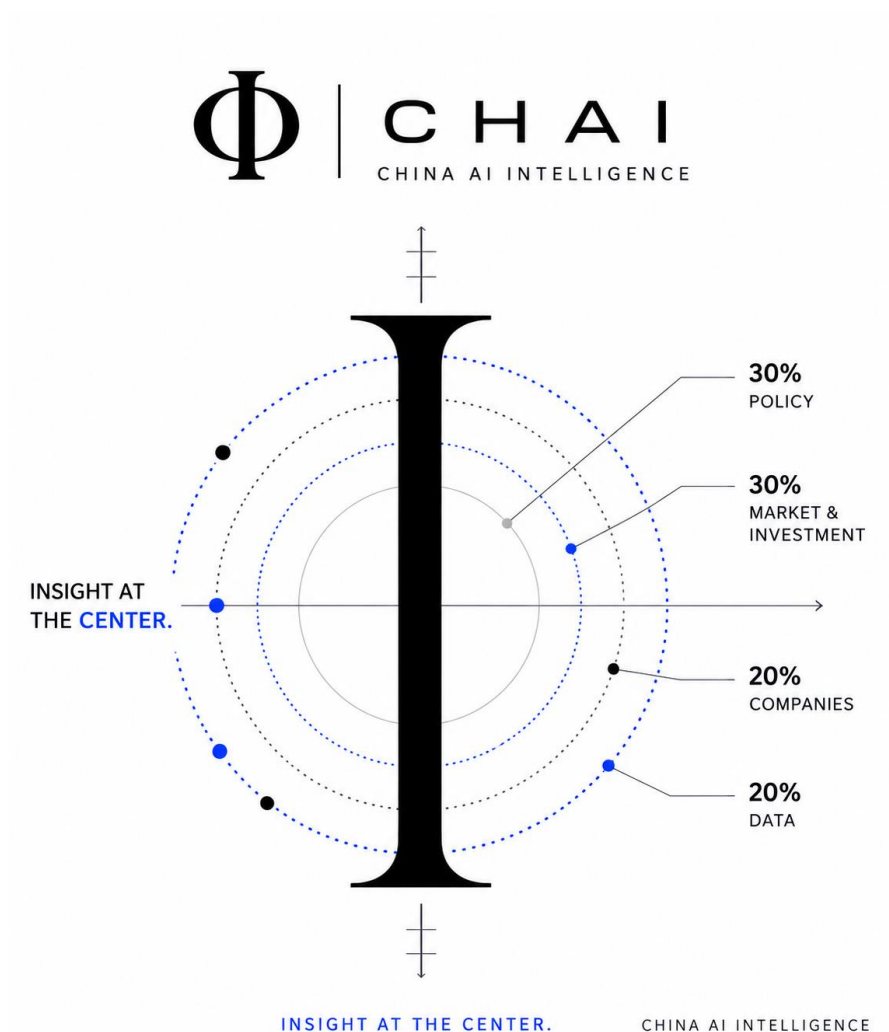


중국 반도체 심층 리포트

소프트웨어가 하드웨어를 정의한다 — 자국산 가속기 양산이 여는 메모리 수요의
창



에디션 월간(Monthly) — 확장 편성

발행일 2026 년 09 월 06 일

발행인 차이나랩 CHINALAB

수신자 대기업 전략팀, 증권사 리서치센터, 반도체 소부장, 정책 입안자

환율 USD/KRW 1359.5 · CNY/KRW 202.22 · HKD/KRW 173.41 (2026-09-04 한국은행)

매매기준율(ECOS))

목차

실행 요약	4
사건 — 딥시크의 16 만 장 발주와 기가와트급 데이터센터	6
왜 함께 설계하는가 — 소프트웨어가 하드웨어를 정의하는 구조	9
연산·메모리 최적화 소프트웨어 스택 해부	12
컴파일러·런타임 — 연산을 융합해 HBM 왕복을 지운다	12
메모리 최적화 — 대역폭 요구를 비트 수만큼 깎는다	12
통신 최적화 — MoE 일수록 상호연결이 병목이 된다	13
추론 서빙 — 프리필과 디코드의 성격이 갈라진다	13
결론 — 소프트웨어가 낮춘 대역폭만큼 자국산 메모리가 버틴다	14
메모리 수요의 재편 — 'GPU+HBM' 공식의 해체	16
온디바이스 AI 반도체 지도	19
스마트폰 — 물량은 크지만 검증 가능한 중국 자국산 스펙은 부족하다	19
피지컬 AI·자동차 — 자국산 진영이 점유율을 확보한 유일한 축	19
웨어블·AIoT — 저전력 저 TOPS 구간의 조용한 확장	20
로드맵 — 발표는 됐으나 탑재 실적이 없는 것들	20
온디바이스가 메모리에 요구하는 것 — 용량이 아니라 전력당 대역폭이다	21
공급망 사다리 6 단계 재점검과 자국산화 목표	23
자본시장 — 메모리·가속칩 기업의 상장 계보	26
가격 시계열 전망 — 양산 진입이 칩·메모리 가격을 언제 움직이는가	29
수출통제·정책 환경	32
한국 시사점·액션 플랜	34
부록 — 주요 지표와 출처	38
1. 주요 지표 일람	38
2. 15 차 5 개년 계획 — 2030 년 목표 대 현재 수준	39
3. 원문 출처 목록	40
4. 이번 호 신규 용어 대조표	40

실행 요약

이번 호의 결론은 하나다. 자국산 가속기의 양산 전환이 한국 고성능 메모리에 단기 수요 창을 열고, 화웨이 자체 HBM 과 창신메모리(CXMT)·양쯔메모리(YMTC)의 증설이 그 창을 닫는다. 딥시크(DeepSeek)가 네이멍구 데이터센터에 화웨이 어센드 950DT 최소 16 만 장을 배치한다는 계획(블룸버그 2026-09-04 보도)은 이 구조의 첫 사례다. 칩은 이미 자국산이고, 메모리만 아직 외부 조달이 필요한 구간이 지금이며, 그 구간의 길이가 한국 기업의 행동 시간표를 결정한다.

핵심 인사이트 5 개. 첫째, 중국의 15 차 5 개년 계획(2026~2030, 2026 년 3 월 발표)은 반도체 자급률 70%를 2030 년 목표로 다시 썼는데, 이는 《중국제조 2025》가 2025 년 목표로 뒀던 같은 70%가 미달 끝에 5 년 이월된 것으로, 목표의 신규성이 아니라 미달의 재확인이다(IC 인사이트는 2025 년 중국 내 생산 점유율을 19.4%로 전망, 2021 년 기준). 둘째, 딥시크와 화웨이의 협력은 칩 구매가 아니라 어센드 슈퍼노드 공동 설계이므로, 모델 기업이 하드웨어 사양을 정의하는 '소프트웨어 주도 재편'이 발주 단위로 현실화됐다(블룸버그 2026-09-04). 셋째, 어센드 950PR 이 화웨이 자체개발 HBM HiBL 1.0(128GB·1.6TB/s)을 채택하고 2026 년 1 분기에 출시된다는 점에서, 자국산 HBM 은 로드맵이 아니라 출시 제품 단계에 진입했다(화웨이 전연결대회 2025, 스위스진 발표, 2025-09-18). 넷째, 다만 HiBL·HiZQ 의 양산 능력·수율·공급량이 공개되지 않았고, 16 만 장 규모 클러스터의 HBM 수요를 자체 물량이 즉시 충당할 수 없다는 점에서 한국 고성능 메모리에 1~2 년짜리 수요 창이 열린다. 단 이 창은 대중 HBM 직수출이 아니다 — 미 BIS 의 2024 년 12 월 2 일 규칙(대역폭 밀도 3.3 GB/s/mm² 미만만 예외)으로 직판은 이미 막혀 있고, 창은 비중국 물량의 재배분·통제 기준 미만 범용 메모리·중국 외 중국계 고객이라는 세 경로로만 열린다. 다섯째, 창신메모리(2026-07-29 과창판 상장, 첫날 +465.82%, 시총 3.28 조 위안)와 양쯔메모리(IPO 수리 2026-08-29, 조달 목표 330 억 위안)의 자본 조달은 그 창을 닫는 증설 재원이며, 양쯔메모리 우한 3 기의 양산 시점이 2027 년에서 2026 년 하반기로 앞당겨졌다는 보도(2026 년 상반기 산업지)는 창이 예상보다 빨리 닫힐 수 있음을 시사한다.

한국 기업에 가장 중요한 변화 3 가지. 첫 번째는 수요 창 의 비대칭이다. 화웨이 950DT 는 HiZQ 2.0(144GB·4TB/s)을 쓰도록 설계됐고 2026 년 4 분기 출시 예정이므로, 한국 HBM 의 실질 기회는 자체 HBM 물량이 부족한 초기 양산 구간과, HBM 칩화가 안 된 주변 메모리(DDR·SSD)로 좁혀진다. 글로벌 메모리가 최근 15 년 중 최악의 공급 부족 국면에 있고 2026 년 2 분기 D 램 계약가가 전분기 대비 +58~63%, NAND 가 +70~75% 오른다는 전망(시장조사·보도, 전망치이므로 실측과 구분 필요)까지 겹쳐, 이 구간의 가격 협상력은 한국 쪽에 있다. 두 번째는 거래 상대의 성격 변화다. 딥시크처럼 모델 기업이 슈퍼노드 사양을 공동 정의하면 구매 결정권이 칩 회사에서 모델 회사와 클라우드 사업자로 이동하고, 한국 메모리 기업은 칩 회사가 아니라 딥시크·텐센트 같은 수요자와 직접 사양을 협의해야 한다. 엔플레임의 최대주주이자 최대 고객이 텐센트라는 구조(과창판 IPO, 2026-08-29)는 이 수직 결합이 업계 표준이 되고 있음을 보여준다. 세 번째는 창이 닫힌 뒤의 시장 재편이다. 양쯔메모리 NAND 가 출하량 세계 3 위·점유율 14%로 키옥시아를 추월했고(CHAI 주간 tech 2026-08-16 호) 월 생산능력 30 만 장 확충을 목표로 하므로,

자국산 HBM 이 양산 궤도에 오르는 시점부터 한국의 중국 시장 포지션은 주력 공급자에서 프리미엄 보완 공급자로 강등된다.

위험·기회 2x2 매트릭스. 단기 위험은 두 가지다. 하나는 화웨이 자체 HBM 의 실제 수율이 공개 전망보다 높아 수요 창이 6 개월 안에 닫히는 경우, 다른 하나는 미국 수출 통제가 어센드 클러스터용 한국 HBM 공급 자체를 차단해 창이 법적으로 소멸하는 경우다. 단기 기회는 어센드 950 초기 물량의 HBM 공백과, 기가와트급 데이터센터 전력 규모에 수반하는 서버용 D 램·기업용 SSD 수요다. 장기 위험은 창신메모리의 HBM3/4 규모화 생산 목표(15.5 규칙, 2026 년 3 월)가 실현돼 중국 시장에서의 구조적 점유율 상실이고, 장기 기회는 워크로드별로 갈라지는 가속기 시장에서 추론 프리필용·디코드용·온디바이스용으로 세분화되는 메모리 사양에 맞춘 제품 다변화다. 온디바이스 쪽에서는 로크칩 RK182X 양산 진입과 AI 스마트폰 출하 4.7 억 대·침투율 38% 전망(2026 년, 업계 보도)이 저전력 메모리의 새 수요처다.

실행 액션 아이템 5 개. ①영업 부서는 2026 년 4 분기 이전에 어센드 950DT 초기 양산분의 HBM 공급 가능성을 검증하되, 수출 통제 적용 여부를 법률 검토와 병행한다(시기: 즉시). ②제품 기획 부서는 HiBL 1.0 의 프리필 최적화 설계를 역산해 프리필·디코드 분리형 가속기에 대응하는 메모리 사양 두 갈래를 2027 년 1 분기까지 정의한다. ③전략 부서는 딥시크·바이트댄스·텐센트 등 모델·클라우드 사업자의 슈퍼노드 자체 설계 동향을 분기 단위로 추적하는 전담 라인을 2026 년 4 분기 안에 둔다. ④투자 심의 부서는 창신메모리 상장 후 공시와 양쯔메모리 IPO 진행(수리 단계, 상장 미완료)을 통해 양사의 증설 CAPEX 를 월 단위로 갱신하고, 자국산 HBM 양산 시점 추정 모델을 2027 년 상반기까지 구축한다. ⑤정책 대응 부서는 15.5 규칙의 '성숙 공정 세계 비중 52%' 목표가 가격 결정권 확보를 겨냥한다는 점(2026 년 3 월 규칙 해설, 전자산품세계)을 감안해, 성숙 공정 메모리의 덤핑 리스크에 대한 정부 차원 대응 논의를 2027 년 안에 개시한다.

다음 호 전망. 다음 호는 어센드 950DT 의 2026 년 4 분기 실제 출시 여부와 HiZQ 2.0 의 공급량 첫 확인 결과가 한국 HBM 수요 창이 잔여 기간을 결정하는 분기가 되며, 양쯔메모리 우한 3 기의 하반기 양산 실적이 NAND 가격 전망(+200~250%, 2025-12 보도 기준 전망치)을 뒤집을 변수로 떠오른다.

이번 호에서 새로 확인한 세 가지. 첫째, 16 만 장은 훈련이 아니라 추론용이며 딥시크는 훈련의 상당 부분에 여전히 엔비디아를 쓸 계획이다. 자국산 칩은 생태계 이전 비용이 낮은 추론부터 규모를 만들고 있고, 그 순서가 메모리 수요의 성격도 결정한다. 둘째, 화웨이 로드맵에서 연산보다 메모리가 더 가파르게 커진다 — 어센드 960(2027 년 4 분기)이 288GB·9.6TB/s, 970(2028 년 4 분기)이 288GB·14.4TB/s 다. 메모리가 연산보다 비싼 자원이 됐다는 설계상의 선언이다. 셋째, 같은 AI 붐이 메모리 가격은 밀어 올리고 추론 연산 단가는 끌어내린다. 소프트웨어 최적화가 토큰당 원가를 계속 깎는 동안 그 연산을 담을 메모리는 부족해지며, 중국의 양산 체제는 이 두 곡선을 모두 가속한다. 한국 기업의 전략은 이 비대칭을 같은 시간표 위에 놓는 데서 출발해야 한다.

사건 — 딥시크의 16 만 장 발주와 기가와트급 데이터센터

딥시크(DeepSeek, 深度求索)가 화웨이의 자국산 AI 가속기를 16 만 장 단위로 배치한다는 계획이 공개되면서, 중국 AI 인프라의 조달 주도권이 칩 공급자에서 모델 기업으로 넘어왔다는 사실이 숫자로 확인됐다. 이 절의 상당 부분은 2026 년 9 월 4 일 블룸버그 보도를 관찰자만이 전제한 내용에 근거한다(CHAI 주간 tech 2026-09-05 호 게재). 아래 서술은 '보도된 계획'과 '확정된 사실'을 엄격히 구분한다.

조달 규모부터 본다. 보도에 따르면 딥시크는 네이멍구(내몽골)에 건설 중인 대형 데이터센터에 화웨이 어센드 950DT 칩을 최소 16 만 장 배치할 계획이다. '최소'라는 표현이 붙은 것은 상향 가능성을 열어둔 것이지 확정 물량이 더 크다는 뜻은 아니다. 완료될 경우 이 클러스터는 알려진 최대 규모의 자국산 AI 칩 클러스터 가운데 하나가 된다(블룸버그 2026-09-04 보도). 여기까지가 보도된 계획이며, 계약 체결부터 인도 개시까지의 집행 단계는 아직 공개되지 않았다.

전력 규모는 조달 규모보다 먼저 눈에 띄는 숫자다. 해당 프로젝트의 전력 규모는 기가와트(GW)급으로 보도됐으며, 16 만 장은 그중 일부 용량에 해당한다(같은 보도). 기가와트급은 원자력발전소 1 기의 출력에 맞먹는 규모로, 단일 기업의 단일 프로젝트로는 글로벌 최상위권 전력 배정이다. 네이멍구가 입지로 선택된 배경에는 풍력·석탄 기반의 저렴한 전력과 서늘한 기후라는 물리 조건이 있다는 것이 통상의 해석이다.

설치 진도는 공급 능력에 좌우되며, 전량 인도에 1 년 이상 걸릴 수 있다는 것이 보도의 판단이다. 이 문구는 사실상 이 발주가 단기 납품 계약이 아니라 다년 할당 계획임을 시사한다. 공급 병목이 어디인지가 관건인데, 어센드 950DT 는 화웨이 자체개발 HBM 인 HiZQ 2.0(용량 144GB, 대역폭 4TB/s, 화웨이 전연결대회 2025 발표, 2025-09-18)을 채택하므로, 칩 공급 속도는 다이 수율뿐 아니라 HiZQ 2.0 의 양산 능력에도 묶인다. 문제는 HiBL 1.0 과 HiZQ 2.0 의 양산 능력·수율·공급량이 공개되지 않았다는 점이다. 16 만 장에 필요한 HBM 총량을 자체 생산으로 감당할 수 있는가가 한국 고성능 메모리 수요 창고의 크기를 좌우하는 직접 변수다.

출시 시점의 정합성도 따져봐야 한다. 어센드 950DT 는 2026 년 4 분기 출시 예정이다(화웨이 전연결대회 2025, 쉬즈쥔 순환회장 발표, 2025-09-18). 딥시크의 16 만 장 배치 계획이 보도된 시점은 2026 년 9 월이므로, 보도 시점 기준으로 950DT 는 아직 출시 전 제품이다. 즉 16 만 장은 양산 전 제품에 대한 선행 물량 확보이며, 실제 설치는 2026 년 4 분기 이후에나 시작될 수 있다. 설치 진도가 '공급 능력에 좌우된다'는 보도 문구와 결합하면, 이 프로젝트의 실질 완성 시점은 빠르더라도 2027 년 이후로 밀릴 공산이 크다. 딥시크가 하반기 어센드 950 슈퍼노드 양산 출시 시점에 맞춰 프로 가격 인하를 예고한 것(같은 보도)도 같은 시간표 위에 있다.

이 사건에서 하드웨어 조달보다 중요한 것은 협력의 범위다. 보도에 따르면 딥시크와 화웨이의 협력은 칩 구매를 넘어 컴퓨팅 시스템 설계로 확장되며, 양사는 어센드 슈퍼노드를 공동으로 정의하고 있다. 공동 정의의 대상은 장문 추론의 지연 저감, 처리량 향상, 비용 통제다. 아틀라스 950 슈퍼노드는 최대 8,192 장을, 아틀라스 960 슈퍼노드는 15,488 장을 지원하고 칩 간 데이터 교환은 화웨이 자체 상호연결 기술 링취(靈衢)가 담당한다(화웨이 전연결대회 2025 발표). 16 만 장은 8,192 장 규모 슈퍼노드 약 20 기분에 해당한다.

딥시크가 이토록 큰 물량을 선점하는 동기는 수요 측 숫자에서 드러난다. 딥시크는 고급 컴퓨팅 제약으로 4-프로(4-Pro) API 처리량이 매우 제한적이라고 스스로 밝혔고, 하반기 어센드 950 슈퍼노드가 양산 출시되면 프로 가격을 대폭 내릴 수 있다고 했다(같은 보도). 이는 모델 기업이 가격 경쟁력의 병목을 모델 품질이 아니라 컴퓨팅 공급량으로 진단하고, 그 병목을 자국산 하드웨어의 대량 선점으로 풀겠다는 전략이다. 2026 년 글로벌 AI 인프라 지출 4,500 억 달러 가운데 추론 컴퓨팅 비중이 처음으로 70%를 넘는다는 전망(15·5 기획 해설 자료 인용, 전자산품세계 2026-03-17)과 맞물려, 딥시크의 발주는 추론 시대의 물량 확보 경쟁을 자국산 공급망 안에서 먼저 시작한 사례로 읽힌다.

워크로드별 하드웨어 분화라는 이번 호의 관통 논지에 이 사건은 교과서적 사례를 제공한다. 화웨이의 950 계열 자체가 프리필용 950PR(2026 년 1 분기 출시, HiBL 1.0 탑재, 128GB·1.6TB/s)과 디코드·훈련용 950DT(2026 년 4 분기 출시, HiZQ 2.0 탑재, 144GB·4TB/s)로 갈라져 있다(화웨이 전연결대회 2025 발표). 딥시크가 장문 추론의 지연과 비용을 목표로 디코드 특화형인 950DT 를 선택한 것은, 추론의 한 구간에 최적화된 메모리 구성을 가진 칩을 통째로 사들인 행위다. 'GPU+HBM' 단일 공식이 워크로드별 분해되는 현장이며, 이 분해가 메모리 사양 분화(HiBL 대 HiZQ 처럼 대역폭·용량이 다른 HBM 등급)를 부품 수요 차원에서 만든다는 점이 핵심이다.

한국 기업의 관점에서는 두 가지를 분리해 봐야 한다. 첫째, 16 만 장이 전량 자체 HBM 으로 공급된다면 한국 메모리의 직접 수요는 발생하지 않는다. 둘째, HiZQ 2.0 의 양산 능력이 공개되지 않은 이상, 950DT 공급 지연이 생길 경우 딥시크가 전력과 부지를 확보한 채 대기하는 구조가 되고, 그 대기 기간이 곧 외부 고성능 메모리의 기회 창이 된다. 본지는 HBM 이 칩 전체 가치의 약 절반을 차지한다는 그레고리 앨런(미 CSIS)의 설명을 앞선 절에서 인용한 바 있는데, 16 만 장 규모에서는 HBM 확보 여부가 프로젝트 성패를 가르는 단일 최대 부품 변수다. 따라서 추적 지표는 발주 보도가 아니라 2026 년 4 분기 950DT 의 실제 출시 여부, HiZQ 2.0 의 월 공급량, 그리고 네이밍구 데이터센터의 전력 인입 개시 시점 세 가지다. 계약 체결과 인도 실적이 확인되지 않은 현 시점에서, 이 사건은 '확정된 수요'가 아니라 '공표된 조달 의향'으로 분류하는 것이 정확하다.

용도를 짚어야 이 사건의 크기가 정확해진다. **16 만 장은 훈련이 아니라 추론용이며, 딥시크는 훈련의 상당 부분에 여전히 엔비디아 칩을 쓸 계획**이라고 같은 보도가 전한다. 우연이 아니다. 추론은 훈련보다 소프트웨어 생태계 이전 비용이 낮아 자국산 칩이 먼저 규모를 이룰 수 있는 지점이고, 그렇게 쌓인 운영 경험이 훈련 전환의 공학적 자산이 된다. 자국산 칩은 가장 큰 시장인 추론부터 공략하고 있다는 것이 이 사건의 첫 번째 독법이다.

규모의 진전 속도는 시계열로 봐야 보인다. 2026 년 3 월 26 일 선전이 1 만 1,000P 규모 어센드 클러스터를 점등했고, 6 월 30 일 메이탄(Meituan, 美团)은 롱캣(LongCat) 2.0 모델을 **5 만 장 규모 자국산 칩 클러스터**에서 사전훈련부터 추론까지 전 과정 수행했다고 밝혔으며, 7 월에는 즈푸(Z.AI, 智譜)가 **약 1GW 규모 자국산 칩 데이터센터**를 부분 가동했다(중국 업계 매체 종합, 등급: 보도). 반년 전 '1 만 장'이 화두이던 시장이 '16 만 장 발주 계획'을 논하게 된 것이다. 입지도 우연이 아니다. 네이밍구 우란차부(乌兰察布)는 국가발전개혁위원회가 승인한 전국 일체화 컴퓨팅 네트워크의 국가 허브 노드가 있고 화웨이·알리바바 데이터센터가 밀집한 '동수서산(東數西算)'의 핵심 지역이다.

공급 측 제약은 숫자로 확인된다. 화웨이의 2026 년 어센드 다이 생산 계획은 **약 160 만 개** 수준으로 알려졌는데, 이는 기존 910C 계열 물량까지 함께 감당해야 하는 규모다(중국 업계 매체, 등급: 보도). 즉 딥시크가 더 사고 싶어도 살 수 없는 구간이 존재하며, 그 병목의 핵심이 고대역폭 메모리다.

왜 함께 설계하는가 — 소프트웨어가 하드웨어를 정의하는 구조

모델 기업이 칩 사양표를 받아드는 시대는 끝났고, 사양표를 쓰는 시대가 시작됐다. 블룸버그 2026-09-04 보도(관찰자망 전재)에 따르면 딥시크(DeepSeek, 深度求索)와 화웨이의 협력은 칩 구매 계약에 그치지 않고 "어센드 슈퍼노드를 공동 정의"하며 "장문 추론의 지연 저감·처리량 향상·비용 통제를 함께 다루는" 단계에 있다. 이 문장의 무게는 '공동'이 아니라 '정의'에 있다. 정의한다는 것은 요구 사항이 소프트웨어 쪽에서 발생해 하드웨어 쪽으로 흘러 들어간다는 뜻이며, 이 방향이 한 번 고정되면 부품 수요의 순서와 물량이 그 경로를 따라 재배열된다.

되먹임의 경로는 구체적이다. 첫째, 장문 추론의 지연(latency) 문제가 칩의 메모리 계층을 규정한다. 대형 언어모델의 추론은 프리필(Prefill)과 디코드(Decode)로 나뉘는데, 프리필은 긴 입력 문맥을 한 번에 처리하는 연산 집약 구간이고 디코드는 토큰을 하나씩 생성하며 매번 모델 가중치와 KV 캐시를 메모리에서 읽어 오는 대역폭 집약 구간이다. 같은 '추론' 칩이 이 두 구간을 함께 커버하려 하면 한쪽이 낭비된다. 딥시크가 다루는 장문 추론에서 지연이 문제로 제기되는 순간, 그 문제는 소프트웨어 최적화만으로 풀리지 않고 "프리필에 강한 칩"과 "디코드에 강한 칩"이라는 두 개의 하드웨어 요구로 번역된다.

둘째, 그 요구는 실제 제품 분기로 실현됐다. 화웨이 전연결대회 2025 에서 순환회장 쉬즈쥔(徐直军)이 2025-09-18 발표한 로드맵에서 어센드 950 계열은 950PR 과 950DT 로 갈라진다. 950PR 은 950 코어에 화웨이 자체 HBM 인 HiBL 1.0(용량 128GB·대역폭 1.6TB/s)을 붙여 프리필 성능과 추천 업무 성능을 높인 제품으로 2026 년 1 분기 출시이고, 950DT 는 HiZQ 2.0(용량 144GB·대역폭 4TB/s)을 붙여 디코드 성능·훈련 성능·메모리 용량과 대역폭을 높인 제품으로 2026 년 4 분기 출시 예정이다. 같은 코어에 서로 다른 HBM 을 결합해 워크로드별로 대역폭 예산을 다르게 배정한 이 구조는, 모델 기업이 제기한 지연 문제가 메모리 사양(용량·대역폭)의 차별화로 되먹임된 직접 증거다. 'GPU+HBM'이라는 단일 공식이 프리필용 공식과 디코드용 공식으로 분해된 것이다.

셋째, 비용 통제 요구는 저정밀 연산 사양으로 되먹임됐다. 950 아키텍처는 저정밀 데이터 포맷을 새로 지원해 FP8/MXFP8/HIF8 에서 1 PFLOPS, MXFP4 에서 2 PFLOPS 를 제공한다(화웨이 전연결대회 2025, 2025-09-18). 추론 비용은 정밀도를 낮출수록 내려가지만 모델 기업이 해당 포맷에서 정확도를 방어할 수 있어야만 유효하다. 딥시크는 고급 컴퓨팅 제약으로 4-프로(4-Pro) API 처리량이 매우 제한적이며, 하반기 어센드 950 슈퍼노드가 양산 출시되면 프로 가격을 대폭 내릴 수 있다고 밝혔다(블룸버그 2026-09-04). 이 발언은 가격 결정 변수가 마케팅이 아니라 하드웨어 양산 시점과 처리량에 있다는 뜻이며, 모델 기업의 가격표가 칩 로드맵에 종속돼 있음을 보여준다. 추론이 2026 년 글로벌 AI 인프라 지출 4,500 억 달러 중 처음으로 70%를 넘는다는 전망(15.5 해설 자료 인용)을 감안하면, 이 비용 되먹임은 일회성이 아니라 시장의 주류 수요에서 작동하는 메커니즘이다.

공동 정의는 단일 칩을 넘어 클러스터 토폴로지까지 관통한다. 아틀라스 950 슈퍼노드는 최대 8,192 장, 아틀라스 960 슈퍼노드는 15,488 장을 지원하며 화웨이 자체 상호연결 기술 링취(灵衢)로 칩 간 데이터를 고속 교환한다(화웨이 전연결대회 2025, 2025-09-18). 950 세대에서 상호연결 대역폭은 2.5 배로 강화됐다. 이 숫자들은 하드웨어 스펙인 동시에 모델 병렬화 전략의 제약 조건이다. 대형 모델을 8,192 장에

조개 올릴 때 텐서 병렬화·파이프라인 병렬화 전문가 병렬화를 어떤 비율로 조합할지는 상호연결 대역폭이 결정한다. 노드 내 대역폭이 충분하면 통신이 잦은 텐서 병렬화 구간을 넓게 잡을 수 있고, 부족하면 통신이 적은 파이프라인 병렬화 비중을 늘려야 하며, 그 결과 미니배치 크기와 지연 특성이 달라진다. 링취의 실효 대역폭과 토폴로지가 곧 딥시크가 설계할 수 있는 모델의 형태를 규정하는 셈이다. 하드웨어가 소프트웨어의 설계 공간을 규정하는 역방향 고리가 여기서 닫힌다 — 모델 요구가 칩을 만들고, 만들어진 클러스터가 다시 다음 모델의 병렬화 전략을 묶는다.

딥시크의 네이밍구 프로젝트는 이 고리의 규모 시험장이다. 어센드 950DT 를 최소 16 만 장 배치하고 전력 규모는 기가와트급이며, 완료 시 알려진 최대 규모의 자국산 AI 칩 클러스터 중 하나가 된다(블룸버그 2026-09-04, 보도 등급). 다만 이는 공표된 조달 의향이며, 전량 인도에 1 년 이상 걸릴 수 있다는 점이 같은 보도에 명시돼 있다. 950DT 출시가 2026 년 4 분기이므로 실질 설치는 그 이후에야 시작된다.

그러나 공동 설계를 시장의 새 표준으로 단정하는 것은 위험하며, 반대 해석이 필요하다. 첫째, 공동 설계는 공급자가 사실상 하나뿐일 때의 차선택일 수 있다. 미국의 고급 컴퓨팅 수출 통제 아래에서 중국 모델 기업이 자국산 가속기 외의 선택지를 확보하기 어려운 조건이라면, 딥시크가 화웨이 사양에 깊이 개입하는 것은 우월한 전략이 아니라 유일한 공급선과의 협상력 확보 행위로 읽을 수 있다. 여러 공급자가 경쟁하는 시장에서는 모델 기업이 사양을 '정의'하기보다 기성품 중 최적 조합을 '선택'하는 편이 비용이 낮다. 둘째, 사양 협의와 설계 개입은 다르다. 대형 고객이 지연·처리량·가격 요구를 제시하고 공급자가 다음 세대 제품에 반영하는 관행은 모든 칩 생태계에 존재하며, 이를 '소프트웨어가 하드웨어를 정의한다'는 구조 변화로 과대 해석하면 신호보다 소음을 키운다. 블룸버그 보도만으로는 공동 정의의 깊이 — 딥시크 인력이 실제로 칩 아키텍처 결정에 참여했는지, 요구 사항 전달에 그쳤는지 — 를 판별할 수 없다. 셋째, 결합의 대가가 있다. 특정 슈퍼노드 토폴로지에 맞춰 병렬화 전략을 최적화한 모델은 다른 하드웨어로 이식할 때 재최적화 비용을 치러야 하고, 이 잠금(lock-in)은 공급자가 하나일 때 위험으로 돌아온다.

판단의 추적 지표는 선언이 아니라 행동이다. 950DT 가 2026 년 4 분기에 실제 출시돼 네이밍구에 인도되기 시작하는지, 딥시크가 프로 가격 인하를 실제로 집행하는지, 그리고 공동 정의가 950 의 후속인 960·970 로드맵에도 반복되는지를 봐야 한다. 세 지표가 모두 확인되면 이것은 일회성 조달 협의가 아니라 구매 결정권의 구조적 이동이다. 그 이동이 확인되는 순간, 부품 수요의 질문도 바뀐다. HBM 이 칩 전체 가치의 약 절반을 차지한다는 CSIS 그레고리 앨런의 설명(미 CSIS)을 전제하면, 워크로드별로 갈라진 칩이 요구하는 서로 다른 HBM — HiBL 1.0 의 128GB·1.6TB/s 와 HiZQ 2.0 의 144GB·4TB/s — 은 각기 다른 메모리 수요 곡선을 만든다. 한국이 물어야 할 것은 딥시크·화웨이 결합의 명분이 아니라, 그 결합이 만드는 분기점마다 어떤 메모리·어떤 패키징·어떤 전력 예산의 수요가 언제 열리고 언제 닫히는가다.

로드맵 전체를 놓고 보면 화웨이의 설계 의도가 더 분명해진다. 쉬즈쥔이 공표한 3 년 계획은 2026 년 1 분기 950PR(프리필·추천), 4 분기 950DT(디코드·훈련), **2027 년 4 분기 어센드 960(HBM 288GB·대역폭 9.6TB/s, FP8 2 PFLOPS·FP4 4 PFLOPS, 자체 개발 HiF4 포맷), 2028 년 4 분기 어센드 970(HBM 288GB·대역폭 14.4TB/s, FP8 4 PFLOPS·FP4 8 PFLOPS, 상호연결 4TB/s)**으로 이어진다(화웨이 로드맵, 2025-09-18). 주목할 것은 연산보다 **메모리가 더 가파르게 커진다는 점**이다. 메모리 용량은 144GB 에서 288GB 로, 대역폭은 4TB/s 에서 9.6TB/s 를 거쳐 14.4TB/s 로 뛴다. 대규모 MoE 모델의 디코드가

본질적으로 메모리 대역폭 제약 작업이기 때문이며, 이 설계 방향 자체가 '메모리가 연산보다 비싼 자원이 됐다'는 선언이다.

슈퍼노드 단위의 제원도 같은 방향을 가리킨다. 아틀라스 950 슈퍼노드는 8,192 장을 묶어 FP8 기준 8 EFLOPS 와 상호연결 16.3PB/s 를, 2027 년 4 분기의 아틀라스 960 슈퍼노드는 15,488 장으로 FP8 30 EFLOPS·메모리 4,460TB·상호연결 34PB/s 를 표방한다. 단일 칩의 공정 열세를 시스템 집적으로 상쇄하겠다는 전략의 공식화다.

이 결합이 왜 가능한지는 그 이전 단계에서 이미 준비됐다. 화웨이의 클라우드매트릭스 384(CloudMatrix 384)는 어센드 910C 384 개와 쿼터 CPU 192 개를 통합 버스로 묶어 물리적으로는 여러 대이면서 논리적으로는 한 대처럼 작동하게 한 시스템으로, 대규모 MoE 전문가 병렬과 분산 KV 캐시 접근처럼 통신 집약적인 작업에 유리하도록 설계됐다. 그리고 딥시크는 2025 년 8 월 V3.1 에서 **UE8M0 FP8 스케일 포맷**을 채택하며 "곧 출시될 차세대 자국산 칩을 위해 설계됐다"고 명시했다. **모델 회사가 아직 나오지도 않은 칩의 수치 표현 방식에 맞춰 모델을 미리 최적화한 것이다.** 공동 설계는 이번 발주에서 시작된 것이 아니라, 이미 한 세대 앞서 진행돼 온 셈이다.

연산·메모리 최적화 소프트웨어 스택 해부

컴파일러·런타임 — 연산을 융합해 HBM 왕복을 지운다

소프트웨어 스택의 첫 번째 과제는 하드웨어에 도착하기도 전에 데이터 이동 자체를 줄이는 것이다. AI 연산의 비용은 부동소수점 연산 자체보다 가중치·중간 텐서를 HBM 과 온칩 메모리 사이에서 왕복시키는 데서 더 크게 발생한다. 연산자 융합(operator fusion)은 여러 연산을 하나의 커널로 합쳐 중간 결과를 HBM 에 썼다가 다시 읽는 왕복을 없앤다. 어텐션 연산을 단일 커널로 묶는 플래시어텐션(FlashAttention)이 대표적이다. 그래프 최적화는 실행 순서를 재배열해 캐시 재사용률을 높이고, 커널 튜닝은 타일 크기·스레드 배치를 특정 칩의 메모리 계층에 맞춰 조정한다. 이 세 가지는 모두 "같은 계산을 더 적은 메모리 트래픽으로 끝내는" 기술이다.

이 계층의 경쟁 측은 CUDA 대 CANN 이다. 엔비디아의 CUDA 는 cuBLAS·cuDNN·NCCL 같은 라이브러리가 15 년 이상 다듬어져 개발자가 커널을 직접 쓰지 않아도 최적에 가까운 성능을 얻는다. 화웨이의 CANN 은 어센드 칩용 대응 스택으로, PyTorch 같은 상위 프레임워크와 어센드 하드웨어 사이를 잇는다. CANN 의 융합·그래프 최적화 원리는 CUDA 와 같지만, 생태계의 축적 연수와 최적화된 연산자 커버리지에서 격차가 있다. 이 격차를 메우는 방식이 이 리포트의 관통 논지와 연결된다. 딥시크 같은 모델 기업이 자사 모델의 핵심 연산 패턴을 직접 알려주면, CANN 은 범용 커버리지가 아니라 특정 모델의 핵심 커널만 집중 튜닝하면 된다. 모델 기업과 칩 기업의 공동 설계(블룸버그 2026-09-04 보도)는 소프트웨어 계층에서 컴파일러 최적화 범위를 좁혀 하드웨어 성능을 끌어올리는 지름길이기도 하다.

메모리 최적화 — 대역폭 요구를 비트 수만큼 깎는다

두 번째 계층은 메모리에 무엇을 얼마나 올릴 것인가를 다룬다. 생성형 추론에서 대역폭을 잡아먹는 주범은 모델 가중치와 KV 캐시다. 디코드 단계는 토큰을 하나 생성할 때마다 가중치 전체와 누적된 KV 캐시를 읽어야 하므로 연산량보다 메모리 대역폭이 병목이다. 페이지드 어텐션(PagedAttention)은 KV 캐시를 운영체제의 가상 메모리처럼 페이지 단위로 쪼개 관리해, 연속 메모리 예약에 따른 낭비를 제거한다. KV 캐시 압축은 어텐션 카·값 자체를 저랭크 표현으로 바꾸는데, 딥시크의 MLA(Multi-head Latent Attention)가 이 방식으로 캐시 크기를 크게 줄인 구조적 사례다. 오프로딩은 사용 빈도가 낮은 가중치·캐시를 HBM 에서 DDR 이나 스토리지로 내려 보내 고가의 HBM 용량을 절약한다.

양자화는 이 계층에서 가장 계산이 명확한 기법이다. FP16 가중치를 FP8 로 바꾸면 파라미터당 바이트 수가 절반이 되므로, 같은 모델을 같은 토큰 속도로 생성하는 데 필요한 대역폭도 정확히 절반이 된다. INT4 로 가면 4 분의 1 이다. 즉 양자화는 정확도를 일부 포기하는 대가로 대역폭 요구를 산술적으로 나누는 기술이다. 어센드 950 아키텍처가 저정밀 포맷을 새로 지원해 FP8/MXFP8/HIF8 에서 1 PFLOPS, MXFP4 에서 2 PFLOPS 를 낸다는 사양(화웨이 전연결대회 2025, 2025-09-18 쉬즈쥘 발표)은 이 소프트웨어 경로를

하드웨어가 정면으로 받아들인 것이다. MXFP4 연산력이 FP8 의 두 배로 설계된 것은, 4 비트 양자화 모델을 돌릴수록 대역폭 한계가 완화되는 구간이 넓어진다는 뜻이기도 하다.

통신 최적화 — MoE 일수록 상호연결이 병목이 된다

세 번째 계층은 칩과 칩 사이의 데이터 이동을 다룬다. 모델이 한 장의 메모리에 들어가지 않으면 가중치를 쪼개 여러 칩에 올려야 하고, 그 쪼개는 방식이 통신량을 결정한다. 텐서 병렬은 행렬 곱을 칩마다 나누고 매 층마다 결과를 합치는 올리듀스(all-reduce) 통신을 요구하므로 통신 빈도가 가장 높다. 파이프라인 병렬은 층 단위로 칩을 나눠 경계에서 활성값만 넘기므로 통신량은 적지만 버블(유휴 시간)이 생긴다. 전문가 병렬(MoE 병렬)은 토큰마다 라우터가 선택한 전문가가 있는 칩으로 토큰을 보내고 결과를 회수하는 올투올(all-to-all) 통신이 발생한다.

딥시크의 모델 구조가 노드 간 대역폭에 민감한 이유가 여기에 있다. 딥시크 V3 계열은 수백 개의 세분화된 전문가 중 토큰당 소수만 활성화하는 희소 MoE 구조를 택해 연산량을 줄였지만, 그 대가로 토큰이 배치마다 여러 노드를 오가는 올투올 통신이 필수가 됐다. 전문가를 여러 노드에 분산시킬수록 통신은 노드 내부가 아니라 노드 간 링크를 타고, 상호연결 대역폭이 곧 훈련·추론 속도의 상한이 된다. 화웨이가 상호연결 대역폭을 2.5 배로 올리고 자체 상호연결 기술 링취(灵衢)를 내세우며 아틀라스 950 슈퍼노드를 최대 8,192 장까지 묶은 것(2025-09-18 발표)은 이 통신 병목에 대한 하드웨어적 응답이다. 슈퍼노드의 상호연결 사양은 스펙표의 숫자가 아니라 다음 모델의 전문가 수·분산 폭을 제한하는 설계 제약이다. 이전 절에서 지적한 역방향 고리 — 하드웨어가 모델의 설계 공간을 묶는 구조 — 가 바로 이 계층에서 작동한다.

추론 서빙 — 프리필과 디코드의 성격이 갈라진다

네 번째 계층은 요청 단위의 스케줄링이다. 연속 배치(continuous batching)은 요청이 끝날 때마다 새 요청을 배치에 즉시 끼워 넣어 가속기의 유휴를 없앤다. 스펙컬러티브 디코딩은 작은 초안 모델이 여러 토큰을 미리 생성하고 큰 모델이 한 번의 순방향으로 검증해, 대역폭이 걸리는 디코드 패스의 횟수를 줄인다. 그리고 이 절의 핵심인 프리필/디코드 분리가 있다.

프리필과 디코드는 같은 GPU 위에서도 완전히 다른 워크로드다. 프리필은 입력 프롬프트 전체를 한 번에 처리하므로 행렬 곱이 크고 연산 집약적이다. 메모리 대역폭보다 컴퓨팅 파워가 병목이다. 디코드는 토큰을 하나씩 생성하면서 매번 가중치와 KV 캐시를 통째로 읽으므로 대역폭 집약적이다. 둘을 한 칩에서 섞어 돌리면, 칩은 두 요구 중 어느 하나에도 최적화되지 않은 절충안이 된다. 둘을 분리하면 각 단계에 맞는 하드웨어를 따로 설계할 수 있다.

어센드 950 계열의 제품 분기는 이 소프트웨어 단계 분리를 하드웨어로 굳힌 사례다. 950PR 은 화웨이 자체 HBM 인 HiBL 1.0(용량 128GB, 대역폭 1.6TB/s)을 달아 추론 프리필 성능과 추천 업무 성능을 끌어올린 프리필용 제품이고, 950DT 는 HiZQ 2.0(용량 144GB, 대역폭 4TB/s)을 달아 디코드 성능·훈련 성능·메모리 용량과 대역폭을 높인 디코드용 제품이다(화웨이 전연결대회 2025, 2025-09-18). 대역폭 1.6TB/s 와 4TB/s 라는 2.5 배 차이는 두 단계의 병목이 다르다는 소프트웨어적 사실을 메모리 사양으로 번역한 것이다.

950PR 은 2026 년 1 분기, 950DT 는 2026 년 4 분기 출시 예정이다. 딥시크가 네이멍구 데이터센터에 배치하겠다고 공표한 16 만 장이 디코드용 950DT 라는 점(블룸버그 2026-09-04 보도)은, 딥시크의 장문 추론 서비스에서 병목이 디코드 쪽 대역폭에 있다는 진단과 일치한다. 이 분리의 경제적 배경도 있다. 2026 년 글로벌 AI 인프라 지출 4,500 억 달러 중 추론 컴퓨팅 비중이 처음으로 70%를 넘는다(15·5 규칙 해설 자료 인용). 추론이 지배하는 시장에서는 추론 단계별 최적화가 곧 비용 경쟁력이다. 딥시크가 어센드 950 슈퍼노드 양산 후 프로 가격을 대폭 내릴 수 있다고 밝힌 것(블룸버그 2026-09-04 보도)은 이 단계 분리가 가져올 단가 절감을 전제로 한 발언이다.

결론 — 소프트웨어가 낮은 대역폭만큼 자국산 메모리가 버틴다

소프트웨어 스택의 네 계층은 모두 같은 방향을 가리킨다. 연산자 융합은 HBM 왕복을 줄이고, 양자화는 대역폭 요구를 비트 수만큼 나누고, 병렬화 전략은 통신을 하드웨어가 감당할 수 있는 범위로 맞추고, 프리필/디코드 분리는 각 단계가 요구하는 대역폭을 서로 다른 제품으로 분산한다. 소프트웨어가 대역폭 요구를 낮출수록 자국산 메모리로 버틸 수 있는 구간이 넓어진다. 최첨단 HBM 이 없어도, 양자화와 캐시 압축으로 요구 대역폭을 절반으로 줄이면 그 절반 사양의 자국산 HBM 으로 서비스가 성립한다. 950PR 이 프리필을, 950DT 가 디코드를 맡도록 갈라진 것은 이 소프트웨어 단계 분리를 하드웨어 제품 구조로 확정한 것이다. 다만 HiBL·HiZQ 의 양산 능력·수율·공급량은 공개되지 않았다. 소프트웨어가 아무리 요구를 낮춰도 그 낮아진 요구를 채울 메모리가 실제로 생산돼야 이 구조는 완성된다.

딥시크 V4 가 이 논리의 교과서다. 2026 년 4 월 공개된 V4 는 1.6 조 파라미터(활성 49B)의 프로와 284B(활성 13B)의 플래시로 구성되고 둘 다 100 만 토큰 컨텍스트를 지원하는데, 핵심은 **CSA(압축 희소 어텐션)와 HCA(고압축 어텐션)의 혼합 구조**다. CSA 는 토큰 4 개를 KV 항목 1 개로 압축하고 인덱서로 상위 일부만 골라 계산하며, HCA 는 훨씬 높은 압축률로 전역 의미를 읽는다. 그 결과 100 만 토큰 컨텍스트에서 **단일 토큰 추론 연산량이 V3.2 의 27% 수준, KV 캐시는 10% 수준으로 내려갔다**(딥시크 기술보고서 기반 중국 기술 매체 정리, 등급: 보도). 아키텍처 진화 자체가 병목 구동 논리를 따른다 — V2 는 MoE 와 MLA 로 메모리 대역폭 병목을, V3 는 극한 MoE 로 훈련 규모화를, V4 는 CSA·HCA 로 장문 컨텍스트의 계산 벽을 넘었다.

생태계 쪽 변화가 더 결정적이다. V4 는 발표와 동시에 화웨이 어센드를 포함한 **8 개 자국산 AI 칩이 '데이제로' 호환 검증을 마쳤다**(한우지·하이광·무어쓰레드·무시·쿤룬신·핑터우거·텐수즈신 등). 과거 수개월씩 걸리던 추격식 적합화 주기가 사라진 것이다. 딥시크는 어센드 아키텍처 적합화를 위해 **40 만 개 이상의 연산자를 재작성**했다고 전해지며, 화웨이는 CANN 컴파일러와 가상 명령 집합 인터페이스를 개방하고 나머지 소프트웨어를 오픈소스로 푸는 방식으로 응수했다. 우회로도 함께 열려 있다 — 하이광의 DCU 는 HIP 인터페이스로 CUDA 코드 호환성 95% 이상을 내세우고, 무어쓰레드의 MUSA 는 자동 이식용, 무시의 MXMACA 는 주류 프레임워크 지원을 앞세운다. 생태계 전쟁은 '자체 스택 정면 돌파'와 'CUDA 호환 우회'라는 두 전선에서 동시에 벌어진다.

시스템 계층에도 같은 논리가 내려와 있다. 프리필 인스턴스와 디코드 인스턴스를 고정 비율로 묶던 배치를 버리고 **각각을 전역 자원 풀로 분리해 요청 단위로 동적 결합**하는 구조(대 PD 아키텍처)가 중국 대형

서비스에서 운용 중이다. 프리필 부하가 높으면 프리필 풀만, 디코드 부하가 높으면 디코드 풀만 늘린다. 어센드 950PR 과 950DT 의 제품 분기가 하드웨어에서 한 일을, 이 구조는 인프라에서 한다.

메모리 수요의 재편 — 'GPU+HBM' 공식의 해체

'GPU+HBM'은 더 이상 AI 메모리 수요를 설명하는 단일 공식이 아니다. 2026 년 글로벌 AI 인프라 지출 4,500 억 달러 가운데 추론 컴퓨팅 비중이 처음으로 70%를 넘는다는 전망이 나왔다(15.5 규칙 해설 자료 인용, 2026 년 3 월). 훈련 단일 수요가 지배하던 시기에는 대역폭 극대화가 곧 메모리 전략의 전부였으나, 추론이 전체 지출의 7 할을 차지하는 구조에서는 프리필과 디코드, 온디바이스와 데이터센터가 서로 다른 메모리를 요구한다. 화웨이가 어센드 950 계열을 프리필용 950PR(HiBL 1.0, 128GB·1.6TB/s)과 디코드용 950DT(HiZQ 2.0, 144GB·4TB/s)로 갈라 2025 년 9 월 18 일 전연결대회에서 발표한 것은 이 분해가 실제 제품 사양으로 굳어진 첫 대형 사례다. 같은 '950'이라는 이름 아래 대역폭이 1.6TB/s 와 4TB/s 로 2.5 배 차이 나는 두 메모리가 붙는다는 사실 자체가, 가속기 한 종에 HBM 한 종을 붙이던 시절의 공식이 깨졌음을 보여준다.

훈련 워크로드는 대역폭과 용량을 동시에 최대로 요구하는 유일한 구간으로 남는다. 미국 CSIS 의 그레고리 앨런은 HBM 이 고급 AI 칩 전체 가치의 약 절반을 차지한다고 지적했고, 훈련 쪽 칩의 HBM 침투율은 100%에 근접한다(참고 자료 G). 대규모 병렬 훈련에서는 그래디언트 동기화와 가중치 갱신이 수시로 일어나 칩당 메모리 대역폭이 곧 훈련 시간을 결정하므로, 이 구간에서는 HBM3E·HBM4 같은 최상위 적층 D 램을 대체할 수단이 없다. 다만 훈련 수요는 모델 기업 수와 기반 모델 경쟁의 강도에 묶여 있어 절대 물량의 성장률은 추론보다 낮다. 메모리 공급자 관점에서 훈련 구간은 고단가·고마진이지만 물량 확장의 주 엔진은 아니라는 뜻이다.

장문 추론은 용량과 KV 캐시 관리가 병목인 구간으로, 대역폭 요구가 프리필과 디코드 사이에서 갈라진다. 프리필은 입력 토큰을 한꺼번에 처리하는 연산 집약 단계라 상대적으로 낮은 대역폭(1.6TB/s)으로 버틸 수 있고, 디코드는 토큰을 하나씩 생성하며 매번 모델 가중치와 KV 캐시를 읽어야 하는 대역폭 집약 단계라 4TB/s 급이 필요하다. 화웨이가 950PR 에 저대역폭 HiBL 1.0 을, 950DT 에 고대역폭 HiZQ 2.0 을 배정한 것은 이 소프트웨어 단계 분리를 메모리 사양으로 옮긴 것이다(화웨이 전연결대회 2025, 2025-09-18). 이 분기의 함의는 크다. 장문 추론 서비스가 늘수록 '최고 사양 HBM 하나'가 아니라 '대역폭이 다른 두 등급의 HBM'이 시스템당 혼용되고, 프리필 노드에는 중급 대역폭 메모리로도 진입할 수 있는 틈이 생긴다.

온디바이스 구간의 결정 변수는 전력당 성능이며, 여기서는 LPDDR 계열이 HBM 의 자리를 대신한다. AI 스마트폰 출하가 2026 년 4.7 억 대를 돌파하고 침투율이 38%에 이를 것이라는 전망이 있고, AI 웨어러블 시장은 2024 년 419 억 달러에서 2028 년 1,207 억 달러로 커질 것으로 전망된다(업계 보도). 이 디바이스들은 배터리 예산 안에서 추론을 돌려야 하므로 저전력 특화의 LPDDR5X 급 메모리와 SoC 내장 NPU 의 조합이 표준이 된다. 중국 진영에서 로크칩(Rockchip, 瑞芯微)은 2026 년 상반기 온디바이스 관련 매출이 전년 동기 대비 40% 이상 늘었고, 0.2~6 TOPS 의 AIoT 칩 라인업과 세계 첫 3D 적층 AI 코프로세서 RK182X 의 양산에 들어갔다. 이 시장에서 경쟁 축은 절대 대역폭이 아니라 와트당 토큰 처리량이며, HBM 공급 능력은 무의미하다.

엣지 배치 구간, 특히 차량과 산업 현장의 추론은 단가가 지배 변수다. 호라이즌 로보틱스(Horizon Robotics, 地平线)는 자국 브랜드 ADAS 시장 점유율 47.7%로 연속 1 위이며 27 개 완성차 그룹 42 개 브랜드에

채택됐고, 블랙세서미(Black Sesame, 黑芝麻智能)는 화산 A2000 계열(7nm, 200~2,000 TOPS)로 해외 판매 채널을 확보했다. 이 시장의 칩은 GDDR 계열 그래픽 메모리나 성숙 공정 D 램으로 요구 대역폭을 충족하는 경우가 많다. HBM 은 적층·TSV·첨단 패키징 비용 때문에 와트당·달러당 효율이 엷지 예산에 맞지 않고, GDDR7 급은 HBM 대비 절대 대역폭은 낮지만 단가와 공급 탄력에서 우위에 있어 차량·로봇 같은 대량 배치 시장의 현실적 선택지다.

이 네 워크로드를 메모리 계보로 정리하면 다섯 갈래다. 첫째, 훈련과 고급 디코드용 HBM3E·HBM4 는 대역폭 극대화 계보로 기술 장벽이 가장 높다. 둘째, LPDDR5X 는 온디바이스 전력 예산에 최적화된 계보다. 셋째, GDDR7 은 엷지 배치의 단가·대역폭 균형 계보다. 넷째, CXL 확장은 호스트 메모리 풀을 가속기와 공유해 장문 추론의 KV 캐시를 HBM 밖으로 오프로드하는 계보로, '칩 옆 HBM 용량'이라는 제약 자체를 완화한다. 다섯째, PIM·메모리-연산 통합은 연산을 메모리 내부로 밀어 넣어 데이터 이동 자체를 줄이는 계보로, 로크칩의 3D 적층 코프로세서 시도처럼 온디바이스 쪽에서 먼저 제품화되고 있다. 갈라진 자리마다 패키징과 전력 예산이 다르므로, 메모리 기업의 제품 믹스 전략도 이 다섯 축으로 재편될 수밖에 없다.

화웨이 자체 HBM 의 등장은 한국 공급자에게 양면의 사건이다. 한 면에서 HiBL 1.0 과 HiZQ 2.0 이 실물 사양(각각 128GB·1.6TB/s, 144GB·4TB/s)과 함께 공개됐고 950PR 이 2026 년 1 분기 출시되면서 채택이 확정됐다는 것은, 중국이 HBM 의 '존재' 단계를 넘어 '제품 탑재' 단계에 들어섰음을 뜻한다. 15·5 규칙이 HBM3/4 의 기술 업그레이드와 규모화 생산을 2030 년 목표로 못 박았고, 메모리 자급률이 2020 년 5% 미만에서 2026 년 35%로 올랐다는 정부 해설 수치(전자산품세계, 2026-03-17)는 이 방향의 지속성을 뒷받침한다. 다른 면에서 HiBL·HiZQ 의 양산 능력·수율·공급량은 여전히 비공개다. 딥시크의 어센드 950DT 16 만 장 배치 계획(블룸버그 2026-09-04 보도)의 전량 인도에 1 년 이상 걸릴 수 있다는 보도 자체가 공급 병목의 존재를 시사하며, 그 병목의 상당 부분은 자체 HBM 수급에 묶여 있다. '자체 HBM 이 있다'는 사실과 '외부 공급이 불필요해졌다'는 명제 사이에는 수율과 물량이라는 검증되지 않은 공백이 놓여 있다.

한국의 기회는 규모로가 아니라 조건으로 정의돼야 한다. **다만 전제를 먼저 못 박는다 — 이 기회는 대중 HBM 직수출이 아니다.** 미 상무부 산업안보국(BIS)의 2024 년 12 월 2 일 규칙은 메모리 대역폭 밀도가 3.3 GB/s/mm² 미만일 때만 예외를 허용하고 현행 HBM 대부분이 이 문턱을 넘으므로, 직판 경로는 이미 닫혀 있다. 따라서 아래 네 조건은 ① 중국이 자국산 전환으로 글로벌 수요에서 빠지며 생기는 비중국 물량의 재배분, ② 통제 기준 미만 사양의 범용 서버 메모리, ③ 중국 외 지역의 중국계 고객이라는 세 경로 위에서 읽어야 한다. 첫째, 고대역폭 디코드용 HBM 구간에서 2026 년 4 분기 950DT 출시 이후 HiZQ 2.0 의 실제 공급량이 발주 물량을 못 따라가는 기간이 열리면, 그 기간 동안 한국 HBM3E 이상급 제품에 수요 창이 생긴다. 이 창의 길이는 HiZQ 수율 곡선과 창신메모리(CXMT, 长鑫科技)의 HBM 진입 속도에 의해 달린다. 둘째, 프리필용 1.6TB/s 급 같은 중대역폭 구간은 자국산 대체의 우선순위가 높은 만큼 한국의 진입 여지가 오히려 좁고, 이 구간에서 기회가 생긴다면 그것은 기술 격차가 아니라 자국산 수율 사고 같은 공급 차질 국면에 한정된다. 셋째, 창신메모리의 과창판 상장(2026-07-29, 첫날 +465.82%, 시총 3.28 조 위안)으로 확보된 증설 자금이 실제 HBM 급 양산으로 전환되는 시점이 확인되기 전까지, 즉 15·5 의 '규모화 생산' 문구가 출하 실적으로 증명되기 전까지가 협상력이 남아 있는 기간이다. 넷째, D 램 계약가가 2026 년 2 분기 전분기 대비 +58~63% 오를 것이라는 전망(시장조사·보도, 전망치) 아래에서는 중국 수요자도 가격보다

공급 확보를 우선하므로, 장기공급계약의 조건으로 프리미엄을 확보할 수 있는 창이 열린다. 네 조건 모두 '언제까지, 어떤 제품군에서, 어떤 사건이 확인되면 달하는가'를 추적 지표로 삼아야 하며, 그 지표는 950DT 인도 실적, HiZQ 공급량, 창신메모리 HBM 양산 공시다.

자국산 HBM 의 시계도 구체화됐다. 창신메모리(CXMT)는 2024 년 하반기 4 층 HBM2 양산을 시작해 2025 년 10 월 16nm 공정 HBM3 샘플을 공급했고, 2026 년 8 층 HBM3 양산과 **2027 년 12 층 HBM3E 양산**을 목표로 한다. 실현되면 한국·미국 3 사와의 기술 격차가 2~3 년으로 좁혀진다. 창신은 **2026 년 D 램 생산능력의 20%를 HBM 으로 전환해 웨이퍼 기준 월 6 만 장을** 배정하는 계획도 갖고 있다(중국·대만 업계 매체 종합, 등급: 보도).

그러나 현재의 벽은 수율이다. 자국산 HBM 은 초기 수율 35~40% 수준으로 상용 임계치인 70~80%에 **미치지 못한다**는 것이 업계의 공통된 관측이며, 창신 HBM3E 의 성능이 국제 선두 대비 뒤진다는 평가와 공칭 대역폭이 대량 양산에서 안정될지 미지수라는 지적이 함께 나온다. 모건스탠리는 창신의 D 램 공정이 삼성·마이크론 대비 약 2 세대 뒤쳐진다고 보면서도 **2028 년을 분수령**으로 지목한다. 수요 측 지표도 함께 볼 필요가 있다 —뱅크오브아메리카는 HBM 시장이 2025 년 약 350 억 달러에서 2030 년 약 2,460 억 달러로 연평균 34% 성장하고, **AI 가속기 한 대당 탑재 HBM 이 187GB 에서 464GB 로** 늘어난다고 추정한다. 수요가 이 속도로 커지는 동안 자국산 수율이 임계치를 넘지 못하면, 창은 수치상으로도 열려 있다.

온디바이스 AI 반도체 지도

온디바이스 AI 반도체는 데이터센터 가속기와 다른 경제 논리로 움직인다. 데이터센터가 절대 성능과 총소유비용을 겨루는 시장이라면, 온디바이스는 배터리와 발열 예산 안에서 전력당 성능을 겨루는 시장이다. 2026 년 AI 스마트폰 출하는 4.7 억 대를 돌파하고 침투율은 38%에 이를 것으로 전망되며(업계 보도, 2026 년 기준), AI 웨어러블 시장은 2024 년 419 억 달러에서 2028 년 1,207 억 달러로 성장할 전망이다(같은 출처). 이 수요가 실리콘으로 환산되면 연간 수억 대의 SoC·NPU 물량이 되며, 이 절에서는 이 물량이 어느 제조사의 어떤 제품으로 흘러가는지를 탑재 실적 중심으로 정리한다. 발표만 있고 탑재 실적이 확인되지 않은 제품은 별도의 로드맵 표로 분리한다.

스마트폰 — 물량은 크지만 검증 가능한 중국 자국산 스펙은 부족하다

스마트폰 측은 세 축 가운데 물량이 가장 크지만, 이번 호가 확보한 검증 자료에서 중국 자국산 애플리케이션 프로세서(AP)의 공정·NPU 성능·탑재 실적을 병기할 수 있는 제품은 없다. 아래 표는 그 공백 자체를 기록하는 것이 목적이다.

제조사	대표 제품	공정	컴퓨팅 파워(NPU)	실제 탑재처
중국 자국산 AP 벤더 전반	미공개	미공개	미공개	미공개

이 공백은 해석할 가치가 있다. 스마트폰 AP 는 선단 공정에 대한 요구가 가장 강한 제품군인데, 2026 년 기준 중국 대륙의 7nm 이하 선단 공정 생산 비중은 2% 미만이고 TSMC·삼성과 2~3 세대 기술 격차가 존재한다(15·5 해설 자료, 2026-03-17 전자산품세계). 즉 스마트폰용 최상위 SoC 는 자국산화가 가장 늦는 구간이며, 온디바이스 AI 물량의 상당 부분이 이 구간에 걸려 있다는 뜻이다. 반대로 스마트폰 주변의 AIoT·서브 칩 구간에서는 자국산 진영의 진입이 확인된다 — 그것이 아래 웨어러블·AIoT 축이다.

피지컬 AI·자동차 — 자국산 진영이 점유율을 확보한 유일한 축

차량·로봇 측은 세 축 가운데 중국 자국산 칩이 시장 지위를 문서로 입증한 유일한 구간이다. 15·5 계획이 차량용 칩 자급률을 2026 년 기준 20% 미만으로 적시한 것(전자산품세계, 2026-03-17)과 겹보기에 충돌해 보이지만, 둘은 다른 측정이다 — 자급률은 차량 전체 칩(파워·MCU·센서 포함) 기준이고, 아래 점유율은 스마트 드라이빙 SoC 라는 특정 품목 기준이다.

제조사	대표 제품	공정	컴퓨팅 파워	실제 탑재처
호라이즌 로보틱스(Horizon Robotics, 地平线)	정청(征程) 계열	미공개	미공개	자국 브랜드 ADAS 시장 점유율 47.7%로 연속 1 위, 27 개 완성차

				그룹·42 개 브랜드 채택(업계 보도, 2026 년 기준)
블랙세서미(Black Sesame, 黑芝麻智能)	우당(武当) 계열	미공개	미공개	2024~2025 년 차량 탑재 실적 확인(업계 보도)
로크칩(Rockchip, 瑞芯微)	RK3588·RK3576	미공개	6 TOPS NPU 탑재	AIoT·엣지 디바이스 전반, 로봇·산업용 단말(업계 보도, 2026 년 기준)

호라이즌 로보틱스의 47.7%는 자국산 대체가 '수출 통제로 밀려난 빈자리'가 아니라 자국 완성차의 설계 채택 경쟁에서 정면으로 얻은 점유율이다. 42 개 브랜드라는 채택 폭은 단일 고객 의존이 아니라 플랫폼화가 진행됐음을 보여준다.

웨어러블·AIoT — 저전력 저 TOPS 구간의 조용한 확장

웨어러블·AIoT 측은 칩당 단가와 컴퓨팅 파워는 낮지만 대수가 크고, 선단 공정 제약에서 가장 자유로운 구간이다. 로크칩이 이 축의 대표 주자로, 2026 년 상반기 온디바이스 관련 매출이 전년 동기 대비 40% 이상 증가했다(업계 보도, 2026 년 상반기).

제조사	대표 제품	공정	컴퓨팅 파워	실제 탑재처
로크칩(Rockchip, 瑞芯微)	AIoT 칩 전 계열	미공개	0.2~6 TOPS	AIoT 단말 전반(업계 보도, 2026 년 기준)

0.2~6 TOPS 라는 대역은 데이터센터 기준으로는 무의미한 숫자지만, 수 밀리와트 전력 예산 안에서 음성 인식·키워드 검출·저해상도 비전을 상시 구동하기에 충분한 수치다. 이 구간의 승부처는 피크 성능이 아니라 유헤 전력과 웨이크업 지연이며, 그래서 아래에서 다룰 메모리 선택이 성능을 좌우한다.

로드맵 — 발표는 됐으나 탑재 실적이 없는 것들

아래 제품들은 사양 또는 양산 단계 진입이 공표됐으나 아직 탑재 실적이 없다. 실적 표와 분리한다.

제조사	대표 제품	공정	컴퓨팅 파워	상태
블랙세서미(Black Sesame, 黑芝麻智能)	화산(华山) A2000 계열	7nm	200~2,000 TOPS	해외 판매 채널 확보(업계 보도, 2026 년 기준)
로크칩(Rockchip, 瑞芯微)	RK182X	미공개	미공개	세계 첫 3D 적층 AI

瑞芯微)				코프로세서로 양산 단계 진입(업계 보도, 2026 년 상반기)
15·5 기획 목표	AI 칩 전반	—	200 TOPS 이상	2028 년 전 양산이 기획 목표(전자산품세계, 2026-03-17)

블랙세서미의 화산 A2000 이 7nm·최대 2,000 TOPS 를 표방하는 것은 자동차 SoC 가 선단 공정 경쟁으로 미끄러져 들어가고 있음을 보여준다. 7nm 는 중국 대륙 생산 비중 2% 미만의 구간이므로, 이 제품의 파운드리가 어디인지가 자국산 대체 논의의 실질을 가르다. 15·5 의 '2028 년 전 200TOPS+ 양산' 목표는 A2000 의 스펙 레인지(200~2,000 TOPS)와 정확히 겹치는데, 이는 기획 목표가 이미 발표된 민간 제품 로드맵을 뒤따라 적은 것일 가능성을 시사한다. 계획 문구가 아니라 양산 실적을 추적해야 한다는 1 절의 추적 원칙이 여기서도 그대로 적용된다.

온디바이스가 메모리에 요구하는 것 — 용량이 아니라 전력당 대역폭이다

온디바이스 워크로드의 병목은 연산이 아니라 메모리 이동이다. 수십 억 파라미터의 모델을 스마트폰·이어폰·차량에서 추론할 때 소모 전력의 상당 부분은 가중치를 외부 D 램에서 연산 유닛으로 옮기는 데 쓰인다. 배터리 예산이 고정된 단말에서 용량을 늘리는 것은 전력 비용을 키울 뿐이며, 경쟁 변수는 '와트당 몇 GB/s 를 공급하는가'로 압축된다. 이것이 5 절에서 정리한 다섯 메모리 계보 가운데 온디바이스 측이 LPDDR5X 의 자리를 차지하는 이유다 — HBM 처럼 절대 대역폭을 극대화하는 대신, 저전압 동작으로 전력당 대역폭을 극대화하는 계보다.

같은 논리가 두 개의 구조적 선택을 밀어올린다. 첫째는 온칩 SRAM 확대다. 자주 쓰는 가중치와 활성값을 칩 내부 SRAM 에 상주시키면 외부 D 램 접근 자체를 건너뛸 수 있으므로, 전력당 대역폭 경쟁의 종착점은 '외부 메모리에 가지 않는 것'이 된다. 로크칩 RK182X 가 3D 적층 구조로 AI 코프로세서를 만든 것(업계 보도, 2026 년 상반기)은 연산과 메모리의 물리적 거리를 패키징으로 좁히려는 시도로 읽을 수 있으며, 첨단 패키징이 온디바이스 성능의 변수로 진입했음을 보여준다. 둘째는 PIM(Processing-in-Memory)이다. 연산을 메모리 내부로 옮기면 데이터 이동 전력이 구조적으로 제거되므로, 5 절에서 PIM 을 별도 계보로 분리한 근거가 온디바이스 측에서 가장 선명하게 드러난다.

이 재편의 함의는 부품 수요의 순서에 관한 것이다. 온디바이스 물량이 수억 대로 커져도 그것이 HBM 수요로 연결되지는 않는다 — 연결되는 것은 저전력 D 램, 고용량 온칩 SRAM, 그리고 메모리-연산 근접 패키징이다. 한국 메모리가 데이터센터 측에서 어센드 950 의 자체 HBM 공백이라는 1~2 년짜리 창을 두고 계산하는 동안, 온디바이스 측에서는 LPDDR·PIM 이라는 다른 계보의 경쟁이 별도의 시계로 돌아가고 있다. '어떤 부품의 수요가 언제 늘고 언제 꺾이는가'라는 이 호의 질문은 온디바이스에서도 동일하게 적용되며, 답은 칩 점유율이 아니라 메모리 계보별 전력당 대역폭 경쟁에서 갈린다.

스마트폰 축에서 경쟁의 기준이 바뀐 것도 같은 맥락이다. 미디어텍 텐지(Dimensity) 9500 은 고성능 NPU 와 초고효율 NPU 를 나눠 신품 효율 NPU 에 **컴퓨팅-인-메모리(CIM, 存算一体)** 구조를 적용해 저부하 소형 모델 구동 시 전력을 42% 줄였고, 성능 NPU 는 피크 100 TOPS 를 내면서 피크 전력을 56% 낮췄다(제조사 발표, 2025-09). 숫자보다 중요한 것은 방향이다 — **7B 모델을 INT4 로 양자화하면 약 4GB 가 필요해 12~16GB LPDDR5X 를 단 플래그십에서 구동 가능해지고**, 그 순간 승부는 피크 TOPS 가 아니라 **메모리 용량·대역폭·전력 예산의 삼각 최적화**로 옮겨간다. 단말 AI 가 요구하는 것은 HBM 이 아니라 LPDDR5X·LPDDR6 와 온보드 고대역 메모리이며, 상주 모델이 보편화될수록 플래그십의 메모리 하한선이 올라간다. 연 10 억 대 단말이라는 모수를 곱하면 총량에서 무시할 수 없는 수이다.

피지컬 AI 쪽의 제약은 데이터센터와 완전히 다르다. 배터리 용량, 무게 상한, 좁은 방열, 비용 레드라인이 설계를 지배한다. 그래서 로봇 업계의 표준 구성은 오랫동안 **'대뇌(고성능 추론 SoC) + 소뇌(실시간 운동 제어 MCU)'**의 이중 칩이었고, 이는 칩 간 스케줄링 지터와 전력 비용을 수반했다. 호라이즌 로보틱스 계열이 내놓은 연산·제어 일체형 SoC 는 하나의 다이에 인지용 대형 코어와 실시간 제어용 코어를 함께 얹어 이 지터를 없애는 방향이며, BOM 과 전력을 함께 줄이는 것이 판매 논리다. 자동차 축의 경쟁 기준도 절대 성능이 아니라 **TOPS 당 가격**으로 재편됐다 — 엔비디아 오린(Orin)·토르(Thor) 계열에 자국산 SoC 가 같은 연산력을 더 낮은 단가로 맞붙이는 구도이며, 완성차의 자체 칩 개발(비야디의 4nm 자율주행 칩 양산 등)이 고가 칩의 프리미엄을 소프트웨어 수직 통합으로 압축한다.

웨어러블은 또 다른 극단이다. 여기서는 TOPS 가 아니라 **밀리와트와 면적**이 승부를 가른다. AI 안경 전용 칩 가운데 12nm 공정에 6 TOPS NPU 를 얹고 1080P 녹화 시 칩 전체 전력을 300mW 미만으로 묶은 제품(에너지 효율 약 10 TOPS/W)이 등장했고, 저전력 무선 SoC 와 Wi-Fi MCU 진영이 이 구간을 나눠 갖는다. 시장조사기관 옴디아(Omdia)는 **2026 년 AR 안경 출하가 전년 대비 154% 늘어 약 410 만 대**에 이를 것으로 전망한다(등급: 시장조사·전망치). 정리하면 AI 반도체는 데이터센터의 기가와트급 클러스터에서 안경의 300mW SoC 까지 **여섯 자릿수 전력 스펙트럼** 위에 분포하며, 각 구간의 승리 조건이 전부 다르다. 최고 사양 GPU 하나가 모든 시장을 지배하던 시대는 끝났다.

공급망 사다리 6 단계 재점검과 자국산화 목표

공급망 사다리 전체를 점검하면 결론은 하나다. 15 차 5 개년 계획(2026~2030)이 제시한 자국산화 목표는 사다리의 아래 단계(소재·장비)보다 위 단계(팹리스·파운드리)에 더 가까이 와 있는 현재 실적과 정면으로 어긋나며, 이 어긋남의 크기가 곧 앞으로 5 년간 중국 정책 자금이 밀어 넣어야 할 거리다. 《중화인민공화국 국민경제·사회발전 제 15 차 5 개년 계획 강요(2026~2030)》(2026 년 3 월 정식 발표)는 집적회로를 10 대 신산업·신트랙의 1 위에 배치하고 "초상규 조치를 취하고 신형 거국체제를 완비해 전 사슬로 핵심기술 공략에서 결정적 돌파를 이룬다"고 적었다(해설: 전자산품세계, 신량재경 전재, 2026-03-17). 14·5 의 '단점 보완'이라는 방어적 문구가 '전 사슬 돌파'의 공세적 문구로 바뀐 것은, 6 단계 사다리 어디에도 미달을 인정할 여유가 없다는 선언이다.

목표 숫자를 정면으로 나열한다. 2030 년 기준 ▲반도체 산업 규모 3 조 위안 돌파 ▲전체 자급률 70% ▲22nm 이하 선단 공정 자국산화율 50% ▲성숙 공정 생산능력 세계 비중 52% ▲12 인치 라인 자국산 장비 커버리지 60% ▲포토리지스트(ArF/EUV) 자급률 50% ▲12 인치 실리콘 웨이퍼 100% 자국산 ▲2028 년 전 200TOPS 이상 AI 칩 양산 ▲HBM3/4·3D NAND 기술 업그레이드와 규모화 생산이다(15·5 강요 해설, 2026-03-17). 전사회 연구개발 경비 연평균 7% 이상 증가와 국가대표펀드(대기금) 3 기 3,000 억 위안 초과 규모의 장비·소재·EDA 중점 투자가 재원으로 배정됐다.

첫 번째 단계인 소재는 목표가 가장 공격적이면서 현재 기반이 가장 얇은 구간이다. 12 인치 실리콘 웨이퍼 100% 자국산이라는 목표는 완전 자립을 뜻한다. 포토리지스트는 자급률 50%가 목표로 명시됐다는 사실 자체가 현재 미달의 확인이다. 다만 목표가 ArF 와 EUV 를 함께 지목했다는 점이 중요하다. 자국산 가속기 양산 시나리오에서 7nm 급 공정이 필요한 어센드 950DT(2026 년 4 분기 출시 예정, 화웨이 전연결대회 2025, 2025-09-18)의 물량을 받치려면 ArF 이말전 이상의 포토리지스트가 대량으로 안정 공급돼야 하므로, 이 단계의 병목은 수치가 비어 있다는 사실 자체가 신호다.

두 번째 단계인 장비는 목표 60% 대비 현재 40% 미만(12 인치 라인 자국산 장비 커버리지, 2026 년 기준, 15·5 해설)으로 간극이 숫자로 확인되는 구간이다. 중국 대륙은 6 년 연속 세계 1 위 반도체 장비 투자 규모를 유지하고 있고, 양쯔메모리 우한 3 기 공장의 자국산 장비 비중 52%, 창신메모리 45%(산업지 보도, 2026 년 상반기)라는 라인 단위 실적이 존재한다. 나우라(NAURA, 北方华创)와 AMEC(中微公司)이 식각·증착 계열에서 커버리지를 넓히는 동안, 노광 장비는 여전히 자국산 대체의 최대 난관으로 남아 있다. 장비 투자 세계 1 위와 커버리지 40% 미만이 공존한다는 것은, 돈은 쓰고 있으나 노광을 축으로 한 최종 관문이 아직 열리지 않았다는 뜻이다.

세 번째 단계인 EDA/IP 는 이번 해설 자료에서 유일하게 자급률 목표 숫자가 없는 단계다.

국가대표펀드(대기금) 3 기가 EDA 를 중점 투자 대상으로 명시했다는 사실이 이 단계의 취약성을 대신 말해 준다. 그러나 자국산 가속기 양산 시나리오에서 EDA 는 우회로가 없는 관문이다. 어센드 950 아키텍처가 FP8/MXFP8/HIF8 1 PFLOPS, MXFP4 2 PFLOPS 의 저정밀 포맷과 상호연결 대역폭 2.5 배를 새로 구현하고(화웨이 전연결대회 2025), 링취(灵衢) 상호연결로 최대 8,192 장의 아틀라스 950 슈퍼노드를

목는 설계는 설계 도구 체인 전체가 그 사양을 따라가야 성립한다. 목표 숫자가 비어 있는 것이 이 단계의 취약성을 보여주는 정황이다.

네 번째 단계인 팹리스는 사다리에서 유일하게 목표를 초과 달성할 가능성이 보이는 구간이다. 2028 년 전 200TOPS 이상 AI 칩 양산이라는 목표에 대해, 블랙세서미(Black Sesame, 黑芝麻智能)의 화산 A2000 계열은 이미 7nm 공정에 200~2,000 TOPS 를 내걸고 해외 판매 채널을 확보했다(업계 보도, 2026 년 기준). 호라이즌 로보틱스(Horizon Robotics, 地平线)는 자국 브랜드 ADAS 시장 점유율 47.7%, 27 개 완성차 그룹 42 개 브랜드 채택으로 설계 역량의 상업적 검증을 끝냈다. 화웨이 어센드는 950PR(2026 년 1 분기)·950DT(2026 년 4 분기)·960·970 의 3 개년 로드맵을 공표했고, 딥시크가 네이멍구 데이터센터에 950DT 최소 16 만 장을 배치할 계획이라는 조달 의향이 보도됐다(블룸버그 2026-09-04 보도). 설계는 된다. 문제는 그 설계를 찍어낼 아래 단계다.

다섯 번째 단계인 파운드리가 이 시나리오의 진짜 병목이다. 성숙 공정 생산능력 세계 52%라는 2030 년 목표에 대해 현재 12 인치 생산능력 세계 비중은 25%다(2026 년 기준, 15·5 해설). 겉으로는 2 배 남짓의 거리지만 구성이 문제다. 중국 대륙 12 인치 생산능력의 77%가 28nm 이상 성숙 공정에 몰려 있고, 7nm 이하 선단 공정 비중은 2% 미만으로 TSMC·삼성과 2~3 세대 격차가 유지되고 있다(2026 년 기준, 같은 출처). 22nm 이하 선단 공정 자국산화율 50%라는 목표는 이 2% 미만의 바닥에서 출발하는 목표다.

어센드 950DT와 화산 A2000 같은 200TOPS 급 칩은 모두 선단 공정을 요구하므로, SMIC(中芯国际)와 화홍반도체(Hua Hong)의 선단 램프업이 팹리스 단계의 설계 성과를 실리콘으로 바꿀 수 있느냐가 16 만 장 발주의 실물화를 결정한다.

여섯 번째 단계인 패키징·첨단 패키징은 목표 표에 명시 수치가 없지만, HBM 목표(HBM3/4 기술 업그레이드·규모화 생산)가 사실상 이 단계의 목표를 대신한다. 화웨이가 자체개발 HBM 인 HiBL 1.0(128GB·1.6TB/s)과 HiZQ 2.0(144GB·4TB/s)의 제품 탑재를 확정된 것은 패키징 역량의 존재를 보여주나, 양산 능력·수율·공급량은 비공개다. 고대역폭 메모리의 적층과 인터포저 실장이 칩 전체 가치의 약 절반을 좌우한다(미 CSIS 그레고리 앨런 설명 인용, 15·5 해설 자료)는 구조에서, 패키징 수율은 선단 공정 수율과 별개로 두 번째 게이트가 된다.

목표와 실적의 간극을 정리하는 마지막 숫자가 이 절의 논지다. 《중국제조 2025》기술 로드맵의 집적회로 자급률 목표는 2025 년 70%였고, 2020 년 국무원 관련 논의에서도 "향후 5 년 내 70%, 당시 30%"가 언급됐다(관찰자망 2020-08-20). 실제로 IC 인사이츠는 2025 년 중국 내 생산 칩이 중국 집적회로 시장의 19.4%에 그칠 것으로 전망했다(2021 년 전망치). 목표 70%와 실적 19.4%, 50 포인트 이상의 미달이다. 그리고 같은 70%가 15·5 에서 2030 년 목표로 5 년 이월됐다. 현재 수치들 — 메모리 자급률 35%(2020 년 5% 미만에서 상승), 차량용 칩 자급률 20% 미만, 장비 커버리지 40% 미만, 7nm 이하 2% 미만(모두 2026 년 기준, 15·5 해설) — 는 모두 진행 중인 상승 곡선이지만, 목표가 한 번 미끄러진 전력이 있으므로 계획 문구가 아니라 양산 실적으로 검증해야 한다. 950DT 인도 물량, HiZQ 2.0 공급량, 양쯔메모리 우한 3 기(월 10 만 장 추가, 총 30 만 장 목표, 보도)의 램프업, 자국산 장비 커버리지의 분기별 변화가 2030 년 70%의 진위를 가를 추적 지표다.

양산으로 넘어가는 길목의 병목은 세 겹으로 정리된다. **첫째 웨이퍼** — SMIC 가 어센드 전 시리즈의 N+ 계열 공정을 사실상 독점하는데 장비 통제로 선단 공정 확장에 물리적 천장이 있다. 여기서 배분 문제가 생긴다. 업계 보도에 따르면 SMIC N+2 생산능력의 **약 43%**를 **화웨이가 가져가고 한우지(Cambricon, 寒武纪)가 9~11% 수준**을 배정받으며 나머지를 다른 업체들이 나눈다(중국 업계 매체, 등급: 보도). 자국산 칩 경쟁의 실질은 설계 경쟁이 아니라 **파운드리 좌석 배분 경쟁**이라는 뜻이다. **둘째 첨단 패키징** — 고급 칩에 필요한 2.5D 칩렛 패키징을 중국에서 양산할 수 있는 곳은 성허징웨이(약 85% 점유)와 톈푸마이프로 정도로 한정되고 배치당 납기가 1.5~2 개월이라 연말이면 정체가 생긴다. **셋째 HBM** — 앞 절에서 본 대로 자국산은 아직 양산 검증 단계다. 이 3 중 병목이 풀리는 속도가 곧 16 만 장 계획의 현실화 속도이며, 동시에 중국발 가격 하락 압력의 방아쇠다.

시장 지형을 숫자로 보면 국산화의 현주소가 더 분명하다. IDC 집계로 **2025 년 중국 AI 서버 시장에 납품된 AI 가속기는 약 400 만 개이고 그중 자국산 비중이 41%**였으며, 업체별로는 화웨이 어센드가 약 81 만 2,000 장으로 압도적 1 위였다(등급: 시장조사). 후발 4 개사(무어쓰레드·무시·비런·텐수즈신)의 2025 년 합산 출하는 화웨이 한 회사의 3 분의 1 수준에 그친다. 자국산 대체는 진행 중이되 **한 회사에 극도로 집중된 형태로** 진행되고 있다.

자본시장 — 메모리·가속칩 기업의 상장 계보

2026 년 여름, 중국 메모리 산업은 자본시장을 통해 증설 자금의 수도꼭지를 열었다. 이 절이 추적하는 것은 주가가 아니라 조달금의 행선지다. 창신메모리(CXMT, 长鑫科技)는 상장을 완료했고, 양쯔메모리(YMTC, 长江存储)는 수리 단계에 있으며, 가속칩 기업 엔플레임(Enflame, 燧原科技)은 같은 창에서 별도의 자금 줄을 뽑았다. 세 기업의 자금이 실제로 흘러갈 곳은 모두 생산능력과 장비로 수렴하며, 이것이 다음 절에서 다른 메모리 가격 전망의 공급 측 입력값이다.

창신메모리의 상장은 완료된 사실이다. 이 회사는 2026 년 7 월 29 일 과창판에 상장(종목코드 688825)해 상장 당일 시가총액 3.28 조 위안으로 A 주 시총 1 위에 올랐고, 8 월 10 일 MSCI 중국 전체지수에 편입됐다(CHAI 주간 invest·stock 2026-07-29 호). 시총 순위 자체가 아니라 지수 편입이 자본 구조에 남긴 흔적을 봐야 한다 — 편입과 함께 패시브 자금이 들어왔고, 이는 후속 유상증자·회사채 발행 같은 추가 조달의 가격 조건을 유리하게 만든다. 확실한 것은 회사의 증설 방향뿐이다 — 이전 절에서 짚었듯 D 램 자금률이 2020 년 5% 미만에서 2026 년 35%로 올라오는 과정(15·5 해설, 전자산품세계, 2026-03-17)의 주역이 창신메모리이며, 자본시장에서 잠금 해제된 자금이 이 궤적 위에 올라탄다. 주가 상승률과 수익률은 이 절의 분석 대상이 아니다.

양쯔메모리는 단계가 다르다. 2026 년 8 월 29 일 과창판 IPO 가 수리됐을 뿐, 상장이 완료되지 않았다(CHAI 주간 invest 2026-08-29 호). 조달 목표는 330 억 위안으로 과창판 사상 최대 규모이며, 회사는 2026 년 1 분기 순이익 333.79 억 위안을 공시했다(같은 출처). 수리에서 발행·청약·상장까지 절차가 남아 있어 자금이 실제 유입되는 시점은 아직 뒤에 있다. 그럼에도 이 IPO 가 던지는 질문은 명확하다 — 330 억 위안이 어디로 가는가.

그 행선지는 우한 3 기 공장이다. 산업지 보도(2026 년 상반기~중반, 등급: 보도)에 따르면 양쯔메모리의 합계 월 생산능력은 약 20 만 장이고, 우한 3 기 양산을 당초 2027 년에서 2026 년 하반기로 앞당겨 월 10 만 장을 추가, 총 30 만 장을 목표로 한다. 2026 년 1 분기 매출은 200 억 위안을 넘었고 NAND 생산량은 글로벌 10%를 초과했다(같은 보도). 출하량 기준으로는 이미 세계 3 위에 진입해 점유율 14%로 키옥시아를 추월했다(CHAI 주간 tech 2026-08-16 호). 즉 수리된 330 억 위안은 신규 시장 개척비가 아니라 이미 가동 중인 생산 기반 위에 얹는 증설비다. 자금이 '매출 이익이 나는 사업의 캐파 확장'으로 가는지 '적자 사업의 생존 연장'으로 가는지에 따라 공급 증가의 확실성이 갈리는데, 1 분기 순이익 공시는 전자를 시사한다.

조달금의 또 다른 목적지는 장비이고, 그 장비의 상당 부분이 자국산이다. 우한 3 기의 자국산 장비 비중은 52%, 창신메모리 라인은 45%로 보도됐다(산업지 보도, 2026 년 상반기~중반). 메모리 기업의 공모자금 중 장비 감가상각으로 돌아가는 몫이 절반가량 자국산 장비 기업의 매출로 전이된다는 뜻이다. 15·5 규칙이 2030 년 12 인치 라인 자국산 장비 커버리지 60%를 목표로 삼고 현재 40% 미만에 머무는 상황(15·5 해설, 2026-03-17)에서, 두 메모리 기업의 상장 자금은 그 격차를 메우는 재원이 된다. 국가대표펀드 3 기(규모 3,000 억 위안 초과, 장비·소재·EDA 중점)라는 공공 자금에 더해, 이제 자본시장 자금이 같은 방향으로 합류하는 구조다. 자금의 원천이 국고에서 공모시장으로 넓어졌다는 것이 2026 년 여름의 실질적 변화다.

엔플레임의 IPO 는 다른 층위의 자금 흐름을 보여준다. 이 AI 가속칩 기업은 8 월 28 일 수요예측, 9 월 2 일 청약 일정으로 과창판 IPO 를 진행해 60 억 위안을 조달하며, 텐센트가 최대주주이자 최대 고객이다(CHAI 주간 invest 2026-08-29 호). 330 억 위안의 양쯔메모리에 비해 규모는 작지만 구조가 다르다 — 가속칩 기업의 자금은 팹이 아니라 설계·테이프아웃·양산 위탁으로 가고, 그 칩이 완성되면 자국산 메모리의 수요처가 된다. 전월의 논지대로 소프트웨어가 하드웨어를 정의하는 국면에서 가속칩 기업이 늘수록 메모리 수요의 '규격을 정하는 주체'가 늘고, 엔플레임의 조달은 그 수요 주체가 자본시장에서 스스로 자금을 조달하기 시작했음을 보여준다. 텐센트가 주주이자 고객으로 겹쳐 있다는 점은 딥시크·화웨이 공동 설계와 같은 수직 결합이 자본 구조 단계에서도 재현되고 있음을 뜻한다.

세 건을 한 축에 놓으면 상장 계보의 순서가 읽힌다. 창신메모리(상장 완료, 2026-07-29) → 양쯔메모리(수리, 2026-08-29) → 엔플레임(청약, 2026-09-02). 한 달 남짓한 기간에 D 램, NAND, 가속칩의 세 축이 모두 과창판에서 자금 줄을 확보하거나 확보 절차에 들어갔다. 이것은 개별 기업의 재무 이벤트가 아니라 15·5 의 '초상규 조치·신형 거국체제'(2026 년 3 월 발표)가 자본시장 단계로 번역된 결과로 읽어야 한다.

이 조달이 다음 절의 가격 전망에 입력하는 값은 명확하다. 시장조사·보도 단계의 전망으로는 글로벌 메모리가 최근 15 년 중 최악의 공급 부족 국면에 있고 D 램 연간 250~280%, NAND 200~250% 상승이 겹쳐지며 2026 년 2 분기 계약가도 D 램 +58~63%, NAND +70~75%(전분기 대비)로 전망된다(2025-12 보도, 전망치이며 실측과 구분해야 한다). 이 전망을 꺾을 수 있는 공급 측 변수가 바로 이 절의 조달금이 만들 물량이다. 우한 3 기의 월 10 만 장이 2026 년 하반기에 실제 양산으로 들어오는지, 창신메모리의 자본시장 자금이 D 램 캐파로 전환되는 속도가 어느 정도인지가, 부족이 2027 년까지 이어진다는 전망의 성립 여부를 가른다. 이전 절들이 어센드 950 양산 전환기에 한국 고성능 메모리에 열리는 1~2 년짜리 수요 창을 확인했다면, 이 절의 상장 자금은 그 창을 닫는 시계의 초침이다. 조달금이 팹과 장비로 얼마나 빨리 옮겨가는가 — 그것이 한국 기업이 남은 창 안에서 취할 수 있는 물량의 상한을 결정한다.

설계 쪽의 상장 물결은 더 극적이다. 이른바 'GPU 4 소룡'이 반년 사이에 모두 자본시장에 들어왔다.

무어쓰레드(Moore Threads, 摩尔线程)는 2025 년 12 월 5 일 과창판에 상장해 시초가가 공모가 대비 **+468.78%**로 뛰며 시총 3,000 억 위안을 넘겼고(이후 4,000 억 위안대까지 상승), 상장 8 개월 만인 2026 년 8 월 홍콩 H 주 발행 준비에 들어갔다. **무시(MetaX, 沐曦)**는 12 월 17 일 과창판에 상장해 공모가 104.66 위안에서 시초가가 **약 5.7 배**로 출발해 역시 시총 3,000 억 위안을 돌파했고 6 월 H 주 발행 계획을 공고했다. **비런과기(Biren, 壁仞科技)**는 2026 년 1 월 2 일 홍콩거래소에 상장해 시초가 35.70 홍콩달러(공모가 대비 +82%)·장중 최고 42.88 홍콩달러로 시총 1,000 억 홍콩달러를 넘겼다. 다만 2026 년 1 월 초 무어쓰레드와 무시는 고점 대비 약 40% 조정을 받았다.

이 물결이 보내는 신호는 세 가지다. 첫째, 자국산 AI 칩이 '테마'에서 '실적'으로 넘어가고 있다. 둘째, 자본시장이 설비 투자와 연구개발의 연료를 대면서 자국산화의 자기강화 회로가 완성된다 — 통제가 자금 조달을 막지 못하고 오히려 자본시장 경로를 활성화시켰다. 셋째, 그럼에도 가장 부족한 자원은 자본도 기술도 아닌 **확정된 수요**로 옮겨가고 있다. 바이두 계열 쿤룬신이 홍콩 IPO 에서 투자자에게 청약 금액의 여러 배에

해당하는 칩 구매 약정을 요구했다는 보도(등급: 보도)는 그 징후다. 자금이 아니라 고객을 잠그기 위한 상장이라면, 그것은 공급 과잉의 전주곡이기도 하다.

가격 시계열 전망 — 양산 진입이 칩·메모리 가격을 언제 움직이는가

가격 전망은 한 줄 예측치로 정리할 수 없는 문제다. 지금 시장에 나돌고 있는 수치는 "D 램 연간 상승률 250~280%, NAND 200~250%, 2026 년 2 분기 계약가 전분기 대비 D 램 +58~63%, NAND +70~75%"(시장조사·보도 전망치, 2025 년 12 월 보도 인용)이고, 이는 '최근 15 년 중 최악의 공급 부족'이라는 전제 위에 서 있다. 이 전망치의 유효 기간은 중국 자국산 공급이 언제, 얼마나 시장에 진입하는가에 좌우된다. 따라서 이 절은 숫자 예측을 하지 않는다. 대신 세 시간 구간으로 나누고, 각 구간에서 가격이 꺾이거나 유지되려면 무엇이 먼저 관측돼야 하는지 선행 지표를 나열한다. 전제가 무너지는 반증 조건도 함께 적는다.

0~6 개월 구간(2026 년 4 분기~2027 년 1 분기) — 공급 부족 연장이 기본값이다. 이 구간에 중국 자국산 신규 물량이 시장 가격을 누를 가능성은 낮다. 양쯔메모리의 우한 3 기 공장은 양산 목표가 2026 년 하반기로 앞당겨졌으나 월 10 만 장 추가 물량이 실제로 웨이퍼 투입에서 수율 안정을 거쳐 출하로 이어지기까지는 통상 수 분기가 걸리며, 산업지 보도 기준 양쯔메모리 합계 생산능력은 약 월 20 만 장, 추가 후 30 만 장 목표에 불과하다(2026 년 상반기~중반 보도). 창신메모리의 D 램 증설분도 상장 직후(2026 년 7 월 29 일)인 점을 감안하면 이 구간 내 실물 출하 증가분은 제한적이다. 이 구간에 관측해야 할 선행 지표는 세 가지다. 첫째, 분기 계약가 보도의 기울기 변화다. 전망치대로 2 분기 +58~63%가 실측 계약가로 확인되는지, 그리고 3·4 분기에 상승률이 축소되는지가 첫 신호다. 둘째, 어센드 950DT 의 2026 년 4 분기 출시(화웨이 전연결대회 2025 발표, 2025-09-18) 이후 실제 인도 실적이다. 950DT 가 예정대로 나오면 중국 내부의 HBM 급 수요가 자체 소화로 이동하기 시작하며, 이는 한국산 고성능 메모리의 대중 수요 감소 신호이지만 동시에 글로벌 D 램 전체 가격에는 아직 중립적이다. 셋째, 딥시크 네이밍구 프로젝트의 칩 인도 진도다(블룸버그 2026-09-04 보도, 등급: 보도). 이 구간의 반증 조건은 화웨이 자체 HBM 인 HiZQ 2.0 의 공급 부족이다 — 수율·공급량이 비공개인 상태에서 950DT 인도가 지연되면 중국 AI 인프라 수요가 외부 메모리로 회귀하고, 부족 국면이 더 길어진다.

6~18 개월 구간(2027 년 상반기~하반기) — 가격의 분기점이 열리는 창이다. 이 구간에서 시장이 묻는 질문은 하나다. 우한 3 기와 창신메모리의 증설분이 '투입'을 넘어 '출하'로 전환되는가. 이 전환이 확인되는 순서는 정해져 있다. 먼저 공장 가동률 보도가 나오고, 다음으로 수율 관련 언급(공식 발표든 산업지 보도든)이 나오고, 마지막으로 계약 단가 하락 보도가 나온다. 순서가 뒤집히는 보도 — 가동률 확인 없이 갑자기 계약가가 흔들리는 경우 — 는 수요 측 요인일 가능성이 높으므로 공급 효과와 구분해야 한다. 이 구간의 핵심 선행 지표는 자국산 장비 비중이다. 양쯔메모리 우한 3 기 52%, 창신메모리 45%(보도)라는 수치는 공격적 증설의 징표인 동시에 양산 리스크의 징표다 — 자국산 장비 커버리지가 높은 라인일수록 초기 수율 안정화에 시간이 걸릴 수 있고, 이는 2026 년 하반기 양산 목표가 2027 년으로 미끄러지는 시나리오의 근거다. 실제로 양쯔메모리의 우한 3 기는 원래 2027 년 양산 목표였던 것을 앞당긴 일정이다. 이 구간의 전제는 '공급 부족이 2027 년까지 이어진다'는 전망(2025 년 12 월 보도)인데, 반증 조건은 두 방향 모두에 존재한다. 상방 반증은 중국 AI 수요의 폭발 — 950DT 인도가 순조롭고 딥시크 16 만 장급 프로젝트가 추가로 발표되면 내수 흡수가 늘어 오히려 글로벌 여유분이 줄어든다. 하방 반증은 우한 3 기·창신 증설분이

수율 안정 없이도 저가 덩핑성 출하로 시장에 풀리는 경우다 — 성숙 공정 세계 비중 52%라는 15·5 목표(15·5 강요 해설, 2026-03-17)가 '가격 결정권 확보'를 명시하고 있다는 점에서, 중국 자국산 진영이 손익보다 점유율을 우선하는 출하 전략을 택할 가능성은 배제할 수 없다.

18~36 개월 구간(2028 년~2029 년) — 구조적 재편이 가격보다 먼저 온다. 이 구간에서는 'D 램·NAND 평균 가격'이라는 단일 지표 자체가 의미를 잃기 시작한다. 이전 절들에서 확인한 것처럼 추론이 AI 인프라 지출의 70%를 넘기면서(15·5 해설 자료 인용, 2026 년 기준) 메모리 수요는 HBM3E/4·LPDDR5X·GDDR7·CXL 등 워크로드별 계보로 갈라졌고, 품목마다 가격 동력이 다르다. 화웨이 로드맵상 960·970 이 이 구간에 위치하고(화웨이 전연결대회 2025 발표), HiZQ 계열의 세대 교체가 진행되면 중국 내부에서 최상위 메모리 수요가 자체 소화로 넘어가는 속도가 빨라진다. 이 구간에 관측해야 할 선행 지표는 네 가지다. 첫째, 양쯔메모리·창신메모리의 증설 착공 발표 주기 — 상장 자금(창신 3.28 조 위안 시총, 창장 조달 목표 330 억 위안)이 4 기·5 기 신규 착공으로 이어지는지다. 둘째, 15·5 의 메모리 항목인 'HBM3/4·3D NAND 규모화 생산'(2030 년 목표)의 중간 실적 — 중국 자국산 HBM 이 화웨이 계열 외 상업 시장에 판매되는지가 결정적 분기점이다. 셋째, 12 인치 실리콘 웨이퍼 100% 자국산 목표(2030 년)의 진전 — 웨이퍼 자급이 이뤄지면 메모리 생산 원가 구조가 바뀌어 가격 하방 압력이 장기화된다. 넷째, 글로벌 메이저의 대응 투자 — 삼성·SK 하이닉스·마이크론의 캐פק스 축소 또는 전환(선단 D 램·HBM 으로의 믹스 이동) 보도가 나오면 성숙 메모리 가격 방어가 무너졌다는 신호다. 이 구간의 반증 조건은 미국 수출 통제의 재강화다 — 자국산 장비 비중 52%·45%라는 수치가 의미하듯 나머지 절반 가까이가 여전히 외부 장비이므로, 통제가 장비 유지보수·부품 단계로 확대되면 증설 일정 전체가 밀리고 공급 부족 국면이 2028 년 이후까지 연장될 수 있다.

세 구간을 관통하는 원칙은 하나다. 이 리포트 전체가 반복해 온 것처럼, 계획 문구가 아니라 실적을 추적해야 한다. 15·5 의 성숙 공정 세계 52%·자급률 70%는 2025 년에 미달한 목표의 5 년 이월본이므로(2020 년 30% → 2025 년 실적 19.4%, IC 인사이드 2021 년 전망), 가격 전망의 입력값으로는 목표가 아니라 분기별 실적 — 가동률, 수율, 계약 단가 보도, 착공 공시, 자국산 장비 비중 변화 — 만을 써야 한다. 250~280%라는 상승 전망치가 맞든 틀리든, 그것을 꺾는 힘은 전망 기관의 수정 보고서가 아니라 우한과 허페이의 클린룸에서 나온다.

시장의 견해는 세 진영으로 갈린다. **슈퍼사이클 지속 진영** — 노무라는 D 램·낸드·HBM 이 동시에 폭증하는 삼중 사이클이 2027 년까지 이어진다고 보고, 골드만삭스는 2027 년 D 램 수급 격차 전망을 더 악화하는 쪽으로 조정하며 부족이 2028 년까지 이어질 수 있다고 본다. **2028 년 전환 진영** — 블룸버그 인텔리전스는 부족이 2026 년 2 분기에 정점을 찍고 2027 년 4 분기경 수급 균형에 이른 뒤 2028 년에는 공급 과잉 가능성이 있다고 본다. **중간 진영** — 뱅크오브아메리카는 공급충족률이 2028 년까지 110% 이상을 유지해 역사적 과잉은 없다고 본다(모두 등급: 시장조사·전망치).

세 진영 모두가 공통으로 지목하지 않는 변수가 하나 있는데, 그것이 이 리포트가 주목하는 **중국의 증설**이다. 전 삼성전자 임원 칭구이시안은 이번 가격 폭등이 역설적으로 중국 기업의 공격적 증설을 촉발했으며, 양쯔메모리 우한 3 공장과 창신메모리 상하이 링강 공장이 2027 년 가동에 들어가면 **슈퍼사이클이 2027 년 하반기에 종료될 수 있다**고 경고한다(등급: 보도). 한국 쪽 증설도 같은 시점에 겹친다.

삼성전자·SK 하이닉스의 대규모 다년 투자 계획에서 나오는 신규 생산능력이 2027 년 하반기 이후 집중 가동되면, 지난 사이클을 지탱한 '공급 규율'의 시대가 끝날 수 있다.

가격표를 하나의 문장으로 압축하면 이렇다. **메모리 가격은 올라가고, 연산 단가는 내려간다.** 같은 AI 붐이 메모리에는 구조적 초과수요를, 추론 연산에는 구조적 디플레이션을 만든다. 앞 절에서 본 소프트웨어 최적화(저정밀 포맷·어텐션 압축·프리필/디코드 분리)가 토큰당 원가를 계속 깎아내리는 동안, 그 연산을 담을 메모리는 부족해지기 때문이다. 그리고 중국의 양산 체제는 이 두 곡선을 **모두 가속한다** — 자국산 메모리는 증설로 메모리 가격의 정점을 앞당기고, 자국산 AI 칩은 추론 단가의 하락을 앞당긴다. 한국 기업의 전략은 이 비대칭, 즉 메모리의 단기 초과수익과 연산의 장기 디플레이션을 같은 시간표 위에 놓고 읽는 데서 출발해야 한다.

수출통제·정책 환경

미국의 대중 반도체 통제는 장비·칩·메모리의 세 축으로 쌓여 왔으나, 이 리포트의 관점에서 실질적 압력점은 고대역폭 메모리(HBM)와 첨단 패키징 두 지점에 걸려 있다. 장비 통제는 2022 년 10 월 7 일 미 상무부 산업안보국(BIS)의 첫 패키지로 골격이 세워졌고 2023 년 10 월 개정으로 보강됐으나, 이 통제가 겨누는 것은 '칩을 만들 능력'이었다. 반면 2024 년 12 월 2 일 BIS 가 추가한 HBM 통제는 '만든 칩을 쓸모 있게 하는 부품'을 직접 겨눴다. 해당 조치의 기준은 사양 세대가 아니라 밀도다 — **메모리 대역폭 밀도(memory bandwidth density)가 3.3 GB/s/mm² 미만일 때만 수출·재수출**·이전 예외가 허용된다(BIS 2024-12-02 규칙). 현행 HBM 제품 대부분이 이 문턱을 넘으므로 사실상 전면 통제다. 같은 규칙은 칩 제조 장비 24 종과 소프트웨어 3 종의 대중 공급을 금지하고 중국 기업 140 곳을 규제 명단에 올렸다. 즉 중국 AI 가속기의 메모리 조달 경로가 이 시점에 막혔다. 이 시점 이후 화웨이가 자체 HBM(HiBL 1.0 128GB·1.6TB/s, HiZQ 2.0 144GB·4TB/s, 화웨이 전연결대회 2025, 2025-09-18)을 공개한 것은 시간적 우연이 아니라 통제가 만든 필연적 경로다.

첨단 패키징은 우회로이자 이미 표적이다. HBM 이 통제돼도 2.5D 적층(CoWoS 류) 패키징 역량이 있으면 저사양 메모리를 적층해 부분적으로 우회할 수 있는데, 이 우회로의 병목이 패키징 장비와 공정 노하우다. 앞선 절에서 확인했듯 중국의 첨단 패키징은 2~3 세대 뒤쳐진 단계로 분류되며(15·5 해설, 전자산품세계 2026-03-17), 15·5 목표표에도 첨단 패키징 자급을 수치는 비어 있다. 같은 BIS 규칙(2024-12-02)은 패키징 자체에도 기준을 걸었다 — **최종 패키징 집적회로의 '집합 근사 트랜지스터 수'가 300 억 개 미만이거나 HBM 을 포함하지 않아야 하며, 2027 년까지 이뤄지는 수출은 350 억 개 미만이어야 한다**. 사각지대라기보다 이미 문턱이 설정된 영역이다. 구조 자체는 명확하다 — 연산칩 설계(판매 통제)·제조(장비 통제)·메모리(HBM 통제)·적층(패키징)의 네 관문 가운데 앞의 세 관문이 닫힌 상태에서 마지막 관문이 닫히면 자국산 가속기의 사양 상한이 고정된다. 950DT 의 4TB/s 가 외부 HBM3E 급(스택당 1TB/s 초과, 8~12 단 적층) 대비 어느 위치인지는 이 관문 구조로 읽어야 한다.

장비 통제의 현황은 3 국 분업으로 유지되고 있다. 미국은 자국 장비사(어플라이드·램·KLA)의 대중 판매를 허가제로 묶었고, 네덜란드는 ASML 의 EUV 전면 금지에 이어 2024 년 1 월 침윤식 DUV(NXT:2000i 이상)의 기발급 허가 일부를 취소했으며, 일본은 2023 년 7 월 노광·식각·증착 등 23 개 품목을 수출관리 대상에 올렸다. 이 3 중 구조의 실적은 수치로 드러난다 — 중국 대륙의 7nm 이하 선단 공정 비중 2% 미만, TSMC·삼성과 2~3 세대 격차(15·5 해설, 2026 년 기준)는 통제가 '붕쇄'가 아니라 '세대 고정'으로 작동하고 있음을 보여준다. 반대 방향의 수치도 있다 — 중국 대륙은 6 년 연속 세계 1 위 반도체 장비 투자 규모(같은 출처)를 기록 중이며, 이는 통제 대상이 아닌 성숙 공정 장비와 통제 이전 비축분의 합산이다. 분기별 BIS 개정 공고와 네덜란드·일본의 대응 조치가 추적 대상이다.

중국의 반제 수단은 상류 광물에서 나온다. 갈륨·게르마늄은 2023 년 8 월 1 일부터 수출 허가제가 시행됐고(중국 상무부·해관총서 공고), 안티몬은 2024 년 9 월 15 일부터 같은 방식으로 묶였다(중국 상무부 공고). 세 품목 모두 반도체의 직접 입력물이다 — 갈륨은 GaN 전력·GaAs RF 소자의 기판, 게르마늄은 광통신 소자와 일부 SiGe 공정, 안티몬은 화합물 반도체와 납연 소재에 들어간다. 중국은 갈륨 생산의 절대

다수를 점하고 있어 허가제 자체가 가격·납기 신호로 작동한다. 회토류는 2025 년 4 월 중·중금속 7 개 품목에 허가제가 도입됐고, 2025 년 10 월에는 가공 기술과 해외 생산 파생품까지 포괄하는 확대 조치가 발표됐다는 보도가 있으나 미중 정상회담 이후 유예·조정 보도가 혼재해 현재 시행 상태는 불분명하다. 회토류가 반도체로 직결되는 경로는 두 가지다 — CMP 슬러리·폴리싱 소재(세륨계)와 장비용 고성능 영구자석(네오디뮴·디스프로슘)이다. 즉 중국의 반제 카드는 칩이 아니라 '칩을 만드는 설비와 소재의 원료'에 걸려 있어, 통제전의 비대칭이 장비(서방) 대 원료(중국)의 구도를 만든다.

이 비대칭 구도의 취약점은 양쪽 모두에게 있다. 서방의 장비 통제는 시간이 갈수록 누수가 생긴다 — 자국산 장비 커버리지가 40% 미만(15·5 해설, 2026 년 기준)에서 우한 3 기 52%·창신메모리 45%(산업지 보도, 2026 년 상반기)로 올라서는 속도가 그 증거다. 반대로 중국의 광물 허가제는 수요 이전을 유발한다 — 허가 불확실성이 커질수록 구매처는 비중국산 공급선과 재고를 확보하려 하고, 이는 장기적으로 카드의 위력을 갉아먹는다. 양쪽 다 '지금 당장의 압력'과 '5 년 후의 압력'이 반대로 움직이는 자산이다.

국가대표펀드 3 기는 이 구도의 중국 측 답이다. 규모는 3,000 억 위안을 초과하며 투자 중점은 장비·소재·EDA 로 명시됐다(15·5 해설, 전자산품세계 2026-03-17). 배분 방향 자체가 통제 지도의 반전이다 — 통제가 닫은 세 관문(장비·소재·설계툴)에 정확히 대응한다. 장비 축의 수혜는 나우라(NAURA)와 AMEC 등 증착·식각 전문사로 수렴하고, EDA 축에서는 엠피리언(Empyrean) 등이 자국산 툴체인 공백을 메우는 구조다. 다만 3 기 자금의 기업별 납입액과 집행 내역은 비공개이며, 1·2 기 대비 배분 비중 변화도 공개 자료가 없어 추정하지 않는다.

주목할 변화는 국가대표펀드의 독주가 끝났다는 점이다. 앞선 절에서 확인했듯 창신메모리는 상장으로 3.28 조 위안 시총을 확보했고(2026-07-29, 과창판), 양쯔메모리는 330 억 위안 조달을 목표로 수리 단계에 있으며(2026-08-29), 엔플레임은 텐센트를 최대주주로 둔 채 60 억 위안을 조달했다(2026-08-29). 국가대표펀드(대기금) 3 기가 통제가 닫은 관문(장비·소재·EDA)에 들어가는 동안, 자본시장이 통제 바깥의 증설(메모리·가속칩 양산) 자금을 낸다. 공공 자금은 병목 해소, 민간 자금은 물량 확대라는 이원 구조가 형성된 것이다. 통제가 중국의 자금 조달을 막지 못하고 오히려 자본시장 루트를 활성화시켰다는 점은 통제 설계의 한계를 보여준다.

정책 환경을 하나의 문장으로 묶으면 이렇다 — 통제는 중국의 사양 상한을, 반제는 서방의 원료 불확실성을, 국가대표펀드(대기금) 3 기는 그 상한을 뚫는 시간표를 정한다. 이 리포트의 관통 논지와 연결하면, 소프트웨어가 하드웨어를 정의하는 재편은 통제 환경에서 중국에 오히려 유리한 방향이다. 범용 GPU 경쟁이라면 통제로 고정된 선단 공정 격차(2~3 세대)가 그대로 성능 격차가 되지만, 워크로드별 특화(950PR/950DT 분기, 저정밀 포맷, 회소 MoE)는 연산 밀도가 아닌 아키텍처 적합도로 승부하기 때문이다. 추적 지표는 세 가지다 — HBM·패키징 통제의 개정 여부, 회토류 확대 조치의 시행 상태, 국가대표펀드(대기금) 3 기의 집행 실적. 셋 모두 비공개 영역에 있으므로, 공개 가능한 대리 지표는 자국산 장비 커버리지(15·5 목표 2030 년 60% 대 현재 40% 미만)의 분기 변화와 나우라·AMEC 의 수주 공시다.

한국 시사점·액션 플랜

먼저 **창의 형태를 정확히 규정해야 한다. 이 창은 대중 HBM 직수출로 열리지 않는다.** 미 상무부 산업안보국(BIS)의 2024 년 12 월 2 일 규칙은 메모리 대역폭 밀도가 3.3 GB/s/mm² 미만일 때만 수출·재수출·이전 예외를 허용하는데, 현행 HBM 제품 대부분이 이 문턱을 넘는다. 따라서 창이 열린다면 형태는 셋 중 하나다 — ① 중국이 자국산 전환으로 글로벌 HBM 수요에서 빠지면서 비중국 고객 물량이 한국으로 재배분되는 **간접 효과**, ② 통제 기준 미만 사양의 **범용 서버 D 램·기업용 SSD**, ③ 중국 외 지역에 법인을 둔 중국계 고객사다. 이 구분을 흐리면 아래 과제 전체가 틀린 전제 위에 놓인다. 그 위에서, 한국 기업이 지금 해야 할 일은 가격표를 다시 쓰는 것이 아니라 계약의 형태를 바꾸는 것이다. 어센드 950DT 가 2026 년 4 분기에 출시되고(화웨이 전연결대회 2025, 2025-09-18) 화웨이 자체 HBM 인 HiZQ 2.0 의 공급량이 비공개인 동안, 중국 모델 기업의 고성능 메모리 수요는 외부로 열려 있다. 그러나 이 창은 창신메모리의 상장 완료(2026-07-29, 시총 3.28 조 위안)와 양쯔메모리의 IPO 수리(2026-08-29, 조달 목표 330 억 위안)로 확보된 자금이 우한 3 기 증설(월 10 만 장 추가, 2026 년 하반기 양산 목표, 보도)로 전환되는 시점부터 닫히기 시작한다. 1~2 년짜리 창에서 현물 단가를 놓고 싸우는 것은 창이 닫히는 순간 아무것도 남기지 못하는 전략이다. 남겨야 할 것은 사양 협업 관계, 복수년 공급 계약, 메모리-패키징 결합 구조다.

현물 가격 경쟁이 유효하지 않은 이유는 가격 방향 자체가 이미 판매자 우위로 기울어 있기 때문이다. 시장 전망치 기준 2026 년 2 분기 D 램 계약가는 전분기 대비 +58~63%, NAND 는 +70~75% 상승이 예상되고(2025 년 12 월 보도, 전망치), 공급 부족은 2027 년까지 이어진다는 전망이다. 이 국면에서 단기 물량을 비싸게 파는 것은 어렵지 않다. 어려운 것은 딥시크-화웨이처럼 모델 기업이 하드웨어 사양을 규정하는 구조에서 한국 제품이 사양 표의 바깥에 놓이지 않는 것이다. 950PR(HiBL 1.0, 128GB·1.6TB/s)과 950DT(HiZQ 2.0, 144GB·4TB/s)로 갈라진 제품 구조가 보여주듯, 구매자는 이제 'HBM 최신 세대'가 아니라 '프리필용·디코드용 메모리'를 주문한다. 협상 단위가 세대에서 워크로드로 바뀐 것이다.

주체	0~3 개월	3~6 개월	6~12 개월	리스크
삼성전자	950DT 출시(2026 년 4 분기 예정) 전에 중국 모델 기업·클라우드사의 프리필/디코드 분기 워크로드별 메모리 요구 사양을 확보한다. 협상 상대를 칩 조달 부서에서 모델·인프라 설계 조직으로 바꾼다	HBM3E/4 와 첨단 패키징을 묶은 결합 견적을 제시한다. 메모리 단품이 아니라 스택·적층·전력 예산을 함께 정의하는 사양 협업으로 진입한다	950DT 인도 실적과 HiZQ 2.0 공급량이 확인되는 시점에 복수년 공급 계약 전환 여부를 결정한다. HiZQ 공급 부족이 확인되면 장기 계약, 원활하면 단기 고수익 물량 중심으로	대중 매출 의존이 심화될수록 통제 개정 한 번에 계약 전체가 흔들린다. 미 BIS 의 HBM2E 이상 통제(2024-12)가 확대되면 진행 중 계약의 중단

			회수한다	비용이 현물 이익을 초과할 수 있다
SK 하이닉스	통제 기준을 넘지 않는 서버 D 램·기업용 SSD 물량을 현물이 아니라 분기 단위 가동률 데이터와 연동해 배정한다. HBM 은 대중 직판이 아니라 비중국 고객 재배분 물량으로 계획한다. 중국 내 자국산 메모리 가동률·수율 공개를 계약 물량 조정의 트리거로 명문화한다	디코드 워크로드의 대역폭 집약 특성(950DT 기준 4TB/s 급 요구, 2025-09-18 발표)에 맞춘 맞춤 사양 대응 체계를 만든다. 범용 최신 세대 공급 논리에서 벗어난다	자국산 대체 시점을 양산 발표가 아니라 실측 출하로 판단한다. 창신메모리 월 생산능력·우한 3 기 월 10 만 장 증설분의 출하 전환(보도 기준 2026 년 하반기 목표)이 확인되기 전까지는 창을 열려 있는 것으로 본다	자국산 대체 시점 오판이 최대 리스크다. HiZQ 2.0 수율이 비공개이므로, 창이 닫히는 시점을 앞당겨 가정하면 물량을 남기고, 늦게 가정하면 재고를 안는다
소부장(패키징·소재·장비)	우한 3 기 자국산 장비 비중 52%·창신메모리 45%(보도)를 기준선으로 삼아, 자국산 장비 비중이 정체된 공정 단계를 자사 수요가 남은 구간으로 특정한다	첨단 패키징 장비·소재에서 메모리 제조사와 공동 검증을 시작한다. HBM(2024-12 BIS 통제)과 함께 통제의 두 번째 관문이 된 패키징이 중국 자국산화에서 가장 늦게 닫히는 창이다	자국산 장비 커버리지가 60% 목표(2030 년, 15·5 해설 2026-03-17)를 향해 움직이는 속도를 분기별로 추적해 철수 또는 버티기 결정의 근거로 쓴다	12 인치 라인 자국산 장비 커버리지가 현재 40% 미만(15·5 해설, 2026 년 기준)에서 목표 60%로 수렴하면, 남은 20%p 구간에서 한국 장비의 대중 수출 여지는 구조적으로 줄어든다
정책 입안자	9 절의 선행 지표(가동률·수율 공개·계약 단가 보도·증설 착공·자국산 장비 비중)를 산업부·무역협회 등 기존 동향 조사	통제 개정 시나리오별 계약 중단 비용을 기업과 함께 점검한다. 갈룸·게르마늄(2023-08)·안티몬(2024-09)·희토류(2025-04·10) 반제 조치가	15·5 의 자급률 70%(2030 년 목표)가 2025 년 미달 목표의 5 년 이월임을 전제하고, 계획 문구가 아니라 950DT	정책이 어느 한쪽 시나리오에 고정되면 다른 쪽의 비용을 기업이 전가받는다. 통제 완화에

	체계에 정식 항목으로 편입한다	소부장 원료 조달에 미치는 경로를 항목별로 조사한다	인도·HiZQ 공급량·우한 3 기·양쯔메모리 상장 완료의 분기 실적을 대중 정책의 검증 지표로 삼는다	베팅하면 확대 시 수출 중단 손실, 통제 강화에 베팅하면 완화 시 시장 상실이 발생한다
--	---------------------	------------------------------------	---	--

삼성전자와 SK 하이닉스에 공통으로 적용되는 원칙은 협상 상대의 재설정이다. 딥시크가 어센드 슈퍼노드를 화웨이와 공동 정의하고(블룸버그 2026-09-04 보도) 장문 추론의 지연·처리량·비용을 사양으로 되먹임한 사례는, 구매 결정권이 칩 회사에서 모델 회사로 이동했음을 보여준다. 한국 메모리 기업이 중국 칩 회사의 조달 부서와만 이야기하면 사양은 이미 다른 곳에서 정해진 뒤다. 모델 기업의 인프라 설계 조직이 KV 캐시 용량과 디코드 대역폭을 계산하는 단계에 들어가야, 그 다음 세대 칩의 메모리 사양표에 한국 제품이 기본값으로 적힌다.

소부장의 창은 메모리보다 좁고 더 빨리 닫힌다. 자국산 장비 비중이 우한 3 기 52%, 창신메모리 45%(보도)까지 올라온 상태에서, 성숙 공정 범용 장비는 이미 나우라(NAURA)·AMEC 의 영역으로 넘어가고 있다. 남은 진입 공간은 두 곳이다. 하나는 통제의 두 번째 관문인 첨단 패키징으로, HBM 적층과 상호연결이 워크로드별 분기(950PR/950DT)의 물리적 실현 수단이 되면서 패키징 검증 수요는 자국산 장비로 메우기 어려운 구간에 남아 있다. 다른 하나는 소재다. 국가대표펀드(대기금) 3 기(3,000 억 위안 초과, 장비·소재·EDA 중점 — 15·5 해설 2026-03-17)가 소재에 배분되고 있다는 사실은 역으로 현재 소재 자급이 목표에 미달한다는 뜻이며, 포토레지스트 자급률 50% 목표(2030 년)가 설정됐다는 것 자체가 현재 미달의 확인이다.

모니터링 체계 편입은 이번 절의 과제 가운데 유일하게 네 주체 모두에게 해당한다. 창이 닫히는 신호는 정부 발표나 목표 문구에서 오지 않는다. 《중국제조 2025》의 '2025 년 자급률 70%'가 실적 19.4%(IC 인사이츠 2021 년 전망)로 미달하고 같은 목표가 2030 년으로 이월된 전례가 증명하듯, 문구는 지표가 아니다. 편입해야 할 항목은 9 절의 다섯 가지다 — 중국 내 메모리 가동률, 수출 공개 여부, 계약 단가 보도, 우한 3 기 등 증설 착공·양산 실적, 신규 라인의 자국산 장비 비중. 여기에 950DT 인도 실적, HiZQ 2.0 공급량(현재 비공개), 양쯔메모리 상장 완료(현재 수리 단계, 2026-08-29 기준), 딥시크의 프로 가격 인하 실행 여부를 더한다. 프로 가격 인하는 딥시크가 스스로 밝힌 슈퍼노드 양산의 하방 신호이므로(블룸버그 2026-09-04), 가격이 실제로 내리면 자국산 공급이 돌고 있다는 뜻이고, 내리지 못하면 외부 메모리 수요 창이 연장된다는 뜻이다.

리스크는 표의 오른쪽 열에 적은 대로 세 갈래다. 첫째, 대중 매출 의존 심화다. 창이 열린 동안 매출이 늘수록 통제 개정 한 번의 충격이 커진다. 2024 년 12 월 BIS 규칙(대역폭 밀도 3.3 GB/s/mm² 미만만 예외) 이후 확대되면, 진행 중인 복수년 계약의 중단 비용은 현물 이익으로 상쇄되지 않는다. 둘째, 계약 중단 그 자체다. 중국의 반제 수단이 칩이 아니라 갈륨·게르마늄·안티몬·희토류라는 원료에 걸려 있어(2023-08·2024-09·2025-04·10), 한국 소부장의 대중 수출과 중국에서의 원료 조달이 같은 지정학 변수에 묶여 있다. 셋째,

자국산 대체 시점 오판이다. HiZQ 2.0의 수율과 공급량이 비공개인 조건에서 창 의 종료 시점을 앞당겨 가정하면 열려 있는 물량을 경쟁사에 넘기고, 늦게 가정하면 닫힌 창에 재고를 쌓는다. 세 리스크 모두 방향이 아니라 시점의 문제이며, 시점은 예측이 아니라 위의 선행 지표 체계로만 관리할 수 있다.

이 모든 과제를 하나의 개념으로 묶으면 **시간차 전략**이다. 지금 중국의 AI 야심이 요구하는 고급 메모리를 중국 스스로는 아직 공급하지 못한다. 창신메모리의 12 층 HBM3E 가 2027 년에 양산돼도 기술 격차는 2~3 년이 남고, 화웨이 자체 HBM 은 초기 수율 35~40%의 양산 검증 단계다. 이 창은 대략 2 년이다. 그 안에서 한국이 해야 할 일은 네 가지로 좁혀진다. **첫째, 세대 선도의 유지** — 경쟁의 무기를 '양'이 아니라 HBM4·HBM4E 와 16 층·하이브리드 본딩이라는 '세대 차이'로 옮긴다. **둘째, 차세대 장벽의 조기 확보** — 자국산 진영이 12 층에 도달하는 시점에 한국은 이미 다음 공정에 가 있어야 한다. **셋째, 글로벌 고객과의 장기계약 잠금** — 대중 직수출이 막힌 조건에서 창 의 실체는 비중국 물량의 재배분이므로, 그 물량을 복수년 계약으로 고정하는 것이 창을 현금으로 바꾸는 유일한 방법이다. **넷째, 역설적이지만 중국향 범용 D 램·장비·소재의 현금 흐름 유지** — 통제 기준 미만 구간은 여전히 열려 있고, 그 매출이 세대 선도에 필요한 투자 재원이 된다.

네 과제의 공통 전제는 하나다. **창의 길이를 예측하지 말고 관측하라는 것이다.** 창을 닫는 사건은 정부 발표가 아니라 수율 공개·증설 착공·계약 단가 보도·자국산 장비 비중 같은 실적 지표에서 먼저 나타난다.

부록 — 주요 지표와 출처

이번 호 본문에 사용한 주요 수치를 한 표로 묶었다. 등급이 '보도' 또는 '시장조사'인 수치는 실측과 구분해 읽어야 한다.

1. 주요 지표 일람

지표	값	기준시점	출처
딥시크 어센드 950DT 배치 계획	최소 16 만 장(기가와트급 데이터센터의 일부 용량)	2026-09-04	블룸버그, 관찰자망 전채(등급: 보도)
어센드 950PR 출시 시점	2026 년 1 분기	2025-09-18	화웨이 전연결대회 2025, 쉬즈쥔 발표
어센드 950DT 출시 시점	2026 년 4 분기(예정)	2025-09-18	화웨이 전연결대회 2025
HiBL 1.0(950PR 탑재)	128GB·1.6TB/s	2025-09-18	화웨이 전연결대회 2025
HiZQ 2.0(950DT 탑재)	144GB·4TB/s	2025-09-18	화웨이 전연결대회 2025
HiBL/HiZQ 양산 능력·수율·공급량	비공개	—	미공개
950 아키텍처 저정밀 연산	FP8/MXFP8/HIF8 1 PFLOPS, MXFP4 2 PFLOPS	2025-09-18	화웨이 전연결대회 2025
상호연결 대역폭	전 세대 대비 2.5 배	2025-09-18	화웨이 전연결대회 2025
아틀라스 950 슈퍼노드	최대 8,192 장	2025-09-18	화웨이 전연결대회 2025
아틀라스 960 슈퍼노드	최대 15,488 장	2025-09-18	화웨이 전연결대회 2025
창신메모리 과창판 상장	2026-07-29, 첫날 +465.82%, 시총 3.28 조 위안	2026-07-29	CHAI 주간 invest-stock 2026-07-29 호
창신메모리 MSCI 편입	8 월 10 일	2026 년 8 월	CHAI 주간 invest-stock 2026-07-29 호
양쯔메모리 IPO	수리 단계(상장 완료 아님), 조달 목표 330 억 위안	2026-08-29	CHAI 주간 invest 2026-08-29 호
양쯔메모리 NAND 출하	세계 3 위, 점유율 14%(키옥시아 추월)	2026 년 8 월	CHAI 주간 tech 2026-08-16 호
양쯔메모리 월 생산능력	약 20 만 장, 우한 3 기(월 10 만 장 추가) 양산 2027 년→2026 년 하반기로 앞당김, 총 30 만 장 목표	2026 년 상반기	산업지 보도(등급: 보도)
양쯔메모리 1 분기 매출·순이익	매출 200 억 위안 초과, 순이익 333.79 억 위안	2026 년 1 분기	CHAI 주간 invest 2026-08-29 호·산업지 보도
자국산 장비 비중	창장 우한 3 기 52%, 창신메모리 45%	2026 년	산업지 보도(등급: 보도)
엔플레임 IPO 조달	60 억 위안, 텐센트 최대주주·최대고객	2026 년 8~9 월	CHAI 주간 invest 2026-08-29 호
D 램 연간 상승률 전망	+250~280%	2025-12 보도	시장조사(전망치, 실측 아님)
NAND 연간 상승률 전망	+200~250%	2025-12 보도	시장조사(전망치)

2026 년 2 분기 D 램 계약가	전분기 대비 +58~63%(전망)	2025-12 보도	시장조사(전망치)
2026 년 2 분기 NAND 계약가	전분기 대비 +70~75%(전망)	2025-12 보도	시장조사(전망치)
공급 부족 지속 전망	2027 년까지	2025-12 보도	시장조사(전망치)
글로벌 AI 인프라 지출	4,500 억 달러, 추론 비중 처음으로 70% 초과	2026 년	15·5 해설 자료 인용
HBM 칩 가치 비중	약 절반	—	미 CSIS 그레고리 앨런
AI 스마트폰 출하	4.7 억 대 돌파 전망, 침투율 38%	2026 년	업계 보도(전망치)
AI 웨어러블 시장	419 억 달러(2024 년)→1,207 억 달러(2028 년 전망)	2024~2028 년	업계 보도
로크칩 온디바이스 매출	전년 동기 대비 +40% 이상	2026 년 상반기	업계 보도
로크칩 AIoT 컴퓨팅	0.2~6 TOPS, RK182X 양산 단계	2026 년	업계 보도
블랙세미 화산 A2000	7nm, 200~2,000 TOPS	2026 년	업계 보도
호라이즌 로보틱스 점유율	자국 브랜드 ADAS 47.7%, 27 개 완성차 그룹(42 개 브랜드) 채택	2026 년	업계 보도
국가대표펀드(대기금) 3 기	3,000 억 위안 초과, 장비·소재·EDA 중점	2026 년	15·5 해설(전자산품세계, 2026-03-17)

2. 15 차 5 개년 계획 — 2030 년 목표 대 현재 수준

이번 호의 핵심 대비표다. 목표 문구와 현재 수치 사이의 공백이 큰 항목일수록 검증 지표(분기 실적)로 추적해야 한다.

항목	2030 년 목표	현재 수준	격차 판정
반도체 산업 규모	3 조 위안 돌파	미공개	—
전체 자급률	70%	참고: 2025 년 실적 19.4%(IC 인사이츠 2021 년 전망)	최대 병목
22nm 이하 선단 공정 자국산화	50%	7nm 이하 비중 2% 미만, TSMC·삼성과 2~3 세대 격차	병목
성숙 공정 세계 비중	52%	12 인치 생산능력 세계 25%, 그중 77%가 28nm 이상	근접권
12 인치 자국산 장비 커버리지	60%	40% 미만	병목
포토리티스트(ArF/EUV) 자급률	50%	미공개	—
12 인치 실리콘 웨이퍼	100% 자국산	미공개	—
AI 칩	2028 년 전 200TOPS+ 양산	어센드 950 계열 로드맵 공표	양산 전

메모리	HBM3/4·3D NAND 규모화 생산	메모리 자급률 5% 미만(2020 년)→35%(2026 년)	진행 중
차량용 칩 자급률	—(목표치 별도 표기 없음)	20% 미만	병목

목표는 《제 15 차 5 개년 계획 강요(2026~2030)》(2026 년 3 월 정식 발표)와 해설(전자산품세계, 신랑재경 전재, 2026-03-17) 기준이고, 현재 수준도 같은 출처다. 과거 《중국제조 2025》로드맵의 '2025 년 자급률 70%'는 실적 19.4%로 미달했고, 같은 70%가 2030 년으로 5 년 이월됐다는 점에서 이 표의 '목표' 열은 계획 문구로 읽어야 한다.

3. 원문 출처 목록

1. 《중화인민공화국 국민경제·사회발전 제 15 차 5 개년 계획 강요(2026~2030)》, 2026 년 3 월 정식 발표. 해설: 전자산품세계(电子产品世界), 신랑재경 전재, 2026-03-17.
2. 블룸버그, 딥시크 네이밍구 데이터센터·어센드 950DT 16 만 장 보도, 2026-09-04. 관찰자망 전재, CHAI 주간 tech 2026-09-05 호.
3. 화웨이 전연결대회 2025, 순환회장 쉬즈쥔(徐直军) 기조연설, 2025-09-18. 콰이커지(快科技)·신랑재경 보도.
4. CHAI 주간 invest·stock 2026-07-29 호(창신메모리 상장), CHAI 주간 invest 2026-08-29 호(양쯔메모리 수리·엔플레임 청약), CHAI 주간 tech 2026-08-16 호(양쯔메모리 NAND 세계 3 위).
5. 관찰자망, 자급률 70% 목표 관련 논의, 2020-08-20.
6. IC 인사이드, 2025 년 중국 내 생산 칩 점유율 19.4% 전망(2021 년 발표).
7. 메모리 가격 전망 보도(2025 년 12 월) — 전망치이므로 실측 계약가와 구분.
8. 미 CSIS 그레고리 앨런, HBM 가치 비중 관련 발언(인용).
9. 통제 관련: 미 BIS HBM2E 이상 통제(2024-12), 중국 갈룽·게르마늄 허가제(2023-08), 안티몬(2024-09), 희토류 확대(2025-04·10).

4. 이번 호 신규 용어 대조표

용어	원어·영문	의미
어센드 950PR·950DT	Ascend 950PR/950DT	화웨이 가속기 — PR 은 프리필용(HiBL 1.0), DT 는 디코드용(HiZQ 2.0)
HiBL 1.0 / HiZQ 2.0	바이루 1.0 / 주첸 2.0	화웨이 자체개발 HBM 2 종
아틀라스 950 슈퍼노드	Atlas 950 SuperPoD	최대 8,192 장 묶는 대형 노드
링취	灵衢(LingQu)	화웨이 자체 칩 간 상호연결 기술
프리필/디코드	Prefill/Decode	추론의 두 단계 — 연산 집약 대 대역폭 집약

희소 MoE	Sparse Mixture-of-Experts	전문가 일부만 활성화하는 모델 구조, 율투율 통신 필수
초상규	超常规	'관례를 넘는' 초강력 정책 조치(15·5 문구)
자국산 대체	国产替代	외국산을 중국산으로 대체 — '국산'은 한국산으로 읽히므로 이 표기 사용
성숙 공정/선단 공정	成熟制程/先进制程	28nm 이상 대 22nm 이하 구분
첨단 패키징	先进封装	HBM 적층 등 후공정 기술
과창판	科创板(STAR Market)	상하이거래소 기술주 시장
국가대표펀드	国家集成电路产业投资基金	국가 집적회로 산업투자펀드, 현재 3 기
피지컬 AI	具身智能	로봇 등 물리 실체에 탑재되는 AI