오픈웨이트와 AI 수요의 저변 확장

글로벌 테크 박재환
6158 / jaehwan124@

Glossary

용어	정의
오픈웨이트(Open-Weight) 모델	학습된 가중치를 공개해 사용자가 자체 인프라에서 구동·변형할 수 있는 모델. 주로 중국 AI 모델이 오픈웨이트 방식으로 출시
클로즈드(Closed) 모델	가중치를 공개하지 않고 API를 통해서만 접근을 허용하는 모델
프론티어(Frontier) 모델	당대 최고 수준의 성능을 보유한 최선단 AI 모델. 주로 앤트로픽, 오픈AI 등을 프론티어 모델랩으로 통칭
API(Application Programming Interface)	외부 사용자가 모델랩의 인프라에 접속해 모델을 호출하는 인터페이스
가중치(Weight)	학습을 통해 최적화된 모델의 파라미터 값
토큰(Token)	AI 모델이 데이터를 처리하는 최소 단위
컨텍스트 윈도우(Context Window)	모델이 한 번에 참조할 수 있는 최대 문맥의 크기
KV 캐시(KV Cache)	추론 과정에서 과거 문맥의 정보를 저장한 데이터
CUDA	엔비디아 GPU에서 범용 연산을 수행하기 위한 컴퓨트 플랫폼
하이퍼스케일러(HyperScaler)	초대형 클라우드·데이터센터를 운영하는 빅테크 기업
네오클라우드(NeoCloud)	GPU 중심 인프라를 임대하는 AI 특화 클라우드 사업자
소버린 AI(Sovereign AI)	국가·관할권 단위로 데이터와 연산 주권을 확보하는 AI 인프라
온프레미스(On-Premise)	고객이 인프라를 직접 소유·운영하는 방식

Contents

Summary

오픈웨이트와 AI 수요의 저변 확장

- **오픈웨이트 모델의 확산은 프론티어 모델의 수요를 대체하기보다 AI 수요의 저변을 넓히는 방향으로 작용할 전망이다.** 최근 토큰 지출 인덱스 하락에도 오픈 AI·앤트로픽의 ARR 과 전체 토큰 사용량은 동반 증가하고 있으며, 이는 프론티어 모델 수요 하락이 아닌 저가형 모델로의 증분 수요 분산을 시사한다. 오픈웨이트는 낮은 토큰 가격을 통해 총 사용량을 확대하는 제본스의 역설을 촉발하는 동시에, 높은 커스터마이징 자유도와 데이터 주권을 기반으로 기존 클라우드를 경유하지 않던 엔터프라이즈·온프레미스 수요까지 신규로 유입시킬 것으로 판단한다.

- **낮은 토큰 가격이 낮은 인프라 수요를 의미하지는 않는다.** 오픈웨이트 모델 역시 추론 단계에서는 대규모 파라미터와 긴 컨텍스트를 처리하기 위한 가속기·메모리 인프라가 필요하며, 사용량 확대는 결국 추가적인 컴퓨팅 수요로 연결된다. 동시에 저가형 모델의 확산은 프론티어 모델랩의 가격 경쟁과 마진 압박을 심화시켜, 동일 비용 대비 더 높은 토큰 처리량을 제공하는 차세대 GPU 로의 전환을 가속할 전망이다. **결과적으로 오픈웨이트의 확산은 총 AI 사용량 확대와 신규 인프라 수요 창출, 프론티어 진영의 차세대 추론 인프라 투자 확대를 동시에 유도하는 요인으로 판단한다.**

- **이러한 변화의 수혜주로 엔비디아(Nvidia), 델 테크놀로지스(Dell Technologies), 네비우스(Nebius)를 선호주로 제시한다.** 엔비디아는 다양한 커스텀 모델을 수용할 수 있는 GPU 의 범용성과 CUDA 생태계의 Lock-in, 턴키 AI 인프라 역량을 기반으로 오픈웨이트 생태계 확장의 최대 수혜가 기대된다. 델은 AI Factory 와 Enterprise Hub 를 통해 오픈웨이트 모델의 온프레미스 배포를 상품화하고 있어 엔터프라이즈 AI 수요 확대에 가장 직접적으로 노출되어 있다. 네비우스는 모델 중립성·엔비디아 기반 클러스터·Token Factory 를 기반으로 오픈웨이트 서빙 수요를 흡수하는 동시에, 엄격한 선수금 구조와 중단기 계약 확대를 통해 성장성과 재무 안정성을 함께 확보하고 있다는 점에서 네오클라우드 업종 내 선호주로 제시한다.

업종	기업명	시가총액 (십억달러)	12M Fwd PER (배)	투자포인트
팹리스	엔비디아	5,552	18	■ 오픈웨이트 생태계 확장의 최대 수혜 ■ 범용성·CUDA 락인·통합 솔루션 3중 해자
서버 ODM	델 테크놀로지스	339	18.6	■ 오픈웨이트 온프레미스 배포 상품화 ■ 네오클라우드·엔터프라이즈·소버린 동시 노출 ■ 범용서버·스토리지 확대로 마진 방어
네오클라우드	네비우스	65	-	■ 모델 중립성 기반 오픈웨이트 서빙 수요 확보 ■ 엄격한 선수금 구조로 재무 안정성 확보 ■ Token Factory로 오픈웨이트 토큰 서빙 확장

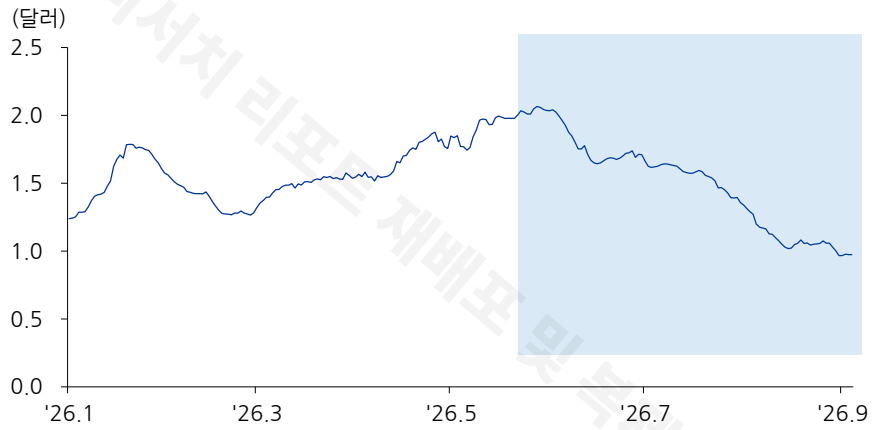
I. 오픈웨이트 모델 확산의 시사점

토큰 지출 인덱스와 모델랩 ARR 의 디커플링

토큰 지출 인덱스는 5 월 말 고점을 기록한 이후 하락세를 이어가고 있다. 이는 기업들의 토큰 지출 부담이 누적되며 오픈웨이트를 비롯한 저가형 모델로 수요가 이동한 결과로 판단한다. 실제로 모델 게이트웨이 서비스를 운영하는 Vercel AI 에 따르면 오픈웨이트 모델의 토큰 점유율은 6 월 말 30%에서 8 월 말 62.5%로 두 달 만에 두 배 이상 확대됐다.

토큰 지출 인덱스는 5 월 고점 이후 지속 하락세

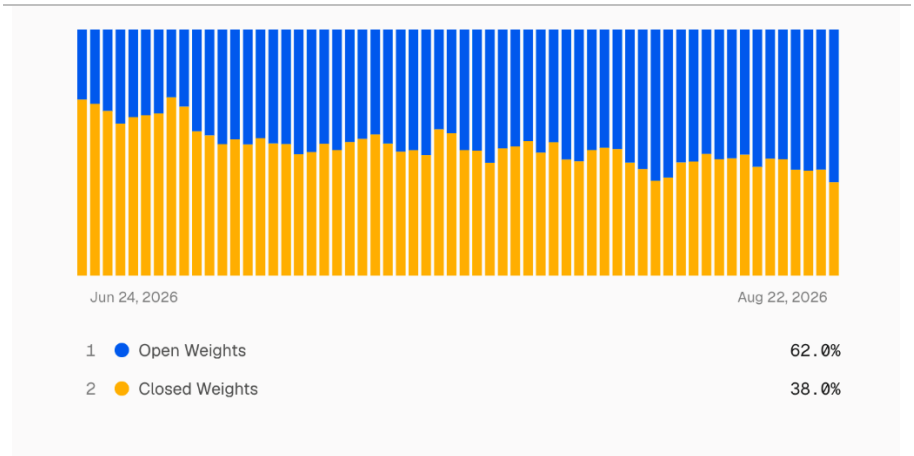
토큰 지출 인덱스 추이



자료: Bloomberg, 유진투자증권

Vercel AI 상 오픈웨이트 토큰 점유율은 62%까지 증가

모델 생태계별 토큰 점유율 추이



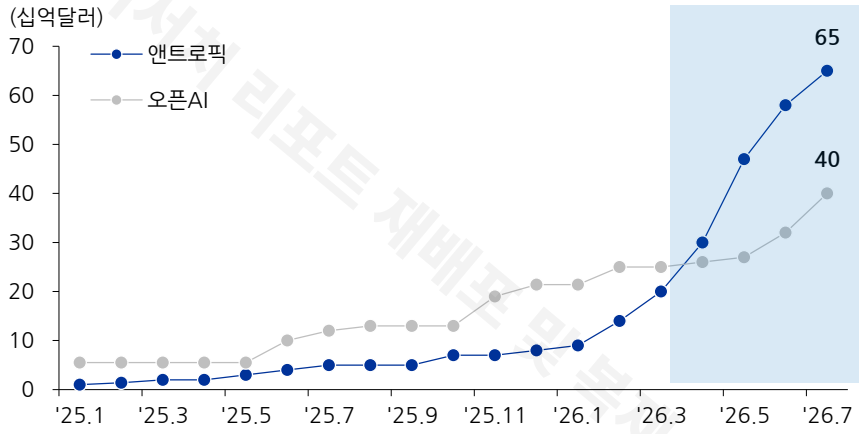
자료: Vercel AI, 유진투자증권

같은 기간 앤트로픽의 7 월 ARR(연환산매출)은 약 650 억달러로 시장 기대치를 소폭 하회하며, 토큰 지출 인덱스의 하락과 맞물려 프론티어 모델랩에 대한 우려를 키웠다. 다만 앤트로픽은 여전히 전년 동월 대비 10 배 이상의 ARR 성장을 유지하고 있다. 기대치 하회의 주된 원인은 7 월 출시된 오픈 AI GPT-5.6 의 호조로 판단하며, 오픈 AI ARR 의 재가속이 이를 방증한다. 양사 합산 ARR 은 7 월 기준 1,000 억달러를 상회하며 전년비 5 배 이상 성장했으며, 주요 프론티어 모델 합산 토큰 사용량 또한 급격하게 증가하고 있는 양상이다.

따라서 토큰 지출 인덱스의 하락은 프론티어 진영의 절대 수요 감소로 해석하기 어렵다. **① 앤트로픽과 오픈 AI 의 ARR 이 동반 성장하고 있고, ② OpenRouter 기준 토큰 사용량 역시 증가세를 유지하고 있는 점을 감안하면, 저가형 모델로의 '증분 수요 분산'으로 보는 것이 적합하다.**

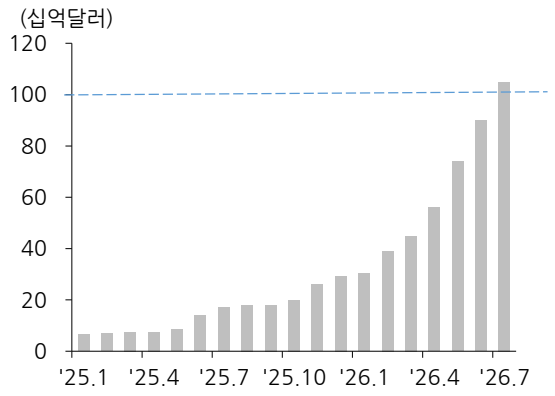
앤트로픽과 오픈 AI ARR 은 동반 상승세를 유지

앤트로픽, 오픈 AI ARR 추이



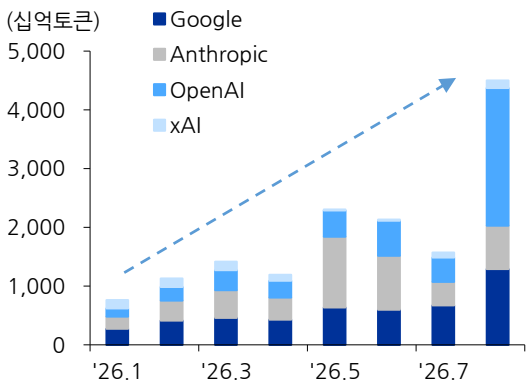
자료: 언론 종합, 각 사, 유진투자증권

앤트로픽, 오픈 AI 합산 ARR 추이



자료: 언론 종합, 각 사, 유진투자증권

주요 프론티어 모델 토큰 사용량 추이



주 1: OpenRouter 플랫폼을 거치는 토큰 사용량만을 산정
주 2: 매월 말 기준 토큰 사용량을 산정
자료: OpenRouter, 유진투자증권

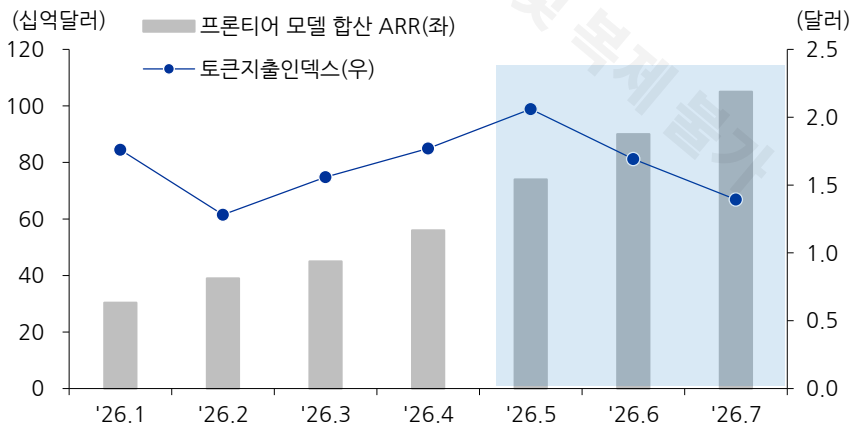
그간 AI 밸류체인 투자자들은 전방 수요를 가늠하는 지표로 앤트로픽의 ARR 성장률에 크게 의존해 왔다. 앤트로픽은 API 와 엔터프라이즈 매출 비중이 높아 기업의 AI 지출 강도를 가장 직접적으로 반영하는 지표로 인식되어 왔기 때문이다.

그러나 토큰 지출 인덱스와 프론티어 모델랩 ARR 이 디커플링되는 현 국면은 증분 수요가 오픈웨이트 생태계로 분산되고 있음을 시사한다. 오픈웨이트 모델은 외부 인프라에서 구동될 수 있는 만큼, 토큰 사용량이 증가하더라도 모델랩의 매출로 인식되는 부분은 제한적이다. **즉, 프론티어 ARR은 여전히 유효한 지표이나, AI 전방 수요의 일부만을 포착하는 지표로 그 위상이 변화하고 있다고 판단한다.**

따라서 향후 AI 전방 수요는 모델랩의 매출보다 인프라 단의 지표를 통해 측정하는 것이 적합하다고 판단한다. 구체적으로는 ① 플랫폼 전반의 토큰 사용량, ② 클라우드·서버 사업자의 백로그와 AI 수주, ③ GPU 렌탈가 추이, ④ 데이터센터 계약 용량 등이 대안 지표로 기능할 수 있다. 실제로 이들 지표는 토큰 지출 인덱스가 하락한 기간에도 동반 상승하고 있다. 이러한 관점에서 향후 앤트로픽의 ARR 성장세가 둔화되더라도 이는 프론티어 모델 간 경쟁 심화, 오픈웨이트 모델로의 수요 확산 등으로 인한 결과일 가능성이 높으며, 전체 컴퓨팅 수요 감소의 신호로 해석하기는 어려울 것으로 판단한다.

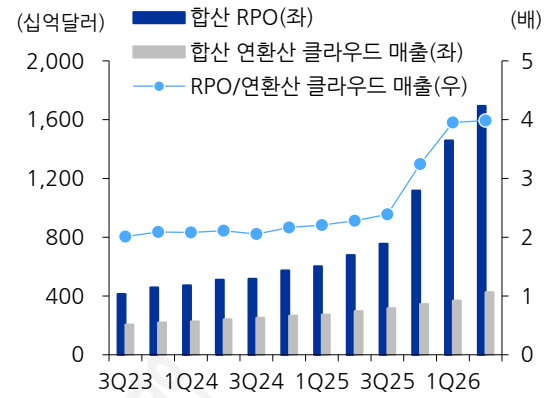
토큰 지출 인덱스와 프론티어 모델랩 합산 ARR 은 디커플링

토큰지출인덱스와 프론티어 모델랩 합산 ARR 비교



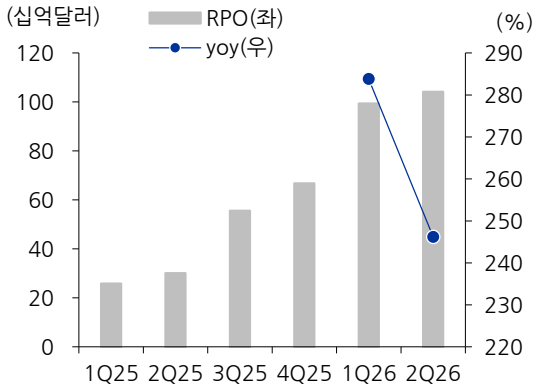
자료: Bloomberg, 유진투자증권

하이퍼스케일러 3 사 RPO 추이



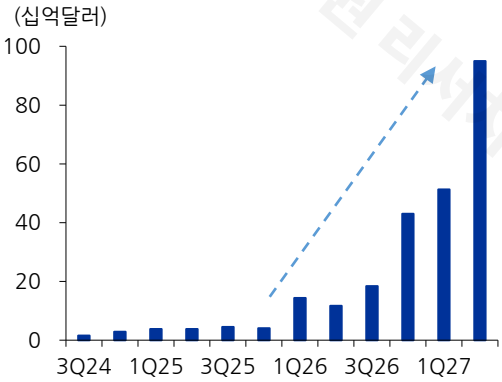
자료: Bloomberg, 유진투자증권

코어위브 RPO 추이



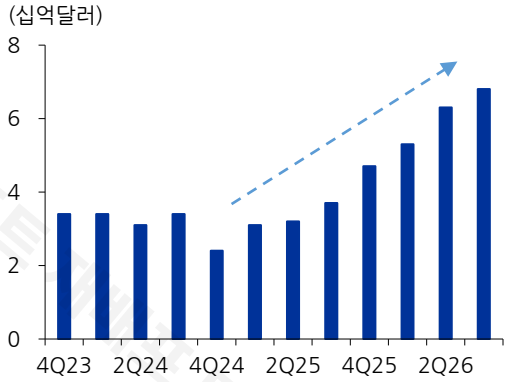
자료: Bloomberg, 유진투자증권

델 AI 서버 백로그 추이



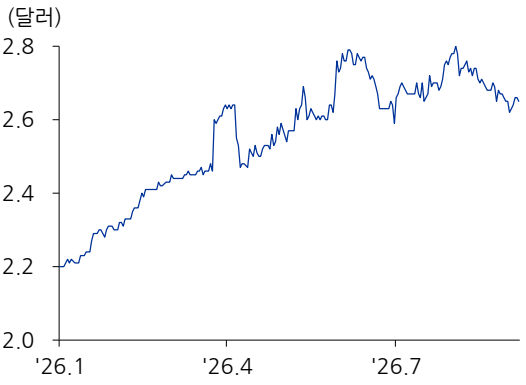
자료: Bloomberg, 유진투자증권

HPE AI 시스템 백로그 추이



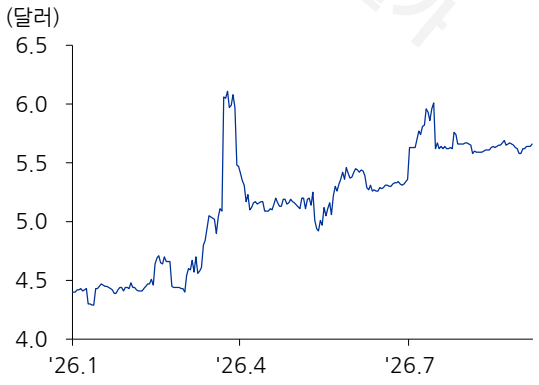
자료: Bloomberg, 유진투자증권

H100 렌탈가 추이



자료: Bloomberg, 유진투자증권

B200 렌탈가 추이



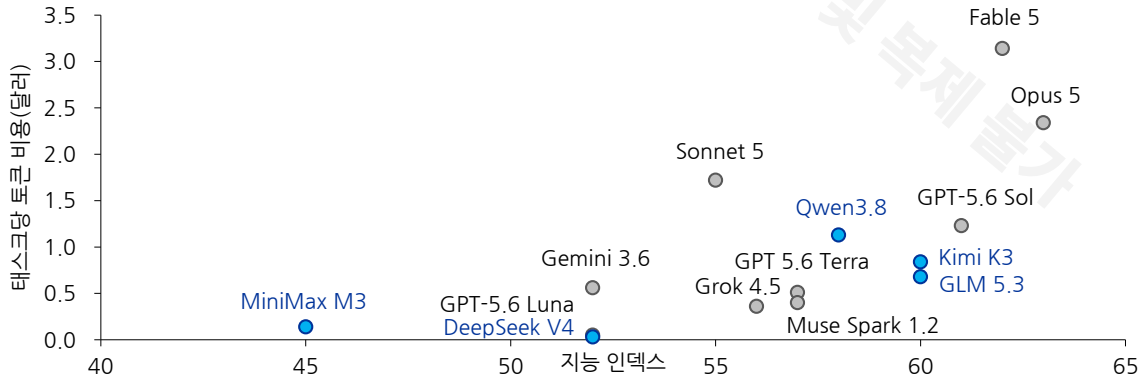
자료: Bloomberg, 유진투자증권

추론 앞에서는 모두가 공평하다

오픈웨이트 모델은 프론티어 모델에 근접한 지능 인덱스를 확보한 반면, 태스크당 비용은 최대 80% 저렴하다. 이에 2025 년 초 딥시크 모먼트와 유사하게 AI 인프라 과잉 투자에 대한 우려가 재차 부각됐다.

그러나 증류를 통해 학습 비용을 절감했다 하더라도, 추론 단계에서는 사정이 다르다. 추론은 이미 고정된 파라미터를 기반으로 새로운 토큰을 생성하는 과정이기에, 요구되는 인프라의 규모는 파라미터 수에 직접적으로 좌우된다. MoonShot 의 Kimi K3 는 2.8 조개의 파라미터를 보유해 MXFP4 기준 파라미터 저장에만 약 1.5TB 의 가속기 메모리가 필요하다. 여기에 주요 프론티어 모델과 동등한 100 만 토큰의 컨텍스트 윈도우는 HBM 을 넘어 DRAM, NAND 등 추가적인 메모리 계층까지 요구한다. MoonShot 이 Kimi K3 구동을 위해 64 개 이상의 가속기로 구성된 SuperNode 사용을 권고한 점 역시 이를 뒷받침한다. 조 단위 파라미터를 보유한 Qwen3.8 과 DeepSeek V4 에도 유사한 강도의 추론 인프라 부담이 작용할 것으로 판단한다.

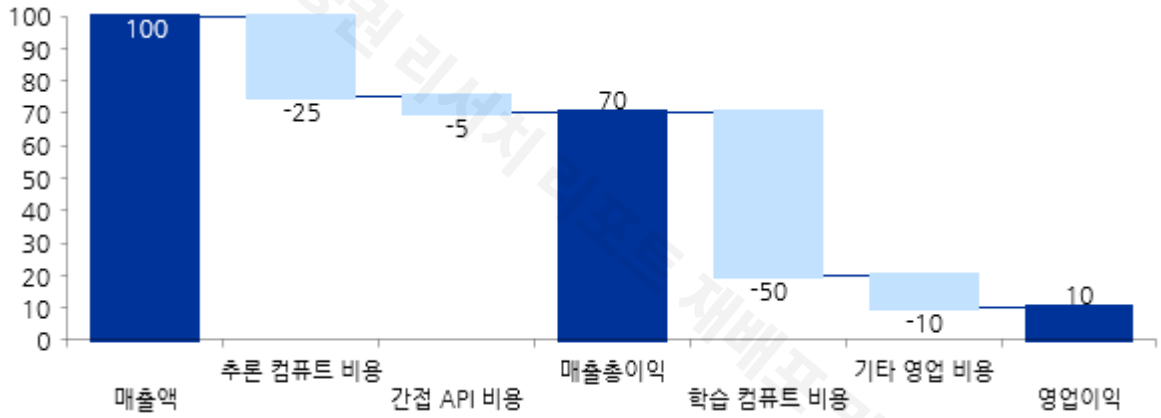
주요 AI 모델 지능인덱스, 태스크당 토큰 비용 추이



자료: Artificial Analysis, 유진투자증권

AI 모델랩의 매출총이익은 '토큰 가격(매출) - 토큰당 추론 인프라 비용(매출원가)'으로 단순화된다. 따라서 토큰당 추론 비용이 유사하다는 전제 하에, 기존 토큰 가격 수준에서는 서빙을 확대할수록 수익성이 악화되는 구조이다. 여기에 차세대 모델 개발을 위한 학습 컴퓨트 비용은 비용 부담을 추가로 가중시킨다. 이는 ① MoonShot 이 Kimi K3 공개 직후 신규 구독을 일시 중단한 점, ② DeepSeek 가 V4 의 피크 타임 토큰 가격을 최대 4 배 이상 인상한 점에서 확인된다. **결국 오픈웨이트 모델의 낮은 토큰 가격이 곧 낮은 추론 인프라 부담을 의미하지는 않으며, 오픈웨이트 모델랩이 자체 API 로 점유율을 확대하기 위해서는 추가적인 대규모 AI 인프라 확보가 불가피하다고 판단한다.**

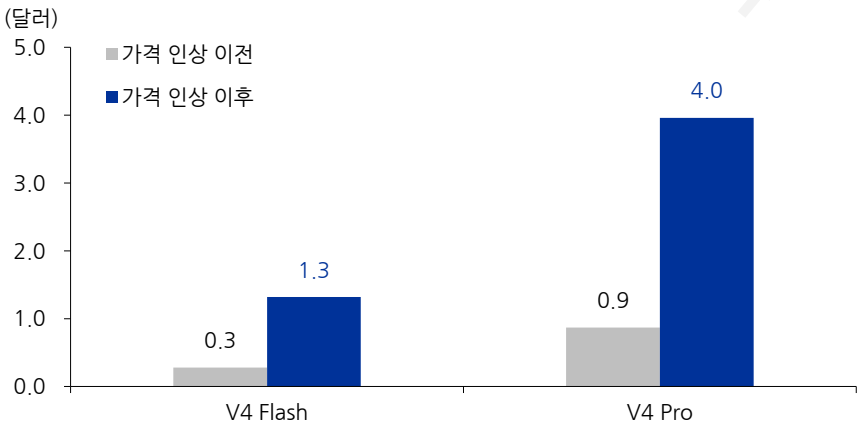
가상의 AI 모델랩 손익계산서



자료: 유진투자증권

DeepSeek 는 V4 모델의 피크타임 가격을 최대 4 배 이상 인상

DeepSeek 피크타임 토큰 가격 인상



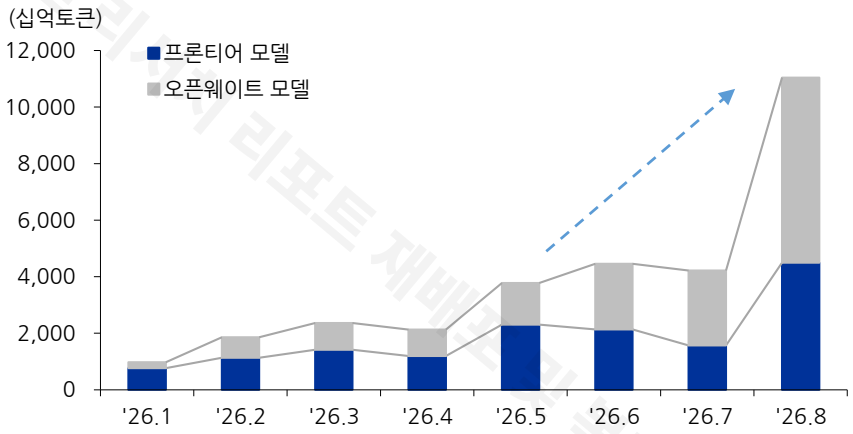
자료: DeepSeek, 유진투자증권

제본스의 역설: 저변의 확장

오픈웨이트 모델의 확산은 토큰 가격 하락을 통해 전체 사용량을 끌어올리는 제본스의 역설을 촉발할 것으로 판단한다. 실제로 오픈웨이트 모델의 확산과, 오픈AI 등 주요 프론티어 모델의 토큰 가격 인하가 맞물리며 OpenRouter 기준 토큰 사용량은 다시 가파른 증가세로 전환했다. 이는 AI 의 수요가 토큰 가격에 민감하게 반응한다는 점을 시사한다. 즉 그간의 토큰 가격 수준이 잠재 수요를 온전히 발현시키지 못하는 제약으로 작용해 왔으며, **가격 하락은 그 제약을 해소하며 AI 산업의 저변을 넓히는 요인으로 작용할 전망이다.**

프론티어와 오픈웨이트 모델 토큰 사용량은 동반 상승세

OpenRouter 프론티어 모델, 오픈웨이트 모델 토큰 사용량 추이



주 1: OpenRouter 플랫폼을 거치는 토큰 사용량만을 산정
주 2: 매월 말 기준 토큰 사용량을 산정
자료: OpenRouter, 유진투자증권

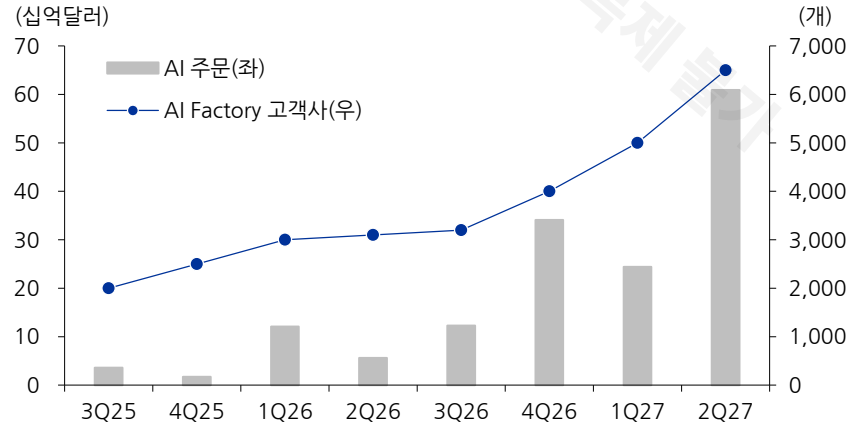
가격 하락만이 저변 확장의 경로는 아니다. 오픈웨이트 모델은 가중치가 공개되어 있어 사용자가 자체 데이터로 파인튜닝이나 강화학습을 적용해 모델을 커스터마이징하기에 용이하다. **이는 사내 워크로드에 최적화된 모델의 부재로 AI 의 효용을 체감하지 못했던 신규 수요층을 유입시키는 유인으로 작용한다.**

이러한 커스터마이징은 필연적으로 자사 데이터를 모델에 투입하는 과정을 수반한다. 클로즈드 모델은 API 호출 과정에서 데이터가 외부 클라우드로 유출될 수 있는 구조인 반면, 오픈웨이트 모델은 가중치를 확보해 자체 인프라에서 구동할 수 있어 이 제약에서 자유롭다. 관련해서 델은 전 세계 데이터의 83%가 온프레미스에 소재하고 있어, 데이터를 클라우드로 옮기기보다 데이터가 있는 곳으로 연산을 가져가는 편이 경제성 측면에서도 유리하다고 언급했다. **즉 오픈웨이트의 확산은 기존 수요를 키우는 데 그치지 않고, 하이퍼스케일러를 경유하지 않는 별도의 수요 채널을 형성한다는 점에서 의미가 있다.**

온프레미스 수요는 이미 주요 기업들의 실적 발표에서도 확인된다. 델의 AI Factory 도입 기업 수는 2025년 3월 2,000개에서 2026년 9월 6,500개로 18개월 만에 3 배 이상 확대되었으며, 특히 첫 3,200 개 고객을 확보하는 데 8 개 분기가 소요된 반면, 이후 3,300 개는 3 개 분기 만에 추가됐다. 델 경영진은 이 가속의 주체가 엔터프라이즈임을 명시했다.

엔터프라이즈 수요 강세로 AI Factory 고객사는 지속 증가하는 양상

델 AI 주문, AI 고객사 추이

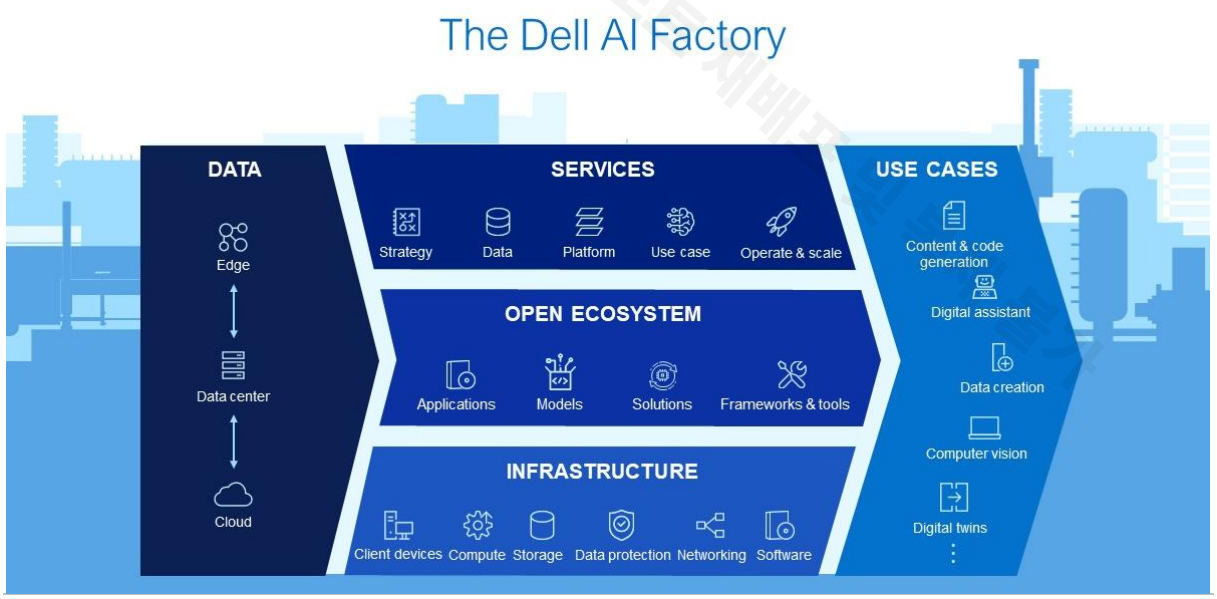


주: AI Factory 고객사는 델 어닝콜을 기반으로 산정했으며, 일부 추정이 포함됨
자료: Dell Technologies, 유진투자증권

공급 측면에서도 진입 장벽이 낮아지고 있다. 오픈웨이트 모델이 가중치가 공개되어 있다고 하더라도, 실제 구동에는 컨테이너 구성, 병렬화, 양자화 설정 등 상당한 간극이 존재해 왔다. 델은 자사 온프레미스 AI 배포 턴키 솔루션인 AI Factory 내에 허깅페이스와 공동 운영하는 Dell Enterprise Hub 플랫폼을 운영 중이다. 이는 DeepSeek, Kimi, GLM 등 최신 오픈웨이트 모델을 플랫폼별 최적화가 완료된 배포 패키지 형태로 제공하며, 공개 릴리스 수시간 내에 검증된 컨테이너가 준비된다. 자체 역량이 부족한 엔터프라이즈에게도 오픈웨이트 도입 경로가 열리고 있다는 의미이다.

종합하면 오픈웨이트 모델의 확산은 ① 토큰 가격 하락을 통해 총 사용량을 확대하고, ② 커스터마이징 용이성을 통해 신규 수요층을 유입시키며, ③ 그 수요가 온프레미스라는 별도 채널로 발현되면서 하이퍼스케일러 캐팩스로는 포착되지 않던 신규 인프라 수요를 창출한다는 점에서, AI 인프라 수요에 긍정적인 측면이 크다고 판단한다.

델 AI Factory



자료: Dell Technologies, 유진투자증권

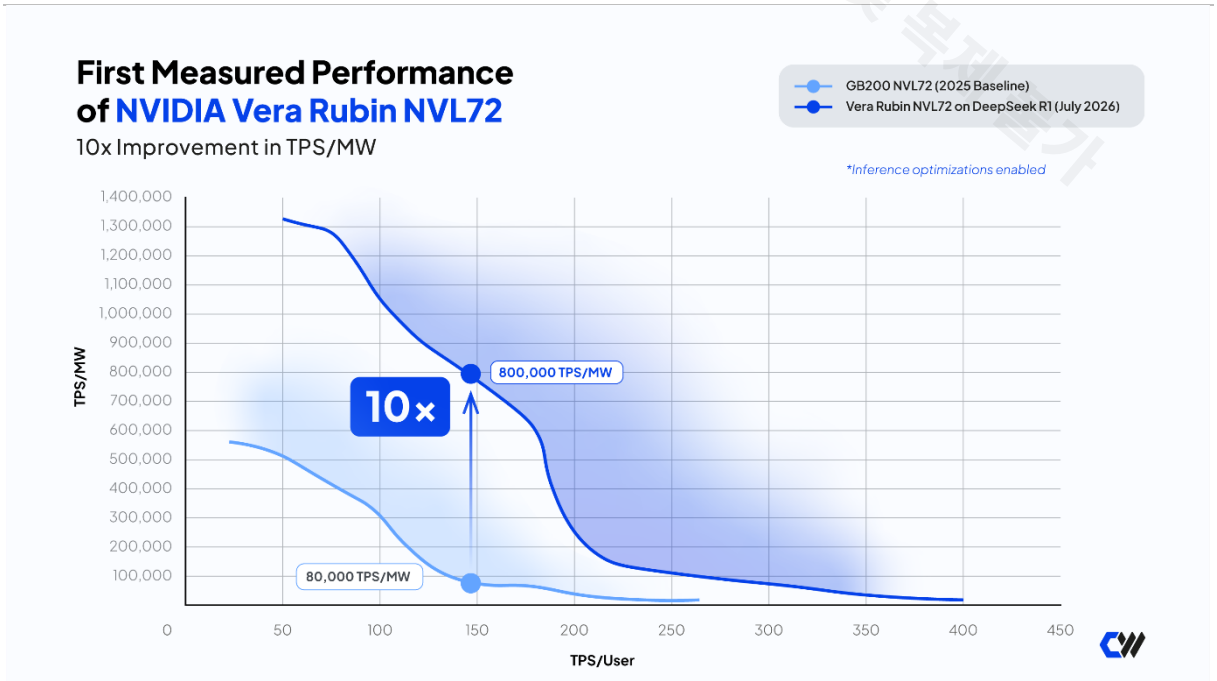
싸지는 토큰이 비싼 하드웨어를 부른다

다만 오픈웨이트 모델의 확산이 프론티어 모델랩에 반가운 소식만은 아니다. 저가형 오픈웨이트 모델의 점유율 확대는 프론티어 모델랩의 차세대 모델 가격 책정에 직접적인 압박으로 작용할 가능성이 높다. 오픈AI가 최근 GPT-5.6 Luna의 토큰 가격을 최대 80% 인하한 것 역시 이러한 경쟁 환경을 선제적으로 반영한 결정으로 판단한다.

가격 하락 압력에 노출된 프론티어 모델랩이 마진을 방어하기 위해서는 결국 토큰당 추론 인프라 비용, 즉 추론 원가를 낮춰야 한다. 기존 추론은 학습 대비 연산 요구 강도가 낮았기에 H100, H200 등 레거시 GPU 중심으로 인프라가 구성되는 경향이 있었다. 그러나 추론 원가 절감이 수익성의 핵심 변수로 부상하면서, 레거시 GPU 중심의 인프라로는 대응이 어려워질 것으로 판단한다. **이는 동일 ASP(혹은 W) 대비 더 높은 토큰 처리량을 확보한 블랙웰·루빈 등 차세대 칩 수요를 자극하는 결과로 이어질 것으로 판단한다.**

결과적으로 오픈웨이트 모델의 부상은 ① 제본스의 역설을 통해 AI의 저변을 확장하고, ② 프론티어 모델랩의 차세대 추론 인프라 수요를 자극한다는 점에서, AI 인프라 사이클에 긍정적인 요인으로 작용할 것으로 판단한다.

GB200 NVL72, Vera Rubin NVL72 토큰 처리량 비교



자료: Coreweave, 유진투자증권

II. 오픈웨이트 생태계 확장의 수혜자들

모델 진영의 분화: 폐쇄형과 개방형

오픈웨이트 생태계에서 인프라 사업자의 모델 중립성은 핵심 경쟁 요소로 부상하고 있다. 우선 오픈웨이트 생태계의 수요는 본질적으로 파편화되어 있다. 사용자는 공개된 모델 가중치를 자사 데이터로 파인튜닝하거나 강화학습을 적용해 워크로드에 맞게 변형한다. 이는 소수의 표준화된 모델을 대규모로 서빙하는 프론티어 모델 구조와 근본적으로 다르며, 인프라 사업자에게는 다수의 커스텀 모델을 동시에 수용할 수 있는 유연성이 요구된다.

또한 자체 모델을 보유했거나 특정 프론티어 모델랩과 긴밀히 결합된 사업자는 오픈웨이트 서빙에 이해 상충이 발생한다. 오픈웨이트 모델은 프론티어 모델과 직접적인 경쟁 관계에 있기 때문이다. 메타가 2 분기 실적 발표에서 외부 컴퓨팅 임대 요청이 다수 있었음에도 "단기 수익을 위해 컴퓨팅을 수익화하는 것은 어리석은 판단"이라며, 컴퓨팅에 자사 모델을 결합해야 최대 부가가치가 창출된다고 언급한 것은 이러한 구조를 명확히 보여준다. 하이퍼스케일러에게 컴퓨팅은 완전히 독립적인 상품이 아니라 자사 모델 경쟁력을 뒷받침하는 수단에 가깝다.

이에 따라 AI 모델 진영은 크게 프론티어 모델·하이퍼스케일러가 결합된 폐쇄형 축, 그리고 오픈웨이트 생태계와 엔비디아·네오클라우드가 결합된 개방형 축으로 분화되고 있다고 판단한다. **오픈웨이트 생태계의 확장 국면에서는 개방형 밸류체인이 상대적으로 유리한 위치를 점할 것으로 본다.**

하이퍼스케일러는 자체 모델을 서빙 혹은 개발중인 상황

하이퍼스케일러의 자체 모델

하이퍼스케일러	클라우드 플랫폼	모델명
구글	Google Cloud	Gemini
마이크로소프트	Azure	MAI
메타	-	Muse Spark
아마존	AWS (Bedrock)	미정(프론티어 모델 개발 중)

자료: 각 사, 유진투자증권

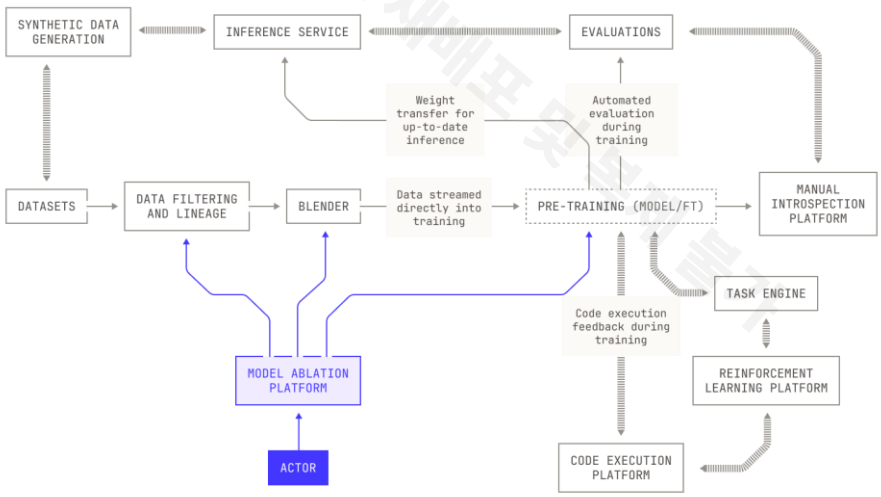
엔비디아: 개방형 진영을 설계하다

엔비디아는 오픈웨이트 생태계 확장에 능동적으로 개입하고 있다. Nemotron 으로 대표되는 자체 오픈웨이트 모델을 출시한 데 이어, 미국 오픈웨이트 스타트업 Poolside AI 에 소프트웨어 라이선스 대가로 60 억달러를 지불하고 10 억달러를 추가 지분 투자했다. 주목할 점은 라이선스 대상이 완성된 모델이나 가중치가 아니라 Poolside 의 AI 모델 개발 플랫폼 Model Factory 라는 점이다. 이는 엔비디아가 모델 개발 단계에서부터 생태계에 관여하며, 규제와 보안 측면에서 상대적으로 자유로운 미국 오픈웨이트 진영을 확립하려는 의지로 해석된다.

이에 더해 엔비디아는 AI 모델 플랫폼인 허깅페이스(Hugging Face)를 129 억달러에 인수하기로 합의했다. 허깅페이스는 100 만개 이상의 커스텀 오픈웨이트 모델이 업로드되어 있으며, AI 워크로드를 위한 데이터셋을 제공하는 플랫폼이다. 엔비디아의 허깅페이스 인수는, 오픈웨이트 생태계의 중심 좌표 역할을 하는 플랫폼의 인수를 통해 생태계를 확장하려는 의도로 해석된다.

Model Factory 는 AI 모델 개발 프로세스 전반을 지원하는 플랫폼

Poolside AI Model Factory



자료: Poolside AI, 유진투자증권

엔비디아가 오픈웨이트 생태계를 적극적으로 확장하는 이유는, 당연하게도 엔비디아가 곧 오픈웨이트 생태계 확장의 최대 수혜자이기 때문이다.

오픈웨이트 생태계는 커스텀 모델 수요가 지배적이다. 특정 모델 아키텍처에 최적화된 ASIC 과 달리, 엔비디아 GPU 는 다양한 모델 구조와 워크로드에 유연하게 대응할 수 있다. **파편화된 수요에 대응해야 하는 환경에서 범용성은 그 자체로 경쟁 우위로 작용한다.**

또한 현존하는 오픈웨이트 모델의 절대 다수는 CUDA 기반으로 학습되었으며, 학습·추론 전 과정에서 CUDA 라이브러리에 의존하고 있다. 최근 GLM 과 Qwen 은 중국계 가속기에 추론 워크로드를 공격적으로 포팅하고 있는 것으로 확인되나, 파인튜닝과 강화학습이 적용된 커스텀 모델까지 모두 대응할 수 있는 여력은 제한적일 것으로 판단한다. **자체 생태계 구축 여력이 제한적이고 모델 커스텀 강도가 높은 오픈웨이트 진영은 엔비디아 생태계를 중심으로 구동될 가능성이 높으며, 오픈웨이트 모델이 확산될수록 AI 산업 전체의 엔비디아 아키텍처 Lock-in 은 구조적으로 강화될 것으로 판단한다.**

주요 오픈웨이트 모델의 절대 다수는 CUDA 기반으로 학습되었으며 CUDA Library 에 의존

주요 오픈웨이트 모델의 CUDA 의존 구조

모델명	CUDA 사용 여부	CUDA 의존도	CUDA 기반 기능
Kimi K3	O	높음	MoonEP, FlashKDA, CUTLASS 등
Qwen3.8	O	높음	Pytorch CUDA, FlashQDA, TileLang 등
DeepSeek V4	O	높음	DeepGEMM, Mega MoE CUDA 등
GLM-5.3	△	중간	cuDNN, Megatron 등

자료: 각 사, Github, 유진투자증권

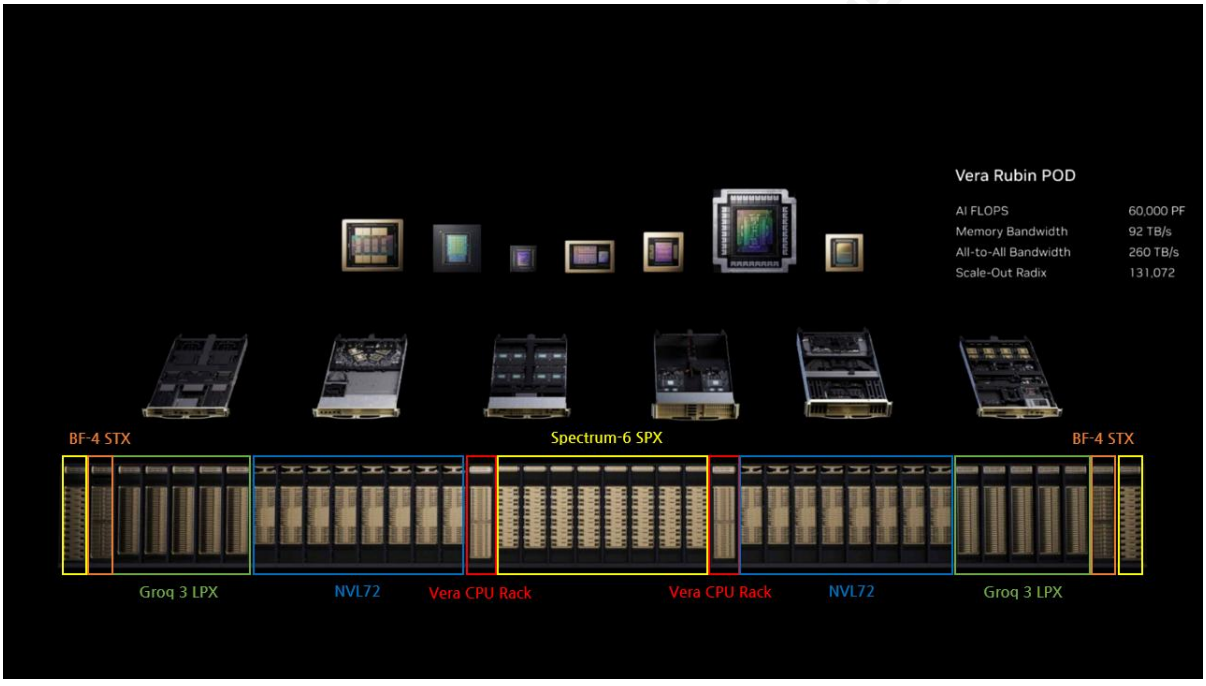
아울러 데이터센터 운영 경험이 풍부하고 자체 ASIC으로 추론 수요에 대응 가능한 하이퍼스케일러와 달리, 소버린 AI 프로젝트와 엔터프라이즈 온프레미스 수요처는 자체 생태계가 빈약해 즉시 가동 가능한 통합 솔루션에 대한 니즈가 지속적으로 상승할 전망이다. **엔비디아는 GPU 를 넘어 NVLink 스케일업 시스템, Grace·Vera CPU, Spectrum-X, Groq LPU, BlueField-4 DPU 까지 학습부터 에이전트 추론에 이르는 전 워크로드 사이클에 대응 가능한 유일한 사업자이기에, 통합 솔루션 수요에 가장 효과적으로 대응할 수 있을 것으로 판단한다.**

엔비디아 베라 루빈 플랫폼구성

구분	VR NVL72	Vera CPU 랙	Groq 3 LPX	BlueField-4 STX	Spectrum-6 SPX
주요 프로세서	Rubin GPU	Vera CPU	Groq 3 LPU	BlueField-4	Spectrum-6 CPO
주요 프로세서 탑재량	72	256	256	32	변동
주요 메모리	HBM	LPDDR	SRAM	SSD	-
역할	AI 연산	시스템 오케스트레이션	추론 Decode 연산 분담	KV 캐시 저장 및 전송	스케일아웃 네트워크

자료: Nvidia, 유진투자증권

Vera Rubin Pod 사진

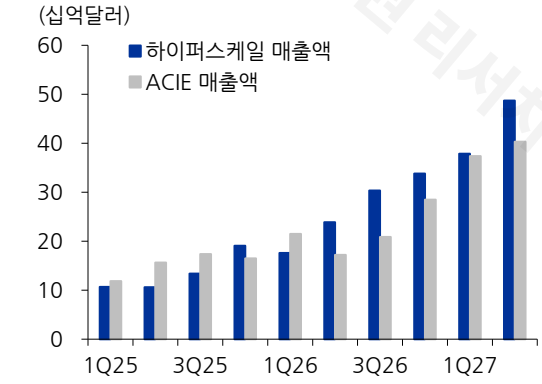


자료: Nvidia, 유진투자증권

이러한 수혜는 실적으로도 확인된다. 2 분기 ACIE(네오클라우드·소버린 AI·엔터프라이즈 등) 매출은 403 억달러로 전년비 134% 성장하며 하이퍼스케일 매출 성장률(+104%yoy)을 상회했다. AI 수요가 프론티어 모델 중심에서 산업 전반으로 저변을 넓히면서, 엔비디아의 수요처 역시 ACIE 를 축으로 빠르게 확장되고 있다.

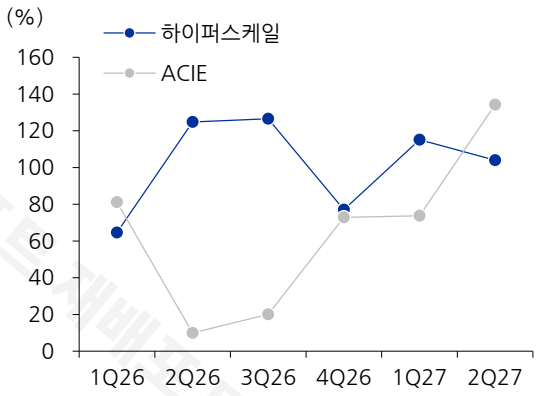
정리하면 오픈웨이트 생태계의 확장은 엔비디아에게 ① 하이퍼스케일러 ASIC 확대로 인한 수요 잠식 리스크를 분산시키고, ② ACIE 중심의 신규 수요처를 창출하며, ③ 프론티어 모델랩의 차세대 추론 인프라 수요를 동시에 자극한다. 따라서 엔비디아와 하위 밸류체인에 대한 매수 관점 접근이 여전히 유효하다고 판단한다.

엔비디아 전방 산업별 매출액 추이



자료: Bloomberg, 유진투자증권

엔비디아 데이터센터 매출 성장률 추이



자료: Bloomberg, 유진투자증권

네오클라우드: 중립성이라는 무기

네오클라우드 역시 오픈웨이트 생태계 확장의 직접적인 수혜자로 판단한다.

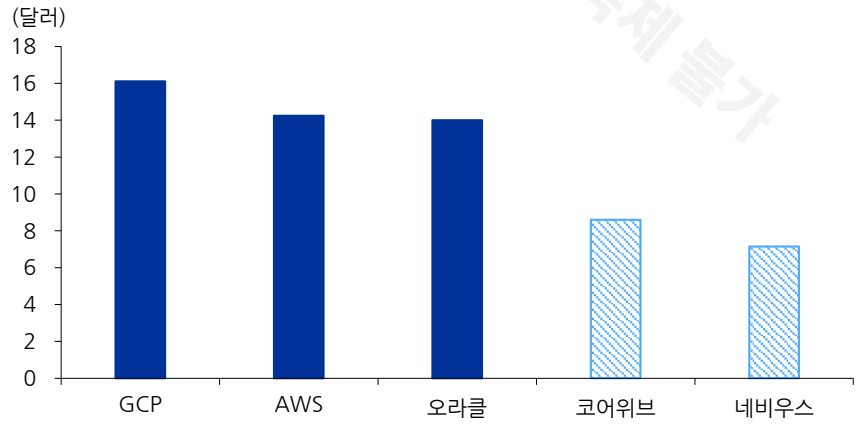
우선 네오클라우드는 하이퍼스케일러 대비 모델 이해관계가 중립적이다. 하이퍼스케일러는 자사 모델 및 프론티어 모델랩과의 이해관계로 인해 오픈웨이트 서빙에 구조적 제약이 있는 반면, 네오클라우드는 특정 모델에 대한 이해관계가 없어 다수의 커스텀 모델을 제약 없이 수용할 수 있다.

또한 네오클라우드는 범용성에 특화된 엔비디아 GPU 중심의 순수 클러스터를 보유하고 있다. 자체 ASIC 과 GPU 를 병행 운영하는 하이퍼스케일러와 달리, 네오클라우드는 CUDA 생태계 위에서 구축된 단일 아키텍처 클러스터를 운영한다. 이는 CUDA 기반으로 학습된 다양한 커스텀 모델을 별도의 포팅 없이 신속하게 서빙할 수 있는 기반이 된다.

아울러 오픈웨이트 모델은 프론티어 모델 대비 토큰 가격 메리트를 핵심 경쟁력으로 삼는 만큼, 서빙 인프라의 단가와 가용성이 사업성을 좌우하는 비중이 크다. **상대적으로 합리적인 컴퓨팅 가격을 제시하는 네오클라우드는 오픈웨이트 진영의 저가 전략과 자연스러운 시너지를 형성할 것으로 본다.**

네오클라우드는 상대적으로 합리적인 컴퓨팅 가격을 제시

하이퍼스케일러, 네오클라우드 B200 온디맨드 렌탈가 차이



자료: 각 사, 유진투자증권

아울러 네오클라우드는 서빙 플랫폼을 통해 부가가치를 확장하고 있다. 코어위브는 Managed Inference, 네비우스는 Token Factory 를 통해 오픈웨이트 모델을 직접 서빙 중이다. 이는 오픈소스·오픈웨이트 모델에 자사 소프트웨어·운영 레이어를 얹어 토큰 사용량 기반으로 과금하는 구조로, 기존 네오클라우드보다 하이퍼스케일러의 사업 모델에 가깝다. 이는 네오클라우드의 수익 구조를 단순 시간당 과금에서 사용량 기반 과금으로 확장시키며, 계약이 만료되어 감가상각이 종료된 구세대 GPU 를 토큰 서비스로 재배치해 부가가치를 극대화하는 효과까지 창출한다. 코어위브의 Managed Inference ARR 은 출시 수개월 만에 1 억달러를 돌파했으며, 연말까지 2.5 억달러 달성을 가이던스로 제시했다.

모델 서빙은 단순 GPU 렌탈 대비 부가가치를 극대화할 수 있음

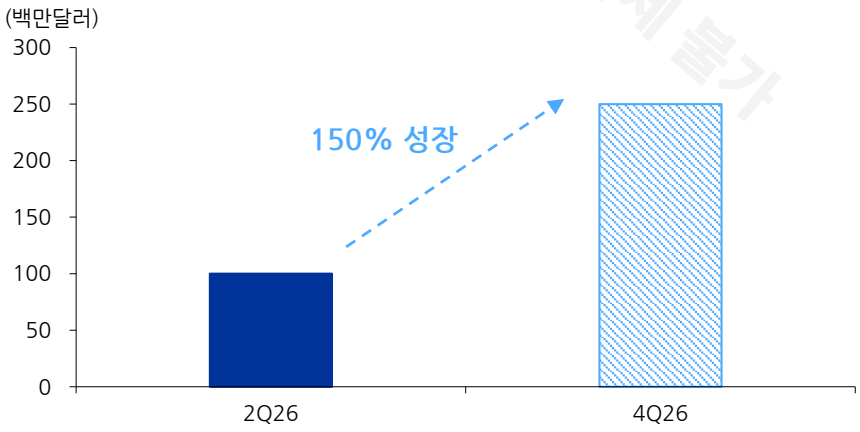
네오클라우드 GPU 렌탈과 모델 서빙 사업 구조 차이

구분	GPU 렌탈	모델 서빙
과금 구조	$/GPU-Hour(사용 시간)	$/1M Tokens(사용량)
매출 산식	GPU 임대 단가 x 가동률	토큰 판매 단가 - 추론 컴퓨트 비용
마진	상대적으로 낮음	높음
주요 고객	모델랩, AI 스타트업 등	엔터프라이즈 등
경쟁력	GPU 확보 능력, 단가 등	추론 최적화 SW, 단가 등

자료: 각 사, 유진투자증권

코어위브는 연말 Managed Inference ARR 을 2.5 억달러로 제시

코어위브 Managed Inference ARR 가이던스

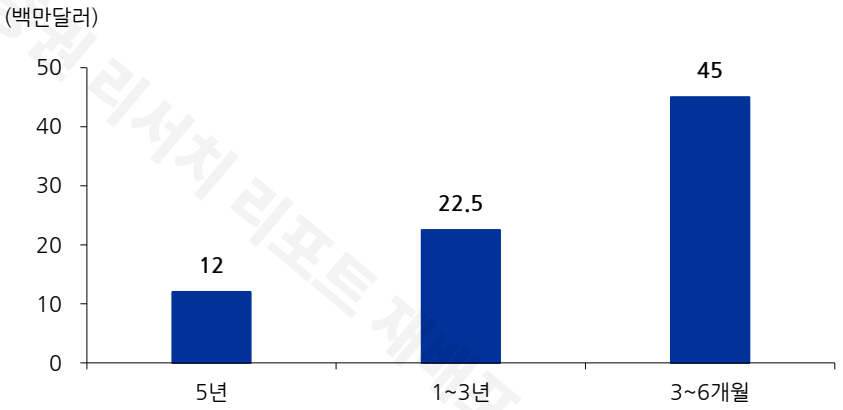


자료: 각 사, 유진투자증권

수요 환경 또한 우호적이다. 구글은 2 분기 실적 발표에서 컴퓨팅 캐파 부족을 해소하기 위해 3 분기 중 제 3 자 임대를 확대하겠다고 언급했으며, 이는 하이퍼스케일러조차 단기 수요를 충족시키지 못해 네오클라우드의 컴퓨트 캐파에 의존하는 상황임을 의미한다. **이에 더해 우호적인 GPU 렌탈가 환경이 지속되는 가운데 단기 컴퓨팅 수요가 급증하며, 계약 기간이 단축될수록 MW 당 ACV(연간 계약 가치)가 상승하는 구조가 형성되고 있다. 이는 주요 네오클라우드 업체들이 하이퍼스케일러 대비 불안정한 재무 구조로 인해 빠른 탑라인 성장이 요구되는 만큼 긍정적인 모멘텀으로 작용할 수 있다.**

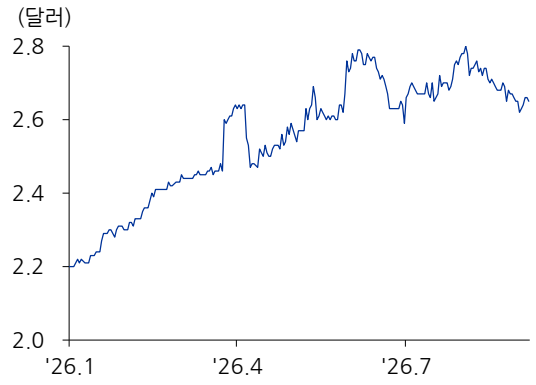
단기 수요 급증으로 인해 계약 기간이 단축될수록 MW 당 ACV 는 상승세

네비우스 계약 기간별 ACV



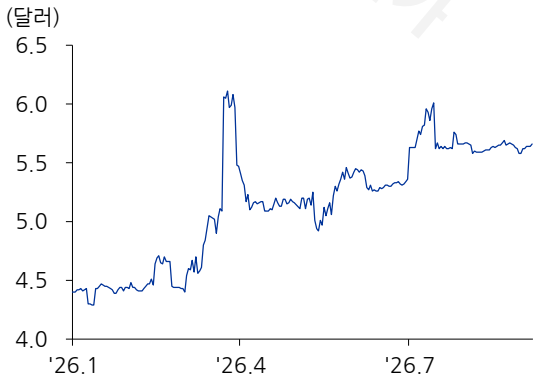
자료: Nebius, 유진투자증권

H100 렌탈가 추이



자료: Bloomberg, 유진투자증권

B200 렌탈가 추이



자료: Bloomberg, 유진투자증권

결론 및 투자 아이디어

오픈웨이트 모델의 확산은 거스를 수 없는 흐름이며, 이는 AI 산업의 저변을 구조적으로 확장시키는 요인이다. 핵심은 이러한 흐름이 **① 토큰 가격 하락으로 총 토큰 사용량을 증가시키는 제본스의 역설 실현, ② 커스터마이징 용이성과 데이터 주권 확보로 엔터프라이즈 온프레미스 수요 창출, ③ 프론티어 모델랩의 마진 압박으로 인한 차세대 추론 인프라 수요 확대로 이어질 수 있다는 점에서 AI 인프라 수요에 긍정적인 측면이 크다고 본다.**

해당 국면에서는 개방형 생태계에 최적화된 **엔비디아(Nvidia)**와 하위 밸류체인 및 네오클라우드 업체들에 대한 매수 관점이 유효할 것으로 판단한다.

엔비디아 하위 밸류체인에서는 **델 테크놀로지스(Dell Technologies)**를 선호주로 제시한다. 동사는 ① 견조한 AI 서버 매출 성장세를 보이고 있으며, ② 6,500 개를 상회하는 분산된 고객 기반과, ③ AI Factory, Enterprise Hub 등을 통해 오픈웨이트 모델의 온프레미스 배포를 상품화한 유일한 서버 사업자로, 오픈웨이트 모델의 온프레미스 확산 수혜 강도가 가장 높은 업체 중 하나로 판단한다.

네오클라우드 업종 내에서는 **네비우스(Nebius)**를 선호주로 제시한다. 모델 이해관계가 중립적이고 엔비디아 인프라 아키텍처 기반 클러스터를 운영한다는 업계 공통의 강점에 더해, ① 엄격한 선수금 수취로 재무 구조가 상대적으로 건전하고, ② 공격적인 계약 중단기 믹스 전환을 통해 MW 당 ACV 를 개선하고 있으며, ③ Token Factory 라는 오픈웨이트 서빙 플랫폼을 보유하고 있다는 점에서 경쟁사와 차별화된다. 경쟁사 대비 높은 밸류에이션에도 불구하고, 업종 내 가장 매력적인 투자처로 판단한다.

선호주 리스트

업종	기업명	시가총액 (십억달러)	12M Fwd PER (배)	투자포인트
팹리스	엔비디아	5,552	18	■ 오픈웨이트 생태계 확장의 최대 수혜 ■ 범용성·CUDA 락인·통합 솔루션 3중 해자
서버 ODM	델 테크놀로지스	339	18.6	■ 오픈웨이트 온프레미스 배포 상품화 ■ 네오클라우드·엔터프라이즈·소버린 동시 노출 ■ 범용서버·스토리지 확대로 마진 방어
네오클라우드	네비우스	65	-	■ 모델 중립성 기반 오픈웨이트 서빙 수요 확보 ■ 엄격한 선수금 구조로 재무 안정성 확보 ■ Token Factory로 오픈웨이트 토큰 서빙 확장

자료: 유진투자증권

III. Appendix

클로즈드·오픈웨이트·오픈소스 모델의 구분

클로즈드(Closed) 모델은 가중치를 공개하지 않고 API 를 통해서만 접근을 허용하는 모델이다. 오픈 AI 의 GPT 시리즈, 앤트로픽의 Claude, 구글의 Gemini 가 대표적이다. 사용자는 토큰 사용량 기준으로 과금하며, 모델은 개발사의 인프라 혹은 제 3 자 클라우드 인프라 위에서 구동된다.

오픈웨이트(Open-Weight) 모델은 학습이 완료된 파라미터, 즉 가중치를 공개하는 모델이다. 사용자는 이를 내려받아 자체 인프라에서 구동하거나 변형할 수 있다. 다만 학습에 사용된 데이터셋과 학습 레시피는 공개되지 않는 경우가 대부분이다. DeepSeek, Kimi, Qwen, GLM 등 최근 부상한 모델들이 여기에 해당한다. 가중치가 무료로 배포되더라도 이를 구동하기 위한 가속기, 메모리, 전력, 운영 인력은 사용자가 부담한다. 즉 클로즈드 모델에서 모델랩이 지불하던 추론 인프라 비용이 사용자 측으로 이전되는 구조에 가깝다.

오픈소스(Open-Source) 모델은 가중치뿐 아니라 학습 데이터와 코드까지 공개해 학습 과정 전체의 재현이 가능한 모델이다. 주로 허깅페이스 등의 플랫폼을 통해 배포되며, 아직까지 대규모 모델에서는 사례가 제한적이다.

오픈소스·오픈웨이트·클로즈드 모델의 차이

구분	클로즈드	오픈웨이트	오픈소스
가중치	비공개	공개	공개
학습 데이터	비공개	비공개	공개
자체 인프라 구동	불가능	가능	가능
커스터마이징	제한적	가능	가능
과금 구조	토큰 사용량	인프라 비용	인프라 비용
예시	GPT, Claude	Kimi, Qwen, GLM	OLMo

자료: 유진투자증권

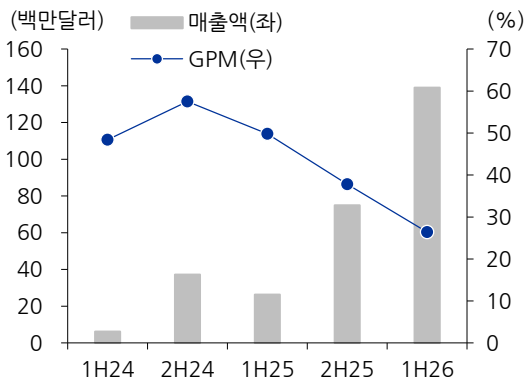
오픈웨이트 모델랩의 비즈니스 모델

오픈웨이트 모델랩은 가중치를 무상 공개하면서도 동시에 자체 API 를 통해 매출을 창출한다. GLM 개발사 Z.ai 의 매출은 크게 자체 API 서비스와 엔터프라이즈 온프레미스 구축 서비스로 구분된다. 전자는 자사 인프라에서 모델을 호스팅해 토큰 사용량 기준으로 과금하는 구조이며, 후자는 고객사 자체 인프라에 모델을 배포·최적화해주는 프로젝트성 서비스다.

모델랩의 수익원이 될 수 있는 가중치를 공개하는 것은, 후발 주자의 유통 전략으로 해석하는 것이 합리적이다. 프론티어 모델랩 대비 브랜드와 신뢰도가 열위한 상황에서, 무상 배포는 채택률을 빠르게 끌어올려 사실상의 표준 지위를 확보하는 수단이 된다. 실제로 허깅페이스 기준 Qwen 계열 파생 모델은 15 만개를 상회하는데, 이는 원 개발사가 아닌 외부 개발자들이 만든 결과물이며, 파생 모델이 축적될수록 해당 계열이 개발자들의 기본 선택지로 자리 잡는 구조이다. 다만 최근에는 일부 오픈웨이트 모델랩들이 대형 기업을 대상으로 라이선스 비용을 받기 시작하며, 확보한 생태계를 수익화하는 단계로 넘어가는 움직임도 포착된다.

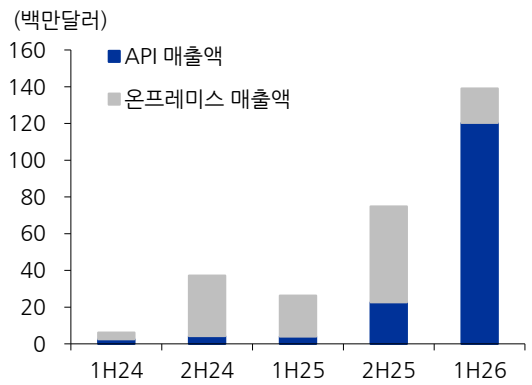
오픈웨이트 모델랩은 아직까지 모델 계층에서의 수익화 폭이 구조적으로 제한적이다. 이는 엔터프라이즈 자체 인프라에서 구동되는 사용량은 모델랩의 매출로 잡히지 않기 때문이다. **즉 모델 사용이 확산될수록 창출되는 가치가 모델 계층이 아니라 인프라 계층으로 이전되는 구조다. 본 리포트가 오픈웨이트 확산이 AI 인프라 투자에 긍정적이라고 보는 근본적인 배경이기도 하다.**

Z.ai 매출액, GPM 추이



자료: Bloomberg, 유진투자증권

Z.ai 세그먼트별 매출액 추이



자료: Bloomberg, 유진투자증권

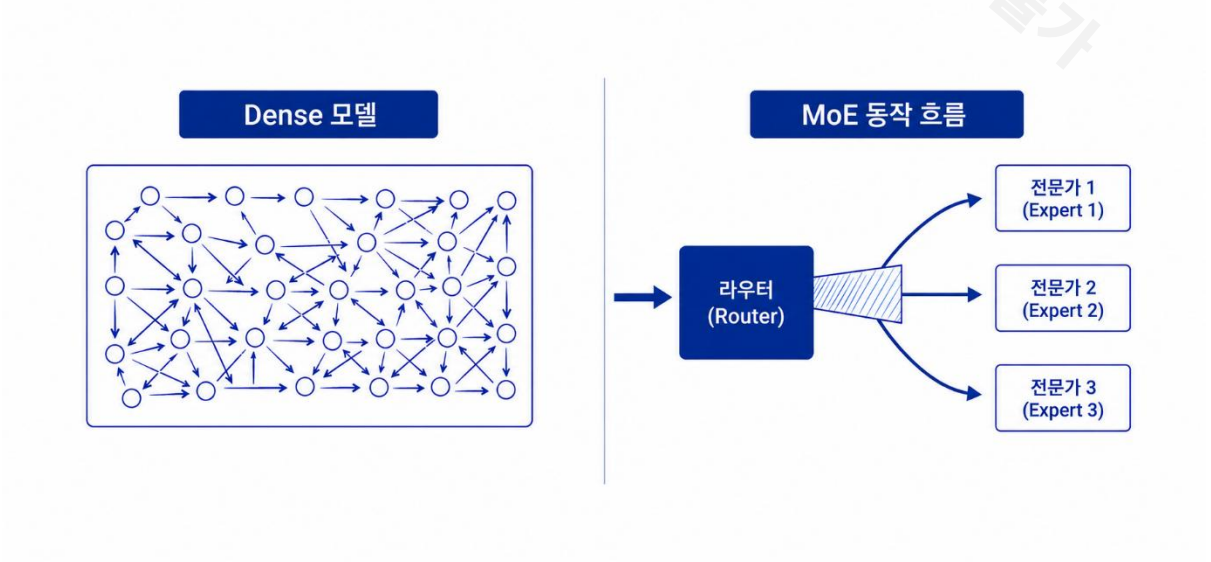
MoE(Mixture of Experts) 기법

MoE 는 모델을 다수의 전문가(Expert) 네트워크로 분할하고, 입력 토큰마다 라우터(Router)가 소수의 전문가만을 선택해 활성화하는 구조다. 전체 파라미터를 모든 토큰에 대해 연산하는 밀집(Dense) 구조와 달리, MoE 는 토큰별로 필요한 전문가만 사용하기에 동일한 파라미터 규모에서 연산량을 크게 절감할 수 있다.

이에 따라 MoE 모델에서는 총 파라미터(Total Parameter)와 활성 파라미터(Active Parameter)가 분리된다. 총 파라미터는 모델이 보유한 전체 파라미터의 수를, 활성 파라미터는 토큰 하나를 처리할 때 실제로 연산에 참여하는 파라미터의 수를 의미한다. 최근 조 단위 총 파라미터를 보유한 오픈웨이트 모델들이 상대적으로 낮은 토큰 가격을 제시할 수 있는 이유가 여기에 있다.

다만 인프라 관점에서 핵심은 두 파라미터가 서로 다른 자원을 요구한다는 점이다. 활성 파라미터는 연산량(FLOPs)을 결정하는 반면, 요구되는 메모리 용량은 총 파라미터가 결정한다. 어떤 토큰이 어떤 전문가로 라우팅될지 사전에 예측할 수 없으므로, 전체 가중치가 항상 가속기 메모리에 상주해야 하기 때문이다. **즉, MoE는 연산 부하는 일부 줄이는 효과가 있지만 전체적인 메모리 부담은 줄이지 못하는 구조이다.**

MoE 동작 흐름



자료: Maarten Grootendorst, 유진투자증권

사후학습(Post Training)의 3 가지 방식

AI 가 다양한 작업에 활용되기 시작함에 따라 단순 학습을 넘어 학습이 완료된 모델을 특정 목적에 맞게 조정하는 사후 학습(Post-Training)의 중요성이 높아지고 있다.

지식 증류(Knowledge Distillation)는 성능이 우수한 대형 교사(Teacher) 모델의 출력을 소형 학생(Student) 모델이 모방하도록 학습시키는 기법이다. 처음부터 대규모 데이터로 사전학습하는 것에 비해 학습 비용을 크게 절감할 수 있어, 후발 모델랩이 빠르게 성능을 확보하는 경로로 활용된다. 다만 유의할 점은 증류가 절감하는 것은 학습 비용이며 추론 비용과는 무관하다는 것이다. 증류를 통해 확보된 모델이라고 하더라도 파라미터 규모가 크다면 추론 단계에서 요구되는 인프라 부담은 동일하게 발생한다.

파인튜닝(Fine-Tuning)은 사전학습된 모델의 파라미터를 특정 도메인 데이터로 추가 학습시켜 조정하는 과정이다. 전체 파라미터를 갱신하는 방식과, 소수의 추가 파라미터만 학습시켜 연산 부담을 낮추는 방식으로 나뉜다. 후자의 경우 상대적으로 소규모 인프라로도 수행이 가능해, 엔터프라이즈가 자사 데이터로 모델을 커스터마이징하는 주된 경로가 된다.

강화학습(Reinforcement Learning)은 정답 데이터를 직접 학습시키는 대신, 모델의 출력에 대한 보상 신호를 기반으로 행동을 교정하는 방식이다. 인간 피드백을 보상으로 활용하는 RLHF 와, 수학·코드처럼 정답 검증이 가능한 영역에서 자동 채점 결과를 보상으로 활용하는 RLVR 등이 있다. 최근 추론 특화 모델의 성능 향상은 상당 부분 강화학습 단계의 고도화에 기인한다. 특히 에이전트 워크로드에서 태스크 결과에 대한 보상만으로도 이후 작업의 성능을 개선시킬 수 있다는 점에서, 강화학습은 에이전트 성능 확보의 핵심으로 기능한다.

기업분석

Nvidia(NVDA.US)
개방형 진영의 설계자

Dell Technologies(DELL.US)
온프레미스 AI의 관문

Nebius Group(NBIS.US)
성장과 안정성을 모두 잡다

CoreWeave(CRWV.US)
빠른 길, 더 무거운 짐

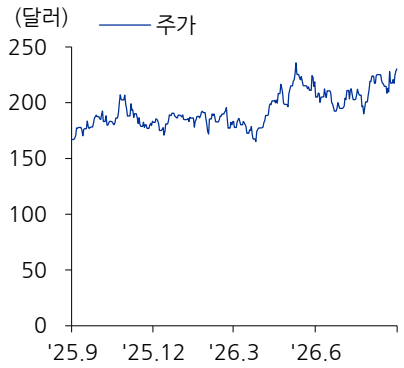
(NVDA.US)

Nvidia

개방형 진영의 설계자

투자의견	NR
현재주가	230.36 USD (9/4)
시가총액	5,552(십억달러) / 7,470 (조원)

- 엔비디아의 2분기 실적은 매출 962.2억달러(+106%yoy), GPM 75%, EPS 2.22달러로 컨센을 상회. 데이터센터 매출은 890.2억달러(+117%yoy)를 기록하며 4분기 연속 성장세가 가속화. 컴퓨트 매출액은 전년비 111%, 네트워킹 매출액은 전년비 141% 성장한 것으로 추정. 3분기 가이던스는 1,080억달러(+90%yoy)를 제시하며 컨센을 상회했으나, **GPM은 74%를 제시하며 컨센을 하회. 메모리 가격 상승으로 인해 마진 압박이 지속되고 있으며, 4분기 71.5% 저점 형성 이후 F28에는 72.5%로 일부 회복할 전망.**
- **전방위적 AI 인프라 수요로 인해 F28 매출 성장률을 70% 수준으로 제시하며 컨센 45%를 크게 상회.** 이는 강도 높은 루빈 플랫폼 수요와 서버 가격 인상분이 동시에 반영된 수치로 추정. 고객들의 수요는 100% 수준이나, 타이트한 공급 환경으로 70% 달성이 최대치일 정도로 수요가 강함을 언급.
- Vera CPU와 Groq LPU의 수요 환경 또한 긍정적. 견조한 에이전틱 워크로드 수요로 CPU 단독 매출 200억달러 가이던스를 재확인하였으며, F28에는 매출이 400억달러 이상으로 성장할 것을 언급. Groq LPU는 3분기부터 정상 출하 예정이며, 네비우스를 첫 고객으로 확보.
- 2분기 ACIE 매출은 403억달러로 전년비 134% 성장. 주요 네오클라우드 업체들의 데이터센터 용량은 전년 3GW에서 올해 말 8GW 수준으로 확대될 예정. 우호적인 GPU 렌탈가 환경이 지속됨에 따라 MW 당 ACV(연간 계약 가치)가 지속 상승하는 중. 네오클라우드 업체들의 성장에 따라 엔비디아 기반 인프라 수요 역시 동시에 확대되는 구조.
- **오픈웨이트 모델 생태계의 빠른 확장 역시 엔비디아에 우호적.** 오픈웨이트 생태계는 사용자가 자체 데이터로 변형한 커스텀 모델 수요가 지배적이어서, 특정 모델 구조에 최적화된 ASIC 대비 GPU의 범용성이 부각되는 환경. 또한 주요 오픈웨이트 모델 대부분이 엔비디아 소프트웨어 스택 위에서 학습·최적화되고 있으며, 자체 생태계가 전무한 네오클라우드·소버린·엔터프라이즈가 주된 수요처라는 점에서 턴키 AI 인프라 솔루션 역량을 보유한 엔비디아의 수혜 강도가 가장 높다고 판단.
- 하이퍼스케일러의 ASIC 개발과 주요 프론티어 모델랩들의 ASIC 사용량 증가는 엔비디아에게 지속적인 우려로 작용. **다만 ① 네오클라우드·소버린AI·오픈웨이트 생태계가 확산되며 AI 산업 전반의 엔비디아 아키텍처 Lock-in은 더욱 강화되고 있으며, ② 우월한 공급망 관리 능력에 기반한 F28까지의 성장 가시성을 확보한 점 등을 감안하면 엔비디아의 12M Fwd PER은 18배 수준으로 여전히 매수하기에 매력적인 밸류에이션으로 판단.**



현지명	Nvidia Corp
한글명	엔비디아
시가총액(십억달러/조원)	5,552/7,470
설립연도	1993년
설립자	Jensen Huang (공동 창업)
본사 위치	미국 캘리포니아
현 CEO	Jensen Huang
52주 최고/최저(달러)	237/164
배당수익률(26E, %)	0.1
주요주주 지분율(%)	
뱅가드	9.5
블랙록	8.0
스테이트 스트리트	4.1

주가상승률(%)	1M	6M	YTD
	5.2	29.5	23.5

(Non GAAP)	FY2026	FY2027E	FY2028E
매출액(백만달러)	215,938	408,098	675,969
영업이익(백만달러)	137,300	267,782	446,350
당기순이익(백만달러)	116,997	228,198	376,301
EPS(달러)	4.93	9.35	15.42
증감률(%)	59.5	95.9	65.0
PER(배)	33.1	24.6	14.9
ROE(%)	117.2	95.2	79.3
PBR(배)	24.3	17.5	9.9
EV/EBITDA(배)	33.6	19.8	11.8

엔비디아의 턴키 AI 솔루션

엔비디아의 **베라 루빈 팟(Vera Rubin Pod)**은 ① AI 연산을 담당하는 Vera Rubin NVL72, ② 시스템 오케스트레이션을 담당하는 Vera CPU 랙, ③ Decode 연산을 분담하는 Groq 3 LPX, ④ KV 캐시를 오프로딩하는 BlueField-4 STX, ⑤ 스케일아웃 네트워킹을 담당하는 Spectrum-6 SPX로 구성된다. 루빈 플랫폼은 AI 모델의 사전·사후학습부터, 추론, 에이전트 워크로드까지 AI 생태계 전반을 지원한다.

이에 더해 엔비디아는 AI 팩토리 설계·구축 플랫폼인 **DSX**를 공개했다. DSX 플랫폼은 단순 AI 반도체 시스템을 넘어 AI 인프라의 설계안, 시뮬레이션, 워크로드 조정, 운영 OS 등을 제공한다. 엔비디아는 단순 반도체 판매를 넘어 설계·시뮬레이션·운영·전력 최적화까지 제공하는 통합 AI 인프라 기업으로 확장 중이다.

엔비디아 베라 루빈 플랫폼구성

구분	VR NVL72	Vera CPU 랙	Groq 3 LPX	BlueField-4 STX	Spectrum-6 SPX
주요 프로세서	Rubin GPU	Vera CPU	Groq 3 LPU	BlueField-4	Spectrum-6 CPO
주요 프로세서 탑재량	72	256	256	32	변동
주요 메모리	HBM	LPDDR	SRAM	SSD	-
역할	AI 연산	시스템 오케스트레이션	추론 Decode 연산 분담	KV 캐시 저장 및 전송	스케일아웃 네트워크

자료: Nvidia, 유진투자증권

DSX 플랫폼 구성 요소

구분	구성 요소	역할
설계 표준	DSX Reference Design	세대별 검증된 AI 팩토리 설계안 제공
사전 검증	DSX Sim	구축 전 AI 팩토리 시뮬레이션 진행
전력 연계	DSX Flex	전력망 신호에 따른 워크로드 조정
신호 통합	DSX Exchange	IT·OT·전력·냉각 신호 통합
전력 효율	DSX MaxLPS	고정 전력 내 토큰 성능 극대화
운영 체계	DSX OS	AI 팩토리 운영 자동화 소프트웨어

자료: 엔비디아, 유진투자증권

이러한 시스템 단위 접근은 특히 네오클라우드·엔터프라이즈·소버린 수요처에서 강력한 경쟁력으로 작용한다. 세 수요처의 공통점은 하이퍼스케일러와 달리 자체 인프라 설계 역량과 소프트웨어 생태계가 빈약하다는 점이다. 하이퍼스케일러는 수년간 축적한 데이터센터 운영 경험과 자체 ASIC·네트워크 스택을 보유하고 있어 개별 구성 요소를 조합해 클러스터를 설계할 수 있으나, 신규 수요처는 그렇지 않다. 이들에게는 랙 배치 비율부터 네트워크, 냉각·전력 설계까지 검증된 레퍼런스가 필요하며, 엔비디아 플랫폼은 사실상 그 레퍼런스로 기능한다.

범용성 역시 핵심 해자로 작용한다. 오픈웨이트 생태계에서는 사용자가 자체 데이터로 변형한 커스텀 모델 수요가 지배적이며, 워크로드 또한 학습·Prefill·Decode·에이전트 루프로 파편화되어 있다. 특정 모델 구조에 최적화된 ASIC 과 달리 엔비디아 아키텍처는 팟 내부에서 워크로드별로 연산 자원을 분담하는 구조를 취하고 있어, 사전에 워크로드를 특정하기 어려운 환경에서 대응력이 높다.

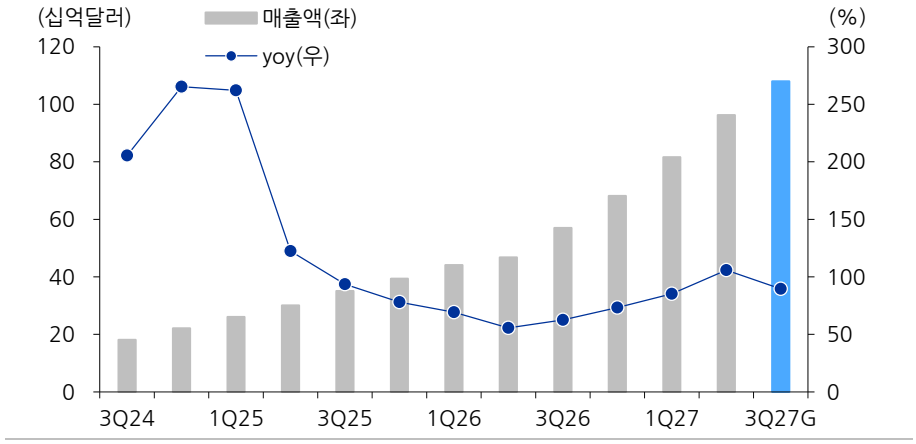
결론적으로 엔비디아는 턴키 솔루션, 검증된 생태계, 워크로드 범용성이라는 세 요소를 결합함으로써, 자체 생태계가 부재한 신규 수요처의 진입 장벽을 낮추는 역할을 수행한다. **이는 엔비디아 아키텍처의 락인 효과를 강화하는 동시에, ACIE 부문의 구조적 성장을 뒷받침하는 요인으로 작용할 전망이다.**

엔비디아 DSX 플랫폼

자료: 엔비디아, 유진투자증권

루빈 플랫폼 램프가 본격화되는 4 분기부터 매출 성장세 재가속 전망

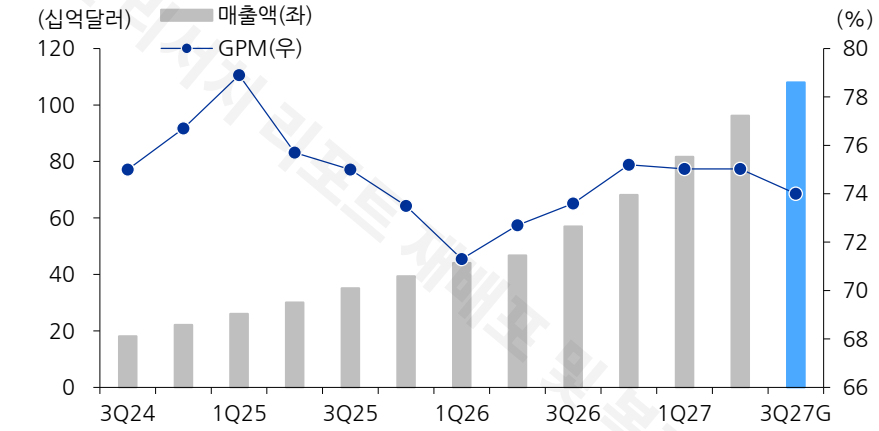
엔비디아 분기별 매출액, yoy 성장률 추이 및 가이던스



자료: Bloomberg, 유진투자증권

메모리 인플레이션으로 인해 3 분기 GPM 가이던스는 74%로 컨센 하회

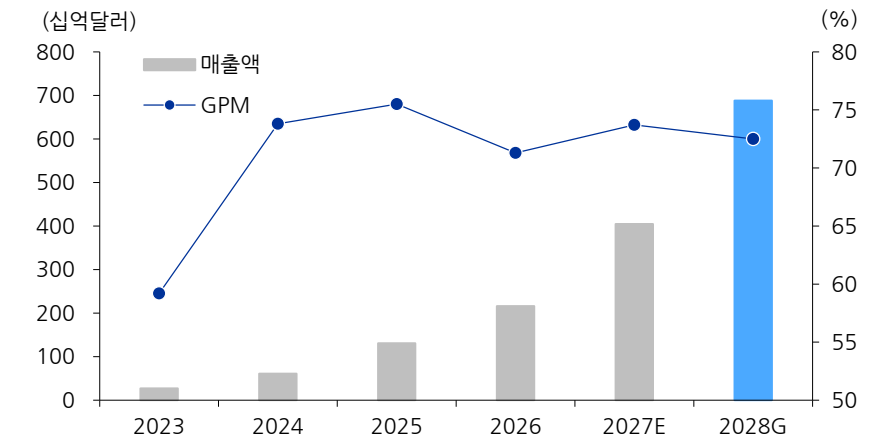
엔비디아 분기별 매출액, GPM 추이 및 가이던스



자료: Bloomberg, 유진투자증권

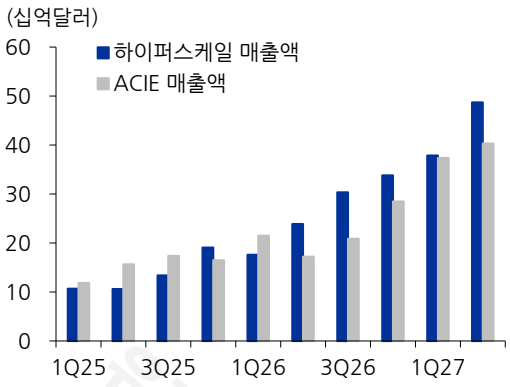
F28 70%의 매출 성장이 예상되나, 메모리 인플레이션으로 인해 GPM 은 72.5% 제시

엔비디아 연간 매출액, GPM 추이 및 가이던스



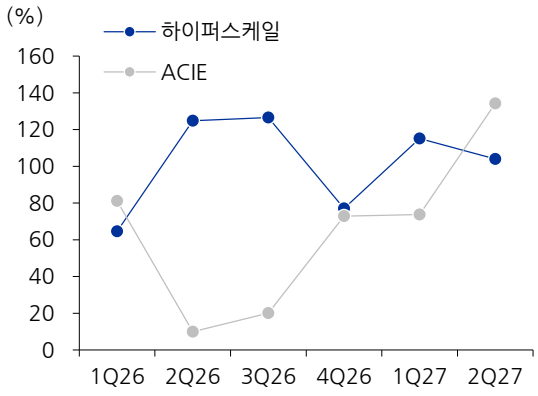
주: 매출 기준 FY3Q27, FY2028 매출액은 가이던스, FY4Q27 은 컨센서스를 적용
자료: Bloomberg, 유진투자증권

엔비디아 전방 산업별 매출액 추이



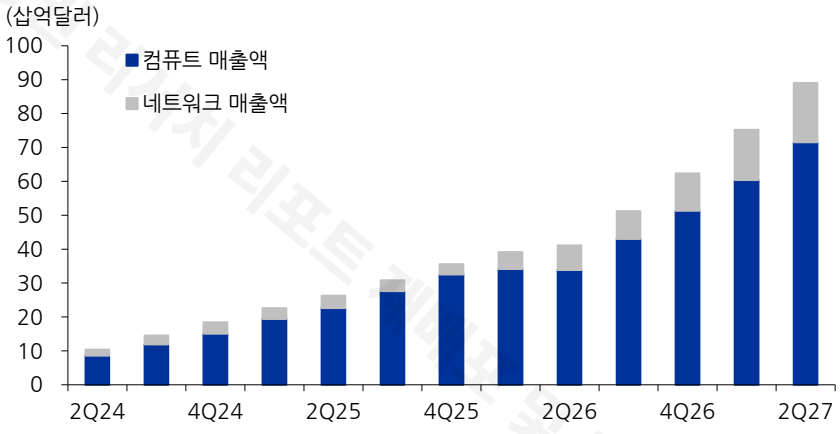
자료: Bloomberg, 유진투자증권

엔비디아 전방 산업별 매출 성장률 추이



자료: Bloomberg, 유진투자증권

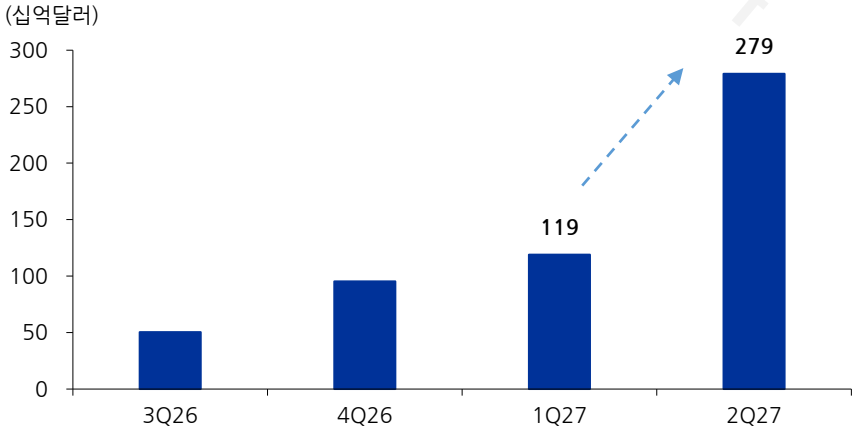
엔비디아 데이터센터 매출액 세그먼트별 추이



자료: Bloomberg, 유진투자증권

메모리 물량 확보로 인해 2 분기 구매 약정은 1,600 억달러 증가

엔비디아 생산·공급 관련 구매 약정 추이



자료: EDGAR, 유진투자증권

글로벌 소버린 AI 프로젝트는 수십 GW 규모로 확대

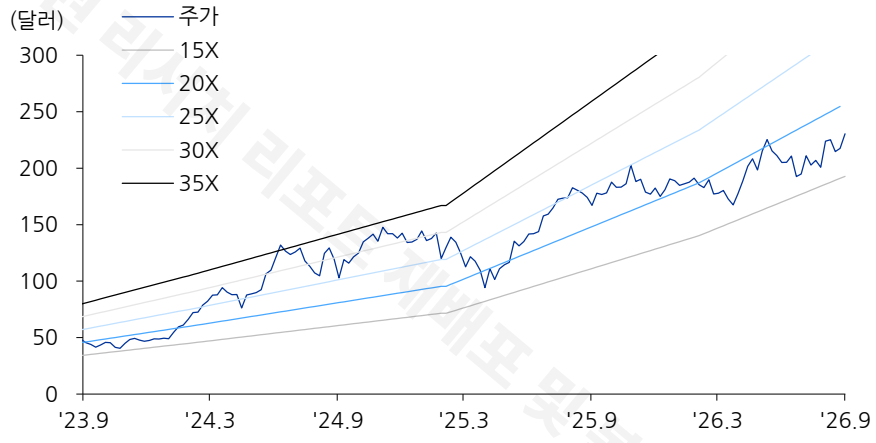
주요 소버린 AI 프로젝트

지역·국가	프로젝트 주체	하드웨어(예상)	목표 규모(예상)
아르메니아	Firebird AI	블랙웰·루빈	2GW
아프리카·발칸	Thinking Machines Lab	루빈	1GW
대만	GMI Cloud	블랙웰·루빈	1GW
인도	Yotta Data Services	블랙웰	2GW
말레이시아	YTL AI Cloud	블랙웰	600MW
대한민국	SKT·네이버 등	블랙웰·루빈	10GW 이상
프랑스	Mistral AI	블랙웰	1GW
UAE	Core42	블랙웰	1GW
사우디아라비아	HUMAIN	블랙웰	6.6GW

자료: 언론 종합, 유진투자증권

엔비디아의 Fwd PER 은 18 배 수준으로 주가 상승에도 불구하고 여전히 매력적

엔비디아 12M Fwd PER 밴드



자료: Refinitiv, 유진투자증권

엔비디아 실적 추이 및 컨센서스

Non-GAAP(십억달러)	1Q27	2Q27	3Q27E	4Q27E	1Q28E	2Q28E	3Q28E	4Q28E	2025	2026	2027E	2028E
매출액	81.6	96.2	108.2	123.2	142.8	162.0	182.2	203.9	130.5	215.9	408.1	676.0
yoy	*85%*	*106%*	*90%*	*81%*	*75%*	*68%*	*68%*	*65%*	*114%*	*65%*	*89%*	*66%*
데이터센터 매출액	75.2	89.0	100.9	115.4	134.6	152.8	171.0	190.8	115.2	193.7	388.6	658.1
yoy	*92%*	*117%*	*97%*	*85%*	*79%*	*72%*	*70%*	*65%*	*142%*	*68%*	*101%*	*69%*
매출총이익	61.2	72.2	80.1	89.4	103.8	118.1	133.1	149.1	98.5	154.0	302.0	493.1
GPM	75.0%	75.0%	74.0%	72.6%	72.7%	72.9%	73.0%	73.1%	75.5%	71.3%	74.0%	72.9%
영업이익	53.8	64.0	71.5	79.2	92.8	105.8	120.3	135.7	86.8	137.3	267.8	446.3
OPM	65.9%	66.5%	66.1%	64.3%	65.0%	65.3%	66.0%	66.5%	66.5%	63.6%	65.6%	66.0%
순이익	45.5	54.0	60.0	66.8	77.7	88.3	99.0	111.2	74.3	117.0	228.2	376.3
NPM	55.8%	56.1%	55.5%	54.2%	54.4%	54.5%	54.4%	54.5%	56.9%	54.2%	55.9%	55.7%
EPS (달러)	1.87	2.22	2.46	2.73	3.17	3.62	4.09	4.59	2.99	4.77	9.35	15.42
yoy	*95%*	*111%*	*89%*	*69%*	*69%*	*63%*	*66%*	*68%*	*131%*	*60%*	*96%*	*65%*

주: 기간은 FY 기준을 적용
자료: Bloomberg, 유진투자증권

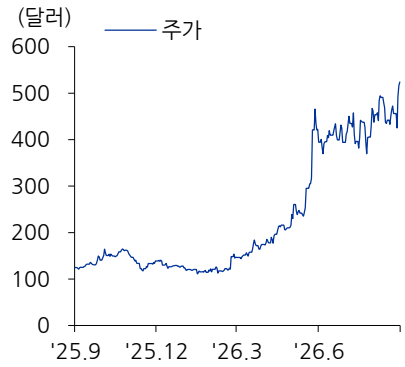
편집상의 공백페이지입니다

(DELL.US)

Dell Technologies

온프레미스 AI의 관문

투자의견	NR
현재주가	524.1 USD(9/4)
시가총액	339(십억달러)/456(조원)



- 델의 2분기 매출은 470억달러(+58%yoy), GPM 21.1%, OPM 12.6%, EPS 7.04로 컨센을 상회. AI 서버발 탑라인 성장으로 인한 영업 레버리지, 스토리지 믹스 개선으로 수익성 개선이 두드러짐. 3분기 가이던스도 매출 490억달러(+81%yoy), EPS 6.5달러를 제시하며 컨센을 상회했으며, F27 연간 가이던스는 매출 1,920억달러, EPS 25.5달러로 상향 조정.
- AI 서버 매출은 164억달러(+100%yoy)를 기록하며 탑라인 성장을 견인. **전방위적 AI 수요 강세로 인해 델이 강점을 가진 네오클라우드·엔터프라이즈·소버린이 가파르게 성장 중.** AI 서버 주문은 609억달러로 전분기 대비 150% 증가하며 역대 최고치를 달성했고, 분기말 AI 서버 백로그는 950억달러로 전년비 7배 이상 증가.
- 엔터프라이즈향 IT 서버 교체 사이클과 AI 데이터센터향 에이전틱 워크로드로 인한 범용 서버 수요가 가중되며 범용 서버 매출은 105억달러(+122%yoy)를 기록. 이에 더해 추론 워크로드의 KV 캐시 급증으로 인해 스토리지의 중요성이 상승하는 가운데, Dell-IP 기반 스토리지로의 믹스 전환이 이루어지는 중. AI 서버 외 매출까지 강세를 보이는 동시에 스토리지 믹스 개선이 이루어지며 ISG OPM은 15%로 전분기 대비 4%p 개선.
- **대형 고객사향 AI 서버 수요도 지속적으로 높아지는 상황.** 델의 주요 고객사인 스페이스X는 2027년까지 10GW 용량 가이던스를 제시했으며, 코어위브는 2027년까지 3GW 이상의 용량을 구축할 것으로 추정.
- **또한 오픈웨이트 모델의 확산으로 AI 산업의 저변 확장이 기대되는 국면에서, 델은 수혜 강도가 가장 높은 서버 사업자로 판단.** 델은 오픈웨이트 온프레미스 배포를 상품화한 유일한 서버 사업자로, 자체 IP 및 설계 역량으로 직접 아키텍처 설계가 어려운 네오클라우드·엔터프라이즈 고객 대응이 용이하다는 점도 AI Factory와의 시너지 요인.
- 엔터프라이즈 AI 도입은 변곡점을 통과한 것으로 판단. AI Factory 누적 고객 수는 2분기 기준 6,500개사를 돌파했는데, 첫 3,200개 확보에 8개 분기가 소요된 반면 이후 3,300개는 3개 분기 만에 추가. 경영진 또한 가속의 주체가 엔터프라이즈임을 명시.
- 델의 12M Fwd PER은 18.6배 수준으로 경쟁사(HPE, SMCI 등) 대비 높은 밸류에이션을 받고 있음. 델 역시 서버 ODM의 고질적 리스크인 메모리 인플레이션으로 인한 잠재적 마진 부담, 백로그 이행 속도 등에서 자유롭지는 못함. **다만 견조한 대형 AI 고객사의 캐팩스 투자와 더불어 엔터프라이즈향 오픈웨이트 모델 확산 수혜에 가장 크게 노출되어 있다고 판단하며, 리스크에도 불구하고 가장 높은 성장 포텐셜을 보유하였다고 판단.**

현지명	Dell Technologies Inc.
한글명	델 테크놀로지스
시가총액(십억달러/조원)	339/456
설립연도	2013년
설립자	Michael S. Dell
본사 위치	미국 텍사스
현 CEO	Michael S. Dell
52주 최고/최저(달러)	535/110
배당수익률(26E, %)	0.4
주요주주 지분율(%)	
뱅가드	9.5
블랙록	7.4
마이클 델	5.8

주가상승률(%)	1M	6M	YTD
	19.8	257.8	316.4

(Non GAAP)	FY2026	FY2027E	FY2028E
매출액(백만 달러)	113,538	193,211	233,640
영업이익(백만 달러)	9,991	21,849	25,347
당기순이익(백만 달러)	7,046	16,651	19,392
EPS(달러)	10.30	25.87	30.29
증감률(%)	26.5	151.1	17.1
PER(배)	30.7	20.3	17.3
ROE(%)	N/A	393.9	63.0
PBR(배)	N/A	129.1	33.3
EV/EBITDA(배)	7.7	15.0	12.9

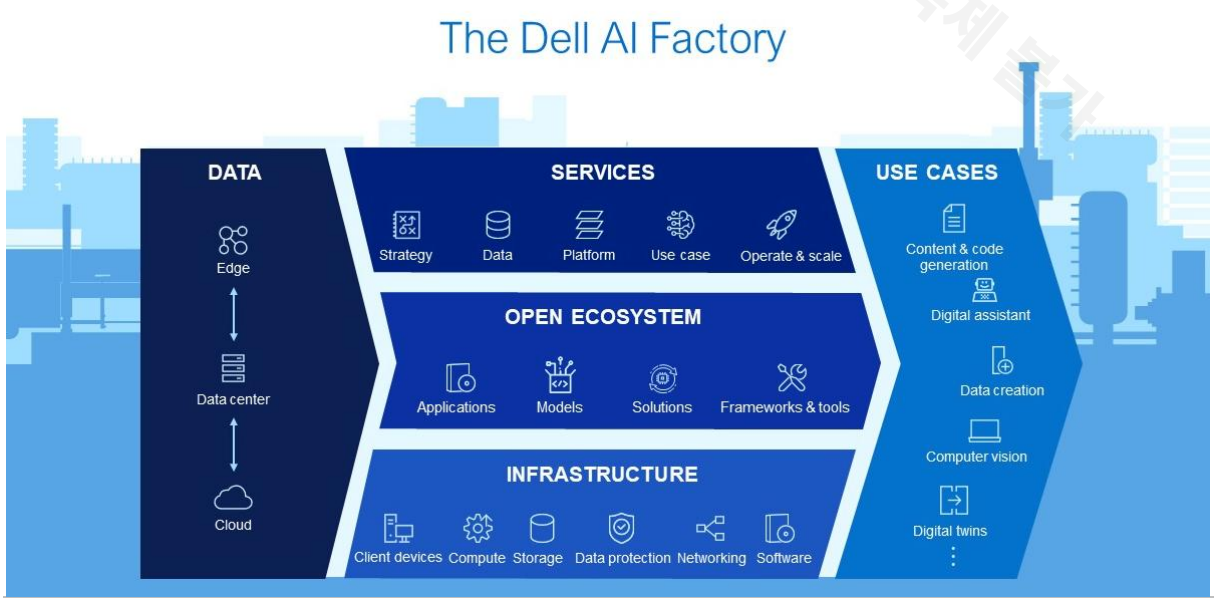
델 AI Factory

델의 AI Factory는 단일 제품이 아니라 AI 인프라 도입에 필요한 하드웨어·소프트웨어·서비스를 통합한 턴키 프레임워크다. AI 서버와 스토리지, 네트워킹, 냉각·전력 설비 등 하드웨어 계층부터 데이터 플랫폼 등 소프트웨어 계층, 나아가 컨설팅·배포·운영 서비스와 파이낸싱까지 포괄한다. 엔비디아와 공동 검증된 구성과 AMD 기반 오픈 아키텍처 구성을 병행 제공하며, 오픈 AI·구글·팔란티어 등 주요 모델·소프트웨어 사업자와의 파트너십을 통해 생태계를 확장하고 있다.

이 중 오픈웨이트 확산 국면에서 가장 주목할 요소는 허깅페이스와 공동 운영하는 Dell Enterprise Hub 다. 가중치가 공개되어 있다는 것과 실제로 구동할 수 있다는 것 사이에는 컨테이너 구성, 병렬화 설정, 양자화 선택 등 상당한 기술적 간극이 존재해 왔다. Enterprise Hub 는 DeepSeek, Kimi, GLM 등 최신 오픈웨이트 모델을 플랫폼별 최적화와 소프트웨어 종속성이 통합된 배포 패키지 형태로 제공하며, 공개 수시간 내에 검증된 컨테이너를 준비한다.

AI Factory 는 자체 AI 인프라 인력을 보유하지 못한 엔터프라이즈에게도 오픈웨이트 도입 경로가 열린다는 점에서 의미가 있으며, 델이 오픈웨이트 확산 국면에서 강도 높은 수혜를 받을 수 있는 근거로 작용한다.

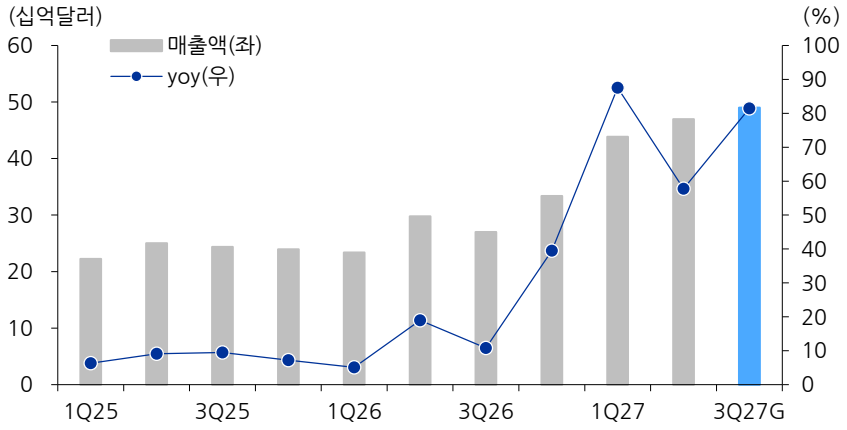
델 AI Factory



자료: Dell Technologies, 유진투자증권

AI 서버 수요 강세로 2 분기 매출은 전년비 58%를 기록, 3 분기 가이던스는 전년비 81% 성장을 제시

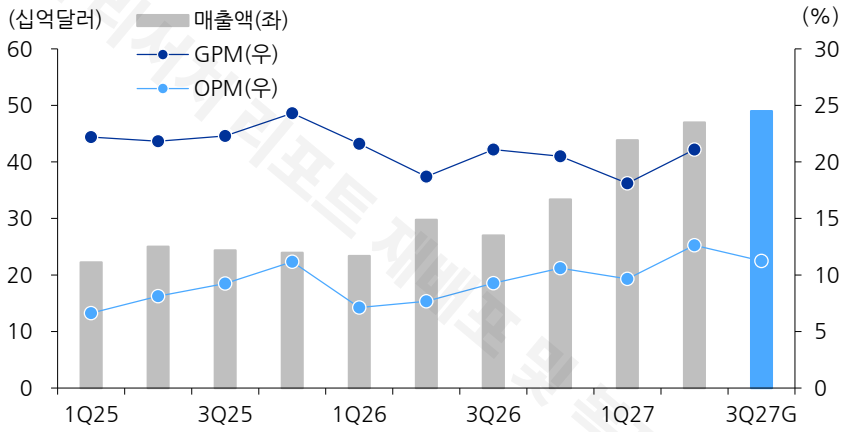
델 매출액, yoy 성장률 추이



자료: Bloomberg, 유진투자증권

메모리 인플레이션에도 양호한 마진 방어, 스토리지 IP 믹스전환으로 ISG 마진 개선 중

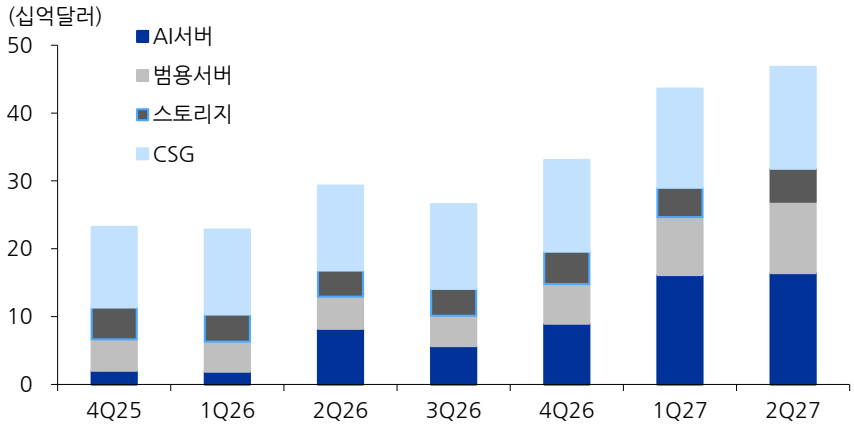
델 매출액, 마진율 추이



자료: Bloomberg, 유진투자증권

AI 서버와 범용 서버 강세가 탑라인 성장을 견인하는 양상

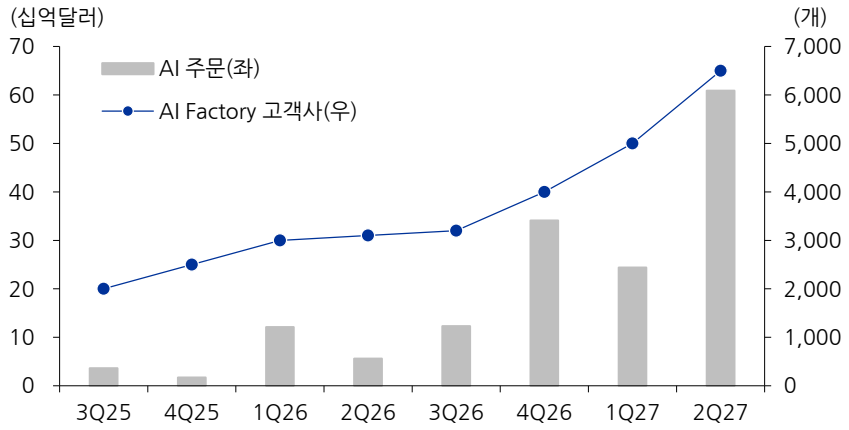
델 세그먼트별 매출액 추이



자료: Bloomberg, 유진투자증권

엔터프라이즈 수요 강세로 AI Factory 고객사는 지속 증가하는 양상

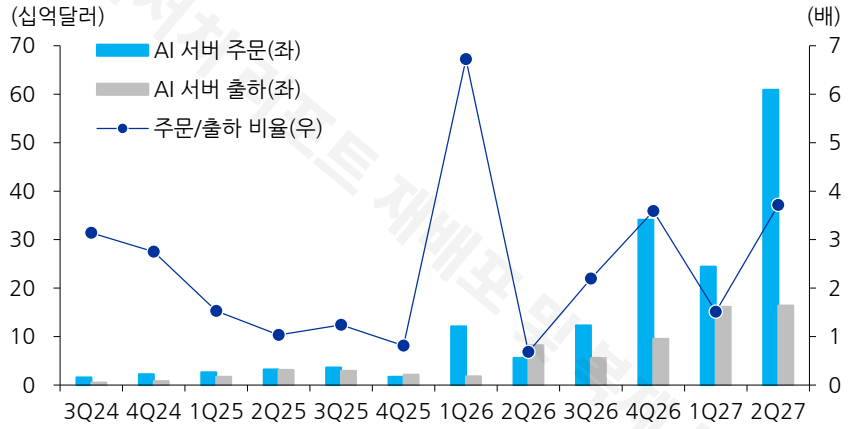
델 AI 주문, AI Factory 고객사 추이



주: AI Factory 고객사는 델 어닝콜을 기반으로 산정했으며, 일부 추정이 포함됨
자료: Dell Technologies, 유진투자증권

AI 서버 주문 급증으로 주문/출하 비율은 3.7 배까지 급증, 주문 이행 속도가 성장의 핵심 요소

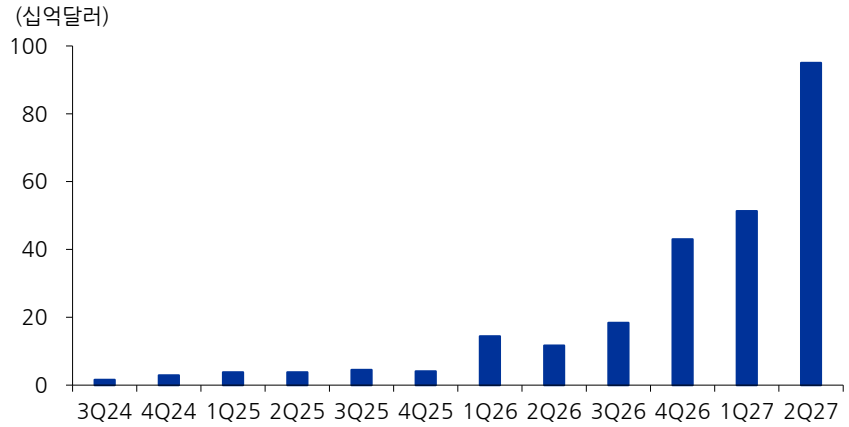
델 AI 서버 주문·출하 추이



자료: Bloomberg, 유진투자증권

분기말 AI 서버 백로그 950 억달러로 역대 최고치를 기록, 전년비 7 배 이상 증가

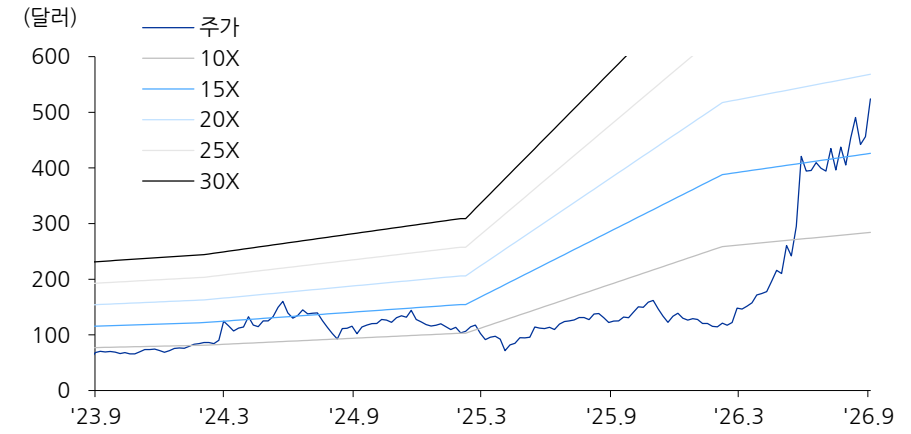
델 AI 서버 백로그 추이



자료: Bloomberg, 유진투자증권

12M Fwd PER 은 18.6 배 수준으로, Peer 대비 높은 프리미엄을 받고 있으나, 여전히 매수 관점 접근 가능

델 12M Fwd PER 밴드



자료: Bloomberg, 유진투자증권

델 실적 추이 및 컨센서스

Non-GAAP(십억달러)	1Q27	2Q27	3Q27E	4Q27E	1Q28E	2Q28E	3Q28E	4Q28E	2025	2026	2027E	2028E
매출액	**43.8**	**47.0**	**49.2**	**53.7**	**54.6**	**57.4**	**60.1**	**63.2**	**95.6**	**113.5**	**193.2**	**233.6**
yoy	*88%*	*58%*	*82%*	*61%*	*25%*	*22%*	*22%*	*18%*	*8%*	*19%*	*70%*	*21%*
AI 서버 매출액	**16.1**	**16.4**	**19.2**	**22.8**	**24.1**	**24.8**	**26.9**	**29.1**	**27.1**	**44.2**	**95.4**	**117.6**
yoy	*757%*	*100%*	*240%*	*155%*	*49%*	*51%*	*40%*	*27%*		*166%*	*202%*	*41%*
매출총이익	**7.9**	**9.9**	**9.4**	**10.3**	**10.0**	**10.6**	**11.1**	**11.6**	**21.8**	**23.2**	**37.5**	**43.3**
GPM	18.1%	21.1%	19.2%	19.1%	18.3%	18.4%	18.4%	18.4%	22.8%	20.4%	19.4%	18.5%
영업이익	**42.4**	**59.3**	**55.5**	**61.5**	**57.0**	**61.6**	**66.3**	**70.8**	**8.5**	**10.0**	**21.8**	**25.3**
OPM	9.7%	12.6%	11.3%	11.5%	10.4%	10.7%	11.0%	11.2%	8.9%	8.8%	11.3%	10.8%
순이익	**3.2**	**4.6**	**4.2**	**4.7**	**4.4**	**4.7**	**5.1**	**5.5**	**5.9**	**7.0**	**16.7**	**19.4**
NPM	7.3%	9.8%	8.6%	8.8%	8.0%	8.3%	8.6%	8.7%	6.1%	6.2%	8.6%	8.3%
EPS (달러)	**4.86**	**7.04**	**6.53**	**7.40**	**6.82**	**7.44**	**8.01**	**8.65**	**8.14**	**10.30**	**25.87**	**30.28**
yoy	*214%*	*203%*	*152%*	*90%*	*40%*	*6%*	*23%*	*17%*	*14%*	*27%*	*151%*	*17%*

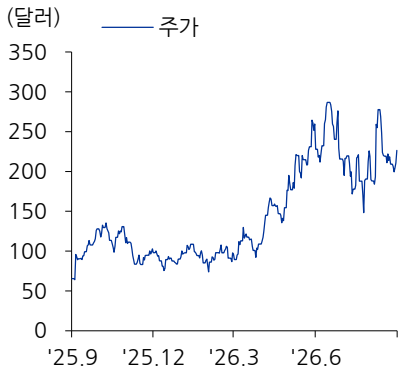
주: 기간은 FY 기준을 적용
자료: Bloomberg, 유진투자증권

(NBIS.US)

Nebius Group

성장과 안정성을 모두 잡다

투자의견	NR
현재주가	226.4 USD (9/4)
시가총액	65 (십억달러) / 88 (조원)



- 네비우스의 2분기 그룹 매출은 5.82억달러(+454%yoy), AI Cloud 매출은 5.75억달러(+514%yoy)를 기록하며 컨센을 상회. ARR은 6월 말 기준 30억달러(+598%yoy), RPO는 400억달러를 달성했으며, EBITDA 마진은 41%로 전분기(32%) 대비 큰 폭으로 개선. 연간 가이던스는 매출 32억달러, ARR 80억달러, EBITDA 마진 40%로 유지했으며, 바인랜드 데이터센터 지연 우려에도 달성 의지를 재확인한 점이 긍정적. 연간 계약용량 가이던스는 4GW에서 5GW로 상향.
- 고객 선수금 비중도 사상 최고치를 기록. 2분기 체결 계약의 약 70%에 선수금 조건이 포함되었으며, 2026년 전체 선수금은 90억달러(연간 캐팩스 가이던스의 40% 수준)로 예상. **비중 높은 선수금 기반의 캐팩스 투자로 상대적으로 안정적인 재무상태를 유지한다는 점이 네비우스가 받는 밸류에이션 프리미엄의 핵심 근거.**
- 2분기에는 10억달러 이상 규모 계약을 네 건 체결. **1~3년 단위 중기계약 ACV는 MW 당 2,000~2,500만달러 수준으로 장기계약 단가(1,200만달러) 대비 70% 가까이 상승. 6개월 이내 단기 계약 ACV는 최대 4,000~5,000만달러까지 상승했으며 해당 단가로 첫 계약을 체결.** 상당수 계약이 연말 가동 예정 용량에 배정되어 2027년부터 본격적인 실적 모멘텀으로 이어질 전망.
- **오픈웨이트 생태계 확장의 직접 수혜도 기대.** 네오클라우드는 자사 모델을 보유한 하이퍼스케일러와 달리 모델 이해관계가 중립적. 또한 엔비디아 인프라 기반 클러스터를 보유해, 파편화된 커스텀 모델 수요를 제약 없이 수용 가능. 아울러 하이퍼스케일러 대비 저렴한 GPU 인프라 비용으로 비용 효율적인 오픈웨이트 모델과 시너지 효과 창출 가능. 네비우스는 Token Factory를 통해 오픈웨이트 모델을 직접 서빙하고 있으며, 이는 수익 구조를 기존 시간당 과금에서 토큰 사용량 기반 과금으로 확장시키는 동시에 감가상각이 종료된 구세대 GPU를 재배치해 ROI를 극대화하는 효과.
- 급증하는 AI 컴퓨팅 수요로 GPU 렌탈가 상승세가 지속되고 있으나, 기존 5년 단위 장기계약 구조에서는 가격 상승분 반영에 구조적 한계가 존재. 다만 최근 네오클라우드 업체들이 계약 형태를 다양화하며 MW 당 ACV 산정 기준 자체가 구조적으로 재편되는 국면. **① 엄격한 선수금 전략으로 재무 건전성을 확보하고 있고, ② 기존 캐파 대비 증분 캐파 비중이 높아 계약 믹스 전환의 레버리지 강도가 크며, ③ 모델 중립성과 Token Factory를 통해 오픈웨이트 확산 수혜를 직접 수취할 수 있는 네비우스를 네오클라우드 업종 내 최선호주로 제시.**

현지명	Nebius Group NV
한글명	네비우스 그룹
시가총액(십억달러/조원)	65/88
설립연도	1989 년
설립자	Arkady Volozh
본사 위치	네덜란드 암스테르담
현 CEO	Arkady Volozh
52 주 최고/최저(달러)	300/64
배당수익률(26E, %)	0.0
주요주주 지분율(%)	
골드만삭스	9.1
블랙록	4.5
인베스코	2.7

주가상승률(%)	1M	6M	YTD
	19.2	153.4	170.5

(Non GAAP)	FY2025	FY2026E	FY2027E
매출액(백만달러)	530	3,275	11,308
영업이익(백만달러)	-596	-886	-410
당기순이익(백만달러)	-447	-686	-848
EPS(달러)	-1.77	-1.90	-2.13
증감률(%)	적지	적지	적지
PER(배)	N/A	N/A	N/A
ROE(%)	0.9	-4.4	-27.8
PBR(배)	6.0	8.3	9.3
EV/EBITDA(배)	88.4	57.5	15.3

네비우스 Token Factory

Token Factory 는 네비우스가 2025 년 11 월 출시한 오픈웨이트 모델 전용 추론 플랫폼이다. 단순히 모델을 호스팅하는 것을 넘어, 추론과 사후학습(Post-Training), 접근 권한 관리를 하나의 플랫폼에 통합한 구조를 취한다. Kimi, DeepSeek, GLM, Qwen, Nemotron 등 40 개 이상의 오픈소스 모델을 지원하며, 고객이 자체 보유한 커스텀 모델을 호스팅하는 것도 가능하다.

핵심 차별점은 커스텀 모델 대응 역량이다. Token Factory 는 다양한 파인튜닝 파이프라인을 제공하며, 증류와 사용자 워크로드에 특화된 학습까지 지원한다. 네비우스는 이를 통해 추론 비용과 지연시간을 최대 70% 절감할 수 있다고 밝히고 있다. 오픈웨이트 생태계의 수요가 소수의 표준 모델이 아니라 사용자별로 변형된 다수의 커스텀 모델로 파편화되어 있음을 감안하면, 파인튜닝부터 배포까지의 전 주기를 단일 플랫폼에서 처리하는 구조는 유의미한 경쟁 요소로 작용할 수 있다.

네비우스 Token Factory

자료: Nebius, 유진투자증권

네비우스 Token Factory 지원 모델 및 토큰 가격

모델명	토큰 가격 (달러)
Kimi K3	15
DeepSeek V4 Pro	3.5
GLM-5.3	0.5
Nemotron-3.5	0.24
MiniMax M3	1.2

주: 토큰 가격은 1M 출력 토큰을 기준으로 산정
자료: Nebius, 유진투자증권

투자 관점에서 Token Factory 의 의미는 네비우스의 비즈니스 모델 확장에 있다. 기존 네오클라우드의 수익 구조는 GPU 임대 단가와 가동률의 곱으로 결정되는 시간당 과금이었으나, Token Factory 는 토큰 사용량 기준으로 과금하는 구조다. **이는 ① 소프트웨어·운영 레이어를 얹어 마진율이 상대적으로 높고, ② 계약이 만료되어 감가상각이 종료된 구세대 GPU 를 토큰 서비스로 재배치해 자산 수익률을 개선하며, ③ 대규모 인프라 계약이 어려운 엔터프라이즈까지 고객 저변을 확장할 수 있다는 점에서 긍정적이다.** 다만 현재까지 Token Factory 의 매출 기여도와 고객 수는 별도로 공시되지 않고 있어, 향후 실적 발표에서의 정량 지표 공개 여부를 점검할 필요가 있다.

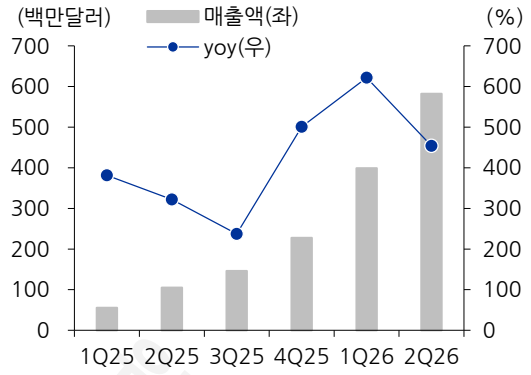
모델 서빙은 단순 GPU 렌탈 대비 부가가치를 극대화할 수 있음

네오클라우드 GPU 렌탈과 모델 서빙 사업 구조 차이

구분	GPU 렌탈	모델 서빙
과금 구조	$/GPU-Hour(사용 시간)	$/1M Tokens(사용량)
매출 산식	GPU 임대 단가 x 가동률	토큰 판매 단가 - 추론 컴퓨트 비용
마진	상대적으로 낮음	높음
주요 고객	모델랩, AI 스타트업 등	엔터프라이즈 등
경쟁력	GPU 확보 능력, 단가 등	추론 최적화 SW, 단가 등

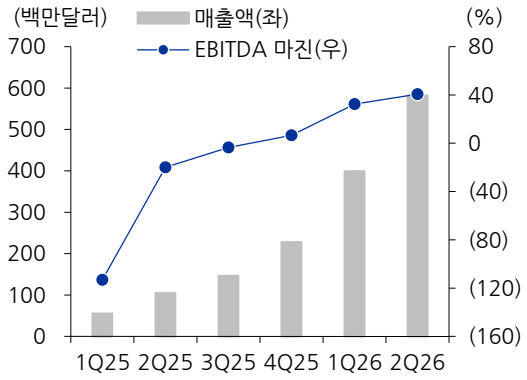
자료: 각 사, 유진투자증권

네비우스 매출액, yoy 추이



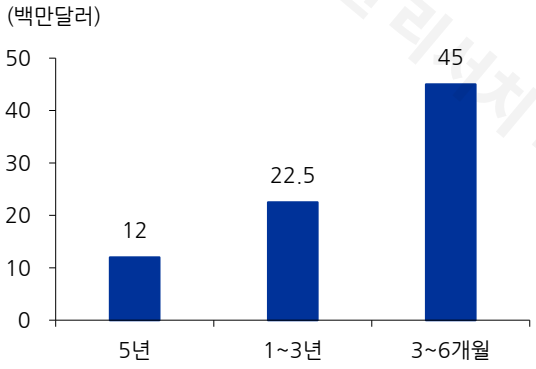
자료: Bloomberg, 유진투자증권

네비우스 매출액, EBITDA 마진 추이



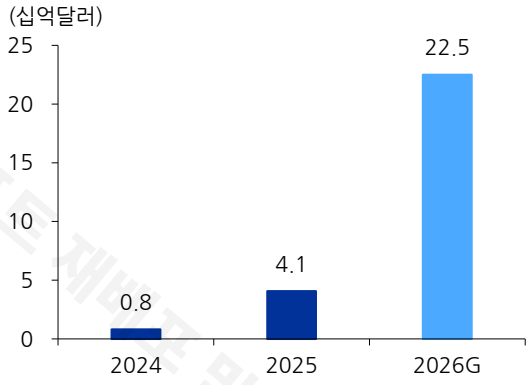
자료: Bloomberg, 유진투자증권

네비우스 기간별 ACV



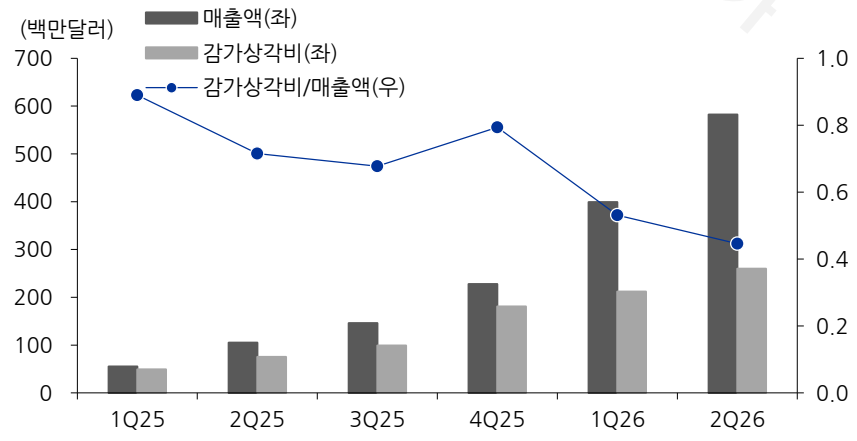
자료: Nebius, 유진투자증권

네비우스 연간 캐팩스 추이



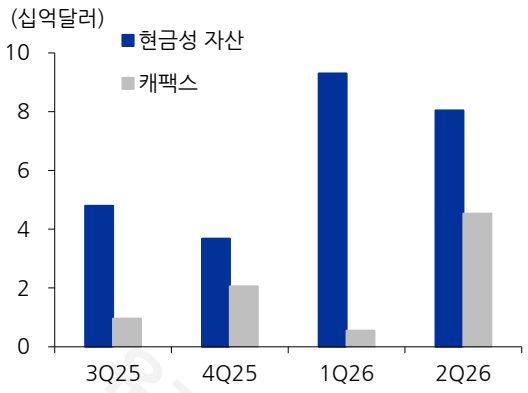
자료: Bloomberg, 유진투자증권

네비우스 매출액, 감가상각비 추이



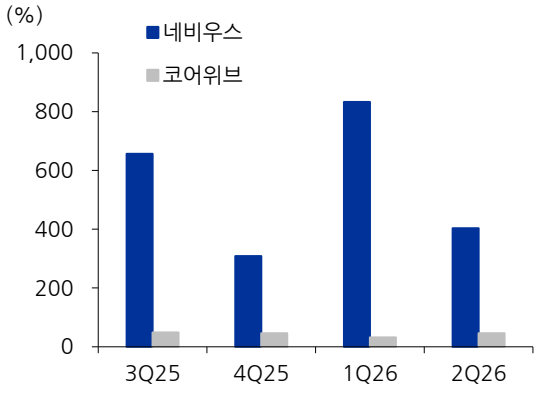
자료: Bloomberg, 유진투자증권

네비우스 현금성자산, 캐팩스 추이



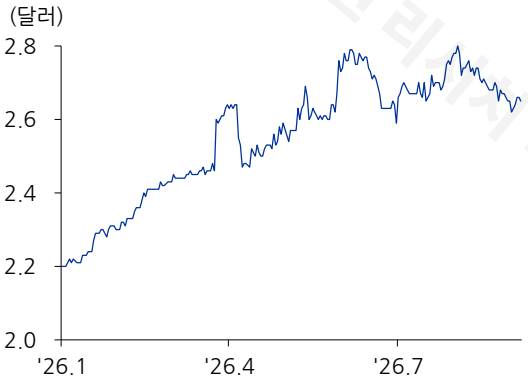
자료: Bloomberg, 유진투자증권

네비우스, 코어위브 유동비율 비교



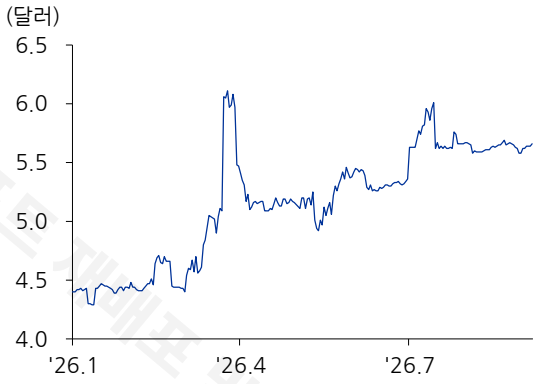
자료: Bloomberg, 유진투자증권

H100 렌탈가 추이



자료: Bloomberg, 유진투자증권

B200 렌탈가 추이



자료: Bloomberg, 유진투자증권

네비우스 실적 추이 및 컨센서스

Non-GAAP(백만달러)	1Q26	2Q26	3Q26E	4Q26E	1Q27E	2Q27E	3Q26E	4Q26E	2024	2025	2026E	2027E
매출액	**399**	**582**	**866**	**1,390**	**2,093**	**2,569**	**3,273**	**4,030**	**118**	**530**	**3,275**	**11,308**
yoy	*622%*	*454%*	*493%*	*511%*	*425%*	*341%*	*278%*	*190%*		*351%*	*518%*	*245%*
EBITDA	**130**	**236**	**339**	**646**	**1,123**	**1,437**	**1,836**	**2,330**	**-266**	**-65**	**1,340**	**6,345**
EBITDA 마진	32%	41%	39%	46%	54%	56%	56%	58%	-227%	-12%	41%	56%
영업이익	**-83**	**-176**	**-286**	**-399**	**-273**	**-249**	**-159**	**44**	**-347**	**-596**	**-887**	**-410**
OPM	-21%	-30%	-33%	-29%	-13%	-10%	-5%	1%	-295%	-113%	-27%	-4%
순이익	**-100**	**-33**	**-286**	**-373**	**-304**	**-335**	**-392**	**-330**	**-247**	**-447**	**-685**	**-848**
NPM	-25%	-6%	-33%	-27%	-15%	-13%	-12%	-8%	-210%	-84%	-21%	-7%
EPS(달러)	**-0.32**	**-0.12**	**-0.90**	**-1.24**	**-0.99**	**-1.17**	**-1.41**	**-1.26**	**-0.83**	**-1.77**	**-1.90**	**-2.13**

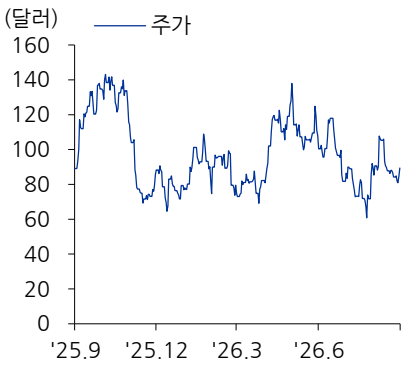
주: 기간은 FY 기준을 적용
자료: Bloomberg, 유진투자증권

(CRWV.US)

Coreweave

빠른 길, 더 무거운 짐

투자의견	NR
현재주가	89.36 USD (9/4)
시가총액	50 (십억달러) / 67 (조원)



- 코어위브의 2분기 매출은 25.8억달러(+112%yoy)를 기록하며 컨센을 상회. EBITDA 마진은 59%, OPM은 5%를 기록하며 전분기 대비 뚜렷한 개선세를 시현. 순손실은 5.67억달러로 전분기 대비 소폭 축소. 급증하는 AI 컴퓨팅 수요로 RPO는 1,042억달러(+246%yoy)로 증가.
- 2분기 캐팩스는 94억달러, 계약용량은 4.2GW를 확보. **활성용량은 1.5GW로 전분기 실적발표 시점 대비 500MW 증가하며 실행력을 재확인, 연말 활성용량 가이던스도 기존 1.7GW에서 1.85GW로 상향 조정.** 3분기 가이던스는 매출 35.3억달러(+158%yoy), OPM 6.5%로 외형성장 가속화와 마진 개선을 동시에 제시했으며, 4분기에는 OPM이 10%대 초반까지 확대될 것으로 언급.
- **GPU 렌탈가 상승 기조 지속. 블랙웰과 루빈 플랫폼 가격은 사상 최고치를 기록했으며, 레거시 A100까지 전년과 유사한 ASP로 2029년까지 계약이 연장.** 루빈 플랫폼 도입으로 신규 계약 마진은 5~10%p 개선되며, 하반기 본격 램프 시 추가 개선을 기대 가능. 레거시 GPU는 감가상각 종료, 선단 GPU는 ASP 상승으로 수익성이 극대화되는 구조.
- **Managed Inference 플랫폼은 출시 수 개월만에 ARR 1억달러를 초과 달성했으며, 연말까지 2.5억달러 가이던스를 제시.** 네비우스의 Token Factory와 유사하게 오픈소스·오픈웨이트 모델에 자사 소프트웨어·운영 레이어를 얹어 토큰 사용량 기반으로 과금하는 구조로, 네오클라우드의 사업 모델을 단순 인프라 임대에서 토큰 서비스로 확장시키는 시도. 오픈웨이트 확산 국면에서 업종 전반의 수익 구조 변화를 선도하는 긍정적 요인.
- 2분기 채권·전환사채·주식 조합으로 약 180억달러를 조달했으며, 이자비용이 매출의 25% 수준을 유지 중. 캐팩스 누적으로 감가상각비 또한 매출의 50%를 상회하고 있어, 계약 만기 이후 구세대 GPU 재배치를 통한 잔존가치 실현이 투자 회수의 전제 조건. **금리 환경이 악화되거나 GPU 렌탈가가 하락 전환할 경우 수익성 훼손 속도가 빠를 수 있는 구조.**
- 코어위브는 ① 우호적인 컴퓨팅 수요로 인한 MW 당 ACV 증가와 ② 오픈웨이트 확산 수혜라는 업종 공통의 포인트를 공유. **다만 이자비용 부담이 경쟁사 대비 높아 매크로 환경에 민감하며, 대형 고객사 의존도가 높아 하이퍼스케일러 및 프론티어 모델랩의 컴퓨트 수요 환경 변동에 따른 실적 변동성이 크다는 개별적 특성은 부담 요인으로 작용. 업종 내에서는 네비우스가 코어위브 대비 매력적이라고 판단.**

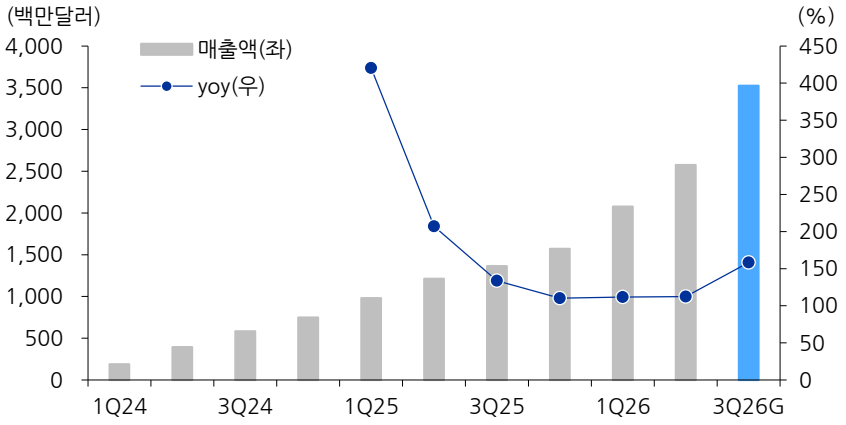
현지명	COREWEAVE INC
한글명	코어위브
시가총액(십억 달러/조원)	50/67
설립연도	2017년
설립자	Michael Intrator
본사 위치	미국 뉴저지
현 CEO	Michael Intrator
52주 최고/최저(달러)	153/61
배당수익률(26E, %)	0.0
주요주주 지분율(%)	
마그네타 파이낸셜	10.4
엔비디아	10.3
골드만삭스	7.8

주가상승률(%)	1M	6M	YTD
	4.7	22.4	24.8

(Non GAAP)	FY2025	FY2026E	FY2027E
매출액(백만달러)	5,131	12,920	26,564
영업이익(백만달러)	666	1,070	4,168
당기순이익(백만달러)	-606	-2,227	-1,136
EPS(달러)	-1.39	-3.90	-1.78
증감률(%)	적지	적지	적지
PER(배)	N/A	N/A	N/A
ROE(%)	-43.6	-55.7	-33.1
PBR(배)	9.8	12.2	11.0
EV/EBITDA(배)	19.0	12.0	6.9

2 분기 매출 25.8 억 달러로 컨센 상회, 3 분기 가이던스 성장 가속 제시

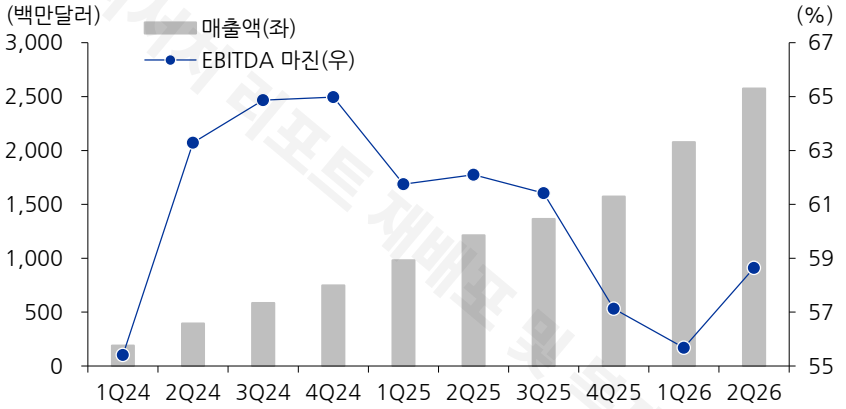
코어위브 매출액, yoy 성장률 추이



자료: Bloomberg, 유진투자증권

EBITDA 마진 59%로 전분기 대비 개선

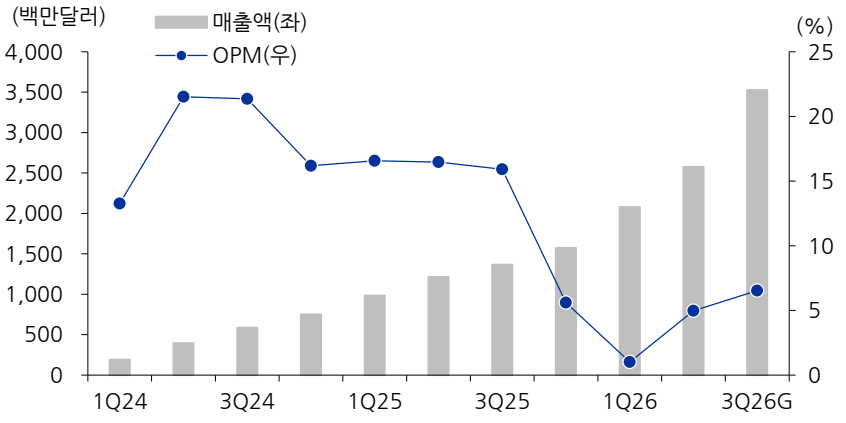
코어위브 매출액, EBITDA 마진 추이



자료: Bloomberg, 유진투자증권

OPM 5%로 전분기 대비 개선, 2 분기 연속 개선세

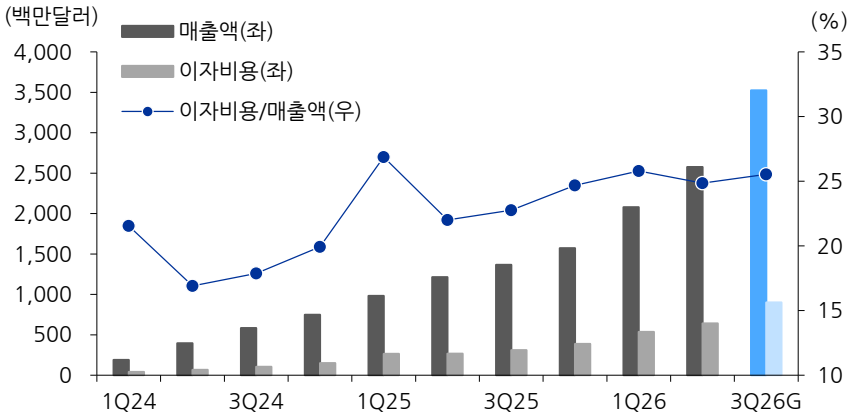
코어위브 매출액, OPM 추이



자료: Bloomberg, 유진투자증권

이자 비용은 여전히 매출의 25% 수준을 유지, 리스크 요인

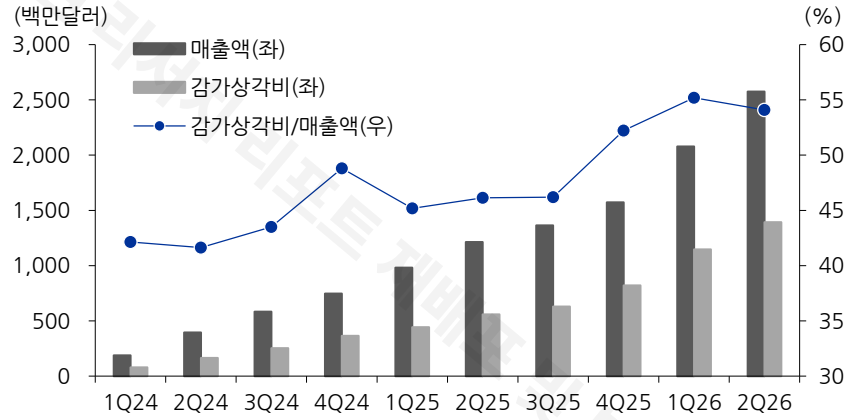
코어위브 매출액, 이자비용 추이



자료: Bloomberg, 유진투자증권

캐팩스 누적으로 인해 감가상각비 또한 매출의 50% 이상으로 누적

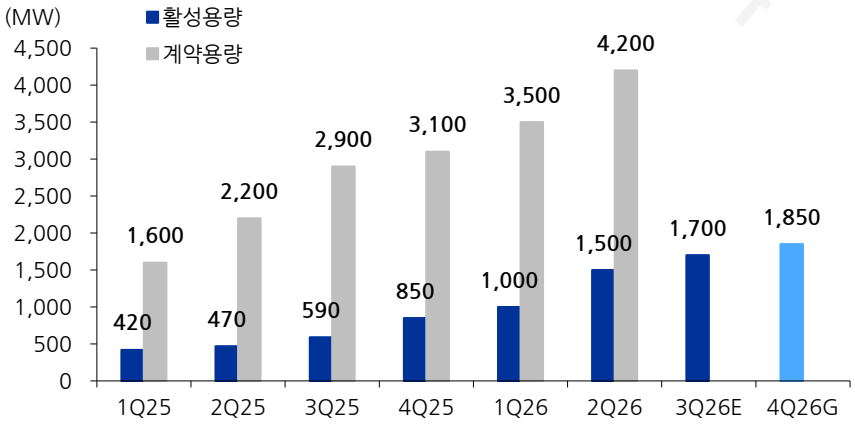
코어위브 매출액, 감가상각비 추이



자료: Bloomberg, 유진투자증권

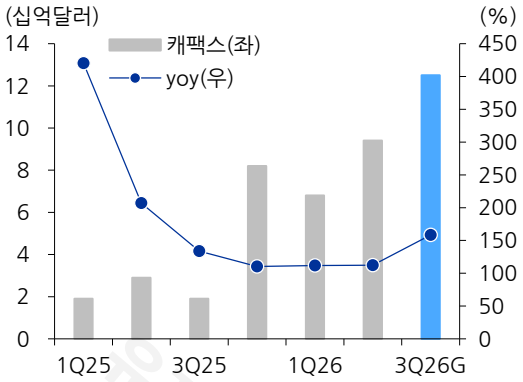
2 분기 활성용량 1.5GW 연말 활성용량 가이던스 1.85GW 로 상향

코어위브 활성/계약용량 추이



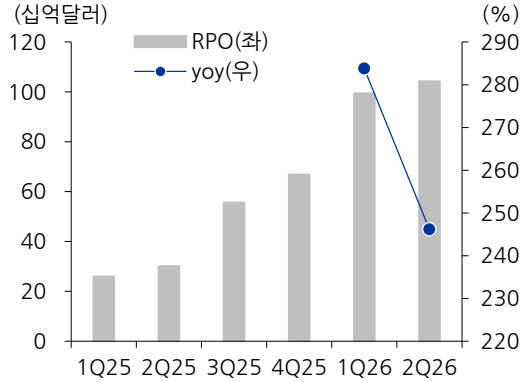
자료: Bloomberg, 유진투자증권

코어위브 캐팩스 추이



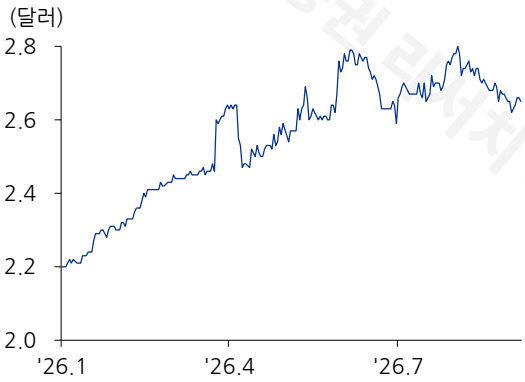
자료: Bloomberg, 유진투자증권

코어위브 RPO 추이



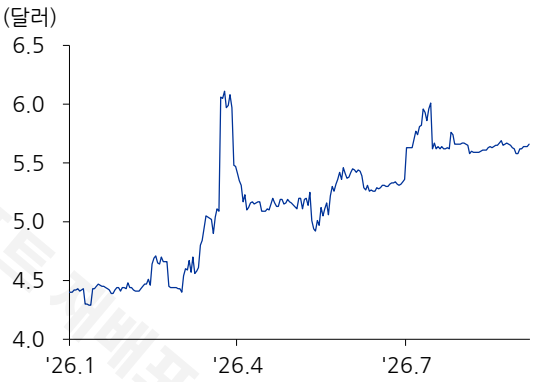
자료: Bloomberg, 유진투자증권

H100 렌탈가 추이



자료: Bloomberg, 유진투자증권

B200 렌탈가 추이



자료: Bloomberg, 유진투자증권

코어위브 실적 추이 및 컨센서스

Non-GAAP(십억달러)	1Q26	2Q26	3Q26E	4Q26E	1Q27E	2Q27E	3Q27E	4Q27E	2024	2025	2026E	2027E
매출액	2.08	2.58	3.55	4.71	5.31	6.10	7.05	8.09	1.92	5.13	12.92	26.56
yoy	*112%*	*112%*	*160%*	*199%*	*155%*	*137%*	*99%*	*72%*		*168%*	*152%*	*106%*
EBITDA	1.16	1.51	2.05	2.87	3.27	3.82	4.42	5.15	1.22	3.09	7.58	16.71
EBITDA 마진	56%	59%	58%	61%	62%	63%	63%	64%	64%	60%	59%	63%
영업이익	0.02	0.13	0.25	0.68	0.74	0.92	1.14	1.40	0.36	0.67	1.07	4.17
OPM	1%	5%	7%	14%	14%	15%	16%	17%	19%	13%	8%	16%
순이익	-0.59	-0.57	-0.67	-0.39	-0.36	-0.31	-0.29	-0.14	-0.06	-0.61	-2.23	-1.14
NPM	-28%	-22%	-19%	-8%	-7%	-5%	-4%	-2%	-3%	-12%	-17%	-4%
EPS (달러)	-1.12	-1.03	-1.19	-0.60	-0.63	-0.53	-0.48	-0.45	-5.96	-1.39	-3.90	-1.77

주: 기간은 FY 기준을 적용
자료: Bloomberg, 유진투자증권

Compliance Notice

당사는 자료 작성일 기준으로 지난 3 개월 간 해당종목에 대해서 유가증권 발행에 참여한 적이 없습니다
당사는 본 자료 발간일을 기준으로 해당종목의 주식을 1% 이상 보유하고 있지 않습니다
당사는 동 자료를 기관투자가 또는 제 3 자에게 사전 제공한 사실이 없습니다
조사분석담당자는 자료작성일 현재 동 종목과 관련하여 재산적 이해관계가 없습니다
동 자료에 게재된 내용들은 조사분석담당자 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 작성되었음을 확인합니다

동 자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료 및 정보로부터 얻어진 것이나, 당사는 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 어떠한 경우에도 자료는 고객의 주식투자의 결과에 대한 법적 책임소재에 대한 증빙자료로 사용될 수 없습니다

투자기간 및 투자등급/투자의견 비율

종목추천 및 업종추천 투자기간: 12 개월 (추천기준일 종가대비 추천종목의 예상 목표수익률을 의미함)		당사 투자의견 비율(%)
· STRONG BUY(매수)	추천기준일 종가대비 +50%이상	3%
· BUY(매수)	추천기준일 종가대비 +15%이상 ~ +50%미만	94%
· HOLD(중립)	추천기준일 종가대비 -10%이상 ☒ +15%미만	3%
· REDUCE(매도)	추천기준일 종가대비 -10%미만	0%

(2026.06.30 기준)