중국 유망 산업 / 기업 전망

미중 AI 생태계 경쟁, 최종 승자는 누가 될 것인가



2026.11.19 | 중국 기업분석 백승혜 sh_baek@hanafn.com RA 안기량 giryangan@hanafn.com

하나증권

Contents

Part 1.	화웨이 AI칩 기술력 및 중국의 연산능력 점검	5
Part 2.	중국 HBM 기술 동향과 글로벌 영향력 점검	14
Part 3.	2027년 중국의 AI 연산능력 규모 전망	25
Part 4.	중국의 AI 모델 개발 전략	32
Part 5.	2027년 중국 주식 투자 유망종목	45

하나증권

Compliance Notice

본 조사자료는 고객의 투자에 정보를 제공할 목적으로 작성되었으며, 어떠한 경우에도 무단 복제 및 배포 될 수 없습니다. 또한 본 자료에 수록된 내용은 당사가 신뢰할 만한 자료 및 정보로 얻어진 것이나, 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 최종결정을 하시기 바랍니다. 따라서 어떠한 경우에도 본 자료는 고객의 주식투자의 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다.

Summary

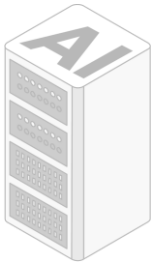
미중 AI 생태계 경쟁, 최종 승자는 누가 될 것인가

AI 생태계 전반으로 경쟁 심화됨에 따라 美 견제 범위와 강도 확대될 가능성

- 미국과 중국의 AI 경쟁은 AI 모델과 애플리케이션부터 피지컬 AI, 데이터센터 인프라, 반도체 칩 및 제조장비에 이르기까지 기술 생태계 전반에서 진행되고 있음. 2026년 기점으로 중국의 AI 모델, 휴머노이드 제품의 글로벌 영향력이 빠르게 확산되고 있어, 향후 관련 영역으로 미국의 견제 범위가 더욱 확대될 가능성이 높음
- 중국 정부는 AI 모델의 빠른 글로벌 확산과 AI 생태계 영향력 확대를 통해 중국 AI 산업의 글로벌 위상을 높이고, 궁극적으로 글로벌 시장의 중국 AI 기술 의존도를 높이는 것을 목표하고 있음. 따라서 중국은 오픈웨이트 AI 모델의 가격 경쟁력과 개방성을 활용해 글로벌 시장에서 영향력을 확대하고, 이를 중국산 AI 서비스, 디바이스, 반도체, 데이터센터 등 AI 기술 스택 전반에 대한 의존도 확대로 연결하려 할 가능성이 높음. Mythos급 AI 모델의 등장으로 우월한 AI 역량을 확보한 국가가 경쟁국에 대해 상당한 비대칭적 우위를 점하게 될 것으로 예상되며, 최첨단 AI 모델의 기술 경쟁력은 앞으로도 미중 패권 경쟁을 좌우하는 핵심 요인으로 작용할 전망

CXMT의 HBM 공급 규모가 중국 AI 연산능력 확대 속도를 좌우할 것

- 단기적으로 AI 공급망에서 가장 큰 병목은 HBM으로 CXMT의 HBM 공급 능력이 중국 연산능력 확대 속도를 좌우하게 될 것. HBM과 AI 반도체가 차세대 기술을 통해 EUV 없이 최첨단 성능 구현과 대량양산이 가능해진다면 중국 AI 연산능력 증가 속도는 가팔라질 것.미중 AI 경쟁이 본격화됨에 따라 중국은 해외 DC 접근, 밀수, 미국산 AI 모델 증류 등 비공식적인 방식까지 동원해 연산능력 부족으로 인한 약점을 보완해왔으나, 이와 관련해 앞으로 미국이 규제 수위를 높이면, 중국의 AI 공급망 자립, 특히 HBM, EUV 등 중국이 가장 취약한 영역에서의 기술 자립 속도는 더욱 높아질 것



I. 화웨이 AI칩 기술력 및 중국의 연산능력 점검

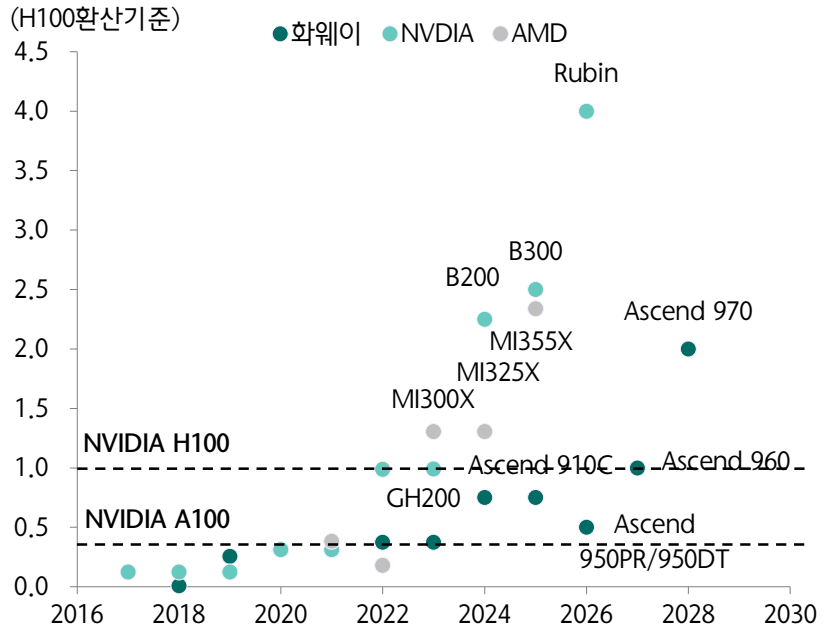
하나증권 리서치센터 해외주식분석실

중국기업분석 백승혜

미국 규제로 중국 AI 칩 연산능력은 여전히 H100 이하

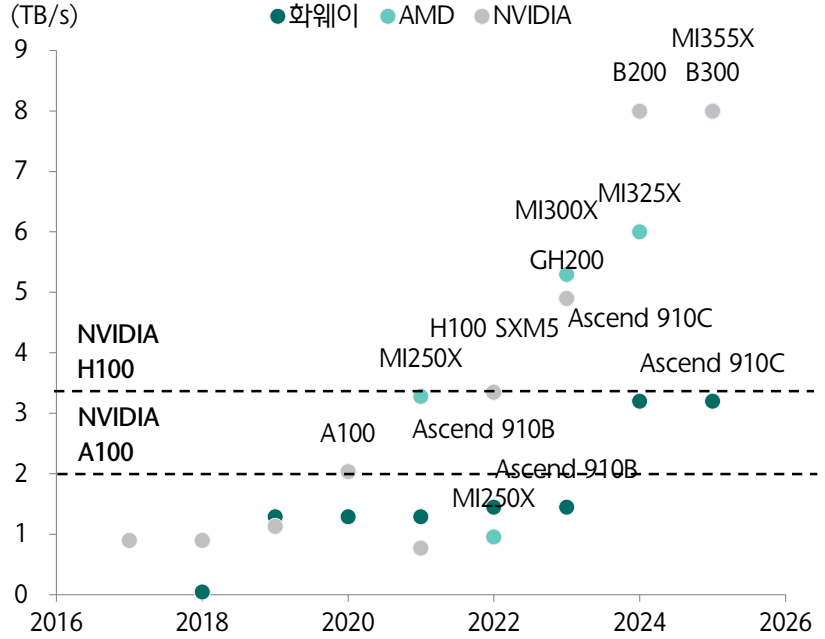
- 중국 AI 산업에 대한 미국의 규제는 중국의 AI 연산능력 제약에 집중해왔으며, 이로 인해 반도체 수출통제가 핵심 규제 수단이 됐음
- 미국 규제로 인해 화웨이를 제외한 대부분의 중국 AI 반도체 팹리스는 최대 연산 성능 기준으로 엔비디아의 A100 수준 또는 2022년 H100의 약 30% 이하 수준에 머물러 있으며, 화웨이의 2026년 최신 칩도 H100 이하 수준의 연산능력과 메모리 성능

화웨이와 미국 AI 칩 연산능력 비교



자료: Epoch AI, 하나증권

화웨이와 미국 AI 칩 메모리 대역폭 비교



자료: Epoch AI, 하나증권

화웨이 개별 칩 성능은 엔비디아와 격차 큰 상황

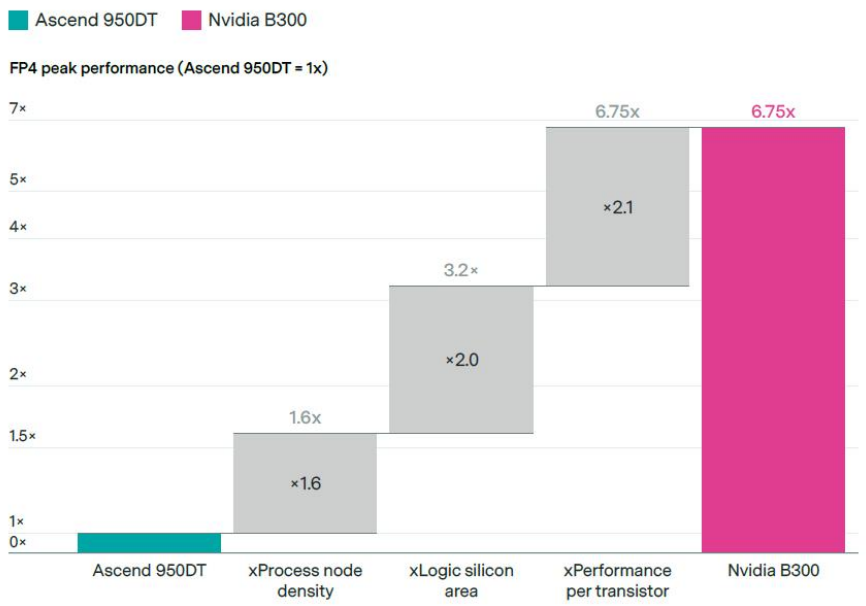
- 화웨이가 2026년 4분기 출시된 950DT는 프리필(Prefill) 단계에 최적화된 950PR과 동일한 연산 다이를 사용하나, 메모리 용량과 대역폭을 각각 144GB, 4TB/s로 확대. FP4 기준 최대 연산 성능은 B300이 Ascend 950 대비 약 6.75배로 엔비디아 대비 성능 격차 존재
- 이러한 차이는 로직 다이 면적, 트랜지스터 밀도뿐 아니라 텐서코어 설계, 클럭 속도, 전력 한도, 데이터 처리 경로 등 요인에서 발생

2026년 화웨이가 출시한 Ascend 950DT/950PR 칩 성능

구분	Ascend 950PR	Ascend 950DT	NVIDIA B300 (HGX)	Blackwell Ultra (GB300)
FP8 연산성능 (PFLOPS)	1.0	1.0	9.0	10.0
FP4 연산성능 (PFLOPS)	2.0	2.0	13.5	20.0
메모리 용량 (GB)	128.0	144.0	270.0	279.0
메모리 대역폭 (TB/s)	1.6	4.0	7.7	8.0
메모리 종류	HiBL 1.0	HiZQ 2.0	HBM3E	HBM3E

자료: Epoch AI, 하나증권

B300과 Ascend 950DT의 성능 격차 분해



주: 제조공정 밀도 차이는 SMIC N+3와 TSMC N4P의 단위 면적당 트랜지스터 수, 로직 실리콘 면적은 Ascend 950DT 약 800mm², B300 약 1,600mm²로 비교, 트랜지스터당 성능 차이는 클럭 속도, FP4 데이터 처리 경로, 텐서코어 설계, 전력 한도, 연산 코어 면적 등 반영
자료: Epoch AI, 하나증권

화웨이 자체 개발 HBM 성능은 HBM3E 이하, 생산규모도 매우 제한적

- HiBL 1.0과 HiZQ 2.0은 2025년 9월 화웨이가 공개한 자체 개발 HBM 기술로 Ascend 950 시리즈에 첫 탑재
- 950DT는 4개의 HBM 스택 사용, HiZQ 2.0의 스택당 메모리 용량 36GB, 대역폭 1.0TB/s로 글로벌 HBM3E보다 다소 낮은 수준
- 화웨이 자체 DRAM 생산능력은 Swaysure 포함 월 10만장 수준, CXMT의 약 30~40% 수준으로 추정되나, HBM 생산능력 규모는 매우 제한적인 수준으로 예상. 또한, HiBL·HiZQ의 실제 제조사가 공개되지 않아 CXMT 등이 공급에 참여했을 가능성 제기되고 있음

화웨이 자체 개발 HBM 제품 성능

구분	HiBL 1.0	HiZQ 2.0
탑재 AI 칩	Ascend 950PR	Ascend 950DT
메모리 용량	128GB	144GB
메모리 대역폭	1.6TB/s	4.0TB/s
성능 기준	AI 칩 전체	AI 칩 전체
주요 용도	AI 추론 Prefill, 추천 시스템	AI 추론 Decode, 모델 학습

자료: Epoch AI, 하나증권

글로벌 3사 HBM과 성능 비교 (HBM 스택당 기준)

기업·제품	메모리 용량	메모리 대역폭
Huawei HiZQ 2.00 (950DT)	36GB	1.0TB/s
삼성전자 HBM3E 12단	36GB	1.28TB/s
SK하이닉스 HBM3E 12단	36GB	약 1.23TB/s
마이크론 HBM3E 12단	36GB	1.2TB/s 이상
삼성전자 HBM4 12단	36GB	최대 3.3TB/s
SK하이닉스 HBM4 12단	36GB	약 3.0TB/s
마이크론 HBM4 12단	36GB	2.8TB/s 이상

자료: 각 사, 하나증권

대규모 시스템은 개별 칩 성능 열세를 완화하는 중요한 수단

- 화웨이는 스케일업 시스템과 네트워킹에서 상대적으로 높은 경쟁력 보유, 많은 칩을 연결해 전체 시스템 차원에서 보완하려는 전략
- 화웨이는 엔비디아 대비 스케일업 규모를 공격적으로 확대 중, 4Q26 최대 8,192개 칩 연결하는 대규모 컴퓨팅 시스템을 출하할 계획?
- 실제 시스템 경쟁력에서 화웨이 UnifiedBus 등 초고속 네트워크 기술이 엔비디아의 네트워킹 성능을 얼마만큼 구현 가능한가가 중요

화웨이와 엔비디아의 스케일업 시스템 비교

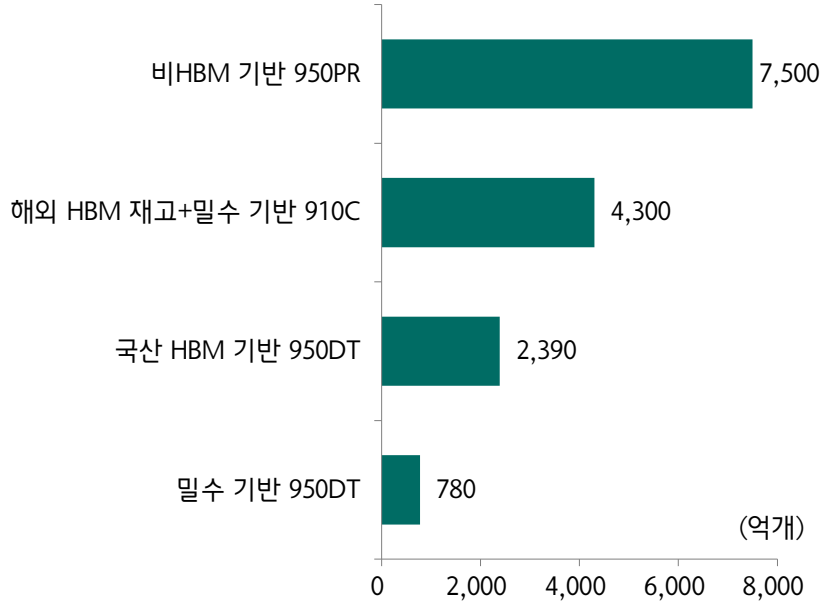
시스템	총 연산능력	HBM 총용량	HBM 총대역폭	총 전력소모
화웨이 CloudMatrix 384	약 300 PFLOPS (BF16)	49.2 TB	1,229 TB/s	약 559~600 kW
화웨이 Atlas 950 SuperPoD	8 EFLOPS (FP8)	1,152 TB	약 32,768 TB/s	약 9,500 kW
화웨이 Atlas 960E SuperPoD	8 EFLOPS (FP8)	최대 1 PB	-	-
엔비디아 NVL72 (Blackwell)	360 PFLOPS (FP8 Dense)	13.4 TB	576 TB/s	약 145 kW
엔비디아 NVL72 (Rubin)	1,260 PFLOPS (FP8 Dense)	20.7 TB	1,400 TB/s	약 225~230 kW
엔비디아 NVL576 (Rubin Ultra)	약 5 EFLOPS	-	약 4,600 TB/s	-

자료: 각 사, 하나증권

화웨이는 생산량 확대를 통해 격차를 만회할 수 있을까?

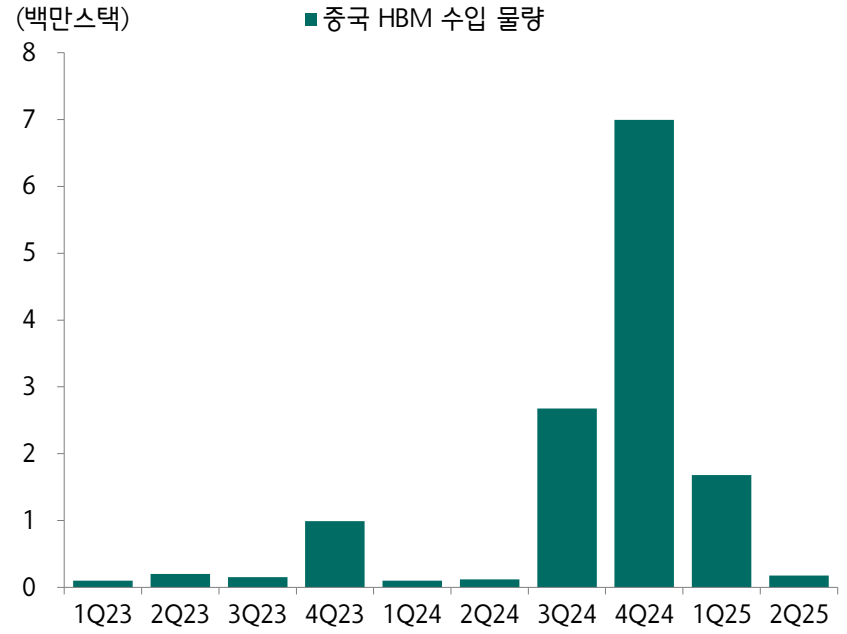
- 화웨이가 낮은 개별 칩 성능을 생산량 확대로 보완할 수 있는지는 공급업체들의 생산능력 확대 역량에 달려 있으나, 단기적으로는 개별 칩 성능보다 HBM 생산량 자체가 화웨이에 더 시급한 제약 요인
- 중국은 2025년 불법 밀수한 물량과 2024년 12월 수출통제 시행 전 확보한 해외산 HBM 재고 1,000만스택 이상을 활용해왔으, 해당 재고는 2026년 중 소진될 전망. 비축 HBM 소진 시 중국 AI 칩 생산 규모는 CXMT 생산능력과 밀수 물량에 의해 결정될 수밖에 없음

2026년 화웨이 HBM 확보 루트별 Ascend 칩 출하량



자료: Epoch AI, 하나증권

미 제재로 2024년 3~4분기 HBM 수입 물량 급증 이후 크게 감소

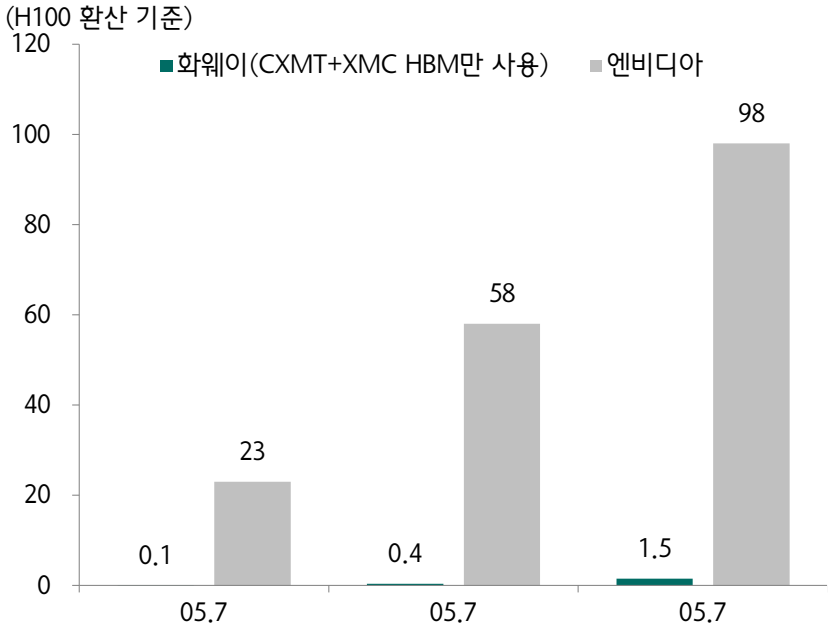


자료: SemiAnalysis, 하나증권

중국 HBM만으로는 화웨이의 생산량 엔비디아의 1% 수준에 머무를 전망

- 중국산 HBM만 사용할 경우 화웨이의 AI 연산능력 생산 규모는 2028년까지 엔비디아의 약 1% 수준에 머무를 가능성이 있음
- HBM 생산에 할당된 DRAM 웨이퍼 규모와 수율 차이를 고려하면, 2026년 엔비디아는 화웨이보다 약 73배 많은 HBM 확보할 전망
- 해외 확보 HBM 재고, 저대역폭 대체 메모리, 밀반입 HBM을 활용해도 화웨이의 2026년 생산 규모는 엔비디아의 4% 미만에 그칠 것

2026~28년 국산 HBM만 사용한 화웨이의 생산량 전망



자료: Epoch AI, 하나증권

엔비디아와 화웨이의 연산능력 생산 격차 분해

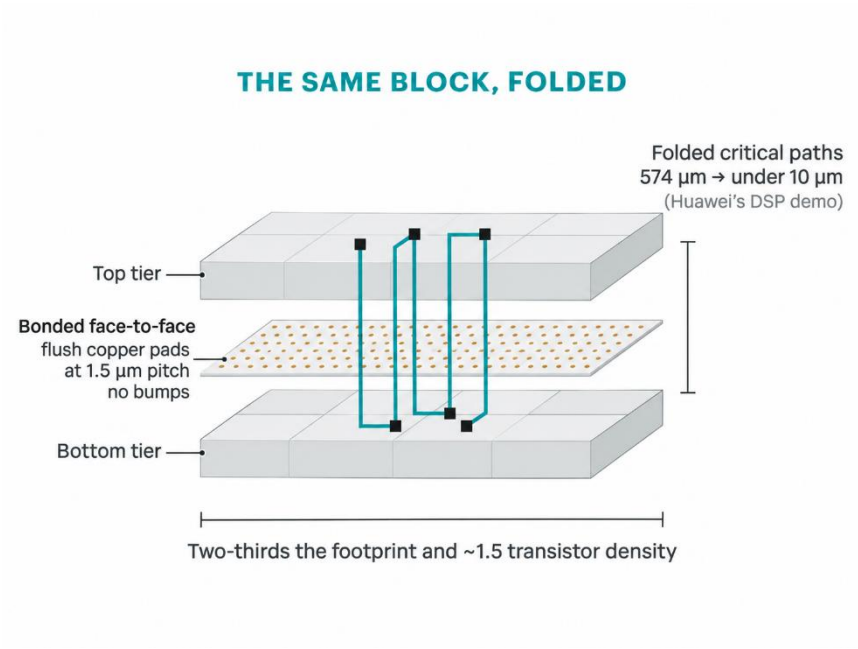
기준	2026년	2028년
사용 가능한 HBM 공급량 격차	73배	11배
HBM 용량당 연산성능 격차	2.7배	6.1배
총 AI 연산능력 생산 격차	약 200배	약 65배

자료: Epoch AI, 하나증권

2030년 이후 화웨이의 핵심 전략은 'LogicFolding'

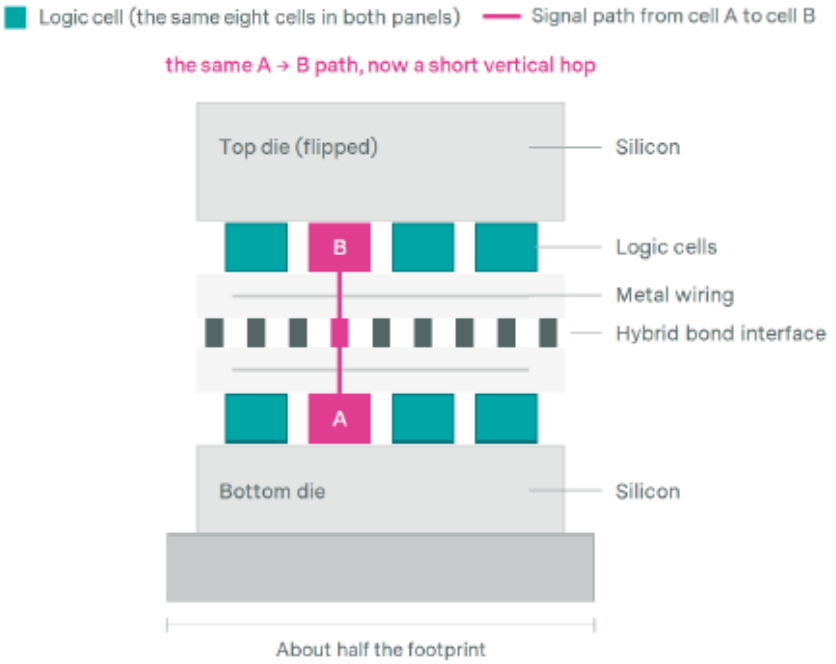
- 화웨이는 장기적으로 'LogicFolding', 즉 로직 다이를 수직 적층해 첨단공정 수준의 트랜지스터 밀도 확보하는 기술 확보를 목표 중
- EUV 장비 규제로 인한 5nm급 제조공정 한계를 넘어 연산성능을 높이고, 메모리 적층을 통해 대역폭을 확대에 기여할 수 있음
- 이러한 3D 적층 기술을 성공적으로 구현하고 대규모로 양산하는 것이 2030년 이후 화웨이 경쟁력의 핵심이 될 전망
- LogicFolding이 Ascend 제품군에 적용되는 시점은 2030년, Ascend 950부터 980까지는 3D 적층 기술이 적용되지 않을 전망

화웨이 '로직폴딩'은 로직 회로를 2개 층으로 나누어 적층하는 기술



자료: 화웨이, 하나증권

'로직폴딩'의 3D 적층 기술은 긴 수평 배선을 짧은 수직 연결로 대체

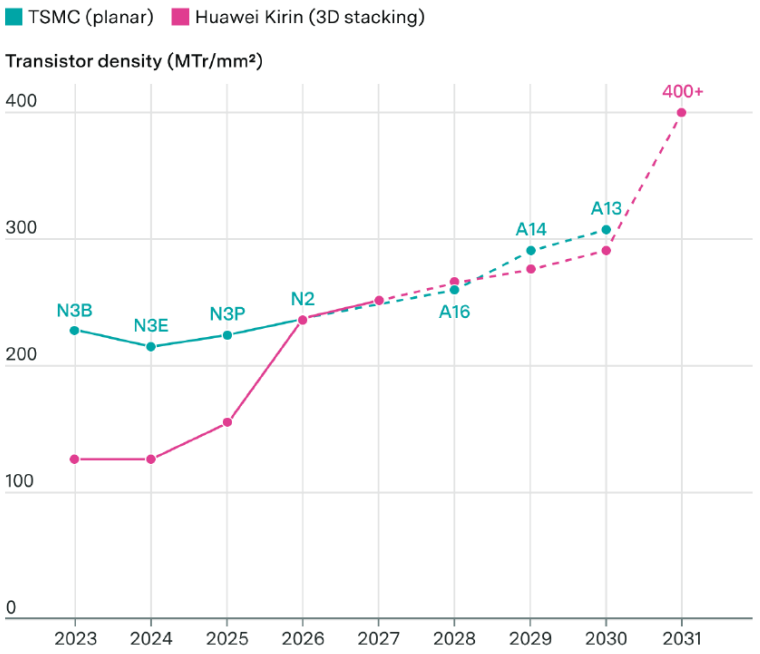


자료: 화웨이, 하나증권

LogicFolding으로 2030년까지 TSMC 평면 트랜지스터 밀도 확보할 가능성

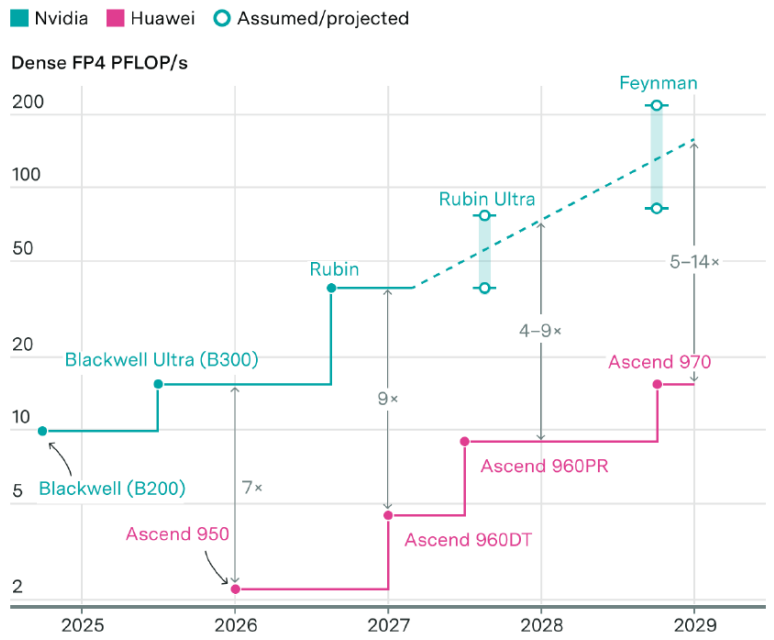
- LogicFolding을 활용하면 화웨이는 2030년까지 TSMC의 최첨단 제조공정이 달성하는 평면 트랜지스터 밀도 수준을 따라갈 가능성
- 이는 엔비디아가 1.6nm급 공정에서 제조한 적층 로직을 적용한 Feynman을 출시할 것으로 예상되는 시점보다 2년 늦음
- 2031년까지 1.4nm 상당의 트랜지스터 밀도 구현 목표는 수율, 장비, EDA, 메모리 기술의 상당한 발전을 전제로 하고 있음
- 장기적으로 엔비디아와 격차 축소 위해서는 뛰어난 수직 적층 기술, CUDA와의 하드웨어 활용률 격차 해소, 첨단 공정 접근 등 필요

3D 수직 적층으로 TSMC 평면 트랜지스터 집적도 수준 근접할 전망

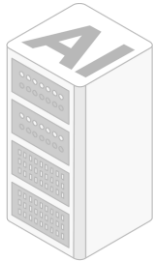


자료: Epoch AI, 하나증권

화웨이 Ascend와 NVIDIA의 AI 반도체 연산성능 로드맵 비교



자료: Epoch AI, 하나증권



II. 중국 HBM 기술 동향과 글로벌 영향력 점검

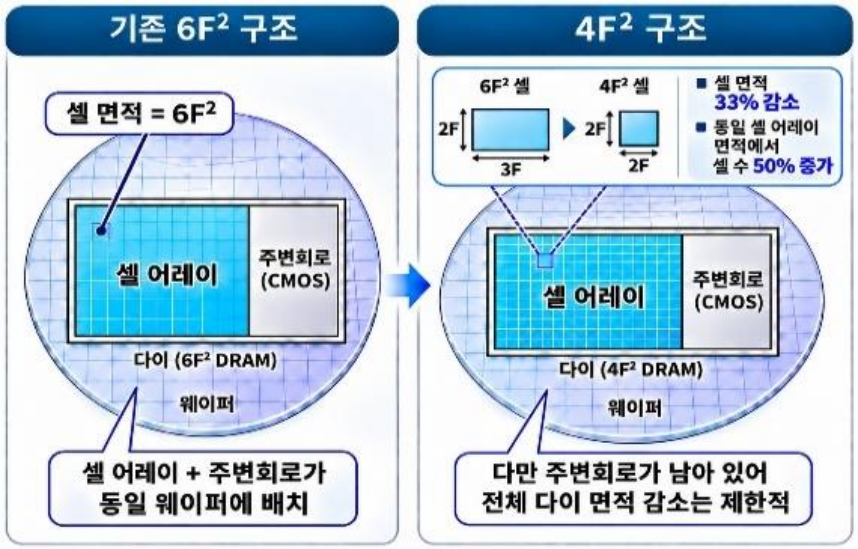
하나증권 리서치센터 해외주식분석실

중국기업분석 백승혜

CXMT의 4F² 기술 조기 도입은 DRAM 미세화를 위해 불가피한 선택

- 미국 규제로 EUV 장비 사용이 불가한 CXMT는 DRAM 셀의 최소 선폭(F) 미세화 한계 돌파하기 위해 지난 수년간 4F² 기술 개발 진행
- 기존 6F²에서 4F² 구조로 전환 시 셀 면적 감소해 더 많은 메모리 셀 집적 가능, EUV가 가장 필요한 콘택트/랜딩패드 층 제거 가능
- 이론적으로 셀 면적이 약 33% 감소하고, 단위 면적당 셀 집적도는 약 50% 증가하나, 6F² 대비 수율 하락, 주변회로용 추가 웨이퍼 투입에 따른 생산능력 감소로 실제 셀 집적도는 이론적 50% 증가 효과 대비 낮을 전망

기존 6F²와 차세대 4F² DRAM 다이 평면 구조



자료: 하나증권

6F²와 4F² 아키텍처 주요 차이점 비교 표

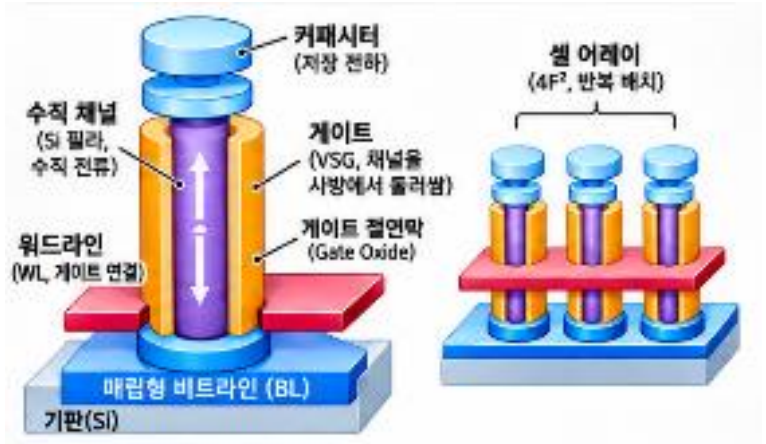
구분	6F² (현재 주류 구조)	4F² (수직 채널 구조)
셀 면적	6F²	동일한 F 기준 셀 면적 33% 감소, 셀 집적도 최대 50% 증가
접근 트랜지스터	매립 채널 어레이 트랜지스터 구조. 채널이 웨이퍼 표면과 수평으로 배치	수직 채널 트랜지스터(VCT) 적용, 채널이 웨이퍼 표면과 수직으로 배치
비트라인	트랜지스터 상부에 배치	수직 채널 하부에 매립형 비트라인 배치
커패시터 연결	비트라인 콘택트, 스토리지 노드 콘택트 및 랜딩 패드 필요	수직 채널 상부에 커패시터를 직접 배치해 연결 구조 단순화
EUV 필요성	1α 공정부터 랜딩 패드 및 커패시터 홀 등 핵심 공정에 EUV 적용 확대	셀 어레이의 EUV 의존도 감소. 주변회로는 고해상도 노광공정 필요
주변회로	셀 어레이와 동일한 웨이퍼에 배치	주요 구현 방식에서는 주변회로를 별도 웨이퍼에 형성, 본딩을 통해 셀 어레이 하부에 배치
주요 기술 과제	물리적 미세화 한계에 근접하면서 EUV 적용 공정 수 및 비용 증가	플로팅 바디 효과에 따른 데이터 보존 특성 저하, 비트라인 간 간섭, 웨이퍼 본딩 수율 및 제조비용 부담

자료: ToAlpha, 하나증권

CXMT 4F² 아키텍처의 핵심은 VSG와 하이브리드 본딩 기술

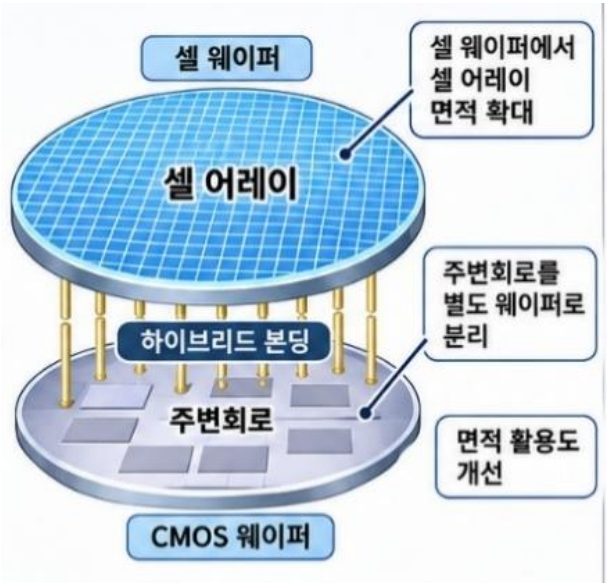
- 4F² 구조로 미세화할 경우 트랜지스터 크기가 작아지면서 메모리 셀의 누설전류(leakage current)에 더욱 취약해지는 문제가 발생
- 이를 위해 CXMT는 VSG(Vertical Surrounding Gate) 트랜지스터와 W2W(Wafer-to-Wafer) 하이브리드 본딩 기술을 결합할 계획
- VSG는 트랜지스터 채널을 수직 배치해 셀 평면 면적을 축소하고, 게이트가 채널을 완전히 둘러싸도록 설계해 누설전류 줄이는 기술
- 하이브리드 본딩 기술은 주변회로를 별도 웨이퍼로 분리해 셀 어레이 아래로 수직 적층, 주변회로 이동을 통해 전체 다이 면적 축소

CXMT의 VSG 기술 구조도



자료: 하나증권

하이브리드 본딩 구조도

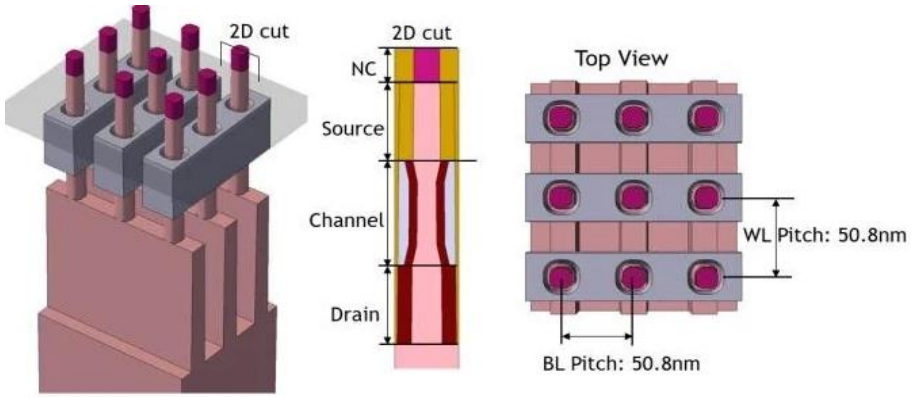


자료: 하나증권

4F² 기술은 CXMT가 장기간 축적해온 연구개발의 연장선

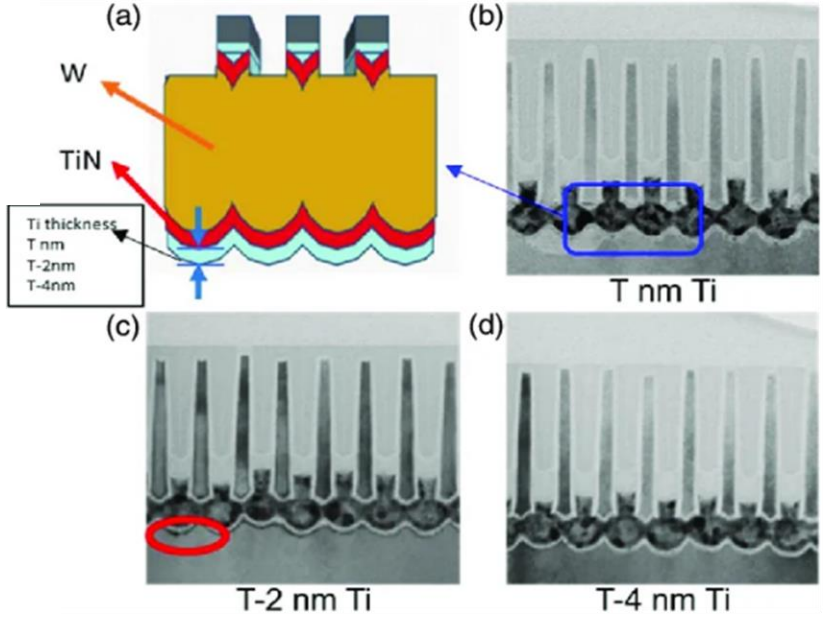
- CXMT는 이미 2021년 193nm ArF 액침 노광과 자기정렬 이중 패터닝(SADP)을 활용해 41 × 47.3nm 피치의 수직 기둥 배열을 구현
- 2023년 이후 전체 배열 수직 채널 트랜지스터, 접합 없는 GAA 트랜지스터, 매립형 금속 비트라인, 고일함수 금속 게이트 등 핵심 기술 연구 결과를 지속 발표. 따라서 이번 4F² 발표는 단발성 기술 성과가 아니라 장기간 축적된 연구개발의 연장선이라고 볼 수 있음

2023년 공개한 CXMT 4F² VCT 셀 어레이 3D 구조



주: CXMT가 제시한 9개 수직 채널 트랜지스터의 3차원 배열, 수직 단면과 평면도로 50.8nm 피치의 워드라인·비트라인 배치 포함
자료: CXMT <IEEE IMW 2023>, 하나증권

2024년 공개한 CXMT 수직 기둥 및 매립형 금속 비트라인 TEM 사진



주: 실리콘 수직 기둥 하부에 $TiSi_x$ 금속 실리사이드를 형성하는 구조와 실제 TEM 관찰 결과
자료: CXMT <Japanese Journal of Applied Physics 2024>, 하나증권

$4F^2$ 기술의 핵심 쟁점은 생산성과 경제성

- $4F^2$ 구조는 수직 트랜지스터의 균일한 제조, 축소된 면적 내 커패시터 충분한 전하 저장, 배선 간 간섭 최소화 위해 정밀한 설계 필요
- 또한 하이브리드 본딩은 추가 공정, 두 장의 웨이퍼 사용으로 본딩 비용 추가되며, 메모리 셀과 주변회로 사이의 정밀한 정렬이 요구
- 이러한 기술적 난제는 수율 하락과 제조비용 상승으로 이어질 수 있으며, 동일 면적에서 셀 집적도를 높이는 장점을 상쇄할 수 있음
- CXMT는 '26년 말까지 $4F^2$ 적용 제품 샘플 제작이 목표, 샘플 용도(연구개발/고객사 검증), 수율, 원가, 대량생산 일정 등 확인 필요

향후 글로벌 DRAM 산업 주요 관전 포인트

시점	주요 이벤트	핵심 모니터링 지표
4Q26	CXMT, $4F^2$ DDR5 RDIMM 출시	메모리 용량(Gb) 및 속도 사양, 칩 전체에 $4F^2$ 구조 적용 여부, 서버 고객사 인증 여부
1H27	ISSCC 2027, VLSI 2027 개최	CXMT의 $4F^2$ 제품 수준 기술 논문 발표 여부 (데이터 보존 특성, 리프레시, 수율)
2027	CXMT 상하이 팹 가동, G5 제품 본격 출하	G5 공정 수율 개선 속도, 연말 웨이퍼 생산능력 월 38만~42만장 도달 여부
2H27	마이크론 ID1, SK하이닉스 용인 공장, 난야 Fab 5A 웨이퍼 투입 개시	신규 생산능력의 계획 대비 적기 확보 여부
2027~2028	삼성전자, 수직 채널 기반 $4F^2$ 구조의 고객사 인증 및 양산 계획	삼성전자 또는 SK하이닉스의 $4F^2$ 도입 일정 조기화 여부
2028	글로벌 DRAM 수급 균형 근접	CXMT 전체 웨이퍼 투입량 중 $4F^2$ 비중 25% 이상 도달 여부 및 기술적 도약 시나리오의 실현 가능성
장기	미국 MATCH Act 및 Entity List 규제	DUV 액침 노광장비의 공급 및 유지보수 제한 여부

자료: ToAlpha, 하나증권

한국 기업은 차세대 기술 도입을 서두를 유인이 크지 않음

- 삼성전자도 2026년 말까지 $4F^2$ 기반 10a DRAM 개발을 완료하고, 2027년 고객사 인증을 거쳐 2028년 양산에 진입할 계획
- SK하이닉스도 10nm 이하 공정에서 수직 게이트와 본딩 기술 도입 계획, $6F^2$ 미세화 최대한 연장한 후 $4F^2$ 또는 3D DRAM 도입 전망
- 한국은 EUV 등 미세화 수단을 이미 보유, 차세대 기술 도입에 따른 수율 저하와 비용 상승 감수하지 않고도 집적도와 성능 개선 가능
- $4F^2$는 최종 기술이 아닌 3D DRAM으로 가는 과도기적 선택, 장기적으로 3D DRAM 전환에 성공한 기업이 기술경쟁 주도권 재편할 것

글로벌 DRAM 제조사 차세대 DRAM 기술 로드맵 비교

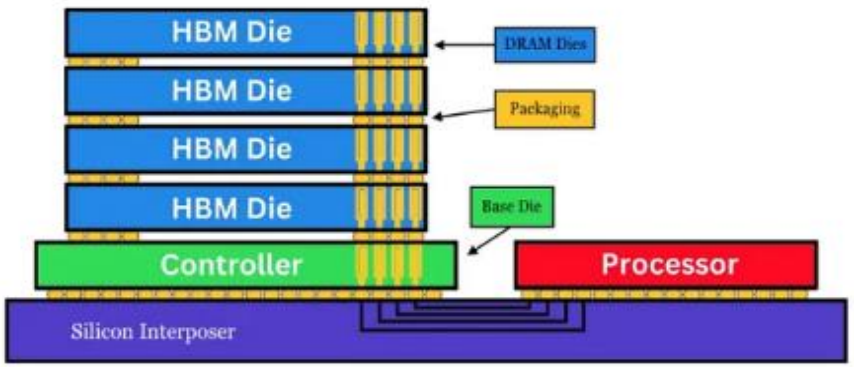
업체	핵심 기술 및 개발 전략	도입 및 양산 전망
CXMT	$4F^2$ VSG + 웨이퍼 본딩. EUV 접근 제약을 셀 구조 혁신과 DUV 기반 공정으로 보완	2026년 말 DDR5 RDIMM 고객사 샘플 공급 목표. 이후 서버용 DRAM부터 도입하고 다른 제품군으로 확대. 양산은 고객 인증·수율·시장 수요에 따라 점진적으로 추진
SK하이닉스	$4F^2$ VG(Vertical Gate) + 웨이퍼 본딩 병행 개발. 기존 $6F^2$ 미세화를 최대한 연장한 후 $4F^2$ 또는 3D DRAM 도입	공급망 조사상 10b 세대부터 VG·본딩 적용 전망. 일부 시장에서는 2027~2028년 도입 예상. 다만 회사는 구체적 양산 시점을 확정하지 않고 경제성 확보 이후 적용한다는 방침
삼성전자	$4F^2$ VCT + W2W 하이브리드 본딩 개발. 2026년 9월부터 본딩 우선 도입, VCT 별도 개발 전략으로 수정	기존 공급망 전망: 2027년 VCT 인증, 2028년 양산. 다만 로드맵 수정 이후 본딩 기반 D0a 상용화를 먼저 추진하며, VCT 기반 $4F^2$ 양산 일정은 재확인 필요
마이크론	$6F^2$ 미세화 연장 → 웨이퍼 본딩 → 3D DRAM. $4F^2$ 수직 트랜지스터는 현재 로드맵에서 제외	2027년 하반기 1δ 양산, 이후 1ε에서 웨이퍼 본딩 도입. 1ζ·1η 후속 세대와 병행해 3D DRAM 개발. 본딩 및 3D DRAM 양산 연도는 미확정

자료 각 사, 언론보도, 하나증권

차세대 DRAM 기술만으로는 여전히 한국을 이길 수 없다

- CXMT의 차세대 HBM 경쟁력은 DRAM 셀 미세화 기술뿐 아니라 첨단 로직 공정 기반의 "베이스 다이" 확보 여부에도 좌우될 수 있음
- DRAM 다이의 성능이 아무리 뛰어나더라도 베이스 다이의 데이터 처리 능력 부족 시 HBM 전체 대역폭과 전력효율이 제한될 수 있음
- HBM4부터 데이터 입출력 인터페이스 1,024비트에서 2,048비트로 확대, 베이스 다이에 5nm 이하 첨단 파운드리 공정 적용 시작
- CXMT도 TSMC와 협력 시도 예상되나 美 견제로 TSMC가 거절할 가능성, SMIC 7nm 수준 공정 활용 시 대역폭 경쟁력 제한 가능성

HBM의 3대 핵심 구성요소 및 패키징 구조



자료: 산업자료, 하나증권

HBM 세대별 베이스다이 공정 노드

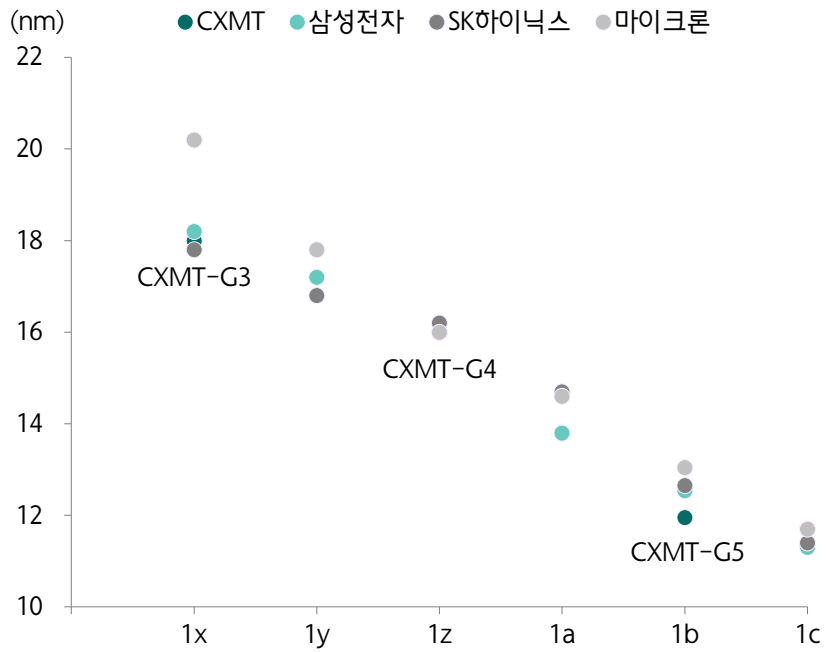
Specification	HBM2	HBM2E	HBM3	HBM3E	HBM4	HBM4E
베이스 다이 공정 노드	1x	1y/1z	1z	1a/1b	삼성 4nm 하이닉스 TSMC 12nm	삼성 2nm (엔비디아 맞춤 HBM TSMC 3nm 예상) 하이닉스 TSMC 3~5nm 예상
입출력 데이터 폭 (bits)	1,024	1,024	1,024	1,024	2,048	2,048
데이터 전송속도 (Gbps/Pin)	2.4	3.6	6.4	9.8	8.0	12.0

자료: 각 사, 하나증권

최신 DRAM 공정의 기술적 진전은 고무적

- CXMT는 2026년 9월 5세대 DRAM 공정(G5) 양산 발표, G5 공정은 최소 선폭 측면에서 선두업체와 동등한 수준에 도달했음을 시사
- CXMT G5 공정의 활성 영역 하프피치는 11.95nm로 SK하이닉스, 삼성전자 등 글로벌 선두업체들의 1b 공정과 유사한 수준
- 다만, G5 공정은 4중 패터닝 기술을 통해 최소 선폭을 축소, 노광 공정 단계가 더 많아진 만큼 수율 손실 위험도 높아질 수 있음
- G5 공정은 웨이퍼당 총 다이 생산량이 G4 대비 최소 50% 증가 예상, 향후 대량 출하 실적, 수율, 제조원가 등 통해 경쟁력 확인 필요

G5 공정의 활성 영역 하프피치는 글로벌 1b 공정 수준



주: 통상적으로 DRAM에서 가장 작은 치수인 인접한 두 활성 영역(Active Area) 사이의 거리, 즉 피치(Pitch)의 절반인 하프피치(Half-pitch)를 기준으로 공정 세대를 구분
자료: TechInsights, 하나증권

글로벌 DRAM 업체 세대별 공정

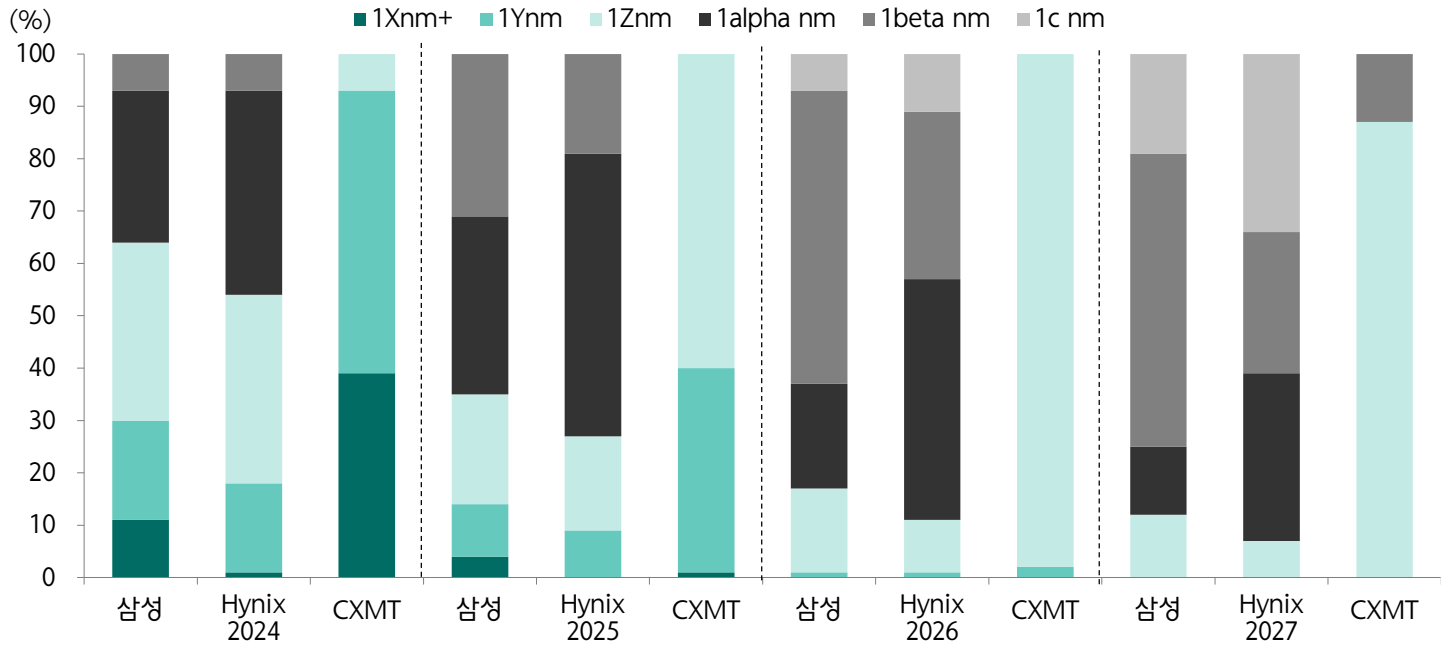
	2025 MS	2018	2019	2020	2021	2022	2023	2024	2025	2026E	2027E	2028E
삼성전자	34%	17 nm (1y)			15/1β nm (1z)*		13/14 nm (1a)		12 nm (1b)	1c nm		1d nm
SK 하이닉스	36%	18 nm (1x)		17 nm (1y)	15/16 nm (1z)		14 nm (1a)*		12 nm (1b)	1c nm		1d nm
Micron	22%	17 nm (1x)		17 nm (1y)	15/16 nm (1z)	14 nm (1a)		13 nm (1β)		1γ nm*		1δ nm
CXMT	6%		G1 (2y nm)				G3 (1x-1ynm)		G4 (1z-1anm)			G5 (1β nm)
Nanya	1%	2x nm					1A nm			1B nm		1C nm

자료: TechInsights, 하나증권

현재는 1z 공정 중심, 2027년부터 G5(1b) 비중 점차 확대될 전망

- 2027~28년 CXMT의 웨이퍼 생산능력 확대는 고사양 DDR5, DDR6 및 HBM 생산을 위한 차세대 G5 공정에 집중될 전망
- CXMT는 현재 대부분의 DRAM을 1z 공정에서 생산 중, 2027년 상반기까지 G5(1b) 공정 생산 비중은 매우 제한적인 수준 예상
- 2028년까지 G5 공정의 수율은 기존 세대(90% 이상)보다 낮은 70~85%에 그칠 수 있으나, 공정 전환과 웨이퍼 생산능력 확대에 힘입어 2027~28년 DRAM 비트 출하량은 연간 30% 이상 증가할 것으로 전망되고 있음

한국기업과 CXMT의 DRAM 제조공정별 생산비중

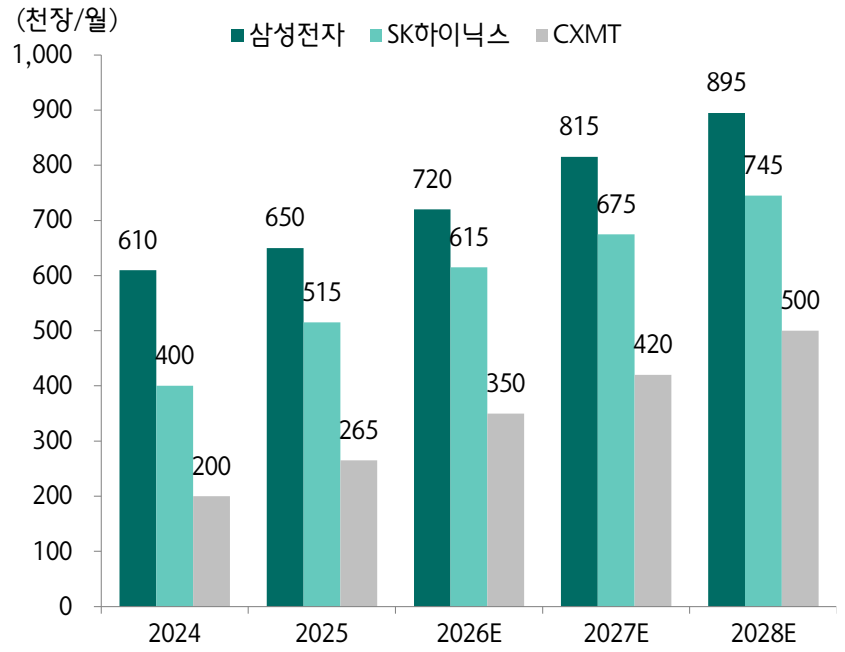


자료: Trendforce, 각 사, 하나증권

2028년까지 증설 속도는 한국 업체 대비 공격적일 전망

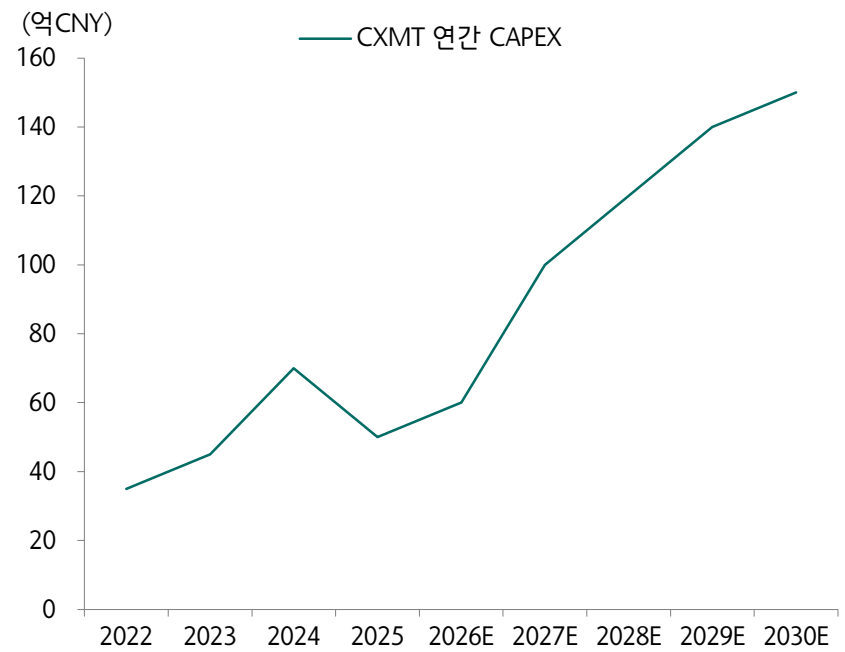
- 2027~30년 신규 생산공장 건설이 완료될 경우, CXMT는 연간 월 8만~9만 장 규모의 웨이퍼 생산능력을 추가 확보할 것으로 예상
- 다만 현재 확인 가능한 클린룸 공간의 제한적인 확장 여력과 핵심 장비/소재 확보에 대한 잠재적 규제를 고려하면, 연간 월 10만 장 이상의 생산능력 증설 가능성은 아직 제한적인 상황
- 설비투자 측면에서는 2027~28년 신규 생산능력 확대를 위한 장비 투자에 연간 약 100억위안 이상 필요할 것으로 예상

CXMT와 한국 DRAM 업체 CAPA 전망



자료: Trendforce, SemiAnalysis, 하나증권

CXMT Capex 전망

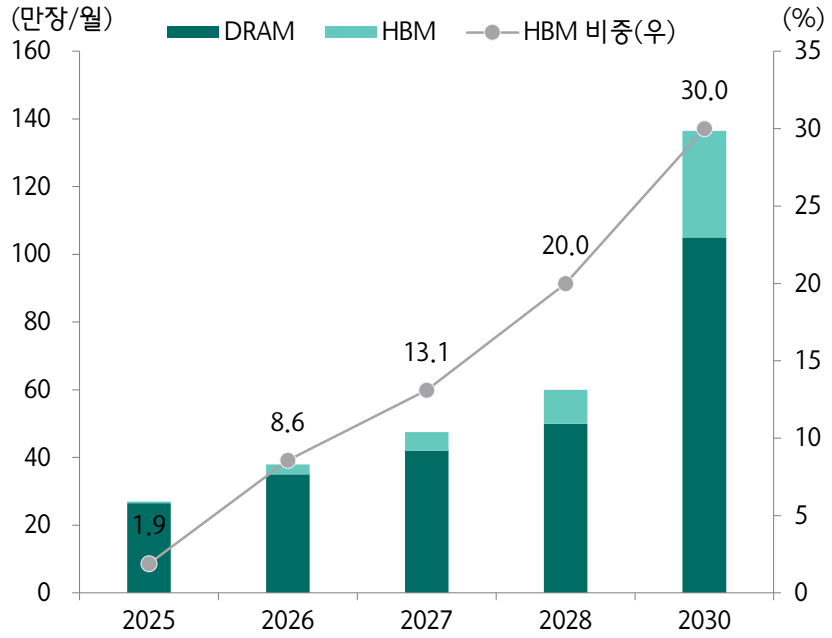


자료: Trendforce, SemiAnalysis, 하나증권

CXMT 증설에도 중국 HBM 수요 충족시키기엔 역부족

- DRAM 생산능력 확대와 HBM 생산 비중 증가로 CXMT의 HBM 생산능력은 2026년 말 대비 2030년 약 10배 증가할 전망
- 다만, CXMT 공급량만 고려 시 중국 HBM 수요 대비 공급 부족분은 2027년/28년에도 여전히 40~50%에 달할 것으로 예상
- XMC, JHICC, Swaysure 등 다른 중국 메모리 제조업체들도 HBM 생산능력을 지속적으로 확보할 것으로 예상되며, 현지 업계 추정에 따르면 2030년 말 중국 전체 DRAM 웨이퍼 생산능력은 월 140만 장에 달할 전망

CXMT DRAM 및 HBM 생산능력 전망

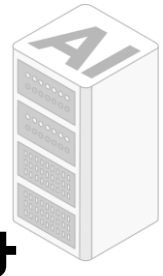


자료: Epoch AI, SemiAnalysis, 하나증권

중국 HBM 수요와 공급 부족분 전망 (CXMT 공급량만 고려)

구분	2026E	2027E	2028E
중국 HBM 수요	547	1,218	3,086
CXMT HBM 공급	125	687	1,179
자국산 공급 부족분	422	531	1,907
자국산 HBM 수요 충족률	23%	56%	38%
자국산 공급 부족률	77%	44%	62%

주: 공급 부족분은 중국 전체 수요에서 CXMT 공급을 차감한 값. 해외산 HBM 재고 및 추가 수입을 포함한 실제 미충족 수요는 아님
자료: Epoch AI, SemiAnalysis, 하나증권



III. 2027년 중국의 AI 연산능력 규모 전망

하나증권 리서치센터 해외주식분석실

중국기업분석 백승혜

중국의 전체 AI 연산능력 규모는 2027년 YoY 80% 증가 전망

- 중국의 AI 연산자원은 2027년 전년 대비 약 두 배 확대될 전망이지만, 여전히 약 64%는 미국산 반도체에 의존할 전망
- 중국 기업이 AI 칩을 확보할 수 있는 4개 경로는 1)중국 자체 생산, 2)미국산 AI 칩의 합법 수입과 3)불법 반입, 또는 4)해외 DC 이용
- 최근 미중 정부 모두 미국산 GPU 접근 규제 강화 중, 대형 DC 계약 취소, 규제 추가 강화 시 중국 자체 생산 의존도 높아질 가능성
- '27년 미중 연산능력은 모두 빠르게 증가할 전망이나, 미국이 2배 빠르게 증가해 연산능력 격차는 9배에서 14배로 더욱 확대될 전망

중국 AI 칩 공급 경로별 누적 확보 가능 물량 (만 H100 환산 기준)

공급 경로	누적 연산능력 규모		비중(%)		증가율 (YoY)
	26E	27E	26E	27E	
중국 자체 생산	173	303	37.6	36.2	+75%
미국산 합법 수입	85	135	18.5	16.1	+59%
미국산 비공식·불법 반입	100	160	21.7	19.1	+60%
해외 데이터센터 원격 이용	105	238.2	22.8	28.5	126.8571
전체 연산자원	**약 460**	**약 836.2**	**100.7**	**100.0**	**약 +82%**

주: 각 연도 말 누적 보유/접근 가능 용량 기준, 2027년 해외 데이터센터 연산능력에는 텐센트-오라클 DC 제외
자료: Trendforce, Epoch AI, SemiAnalysis, WSJ, IFP, 하나증권

미중 전체 누적 연산능력 규모 전망 (만 H100 환산 기준)

	2026E	2027E	YoY(%)
중국 AI 연산자원	460	836	81.8
미국 AI 연산자원	4,150	11,700	181.9
미중 격차	**약 9배**	**약 14배**	-

주: 각 연도 말 누적 보유/접근 가능 용량 기준
자료: Trendforce, Center for Technology and Statecraft, IFP, 하나증권

중국 AI 반도체 자체 생산의 핵심 병목은 HBM

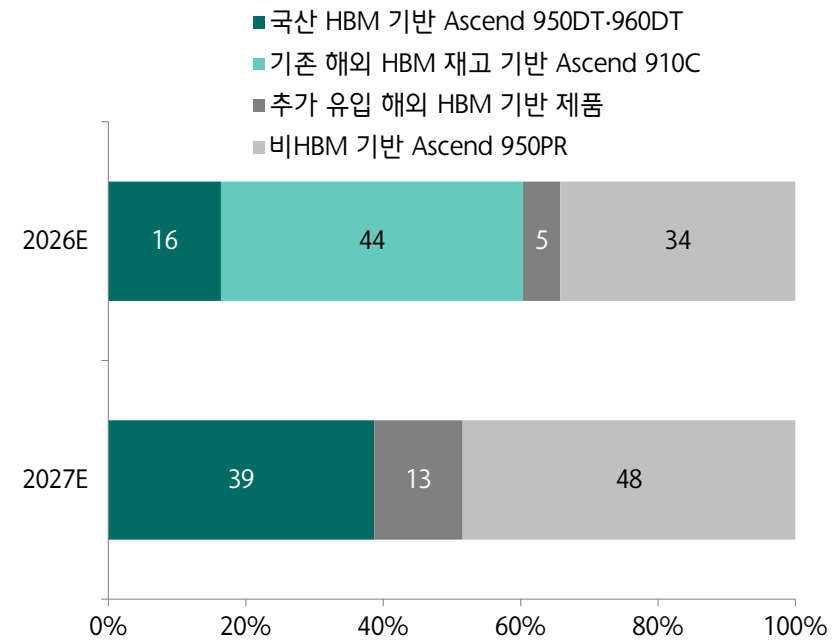
- 중국 자체 생산 AI 칩 공급량은 2027년 32% 증가 전망, 여전히 미국 연간 신규 생산량의 약 1~2% 수준에 불가
- 2027년에도 화웨이 생산 비중은 70% 수준 전망, 화웨이의 중국 HBM 의존도는 16%에 불과, 나머지 84%는 기존 해외 재고 또는 비공식 경로 반입. 2026년까지 해외 HBM 재고 활용한 Ascend 910C 생산 비중 압도적, 2027년부터 중국산 HBM 사용한 950DT/960DT와 비HBM 제품 중심 변화 예상

중국 AI 반도체 생산 규모 전망

	연간 신규 생산		증가율 (YoY,%)
	2026E	2027E	
화웨이	74	90	21.8
캠브리콘	6	10	71.4
기타 중국 AI 반도체 업체	19	30	56.5
중국 전체 생산량	**99**	**130**	**31.8**
중국 전체 생산량 (HBM 제약 없는 경우)	400	1,100	175.0
미국 생산량	**4,520**	**12,510**	**176.8**
미국 생산량 대비 중국 비중(%)	**2.2%**	**1.0%**	-
미국 생산량 우위(배)	**46배**	**96배**	-

주: 연간 신규 생산 기준, 캠브리콘은 Siyuan 590 생산량 기준
자료: Trendforce, Center for Technology and Statecraft, IFP, 하나증권

화웨이 HBM 공급 루트별 AI 칩 생산 비중



주: 연간 신규 생산 기준자료
자료: Epoch AI, SemiAnalysis, WSJ, Nikkei Asia, IFP, 하나증권

엔비디아 GPU 합법 수입 규모①: AI 전용 반도체

- 2026년 1월 미국 산업안보국(BIS)은 H200과 MI325X 중국 수출을 허용, 다만 미중 양국 규제로 실제 판매량 제한적
- H200 주문량이 200만개 이상으로 보도, '26년 10월까지 실제 공급량 약 2만개로 바이트댄스/텐센트가 각 1만개씩 확보 추정
- 엔비디아는 지난 7월 H200 실제 판매량 적어 초과 재고 발생, 관련해 4억달러 평가손실 인식했다고 공시한 바 있음
- 2026년 말까지 H200 누적 공급량 약 10만개 전망, 중국 AI 연산능력은 2025년 대비 YoY 23% 증가 효과 예상

미국/중국 정부 미국산 AI 반도체 수출입 규제 동향

발표일자	발표기관	규제/정책 주요 내용
2025-12-08	미국 대통령	트럼프 대통령이 승인된 중국 고객에 대해 NVIDIA H200 수출 허용 방침 발표
2026-01-13 발표 / 01-15 발효	미국 BIS	중국·마카오 대상 H200·MI325X(또는 동급 이하) 수출허가 심사를 원칙적 거부에서 건별 심사로 변경. TPP < 21,000; 총 DRAM 대역폭 < 6,500GB/s 기준
2026-01-14 발표 / 01-15 발효	미국 백악관	무역확장법 232조에 따라 특정 고성능 반도체(H200·MI325X)의 미국 수입에 25% 관세 부과해 중국향 H200 재수출 관세 부담 증가, 미국 내 사용 목적 제품은 면세
2026-01-14	중국 세관·당국	중국 세관에 H200 반입 금지 조치, 중국 기업에 불가피하지 않으면 구매하지 말라는 정부 지침 있어 사실상 수입 제한 조치
2026-03-03	미국 정부 (보도)	블룸버그는 미국 당국이 중국 고객사별 H200·MI325X 합산 구입 상한 7.5만개를 검토 중이라고 보도, 이후 실제 이행 중인 상황
2026-05~07	미국 BIS	중국 일부 기업 일부 구매 허가 발급, 7월 14일 BIS 고위 관계자가 의회에서 H200 대중국 출하 시작을 확인, 다만 수량은 극히 제한적이라고 설명
2026-09-27	중국 MIIT (보도)	중국 공업정보화부가 바이트댄스·알리바바 등에 RTX PRO 5500 구매 계획 제출 요구, 일부 구매 승인 가능성을 시사
2026-11-05	중국 정부·세관 (보도)	로이터는 엔비디아의 H200 AI 반도체 약 60만개가 중국 정부의 대규모 수입 승인 지연으로 홍콩에 보관 중이라고 보도

자료: 하나증권

엔비디아 GPU 합법 수입 규모②: 게임/워크스테이션용 반도체

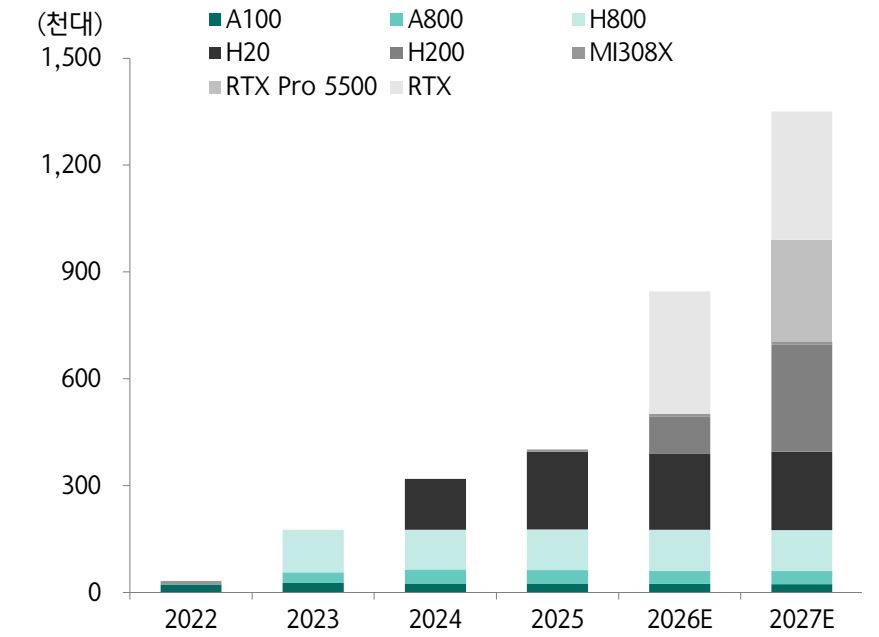
- 중국은 AI 전용 반도체 외에 게임/워크스테이션용 엔비디아 GPU도 대규모로 구매해 추론용 연산능력 확대 노력 중
- 2026년 중국은 엔비디아 RTX 시리즈 120만개 수입 예상, 연간 H200 수입 통해 신규 확보한 연산능력의 3.4배 수준
- 1Q27 출하 예정인 RTX Pro 5500 중국 주문량 600만개 상회, 엔비디아 공급 역량 부족으로 최대 100만개 수준 확보 예상
- 2027년 H200 추가 공급과 RTX Pro 5500 확보로 합법 수입 물량 통한 중국 AI 연산능력은 YoY 60% 증가 효과 전망

연도별 중국의 미국산 AI/게임용 반도체 수입 모델 및 규모

수입년도	모델명	수입량(개)	H100 환산기준 연산능력
2022년	A100	72,696	31,791
	A800	28,132	
2023년	A800	93,286	144,842
	H800	115,429	
2024년	H20	958,410	143,348
2025년	H20	474,712	81,525
	MI308X	31,859	
2026년 (예상)	H200	100,000	100,000
	RTX Pro 6000D/ RTX Pro 5000/ RTX 5090D	1,200,000	344,400
2027년 (예상)	H200	-	200,000
	RTX Pro 5500	1,000,000	300,000

자료: Epoch AI, Financial Times, IFP, 하나증권

중국 GPU별 누적 수입량 (H100 환산기준 연산능력)

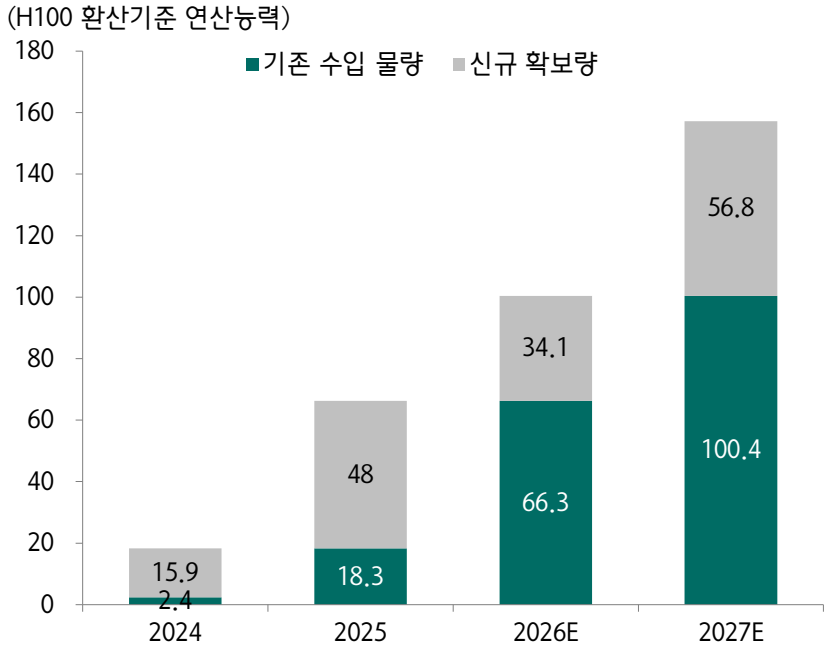


자료: Epoch AI, Financial Times, IFP, 하나증권

미국산 AI 반도체 비공식 유입 규모: 2027년 기여도 약 20% 예상

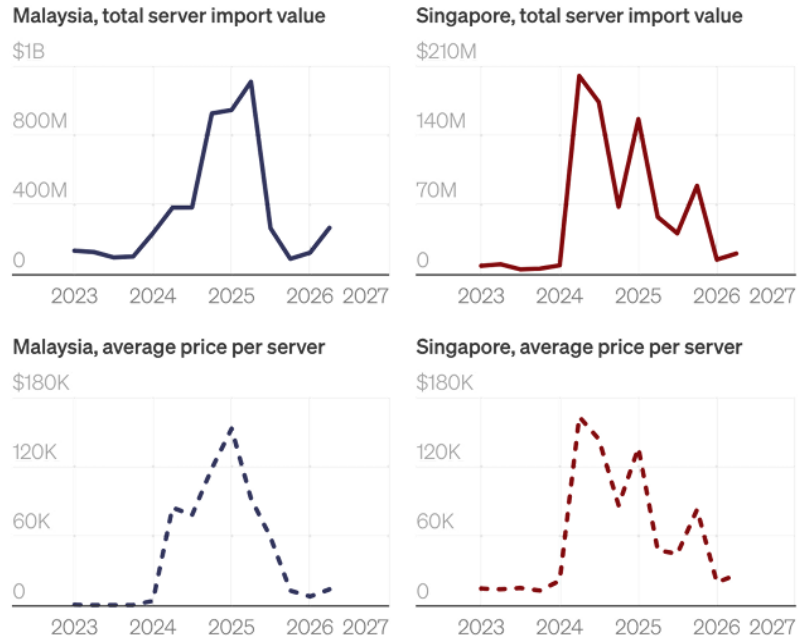
- 2023년 10월 미국의 대중국 반도체 규제 강화 이후 동남아 지역을 통한 중국의 AI 칩 밀수 규모 크게 증가
- 2026년 중국 전체 AI 연산능력의 약 20%가 비공식 경로 통해 확보 추정, 2027년은 전년비 60% 증가해 기여도 유사할 전망
- 말레이시아 정부의 AI 칩 재수출/환적 규제 강화, 엔비디아의 구매자 검증 강화로 2026년 신규 유입 전년비 감소 전망
- 중국 밀수 규모가 매년 미국의 AI 칩 생산능력의 약 3% 수준 차지, 미국 공급량 확대 시 밀수 규모도 함께 확대될 가능성

중국의 AI 반도체 비공식 유입 규모 추정치



자료: Epoch AI, Financial Times, IFP, 하나증권

말레이시아/싱가포르 서버 수입액 및 평균 가격



자료: Epoch AI, IFP, 하나증권

해외 데이터센터 원격 이용 규모: 중국 AI 인프라 확대의 핵심 경로

- 중국 AI 연산능력 확보의 마지막 경로는 중국 외 지역에서 첨단 GPU 기반 AI 서버를 직접 설치하거나 임대하는 방식
- 2027년 해외 DC 통한 AI 연산능력은 중국의 4개 경로 중 가장 빠르게 증가(2.4배)해 '27년 비중 약 29%('26년 23%)로 확대될 전망
- 최근 미국/중국 정부 압박으로 텐센트x오라클의 GPU 10만개 임대 계약 취소된 점 감안 시 당분간 규제 강도 강화 흐름 예상

중국 AI 칩 공급 경로별 누적 확보 가능 물량 (만 H100 환산 기준)

지역	H100 환산기준 연산능력		비중	
	26년	27년	26년	27년
동남아시아	79.5만	200만	75%	72%
기타 아시아태평양	8.9만	24.4만	8%	9%
유럽·중동·아프리카	6.4만	25.5만	6%	9%
미국 및 중남미	10.7만	29.9만	10%	11%
합계	110만	280만	100%	100%

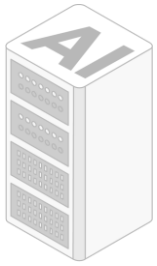
주: H100(FP8) 환산기준 연산능력
자료: SemiAnalysis, IFP, 하나증권

바이트댄스/텐센트 해외 DC 연산자원 현황

클러스터	지역	GPU 구성	가동 예상 시점	연산능력 (H100 환산기준)
텐센트 × 오라클	동남아시아	GPU 10만개	2027년 말	45.2만(80%)/ 56.5만(100%)
바이트댄스 × 오라클	동남아시아	NVIDIA B200	2026년 말	26.1만
바이트댄스 × Aolani	동남아시아	NVIDIA B200 3.6만개	2027년 말	9.1만
텐센트 × Datasection(시드니)	호주	NVIDIA B300 1만개	2027년 말	2.5만
텐센트 × Datasection(지바)	일본	NVIDIA B300 5,080개	2027년 말	1.3만
바이트댄스 × Nscale	노르웨이	NVIDIA B200 2,304개	2026년 말	0.6만
구글 프라이어 노스	미국	Google TPU v5p·v5e	가동 중	63.7만
메타 프로메테우스	미국	NVIDIA B300	가동 중	53.6만

주: FT와 로이터에 따르면 텐센트-오라클 DC는 미국 트럼프 행정부의 압박으로 GPU 임대 계약을 취소
자료: SemiAnalysis, IFP, 하나증권

IV. 중국의 AI 모델 개발 전략



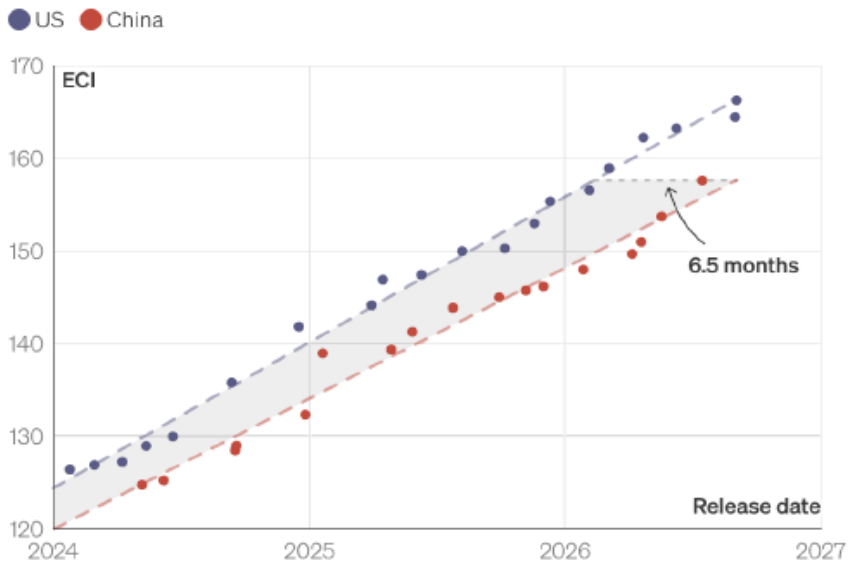
하나증권 리서치센터 해외주식분석실

중국기업분석 백승혜

미중 AI 모델 종합 성능 격차는 약 6.5개월에 불과

- 미중 연산능력 격차는 2027년 14배로 더욱 확대될 전망이나, 최상위 AI 모델의 종합적인 성능 격차는 약 6.5개월에 불과
- IFP는 여러 AI 벤치마크 결과를 통합해 성능 과대평가 가능성을 배제한 Epoch Capabilities Index(ECI)를 활용해 격차를 측정
- 소프트웨어 엔지니어링 역량, 즉 개발 업무 자율 수행 능력(Time Horizon) 기준으로 측정해도 미중 격차는 약 6.6개월로 유사
- 다만, 미국의 ECI 점수 상승 속도가 중국보다 약 11% 높기 때문에 장기적으로 격차가 조금씩 벌어지는 추세가 관찰

Epoch AI 미중 모델 성능 격차 추이



주: ECI(Epoch Capabilities Index)는 특정 벤치마크에 의존하지 않고 다양한 벤치마크를 활용해 AI 모델의 종합적인 역량을 추정. 모델의 성능 왜곡이나 과대평가를 야기하는 데이터 오염, 벤치마크 포화 등의 문제를 최소화하도록 설계
자료: Epoch AI, IFP, 하나증권

미중 주요 AI 모델의 성능 발전 속도

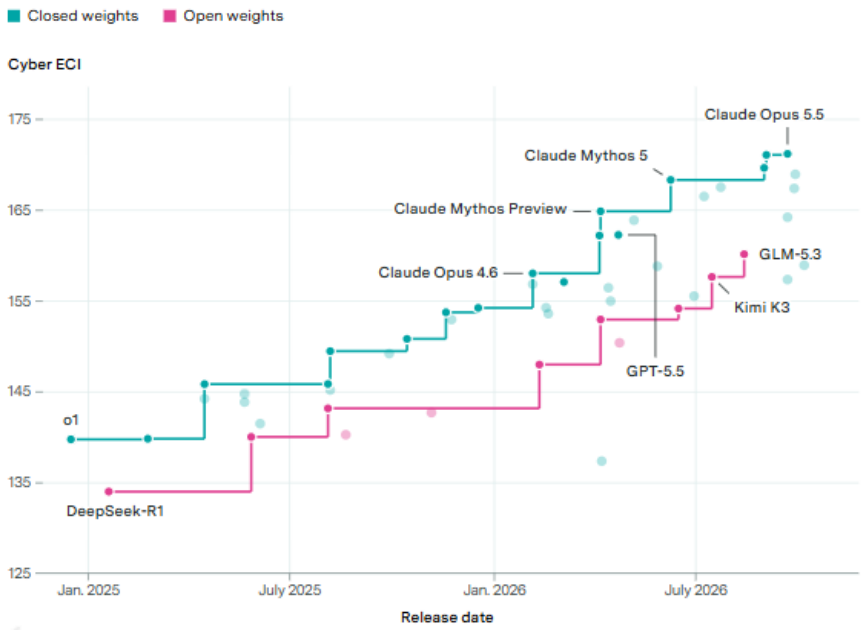
구분	ECI (범용 AI 성능)	METR (소프트웨어 엔지니어링)
평가 내용	추론·코딩·과학 등 종합 성능	자율적인 소프트웨어 개발 작업 수행 능력
미국의 성능 발전 속도	월 1.31 ECI점 상승	3.3개월마다 수행 시간 2배
중국의 성능 발전 속도	월 1.18 ECI점 상승	3.7개월마다 수행 시간 2배
중국의 미국 대비 기술 시차	6.5개월	6.6개월
분석 기간	2024~26년	2024년 3월~2026년 4월
최신 분석 모델	GPT-6 Astra(26.9), Kimi K3(26.7)	Claude Mythos Preview (26.4), Kimi K2 Thinking (25.11)

자료: Epoch AI, METR, IFP, 하나증권

Mythos급 사이버 역량 갖춘 모델은 2026년 말 출시 전망

- 2026년 4월 공개된 Claude Mythos는 사이버보안 성능의 비약적인 발전으로 국가안보에 심각한 우려를 불러일으킨 최초의 모델
- IFP는 종합 AI 성능과 SW 엔지니어링 성능 발전 속도(6.5~6.6개월) 기준 추정 시 중국이 '26년 말까지 Mythos 수준 AI 모델 출시 전망
- '26년 10월 Epoch AI도 GLM-5.3 포함 중국 최상위 모델들이 사이버보안 역량에서도 미국 모델 대비 약 6~7개월 뒤처져 있다고 분석

Epoch AI의 미중 주요 모델 사이버보안 성능 비교



주: Cyber ECI Index는 여러 사이버보안 벤치마크 결과를 종합해 산출한 AI 모델의 사이버보안 역량 지수
자료: Epoch AI, 하나증권

주요 AI 모델 사이버보안 성능 지표 순위

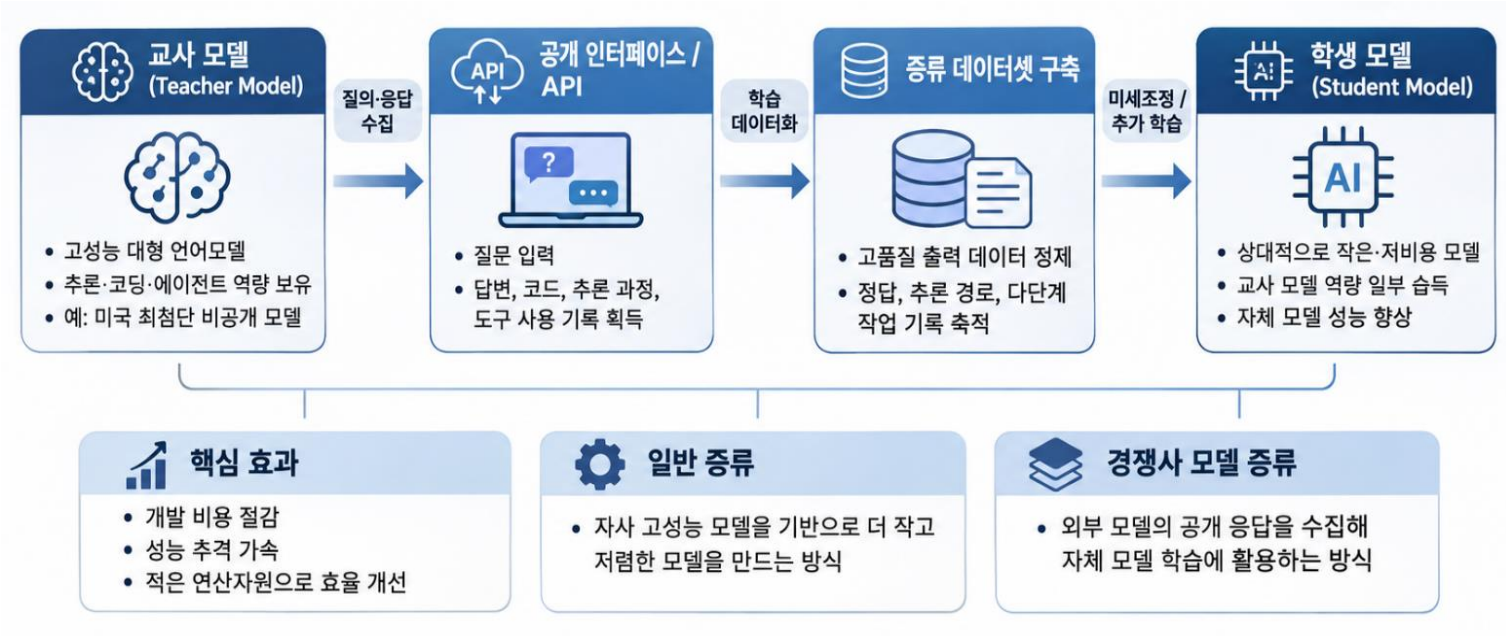
모델명	개발사	Cyber ECI 점수	출시일
Claude Opus 5.5	Anthropic	171.1	2026년 9월
GPT-6 Astra	OpenAI	171.0	2026년 9월
Claude Mythos 5	Anthropic	168.3	2026년 6월
Claude Mythos Preview	Anthropic	164.8	2026년 4월
GLM-5.3	Z.ai	160.1	2026년 8월
Claude Opus 4.6	Anthropic	158.0	2026년 2월
Kimi K3	Moonshot	157.6	2026년 7월
DeepSeek-V4-Pro	DeepSeek	150.4	2026년 4월
GLM-5	Z.ai	148.0	2026년 2월
Kimi K2 Thinking	Moonshot	142.7	2025년 11월
o1	OpenAI	139.7	2024년 12월
DeepSeek-R1	DeepSeek	134.0	2025년 1월

주: Cyber ECI Index는 여러 사이버보안 벤치마크 결과를 종합해 산출한 AI 모델의 사이버보안 역량 지수
자료: Epoch AI, 하나증권

중국 모델 기술격차 축소 방안(1): 증류(Distillation) ①

- 중국 AI 모델 개발업체들은 컴퓨팅 자원 부족을 해결하기 위해 합법적 수단과 불법적 수단을 모두 포함한 다양한 전략을 활용 중
- 중국은 모델 증류를 통해 미국 기업이 막대한 연산자원을 투입해 개발한 최첨단 AI 모델의 역량을 적은 컴퓨팅 자원만으로 확보 중
- 증류는 학습 완료된 최첨단 AI 모델의 출력 결과로 다른 AI 모델을 학습시키는 방식, 초기 학습에 필요한 연산자원, 비용 절감 가능
- 미국의 연산능력 우위가 커질수록 중국은 미국 모델 증류에 대한 유인과 이를 통해 얻을 수 있는 잠재적 효과가 더 커질 수 있음

AI 모델 증류 기술 구조도



자료: 하나증권

중국 모델 기술격차 축소 방안(1): 증류(Distillation)②

- 실제로 2026년 오픈AI와 앤트로픽은 딥시크, 문샷AI, 미니맥스가 자사 AI 모델을 대상으로 증류 활동을 수행했다고 보고
- 앤트로픽 보고서에 따르면 중국 AI 기업들이 Claude 모델의 추론 과정, 에이전트 코딩/도구 활용 역량 추출에 집중했다고 주장
- 또한, 미국 정부 기관 보안 권고문에도 중국 6개 기업이 2024년 말 이후 미국 AI 모델에서 수십억개의 토큰을 추출했다고 밝힘

앤트로픽이 공개한 중국 기업 Claude 증류 관련 상호작용 규모

기업명	증류 상호작용 규모	증류 대상
DeepSeek	15만회 이상	추론/보상 모델 기반 평가 등
Moonshot AI	340만회 이상	에이전트 추론/코딩/ 컴퓨터 활용 역량 등
MiniMax	1,300만회 이상	에이전트 코딩/ 도구 활용 역량

자료: 앤트로픽(2026.2.23), 하나증권

미국 정부 기관이 발표한 중국 AI 기업 리스트 및 증류 대상

중국 기업	미국 측이 지목한 주요 대상 모델	증류 목적·대상 역량
DeepSeek	GPT-4·5, Claude, Gemini, Grok	추론, 전문지식, 에이전트, SFT
Moonshot AI	GPT, Claude, Gemini, Grok	추론, 코딩, 수학, RL
Alibaba	GPT-5, Claude	소프트웨어 개발, 고객응대, 에이전트
MiniMax	Claude, Gemini, GPT	코딩, 에이전트, CoT 추론
StepFun	GPT-5 계열, Claude	코딩, 에이전트
Z.ai(즈푸AI)	GPT-5.5, Claude	추론 과정(CoT) 학습

자료: CISA(2026.9.8), 하나증권

중국 모델 기술격차 축소 방안(2): 아키텍처/알고리즘 혁신

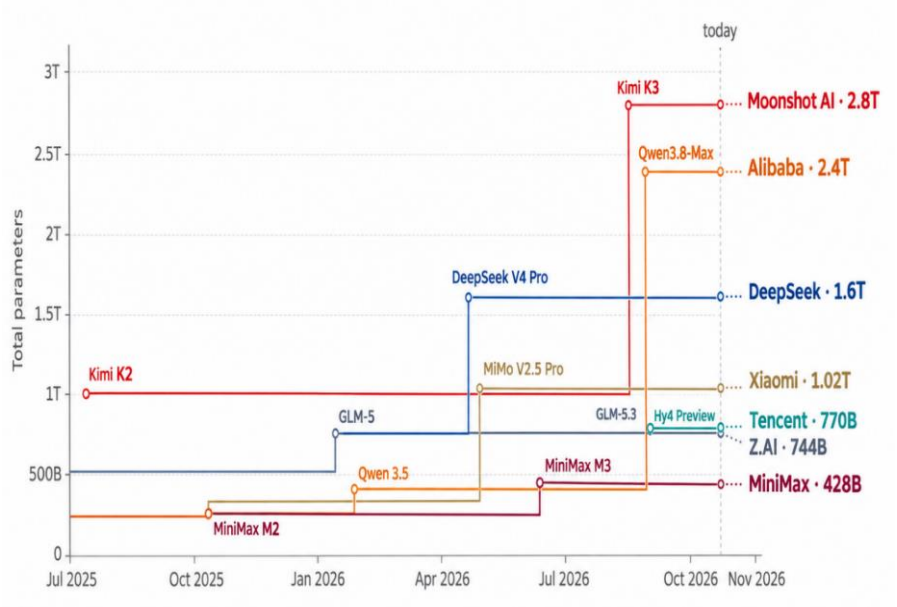
- 중국의 주요 AI 개발사들은 지난 몇년간 AI 모델의 효율성을 개선, 연산자원 소비를 줄이기 위해 아키텍처 및 알고리즘 혁신을 추진
- 2025년 초 딥시크 R1 모델 혁신으로 인한 글로벌 충격 이후 2026년 7월 문샷AI의 Kimi K3 모델의 아키텍처 혁신이 세계적으로 화두
- Kimi K3는 미국 최첨단 모델에 근접한 성능을 구현한 총 2.8조 파라미터의 초대형 오픈웨이트 모델로 출시 직후 다운로드 수요 급증
- Kimi의 Attention Residual 기술은 동일 성능 모델 대비 연산량 20% 절감, 학습 효율성 25% 개선, 추가 지연시간 2% 미만으로 억제

중국의 주요 AI 모델 아키텍터/알고리즘 혁신 사례

기업	모델명	핵심 기술
DeepSeek	DeepSeek-V3	MLA, DeepSeekMoE
	DeepSeek-R1	강화학습(RL), GRPO
Moonshot AI	Kimi K3	Attention Residuals, KDA (Kimi Delta Attention), Stable Latent MoE
Alibaba	Qwen3	MoE, Hybrid Reasoning
ByteDance	FlexPrefill	Sparse Attention

자료: 하나증권

중국 주요 오픈웨이트 모델 파라미터 비교



자료: Kimi.ai, 하나증권

중국 모델 기술격차 축소 방안(2): 아키텍처/알고리즘 혁신

중국의 주요 AI 모델 아키텍터/알고리즘 혁신 사례

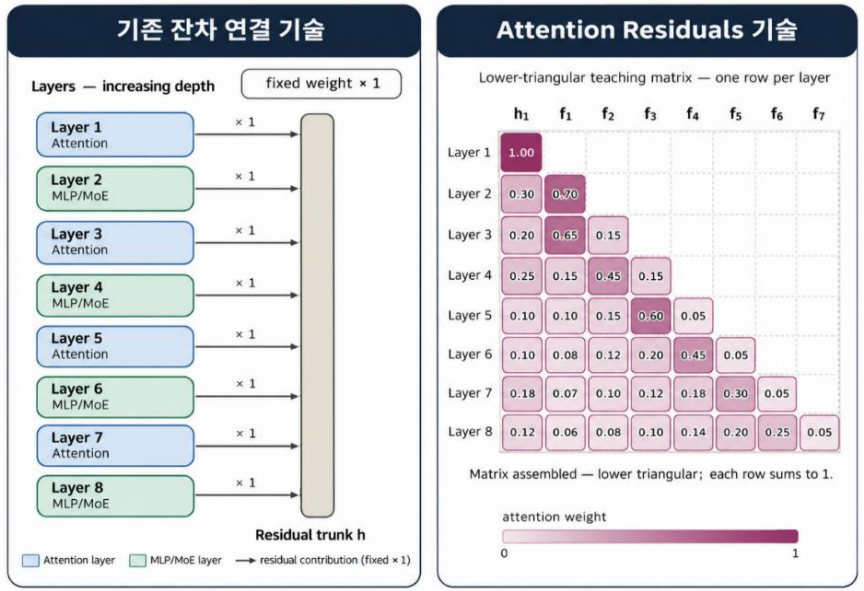
출시 시점	기업	모델명	주요 혁신 기술	기술 설명
2024.12	DeepSeek	DeepSeek-V3	MLA, DeepSeekMoE	총 6,710억 개 파라미터 중 토큰당 370억 개만 활성화하는 MoE 구조를 도입. MLA(Multi-head Latent Attention)를 통해 KV 캐시 부담을 축소하고, FP8 혼합정밀도 학습 등으로 연산 효율을 개선. 1,480억이 아닌 14.8조 개 토큰을 사전 학습하는 데 약 278.8만 H800 GPU 시간을 사용했다고 발표.
2025.01	DeepSeek	DeepSeek-R	강화학습, GRPO	대규모 인간 작성 추론 예시 없이도 강화학습을 통해 추론 능력을 개선할 수 있음을 제시. GRPO(Group Relative Policy Optimization) 등을 활용해 추론 경로의 품질을 개선하고, 별도로 R1의 추론 능력을 작은 모델에 증류하는 방식도 적용. OpenAI o1에 근접한 추론 성능을 구현하면서도 API 이용료를 약 96% 낮춘 파격적인 가격 경쟁력으로 글로벌 AI 시장에 충격. 출시 후 한 달간 허깅페이스 다운로드 약 420만 회 기록, 중국 오픈웨이트 AI 모델 글로벌 확산을 촉발.
2025.02	ByteDance	FlexPrefill	Sparse Attention	긴 입력 문맥에서 중요도가 높은 토큰 간 관계를 선택적으로 계산하는 희소 어텐션 기술. 입력 내용과 어텐션 헤드 특성에 따라 연산 범위를 동적으로 조절해 긴 문맥 추론의 연산 부담을 줄임.
2025.04	Alibaba	Qwen3	MoE, Hybrid Reasoning	Qwen3-235B-A22B는 총 2,350억 개 파라미터 중 220억 개만 활성화. 복잡한 문제에는 Thinking Mode, 단순 작업에는 Non-thinking Mode를 활용해 추론 비용을 최적화.
2026.07	Moonshot AI	Kimi K3	Attention Residuals, KDA, Stable Latent MoE	Attention Residuals·KDA·Stable LatentMoE 등 아키텍처 혁신으로 전작 대비 학습 스케일링 효율 약 2.5배 개선. 미국 최첨단 비공개 모델에 근접한 성능을 구현한 총 2.8조 파라미터의 초대형 오픈웨이트 모델로, 출시 직후 다운로드 수요가 급증해 서비스 인프라에 과부하가 발생할 정도로 글로벌 관심이 집중. Attention Residuals로 계층 간 정보 활용을 개선하고, Kimi Delta Attention으로 긴 문맥 처리 효율을 향상. Stable LatentMoE는 전문가 선택과 학습 안정성을 개선.

자료: 하나증권

Kimi K3의 핵심 아키텍처 기술인 "Attention Residuals"

- "Attention Residuals"은 Kimi K3에 적용된 핵심 기술 중 하나로, 동일한 GPU 연산자원으로 더 높은 AI 모델 성능 구현에 기여
- 기존 Transformer 모델은 일반적으로 잔차 연결 구조 사용, 신경망 계층 통과 시 이전 계층의 정보 모두 누적해 다음 계층으로 전달
- Attention Residuals는 가중치에 기반해 이전 계층 정보를 무조건 누적하지 않고 필요한 정보를 선택적으로 활용하도록 개선한 기술
- 또한, 여러 계층을 하나의 블록(Block)으로 묶어, 개별 계층의 결과 대신 블록 단위의 정보를 저장/활용해 연산자원 사용을 줄임

기존 잔차 연결 구조와 Attention Residuals 기술 비교



주: 왼쪽 그림은 AI 모델이 8개 계층을 거치면서 생성한 정보를 하나의 잔차 경로에 계속 누적하는 구조, 오른쪽 그림은 각 계층이 이전 계층의 정보를 얼마나 활용하는지를 나타낸 어텐션 가중치 행렬, 각 행은 현재 계층, 각 열은 참조할 수 있는 이전 계층의 정보를 의미
자료: 하나증권

Kimi K3 주요 벤치마크 성능도 개선 효과

벤치마크	기존 잔차연결 기술	Attention Residuals 기술	성능개선 폭
GPQA Diamond (과학추론)	36.9%	44.4%	+7.5%p
MATH (수학)	53.5%	57.1%	+3.6%p
HumanEval (코딩)	59.1%	62.2%	+3.1%p
MMLU (종합지식)	73.5%	74.6%	+1.1%p

자료: Kimi.ai, 하나증권

오픈웨이트 모델로 빠른 글로벌 확산과 중국 AI 생태계 영향력 확대

- 오픈웨이트 모델은 오픈소스 개발/배포 플랫폼, 제3자 추론 플랫폼, 클라우드 등을 통해 빠른 확산이 가능한 점이 가장 큰 장점
- 중국 개발사 서버에 데이터 보내지 않고 자체 서버, 제3자 호스팅 통해 활용 가능해 기업, 정부 기관의 정치적/운영상 리스크를 낮춤
- 또한, 연산자원 제약 때문에도 중국 기업 자체 GPU 인프라만으로 글로벌 추론 수요 감당이 어려워 오픈웨이트 전략은 불가피한 선택
- 정책적으로도 중국 AI 모델의 글로벌 영향력 확대를 통한 중국 AI 산업의 글로벌 위상과 중국 기술 의존도 강화를 목표 중

중국의 오픈웨이트 모델 보급 비중 압도적

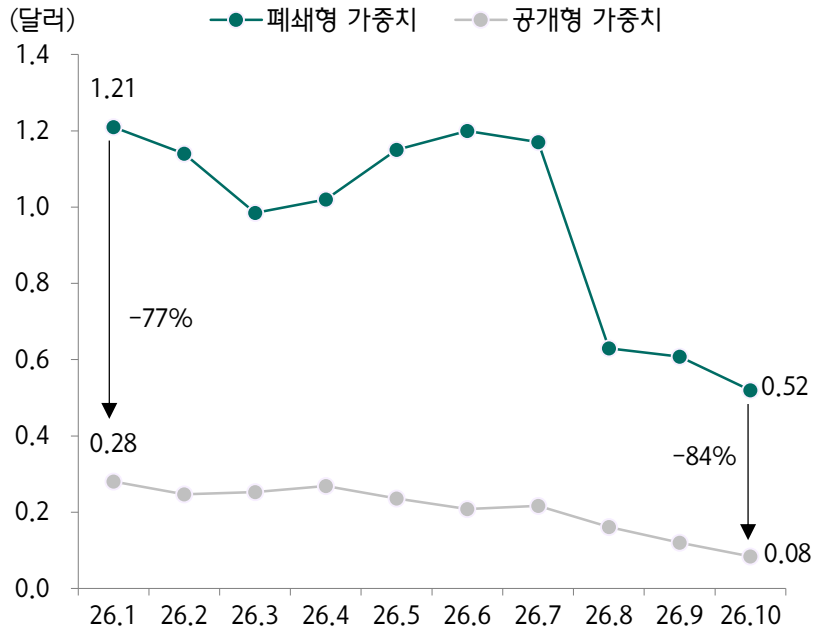
국가	기업	전체 모델 수	오픈웨이트 수	비중
중국	DeepSeek	8	8	100.0%
	MiniMax	4	4	100.0%
	Z.ai (Zhipu)	4	4	100.0%
	Alibaba Qwen	9	8	88.9%
	Moonshot AI	5	4	80.0%
미국	Meta	14	11	78.6%
	Google	15	6	40.0%
	xAI	7	2	28.6%
	OpenAI	27	3	11.1%
	Anthropic	24	0	0.0%

주:2026년 10월까지 역대 출시된 총 159건의 글로벌 주요 LLM 분석 기준
자료: agitracker.io, 하나증권

중국 AI 모델의 저렴한 추론 비용은 글로벌 기업에 너무 매력적

- 최근 기업들은 AI 모델 사용량을 줄이지 않으면서 비용 통제를 하기 위해 복잡도가 낮은 반복 작업에 저렴한 중국 모델 도입 확대 중
- 미국 코인베이스는 중국 GLM5.2/Kimi2.7 모델 사용으로 개발자들의 토큰 사용량 급증에도 사내 AI 지출을 약 50% 줄였다고 발표
- OpenRouter에 따르면 중국 오픈웨이트 모델 실제 이용 단가는 앤트로픽/오픈AI 주요 모델 대비 약 60~90% 저렴한 수준으로 추정
- 중국 최상위 모델에 속하는 GLM5.3 입력 토큰 100만개당 가격은 1.4달러로 GPT5.5/Opus5.5의 5달러/4달러 대비 72%/65% 저렴

실효 API 이용 단가: 폐쇄형 Vs. 오픈웨이트 가격 격차 더욱 확대



주: 실제 이용 단가(Effective Price)는 OpenRouter에서 개별 모델별로 실제 청구된 API 사용금액을 전체 입력·출력 토큰 수로 나눈 실효 평균 가격으로 실제 사용량, 모델 구성, 캐시 할인 등을 모두 반영
자료: OpenRouter, 하나증권

미중 주요 최상위 AI 모델 API 가격 비교

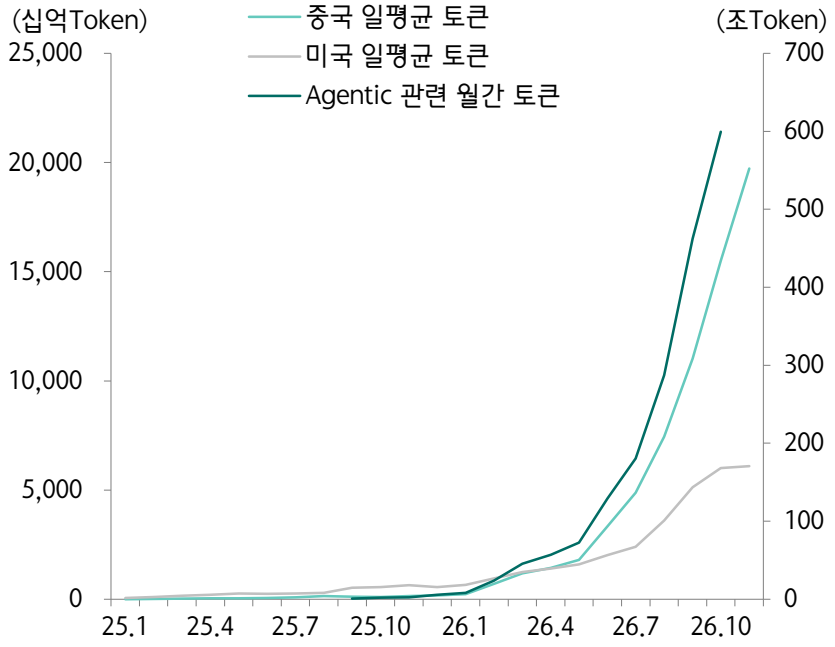
AI 모델	개발사	입력 토큰 100만 개당 가격	출력 토큰 100만 개당 가격	출시일
중국				
Kimi K3	Moonshot AI	$3.00	$15.00	2026.7
Qwen3.8-Max	Alibaba	$1.65	$4.95	2026.8
GLM-5.3	Z.ai	$1.40	$4.40	2026.8
V4 Pro	DeepSeek	$1.32	$3.96	2026.4
M3	MiniMax	$0.30	$1.20	2026.6
미국				
GPT-6 Astra	OpenAI	$10.00	$50.00	2026.9
Claude Fable 5.2	Anthropic	$8.00	$40.00	2026.11
GPT-5.5	OpenAI	$5.00	$30.00	2026.4
Claude Opus 5.5	Anthropic	$4.00	$20.00	2026.9

주: 일반 API 표준 요금 기준
자료: 각 사, 하나증권

2026년 Agentic AI 발전과 함께 중국 모델 수요 폭발적으로 증가

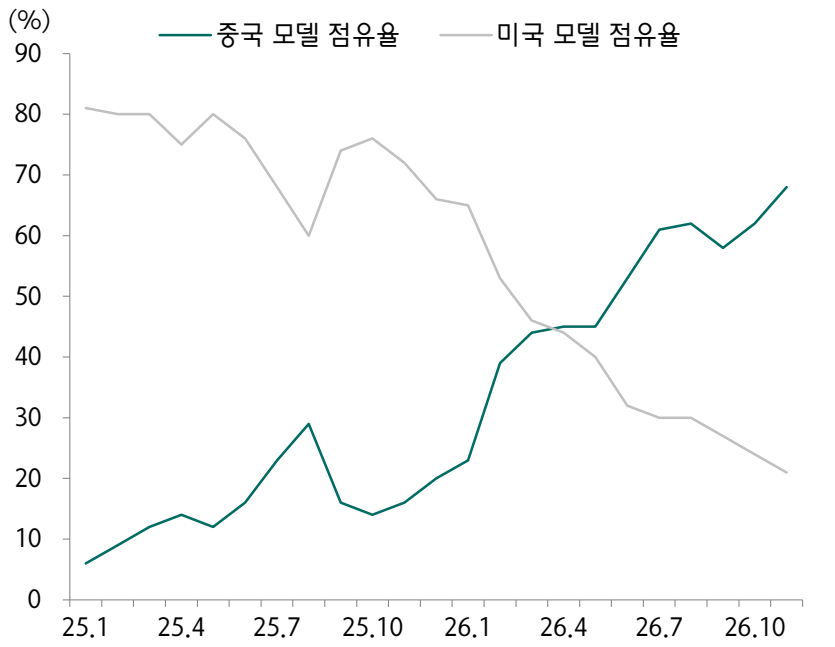
- 중국의 AI 모델 수요는 2026년 초 AI 에이전트 플랫폼의 빠른 확산과 함께 글로벌 개발자/기업들의 사용량이 급증하는 흐름
- AI 에이전트는 AI 모델을 반복적으로 호출해 작업 계획, 코드 작성, 테스트/수정하며 챗봇 대비 훨씬 많은 토큰을 소비
- AI 모델 호출 횟수가 크게 증가하면서 토큰당 가격이 기업의 AI 운영 예산을 결정하는 중요한 변수이자 제약 요인으로 부상
- 이는 글로벌 기업과 개발자들이 저비용, 고성능의 중국 AI 모델 채택을 빠르게 확대하는 흐름으로 이어지고 있음

중국 AI 모델 오픈라우터 토큰 처리량 Agentic 수요와 함께 급증



자료: OpenRouter, 하나증권

오픈라우터 중국 토큰 점유율 4월 미국 역전 이후 11월 68%까지 상승

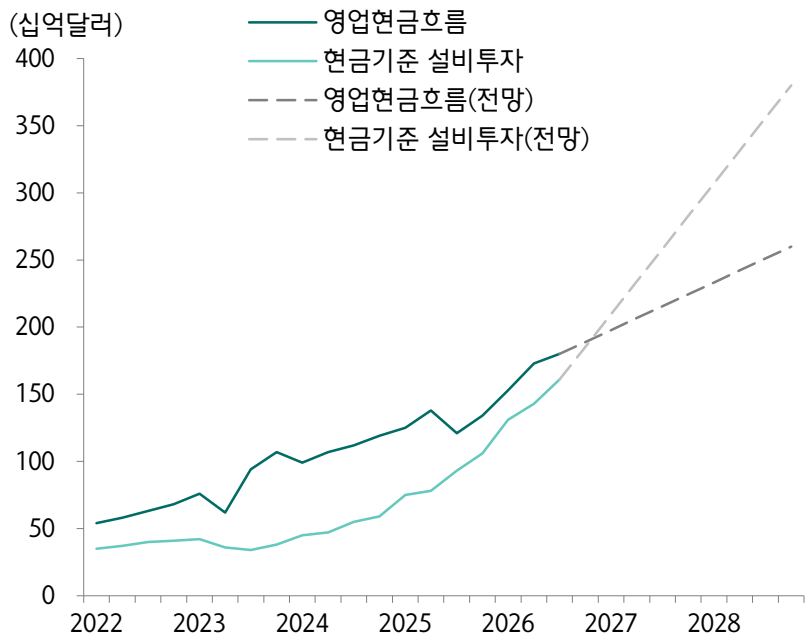


자료: OpenRouter, 하나증권

중국 AI 모델의 가격 경쟁력이 미국 하이퍼스케일러에 미치는 영향

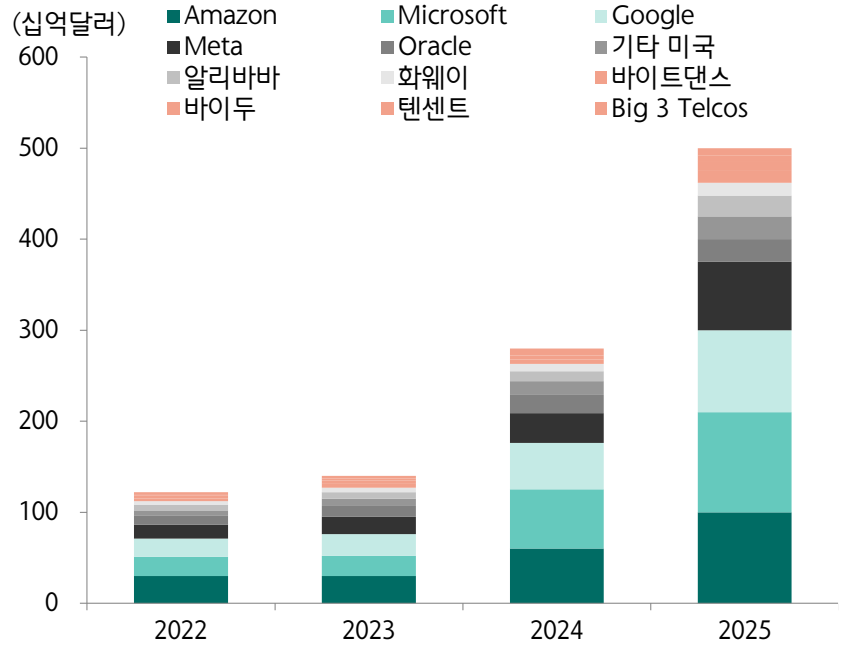
- AI 에이전트 확산으로 토큰 사용량이 빠르게 증가해도 중국 오픈웨이트 모델의 가격 경쟁으로 토큰당 매출 하락 시 미국 하이퍼스케일러들이 가정한 매출 성장률에 하방 압력 발생, 막대한 투자액에 대한 자본회수 기간, 감가 부담, 부채상환능력 전망에 영향 예상
- 다만 AI 추론 단가 하락은 신규 수요를 창출하고 기존 작업의 AI 전환을 촉진해 전체 토큰 사용량을 추가로 확대할 수 있음
- 따라서 최종 매출 영향은 토큰당 가격 하락률과 실제 연산 수요 증가율의 상대적인 크기에 의해 결정될 것

MS/아마존/알파벳/메타/오라클 합산 현금흐름 및 설비투자 추이



자료: Epoch AI, 하나증권

글로벌 주요 빅테크 기업의 데이터센터 인프라 설비투자 규모



자료: Rhodium Group, Omdia, 하나증권

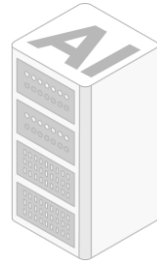
중국 AI 모델 궁극적 목표는 매출보다 AI 생태계 확대

- 중국 정부는 AI 모델의 빠른 글로벌 확산과 AI 생태계 영향력 확대를 통해 중국 AI 산업의 글로벌 위상을 높이고, 궁극적으로 글로벌 시장의 중국 AI 기술 의존도를 높이는 것이 전략적 목표
- 따라서 중국은 오픈웨이트 AI 모델의 가격 경쟁력과 개방성을 활용해 글로벌 시장에서 영향력을 확대하고, 이를 중국산 AI 서비스, 디바이스, 반도체, 데이터센터 등 AI 기술 스택 전반에 대한 의존도 확대로 연결하려 할 가능성이 높음

중국 국무원, 2025년 8월 <AI+ 정책> 오픈소스 생태계 육성

제11항: 오픈소스 생태계 발전 촉진
• 오픈소스 기여 인센티브: AI 오픈소스 기여도를 평가하고 보상하는 제도 마련
• 대학·연구기관 지원: 학생의 오픈소스 기여를 학점으로 인정하고, 교원의 연구성과 평가에 반영하도록 장려
• 기업의 오픈소스 활용: 기업·대학·연구기관의 효율적인 오픈소스 AI 활용 방식 개발 지원
• 글로벌 생태계 육성: 전 세계에 개방된 기술체계와 커뮤니티를 구축, 국제적 영향력 갖춘 오픈소스 프로젝트/개발 도구 육성
제6항: AI+ 글로벌 협력
• AI 기술의 오픈소스화와 접근성 확대를 추진하고, 컴퓨팅 자원·데이터·인재 등 분야에서 국제협력을 강화

자료: 중국국무원, 하나증권



V. 2027년 중국 주식 투자 유망종목

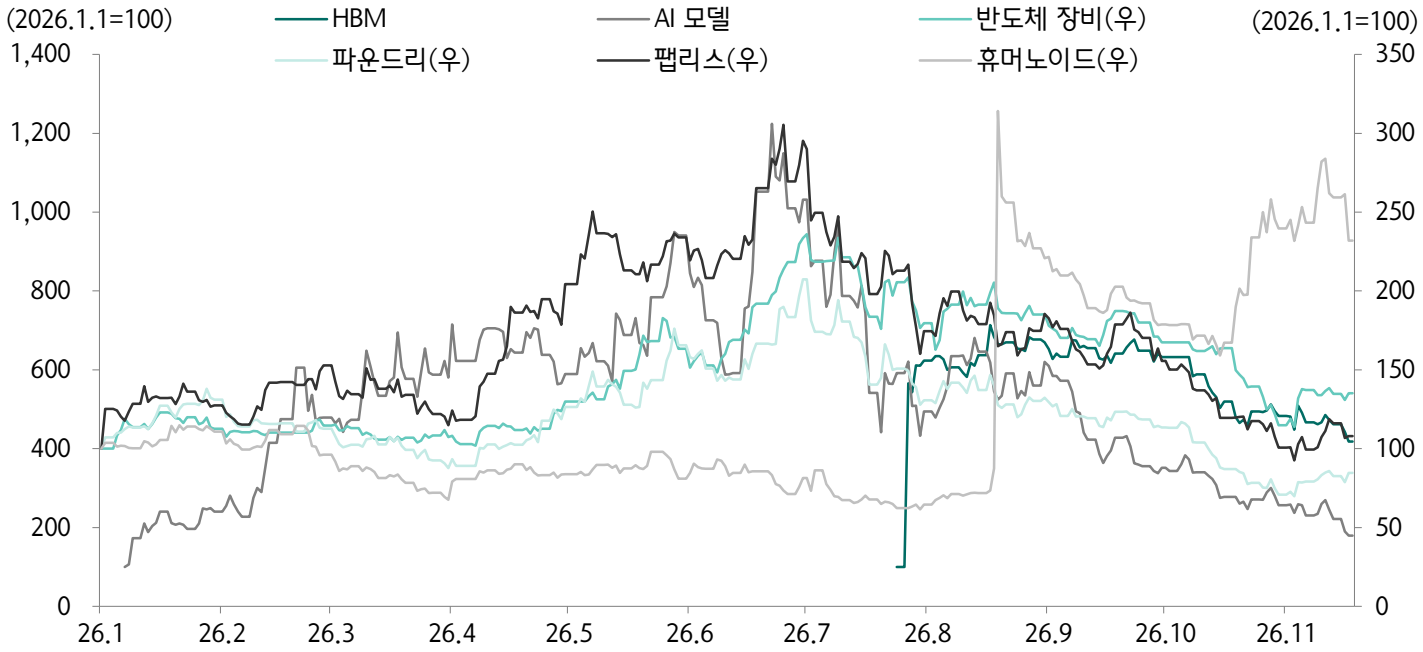
하나증권 리서치센터 해외주식분석실

중국기업분석 백승혜

중국 AI 밸류체인 선호도는 장비>메모리>휴머노이드>팹리스 순

- CXMT/YMTC 중심으로 메모리 업종 대규모 증설과 장비 국산화 가속화 기대, Piotech/AMEC은 대표적인 수혜주로 지속 관심 필요
- CXMT는 HBM 생산 확대, 차세대 기술 기대감이 주가에 긍정적으로 반영될 전망, 대주주 보호예수 해제로 상반기 변동성 확대 경계
- AI 반도체 팹리스 업종은 '27년 화웨이를 제외한 나머지 기업들의 HBM 공급 병목 더욱 심화될 것으로 예상, 출하량 증가 전망 보수적
- 휴머노이드는 출하량 기대감, 피지컬 AI 상장사 프리미엄 고려해 유니트리 선호, AI 모델 업종은 수익화 난제가 주된 하방 요인

중국 AI 밸류체인 섹터별 연초 이후 주가 흐름



자료: Bloomberg, 하나증권

중국 AI 밸류체인 추천 종목 리스트

업종	기업명	티커	투자 포인트
반도체 장비	Piotech	688072 CH	• PECVD 중심에서 ALD·Flowable CVD 등 고부가가치 증착장비로 제품군 확대, 첨단 공정 대응력 강화 • 하이브리드 본딩 장비 상용화를 통해 기존 전공정에서 첨단 패키징 시장으로 사업 영역 확장 • 신규 장비의 고객사 양산 적용 본격화에 따른 매출 성장 및 제품 믹스 개선 기대
	AMEC	688012 CH	• 3D NAND 고단화에 따른 고종횡비 식각 공정의 기술적 난도 상승으로 첨단 식각장비의 중요성 부각 • 첨단 식각 기술 경쟁력을 바탕으로 글로벌 장비 업체가 주도하던 고부가가치 공정으로 공급 영역 확대 • 중국 반도체 제조사의 첨단 공정 전환과 신규 식각장비 채택 확대가 중장기 실적 성장 동력
	북방화창	002371 CH	• 식각·증착·열처리·세정 등 전공정 전반을 아우르는 종합 장비 포트폴리오를 기반으로 고객사 내 공급 범위 확대 • CXMT·YMTC의 신규 생산라인 투자에 따른 다양한 공정 장비의 동시 공급으로 수주 규모 증가 기대 • 다각화된 제품군과 고객사 기반을 바탕으로 중국 반도체 장비 시장 내 입지 강화 및 안정적인 성장 기반 확보
메모리	CXMT	688825 CH	• 글로벌 DRAM 3사 대비 낮은 경쟁 강도와 중국 내 서버용 DRAM 수요 확대 기반 '28년까지 높은 영업이익률 유지 전망 • DRAM 생산능력 2Q26 월 27만 장에서 3Q26 월 28만 장으로 확대, 차세대 $4F^2$ DRAM 구조 도입에 따른 원가 경쟁력 강화 • 2027년 예상 PER 8.7배로 중국 반도체 업종 평균 34배 대비 저평가되어 있어 밸류에이션 매력 부각
휴머노이드 로봇	유니트리	688836 CH	• 휴머노이드 핵심 부품 내재화와 높은 원가 경쟁력을 기반으로 글로벌 시장점유율 확대 기대 • 연구·교육용 중심에서 제조·물류 등 산업용 시장으로 적용 범위 확대에 따른 출하량 증가 전망 • 휴머노이드 양산 확대와 실제 작업 데이터 축적을 통한 AI 동작제어 기술 고도화 • 로봇 기업에 대한 IPO 심사 강화 추세, 휴머노이드 상장사로서의 희소성 프리미엄 유지

자료: 하나증권

중국 AI 밸류체인 주요 밸류에이션 지표

섹터	기업명	티커	시총 (조원)	YTD 수익률 (%)	매출 성장률 (26~28F CAGR)	순이익 성장률 (26~28F CAGR)	PER(배)		ROE(%)	
							2026F	2027F	2026F	2027F
반도체 장비	북방화창	002371 CH	82.4	23.9	32.4	48.3	57.2	37.3	16.5	20.9
	AMEC	688012 CH	45.6	30.3	40.2	33.3	47.6	41.2	16.7	16.7
	Piotech	688072 CH	29.1	51.5	52.4	43.1	58.9	46.9	23.2	23.0
파운드리	SMIC	0981 HK	107.8	(31.1)	18.7	27.8	34.6	26.1	6.1	7.4
	화홍반도체	1347 HK	45.0	0.4	22.0	49.1	61.3	37.1	4.2	6.6
메모리	CXMT	688825 CH	490.2	317.9	45.1	48.9	12.6	7.7	86.8	67.2
	기가디바이스	603986 CH	30.1	(1.5)	20.9	12.2	10.1	8.9	44.8	35.0
팹리스	캠브리콘	688256 CH	113.5	(0.5)	91.8	97.8	96.4	43.7	37.8	51.9
	해광정보	688041 CH	84.6	(18.7)	43.3	58.5	95.4	60.6	17.0	21.3
	일루바타 코어엑스	9903 HK	9.1	40.6	108.8	흑전	-	41.5	(9.2)	13.3
	비렌테크	6082 HK	10.0	10.2	204.4	흑전	-	26.4	(3.3)	12.1
AI 모델	즈푸AI	2513 HK	27.7	188.3	150.0	적지	-	-	(10.1)	(9.4)
	미니맥스	0100 HK	6.9	(29.7)	149.9	12.0	-	-	13.5	(1.9)
휴머노이드	유니트리	688836 CH	50.0	311.0	71.1	79.3	399.3	235.5	11.0	12.4
	유비테크	9880 HK	5.8	(47.1)	59.5	흑전	-	131.7	(3.8)	4.4

주: 12M 선행 기준
자료: Bloomberg, 하나증권