

2026.09.07

US Issue Analysis

DS 미국주식



이규원 해외주식/ETF
02-709-2674
kw.lee@ds-sec.co.kr

강근재 연구원
02-709-2667
Kgg812@ds-sec.co.kr

GPT-6 Astra: OpenAI가 최고 성능을 찍은 진짜 배경

이슈 #1 GPT-6 Astra: 고작 벤치마크 성능 향상이 전부였다면 GPT 5.7였겠지

GPT-6.0 Astra는 (1) Recurrent Depth(순환 깊이)라는 새로운 스케일링 방식을 도입하고 (2) 10만장 이상의 GPU를 활용한 OpenAI 사상 최대 규모의 사전학습 진행했다는 점에서 기존 모델과 차별화된다. Recurrent Depth는 동일 파라미터를 반복 연산해 모델 성능을 높이는 방식으로 AI HW 경쟁의 무게중심을 일부 연산성능으로 되돌릴 가능성이 있다. 다음으로 Astra는 B200, B300이 각각 약 10만장 배치된 Stargate Abilene에서 대규모 사전훈련을 거친 것으로 보인다. Anthropic이 Trainium 기반 대규모 학습을 통해 Fable 계열의 경쟁력을 높였듯이 OpenAI도 GPT-5.5부터 B300을 활용한 이후 Astra에 이르러 다시 모델 성능 경쟁의 선두에 올라섰다. 이는 대규모 학습 인프라가 여전히 모델 성능의 핵심이며, 학습 영역에서 NVIDIA GPU의 경쟁력이 유효함을 보여주는 사례이다. 연말까지 Abilene의 B300이 약 20만장 추가 배치될 것으로 예상되는 만큼 2Q26~1Q27에도 OpenAI의 주도권을 전망한다.

이슈 #2 칩을 강화하는 Alphabet와 OpenAI, 모델을 강화하는 NVIDIA

OpenAI의 첫 자체 추론 ASIC인 Jalapeño는 일부 추론 워크로드에서 NVIDIA B300 GPU를 상회하는 성능과 전력효율을 기록하며 Custom ASIC의 경쟁력이 단순한 비용 절감에 그치지 않음을 보여줬다. Google 역시 TPU 신제품 출시 주기를 기존 약 2년에서 반기 수준으로 단축하며 AI가속기 경쟁에 속도를 내고 있다. 반면 NVIDIA는 여러 AI기업과 Nemotron 연합을 결성해 자사 HW, SW에 최적화된 범용 오픈모델 생태계를 확대하고 있으며, 이를 통해 AI 모델, GPU를 AI 컴퓨팅의 표준 플랫폼으로 유지하려는 전략으로 판단한다.

US Market View: 12MF EPS 성장률 YTD 27.8%. 진짜 금리가 중요할까?

S&P500 12MF EPS 성장률은 연초 대비 27.8% 상향됐으며 견조한 이익 성장을 바탕으로 지수는 최고점 부근에 와있다. 올해 미국 10년물 금리가 시장의 심리적 임계선인 4.3%, 4.5%를 돌파할 때 주가는 부진했지만, 2~3주만에 고점을 다시 경신했다. 이번에도 크게 다르지 않을 것으로 본다. 성장이 동반되는 금리 상승은 주식시장에 반드시 악재로 작용하지 않는다. 미국 AI기업의 12MF EPS 전망치가 연초 대비 61% 상향되는 현재의 이익 사이클에서는 금리 상승 자체보다 이를 상쇄할 수 있는 기업이익의 성장 속도에 주목해야 하며, 금리에 대한 과도한 우려는 지양할 필요가 있다.



고작 벤치마크 성능 향상이 전부였다면 GPT 5.7였겠지

새로운 스케일링 법칙을 제시: Recurrent Depth(순환 깊이)

주요 벤치마크 상단을 뚫은
GPT-6.0 Astra

주목할 변화

- (1) 새로운 스케일링 법칙:
Recurrent Depth
- (2) 최대규모의 사전학습
- (3) 영역을 확장한 Agent

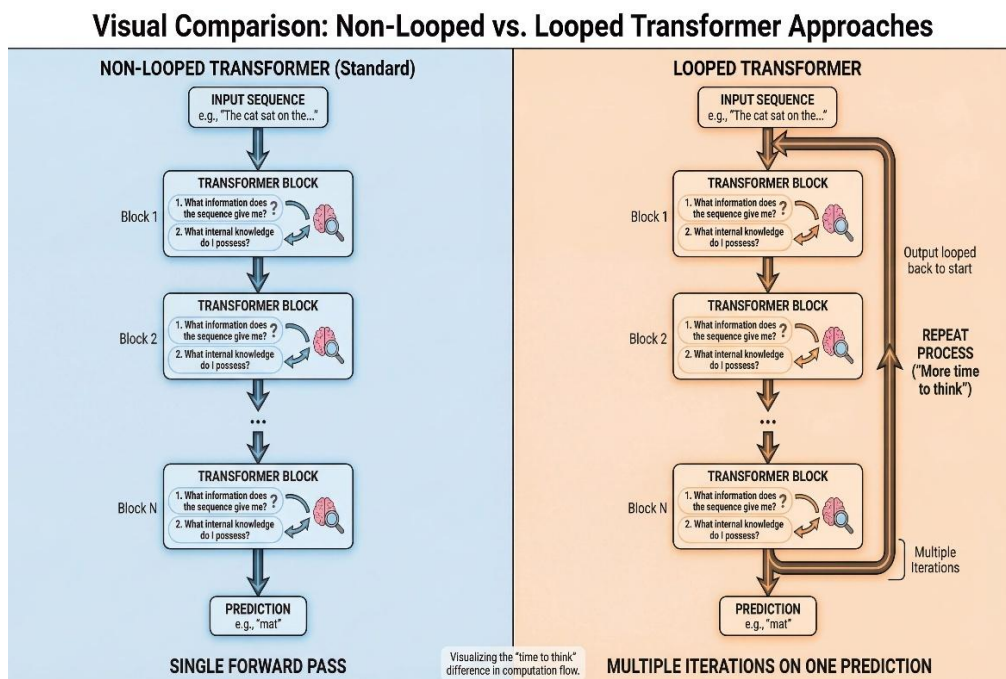
Recurrent Depth

:파라미터 규모를 키우지 않
고 동일한 Transformer 블
록을 반복 사용해 성능 향상

GPT-6 Astra는 최근 출시된 Anthropic의 Fable 5.1을 Terminal-Bench 4.0(57.9% vs 55.8%), DeepSWE v1.1(74.1% vs 67.4%) 등 주요 코딩 벤치마크에서 앞서며 OpenAI를 다시 모델 경쟁의 선두에 위치시켰다. 그럼에도 성능 향상이 전부였다면 GPT-5.6 Sol에 이은 5.7이었을 것이다. OpenAI가 Astra를 GPT-6.0이라는 새로운 세대(Generation)의 시작점으로 분류한 배경에는 (1) Recurrent Depth라는 새로운 스케일링 방식의 도입 (2) 10만개 이상의 GPU를 투입한 OpenAI 사상 최대 규모의 사전학습 (3) Coding을 넘어 컴퓨터 활용, 리서치, 전문 업무까지 확장된 Agent의 활용 영역 등이 있었을 것으로 판단하며 본 리포트에서는 각각의 변화를 다뤄보았다.

GPT-6 Astra는 ‘파라미터 수’와 ‘Reasoning(내부 추론)’이라는 기존의 스케일링 법칙에 더해 ‘Recurrent Depth(순환 깊이)’를 새로운 스케일링 축으로 제시했다. 기존의 모델은 (1) Transformer 레이어(또는 블록) 수를 늘려 파라미터 크기를 확대하거나 (2) 답을 내기 전 더 많은 추론 토큰을 생성(Reasoning)해 연산량을 늘리는 방식으로 지능을 높여왔다면, Recurrent Depth는 동일한 레이어 블록을 여러 차례 반복 사용해 하나의 토큰을 생성하기 전 수행하는 내부 연산의 깊이를 확대하는 방식이다. 즉 파라미터의 크기를 확대하지 않고도 토큰당 투입되는 연산량을 늘려 토큰당 지능을 높이는 부분이 기존 방식과 가장 큰 차이이다.

그림1 Recurrent Depth: Transformer Block를 늘리는 대신(=파라미터 크기를 확대하는 대신), Transformer Block을 반복 사용하는 방식.



자료: Medium, DS투자증권 리서치센터

보다 직관적으로 비교해보면 기존 Reasoning은 ‘입력→Layer 1→Layer 2→...→Layer 20→내부 추론→답변 출력’ 처럼 여러 Transformer 레이어를 순차적으로 통과한 뒤, 내부 추론을 더해 생각의 깊이를 늘리는 방식이다. 반면 Recurrent Depth는 ‘입력→Reasoning Block(Layer 1→Layer 20)→Reasoning Block(Layer 1→Layer 20)→Reasoning Block(Layer 1→Layer 20)→출력’처럼 동일한 Reasoning 블록을 반복 사용해 하나의 토큰을 생성하기 전 내부 연산의 깊이를 확대하는 방식이다. 즉 기존 Reasoning이 생각의 깊이를 늘리는 방식이라면, Recurrent Depth는 생각의 깊이를 늘리는 방식에 가깝다.

Recurrent Depth는 다양한 연구에서 새로운 스케일링 방식으로 주목받고 있다. VLD(Virtual Logic Depth) 연구는 가중치를 반복 사용해 파라미터 수를 늘리지 않고도 유효 연산 깊이(Effective Algorithmic Depth)를 확대할 수 있음을 보였으며, 이를 추론 토큰을 늘리는 기존 Token-wise Scaling과 구분한다. Looped Transformer 계열 연구 역시 동일한 블록의 가중치를 반복적으로 재사용함으로써 파라미터 수를 늘리지 않고도 실제 연산 깊이를 확대할 수 있음을 보여준다. Recurrent Depth는 파라미터 확대를 통해 모델을 키우는 스케일링에서 벗어나, 동일 파라미터에 더 많은 연산을 투입하는 새로운 스케일링 축을 제시한 것이다.

파라미터가 저장되는
HBM 용량 수요에는
부정적인 영향

다만 파라미터 크기는
성능에 매우 중요한만큼
실제 수요 하향을
예상하기 어려움

이러한 변화는 AI HW에도 중요한 시사점을 갖는다. 파라미터 크기를 키우는 기존 스케일링에서는 파라미터가 커질수록 저장해야 하는 가중치가 늘어나기에 GPU 성능과 함께 HBM 용량 확대가 중요했다. 반면 Recurrent Depth 방식은 동일한 가중치를 반복 사용하기 때문에 HBM 용량 수요는 상대적으로 낮추면서 GPU 연산량은 늘리는 구조이다. 이에 따라 GPU의 연산성능과 전력당 연산효율의 중요성은 높아질 수 있다. 다만 동일한 가중치를 반복적으로 불러와 연산하는 과정에서는 높은 메모리 대역폭이 요구되고, 반복 단계별로 KV Cache를 별도로 저장해야 할 가능성도 높아 HBM 병목이 사라질 가능성은 낮다. 더 나아가 Recurrent Depth는 HBM 용량을 감소시키기보다 동일한 HBM 하에서 성능을 끌어내는 방향으로 발전해갈 전망이다. 이는 모델 성능 경쟁이 치열한 상황에서 파라미터 규모와 직결된 HBM은 가능한 많이 탑재해 파라미터 규모를 키우고 그 위에 Recurrent Depth를 통해 연산 깊이를 추가로 확대하는 방향을 의미한다.

표1 Recurrent Depth 스케일링 확산 시 주요 AI HW의 수혜 논리

AI HW	영향	논리
GPU / AI가속기	수혜	- 같은 가중치를 여러 번 실행하므로 출력 토큰당 FLOPs가 증가. Parameter 수보다 Compute/Token이 중요한 스케일링으로 이동
HBM 용량	상대적 중요도 감소	- 가중치를 확대시키는 방식이 아니기 때문에 가중치가 저장되는 HBM 용량의 상대적 중요도 감소. 단 가중치는 AI모델의 성능의 핵심이라는 점에서 절대적 중요도는 여전히 높을 것
HBM 대역폭	수혜	- 반복마다 가중치와 KV Cache를 다시 읽어와 할 수 있어 데이터 이동량이 증가. 데이터 이동 통로인 HBM 대역폭 중요성 ↑
Scale-up Network	수혜	- 모델이 다수의 GPU에 분산된 경우 한 번의 Transformer Block을 반복 실행할 때마다 GPU 간 통신 발생
DRAM/NAND	소폭 수혜	- 반복 과정에서 KV Cache 저장량이 증가하기에 메모리 용량 수요 증가
CPU	중립	- Recurrent Depth 자체는 GPU 연산량을 증가시키기에 CPU의 상대적 수혜는 제한적. 다만 AI 성능 향상으로 Agentic AI가 확산될수록 CPU 수요에는 긍정적

자료: DS투자증권 리서치센터

다시 주도권을 잡은 OpenAI 그리고 NVIDIA

OpenAI 사상 최대 규모의 학습(GPU 10만개 이상)

B300 활용이 예상되기에 HW가 모델 성능에 절대적임을 다시 입증

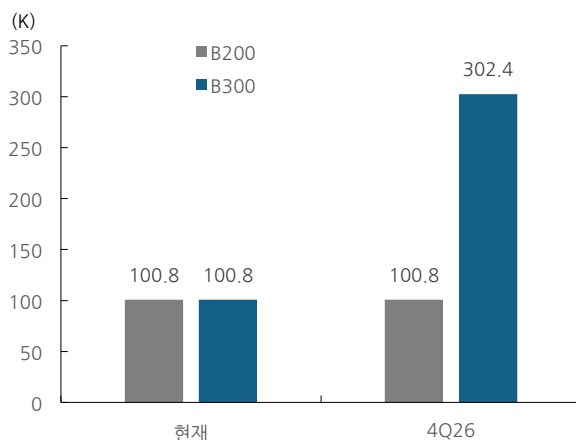
모델 훈련에서 NVIDIA GPU의 경쟁력을 재확인

Anthropic Fable 계열은 AWS Trainium 기반 훈련으로 알려짐

Astra는 OpenAI의 프론티어 모델 학습이 본격적으로 '10만+ GPU 시대'에 진입했음을 보여주는 모델이라는 점에서 의미가 있다. Axios에 따르면 Astra는 텍사스 Stargate에서 10만개 이상의 GPU를 활용한 OpenAI 사상 최대 규모의 학습을 거쳤으며, 해당 시설은 Stargate Abilene으로 추정된다. AI인프라 조사업체 Epoch AI는 현재 Abilene에 B200, B300이 각각 약 10.1만개씩, 총 20.2만개 배치돼 있는 것으로 추정하기 때문에 10만개 이상의 활용되었을 가능성을 배제할 수 없으며, 이 규모는 H100 기준 약 50.9만개 수준의 연산능력으로 평가된다. 이처럼 대규모 사전학습 컴퓨팅을 투입한 결과 Astra는 기존 세대 대비 복잡한 추론과 Agentic 업무 수행 전반에서 성능이 크게 개선되며 프론티어 모델 경쟁력을 다시 끌어올린 것으로 판단한다.

GPT 5.6 Sol 전후로 부활한 OpenAI 모델의 경쟁력은 NVIDIA GPU의 경쟁력을 확인하는 사례이기도 하다. 프론티어 모델의 성능은 알고리즘, 데이터 뿐 아니라 사전학습 클러스터의 규모와 이를 구성하는 AI HW에도 상당한 영향을 받는다. 이를 단적으로 보여주는 사례가 Anthropic으로, Fable(Mythos) 계열을 통해 빠르게 모델 성능을 높인 배경에는 AWS의 Project Rainier가 상당한 역할을 했을 것으로 판단한다. Rainier는 2025년 11월 약 50만개의 Trainium2로 가동을 시작했고, 올해 4월부터 100만개 이상의 Trainium2를 Claude 학습 및 서비스에 활용되고 있는 것으로 알려져 있다. 같은 기간 OpenAI, Grok 등 NVIDIA 칩 진영의 모델 주도권은 약화되는 구도였는데, GB300이 학습에 활용되기 시작한 GPT-5.5부터 OpenAI의 모델 경쟁력이 다시 강화되고 있다. 향후 더 많은 GB300이 대규모 학습에 투입되면서 NVIDIA GPU 기반 모델이 다시 성능 경쟁의 전면에서 설 가능성에 주목한다.

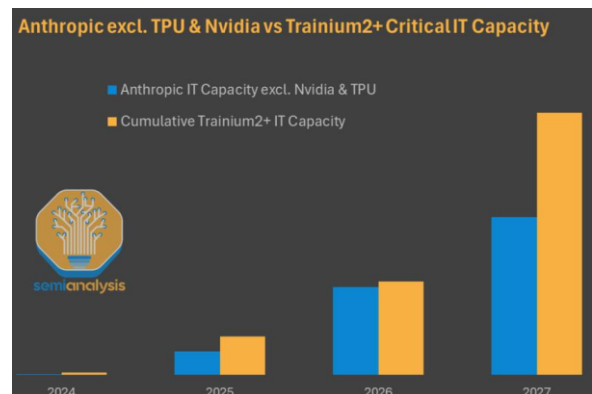
그림2 Stargate Abilene에 배치된 GPU 현황 및 전망



자료: EpochAI, DS투자증권 리서치센터

그림3 대규모 학습 인프라와 모델 경쟁력이 관련성

: '25년 본격 공급된 Trainium이 Anthropic 모델 성능 개선에 기여



자료: SemiAnalysis, DS투자증권 리서치센터

주: 파란 막대는 NVIDIA GPU, TPU를 제외한 Anthropic의 AI가속기 수요

주: 주황 막대는 Anthropic이 누적 배치한 AWS Trainium 용량

Stargate Abilene은
연말 GPU 40만개 규모로
확대되고 대부분 B300

하반기~내년 상반기에
모델 성능 향상 두드러질 것

약한 의미의 재귀적 자기개
선은 이미 시작됨

이러한 관점에서 Stargate Abilene은 Astra보다 오히려 그 이후 모델에서 더 큰 의미를 가질 가능성이 있다. AI시장조사업체 Epoch AI는 Abilene의 GPU가 현재 약 20만개에서 연말 약 40만개까지 2배로 확대되고 증가분 대부분이 B300일 것으로 추정한다. OpenAI가 GPT-5.5부터 GB300을 학습에 활용한 후 GPT-5.6 Sol과 6.0 Astra에서 향상된 경쟁력을 시장에서 인정받았듯이 향후 더 많은 B300을 투입한 사전학습이 본격화되면서 Astra 이후 모델에서도 성능 주도권은 OpenAI에 있을 가능성이 높다. 실제로 OpenAI의 샘 알트먼은 최근 사이버 보안 우려로 개발이 일시 중단된 모델이 'Astra'가 아님을 분명히 밝혔는데, Astra 외에도 높은 Agent 능력을 가진 모델이 내부에서 개발되고 있음을 암시한다.

대규모 학습 인프라는 AI가 다음 세대 AI의 개발을 가속하는 구조와 결합될 때 더욱 강한 의미를 갖는다. Axios에 따르면 Astra는 다른 AI 모델이 학습 감독에 중요한 역할을 맡은 OpenAI의 첫 모델로, AI가 단순한 연구 보조를 넘어 다음 모델 개발 과정 자체에 개입하기 시작했음을 보여준다. 더 좋은 AI가 코드 작성, 실험 설계, 평가와 모니터링을 자동화하면 연구자 한 명이 동시에 수행할 수 있는 실험의 수가 증가하고, 이는 다시 더 많은 학습과 추론 연산으로 이어질 수 있다. 강한 의미의 재귀적 자기개선(RSI)과는 다르지만 더 좋은 AI → 연구 자동화 → 실험량 증가 → 다음 모델 개선이라는 약한 자기개선 구조는 이미 형성되기 시작한 것으로 판단한다. 결국 Stargate의 확대는 단순히 더 큰 모델을 학습하기 위한 인프라를 넘어 AI가 AI 개발 속도 자체를 높이는 피드백 루프를 가속하는 기반이 될 가능성이 있다.

Coding을 넘어서 화이트칼라 전체를 겨냥

Coding 외 사무 분야에서
경쟁력이 입증됨

“A step change in
professional work” 강조

GPT-6 Astra는 Agentic AI의 적용 범위를 Coding에서 화이트칼라 업무 전반으로 확장했다는 점에서 의미가 있다. OpenAI는 Astra를 단순한 기능 개선이 아니라 ‘전문 업무의 단계적 변화(A step change in professional work)’로 표현하며, 컴퓨터 사용, 리서치, 과학 및 전문 업무까지 Agent의 활용 영역이 확대됐음을 강조했다. 벤치마크 성능도 이를 뒷받침 하는데 전문 업무 자동화 능력을 측정하는 AutomationBench에서 Astra는 41.4%로 GPT-5.6 Sol의 18.1%, Claude Fable 5.1의 31.4%를 크게 상회했으며, 전문 소프트웨어를 활용한 복잡한 업무 수행 능력을 평가하는 Agents’ Last Exam에서도 59.3%로 GPT-5.6 Sol의 53.6%, Claude Opus 5의 55.5%를 상회했다. 이는 Astra의 성능 개선이 Coding을 넘어 여러 소프트웨어를 직접 사용해 하나의 업무를 끝까지 수행하는 Agent 능력의 확장에서 나타나고 있음을 보여준다.

AI 경제성 평가단위가
‘토큰’에서 ‘완료된 업무’로
이동

특히 주목할 부분은 AI의 경제성을 평가하는 단위가 토큰에서 ‘완료된 업무(Task)’로 이동하고 있다는 점이다. OpenAI는 Astra의 토큰 가격이 높지만 다수의 벤치마크에서 더 적은 토큰을 출력하고 업무 수행시간을 단축하면서도 업무를 완수하여 업무당 추정 API 비용(Cost per Task)은 오히려 낮아졌다고 강조한다. 실제 Agents’ Last Exam에서는 Claude Opus 5보다 약 65% 적은 출력 토큰을 사용했고, OSWorld 2.0에서는 GPT-5.6 Sol보다 높은 성능을 내면서 업무 수행시간을 약 47% 단축했다.

AI의 가치가 토큰 생성량보다 완료된 업무를 기준으로 평가되기 시작하면 소프트웨어의 과금 방식도 사용자 수(\$/Seat)에서 업무량이나 성과(\$/Completed Task) 중심으로 이동할 가능성이 높다. 이는 소프트웨어 기업의 가격 결정력을 제약하는 요인이다. 과거에는 고객이 A 기능만 필요하더라도 A+B+C를 하나의 옵션으로 묶여 판매하면서 과금 금액을 높일 수 있었지만, Agent가 필요한 기능만 선택적으로 활용하고 업무 성과를 기준으로 비용을 평가하면 이러한 번들링의 효과는 약화될 수 있다. 여기에 GB300 기반 대규모 사전학습을 통해 AI모델의 지능이 향상되고 활용 영역 확대될 개연성이 높은 상황에서 개별 기능에 특화된 소프트웨어(Adobe 등)의 상대적 경쟁력도 낮아질 전망이다. 반면 기업 데이터와 접근 권한을 보유하고 여러 시스템의 워크플로우를 통합, 통제해 Agent의 업무 수행을 지원하는 소프트웨어의 가치는 오히려 높아지면서 업종 내 양극화가 더욱 뚜렷해질 가능성이 높다.

소프트웨어 업종의
옥석가리기 중요

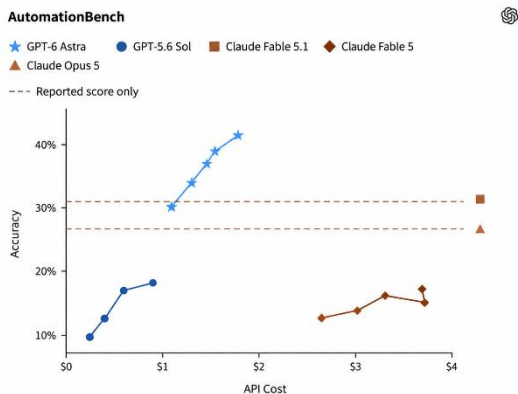
사이버보안은 Agent 확산
의 대표적인 수혜 분야

최근 소프트웨어 업종 전반의 강세 속에서 옥석가리기가 매우 중요하다는 시각 속에 사이버보안은 Agent 확산의 명확한 수혜 영역으로 전망된다. Astra는 OpenAI의 내부 안전 기준에서 처음으로 ‘Critical’ 사이버 역량에 도달한 모델로 알려지지 않은 취약점 탐색부터 공격 방법 개발까지 상당 부분 자동화할 수 있다. 이는 공격자 한 명이 수행할 수 있는 공격의 빈도, 속도, 병렬성을 크게 높여 기업의 보안 대응 역시 AI 기반 자동 탐지, 조사, 차단으로 전환될 필요성을 높인다. 동시에 기업 내부에서는 수많은 Agent가 이메일, CRM, 클라우드 접근 권한

을 가진 비인간 계정(Non-human Identity)으로 추가되면서 새로운 공격 표면도 크게 늘어나게 되면 Agent의 신원, 접근 권한을 관리하고 여러 보안 데이터를 통합해 자동 대응까지 수행하는 보안 플랫폼의 가치가 높아질 전망이다.

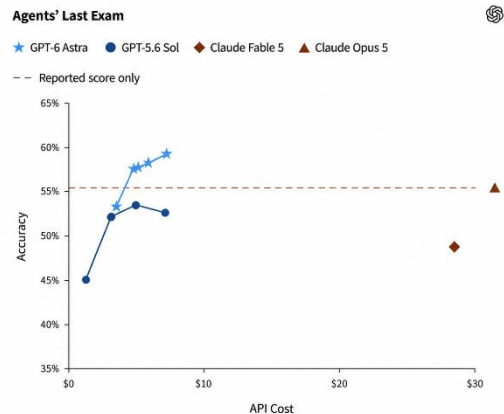
결국 Astra는 AI가 기존 소프트웨어 내 하나의 기능에서 여러 소프트웨어를 직접 조작하는 상위 업무 수행 계층으로 이동하고 있음을 보여주는 모델이다. 이 과정에서 코드 편집, 문서 작성처럼 생성 기능 자체를 판매하던 소프트웨어의 부가가치는 낮아지는 반면, 기업의 원천 데이터와 업무 기록을 보유하거나 권한·보안·업무 흐름을 통제하는 소프트웨어의 중요성은 오히려 높아질 수 있다. 특히 Astra의 성능 향상이 단순 답변 정확도보다 컴퓨터 사용과 전문 업무의 자율성에서 두드러졌다는 점을 고려하면, 향후 소프트웨어 산업의 핵심 경쟁은 사람에게 더 좋은 도구를 제공하는 것에서 AI가 얼마나 많은 인간 업무를 직접 완결할 수 있도록 지원하는가로 이동할 가능성이 높다.

그림4 AutomationBench: 전문업무 자동화 능력을 측정



자료: OpenAI, DS투자증권 리서치센터

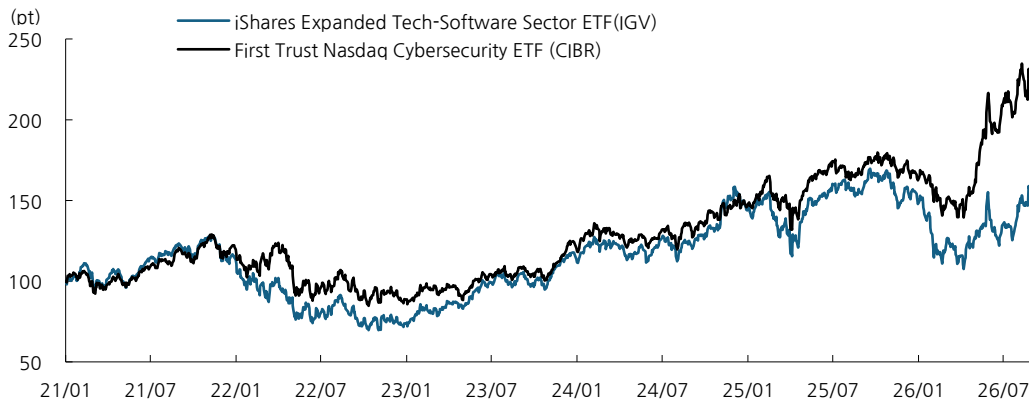
그림5 Agents' Last Exam: 전문 SW를 활용한 업무 수행 능력 평가



자료: Broadcom, DS투자증권 리서치센터

주: 2026년은 초기 출하로 0.1GW 미만 추정, 2028년은 후속 세대 합산 5GW 이상

그림6 소프트웨어 ETF(IGV) 반등, 사이버보안 ETF(CIBR) 고점 경신. AI 시대에 생사가 갈릴 SW 섹터는 특히 옥석가리기가 중요



자료: Bloomberg, DS투자증권 리서치센터

이슈 2: 칩을 강화하는 Alphabet와 OpenAI, 모델을 강화하는 NVIDIA

Custom ASIC의 가능성을 증명한 OpenAI의 'Jalapeño'

B300 이상의 성능을 보인
OpenAI의 'Jalapeño'

OpenAI의 첫 자체 추론 ASIC인 Jalapeño는 Custom ASIC의 경쟁력이 단순한 비용 절감에 그치지 않음을 보여줬다. OpenAI가 공개한 InferenceX 결과에서 Jalapeño는 GPT-OSS 120B, DeepSeek R1 670B, Kimi K2.5 1T 등 외부 모델에서도 초당 와트당 토큰 처리량의 파레토 프론티어를 갱신했으며, DeepSeek R1과 Kimi K2.5에서는 GB300 대비 Peak Throughput/kW가 각각 약 1.7배, 1.5배 높았다. 특히 Jalapeño에 맞춰 개발되지 않은 외부 모델에서도 경쟁력을 보였고 OpenAI 내부 모델에서는 격차가 더 확대됐다는 점을 감안하면, 특정 모델 하나가 아니라 LLM 추론이라는 워크로드 자체에 특화된 ASIC의 가능성을 입증한 사례로 충분히 평가 가능하다.

모델 서빙 환경을 기준으로
칩 전반을 설계 가능한 것이
장점

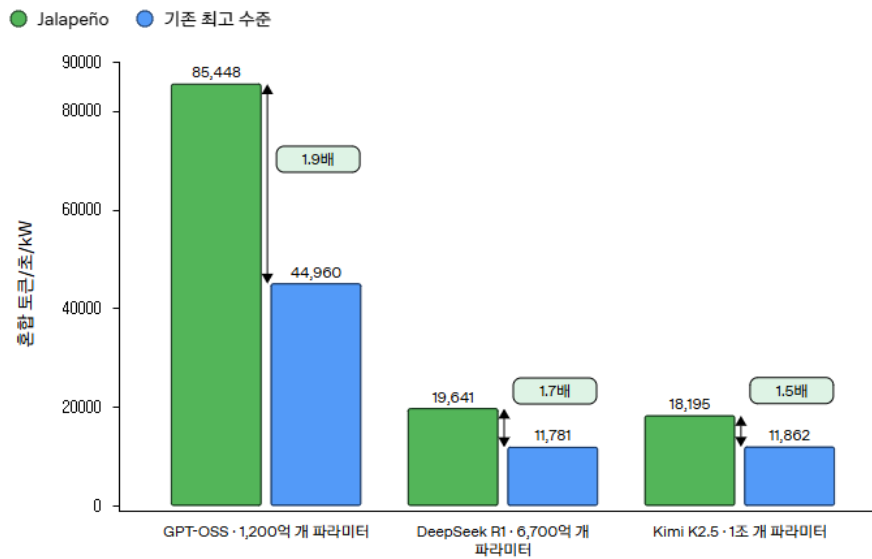
Jalapeño의 높은 효율은 모델사가 자신의 워크로드를 가장 잘 이해한다는 Custom ASIC의 구조적 장점에서 나온다. GPU가 다양한 모델과 연산을 처리하기 위한 범용성을 확보해야 하는 것과 달리 OpenAI는 ChatGPT, Codex, Agent의 실제 모델 서빙 환경을 기준으로 연산부, 메모리, KV Cache, 네트워크와 소프트웨어를 처음부터 함께 설계할 수 있다. 특히 Agentic AI에서는 검색과 도구 호출, 추론이 반복되면서 Prefill과 Decode 비중이 계속 달라지고 지연시간이 업무 완료시간에 직접 영향을 미친다. 결국 가속기 경쟁의 기준도 PFLOPs(연산 성능 지표)보다 전력당 토큰 생성량과 지연 시간, 나아가 하나의 업무를 완료하는데 필요한 비용(Cost per Completed Task)으로 이동할 가능성이 높다.

NVIDIA 대비 1~2세대 뒤쳐
졌기에 경쟁은 어려움

다만 첫 제품임에도
생각보다 괜찮은 성능
브로드컴의 경쟁력 재확인

다만 Jalapeño를 당장 NVIDIA의 대체재로 해석하기는 어렵다. Jalapeño는 NVIDIA Vera Rubin보다 성능에서 열위를 보이며 공동 개발사인 Broadcom이 언급한 본격 배치 시점인 2027년~2028년에는 더 성능이 우월한 Vera Rubin Ultra가 출시될 예정이다. NVIDIA 제품 대비 약 1~2세대 뒤쳐진 것으로 해석되며, OpenAI 역시 학습과 범용 추론에서는 NVIDIA GPU를 대규모로 사용할 전망이다. 오히려 중요한 것은 Jalapeño가 1세대부터 경쟁력을 확보한 데 이어 2, 3세대까지 동시에 개발되고 있다는 점인데 훈련에서는 NVIDIA GPU, 대규모 반복 추론에서는 자체 ASIC이라는 역할 분화가 공고해질 경우 가장 압박받는 영역은 AMD나 외부판매를 시작한 TPU와 같은 범용 AI가속기일 가능성이 높다. 반면 기술력을 입증한 Broadcom은 추후에 AI모델사 등이 자체 ASIC 개발 시 가장 먼저 찾을 가능성이 높다는 점에서 긍정적이며 ASIC 생태계가 확대될수록 첨단 파운드리, HBM·패키징 공급망 시장에서 TSMC의 대체 기업을 찾으려는 시도가 확대될 것이다.

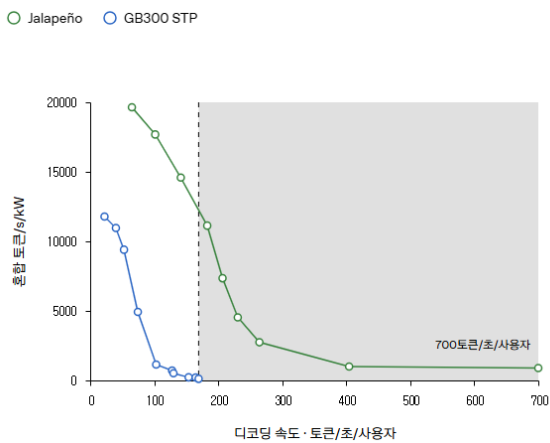
그림7 OpenAI모델은 물론 타 오픈소스 모델에서도 B300보다 우수한 성능을 보여준 Jalapeño



자료: OpenAI, DS투자증권 리서치센터

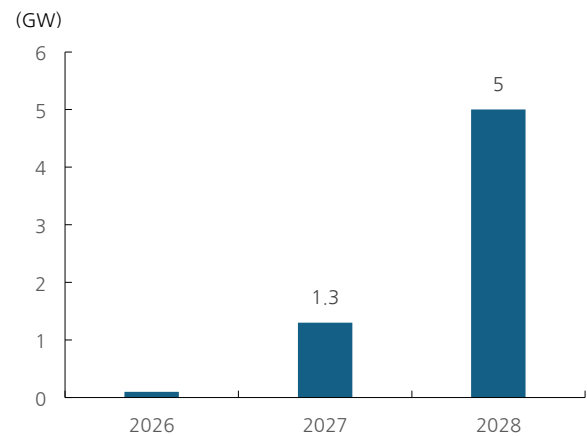
주: DeepSeek R1에서 기존 최고 수준은 GB300, GPT-OSS에서 기존 최고 수준은 GB200

그림8 Decode에서도 강점을 보인 추론 특화 ASIC Jalapeño



자료: OpenAI, DS투자증권 리서치센터

그림9 연도별 Jalapeño 및 후속 세대 배치 계획(브로드컴 언급)



자료: Broadcom, DS투자증권 리서치센터

주: 2026년은 초기 출하로 0.1GW 미만 추정, 2028년은 후속 세대 합산 5GW 이상

Google의 승부수: TPU 신제품 출시 주기를 반기로 단축

Google TPU 제품 출시
주기를 단축(2년 → 6개월)

Jalapeño가 Custom ASIC의 가능성을 보여준 상황에서 Google은 TPU 신제품 출시 주기를 기존 약 2년에서 반기 수준으로 대폭 단축하며 NVIDIA GPU 추격에 속도를 내고 있다. 빠르게 변화하는 AI워크로드에 대응하기 위해 NVIDIA가 A100 → H100 → Blackwell까지 약 2년 주기였던 제품 전환을 Blackwell Ultra부터 연간 주기로 앞당겼듯이, Google 역시 2년 전 설계한 칩 사양만으로 최신 모델과 워크로드 변화에 대응이 어려워지자 개발 주기를 크게 단축하는 모습이다. 결국 Jalapeño의 사례에서 확인했듯이 ASIC의 경쟁력은 단순한 와트당 성능을 넘어 새로운 모델 구조와 AI워크로드의 변화를 얼마나 칩에 반영할 수 있는가까지 포함하는 방향으로 확장되고 있다.

Google의 변화는 출시주기 단축에 그치지 않는다. 8세대 TPU부터 학습용 8t와 추론용 8i를 별도 설계해 워크로드별 최적화를 강화했고, Google DeepMind와 모델, 컴파일러, 네트워크까지 공동 설계하는 구조를 구축했다. 동시에 TPU의 외부 공급을 확대하면서 과거 Google 내부 TCO 절감용 ASIC에서 NVIDIA, AMD와 경쟁하는 외부 AI 가속기로 시장을 넓히고 있다. 신제품 주기가 짧아질수록 높아진 개발 비용을 회수하기 위해 생산 물량을 확대할 유인도 높아지는 만큼 Broadcom, Marvell 등 설계 파트너뿐 아니라 TSMC 파운드리, HBM, 첨단패키징 수요에도 긍정적인 전망이다. 나아가 생산 규모가 커질수록 TSMC에 집중된 공급망을 Samsung, Intel 파운드리로 다변화할 유인도 함께 높아질 전망이다.

그림10 TPU 출시 주기를 반기 주기로 단축한 Google

Google to speed up chip rollout to stay ahead in AI, technology chief says

Company to double down on R&D in Taiwan by expanding local space 60%



자료: NIKKEI Asia, DS투자증권 리서치센터

그림11 신제품 출시가 늘어날수록 물량 확대가 유력
TSMC의 대체 파운드리를 찾을 유인 확대 중

Google in talks with Samsung to make part of next-gen chip, The Information reports

By Reuters

June 11, 2026 10:10 PM GMT+9 · Updated June 12, 2026



자료: Reuter, DS투자증권 리서치센터

모델의 표준까지도 직접 장악하려는 NVIDIA

Nemotron 연합군을 통해
오픈소스 모델까지
직접 장악하려는 NVIDIA

OpenAI와 Google이 자사 모델에 최적화된 칩으로 내려가는 동안 NVIDIA는 반대로 오픈 모델 생태계로 올라가고 있다. 올해 3월 NVIDIA는 Mistral, Thinking Machines Lab, Reflection AI, Black Forest Labs, LangChain, Cursor 등이 참여하는 'Nemotron 연합'을 출범시켰으며 이후 NAVER Cloud 등으로 참여사를 확대했다. 연합은 프론티어 오픈소스 모델의 공동 개발을 목표로 하고 있는데 구체적으로 NVIDIA와 Mistral이 기초 모델을 공동 개발하고 Black Forest Labs가 이미지/비디오 등 멀티모달 역량, LangChain이 Agent 및 장기 추론 역량, Cursor가 실제 코딩 환경의 요구사항과 평가 데이터 등을 제공하는 구조이다. 이렇게 개발된 모델은 향후 Nemotron 4의 기반이 되고 오픈소스로 공개될 예정이다.

GPU에 최적화된 오픈소스
모델이 확산되면
GPU가 표준을 공고히할 수
있다는 계산

이 전략의 핵심은 NVIDIA가 자사 HW와 CUDA에 최적화된 범용 오픈 모델을 개발·확산하며 모델 계층까지 영향력을 넓히는 것이다. 모델사가 자사 워크로드에 최적화된 HW로 수직계열화를 강화하는 상황에서 NVIDIA는 반대로 자사 인프라에서 개발되는 공통 오픈 모델을 확산시켜 오픈 모델 생태계의 기반을 선점하려는 전략으로 해석된다. 여기에 최근 Hugging Face 인수를 통해 모델, 데이터셋, 애플리케이션이 유통되는 최대 개발자 플랫폼까지 확보하면서 모델 개발에서 배포와 활용에 이르는 점점도 넓혔다. 결국 NVIDIA는 단순히 GPU를 판매하는 것을 넘어 어떤 모델이 널리 사용되고 어떤 소프트웨어 스택과 인프라 위에서 최적화될 것인지에 대한 영향력을 확대하려는 모습이다.

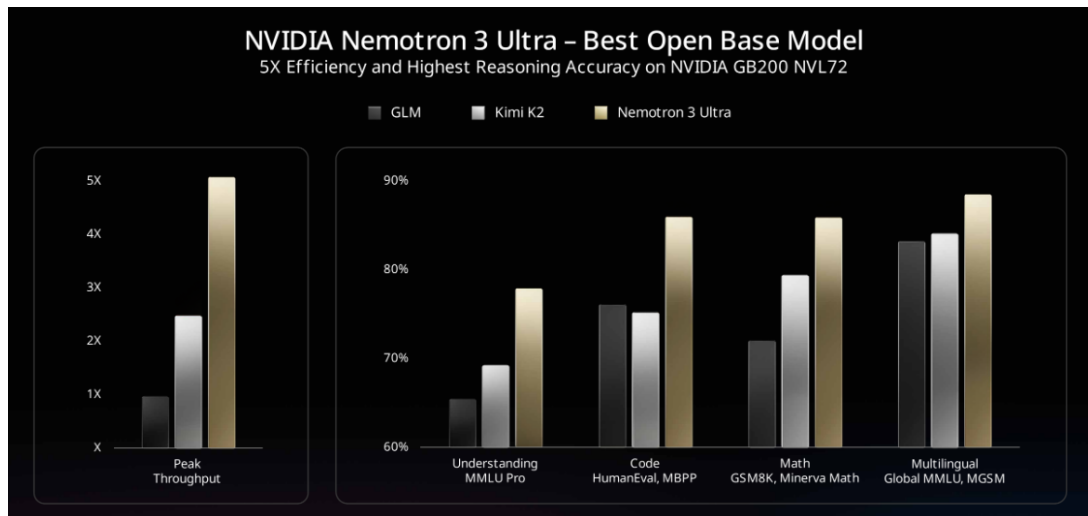
ASIC과의 공존을
불가피하게 선택

HW에서도 NVIDIA는 ASIC의 확산을 막기보다 ASIC 시대에도 NVIDIA의 시스템이 표준에 위치하는 방향을 택했다. NVLink Fusion은 하이퍼스케일러가 자체 XPU를 사용하더라도 NVLink, Spectrum-X, ConnectX, BlueField, Vera CPU와 MGX 랙 등 NVIDIA의 인프라와 결합할 수 있도록 하는 구조다. 즉 Compute의 일부가 자체 ASIC으로 이동하더라도 Scale-up, Scale-out 네트워크, CPU, 랙 등에서는 NVIDIA가 계속 경제적 가치를 가져갈 수 있다. GPU 점유율을 100% 방어하기보다 ASIC의 성장을 인정하면서도 AI Factory의 공통 플랫폼은 NVIDIA가 장악하려는 전략에 가깝다.

AI풀스택을 향한
핵심 플레이어의 레이스는
계속되는 중

결국 AI 가속기 경쟁은 단순한 GPU vs ASIC의 구도에서 벗어나고 있다. OpenAI와 Google은 자신이 가장 잘 아는 모델과 워크로드를 기반으로 AI풀스택을 수직 통합하며 ASIC의 성과와 개발속도를 동시에 높이고 있다. 반대로 NVIDIA는 GPU에서 시작해 네트워크와 랙을 장악한 뒤 이제 모델과 오픈소스 생태계까지 영향력을 확대하고 있다. 향후 승패를 결정하는 것은 개별 칩의 연산 성능보다 모델과 HW를 얼마나 빠르게 공동 최적화해 업무당 비용을 낮출 수 있는가가 될 가능성이 높다. 이 과정에서 ASIC 설계업체와 반도체 공급망은 시장 확대의 수혜가 전망된다.

그림12 2026년 3월 GTC에서 NVIDIA는 Nemotron3 Ultra를 “Best Open Base Model”로 발표



자료: Nvidia, DS투자증권 리서치센터

표2 NVIDIA, Alphabet, OpenAI 의 AI Full-Stack 경쟁력 (◎: 높음, ○: 보통, △: 낮음)

AI Full-Stack Layer	NVIDIA	Alphabet	OpenAI
Energy & Data Center	△ 직접 운영보다 설계, 파트너 투자 중심	◎ 글로벌 데이터센터 및 전력 인프라 직접 운영	○ Stargate 중심으로 대규모 용량 확보, 운영은 파트너사(Oracle, SoftBank)에 의존
AI HW & Systems	◎ GPU, CPU, DPU + 네트워킹(NVLink, InfiniBand, Spectrum-X)	○ TPU, 자체 설계 CPU(Axion) + 자체 Network를 통한 TPU Pod 구축	△ 외부 HW 활용 + 자체 AI HW 개발 착수
Infrastructure Software	◎ CUDA, TensorRT, Dynamo, Run:ai 등 HW 운영부터 추론 최적화까지 폭 넓게 보유	○ Pathways, JAX, XLA, GKE 등 자체 HW와 공동설계된 AI 운영 SW 보유	△ 훈련, 추론을 위한 자체 운영 SW를 개발하지만 내부 사용에 그침
Cloud Platform	○ CoreWeave, Nebius 등 네오클라우드 생태계를 구축하고 DGX Cloud를 통해 하나의 플랫폼으로 연결	◎ Google Cloud를 직접 운영	△ 외부 클라우드 활용
AI Models	○ Nemotron 등 오픈소스 모델	◎ Gemini 계열 프론티어 모델	◎ GPT 계열 프론티어 모델
Applications	△ Omniverse, Isaac, DRIVE 등 산업용 중심 소비자용은 제한적	◎ Search, Youtube, Workspace, Gemini App 등	◎ ChatGPT, Codex 등

자료: DS투자증권 리서치센터

주: 주황색 음영은 최근 각 회사가 강화 중인 레이어

Market View: 12MF EPS 성장률 YTD 27.8%. 진짜 금리가 중요할까?

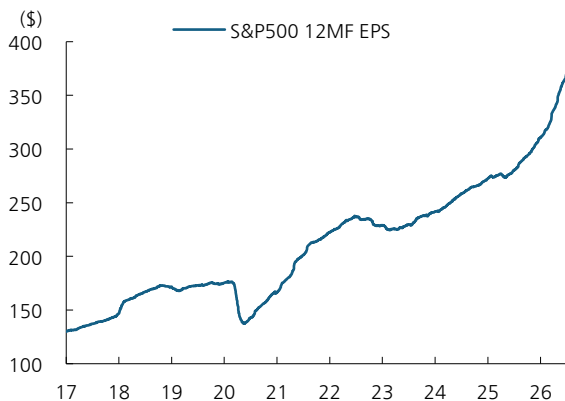
어느새 고점까지 올라온 S&P500. 가속화 중인 EPS 성장률

산적한 악재에도
기업이익 모멘텀에 힘입어
지수는 고점 부근에 위치

S&P500 지수는 9월 4일 기준 7,718.6pt를 기록하며 사상 최고점(7,816.7pt) 대비 불과 약 1% 낮은 수준에 위치해있다. (1) 미국 장기금리 상승과 기준금리 인상 가능성 (2) 지속되는 지정학적 불확실성 (3) 여전한 경기 둔화 신호 등 악재가 산적한 상황에서도 전례 없는 기업이익 모멘텀에 힘입어 지수는 다시 고점 부근까지 올라왔다. 연초 이후 S&P500의 12MF EPS 전망치는 27.8% 상향되며 이례적으로 강한 이익 모멘텀을 보여주고 있다.

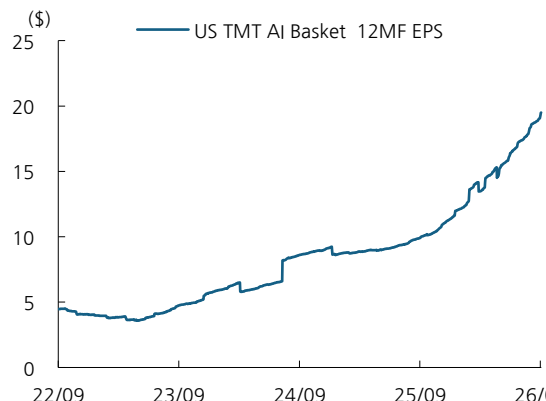
견조한 기업이익의 중심에는 역시 AI가 있다. 미국 내 100여개 AI 관련 기업으로 구성된 Goldman Sachs US TMT AI Basket의 12MF EPS 전망치는 연초 대비 61% 상향되며 AI가 시장의 이익 성장을 주도하고 있음을 보여준다. 2Q26 미국 주요 기업의 실적발표 컨퍼런스콜과 산업 자료를 종합하면 하반기에도 AI 전후방 기업의 이익 성장이 이어지며 증시의 우상향을 뒷받침 할 가능성이 높다

그림13 S&P500 12MF EPS: 전례 없이 가파른 성장(Ytd + 27.8%)



자료: Bloomberg, DS투자증권 리서치센터

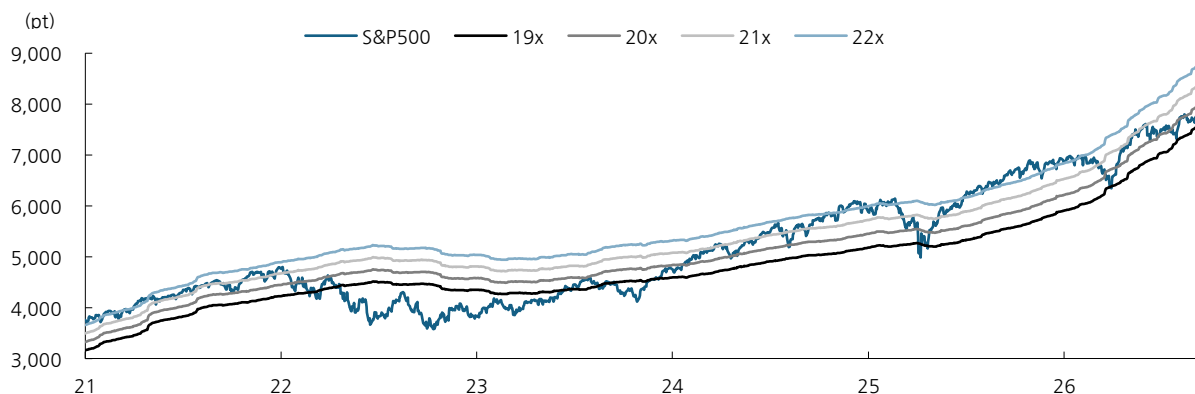
그림14 미국 AI기업 12MF EPS: 더 가파른 성장(Ytd + 61.1%)



자료: Bloomberg, DS투자증권 리서치센터

주: Goldman Sachs U.S. TMT AI Basket은 미국 증시에 상장된 AI기업 100여개를 지수화

그림15 S&P500 12MF PER 밴드



자료: Bloomberg, DS투자증권 리서치센터

2Q26 실적 컨콜을 통해 확인하는 양호한 AI업황: 데이터센터 1GW의 경제학

AI전후방기업의업황이

매우 양호한 상황

현재 미국 AI 전후방 기업의 업황은 매우 양호한 것으로 판단한다. SemiAnalysis는 Anthropic과 같은 AI 모델사가 1GW 규모의 데이터센터를 활용해 추론 서비스를 제공할 경우 연간 약 \$60Bn의 매출을 창출할 수 있다고 추정한다. 추론과 달리 훈련은 매출을 발생시키지 않기에 전체 데이터센터 용량 1GW 당 매출은 \$60Bn을 하회하겠으나 그럼에도 높은 금액이다. 이 정도의 매출이 확보되기에 하이퍼스케일러나 네오클라우드가 AI모델사에 1GW 당 연간 \$20Bn~25Bn 금액(2Q26 Nebius 컨콜)으로 컴퓨팅을 빌려주는 구조도 성립될 수 있다.

AI모델사가돈을버니

밸류체인 전반에 활기가 돌

데이터센터 공급자 입장에서 투자 회수 구조는 매력적이다. NVIDIA는 FY2Q27 컨퍼런스콜에서 1GW 규모 AI데이터센터 구축에 약 \$60Bn이 소요된다고 언급했는데, 이 중 67%를 AI서버 및 네트워킹 장비(내용연수 6년), 33%를 건설 및 부지 비용(내용연수 20년)으로 가정하면 연간 감가상각비는 약 \$7.6Bn 수준이다. 이를 1GW 당 연간 \$20Bn~25Bn의 임대 매출과 비교하면 하이퍼스케일러와 네오클라우드가 데이터센터에 적극 투자하는 이유를 이해할 수 있다. 특히 Nebius는 연초 \$12Bn/1GW였던 컴퓨팅 임대 가격이 약 2배 상승했음을 언급했는데 컴퓨팅 임대 가격이 가파르게 상승하면서 AI HW의 가격 상승을 감내하면서 데이터센터를 구축하려는 것이 현재의 모습이다.

더 중요한 점은 전후방 가격 결정력이 동시에 강화되고 있다는 점이다. AI모델사의 ARR이 증가하고, 하이퍼스케일러와 네오클라우드의 AI컴퓨팅 가격도 상승한다면 앞서 언급했듯 일정 수준의 데이터센터 건설비와 AI 서버 가격 상승을 충분히 감내할 수 있다. 이는 GPU와 네트워킹뿐 아니라 HBM을 포함한 AI HW 전반의 가격과 수요를 지지하는 요인이다. 결국 모델사 → 클라우드·네오클라우드 → AI HW로 이어지는 밸류체인 전반에서 수익성과 투자 여력이 동시에 개선되고 있다는 점에서 현재 AI 업황은 매우 건조한 국면으로 판단한다.

표3 데이터센터 1GW의 경제학

구분	1GW 당 금액 / 년	의미	출처
① AI모델사 추론 매출	~\$60Bn	1GW를 추론에 온전히 활용할 경우 창출 가능한 매출	SemiAnalysis
② AI컴퓨팅 임대료	\$20~25Bn	모델사가 Hyperscaler, Neocloud에 지급하는 Compute 비용(SW 매출 제외)	Nebius 2Q26 컨콜
③ 1GW 데이터센터 구축비용	\$60Bn	서버 및 네트워킹, 건물, 부지 포함	NVIDIA 2Q26 컨콜
↳ AI 서버 및 네트워킹	\$40.2B		업계
↳ 건설, 부지	\$19.8Bn		업계
④ 연간 감가상각비	\$7.6Bn	AI서버 및 네트워킹 내용 연수: 6년 / 건설 및 부지 내용연수: 20년	

[AI 전후방 기업 업황은 매우 우호적]

- 'AI컴퓨팅 임대료 / AI모델사 추론 매출' 이 33~42% 정도로 모델사가 충분히 지불할 수 있는 수준

- 'AI컴퓨팅 임대료 / 연간 감가상각비' 가 2.6~3.2배 정도로 하이퍼스케일러, 네오클라우드 입장에서 감가상각비를 충분히 지불할 수 있는 수준

- AI컴퓨팅 임대 계약의 연초 대비 평균 가격은 2배 상승했고 6개월 이하의 단기 계약은 3~4배 가격에도 체결(Nebius 2Q26 컨콜). AI HW의 가격이 연초대비

30~40% 상승했어도 하이퍼스케일러 등은 데이터센터 구축이 수익성 높음

자료: DS투자증권 리서치센터

금리 걱정하면 돈 못 번다

기준금리 인상하더라도
AI투자사이클을
막지 못할것

최근 미국 장기금리가 빠르게 상승한 가운데 매파적으로 인식된 연준 의장의 코멘트와 견고한 8월 고용지표가 더해지며 시장이 반영하는 9월 기준금리 인상 가능성도 58.6% 과반을 넘어섰다. 그동안 (1) 기준금리 인상만으로 지정학적 리스크와 AI데이터센터 투자발 물가 압력을 억제하기 어렵고 (2) 국채, 회사채 금리 상승을 통해 금융 여건의 긴축이 이미 진행되고 있으며 (3) 경기 둔화 국면에서 임금상승률까지 낮아 물가 상승이 장기적으로 지속되기 어렵다는 점에서 기준금리 인상의 가능성을 낮게 평가해왔다. 다만 연준의장 코멘트와 중앙은행의 선택은 시장의 예상치를 크게 벗어나기 어렵다는 점에서 금리 인상 가능성을 열어둘 필요가 있다. 그럼에도 달라지는 내용은 없다. 수십 년에 한 번 나타나는 AI 투자사이클이 EPS 전망치를 강하게 끌어올리고 12MF PER도 20배 이하로 낮아진 현 국면에서 금리만을 이유로 주식 비중을 축소할 시기는 아니라고 판단한다.

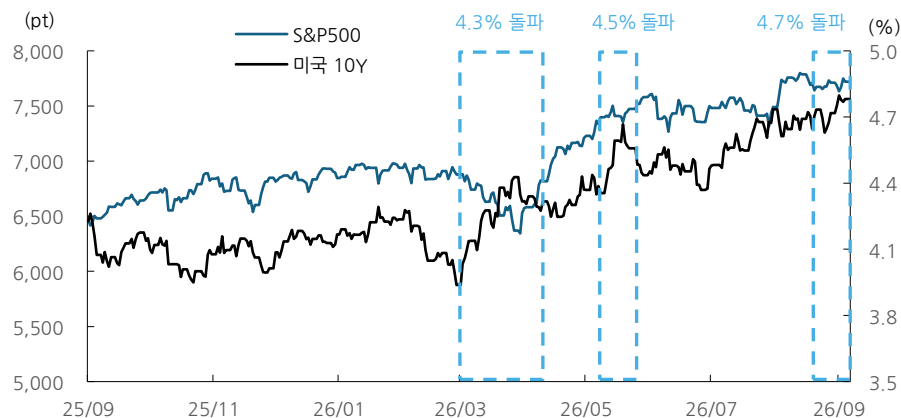
금리가 상승했음에도
지수는 2~3주도 안걸려서
최고점을 경신

성장을 동반하지 않는 금리 상승(주로 장단기 금리 역전으로 이어짐)은 주식시장에는 분명한 악재지만 강한 성장을 동반한 금리 상승까지 동일하게 해석할 필요는 없다. 금리 상승폭을 넘어서는 이익 성장이 발생하는 산업에는 실적 개선과 함께 수급이 쏠리며 주가의 강세폭은 더욱 확대될 수 있다. 최근 미국 10년물 금리와 S&P500의 흐름이 이를 보여준다. 지난 3월과 5월 미국채 10년물 금리가 각각 4.3%, 4.5%라는 심리적 임계치를 상향 돌파하자 ‘금리 상승 → 주식 비중 축소’라는 관성적인 매매가 나타나며 증시가 조정을 받았지만, 불과 2~3주 만에 다시 고점 부근을 회복했다. 금번에는 기준금리 인상 가능성이라는 변수가 추가됐지만 큰 틀의 방향은 유사할 것이다. 시가총액 1위 기업 NVIDIA EPS가 YoY 111% 성장하는 것이 현재 AI 사이클의 강도이다. 금리 상승보다 이익 성장의 속도가 훨씬 빠른 국면에서 금리에 대한 과도한 우려는 지양할 필요가 있다.

Oracle, CoreWeave
크레딧 금리만 보면 됨

유일하게 관심을 갖아야 하는 금리는 Oracle, CoreWeave 등 약한 고리로 인식되는 AI 기업의 크레딧 스프레드이다. 현재 이상 징후가 없어보인다.

그림16 미국채 금리가 임계치를 돌파하면 주가는 부진. 2~3주도 안가서 주가는 고점 경신. 진짜 금리가 중요할까?



자료: Bloomberg, DS투자증권 리서치센터

고금리에 대한 대응으로 AI퀄리티, AI고성장주를 제시

높은 수준을 유지할 금리
성장률 10% 이하 기업은
투자 매력도 낮음

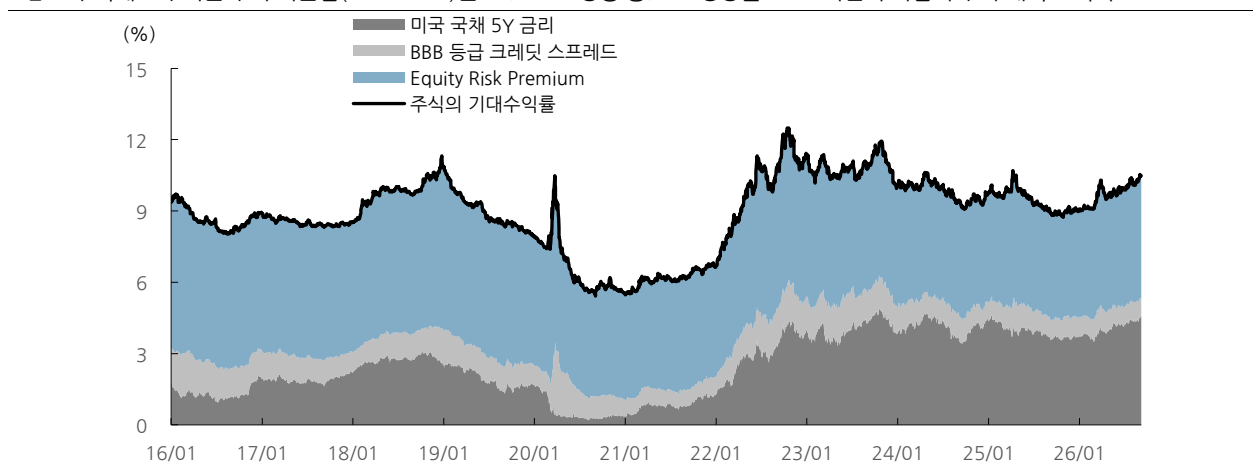
금리 상승 자체가 주식 비중 확대 및 축소의 근거가 되어서는 안 된다고 판단하지만, 섹터 및 스타일 로테이션에는 시사점을 준다. 고금리 환경이 지속될 것으로 전망하는기에 투자 대상 기업에 요구하는 성장률의 기준도 높아져야 한다. 현재 미국 5년물 국채금리 4.51%에 최하위 투자적격등급(IG)인 BBB 회사채의 크레딧 스프레드 약 1.0%와 Equity Risk Premium 5.1%를 더하면 약 10.6%가 산출되며 이를 현 금리 환경에서 주식에 요구할 수 있는 수익률의 기준점으로 판단한다. 이를 다시 말하면 현금흐름이 우수한 퀄리티 기업이라도 EPS 성장률이 10%를 하회한다면 고금리 우량 회사채 대신 리스크를 감수하며 매수할 유인은 크게 낮아진다고 판단한다.

고금리 리스크를 낮추려면
하이퍼스케일러로 대응

수익률을 높이려면
고성장 AI기업
저가매수 기회로 활용

고금리 환경 속에 일시적으로 불확실성이 높아지는 구간에서는 AI퀄리티 주식을 통한 대응이 적절해보이며, 고성장 AI주를 저렴하게 매수할 수 있는 기회로 삼는 것도 좋아 보인다. 우선 일시적인 현금흐름 부담이 있더라도 (1) 높은 컴퓨팅 가격과 견조한 수요 등 업황이 우호적이고 (2) AI 투자 확대가 매출과 이익 성장으로 연결 중인 하이퍼스케일러는 AI신대의 새로운 퀄리티주로 판단된다. Alphabet에 대한 워런 버핏의 투자 역시 이러한 구조적 경쟁력과 장기 현금흐름에 대한 신뢰와 무관하지 않았을 것으로 본다. 동시에 시장 개화 초기에 위치해 중장기 성장 경로가 뚜렷한 고성장 AI주(광통신 등) 역시 대안이 될 수 있다. 할인율 상승이 밸류에이션에 부담이더라도 이익 성장률이 요구수익률을 충분히 상회한다면 주가 상승 여력은 여전히 크기 때문이다.

그림17 주식에 요구되는 투자 기준점(hurdle rate)은 10.6%로 상승 중. EPS 성장률 10% 미만의 기업의 투자 매력도 하락



자료: Bloomberg, DS투자증권 리서치센터

주: Equity Risk Premium은 S&P500 PER의 역수

주: 이론적인 Cost of Equity는 무위험금리 + ERP로 계산되나, 회사채 대신 주식을 매수할 때 주식에 요구하는 실질적인 투자 기준점을 계산하기 위해 '크레딧 금리 + ERP'로 계산

Compliance Notice

- 동 자료는 기관투자가 등 제 3자에게 사전 제공한 사실이 없습니다.
 - 동 자료에 게시된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 작성되었음을 확인합니다. (작성자: 이규원, 강근재)
 - 동 조사자료는 고객의 투자에 참고가 될 수 있는 각종 정보제공을 목적으로 제작되었습니다. 이 조사자료는 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻어진 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 종목 선택이나 투자시기에 대한 최종 결정을 하시기 바랍니다. 따라서 이 조사자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다.
 - 동 조사자료의 지적재산권은 당사에 있으므로 당사의 허락 없이 무단 복제 및 배포 할 수 없습니다.
-