로 확산됐다.

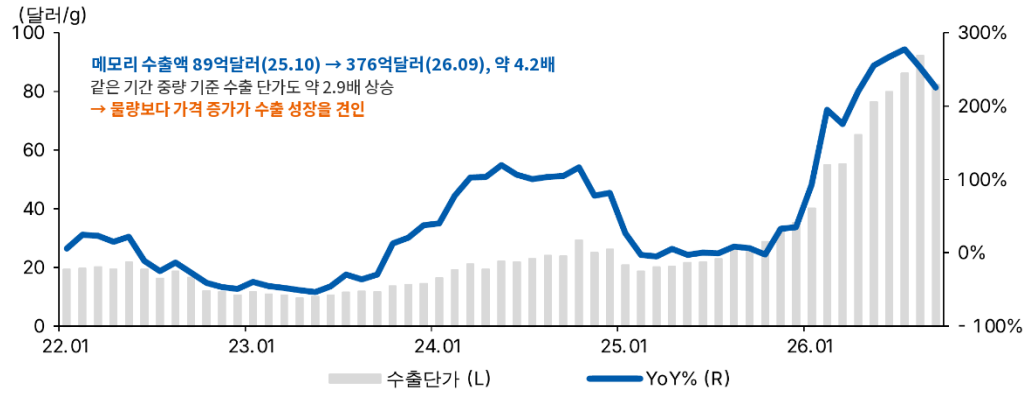
특히 2026년의 특징은 Non-HBM의 가격 상승 강도가 HBM을 압도했다는 점이다. AI Server 증설은 GPU당 HBM 수요뿐 아니라 Host CPU에 탑재되는 RDIMM과 Storage 수요를 동시에 끌어올렸다. 반면 지난 2~3년간 메모리 공급업체들의 Resource가 HBM 신규 투자에 집중되면서 선단 Conventional DRAM의 공급 여력은 크게 제한됐다. 가격은 전년 대비 큰 폭으로 상승했고, Mobile DRAM과 Consumer DRAM까지 Shortage가 확산되면서 AI가 촉발한 공급 부족은 사실상 Memory 전 영역으로 전이됐다.

차트 1. 2026 년 주요 글로벌 반도체 기업 주가 추이



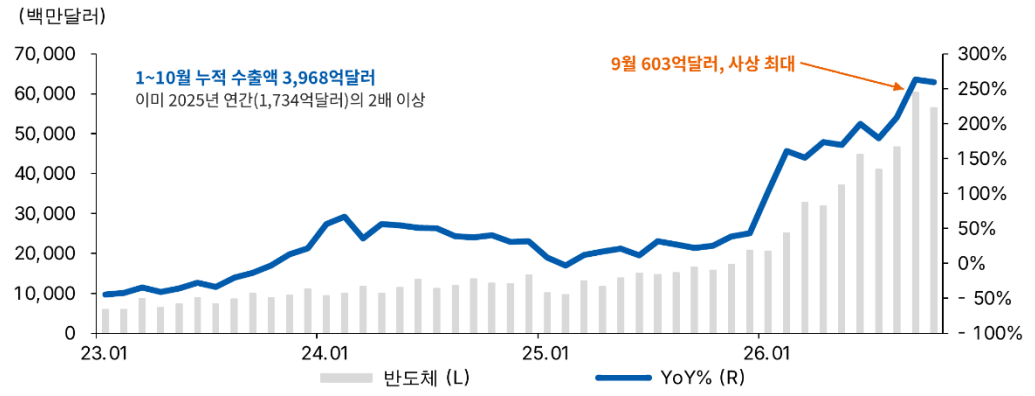

자료: Bloomberg, 유안타증권 리서치센터, 주: 26.11.06 기준 YTD 주가 수익률

차트 2. 메모리 반도체 수출 단가 추이




자료: KITA, 유안타증권 리서치센터

차트 3. 한국 반도체 월별 수출액 추이




자료: 산업통상자원부, 한국무역협회, 유안타증권 리서치센터

## Compute Architecture 의 다변화

2026년에는 AI Compute의 성장축 역시 GPU 단일 구조에서 점차 다변화되기 시작했다. NVIDIA GPU가 Training을 중심으로 AI 인프라의 핵심 지위를 유지하는 가운데, AI 서비스가 Inference 영역으로 확대되면서 모든 워크로드를 동일한 가속기로 처리하는 방식의 비용과 전력 효율 문제가 부각됐다. Hyperscaler들은 GPU 투자를 지속하는 동시에 자체 Custom ASIC과 Host CPU를 활용해 워크로드별 Compute 효율을 높이는 방향으로 인프라 구조를 바꾸기 시작했다.

가장 두드러진 변화는 ASIC의 역할 확대다. Google, Amazon, Meta 등 주요 CSP들은 특정 AI 워크로드에 최적화된 자체 가속기 개발을 확대하며 GPU 의존도를 낮추고 있다. 동시에 AI Server가 다수의 가속기와 Memory, Storage, Networking을 통합하는 시스템으로 발전하면서 이를 제어하는 Host CPU의 전략적 가치도 높아지고 있다. 2026년은 GPU를 다른 Processor가 대체한 해가 아니라, 워크로드의 규모와 종류가 확대되면서 GPU·ASIC·CPU가 역할을 분담하는 Heterogeneous Computing 구조가 형성되기 시작한 해로 판단한다.

주요 CSP들의 ASIC 로드맵에서는 세 가지 변화가 뚜렷하다. 첫째, 학습과 추론의 분리다. Google은 TPU v8을 학습용 8t와 추론용 8i로 나눴고, Microsoft의 Maia 200과 Meta의 MTIA 450·500은 추론 전용으로 설계됐다. 둘째, 설계 파트너의 다변화다. TPU v8t는 Broadcom이, 추론용 v8i는 MediaTek이 설계를 맡으면서 ASIC 공동개발 생태계가 넓어지고 있다. 셋째, 칩당 HBM 탑재량의 빠른 증가다. MTIA는 300의 216GB에서 500의 최대 512GB로, Maia는 100의 64GB에서 200의 216GB로 늘어난다.

표 1. 주요 업체별 ASIC 세대별 라인업

| 업체 | 세대 | 시점 | 사양 | 용도 |
|---|---|---|---|---|
| Google TPU | v7 Ironwood | 2025.4 공개, 2025.11 GA, 설계 Broadcom | HBM 192GB (7.37TB/s), FP8 4.6PF, 최대 9,216칩 | 학습 · 추론 통합 |
| | v8 – 8t (Sunfish) | 2026.4.22 공개, 설계 Broadcom | 칩당 FP4 12.6PF, HBM3e 216GB (6,528GB/s), SRAM 128MB, 슈퍼포드 9,600칩 (121 EF FP4, HBM 2PB) | 학습 |
| | v8 – 8i (Zebrafish) | 2026.4.22 공개, 설계 MediaTek | 칩당 FP4 10.1PF, HBM3e 288GB (8,601GB/s), SRAM 384MB, 포드 1,152칩 | 추론 |
| | v9 (Humufish) | 2027년 하반기 목표 | Intel EMIB-T 채택 보도, 메모리 HBM4 | 차세대 |
| | v10 | 공식 사양 없음 | EMIB-T 지속 채택을 전망 | 차세대 |
| AWS Inferentia, Trainium | Inferentia1 | 2019년, 16nm | NeuronCore-v1 4개, DRAM 8GB (50GB/s), INT8 128 TOPS, FP16/BF16 64 TFLOPS | 추론 |
| | Inferentia2 | 2022~2023년, 7nm | NeuronCore-v2 2개, HBM 32GB (820GB/s), FP16/BF16/cFP8/TF32 190 TFLOPS, INT8 380 TOPS, FP32 47.5 TFLOPS. Inf2 인스턴스는 최대 12칩 (2,280 TFLOPS, 메모리 384GiB) | 추론(LLM, 분산 추론 대응) |
| | Trainium2 | – | HBM3 96GB, 2.9TB/s | 학습 · 추론 |
| | Trainium3 | TSMC 3nm, 2025.12 GA | HBM3e 144GB (4.9TB/s), FP8 2.52PF, 울트라서버 최대 144칩 (362 PF FP8, HBM 20.7TB) | 학습 · 추론 |
| | Trainium4 | 2027년 인도 예상 | Trainium3 대비 FP4 6x, FP8 3x, 대역폭 4x, 메모리 용량 2x, NVLink Fusion 지원 | 학습 · 추론 |
| Meta MTIA | MTIA 300 | 양산 중 (2026.3 공개) | HBM 216GB (6.1TB/s), FP8 1.2PF, TDP 800W | R&R 학습 |
| | MTIA 400 | 랩 테스트 완료, 데이터센터 배치 진행 단계 | HBM 288GB (9.2TB/s), MX4 12PF, FP8 6PF, TDP 1,200W, 스케일업 72칩 | R&R + 범용 GenAI |
| | MTIA 450 | 2027년 초 대량 배치 | HBM 288GB (18.4TB/s), MX4 21PF, FP8 7PF, TDP 1,400W | GenAI 추론 |
| | MTIA 500 | 2027년 대량 배치 | HBM 384~512GB (27.6TB/s), MX4 30PF, FP8 10PF, TDP 1,700W | GenAI 추론 |
| Microsoft Maia | Maia 100 | 2023.11 공개, TSMC 5nm | HBM2E 64GB | 내부용 |
| | Maia 200 | 2026.1.26 공개, TSMC 3nm | HBM3e 216GB (7TB/s), SRAM 272MB, FP4 10PF+, FP8 5PF+, TDP 750W | 추론 전용 |
| | Maia 300 | 2026년 가을 공개 예정, 2027년 인도 목표 | 공식 사양 미공개 | 미공개 |

자료: 업계자료, 언론보도, 유안타증권 리서치센터

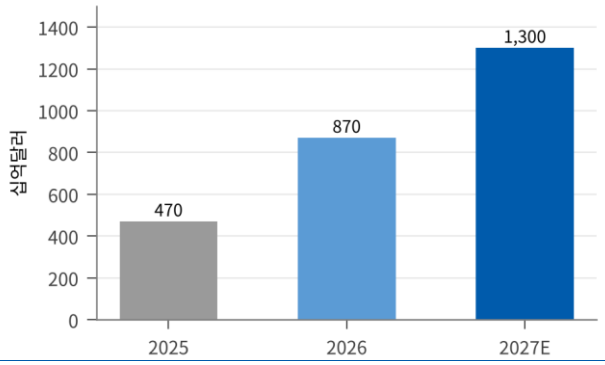
## 2026년이 남긴 질문: 재고 조정이 아닌 아키텍처 조정

2026년 하반기 시장의 화두는 피크아웃이었다. Hyperscaler CapEx의 지속 가능성, Rubin Ultra의 HBM 사양 하향 논의, 메모리 가격 급등에 따른 소비자 기기 수요 위축 등이 번갈아 부각되며 업종 변동성을 키웠다. 그러나 투자 규모 자체는 꺾이지 않았다. S&P Global에 따르면 주요 6개 대형 IT 기업의 설비투자는 2025년 4,700억달러에서 2026년 8,700억달러, 2027년 1.3조달러로 확대될 전망이다.

하반기의 사건들을 다시 보면 공통점이 있다. 메모리 가격이 오르자 고객들은 구매량을 줄이기보다 설계를 바꾸기 시작했다. NVIDIA는 Rubin Ultra의 HBM 단수 하향을 검토했고, CSP들은 자체 ASIC의 HBM 용량을 다시 설계했으며, 서버에는 LPDDR 기반 모듈이, PC에는 SSD 용량 축소가 등장했다. **과거 사이클의 수요 둔화가 재고 조정이었다면, 지금의 변화는 아키텍처 조정이다. 따라서 2027년에 던져야 할 질문은 '수요가 꺾이는가'가 아니라 '고객이 메모리를 어떻게 다르게 쓰기 시작했고, 그 변화가 병목과 이익을 어디로 옮기는가'다.**

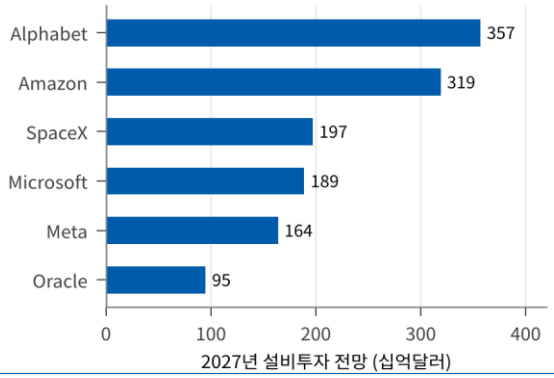
이 질문에 대한 업계의 답은 이미 나오고 있다. **2026년 10월 JEDEC 포럼에서 메모리·로직·장비 업체들이 제시한 Custom HBM, PIM, 서버 메모리의 분화, LPCAMM2는 모두 메모리를 '더 많이'가 아니라 '다르게' 쓰는 방법이다.** 메모리를 쓰는 방식이 바뀌면 병목은 칩 내부에서 메모리 인터페이스, 패키징, 칩 간 연결, 테스트로 옮겨가고, 가격 결정력과 이익도 그 경로를 따라 움직인다. 본 보고서가 2027년 병목이 시스템 전반으로 확산된다는 관점에서 업종 전반을 점검하는 이유다.

차트 4. 주요 6 개 대형 IT 기업 설비투자 합계 추이




자료: S&P Global, 유안타증권 리서치센터 주: 오른쪽 6 개 기업 합산(SpaceX 포함)

차트 5. 2027 년 기업별 설비투자 전망



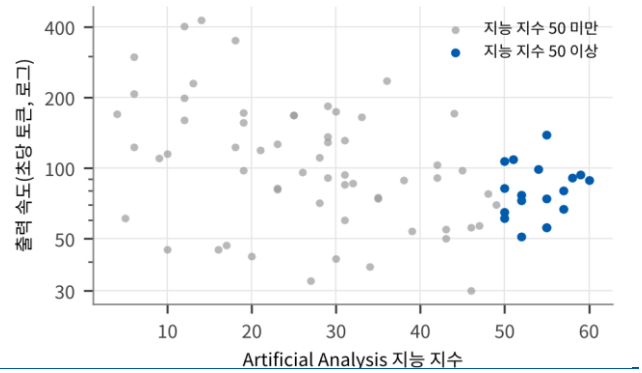

자료: S&P Global, 유안타증권 리서치센터

## 2027년 전망: 토큰이 다시 그리는 반도체 지도

**AI 하드웨어의 근본적인 제약은 '연산'과 '데이터 공급 속도'의 격차다.** Gholami 외(2024)의 'AI and Memory Wall'에 따르면 지난 20년간 서버 하드웨어의 Peak FLOPS는 2년마다 약 3.0배씩 늘어 누적 약 6만배 증가했지만, DRAM 대역폭은 2년마다 1.6배(누적 약 100배), 인터커넥트 대역폭은 1.4배(누적 약 30배) 증가하는 데 그쳤다. Synopsys 역시 GPU 연산 성능은 4년마다 두 자릿수 배수로 늘어나는 반면 메모리 대역폭은 같은 기간 한 자릿수 배수에도 못 미친다고 지적했다.

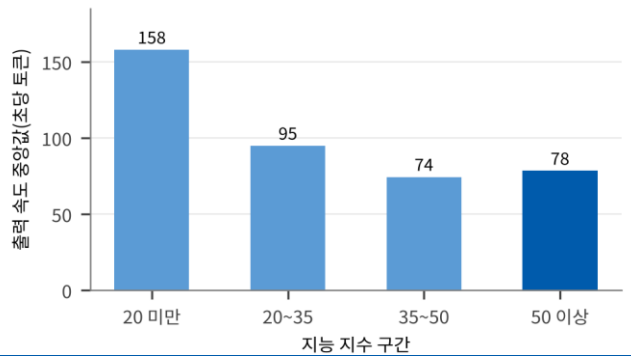
**이 격차는 모델이 똑똑해질수록 더 크게 체감된다.** 삼성전자는 Frontier 모델의 지능 지수(Artificial Analysis Intelligence Index)가 꾸준히 상승하는 반면, 지능이 높은 모델일수록 초당 출력 토큰 수(Output TPS)는 오히려 낮아진다는 점을 제시했다. Artificial Analysis가 집계한 모델별 성능을 보면, 지능 지수 20 미만인 소형 모델의 출력 속도 중앙값은 초당 158토큰인 반면 지능 지수 50 이상인 Frontier 모델은 78토큰으로 절반 수준에 그친다. **모델이 커지고 추론 과정이 길어질수록 연산기가 데이터를 기다리는 시간이 늘어나기 때문이다.**

차트 6. 모델별 지능 지수와 출력 속도




자료: Artificial Analysis(2026.11.5), 유안타증권 리서치센터 주: 모델 설정별 표시

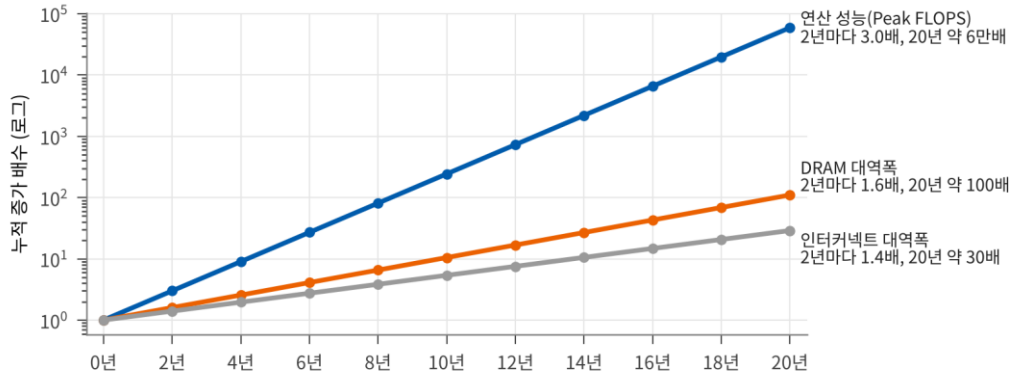
차트 7. 지능 지수 구간별 출력 속도 중앙값




자료: Artificial Analysis(2026.11.5), 유안타증권 리서치센터

**주목할 점은 가장 느리게 개선되는 영역이 메모리가 아니라 연결(Interconnect)이라는 사실이다.** 2026년의 병목이 메모리였다면, 2027년에는 메모리를 포함해 칩과 칩, 패키지와 보드, 랙과 랙을 잇는 모든 데이터 경로가 병목이 된다. 본 장에서는 먼저 병목을 만드는 수요 동인인 Compute의 변화와 추론·에이전트 워크로드를 살펴본 뒤, 실제로 공급이 막히는 메모리와 패키징, 그리고 구조적으로 수요가 늘어나는 테스트와 엣지를 차례로 점검한다. 공정과 연결은 03장에서 다룬다. 병목별 강도 비교와 투자 우선순위는 마지막 투자전략에서 정리한다.

차트 8. AI Memory Wall: 연산 성능과 대역폭의 증가 속도 격차




자료: Gholami et al.(2024) 'AI and Memory Wall', 유안타증권 리서치센터 주: 2 년마다의 평균 증가율을 누적한 지수

## 1. Compute: Domain-Specific 시대와 Rack 단위 경쟁

MediaTek은 컴퓨팅의 진화를 세 번의 파동으로 정리한다. 첫 번째는 싱글코어에서 멀티코어로, Moore's Law 스케일링을 기반으로 한 0.5kW 미만의 범용 CPU 시대다. 두 번째는 대규모 병렬 처리와 2.5D 패키징, 4배 레티클 크기의 칩을 Scale-up·Scale-out으로 묶는 1~2kW급 GPU 시대다. **그리고 현재 진입한 세 번째 파동은 Custom Silicon(ASIC), 3D 이상의 패키징, 웨이퍼 단위 통합, 최대 HBM 대역폭과 용량을 결합한 5kW 이상의 Domain-Specific 시대다. 칩 하나가 처리하는 작업의 범위는 특정 워크로드로 좁아지는 대신, 그 작업에 쏟는 연산 성능과 전력은 폭발적으로 늘어나는 구조다. 범용성을 내주고 전문화를 택하는 방향이다.**

표2. 컴퓨팅 진화의 세 가지 파동

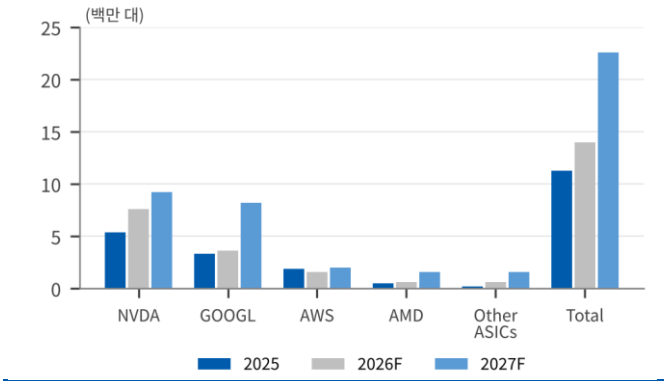
| 구분 | Wave 1: CPU | Wave 2: GPU | Wave 3: Domain-Specific |
|---|---|---|---|
| 핵심 구조 | 싱글코어 → 멀티코어, 가상화 | 대규모 병렬 처리 | Custom Silicon(ASIC) |
| 스케일링 | Moore's Law | Scale-up + Scale-out(디바이스·노드·랙) | 웨이퍼 단위 통합, 양방향 인터커넥트 |
| 패키징·칩 크기 | 1배 레티클 미만 | 2.5D, 4배 레티클 이상 | 3D 이상, 10배 레티클 이상 |
| 메모리 | DDR 중심 | HBM 도입 | HBM 대역폭·용량 극대화, Custom HBM |
| 칩당 전력 | 0.5kW 미만 | 1~2kW | 5kW 이상 |
| 핵심 지표 | 단일 코어 성능 | GPU Peak Performance | Perf/TCO, Perf/Watt(랙 단위) |

자료: MediaTek(JEDEC Industry Forum, 2026.10), 유안타증권 리서치센터

이 변화는 경쟁의 단위를 바꾸고 있다. MediaTek의 표현을 빌리면 'Rack is the new unit of compute'다. NVIDIA의 Vera Rubin NVL72는 72개의 Rubin GPU와 36개의 Vera CPU를 NVLink 6 기반의 하나의 Rack-scale Platform으로 통합하고, AMD 역시 Helios에서 72개의 MI455X GPU와 EPYC CPU, Pensando Networking을 하나의 Rack으로 묶는다. 과거에는 GPU의 Peak Performance와 출하 수량이 투자의 핵심 변수였다면, 앞으로는 얼마나 많은 가속기를 하나의 시스템으로 연결하고 높은 Utilization으로 유지할 수 있는지가 더 중요해진다. 전력·열·신뢰성이 성능을 결정하는 단계로 진입한 것이다.

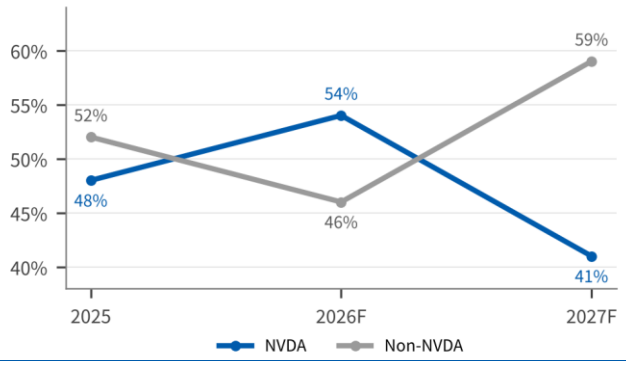
**가속기 시장은 역할 분화와 함께 성장할 전망이다.** 범용 GPU의 강점은 여전히 유연성과 소프트웨어 생태계에 있으며, 모델 구조가 빠르게 바뀌는 Frontier 모델의 학습에서는 GPU의 활용도가 높을 수밖에 없다. 반면 동일한 연산을 막대한 규모로 반복하는 추론 영역에서는 전력 효율과 연산 비용이 중요해지면서 Custom ASIC의 경제성이 재차 부각되고 있다. Google은 차세대 TPU를 학습용 8t와 추론용 8i로 세분화했고, Meta는 MTIA를 추천·추론에서 학습 영역까지 확대하고 있다. **ASIC 확대는 Big Tech 내부에서 모든 부가가치를 흡수하는 것이 아니라, 설계·IP·네트워크·파운드리·첨단 패키징으로 이어지는 외주 생태계를 동시에 키우고 있다.**

차트 9. 주요 업체별 가속기 출하량 추이




자료: 업계자료, 유안타증권 리서치센터

차트 10. NVDA vs Non-NVDA 진영 가속기 점유율 추이




자료: 업계자료, 유안타증권 리서치센터

**GPU와 ASIC의 동반 팽창은 이들을 총괄할 Host CPU 수요를 함께 끌어올린다.** 상위 5개 공급사 기준 글로벌 서버 CPU 공급량은 2025년 2,470만대에서 2027년 5,000만대 규모로 늘어날 전망이다. 에이전트 AI는 한 번의 추론으로 작업이 끝나지 않고 외부 도구 호출(Tool-call), 코드 실행(Code Execution), 정보 검색(RAG Retrieval) 등 여러 단계를 반복한다. 이 단계들은 GPU가 아니라 CPU에서 처리되기 때문에, 단계 사이의 흐름을 관리하는 CPU의 작업 조율(Orchestration)이 늦어지면 GPU는 다음 작업을 기다리며 쉬게 된다. **CPU 성능이 시스템 전체의 처리량과 GPU 가동률을 좌우하는 이유다.** Google Axion, AWS Graviton, NVIDIA Vera 등 ARM 기반 자체 CPU의 확산과 AMD EPYC의 점유율 확대가 2027년 서버 CPU 시장의 핵심 변화가 될 것이다.

표3. 에이전트 AI 의 주요 작업 단계와 용어 설명

| 용어 | 뜻 | 예시 | 주로 쓰이는 하드웨어 |
|---|---|---|---|
| Tool-call (도구 호출) | AI 모델이 직접 답하지 않고 외부 프로그램이나 서비스(API)를 불러 기능을 쓰는 것 | 쇼핑몰 상품 검색, 항공권 조회, 일정 등록 | CPU, 네트워크 |
| Code Execution (코드 실행) | AI 가 작성한 코드를 실제로 실행해 계산이나 데이터 처리 결과를 얻는 것 | 엑셀 데이터 분석, 차트 생성, 수치 계산 | CPU, 서버 DRAM |
| RAG Retrieval (정보 검색) | 답하기 전에 사내 문서나 데이터베이스에서 관련 정보를 찾아 모델에 함께 넣는 것 | 사내 규정 검색 후 답변, 최신 뉴스 반영 | CPU, 서버 DRAM, SSD |
| Orchestration (작업 조율) | 추론, 도구 호출, 검색, 재추론으로 이어지는 여러 단계의 순서와 흐름을 관리하는 것 | 예약 에이전트가 검색 → 비교 → 결제를 차례로 진행 | Host CPU |
| GPU 가동률 (Utilization) | GPU 가 실제로 연산하는 시간의 비율. CPU 단계가 늦어지면 GPU 가 대기하며 낮아짐 | - | GPU |

자료: 유안타증권 리서치센터

## 2. 추론의 경제학: 토큰 한 개의 손익계산서

AI 추론은 성격이 전혀 다른 두 단계로 나뉜다. 추론을 입력을 처리하는 Prefill과 토큰을 하나씩 생성하는 Decoding이다. Prefill은 행렬-행렬 곱셈(GEMM)이 주를 이루는 Compute-bound 구간으로 데이터 재사용률이 높아 GPU·NPU 등 기존 가속기가 효율적으로 처리한다. 반면 Decoding은 행렬-벡터 곱셈(GEMV)이 주를 이루는 Memory-bound 구간으로, 토큰을 하나 만들 때마다 가중치 데이터를 메모리에서 반복해 읽어와야 한다. **GEMM 가속 하드웨어는 충분히 발전했지만, GEMV에 특화된 하드웨어는 거의 없어, Decoding 구간에서 가속기 Utilization이 낮아지고 메모리 대역폭이 추론 성능의 병목이 된다.**

표4. AI 추론의 두 단계: Prefill vs Decoding

| 구분 | Prefill | Decoding |
|---|---|---|
| 주요 연산 | GEMM(행렬 × 행렬) | GEMV(행렬 × 벡터) |
| 병목 | Compute-bound | Memory-bound |
| 데이터 재사용 | 높음 | 낮음(가중치 반복 이동) |
| 적합한 하드웨어 | GPU, NPU, TPU 등 기존 가속기 | 전용 하드웨어 부재<br>→ PIM, Custom HBM 등 대안 부상 |
| 토큰 단가 | 입력 토큰(상대적으로 저렴) | 출력 토큰(상대적으로 비쌈) |

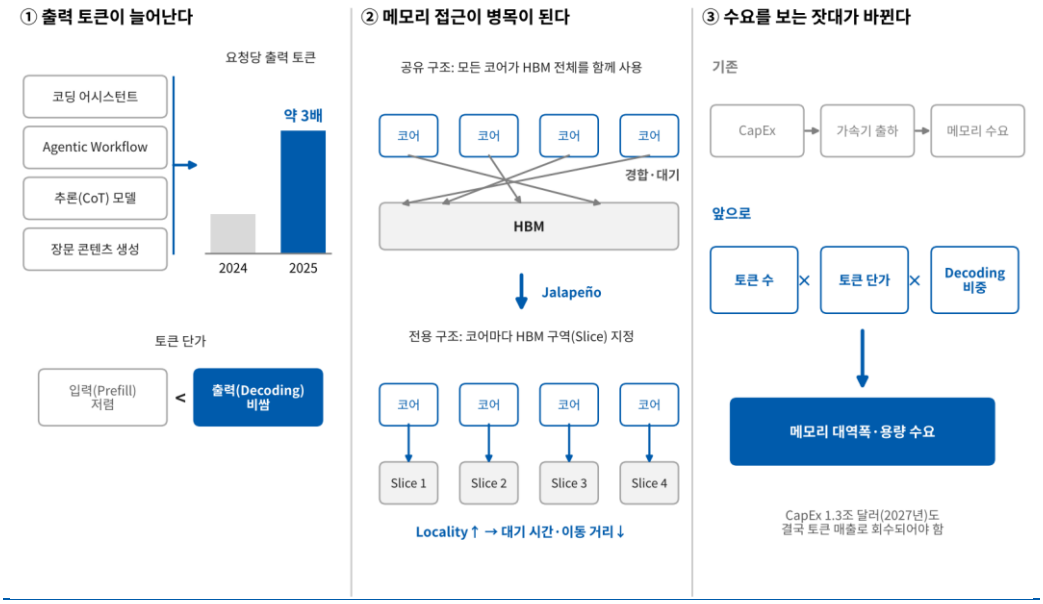
자료: SK 하이닉스(JEDEC Industry Forum, 2026.10), Artificial Analysis, 유안타증권 리서치센터

**문제는 Decoding의 비중이 빠르게 커지고 있다는 점이다.** OpenRouter에 따르면 요청당 출력(Completion) 토큰 수는 2024년부터 2025년 사이 약 3배 증가했다. 코딩 어시스턴트의 확산, 단순 채팅에서 Agentic Workflow로의 전환, 추론(Chain of Thought) 모델의 확대, 장문 콘텐츠 생성 수요가 동시에 작용한 결과다. **Artificial Analysis가 집계한 주요 모델의 가격표에서도 출력 토큰 단가는 입력 토큰보다 일관되게 높다. 사용자에게도 Decoding이 Prefill보다 비싼 자원이라는 의미다.**

**이는 메모리 수요를 설명하는 변수가 CapEx에서 토큰으로 옮겨가고 있음을 뜻한다. 출력 토큰이 늘수록 Decoding 구간이 길어지고, 그만큼 메모리 대역폭과 데이터 이동 효율이 추론 비용을 결정하기 때문이다.** OpenAI가 개발 중인 AI 전용 칩(Jalapeño)의 설계도 같은 고민에서 출발한다. 일반적인 가속기에서는 여러 연산 코어가 HBM 전체를 함께 쓰기 때문에, 여러 코어가 동시에 같은 메모리에 접근하면 순서를 기다리는 경합이 생긴다. Jalapeño는 HBM을 여러 구역(Slice)으로 나눠 코어마다 전용 구역을 정해주는 구조로 알려져 있다. 코어가 자기 옆의 정해진 메모리만 쓰면 기다리는 시간이 줄고 데이터가 이동하는 거리도 짧아진다. **필요한 데이터를 연산 유닛 가까이에 두는 이 원리를 Locality(지역성)라고 부른다. 2027년 AI 가속기 설계의 경쟁력은 연산기 숫자보다 메모리 접근 구조를 얼마나 최적화하느냐에서 갈릴 가능성이 높다.**

수요를 추정하는 기준도 바뀌어야 한다. S&P Global 전망 기준 주요 6개 기업의 설비투자는 2027년 1.3조달러로 늘어나지만, **이 가운데 같은 해 잉여현금흐름이 플러스로 예상되는 곳은 한 곳에 불과하다. 결국 이 투자는 토큰 매출로 회수돼야 한다는 뜻이다.** **CSP 설비투자에서 출발하는 하향식 추정만으로는 한계가 있으며, 출력 토큰 증가율, 토큰당 단가, Decoding 비중처럼 추론의 경제성을 설명하는 변수로 메모리 수요를 함께 점검할 필요가 있다.**

그림 1. 토큰이 메모리 수요를 결정하는 경로




자료: OpenRouter, Artificial Analysis, S&P Global, 업계자료, 유안타증권 리서치센터 주: Jalapeño 구조는 알려진 내용 기준 개념도

## 토큰이 매출이 되는 여섯 가지 경로

AI 투자가 결국 토큰 매출로 회수돼야 한다면, 가장 먼저 확인해야 할 것은 토큰이 실제로 돈이 되고 있느냐다. 2026년 하반기에 들어 그 답이 숫자로 확인되기 시작했다. 토큰이 매출이 되는 경로는 하나가 아니라 여섯 갈래로 동시에 열릴 전망이다. 기업이 토큰을 직접 사는 종량제, 무료 사용자의 토큰을 광고로 바꾸는 경로, 에이전트가 거래를 만들어 수수료를 받는 경로, 소프트웨어가 사람의 일을 대신하고 그 결과로 돈을 받는 경로에 더해, 에이전트끼리 서로 값을 치르는 경로와 국가가 직접 토큰을 사는 경로가 새로 열릴 것이다.

표5. 토큰이 매출이 되는 여섯 가지 경로

| 경로 | 과금 방식 | 대표 사례와 근거 | 성격 |
|---|---|---|---|
| 기업이 토큰을 직접 산다 | 사용량 기반(종량제) | Anthropic 2026년 상반기 매출의 80%가 사용량 기반, 매출총이익률 85% | 가장 크고 이미 확인된 경로 |
| 토큰이 광고가 된다 | 답변 내 광고 | ChatGPT 광고, 출시 약 200일 만에 연환산 10억달러 | 무료 사용자의 수익화 |
| 토큰이 거래를 만든다 | 거래 수수료 | 메타 Muse의 판매자 수수료 모델 | 소비자 영역의 다음 단계 |
| 토큰이 일을 대신한다 | 성과 기반 과금 | 해결 건수·업무 단위 과금 확산(Salesforce 등) | 소프트웨어 가격 체계의 변화 |
| 에이전트끼리 값을 치른다 | 건당 소액 결제 | Visa·Artemis, 에이전트 결제를 사람 대신 사는 거래와 소프트웨어 간 소액 결제로 구분, Visa·Mastercard·Stripe 등 스테이블코인 OUSD 출시(2026.6) | 사람의 시간과 무관한 상시 수요 |
| 국가가 토큰을 산다 | 국가 단위 인프라·사용 계약 | 2026년 각국 소버린 AI 연산 투자 약정 1,000억달러 이상(Deloitte) | 수요처의 분산 |

자료: Anthropic S-1, OpenAI, Meta, Salesforce, Visa·Artemis, Deloitte, 유안타증권 리서치센터

**가장 크고 이미 확인된 경로는 기업이 쓴 만큼 돈을 내는 종량제다.** 10월 공개된 Anthropic의 상장 신고서(S-1)에 따르면 2026년 상반기 매출 163억달러 가운데 80%인 131억달러가 사용량 기반 매출이었다. 분기 매출은 1분기 47억달러에서 2분기 116억달러로 늘었고, 회사가 투자자에게 제시한 3분기 매출은 168억달러로 연환산 약 670억달러 수준이다. 더 중요한 것은 수익성이다. 매출총이익률은 2025년 상반기 12%에서 2026년 상반기 85%로 높아졌다. 토큰을 팔아 이익을 남기기 시작했다는 뜻으로, 'AI는 돈이 되지 않는다'는 의구심에 대한 가장 직접적인 반론이다. 연간 10만달러 이상을 쓰는 고객 수도 1년 만에 약 950곳에서 1만 1,900곳 이상으로 늘었다.

**반대편에는 2,656억달러 규모의 고정 컴퓨트 약정이 있다. 토큰 매출이 늘어나는 속도가 곧 이 약정, 그리고 그 안에 담긴 메모리 수요가 회수되는 속도다.** 무료 사용자의 토큰도 돈이 되기 시작했다. OpenAI는 2월 ChatGPT 답변 안에 광고를 넣기 시작해 약 200일 만에 연환산 광고 매출 10억달러를 넘었고, 연말 목표를 25억달러로 제시했다. 기업용 소프트웨어에서는 사용자 수에 따라 받던 가격이 처리한 업무 단위나 해결 건수에 따라 받는 성과 기반 과금으로 옮겨가고 있다. 토큰 사용량 자체도 폭발적으로 늘고 있다. 구글이 처리하는 토큰은 2024년 월 9.7조 개에서 2025년 480조 개, 2026년 3,200조 개로 늘었다.

**새로 열리는 첫 번째 경로는 에이전트끼리 값을 치르는 시장이다.** 지금의 에이전트 커머스가 사람을 대신해 물건을 사는 거래라면, **그다음 단계는 에이전트가 다른 에이전트의 API나 연산을 건당 1달러에도 못 미치는 금액으로 수천 번씩 사들이는 거래다.** Visa는 7월 Artemis와 함께 낸 보고서에서 에이전트 결제를 사람 대신 사는 매크로 커머스와 소프트웨어끼리 오가는 마이크로 커머스로 나누고, 카드 수수료가 부담스러운 소액 결제는 스테이블코인이 맡게 될 것으로 봤다. 6월에는 Visa와 Mastercard, Stripe 등이 함께 스테이블코인 Open USD(OUSD)를 내놨다.

이 시장의 토큰 수요는 사람이 깨어 있는 시간과 상관없이 24시간 이어진다. 에이전트가 쉬지 않고 작업을 나누고 서로를 호출하는 만큼, 작업 흐름을 관리하는 Host CPU와 그 옆의 서버 DRAM 수요가 함께 커진다. 앞서 에이전트 워크로드가 GPU 밖의 메모리 수요를 키운다고 설명한 논리가 그대로 이어지는 경로다.

**두 번째는 국가가 직접 토큰을 사는 경로다. 소버린 AI는 각국 정부가 자국의 데이터와 언어, 안보를 위해 자체 AI 인프라와 모델을 갖추려는 움직임으로, 토큰을 사는 주체를 빅테크와 기업 밖으로 넓힌다.** Deloitte는 2026년 각국이 소버린 AI 연산 구축에 1,000억달러 이상을 약정할 것으로 추정했고, 현재 전 세계 AI 연산의 약 10%에 불과한 미국과 중국 밖의 비중이 2030년까지 두 배 가까이 늘 것으로 봤다. 일본의 2030년까지 650억달러 규모 투자 패키지 제안, 유럽의 200억유로 규모 AI 기가팩토리 계획, 중동의 230억달러 규모 파트너십이 대표적이다. 반도체 수요가 소수 빅테크의 투자에 쏠려 있다는 우려를 덜어주는 경로이다.

다만 모든 토큰이 매출이 되는 것은 아니다. 대표적인 예가 영상 생성이다. OpenAI는 3월 영상 생성 서비스 Sora의 종료를 발표했는데, 하루 약 100만달러에 이르는 연산 손실이 배경으로 보도됐다. 10초짜리 영상 한 편을 만드는 데 약 1.3달러가 들어 텍스트 질의 수천 건에 해당하는 자원을 썼고, 함께 추진되던 Disney의 10억달러 투자 논의도 무산됐다. 토큰 소비로만 보면 영상 생성은 가장 큰 시장이지만, 원가가 매출을 앞서는 한 매출로 이어지지 않는다. 토큰을 많이 쓰는 시장일수록 토큰당 원가를 낮추는 메모리 대역폭과 추론 효율이 먼저 해결돼야 열린다는 뜻이다. 거꾸로 보면 하드웨어 개선으로 토큰 원가가 충분히 낮아지는 시점에는 영상 생성이 가장 큰 토큰 매출 시장으로 다시 열릴 수 있다는 의미이기도 하다.

### 에이전트 커머스: 소비자 영역에서 토큰이 거래가 되는 경로

**여섯 갈래 가운데서도 반도체 수요에 미치는 함의가 가장 큰 것은 에이전트 커머스다.** 소비자 영역에서 AI가 처리한 작업이 곧바로 거래로 이어지는 경로로, AI 수익화에 대한 의구심에 가장 구체적인 답을 내놓기 시작했다. 메타가 9월 8일 미국과 캐나다에 출시한 AI 에이전트 Muse는 출시 6주 만에 1,000만명 이상의 사용자를 확보했고, 사용자의 약 4분의 1이 매일 이를 쓰고 있다. 저커버그는 10월 실적 발표에서 쇼핑이 Muse의 가장 많은 사용처이며 평균 구매액이 200달러를 넘는다고 밝혔다. 수요가 예상을 웃돌면서 출시 첫 주에 결제 파트너인 Stripe와 신용 한도를 상향해야 했을 정도다. OpenAI도 유사한 에이전트 서비스 Dots를 내놓으며 경쟁에 뛰어들었다.

의미 있는 것은 수익 구조다. Muse는 광고가 아니라 에이전트가 거래를 성사시켰을 때 판매자에게 수수료를 받는 모델이다. 소비자는 상품 가격 외에 별도 이용료를 내지 않으며, 더 많은 토큰이나 고급 모델을 원할 때만 비용을 낸다. 지금까지 AI 서비스의 수익화가 구독료나 광고 효율 개선이라는 간접 경로에 머물렀다면, 에이전트 커머스는 소비자 영역에서 AI가 처리한 작업이 거래 단위로 매출이 되는 첫 사례다. 앞서 설비투자가 결국 토큰 매출로 회수돼야 한다고 언급한 바 있는데, 그 회수 경로가 처음으로 숫자로 확인되기 시작한 것이다.

에이전트 커머스는 광고 예산이 흐르는 자리도 바꾼다. 지금의 디지털 광고는 사람이 검색하고 비교하고 클릭하는 과정에 광고를 끼워 넣는 구조다. 에이전트가 검색과 비교, 결제까지 대신하면 사람이 광고를 볼 기회가 줄어들고, 구매 의도가 분명한 검색광고와 가격비교 트래픽이 가장 먼저 영향을 받는다. 아마존이 외부 AI 쇼핑 에이전트와의 소송을 불사하면서도 자사 쇼핑 에이전트 육성에 집중하고, 외부 에이전트와는 자사 조건에 맞는 협력만 열어두는 것도 사람이 아마존 안에 머물러야 리테일 미디어 매출이 유지되기 때문일 것으로 추정한다. 아마존은 3분기 실적발표에서 외부 에이전트 쇼핑이 광고 매출에 아직 의미 있는 영향을 주지 않는다고 밝혔지만, 이 질문에 직접 답해야 할 만큼 시장의 관심은 커지고 있다.

다만 광고 예산이 사라지기보다는 형태를 바꿔 이동할 가능성이 높다. 판매자가 내는 거래 수수료는 본질적으로 성과가 났을 때만 비용을 내는 광고와 같고, 과금 기준이 클릭에서 구매로 옮겨가는 것으로 볼 수 있다. **에이전트의 추천 순서에 비용을 지불하는 형태의 광고가 등장할 가능성도 있다. 구독 모델도 갈라질 것이다.** 기본 기능을 무료로 제공하고 판매자에게서 수익을 얻는 모델이 확산되면 일반 소비자 대상 구독료의 설득력은 약해지는 반면, 코딩·업무 자동화·기업용 에이전트처럼 **사용자가 직접 생산성을 얻는 영역에서는 구독과 사용량 과금이 유지될 것이다. 반도체 관점에서는 어느 쪽이 이기든 결과가 같다. 사람이 광고 몇 개를 보고 끝내던 구매 과정을 에이전트가 수십 번의 추론과 도구 호출로 처리하는 순간, 수익이 광고주에게서 나오든 판매자에게서 나오든 거래 한 건에 필요한 컴퓨트와 메모리는 늘어난다.**

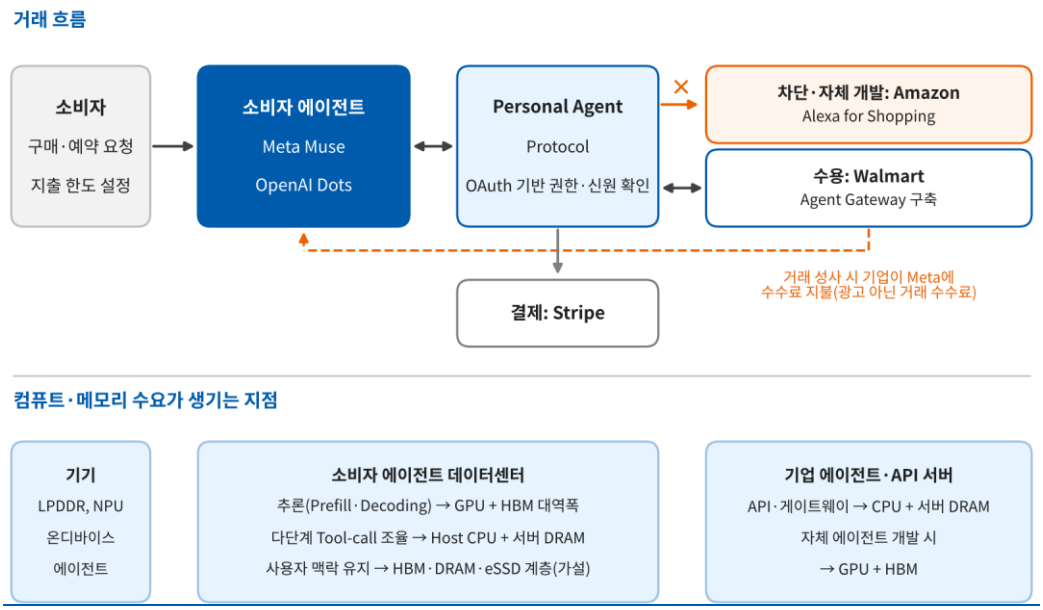
표6. 아마존의 AI 에이전트 대응: 선택적 개방

| 시점 | 내용 | 의미 |
|---|---|---|
| 2026.2 | OpenAI에 500억 달러 투자하는 전략적 제휴 발표. 쇼핑 연동 포함 여부는 미공개 | 직후 ChatGPT Instant Checkout 중단, 쇼핑 협력 가능성 거론 |
| 2026.3 | Rufus·Buy for Me로 아마존에 없는 상품을 외부 쇼핑몰에서 대신 구매하는 기능 확대 | 외부 에이전트를 들이기보다 자사 에이전트가 밖으로 나가는 방식 |
| 2026.8경 | Perplexity Comet 소송 항소심에서 1심 가처분 취소 | 외부 에이전트를 기술적으로 막는 전략의 법적 근거 약화 |
| 2026.9 | 판매자용 Claude 플러그인 출시, Claude에서 매장 관리 가능 | 소비자 쇼핑이 아닌 판매자 도구 쪽 개방 |
| 2026.10 | 3Q26 실적발표: 외부 범용 에이전트와 협력 예상하나 자사 Alexa for Shopping이 주력, 광고 매출 영향은 아직 미미 | 자사 조건에 맞춘 협력과 자사 에이전트 우선 |

자료: Amazon, Forrester, TechCrunch, Retail Brew, 언론 보도 종합, 유안타증권 리서치센터

반도체 관점에서 더 중요한 발언은 따로 있다. 저커버그는 Muse가 컴퓨트를 매우 많이 쓰는 서비스이며, 출시 지역을 미국과 캐나다로 제한한 이유 중 하나도 컴퓨트 부담이라고 설명했다. **서비스 확대를 막고 있는 것이 수요가 아니라 컴퓨트 공급이라는 뜻이다. 메타는 1년 안에 효율을 10배 이상 높이고, 장기적으로 10억명 이상에게 서비스하겠다는 목표를 제시했다. 사용자가 1,000만명에서 10억명으로 100배 늘어나는 동안 효율 개선이 10배에 그친다면, 1인당 사용량이 지금 수준에 머문다고 가정해도 필요한 컴퓨트는 약 10배 늘어난다. 효율 개선은 수요를 줄이기보다 서비스 대상을 넓히는 데 쓰일 가능성이 높다.**

그림 2. 에이전트 커머스의 거래 흐름과 컴퓨트 · 메모리 수요 발생 지점




자료: Meta, Sierra, Forbes, 유안타증권 리서치센터 주: 사용자 맥락 유지에 따른 메모리 계층 수요는 당사 가설

**커머스 업체의 대응은 둘로 갈리고 있지만, 어느 쪽이든 컴퓨트 수요로 이어진다는 점이 핵심이다.** Walmart는 Muse를 받아들이는 쪽을 택했지만, 문을 그냥 열어준 것이 아니라 에이전트를 맞을 관문인 Agent Gateway를 직접 구축했다. 에이전트별 지출 한도와 건별 승인, 결제 정보 접근 차단 등을 관리하는 이 관문 자체가 기업 측의 새로운 서버·컴퓨트 투자다. 반면 Amazon은 봇 차단으로 Muse의 접근을 막고 자체 쇼핑 에이전트인 Alexa for Shopping을 키우고 있다. 메타 CFO Susan Li도 고객 관계를 내주는 데 신중한 업체들이 있어 도입은 고르지 않을 것이라고 인정했다.

그러나 받아들이는 업체는 에이전트를 맞을 게이트웨이와 API 서버를, 거부하는 업체는 자체 에이전트를 만들어야 한다. 거부하는 쪽도 계속 문을 닫아두기는 어렵다. 아마존은 3분기 실적발표에서 외부 범용 에이전트와도 협력할 것으로 예상한다고 밝혔고, 항소심에서 외부 에이전트 차단의 법적 근거도 약해졌다. 결국 거부하는 업체도 자체 에이전트에 더해 외부 에이전트를 맞을 관문까지 갖추게 될 가능성이 높다. 활동 지표를 높여야 하는 후발 업체일수록 외부 에이전트를 적극적으로 받아들일 가능성이 높아, 도입 수요 자체도 충분할 것으로 판단한다. **다만 수요의 크기는 같지 않다. 게이트웨이와 API 서버는 주로 CPU와 서버 DRAM을 쓰는 상대적으로 가벼운 부하인 반면, 자체 에이전트는 GPU와 HBM이 필요한 추론 부하이며, 가장 큰 수요는 여전히 소비자 에이전트를 운영하는 플랫폼의 데이터센터에서 발생한다.**

주체별로 수요의 크기를 나눠 보면 다음과 같다. **첫째, 가장 큰 수요는 소비자 에이전트 플랫폼에서 나온다.** 사용자의 요청 한 건마다 요청 이해, 여러 차례의 추론, 응답 생성이 모두 플랫폼의 데이터센터에서 이뤄지고, 출력 토큰이 많은 Decoding 비중이 높아 HBM 대역폭 부담이 크다. 여기에 다단계 작업 조율을 위한 CPU와 서버 DRAM, 사용자 맥락을 담을 메모리 계층까지 더해진다. 사용자 수가 늘어나는 만큼 수요가 비례해 늘어나는 구조이며, 메타가 컴퓨트 부담 때문에 출시 지역을 제한했다는 사실이 그 크기를 방증한다.

**둘째, 기업의 자체 에이전트는 같은 추론 부하를 기업이 직접 지는 구조다.** 자체 개발 여력이 있는 소수 대형 업체에 한정되지만, 아마존처럼 사용자 기반이 큰 업체가 포함돼 있어 규모가 작지 않다. 아마존은 자체 쇼핑 에이전트 Alexa for Shopping의 누적 이용자가 4억명을 넘었다고 밝혔다. 다만 질문 응답처럼 가벼운 사용까지 포함된 수치여서, 다단계 작업을 대신 수행하는 플랫폼 에이전트보다 사용자당 부하는 가벼울 가능성이 높다.

**셋째, 게이트웨이와 API 서버는 단위 부하가 가장 가볍다.** AI 모델을 직접 돌리지 않고 권한 확인과 상품·재고·결제 조회 요청에 응답하는 부하이기 때문이다. 사람의 웹 트래픽 일부가 에이전트 트래픽으로 바뀌는 대체 성격도 있다. 다만 에이전트는 가격을 비교하기 위해 여러 판매처를 동시에 조회하기 때문에, 구매 한 건당 호출 수는 사람이 직접 쇼핑할 때보다 늘어날 가능성이 높다.

**마지막으로 기기는 중장기 변수다.** 플랫폼이 컴퓨트 비용 일부를 기기로 넘기려 할수록 스마트폰과 PC의 메모리 탑재량 기준이 높아지겠지만, 2027년 실적에 미치는 영향은 데이터센터보다 작을 것으로 본다.

표7. 에이전트 커머스 주체별 컴퓨트·메모리 수요 비교

| 주체 | 수요 발생 지점 | 주요 하드웨어 | 수요 성격 | 상대적 크기 |
|---|---|---|---|---|
| 소비자 에이전트 플랫폼 (Meta Muse, OpenAI Dots) | 요청 이해, 다단계 추론·응답 생성, 작업 조율, 사용자 맥락 유지 | GPU+HBM, Host CPU+서버 DRAM, eSSD | 전량 신규, 사용자 수에 비례 | 가장 큼 |
| 기업 자체 에이전트 (Amazon Alexa for Shopping) | 자체 모델 추론, 상품 추천 | GPU+HBM | 대부분 신규, 소수 대형 업체이나 사용자 기반 큼 | 큼 |
| 기업 게이트웨이·API 서버(Walmart Agent Gateway) | 권한 확인, 상품·재고·결제 조회 응답 | CPU+서버 DRAM | 일부 대체 + 건당 호출 수 증가 | 상대적으로 작음 |
| 기기(온디바이스 에이전트) | 일부 연산의 기기 이전 | LPDDR, NPU | 탑재량 기준 상향 | 중장기 변수 |

자료: Meta, Amazon, Forbes, 유안타증권 리서치센터 주: 상대적 크기는 당사 판단

양측을 잇는 표준도 만들어지고 있다. Sierra가 주도하고 메타, Shopify, Stripe, Walmart 등이 참여한 Personal Agent Protocol은 10월 공개된 개방형 표준으로, OAuth 기반으로 사용자가 에이전트에 권한을 주고 회수할 수 있으며 기업은 에이전트가 쓸 수 있는 기능을 직접 정할 수 있다. 기업이 어떤 에이전트가 누구를 대신해 어떤 권한으로 움직이는지 확인할 수 있게 되면, 에이전트 트래픽은 차단 대상에서 관리 가능한 거래 채널로 바뀐다. 아마존의 Buy for Me처럼 기업 에이전트가 다른 쇼핑몰에 들어가 대신 구매하는 사례도 나오고 있다. 소비자 에이전트와 기업 에이전트, 그리고 기업 에이전트끼리 서로 통신하는 구조가 자리 잡을수록 각 참여자의 데이터센터 부하는 함께 늘어난다.

에이전트 커머스는 HBM만이 아니라 CPU와 서버 DRAM 수요를 동시에 키우는 워크로드다. 에이전트는 요청 하나를 처리하기 위해 상품 검색, 가격 비교, 재고·예약 조회, 결제를 여러 단계에 걸쳐 반복한다. 모델 추론은 GPU와 HBM이 맡지만, 외부 앱 호출과 단계별 조율은 Host CPU가 담당하고 CPU마다 서버 DRAM이 붙는다. 앞서 살펴본 에이전트 AI의 Host CPU 수요 증가와 같은 맥락이다. 여기에 비교·추론·응답 생성 단계는 출력 토큰이 많은 Decoding 구간이라 HBM 대역폭이 병목이 된다. 개인 에이전트가 사용자의 취향과 이전 대화, 구매 이력을 계속 기억하고 활용할수록 이를 담을 HBM·DRAM·eSSD 메모리 계층 수요도 늘어날 것으로 예상한다.

표8. 에이전트 커머스 처리 단계별 하드웨어 부하

| 단계 | 하는 일 | 주요 연산 | 병목 하드웨어 | 관련 메모리 |
|---|---|---|---|---|
| ① 요청 이해·계획 | 요청 해석, 작업 단계 설계 | Prefill(GEMM) | GPU·가속기 | HBM 용량 |
| ② 검색·외부 앱 호출 | 상품 검색, 가격 비교, 재고·예약 조회 | Tool-call, API 통신 | Host CPU, 네트워크 | 서버 DDR5, SOCAMM |
| ③ 비교·추론·응답 | 후보 비교, 추론, 결과 설명 | Decoding(GEMV) | 메모리 대역폭 | HBM 대역폭 |
| ④ 맥락 유지 | 취향·이전 대화·구매 이력 활용 | KV 캐시·맥락 저장과 로드 | 메모리 계층 | HBM·DRAM·eSSD(가설) |
| ⑤ 결제·확인 | 권한 확인, 결제, 사후 알림 | 인증·트랜잭션 처리 | CPU, 보안 | 서버 DRAM |

자료: Meta, Sierra, SK 하이닉스, 유안타증권 리서치센터 주: 단계 구분과 하드웨어 매핑은 당사 추정

## 3. 메모리의 파운드리화: 존재감을 드러내는 새로운 메모리

### HBM: 베이스다이의 로직화와 Custom HBM

**HBM4는 HBM의 변곡점이다.** HBM2부터 HBM3E까지 데이터버스 폭은 1,024bit로 고정된 채 핀당 속도만 2.4Gbps에서 9.6Gbps로 높아졌고, 베이스다이는 DRAM 공정으로 만들어졌다. HBM4부터는 데이터버스 폭이 2,048bit로 두 배 넓어진다. 데이터가 오가는 통로가 두 배로 늘어나는 만큼 쇼어라인(Shoreline) 부담도 커진다. 쇼어라인은 칩 가장자리에서 가속기와 신호를 주고받는 입출력 회로가 배치되는 테두리 공간으로, 칩을 섬에 비유하면 바깥과 연결되는 해안선에 해당한다. 칩 크기를 그만큼 키울 수 없기 때문에, 같은 쇼어라인에 두 배의 입출력 회로를 촘촘하게 배치하면서 더 빠른 속도로 구동해야 한다. DRAM 공정으로는 이 고밀도·고속 회로를 효율적으로 만들기 어려워, 베이스다이가 미세 회로에 유리한 로직 공정으로 넘어간다. MediaTek은 Custom HBM이 데이터버스 폭 2,048bit 이상, 핀당 16~32Gbps 이상으로 진화하고, 궁극적으로는 xPU끼리 cHBM을 통해 서로 연결되는 구조를 제시한다. **HBM이 단순한 메모리를 넘어 데이터 경로의 일부가 되는 것이다.**

삼성전자는 HBM의 표준화 구조가 세대를 거치며 바뀌고 있다고 설명한다. HBM2~HBM5까지의 표준 HBM은 코어다이와 베이스다이 모두 JEDEC 표준을 따른다. HBM4부터 가능해진 cHBM은 코어다이는 표준을 유지하되 베이스다이만 고객 요구에 맞춰 설계하는 Semi-custom 구조다. 삼성전자가 개념 단계로 제시한 zHBM은 코어다이와 베이스·인터레이어 다이를 모두 맞춤 설계하고 GPU 위에 HBM을 3D로 직접 적층하는 Full-custom 구조로, HBM4E 대비 GPU당 성능 최대 8배, 전력당 대역폭 3배, 열저항 0.1~0.25배 수준을 목표로 한다. 현재 zHBM에 대한 JEDEC 표준은 없다.

표9. HBM 세대별 사양과 표준화 구조의 변화

| 구분 | 데이터버스 폭 | 핀당 속도 | 베이스다이 공정 | 표준화 구조 |
|---|---|---|---|---|
| HBM2 | 1,024bit | 2.4Gbps | DRAM | 표준(코어·베이스 모두 JEDEC) |
| HBM2E | 1,024bit | 3.6Gbps | DRAM | 표준 |
| HBM3 | 1,024bit | 6.4Gbps | DRAM | 표준 |
| HBM3E | 1,024bit | 9.6Gbps | DRAM | 표준 |
| HBM4 | 2,048bit | 10Gbps 내외 | Logic | 표준 / cHBM(베이스만 맞춤) 가능 |
| HBM4E | 2,048bit | 12Gbps 이상 | Logic | 표준 / cHBM 가능 |
| Custom HBM(비전) | 2,048bit 이상 | 16~32Gbps 이상 | Logic | Semi-custom, xPU 간 연결 경로로 활용 |
| zHBM(개념) | - | - | Custom(베이스·인터레이어) | Full-custom, GPU 위 3D 적층, JEDEC 표준 없음 |

자료: MediaTek, 삼성전자(JEDEC Industry Forum, 2026.10), 유안타증권 리서치센터 주: HBM4·4E 속도는 발표 자료상 잠정치

NVIDIA가 공개한 NVHBM도 같은 방향을 가리킨다. 기존 xPU 다이에 있던 Memory Controller를 HBM 베이스다이로 옮겨 xPU의 연산 면적을 확보하는 구조로, NVIDIA는 HBM4E 대비 최대 30% 높은 대역폭, 15% 낮은 HBM 전력, 최대 25%의 xPU 연산 면적 확보 효과를 제시했다. Rubin Ultra의 HBM 사양 조정도 이 맥락에서 읽을 필요가 있다. 삼성전자와 SK하이닉스는 Rubin Ultra용 8단 HBM4를 2027년 2분기부터 양산·출하하는 것을 목표로 하고, NVIDIA는 8단과 12단 HBM4·HBM4E를 혼용하는 방안을 검토 중인 것으로 파악된다. 이는 수요 둔화라기보다 제한된 DRAM 웨이퍼로 더 많은 가속기를 생산하고 공급 안정성을 높이기 위한 최적화로 판단한다.

**투자 관점에서 HBM의 맞춤화는 양면적이다.** 베이스다이가 로직 공정으로 넘어가면서 부가가치의 일부가 파운드리와 ASIC 디자인하우스로 이동하는 것은 사실이다. 그럼에도 불구하고 당사는 긍정적인 측면에 재차 주목한다. 메모리가 범용 Commodity에서 고객별로 설계되는 수주형 제품으로 바뀌고, 앞서 언급한 대로 HBM이 단순한 메모리를 넘어 데이터 경로의 일부가 된다는 점이다. 고객별 설계는 전환 비용을 높이고 장기 계약을 강화해 이익의 변동성을 낮춘다.

**한편 HBM의 맞춤화는 중국 업체에게는 넘기 어려운 벽이 된다.** 당사가 올해 3월 'Semicon China 2026' 보고서에서 밝혔듯이, 베이스다이가 파운드리 선단 공정으로 넘어가면서 HBM4부터는 설계 단계에서부터 파운드리·팹리스·메모리 업체가 IP를 공유하고 공정을 함께 최적화해야 한다. **SK하이닉스·TSMC·NVIDIA로 이어지는 연합 모델이든 삼성전자의 파운드리·메모리 턴키 모델이든, 승패를 가르는 것은 선단 파운드리 공정에 대한 접근권과 로직 IP 설계 역량이다.** 미국의 규제로 7nm 이하 공정과 핵심 로직 IP 도입이 막힌 중국 업체에게 HBM4는 자본만으로는 넘기 어려운 '공정과 생태계의 복합 장벽'이다. 7월 상장한 CXMT의 투자설명서에 HBM 전용 프로젝트가 빠져 있고, 8단 HBM3의 수율이 25% 수준에 머물러 있다는 추정이 나오는 것도 같은 맥락이다. **HBM이 Custom으로 갈수록 한국과 중국 업체의 격차는 좁혀지기보다 벌어질 가능성이 높다는 판단이다.**

표준화 기구의 역할도 바뀌고 있다. MediaTek과 삼성전자는 모든 아키텍처를 하나의 표준으로 정의하기는 어려워지고 있다는 데 공감하고 있다. 대신 Processing Element, RAS, Memory Controller, Memory BIST, Interface·Protocol 등 핵심 기능만 모듈 단위로 표준화하고 나머지는 선택 기능으로 남겨 호스트 업체의 설계 자유도를 보장하는 방향이 현실적이라는 의견이 지배적인 것으로 파악된다. **JEDEC 표준이 완성된 제품 규격에서 공통 인터페이스와 모듈을 정의하는 프레임워크로 바뀌는 과정인 것이다.**

표10. 모듈 단위 표준화 논의 대상 기능 설명

| 기능 | 뜻 | 표준화가 필요한 이유 |
|---|---|---|
| Processing Element(연산 유닛) | 메모리 안이나 베이스다이에 넣는 작은 연산 회로. PIM 이나 Compute Near Memory 에서 곱셈·덧셈(MAC) 같은 단순 연산을 메모리 가까이에서 처리 | 호스트마다 원하는 연산 종류와 데이터 형식이 달라, 공통 명령 체계가 없으면 메모리 업체가 고객마다 다른 칩을 만들어야 함 |
| RAS(Reliability, Availability, Serviceability) | 세 단어의 약자로, 메모리 오류에 대응하는 기능 묶음. R(Reliability, 신뢰성): 오류를 찾아 고치는 능력(ECC, 리페어) / A(Availability, 가용성): 고장이 나도 시스템이 멈추지 않고 계속 돌아가는 능력 / S(Serviceability, 유지보수성): 문제를 진단하고 부품을 교체하기 쉬운 정도 | 오류 보고·처리 방식이 같아야 데이터센터가 여러 업체의 메모리를 같은 방식으로 관리할 수 있음 |
| Memory Controller(메모리 컨트롤러) | 프로세서가 메모리에 데이터를 읽고 쓸 때 명령 순서와 타이밍을 관리하는 회로. 기존에는 GPU 등 프로세서 안에 있었으나 NVHBM 처럼 HBM 베이스다이로 옮기는 시도가 나타남 | 컨트롤러가 메모리 쪽으로 넘어오면 프로세서와 메모리 사이의 역할 분담 기준이 새로 필요함 |
| Memory BIST(Built-In Self-Test) | 메모리가 스스로 결함을 검사하는 내장 테스트 회로. GPU 위에 적층된 HBM 처럼 외부 장비로 직접 검사하기 어려운 구조에서 필수 | 검사 방식이 통일돼야 양산 검증과 불량 분석이 가능함(2장 4. 패키징·테스트 참조) |
| Interface·Protocol(인터페이스·통신 규약) | 프로세서와 메모리가 신호를 주고받는 물리적 연결 방식과, 명령·데이터를 주고받는 형식에 대한 약속 | 규약이 같아야 서로 다른 업체의 프로세서와 메모리를 연결할 수 있음 |

자료: 삼성전자, MediaTek(JEDEC Industry Forum, 2026.10), 유안타증권 리서치센터

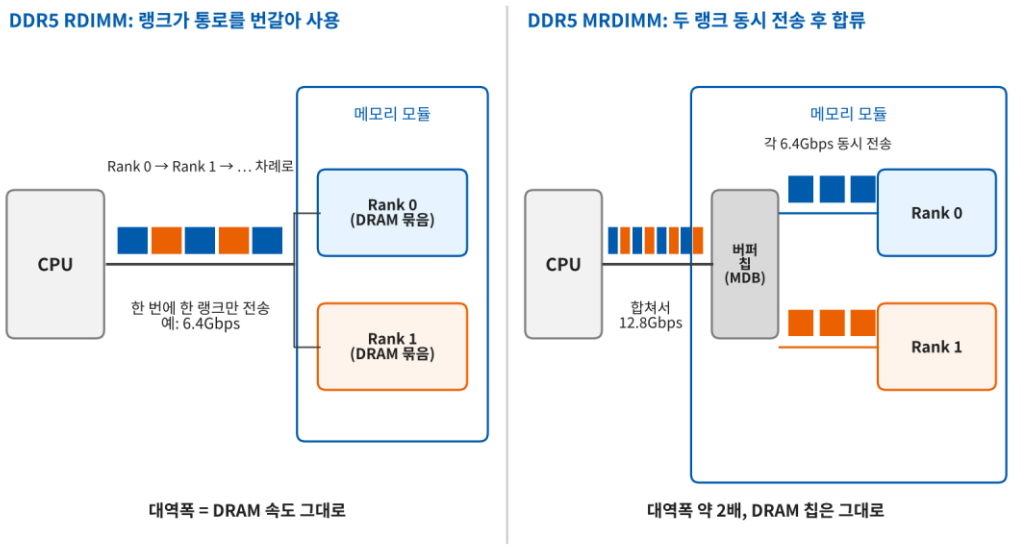
## 서버 메모리의 분화: RDIMM, MRDIMM, SOCAMM2

**서버 메인 메모리도 하나의 표준에서 여러 갈래로 나뉘고 있으며, 이 분화는 시장의 크기 자체를 넓히는 변화다.** Synopsys는 이를 'Server Memory Fragmentation(서버 메모리 분화)'으로 정의했다. 지금까지 서버 메모리의 표준은 DDR5 RDIMM이다. RDIMM은 모듈 하나에 최대 512GB의 용량을 담을 수 있고, DRAM 칩 하나가 통째로 고장 나도 데이터를 복구하는 Chipkill/SDDC 등 강력한 RAS(신뢰성·가용성·유지보수성) 기능을 갖춰 앞으로도 서버의 기본 메모리로 남을 전망이다. 문제는 대역폭이다. 속도가 높아질수록 신호 품질을 지키기 위해 메모리 채널 하나에 꽂을 수 있는 모듈 수가 줄어들고 있어, 모듈을 더 꽂는 방식으로는 대역폭을 늘리기 어려워졌다.

이 한계를 메우기 위해 등장한 MRDIMM과 SOCAMM2는 RDIMM을 밀어내는 제품이라기보다 RDIMM으로는 채우지 못하던 수요를 새로 담는 제품이다. 일반 서버는 RDIMM을 그대로 쓰고, 대역폭이 필요한 서버에는 MRDIMM이, 전력이 빠듯한 AI 서버에는 SOCAMM2가 더해진다. 스마트폰과 PC에 머물던 LPDDR이 서버라는 새로운 시장에 들어가고, 모듈마다 버퍼 칩과 PMIC 같은 부품이 새로 붙는다. 하나의 제품을 두고 가격과 물량을 다투던 서버 메모리 시장이, 용도별로 여러 제품이 공존하면서 판매 기회와 부가가치가 함께 늘어나는 시장으로 바뀌고 있는 것이다.

MRDIMM(Multiplexed Rank DIMM)은 이 문제를 모듈 안에서 푼다. 모듈 위의 DRAM 칩들은 랭크(Rank)라는 묶음으로 나뉘는데, 기존 RDIMM에서는 여러 랭크가 하나의 통로를 번갈아 쓰기 때문에 한 번에 한 랭크만 데이터를 보낸다. MRDIMM은 모듈에 버퍼 칩(MRCD·MDB)을 추가해 두 랭크가 동시에 데이터를 내보내게 하고, 버퍼 칩이 두 흐름을 하나로 합쳐 두 배 빠른 속도로 프로세서에 전달한다. 두 차선의 차량을 합류 지점에서 두 배 빠른 고속도로 한 차선으로 흘려보내는 것과 비슷하다. DRAM 칩 자체의 속도를 높이지 않고도 프로세서가 받는 대역폭을 약 2배로 늘릴 수 있고, 속도는 Gen2 12.8Gbps, Gen3 16.0Gbps까지 높아진다. RDIMM의 RAS 기능도 그대로 유지된다.

그림 3. DDR5 RDIMM vs MRDIMM 데이터 전송 구조




자료: Synopsys(JEDEC Industry Forum, 2026.10), 유안타증권 리서치센터 재구성 주: DRAM 6.4Gbps 기준 예시

**반대편에서는 LPDDR이 서버에 진입하고 있다.** LPDDR5X 기반 SOCAMM2는 DDR5 RDIMM 대비 면적과 전력이 약 3분의 1 수준이면서 같은 용량과 더 높은 대역폭을 제공한다. Synopsys는 AI 추론 시나리오에서 LPDDR5X가 표준 DDR5 대비 DRAM 전력 효율을 최대 약 70% 개선할 수 있다고 설명했다. **다만 Chipkill 수준의 RAS를 지원하지 못하고 호스트 SoC에 더 많은 PHY·Controller 면적이 필요하다는 한계가 있다.**

결국 서버 메모리는 용량과 RAS가 중요한 일반 서버(RDIMM), 대역폭이 필요한 고성능 서버(MRDIMM), 전력과 대역폭을 위해 RAS를 일부 타협할 수 있는 AI 추론 서버(SOCAMM2)로 분화될 것이다. **제품이 늘어날수록 수혜 범위도 메모리 업체를 넘어 버퍼·PMIC 등 인터페이스 칩과 모듈 테스트 업체까지 넓어진다. 지금 메모리 업계는 창조적 시행착오의 한가운데 있다.** Custom HBM과 zHBM, PIM, SOCAMM2, MRDIMM 등 수많은 구조가 동시에 제안되고 있고, LPDDR5X를 MRDIMM 구조에 적용하려던 LP-MRDIMM처럼 생태계 지원 부족으로 표준화 단계에서 사라지는 시도도 나온다. 어떤 구조가 살아남을지는 아직 알 수 없지만, 시행착오가 많을수록 메모리와 로직의 경계에서 새로운 가치가 만들어지고 있다는 점은 분명하다.

표11. 서버 메인 메모리 규격 비교

| 구분 | DDR5 RDIMM | DDR5 MRDIMM | LPDDR5X SOCAMM2 |
|---|---|---|---|
| 최대 속도 | 9.2Gbps | 12.8Gbps(Gen2),<br>16.0Gbps(Gen3) | 10.7Gbps |
| 구조 | RCD + DDR5 | MRCD·MDB 로 2개 랭크 다중화 | x32 LPDDR5X 4개, 독립 인터페이스 |
| 장점 | 대용량(최대 512GB), Chipkill RAS | 대역폭 약 2배, 서버급 RAS 유지 | RDIMM 대비 면적·전력 약 1/3 |
| 한계 | 고속화 시 채널당 DIMM 수 감소 | 버퍼 칩 비용 추가 | Chipkill 미지원, SoC PHY 면적 증가 |
| 주요 용도 | 범용·엔터프라이즈 서버 | HPC, 고대역폭 서버 | AI 추론 서버, 엣지 데이터센터 |

자료: Synopsys(JEDEC Industry Forum, 2026.10), 유안타증권 리서치센터

## PIM: Decoding 병목에 대한 메모리 업체의 해법

PIM(Processing-In-Memory)은 앞서 설명한 Decoding 병목에 대한 메모리 업체의 해법이다. DRAM 뱅크 옆에 연산 유닛을 배치해 가중치가 저장된 자리에서 GEMV를 직접 수행하고, 호스트에는 결과만 보내 데이터 이동을 줄이는 방식이기 때문이다. **앞서 본 HBM이 고객별 설계로, 서버 메모리가 용도별 분화로 맞춤화되고 있다면, PIM은 메모리가 맡는 기능 자체가 늘어나는 방식의 변화다.** 메모리가 데이터를 보관하는 부품에서 연산의 일부를 맡는 부품으로 바뀌는 것이다.

**PIM이 업체별 시도를 넘어 표준 규격으로 들어간다는 점이 중요하다.** JEDEC 표준 기준으로 LPDDR5X는 PIM을 지원하지 않았지만, LPDDR6부터는 표준 규격 안에 PIM이 들어갈 예정이며 2026년 4분기 LPDDR6-PIM 1.0 규격 발표를 목표로 표준화가 진행 중이다. SK하이닉스가 주도하는 LPDDR6-PIM은 외부 인터페이스와 용량을 그대로 유지하면서 내부 대역폭을 약 4배(SEFF 모드 최대 8배)로 높이고, 데이터 이동 에너지를 표준 LPDDR6 대비 20~25% 수준으로 낮추는 것을 목표로 한다. 표준 이전의 자체 시도로는 삼성전자의 LPDDR5X-PIM이 있다. Llama-3.1-8B 기준 처리 속도는 LPDDR5X 대비 3.8배 높아졌고, 전력은 1.4배 늘었지만 전력당 처리량은 2.8배 개선됐다. **표준이 없으면 메모리 업체가 고객마다 다른 칩을 만들어야 하지만, 규격이 정해지면 AP·NPU 업체가 공통 방식으로 PIM을 쓸 수 있어 채택 속도가 빨라진다.** 다만 PIM의 채택 여부는 결국 호스트 칩 업체가 결정한다. PIM은 칩이 비효율적인 Decoding 구간만 덜어주는 분업 구조인 데다 표준 규격으로 들어가는 만큼, 칩 업체도 받아들일 가능성이 높다.

**PIM은 엣지에서 먼저 시작될 전망이다.** 스마트폰과 PC는 데이터센터보다 전력과 메모리 대역폭이 훨씬 빠듯하면서도, 수십억 파라미터 규모의 모델을 기기 안에서 상시 구동해야 한다. **데이터를 옮기지 않고 메모리 안에서 처리하는 PIM의 장점이 가장 크게 드러나는 환경이다.** 연산의 일부가 메모리로 넘어오는 만큼 메모리 업체가 가져가는 부가가치도 커진다. 엣지 기기에서의 구체적인 성능 효과는 '5. 엣지'에서 다시 살펴본다.

**PIM을 담는 그릇인 LPDDR6는 그 자체로도 서버 시장을 넓힌다.** 채널 폭이 24bit로 넓어져 채널당 대역폭이 두 배로 늘고, 메타데이터 기반 RAS가 더해진다. 32바이트 데이터마다 4바이트의 여유 공간에 오류 검사 코드를 실어 보내 칩 내부 ECC보다 한 단계 강한 오류 보호가 가능해지는 구조다. 앞서 SOCAMM2의 한계로 지적한 RAS 부족을 일부 메워주는 셈이며, LPDDR의 서버 진입을 넓혀줄 변화다.

표12. LPDDR6·PIM 주요 사양 비교

| 구분 | LPDDR5X | LPDDR6 |
|---|---|---|
| PIM | 표준 미지원(삼성 자체 LPDDR5X-PIM 존재) | 표준 지원 예정(LPDDR6-PIM) |
| 최대 전송속도 | 10.67Gbps | 14.4Gbps |
| 채널 폭 | 16bit | 24bit(12bit 서브채널 2개) |
| 채널당 최대 대역폭 | 21.3GB/s | 43.2GB/s |
| 패키지당 최대 용량 | 64GB | 128GB |
| SOCAMM2당 최대 용량 | 256GB | 512GB |
| 전력 효율 | 기준 | 약 20~21% 개선 |
| RAS | 표준 ECC | 메타데이터 기반 Enhanced RAS |

자료: Synopsys, SK 하이닉스(JEDEC Industry Forum, 2026.10), 유안타증권 리서치센터

## 4. 패키징·테스트: 가격으로 번지는 병목

### 패키징: 레티클 한계를 넘는 가속기와 수요의 확산

**메모리 다음으로 2027년 실적에서 병목이 확인될 영역은 첨단 패키징이다.** AI 가속기는 이미 노광 장비가 한 번에 찍을 수 있는 최대 면적인 레티클의 한계를 넘어섰다. 여러 개의 연산 다이와 HBM을 하나의 패키지에 담아야 하기 때문이다. TSMC는 3분기 실적발표에서 레티클 5배 크기의 패키지를 실제 양산하고 있으며, 9배와 14배 크기도 개발 중이라고 밝혔다. TSMC의 CoWoS 생산능력은 2027년 월 18만장 수준까지 늘어날 전망이지만, 가속기 출하가 60% 이상 늘어나는 2027년 수요를 감안하면 여유롭지 않을 것으로 추정한다. 당사 채널체크에 따르면 **CoWoS 생산능력은 여전히 빠듯하고, 고객사와 15~20% 수준의 패키징 가격 인상이 논의되고 있다. 병목이 물량을 넘어 가격으로 나타나기 시작한 것이다.** HBM이 8·12단에 머물면 같은 용량을 담기 위해 스택 수가 늘어나고, 그만큼 인터포저 면적과 패키징 부담이 커진다는 점도 수요를 더한다.

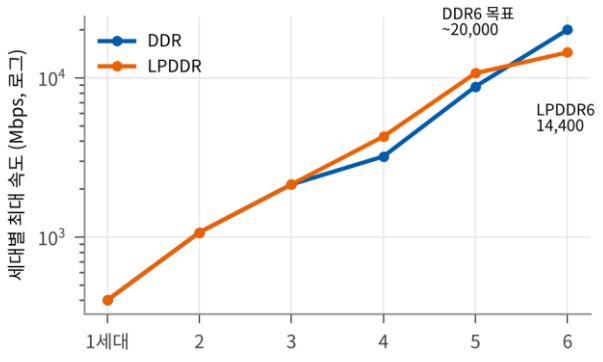
첨단 패키징 수요는 가속기에서 서버 CPU로도 확산되고 있다. **AMD Venice와 NVIDIA Vera 등 차세대 서버 CPU의 대량 양산이 시작되면서 TSMC의 CoWoS 생산능력만으로는 수요 대응이 어려워지고, ASE/SPIL과 Amkor 등 상위 OSAT로의 외주가 확대될 전망이다.** 과거 단순 조립·테스트를 담당하던 OSAT는 패널 레벨 패키징, 실리콘 포토닉스 패키징, System Level Test 역량을 갖춘 솔루션 파트너로 체질을 바꾸고 있으며, 첨단 패키징에 진입한 상위 업체와 레거시 업체 간의 실적·밸류에이션 양극화는 더욱 뚜렷해질 것이다. 2028년 이후의 패키징은 원형 웨이퍼를 넘어 사각 패널로, 칩을 옆에 놓는 방식에서 위로 쌓는 방식으로 바뀐다. 이 변화는 다음 장에서 자세히 살펴본다.

### 테스트: 고속화가 만든 새로운 병목

메모리 고속화는 더 이상 메모리 칩 자체의 개선만으로 해결되지 않는다. 과거에는 세대가 바뀔 때마다 DRAM 칩의 미세공정과 회로 설계를 개선하는 것만으로 속도를 끌어올릴 수 있었다. **Advantest에 따르면 LPDDR은 이런 방식으로 LPDDR1부터 LPDDR5까지 세대마다 대략 두 배씩 속도를 높여왔지만, LPDDR6는 약 10~13Gbps 수준에서 증가폭이 제한되고 있다. 병목이 칩 안이 아니라 칩 밖으로 옮겨갔기 때문이다.**

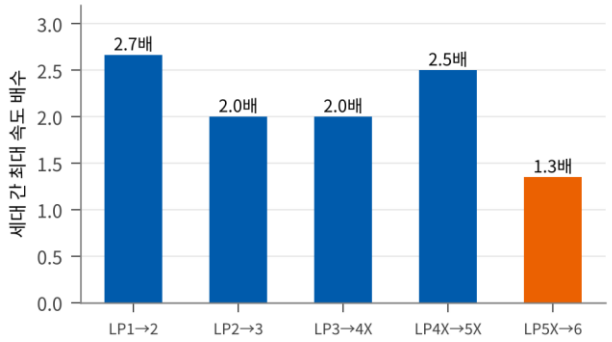
**속도가 높아질수록 칩에서 나온 신호가 패키지와 보드, 인터페이스를 지나는 동안 약해지거나 타이밍이 어긋나기 쉬워지는데,** 이렇게 신호를 온전하게 전달하는 능력을 신호 무결성(SI, Signal Integrity)이라고 한다. DRAM 칩을 아무리 빠르게 만들어도 시스템 전체의 신호 무결성이 받쳐주지 못하면 속도를 높일 수 없다. LPDDR6가 다이당 DQ 수를 x12로 다시 늘린 것도 속도만으로 대역폭을 높이는 데 한계가 왔다는 방증이다. GDDR7은 PAM3 다단 신호를 도입해 전체 대역폭은 늘렸지만 같은 데이터 속도에서 필요한 심볼 속도(보드레이트)는 오히려 낮아졌다. DDR6는 최대 약 20,000Mbps를 목표로 하지만, 같은 세대 안에서 지원해야 하는 속도 범위도 함께 넓어지고 있다.

차트 11. DDR · LPDDR 세대별 최대 전송속도




자료: JEDEC, Advantest, Synopsys, 유안타증권 리서치센터 주: DDR6 는 목표치

차트 12. LPDDR 세대 간 최대 전송속도 증가 배수




자료: JEDEC, Synopsys, 유안타증권 리서치센터 주: LP4X · LP5X 기준

**이는 테스트의 난이도와 강도를 동시에 높인다.** 먼저 Unmatched RX 문제가 있다. 메모리는 데이터를 비트 단위로 일정한 시간 칸에 나눠 보내는데, 비트 하나에 할당된 이 시간 칸을 UI(Unit Interval)라고 한다. 18Gbps 속도라면 1 UI는 약 55피코초(1조분의 55초)에 불과하다. 여러 핀으로 동시에 보낸 신호는 같은 기준 시계(클록)에 맞춰 같은 칸에 도착해야 하는데, 속도가 높아지면서 수신(RX) 측에서 핀마다 도착 시점이 1 UI 이상 어긋나는 현상이 나타나고 있다. 이를 Unmatched RX라고 한다. 어긋남이 한 칸을 넘으면 앞뒤 비트를 혼동할 수 있기 때문에, 자동 테스트 장비(ATE, Automatic Test Equipment)가 테스트 전에 핀별 도착 시점을 측정해 보정하는 학습(Training) 과정을 먼저 거친 뒤 기능 테스트를 이어가야 한다.

PAM3 같은 비이진 신호는 하나의 비트 오류가 여러 심볼 오류로 번지기 때문에 단순 Pass/Fail이 아니라 패킷을 인코딩·디코딩하고 실패 비트를 추적하는 분석이 필요하다. Advantest는 양산 테스트의 병렬성을 확보하려면 향후 테스트 장비 한 대에 2만 5천개 이상의 핀 자원이 필요하다고 제시했다. 여기서 핀 자원은 테스트 장비 내부에서 칩의 각 단자에 신호를 넣고 결과를 읽어내는 전자 회로 채널을 뜻한다. 칩의 입출력 단자 수가 늘어나는 데다, 고속 신호에서는 채널 하나를 여러 단자가 나눠 쓰면 신호 품질이 떨어져 단자마다 전용 채널을 배정해야 하기 때문이다.

표13. 차세대 DRAM 신뢰성 기능별 테스트 표준화 현황

| 영역 | 기능 | 표준화된 기능 테스트 방법 |
|---|---|---|
| Cell Data | On-die ECC | 없음(업체별 자체 테스트 회로·모드 활용) |
| Row Hammer 대응 | RFM(Refresh Management) | - |
| Row Hammer 대응 | PRAC(Per Row Activation Counting) | 전용 테스트 모드 준비 |
| 리페어 | PPR(hPPR, sPPR, mPPR) | 없음 |
| 오류 검사 | ECS(Error Check and Scrub) | 없음 |
| 링크 보호 | Read/Write CRC | 오류 주입(Error Injection) 방식 검토 중 |

자료: Advantest(JEDEC Industry Forum, 2026.10), 유안타증권 리서치센터

신뢰성 기능은 계속 늘어나는데 이를 양산 현장에서 검증할 표준 방법은 뒤따라오지 못하고 있다. On-die ECC, PPR, ECS 등 상당수 기능에 표준화된 기능 테스트 방법이 없어 벤더별 DFT에 의존하는 상황이다. DFT(Design for Test)는 칩을 설계할 때부터 나중에 검사하기 쉽도록 넣어두는 테스트 전용 회로와 모드를 말한다. 표준 테스트 방법이 없다 보니 메모리 업체마다 자체 방식의 테스트 모드를 만들어 쓰고 있고, 테스트 장비와 검사 절차도 업체별로 맞춰야 한다. 테스트가 그만큼 복잡해지고, 장비 업체에는 고객별 맞춤 대응 역량이 요구된다.

Keysight 역시 고속 메모리 검증이 PHY 단독 테스트에서 채널·트레이닝·프로토콜을 포함한 Full-system 측정으로 확대돼야 한다고 강조했다. SEMI는 2026년 반도체 테스트 장비 매출이 전년 대비 31.0% 늘어난 153억달러, 2028년에는 208억달러에 이를 것으로 전망했는데, 2026년 조립·패키징 장비 성장률(9.6%)을 크게 웃돈다. 칩 하나당 테스트 시간과 장비 투입이 늘어나는 구조적 변화로 판단하며, 그 수혜는 메모리 테스트에 더 집중될 것으로 본다. 앞서 살펴본 고속화에 따른 신호 무결성 문제와 신뢰성 기능 증가, 표준 테스트 방법의 부재가 모두 메모리에서 가장 먼저 나타나고 있기 때문이다. 메모리 테스터, 번인, 소켓, 핸들러 등 메모리 테스트 밸류체인의 수혜를 예상한다.

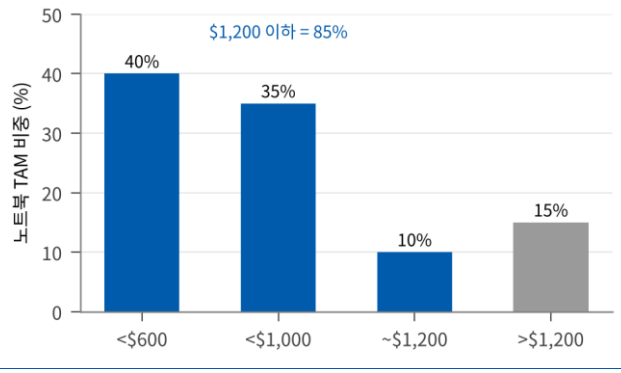
**앞서 살펴본 메모리 업계의 시행착오는 테스트 수요의 관점에서 보면 기회다.** 어느 구조가 살아남을지에 베팅하지 않더라도, 새로운 규격과 구조가 나올 때마다 신호 검증과 신뢰성 테스트, 양산 테스트가 새로 필요하기 때문이다. **Custom HBM은 고객별로, PIM은 메모리 셀에 더해 내부 연산 유닛과 레지스터까지 검사하는 로직 테스트를 함께, SOCAMM2와 MRDIMM은 모듈 단위로 각각 다른 검증 과정을 요구한다.** 구조가 다양해질수록 칩 하나를 시장에 내놓기까지 거쳐야 하는 검증의 종류와 횟수도 늘어난다. 테스트를 병목 순위와 별개로 상시 수혜 영역으로 보는 이유다.

## 5. 엣지: 메모리 원가가 시험하는 소비자 기기

2027년 PC 시장은 2026년 큰 폭의 역성장으로 낮아진 기저 위에서 On-premise AI PC 중심의 Mix 개선과 지연된 Commercial PC 교체 수요가 맞물리며 회복할 전망이다. **기업들은 클라우드 비용, 지연시간, 데이터 보안 문제를 해결하기 위해 일부 AI 연산을 개별 PC로 옮기고 있으며, 이는 PC 메모리의 용량과 대역폭 기준을 동시에 높이고 있다.** 기존 PC가 8~16GB, 50~90GB/s 수준이었다면 Copilot+ 기준 AI PC는 32GB 권장 사양과 100~128GB/s를, 로컬 모델을 구동하는 고사양 AI PC는 64~128GB 이상과 256GB/s 이상의 대역폭을 요구한다.

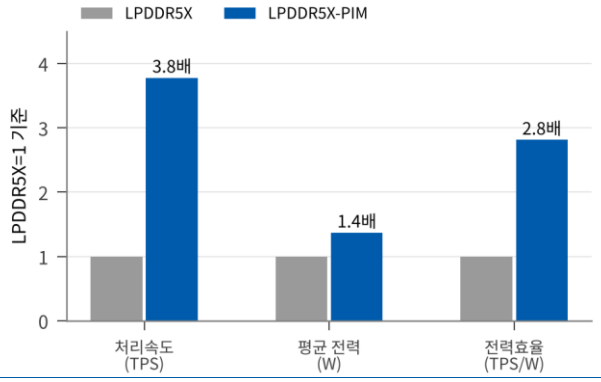
관건은 이 성능을 메인스트림 가격대로 끌어내리는 것이다. Intel은 JEDEC 포럼에서 현재 Halo급 성능은 노트북 시장 상위 5~15%에만 존재하며, 전체 노트북 TAM의 85%는 1,200달러 이하 가격대에 있다고 설명했다. **Intel은 향후 5년간 메인스트림 노트북의 메모리 대역폭을 50~100% 높여 600달러 미만 150GB/s, 1,000달러 이하 250GB/s, 1,200달러 수준 350GB/s까지 끌어올리는 것을 목표로 제시했다.** 다만 현재 노트북의 지속 가능한 전력 예산은 필요 수준보다 약 50% 부족해, 전력 효율과 열 관리가 대역폭 확대의 핵심 제약이 된다. LPCAMM2가 DDR5 SODIMM 대비 약 60%의 공간을 절약해 배터리와 냉각에 공간을 돌려줄 수 있다는 점에서 대안으로 주목받는 이유다.

차트 13. 노트북 가격대별 TAM 비중




자료: Intel(JEDEC Industry Forum, 2026.10), Statista, 유안타증권 리서치센터

차트 14. LPDDR5X-PIM 성능 평가(Llama-3.1-8B)




자료: 삼성전자(JEDEC Industry Forum, 2026.10), 유안타증권 리서치센터 주: 삼성 엣지 AI 가속기

스마트폰은 메모리 원가 상승에도 불구하고 프리미엄 제품을 중심으로 메모리 사양이 오히려 상향되고 있다. On-device AI가 상시 구동될수록 더 큰 모델과 컨텍스트를 담을 용량과 실시간 처리를 위한 대역폭이 함께 필요하기 때문이다. **프리미엄 스마트폰의 메모리 탑재량은 8~12GB에서 12~16GB 이상으로 이동하고, LPDDR5X에서 LPDDR6로의 전환도 본격화될 전망이다.** SK하이닉스는 AI 스마트폰과 AI PC의 비중이 2031년까지 꾸준히 상승해 AI PC가 2030년 70% 이상에 이를 것으로 보고 있으며, 이것이 LPDDR6-PIM 개발의 직접적인 배경일 것으로 추정한다.

PIM은 엣지 AI의 성능 격차를 줄이는 수단이 될 수 있다. 삼성전자는 클라이언트 PC에서 GDDR7을 쓰는 외장 GPU가 40 TPS, LPDDR5X를 쓰는 외장 NPU가 25 TPS 수준인 반면, LPDDR5X-PIM을 적용한 외장 NPU는 103 TPS를 구현해 단순 질의를 넘어 코딩과 에이전트 AI까지 처리할 수 있다고 제시했다. 다만 메모리 가격 급등은 엣지 수요의 부담 요인이기도 하다. **채널체크에 따르면 스마트폰 재고는 8~14주 수준이고 올해 출하는 전년 대비 6~10% 감소가 예상되며,** 중저가 중국 브랜드는 원가를 줄이기 위해 중고 메모리를 대거 사용하고 있는 것으로 파악된다. PC는 상반기에 과도하게 풀린 물량의 소화가 2027년 상반기까지 이어질 전망이다. 2027년 엣지 수요 회복의 강도는 AI 기능 확산 속도와 메모리 가격의 줄다리기에 달려 있다.

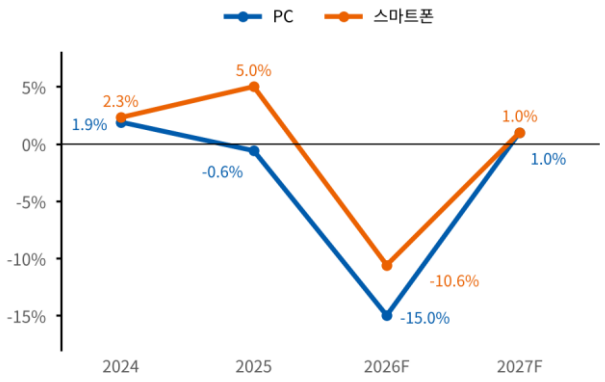
에이전트가 기기 안으로 들어오면 컨슈머 기기의 사양 기준도 달라진다. 지금의 에이전트는 대부분 데이터센터에서 구동되지만, 메타가 밝힌 것처럼 컴퓨트 비용이 서비스 확대의 제약이 되는 상황에서 일부 연산을 기기로 넘기려는 유인은 커질 수밖에 없기 때문이다. **사용자의 일정, 결제, 앱 사용 맥락을 상시 다루는 개인 에이전트가 기기에서 돌아가려면 스마트폰과 PC도 서버처럼 충분한 DRAM과 CPU·NPU 성능을 기본으로 갖춰야 한다. 엣지가 단순한 교체 수요가 아니라 새로운 메모리 탑재량 기준을 만드는 시장이 될 수 있는 이유다.** 다만 이 전환의 속도는 메모리 가격이 소비자 기기 원가에 주는 부담과 맞물려 결정될 것이다.

표 14. 기존 PC vs AI PC 메모리 비교

| 항목 | 기존 PC | AI PC (Standard / Copilot+) | AI PC (Local Pro / Dev) |
|---|---|---|---|
| 최소 용량 | 8 GB – 16 GB | 16 GB baseline<br>(32 GB recommended) | 64 GB – 128 GB+ |
| 주요 메모리 유형 | DDR4 / DDR5 (modular) | DDR5 / LPDDR5X | Unified LPDDR5X /<br>High-Bandwidth Pool |
| 메모리 버스 / 스피드 | 3200 MT/s (DDR4) / 4800+ MT/s<br>(DDR5) | Fast multi-channel DDR5 | 256-bit bus, up to 8000 MT/s<br>(e.g., Ryzen AI Max) |
| 대역폭 | ~50–90 GB/s | ~100–128 GB/s | 256 GB/s+ |
| AI 워크로드 범위 | Cloud-dependent<br>(Copilot web/light features) | Local light models<br>(7B–13B quantized) | Large open-weight models<br>(200B+ on unified platforms) |
| 게이트키퍼 | System RAM for apps/OS | NPU (40+ TOPS) + 16GB RAM<br>minimum | VRAM/Unified capacity for model<br>weights |

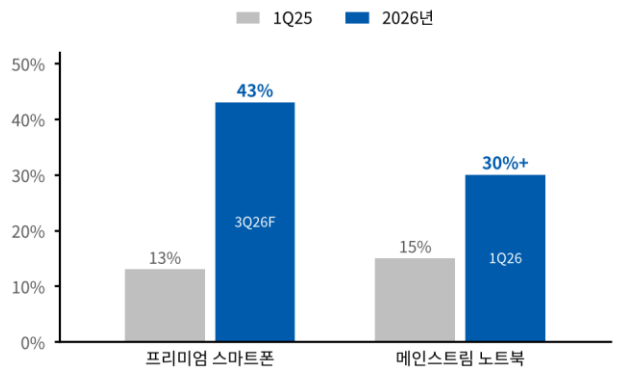
자료: 업계자료, 유안타증권 리서치센터

차트 15. 글로벌 PC · 스마트폰 출하량 증감률 추이




자료: 업계자료, 유안타증권 리서치센터

차트 16. 기기별 원가(BOM) 내 메모리 비중




자료: Counterpoint, TrendForce, 유안타증권 리서치센터 주: 노트북은 900 달러 기준

# 2028~2030년: 추론 다음의 수요, 미세화 다음의 기술

2028년 이후에 대한 질문은 두 가지다. 수요가 2028년 이후에도 이어지는가, 그리고 기술이 그 수요를 따라갈 수 있는가다. **수요 측면에서 당사는 추론 다음 단계로 '행동'에 주목한다.** AI는 배우는 단계(학습)에서 생각하는 단계(추론)로 넘어왔고, 다음은 스스로 실행하고 움직이는 단계다. 에이전트가 거래와 업무를 직접 실행하는 디지털 행동과, 로봇과 자율주행으로 대표되는 물리적 행동이 여기에 해당한다. 행동은 추론과 달리 멈추지 않고 돌아가며, 지연을 허용하지 않고, 겪은 결과를 기억으로 쌓아야 한다. **수요를 설명하는 변수도 토큰에서 실행 건수와 기기 대수, 가동 시간으로 넓어질 전망이다.** 다만 물리적 행동이 실적으로 확인되는 시점은 2028년 이후로 판단한다. 공급 측면에서는 미세화만으로 성능을 높이던 방식이 한계에 이르면서 공정과 패키징, 연결, 전력에서 기술 전환이 동시에 진행된다. 이하에서는 2030년까지의 기술 로드맵을 따라 다음 병목이 생길 지점을 점검하고자 한다.

표15. AI 의 단계별 수요 변수와 중심 반도체

| 단계 | 사람에 빗대면 | 수요를 설명하는 변수 | 중심 반도체 |
|---|---|---|---|
| 학습 | 배운다 | 설비투자(CapEx), GPU 대수 | GPU, HBM |
| 추론 | 생각한다 | 토큰 | 메모리 대역폭, ASIC, 첨단 패키징 |
| 행동 | 움직인다 | 실행 건수, 기기 대수, 가동 시간 | 기기용 저전력 메모리(LPDDR·PIM), 센서, 전력반도체, 저지연 연결 |
| 기억 | 겪은 것을 쌓는다 | 저장되는 경험의 양 | 서버 DRAM, Enterprise SSD |

자료: 유안타증권 리서치센터

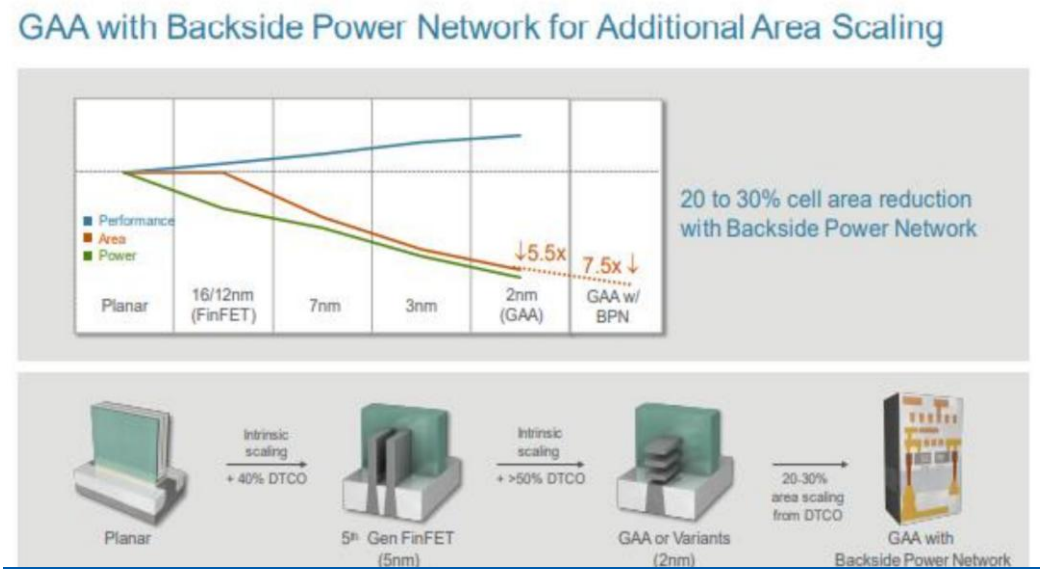
## 1. 공정: 줄이는 시대에서 세우고, 뒤집고, 붙이는 시대로

### 로직 공정: 트랜지스터, 배선, 노광이 동시에 바뀐다

2nm 이후의 선단 공정은 세 가지 변화가 겹친다. 트랜지스터 구조는 FinFET에서 GAA로, 전원 배선은 웨이퍼 앞면에서 뒷면으로, 노광은 기존 EUV에서 High-NA EUV로 넘어간다. 과거에는 세대마다 한 가지 변화만 소화하면 됐지만, 2025~2030년에는 이 세 가지가 2~3년 간격으로 잇따라 도입된다. 공정 난이도가 높아지는 만큼 같은 웨이퍼 한 장을 만드는 데 필요한 장비와 소재, 공정 단계도 늘어난다. TSMC는 2026년 설비투자를 600~640억달러로 제시했고, 2026~2028년 설비투자가 직전 3년보다 대폭 늘어날 것으로 제시하고 있다.

첫 번째 변화인 GAA는 이미 양산 궤도에 올랐다. 트랜지스터는 게이트가 채널을 감싸 전류를 켜고 끄는 스위치다. FinFET은 지느러미 모양의 채널을 게이트가 세 면에서 감쌌지만, 선폭이 더 줄어들면 꺼진 상태에서도 전류가 새는 누설이 커진다. GAA(Gate-All-Around)는 채널을 얇은 판(Nanosheet) 여러 장으로 쌓고 게이트가 네 면을 모두 감싸 누설을 줄이는 구조다. **TSMC N2는 2025년 4분기부터 양산을 시작했고, 2026년 3분기 기준으로는 매출의 8%를 차지했다. N2 생산능력은 2027년에 두 배 이상 늘어날 것으로 예상한다.** 동시에 채널을 여러 층으로 쌓고 그 사이를 정밀하게 깎아내야 하는 만큼 공정 단계가 크게 늘어난다. **원자 단위로 막을 쌓는 ALD 장비와 특수 전구체, 특정 물질만 골라 깎는 선택적 식각 장비의 수요가 늘어나는 이유다. 나아가서는 고출력 공정 환경을 견뎌야 하는 SiC Ring, Quartz 등 소모성 부품의 교체 주기도 짧아진다.**

그림 4. 트랜지스터 구조의 진화: Planar → FinFET → GAA → GAA+후면 전력 공급

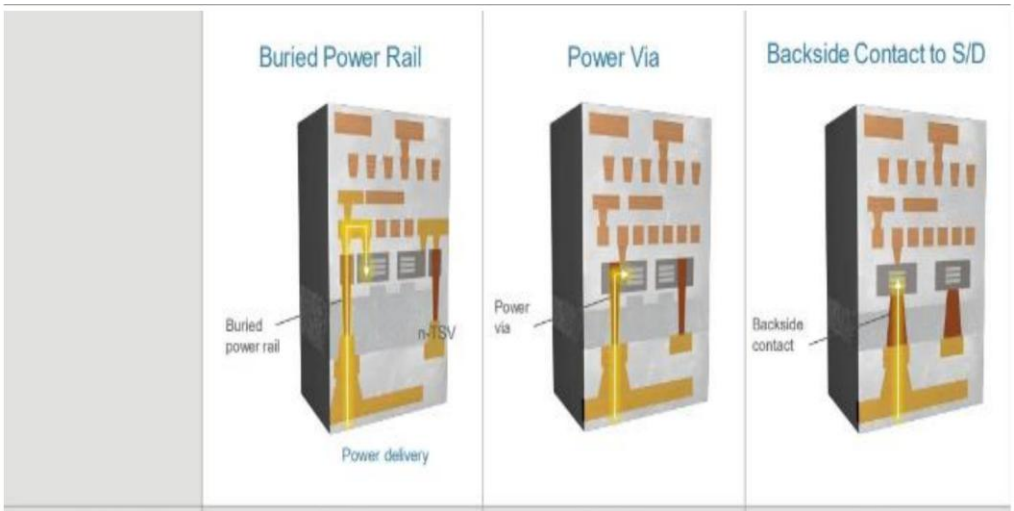



자료: Applied Materials, 유안타증권 리서치센터

두 번째 변화는 전원 배선을 웨이퍼 뒷면으로 옮기는 후면 전력 공급(BSPDN)이다. 지금까지는 신호선과 전원선이 모두 트랜지스터 위쪽의 앞면 배선층에 섞여 있었다. 선폭이 줄어들수록 배선이 빽빽해지고, 전원선을 따라 전압이 떨어지는 문제가 커진다. BSPDN은 전원선을 웨이퍼 뒷면으로 빼서 앞면은 신호선만 쓰게 하는 방식이다. Intel은 18A 공정에 PowerVia라는 이름으로 이를 먼저 적용해 양산 중이고, TSMC는 A16(1.6nm)에 Super Power Rail을 적용하는데, 2026년 4월 기술 심포지엄에서 본격 양산 시점을 2027년으로 늦췄고, 데이터센터용 후속 공정인 A12(2029년)에서 차세대 Super Power Rail을 이어갈 계획이다. TSMC가 제시한 A16의 목표치는 N2P 대비 같은 전력에서 속도 8~10% 향상, 같은 속도에서 전력 15~20% 절감, 집적도 최대 10% 향상이다. 전력 소모와 발열이 가장 큰 AI·HPC 칩이 첫 수요처가 될 전망이다.

BSPDN을 구현하는 방식은 크게 세 가지다. 전원선을 트랜지스터 사이에 묻는 Buried Power Rail, 전원선을 뒷면에 두고 나노 단위의 관통 전극(Nano-TSV)으로 트랜지스터 옆을 연결하는 PowerVia, 뒷면에서 트랜지스터의 소스·드레인에 직접 연결하는 Backside Contact다. 뒤로 갈수록 배선이 짧아지고 면적 이득이 커지지만 공정 난도도 높아진다. Intel 18A가 두 번째, TSMC A16의 Super Power Rail이 세 번째 방식이며, 삼성전자도 2nm 공정의 후속 버전(SF2Z)에 후면 전력 공급을 도입할 계획을 밝힌 바 있다. 공정 순서를 보면 수혜 영역이 분명해진다.

그림 5. 후면 전력 공급(BSPDN)의 세 가지 구현 방식

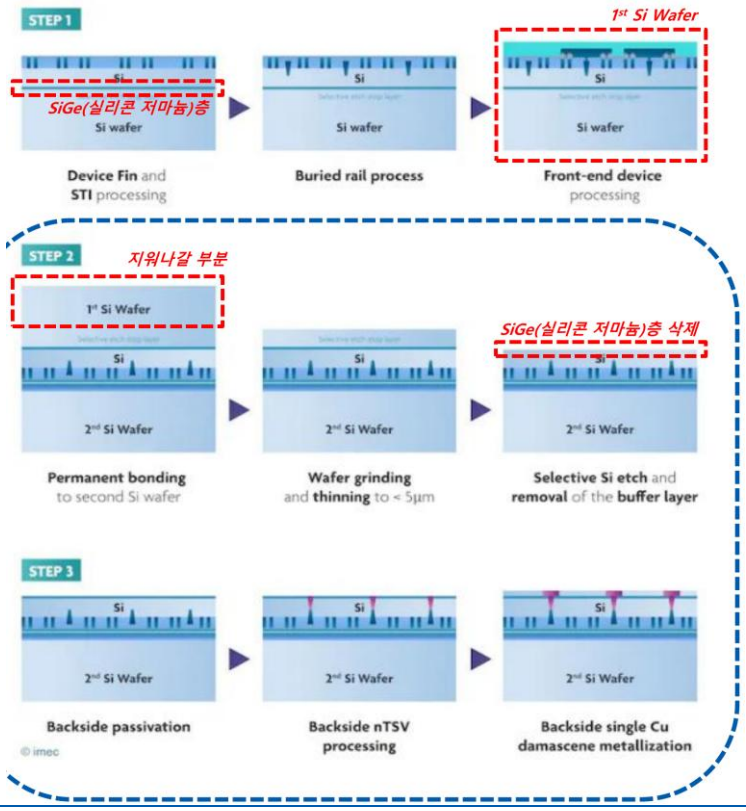



| | | | |
|---|---|---|---|
| Area Scaling | Good | Better | Best |
| Process Complexity | Low | Medium | High |

자료: Applied Materials, 유안타증권 리서치센터

먼저 웨이퍼 앞면에 트랜지스터와 신호 배선을 모두 만든 뒤, 그 위에 빈 실리콘 웨이퍼(캐리어 웨이퍼)를 붙이고(Fusion Bonding) 통째로 뒤집는다. 이어 원래 웨이퍼의 바닥면을 Grinding, CMP, 선택적 식각을 반복해 10㎛ 미만까지 깎아내고, 드러난 뒷면에 전원 배선층을 새로 쌓는다. **웨이퍼 본딩 장비, CMP 장비와 슬러리, 선택적 식각 장비, 얇아진 웨이퍼를 다루기 위한 임시 본딩·디본딩 소재가 새로 필요해지고, 앞면 트랜지스터와 뒷면 전원선을 10nm 이하의 오차로 맞춰야 하기 때문에 오버레이 계측의 중요성도 커진다.** 당사가 2023년 BSPDN을 DTCO(설계·공정 동시 최적화) 시대의 핵심 기술로 제시했을 때 연구 단계였던 공정들이, 2027~2028년 본격적인 양산 투자로 이어질 가능성도 배제할 수 없다.

그림 6. BSPDN 공정 흐름: 붙이고, 뒤집고, 깎고, 다시 배선한다



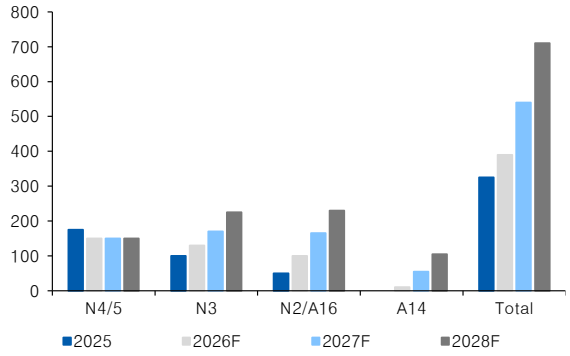

자료: IMEC, 유안타증권 리서치센터

세 번째 변화는 High-NA EUV다. EUV 노광 장비의 해상도는 빛을 모으는 렌즈의 개구수(NA)에 따라 정해진다. 기존 EUV의 NA는 0.33, High-NA EUV는 0.55로, 더 가는 선을 한 번에 찍을 수 있다. 기존 장비로는 같은 패턴을 여러 번 나눠 찍어야 하는 층을 한 번에 끝낼 수 있다는 것이 장점이지만, 장비 가격이 기존 EUV보다 크게 비싸고 한 번에 찍을 수 있는 면적이 절반으로 줄어든다는 부담이 있다. 그래서 업체마다 도입 시점이 갈린다. Intel은 14A에서 High-NA EUV를 처음으로 양산에 쓸 계획이며, 최근에는 위험 생산을 2028년, 양산을 2029년으로 제시했다. 삼성전자는 기흥 NRD-K 연구단지에 High-NA EUV를 설치하고 2029년 양산을 목표로 하는 1.4nm(SF1.4)부터 일부 층에 적용할 계획이다. **반면 TSMC는 2027년 위험 생산, 2028년 양산 예정인 A14까지 기존 EUV를 유지하고, High-NA EUV는 2030년부터 쓸 계획이라고 밝혔다.**

**TSMC의 선택은 장비 시장에 의미가 있다.** **High-NA EUV 없이 A14를 만들려면 기존 EUV로 여러 번 나눠 찍는 멀티 패터닝이 늘어나고, 그만큼 증착·식각·세정 공정의 반복 횟수도 늘어난다.** 선단 공정 웨이퍼 한 장에 들어가는 장비 투자는 늘어나는 구조다. 2030년 이후에는 nMOS와 pMOS 트랜지스터를 위아래로 쌓는 CFET(Complementary FET)이 다음 구조로 거론되지만, 그 전까지는 트랜지스터 자체를 작게 만들기보다 칩을 나눠 만든 뒤 쌓는 방향이 성능 향상의 주된 수단이 된다. A14 단계에서는 연산부만 최선단 공정으로 만들고 나머지를 성숙 공정으로 나눈 뒤 3D로 집적하는 칩렛 설계가 표준이 될 전망이며, 공정이 고도화될수록 표준 셀 라이브러리, 고속 인터페이스 IP, 디자인하우스의 역할도 커진다.

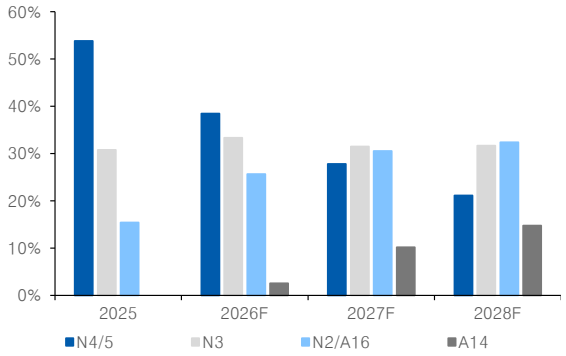
**경쟁 구도에서는 TSMC의 우위가 이어지는 가운데 삼성전자의 회복 여부가 관건이다.** TSMC는 미국 애리조나 2공장의 양산 시점을 2027년 3분기로 앞당겼고, AI 가속기 매출의 2024~2029년 연평균 성장률이 기존에 제시한 50% 중후반보다 높을 수 있다고 밝히고 있다. Intel은 18A 양산을 시작했지만 수율과 공급은 아직 안정화 단계이며, 14A는 의미 있는 외부 고객을 확보하지 못하면 개발을 중단할 수 있다고 밝힌 만큼 외부 고객 확보가 변수다. **삼성전자 파운드리는 2nm에서 회복의 실마리를 찾고 있다.** 1.4nm 양산 목표를 2029년으로 늦추고 2nm 수율 안정화에 집중하는 전략이다. 2세대 2nm 공정(SF2P)의 첫 제품인 Exynos 2700은 10월 초 웨이퍼 투입을 시작한 것으로 파악되며, 1세대 2nm 대비 성능 약 12% 향상, 전력 약 25% 절감, 면적 약 8% 축소를 목표로 한다. 삼성전자는 10월 말 3Q26 실적 발표에서 2nm 수주 과제 수가 전년 대비 두 배 이상이라고 밝혔다. 4분기에는 2nm 2세대 모바일 제품과 주요 고객 4nm NPU의 본격 양산을, 2027년에는 2nm 주요 고객의 하반기 양산과 8nm 이하 공정 가동률 100% 초과를 제시했다. 2025년 체결한 Tesla향 대규모 파운드리 계약과 함께 2nm 수율 안정화가 확인될 경우, 2027년은 삼성전자 파운드리의 실적 턴어라운드 원년이 될 수 있다.

차트 17. TSMC 공정별 생산능력 추이 및 전망




자료: TSMC, 유안타증권 리서치센터

차트 18. TSMC 공정별 생산 비중 추이 및 전망



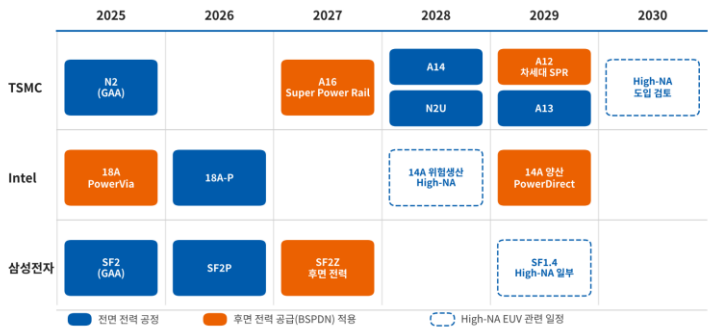

자료: TSMC, 유안타증권 리서치센터

표 16. N2 vs A16 세부 설명

| 구분 | N2 (2nm) | A16 (1.6nm) |
|---|---|---|
| 양산 타임라인 | 2025년 하반기 양산 개시 | 2026년 하반기 위험 생산 -> 2027년 본격 양산 |
| 트랜지스터 구조 | GAA (Nanosheet) 최초 적용 | GAA (Nanosheet) 고도화 |
| 전력 배선 구조 | • 전면 배선 (Frontside Power)<br>• 기존 방식 유지 (신호선과 전원선이 전면에 혼재) | • 후면 전력 배선 (BSPDN / Super Power Rail)<br>• 전원선을 웨이퍼 후면으로 완벽히 분리 |
| 성능/전력 효율 | N3E 대비 성능 +10~15%, 전력 -25~30% 개선 | N2P 대비 성능 +8~10%, 전력 -15~20% 추가 개선 |
| 배선 밀도 (Density) | 표준 GAA 배선 밀도 확보 | 후면 전원 분리로 배선 혼잡도 완화<br>→ 칩 면적 최대 10% 추가 축소 |
| 주요 난제/수율 | GAA 나노시트 패턴 정밀도 제어 | 웨이퍼 후면 CMP 연마 및 nTSV/본딩 수율 확보 |
| 핵심 타겟 | 모바일 AP (Apple A20 등), 메인스트림 AI 칩 | 초고성능 AI 서버 가속기 (HPC/Data Center ASIC) |

자료: 업계자료, 유안타증권 리서치센터

그림 7. 파운드리 3 사 선단 공정 로드맵(2025~2030 년)



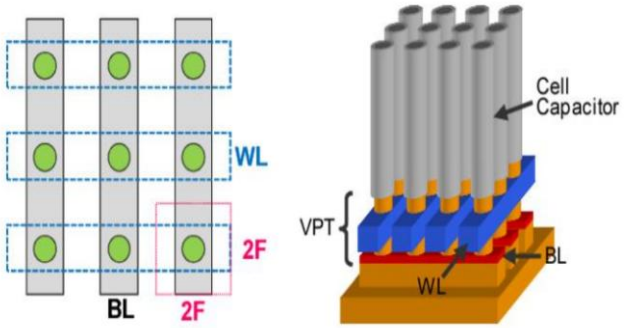

자료: TSMC(2026 North America Technology Symposium), Intel, 삼성전자, Tom's Hardware, 유안타증권 리서치센터 주: 회사 발표와 언론 보도 종합

### 메모리 공정: DRAM 은 세우고, NAND 는 붙인다

**메모리 공정에서도 2028년 전후로 구조 변화가 시작된다.** DRAM은 지금까지 트랜지스터를 눕힌 채 선폭을 줄여 집적도를 높여왔지만, 10nm대 초반에서 이 방식이 한계에 다다랐다. 셀마다 데이터를 담는 커패시터의 용량을 확보하기 어려워지고, 셀 사이 간격이 좁아지며 누설 전류와 간섭이 커지기 때문이다. **해법은 트랜지스터를 세우는 것이다. 채널을 수직으로 세운 VCT(Vertical Channel Transistor) 구조를 쓰면 셀 하나가 차지하는 면적을 현재의 $6F^2$에서 $4F^2$로 약 3분의 1 줄일 수 있다.**

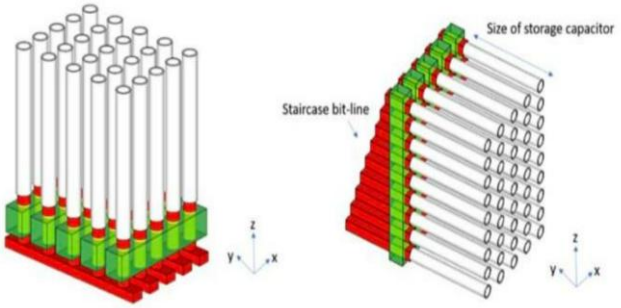
삼성전자는 2026년 3월 $4F^2$·VCT 구조의 10nm 미만(10a) DRAM에서 정상 작동하는 다이를 확보했고, 2027년 품질 테스트를 거쳐 2028년 양산을 목표로 하고 있다. SK하이닉스는 그다음 세대인 10b부터 $4F^2$ 구조를 적용할 것으로 알려졌다. 셀을 수직으로 여러 층 쌓는 3D DRAM은 그 이후인 10d 세대부터 거론되고 있어 2030년대 초반이 될 전망이다. 당사는 2023년 VCAT(현재의 VCT) 도입을 2026년, 3D DRAM의 원가 경쟁력 확보를 2027~2028년으로 전망했는데, 방향은 그대로 진행되고 있지만 시점은 1~2년가량 늦어지고 있다.

그림 8. $4F^2$ VCT(수직 채널 트랜지스터) 구조 모식도




자료: 업계 자료, 유안타증권 리서치센터

그림 9. 2D DRAM vs 3D DRAM 모식도



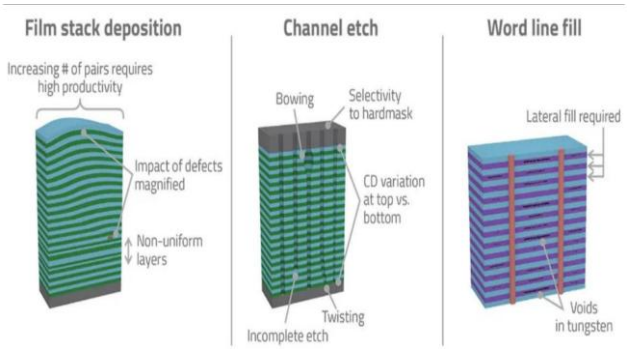

자료: 업계 자료, 유안타증권 리서치센터

**$4F^2$ 구조에서는 셀 아래 공간이 부족해, 셀을 구동하는 주변 회로(Peri)를 별도의 웨이퍼에 만들어 셀 웨이퍼와 붙이는 방식이 거론된다.** SK하이닉스가 주변 회로용 12인치 로직 라인 확보를 두고 자체 구축과 외주를 저울질하고 있다는 보도도 이 때문일 것으로 추정한다. 지금까지 패키징의 영역이던 웨이퍼 본딩이 DRAM 전공정의 핵심 단계로 들어올 가능성이 높아지고 있다.

NAND는 웨이퍼를 붙이는 방향으로 간다. NAND는 셀을 위로 쌓는 단수 경쟁을 이어왔지만, 400단에 가까워지면서 쌓을수록 웨이퍼가 휘고, 깊은 구멍을 뚫는 식각이 어려워지며, 셀 아래에 둔 주변 회로가 고온 공정에서 손상되는 문제가 커졌다. 이에 셀과 주변 회로를 각각 다른 웨이퍼에 만든 뒤 붙이는 웨이퍼 본딩이 표준이 되고 있다.

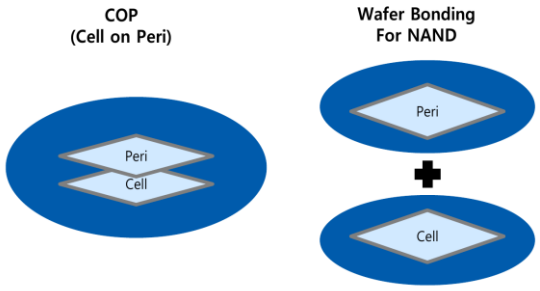
Kioxia는 이미 CBA(CMOS Directly Bonded to Array) 방식으로 이를 양산 중이고, 삼성전자는 400단급 V10 NAND에 웨이퍼 간(W2W) 본딩을 처음 도입할 예정이며, SK하이닉스도 차세대 NAND에 하이브리드 본딩을 앞당겨 2027년 양산을 추진하는 것으로 알려졌다. 400단급의 깊은 채널 홀을 뚫기 위해 삼성전자는 영하의 온도에서 식각하는 극저온 식각 장비 도입을 준비하고 있고, SK하이닉스는 워드라인 일부의 소재를 텅스텐에서 몰리브덴으로 바꾸고 있는 것으로 파악된다. 당사가 2023년 'NAND는 단기적으로 단수 경쟁이 이어지되 중장기적으로 COP(Cell on Peri)에서 웨이퍼 본딩으로 전환할 것'이라고 제시한 흐름이 현실화되고 있다.

그림 10. NAND 단수 증가에 따른 공정 난제




자료: 업계 자료, 유안타증권 리서치센터

그림 11. NAND 구조 변화: COP → 웨이퍼 본딩



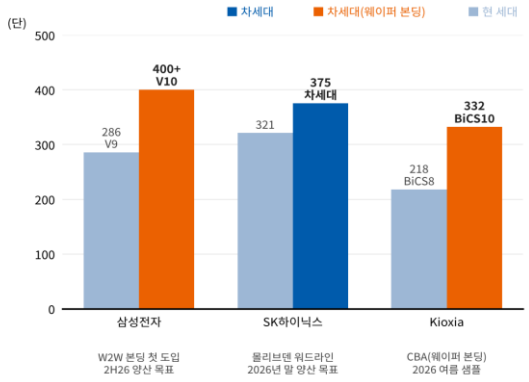

자료: 업계 자료, 유안타증권 리서치센터

그림 12. DRAM 셀 구조 로드맵: $6F^2$ → $4F^2$ VCT → 3D DRAM




자료: TrendForce(2026.4), 업계 자료, 유안타증권 리서치센터

차트 19. 주요 업체별 NAND 적층 단수: 현 세대 vs 차세대



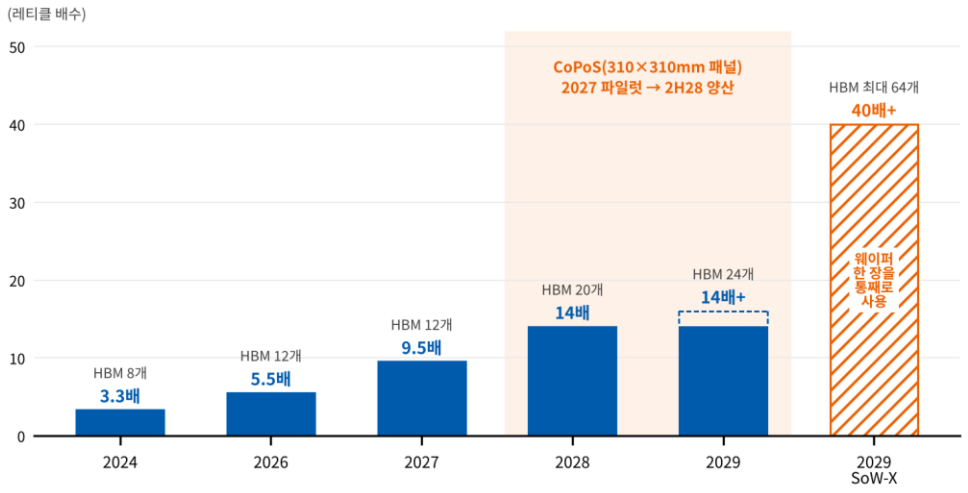

자료: TrendForce(2026.6), Kioxia, 업계 자료, 유안타증권 리서치센터

**메모리 공정의 변화는 2028~2030년 장비 수요의 성격을 바꾼다.** DRAM과 NAND 모두에서 웨이퍼 본딩과 정밀 정렬, 오버레이 계측의 수요가 커지고, 수직 구조를 만들기 위한 원자층 증착(ALD)과 선택적 증착, 극저온 식각 같은 고난도 공정의 비중이 높아진다. **이러한 구조 전환은 기술 진입 장벽을 높이고, 제한적인 생산을 야기한다면, 메모리 공급 제약이 2028년 이후에도 쉽게 풀리지 않는 또 하나의 이유가 될 가능성도 배제할 수 없다는 판단이다.**

## 2. 패키징: 반도체의 공통 언어, 본딩

AI 가속기의 크기는 이미 단일 레티클의 한계를 훌쩍 넘어섰다. MediaTek은 더 나은 Perf/Watt와 Perf/TCO를 위해 레티클의 10배가 넘는 대형·복합 통합이 필요하다고 강조하며, 복수의 컴퓨트 다이와 메모리 인터페이스 다이, I/O 칩렛, HBM, 광·전기 인터커넥트를 하나의 패키지에 담는 xPU 구조를 제시했다. 자사 ASIC 로드맵에서도 CoWoS-S/R/L 기반 최대 120×150mm 이상 크기의 패키지, 448G SerDes, PCIe Gen7, 1.6T/3.2T 이더넷을 핵심 기술로 꼽았다. 장기적으로는 SiP를 넘어 웨이퍼 하나를 시스템으로 쓰는 System on Wafer로의 진화를 목표로 한다.

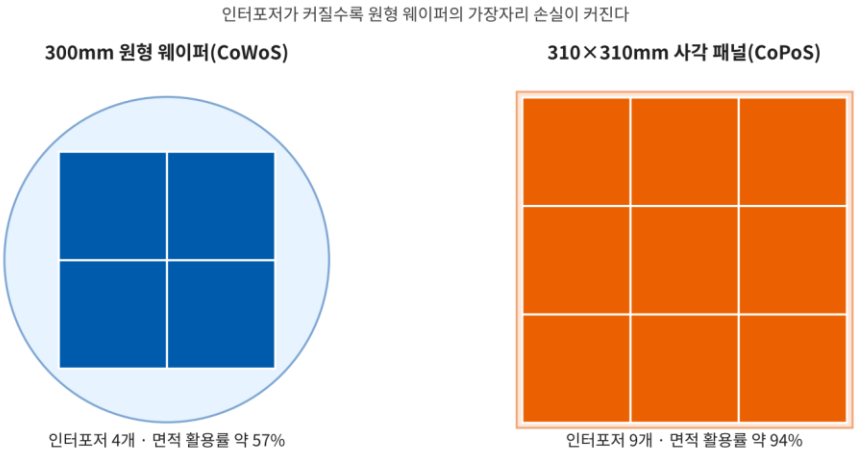
차트 20. TSMC CoWoS 인터포저 크기 로드맵과 HBM 탑재 개수




자료: TSMC(2025 · 2026 Technology Symposium), TrendForce, 유안타증권 리서치센터 주: 레티클 배수는 노광 1 회 최대 면적(약 858mm² ) 대비 크기,

이에 따라 기존 300mm 원형 웨이퍼 기반 인터포저 패키징은 물리적·경제적 한계에 도달하고 있다. TSMC의 CoWoS가 2027년 월 18만장 이상까지 확장되는 가운데 WMCM과 3D 적층 기술인 SoIC의 비중이 빠르게 확대될 전망이며, **사각 패널 기반 CoPoS(FOPLP)도 2027년 파일럿 라인 가동을 거쳐 2028년 하반기 양산에 들어갈 것으로 예상한다.** 이는 2028년 이후 등장할 레티클 10배 이상의 초대형 AI 칩을 수용하기 위한 기술적 전제다. 패널 전용 노광기, RDL 장비, 패널 휘어짐을 제어하는 고내열·저수축 소재가 새로운 수혜 영역이다.

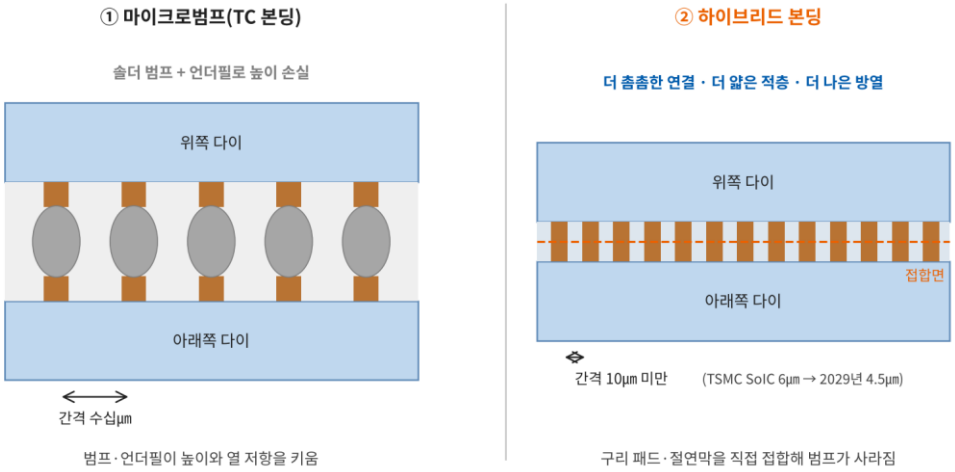
그림 13. 원형 웨이퍼 vs 사각 패널: 대형 인터포저의 면적 활용




자료: TSMC, 업계 자료, 유안타증권 리서치센터 주: 100×100mm 인터포저, 가장자리 여유 3mm 가정 시 예시

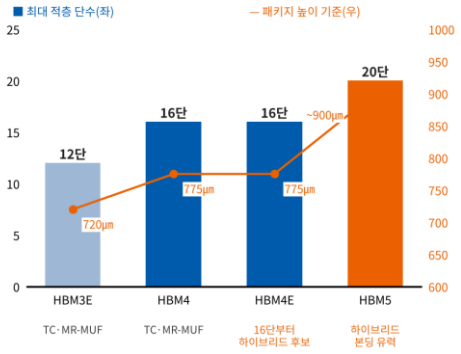
칩을 옆에 놓는 방식에서 위로 쌓는 방식으로의 전환도 2028년부터 본격화된다. 칩과 칩을 범프(작은 납땜 돌기) 없이 구리 패드끼리 직접 붙이는 하이브리드 본딩은 이미 로직 칩에서 양산 중이다. **TSMC의 3D 적층 기술인 SoIC는 연결 간격(피치)을 2023년 9㎛에서 2025년 6㎛로 좁혔고, 2029년 4.5㎛를 목표로 한다. 쌓는 칩의 공정도 높아져, 현재 N3P 위에 N4를 쌓는 단계에서 2028년 N2P 위에 N2P, 2029년 A14 위에 A14를 쌓는 단계로 간다.** NVIDIA가 2028년 Feynman 세대에서 GPU 다이를 적층하겠다고 밝힌 것도 같은 흐름이다. 칩을 쌓으면 같은 면적에 더 많은 연산을 담을 수 있지만, 열이 위아래로 빠져나가기 어려워지고 쌓은 칩 하나만 불량이어도 전체를 버려야 해 테스트와 열 관리의 중요성이 커진다.

그림 14. 마이크로범프 vs 하이브리드 본딩 단면 비교



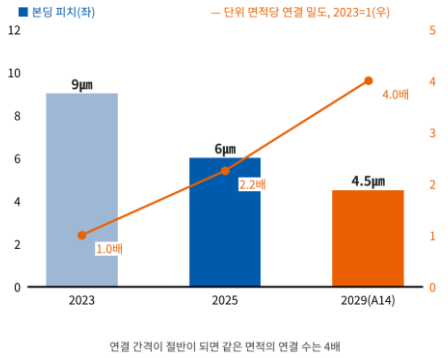

자료: TSMC, 업계 자료, 유안타증권 리서치센터 재구성 주: 개념도

차트 21. HBM 세대별 적층 단수와 패키지 높이 기준




자료: JEDEC, TrendForce(2026.7), 유안타증권 리서치센터 주: HBM5 높이는 JEDEC 논의 중인

차트 22. TSMC SoIC 본딩 피치와 연결 밀도




자료: TSMC, 유안타증권 리서치센터 주: 연결 밀도는 피치 제곱에 반비례로 단순 환산

반면 HBM의 하이브리드 본딩 도입은 예상보다 늦어지고 있다. JEDEC이 2026년 1월 HBM 패키지 높이 기준을 720㎛에서 775㎛로 완화하면서, 16단 HBM4도 기존 마이크로범프 방식으로 쌓을 수 있게 됐기 때문이다. HBM4 수준의 패드 간격에서는 아직 하이브리드 본딩의 비용 효율이 낮다는 판단도 작용했다. SK하이닉스는 16단 HBM4에 기존 MR-MUF 방식을 유지하고 하이브리드 본딩은 후속 세대에서 검증하고 있어, HBM의 하이브리드 본딩은 HBM4E~HBM5 세대를 전후해 도입 준비에 들어갈 가능성이 있다.

표17. 차세대 반도체에 요구되는 본딩 기술

| 디바이스 | 붙이는 대상 | 방식 | 단계(2026년 기준) |
|---|---|---|---|
| 3D NAND | 셀 + 주변 회로 | W2W, 하이브리드 | 양산(Kioxia CBA) → 2026~27년 확대(삼성전자·SK 하이닉스 차세대) |
| DRAM(4F²) | 셀 + 주변 회로 | W2W, 하이브리드 | 개발 → 2028년 양산 목표(삼성전자 10a) |
| HBM | DRAM 다이 20단 이상 적층 | D2W, 하이브리드 | 검증 → HBM4E~HBM5(2020년대 말) |
| 로직 3D 적층(SoIC) | 로직 + 로직·SRAM | D2W·W2W, 하이브리드 | 양산(피치 6㎛ → 2029년 4.5㎛) |
| BSPDN | 소자 웨이퍼 + 캐리어 웨이퍼 | W2W, Fusion | 양산(Intel 18A) → 2026~27년 확대(TSMC A16, 삼성전자 SF2Z) |

자료: 각 사, TrendForce, Tom's Hardware, 유안타증권 리서치센터 주: 유안타증권 '2024년 반도체 전망'(2023.11) 표를 2026년 기준으로 업데이트

표 18. CoWoS (Chip-on-Wafer-on-Substrate) vs CoPoS (Chip-on-Panel-on-Substrate) 세부 설명

| 구분 | CoWoS (Chip-on-Wafer-on-Substrate) | CoPoS (Chip-on-Panel-on-Substrate) |
|---|---|---|
| 기본 개념 및 공법 | • 2.5D 실리콘 인터포저 기반 웨이퍼 레벨 패키징 (CoW)<br>• 300mm 원형 실리콘 웨이퍼 위에서 칩을 집적 | • 사각 패널 기반 팬아웃 패키징 (FOPLP)<br>• 대면적 사각형 패널(Panel RDL) 위에서 칩을 집적 |
| 기판/캐리어 형태 | 300mm (12인치) 원형 실리콘 웨이퍼 | 310×310mm / 510×515mm 등 대면적 사각 패널 |
| 중간 연결층 (Interposer) | • Silicon Interposer (TSV 공정 적용)<br>• 초미세 배선(Line/Space < 1㎛) 구현 및 고밀도 I/O | • Panel RDL (Redistribution Layer)<br>• 유기 재배치층 배선 활용 (TSV 미사용, RDL 기반) |
| 면적 이용률 (Efficiency) | • 낮음 (~60~70%)<br>• 원형 가장자리(Edge) 부분의 미사용 버림 영역 (Waste) 발생 | • 매우 높음 (>95%)<br>• 직사각형 구조로 패널 모서리까지 칩 배치 가능 |
| 레티클 면적 한계 | • 현재 5배 양산, 9배·14배 개발 중(CoWoS-L)<br>• 인터포저 대면적화 시 웨이퍼 수율 급락 및 단가 폭증 | • 9.5x ~ 15x Reticle Size 지원 가능<br>• 초대형 모듈 및 멀티 다이(Multi-die) 집적에 용이 |
| 물성 한계 (열/열변형) | • 열 팽창 계수(CTE) 정합 우수 (실리콘 간 결합)<br>• 면적 확장 시 인터포저 얇아짐에 따른 균열 위험 | • 열 변형(Warpage) 제어가 최우선 과제<br>• 사각 패널 면적 확장 시 열 응력에 의한 휘어짐 발생 |
| 제조 공정 및 장비 | • 전공정(FEOL) 기반 첨단 스티칭/노광 장비<br>• EUV/ArFi 노광 및 300mm 웨이퍼 핸들링 라인 | • 후공정(BEOL) 및 대량 패널 전용 라인<br>• 대면적 대량 스퍼터/라미네이션 및 패널 노광기 |
| 제조 원가 (Cost) | • 고비용 (High CAPEX / High OPEX)<br>• 비싼 실리콘 인터포저 가공 및 TSV 공정 비용 발생 | • 상대적 저비용 (원가 20~30% 절감 가능)<br>• 원형 대비 패널당 유효 다이 수율 증가 및 대량 생산성 |
| 주요 채용 AI 가속기 | • NVIDIA Hopper (H100/H200), Blackwell (B200)<br>• AMD Instinct MI300 시리즈 | • 2028년 이후 차세대 초대형 Custom ASIC |
| 상용화 타임라인 | "• 현재 양산 메인스트림<br>• 2027년 월 18만장 이상으로 확대 | • 2027년 하반기 파일럿/초기 상용화 (월 8천 장)<br>• TSMC 200mm Fab5 리퍼포싱을 시작으로 2028년 본격화 |

자료: 업계자료, 유안타증권 리서치센터

## 3. 연결: 빨라진 신호가 구리선을 떠나는 곳

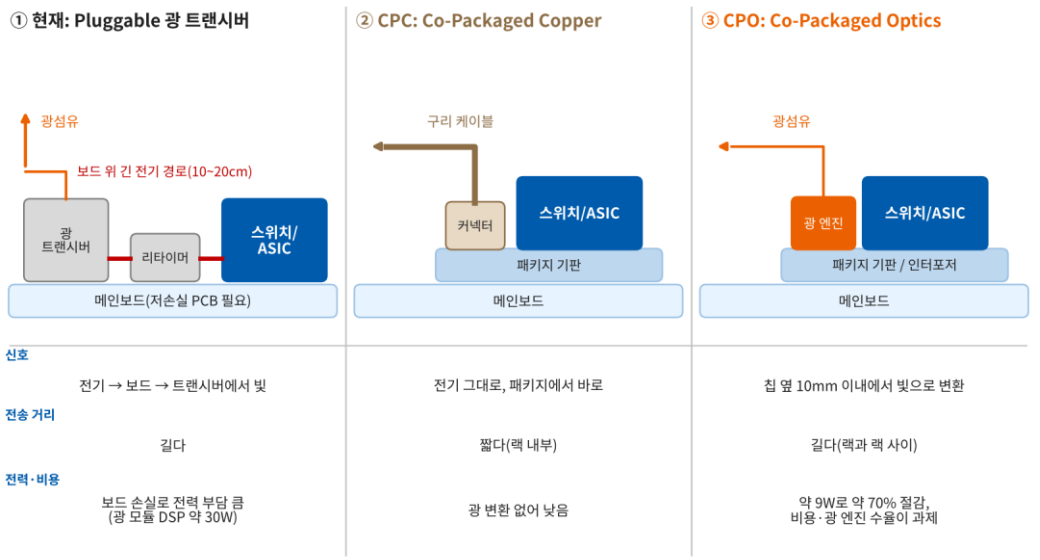
초대형 AI 클러스터의 다음 한계는 칩 내부 연산이 아니라 칩 밖으로 데이터를 옮기는 연결이다. 앞서 본 것처럼 인터커넥트 대역폭은 지난 20년간 2년마다 1.4배 증가하는 데 그쳐 가장 느리게 개선된 영역이다. 연결의 병목은 거리에 따라 두 곳에서 동시에 나타난다. 랙과 랙, 스위치와 GPU를 잇는 수 m 이상의 연결에서는 구리선을 빛으로 바꾸는 CPO가, 칩과 메모리를 잇는 보드 위 수 cm의 연결에서는 메모리가 신호 방식과 모듈 구조를 바꾸며 스스로 해법을 찾고 있다. 두 영역 모두 신호가 빨라질수록 구리선 위에서 더 많이 약해진다는 같은 문제에서 출발한다.

### 랙 간 연결: 구리에서 빛으로

MediaTek은 대형 패키지에서 200/400Gbps 신호를 통과시키는 것이 점점 어려워지고 있으며, 업계가 향후 2~3년 내 Co-Packaged Copper(CPC)와 Co-Packaged Optics(CPO)로 이동할 것이라고 밝혔다. **두 기술 모두 칩의 신호를 메인보드 배선을 거쳐 멀리 보내지 않고 칩 패키지에서 곧바로 내보낸다는 점은 같다. 차이는 신호를 무엇으로 보내느냐다.**

CPC는 패키지에 구리 케이블을 직접 연결해 전기 신호 그대로 보내는 방식으로, 광 변환이 없어 비용과 전력 부담이 작지만 전송 거리가 짧아 랙 내부 연결에 적합하다. CPO는 전기 신호를 빛으로 바꾸는 광 엔진을 칩 바로 옆 패키지 안에 넣어 광섬유로 내보내는 방식으로, 비용은 높지만 먼 거리까지 대역폭을 유지할 수 있어 랙과 랙을 잇는 연결에 쓰인다.

그림 15. 칩 간 연결 방식의 변화: Pluggable → CPC → CPO




자료: MediaTek(JEDEC Industry Forum, 2026.10), Singapore IME(SEMICON China 2026), TrendForce, 유안타증권 리서치센터 재구성

**CPO는 이미 상용화 단계에 진입했다.** TrendForce에 따르면 NVIDIA는 TSMC의 COUPE 패키징을 적용한 최대 400Tb/s급 Spectrum-X CPO 스위치를 일부 고객에 출하하기 시작했고, Broadcom은 51.2T Bailly에 이어 레인당 속도를 200Gbps로 두 배 높인 102.4T급 Tomahawk 6-Davisson CPO 스위치를 내놨고, 10월 OCP 행사에서는 이를 탑재한 Celestica의 수냉식 스위치가 공개됐다. 업계는 CPO 스위치 시장이 2027~2028년 본격적인 물량 확대 구간에 진입할 것으로 전망한다. 전력 절감 효과가 분명하기 때문이다. 싱가포르 IME에 따르면 플러거블 광 모듈은 전기 신호가 보드를 따라 10~20cm를 이동해야 해 DSP 전력이 약 30W에 이르지만, 광 엔진을 칩 옆 10mm 이내에 붙이는 CPO는 이를 약 9W로 낮출 수 있다. Broadcom도 Davisson의 광 연결 전력이 플러거블 대비 약 70% 낮다고 밝혔다.

**다만 초기 상용화가 곧 본격적인 확산을 뜻하지는 않는다.** 고객 인증과 유지보수 편의성, 생산 수율, 의미 있는 출하 물량이 확인돼야 하는 단계다. 동시에 광 엔진의 수율과 열 관리, 실리콘 포토닉스 웨이퍼 공정의 정밀도, AI 칩과 같은 2.5D/3D 패키징 자원을 두고 경쟁해야 한다는 점이 확장 속도를 제약하는 요인이다.

**CPO의 성능은 결국 패키징 기술에서 갈린다.** 당사가 3월 SEMICON China 2026에서 확인한 고성능 CPO의 4대 핵심 기술은 본딩, 인터포저, 광 결합, 전송 속도다. 광 엔진의 전자 칩(EIC)과 광학 칩(PIC)을 D2W 하이브리드 본딩으로 붙여 칩 사이 간격을 극한으로 좁히고, 절연체를 관통하는 전극(TDV, Through Dielectric Via) 기반 인터포저로 신호 통로를 만든다. 여기에 광섬유와 칩 사이의 빛 손실을 최소화하는 수직·수평 광 결합(Coupling) 기술이 더해져야 레인당 400Gbps 이상의 전송 속도를 구현할 수 있다. 네 가지 모두 AI 칩의 첨단 패키징과 같은 장비와 공정을 쓴다는 점에서, CPO의 확산은 앞서 살펴본 하이브리드 본딩과 패키징 생산능력에 대한 수요를 한층 키운다. 다음 단계는 인터포저 자체가 빛을 전달하는 포토닉 인터포저다. 기존 구조 대비 기생 성분을 5배 이상 줄여 신호 품질을 높일 수 있어, 광 연결의 병목이 정점을 지나는 2028년 이후의 기술 과제가 될 전망이다.

**CPO 확산은 연결 병목이 해소되는 과정이자 새로운 병목이 만들어지는 과정이다.** 광 엔진과 레이저 광원, 광학 패키징이 새로운 공급 제약 지점이 되고, 첨단 패키징 캐파 수요는 AI 칩과 스위치 양쪽에서 동시에 늘어난다. 중장기적으로 광반도체 IP, 레이저 광원 모듈, 광학 패키징 밸류체인의 리레이팅을 기대한다.

표19. CPO 주요 부품별 밸류체인

| 핵심 부품 | 역할 | 글로벌 주요 업체 | 중국 주요 업체 |
|---|---|---|---|
| 광 엔진(실리콘 포토닉스) | 전기 신호를 빛으로 바꾸는 CPO 의 핵심. 레인당 속도 경쟁(Broadcom 은 200Gbps 세대) | Broadcom, NVIDIA, Marvell, Intel, Ayar Labs | HGTECH, Innolight |
| 레이저 광원(InP) | 광 엔진에 빛을 공급하는 연속파(CW) 레이저. 고출력·저발열이 관건 | Lumentum, Coherent | Yuanjie Tech, Source Photonics |
| 광도파로·패키징 플랫폼 | 칩과 광 엔진을 하나의 패키지로 연결. TSMC COUPE 가 가장 앞선 플랫폼 | TSMC(COUPE), GUC | TFC Optical, JCET, Tongfu |
| 광섬유 정렬(FAU)·조립 | 광섬유를 나노 단위 오차로 칩에 부착 | Teramount, Silitronics | T&S Communications, Accelink |

자료: 각 사, SEMICON China 2026, 업계 자료, 유안타증권 리서치센터 주: 유안타증권 'Semicon China 2026'(2026.3) 표를 업데이트

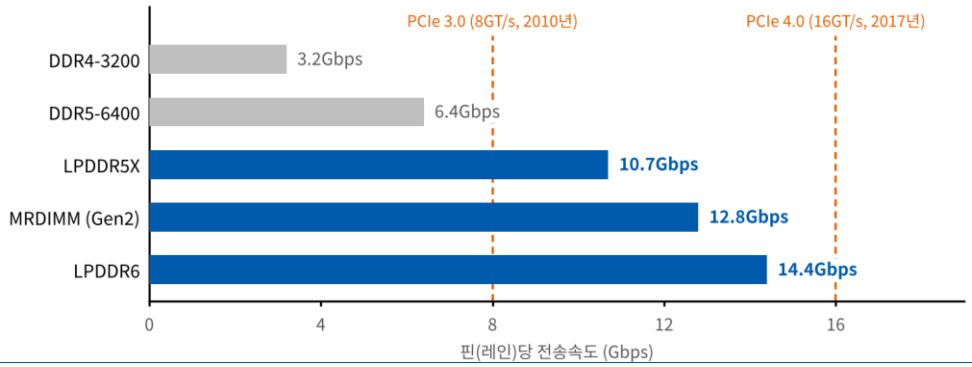
**2028년은 광 연결이 GPU 사이의 연결로까지 들어오는 해다.** NVIDIA는 2028년 Feynman 세대에서 CPO를 적용한 NVLink 스위치로 광 NVLink를 구현해, 하나로 묶을 수 있는 GPU 패키지 수(Scale-up 도메인)를 Oberon 랙 576개, Kyber 랙 1,152개까지 늘릴 계획이다. 지금까지 광 연결이 랙과 랙 사이의 Scale-out 영역에 머물렀다면, Feynman부터는 GPU끼리 직접 묶는 Scale-up 영역까지 빛으로 바뀌는 것이다. 투자전략에서 광 연결의 병목이 2028년에 정점을 맞이한다고 보는 이유이다.

## 보드 위 연결: 메모리가 찾는 해법

랙 사이의 연결이 빛으로 넘어가는 동안, 칩과 메모리를 잇는 보드 위의 짧은 구리 배선에서도 같은 문제가 나타나고 있다. 전기 신호는 기판의 구리선을 지나면서 조금씩 약해지는데, 신호가 빠를수록, 선이 길수록 더 많이 약해진다. 그래서 아주 빠른 신호가 지나가는 구간에는 일반 기판 대신 신호 손실이 적은 고가의 소재(저손실 CCL)를 쓰고 층수를 늘린다. 지금까지 이런 저손실 기판이 필요했던 곳은 서버의 네트워크(이더넷)나 GPU·SSD를 잇는 PCIe처럼, 데이터를 한 줄로 매우 빠르게 보내는 직렬 통신(SerDes) 구간이었다. 메모리는 속도가 상대적으로 느리고 CPU와의 거리도 짧아 일반 기판으로 충분했다.

MediaTek은 JEDEC 포럼에서 이 구도가 바뀌고 있다고 지적했다. LPDDR5X는 핀당 약 10Gbps, LPDDR6는 14.4Gbps, 서버용 MRDIMM은 12.8Gbps까지 빨라지며 몇 년 전 PCIe와 비슷한 속도에 이르렀다. 게다가 메모리는 수십에서 백여 개의 선을 나란히 깔아야 한다. 두 선을 짝지어 잡음을 상쇄하는 통신용 신호와 달리 선 하나로 신호를 보내기 때문에 잡음에도 약하다. 이제는 메모리 배선이 메인보드에 비싼 기판을 요구하는 주된 이유가 되고 있다는 것이다. 서버뿐 아니라 LPDDR을 쓰는 노트북과 스마트폰 보드도 예외가 아니다.

차트 23. 메모리 핀당 전송속도와 PCIe 레인 속도 비교




자료: JEDEC, PCI-SIG, MediaTek(JEDEC Industry Forum, 2026.10), 유안타증권 리서치센터

**메모리도 이 병목을 기판에만 맡겨두지 않고, 신호를 덜 약해지게 보내는 방법을 스스로 찾고 있다.** 해법은 두 가지다. 첫째는 Multi-PAM 신호 방식이다. 전압을 두 단계(0과 1)가 아니라 세 단계나 네 단계로 나눠 한 번에 더 많은 비트를 보내면, 같은 데이터 속도를 더 낮은 주파수로 낼 수 있어 기판에서의 손실이 줄어든다. 그래픽 메모리인 GDDR7이 이미 세 단계 방식(PAM3)을 쓰고 있다. 둘째는 버퍼드 메모리다. 메모리 모듈 위에 신호를 다시 정리해 내보내는 버퍼 칩을 올리면, 메인보드의 긴 배선은 수많은 DRAM 대신 버퍼 칩 하나만 상대하면 된다. 서버 RDIMM의 RCD, MRDIMM의 버퍼 칩, PC용 CUDIMM의 클럭 드라이버(CKD)가 이런 역할을 한다.

**투자 관점에서 주목할 점은 연결이라는 병목에서 메모리도 해법을 만드는 쪽에 서 있다는 것이다. 신호 방식과 모듈 구조가 바뀔수록 메모리 모듈에 들어가는 버퍼 칩, 클럭 드라이버 등 인터페이스 칩의 탑재량이 늘어나고, 이를 함께 설계해야 하는 메모리 업체의 역할도 커진다.** 저손실 기판 수요가 네트워크 장비를 넘어 메모리 채널로 넓어지는 것은 같은 변화의 다른 한 면이다.

## 4. 전력: 마지막 병목이자 새로운 반도체 수요처

**2030년을 향한 마지막 병목은 전력이다. 전력이 마지막 병목인 이유는 반도체 업체가 스스로 풀 수 없는 문제이기 때문이다.** 앞서 본 메모리와 패키징, 연결의 병목은 반도체 업체가 공장을 짓고 기술을 바꾸면 2~3년 안에 풀어갈 수 있다. 반면 데이터센터에 전기를 대려면 발전소와 송전망, 변전소가 함께 늘어나야 하고, 이를 새로 짓는 데는 길게는 10년 가까이 걸린다. 칩을 아무리 빨리 만들어도 꽂을 전기가 없으면 데이터센터를 돌릴 수 없다. Microsoft도 2025년 말 확보한 GPU 일부를 전력과 건물이 부족해 바로 설치하지 못하고 있다고 밝힌 바 있다.

숫자로 보면 격차가 분명하다. TrendForce는 전 세계 데이터센터 전력 수요가 2025년 122.9GW에서 2026년 161GW로 31% 늘고, 2030년에는 490.7GW에 이를 것으로 전망했다. 1GW가 대형 원자력 발전소 1기의 발전량과 비슷하다는 점을 감안하면, 5년 사이 원전 350기 이상에 해당하는 전력 수요가 새로 생기는 셈이다. 반면 같은 해 데이터센터가 쓸 수 있는 전력망 용량은 222.6GW로, 수요와의 격차가 268GW에 달한다. 부지 내 자가발전을 감안하면 실제 부족분은 이보다 작겠지만, 전력망 연결 지연만으로도 데이터센터 건설 속도가 제약받을 수 있다. 투자 관점에서 이는 AI 사이클의 형태를 바꾼다. **전력이 부족하면 AI 투자가 한꺼번에 몰렸다가 꺼지기보다, 전력이 확보되는 속도에 맞춰 나눠 집행된다. 단기 수요의 폭발력은 줄지만 수요의 지속 기간은 길어지는 셈이다.**

**전력은 그 자체로 반도체의 새로운 수요처가 된다.** 전기는 발전소에서 칩까지 오는 동안 여러 번 전압을 바꿔야 하고, 바꿀 때마다 일부가 열로 버려진다. 같은 전력을 보낼 때 전압이 높을수록 전류가 줄어 전선에서 버려지는 열도 줄어든다. 랙 하나의 전력이 수백 kW를 넘어서면서 데이터센터가 랙 내부 전압을 기존 48~54V에서 800V 직류(800VDC)로 높이기 시작한 이유다. **전압이 높아질수록 높은 전압을 견디고 전환 손실이 작은 SiC(탄화규소)와 GaN(질화갈륨) 전력반도체의 역할이 커진다.** onsemi는 11월 3분기 실적발표에서 AI 데이터센터 매출이 전년 대비 175% 늘었고, 2026년 약 6.2억달러에서 2027년 두 배로 늘어날 것으로 전망했으며, 2030년 목표를 30억달러 이상으로 높였다. 전원 공급 장치부터 프로세서 바로 앞의 저전압 변환까지 전력 경로 전체에서 반도체 탑재량이 늘어나는 구조다.

**전력반도체 수혜는 전기가 지나가는 경로를 따라 세 단계로 나눠 볼 수 있다.** 첫 단계는 랙 입구에서 전력망의 교류를 800V 직류로 바꾸는 구간이다. 높은 전압을 견뎌야 해 SiC가 주로 쓰이며, Infineon, onsemi, STMicroelectronics, ROHM이 주요 공급사다. 두 번째는 800V를 보드 위에서 50V, 12V, 6V로 낮추는 구간이다. 작은 공간에서 빠르게 전환해야 해 GaN의 역할이 커지며, Navitas, Texas Instruments, STMicroelectronics가 GTC 2026에서 800V를 6V로 바로 낮추는 컨버터를 공개했다. 마지막은 GPU 바로 앞에서 전압을 1V 이하로 낮춰 공급하는 구간으로, 실리콘 기반 전원 IC를 만드는 MPS, Infineon, Renesas 등이 맡는다. NVIDIA가 2025년 10월 공개한 800VDC 칩 파트너만 14곳에 이를 만큼, 전력 경로 전체에서 반도체 공급망이 새로 짜이고 있다.

**관전 포인트는 세 가지다.** 첫째, 실적은 GPU 바로 앞 단계에서 먼저 확인되고 있다. MPS의 3분기 데이터센터 부문 매출은 5.07억달러로 전년 대비 165% 늘었다. 반면 800V 관련 매출은 아직 초기 단계로, 800VDC 대표 수혜주로 꼽히는 Navitas의 3분기 매출은 1,400만달러에 불과하다. 둘째, 800VDC의 실제 채택 속도다. NVIDIA는 Kyber 랙이 출하되는 2027년부터 800VDC 채택이 본격화될 것으로 보고 있어, 2027년이 SiC·GaN 업체 실적의 분기점이 될 전망이다. 셋째, 전기차 수요 둔화로 부진했던 SiC 업체들에게 AI 데이터센터가 대체 수요가 될 수 있는지다. Infineon은 올해 2월 AI 관련 매출이 FY2026 15억유로에서 FY2027 25억유로로 늘어날 것으로 제시했다. 다만 800VDC 생태계에 참여한 한국 기업은 아직 없고 직접 고객도 Delta, Lite-On 등 대만 전원장치 업체여서, 국내 업체의 직접 수혜는 제한적이다.

표20. 데이터센터 전력 경로별 주요 전력반도체 업체

| 단계 | 역할과 주력 소재 | 주요 업체 | 관전 포인트 |
|---|---|---|---|
| ① 랙 입구 (AC → 800VDC) | 전력망의 교류를 고압 직류로 변환. 고전압을 견디는 SiC 중심 | Infineon, onsemi, STMicroelectronics, ROHM / 전원장치: Delta, Lite-On | 전기차 둔화로 부진한 SiC 수요를 AI 데이터센터가 얼마나 대체하는지 |
| ② 보드 변환 (800V → 50·12·6V) | 고압 직류를 보드 위에서 낮춤. 작고 빠른 GaN 의 역할 확대 | Navitas, Texas Instruments, STMicroelectronics, Infineon, Innoscience, EPC | 2027년 Kyber 랙 출하와 함께 800VDC 채택이 일정대로 늘어나는지 |
| ③ GPU 바로 앞 (→ 1V 이하) | GPU 가 쓰는 저전압으로 최종 변환. 실리콘 전원 IC | MPS, Infineon, Renesas, ADI, AOS, Richtek | 현재 실적이 가장 먼저 확인되는 구간(MPS 데이터센터 매출 +165% YoY) |

자료: NVIDIA(OCP Global Summit 2025), 각 사, The Elec, 유안타증권 리서치센터

표21. 2027~2030년 병목 이동 타임라인

| 시점 | 주요 기술·제품 변화 | 병목·수혜 영역 |
|---|---|---|
| 2027 | Rubin Ultra·Kyber 랙, HBM4E, N2 생산능력 2배 이상, CoWoS 월 18만장 | 메모리 공급, 첨단 패키징, 테스트 |
| 2028 | Feynman(다이 적층·Custom HBM·광 NVLink), TSMC A14 양산, N2P 적층 | 광 연결(CPO), 3D 적층·하이브리드 본딩(로직) |
| 2029 | Intel 14A·삼성 SF1.4 양산(High-NA EUV), SoIC 4.5㎛, A14 적층 | High-NA EUV 밸류체인, 신규 메모리 팹 가동 시작 |
| 2030 | TSMC High-NA EUV 도입, HBM 하이브리드 본딩 본격화, 데이터센터 전력 490GW | 전력(800VDC·SiC·GaN), 메모리 공급 증가 본격화 |

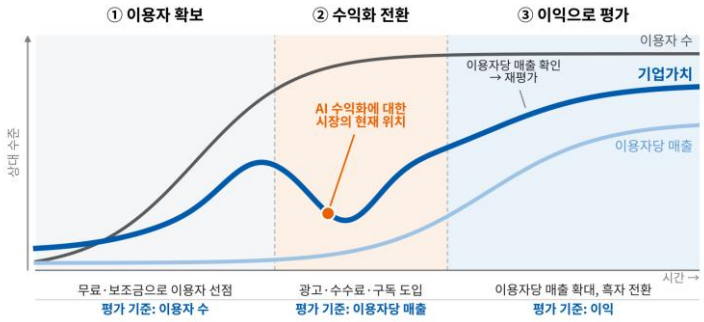
자료: TSMC, NVIDIA, Intel, 삼성전자, SK 하이닉스, TrendForce, Tom's Hardware, 유안타증권 리서치센터

# 반도체 업종 투자전략

2027년 반도체 업종을 바라보는 시장의 시선에는 기대와 불안이 함께 있다. 2026년 업종을 끌어올린 호재는 틀리지 않았다. 메모리 공급 부족은 실적으로 확인됐고, AI 투자는 계획보다 커졌다. 다만 연초 이후 11월 6일까지 샌디스크가 203% 오르는 동안 엔비디아는 1% 하락했을 만큼 같은 AI 밸류체인 안에서도 쏠림이 극단적이었고, 하반기에는 하이퍼스케일러의 현금흐름과 자금 조달, HBM 디스펙, 공급 과잉 우려가 번갈아 부각되며 업종 변동성을 키웠다.

당사는 2027년을 피크아웃이 아니라, 시장이 불편해하는 우려가 하나씩 걷히면서 업종이 균형을 맞춰 우상향하는 해로 전망한다. **시장이 가장 불편해하는 것은 AI 수익화이며, 당사 리서치센터는 그 경로가 인터넷 플랫폼이 걸어온 길과 유사할 것으로 예상한다.** 플랫폼은 무료로 이용자를 먼저 확보한 뒤 수익화에 나섰고, 평가 기준도 이용자 수에서 이용자당 매출, 이익 순으로 옮겨갔다. 수익화로 넘어가는 구간에서는 의심이 주가의 변동성을 야기했다. 과거 Facebook은 2012년 상장 직후 모바일 수익화 의심으로 주가가 절반 수준까지 하락했지만, 모바일 광고 매출이 확인되자 1년여 만에 상장가를 회복하고 기업가치는 우상향했다. **AI 수익화에 대한 시장의 현재 위치는 이 의심 구간의 후반부로 판단한다.** 과거와 다른 점은 반도체다. 과거 플랫폼 기업이 급락하던 시기에 반도체는 스마트폰과 메모리 수급에 따라 움직였지만, 지금은 AI 투자가 반도체 수요의 중심에 있어 의심도 확인도 반도체 주가에 직접 반영된다. **수익화가 숫자로 확인된다면 빅테크와 반도체 주가는 정상화를 넘어 재평가 국면에 들어설 수 있을 것이다.**

그림 16. 플랫폼 수익화 단계와 AI 수익화에 대한 시장의 현재 위치




자료: 유안타증권 리서치센터 주: 인터넷 플랫폼의 수익화 경로를 단순화한 개념도

수익화가 확인되어 빅테크와 반도체 주가가 정상화된다면, 그다음 질문은 어느 시장을 봐야 하느냐다. 당사는 메모리를 가장 먼저 점검한다. 메모리는 AI 반도체 안에서 입지와 중요도가 구조적으로 높아지고 있으며, 물량 확보 자체가 경쟁력이 되는 전략 자산이기 때문이다. 이에 피크아웃 우려를 먼저 짚은 뒤 메모리 수급을 살펴보고, 메모리를 넘어선 투자 기회는 병목 대응 전략으로 정리한다.

## 1. 리스크 포인트 점검: 하나씩 걷히는 피크아웃 우려

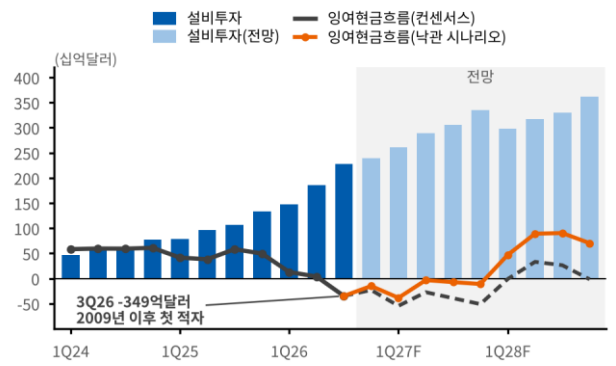
표22. 피크아웃 우려 점검표

| 시장의 우려 | 당사 판단 | 근거 | 점검할 신호 |
|---|---|---|---|
| 빅테크 설비투자 둔화 (현금흐름 악화) | 2027년 투자 확대 지속. 수익화 경로 확인이 관건 | 주요 6개사 설비투자 2026년 8,700억달러 → 2027년 1.3조달러 전망, 에이전트 커머스의 거래 수수료 모델 등장 | CSP 의 2027년 설비투자 가이던스 하향, 에이전트 서비스 확대 지연 |
| AI 자금 조달 경색 (네오클라우드 신용·순환금융) | 수요가 아닌 자금 조달 주체의 문제. 하이퍼스케일러와 계약 수요 중심으로 재편 | AI 관련 하이일드 스프레드 360bp 확대 vs AI 외 2bp, 엔비디아 데이터센터 보증·약정 최대 1,645억달러 | 하이퍼스케일러 신용 스프레드 확대, 프론티어 모델 기업 상장 부진, 엔비디아 약정의 실제 지급 확대 |
| HBM 디스펙 (Rubin Ultra 단수 하향) | 수요 둔화가 아닌 공급 제약 아래의 최적화 | 8단 HBM4 2Q27 양산 목표, HBM 비트 출하 +50~60%에도 수요 미충족 전망, 2027년 HBM ASP +121% 전망 | 가속기 출하 계획 자체의 하향, HBM 계약가격 상승률 둔화 |
| 소비자 기기 수요 위축 | 범용 메모리 일부에 부담이나 서버 수요가 상쇄. | 노트북·스마트폰 원가 압박과 PC SSD 용량 축소 vs 서버 DDR5·eSSD 강세 | 서버 DRAM 계약가격 상승률 둔화, 모바일·PC 재고 증가 |
| 신규 증설에 따른 공급 과잉 | 2027년은 해당 없음. 2029년 이후 점검 대상 | SK 하이닉스 신규 팹 2028년 12월·2029년 6월 가동 예정, 2027~28년 생산능력 50% 이상 장기계약 | 증설 일정 앞당김, 장기계약 조건 재협상 |
| 중국 메모리 증설(CXMT·YMTC 상장) | DRAM 은 2027년까지 영향 제한적, NAND 는 2028년 이후 변수 | CXMT 기술 17.5nm 정체·증설분 HBM 과 내수 집중, YMTC 캐파 월 20만장 → 27만장 | CXMT 범용 DRAM 수율 개선과 해외 출하 확대, YMTC 의 eSSD 시장 진입 |
| 모델 효율화에 따른 하드웨어 수요 감소 | 효율 개선은 사용자 확대로 연결(제번스 역설) | Meta Muse 1년 내 효율 10배 이상 목표 vs 사용자 1,000만명 → 10억명 이상 목표 | 토큰 사용량 증가율 둔화, 에이전트 서비스 사용자 정체 |

자료: 각 사, S&P Global, TrendForce, 유안타증권 리서치센터 주: 당사 판단

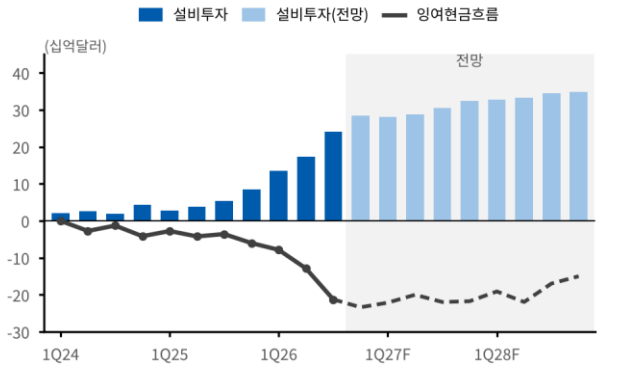
첫째, 빅테크 설비투자의 지속 가능성이다. 시장이 가장 먼저 묻는 것은 현금흐름이 악화되는 가운데 투자가 계속될 수 있느냐다. S&P Global 전망 기준 주요 6개 기업의 설비투자는 2026년 8,700억달러에서 2027년 1.3조달러로 늘어나지만, 같은 해 잉여현금흐름이 플러스로 예상되는 곳은 한 곳에 불과하다. 미국 5대 하이퍼스케일러의 합산 잉여현금흐름은 3Q26에 349억달러 적자를 기록해 2009년 이후 처음 적자로 돌아섰다. 그러나 투자를 회수할 경로는 빠르게 넓어지고 있다. Anthropic의 상반기 매출 중 80%가 사용량 기반이고 매출총이익률은 85%까지 높아졌으며, 광고와 에이전트 커머스의 거래 수수료, 소버린 AI까지 토큰이 매출이 되는 경로가 여섯 갈래로 열리고 있다. 2027년에도 투자 확대는 이어지되, 관건은 투자 규모보다 수익화 경로의 확인이다.

차트 24. 미국 5 대 하이퍼스케일러 설비투자와 잉여현금흐름




자료: Bloomberg, 유안타증권 리서치센터 주: Amazon · Microsoft · Alphabet · Meta · Oracle 합산

차트 25. 네오클라우드 · AI 데이터센터 8 사 설비투자와 잉여현금흐름

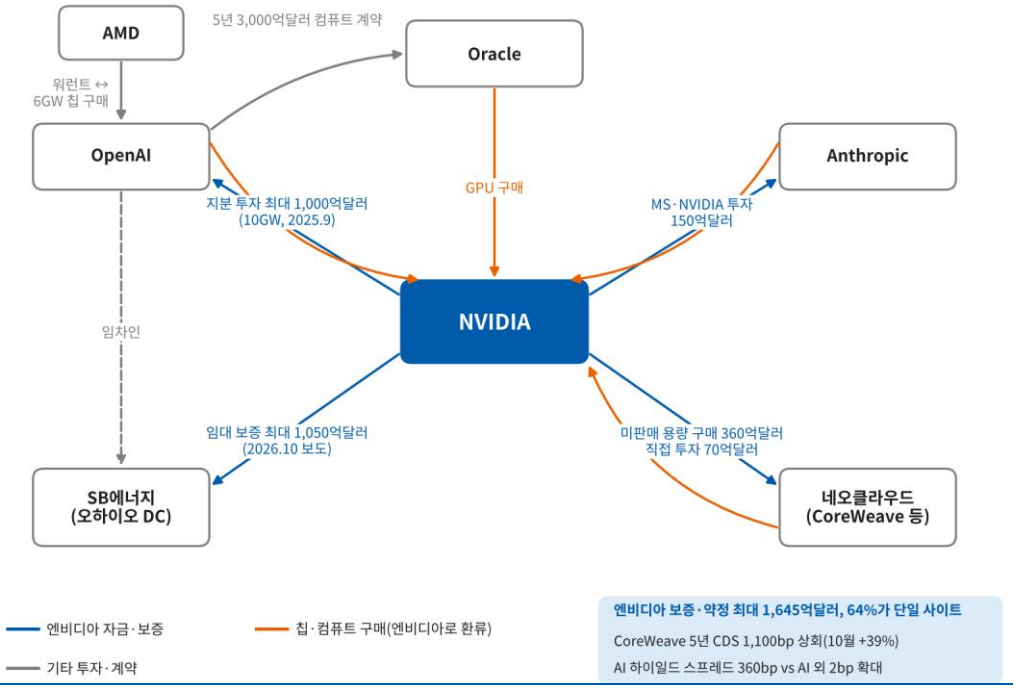



자료: Bloomberg, 유안타증권 리서치센터 주: CoreWeave · Nebius · IREN · Applied Digital · Cipher · TeraWulf · Core Scientific · Hut 8 합산

둘째, AI 투자의 자금 조달이다. 10월에는 부채로 GPU를 확보해 온 네오클라우드의 신용 위험이 부각됐다. CoreWeave의 5년물 CDS 스프레드는 10월에만 39% 올라 1,100bp를 넘어섰고, 메타·오라클·엔비디아의 CDS 스프레드도 사상 최고 수준을 기록했다. 엔비디아가 고객사의 데이터센터 임차와 캐파 구매를 보증하거나 약정한 규모가 최대 1,645억달러에 이르고, 그중 1,050억달러가 단일 사이트에 몰려 있다는 보도가 나오면서 '순환금융' 논란도 커졌다. 그러나 Goldman Sachs에 따르면 AI 관련 하이일드 스프레드가 360bp까지 벌어지는 동안 AI와 무관한 하이일드 스프레드는 2bp 확대에 그쳤다. 신용시장 전체가 아니라 부채 의존도가 높은 일부 고리의 문제라는 뜻이다.

쟁점은 컴퓨트 수요가 아니라 '누가 돈을 대느냐'이며, 투자는 현금 창출력이 큰 하이퍼스케일러와 계약 수요가 확인된 사업자 중심으로 재편될 가능성이 높다. 재편 과정에서 일부 고리의 변동성은 불가피하지만, 이는 수요의 붕괴가 아니라 투자 주체의 교체에 가까울 것이다. 엔비디아가 SK하이닉스와의 공급 계약을 포함해 최대 1,600억달러 규모의 메모리 조달을 약정한 것은 이런 환경에서도 메모리 수요의 가시성이 유지되고 있음을 보여준다. 연말로 예정된 Anthropic 등 프론티어 모델 기업의 상장은 AI 투자의 자금 조달 구조가 시장에서 검증받는 첫 번째 시험대가 될 것이다. 하이퍼스케일러의 신용 스프레드 확대와 프론티어 모델 기업의 상장 부진, 엔비디아 약정의 실제 지급 확대가 점검할 신호다.

그림 17. AI 투자의 순환금융 구조




자료: The Information(2026.10), 각 사, Goldman Sachs, Bloomberg, 언론 보도 종합, 유안타증권 리서치센터 주: 금액은 발표 · 보도 기준 최대치

셋째, HBM 사양 하향(디스펙)이다. NVIDIA가 Rubin Ultra에서 8단과 12단 HBM을 혼용하는 방안을 검토한다는 소식은 HBM 수요 둔화로 읽혔다. 그러나 배경은 수요 둔화가 아니라 HBM 확보의 어려움이다. GPU 하나에 들어가는 HBM을 줄여야 한정된 물량으로 더 많은 GPU를 만들 수 있기 때문이다. 당사는 2027년 HBM 공급 Bit 출하가 50~60% 늘어나더라도 수요를 충족하기 어렵고, HBM 평균판매가격은 121% 오를 것으로 전망한다. 물량을 줄인 것이 아니라 공급 제약 아래에서 설계를 바꾼 최적화로 봐야 한다.

넷째, 소비자 기기 수요 위축이다. 메모리 원가 상승으로 노트북과 스마트폰의 원가 압박이 커졌고, PC에서는 SSD 용량을 줄이는 사례도 나타나며 2026년 PC 출하는 큰 폭으로 역성장했다. 그러나 2027년을 들어가는 시점에서 보면 이는 오히려 기저가 낮아졌다는 의미다. 이미 낮아진 출하 눈높이 위에서 On-premise AI PC로의 Mix 개선과 지연된 Commercial PC 교체 수요가 맞물리고, 프리미엄 스마트폰은 메모리 원가 부담에도 불구하고 탑재량을 12~16GB 이상으로 늘리고 있다. 수요의 무게중심이 서버로 옮겨가 서버 DDR5와 eSSD의 강세가 범용 메모리를 받쳐주는 가운데, 소비자 기기에서는 낮아진 기대를 웃도는 서프라이즈가 나올 가능성도 배제할 수 없다.

다섯째, 신규 증설에 따른 공급 과잉이다. 메모리 업체들의 투자가 가속하면서 2028년 이후의 공급 과잉을 걱정하는 시각이 있다. 그러나 신규 팹의 가동 일정을 따져보면 공급 과잉의 시점은 시장의 우려보다 뒤로 밀려 있다. SK하이닉스의 청주 M17과 용인 Y2는 각각 2028년 12월과 2029년 6월에야 첫 클린룸이 가동되고, 삼성전자의 평택 P5 1라인도 2027년 하반기 장비 반입을 거쳐 2028년에 가동을 시작한다. 삼성전자 용인 클러스터의 첫 팹은 2029년 가동이 목표다. 클린룸이 열린 뒤 장비를 채우고 수율을 끌어올려 의미 있는 비트 공급으로 이어지기까지 통상 1년 이상이 걸린다는 점을 감안하면, 신규 팹의 공급 효과는 2029년 이후에야 본격화된다. 2027~2028년 생산능력의 50% 이상이 이미 장기공급계약으로 묶여 있다는 점도 공급 과잉이 2027년의 우려가 아닌 이유다.

공급이 늘어나는 시점에도 과거처럼 쉽게 공급 과잉으로 넘어가지 않을 장치가 있다. 첫째는 HBM의 웨이퍼 잠식 속도다. HBM은 범용 DRAM 대비 3~4배의 웨이퍼를 쓰고, HBM4E와 Custom HBM으로 갈수록 그 부담은 더 커진다. 업계 DRAM 웨이퍼 캐파에서 HBM이 차지하는 비중이 2027년 30%에 가까워지는 만큼, 신규 팹의 상당 부분은 범용 비트 증가가 아니라 HBM 전환에 흡수된다. 둘째는 Non-HBM 제품의 다변화다. 서버 DDR5와 MRDIMM, SOCAMM2, LPDDR6, eSSD처럼 용도별로 제품이 갈라지면서, 메모리 공급은 하나의 범용 제품을 두고 가격과 물량을 다투던 구조에서 고객별·용도별로 나뉘는 구조로 바뀌고 있다. 생산능력이 여러 제품으로 분산되고 장기계약 비중이 높아질수록, 한 제품의 과잉이 업황 전체의 가격 급락으로 번지던 과거의 사이클은 반복되기 어렵다. 공급 과잉은 2029년 이후에나 점검할 대상이며, 그 시점에도 과거와 같은 급격한 하락보다는 완만한 조정에 가까울 것으로 판단한다.

여섯째, 중국 메모리 업체의 증설이다. 당사는 3월 'Semicon China 2026' 보고서에서 CXMT와 YMTC의 상장이 '제2의 치킨게임'의 신호가 아니라, 중국 내수 AI 인프라에 공급할 안정적인 물량을 확보하기 위한 '내수 안전판'의 성격이 강하다고 판단한 바 있다. 한국 업체들이 HBM과 서버용 DDR5 등 고부가 제품으로 생산능력을 옮기면서 생긴 중국 내수 범용 메모리 시장을 중국 업체들이 흡수하고, 그 결과 한국과 중국 업체 간의 직접적인 가격 충돌 구간은 오히려 줄어드는 '시장의 구조적 이원화'가 진행되고 있다는 판단도 유효하다. 한편 CXMT의 DRAM은 17.5nm 수준에 머물러 있고 신규 생산능력의 상당 부분이 HBM에 투입되고 있어, 글로벌 시장에 대한 실효 공급은 제한적일 것으로 추정한다.

NAND에서는 YMTC가 실질적인 변수다. 자체 Xtacking 기술로 장비 규제를 우회해 256단 제품을 생산하고 있으며, 상장 자금을 기반으로 330억위안을 조달할 계획인 것으로 파악된다. 채널체크에 따르면 생산능력은 2026년 말 월 20만장에서 2027년 말 27만장으로 늘고 이후에도 매년 6~7만장씩 증가할 예정이다. 여전히 중국 내수 스마트폰 위주이고 Enterprise SSD 시장의 진입 장벽은 높다는 점을 재차 강조한다.

일곱째, 모델 효율화에 따른 하드웨어 수요 감소다. 모델이 효율화되면 같은 일을 하는 데 필요한 컴퓨트가 줄어든다는 우려다. 그러나 효율 개선은 수요를 줄이기보다 서비스 대상을 넓히는 데 쓰인다. 메타는 Muse의 효율을 1년 안에 10배 이상 높이겠다고 밝혔지만, 동시에 사용자를 1,000만명에서 10억명 이상으로 늘리겠다는 목표를 제시했다. 사용자가 100배 늘어나는 동안 효율 개선이 10배에 그친다면 필요한 컴퓨트는 오히려 약 10배 늘어난다. 토큰 단가가 떨어질수록 사용량이 늘어나는 제번스 역설이다.

## 2. 메모리: 비중 확대 유지, 늘어나는 토큰과 모자란 웨이퍼

### 수요: 토큰을 메모리로 환산한다

메모리 수요를 보는 잣대는 AI 설비투자에서 토큰으로 바뀌고 있다. AI 투자가 토큰 매출로 회수되는 구조라면 메모리 수요도 토큰이 얼마나 늘고, 얼마에 팔리며, 그중 얼마가 Decoding에서 처리되는지로 설명해야 한다. 세 변수는 각각 메모리의 다른 부분을 움직인다.

출력 토큰 증가율은 HBM의 대역폭과 용량을 움직인다. Decoding은 토큰 하나를 만들 때마다 모델 가중치와 지금까지의 대화 맥락(KV 캐시)을 메모리에서 다시 읽어야 하는 구조여서, 출력 토큰이 늘면 연산보다 메모리 대역폭 수요가 먼저 늘어난다. 요청당 출력 토큰은 2024년에서 2025년 사이 약 3배로 늘었고, 에이전트는 하나의 요청을 처리하기 위해 수십 번의 추론과 도구 호출을 반복한다. 구글이 처리하는 토큰이 1년 만에 월 480조 개에서 3,200조 개로 늘어난 것은 이 변화가 이미 물량으로 나타나고 있음을 보여준다.

Decoding 비중은 메모리의 계층을 넓힌다. 대화가 길어지고 에이전트가 맥락을 오래 들고 있을수록 KV 캐시는 HBM에 다 담기지 않고, 서버 DRAM과 SOCAMM 같은 LPDDR 모듈, 나아가 Enterprise SSD로 내려간다. 채널체크에 따르면 미국 CSP의 Enterprise SSD 추가 주문 가운데 70~80%가 긴 대화 기록과 Context 저장 등 AI 실사용에서 비롯됐다. 토큰의 성격이 바뀌면 HBM뿐 아니라 서버 DRAM과 NAND까지 같은 논리로 수요가 늘어나는 중이다. 채널체크에 따르면 미국 CSP들은 기존 장기공급계약 물량에 더해 Enterprise SSD 구매량을 두 배로 늘리고 있다. Enterprise SSD 내 QLC 비중도 2026년 약 30%에서 2027년 50% 이상으로 높아질 전망이며, QLC 비중이 91%에 달하는 SK그룹의 Solidigm의 수혜 강도가 가장 높을 전망이다.

**토큰당 단가는 수요의 지속성과 가격 결정력을 정한다.** 제번스의 역설에 따르면 모델 효율화로 토큰 단가가 떨어지면 사용량이 늘어 총 메모리 수요는 오히려 커진다. 동시에 추론 원가에서 메모리가 차지하는 비중이 높아지면서, 고객은 가격보다 물량 확보를 우선하게 된다. 실제로 DRAM 출하에서 장기공급계약이 차지하는 비중은 40%에서 60%로 높아지고 있다. **중장기적 관점에서 토큰 판매가 이익을 내기 시작하면 메모리 가격 상승을 흡수할 여력이 생길 것으로 예상한다.**

### 공급: HBM 고도화가 늦추는 공급 정상화

**2027년 DRAM 공급 측면에서 가장 중요한 변수는 HBM의 웨이퍼 잠식이다.** 삼성전자는 HBM이 업계 DRAM 웨이퍼 캐파에서 차지하는 비중이 현재 약 20%에서 2027년 30% 가까이로 높아질 것으로 전망했다. HBM은 만드는 데 범용 DRAM 대비 약 3~4배의 웨이퍼 면적을 사용하기 때문에, HBM 비중 확대는 그 이상의 속도로 범용 DRAM의 공급 여력을 제한한다. **삼성전자의 월평균 HBM 웨이퍼 투입량은 올해 약 17만장에서 2027년 약 27만장으로 60% 가까이 늘어나고, 2027년 HBM 출하에서 HBM4 계열의 비중은 약 80%에 이를 전망이다.** 다이 크기가 더 큰 HBM4로의 전환이 빨라질수록 범용 DRAM에 돌아갈 웨이퍼는 그만큼 줄어든다.

**HBM4부터 베이스다이를 파운드리에서 생산하더라도 DRAM 웨이퍼 부담은 크게 줄지 않는다.** HBM3E까지는 메모리 업체가 HBM 맨 아래의 베이스다이도 자사 DRAM 공정에서 만들었지만, HBM4부터는 SK하이닉스가 TSMC에, 삼성전자가 자사 파운드리 공정에 베이스다이 생산을 맡긴다. 마이크론은 HBM4까지 자체 공정을 유지하고 HBM4E부터 TSMC와 협력할 계획이다.

그러나 12단 HBM 한 스택은 코어 다이 12장과 베이스다이 1장으로 이뤄지기 때문에, 베이스다이를 외부로 넘겨 DRAM 웨이퍼에서 덜어내는 몫은 다이 수 기준으로 많아야 13장 중 1장, 약 7~8%에 그친다. HBM이 웨이퍼를 많이 쓰는 이유는 베이스다이가 아니라 위에 쌓는 코어 다이에 있다. 코어 다이에는 층간 연결을 위한 TSV(실리콘 관통 전극)와 그 주변 보호 영역이 들어가 같은 용량의 범용 DRAM 다이보다 면적이 크고, HBM4에서는 I/O가 1,024개에서 2,048개로 늘어나 TSV도 함께 늘어난다. 여러 장을 쌓는 과정에서 생기는 수율 손실도 정상 제품 하나를 만드는 데 필요한 웨이퍼를 늘린다. **Base die의 외주화는 오히려 HBM 공급을 DRAM 웨이퍼와 파운드리 생산능력이라는 두 가지 제약에 동시에 묶는다는 점에서, 병목을 덜어내기보다 하나 더하는 변화에 가깝다는 판단이다.**

**HBM 고도화는 범용 DRAM의 공급 제약을 장기화시킨다.** HBM4E와 Custom HBM으로 갈수록 선단 공정 적용과 후공정 난이도가 높아지면서 동일 비트 생산에 필요한 웨이퍼와 제조 자원이 늘어난다. HBM 생산 확대는 서버 DDR5와 LPDDR 등 범용 DRAM의 실질 공급 여력을 지속적으로 잠식하며, 신규 생산능력이 가동되더라도 상당 부분을 흡수할 가능성이 높다는 판단이다.

### 가격: 우상향 트렌드는 지속될 전망

2027년 메모리 가격은 상승 폭이 줄더라도 상승 흐름 자체는 이어질 전망이다. 범용 DRAM 평균 기준으로 분기마다 한 자릿수 상승이 이어지고, 장기공급계약(LTA) 가격은 기존보다 20~30% 이상 높은 수준에서 협상되고 있는 것으로 파악된다.

**제품별로는 온도 차이가 있다. 1H27 예상 가격 기준으로 가장 강한 것으로 전망되는 제품은 LPDDR이다. 채널체크에 따르면 LPDDR 계약가격은 2027년 1분기에 약 20% 이상 오를 전망인데, LPDDR이 SOCAMM2 형태로 서버에 들어가면서 스마트폰과 PC, 서버가 같은 물량을 두고 경쟁하게 될 것이기 때문이다.** HBM도 HBM4 비중 확대와 함께 단가가 오르고 있다. 반면 NAND는 Enterprise SSD 중심의 수요가 이어지지만, 2026년 하반기 급등에 따른 부담과 소비자용 수요 부진으로 상반기 상승 폭이 다소 둔화될 전망이다.

### 투자의견: 비중 확대를 유지하는 세 가지 근거

**첫째, 2028년은 AI 반도체 수요의 정점이 아니라 기준선이 한 단계 높아지는 해다.** 시장의 가장 큰 우려는 2027년까지 이어진 메모리 가격 상승과 설비투자 확대가 2028년에 꺾이는 시나리오다. 일각에서, 메모리 공급업체의 DRAM 평균판매가격이 2027년 말 고점을 찍은 뒤 2028년 고점 대비 약 53% 하락할 것으로 제시한 바 있다. **그러나 그와 같은 최악의 추정에서 2028년 가격은 여전히 2025년의 약 두 배 이상의 수준이다. 공급은 절제된 속도로 따라오고, 수요처는 세대가 바뀔 때마다 커지기 때문이다.**

**실제 메모리 공급업체들의 CapEx discipline 의지는 강하다.** SK하이닉스는 3분기 실적발표에서 메모리 공급 부족이 2030년까지 이어질 것으로 전망했고, 삼성전자도 최소 2028년까지로 봤다. 동시에 최태원 SK그룹 회장이 밝힌 향후 5년간 웨이퍼 생산능력 두 배 확대는 연평균 15% 안팎의 속도이며, 채널체크에 따르면 SK하이닉스는 HBM 수요의 60~70%만 충족하는 수준으로 증설 속도를 조절할 가능성도 열려 있다. TSMC 역시 여러 수요 시나리오를 바탕으로 생산능력을 관리하고 있는 것으로 파악된다. 앞서 살펴본 신규 팹의 가동 일정까지 감안하면, 의미 있는 공급 증가는 2029년 이후에 집중된다.

**수요처의 세대 교체도 매년 이어진다.** NVIDIA는 2027년 Rubin Ultra와 144개 GPU를 묶는 Kyber 랙, 2028년 Feynman과 자체 CPU인 Rosa를 출시할 계획이다. Feynman은 GPU 다이 적층, Custom HBM, 광 NVLink를 동시에 도입하는 세대로, 메모리와 패키징, 광 연결의 수요를 한꺼번에 키운다. 세대가 바뀔 때마다 칩 하나에 들어가는 메모리와 패키징 면적, 연결 대역폭이 늘어나는 구조에서는 출하 대수가 정체되더라도 반도체 수요의 기준선은 올라간다. **새로운 수요처도 등장한다.** 업계에 따르면 2028년 이후 본격화될 휴머노이드 로봇은 대당 100GB 이상의 DRAM·NAND를 탑재해 모바일 기기의 10배에 이를 것으로 예상되고, NAND 기반 고대역폭 메모리인 HBF(High Bandwidth Flash)는 스마트폰에서 먼저 도입될 가능성이 거론된다.

**둘째, 메모리는 사이클이 아니라 계약의 산업으로 바뀌고 있다.** 가격이 분기마다 시장에서 정해지던 산업에서, 수년 단위의 계약으로 물량과 가격이 미리 정해지는 산업으로 바뀌는 것이다. 가격 결정의 중심이 분기 단위 고정거래가격에서 장기공급계약으로 옮겨가면서, 이익의 수준만큼이나 이익의 변동성이 낮아진다는 점이 2027년 메모리 투자의 핵심이다. 메모리 업체의 가격 결정력은 범용 제품의 공급 부족 프리미엄에서 고객별 맞춤 설계에 참여하는 데서 오는 프리미엄으로 옮겨가는 중이며, Logic 베이스다이 역량과 파운드리 협업 구조가 업체 간 차별화 요인이 될 것이다.

**셋째, 이익이 주주에게 돌아오는 구조가 만들어지고 있다.** 이익이 커진 만큼 이를 어떻게 나누느냐가 새로운 변수로 떠올랐는데, 먼저 노사 간 성과급 배분이다. SK하이닉스는 2025년 영업이익의 10%를 성과급(PS) 재원으로 하고 상한을 없애는 제도를 10년간 유지하기로 합의했고, 삼성전자도 2026년 5월 총파업 직전 DS부문 영업이익에 연동한 특별성과급을 포함해 약 12% 수준의 성과급을 10년간 제도화하는 데 합의했다. 영업이익의 10% 안팎이 매년 인건비로 나가는 구조가 고정된 셈이다. **다만 두 회사 모두 성과급의 상당 부분을 자사주로 지급하는 방향으로 바꾸고 있다는 점에 주목한다.** 삼성전자는 특별성과급 전액을 세후 자사주로, SK하이닉스는 성과급의 50~60%를 자사주로 지급하기로 했다. **현금 유출을 줄이는 동시에 임직원과 주주의 이해를 같은 방향으로 맞추는 장치다.**

주주의 몫도 늘어날 것으로 기대한다. 삼성전자는 '27년 CAPEX가 108.0조원으로 올해(84.2조원) 대비 큰 폭 증가가 예상되고 파운드리 비즈니스를 보유하고 있어 Capital intensity가 높다는 점을 고려해야 하지만, 메모리 비즈니스 기반한 추가 FCF 여력을 바탕으로 '27년 1월 말 발표될 차기 주주환원 정책이 추가 모멘텀이 될 것으로 기대한다. SK하이닉스 역시 2027년 예상 영업이익이 전년비 2배 가까이 증가할 전망인 만큼 환원 규모의 추가 확대 여력은 충분하다. 현재 주주환원 정책은 2027년까지라는 점을 고려하면 차기 주주환원 정책에 대한 개선된 방안도 기대해본다.

표23. 국내 메모리 2사의 이익 배분: 주주환원과 성과급

| 구분 | 삼성전자 | SK 하이닉스 |
|---|---|---|
| 주주환원 기준 | 2024~2026년 잉여현금흐름의 50% | 2025~2027년 누적 잉여현금흐름의 50% 이상 |
| 2026년 환원 규모 | 약 90~110조원 | 약 100조원(현금배당 30조원+, 자사주 매입·소각 60조원+) |
| 3분기 현금배당 | 약 30조원 | 주당 2만원, 약 14조원(특별배당 포함) |
| 기타 | 잔여 환원 방식은 2027년 1월 확정, 임직원 보상용 자사주 약 15조원 별도 | Kioxia 매각 차익 약 20조원 특별배당, Solidigm 미국 상장 검토 중단(10월 23일) |
| 성과급 제도 | DS 부문 영업이익 연동 성과급 약 12%, 10년 제도화(2026.5) | 영업이익의 10% PS, 상한 폐지, 10년 유지(2025) |
| 성과급 지급 방식 | 특별성과급 전액 세후 자사주 | 현금 40%, 자사주 60% |

자료: 각 사, 언론 보도 종합, 유안타증권 리서치센터

## 3. 병목 대응 전략: 대체 수요와 가격 결정력

병목이 확산되고 있다. 투자 기회는 병목에 대한 고객의 대응에서 나온다. 병목이 심해지면 고객은 높은 가격을 감수하거나 대안을 채택한다. 당사는 병목 대응 전략을 두 갈래로 제시한다. 첫째는 병목을 피하려는 고객이 대안을 채택하면서 생기는 '대체 수요'이고, 둘째는 대안이 없어 공급자가 단가를 올릴 수 있는 '가격 결정력'이다. **두 갈래를 가르는 기준은 우회 가능성이다.**

2027년 가장 심한 병목은 메모리 생산능력이며, 첨단 패키징과 광 연결이 뒤를 잇는다. 당사는 병목별 강도를 수급 격차, 증설을 통한 해소까지의 리드타임, 우회 가능성, 2027년 이익 영향의 네 가지 기준으로 비교했다.

메모리 생산능력은 신규 팹의 의미 있는 가동이 2028~2029년에 집중되고 LPDDR·PIM 같은 우회 수단도 2027년 안에는 규모가 제한적이어서, 네 가지 기준 모두에서 강도가 가장 높다. 첨단 패키징은 수급 격차는 메모리 못지않지만, OSAT 외주로 일부를 분산할 수 있고 증설 리드타임이 1~2년으로 짧아 두 번째에 위치한다. 광 연결은 우회가 어렵지만 물량 확대가 2027~2028년에 걸쳐 진행되는 만큼 2027년 이익 영향은 아직 중간 수준이다. 반면 AI 가속기는 공급이 개선되고 ASIC으로 다변화되면서 병목에서 멀어질 수 있다는 판단이다. 테스트는 공급 병목이 아니라 칩당 테스트 강도가 높아지는 구조적 수요 증가 영역으로 구분한다. 결국 메모리 생산능력과 레티클 10배 이상의 최상단 첨단 패키징, 테스트는 가격 결정력에 가깝고, 범용 CoWoS와 선단 공정은 병목의 일부가 OSAT와 삼성전자 2nm 같은 대안으로 흘러가는 대체 수요의 영역이다.

표24. 2027년 병목 강도 비교

| 구분 | 수급 격차 | 해소 리드타임 | 우회 가능성 | 2027년 이익 영향 | 종합 |
|---|---|---|---|---|---|
| 메모리 생산능력(HBM·DRAM) | 상 | 김(신규 팹 2028~29년) | 낮음(LPDDR·PIM 규모 제한) | 상 | ① 가장 심함 |
| 첨단 패키징(CoWoS·SoIC) | 상 | 중간(1~2년, OSAT 분산) | 중간 | 중상 | ② 높음 |
| 광 연결(CPO) | 중상 | 중간(광 엔진 수율) | 낮음 | 중(정점 2028년) | ③ 중상 |
| 선단 공정(2nm) | 중 | 중간 | 중간(삼성 2nm 대안) | 중 | ④ 중간 |
| AI 가속기(Compute) | 하 | 짧음 | 높음(ASIC 다변화) | 하 | ⑤ 낮음 |
| 테스트 | 하(공급) | 짧음 | 낮음 | 중상(수요) | 구조적 수요 증가 |

자료: 유안타증권 리서치센터 주: 당사 판단 기준이며 공급 병목과 수요 증가 영역을 구분

시간 축으로 보면 병목은 사라지지 않고 정점을 옮겨간다. 메모리 생산능력이 2026~2027년 정점을 지나는 동안 첨단 패키징은 2027년, 광 연결은 Feynman 세대의 광 NVLink가 도입되는 2028년, 데이터센터 전력은 2030년에 각각 정점을 맞을 전망이다.

메모리 생산능력은 신규 팹 가동이 본격화되는 2029년 이후 점차 완화되지만, 첨단 패키징은 레티클 10배 이상의 초대형 칩과 하이브리드 본딩 도입으로 2030년까지 높은 강도를 유지한다. 선단 공정은 2028년 대안이 늘며 잠시 완화된 뒤 2029~2030년 High-NA EUV 도입과 함께 다시 높아지고, AI 가속기는 2028년 이후 병목에서 벗어난다. 테스트 수요는 병목이 옮겨갈 때마다 새로운 구조가 등장하는 만큼 2028년 이후 가속될 전망이다.

그림 18. 2026~2030 년 병목 이동 지도

| 공급 병목 | 2026 | 2027E | 2028E | 2029E | 2030E |
|---|---|---|---|---|---|
| 메모리 생산능력(HBM·DRAM) | 정점 | 정점 | 높음 | 높음 | 중간 |
| 첨단 패키징(CoWoS·SoIC·본딩) | 높음 | 정점 | 높음 | 높음 | 높음 |
| 광 연결(CPO) | 낮음 | 높음 | 정점 | 높음 | 높음 |
| 선단 공정(2nm·High-NA) | 높음 | 중간 | 중간 | 높음 | 높음 |
| 전력(800VDC·SiC·GaN) | 중간 | 중간 | 높음 | 높음 | 정점 |
| AI 가속기 | 높음 | 낮음 | 낮음 | 낮음 | 낮음 |
| **구조적 수요 증가** | | | | | |
| 테스트(수요 강도) | 증가 | 확대 | 가속 | 가속 | 가속 |

주: 당사 판단. 공급 병목은 수급 격차·해소 리드타임·우회 가능성 종합, 테스트는 칩당 테스트 강도 기준

자료: 유안타증권 리서치센터 주: 당사 판단

**대체 수요는 병목이 가장 심한 곳 바로 옆에서 열린다.** 강도가 가장 큰 것은 메모리 후공정 외주다. HBM 캐파 확장이 이어지는 가운데 신규 팹의 클린룸은 2028년 말 이후에야 열리는 만큼, 메모리 업체는 제한된 클린룸을 전공정에 우선 배정하고 후공정을 OSAT로 넘길 수밖에 없다. 가장 강한 병목을 우회하는 수요인 만큼 수혜의 강도도 크다. 이와 함께 HBM 생산능력이 부족할수록 LPDDR 기반의 SOCAMM2와 PIM이, CoWoS가 빠듯할수록 OSAT 외주와 패널 레벨 패키징이, TSMC N2가 타이트할수록 삼성전자 2nm가 대안으로 채택된다. 한편 가격 결정력은 대안이 없는 곳에서 나온다. 메모리 생산능력, 레티클 10배 이상의 초대형 칩을 수용할 수 있는 최상단 첨단 패키징, 어떤 구조를 택하든 거쳐야 하는 테스트가 대표적이다.

대체 수요는 채택이 확인되는 시점에 멀티플 리레이팅이 나타나는 변곡형이고, 가격 결정력은 병목이 지속되는 동안 이익 추정치 상향이 쌓이는 누적형이다. 특히 테스트는 대안 구조가 늘어날수록 검증 수요도 함께 늘어난다는 점에서 두 갈래의 수혜를 모두 받는다는 점이 긍정적이다. 두 갈래 모두에서 남은 질문은 결국 기업의 선택이다. 대체 수요에서는 누가 채택을 실제 수주로 연결하는지, 가격 결정력에서는 누가 병목이 완화된 이후에도 단가를 지켜내는지가 수혜의 크기를 가른다. 병목이 이동하는 속도만큼 수혜 기업의 순위도 바뀔 것이며, 당사는 이 질문에 대한 답을 영역별 기업 분석에서 이어갈 예정이다.

표25. 병목 대응 전략: 대체 수요와 가격 결정력

| 구분 | 병목 → 대응 영역 | 근거 | 체크포인트 |
|---|---|---|---|
| 대체 수요 | 메모리 클린룸 →<br>메모리 후공정 외주(OSAT) | 신규 팹 첫 클린룸 가동 2028년 말~2029년,<br>HBM 캐파 확대로 내부 클린룸을 전공정에 우선 배정 | 메모리 외주 비중,<br>OSAT 이익 레버리지 확대 구간 |
| | HBM 생산능력 →<br>LPDDR·SOCAMM2·PIM | LPDDR 계약가격 1Q27 +20% 이상,<br>LPDDR6-PIM 표준화로 채택 가속 | 칩 업체의 PIM 지원,<br>SOCAMM2 서버 탑재 |
| | CoWoS 생산능력 →<br>OSAT·패널 레벨 패키징 | AMD Venice·NVIDIA Vera 로 OSAT 외주 확대, CoPoS 2H28 양산 | OSAT 수주, CoPoS 양산 일정 |
| | TSMC N2 → 삼성전자 2nm | SF2P 첫 제품 웨이퍼 투입, Tesla 향 대규모 파운드리 계약 | 2nm 수율 안정화, 신규 고객 확보 |
| 가격 결정력 | 메모리 생산능력<br>→ HBM·서버 DRAM | 신규 팹 가동 2028년 말~2029년,<br>우회 수단은 2027년 안에 규모 제한 | 장기공급계약 확대 |
| | 최상단 첨단 패키징<br>→ 대형 CoWoS·SoIC | 레티클 10배 이상 초대형 칩은 대안 부재,<br>패키징 가격 15~20% 인상 논의 | 패키징 가격 인상 실현 |
| | 테스트<br>→ 메모리 테스터·번인·소켓 | 구조가 다양해질수록 검증 횟수 증가,<br>Chroma ATE +180%·WinWay +153%(3Q26 YoY) | 칩당 테스트 시간, 신규 규격 수 |

자료: 각 사, TrendForce, 대만증권거래소(MOPS), 유안타증권 리서치센터 주: 본문 내용을 바탕으로 한 당사 판단

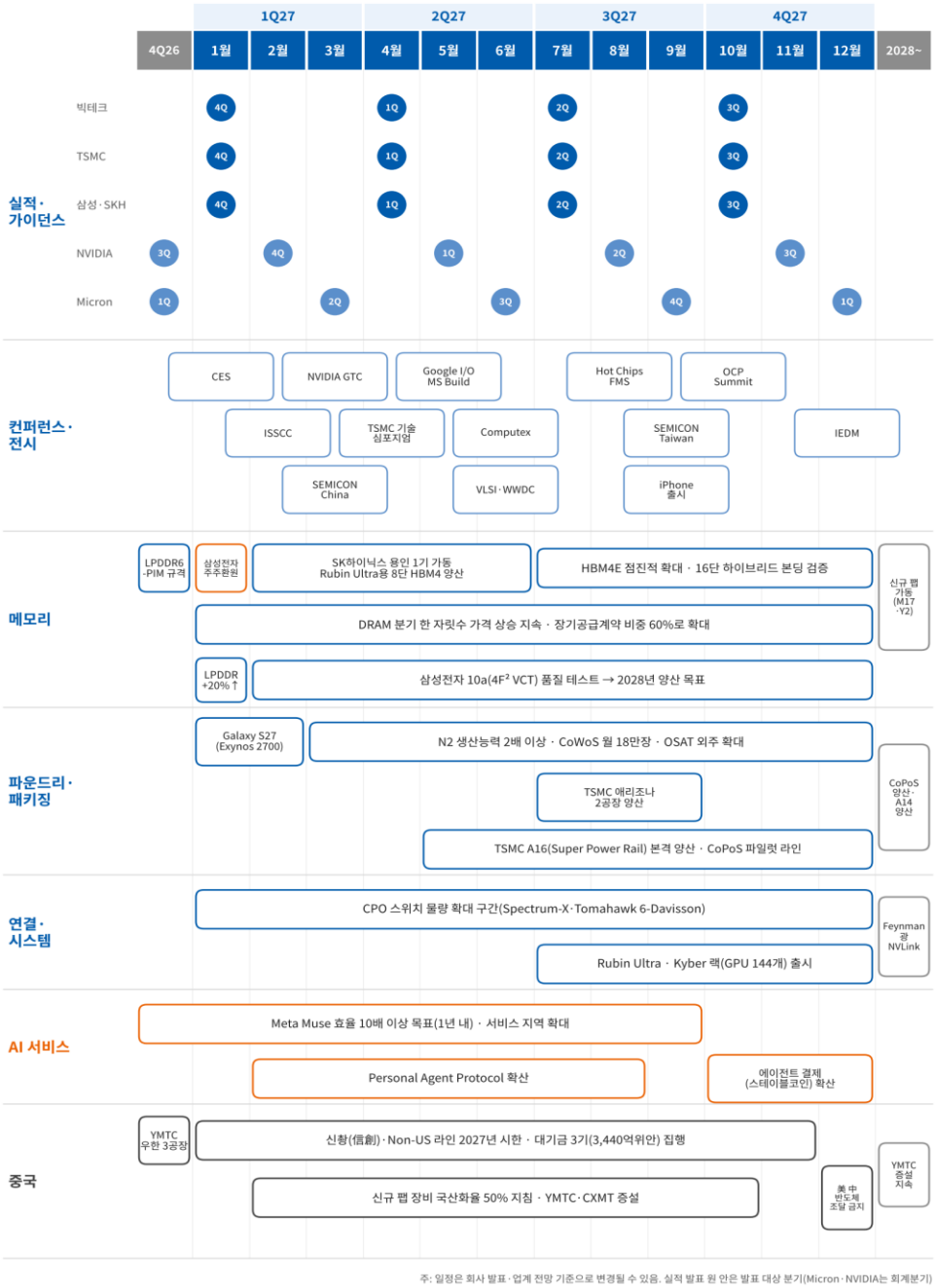
표26. 2027~2030년 병목 이동과 수혜 영역

| 병목·변화 | 시점 | 핵심 근거 | 수혜 영역 |
|---|---|---|---|
| ① 메모리 생산능력 | 2027년 가장 심함,<br>2029년 이후 완화 | HBM 웨이퍼 비중 20%→30% 근접, 삼성전자 HBM 웨이퍼 투입 월 17만→27만장, 베이스다이 외주화로 DRAM 웨이퍼와 파운드리 생산능력 두 제약에 동시에 묶임, 신규 팹 가동 2028년 말~2029년, LPDDR 계약가격 1Q27 +20% 이상 | 메모리 업체, 전공정 장비(증착·식각·세정), 소모성 부품, 메모리 후공정 외주(OSAT) |
| ② Compute 구조 변화(ASIC·Host CPU) | 2027년 | ASIC 확대가 설계·IP·네트워크·파운드리·첨단 패키징의 외주 생태계를 키움, 서버 CPU 2,470만대(2025)→5,000만대(2027), CPU 성능이 시스템 처리량과 GPU 가동률을 좌우 | 디자인하우스·IP, 서버 CPU, 서버 DRAM |
| ③ HBM 맞춤화(메모리의 파운드리화) | 2027년 HBM4부터 | HBM 이 단순한 메모리를 넘어 데이터 경로의 일부가 됨, 승패는 선단 파운드리 접근권과 로직 IP 설계 역량, Custom 으로 갈수록 한국과 중국 업체의 격차 확대 | 로직 베이스다이, 메모리·파운드리 턴키, 파운드리 협업 |
| ④ HBM 고단화·본딩 | 2027~2028년 | 16단 HBM4·HBM4E 고단화, 하이브리드 본딩은 HBM4E~HBM5 세대 전후 도입 준비 가능성 | TC 본더, 하이브리드 본더 |
| ⑤ 서버 메모리 분화·PIM·보드 연결 | 2027년부터 구조적 | 서버 메모리 분화가 시장 크기 자체를 넓힘, 버퍼·PMIC 등 인터페이스 칩 탑재량 증가, LPDDR6-PIM 표준화로 채택 가속, 메모리도 연결 병목의 해법을 만드는 쪽 | MRDIMM·SOCAMM2·LPDDR6, 인터페이스 칩, 모듈·SSD 테스트 |
| ⑥ 첨단 패키징 | 2027년 두 번째로 심함 | CoWoS 2027년 월 18만장 이상에도 빠듯, 패키징 가격 15~20% 인상 논의(병목이 가격으로), AMD Venice·NVIDIA Vera 로 OSAT 외주 확대, SoIC 피치 6→4.5㎛, CoPoS 2H28 양산 | OSAT, 패널 레벨 패키징, 3D 적층·본딩 |
| ⑦ 테스트 | 병목 순위와 별개, 상시 수혜 | 메모리 업계의 시행착오가 테스트 수요의 기회, 고속화·PAM3·PIM 으로 테스트 강도 상승, 수혜는 메모리 테스트에 집중, 대만 Chroma ATE +180%·WinWay +153%(3Q26 YoY) | 메모리 테스터, 번인, 소켓, 핸들러 |
| ⑧ 엣지 | 2027년 낮아진 기저에서 회복 | AI PC 메모리 대역폭 5년간 50~100% 상향(Intel), 프리미엄 스마트폰 12~16GB 이상·LPDDR6 전환, 개인 에이전트가 새로운 메모리 탑재량 기준을 만듦 | 모바일·PC DRAM, LPDDR6 |
| ⑨ 광 연결(CPO) | 2028년 정점 | CPO 상용화 진입, 성능은 패키징 기술에서 갈림, 2028년 Feynman 광 NVLink 로 Scale-up 까지 확장 | 광 엔진, 레이저 광원, 광학 패키징 |
| ⑩ 선단 공정(로직) | 2027년 중간, 2029~2030년 High-NA 로 재부상 | TSMC N2 매출 비중 8%(3Q26), 2027년 N2 생산능력 두 배 이상, GAA 로 ALD·전구체·선택적 식각 수요, SiC Ring·Quartz 교체 주기 단축, BSPDN 으로 웨이퍼 본딩·CMP·오버레이 계측, High-NA 미도입으로 멀티 패터닝 증가, 삼성전자 2nm 회복 | 증착·식각·세정 장비, 소모성 부품, CMP, 계측 |
| ⑪ 메모리 공정 전환 | 2028년 전후 시작 | DRAM 4F² VCT 로 셀 면적 약 1/3 축소, 주변 회로 웨이퍼 본딩, NAND W2W 본딩·극저온 식각·몰리브덴, 구조 전환 자체가 공급 제약 요인 | 웨이퍼 본딩, ALD, 극저온 식각, 오버레이 계측 |
| ⑫ 전력 | 2029~2030년 마지막 병목 | 반도체가 스스로 풀 수 없는 병목, AI 투자를 나눠 집행되게 만들어 사이클을 장기화, 800VDC 로 SiC·GaN 역할 확대 | SiC·GaN 전력반도체, 전원 IC |

자료: 각 사, 업계 자료, Bloomberg, 대만증권거래소(MOPS), 유안타증권 리서치센터 주: 본문 내용을 바탕으로 한 당사 판단

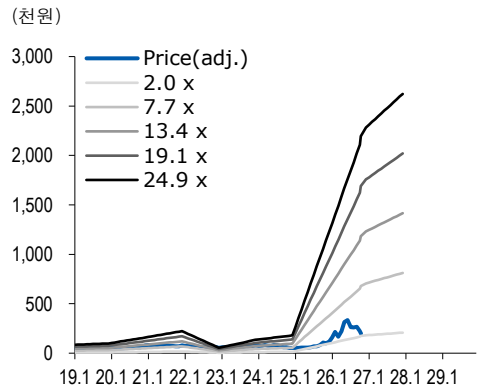
# Appendix

그림 19. 2027 년 영역별 체크포인트

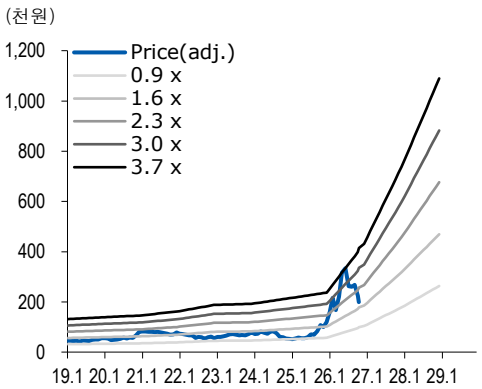



자료: 각 사, TrendForce, Federal Register, 언론 보도, 유안타증권 리서치센터 주: 일정은 회사 발표 및 업계 전망 기준이며 변경될 수 있음

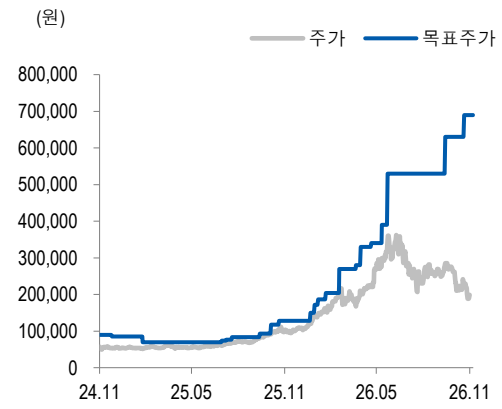
P/E band chart




P/B band chart




삼성전자 (005930) 투자등급 및 목표주가 추이




| 일자 | 투자 의견 | 목표가 (원) | 목표가격 대상시점 | 괴리율 | |
|---|---|---|---|---|---|
| | | | | 평균주가 대비 | 최고(최저) 주가 대비 |
| 2026-11-11 | BUY | 690,000 | 1년 | | |
| 2026-10-30 | BUY | 690,000 | 1년 | | |
| 2026-10-29 | BUY | 630,000 | 1년 | -61.71 | -56.19 |
| 2026-10-29 | BUY | 690,000 | 1년 | -67.39 | -67.39 |
| 2026-09-23 | BUY | 630,000 | 1년 | -61.71 | -56.19 |
| 2026-06-01 | BUY | 530,000 | 1년 | -47.30 | -31.60 |
| 2026-05-21 | BUY | 390,000 | 1년 | -22.31 | -18.72 |
| 2026-04-30 | BUY | 340,000 | 1년 | -19.46 | -12.94 |
| 2026-04-09 | BUY | 330,000 | 1년 | -34.51 | -31.52 |
| 2026-01-30 | BUY | 204,000 | 1년 | -13.68 | -0.25 |

자료: 유안타증권

주: 괴리율 = (실제주가* - 목표주가) / 목표주가 X 100

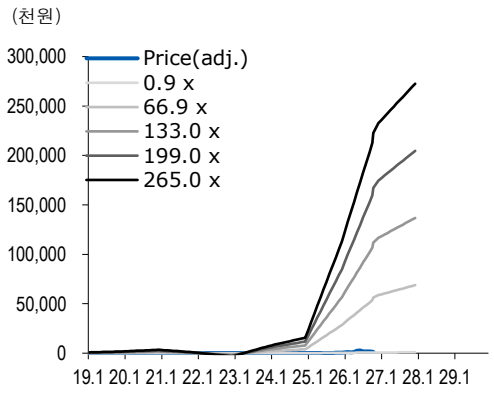
* 1) 목표주가 제시 대상시점까지의 "평균주가"
2) 목표주가 제시 대상시점까지의 "최고(또는 최저) 주가"

| 구분 | 투자의견 비율(%) |
|---|---|
| Strong Buy(매수) | 0 |
| Buy(매수) | 96.9 |
| Hold(중립) | 3.1 |
| Sell(비중축소) | 0 |
| 합계 | 100.0 |

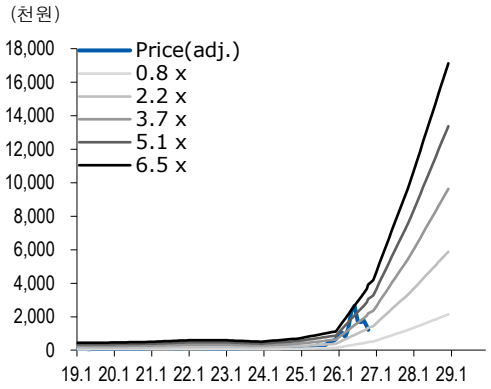
주: 기준일 2026-11-10

※해외 계열회사 등이 작성하거나 공표한 리포트는 투자등급 비율 산정시 제외

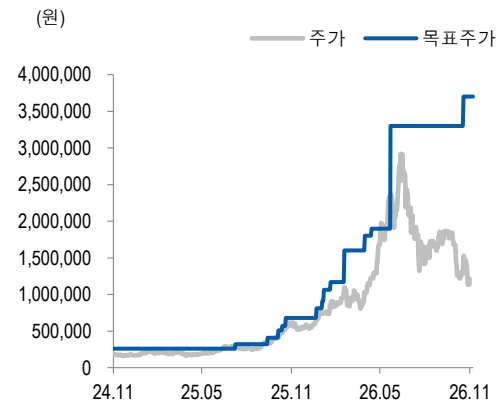
P/E band chart




P/B band chart




SK 하이닉스 (000660) 투자등급 및 목표주가 추이




| 일자 | 투자 의견 | 목표가 (원) | 목표가격 대상시점 | 괴리율 | |
|---|---|---|---|---|---|
| | | | | 평균주가 대비 | 최고(최저) 주가 대비 |
| 2026-11-11 | BUY | 3,700,000 | 1년 | | |
| 2026-10-28 | BUY | 3,700,000 | 1년 | | |
| 2026-06-01 | BUY | 3,300,000 | 1년 | -43.83 | -11.55 |
| 2026-04-23 | BUY | 1,900,000 | 1년 | -7.58 | 22.79 |
| 2026-04-09 | BUY | 1,800,000 | 1년 | -37.02 | -32.00 |
| 2026-02-26 | BUY | 1,600,000 | 1년 | -41.55 | -33.69 |
| 2026-01-29 | BUY | 1,169,000 | 1년 | -22.82 | -12.92 |
| 2026-01-15 | BUY | 1,066,000 | 1년 | -28.06 | -21.11 |
| 2026-01-12 | BUY | 910,000 | 1년 | -18.68 | -18.46 |
| 2025-10-22 | BUY | 570,000 | 1년 | -10.33 | -6.14 |

자료: 유안타증권

주: 괴리율 = (실제주가* - 목표주가) / 목표주가 X 100

* 1) 목표주가 제시 대상시점까지의 "평균주가"
2) 목표주가 제시 대상시점까지의 "최고(또는 최저) 주가"

| 구분 | 투자의견 비율(%) |
|---|---|
| Strong Buy(매수) | 0 |
| Buy(매수) | 96.9 |
| Hold(중립) | 3.1 |
| Sell(비중축소) | 0 |
| 합계 | 100.0 |

주: 기준일 2026-11-10

※해외 계열회사 등이 작성하거나 공표한 리포트는 투자등급 비율 산정시 제외

## Appendix

- 이 자료에 게재된 내용들은 본인의 의견을 정확하게 반영하고 있으며 타인의 부당한 압력이나 간섭 없이 작성되었음을 확인함. (작성자: **백길현**)
- 당사는 자료공표일 현재 동 종목 발행주식을 1%이상 보유하고 있지 않습니다.
- 당사는 자료공표일 현재 해당 기업과 관련하여 특별한 이해관계가 없습니다.
- 당사는 동 자료를 전문투자자 및 제 3자에게 사전 제공한 사실이 없습니다.
- 동 자료의 금융투자분석사와 배우자는 자료공표일 현재 대상법인의 주식관련 금융투자상품 및 권리를 보유하고 있지 않습니다.
- 종목 투자등급 (Guide Line): 투자기간 12개월, 절대수익률 기준 투자등급 4단계(Strong Buy, Buy, Hold, Sell)로 구분한다
- Strong Buy: +30%이상 Buy: 15%이상, Hold: -15% 미만 ~ +15% 미만, Sell: -15%이하로 구분
- 업종 투자등급 Guide Line: 투자기간 12개월, 시가총액 대비 업종 비중 기준의 투자등급 3단계(Overweight, Neutral, Underweight)로 구분
- 2014년 2월21일부터 당사 투자등급이 기존 3단계 + 2단계에서 4단계로 변경

본 자료는 투자자의 투자를 권유할 목적으로 작성된 것이 아니라, 투자자의 투자판단에 참고가 되는 정보제공을 목적으로 작성된 참고 자료입니다. 본 자료는 금융투자분석사가 신뢰할만 하다고 판단되는 자료와 정보에 의거하여 만들어진 것이지만, 당사와 금융투자분석사가 그 정확성이나 완전성을 보장할 수는 없습니다. 따라서, 본 자료를 참고한 투자자의 투자의사결정은 전적으로 투자자 자신의 판단과 책임하에 이루어져야 하며, 당사는 본 자료의 내용에 의거하여 행해진 일체의 투자행위 결과에 대하여 어떠한 책임도 지지 않습니다. 또한, 본 자료는 당사 투자자에게만 제공되는 자료로 당사의 동의 없이 본 자료를 무단으로 복제 전송 인용 배포하는 행위는 법으로 금지되어 있습니다.

유안타증권
Yuanta Securities (Korea)


Head of Research Center 최현재
3770-2553 / hyunjae.choi@yuantakorea.com

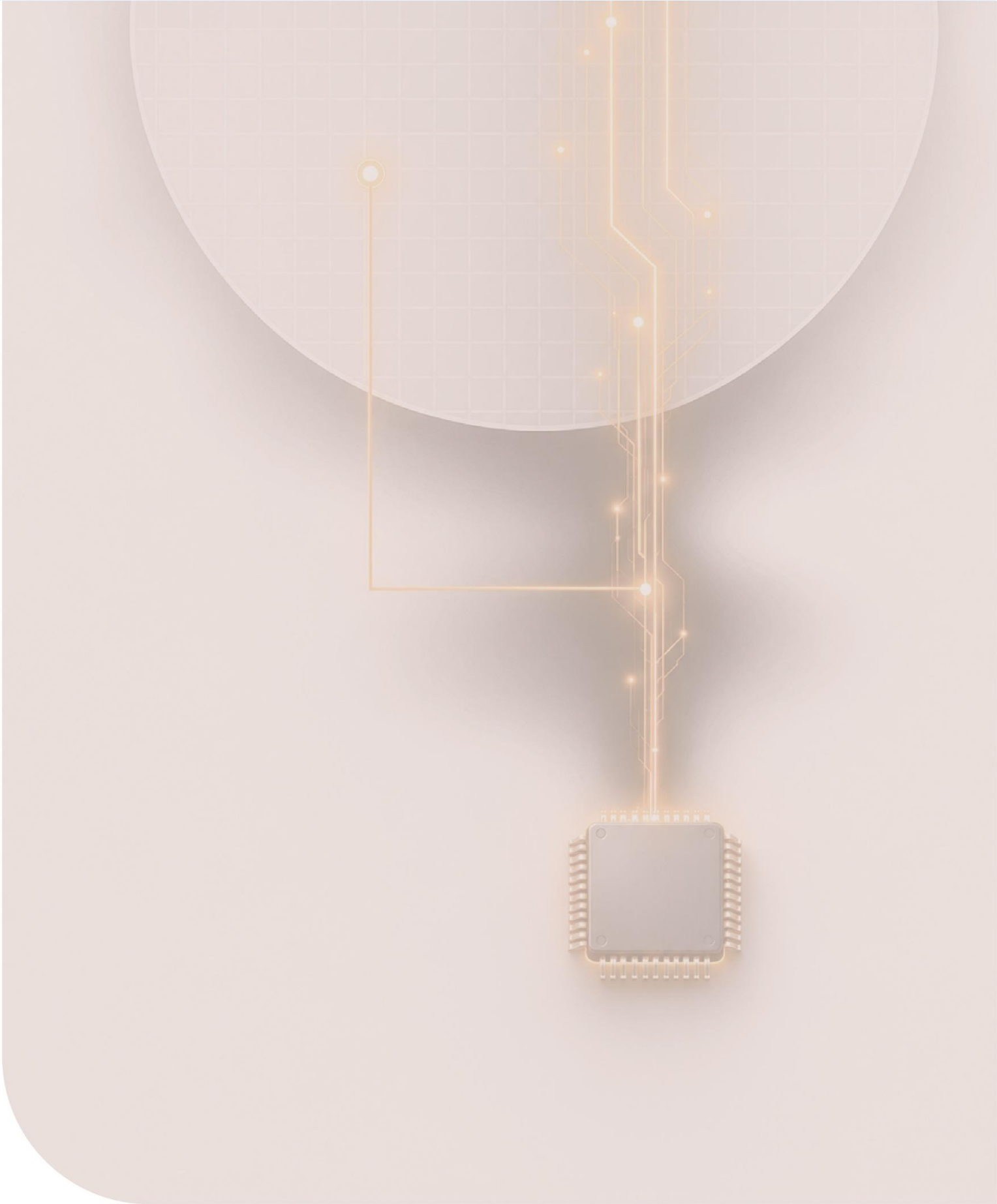
부센터장 2차전지/신에너지 이안나
3770-5599 / anna.lee@yuantakorea.com


## 투자전략

| | | | |
|---|---|---|---|
| 팀장 Strategist | 김용구 | 3770-3521 | yg.kim@yuantakorea.com |
| Fixed Income Strategist | 이재형 | 5579 | jaehyung.lee@yuantakorea.com |
| Passive/ ETF Analyst | 고경범 | 3625 | gyeongbeom.ko@yuantakorea.com |
| Credit Analyst | 이현수 | 5718 | hyunsoo.yi@yuantakorea.com |
| Global Strategist | 민병규 | 3635 | byungkyu.min@yuantakorea.com |
| Economist/ESG | 김호정 | 3630 | hojung.kim@yuantakorea.com |
| Quant Analyst | 신현용 | 3634 | hyunyong.shin@yuantakorea.com |
| Market Analyst | 이재원 | 5719 | jaewon2.lee@yuantakorea.com |
| Research Assistant | 임지윤 | 3527 | jiyoon.lim@yuantakorea.com |
| Research Assistant | 김혜원 | 3526 | hyewon.kim@yuantakorea.com |
| Research Assistant | 김세빈 | 3646 | sebin2.kim@yuantakorea.com |

## 기업분석

| | | | |
|---|---|---|---|
| 팀장 2차전지/신에너지 | 이안나 | 3770-5599 | anna.lee@yuantakorea.com |
| 인터넷/SW | 이창영 | 5596 | changyoung.lee@yuantakorea.com |
| 정유/화학 | 황규원 | 5607 | kyuwon.hwang@yuantakorea.com |
| 스몰캡 | 권명준 | 5587 | myoungchun.kwon@yuantakorea.com |
| 화장품/의료기기/유통 | 이승은 | 5588 | seungeun.lee@yuantakorea.com |
| 제약/바이오 | 하현수 | 2688 | hyunsoo.ha@yuantakorea.com |
| 통신/지주 | 이승웅 | 5597 | seungwoong.lee@yuantakorea.com |
| 조선/자동차 | 김용민 | 5606 | yongmin.kim@yuantakorea.com |
| 미디어/엔터/디지털자산 | 이환욱 | 5590 | hwanwook.lee@yuantakorea.com |
| US Market | 황병준 | 3523 | byeongjun.hwang@yuantakorea.com |
| 유틸리티/음식료 | 손현정 | 5595 | hyunjeong.son @yuantakorea.com |
| 반도체 | 백길현 | 5635 | gilhyun.baik@yuantakorea.com |
| 금융 | 우도형 | 5589 | dohyeong.woo@yuantakorea.com |
| 전기전자 | 고선영 | 3525 | sunyoung.kou@yuantakorea.com |
| 방산/우주/AI/로보틱스 | 백종민 | 5598 | jongmin.baik@yuantakorea.com |
| 운송/게임 | 최지운 | 3640 | jiyun.choi@yuantakorea.com |
| 건설/기계 | 김도엽 | 5580 | doyub.kim@yuantakorea.com |
| Research Assistant | 서석준 | 5585 | seokjun.seo@yuantakorea.com |
| Research Assistant | 한동우 | 3647 | dongwoo.han@yuantakorea.com |
| Research Assistant | 임석민 | 3648 | seokmin.lim@yuantakorea.com |
| Research Assistant | 김고은 | 3649 | koeun2.kim@yuantakorea.com |
| Research Assistant | 배종성 | 3643 | jongsung.bae@yuantakorea.com |
| Research Assistant | 신승우 | 5594 | sungwoo.shin@yuantakorea.com |
| Research Assistant | 김영민 | 5602 | yeongmin.kim@yuantakorea.com |
| Research Assistant | 조계철 | 2665 | gyecheol.jo@yuantakorea.com |



2027년 연간전망

# 수요의 재설계

AI 수익화와 메모리 접근 구조의 변화

유안타증권
Yuanta Securities (Korea)