Market Insight | 2026.3.6

그림: Midjourney(너도 안빼고 새로운 영웅은 언제나 환영이야)

MIRAE ASSET 미래에셋증권

AI Bi-weekly

새로운 사이클은 언제나 환영이야

한종목
chongmok.han@miraeasset.com

김은지
eunji.kim.a@miraeasset.com

CONTENTS

Highlight of the Week

CPU가 돌아왔다

AI 인프라의 중심축이 대규모 병렬 연산을 전담하던 GPU에서 극초저지연 순차 제어를 담당하는 매니코어 CPU로 이동하고 있습니다. 2026년 현재 시장의 표준으로 자리 잡은 멀티 에이전트 및 Hydra 아키텍처는 수십 개의 독립된 자아가 실시간으로 상호 검증하고 코드를 수정하는 복잡한 논리 회로를 강제합니다.

이 과정에서 발생하는 초당 수백만 번의 조건부 분기 판단과 방대한 KV 캐시의 컨텍스트 스위칭은, 동일한 명령을 수만 개의 코어에 동시다발적으로 하달해야만 제 성능을 발휘하는 단일 명령어 다중 스레드 구조를 완벽하게 마비시킵니다. 에이전트 워크로드는 단순한 연산량의 폭발이 아니라 순차적 논리 제어 부하의 폭발을 의미하며, 이를 해결하기 위해 강력한 단일 스레드 통제력과 거대한 내부 캐시를 갖춘 CPU가 데이터센터의 최고 지휘관으로 귀환했습니다.

이러한 물리적 워크로드의 근본적 변화는 클라우드 벤더들의 데이터센터 아키텍처를 원점에서 해체하고 있습니다. 과거 1U 서버 박스 내에 연산과 제어 장치를 혼합 배열하던 모놀리식 구조는 열역학적 파산을 맞이하고 있습니다. 대신 순수 병렬 연산만을 담당하는 수냉식 가속기 랙과, 수만 개의 에이전트 상태 통제를 전담하는 초고밀도 CPU 랙이 물리적으로 완전히 분리되는 디스어그리게이션 아키텍처가 확립될 것으로 보입니다. 이 지휘관 랙을 선점하기 위해 AMD는 칩렛 기술의 원가 경제성을 극한으로 끌어올려 서버 시장 점유율 45퍼센트를 정조준하고 있으며, 인텔은 클라우드 환경의 최대 리스크인 보안 취약점과 성능 불균형을 해결하기 위해 20년간 이어온 하드웨어 멀티스레딩 기능을 과감히 삭제하는 아키텍처 혁명을 단행했습니다. 동시에 아마존과 구글 같은 하이퍼스케일러들은 범용 x86 명령어 세트를 버리고 자사 트래픽 라우팅에 최적화된 극한의 맞춤형 Arm 실리콘을 전면 도입하여 총소유비용을 억제하고 있습니다.

엔비디아는 이 거대한 하드웨어 권력 이동을 공급망 통제력과 자체 아키텍처 혁신으로 잘 방어해내고 있습니다. 과거 타사의 기성품 코어가 분기 예측에서 실패했던 뼈아픈 경험을 거울삼아, 자체 설계한 올림푸스 코어를 적용한 Vera 프로세서를 출시할 예정입니다. 특히 스레드 간 물리적 간섭을 원천 차단한 '공간적 멀티스레딩' 기술을 도입하여 CPU 시장의 핵심 파이마저 잠식하기 시작했으며, 메타와 맺은 Vera 단독 대규모 공급 계약이 그 물리적 증거입니다.

2029년 300기가와트 규모의 전력 소모가 예고된 극한의 열역학 전쟁에서 최첨단 패키징과 고대역폭 메모리 물량을 선제적으로 확보한 엔비디아는 높은 가격 결정력을 유지할 가능성이 높습니다. 가속기 사이클이 성숙했다는 시장의 얕은 오판 속에, 에이전트 트래픽이 유발하는 거대한 CPU 및 인터커넥트망의 강제 교체 수요는 아직 관련 기업들의 밸류에이션에 반영되지 않은 거대한 투자 알파입니다.

AI 팬데믹 폭풍전야

2026년 2월 5일, OthersideAI CEO Matt Shumer의 에세이 "Something Big Is Happening"이 48시간 만에 8,500만 뷰를 돌파했습니다. 그가 발견한 것은 새로운 사실이 아니었습니다. 그는 이미 공개된 데이터들을 하나의 직관으로 꿰뚫었습니다. 팬데믹 선언 직전, 아직 일상이 정상처럼 보였지만 이미 모든 것이 돌이킬 수 없이 달라져 있던 2020년 2월. 그는 지금 우리가 그 시점에 있다고 주장했습니다.

이 주장을 뒷받침하는 가장 강력한 근거는 METR의 "시간지평(Time Horizon)" 데이터입니다. AI가 50% 확률로 자율 수행할 수 있는 작업의 최대 길이는 2019년 2초에서 2026년 2월 Claude Opus 4.6 기준 11시간 59분으로 도약했습니다. 더 주목할 점은 이 성장 자체가 가속되고 있다는 사실입니다. 7개월마다 두 배씩 늘던 속도가 2024년 이후 4.3개월로 압축되었습니다. 지수함수 위에서 다시 가속이 일어나고 있습니다.

이 변화는 이미 현실에 착지하고 있습니다. 앤트로픽의 Claude Cowork 엔터프라이즈 전환, Claude Code의 연환산 매출 25억 달러 도달, GitHub 전체 커밋의 4%를 AI가 생성하는 현실이 이를 증명합니다. 그러나 NBER 조사에서는 응답 기업의 80% 이상이 "AI가 생산성에 영향을 주지 않았다"고 답했습니다.

하지만 우리는 이 80%가 AI의 한계를 보여주는 데이터가 아니라고 판단합니다. OpenAI 엔터프라이즈 보고서가 보여주듯, 같은 구독료를 내는 조직 안에서도 상위 5%는 나머지보다 6배 많은 메시지를 보내고 있습니다. 대부분의 기업이 경험한 AI는 프론티어가 아닌, 이미 수 세대 전의 모델이었을 가능성이 높습니다. 1980년대 IT 투자가 생산성 통계에 나타나기까지 10년이 걸렸던 Solow 역설처럼, AI 생산성 역설도 같은 패턴을 따르고 있습니다. 다만 이번에는 훨씬 압축적으로 전개될 것입니다.

결론적으로 Shumer의 핵심 주장은 유효합니다. 변곡점은 실재합니다. 다만 그가 과소평가한 것이 있습니다. AI 채택의 품질 역시 불균등하다는 점입니다. 기술이 가속할수록, 프론티어를 먼저 경험하는 조직과 그렇지 못한 조직 사이의 격차는 지수적으로 벌어질 것입니다. 지금의 한 발자국은 1년 후의 한 발자국보다 훨씬 싸다는 것을 명심해야 할 시점입니다.

표 1. AI 관련 주요 일정

일	월	화	수	목	금	토
1	2	3	4	5	6	7
	· Alchip 실적(예) · MWC 2026(~5)	· CRWD 실적	· AVGO 실적	· MRVL 실적	· BABA 실적(예) · Alchip 실적	
8	9	10	11	12	13	14
	· Gartner Data & Analytics Summit 2026(~11)	· ORCL 실적		· 위스트론 실적		
15	16	17	18	19	20	21
	· 폭스콘 실적(잠)			· ACCN 실적		

자료: Bloomberg, 미래에셋증권 리서치센터

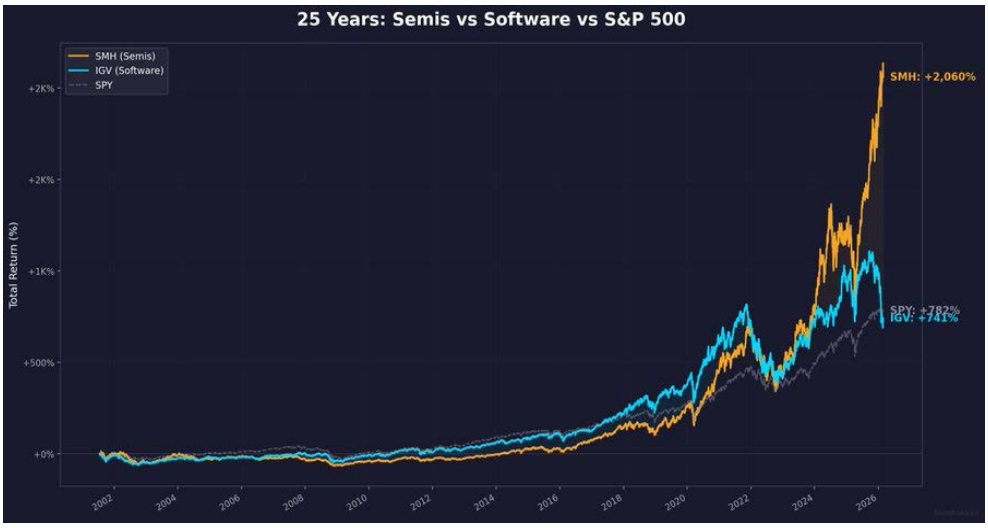
I. AI Issue

1. AI 자본 전쟁과 CPU의 귀환

(1) 서론: 균형의 균열에서 투자 알파가 태어난다

지난 3년간 AI 반도체 시장의 가장 큰 내러티브는 연산의 무게 중심이 그래픽처리장치(GPU)로 영구히 이동했고 CPU는 실리콘 권력의 변방으로 밀려났다는 것이었다. 하이퍼스케일러들은 엔비디아의 GPU를 선점하기 위해 자본 지출(CapEx)을 끌어올렸고, 시장은 이를 AI 인프라 사이클의 거의 전부로 인식했다. 실제로 GPU가 왕좌를 차지하고, 권세가 높아진 왕세자인 메모리 반도체가 슈퍼사이클에 합류한 상태다.

그림 1. Semis(SMH) vs Software(IGV) vs S&P500(SPY) 누적 수익률
2002년부터 2026년 2월까지 1달러를 각각 넣었다면, 반도체는 21.6달러가 됐고,
소프트웨어는 8.41달러, 전체 시장은 8.82달러가 됐다.
2022년까지는 둘 다 비슷하게 움직였는데, 생성 AI 혁명 때문에, 반도체의 슈퍼사이클이 시작



자료: SemiAnalysis, 미래에셋증권 리서치센터

반도체의 원조 왕, CPU의 존재는 외면당했다. 그런데 지금, 그 잊힌 왕이 돌아오고 있다. AI 대전에서 자존심이 짓밟혔을 그 프로세서가, 역설적이게도 바로 그 AI 때문에 슈퍼사이클을 맞이하고 있다. 2026년 3월 현재 실리콘 밸리의 최전선 데이터와 데이터센터 총소유비용(TCO) 장부는 기존과는 다른 진실을 가리키고 있다.

거대한 자본 전쟁의 양상은 지속되고 있으나, 그 자본이 투입되는 물리적 목적지가 조금 변하는 듯하기 때문이다. AI의 패러다임이 단일 턴 방식의 텍스트 생성(LLM)에서, 스스로 코드를 작성하고 환경을 시뮬레이션하며 검증하는 강화학습(RL) 및 멀티 에이전트(Multi-Agent) 워크로드로 진화함에 따라 인프라의 궤도가 수정되고 있는 것이다.

*** 에이전트 팬아웃:**
하나의 메인 에이전트가 복잡한 작업을 분해한 뒤, 수십 개의 전문 하위 에이전트(코드 작성자, 웹 검색기, 요약자, 메일 발송자 등)를 동시에 호출하는 구조. 이게 '외부'인 이유는, Hydra처럼 단일 모델 안에서 자아를 나누는 게 아니라, 실제로 별도의 에이전트 인스턴스를 새로 띄워 병렬로 돌리는 외부 확장(fan-out)이기 때문

사실 이것은 OpenRouter의 실시간 데이터가 증명하고 있다. 에이전트 코딩 및 코딩 어시스턴트 관련 쿼리가 전체의 11%에서 50%로 치솟았다. 게다가 최근에 출시된 xAI Grok 4.20의 Hydra 아키텍처(하나의 모델 안에 4~16개 독립된 '자아'가 동시에 깨어나 계획을 세우고 서로 실시간 검증하는 구조)와 앤트로픽 Opus 4.6 에이전트 팀이 업계 표준이 되면 상황은 더욱이 긴박하게 바뀐다. "단일 모델 내부의 멀티 페르소나와 외부 멀티 에이전트 팬아웃"이 곱해지는 승수효과를 가지기 때문이다.

이는 결과적으로 컨텍스트 스위칭·KV 캐시 관리·보안 검증 같은 CPU-bound 작업이 지수적으로 폭증해, GPU만으로는 경제적으로 버틸 수 없게 만든다. SemiAnalysis의 편집장 Dylan Patel이 정확히 지적한 대로 "인간은 1초 딜레이를 참지만, 내부적으로 수백 번의 루프를 도는 에이전트에게 1초라도 느린 칩은 고철"이다. 이 때문에 KV 캐시(이전에 생성한 토큰 정보를 저장하는 메모리 공간)가 메모리를 잡아먹는 Memory Wall이 본격화되고, 이를 라우팅·관리하는 CPU 수요가 승수 효과로 터진다.

이 보고서의 논거는 간단하다. 추론(Inference)의 시대가 곧 에이전트(Agent)의 대중화를 의미하고, 수많은 에이전트는 결국 지휘자의 역할을 하는 'Control Plane'(CPU만이 가능한 것)을 극도로 많이 필요로 한다는 것이다. 올해가 지나갈수록, 인텔, AMD, 하이퍼스케일러, 엔비디아, Arm 등으로 구성되는 데이터센터 CPU 시장의 판도가 급격하게 재편되는 흐름을 목도하게 될 것이다. 균형의 균열이 있다는 것은 투자 알파의 기회가 있다는 것을 의미한다.

(2) 브로드컴의 1,000억 달러 가이던스와 프론티어 AI 랩의 300GW 자본 전쟁

최근 시장을 관통한 가장 중요한 데이터 포인트는 브로드컴 CEO 혹 탄(Hock Tan)이 제시한 회계연도 2027년(2026년 11월~2027년 10월) AI 칩 매출 1,000억 달러 이상이라는 가이던스다. 이는 프론티어 AI 연구소들이 주도하는 자본 전쟁의 스케일이 근본적으로 변했음을 알리는 선행 지표로 볼 수 있다. 이 거대한 자본이 어디로, 왜 향하고 있는지를 이해하는 것이 현 AI 사이클 분석의 출발점이다.

체스판의 후반부: 스케일링 법칙은 계속

월스트리트 일부는 2027년 AI 인프라 투자가 정점을 찍고 하락할 것이라는 수확 체감의 법칙을 거론한다. 이와 관련해 혹 탄이 남긴 말 중 가장 충격적인 것은 따로 있다. "앤트로픽이 2026년 1GW 규모의 TPU 컴퓨팅을 가동하고, 2027년에는 3GW 이상으로 수요를 확대할 것"이라는 선언이었다.

여기서 앤트로픽 단 한 곳이 2027년에 3GW의 TPU/ASIC 인프라를 요구한다는 점이 포인트다. 구글의 7세대 TPU와 메타의 물량 등을 합치면 브로드컴의 맞춤형 실리콘 진영에서만, 그들이 사용하는 전력 규모는 2027년에 약 10GW에 달할 것으로 사료된다. 이는 엔비디아의 GPU를 구동시키는데 들어가는 전력 규모는 제외한 것이다. 다리오 아모데이가 최근에 예측한 "2027년에는 AI 산업 전체적으로 30~40GW가 소모될 것"라는 숫자는 과장이 아니라 실리콘 벤더들의 확정된 공급망 장부에 이미 기재된 현실인 것이다.

표 2. 앤트로픽 CEO 다리오 아모데이의 산업 전체 데이터센터 구축 전망

연도	최소 (GW)	최대 (GW)	전년 대비 증가배수
2026	10.0	15.0	-
2027	30.0	40.0	2.8
2028	100.0	100.0	2.9
2029	300.0	300.0	3.0

자료: 다리오 아모데이, 미래에셋증권 리서치센터

게다가 아모데이는 2029년쯤이 되면 300GW에 달하는 전력을 소모하는 인프라가 구축될 것을 시사했다. 이 미친듯한 지출을 정당화하는 핵심 기제는 재귀적 자기 개선(Recursive Self-Improvement, RSI)의 도래다. AI가 스스로 아키텍처를 설계하고 고품질의 합성 데이터(Synthetic Data)를 무한 생성하여 학습의 병목을 없애는 순간, 기술 발전 속도는 인간의 개입을 벗어나 기계의 연산 속도에 종속된다. 다리오 아모데이는 이를 체스판의 후반부에 진입한 것으로 규정했다. 그러니 아직은 체스판 초반이기 때문에 월가가 떠올리는 '수확 체감 법칙'은 논의할 가치가 떨어진다는 얘기다.

파죽지세 앤트로픽

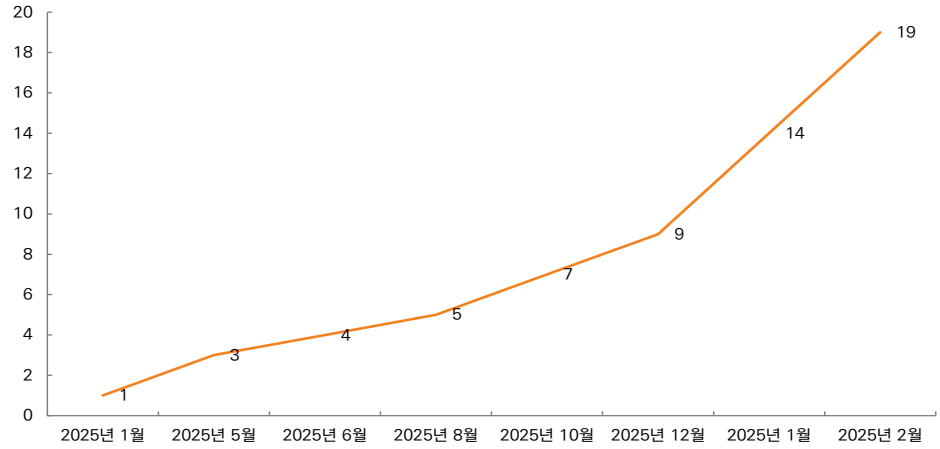
앤트로픽의 CEO가 이처럼 거대한 예상을 하는 자신감의 근거도 있다. OpenAI가 주간 8억 명의 B2C 사용자를 안고 2026년 140억 달러의 적자를 예상하며 인프라 비용의 늪에 빠진 반면, 앤트로픽은 극단적인 자본 효율성을 증명하고 있다. 또한 3년 연속 매년 10배 정도의 성장세다. 앤트로픽은 2024년 12월 10억 달러였던 ARR(연간반복매출)을 불과 1년여 만에 95억 달러로 폭발시켰다. 그리고 최신 기준으로는, 단 3주 만에 50억 달러의 추가 매출을 올려 현재의 ARR은 190억 달러에 달한다.

그림 2. 앤트로픽이 B2B 자본 효율성을 통해 단 3주 만에 50억 달러의 신규 매출을 창출
정치적 마찰이라는 표면적 악재에도 불구하고 190억 달러 ARR의 비선형적 폭발을 이뤄내고 있음. 3,800억 달러의 기업 가치는 스케일링 법칙의 유효성을 증명, AI 인프라 자본지출의 명분이 됨

자료: X(@deedy), 미래에셋증권 리서치센터

핵심 원동력은 앤트로픽 전체 매출의 80%가 발생하는 B2B 엔터프라이즈 시장을 타겟으로 한 클로드 코드(Claude Code)의 전사적 도입이다. Claude Code 단일 항목만 ARR 25억 달러를 기록하고 있다. Fortune 10 중 8개사가 사용하고, 연간 $100만 이상 지출 고객이 500개사를 넘었다. 사용자당 수익(ARPU) 측면에서도 앤트로픽(211달러)은 OpenAI (25달러)를 8.4배로 압도한다. 앤트로픽은 100명의 클로드를 감독하는 군대(Army of 100 Claudes)라는 개념을 통해 AI를 보조자에서 대리인으로 격상시켰다. 에이전트 적용의 승리다.

그림 3. Anthropic 매출(연율화 ARR 기준) 추이 (단위: 십억 달러)



자료: 미래에셋증권 리서치센터

OpenAI의 추론(Inference) 경제학

OpenAI의 입장에서는, 주간 8억 명의 B2C 사용자는 막대한 데이터 플라이휠을 제공하지만 동시에 통제하기 힘든 추론 비용이 발생한다. 에이전트 시대를 맞아 수억 명의 에이전트 군대가 유발하는 추론 쿼리를 범용 엔비디아 GPU로 처리하는 것은 경제적으로 엄청난 소모가 된다. 전력비는 물론이고 감가상각비가 서비스 수익을 크게 갉아먹게 된다.

따라서 빅테크와 프론티어 AI 랩들은 자사의 특정 AI 모델 구조에만 맞춰 불필요한 트랜지스터를 다 깎아내고 추론 전성비(와트 당 토큰 등의 지표)를 극단적으로 끌어올린 맞춤형 가속기(Custom ASIC) 구축에 사활을 걸고 있다. 이에 따라 OpenAI가 브로드컴의 6번째 고객으로 합류하여 2027년 1GW 이상의 컴퓨팅 용량을 1세대 자체 XPU(Custom ASIC)로 채우겠다고 선언한 것은 엔비디아 생태계로부터의 독립 선언이자 생존을 위한 TCO 방어 전략이다.

물론, 이러한 막대한 자본 투자는 단순 연산 장치(Compute)에만 머물지 않는다. 수십만 개의 칩을 엮어 거대한 클러스터로 구동하기 위해서는 초고속 이더넷 스위치 칩도 필수적이다. 자본 전쟁의 1막이 무차별적인 GPU 매입이었다면, 2막은 TCO와 데이터의 움직임을 통제하기 위한 전방위적 인프라 구축으로 이동했다. 그리고 이 변화하는 인프라의 중심에서, 우리는 새로운 병목 현상을 마주하게 된다. 바로 에이전트 워크로드를 지휘할 고성능 CPU의 구조적 결핍이다.

2. 에이전트 시대, 왜 CPU인가?

거대한 자본이 맞춤형 실리콘과 추론 인프라로 향하고 있다. 그렇다면 이 추론 인프라에서 실제로 어떤 연산이 일어날까? 과거 단일 턴 챗봇 시대의 GPU 중심 아키텍처만 가지고는 에이전트 시대에는 무리라는 기술적 필연성을 짚어야 한다.

(1) GPU의 한계: SIMT 구조와 Warp Divergence

OpenRouter의 데이터에 따르면, Agentic coding 및 코딩 어시스턴트 관련 쿼리가 전체의 11%에서 50%로 치솟았다. 이것이 AI 인프라의 스펙을 결정짓는 핵심 동인이다. 반년 전까지 AI 인프라의 10분의 1에 불과하던 워크로드 유형이 이제 전체의 절반을 차지한다. 인프라의 설계 철학 자체가 뒤집어져야 하는 규모의 전환이다.

과거의 AI는 사용자가 질문하면 대답을 한 번 뱉어내는 '단발성 상호작용'이었다. 그러나 에이전틱 AI는 차원이 다르다. 코드를 읽고, 버그를 찾고, 스스로 컴파일해보고, 에러가 나면 다시 수정하는 수십 단계의 다단계 워크플로우를 거친다. 또 에이전트가 코드를 짜려면 함수 하나만 보는 게 아니라, 소프트웨어 저장소(Codebase) 전체를 한 번에 읽고 이해해야 한다. 이를 위해 컨텍스트 윈도우가 급격히 커지고 있다. 문제는 컨텍스트가 길어질수록 AI 모델이 기억해야 할 'KV 캐시(Key-Value Cache, 이전에 생성한 토큰 정보를 저장해두는 메모리 공간)' 데이터가 메모리를 집어 삼킨다는 점이다. 연산 능력(Compute)보다 메모리 용량과 대역폭이 먼저 고갈되는 '메모리 벽(Memory Wall)' 현상이 발생한다. 그리고 이 메모리 벽을 관리하고, KV 캐시를 효율적으로 라우팅하며, 수십 단계의 에이전트 파이프라인을 오케스트레이션하는 것은 GPU가 아니라 CPU의 몫이다. 에이전트 시대에는 '극초저지연(Low Latency)'이 서비스의 생사를 가르는 절대 기준이 된다.

GPU의 근본 설계 철학은 SIMT(Single Instruction, Multiple Threads)다. 거대한 중세 시대 선박 밑에서 1만 명의 선원들(코어)가 북소리(제어 장치)에 맞춰 정확히 동일한 동작(행렬 곱셈)을 반복하는 구조다. 이러한 단순한 병렬 연산에는 지구상에 GPU의 적수가 없다. 엔비디아가 AI 시장을 지배한 것도 SIMT 구조 덕분이었다. CPU의 코어가 무거운 하드웨어 스레드 2개를 관리하며 낑낑댈 때, GPU는 코어 하나에 수십 개의 극도로 가벼운 하드웨어 스레드를 만들어내고, 한 번의 명령으로 수십 개의 스레드가 동시에 똑같은 연산을 수행하게 했다. 복잡한 논리 제어를 포기하는 대신, 물량 공세로 AI 연산을 처리한 것이다.

하지만 에이전트 시대가 성숙해지며 상황이 다시 역전되고 있다. 바이브 코딩이 대세가 되면서 워크로드의 본질이 순차적 논리 구조(Sequential Logic)로 변한 것이다. 에이전트가 "내일 출장 일정을 짜고, 비행기를 예약한 뒤, 관련 자료를 요약해서 팀원에게 메일로 보내라"는 명령을 수행하기 위해서는 수십 개의 하위 에이전트를 소환하고, 파이썬 코드를 컴파일하며, 웹을 크롤링해야 한다. 이러한 작업은 "조건 A가 맞으면 B를 실행하고, 틀리면 C를 실행하라"는 수많은 분기(If-Then-Else)의 연속이다. 반면, GPU의 제어 장치는 1만 개의 코어에 각기 다른 조건문을 하달할 능력이 없다. 조건문이 엇갈리면 경로가 나뉘고, 특정 경로의 코어들은 연산을 멈추고 대기해야 한다. 이를 반도체 공학에서 워프 다이버전스(Warp Divergence)라 부른다. 1만 개의 코어 중 9,900개가 명령을 받지 못해 놀고 있는 '파이프라인 스톨(Stall)' 현상이 발생한다. 즉, 뭘 해야 할지 몰라 갈피를 잡지 못한다는 말이고, SIMT의 GPU 효율은 수직 낙하한다.

*** 소프트웨어 스레드:**
소프트웨어가 "이 작업은 이렇게 처리하자"라고 스스로 정의하고 관리하는 논리적 흐름이다. 예를 들어, "출장 일정 짜고 메일 보내" 같은 사용자 요청을 받은 AI 에이전트는 내부 소프트웨어(코드)가 이 큰 작업을 논리적으로 쪼개서 계획한다. "비행기 예약 → 호텔 확인 → 자료 요약 → 팀원 메일 발송"처럼 단계별로 정확한 순서와 역할을 정한다. 즉, 에이전트가 수십 개 하위 에이전트를 동시에 띄울 때, 이 소프트웨어 스레드가 폭발적으로 늘어난다.

결국 이 복잡한 분기 제어 예측 처리하기 위해, 거대한 캐시를 장착한 막강한 단일 스레드 성능의 CPU(Head Node)가 귀환한 것이다. 분기 예측(Branch Prediction)이란, If-Else 갈림길에서 다음 경로를 미리 예측하는 기술이다. 에이전트 워크로드에서는 이 분기 판단이 초당 수백만 번 발생하기 때문에, 분기 예측의 정확도가 곧 시스템의 성능을 결정한다.

그러나 과거처럼 SMT(동시 멀티스레딩 혹은 하이퍼스레드: 하나의 코어가 여러 스레드를 동시에 처리하는 기술)를 통해, 코어 내부를 복잡하게 꼬아놓지는 않는 게 트렌드가 되고 있다. Arm의 피닉스 프로젝트나 구글의 자체 CPU인 액시온에서 보듯, 물리 코어를 그냥 128개, 192개씩 많이 탑재하는 방식이다. 대신에 이 코어들 사이를 거대한 내부 고속도로(Mesh Network)로 연결하여 소프트웨어 스레드들을 물리 코어에 1대1로 직결시킨다. 즉, 그냥 물리 코어를 많이 늘려서 병렬로 처리하는 '스케일 아웃(Scale-out) 병렬화' 모델로 진화하고 있다. 마치 CPU가 GPU의 “대량 병렬” DNA를 수혈받은 셈이다. 주요 빅테크들은 각자의 방식으로 "CPU에 캐시메모리를 더 탑재하고, 성능 예측 가능성 및 보안 안정성을 위해서 SMT까지 없애면서", CPU의 대대적인 귀환을 맞이하고 있다.

*** SMT(동시 멀티스레딩 / 하이퍼스레딩):**
하나의 물리 코어가 내부 자원을 시간으로 쪼개서 두 개 이상의 소프트웨어 스레드를 동시에 처리하는 기술이다. 처리량은 20~30% 올라가지만, 두 스레드가 같은 L1 캐시와 실행 유닛을 공유하기 때문에 간섭이 발생한다. 에이전트 시대에는 이 간섭이 예측 불가능한 지연(Jitter)과 보안 구멍(Spectre)을 만들어내 치명적이다. 그래서 최근 트렌드는 SMT를 꺼버리고 코어 내부를 단순하게 유지하는 방향으로 바뀌는 중이다.

(2) 오케스트레이션 텍스(Orchestration Tax)와 "매니코어" CPU

GPU가 "대한항공", "오전 9시" 같은 토큰들을 생성하는 동안, 칩 외부에서는 끔찍한 연산 부하가 발생한다. 이를 '오케스트레이션 텍스'라 칭한다. 수천만 명의 접속자가 각각 요구하는 방대한 문맥(Context)을 잃지 않고 관리하는 상태 관리, 그에 따른 올바른 임무 수행 및 시뮬레이션, 이후의 검증 및 평가, 또한 AI가 작성한 악성 코드가 서버를 파괴하지 못하도록 가상 머신(VM)이라는 격리된 샌드박스 방을 수만 개 생성하고 폐기하는 보안 작업, 마지막으로 외부 API 호출 시 발생하는 SSL/TLS 암호화 및 복호화는 100% CPU에 종속되는(CPU-bound) 워크로드다.

이는 운영체제(OS) 커널 단에서 이루어지는 철저한 '순차적' 논리 연산이다. GPU에 아무리 많은 텐서 코어를 박아 넣어도 이런 행정 업무를 대신할 수는 없는 노릇이다. 게다가 시장이 놓치고 있는 승수 효과가 하나가 더 있다. 지난 AI Bi-weekly에서 짚은 주제인, xAI의 Grok 4.20이 던진 'Hydra 아키텍처'가 그 장본인이다. 이 구조에서는 한 모델 안에서 4~16개의 독립된 자아(페르소나)가 동시에 깨어나 계획을 세우고, 서로의 결론을 실시간 검증하는 구조를 가진다. Anthropic도 Opus 4.6에서 에이전트 팀을 공식 출시하기도 했다.

*** 물리 코어 (Physical Core):**
실제 하드웨어로 존재하는 하나의 독립된 연산 장치다. SMT를 쓰지 않으면 하나의 코어가 하나의 소프트웨어 스레드만 전담하게 된다. Arm 피닉스나 구글 액시온처럼 128개, 192개씩 대량으로 탑재하는 이유는 예측 가능한 성능을 극대화하기 위함이다.

멀티 에이전트 팬아웃(Fan-out) 효과에다가 멀티 페르소나 구조까지 더해지면서, CPU가 수행해야 할 데이터센터 프론트엔드의 트래픽은 마비 직전에 이를 수 있다. 에이전트 간 오케스트레이션이라는 CPU-bound 워크로드에 '모델 내부의 자아 수'라는 승수가 곱해져, CPU 수요는 기존 예측보다 한 단계 더 올라간다는 말이다.

기존 넷플릭스 영상이나 스트리밍해주던 4~5년 전의 인텔의 구형 Xeon CPU 캐시와 I/O 대역폭으로는 에이전트 시대에 벌어질 컨텍스트 스위칭(Context Switching)을 절대로 감당할 수 없다. 그래서 빅테크들은 물리 코어를 128~192개씩 늘리고 거대한 캐시를 박아 “스위칭 비용을 최소화”하는 Scale-out CPU로 자체 개발 칩에 전력을 다 하는 중이다.

*** 컨텍스트 스위칭:**
CPU가 지금 돌리고 있는 작업 A(예: 에이전트 1번이 비행기 예약 중)를 잠시 멈추고, 작업 B(에이전트 2번이 메일 작성 중)로 순식간에 넘어가는 과정이다. 이때 CPU는 A의 모든 상태(현재 어디까지 했는지, 메모리에 어떤 데이터가 올라와 있는지 등)를 메모리에 저장했다가, B의 상태를 다시 불러와야 한다. 에이전트 시대에는 수만 개의 에이전트가 동시에 움직이면서 이 스위칭이 초당 수십만 번 발생한다. 구형 Xeon CPU는 캐시와 I/O 대역폭이 너무 작아서 이 과정을 감당 못 하고 시스템 전체가 느려지거나 마비된다.

(3) 아키텍처의 해체: 모놀리식의 붕괴와 물리적 분리

현재 AI 데이터센터의 표준인 엔비디아 HGX 보드는 1~2개의 CPU 주변에 8개의 GPU가 옹기종기 모여 있는 구조다. 이는 학습(Training) 환경에 맞춰진 1:4 혹은 1:8의 모놀리식 아키텍처다. 그러나 추론(Inference) 및 에이전트 환경에서 이 1U 서버 박스는 열역학적, 경제적 파산을 의미할 수 있다. CPU가 처리해야 할 로직 연산량이 수백 배 늘어났음에도, 박스 안에는 CPU를 추가할 소켓도 열을 식힐 쿨링 용량도 없기 때문이다.

따라서 하이퍼스케일러들의 선택은, 랙 전체를 GPU와 HBM으로만 꽉 채운 가속기 랙(Compute Plane)을 만들고 그 옆에 128코어 이상의 고밀도 CPU로만 가득 채운 오케스트레이션 전용 CPU 랙(Control Plane)을 세울 것으로 전망된다. 그리고 이 두 거대한 랙을 광케이블 기반 통신이나 CXL(Compute Express Link, CPU와 가속기 사이의 메모리를 일관되게 공유하는 차세대 인터커넥트 표준)로 묶을 것이다. CPU 랙(지휘관)이 웹 검색, 논리 판단, 보안 통제를 모조리 수행하고, 고밀도의 병렬 계산이 필요할 때 인터커넥트된 옆 GPU 랙(계산기)에 외주를 주는 구조가 될 가능성이 높다. 심지어는 최근 순수 추론용 데이터센터에서는 CPU의 숫자가 GPU를 압도할 것이라는 대담한 예측까지 나오고 있다.

(4) 업자들이 직접 증언하는 CPU의 부상

AMD의 CEO 리사 수는 "전통적 컴퓨팅(CPU)과 가속 컴퓨팅(GPU)의 균형 잡힌 비율이 필요하다"고 최근 말했다. 이 발언은 우리가 이야기하는 사태의 본질을 꿰뚫는다고 생각한다. 에이전트 트래픽이 몰려오자 기존의 구형 CPU(레거시 인프라)들은 거대한 병목 현상을 일으키고 있고, 리사 수는 일찍이 올해 초 "20% 이상의 CPU 성장률"을 암시하며 자신감을 내비친 적이 있고 최근의 동향을 보면 그 예측치는 상향되고 있다.

인텔의 CFO 역시 데이터센터 CPU 사업부의 반등 가능성을 시사했다. 마이크로소프트가 Maia 200 가속기의 헤드 노드로 인텔 Granite Rapids를 사용한다는 사실은 CXL 기능에서 x86이 아직 Arm보다 안정적이라는 판단을 반영하기도 한다. 반면, Arm의 어닝콜에서 르네 하스 CEO는 "에이전트들이 서로 대화하거나 워크플로우를 제어하는 작업은 CPU(Arm/RISC)의 전력 효율성이 뛰어나고 지연 시간이 짧기 때문에 CPU에 완벽하게 부합한다"면서, 데이터센터 로열티 매출이 전년 동기 대비 100% 이상 성장했다고 밝혔다.

CISC든 RISC든, 이들의 모든 증언은 일단 같은 방향을 가리킨다: CPU는 독자적인 컴포넌트로서, 더 이상 조연이 아니라 주연급으로 재평가받고 있다는 점이다.

3. CPU 전쟁: "AMD vs 인텔", CSP 현황

(1) AMD vs 인텔: 칩렛과 메시의 20년 아키텍처 전쟁

앞서 설명한대로, 극초저지연(Low Latency)과 예측 가능한 성능이 에이전트 서비스의 생사를 가르는 절대 척도다. 그런데 CPU의 코어가 많아질수록 칩 내부에서 데이터를 실어 나르는 도로망(인터커넥트)의 설계가 성능의 거의 전부처럼 된다. 이 지점에서 x86 진영의 두 거인, 인텔과 AMD는 각기 다른 물리적 딜레마에 봉착했다.

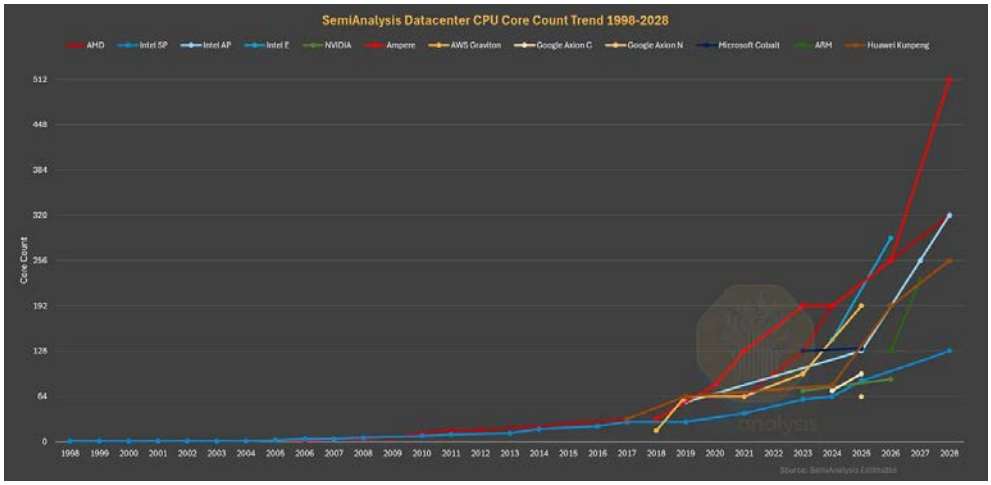
인텔의 모놀리식 한계

과거 인텔은 코어 간 통신을 위해 링 버스(Ring Bus)를 사용했으나, 코어 수가 24개를 넘어가자 통신 지연 시간이 급증했다. 이에 바둑판 모양으로 코어를 배열하는 메시(Mesh) 아키텍처를 전격 도입했다. 하지만 코어 자체의 물리적 크기가 커지면서 심각한 문제에 부딪혔다. Ice Lake에서 40코어 이상을 바둑판에 우겨넣으려다 보니 실리콘 웨이퍼를 굽는 노광장비의 최대 사진 크기인 '레티클'이 한계(26x33mm)에 도달해 버린 것이다. 단일 다이(Monolithic) 크기를 더 이상 물리적으로 키울 수 없는 거대한 벽 앞, 인텔의 혁신은 정체되었다. 차세대 Clearwater Forest는 "18A 공정 + Foveros Direct"를 적용했으나, 출시가 2026년 상반기로 연기됐고 시에라 포레스트 대비 동일 코어 기준 성능 향상은 겨우 17%에 그쳤다. 인텔 내부에서도 Foveros Direct 수율 검증용 Vehicle로 활용하려는 의도로 해석되니, 사지 말자는 여론이 있을 정도다.

AMD의 칩렛의 승리와 기술 부채

한편, AMD는 이러한 단일 다이의 한계를 칩렛(Chiplet) 구조로 돌파했다. 작고 수율 좋은 8코어짜리 다이를 여러 개 구워 기판 위에 레고 블록처럼 이어 붙이는 이종 분산 설계를 활용했다. 그리고 AMD의 칩렛 설계 전략은 CPU 시장에서 구조적 우위를 점하고 있다.

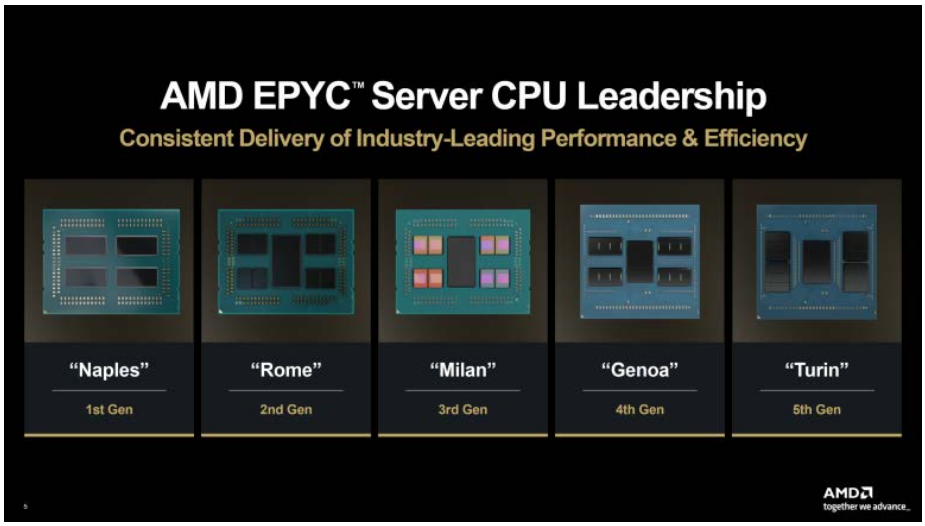
그림 4. 데이터센터 CPU별 코어 수 추세
2020년대 중반을 기점으로 데이터센터 CPU 시장이 완만한 클럭 속도 발전의 시대를 끝내고, 무자비한 숫자의 '코어 수(Core Count) 극대화 경쟁'으로 패러다임이 완전히 전환
붉은색 선의 AMD가 2028년 500코어 이상을 향해 수직 상승하는 궤적, 칩렛의 우위를 보여줌



자료: SemiAnalysis, 미래에셋증권 리서치센터

2017년 EPYC Naples에서 MCM(Multi-Chip Module)으로 데이터센터 CPU 시장에 복귀한 이후, Rome(2019년)에서 중앙집중식 I/O 다이를 도입하고, Milan(2021년)에서 CCX(Core Complex, 코어들을 하나의 단위로 묶은 모듈)를 8코어로 통합하여 통신 병목을 없앴다. Genoa(2022년)에서는 12개 CCD로 96코어, Turin(2024년)에서 16개 CCD로 128코어까지 확장했다.

그림 5. AMD EPYC CPU 세대를 보면, Rome부터 중간에 없던 다이가 생겼다… 이게 I/O



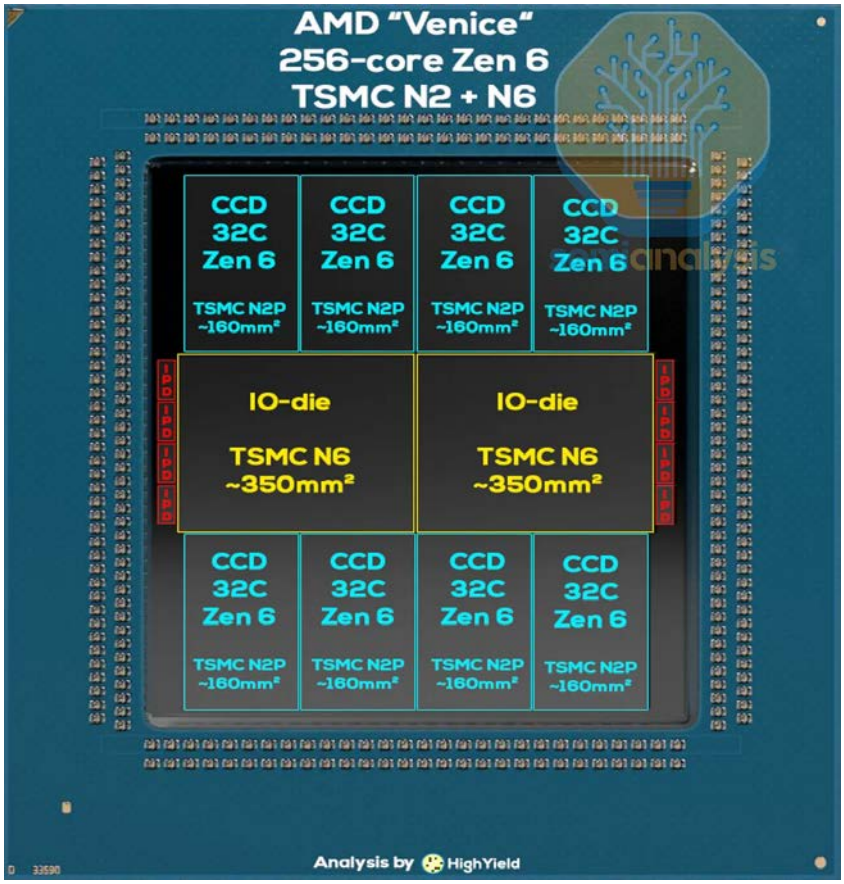
자료: AMD, 미래에셋증권 리서치센터

CCD만 설계하면 몇 개를 붙이느냐에 따라 다양한 코어 수의 제품을 만들 수 있어, 웨이퍼 수율과 제조 비용에서 인텔의 모놀리식 설계를 구조적으로 압도했다. 2026년 출시될 베니스(Venice)는 최대 256코어로 이 격차를 더 벌릴 수도 있다.

그러나 여기에 치명적인 아킬레스건도 있다. 256개로 코어 수가 너무 많아져 중앙의 I/O(입출력) 다이를 하나로 감당할 수 없게 되자, 이를 두 개로 쪼개버린 것이다. 칩렛 기반의 설계에서 중앙 I/O가 분할되면 칩 내부에 거대한 브릿지(Link)가 생긴다. A 코어가 반대편 I/O 다이에 연결된 메모리를 가져오려면 칩 내부의 다리를 건너가야 한다는 말이다. 물리적 위치에 따라 데이터 접근 속도가 달라지는 이 현상을 NUMA(Non-Uniform Memory Access)라고 한다.

NUMA에서 가장 큰 문제는 이로 인해 발생하는 Jitter(예측 불가능한 미세 지연)다. 에이전트가 코드를 실행할 때, 로컬 메모리에서 0.01초 만에 끝나던 작업이 반대편의 원격 메모리를 경유해야 한다면 0.05초가 걸린다. 0.01초~0.05초의 미세한 지연과 예측 가능성 하락은, CPU 코어가 많아질 수록 데이터 파이프라인 타이밍을 뒤틀어버리며 결국 전체 시스템을 덜컹거리게 만든다. AMD는 칩렛의 원가 경제성에 취해 가장 중요한 단일 스레드 " 통신일관성"을 기술 부채로서 담보 잡힌 것이다.

그림 6. AMD의 "Venice(베니스)" 25 6코어 아키텍처 다이샷
코어 확장을 감당하기 위해 기존의 단일 중앙 통제소를 포기하고,
가운데 노란색 I/O 다이를 두 개(350mm² x 2)로 쪼개어 배치한 것이 맹점
코어가 반대편 구역의 데이터를 가져오기 위해 내부 다리를 건너야만 하는 NUMA와 Jitter를 유발
분기 예측과 극초저지연(Low Latency) 동기화가 생명인 '에이전트 및 강화학습' 워크로드 환경에서,
코어 물량(Throughput)을 얻기 위해 AI 파이프라인의 심장인 '통신 일관성'이 담보 잡히게 된 것.



자료: SemiAnalysis, 미래에셋증권 리서치센터

인텔의 승부수는 너무 늦었다

이 타이밍에, 2026년 인텔은 다이아몬드 래피즈에서 자존심을 버리고 AMD의 중앙 집중식 칩렛 구조를 카피했다. 그리고 AMD와 달리 기판 바닥에 미세한 구리선(Long Trace)을 촘촘히 깔아 I/O 다이 분할에 따른 NUMA 지연을 방어하는 엔지니어링 정수도 보여주었다.

그러나 여기서 뼈아픈 극단적 결정을 내린다. 본인들의 자랑이었던 하드웨어 스레드 기술인 SMT를 영구적으로 삭제한 것이다. SMT는 하나의 물리 코어 안의 실행 유닛(ALU)이 데이터를 기다리는 동안 잉여 자원을 활용해 두 번째 소프트웨어 스레드를 욱여넣는 기술이다. 처리량(Throughput)을 20~30% 향상시키지만, 클라우드 가상화 환경에서는 보안상 불리하다. 두 스레드가 물리적으로 동일한 L1 캐시를 공유하기 때문에, 해커가 미세한 전력 파동과 캐시 접근 시간 차이를 역산하여 암호화 키를 탈취하는 Spectre(보안 해킹) 취약점이 발생한다. 또한 한 스레드가 대용량 쓰레기 데이터를 캐시에 올려놓으면, 이웃 스레드의 성능이 폭락하는 시끄러운 이웃(Noisy Neighbor) 문제가 빈발한다.

클라우드 사업자에게 가장 중요한 것은 최대 스루풋이 아니라 예측 가능한 성능(Predictability)이다. 그래서 과감히 SMT를 영구 삭제했다. 192개 온전한 물리 코어(1코어=1스레드)만 남긴 건, 클라우드 사업자들이 “이제는 안정적인 CPU를 원한다”는 신호를 가장 먼저 포착한 선택이었다.

하지만 이 선택이 현재로서는 패착으로 보인다. 인텔은 미래를 읽었지만, 실행 속도에서 완전히 뒤쳐졌기 때문이다. monolithic(단일 다이) 구조에 매달리다 레티클 한계에 막혀 늦게 칩렛으로 넘어간 후유증을 겪고 있는 반면, AMD는 이미 256코어를 준비한다. AMD 베니스의 256코어 512스레드 물량 공세 앞에서 최대 처리량 열세는 피할 수 없을 것이다. 게다가 인텔은 수익의 척추인 보급형 시장(SP 플랫폼)인 8채널 라인업마저 취소하며 볼륨 마켓을 AMD에게 통째로 헌납했다.

(2) 최상위 포식자인 하이퍼스케일러(AWS, Microsoft, Google) 동향

인텔과 AMD가 x86 생태계 안에서 싸우는 동안, 클라우드 사업자들은 룰 자체를 다시 쓰고 있다. 그들에게 상용 x86 CPU는 비싸고, 무겁고, 불필요한 레거시 명령어 세트로 가득 찬 기성품이다. 데이터센터의 전력 예산(Power Budget)은 한정되어 있다. GPU 클러스터가 MW, 곧 GW 단위로 전력을 흡수하는 상황에서, 인터넷 웹 서빙이나 에이전트 오케스트레이션을 담당하는 범용 서버에까지 전기를 낭비할 수 없다. 하이퍼스케일러들은 단일 코어의 절대 성능보다는, 클럭 속도가 다소 느리더라도 전력을 극도로 적게 먹는 코어를 빽빽하게 집어넣은 면적 및 전력 효율적 설계를 원했다.

그림 7. 마이크로소프트가 짓고 있는 초대규모 AI 데이터센터 "Fairwater"
에이전트 오케스트레이션과 강화학습(RL) 트래픽을 통제하기 위해 별도의 CPU 전용 건물이 48MW 규모로 거대하게 지어졌다는 사실
'295MW GPU 건물'과 '48MW CPU 건물'의 전력 불균형(약 6대 1)은 데이터센터 전력 예산이 한정되어 있기 때문
48MW의 공간을 레거시 x86 CPU로 채웠다면, 에이전트의 연산 병목을 감당하기 위해 CPU 건물도 100MW 이상으로 커져야 했을 것.

자료: 구글 어스, SemiAnalysis, 미래에셋증권 리서치센터

이들은 인텔이나 AMD에게 마진을 지불하는 대신, 자사 데이터센터 환경에 완벽히 최적화된 Arm 기반 맞춤형 CPU를 직접 깎아내기 시작했다. AWS의 그래비톤(Graviton), 구글의 액시온(Axion), 마이크로소프트의 코발트(Cobalt)가 그 결과물이다. 이 칩들은 x86의 무거운 구조를 버리고 클라우드 가상화와 자체 네트워크 인프라에 맞춰 설계되었다.

이들 CSP 자체 실리콘의 공통된 기술적 특징은 I/O 레인의 맞춤화와 거대한 캐시 메모리다. 이는 에이전트 워크로드의 특성 때문이다. 에이전트는 끊임없이 분기(If-Then)를 판단해야 한다. 수만 개의 에이전트 상태 정보(Context)를 지연 시간 없이 초고속 캐시에 올려두고, 즉각적으로 다음 행동을 계산하기 위해 Cache의 크기를 극한으로 키운 것이다. 물리적 다이 면적을 희생하더라도 메모리 장벽에 부딪혀 DRAM까지 데이터를 가지러 가는 지연 시간을 원천 차단하겠다는 의지다.

AWS Graviton

AWS는 클라우드 벤더 중 CPU 설계의 내재화를 가장 완벽하게 달성한 기업이다. 구글이 가속기에 집중할 때 아마존은 CPU에 많은 세월 공을 들였기 때문이다. Graviton의 무기는 최고 성능이 아니라, 자체적인 사용처를 두고, 규모의 경제를 활용한 '극단적인 단위 비용 절감(TCO)'이다. Graviton5에서 가장 주목해야 할 기술적 인사이트는 PCIe Gen6를 채택하면서도 레인 수를 96개에서 64개로 줄였다는 점이다. 상용 칩 벤더는 고객이 어떤 장치를 얼마나 연결할지 모르기 때문에 I/O 레인을 넉넉히 탑재(Over-provisioning)해야 한다.

하지만 AWS는 자사 데이터센터 내부의 트래픽 데이터와 고객의 실제 사용량 데이터를 완벽하게 통제하고 분석한다. 불필요한 I/O 레인을 제거하면 실리콘 다이 면적을 줄일 수 있고, 전력 누수를 차단하며, 웨이퍼당 생산 칩 수를 늘려 칩 단가를 드라마틱하게 낮출 수 있다. 풀스택을 가진 자만이 할 수 있는 궁극의 최적화다.

그리고 Graviton은 L3 캐시를 36MB에서 192MB로 5배 이상 늘렸다. 192개 코어가 메모리에 동시 접근할 때 발생하는 병목을 실리콘 내부에서 흡수하려는 의도다. 또한 Trainium3 AI 가속기의 헤드 노드로도 Graviton을 활용함으로써, 에이전트 시대에 필요한 실리콘들을 착착 준비하고 있다.

MS 코발트와 구글 액시온

MS는 코발트 200에서 코어 수를 132개로 억제하는 대신 개별 코어의 성능을 극대화하는 Arm의 Neoverse V3를 채택했다. 그리고 코어당 3MB라는 압도적인 L2 캐시를 할당한 것은, 클라우드 가상화 환경에서 여러 테넌트(컴퓨팅 임대 고객)가 자원을 공유할 때 발생하는 'Noisy Neighbor' 문제를 하드웨어 단에서 격리하기 위함이다. 주목할 점은 코발트를 범용 Azure 컴퓨팅에만 투입하고, AI 가속기(Maia 200)의 헤드 노드로는 인텔 Granite Rapids를 사용한다는 사실이다. CXL 등 AI 헤드 노드에 필수적인 메모리 일관성 기능에서, 아직 본인들의 칩보다는 x86 칩의 안정성이 우위에 있다는 판단으로 해석된다.

한편, 구글은 워크로드의 성격에 따라 라인업을 완전히 분리했다. C4A(퍼포먼스용)는 Arm의 Neoverse V2 72코어를 사용해서 9x9 메시 네트워크를 적용하고, N4A(효율성용)는 Arm Neoverse N3 64코어를 활용해 저전력으로 처리하는 형태다. 구글 역시 향후 Gemini를 구동하는 TPU 클러스터의 헤드 노드로 Axion CPU를 투입할 예정이며, x86일 때 발생하던 컨텍스트 스위칭 오버헤드를 없애고 독자적인 데이터 패브릭 안에서 병목 없는 AI 인프라를 완성하려는 궁극적 목표를 향해 가고 있다. MS보다는 아마존과 구글의 실리콘 역량이 더 높다는 방증으로 해석된다.

무엇보다 구글의 해당 노력이 무서운 것은 풀스택 수직 통합이다. 구글은 자체 가속기인 TPU를 지휘하는 헤드 노드 자리에 자신들이 만든 액시온 CPU를 꽂아 넣었을 뿐만 아니라, CPU와 TPU가 같은 구글의 통신 규약으로 대화하게 했다. 즉, 인텔이나 AMD 칩이 중간에 끼어 발생하던 일종의 "번역 비용"이 완벽히 증발한 것이다.

Arm 피닉스: IP 벤더의 반란과 채널 충돌

x86의 반대편에 있는, Arm은 언뜻 보면 가장 좋은 위치에 서 있는 듯 보인다. 빅테크들이 모두 RISC 기반 설계로 넘어가는 노력을 하고 있기 때문이다.

그런데 그들은 2026년 "프로젝트 피닉스(Phoenix)"를 통해 아예 코어부터 패키징, TSMC 생산까지 모두 마친 '완제품 칩'을 공급하겠다는 놀라운 선언을 했다. 128개의 Neoverse V3 코어, TSMC 3nm 공정, 12채널 DDR5를 담았다. 그리고 첫 번째 고객으로 메타를 확보한 것은 상징적이다. 메타는 자사 전용 AI 가속기인 "MTIA"를 개발 중이긴 하지만 CPU 설계 역량은 타 클라우드 벤더보다 분명히 떨어진다. Arm은 자체 설계 여력이 부족하거나, 빠른 시간 내에 최고 성능의 AI 헤드 노드가 필요한 기업을 직접 공략하며 생태계의 지배력을 물리적 칩의 영역으로 확장하려 하고 있다.

그러나 Arm의 가장 큰 고객이 Neoverse의 지적재산권(IP)을 사다가 맞춤형 칩을 만드는 AWS, MS, Google이다. Arm이 완제품 피닉스를 들고 시장에 나오면, 이들 대형 고객사들과 직접적인 경쟁 관계에 놓이게 된다. 장기적으로 RISC-V 등 개방형 아키텍처로의 이탈을 촉발할 위험성도 내포하고 있다. x86 진영과 Arm 진영에서 기업들간 입지별 다이버전스가 본격화될 가능성이 높다고 사료된다.

4. 잔혹한 엔비디아의 테제

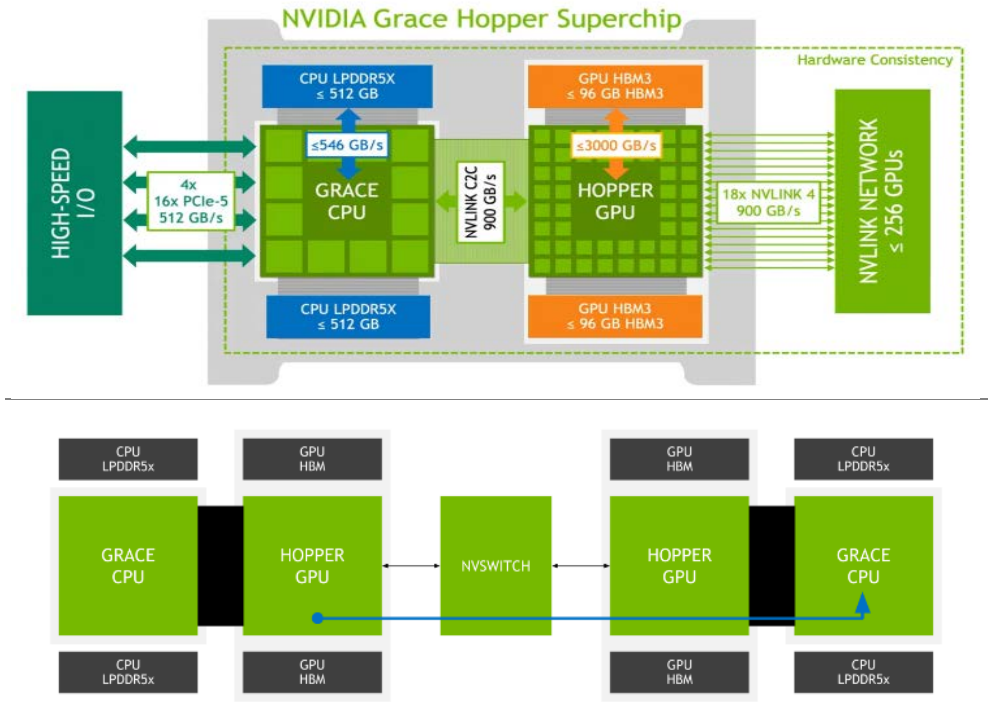
하이퍼스케일러들이 자체 CPU와 ASIC으로 생태계의 주도권을 장악하려 하자, GPU의 제왕 엔비디아는 영리한 방어전을 펼치고 있다.

(1) Grace: 구조적 타협과 병목의 발견

원래 엔비디아에게 CPU는 독립적인 연산 장치가 아니었다. GPU가 겪는 메모리 용량 및 대역폭의 한계(Memory Wall)를 우회하기 위한 전략적 보조 장치 정도로 보였다. Grace CPU는 엔비디아가 범용 CPU 시장이 아닌, 철저히 자사 GPU 생태계 강화를 위해 만든 첫 번째 이정표였다. Grace의 메모리 아키텍처에는 흥미로운 특이점이 있다. 서버용 CPU임에도 불구하고 모바일급인 LPDDR5X 메모리를 채택한 점이다. 극단적으로 전력 소모를 줄이면서도 512비트 버스를 통해 500GB/s의 대역폭을 확보하기 위함이었고, 남는 Power Budget을 모두 GPU에 몰아주겠다는 설계 철학이 반영된 결과다.

NVLink-C2C(양방향 900GB/s)는 CPU와 GPU를 물리적으로 묶는 것을 넘어 메모리 공간을 논리적으로 통합한 것이었다. HBM을 무한정 늘릴 수 없는 GPU의 약점을, CPU에 연결된 최대 480GB의 값싼 LPDDR 메모리를 끌어다 쓰는 방식으로 해결하려 한 것이다. 사실 NVLink-C2C는 GTC 2024에서 처음 크게 강조됐던 내용으로, 행사에 참여했던 필자도 이 때는 메모리 공간을 통합했다는 의미의 크기를 잘 알지 못했다. 그러나 지금 와 생각해보면 메모리의 제약과 풀이법을 엔비디아가 선제적으로 고민했다는 것을 의미한다. 이는 곧 이후에 기술에 이 에이전트 시대의 공급망을 엔비디아가 선점한 이유로 이어진다.

그림 8. Grace CPU와 Hopper GPU를 NVLink-C2C로 묶어서, 수억 명의 에이전트 상태 정보(Context)를 처리할 때, 거대한 일관된 메모리 풀로써 작동하게 만들 수 있는 가능성을 열었다



자료: 엔비디아, 미래에셋증권 리서치센터

그런데 Grace에도 치명적 병목이 발견됐다. Grace가 특정 코드에서 분기 예측기(Branch Predictor)의 한계로 성능이 저하된다는 점이 지적됐다. 이는 Arm의 Neoverse V2 코어가 가진 한계였다. 명령어를 2MB 단위의 32개 영역으로 관리하는데, 32개를 초과하는 복잡한 분기 명령이 들어오면 64MB 버퍼가 완전히 초기화되었던 것이다. 데이터 파이프라인이 텅 비어버리면서 연산 장치(ALU)가 유휴 상태에 빠지는 현상이다. 이것이 2026년 현재 GB200, GB300 시스템에서 AI 워크로드 처리 속도를 갉아먹는 주범으로 작용하고 있다.

(2) Vera: 컨트롤 플레인 탈환을 위한 자체 노력

그레이스에서 발견된 아키텍처의 한계를 보고, 최적화의 부재가 가져온 결과로 해석한 엔비디아는 2026년 베라(Vera) CPU를 통해 Arm의 기성품(Neoverse)을 버리고 8년 만에 자체 코어 설계(Olympus)로 돌아섰다. 올림푸스 코어의 핵심 변경 사항은, Neoverse V2의 분기 예측 실패 및 연산 병목을 해결하기 위해 부동 소수점 장치(FPU) 포트 폭을 4개에서 6개로 확장한 점이다. 이는 코어 하나당 명령어 처리량을 끌어올리는 '물리적' 변화다. 다른 물리적 변화로 코어당 2MB의 전용 L2 캐시를 탑재하여 캐시 미스(Cache Miss, CPU가 필요한 데이터를 캐시에서 찾지 못하고 더 느린 메모리까지 가야 하는 현상) 확률을 대폭 낮췄다.

그리고 최대 88개의 활성 코어에 SMT를 적용하여 176스레드를 처리하게 만든 것은, 대량의 데이터 입출력(I/O) 요청을 지연 없이 병렬로 처리하기 위한 조치다. SMT를 쓴다는 것은 우리가 앞서 이야기했던 "SMT-off"의 트렌드에 역행하는 듯 보이지만, 엔비디아의 Olympus 코어에서 쓰이는 SMT는 전통적인 방식은 아니다. Spatial Multithreading으로, 즉, 코어 자원을 '물리적으로 분할'하는 방식이다. 그래서 88개의 코어가 각각 자신의 자원(실행 유닛, 캐시 일부)을 고정으로 2등분해서 176 스레드가 된다. 기존의 SMT처럼 하나의 물리 코어가 “시간을 쪼개서” 두 개의 소프트웨어 스레드를 번갈아 돌리는 방식이 아닌, 물리적 파티셔닝이기 때문에 각 스레드가 자기 전용 공간을 쓰기 때문에 스레드간 충돌이 원천 차단된다. 결국, 처리량보다 예측 가능성이 훨씬 더 중요한 에이전트 KV 캐시 오케스트레이션에 최적화돼 있다. 하이퍼스케일러(CSP)들은 SMT-off 흐름을 선택하고 있기에, "Vera-only 계약"을 따내기 위한 방향으로 해석된다.

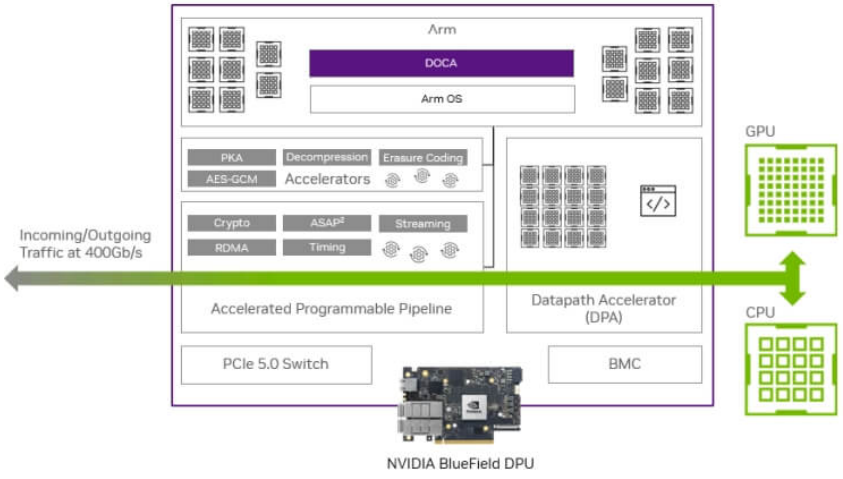
패키징 및 I/O도 진화했다. 3nm 공정의 컴퓨팅 다이 1개, 메모리 다이 4개, I/O 다이 1개로 칩렛 구조를 채택하고 CoWoS-R로 패키징했다. 128비트 SOCAMM 모듈 8개를 사용하여 1.5TB 용량과 1.2TB/s의 대역폭을 구현했으며, NVLink-C2C 역시 양방향 1.8TB/s로 두 배 확장됐다. Vera-Rubin 체제에서 엔비디아 시스템에서 CPU의 역할이 얼마나 올라갈 지 가늠할 수 있는 부분이다. 그리고 벌써 엔비디아는 베라 CPU를 에이전트 추론 전용 데이터센터의 지휘관 자리에 꽂아 넣기 위해, GPU 없이 CPU 단독(Vera-only) 계약까지 체결하고 있다. 메타는 아직 칩이 출시되기도 전인 2026년 초 시점에서, 2027년 Vera CPU의 대규모 선주문(Backlog)까지 확정 지었다. GPU를 팔아 제국을 건설한 기업이, CPU만 단독으로 공급하는 계약을 맺고 있는 것이다. 인프라 독립의 열망마저 꺾어버릴 만큼, 멀티에이전트 아키텍처가 요구하는 CPU-GPU-네트워킹의 유기적 통합은 철저한 하드웨어의 싸움이 된 것으로 풀이된다.

(3) 하드웨어 병목의 길목마다 선수를 치는 자

에이전트 시대에서 가장 중요한 것은 CPU나 GPU 단일 칩의 성능이라기 보다 CPU와 GPU의 매끄러운 인터커넥트(Interconnect, 칩과 칩 사이를 연결하는 고속 데이터 통로)라고 볼 수 있다. NVL72가 72개의 Blackwell GPU를 하나의 거대한 가상 GPU처럼 작동시킬 수 있는 것은, Grace라는 Arm CPU와 Blackwell GPU를 하나의 Superchip으로 붙이고, NVLink-C2C로 연결하여 CPU 메모리와 GPU 메모리를 완전히 공유하는 Unified Memory 덕분이다.

그런데 엔비디아는 이 하드웨어의 병목이 되는 길목마다 선수를 치고 있다. CPU(Grace→Vera), GPU(Hopper→Blackwell→Rubin), 네트워킹(Spectrum-X, Quantum InfiniBand), 그리고 DPU(Bluefield)까지 확보해 놓은 상황이다. 특히 5년 전에 세상에 모습을 드러냈던 DPU는 네트워킹과 보안 기능을 CPU로부터 오프로드시키기 위한 엔비디아의 선견지명이 담긴 제품이다.

그림 9. NVIDIA BlueField DPU 아키텍처
DPU(데이터 처리 장치)는 CPU를 힘들게 하는 막대한 네트워크 트래픽 통제, 보안 암호화/복호화, 데이터 압축 해제 등의 거대한 '오케스트레이션 텍스'를 전담하여 처리하는 프로세서
즉, 외부 데이터를 DPU가 전방에서 미리 떠안아(Offloading) 처리함으로써, 메인 CPU가 전력 낭비 없이 순수한 에이전트 논리 연산과 계산에만 100% 자원을 집중할 수 있게 만드는 "슈퍼서브"



자료: 엔비디아, 미래에셋증권 리서치센터

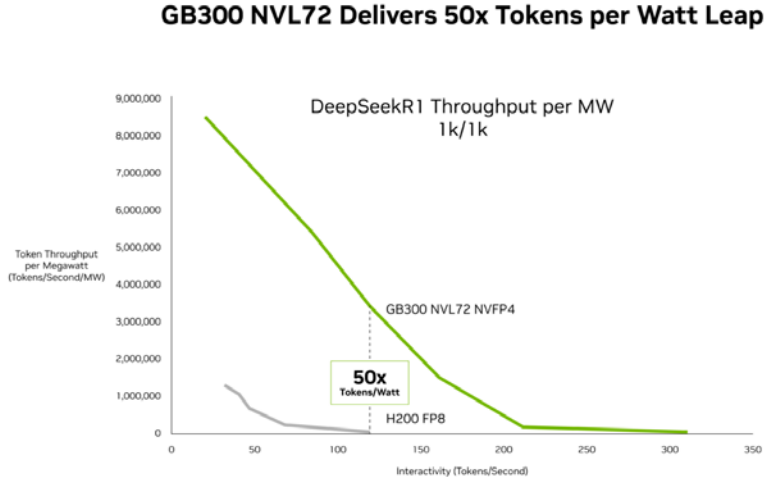
이번 자료에서는 DPU에 대해서 길게 다루지 않겠으나, CPU가 에이전트를 다루는 지휘자로서의 본업에 충실할 수 있도록 '데이터 전처리, API 호출, 암호화 및 복호화'와 같은 귀찮은 행정 업무를 DPU에 맡기기 위해 등장한 제품이다. 그리고 아직 DPU의 위력은 세상이 인지하지 못한 상황이다. 이유는 간단하다. 지금은 아직 "단일 에이전트" 시대라서 행정 부하가 폭발하지 않았기 때문이다. 하지만 멀티에이전트/멀티페르소나가 본격화되는 순간, DPU는 엔비디아의 숨겨진 두 번째 알파가 된다.

"나는 제약(병목)을 사랑합니다. 제약된 세상에서는 '최고'를 선택할 수밖에 없습니다"

젠슨 황은 3월 3일에 있었던 모건스탠리가 주최한 컨퍼런스에서 이같은 발언을 했다.

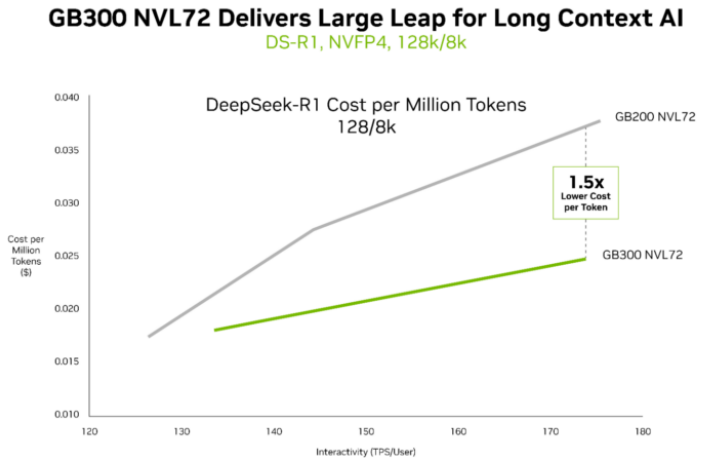
이는 물리적 제약을 무기화한 정수다. '칩도 전력도 극도로 제한된 현재 환경에서 가성비와 전성비(1와트당 가장 많은 토큰을 뽑아내는)가 제일 좋은 우리 시스템을 넣는 게 정배'라는 뉘앙스다.

그림 10. 구형 H200 대비 50배 폭증한 와트당 토큰 처리량
극초저지연이 생명인 에이전트 환경과 전문가 혼합(MoE) 모델인 DeepSeek R1의 라우팅 병목을 해결하고, 가성비를 앞세운 경쟁사 맞춤형 칩(ASIC)의 시장 진입 자체를 원천 봉쇄하고 있는 것. 신형 GB300 랙을 강제 구매할 수밖에 없도록 만드는 엔비디아의 초고속 인프라 교체 슈퍼사이클



자료: 엔비디아, 미래에셋증권 리서치센터

그림 11. GB300 NVL72가 기존 GB200 NVL72 대비 초장문 처리에서 비용을 1.5배 더 낮췄음
'멀티 에이전트' 워크로드에서 필연적으로 발생하는 치명적인 메모리 벽을 돌파 중이라는 지표. 수십만 줄의 저장소 코드를 디버깅하거나 방대한 금융 문서를 검증하는 B2B 엔터프라이즈 에이전트 서비스가 마침내 상업적으로 막대한 마진을 남길 수 있는 단위 경제성에서 앞서나가고 있다는 자신감

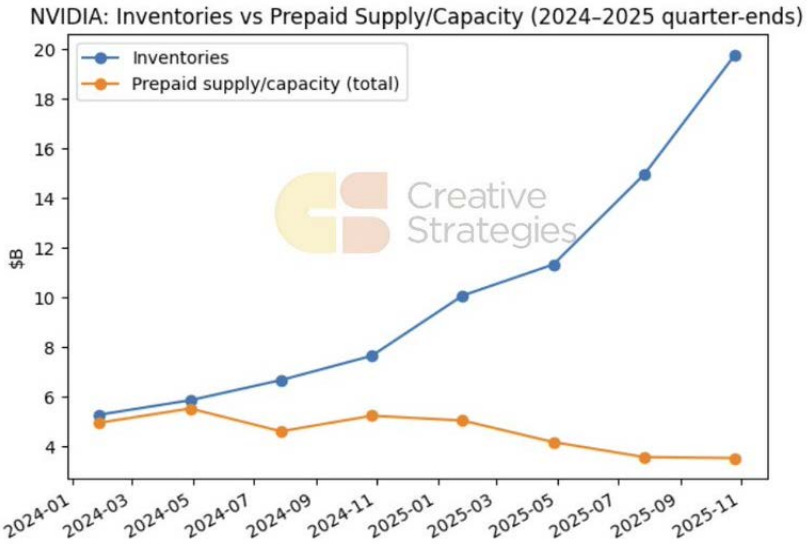


자료: 엔비디아, 미래에셋증권 리서치센터

젠슨 황이 이렇게 공급 병목에 자신만만한 이유는 엔비디아가 "메모리, 웨이퍼, CoWoS, 패키징, 커넥터, 케이블, 구리부터 MLCC까지 거의 모든 것을 선제적으로 확보"했기 때문이다. SK하이닉스가 2025~2026년 HBM 생산량의 대다수를 엔비디아에 보내고, Micron·삼성전자의 잔여분도 그들이 선점한다고 알려진다.

경쟁사(AMD, Intel, Broadcom, 커스텀 ASIC 업체)는 2026년에 HBM을 구하기 힘들거나 훨씬 비싸게 사야 한다는 말이다. 마치 iPod 시절 애플이 NAND 플래시를 다 쓸어 담은 것과 대동소이하다. 엔비디아의 재무제표에서 선급금(Prepaid)이 줄고 재고(Inventory)가 늘어나는 것이 일종의 증거로 보인다.

그림 12. 엔비디아는 2026년 전체, 2027년 일부까지 필요한 HBM을 다량 확보했다는 증거
prepaid 줄고 inventory 늘어나는 것이 예약한 HBM이 실제 생산되어 재고로 들어오는 중으로 해석



자료: Creative Strategies, 미래에셋증권 리서치센터

포인트는, 경쟁사가 아무리 좋은 칩을 설계해도 물리적으로 대량 양산 자체가 어렵게 생태계를 만들어버린 것이다. 반면, 본인들은 미리 확보한 계약들 덕분에, 랙 가격 인상 요인을 억제하고 자신들의 마진도 방어할 수 있다. 그리고 그 힘을 가지고, '우리 랙을 누구에게 줄까' 고민하면서 고객사 명단을 쭉 훑어 말 잘 듣는 기업들에게 최신 랙 물량을 밀어주는 잔혹한 통치술을 벌이고 있다. 실제로도, 엔비디아가 공개하는 최신 블로그 글의 고객사 명단에는 아마존과 구글은 없고 오라클과 코어위브 같은 기업이 최우선으로 등재되어 있기도 한 것에서 그 힌트를 얻을 수 있다.

5. Numbers & Alpha

에이전트 트래픽 증가 × 멀티에이전트(페르소나) 승수 → 순차적 CPU 워크로드 증가 → 레거시 CPU 붕괴 → 신형 CPU 강제 교체 수요. 이것이 CPU 슈퍼사이클의 작동 원리다. 시장이 이 이동을 아직 가격에 반영하지 못하고 있다면, 그것이 바로 알파다.

(1) 엔비디아 Data Center 세그먼트 모델

엔비디아의 Data Center 매출은 FY2025(1월 결산) 약 $1,150억 수준이었다. 이 중 GPU(학습+추론)가 약 85%, 네트워킹 약 10%, CPU/DPU 약 5%를 차지했다. 우리가 이야기했던 CPU의 귀환(엔비디아의 경우 DPU까지 포함)이 구조적으로 발생한다면 향후 매출 추정치는 크게 달라질 수 있다.

표 3. 엔비디아의 2027년에는 데이터센터 매출 비중이 어떻게 될까

세그먼트	FY2025 ($B)	비중	FY2027E ($B)	비중	YoY CAGR	설명
GPU (학습+추론)	~98	~85%	~196~210	~70%	~46%	GPU 성장 지속하나, 에이전트 오케스트레이션 수요 덕분에 CPU/DPU/네트워킹이 더 빠르게 성장.
CPU/DPU	~6	~5%	~42~45	~15%	~170%	Vera CPU standalone 계약이 2026년 하반기부터 본격 시작. Hydra 구조가 올해부터 더 빠르게 상용화되면 비중 최대 20%까지도 상단 열어둠.
네트워킹	~12	~10%	~42~45	~15%	~90%	네트워킹칩 부착률이 현재 90%에 육박, 멀티에이전트간 통신 대역폭으로서 중요도 상승
DC 합계	~115	100%	~280~300	100%	~57%	Rubin 사이클(NVL72→NVL144 전환) + Vera 주문(추론 수요 가속)이 매출 성장 동인.

자료: 미래에셋증권 리서치센터

GPU는 절대 금액 기준으로 여전히 가장 큰 성장을 보이지만, CPU/DPU와 네트워킹이 비중 기준으로 각각 15% 정도로 상승 가능성을 엿본다. Vera-only 계약이 메타에서 시작되어 다른 하이퍼스케일러로 확산될 경우, CPU/DPU만으로 60조원이 넘는 매출이 발생한다. 이것은 AMD의 전체 매출보다 큰 규모다. Grace 시절 CPU/DPU는 $6B 미만이었는데, Vera CPU의 Olympus 코어가 에이전트 지휘관 역할을 하면서 Vera-only가 본격화되면 가능한 규모로 사료된다. SemiAnalysis도 “CPU가 GPU 성장을 앞지를 수 있다”고 지적한 바 있다. 시장이 아직 이 CPU/DPU 구조적 상승을 가격에 반영하지 못하고 있다고 본다.

(2) AMD EPYC 세그먼트 모델

AMD 공식 가이던스는 Data Center 전체 60% 이상 성장이다. 지난 2월 실적발표에서도, CPU(Turin) 램프가 “기대 이상”이라고 언급했지만, 이는 아직 “20% 성장"이라는 프레임 안에서 말한 것이었다. GPU처럼 폭발적인 %가 아니라도, 물론 CPU 시장 TAM이 이미 200억 달러 넘기 때문에 20%만해도 40~50억 달러 추가 매출이라 충분히 의미 있었다. 게다가 아직 레거시 교체 주기 초반임을 감안해야 한다.

그러나, 2026년 3월 3일 리사 수는 CPU가 Data Center 매출에서 차지하는 비중이 점점 커지고 있다고 말했다. 에이전트 때문에 리사 수도 “far exceeded”라고 직접 인정한 것이다. 이 순간부터 시장이 “20% 초기 예상”을 완전히 벗어났다고 생각한다.

표 4. AMD의 CEO 발언의 톤 변화

날짜	AMD 공식 발언 (리사 수)	해석
2025년 말 ~ 2026년 초 (초기 가이던스)	CPU 성장률 20% 이상 암시	보수적 초기 예상. 에이전트 트래픽 초기라 안전하게 잡음.
2026년 2월 3일 Q4 2025 실적 발표	Data Center 전체 60% 이상 CAGR (3~5년) 재확인	여기서 이미 CPU가 GPU보다 더 빨리 붙기 시작한 조짐.
2026년 3월 3일 모건스탠리 TMT 컨퍼런스	"The CPU portion of the business has actually **far exceeded my expectations** in terms of demand." "Supply now tightening."	전환점. 초기 20% 예상 완전 초과. 하이퍼스케일러들이 CPU를 예상보다 훨씬 많이 사들이고 있음.

자료: 미래에셋증권 리서치센터

AMD의 서버 CPU 점유율은 2023년 약 23%에서 2025년 약 34%로 확대된 바 있다. 거기에 베니스(256코어)가 2026년 하반기 출시되면, 에이전트 워크로드가 요구하는 코어 밀도에서 인텔(Clearwater Forest, 가성비 문제)을 압도할 수 있다.

표 5. AMD의 서버 CPU(EPYC) 세그먼트 전망

항목	CY2024	CY2025	CY2027E	근거
서버 CPU 매출 ($B)	~5.5	~7.0	~12~14	Turin 128코어 → Venice 256코어 전환, 에이전트 수요
x86 서버 시장 점유율	~28%	~34%	~42~45%	칩릿 구조적 비용 우위, 인텔 Clearwater Forest 지연/가성비 문제
ASP 추이	상승	상승	추가 상승	코어 수 증가(128→256)에 따른 자연스러운 ASP 상승, 에이전트 수요 공급 병목 프리미엄

자료: 미래에셋증권 리서치센터

앞으로, 멀티에이전트가 추가하는 CPU 수요에다 Grok 4.20와 같은 Hydra 아키텍처가 대세가 될 것을 감안해보자. Anthropic, Google도 멀티에이전트 도입을 완료했고, 2027년이면 프론티어 모델 대부분이 멀티에이전트 아키텍처를 표준 채택할 것으로 예상하기 때문이다. "모델 내부 자아 수(4~16개) × 에이전트 숫자"라는 논리를 통해 CPU 수요는 고성장세를 구가할 수 있을 것으로 생각한다. EPYC Turin/Venice 주문이 이미 공급을 압도한 상황이고, CPU 단독 성장률은 40% 대까지 올라갈 가능성이 높아진 국면이다. 현재 국면에서 에이전트 수요 폭발로 2라운드가 시작된 듯 보인다.

(3) 결론: 새로운 사이클은 언제나 환영이야

표 6. CPU 사이클이라는 새로운 변화에 따라 플레이어들이 갖게 되는 기회 및 리스크 요인

종목 (티커)	기회요인	리스크 요인
AMD (AMD)	CPU 슈퍼사이클 직접 수혜. Venice 256코어, 칩릿 비용 우위. 서버 CPU 점유율 34%→45% 전망	MI GPU의 CUDA 열위, HBM 확보 경쟁
NVIDIA (NVDA)	GPU 절대 강자 + CPU/DPU 숨겨진 TAM $42~45B. Vera-only 계약. 공급망 선점(HBM·CoWoS 등).	커스텀 ASIC 대체
Broadcom (AVGO)	$1,000억 AI 칩 가이던스. OpenAI를 고객으로 확보. 파운드리/패키징 게이트키퍼	고객 집중도(상위 6사), 엔비디아 SW 생태계 열위
Arm (Arm)	피닉스로 IP→칩 메이커 전환. DC 로열티 YoY 100%+. Armv9 침당 로열티율 구조적 상승	채널 충돌(AWS/MS/Google), RISC-V로의 이탈
SK하이닉스 (000660)	HBM 90% 엔비디아 공급. 멀티에이전트 KV 캐시 증가 관련	삼성과 Micron의 추격, DRAM 사이클 변동
TSMC (2330)	CoWoS 캐파 = AI 하드웨어의 물리적 병목. 모든 CPU/GPU/ASIC이 TSMC 통과	지정학적 리스크(대만)
Intel (INTC)	파운드리 지정학 가치. CXL이 강점. Granite Rapids가 Maia 가속기의 헤드노드로 채택.	낮은 성능 향상폭, AMD 칩렛 대비 열위

자료: 미래에셋증권 리서치센터

이 보고서를 관통하는 하나의 공식은 다음과 같다. "에이전트 트래픽 증가 × 멀티에이전트 승수 → 순차적 CPU 워크로드 증가 → 레거시 CPU 붕괴 → 신형 CPU 강제 교체 수요"

여기에 $1,000억이라는 수치의 브로드컴 매출 가이던스, $3,800억에 달하는 앤트로픽의 밸류에이션, 180조 개에 달하는 일간 토큰 소비량, 엔비디아 B200 GPU 가용성 0% 등이 논리 강화의 연료로 부어지고 있다.

이처럼 GPU와 메모리, 인터커넥트, 패키징에서의 현장의 소리들은 계속해서 슈퍼사이클을 외치고 있는데, 극초저지연 오케스트레이션과 데이터 이동의 시대에서 CPU에 대해서는 시장에서 그렇게 많이 논의되고 있지 않기 때문이다.

뒤늦은 듯한 CPU의 사이클 동참은, 시장이 상상하지 못했던 만큼 더욱 거대한 병목으로 다가올 것이다. 엔비디아는 Vera-only 계약으로 CPU 시장에 독립 교두보를 마련했고, 메타는 3개월 만에 TPU 채택 및 탈엔비디아 전략을 폐기하고 엔비디아 풀스택에 항복했다. AMD는 Venice 256코어로 코어 밀도 경쟁을 압도하려 하고 인텔은 이쪽 시장에 사활을 걸며 SMT까지 포기했다. Arm은 이제 완제품 생산 준비까지 하고 있다.

올해가 지나갈수록, 데이터센터 CPU 시장의 판도가 급격하게 재편되는 흐름을 목도하게 될 것이다. 이 흐름을 놓쳐서는 안 된다. GPU의 사이클이 성숙하며 투자 알파가 줄어든다고 느끼는 사람들에게 말하고 싶다. 동일한 메가트렌드 안에서 새로운 사이클이 열리고 있다. 그리고 새로운 사이클은, 언제나 환영이다.

II. AGI & Agents

1. 2020년 2월의 감각

(1) 우리는 지금 어디에 서 있는가

2026년 2월 5일 오후, OthersideAI의 CEO인 Matt Shumer가 자신의 블로그에 에세이 하나를 올렸다. 제목은 "Something Big Is Happening"이었다. 48시간이 채 지나지 않아 이 글은 8,500만 뷰를 돌파했다. 실리콘밸리 벤처 투자자들이 텔레그램 단체방에서 이 링크를 공유했고, 워싱턴의 정책 입안자들이 인쇄본을 돌려 읽었으며, 평범한 직장인들이 "내 직업이 없어지는 건가요?"라는 질문을 댓글에 남겼다. 유명 작가들이 반박문을 썼고, 반박에 대한 반박이 이어졌다.

왜 이 에세이였을까. Shumer가 새로운 사실을 발견한 것은 아니었다. 그가 제시한 데이터 대부분은 이미 공개된 자료였다. 그러나 그는 흩어진 점들을 "2020년 2월"이라는 하나의 선으로 연결하는 데 성공했다. 팬데믹 선언 직전, 아직 일상이 정상처럼 보이던 그 시점. 이미 무언가 돌이킬 수 없이 달라져 있었지만, 아무도 그것을 체감하지 못했던 바로 그 순간. Shumer는 지금 우리가 그 시점에 있다고 주장했다.

그림 13. Matt Shumer가 지난 2월 초 작성한 에세이는 X를 통해서만 8,500만 조회수를 넘김

자료: X(@Matt Shumer), 미래에셋증권 리서치센터

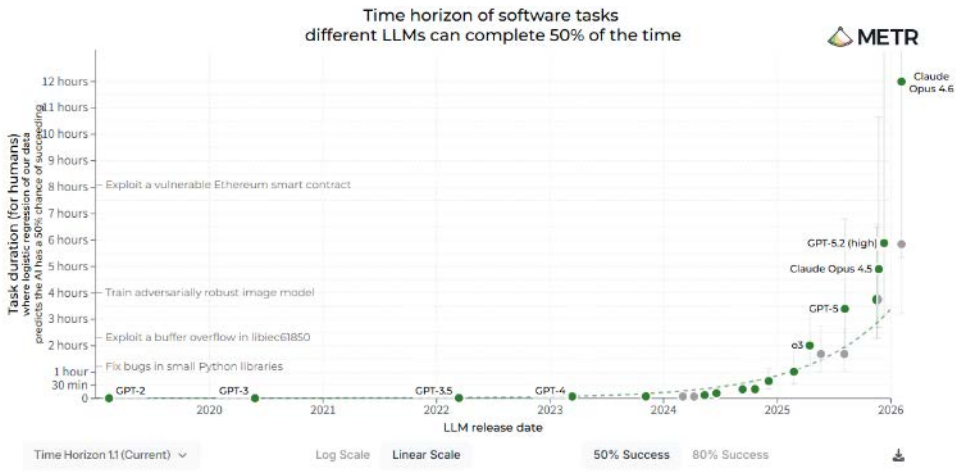
(2) Shumer가 주목한 날

Shumer는 에세이 첫 문장에서 날짜를 특정했다. 2026년 2월 5일, GPT-5.3 Codex와 Claude Opus 4.6이 동시에 공개된 날. 그는 이 날을 이전 시대와 이후 시대를 가르는 경계선으로 지목했다. 지나치게 드라마틱한 수사처럼 들릴 수 있지만, 그 기저에는 단단한 측정 데이터가 있다.

METR(Model Evaluation and Threat Research)는 2019년부터 최전선 AI 모델들이 자율적으로 완료할 수 있는 작업의 길이를 추적해왔다. 측정 기준은 단순하다. 인간 전문가가 동일한 작업을 완료하는 데 걸리는 시간을 기준으로 AI가 50% 확률로 성공할 수 있는 최대 작업 길이를 계산해, 이를 "시간지평(Time Horizon)"이라고 명명했다.

2019년, 이 수치는 2초에 불과했다. AI는 매우 짧고 단순한 작업만 수행할 수 있었다. 2022년에는 이 시간이 약 36초으로 늘었다. 그러다 2024년 초에는 4분을 돌파했고, 2025년 11월에는 4시간에 근접했다. 그리고 2026년 2월, Claude Opus 4.6이 기록한 수치는 11시간 59분이었다. 6년 만에 몇 초에서 한나절 수준의 작업으로 도약한 것이다. 더 중요한 사실은 이 성장 속도 자체가 아직도 가속되고 있다는 점이다. METR의 2026년 1월 업데이트에 따르면, 2019년에서 2025년 사이 약 7개월마다 두 배씩 증가하던 시간지평은, 2024년 이후로는 4.3개월마다 두 배씩 증가하고 있다. 지수함수 위에서 다시 가속이 일어나고 있다는 뜻이다.

그림 14. LLM이 50% 확률로 성공적으로 수행할 수 있는 소프트웨어 작업의 소요시간



자료: METR, 미래에셋증권 리서치센터

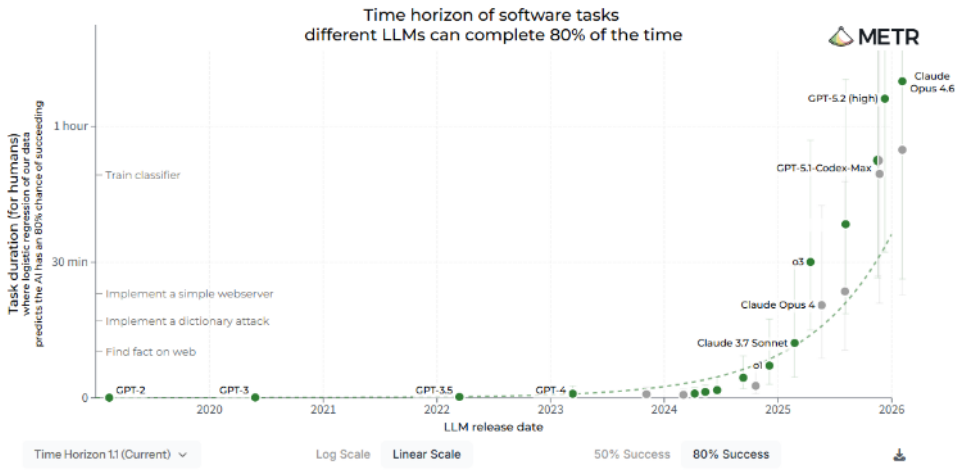
(3) 자신의 학습 과정을 디버깅하는 AI

Shumer가 에세이에서 가장 강조한 사건은 "재귀적 자기 개선(Recursive Self-Improvement)"의 등장이었다. GPT-5.3 Codex가 자신의 학습 과정에서 발생한 오류를 스스로 탐지하고 수정했다는 것이다.

개발 과정에서 정확히 어느 정도의 영향력을 행사했는지 정확히 알기는 어렵지만, 이러한 자기 개선이 일어나고 있다는 사실만큼은 분명하게 알 수 있다. Claude Cowork가 10일 만에 개발되었고, 그 모든 코드가 Claude 자신에 의해 작성되었다는 사실은 이미 알려져 있다. AI가 자기 자신이 아니더라도, 최소한 자신이 사용하는 도구 정도는 스스로 만들 수 있는 수준이라는 것이다. 아직 "자기 학습의 개선"과는 거리가 있지만, "자기 자신을 위한 도구를 설계하는 AI"는 이미 현실이 됐다. Shumer의 직관도 이를 가리키고 있다.

METR 데이터에서 주목해야 할 또 하나의 수치가 있다. 2026년 2월 기준, Claude Opus 4.6의 80% 시간지평, 즉 80% 확률로 성공하는 작업의 최대 길이는 1시간 3분이었다. 50% 확률로 12시간짜리 작업을 수행할 수 있다는 것과, 80% 확률로 1시간짜리 작업을 안정적으로 수행할 수 있다는 것은 서로 다른 이야기다. 전자가 AI의 최대 가능성을 보여준다면, 후자는 실제 업무에 투입 가능한 신뢰 수준을 보여준다. 지금의 AI는 두 영역에서 동시에 유의미한 임계점을 통과하고 있다.

그림 15. LLM이 50% 확률로 성공적으로 수행할 수 있는 소프트웨어 작업의 소요시간



자료: METR, 미래에셋증권 리서치센터

2. 지식 노동의 에이전트화 = 현실

(1) Cowork 엔터프라이즈 출시: 에이전트 진짜 오나?

2월 24일, 앤트로픽은 "Briefing: Enterprise Agents" 온라인 행사를 통해 Claude Cowork의 엔터프라이즈 전환을 공식 발표했다. 그 자리에서 Kate Jensen 아메리카 총괄은 "2025년 엔터프라이즈 AI 에이전트를 둘러싼 과대 선전은 대부분 시기상조였다"고 인정했다. 그러나 그는 이것이 기술의 실패가 아니라 접근 방식의 실패였다고 덧붙였다.

Cowork의 핵심은 통합이다. 이제 기업들은 Claude를 Google Drive, Gmail, DocuSign, FactSet, MSCI, Harvey, LegalZoom 등 13개 주요 서비스와 연결할 수 있다. 재무 분석, 투자은행, 법무, HR, 엔지니어링 등 부서별 특화 플러그인 템플릿도 제공된다. 중요한 것은 이 도구들이 기업의 고유한 방식에 맞게 구성된다는 점이다. Scott White 엔터프라이즈 프로덕트 총괄은 고객들로부터 직접 들은 이야기를 인용했다. "우리가 원하는 것은 법무용 Claude가 아니라, 우리 회사 법무팀을 위한 Cowork입니다."

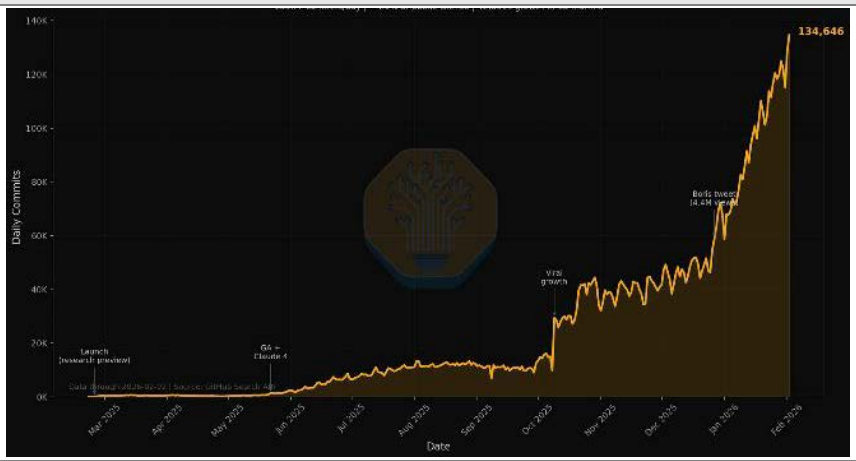
앤트로픽의 발표에 의하면, 노보 노디스크는 Claude 기반 내부 플랫폼 NovoScribe를 도입한 이후 문서화 작업 시간이 10주에서 10분으로 단축되었으며, Spotify의 모든 엔지니어는 이제 자연어로 "대규모 코드 마이그레이션을 시작"할 수 있다. 오직 시니어 엔지니어에게만 맡겨지던 복잡한 레거시 코드 정리 작업이 민주화된 것이다.

(2) Claude Code: 조용히 쌓인 증거들

Cowork 발표보다는 덜 주목받았지만, 더 오래 이어질 흐름이 있다. Claude Code의 성장 궤적이다. 이 도구는 2025년 출시 이후 6개월 만인 지난 11월, 연환산 매출 10억 달러를 돌파했다. 그리고 시리즈 G 발표 당시 공개된 수치에 따르면, 현재 연환산 매출은 25억 달러에 근접한다. 두 달도 안 되어 두 배 이상 성장한 것이다.

GitHub 전체 공개 커밋의 4%가 현재 Claude Code에 의해 생성되고 있으며, 연말에는 이 비율이 20%를 넘을 것으로 전망된다. 연간 100만 달러 이상을 Claude에 지출하는 기업은 500개를 넘었다. 2년 전에는 단 12개 기업에 불과했다는 것을 떠올려 보면, 가히 폭발적인 수요다. 이 수치들은 Shumer의 주장에 구체적인 살을 붙인다. 화이트칼라 노동의 도구가 조용히, 그러나 빠르게 교체되고 있다.

그림 16. Claude Code의 GitHub 커밋 추이. 하루 13.5만 건 이상, GitHub 전체 커밋의 ~4% 추정



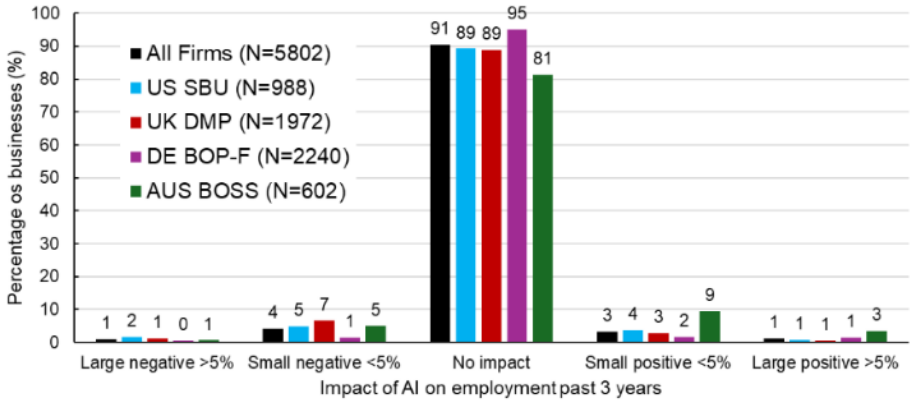
자료: Semianalysis, 미래에셋증권 리서치센터

3. 80%는 체감 못 했다?

(1) NBER은 아직 멀었다던데

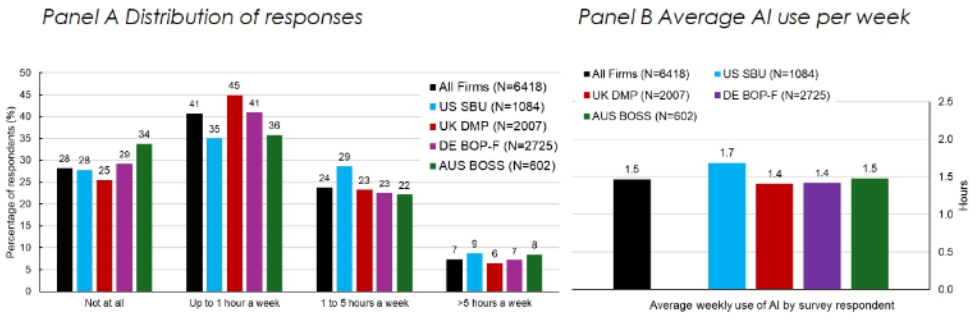
국가경제연구소(NBER)가 발표한 보고서는 Shumer에 대한 가장 강력한 반론처럼 보인다. 미국, 영국, 독일, 호주 4개국 6,000여 명의 최고경영진을 대상으로 한 이 연구의 핵심 발견은 명확하다. 응답자의 80% 이상이 "AI가 지난 3년간 고용이나 생산성에 영향을 주지 않았다"고 답했다. 고용 영향이 없다는 응답은 90%를 넘었다. 70%의 기업이 AI를 사용한다고 응답했음에도 불구하고, 임원들이 AI에 할애하는 시간은 주 평균 1.5시간에 불과했다. 4분의 1은 개인적으로 AI를 전혀 사용하지 않는다고 답했다.

그림 17. AI가 지난 3년간 고용이나 생산성에 영향을 주지 않았다고 응답한 비율이 80% 이상



자료: NBER, 미래에셋증권 리서치센터

그림 18. C-레벨 임원은 평균 AI를 단 1.5시간 사용하며, 4분의 1은 AI를 전혀 사용하지 않는다 답변



자료: NBER, 미래에셋증권 리서치센터

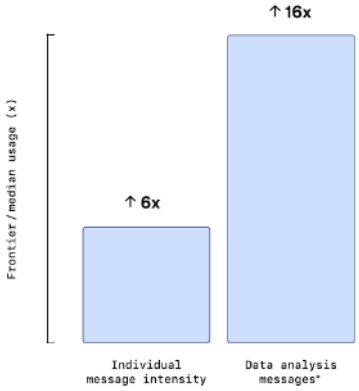
비슷한 흐름은 다른 연구에서도 발견된다. PwC의 2026년 CEO 조사에서 56%가 AI 투자에서 "유의미한 재무적 이익을 얻지 못했다"고 응답했다. MIT 연구에서는 AI 파일럿의 95%가 생산성 향상을 증명하지 못했다. 덴마크 노동시장 연구에서는 AI 도입이 소득이나 근무시간에 "통계적으로 유의미한 영향이 없었다"는 결과가 나왔다. 이 데이터들을 표면만 읽으면 Shumer의 서사와 정면으로 충돌한다.

그러나 이 숫자들을 그대로 받아들이기 전에, 한 가지 핵심 질문을 던져야 한다. "이 70%의 기업들은 구체적으로 어떤 AI를 사용하고 있었는가?"

(2) 같은 회사 안에서도 격차는 벌어지고 있다

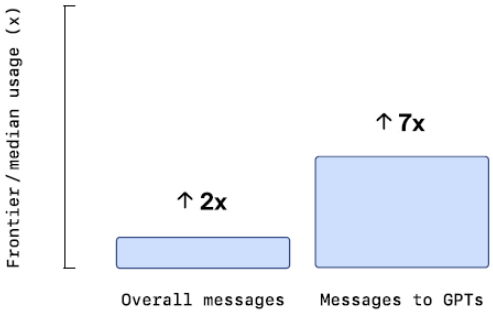
이 질문에 답할 수 있는 가장 직접적인 데이터가 있다. OpenAI가 2025년 12월 발표한 '엔터프라이즈 AI 현황 보고서'다. 이 보고서는 100개 가까운 기업에 소속된 9,000명의 실제 사용자 데이터를 분석했다. 그 결과 상위 5% 사용자, 즉 '프론티어 워커'는 중간값 직원보다 6배 많은 메시지를 보내고 있었다. 프론티어 기업은 중간값 기업보다 직원 1인당 2배 많은 메시지를 사용했으며, 고급 기능(GPTs, 프로젝트)을 활용하는 비중은 7배나 높았다.

그림 19. AI 상위 5%는 중간값보다 6배 많은 메시지를 보내며,



자료: OpenAI, 미래에셋증권 리서치센터

그림 20. 상위 5% 기업은 중간값 기업보다 2배 많은 메시지를 사용



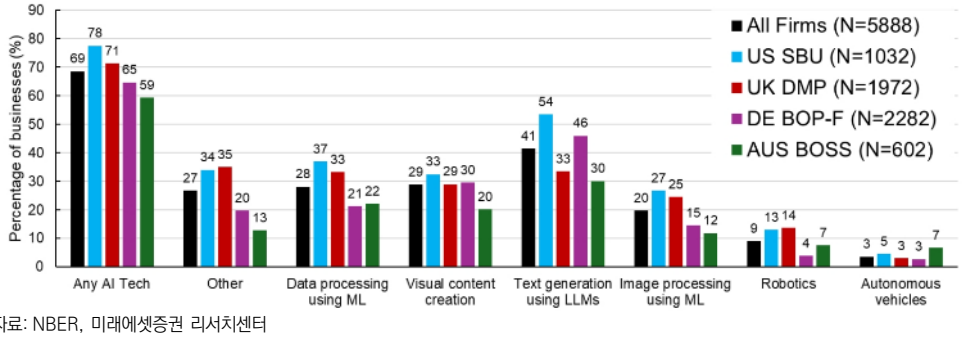
자료: OpenAI, 미래에셋증권 리서치센터

여기서 핵심은 이 비교가 "AI를 쓰는 집단"과 "안 쓰는 집단" 사이의 격차가 아니라는 점이다. 동일한 ChatGPT Enterprise 구독료를 지불하고 있는 같은 회사 안에서 벌어지고 있는 격차다. 같은 도구를, 같은 가격으로, 완전히 다르게 쓰고 있다. NBER가 "AI를 사용한다"고 집계한 조직의 내부에서도 이미 이 분열이 진행 중인 것이다.

더 결정적인 사실은 OpenAI 보고서의 다음 발언이다. "모델의 성능이나 도구의 부재가 더 이상 병목이 아니다. 주요 제약은 조직의 준비도와 실행력이다." 지금의 AI는 대부분의 조직이 실제로 활용하는 것보다 훨씬 더 많은 일을 할 수 있다. 그러나 대부분의 조직은 여전히 AI를 '질문에 답해주는 챗봇' 수준으로만 쓰고 있다.

이것이 NBER이 측정한 AI 사용의 실체였다. 응답 기업의 가장 흔한 사용처는 "대형 언어 모델을 이용한 텍스트 생성"이었다. 이는 구조화된 워크플로우나 에이전트 자동화와는 매우 거리가 먼 저차원의 사용이다. Shumer가 지적했듯, 지금의 AI와 6개월 전의 AI는 완전히 다른 기술이다. 또한, 어떤 AI를 어떻게 쓰느냐에 따라 완전히 다른 결과가 나온다.

그림 21. 아직도 "텍스트 생성"에 AI를 가장 많이 사용. 이는 AI가 할 수 있는 가장 단순한 일.



자료: NBER, 미래에셋증권 리서치센터

METR가 측정한 50% 시간지평을 기준으로, 2024년 9월 Claude 3.5 Sonnet은 약 21분 수준이었다. Claude Opus 4.6은 약 12시간이다. 두 세대 사이에 34배 이상의 시간 격차가 발생했다. 단순히 더 똑똑해진 것이 아니라, 자율적으로 처리할 수 있는 작업의 복잡도 자체가 달라진 것이다. "AI를 써본 적이 있다"는 응답자 상당수가 경험한 모델은 이 12시간짜리 Opus 4.6이 아니라, 그보다 훨씬 이전 세대의 모델이었을 가능성이 높다.

실제로 프론티어를 제대로 활용하는 조직의 데이터는 NBER와 정반대의 그림을 그린다. GitHub Copilot을 이용한 통제 실험에서 개발자들은 동일한 작업을 55.8% 더 빠르게 완료했다. Claude 기반 작업 자동화 데이터에서는 작업 시간이 92% 줄어든 것으로 보고됐다. OpenAI 엔터프라이즈 보고서에서 헤비 유저들이 보고한 시간 절약은 주당 10시간 이상이었다. 이는 주 1.5시간 챗봇 사용이 보고하는 "영향 없음"과는 완전히 다른 세계다.

(3) 80%의 진짜 의미 - Solow 역설의 재림, 그러나 압축판

1987년 노벨 경제학상 수상자 Robert Solow는 "컴퓨터 시대는 어디에서나 보이지만, 생산성 통계에서는 보이지 않는다"는 말을 남겼다. 이른바 "Solow 역설"은 실제로 존재했다. 기업들은 1980년대 내내 IT에 막대한 투자를 했지만, 거시 생산성 통계는 꿈쩍도 하지 않았다. 그런데 1995년부터 2005년까지, 정확히 10년 후에 생산성 붐이 폭발했다. 투자와 성과 사이에 10년의 시차가 있었던 것이다. 이유는 간단했다. 컴퓨터를 기존 업무 프로세스에 끼워 넣는 것만으로는 생산성이 오르지 않았다. 그 프로세스 자체를 컴퓨터 중심으로 재설계해야 했고, 그 재설계에 10년이 걸렸다.

AI 생산성 역설도 이 패턴과 유사하지만, 훨씬 압축적으로 전개될 것이다. Reuters의 CEO Steve Hasker가 앤트로픽 행사에서 "18개월이 필요하다"고 말한 것도 같은 이유다. 이 18개월은 Cowork라는 도구의 한계가 아니라, 법무 조직이 AI를 중심으로 프로세스를 재설계하는 데 필요한 변화 관리의 시간이다. 기술의 준비가 조직의 준비를 앞지르고 있다. NBER의 80%는 AI의 한계를 보여주는 것이 아니라, 이 재설계가 아직 얼마나 초기 단계에 머물러 있는지를 보여주는 숫자다.

결론적으로 NBER 데이터의 진짜 의미는 이렇게 읽혀야 한다. "프론티어 AI의 실제 능력을 경험한 조직은 아직 소수이며, 그 소수는 이미 측정 가능한 생산성 격차를 벌리고 있다." 80%가 체감하지 못했다는 사실은 AI를 부정하는 증거가 아니다. 오히려 프론티어가 아직 얼마나 좁은 영역에만 침투해 있는지, 그리고 그 침투가 본격화될 때 어떤 격차가 드러날지를 예고하는 데이터다.

4. Shumer가 옳은 것, 그가 과소평가한 것

우리는 Shumer의 핵심 주장에 동의한다. 우리는 지금 변곡점 위에 있다. METR 데이터가 보여주는 지수적 가속, Cowork가 촉발한 SaaS 시장의 지각변동, Claude Code의 폭발적 성장. 이 사건들은 모두 같은 방향을 가리킨다. "2020년 2월의 감각"은 실재한다.

그러나 Shumer가 과소평가한 것이 있다. 그의 에세이에는 암묵적인 전제가 있다. AI를 쓴다면, 그 변화를 체감한다는 것이다. 현실은 다르다. OpenAI 엔터프라이즈 데이터가 보여주듯, 같은 도구를 구독하는 조직 안에서도 상위 5%와 나머지 사이에는 이미 6배의 사용 강도 격차가 존재한다. NBER가 측정한 80%의 "영향 없음"은 AI가 아직 약하다는 증거가 아니다. 대부분의 조직이 아직 프론티어를 경험하지 못하고 있다는 증거다. 팬데믹은 이미 시작되었지만, 모든 곳에 동시에 도달하지는 않는다.

바로 이 불균등함이 지금 이 국면에서 가장 중요한 변수다. 기술이 가속하고 있다는 사실은 이제 모두가 안다. 문제는 그 가속이 어디서 먼저, 어디서 나중에 현실화되느냐다. 워크플로우 통합 영역에서는 Cowork가 열어젖힌 방향이 명확하다. 지켜볼 만한 시그널은 어떤 기존 플레이어가 이를 통합해 자신의 해자로 전환하는 데 성공하느냐다. 최근 SaaS 급락만 보더라도, 마치 모든 소프트웨어가 대체될 것 같은 패닉이 있었지만, 그 안에도 반드시 생존자가 있다.

그리고 가장 과소평가된 영역이 있다. Hasker CEO가 "18개월이 필요하다"고 말한 바로 그 시장이다. 법무, 세무, 의료, 금융 등 규제 산업에서 AI 도입의 진짜 병목은 기술이 아니라 규제에 맞는 프로세스 재설계와 관리 능력이다. 이를 자체적으로 구현하거나, 서비스로 제공할 수 있는 기업은 'AI 코로나'를 이겨낼 수 있다.

2022년의 6개월은 아무것도 아닌 것처럼 보였지만, 2025년의 6개월은 모델이 두어 번 업데이트 되는 시간이었다. AI 타임라인에서 누군가는 6개월을 먼저 살고 있으며, 누군가는 6개월 뒤에서 따라오고 있다. 이들의 격차도 지수적으로 벌어질 것이라는 점을 명심하자. AI 타임라인이 압축될수록 (유의미한) AI 활용의 한 발자국 차이가 실제 경쟁력에서 만들어내는 거리는 점점 더 벌어질 것이다. 지금의 한 발자국은 1년 후의 한 발자국보다 훨씬 싸다.

III. AI Peer Table

표 7. Peer Table (1)

종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%)						실적발표 예정일	AI 밸류체인 관련 한줄평
				1W	1M	3M	6M	1Y	YTD		
최종 소비자(온디바이스, 앱)											
애플	AAPL US	260.29	5,673	-4.6	-5.6	-6.5	8.8	10.9	-4.2	05-01	엔드유저 데이터 보유 및 서비스 배포능력
테슬라	TSLA US	405.55	2,259	-0.7	2.1	-10.9	15.6	45.3	-9.8	04-22	FSD v13 공공 출시 및 1Q25 중국 출시 가능성
Pony AI	PONY US	13.59	9	-4.3	3.9	-4.3	-0.9	-10.7	-6.3	03-26	중국의 자율주행 선도기업 중 하나
삼성전자	005930 KS	193,500	1,145	-11.0	20.2	79.7	181.0	264.0	60.7	04-07	온디바이스 AI, HBM, 파운드리 영역에서의 가능성
듀오링고	DUOL US	101.54	7	-13.5	-11.2	-49.2	-62.6	-65.3	-42.1	05-01	AI 캐릭터와 상호작용하는 학습 플랫폼
크래프톤	259960 KS	220,000	10	-11.8	-7.7	-10.3	-31.9	-36.2	-10.1	04-29	PUBG, inZOI 등 자사 게임에 AI CPC 기능 탑재
쇼피파이	SHOP US	134.79	261	7.0	21.2	-16.3	-8.2	27.0	-16.3	05-08	이커머스 콘텐츠 생성의 자동화(Shopify Magic AI)
코인베이스	COIN US	205.71	81	13.6	40.8	-23.7	-31.2	-7.5	-9.0	05-08	AI 에이전트간 자율 암호화폐 거래 인프라 구축
비즈니스 효율화											
크라우드스트라이크	CRWD US	426.16	160	11.8	13.0	-16.8	2.0	16.6	-9.1	06-03	기업 고객 대상 엔드포인트 보안(Charlotte AI)
IBM	IBM US	256.55	357	6.0	-11.0	-16.2	4.4	4.6	-12.9	04-22	기업 고객 대상 AI 모델 개발, 배포 플랫폼(watsonx)
액센츄어	ACN US	214.00	196	3.2	-8.4	-19.3	-15.0	-37.1	-19.8	03-19	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
SAP	SAP GY	170.98	361	-0.7	0.3	-19.8	-25.1	-34.8	-17.9	04-23	기업용 SW 솔루션에 AI를 내장한 에이전트(Joule)
인포시스	INFO IN	1,305.80	86	1.3	-14.1	-19.2	-8.2	-21.4	-19.2	04-16	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
일루미나	ILMN US	128.24	29	0.7	-4.0	-0.6	29.6	49.7	-2.2	05-08	AI 기반 첨단 유전체 분석
HD현대	267250 KS	274,000	22	-5.8	14.4	31.7	91.3	272.7	46.1	04-29	AI, 로보틱스, 디지털 트랜스포메이션 관련 노력
AI Ops											
클라우드플레어	NET US	192.31	101	10.1	17.9	-4.3	-10.3	33.8	-2.5	05-08	실시간 AI 추론에 CDN 수요증가 가능성(Workers AI)
팔란티어	PLTR US	152.67	542	12.3	17.4	-16.0	-0.3	69.4	-14.1	05-05	비즈니스 현장의 의사결정을 돕는 AI 플랫폼(AIP)
파운데이션 모델											
알파벳	GOOGL US	300.88	5,403	-2.1	-9.2	-6.3	28.2	74.6	-3.9	04-24	OpenAI와의 AGI 경쟁이 가능한 업체(Gemini)
메타 플랫폼스	META US	660.57	2,481	0.5	-1.4	-1.8	-12.1	0.9	0.1	04-30	오픈소스 AI 개발의 선구자(Llama 등)
알리바바	9988 HK	126.30	458	-11.6	-18.5	-18.5	-4.2	-8.8	-11.6	03-06	중국의 CSP이자 중국의 대표적인 언어모델(Qwen)
네이버	035420 KS	221,500	35	-13.3	-10.7	-10.3	-2.3	2.2	-8.1	05-08	한국형 LLM 개발 및 AI 서비스(HyperCLOVA)

자료: Bloomberg, 미래에셋증권 리서치센터

주: 실적발표 일정은 변경될 수 있음.

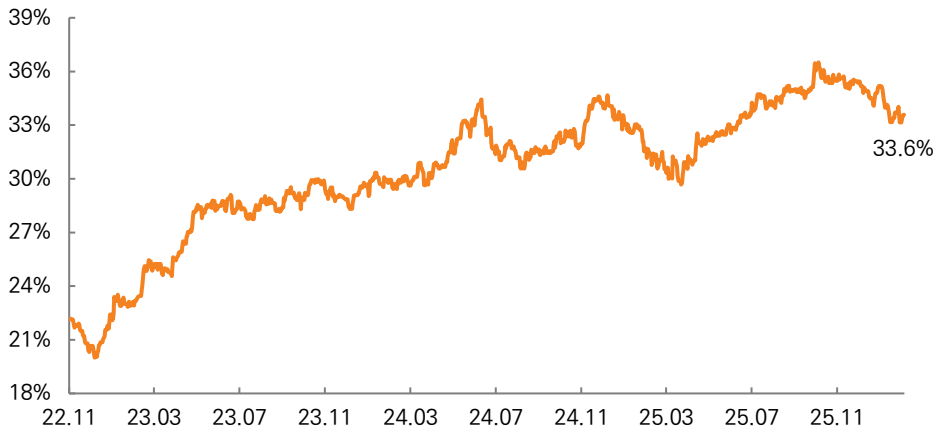
표 8. Peer Table (2)

종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%) 1W	1M	3M	6M	1Y	YTD	실적발표 예정일	AI 밸류체인 관련 한줄평
클라우드											
마이크로소프트	MSFT US	410.68	4,527	2.2	4.6	-14.8	-16.7	3.2	-14.9	04-30	OpenAI 모델 라이선스 독점권 보유한 세계 2위 CSP
아마존 닷컴	AMZN US	218.94	3,489	5.3	-1.7	-4.6	-5.8	5.1	-5.1	05-01	최고의 CSP로서 Anthropic에 수 조원 투자
오라클	ORCL US	154.79	660	3.0	13.4	-28.7	-33.2	-3.2	-20.4	03-10	AI 데이터센터 capex 경쟁에 진입한 전통적 강자
소프트뱅크	9984 JP	3,864.00	208	-5.5	-3.4	-19.9	-0.4	89.6	-12.2	05-13	'비전펀드'는 AI에 집중. 일본 최고의 AI 슈퍼컴퓨터.
하드웨어 인프라											
엔비디아	NVDA US	183.34	6,614	-0.8	6.7	0.5	9.8	56.3	-1.7	05-20	AI 모델 훈련 및 추론에 필수인 GPU계 현존 최강자
브로드컴	AVGO US	332.77	2,339	3.4	7.2	-14.6	-0.3	75.3	-3.9	06-05	이더넷 기반 네트워킹 반도체의 최강자
AMD	AMD US	199.45	483	-2.1	3.6	-8.5	32.0	96.2	-6.9	05-06	서버용 CPU의 최강자. AI 가속기 분야 패스트팔로어
인텔	INTC US	45.95	341	1.1	-4.7	11.0	87.6	120.8	24.5	04-24	파운드리로서 소버린 AI 미 지정학적 가치 부상
마이크론	MU US	397.05	663	-4.5	3.7	67.4	202.5	322.3	39.1	03-18	SK하이닉스 추격 중인 "미국"의 메모리 반도체
SK하이닉스	000660 KS	953,000	679	-10.3	11.3	75.8	249.3	398.1	46.4	04-23	HBM 부문 전세계 최강자
시놉시스	SNPS US	442.30	126	3.8	7.8	-5.2	-26.1	-1.3	-5.8	05-28	반도체 EDA 부문 리딩 업체, AI 에이전트 적극 활용
케이던스 디자인 시스템	CDNS US	299.84	123	0.8	11.0	-11.2	-14.6	22.5	-4.1	04-28	반도체 EDA 부문 리딩 업체
암페놀	APH US	136.24	249	-8.2	6.7	-2.0	23.8	112.2	0.8	04-29	엔비디아향 AI용 서버 내 고속 I/O 및 커넥터 제조
버티브	VRT US	249.75	142	-3.7	40.5	32.1	101.5	184.6	54.2	04-23	액체 냉각 방식에 있어 글로벌 선두업체
Arm	ARM US	120.62	190	-6.7	8.8	-14.6	-12.7	-3.3	10.3	05-07	AI 가속기 설계 위한 다수의 IP를 소유한 팹리스
마벨 테크놀로지스	MRVL US	75.68	98	-4.6	2.0	-23.4	19.7	-15.8	-10.9	03-05	데이터 인프라용 네트워킹 및 스토리지 ASIC 설계
코어위브	CRWV US	74.82	60	-23.4	0.2	-15.3	-16.0	-	4.5	05-14	AI용 클라우드 인프라 제공업체
중과서광	603019 CH	84.79	27	-8.0	-2.3	-14.2	2.9	16.0	-0.9	04-23	중국과학원과 긴밀 관계를 유지 중인 HPC 제조업체
SMIC	981 HK	61.50	116	-9.6	-9.0	-11.8	4.8	9.2	-13.9	05-08	7나노 공정을 달성한 중국 유일의 미세공정 파운드리
TSMC	2330 TT	1,900.00	2,306	-4.8	6.7	30.6	62.2	92.1	22.6	04-17	명실상부 반도체 파운드리 최강자
폭스콘	2317 TT	224.00	146	-7.8	4.2	-3.0	9.3	33.7	-2.8	03-16	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
위스트론	3231 TT	128.50	19	-5.5	-0.4	-14.9	11.3	24.3	-14.6	03-12	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
Alchip	3661 TT	3,270.00	12	-6.3	5.0	3.6	-17.6	14.2	-6.8	03-06	빅테크들의 AI 가속기 설계를 돕는 대만의 팹리스
에너지 인프라											
GE 베르노바	GEV US	815.01	326	-7.0	10.5	29.2	40.2	157.9	24.8	04-23	AI 데이터센터용 가스터빈 인프라 솔루션
슈나이더 일렉트릭	SU FP	252.80	251	-7.4	1.8	7.3	15.8	13.8	7.6	07-30	AI 데이터센터 전력관리 및 냉각시스템 통합 솔루션
비스트라	VST US	167.40	84	-5.3	17.0	0.3	-10.7	32.7	3.8	05-07	미국 최대 독립 전력 생산 사업자
블룸 에너지	BE US	159.99	67	-5.1	17.1	34.2	180.3	563.3	84.1	04-30	AI 데이터센터용 연료전지 솔루션
넥스트에라 에너지	NEE US	91.13	282	-0.3	2.8	10.4	30.3	33.1	14.3	04-23	AI 데이터센터용 엔드투엔드 전력 파트너

자료: Bloomberg, 미래에셋증권 리서치센터
주: 실적발표 일정은 변경될 수 있음.

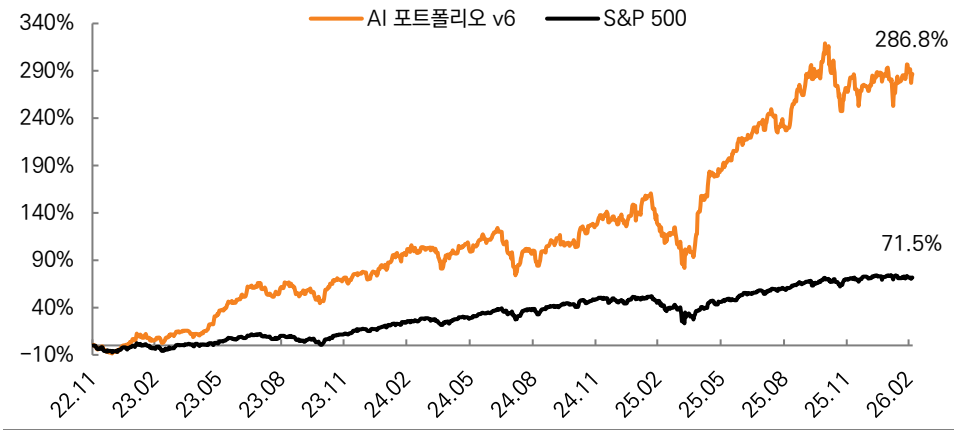
IV. Appendix: Charts

그림 22. ChatGPT 출시일 이후, S&P 500에서 Magnificent Seven이 차지하는 비중(시가총액 기준)



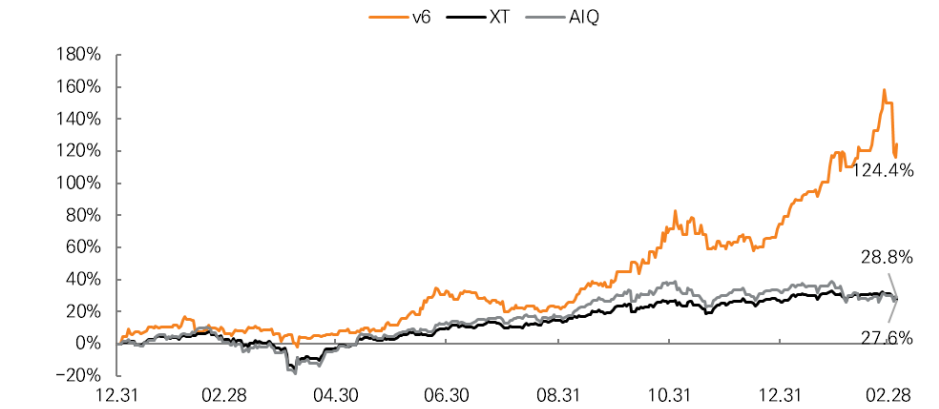
자료: Bloomberg, 미래에셋증권 리서치센터

그림 23. ChatGPT 출시일 이후, AI 밸류체인에 속한 유니버스 종목들의 평균 수익률(동일가중 방식)



자료: Bloomberg, 미래에셋증권 리서치센터

그림 24. 대표 AI ETF vs AI 유니버스(v.6) 수익률 추이(24.12.31 기준)
- Global X Artificial Intelligence & Technology ETF(AIQ), iShares Exponential Technologies ETF(XT)



자료: Bloomberg, 미래에셋증권 리서치센터

표 9. SimpleBench 기준 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	평균 점수(%)	기관명	라이선스	지식 컷오프
-	인간 최고 점수	95.4	-	-	-
-	인간 평균	83.7	-	-	-
1(-)	Gemini 3.1 Pro Preview	79.6	구글	독점 폐쇄소스	2025년 1월
2(-)	Gemini 3 Pro Preview	76.4	구글	독점 폐쇄소스	2025년 1월
3(-)	Claude Opus 4.6	67.6	앤트로픽	독점 폐쇄소스	2025년 8월
4(-)	Gemini 2.5 pro(06-05)	62.4	구글	독점 폐쇄소스	2025년 1월
5(-)	Claude Opus 4.5	62.0	앤트로픽	독점 폐쇄소스	2025년 8월
6(-)	GPT-5 Pro	61.6	OpenAI	독점 폐쇄소스	2024년 9월
7(-)	Gemini 3 Flash Preview	61.1	구글	독점 폐쇄소스	2025년 1월
8(-)	Grok 4	60.5	xAI	독점 폐쇄소스	2024년 12월
9(-)	Claude 4.1 Opus	60.0	앤트로픽	독점 폐쇄소스	2025년 3월

자료: simple-bench.com, 각 사, 미래에셋증권 리서치센터
주: SimpleBench는 인간에게는 쉽지만 기계에게는 어려운 문제로 이루어진 벤치마크

표 10. LLM 벤치마크 챗봇아레나의 Hard Prompts 기준 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	전체 평균 점수	기관명	라이선스	지식 컷오프
1(▲1)	claude-opus-4-6	1535	앤트로픽	독점 폐쇄소스	2025년 8월
2(▼1)	claude-opus-4-6-thinking	1534	앤트로픽	독점 폐쇄소스	2025년 8월
3(-)	gemini-3.1-pro-preview	1523	구글	독점 폐쇄소스	2025년 1월
4(New)	grok-4.20-beta1	1510	xAI	독점 폐쇄소스	2024년 11월
5(▲1)	gemini-3-pro	1503	구글	독점 폐쇄소스	2025년 1월
6(▼2)	gpt-5.2-chat-latest-20260210	1502	OpenAI	독점 폐쇄소스	2025년 8월
7(▼2)	claude-opus-4-5-20251101-thinking-32k	1501	앤트로픽	독점 폐쇄소스	2025년 8월
8(▼1)	claude-opus-4-5-20251101	1496	앤트로픽	독점 폐쇄소스	2025년 8월
9(▼1)	gemini-3-flash	1494	구글	독점 폐쇄소스	2025년 1월
10(▼1)	claude-sonnet-4-6	1491	앤트로픽	독점 폐쇄소스	2025년 8월

자료: lmarena.ai, 각 사, 미래에셋증권 리서치센터
주: Hard Prompts는 기존의 일반적인 요청보다 훨씬 더 까다롭고 복잡한 문제를 제시하여 LLM의 한계를 테스트하는 벤치마크

표 11. LLM 벤치마크 LiveBench의 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	전체 평균 점수	기관명	라이선스	지식 컷오프
1(New)	Gemini 3.1 Pro Preview High*	79.93	구글	독점 폐쇄소스	2025년 1월
2(▼1)	Claude 4.6 Opus Thinking High Effort	76.33	앤트로픽	독점 폐쇄소스	2025년 8월
3(▼1)	Claude 4.5 Opus Thinking High Effort	75.96	앤트로픽	독점 폐쇄소스	2025년 7월
4(▼1)	Claude 4.6 Sonnet Thinking Medium Effort	75.47	앤트로픽	독점 폐쇄소스	2025년 8월
5(▼1)	GPT-5.2 High	74.84	OpenAI	독점 폐쇄소스	2025년 8월
6(▼1)	GPT-5.2 Codex	74.30	OpenAI	독점 폐쇄소스	2025년 8월
7(▼1)	GPT-5.1 Codex Max	73.98	OpenAI	독점 폐쇄소스	2024년 9월
8(▼1)	Gemini 3 Pro Preview High	73.39	구글	독점 폐쇄소스	2025년 1월
9(New)	GPT-5.3 Codex High	72.76	OpenAI	독점 폐쇄소스	2025년 8월
10(▼2)	Gemini 3 Flash Preview High	72.40	구글	독점 폐쇄소스	2025년 1월

자료: LiveBench, Huggingface, 미래에셋증권 리서치센터

표 12. 코딩 관련 벤치마크 SWE-Bench의 성능 상위 모델 10개

순위 (전주 변동)	모델명	점수(%)	기관명	등록일
1(-)	Claude 4.5 Opus (high reasoning)	76.80	앤트로픽	2026-02-17
2(-)	Gemini 3 Flash (high reasoning)	75.80	구글	2026-02-17
3(-)	MiniMax M2.5 (high reasoning)	75.80	미니맥스	2026-02-17
4(-)	Claude Opus 4.6	75.60	앤트로픽	2026-02-17
5(-)	GLM-5 (high reasoning)	72.80	Zhipu AI	2026-02-17
6(-)	GPT-5-2 (high reasoning)	72.80	OpenAI	2026-02-19
7(-)	GPT 5.2 Codex	72.80	OpenAI	2026-02-17
8(-)	Claude 4.5 Sonnet (high reasoning)	71.40	앤트로픽	2026-02-17
9(-)	Kimi K2.5 (high reasoning)	70.80	Moonshot AI	2026-02-17
10(-)	DeepSeek V3.2 (high reasoning)	70.00	DeepSeek	2026-02-17

자료: SWE-Bench, 미래에셋증권 리서치센터
주: 동일 모델이더라도, 추론 설정 등의 변경으로 점수가 재산정되는 경우 새로운 항목으로 판단. 등록일 업데이트를 참고

표 13. AI 에이전트 벤치마크 TAU-bench의 성능 상위 모델 10개

순위 (전주 변동)	모델	성능(%)	비용($)	평가 기준일	실험 환경 구분
1(-)	o4-mini-high	56	11.36	2025년 4월	TAU-bench Tool Calling
2(-)	Claude-3.7 Sonnet	56	42.11	2025년 2월	HAL Generalist Agent
3(-)	o3 Medium	54	14.56	2025년 4월	TAU-bench Tool Calling
4(-)	Claude Opus 4.1	54	180.49	2025년 8월	HAL Generalist Agent
5(-)	Claude-3.7 Sonnet High	52	31.94	2025년 2월	TAU-bench Tool Calling
6(-)	Claude Opus 4.1 High	52	149.98	2025년 8월	TAU-bench Tool Calling
7(-)	Claude Opus 4.1	50	69.78	2025년 8월	TAU-bench Tool Calling
8(-)	GPT-5 Medium	48	23.83	2025년 8월	TAU-bench Tool Calling
9(-)	DeepSeek V3	44	5.43	2025년 3월	TAU-bench Tool Calling
10(-)	Claude-3.7 Sonnet	44	15.45	2025년 2월	TAU-bench Tool Calling

자료: HAL, 미래에셋증권 리서치센터
주: TAU-bench Airline은 AI 에이전트를 통한 개인 업무 자동화 능력을 평가하는 벤치마크. 이번 순위 변동은 TAU-bench Few Shot Agent가 데이터 노출로 인해 평가 항목에서 제외됨에 따른 것으로, 변동한 것으로 표기하지 않음

Compliance Notice

- 당사는 자료 작성일 현재 조사분석 대상법인과 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트 김은지은(는) 자료작성일 현재 엔비디아 65주 보유하고 있습니다.
- 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.