

AI Infra Signal 5호

에이전트 시대 개막: Muse 사용자가 1억명이면 생기는 일

Industry Signal

#1 AI 인프라의 선계약 시대

엔트로픽은 향후 약 10년간 5,180억달러의 컴퓨트/인프라를 약정했고, 이 중 80%는 취소가 제한. 특히 취소 제한 약정의 92.4%가 TPU 등 ASIC에 집중. 미래 AI 수요를 확인한 뒤 투자하는 것이 아니라, 연산 자원을 먼저 확보하고 공급망 투자를 앞당기는 구조로 전환 중.

이는 ASIC, 서버, 데이터센터, 네트워크 공급망의 주문 가시성을 높이는 대신 수요 위험을 AI 기업에 이전. 단기 수요 둔화는 장비 주문 취소보다 고객의 수익성과 현금흐름 부담으로 먼저 나타날 가능성. 향후에는 계약 총액보다 실제 가동률과 약정 소진 속도가 핵심 확인 지표.

#2 에이전트가 늘리는 일반 서버 수요

Muse와 Dots는 사용자가 자리를 비운 뒤에도 클라우드 컴퓨터를 계속 사용. 이에 AI 인프라 수요가 GPU에서 CPU, DRAM, 스토리지까지 확산되고 있으며, 서버 CPU와 DRAM 리드타임은 정상 대비 각각 1.5배, 2.5배까지 늘어난 상황.

당사 추정 기준 메타 Muse와 오픈AI Dots 사용자가 출시 1년 뒤 1.1억명 도달 시, 약 12.9만대의 서버 용량이 필요. 이는 연간 글로벌 서버 증가분의 6.5%에 해당. 핵심은 가입자 수보다 에이전트가 실제로 일하는 시간이 서버 수요를 좌우한다는 점. 자원 공유가 늘수록 DRAM 비중도 66%→79%로 상승 전망.

Company Signal

마이크론 (MU US)	에이전트 확산과 공급 부족으로 일반 서버 DRAM 수혜 부각. HBM 밖 데이터센터 메모리까지 수익성 개선 확산
크레도 테크놀로지 (CRDO US)	ASIC 클러스터 확대와 스케일업 연결 증가 수혜. AEC 중심 고속 링크 확대로 가속기당 연결 콘텐츠 증가
아리스타 네트워크 (ANET US)	ASIC 및 AMD 기반 AI 클러스터 확대로 이더넷 수요 증가. 1.6T와 스케일업 이더넷이 2027년 추가 성장축
관심종목	연결반도체 크레도, 브로드컴, 마벨 네트워크 아리스타, 시스코 서버 델, HPE

자료: 미래에셋증권 리서치센터

Key Takeaways

이번 주 핵심은 AI 수요가 GPU 중심에서 일반 서버 부품까지 넓어지는 동시에, 공급 부족이 기업별 이익 차별화로 이어지기 시작했다는 점.

장기 선계약으로 공급망의 주문 가시성은 높아지나, 에이전트 확산은 CPU, DRAM, 스토리지까지 수요를 넓히는 중. 이미 서버 CPU, DRAM, 기업용 SSD 리드타임은 정상 대비 약 2배까지 늘어나며 부품 조달 부담도 확대되는 상황.

이에 같은 AI 수요 증가라도 이익 레버리지는 밸류체인별로 차별화될 전망. 공급 부족을 가격으로 전환할 수 있는 메모리, ASIC 확대로 연결 콘텐츠가 늘어나는 고속 인터커넥트가 상대적으로 유리. 서버와 네트워크 장비는 부품 확보와 가격 전가력이 실적을 가르는 핵심 변수로 판단.

[Industry Signal #1]

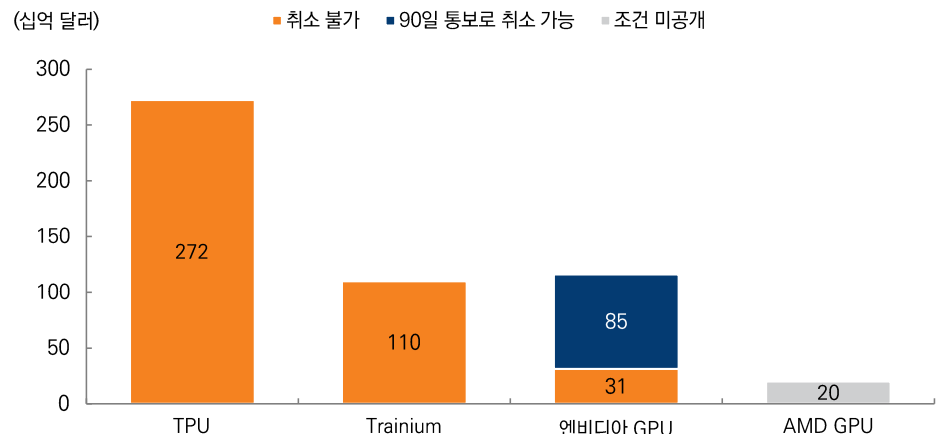
AI 인프라의 선계약 시대

취소 제한 약정의 92%가 ASIC에 집중

지난 9월 29일 로이터는 엔트로픽의 비공개 IPO 문서를 인용해, 회사가 향후 약 10년간 6개 파트너와 5,180억달러 규모의 AI 인프라 구축 및 이용 약정을 체결했다고 보도했다. 이 가운데 약 80%인 4,137억달러는 취소가 어렵거나 실제 사용량과 관계없이 지급해야 한다. 엔트로픽은 향후 AI 성장의 제약이 수요보다 컴퓨팅 자원의 가용성에서 나타날 가능성이 높다고 밝혔다. 수요가 생긴 뒤 설비를 늘리는 것이 아니라 미래 연산 자원을 먼저 계약으로 잠그는 구조다.

약정이 어디에 집중됐는지가 더 중요하다. 취소가 제한된 4,137억달러 중 3,823억달러, **92.4%**가 구글 TPU와 아마존 Trainium 등 ASIC 관련 계약이다. 반면 엔비디아 GPU 약정 1,159억달러 중 72.9%는 상당 부분을 90일 전 통보로 취소할 수 있는 xAI 계약이다. 표 1에서 보이듯 브로드컴, 구글, 아마존, 마이크로소프트 계약은 취소가 제한되는 반면 xAI 계약은 상대적으로 유연하다.

그림 1. 칩 종류별 엔트로픽 컴퓨트 약정액



주: 구글 TPU는 브로드컴 1,612억달러, 구글 1,111억달러로 구성. 엔비디아 GPU는 마이크로소프트 314억달러와 xAI 최대 845억달러
자료: Reuters, 각 사, 미래에셋증권 리서치센터

표 1. 엔트로픽 컴퓨트 약정: 계약 상대별 조건

계약 상대	약정액(십억 달러)	비중(%)	계약 조건	칩, 용량	취소 구분
브로드컴	161.2	31.1	장비 리스, 대부분 취소 불가	TPU, 2027년부터 약 3.5GW	취소 불가
구글	111.1	21.4	약정만큼 못 쓰면 차액 지급	구글 클라우드 TPU	취소 불가
아마존	110.0	21.2	약정만큼 못 쓰면 차액 지급	Trainium2~4, 최대 5GW	취소 불가
마이크로소프트	31.4	6.1	일부 예외 외 취소 불가	Azure 엔비디아 GPU, 최대 1GW	취소 불가
xAI	84.5	16.3	최대 금액, 상당 부분 90일 전 통보로 취소 가능	엔비디아 GPU, 2029년까지	취소 가능
AMD	20.0	3.9	조건 미공개, AMD가 엔트로픽 지분 최대 50억 달러 매입	AMD GPU	미공개
합계	518.2	100.0			
취소 불가 4건	413.7	79.8			

주: 향후 10년 약정. 비중은 6건 합계 대비. AMD는 200억 달러 이상의 하한값 가입

자료: Reuters, 각 사, 미래에셋증권 리서치센터

ASIC은 왜 장기 계약이 필요한가

이 차이는 ASIC과 GPU의 공급 구조에서 나온다. TPU와 Trainium은 특정 클라우드와 고객 수요를 전제로 칩과 서버, 데이터센터 용량을 함께 준비한다. 한 고객을 위해 깔아둔 생산능력과 인프라를 다른 고객에게 즉시 돌리기 어렵기 때문에 공급자는 선투자에 앞서 장기 최소사용 약정을 요구할 유인이 크다.

반면 범용 GPU는 고객과 용도를 바꿔 재배치하기가 상대적으로 쉽다. 물론 마이크로소프트 GPU 계약처럼 장기 약정도 있어 ASIC=장기, GPU=단기로 단순화할 수는 없다. 다만 이번 엔트로픽 계약에서는 장기 수요 위험이 ASIC 쪽 고객에게 훨씬 강하게 이전돼 있다.

미래 수요가 공급망의 현재 주문으로 앞당겨진다

투자 관점에서는 미래 AI 수요가 현재의 공급망 투자로 당겨진다는 점이 중요하다. 대형 고객이 7~10년의 최소 사용량을 확정하면 공급자는 분기별 사용량을 기다리지 않고 ASIC 생산, 서버 조립, 데이터센터 증설과 연결 인프라 발주를 계획할 수 있다.

즉, AI 인프라의 주문 방식이 '수요 발생→장비 발주'에서 '미래 수요 예상→연산 자원 선계약→공급망 선투자'로 바뀌고 있다. 단순히 엔트로픽이 AI CapEx를 많이 쓴다는 뉴스와는 성격이 다른 이유다.

다만 장기 약정이 곧 확정 매출은 아니다. 브로드컴도 2027년부터 공급할 약 3.5GW의 실제 이용 규모가 엔트로픽의 사업 성과에 달렸다고 공시했다. 장기 계약은 공급망의 단기 수요 변동을 줄이는 대신 AI 기업의 매출 성장과 지급 능력에 대한 노출을 높인다. 따라서 계약 총액보다 실제 가동과 약정 소진 속도를 함께 확인할 필요가 있다.

리스크의 위치도 달라진다. 장기 take-or-pay와 하드웨어 선결제처럼 사용량과 관계없이 구매자가 부담을 먼저 지는 계약이 늘면서, 단기 AI 수요 둔화의 충격은 공급망 주문 취소보다 AI 기업의 수익성과 현금흐름 부담으로 먼저 나타날 가능성이 높다. 공급자는 단기 수요 변동에 덜 노출되는 대신 고객의 장기 지급능력에 더 노출되며, 이는 공급 부족 국면에서 공급자의 협상력이 가격을 넘어 계약 조건으로 확대되고 있음을 보여준다.

표 2. AI 인프라 계약 구조 비교

구분	계약 방식	기간	구매자가 먼저 진 부담	수요 위험을 지는 쪽
엔트로픽 (연산 자원 구매)	6개 파트너와 인프라 구축과 이용 약정 5,180억 달러	약 10년	취소 불가 또는 사용량과 무관한 지급 의무 4,137억 달러	구매자(엔트로픽)
마이크론 (메모리 공급)	16개 전략 고객 계약. 약정 물량을 안 사도 대금을 내는 구조(take-or-pay), 가격 상한과 하한 설정. 계약 잔고 약 1,000억 달러	5년(2026~30년)	현금 예치금 약 180억 달러 포함 재무 약정 220억 달러	구매자(고객). 구매 미달 시 예치금 차감. 마이크론은 가격 상단 일부 포기
오라클 (연산 자원 판매)	대형 AI 계약의 GPU 등 하드웨어를 고객이 선결제하거나 직접 공급. 계약 잔고 6,380억 달러	데이터센터 임차 15~19년	하드웨어 선결제와 고객 공급분 750억 달러	하드웨어는 고객. 데이터센터 임차 약정 2,600억 달러는 오라클
코어워브 (연산 자원 판매)	신규 용량을 단기 계약으로 판매, MW당 연 약 4,000만 달러. 6월 말 계약 잔고 1,040억 달러	약 3~6개월 (9월 신규 계약)	계약 기간만큼의 지급 의무	판매자(코어워브). 재계약과 가동을 위험

주: 엔트로픽은 2026년 9월 로이터 보도(비공개 IPO 문서), 마이크론은 FY26 3분기 실적 발표(2026년 6월), 오라클은 FY26 4분기 실적 발표와 연차보고서(2026년 5월 말), 코어워브는 2026년 9월 17일 발표와 2분기 실적 기준. 마이크론 계약 잔고는 16건 중 14건 기준, 코어워브 가격은 연환산 매출을 필요 전력으로 나눈 값

자료: Reuters, 각 사, 미래에셋증권 리서치센터

[Industry Signal #2]

에이전트가 늘리는 일반 서버 수요

메타에 이어 오픈AI까지... 에이전트 전성시대

메타의 Muse에 이어 오픈AI도 9월 29일 Dots를 공개했다. 두 서비스는 클라우드의 가상 컴퓨터에서 사용자가 앱을 닫은 뒤에도 맡은 일을 이어간다. 기존 챗봇이 모델 추론을 위한 GPU 중심이었다면, 에이전트는 브라우저와 프로그램을 실행하기 위해 CPU, 메모리, 스토리지까지 함께 사용한다. 에이전트 환경에서 서버의 CPU 대 GPU 비율이 기존 1대 4~8에서 1대 1~2까지 높아질 것으로 전망되는 상황이다.

초기 사용자 반응도 빠르다. Muse의 출시 21일 차 DAU는 157.5만 명으로 같은 시점 ChatGPT 앱의 91.8만 명을 넘어섰고, 설치자 가운데 매일 사용하는 비율도 31.1%로 ChatGPT 앱의 12.2%보다 높다.

그림 2. 출시 후 일수별 DAU: Muse vs ChatGPT 앱

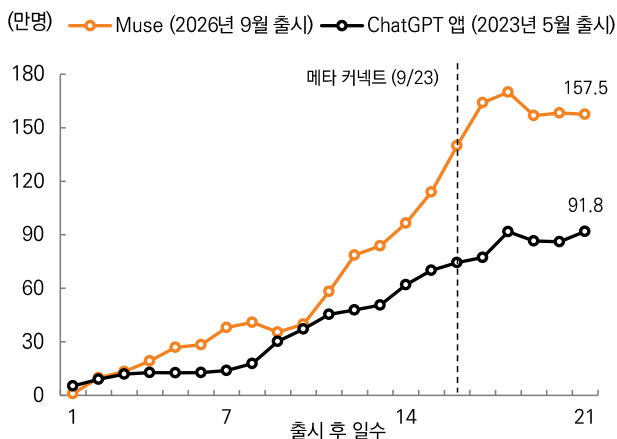
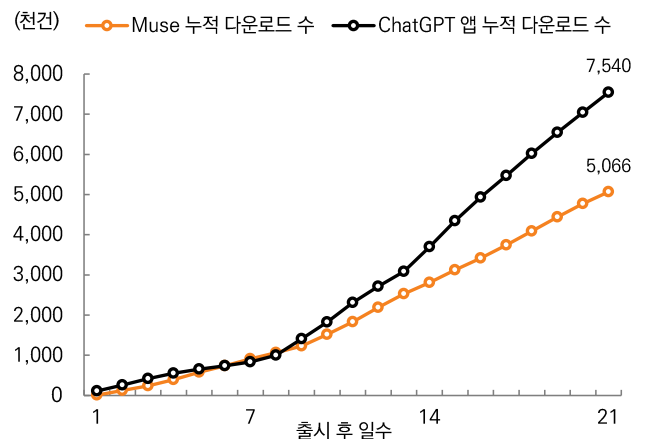


그림 3. 출시 후 일수별 누적 다운로드 수: Muse vs ChatGPT 앱



주: 센서타워 전 세계 추정치. Muse는 2026년 9월 8일 미국과 캐나다에서 iOS, Android 동시 출시. ChatGPT 앱은 2023년 5월 18일 미국 iOS 출시 후 9일 차부터 출시 국가 확대
자료: 센서타워, 미래에셋증권 리서치센터

자료: 센서타워, 미래에셋증권 리서치센터

표 3. Muse와 Dots 비교

구분	메타 Muse	오픈AI Dots
공개일	2026년 9월 8일, 미국(iOS, Android, 웹)	2026년 9월 29일, ChatGPT 안에서
이용 대상	누구나, 대부분 기능 무료	Pro(월 200달러)와 Business Premium 우선, 기업은 베타
실행 환경	사용자별 전용 클라우드 컴퓨터	dot마다 전용 클라우드 컴퓨터
가동 방식	앱을 닫아도 맡긴 일을 수행, 변화가 생기면 알림	24시간 상시 가동, 쉬는 동안에도 할 일을 찾아 처리
사용량 확장	더 쓰려면 구독	dot 추가, 속도와 월 작업량 확대 예정
가상 컴퓨터 사양	가상 CPU 2개, 메모리 8GB, 저장공간 100GB	미공개
초기 이용	출시 21일 차 DAU 157.5만 명	미공개

자료: 각 사, 미래에셋증권 리서치센터

타이트한 부품 공급

문제는 에이전트가 추가로 쓰는 부품의 공급이 이미 빠듯하다는 점이다. TrendForce의 9월 28일 자료에서 서버 CPU 조달 기간은 평상시 16~20주에서 25~30주로 약 1.5배 늘었다. DRAM은 지난 분기 가격이 60%대 초반 오른 뒤에도 조달 기간이 평상시 8주에서 20주로 2.5배 길어졌고, 저용량 모듈 중심으로 올해 4분기부터 내년 1분기까지 공급 부족이 예상된다. 기업용 SSD도 8주에서 16주로 늘었다. 반면 GPU는 30~40주로 평상시와 같아 수급 균형으로 분류됐다.

병목이 GPU보다 **에이전트가 추가로 쓰는 CPU, 메모리, 저장장치** 쪽에 몰려 있다는 뜻이다. 기존 챗봇은 사용자가 질문할 때만 모델을 부르지만, 에이전트는 사용자가 자리를 비운 동안에도 브라우저와 프로그램을 돌린다. 사용량뿐 아니라 컴퓨팅 자원을 점유하는 시간이 길어진다는 점이 다르다.

표 4. 주요 서버 부품 공급 리드타임 현황(26.9.28 기준)

품목	수급 상황	현재 리드타임	정상 리드타임	정상 대비 배수	주요 내용
CPU	타이트	25~30주	16~20주	약 1.5배	- 에이전트형 AI 확산으로 서버 CPU 수요가 강하게 증가 - 주요 CSP도 자체 CPU 개발과 도입을 가속
GPU	수급 균형	30~40주	30~40주	1.0배	- NVIDIA Blackwell(B300/GB300)은 2Q26부터 계획대로 출하 중, 2026년 주력 제품으로 유지될 전망 - Rubin은 HBM 등 upstream 공급 준비가 필요해 서버 물량 확대는 2027년부터 본격화 예상
DRAM	매우 타이트	20주	8주	2.5배	- 특히 저용량 모듈을 중심으로 4Q26~1Q27 공급 부족 - 가격 인상 가능성에 대한 논의 확대
NAND (eSSD)	타이트	16주	8주	2.0배	- NAND 업체들이 미국 CSP의 엔터프라이즈 SSD 수요에 대응해 생산능력 재배치 중 - 그러나 공급 부족 지속
HDD	매우 타이트	50주	16주	3.1배	- 고객의 스토리지 용량 확대 요구에 대응해 HAMR 중심으로 공급 확대 중 - 다만 공급 부족은 최소 2027년까지 지속될 전망
ABF 기판	매우 타이트	48~56주	12주	4.0~4.7배	- AI 칩 다이 면적과 기판 레이어 증가로 수요 확대 - 저CTE 유리섬유 공급은 2027년 완화된 전망 - 그러나 고다층 ABF 기판의 수율 문제와 제한적인 증설로 타이트한 수급 지속 예상
MLCC	타이트	32주	12주	2.7배	- NVIDIA Vera Rubin용 부품 수요 증가로 고사양 범용 MLCC 공급이 타이트 - 반면 구형 TPU 수요는 아직 제한적이라 고객 맞춤형 고사양 MLCC 출하는 상대적으로 더뎠

주: 정상 대비 배수는 현재 리드타임을 정상 리드타임으로 나누어 산출. CPU는 각 구간의 중간값(27.5주/18주) 기준.

자료: TrendForce, 미래에셋증권 리서치센터

서버 수요는 사용자 수보다 작업 시간이 좌우

에이전트가 실제 서버 수요를 얼마나 늘릴지는 사용자 수만으로 판단하기 어렵다. 수요 규모를 가늠하기 위해 출시 1년 뒤 Muse 1억 명, Dots 1,000만 명으로 합산 사용자 1.1억 명을 Base 시나리오로 가정했다. Muse는 과거 ChatGPT의 사용자 성장 궤적을 참고했고, Dots는 오픈AI 유료 구독자 5,000만 명 가운데 20%가 이용하는 경우를 적용했다. 이는 에이전트 확산이 서버 수요에 미치는 영향을 보기 위한 조건부 시나리오다.

사용자당 가상 CPU 2개, 서버당 CPU 256코어를 적용하고 운영 방식을 세 가지로 나눴다. 모든 가상 컴퓨터를 항상 켜두는 경우(A), 사용자의 30%만 동시에 작업하고 물리 코어 하나를 가상 CPU 두 개가 공유하는 경우(B), 동시 작업률을 10%까지 낮추고 코어를 4배 공유하는 경우(C)다.

Base 시나리오에서 필요한 서버는 A 85.9만 대, B 12.9만 대, C 2.1만 대로 계산된다. 같은 사용자 1.1억 명이라도 운영 방식에 따라 **약 40배** 차이가 난다. 현실적인 중간값인 B 기준 12.9만 대는 연간 글로벌 서버 출하 증가분 약 200만 대의 6.5% 수준이다.

핵심은 사용자 수보다 에이전트가 **실제로 동시에 일하는 시간이 서버 수요를 좌우**한다는 점이다. 사용자가 많아도 대부분의 에이전트가 대기 중이면 필요한 물리 서버는 크게 줄어든다. 반대로 사용자를 대신해 브라우저와 프로그램을 장시간 실행할수록 동시 작업률이 높아지고 CPU와 메모리 점유 시간도 길어진다.

따라서 향후 에이전트 인프라를 볼 때는 MAU나 DAU만으로 부족하다. 사용자당 에이전트 작업시간과 동시 실행률, 물리 자원 공유 효율이 실제 서버 수요를 결정하는 핵심 지표가 될 가능성이 높다.

표 5. 운영 방식별 사용자 1명당 서버 계수

운영 방식	동시 작업률(%)	코어 공유(배)	서버 1대당 사용자(명)	1명당 하드웨어(달러)	그중 DRAM 비중(%)	월 서버값(달러)
A 상시 가동	100	1.0	128	616.3	65.5	9.34
B 일부 동시 가동	30	2.0	853	161.8	74.9	2.45
C 적극적 자원 가동	10	4.0	5,120	51.3	78.7	0.78

주1: 가상 컴퓨터용 CPU, 메모리, 저장장치만 포함하고 AI 모델 추론용 GPU는 제외. 월 서버값은 1명당 하드웨어를 서버 내용연수(5.5년)로 나눈 값

주2: 사용자당 가상 CPU 2개, 서버당 CPU 256코어 기준. 모델 추론용 GPU 제외. 기존 여유 설비 차감 전 필요 용량으로 실제 신규 발주량과 차이가 있을 수 있음. DRAM은 2026년 9월 28일 DDR5 RDIMM 64GB 평균 가격(GB당 50.5달러) 기준

자료: DRAMeXchange, DIGITIMES, 메타, 미래에셋증권 리서치센터 추정

표 6. 에이전트 사용자 시나리오별 필요 서버

시나리오	Muse(만 명)	Dots(만 명)	합계(만 명)	A 항상 켜 둠(만 대)	B 30%(만 대)	C 10%(만 대)	A의 연간 서버 증가분 대비(%)
1년 뒤 Bear	5,000	200	5,200	40.6	6.1	1.0	20.3
1년 뒤 Base	10,000	1,000	11,000	85.9	12.9	2.1	43.0
1년 뒤 Bull	20,000	5,000	25,000	195.3	29.3	4.9	97.8
2년 뒤 Bear	14,000	500	14,500	113.3	17.0	2.8	56.7
2년 뒤 Base	28,000	3,000	31,000	242.2	36.3	6.1	121.3
2년 뒤 Bull	40,000	10,000	50,000	390.6	58.6	9.8	195.6

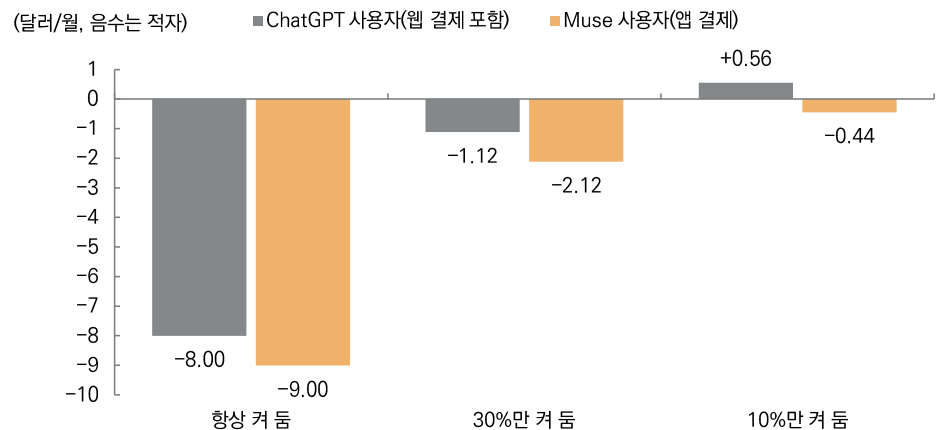
주: 서버 대수는 합계(서버 1대당 사용자(표 5). 연간 서버 증가분은 2025년 1,560만 대에 TrendForce 증가율 12.8% 적용(약 200만 대)

자료: 오픈AI, 메타, TrendForce, DIGITIMES, 미래에셋증권 리서치센터 추정

비용이 자원 공유를 강제하고, 공유할수록 DRAM 비중은 높아진다

에이전트의 운영 방식은 결국 경제성이 결정한다. 사용자 1명당 서버 하드웨어 비용을 메타의 서버 내용연수 5.5년으로 나눠 월 비용으로 환산했다. 대부분이 무료 사용자여서 1인당 월 매출은 ChatGPT 약 1.33달러, Muse 약 0.33달러에 그친다. 가상 컴퓨터를 항상 켜둘 경우 월 서버 하드웨어 비용만 약 9.3달러로, 두 서비스 모두 1인당 매출을 크게 웃돈다. ChatGPT도 동시 실행률을 30%로 낮추고 CPU를 공유해도 1인당 월 1.1달러 적자이고, 10%까지 낮추고 코어를 4배로 나눠 써야 흑자로 돌아선다.

그림 4. 에이전트 컴퓨터 운영 방식별 사용자 1명당 월 손익



주: 1인당 월 매출에서 1인당 월 서버 하드웨어 비용을 차감. 모델 추론용 GPU, 전력, 데이터센터 및 네트워크 비용 제외
자료: 오픈AI, 센서타워, 메타, 미래에셋증권 리서치센터 추정

실제 운영 부담은 이보다 크다. 계산에서 GPU와 전력, 데이터센터 공간, 네트워크 비용을 제외했기 때문이다. 따라서 무료/저가 에이전트를 대규모로 제공하려면 필요할 때만 가상 컴퓨터를 깨우고 CPU를 여러 사용자에게 공유하는 방식으로 자원 효율을 높여야 한다. 오픈AI가 Dots를 월 200달러 Pro와 기업용 상위 요금제에 먼저 제공하는 것도 이러한 비용 구조와 관련있다. 다만 해당 요금제에는 다른 모델과 서비스도 포함돼 있어 구독료 전체를 Dots 매출로 볼 수는 없다.

여기서 일반 서버용 DRAM의 중요도가 높아진다. CPU는 여러 작업이 시간차를 두고 같은 코어를 사용할 수 있어 공유율을 높일수록 사용자당 비용이 빠르게 낮아진다. 반면 실행 중인 에이전트의 작업 상태와 데이터는 메모리에 계속 유지해야 한다. 당사 가정에서 자원 공유를 확대할수록 사용자당 하드웨어 비용은 줄지만 DRAM의 비용 비중은 66%에서 최대 79%까지 상승한다. 에이전트 확산이 HBM뿐 아니라 일반 서버용 DRAM에도 새로운 수요를 더할 수 있다고 보는 이유다.

학습과 기존 챗봇이 GPU와 HBM 중심이었다면, 에이전트는 CPU, DRAM, 스토리지로 구성된 작업 환경을 장시간 유지한다는 점에서 다르다. 초기 Muse 이용자가 빠르게 늘고 CPU와 DRAM 공급도 이미 타이트한 만큼, 에이전트가 실제 업무를 수행하는 시간이 길어질수록 AI 인프라의 병목도 GPU에서 일반 서버 부품으로 넓어질 가능성이 높다.

[Investment Signal]

공급 부족을 이익으로 연결할 기업에 주목

AI 인프라 수요가 GPU에서 CPU, 메모리와 일반 서버로 넓어지는 동시에 부품 공급은 빠듯해지고 있다. 같은 공급 부족이라도 영향은 다르다. 부품 업체에는 가격과 믹스 개선의 기회지만, 서버와 네트워크 장비 업체에는 원가 부담이다. 향후 실적 차별화는 수요 자체보다 공급 부족을 얼마나 이익으로 전환할 수 있는지에 달려 있다고 판단한다.

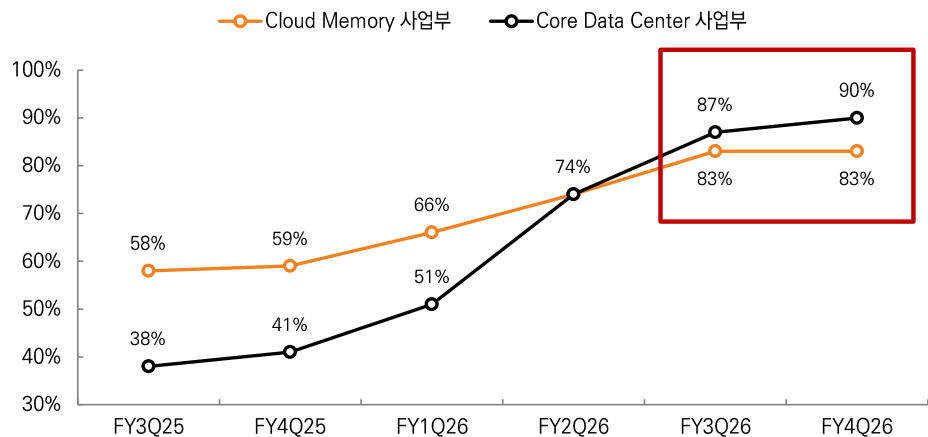
메모리와 서버 CPU: 일반 서버 확산의 직접 수혜

가장 먼저 주목할 영역은 일반 서버용 DRAM이다. 에이전트가 새로운 수요를 만드는 가운데 공급 부족까지 겹치고 있다. 특히 CPU는 가상화와 코어 공유로 사용자당 자원 투입을 줄일 수 있지만, 실행 중인 작업 상태를 유지할 메모리는 상대적으로 덜 줄어든다. AI 메모리 수요가 HBM에서 일반 서버용 DRAM까지 확산될 가능성이 높다.

마이크론의 FY26 4분기 Core Data Center 사업부 매출총이익률이 HBM을 포함한 Cloud Memory 사업부를 웃돈 것도 이러한 흐름을 보여준다. 삼성전자와 마이크론 등 일반 DRAM 비중이 높은 업체들이 대표적인 관련 기업군이다.

서버 CPU도 직접적인 수혜 영역이다. 에이전트가 브라우저와 애플리케이션을 실행하는 시간이 늘수록 CPU 수요도 증가한다. AMD와 인텔 등이 주요 공급사지만, CPU는 코어 공유가 가능해 DRAM보다는 사용자 증가가 수요로 직결되는 정도는 낮다.

그림 5. 마이크론 사업부별 매출총이익률 추이



주: Cloud Memory는 대형 하이퍼스케일러용 메모리와 HBM, Core Data Center는 일반 클라우드, 기업, 서버 제조사용 메모리와 데이터센터 SSD

자료: 마이크론, 미래에셋증권 리서치센터

ASIC 공급망과 네트워크: 연산 다변화가 키우는 연결 수요

AI 백엔드 네트워크 시장이 빠르게 커지는 가운데 ASIC 서버 비중도 상승하고 있다. 엔트로픽의 취소 제한 컴퓨트 약정 중 92.4%가 TPU와 Trainium에 집중된 점을 감안하면, 2027년 이후에도 연산이 엔비디아 GPU에서 ASIC으로 다변화되는 흐름은 이어질 가능성이 높다.

이 과정에서 수혜는 스위치 칩과 장비, 케이블, 광으로 확산된다. 브로드컴은 ASIC과 이더넷 스위치 칩을 함께 공급하고, 크레도는 AEC와 광을 통해 고속 링크 증가에 노출된다. 아리스타와 시스코는 ASIC과 AMD 기반 클러스터에서 표준 이더넷 채택이 늘어날수록 기회가 커질 수 있다.

결국 핵심은 특정 가속기의 점유율보다 가속기 간 연결 수와 속도의 증가다. 2027년에는 1.6T와 스케일업 이더넷이 실제 매출로 이어지는지가 주요 확인 포인트다.

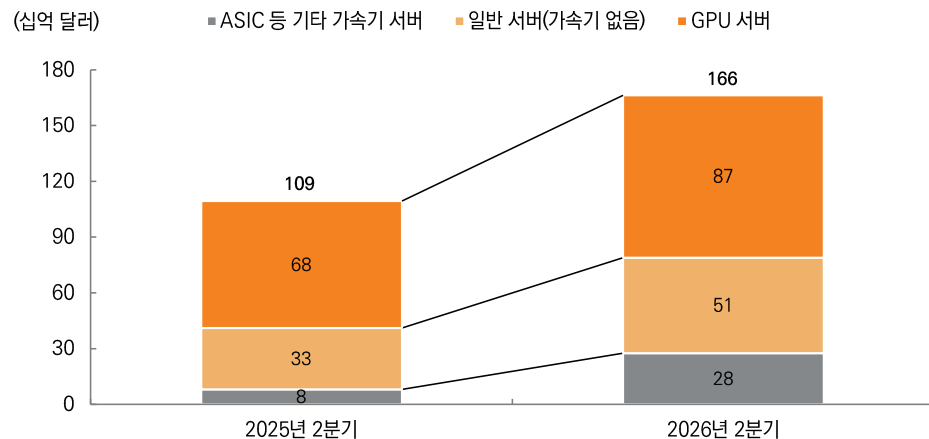
서버 제조사: 일반 서버 회복은 기회, 부품 가격은 부담

에이전트와 추론 확대는 GPU 서버뿐 아니라 CPU, DRAM, 스토리지를 사용하는 일반 서버 수요도 늘린다. 특히 자체 설계 역량이 낮은 기업, 네오클라우드와 소버린 고객이 확대되면 브랜드 서버 업체의 역할도 커질 수 있다.

델은 AI 서버 백로그와 일반 서버 회복이 동시에 나타나고 있다는 점이 강점이다. HPE는 일반 서버 비중이 높고 네트워크 사업까지 보유해 CPU 기반 추론과 에이전트 확산에 대한 노출이 상대적으로 크다.

다만 DRAM, CPU와 스토리지 가격 상승은 원가 부담이다. 따라서 서버 업체의 실적 차별화는 출하 증가보다 부품 확보 능력과 원가를 판매가격에 전가하는 속도에서 나타날 것으로 판단한다.

그림 6. 서버 시장 매출 구성



자료: IDC, 미래에셋증권 리서치센터

표 7. 델과 HPE의 서버 사업 지표

(십억 달러)

항목	델	델 1년 전	델 YoY	HPE	HPE 1년 전	HPE YoY
서버 매출	26.9	12.9	+108.1%	6.8	5.0	+35.3%
AI 서버 매출	16.4	8.2	+99.8%	1.6	1.6	0.0%
일반 서버 매출	10.5	4.7	+122.4%	5.2	3.4	+51.9%
일반 서버 비중(%)	39.1	36.6	+2.5%p	76.4	68.0	+8.4%p
AI 서버 수주	60.9	5.6	+987.5%	2.4	2.1	+14.3%
AI 서버 백로그	95.0	11.7	+712.0%	6.8	3.7	+83.8%
영업이익률(%)	15.0	8.8	+6.2%p	17.0	7.0	+10.0%p

자료: 각 사, 미래에셋증권 리서치센터

AI 인프라 포트폴리오

표 8. AI 인프라 밸류체인 테이블 (26.10.1 종가 기준)

(십억 달러, 달러, %, 배)

구분	기업명	주가	시가총액	매출성장률		영업이익률		EPS 성장률		P/E		EV/EBITDA		수익률			
				26E	27E	26E	27E	26E	27E	12MF	24MF	12MF	24MF	5D	1M	3M	12M
서버 OEM	델	541.7	344.4	70.1	20.9	11.3	10.8	151.2	17.7	18.7	16.0	13.5	11.4	1.1	10.1	37.4	267.6
	HPE	64.6	85.7	35.8	17.5	14.7	15.1	96.8	22.9	14.0	12.1	8.8	8.2	1.7	24.6	56.6	163.8
	슈퍼마이크로	41.9	27.5	71.6	20.5	7.7	7.7	18.8	24.4	9.1	7.5	5.0	4.0	1.0	13.3	54.0	-20.2
네트워크 시스템	엔비디아	230.9	5,563.7	90.3	70.8	65.4	65.2	96.5	69.2	16.8	11.8	13.3	9.1	2.8	2.9	18.5	22.2
	시스코	108.8	428.8	11.8	20.7	24.3	27.4	13.9	30.2	21.0	19.2	15.6	14.2	1.7	-0.6	-3.5	59.2
	아리스타	204.5	257.9	42.8	54.2	44.0	42.5	42.5	50.3	41.7	33.7	31.6	25.0	-0.6	9.9	27.8	41.6
광/DCI (전송 /트랜시버)	코닝	160.4	138.2	16.6	19.4	21.7	23.4	30.3	31.9	39.4	29.2	21.3	16.4	4.1	11.3	-18.5	94.4
	루멘텀	1,045.8	93.8	108.7	53.4	41.2	43.8	149.1	57.8	42.1	27.9	28.1	17.6	12.6	20.1	43.6	516.1
	코히어런트	319.2	62.5	50.3	37.7	23.7	26.0	67.3	53.5	29.9	20.4	19.9	13.5	9.8	18.8	-4.3	183.0
	노키아	10.3	59.4	4.4	7.6	10.9	12.7	16.6	21.0	23.5	19.8	13.9	11.4	1.3	8.0	-14.8	122.5
	시에나	379.1	53.8	34.7	31.7	20.4	25.6	169.2	66.4	33.1	22.3	23.8	16.5	5.8	7.1	-10.3	148.4
	파브리넷	451.4	16.2	30.9	19.1	11.1	11.3	26.5	20.5	24.1	19.5	19.2	15.2	13.3	14.2	-9.6	22.2
인터넷 반도체 (SerDes /DSP /리타이머)	브로드컴	343.6	1,640.4	65.9	64.5	66.9	65.4	71.1	64.9	18.4	11.6	14.2	8.6	-1.9	-6.4	-4.7	1.6
	마벨	268.1	235.1	46.9	51.7	37.1	38.8	48.1	61.8	45.0	28.5	35.2	22.8	3.5	29.8	9.3	211.0
	아스테라 랩스	356.4	61.8	121.1	65.0	36.4	38.5	116.8	62.7	60.7	45.6	57.1	40.3	-1.1	30.0	-12.3	70.0
	크레도	210.2	39.5	109.4	65.8	30.4	36.7	94.9	64.8	27.3	19.3	22.5	14.8	7.2	27.2	-13.1	40.5
	맥컴	301.6	23.0	36.8	35.5	31.7	37.3	58.8	50.5	36.3	30.0	-	-	5.9	17.1	-6.4	135.5
	셈테크	186.6	17.4	41.9	31.9	27.3	36.1	103.8	70.8	36.3	24.4	27.1	19.6	6.9	39.4	37.9	155.8
화이트박스 /ODM	액톤 테크놀로지	373.1	47.0	65.9	73.4	8.3	8.7	85.4	77.1	21.0	14.8	15.3	9.9	-0.1	34.3	11.0	48.7
	관타 컴퓨터	10.3	42.5	90.9	31.4	3.2	3.1	38.3	19.2	10.7	9.0	9.5	8.0	-3.2	-1.9	-10.4	12.4
	셀레스티카	61.2	34.3	61.5	33.4	14.9	15.2	81.3	37.5	17.8	12.9	12.8	9.1	3.4	-8.6	-28.9	81.9

주: 아리스타, 시스코, 크레도는 당사 추정, 그 외는 블룸버그 컨센서스

자료: 블룸버그, 미래에셋증권 리서치센터

Compliance Notice

- 당사는 자료 작성일 현재 해당 회사와 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트는 자료작성일 현재 본 자료에서 매매를 권유한 금융투자상품 및 권리를 보유하고 있지 않습니다.
- 본 자료는 외부의 부당한 압력이나 간섭없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.

본 조사분석자료는 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻은 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 종목 선택이나 투자시기에 대한 최종 결정을 하시기 바랍니다. 따라서 본 조사분석자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다. 본 조사분석자료의 지적재산권은 당사에 있으므로 당사의 허락 없이 무단 복제 및 배포할 수 없습니다.