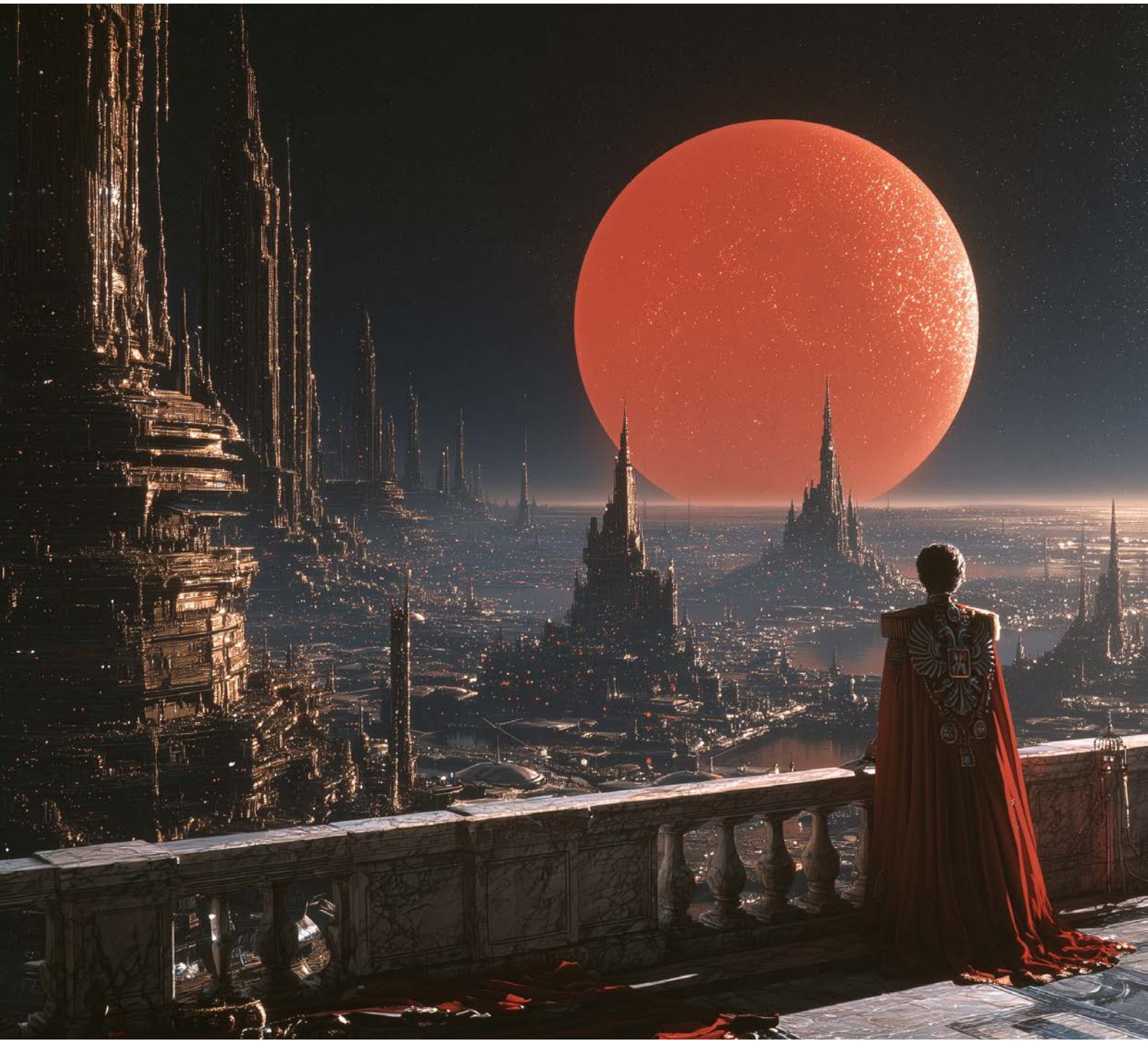
Market Insight | 2026.8.21
그림: Midjourney(토큰값은 내려도 임대료가 오른다)

AI Focus

은하지능전설 2부: The tokens must flow

한종목
chongmok.han@miraeasset.com



CONTENTS

Executive Summary

GPU가 쓰던 겁니까? 헌 걸 갖다가…

AI의 가격이 빠르게 내려가고 있습니다. 그러나 같은 기간 GPU 임대료와 데이터센터의 총 청구서는 오히려 올라갔습니다. 자동차의 연비가 좋아졌다고 도로 위 차량 수와 총주행거리가 줄지는 않는 것과 같습니다. AI 모델 효율이 높아질수록 기업은 더 긴 문맥, 더 많은 추론, 더 복잡한 에이전트를 같은 예산 안에서 사용할 수 있습니다. 한 번의 추론에 필요한 연산량이 줄어도 전체 추론 횟수와 요청당 작업량이 더 빠르게 늘 수 있습니다.

실제 사용량은 이 방향을 가리킵니다. **Ramp 고객사 가운데 AI에 지출하는 기업의 직원 1인당 월간 지출 중앙값은 2024년 1월에 비해 올해 7월에는 4배 넘게 늘었습니다**. OpenRouter에서는 평균 프롬프트 길이가 약 네 배, 평균적인 답변 길이는 약 세 배 증가했습니다. AI가 싸지는 동안 사용의 깊이와 폭이 함께 커졌다는 말입니다.

알고리즘의 효율 향상으로 GPU가 잉여 자원이 되는 것이 아닙니다. 대신 병목을 옮기고 있습니다. 연산칩이 빨라지자 메모리, 광통신, 변전설비, 그리드 인입이 다음 제약으로 떠올랐습니다. AI 인프라 투자는 GPU 한 종류의 구매가 아니라 전력과 메모리, 네트워크를 함께 조달하는 시스템 경쟁입니다.

구형 AI 가속기의 경제적 수명도 예상보다 길어지고 있습니다. **CoreWeave는 2026년 2분기 실적 발표에서 A100 계약을 2029년까지 갱신했다고 밝혔습니다. Google은 출시된 지 7~8년 된 TPU도 완전 가동 중입니다**. 3~5년 뒤의 AI 칩의 잔존가치가 0이라는 가정은 틀리게 되었습니다. 최첨단 모델이 신형 칩을 먼저 가져가더라도 추론·파인튜닝·배치 작업은 가격에 따라 구형 칩으로 이동할 수 있습니다.

가장 최신의 AI 인프라의 전환점은 자본시장이 계산서를 받기 시작했다는 데 있습니다. 특히 **엔비디아는 BlackRock을 위시한 6개 금융기관과 함께 장기적으로 US$500bn 이상의 제3자 자본을 동원하는 플랫폼을 발표했습니다. 또한 OpenAI의 임차불이행 위험을 보증하기 위해 엔비디아는 최대 US$105bn을 지원하는 것을 공식화했습니다**. GPU 수요가 기업들의 설비투자 예산만으로 결정되던 단계에서 프로젝트금융과 리스, 고객 선납금이 공급능력을 결정하는 단계로 넘어가고 있습니다.

물론 자본이 무제한으로 공급되는 것은 아닙니다. 발표되고 있는 AI 인프라 프로젝트는 수 GW 단위입니다. 향후 연산 공급이 계속 늘려면 AI 모델 기업의 매출 성장뿐 아니라 총마진과 현금 회수 속도가 함께 개선돼야 합니다.

앞으로의 승자는 GPU를 가장 많이 사는 회사가 아닙니다. 계약전력을 가장 빨리 가동전력과 매출로 바꾸고, 순이자비용을 감당하며, 고객 선납금과 "저비용" 자본을 확보하는 회사입니다. 그래서 2부의 결론은 꽤 단순명확합니다. **"AI 효율 향상도 인프라 수요를 없애지 않는다. 오히려 더 많은 자본을 조달하고 설비를 더 빨리 가동하려는 경쟁을 부추긴다."**

I. 토큰은 싸졌지만 GPU는 놀지 않는다

1. 기름값과 연비

지난 1부 보고서의 골자는 '모델은 작아지고 연산은 커진다'는 것이었다. 그렇다면 모델 효율이 높아지고 AI의 가격이 내려가는데 반도체는 왜 더 필요한 걸까?

자동차로 생각하면 쉽다. 기름값이 40% 내려가고 연비가 두 배 좋아지면 1km를 달리는 비용은 크게 낮아진다. 그러나 사람들이 더 멀리, 더 자주 이동하고 새 운전자가 도로에 들어오면 전체 연료 소비는 줄지 않을 수 있다.

AI도 마찬가지다. GPU 한 장이 만드는 토큰이 늘면 토큰 하나의 원가는 내려간다. 하지만 기업이 더 긴 문서를 읽히고, 답을 여러 번 검증하고, 에이전트에게 수십 단계의 작업을 맡기면 요청 수와 요청당 연산량이 함께 늘어난다. 결국 총수요는 효율 하나로 결정되지 않는다. "총연산 수요 = 요청 수 × 요청당 연산량"

한 번의 추론이 싸졌다는 사실만으로 GPU가 남아 돌 것이라고 결론 내릴 수 없는 이유다.

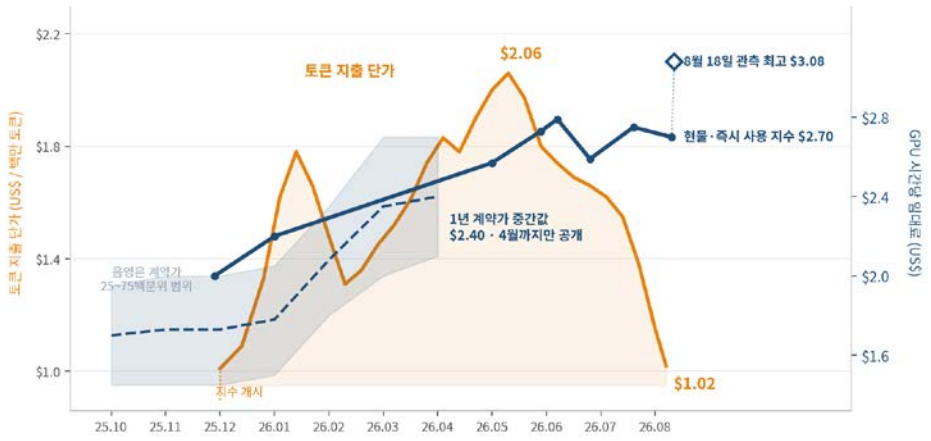
2. 값은 떨어졌는데 임대료는 올랐다

토큰 지출 단가 지수
AI에게 질문하고 답을 받을 때 쓰는 기본 단위인 '토큰' 100만 개당, 시장이 실제로 평균 얼마를 내고 있는지를 매일 계산한 가격. 싼 오픈소스 모델을 많이 쓰면 지수가 내려가고, 비싼 최신 모델을 많이 쓰면 상승. 공식 정가가 아니라, 사람들이 어떤 모델을 얼마나 쓰는지까지 반영한 실효 단가.

AI의 가격이 내려갔다는 데에는 이견이 없다. 같은 수준의 지능을 제공하는 API 가격은 세대가 바뀔 때마다 크게 낮아졌다. 그런데 그 지능을 만드는 물리적 자원의 가격은 반대로 움직였다.

실제 사용자가 낸 가격부터 보자. AI 관련 데이터 어그리게이터 업체인 Silicon Data에 따르면, '토큰 지출 단가 지수'는 2025년 12월 백만 토큰당 US$1.02에서 2026년 5월 US$2.06까지 올랐다가 지난 8월 16일 US$1.02로 되돌아왔다. 여덟 달 동안 두 배가 됐다가 제자리로 온 것이다.

그림 1. 토큰 값은 임대료와 함께 올랐다가 최근에야 갈라졌다
API 정가가 내려가는 것과 사용자가 실제로 내는 돈이 움직이는 방향은 다르다.
토큰이 여덟 달 만의 최저를 찍은 바로 그 시점에 임대료는 한 달 만의 최고를 찍었다.



자료: Silicon Data, SemiAnalysis, 미래에셋증권 리서치센터

GPU 현물·즉시 사용 지수
지금 당장 빌려 쓸 수 있는 엔비디아 H100 GPU 한 장을 1시간 빌리는 데 시장에서 실제로 얼마를 내는지 보여주는 일간 시세. 장기 계약이 아니라, 전문 임대업체들의 즉시(온디맨드) 가격을 표준화한 값. 칩만 있는 게 아니라 전력·냉각·네트워크까지 갖춘 '바로 쓸 수 있는' GPU가 얼마나 부족한지 이 가격에 표시.

반면 'GPU 현물 및 즉시 사용 지수'는 8월 17일 US$2.70이었고, 8월 18일 관측 최고가는 US$3.08로 더 치솟았다. GPU 임대료가 올랐다는 말이다.

그림 1에서 보듯이 토큰값 하락이 곧바로 연산자원 가격 하락으로 이어지는 것이 아니다. AI 서비스 가격은 모델 기업의 경쟁과 사용자가 고른 모델에 달려 있다. 반면 GPU 임대료는 전력과 메모리, 네트워크까지 갖춘 완성형 클러스터가 얼마나 부족한지에 달려 있다.

3. 값이 왜 떨어졌나

우선 토큰값이 내려간 이유도 나눠 봐야 한다. 작은 모델로 같은 일을 처리하는 알고리즘, 같은 칩을 더 잘 쓰는 소프트웨어, 신형 GPU의 성능, 공급자 사이의 가격경쟁이 함께 작용한다.

최근 지수가 내려간 직접적인 이유는 먼저 OpenAI의 가격 인하와 새 오픈소스 모델들의 등장으로 추정된다. 물론, AI 알고리즘의 효율이 좋아진 것도 맞다. 같은 칩에서 더 많은 토큰을 뽑아내는 기법이 계속 나왔기 때문이다. 다만 그 개선폭을 업계 평균으로 잰 공개 자료는 없다. 특정 모델과 특정 조건에서 몇 배가 좋아졌다는 사례는 많이 있지만 측정 조건이 제각각이라 하나의 시계열로 이어 붙일 수 없다.

더 중요한 것은 알고리즘의 효율 개선이라는 게 곧바로 가격 인하로 이어지지 않는다는 점이다. GPU 임대료는 칩뿐 아니라 위치, 전력, 네트워크, 계약기간까지 모두 반영한다. 그래서 투자자로서 집중해야 할 것은 토큰 가격이 몇 배 싸졌는지라기 보다는, 가격이 내려가면 무엇을 얼마나 더 쓰게 되는지다.

4. 물량과 요청당 연산량이 함께 늘었다

Ramp
미국 기업들이 법인카드와 경비로 실제로 얼마를 썼는지를 모아 보는 지출 관리 회사. 설문이나 추정이 아니라 실제 카드 결제액 기준. 약 7만 개 고객사의 결제 내역에서 ChatGPT, Claude, API 사용료 같은 AI 관련 결제만 따로 골라 집계.

Ramp가 7만여 고객사의 결제 데이터를 분석했다. AI에 지출하는 기업의 직원 1인당 월간 지출 중앙값은 2024년 1월 US$2.86에서 2026년 7월 US$11.95로 약 4.2배 늘었다. 게다가 같은 기간 지출 상위 10% 기업군 안에서의 중앙값은 US$611이었다. 전체 중앙값보다 50배 이상 높다. 평균 기업이 선도기업의 사용 수준으로 올라갈 여지가 그만큼 크다는 뜻이다.

그림 2. 값이 싸지는 동안 지출은 급격히 늘었다? 사용량이 가격보다 빨리 늘었다는 것!
상위 10% 기업군의 중앙값이 전체 중앙값의 약 51배.

자료: Ramp AI Index, 미래에셋증권 리서치센터

OpenRouter
여러 AI 모델을 한 곳에서 골라 쓸 수 있게 해주는 중개 플랫폼. 개발자와 기업이 여기서 실제로 보낸 질문·답변 길이와 용도(코딩 비중 등)가 집계됨.

기업들이 수행하는 AI 요청 하나하나도 더 무거워졌다. OpenRouter에 따르면, 평균적인 프롬프트가 약 1,500토큰에서 6,000토큰 이상으로 길어졌다고 한다. 평균 출력은 약 150토큰에서 400토큰 수준으로 늘었다. 전체 시퀀스 길이는 20개월 동안 2,000토큰 미만에서 5,400토큰 이상으로 증가했다. 코딩의 비중도 2025년 초 11%에서 최근 50% 이상으로 올라갔다.

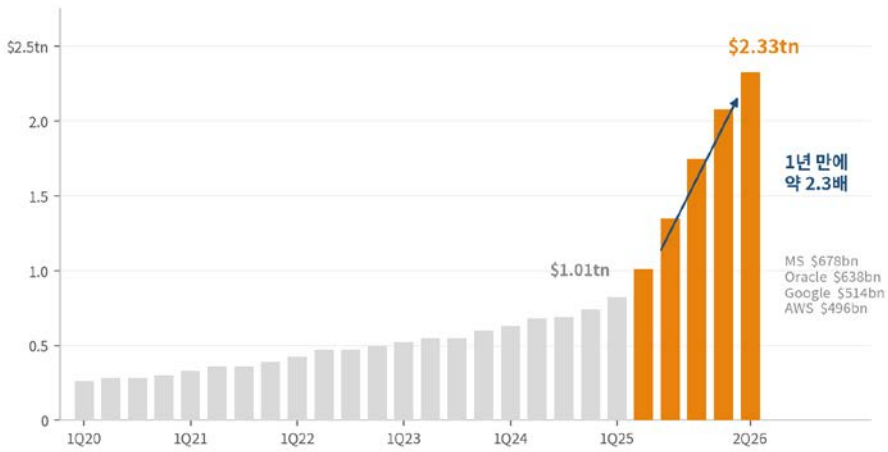
KV cache
AI가 긴 글을 읽을 때, 이미 읽은 앞부분을 다시 처음부터 계산하지 않도록 잠시 저장해 두는 맥락 메모. 문맥이 길어질수록 이 메모가 커져서 메모리를 더 많이 잡아먹음.

긴 문맥은 입력을 읽는 연산과 KV cache를 늘린다. 긴 답변은 출력 토큰을 늘린다. 코딩과 에이전트는 한 번 답하고 끝나지 않는다. 그때그때 필요한 도구를 호출하고 결과도 검증하며 실패하면 다시 시도한다. 쉽게 말해 요청 숫자만 늘어나는 게 아니라 요청 하나가 쓰는 연산도 커진다.

Ramp와 OpenRouter의 두 자료가 함께 보여주는 것은 뚜렷한 방향이다. AI가 싸지는 동안 기업 지출과 요청의 깊이가 함께 늘었다. 대형 클라우드의 계약잔고도 같은 방향을 가리킨다. Microsoft의 상업 RPO, Oracle의 RPO, AWS의 backlog, 구글 Cloud의 backlog를 단순 합산하면 2Q25 약 US$1tn에서 2Q26 US$2.3tn으로 늘었다. 기업은 아직 켜지지 않은 연산용량까지 장기간의 다년계약으로 먼저 확보하고 있는 것이다. 클라우드 사업자는 이 계약을 근거로 부지와 전력, 변압기와 GPU를 선점한다. 수요가 설비투자보다 먼저 계약되고, 그 계약이 다시 다음 설비투자를 당기는 구조다.

RPO, Backlog
이미 계약은 맺었지만 아직 서비스를 제공하지 않아 매출로 잡지 않은 금액. "앞으로 일해 주고받을 돈"의 잔고. 주문이 들어왔다는 뜻이지, 지금 당장 현금이 들어온 것은 아님.

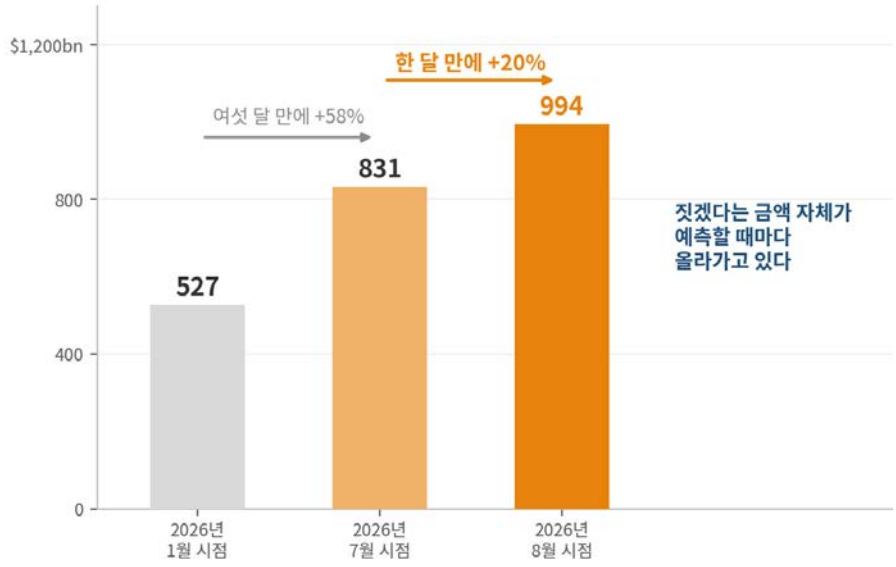
그림 3. 마이크로소프트의 상업 RPO, 오라클의 RPO, AWS와 구글 클라우드의 backlog를 합산. 이미 쓴 돈이 아니라 앞으로 쓰기로 서명한 돈이 1년 만에 약 2.3배가 됐다. 결제 데이터가 과거의 지출이라면 이 숫자는 미래의 지출.



자료: 각사 공시, 미래에셋증권 리서치센터

2027년에 대한 설비투자 전망치도 계속 올라간다. 상위 다섯 회사의 합산 전망치는 2026년 1월 US$527bn에서 7월 US$831bn, 8월 US$994bn으로 높아졌다. 여섯 달 동안 58%, 마지막 한 달에만 20% 상향된 것이다.

그림 4. 2027년 합산 설비투자 전망치가 계속 올라간다(스페이스X는 제외)
여섯 달 만에 58%, 마지막 한 달에만 20% 상승. 짓겠다는 금액이 예측할 때마다 커지고 있다. 수요가 계획을 앞지르고 있다는 신호.



자료: Morgan Stanley, 미래에셋증권 리서치센터

5. 병목은 위로 올라갔다

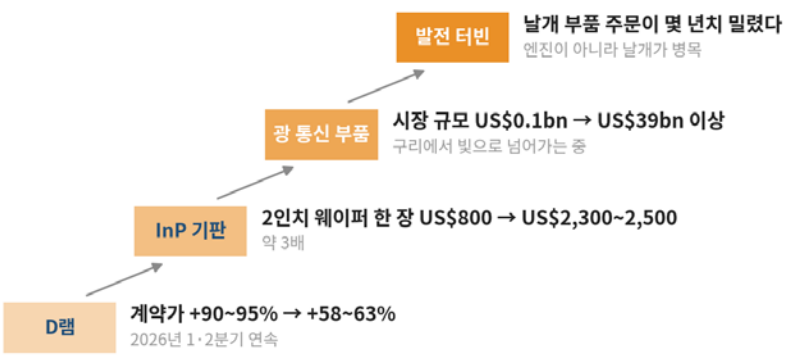
사용량과 GPU 공급이 함께 늘어남과 동시에 각종 제약 사항이 도사리고 있다. 연산칩에 데이터를 넣고 결과를 꺼내는 시스템의 나머지 거의 모든 부분에서 병목이 발생 중이다.

표 1. 여러 가지 병목 사항들

병목	확인된 변화	투자자가 읽어야 할 의미
DRAM 신규 공급	의미 있는 신규 생산능력은 2027년 말 이후에 집중	가격이 내려가기까지 공급 시차가 김
CPO·NPO	시장규모 2025년 약 US$100mn → 2030년 US$40bn 전망	랙 내부와 랙 간 통신이 연산성능의 핵심으로 이동
InP 기판	2인치 웨이퍼 가격이 2025년 약 US$800 → 2026년 4월 US$2,400 보도	광통신 원재료의 공급제약이 가격으로 전이
전력	계약전력과 가동전력 사이에 수년의 시차	GPU 확보보다 그리드 연결과 시운전이 매출 시점을 결정
냉각·변전	고밀도 랙 확대로 설계 난도 상승	기존 데이터센터의 단순 전환이 어려워짐

자료: TrendForce, 미래에셋증권 리서치센터

그림 5. 병목은 사라지지 않고 자꾸 위로 올라간다
GPU가 빨라질수록 그 GPU를 먹여 살릴 메모리와 대역폭과 전기가 더 필요해진다. AI 인프라의 단위는 GPU 한 장이 아니라 전력에 연결된 완성형 클러스터다.



자료: 외신 종합, 미래에셋증권 리서치센터

wafer, packaging, test와 장비 인증을 거쳐 실제 물량이 나오기까지 시간이 걸린다. 메모리에서 시작한 공급제약이 광통신과 전력설비로 올라가는 과정이다. 즉, AI의 효율 향상으로 인프라가 덜 필요해진 것이 아니라는 실물적 증거는 계속 나오고 있다. 더 큰 모델과 긴 문맥, 복잡한 추론을 쓸 수 있게 되면서 병목이 메모리와 광통신, 전력으로 옮겨갔다. AI 인프라의 단위는 GPU 한 장이 아니라 전력에 연결된 완성형 클러스터임을 유념해야 한다.

II. A100은 죽지 않았다

1. Bear들의 논리

GPU 투자에 대한 약세론의 핵심은 짧은 '자산수명'이다. 새로운 AI 가속기 아키텍처가 1~2년마다 나오니 구형 칩은 빠르게 가격을 잃는다는 주장이다. 4~5년짜리 대출로 지은 데이터센터가 대출 만기 전에 경쟁력을 잃을 수 있다는 말이기도 하다.

자산수명은 감가상각과 자산 취득에 따른 이자부담을 동시에 건드린다. GPU를 쓸 수 있는 기간이 짧으면 매년 더 많은 감가상각비를 인식해야 한다. 돈을 빌려주는 대주가 인정하는 담보가치도 낮아지기 때문에 같은 매출을 올려도 주주에게 남는 이익은 줄어든다.

문제는 몇 년 뒤의 중고가격을 지금 알 수 없다는 데 있다. 신형 GPU가 구형을 얼마나 대체할지, 추론수요가 구형 칩으로 얼마나 내려올지에 따라 값이 달라진다. 그래서 약세론은 3~5년 뒤의 AI 가속기 잔존가치를 대개 0으로 두었다. 그러나 이제 그 가정을 흔드는 실제 계약들이 속속 나오기 시작했다.

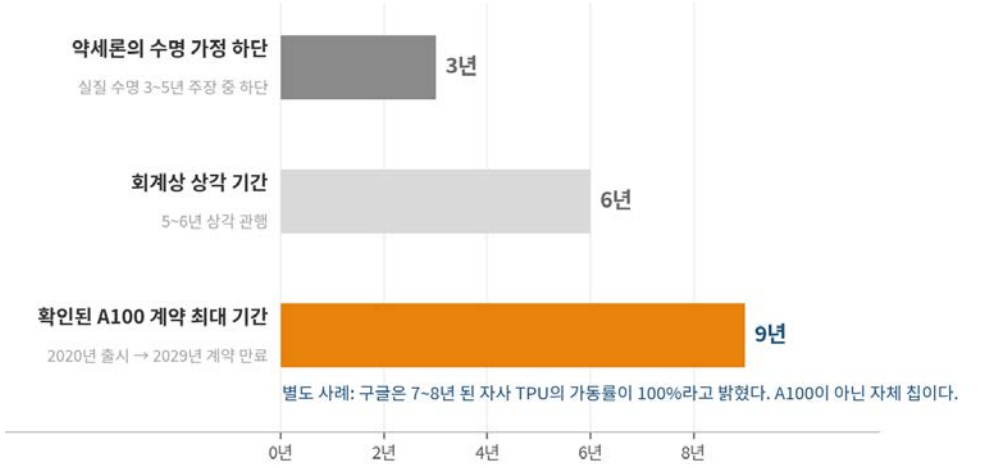
2. 2020년산 칩의 2029년 계약

네오클라우드
빅테크가 아니라, GPU를 빌려주는 데 특화된 전문 임대업체. CoreWeave, Nebius처럼 칩·전력·냉각을 묶어 '바로 쓸 수 있는' AI 컴퓨터를 시간 단위로 판매. 대형 클라우드보다 단가가 낮고 가격이 빨리 움직이는 편.

대표적인 네오클라우드 업체인 CoreWeave는 2026년 2분기 실적설명회에서 엔비디아의 구형 GPU인 "A100" 계약을 2029년까지 갱신했다고 밝혔다. A100은 2020년에 출시됐다. 출시 9년차에도 사용료를 받는 계약이 존재한다는 말이다.

다만, 이 계약 하나로 모든 A100이 9년 동안 같은 가격을 받는다고 말할 수는 없다. 그래도 의미는 분명하다. 3~5년 뒤 A100의 잔존가치가 0이라는 가정에 강력한 반례가 생겼다. 모든 GPU의 수명을 증명할 필요는 없다. 신용모형의 0원 가정을 깨는 데에는 한 건의 계약으로도 충분하다. 재계약 기간이 길어질 수록 대주는 담보의 회수기간과 회수율을 더 길게 잡을 수 있다.

그림 6. 3~5년 뒤 잔존가치가 0이라는 가정은 이미 깨졌다
2020년에 나온 칩이 2029년까지 사용료를 받는 계약이 존재한다.
신용모형의 0원 가정이 깨지면 담보가치가 생기고, 담보가치가 생기면 조달비용이 내려간다.



자료: CoreWeave, Google, 미래에셋증권 리서치센터

물론 CoreWeave는 GPU를 빌려주는 회사기 때문에 자산수명을 길게 말할 유인이 있다. 그래서 이해관계가 다른 사업자의 증언도 필요하다. 이와 관련해서는, 구글의 AI 인프라 총책임자 Amin Vahdat은 이미 2025년 공개 대담에서 "출시된 지 7~8년 된 TPU가 완전 가동 중"이라고 설명한 바 있다는 것을 떠올리면 된다. 신형 칩으로 연산의 최상단을 올리고, 구형 칩은 그 아래 가격대의 작업을 오래 붙잡게 하는 셈법이다.

3. 구형 값도 오르고 있다

오래 쓸 수 있다는 것과 높은 가격을 받을 수 있다는 것은 다른 말이다. 그러나 네오클라우드 업체 중 가장 큰 시가총액을 지닌 Nebius는 2026년 2분기 구형 GPU의 가격이 전분기보다 30% 이상 올랐다고 밝혔다. 구형 칩이 폐기자산이 아니라 가격에 따라 수요가 붙는 대체재라는 점을 확인할 수 있다.

최첨단 학습은 가장 빠른 GPU를 원한다. 반면 추론, 임베딩, 배치 처리, 연구개발 환경은 절대성능보다 시간당 가격과 가용성을 더 본다. 따라서 신형 GPU가 최첨단 학습으로 올라가면 추론 작업은 한 세대 아래 칩으로 내려올 수 있다.

최근에는 사용자가 응답속도와 가격을 직접 고르는 서비스도 늘고 있다. 조금 느려도 싼 경로를 고르면 작은 모델과 구형 GPU로 작업을 보낼 수 있는 것이다. 구형 가속기는 신형과 정면으로 경쟁하지 않으면서, 가격과 (답변) 지연시간을 맞바꾸는 시장에서 계속해서 수명을 늘리고 있다.

4. 감가상각이 끝나도 현금은 남을 수 있다

그래서 GPU의 경제적 수명을 회계상 감가상각연수와 같다고 보면 안 된다. GPU 재계약 매출에서 전력과 유지보수, 네트워크, 영업비용을 빼고 현금이 남는지가 더 중요하다.

구형 가속기의 연장가치 = 재계약 매출 – 전력비 – 직접 운영비 – 추가 자본비용

구형 GPU의 시간당 매출이 내려간다 하더라도 운영비 항목들을 웃돌면 계속 현금을 만들 수 있다. 특히 초기 투자비의 상당 부분을 이미 회수했다면 낮아진 단가는 더욱 의미가 있다.

CoreWeave의 A100 계약이 2029년까지 이어진다는 사실은 첫 계약 이후에도 구형 GPU가 현금을 만들 수 있다는 실제 사례다. 이런 재계약이 반복되면, 대주는 3~5년 뒤의 GPU 가치를 0으로 놓기 어려워진다. 실제로도 CoreWeave의 "DDTL 5.0"은 그러한 사실을 보여준다. CoreWeave는 GPU와 고객계약을 담보로 US$3.1bn을 약 5.5년 만기로 조달했는데(금리는 SOFR+4.50%), 이는 컴퓨트가 장기 금융자산으로 인정받기 시작했다는 것을 내포한다.

종합하면, GPU의 잔존가치는 0이 아니라 실제 회수할 수 있는 현금흐름으로 바뀐다. 그러면 대주의 예상 손실률은 낮아지고 CoreWeave는 같은 자기자본으로 더 많은 설비를 지을 수 있다. 이제 구형 GPU의 수명 연장은 단순한 감가상각 문제가 아니라, 자본비용을 낮추고 증설능력을 키우는 금융상의 이점이 되고 있다.

DDTL 5.0(Delayed Draw Term Loan) 필요할 때마다 나눠 빌려 쓸 수 있는 장기 대출. 설비 공사 일정에 맞춰 돈을 인출하는 대출. CoreWeave가 2026년에 마련한 5번째 DDTL 대출로, GPU와 고객계약을 담보로 여러 은행이 돈을 대고, 나중에 그 대출을 시장에서 사고팔 수 있게 만든 구조.

III. 큰 청구서는 자본시장이 낸다

1. 장부를 열어 보면

그런데 AI 인프라가 자꾸 커지면 현금이 많은 기업도 모든 설비를 영업현금만 갖고는 지을 수 없다. SpaceX의 2026년 2분기 공시는 왜 외부자본이 필요할 수밖에 없는지를 보여준다.

표 2. SpaceX 2026년 2분기 실적자료 하이라이트

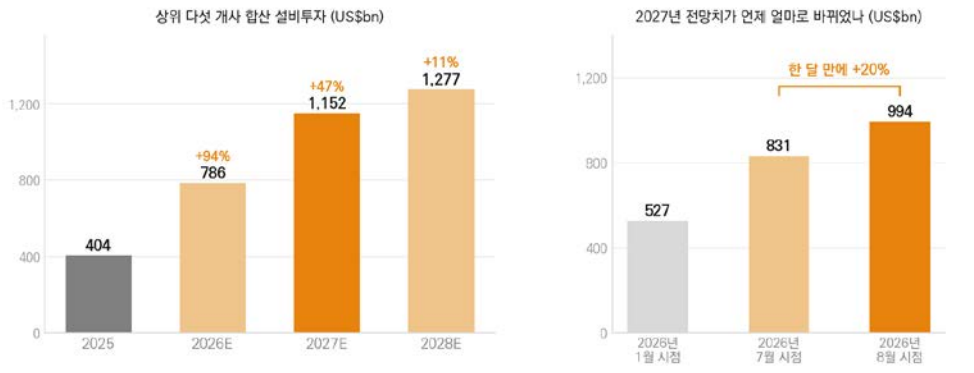
SpaceX 2026년 2분기	금액
전사 매출	US$7,814mn
통신 부문 매출	US$4,291mn
AI 부문 매출	US$2,561mn
우주 부문 매출	US$962mn
AI 부문 영업손실	-US$1,257mn
AI 부문 조정 EBITDA	**US$1,146mn**
AI 부문 설비투자	**US$15,828mn**
보유 연산용량	1.4GW
클라우드 계약 총액	US$14,100mn

자료: SpaceX, 미래에셋증권 리서치센터

AI 부문 설비투자는 같은 분기 매출의 무려 6.2배였다. 조정 EBITDA는 흑자지만 내부 현금만으로 증설을 감당할 수 있는 규모가 아니다. 부족한 돈은 주식과 채권, 리스로 메워야 한다. AI 인프라의 성장속도는 영업성과뿐 아니라 자본시장 접근성에도 달려 있다는 말이 된다.

절대금액도 빠르게 커지고 있다. 자본지출 상위 5개 회사의 합산 설비투자는 2025년 약 US$404bn에서 2026년 US$786bn, 2027년 US$1.152tn, 2028년 US$1.277tn으로 늘어날 전망이다.

그림 7. 짓겠다는 금액이 계속 올라가고 있다
2025년 US$404bn에서 2028년 US$1.277tn으로 세 배가 넘는다.
영업현금만으로 감당할 수 있는 규모가 아니다. 부족한 돈은 주식과 채권과 리스로 메워야 한다.
AI 인프라의 성장속도는 영업성과뿐 아니라 자본시장 접근성이 정한다.

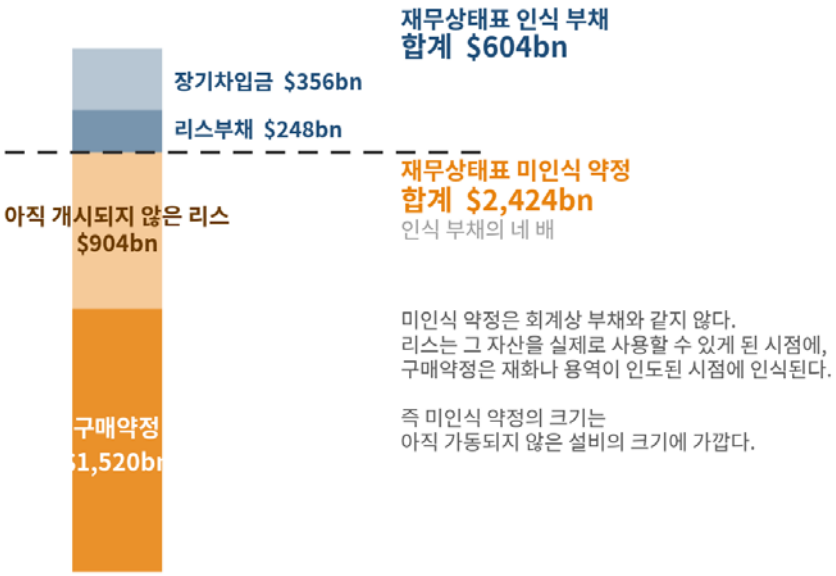


자료: 업계 종합, 미래에셋증권 리서치센터

리스
주로 데이터센터 건물·전력설비·서버 공간 등을 장기간 빌리고 지급하기로 한 임차료. 계약을 맺어도 시설을 아직 사용할 수 없다면 장부 밖에 존재.

또한 장부 밖에도 청구서가 있다. 빅테크들의 현재 재무상태표에 잡힌 금액만 봐서는 산업의 전체 청구서를 알 수 없다. 보도에 따르면 Alphabet·Amazon·Meta·Microsoft의 인식된 리스부채는 US$248bn, 장기차입금은 US$356bn이었다. 그런데 아직 시작되지 않은 리스는 US$904bn, 구매약정은 US$1.52tn에 달한다.

그림 8. 인식된 부채의 네 배가 장부에 잡히지 않는다
재무제표에 잡힌 것만 보면 산업 전체의 청구서를 알 수 없다.
미인식 약정의 크기는 아직 가동되지 않은 설비의 크기에 가깝다.
계약 서명은 끝냈는데 매출을 못 만드는 데이터센터가 그만큼 많다는 뜻이다.



자료: 각사, WSJ, 미래에셋증권 리서치센터

구매약정
서버·GPU·네트워크 장비 등을 앞으로 일정 금액만큼 사겠다고 공급자와 맺은 계약. 계약은 끝났지만 물건을 아직 받지 않았다면 회계상 부채나 자산으로 잡히지 않음.

그런데 장부 바깥에 있다는 것은 아직 회계상 부채로만 잡히지 않았다는 뜻이지, 안 쓰기로 한 돈이라는 뜻은 아니다. 계약서에는 이미 서명이 끝났다. 회계에는 인식 시점 규칙이 있기 때문에 리스는 그 건물을 실제로 쓸 수 있게 된 날부터 부채로 잡힌다. 지금 짓고 있는 데이터센터의 임차료는 계약서에는 적혀 있어도 재무제표에는 안 들어간다. 구매약정이라는 것도 물건이 들어와야 잡힌다. 서버를 주문만 하고 아직 못 받았으면 장부에는 아무것도 없다. 다시 말해 데이터센터를 짓기 전에는 장기계약과 구매약정으로 비용을 미리 확정한다. 시설이 가동되면 그 금액이 리스부채와 자산으로 재무제표에 잡힌다.

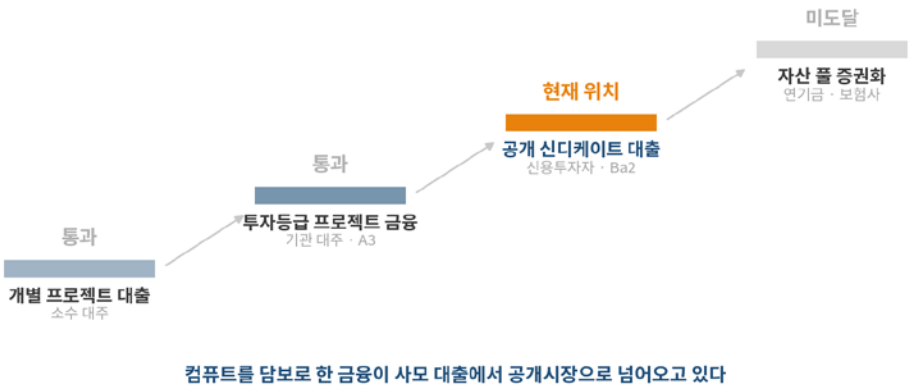
그래서 이 off-balance 숫자는 부채라기보다 발주서에 가깝다. 아직 개시되지 않은 리스 US$904bn과 구매약정 약 US$1.5tn을 합친 US$2.4tn 정도의 금액은 앞으로 몇 해에 걸쳐 지어지고 가동될 설비의 목록이다. 재무상태표에 잡힌 US$604bn의 약 네 배다. 결국 시장이 재무제표에서 보고 있는 AI 인프라 청구서는 이미 확정된 전체 금액의 일부일뿐이다. AI 인프라 투자주기의 실제 크기는 현재 재무제표에 보이는 것보다 훨씬 크다는 점을 알아야 한다.

2. GPU가 담보가 되기까지

돈을 싸게 빌리려면 대주가 담보가치를 계산할 수 있어야 한다. GPU는 오랫동안 이 조건을 충족하기 어려웠다. 건물과 변전소, 냉각설비는 수십 년을 쓰지만 자산가치의 큰 부분을 차지하는 GPU는 교체주기가 짧았기 때문이다.

GPU가 담보가 되려면 세 가지가 필요하다. "일정 기간 현금을 벌어야 하고, 몇 년 뒤 중고 가격을 계산할 수 있어야 하며, 고객이 돈을 갚지 못하면 다른 사용자에게 넘겨 다시 수익을 낼 수 있어야 한다." 이전 장에서 봤던, A100의 2029년 재계약은 첫 번째와 두 번째 조건을 뒷받침한다. 오래된 GPU도 계속 임대료를 받고, 그 계약을 바탕으로 미래가치를 계산할 수 있다는 뜻이다.

그림 9. 컴퓨트는 대출시장까지 왔고 증권시장은 아직이다
컴퓨트를 담보로 한 금융이 사모 대출에서 공개시장으로 넘어오고 있다.
남은 단계는, 여러 자산을 묶어 증권으로 발행하는 단계다.
여기가 열리면 연기금과 보험사의 돈도 들어올 수 있다.



자료: 미래에셋증권 리서치센터

CoreWeave의 DDTL 4.0은 여기서 한 단계 더 나아갔다. 투자등급 고객과의 계약을 담보로 US$8.5bn을 빌렸으며(금리는 SOFR+2.25%), CoreWeave가 아니라 해당 프로젝트의 현금흐름으로 갚는 비소구 대출이었다. 고객계약과 GPU가 독립적인 대출자산으로 인정받기 시작한 것이다.

공개 신디케이트 대출
한 은행이 아니라 여러 은행·투자자가 돈을 나눠 빌려주는 대출. 공개'는 소수와만 조용히 계약하지 않고, 더 많은 투자자에게 조건을 알리고 참여를 받는다는 뜻. 나중에 그 대출 지분을 시장에서 사고팔 수 있어, 은행만의 돈이 아니라 더 넓은 자본시장 돈으로 변환되는 것.

거기에 DDTL 5.0은 대출을 받을 수 있는 범위를 더 넓혔다. 투자등급이 아닌 두 고객의 계약까지 담아서 US$3.1bn을 조달했다. 금리는 SOFR+4.50%, 등급은 Moody's 기준 Ba2였다. 위험이 더 큰 만큼 금리는 높아졌다. 그러나 여러 금융기관이 공동으로 자금을 대고, 대출채권을 기관 간에 양도·거래할 수 있게 설계된 신디케이트 대출이라는 점이 포인트다. 쉽게 말해 대출을 유통시장에서 사고팔 수 있게 된 것이다.

이는 금융시장이 "GPU 및 고객과의 컴퓨팅 임대계약"의 위험을 평가하고 거래하기 시작했다는 뜻이다. 이와 관련해 대주는 고객이 부도나더라도 GPU를 회수해 다른 고객에게 넘길 수 있다는 가능성까지 대출가격에 반영하게 된 것이다. GPU가 담보가 되기 위한 세 번째 조건(실제 부도 이후 다른 고객을 붙여 다시 임대료를 받는 재임대)이 거의 완성된 것이다.

이제 다음 단계는 여러 데이터센터와 GPU 계약을 하나로 묶어 연기금과 보험사에 파는 증권화/자산유동화다. 개별 고객의 위험을 여러 자산으로 분산하면 더 많은 장기자금이 들어올 수 있다. 즉, CoreWeave를 위시한 AI 클라우드 업체들은 점점 더 "컴퓨트 금융을 은행의 개별 프로젝트 대출에서 더 넓은 공개자본시장"으로 옮기고 있다.

3. 엔비디아가 선물한 "8월의 크리스마스"

엔비디아는 이제 GPU만 파는 것이 아니라, 데이터센터를 지을 돈까지 끌어오는 역할을 하고 있다. 지난 8월 11일, BlackRock을 비롯한 여섯 개의 금융기관은 프로젝트별 심사를 거쳐 장기적으로 US$500bn 이상의 자본을 공급해줄 계획으로 엔비디아와 협약을 맺었다. 이때 엔비디아는 GPU 예상잔존가치의 최대 25%까지는 지원하기로 했다. 예를 들어 나중에 GPU가 예상보다 싸게 팔리면, 그 부족분 중 일부를 엔비디아가 메운다는 보증이다. GPU 가격이 예상보다 많이 떨어져도 엔비디아가 손실의 일부를 부담하므로 금융기관은 더 많은 돈을 빌려줄 수 있고, 개발사(데이터센터를 짓는 회사)가 직접 넣어야 하는 자기자본은 줄어든다. 결국 금융회사의 제일 큰 걱정(중고 GPU 값 폭락)을 일부 떠안아, AI 공장을 짓는 돈을 더 잘 끌어오게 만든 장치다.

게다가 이 금융 플랫폼과 별개로, 엔비디아는 8월 17일 미국 오하이오주에서 또 다른 거대 계약을 발표했다. 11일 협력이 외부 기관자본을 끌어오는 장기적인 금융망이라면, 17일 건은 엔비디아가 직접 신용보강에 나선 개별 프로젝트다. 두 거래는 서로 다르지만, 데이터센터 건설자금을 늘리기 위해 엔비디아가 금융시장과 자신의 대차대조표를 동시에 활용하기 시작했다는 공통점을 가진다.

17일 공시 내용을 보면, 계약은 SB Energy가 데이터센터를 짓고 OpenAI가 장기간 빌려 쓰며 "엔비디아가 OpenAI의 임차료 지급을 보강"하는 구조다. 초기 개발용량은 약 4.25GW이고, 엔비디아가 부담할 수 있는 누적 지급액의 상한은 US$105bn이다. 보증은 시설이 완공돼 개별 임대가 시작되는 2028년 이후부터 적용된다.

표 3. 엔비디아가 이번 달에 발표한 두 가지 보증

구분	8월 11일 플랫폼	8월 17일 Ohio 공시
구조	6개 금융기관의 제3자 자본 동원	특정 프로젝트의 조건부 지급의무
규모	장기 목표 US$500bn 이상	초기 4.25GW, 상한 US$105bn
엔비디아 역할	일부 거래에서 최대 25% 잔존가치 지원 가능	OpenAI 불이행 시 잔존가치 부족분 부담
위험 통제	프로젝트별 독립심사	재임대·매각·리스 인수 및 OpenAI 상환청구

자료: 미래에셋증권 리서치센터

돈의 흐름은 간단하다. OpenAI가 임차료를 내지 못하면 임대인은 먼저 시설을 다른 고객에게 빌려주거나 매각한다. 그래도 회수한 금액이 계약서에 정한 최소보장가치보다 적으면 엔비디아가 차액을 메우는 것이다. 엔비디아는 필요하면 리스를 직접 넘겨받거나 새로운 임차인을 찾아 재임대·매각하도록 요구할 수 있다. 이후 OpenAI는 엔비디아가 대신 지급한 금액을 갚아야 한다.

중요한 것은 OpenAI가 임차료를 내지 못해도 임대인과 금융기관의 회수액에는 어떠한 바닥이 생겼다는 점이다. 이 보증이 있기 때문에 땅과 전력, 건물에서 나오는 장기 임대수입을 담보로 더 많은 돈을 빌릴 수 있다. 즉, 엔비디아가 보강한 것은 이번에는 칩 가격이 아니라 OpenAI의 데이터센터 임차료다. GPU 판매를 늘리기 위해 데이터센터 건설자금에까지 자신의 신용을 얹기 시작한 것이다. 엔비디아의 신용이 데이터센터 전체의 자본비용을 낮추는 도구로서 작동하고 있다는 게 현재의 흐름이다.

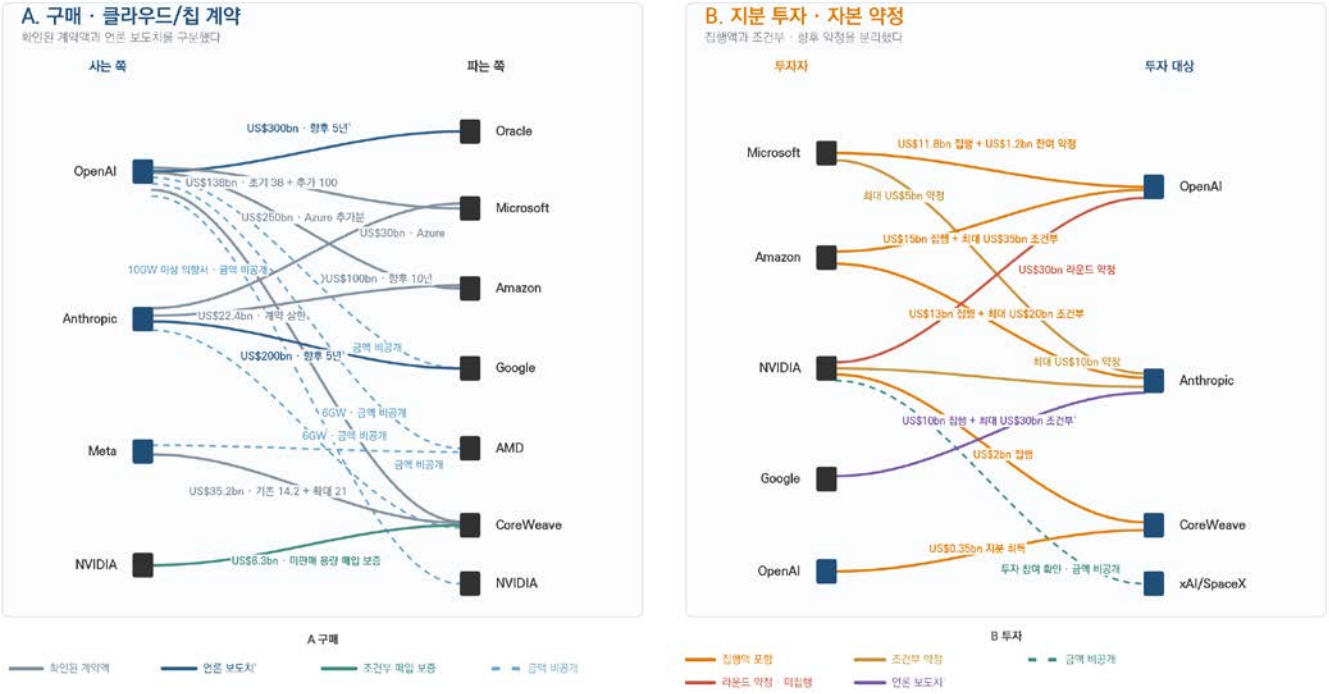
그림 10. 엔비디아가 무엇을 보증했나
일주일 만에 칩값 뿐만 아니라, 땅과 전기와 건물에까지 보증 범위를 확장했다.
오픈AI가 되갚을 능력이 없을 때 엔비디아의 보증이 발동된다.

자료: 엔비디아, 미래에셋증권 리서치센터

4. 순환인가, 새로운 클라우드인가

그런데 칩 공급자가 고객의 금융까지 지원하면 결국 자기 매출을 스스로 만드는 것 아니냐는 의심이 더 강해질 수밖에 없다. 타당한 우려다. 공급자의 보증이 고객의 구매력을 만들고 그 구매가 다시 공급자의 매출이 되면 최종수요보다 매출이 먼저 커질 수 있다는 의심이 피어난다.

그림 11. 같은 회사가 사는 쪽과 파는 쪽에 동시에 있고, 파는 쪽이 사는 쪽의 주주라는 것.
관계가 얽혀 있다는 것은 거래가 곧 허위수요라는 얘기는 아니다.



자료: BIS Working Paper, 각사, 외신 종합, 미래에셋증권 리서치센터

엔비디아의 보증이 고객의 구매력을 높이고 그 구매가 다시 엔비디아 매출로 돌아온다는 점에서는 순환성이 있다. 그러나 엔비디아가 판매대금 전체를 고객에게 빌려주는 구조는 아니다. 실제 건설자금은 여섯 금융기관이 공급하며, 이들은 "고객사 계약과 칩 가동률, 현금흐름, GPU의 중고가치와 재임대 가능성"을 프로젝트별로 각각 직접 다시 계산해 심사한다. 엔비디아의 보증이 구매력을 만들더라도 프로젝트는 외부 신용위원회를 통과해야 한다는 말이다. 고객이 실제 매출을 만들지 못하거나 다른 임차인을 구할 수 없는 프로젝트라면 엔비디아의 제한된 보증만으로 금융기관의 심사를 통과하기 어렵다. 칩 공급자의 매출을 그대로 대출로 바꾸는 구조와 다른 부분이다.

따라서 엔비디아의 보증은 존재하지 않던 최종수요를 만드는 돈이라기보다, 이미 있는 장기수요를 더 낮은 자기자본과 더 긴 대출로 "앞당겨 건설하게 하는" 신용보강으로 봐야 한다. 대신 엔비디아가 맡는 것은 GPU 가격과 고객 신용이 동시에 무너지는 'tail-risk'다.

현재까지 확인된 가격과 계약만 놓고 보면 엔비디아의 베팅은 합리적이다. 출시된 지 6년된 A100도 2029년까지 재계약됐고, 신형 GPU에는 공급부족 프리미엄이 붙고 있다는 것에서 알 수 있다. 엔비디아가 실제로 돈을 잃으려면 OpenAI같은 업체가 임차료를 내지 못하는 동시에 해당 GPU와 데이터센터를 다른 업체에도 빌려주지 못해야 한다. 그러나 이 두 위험이 한꺼번에 발생할 가능성은 현재로서는 적다.

5. AI DC에 적용되는 유동화 기법

사실 엔비디아의 금융기법은 항공기 리스와 닮았다. 항공사가 비행기를 장기간 빌려 쓰다가 파산해도 소유자는 그 비행기를 회수해 다른 항공사에 다시 빌려줄 수 있다. 고객 한 곳이 사라져도 자산이 계속 돈을 벌 수 있기 때문에 금융기관은 비행기를 담보로 장기간 대출해 준다.

물론 데이터센터는 항공기처럼 다른 장소로 옮길 수 없다. 대신 고객이 직접 시설에 올 필요도 없다. 서버와 네트워크, 운영 소프트웨어가 표준화돼 있다면, 기존 고객이 떠나도 다른 고객이 원격으로 같은 시설을 사용할 수 있다. 엔비디아가 AI 데이터센터의 레퍼런스 및 표준화에 사활을 거는 이유다.

그림 12. 컴퓨트가 담보 대열의 문턱에 섰다
자동차·항공기·부동산·데이터센터는 기존 고객이 떠나도 계속 현금을 벌 수 있다는 사실이 확인된 뒤 대출과 유동화 시장이 커졌다. 엔비디아는 표준화된 GPU와 소프트웨어, 잔존가치 지원을 결합해 컴퓨트에도 같은 금융구조를 만들려 한다.
실제로도 임대료가 이어진다면 컴퓨트 대출을 묶어 연기금·보험사에 팔 수 있고,
더 많은 자금이 데이터센터 건설로 들어오게 된다.

자료: 엔비디아, 미래에셋증권 리서치센터

AI 인프라를 담보로 한 대출시장은 이미 열렸다. 다음 단계는 고객이 바뀌어도 임대료가 계속 나오는 여러 데이터센터 대출을 하나로 묶어 증권시장에 판매하는 것으로 사료된다.

이런 사례가 늘어나면 금융기관은 고객이 부도났을 때도 더 많은 돈을 회수할 수 있을 것이다. 대출금리는 내려가고 만기는 길어지며, 연기금과 보험사 같은 장기투자자도 들어올 수 있다. 컴퓨트는 한 회사에 돈을 빌려주는 개별 대출을 넘어, 여러 투자자가 반복해서 사고파는 인프라 자산으로 발전하는 중이다.

IV. 상환능력 점검: 결국 청구서를 갚는 것은 AI 매출

1. 값은 원가가 아니라 용도가 정한다

금융기관이 데이터센터 건설자금을 빌려줄 수는 있다. 그러나 그 대출의 이자를 갚는 돈은 결국 AI를 사용하는 최종 고객에게서 나와야 한다. 고객이 AI 서비스에 돈을 내면 모델 기업이 클라우드와 GPU 사용료를 지급하고, 데이터센터 사업자는 그 돈으로 이자를 갚는다. 따라서 장기적인 GPU 수요를 결정하는 것은 AI 기업이 그 GPU로 얼마나 많은 매출과 이익을 만들 수 있는가다.

고객이 연산에 지불할 수 있는 가격도 AI의 용도에 따라 달라진다. 예를 들어, 기업용 코딩 에이전트나 고부가가치 영상 생성처럼 고객의 시간과 인건비를 크게 줄여주는 서비스는 같은 GPU를 사용해도 훨씬 높은 가격을 받을 수 있다. 영상 1초를 만드는 데 US$0.1~0.2가 들고 한 사용자가 하루 30분을 생성한다면 하루 연산비용은 US$180~360이다. 고객이 그보다 훨씬 큰 가치를 얻는 업무를 가진다면 더 높은 연산비용도 지불할 수 있다는 말이다.

문제는 데이터센터와 전력, 반도체 공급이 수요만큼 빨리 늘지 못한다는 점이다. 가동 가능한 컴퓨팅이 부족하면 더 많은 매출과 비용절감을 만드는 작업이 높은 가격을 지불하고 GPU를 먼저 차지한다. 그러나 최종 고객이 지불한 돈이 모두 인프라 사업자에게 돌아가는 것은 아니다. 클라우드는 '연산시간(컴퓨팅 사용권)'을 팔지만 모델 기업은 그 연산으로 만든 '대답(지능의 사용권)'을 판다. 최종 고객에게 가까운 사업자일수록 같은 연산에서 더 많은 가치를 매출로 가져갈 수 있다.

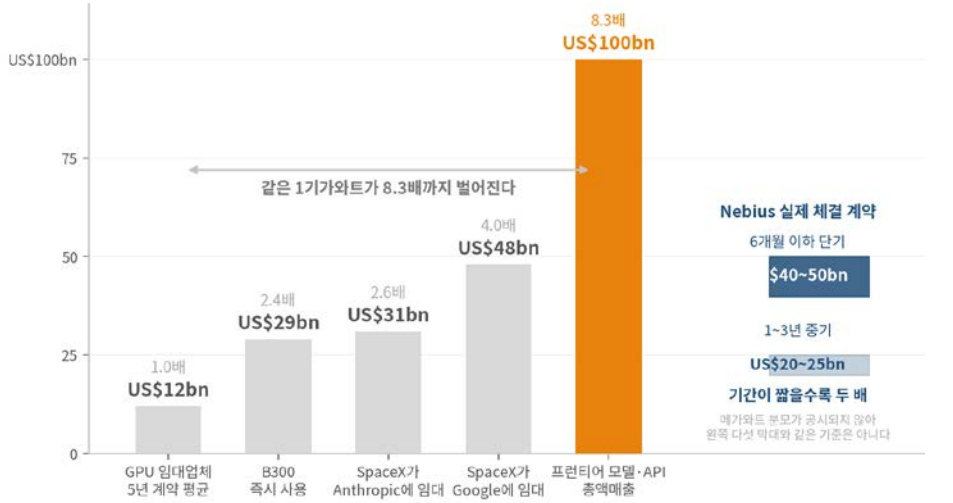
2. 같은 1GW도 매출은 엄청나게 다르다

따라서 같은 1GW를 누가 어떤 상품으로 파는지가 중요하다. 1GW 클러스터에 GPU 약 46만 개가 들어가고 GPU당 시간당 US$3를 받는다고 가정하면 장기 GPU 임대매출은 연간 약 US$12bn으로 환산된다. 그런데 같은 규모의 연산을, 모델과 API를 직접 판매하면 고객이 얻는 업무가치까지 가격에 포함된다.

SemiAnalysis의 비공개 모델 및 추론 시뮬레이터에 따르면,1GW의 연산으로 얻을 수 있는 연매출 상단은 약 US$100bn으로 산출된다. (GB300 클러스터로 OpenAI·Anthropic이 코딩/에이전트 작업을 위한 API 추론 서비스를 직접 파는 경우를 가정) 같은 1GW라도 연산시간을 파느냐, 그 연산으로 만든 지능과 업무결과를 파느냐에 따라 매출이 8.3배까지 벌어지는 것이다.

아래 그림에서 보듯 장기 자산임대에서 모델·API 직접판매까지 가치사슬의 각 층이 1GW에서 얼마의 매출을 가져가는지를 알 수 있다. 오른쪽으로 갈수록 고객이 지불하는 가격이 높아진다. 물론 양 극단의 금액차이인 US$88bn이 전부 모델업체의 이익은 아니다. 이 매출에서 연구개발비와 추론비용, 클라우드 사용료와 판매비용을 내야 한다. 이 비용을 모두 뺀 뒤 남는 총이익과 현금이 데이터센터 임차료와 이자를 갚는 재원이 된다.

그림 13. 같은 전기 1기가와트가 층에 따라 여덟 배까지 벌어진다
오른쪽은 실현 매출이 아니라 수요가 포화됐을 때의 1GW로 달성할 수 있는 매출의 이론적 상단.
왼쪽은 자산임대에 가깝고 위로 갈수록 모델과 소프트웨어의 가치가 붙는다.
네비우스의 계약금액을 보면, 즉시 쓸 수 있는 연산에는 실제로 프리미엄이 붙는다는 것을 확인가능.



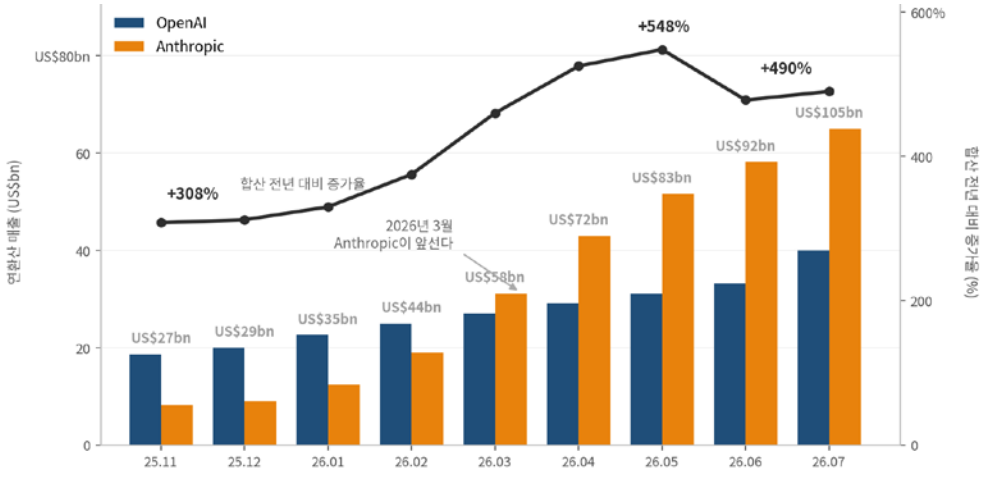
자료: Nebius, SemiAnalysis, 미래에셋증권 리서치센터

따라서, 그 매출에서 비용을 빼고 얼마가 남는지를 봐야 한다. API 가격경쟁으로 판매가격이 내려가거나 긴 추론과 에이전트 작업으로 연산비용이 빨리 늘면 이 "완충폭"은 줄어든다. 그러면 모델 기업이 감당할 수 있는 GPU 임대료가 낮아지고, 클라우드 계약가격과 신규 데이터센터의 대출조건도 함께 압력을 받는다. 모델 기업의 총마진이 AI 인프라 전체의 상환능력을 결정한다.

3. 현재 모델 매출은 설비계획을 얼마나 받쳐주나

이번 달 초까지 보도된 연환산 매출은 Anthropic이 US$65bn 이상, OpenAI가 약 US$40bn이었다. 두 회사의 합계는 최소 US$105bn이다. 2025년 11월 US$27bn에서 불과 여덟 달 만에 약 3.9배로 늘었다.

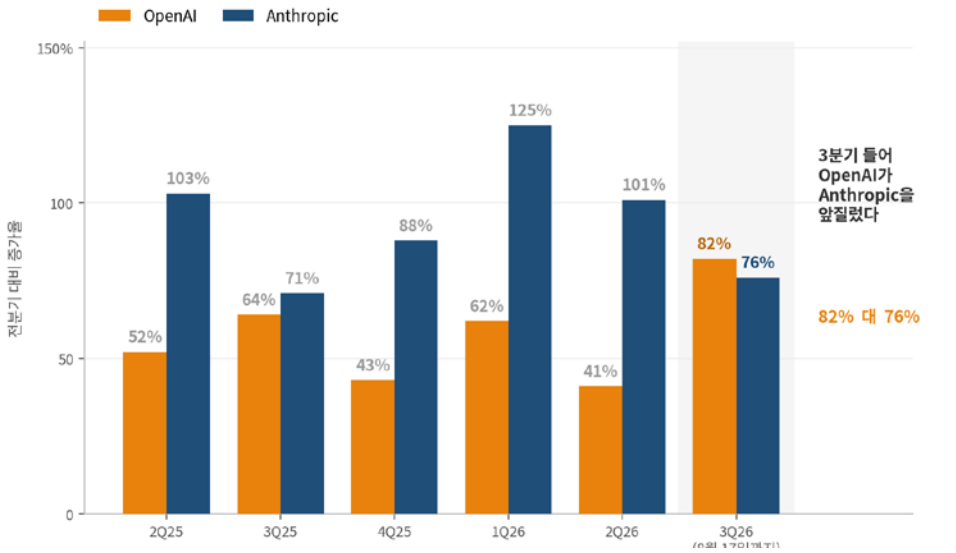
그림 14. 두 회사 합산 매출이 아홉 달 만에 네 배가 됐다.
2025년 11월 US$27bn에서 2026년 7월 US$105bn. 3분기 만에 약 4배.
다만 증가율은 꺾이기 시작했지만 다시 상승하는 국면.



자료: 각 사, 외신 종합, TMT Breakout, 미래에셋증권 리서치센터

특히 7월 합산 매출은 전월보다 약 14% 증가했고 전년 대비 증가율도 490%로 '다시' 높아졌다. 더 최근의 월매출을 연환산한 합산 추정치는 약 US$115bn까지 올라갔을 정도다. API와 구독, 기업계약에서 발생하는 프론티어 AI 모델업자들의 매출이 이미 연 US$100bn을 넘는 최종 수요를 만들고 있고, 성장속도도 다시 빨라지고 있다는 뜻이다.

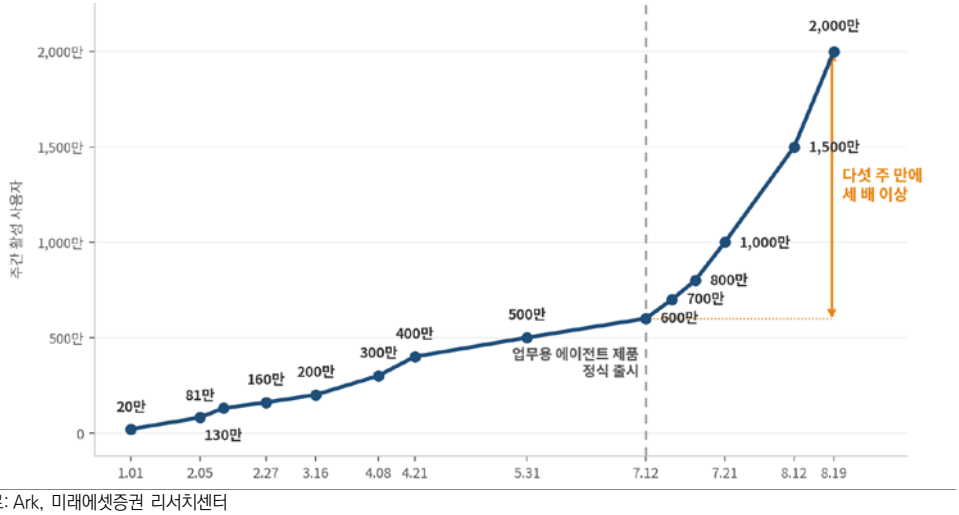
그림 15. 3분기 들어 OpenAI가 Anthropic보다 빨리 크고 있다
API로 연결된 기업의 지출만 집계했으며 구독 요금은 제외.
2Q25에 Anthropic 성장률이 두 배 앞섰고, 1Q26에 차이가 벌어졌다가, 3Q26에 역전됐다.
모델 값이 너무 비싸거나 데이터 보관 조건이 까다로우면 채택이 안 되기 때문이다.



자료: Ramp AI Index, 미래에셋증권 리서치센터

또한, ChatGPT Work와 Codex의 주간 활성 사용자가 8월 19일 2,000만 명에 도달했다고 알려진다. 업무용 제품이 정식 출시된 뒤의 폭발 성장이다. QTD 기준으로 ARR이 35% 넘게 증가했고 API 토큰 처리량은 두 배가 됐다. 사용자 숫자로는 ChatGPT 전체 활성 사용자 약 10억 명에 비해서는 2%에 불과하다. 아직 침투율은 낮은데 증가속도는 빨라지고 있다는 뜻이다.

그림 16. 오픈AI 업무용 챗봇과 코딩 에이전트의 주간 활성 사용자 합계
정식 출시 전까지는 완만했으나 출시 이후 다섯 주 만에 세 배가 넘게 늘었다.

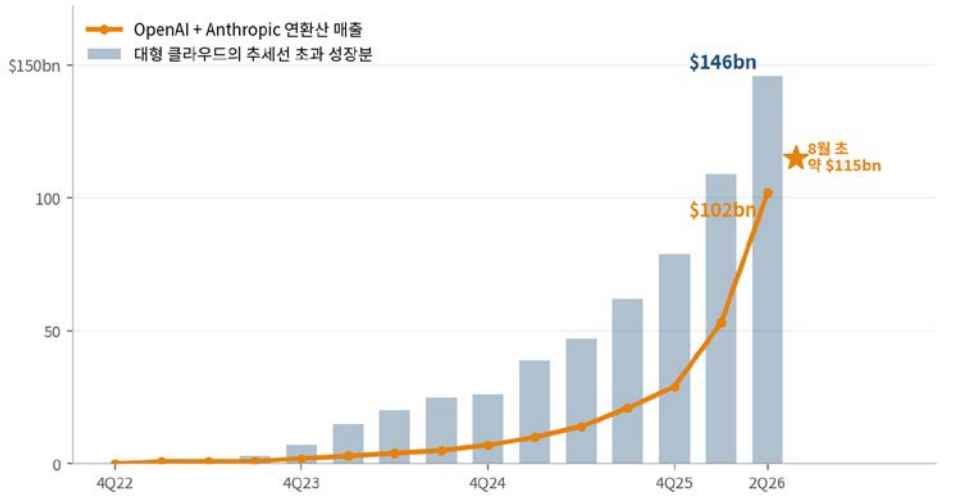


자료: Ark, 미래에셋증권 리서치센터

게다가, 유명 VC 투자자 Gavin Baker에 따르면, Anthropic의 ARR인 US$65bn은 보수적인 기준으로 재계산된 매출일 것이라고 한다. 즉, '총액 기준'에서 클라우드와 유통업체 몫을 뺀 '순액 기준'으로 바꿨다는 것이다. 이것이 사실이라면, 기존의 총액 기준으로 산식을 정렬을 해보면, ARR이 이미 US$80bn을 초과했을 것으로 추정된다.

한편, 3Fourteen의 추정에 따르면, 두 회사의 연환산 매출 US$115bn은 대형 클라우드 매출 가운데 ChatGPT 이전 추세를 웃도는 성장분과 비슷하다. 프론티어 모델 기업은 더 이상 클라우드의 작은 실험고객이 아니다. 현재 클라우드 성장의 상당 부분을 만들고, 다음 데이터센터의 착공 여부를 결정하는 핵심 임차인이 됐다.

그림 17. AI 회사가 클라우드 매출을 끌고 있다
푸른 막대로 표시된 초과 성장분 = 실제 클라우드 매출 – 챗GPT 이전의 성장 추세를 연장한 값
AI가 없었다면 벌지 못했을 클라우드 매출이 그만큼이라는 추정
두 선이 같이 포물선을 그리는 중. 랩의 매출이 클라우드의 초과 성장을 끌고 있다는 뜻으로 해석.



자료: 3Fourteen, 미래에셋증권 리서치센터

업계 추정에 따르면 주요 AI 사업자가 새로 인도받으려는 연산설비는 연간 8~10GW에 이른다. 여기서 8~10GW는 지금 가동 중인 설비가 아니라 앞으로 매년 새로 공급하려는 물량이다. 이 모든 설비가 한날한시에 켜지는 것도 아니다. 먼저 부지를 확보하고 전력계약과 인허가를 마친 뒤 변전설비와 냉각설비, 서버와 GPU를 차례로 설치해야 한다. 계약서에 적힌 8~10GW가 실제 가동전력으로 바뀌기까지는 여러 해가 걸린다.

그래도 각 단계의 공사는 앞 단계에서 들어온 돈을 확인한 뒤 진행된다. 모델 기업의 매출과 현금이 늘면 다음 데이터센터의 대출이 실행되고, 전력과 GPU 주문도 확정된다. 반대로 매출이 계획보다 늦게 늘면 후속 부지는 착공이 미뤄지거나 더 강한 보증과 더 높은 금리를 요구받는다. 결국 연산설비가 얼마나 빨리 늘어날지는 모델 기업이 그 설비의 임차료와 이자를 실제로 낼 수 있는가에 달렸다.

OpenAI와 Anthropic의 성장세는 AI 인프라에 최종수요가 있다는 사실을 이미 보여준다. 그렇다면 이제 확인할 것은 그 수요의 크기다. 두 회사가 현재 부담하는 컴퓨트 예산을 장기 임대료로 바꾸면 몇 GW에 해당하는지, 그리고 앞으로 들어올 8~10GW를 받치려면 매출이 얼마나 더 커져야 하는지를 계산해 볼 수 있다.

4. 2026년 컴퓨트 예산은 약 6GW의 장기 임대료에 해당한다

이를 계산하기 위해서는 우선 실제 OpenAI와 Anthropic의 설비에는 H100·H200·B200 같은 여러 세대의 NVIDIA GPU뿐 아니라 Google TPU와 Amazon Trainium도 섞여 있다는 것을 반영해야 한다. 연산 역시 유료 API에만 쓰이지 않는다. 차세대 모델 훈련과 연구, 무료 사용자 서비스, 안전성 평가에도 함께 투입된다.

지난 5월 5일, 그레그 브록먼 사장은 법정 증언에서 OpenAI가 2026년 한 해 동안 컴퓨팅 파워에 약 US$50bn을 지출할 것으로 예상한다고 말했다. 여기에는 고객의 질문에 답하는 추론뿐 아니라 새 모델을 개발하고 훈련하는 연산도 포함된다.

지난 5월 4일 Forbes의 기고문이 추정한 값에 따르면, Anthropic은 2026년에 모델 훈련에 약 US$12bn, 고객 서비스를 위한 추론에 약 US$7bn을 쓸 것으로 보고 있다. 둘을 합치면 약 US$19bn이다. 과거 Epoch AI가 Anthropic이 2025년에 모델 훈련에 약 US$4.1bn, 추론에 약 US$2.7bn을 쓴 것으로 추정했기 때문에 빠른 매출 증가와 신규 컴퓨트 계약을 반영하면 합리적인 전망으로 보인다.

두 값을 합치면 OpenAI US$50bn과 Anthropic US$19bn, 총 US$69bn이 2026년 컴퓨트 예산 금액으로 추정된다. 이제 이 금액을 1GW의 컴퓨트를 1년 동안 장기로 빌리는 데 드는 금액과 비교해보면 된다. 장기 GPU 임대사업의 1GW당 연매출 기준을 연간 US$12bn/GW로 잡는다면, "US$69bn ÷ US$12bn/GW ≈ 5.8GW"가 된다.

이 말은 OpenAI와 Anthropic의 훈련·추론 예산을 현재의 장기 임대가격으로 바꾸면 약 6GW의 컴퓨팅 규모가 된다는 뜻이다. 따라서 두 회사의 현재 합산 매출을 약 US$115bn로 책정할 때, 컴퓨트 예산 US$69bn은 합산 매출의 60%에 해당한다. US$69bn ÷ US$115bn = 60%. 즉 두 회사는 현재 매출의 60% 남짓을 당장의 서비스와 미래 모델을 위한 연산에 다시 넣고 있다.

이제 앞의 8~10GW와 연결해보면 된다. 9GW의 장기 컴퓨트를 빌리려면 연간 임대료가 약 US$108bn 필요하다. 9GW × US$12bn/GW = US$108. 현재처럼 매출의 60%를 컴퓨트에 쓴다고 가정하면, 9GW를 감당하는 데 필요한 합산 매출은 약 US$180bn으로 계산된다. 합산 매출 US$115bn과 비교하면 9GW는 약 1.6배의 매출을 요구하는 셈이다.

표 4. 현재 모델 업체들의 매출이 감당할 수 있는 컴퓨트 용량 역산

구분	값	산식·기준
OpenAI 컴퓨트 예산	US$50bn	2026년 훈련·추론 등 전체 컴퓨팅 지출 전망
Anthropic 컴퓨트 예산	US$19bn	US$12bn + US$7bn
두 회사 합산 컴퓨트 예산	US$69bn	US$50bn + US$19bn
두 회사 합산 연환산 매출	US$115bn	OpenAI와 Anthropic 합계
매출 대비 컴퓨트 비중	60%	US$69bn ÷ US$115bn
1GW당 연간 장기 임대료	US$12bn/GW	장기 GPU 임대사업의 연매출 기준
현재 예산의 임대료 등가용량	약 5.8GW	US$69bn ÷ US$12bn/GW
업계의 신규 인도계획 중간값	9GW	연간 8~10GW의 중간값
9GW의 연간 임대료	US$108bn	9GW × US$12bn/GW
9GW에 필요한 합산 매출	US$180bn	US$108bn ÷ 60%
현재 대비 필요한 매출 증가	약 1.6배	US$180bn ÷ US$115bn

자료: 각사, SemiAnalysis, 미래에셋증권 리서치센터

물론, 9GW는 업계 전체의 신규 인도계획인데, 여기서는 OpenAI와 Anthropic 두 회사가 그 물량을 전부 부담한다고 가정했다. 실제로는 Google과 Meta, SpaceXAI를 비롯한 다른 모델 기업과 대형 클라우드 사업자도 해당 설비를 사용한다. 따라서 두 회사가 실제로 감당해야 할 매출 문턱은 이 계산보다 낮을 것이다.

5. 수요는 증명됐다. 이제 총이익이 설비보다 빨리 늘어야 한다

해당 계산 결과 현재까지의 판정은 긍정적이다. OpenAI와 Anthropic은 아직 수익모델을 찾지 못한 연구소 수준이 아니다. 8월에는 OpenAI의 매출이 다시 가속하면서 합계가 약 US$110bn 이상으로 가뿐히 올라섰다. 위 계산에 따라 앞으로 9GW 정도를 감당하려면 매출이 지금보다 약 1.6배 정도 더 커지면 된다. 작은 숫자는 아니지만 설비가 여러 해에 걸쳐 순서대로 들어온다는 점을 고려하면 도달할 수 없는 거리는 아니라고 판단한다. 지금처럼 매출이 설비 가동보다 빠르게 늘면 먼저 켜진 데이터센터에서 발생한 매출과 현금이 다음 부지의 대출과 착공을 뒷받침한다.

두 회사 모두 유료 서비스를 팔 때 고객에게 받은 돈보다 답변 생성 비용도 더 큰 상태는 아니다. 다만 회사 전체에서 현금이 빠져나가는 이유는 추론만이 아니라 차세대 모델 훈련과 연구개발, 앞으로 사용할 컴퓨트 용량을 먼저 확보하는 비용까지 동시에 부담하고 있기 때문이다. OpenAI의 연환산 매출은 분기 들어 35% 늘었고 기업 매출은 50% 넘게 증가했다. API 토큰 처리량은 같은 기간 두 배가 됐다. 최종고객과 연산수요가 동시에 늘고 있다는 뜻이다.

종합하면, 현재 매출과 컴퓨트 예산은 약 6GW를 받치고 있으며, 계획된 다음 용량까지의 매출 격차도 1.6배로 줄었다. 현재로서는 데이터센터 증설을 멈출 이유보다 계속할 이유가 더 강하다. 토큰당 원가가 계속 내려가고 기업·에이전트 매출이 지금의 추세처럼 늘어난다면, 먼저 가동된 설비의 현금이 다음 GW의 착공을 계속 밀어 줄 것으로 사료된다.

V. 희소자원을 가진 자가 수익을 가져간다

이제 AI 인프라 수요가 실제로 존재하는지를 확인했다면 그 수요가 누구의 매출과 이익으로 귀속되는지가 중요해진다. 수요가 공급보다 빠르게 늘어날 때 초과수익은 가장 부족한 자원에서 발생한다. AI 인프라에서 부족한 것은 GPU만이 아니다. 전력을 공급받을 수 있는 부지, 변전설비, HBM, 광통신 부품, 냉각설비, 발전터빈과 숙련된 건설인력도 부족하다. 이 가운데 하나라도 없으면 데이터센터는 가동되지 않는다.

그래서 이번 섹션의 포인트를 세 가지로 꼽아봤다. 첫째, '희소자원'을 이미 가진 기업이다. 둘째, 남보다 빨리 확보하고 가동할 수 있는 기업이다. 셋째, 희소자원 자체를 만들어 파는 기업이다. 자본비용이 낮고 실행속도가 빠른 회사가 유리한 이유도 이 두 능력이 희소자원을 더 많이 확보하고 더 빨리 매출로 바꾸게 해주기 때문이다.

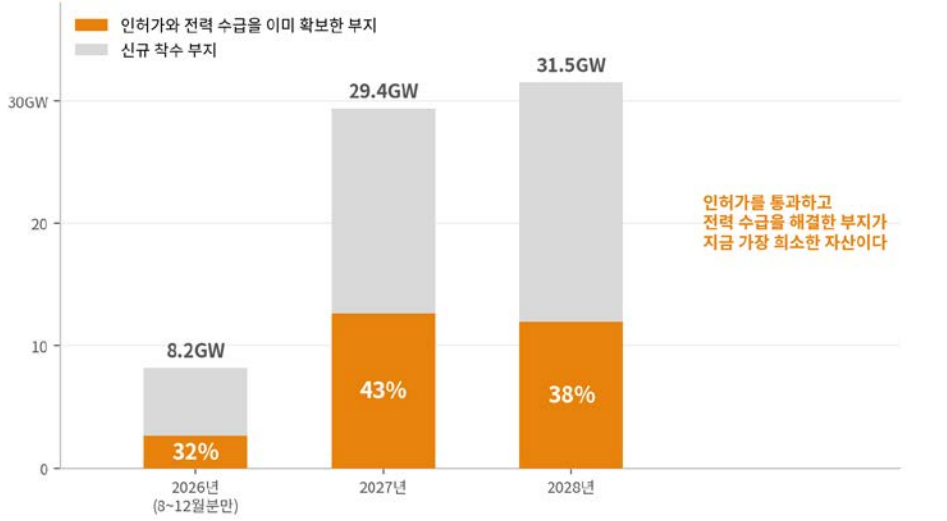
1. 빅테크는 경주의 절반을 끝낸 상태에서 출발한다

전력 인입
데이터센터 땅까지 전기를 실제로 끌어오는 작업. 한전과 계약을 맺었다고 바로 쓸 수 있는 게 아니라, 변전소·전선·변압기를 연결하고 시험까지 끝내야 하는 것. GPU를 사도 인입이 늦으면 매출이 뒤로 밀림.

새 데이터센터를 짓는 데 필요한 땅은 그저 빈 땅을 의미하지는 않는다. 대규모 전력 인입이 가능하고 인허가를 통과했으며 변전소와 광통신망을 연결할 수 있는 땅이다. 이런 부지는 돈만 있다고 바로 만들 수 없다. 전력회사와의 협상, 송전망 보강, 주민 동의와 건설 경험까지 함께 필요하다.

마이크로소프트·아마존·구글·메타 같은 기존 빅테크는 이런 자산을 이미 다량으로 보유하고 있다. 기존 캠퍼스를 확장하면 새로운 부지를 처음부터 개발하는 것보다 인허가와 전력 연결, 네트워크 구축에 걸리는 시간을 훨씬 줄일 수 있다. 2026년부터 2028년까지 미국에서 새로 가동될 예정인 용량 가운데 기존 부지의 확장분이 각각 32%, 43%, 38%로 집계된다.

그림 18. 새로 가동되는 용량의 3분의 1 이상이 기존 부지에서 나온다
기존 캠퍼스를 가진 사업자는 같은 설계와 공급망을 반복해서 쓸 수 있다.
신규 개발보다 전력과 인허가 위험이 낮다.
좋은 부지는 신용과 경험이 좋은 사업자가 먼저 가져가기 때문에 '실행격차'가 벌어지는 구조다.



자료: 전 사업자 미국 인도 예정 물량 집계, 미래에셋증권 AI리서치센터

기존 하이퍼스케일러는 고객도 이미 갖고 있다. 기업 고객의 데이터와 소프트웨어가 클라우드 안에 들어와 있기 때문에 새 AI 서비스를 기존 계약에 붙여 팔 수 있다. 자체 현금흐름으로 먼저 투자하고 나중에 고객을 찾는 것도 가능하다. 높은 신용등급 덕분에 같은 데이터센터를 더 낮은 금리로 지을 수도 있다.

따라서 빅테크의 경쟁력은 단순히 자본이 많다는 데 있지 않다. 부지·전력·고객·자본을 동시에 갖고 있기 때문에 신규 GPU를 실제 매출로 바꾸는 시간이 짧다. AI 인프라 투자가 커질수록 이 네 자산을 이미 가진 사업자의 우위도 커진다.

2. 속도는 새로운 희소자원을 만들어낸다

희소자원을 처음부터 많이 갖고 있지 않아도 승자가 될 수 있다. 남보다 빨리 인허가를 받고 전력을 연결하며 서버를 설치할 수 있다면, 없던 가동용량을 먼저 만들어낼 수 있기 때문이다.

SpaceXAI가 보여주는 강점이 여기에 있다. 부지 조성, 발전과 전력 인입, 서버 설치, 네트워크와 소프트웨어 운영을 여러 조직에 나누기보다 한 조직 안에서 밀어붙인다. 실제로 업계의 "상식"이 100MW 그린필드 기준 약 2년인데 SpaceXAI는 3~4달 수준으로 줄인 바 있다. 남들보다 1년 먼저 켠 1GW는 같은 1GW로 치부될 게 아니다. 공급이 부족한 시기의 높은 임대료와 고객을 먼저 가져갈 수 있기 때문이다.

표 5. 속도전을 펼치는 SpaceXAI

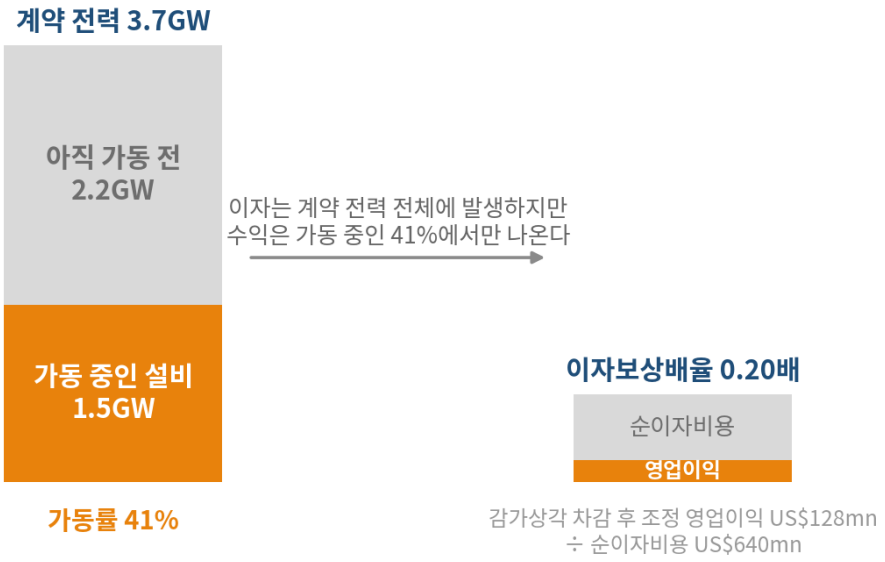
시설	공개된 기간	규모
Colossus 1 (멤피스, 옛 Electrolux 공장)	122일	약 10만 장 H100, 130~150MW(일부는 300MW로 표기)
같은 사이트 1차 확장	92일	약 20만 장까지 증설
Colossus 2 1차 클러스터	91일	약 11만 장 GB200, ~210MW
Colossus 2 2차(GB300)	64일 (추정)	약 11만 장 GB300, ~220MW

자료: SpaceX, 미래에셋증권 리서치센터

이 관점에서 데이터센터 사업자의 가장 중요한 숫자는 계약한 와트가 아니라 "가동 중"인 와트가 된다. CoreWeave는 2026년 2분기 말 기준 약 3.7GW의 전력을 계약했지만 실제 가동 중인 용량은 약 1.5GW였다. 가동률은 41%다. 이미 계약한 전력을 데이터센터와 GPU로 바꾸고 고객의 작업을 올리면 그대로 매출이 발생할 수 있다. 네오클라우드의 투자 thesis는 바로 여기에 있다. 계약전력이 충분하기 때문에 다음 단계의 성장은 수주 경쟁보다 실행속도에 달려 있다고 볼 수 있다.

따라서 투자자는 수주잔고의 크기보다 다음 숫자를 봐야 한다. "분기마다 가동용량이 얼마나 늘었는지, 전력을 연결한 뒤 첫 청구까지 며칠이 걸렸는지, 가동용량의 증가율이 이자비용의 증가율을 앞서는지"다. 계약전력은 가능성을 보여주고, 가동전력은 기업가치를 만든다.

그림 19. CoreWeave의 다음 성장동력은 새로운 계약보다 기존 계약의 가동
이자를 감당하지 못하는 것이 아니라 가동 전 설비의 이자를 먼저 부담하는 것이다



자료: CoreWeave, 미래에셋증권 리서치센터

3. 병목을 만드는 기업은 승자를 고를 필요가 없다

데이터센터 사업자 가운데 누가 이길지 판단하기 어렵다면, 부족한 자원을 직접 생산하는 기업을 사는 방법도 있다. AI 투자경쟁이 계속되는 동안 병목 공급자는 어느 클라우드와 모델 기업이 이기든 매출을 올릴 수 있다.

필자의 6월 보고서 "악마는 병목을 입는다"에서 주장했듯, 병목은 메모리와 광네트워크, 전력설비 등으로 이동하고 있다. 최근에는 이 방면에서 극적인 병목이 더욱 두드러진다. InP 기판 가격은 2인치 웨이퍼당 약 US$800에서 US$2,300~2,500까지 올라갔다. 발전터빈은 엔진보다 날개와 단조품의 공급이 부족해 주문부터 인도까지 몇 년이 걸릴 정도다.

표 6. AI 광네트워트 공급망 병목 점검

부품·공정	확인된 수치	현재 병목	의미
스위치용 메모리	메모리 부족으로 AAOI의 3Q26 100G 매출 US$20~25mn 이연 전망	스위치를 충분히 생산하지 못해 완성된 광모듈도 설치되지 못함	메모리 병목이 광네트워크 매출까지 막고 있다
DSP·TIA	AAOI는 4Q26 800G·1.6T 매출 약 US$330mn을 달성해야 함	DSP·TIA 조달량이 800G·1.6T 출하량을 제한	레이저가 있어도 신호처리 반도체가 없으면 광모듈을 완성할 수 없다
800G·1.6T 광트랜시버	AAOI 월 생산능력: 1Q26 말 10만 개 → 현재 20만 개 → 2026년 말 65만 개 이상 → 2027년 말 93만 개 이상	자동화라인 설치, 수율 안정화, 인증과 최종검사	수요보다 생산능력과 수율이 단기 매출의 상단을 정한다
200G/lane EML	Lumentum의 200G 제품이 EML 매출의 25% 이상. 2026년 말 생산능력을 전년 대비 50% 이상 확대	고속 EML 생산능력과 수율	1.6T 세대교체가 레이저 단계에서 이미 진행되고 있다
CW 레이저	Lumentum은 공급부족과 수율 우위를 바탕으로 가격 인상. FY4Q26 총마진 50.4%	좁은 출력편차와 높은 신뢰성을 가진 레이저 공급 부족	희소한 부품은 물량뿐 아니라 가격과 마진도 함께 올린다
InP 기판	Lumentum은 2031년까지 공급능력을 예약하기 위해 AXT에 총 US$87mn의 보증금 지급	InP 기판 → 에피택시 → 웨이퍼 팹이 모두 필요	병목이 광모듈 조립에서 레이저 칩을 거쳐 원재료 기판까지 거슬러 올라갔다
OCS	Lumentum의 다음 분기 OCS 매출이 처음으로 US$100mn을 넘어설 전망	대규모 AI 클러스터의 광경로 전환장비 공급	GPU 수가 늘수록 서버 연결뿐 아니라 클러스터 내부의 광스위칭 수요도 커진다
UHP 레이저·ELS	Lumentum은 2027년 하반기 공급 확대, 2028년 CPO 본격 배치 전망	고출력 레이저 팹, 외부광원 패키징과 수율	CPO는 광학을 없애지 않고 고출력 레이저와 외부광원이라는 새 병목을 만든다

자료: Lumentu, AAOI, 미래에셋증권 리서치센터

이 기업들의 투자논리는 단순하다. AI 사업자가 더 많은 컴퓨트를 만들려면 해당 부품을 반드시 사야 한다. 공급이 단기간에 늘지 못하면 판매량뿐 아니라 가격도 오른다. AI 사용량 증가를 가장 수익화하는 곳은 모델 기업이나 데이터센터가 아니라, 전체 증설을 막고 있는 부품 공급자라는 것도 유념하자.

4. 희소자원을 더 많이 확보하게 하기 위한 경쟁

희소자원을 발견했다고 해서 모두 자기자본으로 사야 하는 것은 아니다. 고객 선납금과 프로젝트 대출을 끌어올 수 있다면 같은 자기자본으로 더 많은 부지와 GPU를 확보할 수 있다.

신흥 클라우드 업체인 Nebius는 계약의 약 70%에서 고객 선납금을 받고 있다. 고객이 먼저 낸 돈으로 전체 설비투자의 약 50~60%를 충당하며, 이에 따라 "투자금 회수기간도 2~3년에서 약 1년 10개월"로 짧아졌다. 고객계약을 바탕으로 SOFR+2.50%의 썩 괜찮은 금리조건으로 US$775mn을 빌린 사례도 있다.

이에 따라 사업자의 자기자본 부담은 다음과 같이 줄어든다. "필요한 자기자본 = 총설비투자 – 고객 선납금 – 프로젝트 대출". 고객이 절반을 먼저 내고 나머지 일부를 장기대출로 조달하면 사업자는 전체 건설비의 일부만 부담한다. 설비가 가동되고 투자금이 회수되면 그 돈을 다음 캠퍼스에 다시 넣을 수 있다. 같은 자기자본으로 더 많은 설비를 반복해서 지을 수 있는 구조다.

빅테크는 높은 신용등급과 기존 현금흐름으로 이 구조를 만든다. Nebius 같은 사업자는 고객 선납금으로 만든다. CoreWeave는 장기 고객계약을 담보로 비소구 대출과 공개 신디케이트 대출시장을 열었다. 방법은 다르지만 결과는 같다. 자기자본을 적게 묶고 더 많은 희소자원을 경쟁적으로 선점하는 것이다.

하이퍼스케일러는 자본비용이 낮고 고객이 분산되어 있다. 반면 CoreWeave와 Nebius 같은 전문 사업자는 특정 고객과 프로젝트에 집중돼 있지만, 계약전력이 가동될 때 매출이 더 빠르게 증가할 수 있다. 이들은 빅테크보다 위험하지만 성공했을 때 자기자본 수익률도 높다. 반대로 가동이 늦어지면 아직 매출을 만들지 못한 시설의 이자를 먼저 부담해야 한다. 전문 임대사업자의 투자판단은 AI 수요 전망보다 실행지표에 달려 있는 이유다.

5. 병목을 가진 회사와 병목을 푸는 회사를 산다

AI 인프라 수요가 커진다는 이유만으로 모든 데이터센터와 GPU 임대사업자가 잘 될 것이라는 이야기를 하려는 게 아니다. 수익은 희소자원을 소유하거나, 남보다 빨리 확보하거나, 직접 만들어 파는 기업에 집중된다. 반대로 피해야 할 대상은 발표한 계약전력만 크고, 전력 인입과 가동이 늦으며, 비싼 자기자본으로 이자를 버텨야 하는 회사다. 암호화폐 채굴업체에서 AI DC로 전환한 대부분의 업체가 여기에 해당한다. 그래서 바라봐야 할 투자대상은 세 부류다.

첫째, 부지·전력·고객과 낮은 자본비용을 이미 갖춘 기존 빅테크 하이퍼스케일러다. 이들은 증설의 시행착오가 적고 기존 고객에게 AI 서비스를 바로 판매할 수 있다.

둘째, SpaceXAI처럼 의사결정과 건설을 빠르게 밀어붙여 새로운 가동용량을 먼저 만드는 사업자다. 이미 가진 자원보다 확보속도로 이기는 유형이다. 네오클라우드 업체들 중 일부는 계약전력을 빠르게 가동하고 고객자본을 활용한다는 사실을 증명할수록 이 범주에 가까워질 수 있다.

셋째, 메모리·광통신·변전설비·발전터빈과 같이 전체 증설을 가로막는 병목을 생산하는 기업이다. 이들은 최종 사업자를 정확히 고르지 않아도 AI 인프라 경쟁 전체에서 가격결정력을 얻을 수 있다.

AI 인프라 시대의 초과수익은 "희소성을 지배한 기업"이 가져간다. 그리고 그 희소성은 법정화폐가 아니다. 실물 자원이다.

그림 20. 10년 뒤에는 “돈이 중요하지 않을 것”이라고 인터뷰하는 일론 머스크…

자료: The Economist 유튜브, 미래에셋증권 리서치센터

Compliance Notice

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