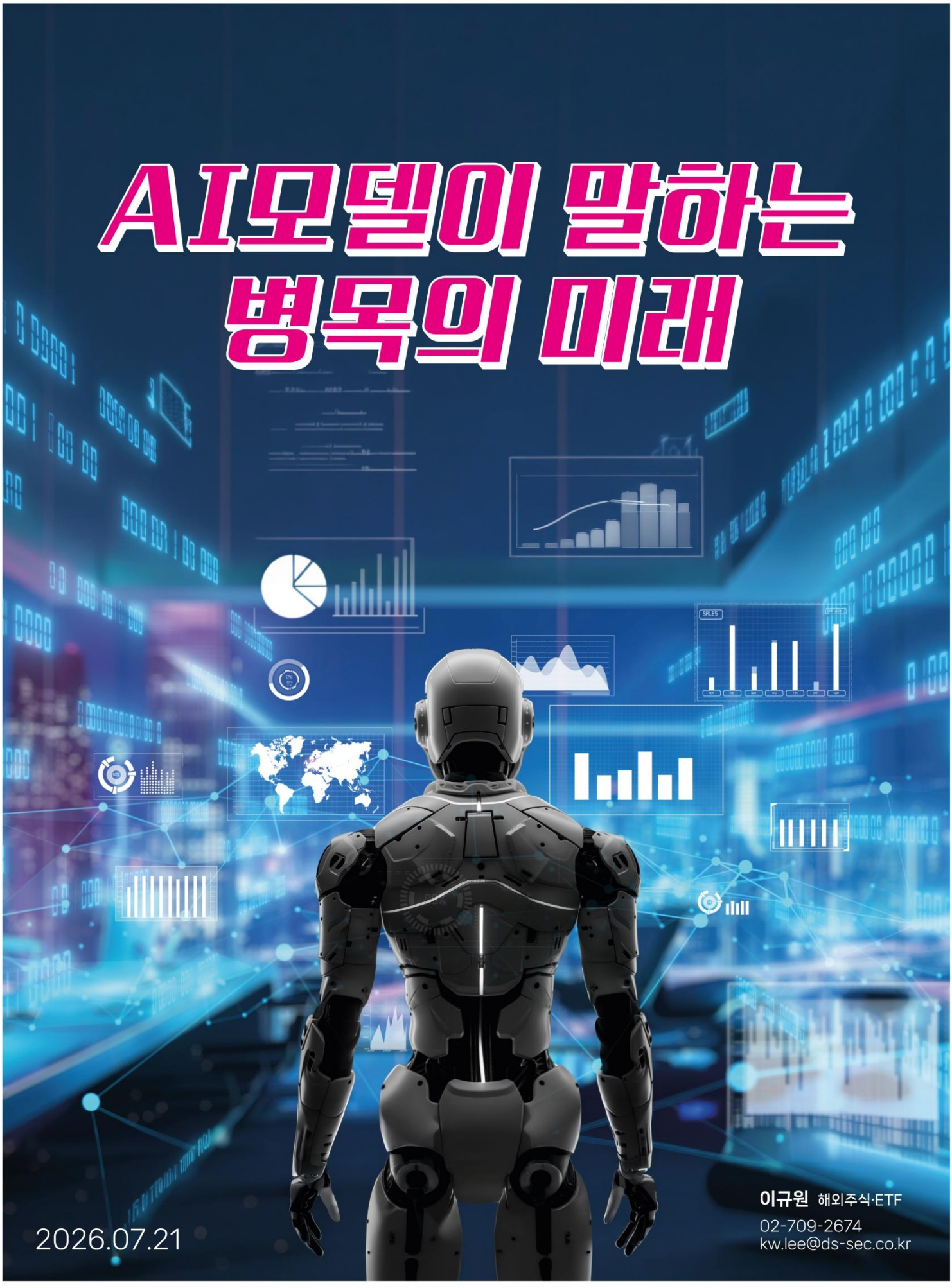
DS INVESTMENT & SECURITIES

글로벌주식

AI모델이 말하는 병목의 미래



2026.07.21

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

Executive Summary

- AI 인프라의 병목은 AI 인프라의 표준을 제시해온 엔비디아의 제품 로드맵을 통해 구체화되어왔지만, 그 출발점은 AI모델의 성능 향상 방식과 AI 워크로드의 변화에 기인한 경우가 많았다. 따라서 AI모델이 어떻게 성능을 향상시켜왔고 향후 어떤 방향으로 발전해갈지를 알면 미래 병목의 아이디를 얻을 수 있을 것이다. 본 리포트는 그러한 관점에서 작성되었다.

- AI모델은 4단계의 스케일링 법칙을 통해 성능을 높여왔으며, 각 방식은 현재도 이어지고 있다. 다만 현시점의 성능 향상은 추론량을 늘려 답변 정확도를 높이는 'Test-time Scaling'이 주도하고 있으며, 계획 수립, 도구 호출, 반복 실행을 통해 업무를 수행하는 'Agentic Scaling'으로 확장되고 있다. 두 방식 모두 토큰과 컨텍스트를 누적해 성능을 높이는 방향으로 비용 효율성이 핵심 과제로 부상했다.

- 향후 AI모델은 발전 방향은 '비용 효율화', 'Agentic 고도화', 'Enterprise AI'로 요약된다. 각 발전 방향에 따라 AI인프라 내 수혜 지점을 고민했으며 네오클라우드, 메모리, ASIC 산업의 수혜를 예상한다. Mamba, dLLM 그리고 Kimi K3의 KDA 방식까지 비싸진 메모리에 대한 반란은 계속 있겠으나 추론이 꽃피운 메모리 수요를 꺾지 못할 것이다.

- 마지막으로 하반기는 컴퓨팅 수요가 높은 가운데 AI를 통한 수익화의 길이 가시화될 전망이다. 하이퍼스케일러 입장에서는 CapEx를 축소할 이유가 없다. CapEx가 흘러가는 병목 찾기는 여전히 유효하다.

Con-
tents

병목은 AI모델이 정한다

AI모델 기초 정리

AI모델과 함께 해온 AI인프라 병목의 역사

AI 모델의 변화가 향후 AI인프라 병목의 힌트를 줄 것

AI가 확장된 이후 현재에 이르기까지 하이퍼스케일러 CapEx의 유입 지점을 선점하는 전략이 유효했으며 하반기에도 견고한 CapEx를 전망하기에 이러한 흐름은 계속될 전망이다. 중요한 것은 CapEx의 방향을 찾는 일이며 본 리포트는 그 단서를 AI 모델의 발전 방향에서 찾는다.

AI투자 사이클이 시작된 이후 주요 병목은 AI HW의 표준을 주도해온 엔비디아의 제품 로드맵을 통해 구체화되어 왔다. (1) 연산을 담당하는 GPU와 패키지에 통합된 HBM을 시작으로 (2) 다수의 GPU를 연결하는 Scale-up/Scale-out 네트워크 (3) 늘어나는 KV Cache를 저장하기 위한 메모리와 스토리지 (4) 향후 Scale-up 네트워크 영역까지 역할을 넓혀갈 광통신 등 엔비디아의 칩, 랙, 플랫폼이 어떤 부품과 시스템을 요구하고 채택하는지가 관련 기업 실적과 주가를 견인하는 핵심 변수로 작용했다.

병목은 엔비디아 제품 로드맵을 통해 구체화되었으나 그 시작점은 AI모델임

다만 엔비디아가 AI인프라의 병목을 독자적으로 만들어 왔다고 해석해서는 안된다. AI HW 변화의 출발점은 근본적으로 AI모델의 성능 향상 방식과 워크로드의 변화이며 엔비디아의 제품은 이를 가장 빠르고 효율적으로 하드웨어 아키텍처로 구현한 결과에 가깝다.

예를 들어 대규모 사전학습을 통해 모델 성능을 향상시켜온 시기에는 대규모 병렬 연산에 최적화된 GPU, HBM이 핵심 병목으로 부상했고, 이후 Reasoning 모델이 하나의 요청에 많은 추론과 토큰을 사용하면서 늘어나는 KV Cache를 처리하기 위한 메모리 용량과 Decode를 지원할 대역폭의 중요성이 높아졌다. 동시에 대형 MoE(Mixture of Expert)모델을 여러 GPU에 분산해 낮은 지연 시간으로 실행하는 모델 병렬 추론(Model Parallelism)이 확대되면서 GPU 간 통신을 담당하는 Scale-up 네트워크도 핵심 인프라로 자리잡았다. 이처럼 많은 사례에서 AI HW는 모델의 발전 방향에 맞춰 진화해왔는데 엔비디아가 자사의 강점으로 '모델 개발사와 협력하여 차세대 모델의 구조, 추론 방식, 메모리 사용량, 지연시간 요구 등을 미리 파악하고 차세대 제품에 반영하는 능력'을 언급한 것도 HW가 모델의 영향을 많이 받음을 의미한다.

특히 최근 OpenAI가 브로드컴과 협업하여 자체 칩을 개발하고, Anthropic이 마이크론과 협업을 발표(메모리, 스토리지 설계 단계에 참여하는 내용 등 포함) 하는 등 모델사는 높아진 체급을 바탕으로 HW에 영향력을 확장해 가고 있다. 따라서 본 리포트에서는 AI모델이 어떤 방향으로 발전해왔고 어디로 향하는지를 살펴보고, 그 변화 속에서 수혜가 전망되는 AI인프라를 분석하고자 한다.

표1 AI모델이 성능을 향상해온 방식과 그에 대응해온 엔비디아의 제품 로드맵(파란색 음영은 유의미한 변화)

Scaling 법칙	Train-time Scaling	Test-time Scaling			Agentic Scaling
AI 모델의 성능 향상 방식	파라미터, 학습데이터, 훈련 연산량 확대	더 많은 추론 연산과 복수 경로의 탐색 및 검증			도구 사용, 모델의 반복 실행, 에이전트 협업
NVIDIA Chip/Package Level					
구분	**H100**	**B200**	**GB200**	**GB300**	**Vera Rubin**
출시일	22년 2H	24년 2H	24년 2H	25년 2H	26년 2H
주요 HW 특징	Transformer 엔진, FP8 → 대형 모델 학습 가속화	저정밀 추론, 메모리 효율 강화	GPU, CPU를 통합 → CPU 메모리를 GPU의 보조 메모리로 활용	Reasoning에 맞춘 변화 (HBM 용량 확대, NVFP4 성능 강화)	연산, 메모리, 네트워킹 구조의 전면 세대 교체
아키텍처	Hopper	Blackwell	Grace Blackwell	Grace Blackwell Ultra	Vera Rubin
제품 단위	GPU 1개	GPU 1개	Grace CPU 1개 + Blackwell GPU 2개	Grace CPU 1개 + Blackwell Ultra GPU 2개	Vera CPU 1개 + Rubin GPU 2개
CPU 코어			Grace 72 코어	Grace 72 코어	Olympus 88 코어
Memory					
GPU 메모리	80GB HBM3	180GB HBM3e	372GB HBM3e	576GB HBM3e	576GB HBM4
GPU 메모리 대역폭	3.35TB/s	8TB/s	16TB/s	16TB/s	44TB/s
CPU 메모리			최대 480GB LPDDR5X	약 480GB LPDDR5X	1.5TB LPDDR5X
Scale-up Network					
NVLink 세대	4세대	5세대	5세대	5세대	6세대
NVLink 대역폭(GPU 당)	0.9TB/s	1.8TB/s	1.8TB/s	1.8TB/s	3.6TB/s
NVLink 대역폭(합산)	0.9TB/s	1.8TB/s	3.6TB/s	3.6TB/s	7.2TB/s
NVLink-C2C			0.9TB/s	0.9TB/s	1.8TB/s
연산 성능					
NVFP4 추론 (Sparse / Dense)	미지원	18 / 9 PFLOPS	40 / 20 PFLOPS	40 / 30 PFLOPS	100 PFLOPS*
NVFP4 학습	미지원				70 PFLOPS
FP8/FP6	3.96 / 1.98 PFLOPS	9 / 4.5 PFLOPS	20 / 10 PFLOPS	20 / 10 PFLOPS	35 PFLOPS 학습
NVIDIA Rack Scale					
구분		**GB200 NVL72**		**GB300 NVL72**	**Vera Rubin NVL72**
주요 HW 특징		초대형 모델의 실시간 추론을 위한 Scale-up 네트워크 강화		HBM 용량 확대 및 Scale-Out Network 강화 → KV Cache 증가에 대응	메모리, 네트워킹 강화 → Context, KV Cache, MoE 통신 병목 완화 → 토큰당 비용↓
제품 단위		Grace CPU 36개 + Blackwell GPU 72개		Vera CPU 36개 + Rubin GPU 72개	
주요 스펙					
GPU 메모리		13.4TB HBM3e		20.7TB HBM3e	20.7TB HBM4
GPU 메모리 대역폭		576TB/s		576TB/s	1,580TB/s
CPU 메모리		17TB LPDDR5X		17TB LPDDR5X	54TB LPDDR5X
NVLink 대역폭(합산)		130TB/s		130TB/s	260TB/s
Scale-Out Network					
Scale-out NIC		ConnectX-7		ConnectX-8	ConnectX-9
Scale-out 대역폭(GPU 당)		400Gb/s		800Gb/s	1.6Tb/s
연산 성능					
NVFP4 추론 (Sparse / Dense)		1,440/720PFLOPS		1,440/1,080PFLOPS	3,600PFLOPS
NVFP4 학습					2,520PFLOPS
FP8/FP6		720/360PFLOPS		720/360PFLOPS	1,260PFLOPS 학습
Attention 성능				GB200 NVL72 대비 2배	

자료: NVIDIA, DS투자증권 리서치센터

들어가기 앞서 1. AI모델 기초 용어 정리

본 장에서는 AI모델의 성능 향상 방식 변화를 살펴보기에 앞서, AI산업과 모델을 이해하는데 필요한 기초 용어와 주류 아키텍처인 Transformer 아키텍처를 정리하고자 한다. 실제 산업에서는 AI모델(ex. GPT 5.5), 모델 계열(ex. GPT), AI서비스(ex. ChatGPT)등이 명확한 구분 없이 혼용되므로 각 용어의 의미를 이해하고 문맥에 맞게 해석할 필요가 있다. 먼저 AI모델과 관련된 주요 기초 용어부터 아래에 정리하였다.

표2 AI모델 기초 용어

#	용어	설명
1. 모델의 구조와 분류		
1	아키텍처(Architecture)	AI모델이 입력을 받아 연산하고 출력을 만드는 기본 구조와 설계 원칙을 의미. Transformer와 RNN 등이 대표적인 모델 아키텍처에 해당
2	AI모델(AI Model)	특정 아키텍처를 기반으로 구성되고, 학습을 통해 파라미터 값이 결정된 계산 모델. 같은 아키텍처를 사용하더라도 파라미터와 학습 방식이 다르면 서로 다른 모델이 됨
3	모델 계열(Model Family)	공통된 설계 원칙과 학습 방식을 공유하면서 크기와 세대, 목적을 달리 개발된 모델의 집합. GPT, Claude, Llama 등이 각각 하나의 모델 계열에 해당
4	LLM(Large Language Model)	대규모 텍스트와 코드 데이터를 학습해 언어를 이해하고 생성하도록 개발된 언어모델(최근 이미지, 음성도 다루긴 함). 어떤 종류의 데이터를 주로 처리하는지를 기준으로 한 분류
5	Foundation Model	방대한 데이터로 폭넓게 사전 학습되어 다양한 업무와 응용서비스의 기반으로 재사용 가능한 모델. 범용적인 활용 가능성을 기준으로 한 분류이며, 주요 LLM 대부분이 이에 해당
6	AI서비스(AI Service)	학습된 AI모델에 사용자 인터페이스, 검색, Tool Calling, Memory 등을 결합해 사용자가 이용하도록 제공된 서비스. ChatGPT는 OpenAI 모델을 기반으로 제공되는 대표 AI서비스
2. 모델 내부의 기본 구성요소		
7	파라미터, 가중치(Parameter, Weight)	모델이 학습 과정에서 데이터로부터 결정하는 숫자 값으로, 입력을 어떻게 처리하고 어떤 출력을 생성할지 결정. 일반적으로 파라미터 수는 모델의 크기를 나타내는 지표로 사용됨
8	토큰(Token)	모델이 텍스트를 읽고 생성하는 기본 처리 단위. 하나의 단어가 여러 토큰으로 나뉘거나, 여러 문자가 하나의 토큰으로 묶일 수 있어 단어와 정확히 일치하지는 않으며, 각 모델 별로 토큰화하는 방식도 상이함
9	임베딩(Embedding)	토큰과 같은 데이터를 모델이 연산할 수 있는 숫자 벡터로 변환한 표현. 의미, 사용 맥락이 비슷한 데이터는 벡터 공간에서도 유사한 위치에 표현되도록 학습
10	레이어(Layer)	입력 정보를 단계적으로 변환하는 모델 내부의 연산 단위. Transformer은 Attention과 FFN 등을 포함한 레이어에서 여러 번 반복해 쌓는 방식으로 구성
11	Context, Context Window	Context는 모델이 현재 답변을 생성할 때 참고하는 입력, 대화, 생성 이력을 의미하며, Context Window는 모델이 한 번의 추론에서 처리할 수 있는 최대 토큰 범위
3. 모델의 학습과 연산		
12	학습(Training)	데이터를 이용해 모델의 파라미터를 결정하거나 조정하는 과정
13	추론(Inference)	학습된 파라미터를 사용해 실제 입력에 답변을 생성하는 과정
14	행렬곱(Matrix Multiplication)	입력 벡터와 모델 가중치를 결합해 새로운 값을 계산하는 AI모델의 핵심 연산. 동일한 계산을 대규모로 병렬처리할 수 있어 GPU가 AI연산에 적합한 주요 이유임

자료: DS투자증권 리서치센터

들어가기 앞서 2. Transformer 아키텍처 정리

현재 주류 LLM은 대부분 Transformer 아키텍처를 기반으로 함

LLM을 대중화한 GPT-3.5 기반 ChatGPT는 2022년 11월 출시됐지만, 현대 LLM의 기반이 되는 Transformer 아키텍처는 이보다 훨씬 이전인 2017년에 처음 제안되었다. 이후 약 10년이 지난 현재도 OpenAI와 Anthropic의 플래그십 모델을 비롯해 Open Source모델, Open Weight 모델(오픈소스 모델보다 공개 범위가 좁으나, 논의의 편의상 오픈소스로 통칭 예정) 대부분이 Transformer 계열을 기반으로 한다는 점에서 그 중요성이 매우 높다.

또한 AI 인프라의 많은 병목은 Transformer의 연산 특성을 효율적으로 지원하고, 구조적 한계를 보완하는 과정에서 형성되어 왔다. 따라서 AI모델의 발전 과정을 살펴보기에 앞서 Transformer 아키텍처의 기본 구조와 연산 메커니즘을 먼저 살펴볼 필요가 있으며 필요한 부분만 간략하게 정리하였다.

Transformer 아키텍처는 Attention과 FFN을 중심으로 설계

Transformer는 2017년 「Attention Is All You Need」 논문에서 처음 제안된 아키텍처로, 기존 RNN의 순환 구조를 제거하고 Attention과 FFN(Feed-Forward Network)을 중심으로 설계되었다. 여기서 Attention은 각 토큰이 다른 토큰의 정보를 얼마나 참고해야 하는지를 계산하고, FFN은 그 결과를 토큰별로 변환, 가공하는 역할을 한다.

Attention, FFN 외에도 입력 토큰을 벡터로 변환하는 Token Embedding과 토큰의 순서를 반영하는 Positional Encoding, 정보 손실과 학습 불안정을 줄이는 Residual Connection 및 Layer Normalization 등이 Transformer의 구성요소이다. 각 구성요소는 그림1과 그림 2(9p)를 통해 직관적으로 확인할 수 있다.

그림1 Transformer 모델 아키텍처

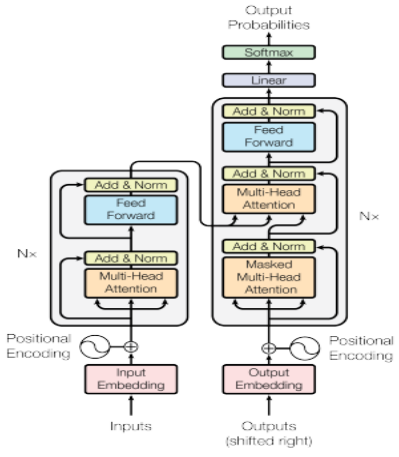


Figure 1: The Transformer - model architecture.

[추가 설명]
- 현재 GPT, Claude 등 주류 LLM 은 좌측 Transformer 아키텍처 중 오른쪽의 Decoder 구조를 기반으로 하는 Decoder-only 구조이며 아래와 같은 순서로 연산

(1) 토큰 입력
(2) Embedding: 입력 토큰을 벡터로 변환
(3) Position Encoding: 토큰의 위치 정보를 벡터에 반영
(4) Masked Multi-Head Attention(Self-Attention)
: 이전 토큰과의 관계를 계산해 문맥 정보를 반영
(5) Feed Forward(FFN): 필요한 정보만을 강조
(4)~(5)를 Layer 의 수만큼 반복
(6) Linear, Softmax: 최종 벡터를 어휘별 점수와 확률로 변환
(7) 다음 토큰 예측

자료: Attention Is All You Need (Ashish Vaswani), DS투자증권 리서치센터

Transformer 핵심 특징
1) 병렬화(Parallelization)
2) Attention

본 리포트에서 주목한 Transformer의 핵심 특징은 (1) 연산 구조와 관련된 병렬화(Parallelization)와 (2) 토큰 간 관계와 문맥을 이해는 방식과 관련된 Attention으로 구분 가능하다

1) 병렬화(Parallelization)
→ 대규모 학습의 기반

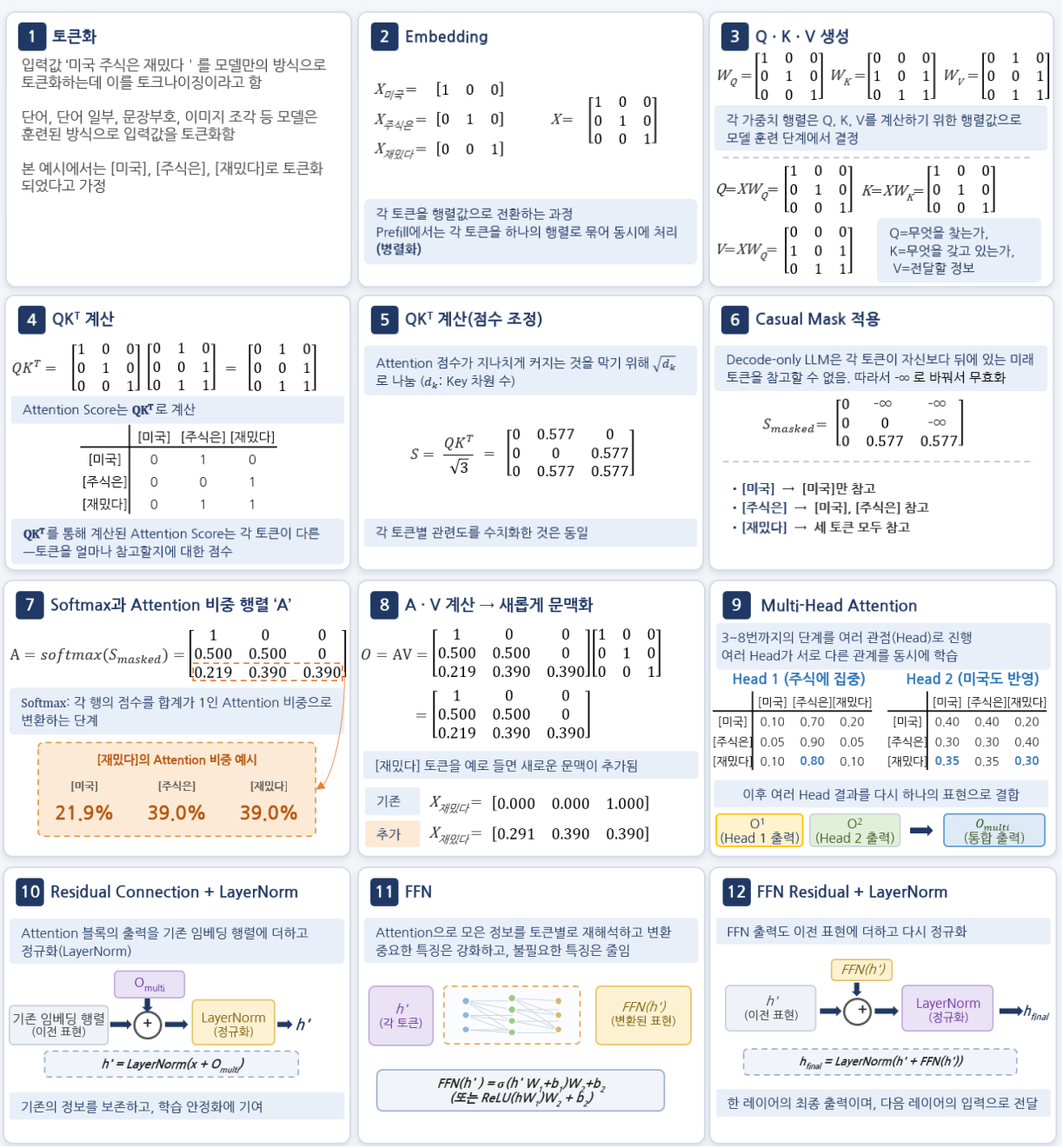
우선 Transformer는 기존 주류 아키텍처인 RNN의 순차적인 연산 구조를 제거하여 학습과 추론의 Prefill 단계에서 여러 토큰을 동시에 처리(=병렬화)할 수 있도록 했다. 이 연산은 대규모 행렬곱 형태로 수행되며 병렬 연산에 강한 GPU의 활용도를 크게 높였다. 병렬화를 주요 연산 처리 방식으로 활용하면서 AI모델의 파라미터를 늘리고 더 많은 데이터로 학습하는 대규모 학습의 기반이 마련되었다.

2) Attention
→ 각 토큰이 다른 토큰을 얼마나 참고할지 계산

다음으로 Attention은 각 토큰이 문장 내 다른 토큰과 얼마나 관련되어 있으며 그 정보를 어느정도 참고해야 하는지를 계산한다. 이를 통해 멀리 떨어진 토큰 간의 관계를 반영하고 같은 단어도 주변 문맥에 따라 서로 다른 의미를 가진 표현으로 변환할 수 있다. 그 결과 Transformer는 문장 전체의 문맥을 반영하며 번역, 질의응답, 코드 생성 등 다양한 언어 작업을 하나의 공통 구조로 처리할 수 있는 기반을 갖추게 됐다.

마지막으로 현재 주요 LLM은 초기 Transformer의 Encoder-Decoder 구조 전체를 사용하기보다, Encoder와 Cross Attention을 제외한 Decoder-only 구조를 주로 사용하고 있다. 이는 그림2의 설명 부분과 이를 시각화한 그림3을 보면 이해가 쉬울 것이다.

그림2 Transformer 아키텍처(Decode-only)의 연산 과정



자료: DS투자증권 리서치센터

AI모델은 어떻게 성능을 향상시켜왔을까?

AI 모델의 성능 향상 과정을 네 단계로 구분

당사는 NVIDIA가 언급한 AI모델의 4단계 Scaling을 근거로 현대 AI모델의 성능 향상 과정을 네 단계로 구분했다. 각 단계는 서로 다른 방식으로 모델 또는 시스템의 성능을 높여왔으며 앞선 방식을 대체하며 순차적으로 전개된 것이 아니라 기존 방식 위에 새로운 성능 향상의 축이 더해지는 형태로 발전해왔다.

표3 AI모델의 성능 향상 과정 4단계

	더 큰 모델	쓸모 있는 모델	생각하는 모델	행동하는 시스템
핵심 목표	범용 지능 향상	유용성과 정렬 향상	문제 해결 능력 향상	실제 업무 수행
성능 향상 방식	**Pre-training Scaling**	**Post-training Scaling**	**Test-time Scaling**	**Agentic Scaling**
대표 모델(서비스)	GPT-3 : 175B 파라미터 모델로 모델 크기와 성능의 관계 입증	ChatGPT(GPT-3.5) : AI의 대중화를 견인	OpenAI o1 : 내부 Reasoning을 통한 지능 향상을 입증	Claude code, Codex : Agentic AI의 시대를 개막
핵심 연산 구조	대규모 병렬 사전학습	사후학습	내부 Reasoning, CoT(Chain of Thought)	계획, 도구 실행, 결과 확인, 수정의 반복적 Loop
모델 성능 측정 지표	모델 크기, 학습 FLOPs, Time-to-Train	사용자 선호	Reasoning Token 당 성능, 달러 당 토큰 처리량, 와트 당 토큰 처리량	토큰 비용 대비 업무 효율, 장기 컨텍스트와 작업상태 유지
주요 HW 병목	GPU, HBM, Scale-up 네트워킹	GPU, HBM	메모리 용량 및 대역폭	CPU, 메모리, 스토리지 등

자료: DS투자증권 리서치센터

Phase 1. 더 큰 모델(성능 향상 방식: Pre-training Scaling)

파라미터 수, 학습 데이터, 훈련 연산량을 늘리면 성능이 향상됨

'더 큰 모델'은 Pre-training Scaling(사전학습 스케일링)을 통해 AI모델의 성능을 높인 단계다. 모델의 파라미터 수, 학습 데이터, 훈련 연산량을 함께 확대해 더 많은 지식과 패턴을 학습하도록 한 것이 핵심이다.

이러한 Scaling이 가능하게 된 배경은 Transformer의 병렬 연산 능력과 자기지도학습이 있다. Transformer는 주요 연산을 대규모 행렬곱으로 처리해 GPU의 병렬 연산 성능을 효율적으로 활용하고, 대규모 학습을 빠르게 수행했다. 그리고 자기지도학습은 웹 문서, 책, 논문, 코드 등 방대한 원본 데이터 자체를 문제와 정답으로 활용해 별도의 수작업 라벨 없이 학습 데이터를 구성할 수 있게 했다.

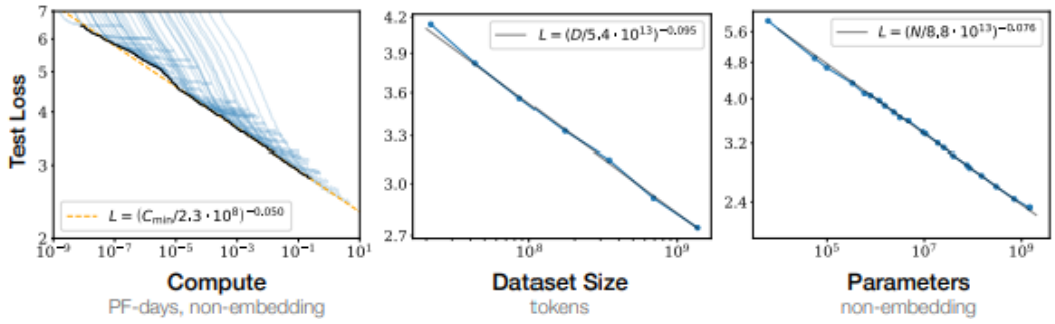
Pre-training Scaling의 경험적 기반은 2020년 발표된 'Scaling Laws for Neural Language Models(Jared Kaplan)' 논문에서 체계화되었다. 해당 논문은 모델의 파라미터, 학습 데이터, 연산량을 확대할수록 사전학습 Loss(손실)이 일정한 규칙에 따라 예측 가능하게 감소한다는 경험적 관계를 제시했다. 사전학습 손실이 감소했다는 것은 주어진 문맥에서 다음 토큰을 더 정확하게 예측한다는 의미이므로, 사전학습 관점에서 모델의 성능이 향상되었다고 봐도 무방하다.

여기에 다양한 텍스트와 코드를 하나의 모델에 사전학습하면서 특정 과제에만 사용되는 모델이 아니라 번역, 질의응답, 코딩, 리서치 등 여러 영역에 재사용할 수 있는 Foundation Model(범용 사전학습 모델)이 본격적으로 발전했다. 그리고 다양한 업무에 활용 가능한 대규모 범용 모델의 등장은 막대한 사전학습 비용을 여러 용도, 서비스에 걸쳐 회수할 수 있다는 경제적 근거도 마련되었다.

[AI HW에 미친 영향]

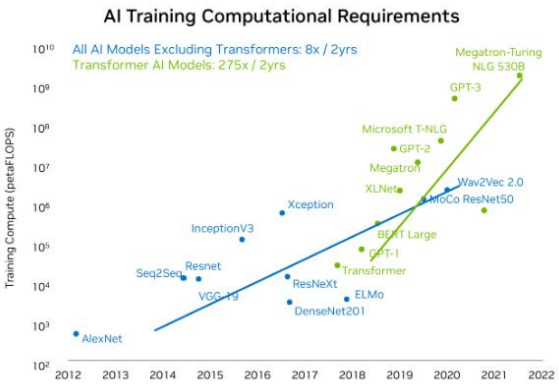
결과적으로 Pre-training Scaling과 Foundation Model의 확산은 AI HW 경쟁의 단위를 단일 GPU에서 대규모 GPU Cluster로 이동시켰다. 모델과 데이터가 커질수록 GPU와 GPU 생산에 필수적인 첨단 파운드리는 물론이고 모델 가중치와 학습 중간 데이터를 저장, 공급하는 HBM, 여러 GPU를 하나의 시스템처럼 연결하는 Scale-up, Scale-out Network의 중요성도 함께 높아졌다. ChatGPT 등장 이후 엔비디아와 GPU 핵심 밸류체인인 TSMC, SK하이닉스가 가장 먼저 주목 받았던 이유도 여기에 있다.

그림3 Compute, Dataset Size, Parameter를 키울수록 사전학습 손실은 감소(=사전학습 관점에서 모델의 성능 향상)



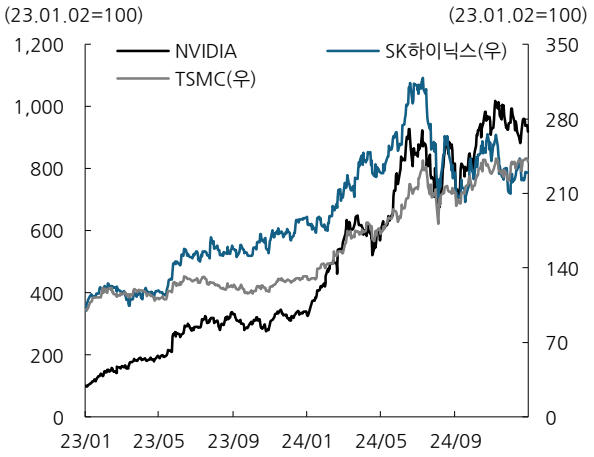
자료: Scaling Laws for Neural Language Models(Jared Kaplan, 2020), DS투자증권 리서치센터

그림4 Transformer 아키텍처가 촉발한 Scaling 경쟁



자료: NVIDIA, DS투자증권 리서치센터

그림5 AI가 부각되자 가장 먼저 오른 GPU 밸류체인 3사



자료: Bloomberg, DS투자증권 리서치센터

Phase 2. 쓸모 있는 모델(성능 향상 방식: Post-training Scaling)

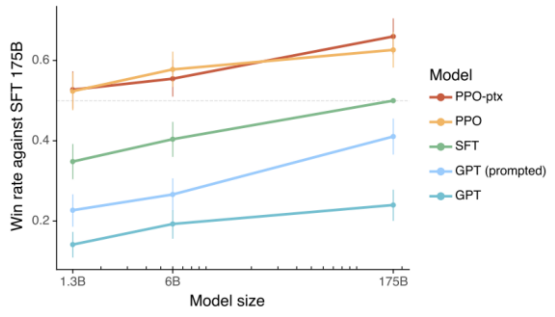
사후학습을 통해 '쓸모 있는 모델'이 등장 이는 현재 AI의 시대를 열었음

'쓸모 있는 모델'은 Post-training Scaling(사후학습 스케일링)을 통해 사전학습된 AI 모델의 능력을 사용자의 지시와 선호에 맞게 조정한 단계로 AI의 확산의 시작점인 GPT-3.5 기반 ChatGPT가 사후학습을 통해 유용성을 늘린 대표 사례이다. Pre-training을 거친 모델은 광범위한 지식과 언어 능력을 갖췄지만, 사용자의 지시를 정확히 따르거나 질문에 직접적이고 유용하게 답하도록 최적화된 상태는 아니었다. 사후학습은 이러한 모델이 사람이 원하는 방식으로 답하도록 정렬했으며, 그 결과 인간의 지시와 요구를 이해한 ChatGPT가 대중적인 성공을 거두며 AI의 시대를 열었다.

앞서 언급했듯이 사전학습을 마친 Foundation Model은 방대한 지식과 언어 능력을 학습했지만 사용자와의 커뮤니케이션에 최적화된 상태는 아니었는데 이를 사후학습을 통해 개선했던 것이다. 2022년 발표된 'Training Language Models to Follow Instructions with Human Feedback(Long Ouyang)' 논문은 모델 크기를 확대하는 것만으로 지시 수행 능력이 자동으로 확보되지 않으며 사람의 지시 예시와 선호 피드백을 활용한 추가 학습이 답변의 유용성과 진실성을 높이고 유해한 응답을 줄일 수 있음을 보여준 연구이다.

그림6은 모델 크기의 확대뿐 아니라 사람의 지시와 선호를 반영한 사후학습이 답변 품질에 큰 영향을 줄 수 있음을 보여준다. X축은 모델의 파라미터 수, Y축은 각 모델의 답변이 175B SFT 모델보다 선호된 비율로, 0.5를 넘으면 비교 대상보다 더 자주 선택됐다는 의미다. 차트는 사전학습만 거친 GPT-3, 프롬프트를 보완한 GPT-3, 지시 데이터로 지도학습한 SFT 모델, 여기에 사람의 선호를 활용한 강화학습을 적용한 PPO 및 PPO-ptx 모델을 비교한다.

그림6 1.3B PPO 계열 모델(주황색)이 175B SFT 모델보다 근소하게 선호되었음



자료: Training Language Models to Follow Instructions with Human Feedback (Long Ouyang), DS투자증권 리서치센터

표4 주요 사후학습(Post-training) 방식 정리

사후학습 방식	설명
Instruction Tuning	- 다양한 지시문과 모범 답변 데이터로 모델을 추가 학습해, 사용자 명령을 더 정확히 이해하고 따르도록 만드는 방식 - 질문 의도 파악, 형식 준수, 답변 유용성이 크게 개선됨
RLHF	- 사람이 여러 답변의 선호 순위를 매기고, 이를 학습한 보상 모델을 기준으로 모델을 강화학습 하는 방식 - 정답률뿐 아니라 친절성, 일관성, 안전성 등 사람이 선호하는 답변 특성을 강화할 수 있음
DPO	- 사람이 선호한 답변과 선호하지 않은 답변을 직접 비교하여 보상모델이나 강화학습 없이 모델을 최적화하는 방식 - RLHF보다 학습 구조가 단순하고 안정적이어서 최근 활용도가 높음

자료: DS투자증권 리서치센터

모델이 쓸모 있으려면 파라미터 크기만큼이나 사후학습이 중요

주목할 점은 파라미터가 1.3B인 PPO 계열 모델이 175B인 SFT 모델과 대등하거나 소폭 높은 선호도를 기록했다는 것이다. 이는 인간이 실제 사용하며 느끼는 실사용 맥락의 답변 품질에서는 모델 크기뿐 아니라 사람의 지시와 선호를 반영하는 사후학습이 중요하다는 점을 보여준다.

대표적인 Post-training 방식으로는 표4에 기재된 Instruction Tuning(지시 튜닝), RLHF(Reinforcement Learning from Human Feedback), DPO(Direct Preference Optimization) 등이 있다.

이러한 사후학습을 거치면서 Foundation Model은 사용자가 자연어로 번역, 질의응답, 코딩, 리서치 등 다양한 작업을 요청할 수 있는 범용적인 AI서비스로 발전됐다. 여기에 질문과 관련된 외부 자료를 찾아 답변의 근거로 활용하는 기술인 RAG가 결합되면서 모델 파라미터에 저장된 지식뿐만 아니라 외부 문서, 데이터베이스, 기업 내부 데이터까지 검색해 활용할 수 있게 되었다. 이에 따라 모델을 다시 학습하지 않고도 최신 정보와 전문 자료를 검색해 답변에 활용할 수 있는 기반이 마련됐다.

[AI HW에 미친 영향]

대규모 추론 서비스 확대로 전력과 냉각 인프라 수요까지 확대

Post-training Scaling과 이를 통한 ChatGPT의 확산(LLM의 대중화)은 AI 인프라 투자의 범위를 대규모 사전학습 중심에서 상시적인 대규모 추론 서비스로 확대했다. 전 세계 사용자의 동시 요청을 처리하기 위해서는 동일한 모델을 여러 GPU, 서버 세트에 복제해 각각의 요청을 병렬로 처리해야 했고, 이 과정에서 Serving Replication (추론 서버 복제)이 본격화됐다. 이는 AI데이터센터 증설로 이어졌고, GPU 등 AI 가속기와 HBM뿐 아니라 이를 운영하는 데 필요한 전력과 냉각 인프라 수요도 함께 증가했다.

동시에 LLM의 대중화는 AI 성능을 평가하는 기준도 변화시켰다. 사전학습 단계에서는 학습 FLOPs(모델 학습에 투입된 총 연산량)와 학습 시간 등 훈련 규모와 효율성이 중요했다면, 대규모 추론 서비스에서는 토큰을 얼마나 빠르고 저렴하게 생성하는지가 핵심이 되었다. 이에 따라 Latency, Throughput, Cost per Token이 주요 지표로 부각됐으며, Tokens/Watt와 Tokens/$로 대표되는 추론 효율성과 비용 효율성이 AI 모델과 시스템 경쟁력의 새로운 축으로 자리 잡았다.

Phase 3. 추론 모델(성능 향상 방식: Test-time Scaling)

더 많은 내부 추론을 통해 성능 향상을 이뤄낸 Test-time Scaling

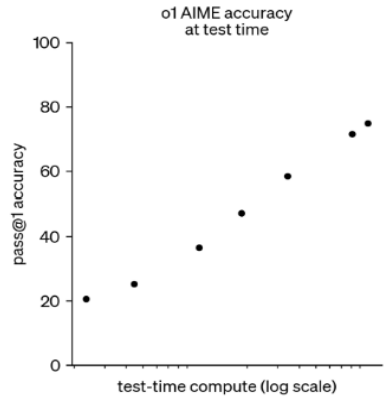
'추론 모델(Reasoning Model)'은 복잡한 문제를 해결하기 위해 추론 과정에서 더 많은 연산을 사용하도록 훈련된 모델로, Test-time Scaling을 통해 요청당 연산량을 늘려 문제 해결 성능을 높인 발전 단계다. 기반 구조는 여전히 Transformer 기반의 Autoregressive(자기회귀형) 언어모델이지만 (1) 사후학습을 통해 문제를 분해하고 중간 결과를 검증, 수정하는 추론 능력을 강화한 뒤 (2) 실제 요청에서도 최종 답변에 도달하기 전 더 많은 Reasoning Token과 연산을 사용한다는 점이 핵심이다.

추론 모델의 대표 사례인 OpenAI o1은 강화학습을 통해 Chain of Thought(생각의 사슬: 문제를 여러 단계로 분해하고 중간 추론을 거쳐 답을 도출하는 방식)를 효과적으로 활용하도록 훈련됐다. OpenAI는 추론 강화학습에 투입되는 Train-time Compute와 실제 요청에서 사용하는 Test-time Compute가 증가할수록 성능이 개선된다고 설명했다. DeepSeek-R1 역시 강화학습 과정에서 자기검증, 반성, 전략 수정과 같은 추론 행동이 나타날 수 있음을 보여줬다. 즉 Test-time Scaling은 단순히 긴 출력을 생성하는 것이 아니라 사후학습된 모델에 추가 추론 연산을 배분해 문제 해결 성능을 높이는 방식이다.

전체 연산 수요를 구조적으로 증폭시킴

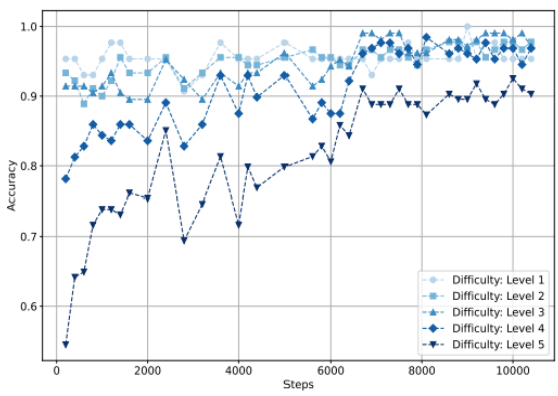
Test-time Scaling(추론 시점의 연산량을 늘려 성능을 높이는 방식)은 최근 AI모델의 발전 가운데 AI워크로드와 HW 수요를 동시에 변화시킨 핵심 요인이다. 총 세 가지 변화로 이어졌으며, 첫째는 AI의 전체 연산 수요를 구조적으로 증폭시켰다는 점이다. 기존에는 연산 수요가 주로 '사용자 수 × 사용자당 요청 수'에 따라 증가했다면, Reasoning Model에서는 여기에 '요청당 추론 연산량'이 새로운 변수로 추가되었다. 이에 따라 사용자 수와 요청 횟수가 동일하더라도 모델이 더 많은 Reasoning Token을 생성하고 검증, 수정을 반복할수록 요청당 컴퓨팅의 양은 크게 증가하는 구조가 형성되었다.

그림7 Test-time Scaling을 통해 답변의 성능 향상을 달성한 o1



자료: OpenAI, DS투자증권 리서치센터

그림8 추론 단계를 늘릴수록 답변의 정확도 상승(DeepSeek R1)



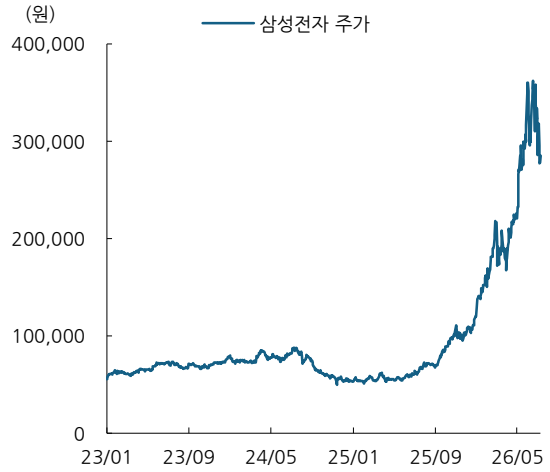
자료: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning(DeepSeek-AI), DS투자증권 리서치센터

파라미터 크기, 훈련 연산량, 학습 데이터 외에 '추론'이라는 성능 향상 경로가 생김

두 번째 변화는 서비스로만 생각되어온 추론에 대한 인식의 전환이다. 기존에는 모델 성능이 주로 파라미터 규모와 훈련 연산량에 의존했기 때문에 더 큰 모델을 학습할 수 있는 GPU와 랙 시스템의 연산성능이 중요했다. 그러나 Reasoning Model은 학습이 끝난 모델이라도 서비스 단계에서 더 많은 추론과 연산을 거치면 답변의 정확도를 높일 수 있다는 Test-time Scaling의 효과가 입증되면서 이는 '더 많은 추론'이라는 새로운 성능 향상 경로가 등장한 것이다. AI 성능 경쟁의 축이 학습에서 학습과 추론 모두로 확장된 것이며, 모델 성능에 절대적이었던 GPU 등 AI가속기의 중요성이 조금은 낮아지게 된 계기이다.

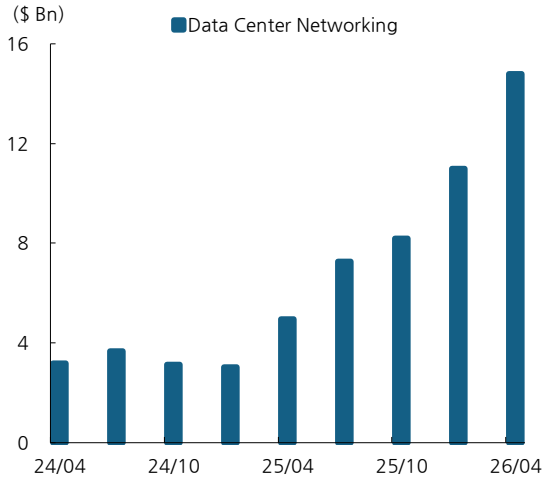
동시에 긴 추론 과정에서 토큰 생성량과 KV Cache가 증가하면서 HW 경쟁의 초점도 최대 연산성능에서 추론 효율 전반으로 확장됐다. 특히 메모리 용량과 대역폭이 토큰 생성 속도와 비용 효율을 좌우하는 핵심 요소로 부상하며 컴퓨팅 성능 중심이던 AI 인프라에서 메모리의 전략적 중요성이 한층 높아졌다. 이에 더해 네트워크 장비도 추론의 효율을 높이는 방법으로 활용되며 그 중요도가 부각되었다.

그림9 추론 모델의 등장으로 부각된 메모리의 전략적 중요성



자료: Bloomberg, DS투자증권 리서치센터

그림10 엔비디아 데이터센터 네트워킹 매출



자료: Bloomberg, DS투자증권 리서치센터

비용 및 Latency 부담이 확대되며 효율성에 대한 고민이 본격화됨

세 번째 변화는 추론 연산 확대가 성능 향상과 함께 비용 및 Latency의 부담도 키우면서 성능과 효율성 간의 균형이 AI 서비스 경쟁력의 핵심으로 부상했다는 점이다. Reasoning Model은 추론 단계에 더 많은 토큰과 연산을 투입할수록 문제 해결 성능을 높일 수 있지만 동시에 응답 시간과 서비스 비용도 증가한다. 특히 답변의 성능이 확연히 개선되면서 AI서비스의 사용자는 계속 증가했기에 비용에 대한 고민이 커질 수밖에 없게 된 것이다.

이에 따라 동일한 성능을 얼마나 적은 토큰과 비용으로 구현하는지 또는 동일한 비용과 시간 안에서 얼마나 높은 성능을 달성하는지가 중요해졌다. 그리고 이러한 변화는 모델 아키텍처와 추론 시스템 전반의 효율화를 촉진했다.

모델 아키텍처 측면에서는 전체 파라미터 중 일부 Expert만 선택적으로 활성화해 총 파라미터 규모를 확대하면서도 토큰당 연산량을 억제하는 MoE가 핵심 구조로 부상했다. 현재 최신 AI모델 다수가 MoE 구조를 채택하고 있다. 또한 추론 시스템 측면에서는 연산 집약적인 Prefill과 메모리 대역폭의 영향을 크게 받는 Decode를 별도의 GPU, 노드에서 처리하는 Prefill-Decode 분리 구조도 확산된다.

한편, 동적 Batching으로 처리량과 지연시간의 균형을 맞추고 증가하는 KV Cache를 효율적으로 배치, 공유, 재사용, 압축 기술도 중요해지고 있다. 현재 모델 구조와 AI HW의 변화를 이끄는 (1) MoE (2) Prefill, Decode 분리 (3) Batching 및 KV Cache 관리 기술을 다음 페이지부터 차례로 살펴볼 예정이다. 향후 모델과 추론 시스템의 여러 변화 역시 상당 부분 이들 기술의 연장선에서 이루어질 가능성이 높다.

Reasoning 모델 심화 학습: MoE, Prefill/Decode 분리, Batching&KV Cache

① MoE(Mixture of Expert) 아키텍처

MoE는 전체 파라미터 대비 활성 파라미터를 작게 하여 연산량 대비 성능을 향상시킴

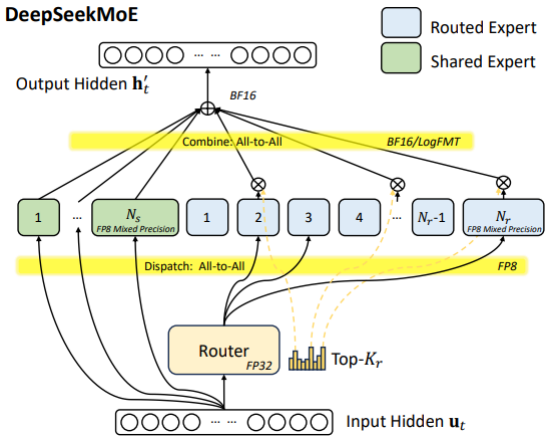
MoE 아키텍처는 전체 파라미터를 확대해 모델의 용량과 표현력은 확대하면서도 토큰당 활성 파라미터와 연산량을 제한해 비용 효율을 개선하는 구조이다. 각 토큰이 Layer 마다 동일한 FFN을 통과하는 Dense모델과 달리 FFN을 여러 Expert로 구성하고 Router가 토큰별로 일부 Expert만 선택해 연산한다. 대표적으로 DeepSeek-V3는 총 671B 개의 파라미터 중 토큰당 약 37B개만 활성화하며 전체 모델 규모를 키우면서도 연산 비용을 억제했다.

메모리, 네트워크 부담을 감소시키는 방식은 아님 오히려 네트워크 부담은 상승

다만 MoE는 연산에 참여하는 파라미터를 제한할 뿐, 메모리와 네트워크 부담까지 같은 비율로 줄이지는 못한다. 일부 Expert만 활성화되더라도 전체 Expert 가중치는 HBM에 분산 저장돼야 하며, Expert가 여러 GPU에 걸쳐 배치되면 토큰을 선택된 Expert로 보내고 결과를 회수하는 Expert Parallelism과 All-to-All 통신은 필수적이다.

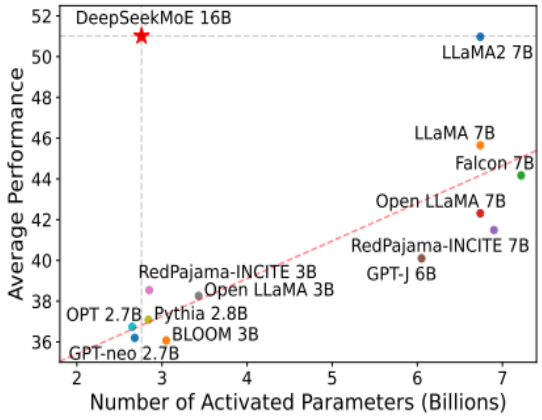
엔비디아는 자사 제품인 NVLink와 NVSwitch를 활용해 NVL72랙 내부의 GPU 72개를 하나의 고대역폭 Scale-up Domain으로 연결한다. 반면 Google은 TPU와 TPU Pod의 고대역폭 인터커넥트를 자체 설계하는 한편, TPU를 비롯한 ASIC 개발에서는 Broadcom 등 외부 파트너도 활용하고 있다. 종합하면 MoE는 모델의 전체 파라미터를 확대하면서도 토큰당 활성 연산비용을 낮추는 대신, 시스템 병목을 HBM 용량, 대역폭과 GPU 간 통신으로 이동시키거나 확대한다.

그림11 DeepSeek MoE 구조



자료: Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures(Chenggang Zhao), DS투자증권 리서치센터

그림12 DeepSeekMoE의 활성 파라미터(2.8B) 대비 우수한 성능



자료: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning(DeepSeek-AI), DS투자증권 리서치센터

② Prefill-Decode 분리(Prefill-Decode Disaggregation)

추론서비스의 두 단계, Prefill과 Decode는 성격이 다름

LLM 추론은 입력 프롬프트를 처리하는 Prefill과 출력 토큰을 생성하는 Decode로 구분된다. Prefill은 입력 토큰 전체를 병렬로 처리해 각 Layer의 표현과 KV Cache를 계산하는 단계로, 대규모 행렬 연산이 집중돼 대체로 GPU의 연산성능에 영향을 받는 Compute-bound 성격이 강하다. 반면 Decode는 저장된 KV Cache를 활용해 토큰을 하나씩 생성하며 매 단계마다 모델 가중치와 KV Cache를 반복적으로 읽기 때문에 메모리 대역폭의 영향을 크게 받는 Memory-bound 성격이 강하다.

KV Cache는 주로 GPU의 HBM에 저장되지만, 긴 Context와 동시 요청 증가로 용량 부담이 커지면서 일부 데이터를 CPU DRAM이나 SSD로 옮겼다가 필요할 때 다시 불러오는 계층형 Offloading 기술도 활용되고 있다. 이처럼 Prefill과 Decode의 병목이 다른 만큼 두 단계를 서로 다른 GPU, 서버에 배치하고 자원을 독립적으로 확장하는 Prefill-Decode 분리가 확산되고 있다.

엔비디아는 이기종 컴퓨팅을 통해 각기 다른 두 단계에 효율적으로 대응

이와 관련해 엔비디아는 Dynamo를 통해 Prefill과 Decode를 별도 Worker로 운영하고, Prefill에서 생성한 KV Cache를 Decode 쪽으로 전달하는 구조를 소프트웨어적으로 구현했다. 더 나아가 차세대 제품인 Vera Rubin 플랫폼에서는 Prefill-Decode 분리를 넘어 Decode 내부의 연산까지 세분화된다. Rubin GPU가 Prefill 전체와 Decode의 Attention을 담당하고, Groq의 LPU 기술을 기반으로 한 NVIDIA Groq 3 LPX가 지연시간에 민감한 Decode의 FFN 및 MoE Expert 연산을 처리해 토큰 생성을 가속한다. 이는 Dynamo의 오케스트레이션 아래 Attention과 FFN의 병목에 적합한 가속기를 조합하는 이종 추론 구조다.

Decode 병목을 줄이기 위한 시도로는 특화 가속기인 Cerebras가 있다. Cerebras는 웨이퍼 전체를 하나의 칩처럼 활용하고 대용량 온칩 SRAM에 모델 가중치를 배치해 외부 메모리 접근을 줄임으로써 빠른 토큰 생성을 지원한다. AWS는 Trainium이 Prefill을, Cerebras CS-3가 Decode를 담당하는 구성을 제시하며 PD 분리와 이종 가속기 결합 가능성을 보여줬다.

특정 연산 단계에 특화된 반도체 개발은 효율화와 직결되므로 확대될 예정이다. 이는 추론용 ASIC과 파운드리 수요 확대가 예상되며, Broadcom이 Google TPU 공동개발에 참여하고 NVIDIA Groq 3 LPU의 생산을 삼성전자가 맡았듯이 관련 기업에 투자의 기회도 많을 것이다.

③ KV Cache와 Batch size(부제: 메모리가 추론의 시대에 대장이 된 이유)

KV Cache는 이전 토큰의 K,V를 저장해 동일 Attention 연산을 반복하지 않도록 하는 공간

[KV Cache] 추론 서비스는 사용자의 입력을 토큰화하고 각 토큰을 벡터로 변환한 뒤, HBM 등에 저장된 모델 가중치를 활용해 여러 Transformer Layer를 통과시키는 Prefill 단계에서 시작된다. 이 과정에서 모델은 입력 토큰 간 관계를 계산해 문맥이 반영된 표현을 생성하는 동시에, Attention Layer에서 만들어진 Key(K)와 Value(V) 벡터를 KV Cache에 저장한다. KV Cache는 이미 처리한 토큰의 K,V를 보관해 Decode 단계에서 동일한 Attention 연산을 반복하지 않도록 하는 메모리 공간이다.

Batch size는 한 번의 연산에 함께 처리하는 요청의 수

[Batch size] 추론 과정에서 한 번의 연산에 함께 처리하는 요청의 수를 Batch Size라고 한다. Batch가 작으면 각 Layer의 가중치를 소수의 토큰 처리에만 활용하게 돼 GPU의 병렬 연산능력과 메모리 대역폭을 충분히 활용하기 어렵다. 반면 여러 요청을 하나의 Batch로 묶으면 한 번 읽어온 가중치를 다수의 토큰 연산에 함께 활용할 수 있어 GPU 활용률과 처리량이 높아지고 토큰당 추론 비용은 낮아진다. 따라서 추론 서비스는 허용 가능한 Latency와 HBM 대역폭 내에서 Batch Size를 조절한다.

Batch size 확대 → 추론비용 하락하지만 KV Cache는 증가 → 메모리 수요 증가

[Batch size와 메모리] Reasoning 모델은 긴 추론 과정에서 많은 출력 토큰을 생성한다. 추론 비용을 낮추기 위해 Batch Size를 키우면 동일한 가중치를 더 많은 토큰 연산에 활용할 수 있지만, 동시에 처리해야 할 KV Cache도 증가한다. KV Cache 크기는 동시 처리하는 시퀀스 수와 각 시퀀스의 누적 길이에 비례하기 때문에 Batch Size가 커질수록 메모리 부담도 빠르게 늘어난다. 결국 GPU의 연산능력이 충분하더라도 KV Cache를 저장할 HBM 공간이 부족하면 Batch Size를 더 확대하기 어려우며, 메모리 용량이 추론 처리량과 비용 효율을 제약하는 핵심 요인으로 작용한다.

KV Cache 재사용으로 추론 비용을 낮출 수 있음

[KV Cache 재사용] 추론 비용을 낮추려면 KV Cache를 재사용하여 Prefill 단계의 중복 연산을 줄이는 것이 중요하다. 여러 요청이 동일한 시스템 프롬프트, 문서, 대화 이력을 공유할 경우 이를 KV Cache로 저장하고 재사용하여 해당 구간의 Prefill 연산을 생략하는 것이다. 일부 AI서비스 사업자는 재사용한 입력 토큰에 최대 90% 낮은 과금 단가를 적용하는데 이는 재사용이 모델사에도 비용효율적임을 보여준다.

다만 Cache 재사용률을 높이려면 KV Cache를 더 많이, 더 오래 보관해야 하므로 메모리 수요도 함께 증가한다. 대규모 서비스에서는 HBM과 DRAM만으로 이를 모두 유지하기 어려워지면서, 사용 빈도가 낮은 KV Cache를 SSD 등 NAND 기반 스토리지에 저장했다가 필요할 때 다시 불러오는 계층형 관리도 도입되고 있다.

종합하면 추론 비용을 낮추려면 메모리가 더 필요. 메모리가 중요해진 이유

Batch Size 확대와 KV Cache 재사용은 추론 비용을 낮추는 동시에 더 많은 메모리 용량과 대역폭을 요구한다. 이는 Reasoning 모델 시대에 메모리가 핵심 하드웨어로 부상한 구조적 배경이다.

Phase 4. 행동하는 시스템 (성능 향상 방식: Agentic Scaling)

Agentic 시스템은 답변 생성에서 나아가 목표가 달성될 때까지 반복하는 구조

'행동하는 시스템(Agentic System)'은 AI가 질문에 답하는 단계를 넘어, 사용자가 제시한 목표를 실제 과업으로 수행하는 시스템이다. Reasoning Model이 주로 '질문 입력 → 내부 추론 → 답변 생성'으로 이어진다면, Agentic System은 '목표 해석 → 계획 수립 → 도구 실행 → 결과 관찰 → 상태 갱신 → 검증, 재계획'의 과정을 목표가 달성될 때까지 반복한다.

Agentic AI는 새로운 모델 아키텍처라기보다 Reasoning 능력을 갖춘 LLM에 RAG, Tool Calling, 외부 실행 환경, Memory, Workflow Engine, Agent Harness, 권한 관리 등을 결합해 복잡한 업무를 계획, 실행, 검증하도록 만든 시스템 아키텍처에 가깝다. 아래 표에서는 Agentic System을 구성하는 각 요소와 주요 역할을 정리했다.

Agentic System은 업무가 복잡해질수록 Orchestrator가 업무를 분해하고 다수의 전문 Agent에 작업을 배정하는 Multi-Agent System으로 확장될 가능성이 높다. Multi-Agent System에서는 업무를 나눠 맡은 각 Agent의 독립적인 작업이 병렬적으로 처리되며, Agent별로 역할에 필요한 Context와 전문화된 도구를 활용할 수 있다. Multi-Agent System는 Agentic System의 향후 발전 방향으로 관련하여 다음 단원('AI모델은 어떻게 발전해갈까?)에 상세히 정리했다.

표5 Agentic AI 시스템을 구성하는 핵심 계층

구성요소(계층)	설명
LLM (판단 계층)	- 사용자의 목표를 이해하고 작업을 하위 단계로 분해하며, 현재 상태에서 어떤 행동을 수행해야 하는지를 결정 - 직접 행동하는 주체보다는 전체 작업의 의사결정을 담당하는 Control Intelligence
RAG (정보 공급 계층)	- 기업 문서, 최신 정보, 데이터베이스 기록 등 Agent가 판단에 필요한 외부 정보를 검색해 모델 Context에 제공 - Agentic RAG는 더 이상 고정된 Retrieval(검색)단계가 아니라, Agent의 계획과 실행 과정에 통합된 동적 정보 획득 메커니즘으로 작동
Tool Calling (행동 요청 계층)	- 모델이 외부 기능을 직접 실행하는 것이 아니라, 어떤 기능을 어떤 입력값으로 실행해야 하는지를 구조화된 형식으로 요청하는 기능 - 모델의 자연어 출력을 외부 시스템이 실행할 수 있는 구조화된 행동 명령으로 변환하는 연결 계층
외부 실행 환경 (실제 행동 계층)	- 모델이 Tool Calling으로 요청한 검색, DB 조회, 코드 실행, 파일 수정 등을 실제 수행하는 계층(모델 자체가 데이터베이스를 조회하거나 코드를 실행하는 것은 아님) - 예를 들어 Coding Agent가 생성한 코드는 Sandbox Server에서 실행되고, 결과와 오류가 다시 모델에 전달돼 다음 판단에 활용
Memory (상태 유지 계층)	- Agentic System은 여러 모델과 도구 실행에 걸쳐 동일한 작업을 이어가야 하므로, 현재까지 진행 상태를 저장하는 Memory 필요. - Memory의 역할은 기억 정도였다면, Agentic AI에서는 작업이 중단되거나 다른 Worker로 이동하더라도 목표와 진행 상황을 잃지 않도록 하는 상태 관리 인프라의 기능도 수행
Workflow Engine, Agent Harness (실행 통제 계층)	- 모델이 수행하는 작업을 시스템적으로 통제하여 의사결정을 담당하는 LLM의 불필요한 반복, 무한 Loop, 단계누락 및 위험행동 방지 - LLM이 "다음에 무엇을 해야 하는가"를 판단한다면 Workflow Engine은 "이 행동을 허용할 것인지, 실패 시 다시 시도할 것인지, 언제 작업을 종료할 것인지" 등 작업 전반을 관리하며, Harness는 확률적으로 행동하는 모델을 제어함으로써 안정성과 재현성을 보완
권한관리, 보안 (행동범위 통제 계층)	- Agent가 어떤 사용자 자격으로 어떤 데이터에 접근하고, 어디까지 행동할 수 있는지를 통제하는 계층. - 부가 기능이 아니라 파일 수정, 결제처럼 모델의 판단이 실제 행동으로 이어질 때 발생하는 위험을 제한하는 필수 실행 계층

자료: DS투자증권 리서치센터

[AI HW에 미친 영향]

Agentic AI의 영향
(1) 연산 수요 증가
(2) KV Cache 부담 증가

Agentic System의 확산으로 AI 워크로드는 단일 추론 요청에서 장기 실행 Workflow 중심으로 변화하고 있으며, 이에 따른 핵심 변화는 (1) 연산 수요 증가와 (2) 장기 작업에 따른 Context 및 KV Cache 저장량 증가가 있다.

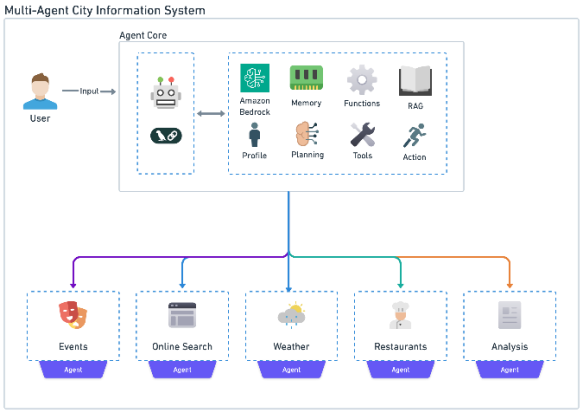
먼저 가장 직관적인 변화는 전체 연산 수요의 증가인데, 연산 수요는 '사용자 수 × 사용자당 작업 수 × 작업당 실행 단계 수 × 단계당 모델 호출 수 × 호출당 평균 연산량'으로 확대된다. 더 나아가 다수의 에이전트가 병렬 작업하는 Multi-Agent System에서는 연산량 및 컴퓨팅 수요가 한층 더 증가하게 된다.

또한 (1) 장기 작업에서 도구 실행 결과와 작업 이력이 쌓이면서 작업 상태의 저장량과 (2) Context가 길어져 발생하는 Prefill 연산 및 KV Cache 부담도 함께 커진다. 전자는 수시간 이상 이어지는 작업에서 모든 상태를 매번 Context Window에 담기 어려워 중간 결과, 진행 상태, 오류 및 복구 지점을 외부 메모리와 스토리지에 저장해야 한다는 내용이다. 그리고 후자는 Context가 길어지는 만큼 저장해야 하는 KV Cache도 증가한다는 내용인데, KV Cache의 저장 및 재사용은 불필요한 Prefill 재연산을 줄일 수 있다는 장점이 있다.

2026년 2월 발표된 'DualPath: Breaking the Storage Bandwidth Bottleneck in Agentic LLM Inference'에서도 이와 관련된 내용이 확인된다. 해당 연구의 특정 Coding Agent의 평균 Context는 32.7K 토큰에 달했지만, 각 단계에서 새로 추가되는 토큰은 평균 429개에 불과했다. 즉 토큰 기준으로 98.7%가 재사용 가능했다는 것이다. 이는 Agentic Workload에서 KV Cache를 보존하고 재사용하는 경제적 가치가 매우 높아졌음을 보여준다. Agentic System의 확산이 메모리, 스토리지 수요를 강화할 것으로 보는 이유이다.

마지막으로 Agentic AI에서 CPU는 데이터 전처리와 Scheduling, Workflow 관리뿐 아니라, 모델이 호출한 도구를 코드 실행, DB 조회, 파일 처리 등 실제 작업으로 연결하는 실행 계층을 담당한다. 이에 따라 CPU는 GPU를 보조하는 Host Processor를 넘어, AI의 판단을 외부 시스템의 행동으로 연결하는 핵심 인프라로 중요성이 높아질 전망이다.

그림13 AWS의 Multi-Agent System 구조 (Agentic System의 미래 지향점)



자료: AWS, DS투자증권 리서치센터

AI모델은 어떻게 발전해갈까?

지금까지 AI모델은 네 단계에 걸친 Scaling 축에 따라 발전해 왔다. 지금도 각 Scaling 은 여전히 모델의 성능 향상에 기여하고 있고, 특히 Reasoning 기반의 Agentic System 을 통해 AI는 수행 가능한 업무의 범위를 확장해가고 있다.

이번 장에서는 주목할 만한 AI의 발전 방향과 이에 따른 수혜, 피해 산업을 살펴볼 예정이다. 앞으로 논의할 Agentic System 고도화와 Enterprise AI 확산은 엄밀히 모델 자체보다 시스템 구조와 수요처의 변화에 가깝다. 다만 Agentic System의 고도화 방향이 모델의 발전 방향에 영향을 미치고, Enterprise AI 확산 과정에서 기업이 요구하는 비용효율성, 보안, 안정성 등이 모델의 발전 방향에 영향을 미치기에 두 흐름을 AI모델을 둘러싼 광의의 발전 방향에 포함했다.

표6 AI모델의 발전 방향과 수혜, 피해 산업

구분	발전 방향	구체적 내용	수혜, 피해 산업
비용 효율화를 위한 노력	(1) Hybrid Architecture	Transformer-Mamba Hybrid 아키텍처	[피해] - KV Cache 감소: Storage, Memory
	(2) Linear Attention (Kimi K3)	- K, V 관계를 고정된 크기로 저장 + MoE 희소성 확대(2.8T 의 거대 파라미터)	[피해] - KV Cache 감소: Storage, Memory
	(3) Conditional Memory	- 정적, 반복적 언어 패턴 등을 매번 연산하지 않기 위해 따로 저장	[수혜] - Engram 저장 수요: Memory, Storage
	(4) 작업에 따라 동적 배분	- 모델, 추론량, Tool, HW 를 작업 난이도에 따라 동적 배분	[수혜] - Routing 역할 확대: Hyperscale - 이기종 컴퓨팅 경쟁력: NVDIA - 특화 칩 수요: ASIC & Foundry
	(5) KV Cache 재사용	- KV Cache 재사용 기술을 향상 → 불필요한 연산 방지	[수혜] - KV Cache 저장 수요: Memory, Storage
	(6) Diffusion 모델	- 다수의 토큰을 병렬 생성하는 방식	[수혜] - 대규모 병렬 연산 확대: NVDIA [피해] - KV Cache 관련 낮은 연산 방식 : Memory, Storage
Agentic System 고도화	(1) 장기 실행 Agent	- 컴퓨팅 증가, 토큰 사용량 증가	[수혜] - 컴퓨팅 수요: Hyperscale, Neo-Cloud
	(2) Vertical Agent	- 특정 산업, 직무에 특화된 Agent	[수혜] - Cloud 서비스 영향력 강화: Hyperscale
	(3) Multi-Agent System	- 전문 Agent 의 독립/병렬적 업무 수행 + 관리자 Agent 의 검증'의 반복	[수혜] - Cloud 서비스 영향력 강화: Hyperscale - 병렬적 업무 수행: NVDIA, ASIC - 독립적 업무 수행(KVC↑): Memory, Storage
Enterprise AI	(1) Enterprise AI 확대	- 기업의 AI 도입 확대	[수혜] - Cloud 서비스 영향력 강화: Hyperscale - 기업의 GPU 클러스터 수요: Neo-Cloud - 자체 데이터센터 구축: NVIDIA

자료: DS투자증권 리서치센터

비용 효율화를 위한 노력

(1) 모델 단의 변화: Hybrid Architecture

성능과 동시에 비용 부담을 완화하려는 시도

1) Hybrid Architecture

2) Engram

Reasoning 모델은 많은 추론 연산을 통해 성능을 향상시켰지만, 토큰 생성량을 크게 증가시켜 서비스 비용 부담으로 이어졌다. Reasoning을 기점으로 비용 효율화 노력이 확대되었으며, 모델 아키텍처 단에서도 비용 효율화를 위한 여러 시도가 있다.

첫번째 방향은 모든 토큰에 동일한 연산을 적용하는 대신, 필요한 연산과 정보만 선택적으로 계산, 조회하는 모델로의 전환이다. 일부 Expert만 활성화하는 MoE가 대표적이며 (1) Attention과 SSM(State Space Model, 상태공간모델)을 결합해 연산 효율을 높이는 Hybrid Architecture와 (2) 반복되는 토큰 패턴을 메모리 테이블에서 조회하는 DeepSeek의 Engram 같은 시도도 주목 받고 있다. 각 시도가 주류 아키텍처로 채택되지 않더라도 향후 모델 발전 방향을 짐작하는데 단서가 될 것이다.

우선 Hybrid Architecture는 기존의 Transformer 아키텍처와 다른 아키텍처를 결합하는 방식으로 최근에는 Transformer와 Mamba아키텍처의 결합하는 시도가 있다. Mamba는 SSM을 기반으로 한 아키텍처인데, SSM은 Attention처럼 컨텍스트 내 모든 토큰의 K,V 값을 저장해 매번 참조하는 대신, 지금까지 입력된 정보를 고정된 크기의 내부 상태(State)에 압축해 유지한다. 그리고 새로운 토큰이 입력되면 이전 상태와 현재 입력을 결합해 상태를 갱신하고 이를 바탕으로 출력을 계산한다.

Mamba의 핵심은 상태 갱신 방식을 입력에 따라 변화시키는 Selective SSM이다. 현재 처리 중인 토큰의 내용에 따라 기존 상태에 압축된 과거 문맥 중 어떤 정보를 얼마나 유지하거나 약화할지, 현재 입력 토큰의 정보를 얼마나 반영할지를 조절해 Attention 없이도 중요한 문맥을 선택적으로 전달한다. 따라서 Mamba는 Prefill에서 Attention Score를 계산하지 않으며, Decode에서도 과거 토큰의 KV Cache를 누적하는 대신 고정된 크기의 내부 상태만 갱신한다. 이에 따라 시퀀스 처리 연산량은 길이에 선형적으로 증가하고(Attention은 제곱으로 증가), 추론 중 유지하는 상태 크기도 Context 길이와 무관하게 고정되어 있다. 다만 Transformer-Mamba Hybrid 모델에서는 Attention Layer에 해당하는 연산과 KV Cache가 여전히 남는다.

Transformer-Mamba Hybrid 아키텍처의 대표 사례로 AI21 Labs의 Jamba와 NVIDIA의 오픈소스 모델인 Nemotron 3 Nano가 있다. Jamba는 Transformer와 Mamba Layer를 교차 배치하고 MoE를 결합해 Attention의 문맥 이해 능력을 유지하면서도, 일반 Transformer 대비 긴 Context에서 더 높은 Throughput과 낮은 메모리 사용량을 구현했다.

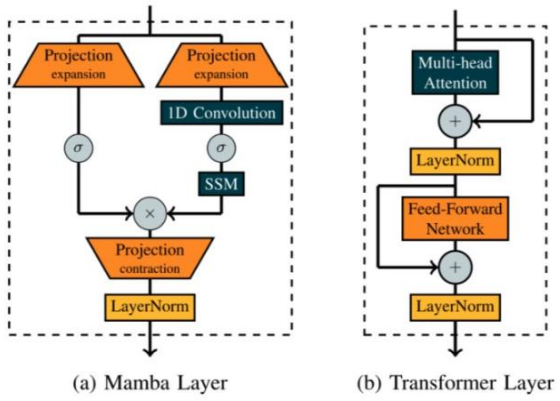
Transformer - Mamba Hybrid 모델은 KV Cache 증가 폭을 낮추는 방향

메모리 용량, 스토리지에 대한 수요에는 부정적

Nemotron 3 Nano 역시 Mamba-2와 MoE를 중심으로 구성하고 일부 Attention Layer만 남겨 추론 효율을 높였으며, 엔비디아에 따르면 H200에서 8K 입력, 16K 출력 기준 Transformer MoE인 Qwen3-30B-A3B보다 약 3.3배 높은 Throughput을 기록했다. 두 모델은 Transformer를 완전히 대체하기보다 연산 및 메모리 부담이 큰 Attention 연산을 필요한 구간에만 사용하고 나머지는 Mamba로 처리함으로써 긴 Context와 많은 Reasoning Token을 KV Cache의 증가 폭을 낮추고 처리량을 높일 수 있음을 보여준다.

Transformer-Mamba Hybrid 모델이 주류로 자리 잡으면 KV Cache가 줄어 KV Cache Offloading 목적의 DRAM, 스토리지 수요는 감소할 가능성이 높다. 반면 모델 가중치와 State를 빠르게 처리하기 위한 고대역폭(HBM) 수요는 여전히 크며, KV Cache가 감소해도 더 긴 Context와 큰 Batch, 동시 요청이 확대되면 절감된 메모리 용량을 일부 재흡수 할 수 있다.

그림14 Mamba와 Transformer의 아키텍처 비교



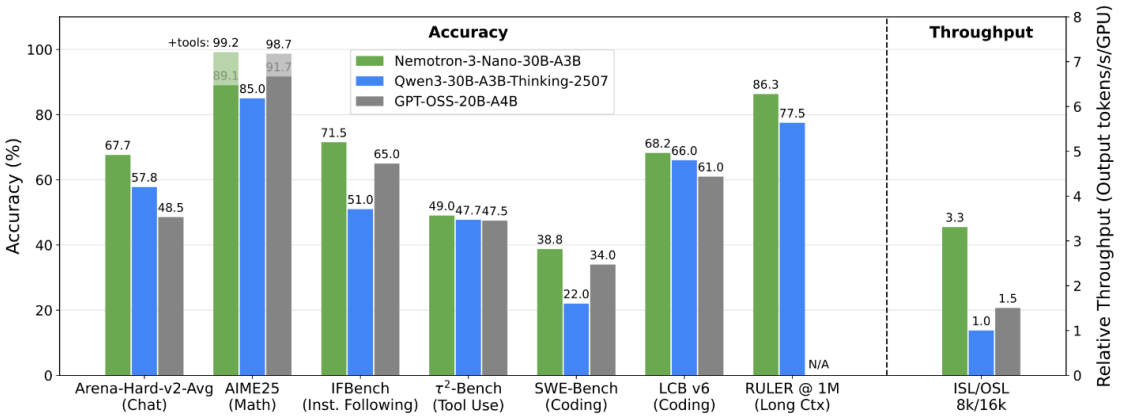
자료: Mamba for Scalable and Efficient Personalized Recommendations (Andrew Starnes), DS투자증권 리서치센터

표7 KV Cache를 현저하게 덜 발생시키는 Jamba

	Available params	Active params	KV cache (256K context, 16bit)
LLAMA-2	6.7B	6.7B	128GB
Mistral	7.2B	7.2B	32GB
Mixtral	46.7B	12.9B	32GB
Jamba	52B	12B	4GB

자료: A Hybrid Transformer-Mamba Language Model(Opher Lieber), DS투자증권 리서치센터
주: Jamba = Transformer-Mamba Hybrid

그림15 Nemotron 3 Nano, 성능과 추론 효율을 동시에 개선



자료: arXiv, DS투자증권 리서치센터

(1) 모델 단의 변화: Linear Attention(Kimi K3)

Kimi K3 가 촉발한 'LLM의 커머디티화'

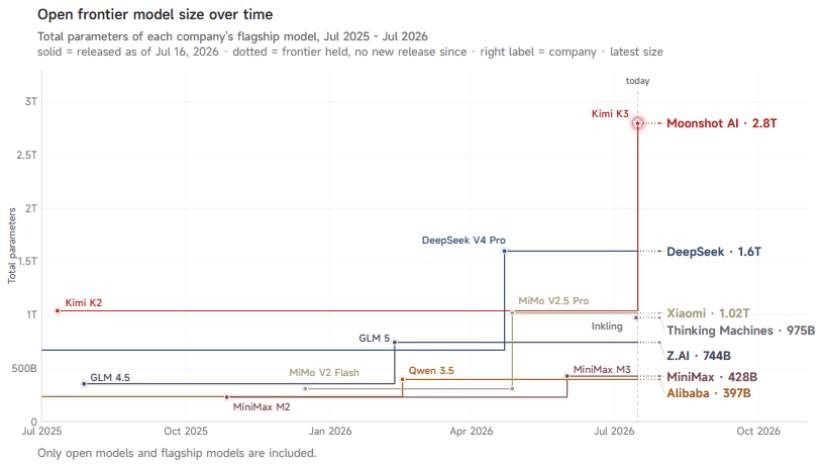
Moonshot AI가 공개한 오픈소스 모델 Kimi-K3가 OpenAI, Anthropic의 플래그십 모델인 GPT-5.6 Sol, Fable5에 근접한 성능을 기록하면서 가격 경쟁력을 갖춘 오픈소스 모델의 발전에 따른 'LLM의 커머디티화' 우려가 부각됐다. 이는 오픈소스 모델의 약진이 OpenAI, Anthropic의 수익성 악화로 이어져 AI인프라 전반의 투자를 축소시킬 수 있다는 우려이다.

'LLM의 커머디티화'와 이에 따른 미국 모델사의 수익성 악화를 벌써부터 AI인프라 투자 축소로 연결 짓는 것은 과도한 우려로 보고 있다. 그 근거는 Anthropic의 매출 성장 및 분기 흑자에 있다. 플래그십 모델이 최고 성능을 기록하면 기업, 사용자가 이를 활용해 수행 가능한 업무 영역을 확장하고 이후에 비용 부담에 따라 오픈소스로 전환해가는 일련의 흐름 속에도 Anthropic은 매출과 이익 측면에서 의미 있는 성장세를 기록 중이다. 이는 발전된 AI 성능이 영역을 확대해가는 시장 규모가 가격 부담에 따른 이탈 손실보다 크다는 점을 의미한다. 플래그십, 오픈소스 모델의 성능 격차와 사용자의 이동을 지속적으로 확인해야겠지만 과도한 우려는 투자 판단에 바람직하지 않다는 생각이다.

Kimi K3의 핵심 메시지는 '병목의 이동 가능성'일 수 있다

Kimi K3의 높은 성능은 시장에 '병목의 이동 가능성'을 암시했다고 판단하며 이 부분을 다뤄보고자 한다. Kimi K3가 타 모델과 차별화되는 지점은 (1) Kimi Delta Attention(이하 'KDA')과 (2) 2.8T에 달하는 파라미터이다. 그리고 이 두 특징을 종합한 결론을 말하면 향후 추론 시스템의 병목이 KV Cache 중심의 메모리 부담에서 가중치 연산과 GPU 간 네트워크로 일부 이동할 가능성이 있다.

그림16 총 파라미터의 크기를 크게 키운 Kimi K3(2.8T 규모 파라미터)



자료: Kimi, DS투자증권 리서치센터

KV Cache를 덜 저장하는 KDA 방식

우선 KDA에 대해 알아보자. KDA는 Moonshot AI가 개발한 Linear Attention 방식으로, 개별 토큰의 K, V를 계속 저장하는 대신 K, V 관계를 고정 크기의 행렬형 상태에 누적한다. 그리고 새로운 토큰이 입력되면 현재 행렬형 상태와 새 토큰의 K를 통해 V'를 추정하고, V'와 새 토큰의 실제 V을 비교해서 예측 오차인 Delta를 반영해 행렬형 상태를 수정한다. 따라서 KDA의 상태는 단순한 정보 저장소라기보다 새 정보에 따라 계속 갱신되는 K, V 관계표에 가깝다. 또한 Decode 단계에서는 현재 토큰의 Q가 이 관계표(행렬형 상태)를 조회해 문맥 정보를 얻고 다음 토큰을 추출하는 것이다.

이 구조는 Context 길이에 비례해 증가해온 KV Cache를 고정 크기의 상태로 대체해 저장 용량과 데이터 이동 부담을 크게 낮춘다. 반면 모든 토큰의 K, V를 개별적으로 보존하지 않기 때문에 정보의 정확성이나 특정 과거 정보의 세밀한 검색에는 제약이 생길 수 있다. 이를 보완하기 위해 Kimi K3는 KDA 층 3개와 Gated MLA 층 1개를 반복적으로 배치해, 대부분의 층에서는 효율성을 높이면서 일부 층에서는 전체 Context에 대한 정밀한 검색 능력을 유지한다. Kimi Linear 연구에서도 이러한 3:1 혼합 구조가 KV Cache 사용량을 최대 75% 줄이면서 Full Attention의 정보 검색 능력을 보완한 것으로 나타났다.

2.8T에 달하는 파라미터 배경은 (1) KDA를 통한 KV Cache 부담 완화와 (2) MoE 희소성 확대

이와 함께 주목해야 할 부분은 2.8T에 달하는 총 파라미터이다. 2026년 4월에 출시된 DeepSeek V4 Pro의 총 파라미터 수가 1.6T이고 미국의 플래그십 모델도 이와 유사한 규모로 추정되는 점을 고려하면 Kimi K3는 파라미터의 크기를 획기적으로 확대한 모델이다. 본 리포트에서는 Kimi K3가 2.8T 규모의 모델을 실용적으로 구현할 수 있었던 배경으로 (1) KDA를 통한 KV Cache 부담 완화와 (2) MoE 희소성 확대를 주목한다.

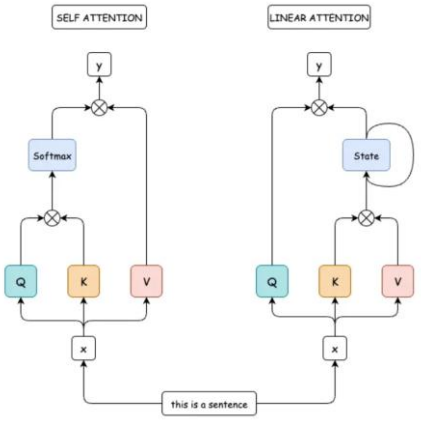
앞서 언급했듯 Kimi K3의 KDA는 K, V 관계를 고정 크기의 행렬형 메모리로 관리하며, 이에 따라 KV Cache의 양이 누적해서 증가하지 않는다. GPU의 HBM에는 모델 가중치와 KV Cache, 기타 실행 메모리가 함께 저장되므로 KV Cache가 차지하는 공간이 줄어들면 그만큼 더 큰 모델 가중치를 수용할 여유가 생기게 된다. 또한 Decode 단계에서 읽어야 할 KV Cache도 감소하여 HBM 대역폭에 여유가 생기고 이는 확대된 모델 가중치를 처리하는 데 필요한 데이터 이동 부담을 흡수할 수 있다. Reasoning 모델에서 추론 비용 효율화를 위해 양자화, 배치 크기 확대를 진행하면 KV Cache가 GPU 메모리의 최대 70%를 차지할 수 있다는 연구 결과를 고려하면, KDA를 통한 KV Cache 부담 완화가 파라미터 확대를 뒷받침했다는 해석은 가능성이 있다

다음으로 MoE 희소성의 확대는 총 파라미터를 늘리면서 토큰당 연산량을 통제하기 위한 방법으로 전체 Expert 가운데 일부만 선택적으로 활성화해 연산량 대비 성능을 향상시킨다. Kimi K3는 MoE를 적극적으로 활용했는데 Expert를 896개까지 늘리면서 토큰당 16개만 활성화해 모델의 전체 크기를 확대했다. MoE는 연산량 대비 성능 향상에 기여하지만 Expert가 여러 GPU에 분산되면 각 토큰을 Expert가 위치한 GPU로 보내고 결과를 다시 받아와야 하는 하는 All-to-All 통신이 발생하게 된다. 이 과정에서 NVLink와 같은 Scale-up 네트워크의 중요성이 높아진다.

GPU, Scale-up 네트워크를 통해 달성한 성능 향상

종합하면 Kimi K3는 KDA를 통해 KV Cache의 저장, 이동 부담을 낮추는 동시에 MoE 희소성을 확대해 총 파라미터를 2.8T까지 늘린 모델이다. 그 결과 기존의 KV Cache 중심 HBM 용량, 대역폭 병목은 일부 완화되는 반면, 확대된 가중치를 처리하기 위한 GPU 연산과 Expert 간 토큰 이동을 담당하는 GPU 네트워크의 중요성은 높아질 수 있다. 물론 엔비디아가 Kimi K3를 기준으로 향후 HW 방향을 결정하지는 않겠지만, 이 구조로 높은 성능과 효율을 구현한 사례가 등장했다는 점은 향후 타 모델에서 유사한 설계가 채택될 가능성을 보여준다.

그림17 Linear Attention(KDA)와 Self-Attention 비교



자료: Medium, DS투자증권 리서치센터

그림18 양자화 및 Batch 크기를 확대하면 메모리에서 KV Cache의 비중이 확대됨

Model	Scenario	KV	Weights	KV/Total	Savings
7B fp16	bs=1, 2K tok	0.5 GB	14 GB	3%	0.3 GB
7B fp16	bs=8, 2K tok	3.8 GB	14 GB	21%	2.3 GB
7B int4	bs=8, 2K tok	3.8 GB	3.5 GB	52%	2.3 GB
70B fp16	bs=8, 4K tok	80 GB	140 GB	36%	48 GB
70B int4	bs=8, 4K tok	80 GB	35 GB	70%	48 GB

자료: Not All Thoughts Need HBM: Semantics-Aware Memory Hierarchy for LLM Reasoning(Aojie Yuan), DS투자증권 리서치센터

(1) 모델 단의 변화: Conditional Memory

Engram은 정적, 반복적 언어 패턴을 매번 연산하지 않기 위해 저장하는 방식

비용 효율화를 위한 또 하나의 시도는 정적, 반복적 언어 패턴을 별도의 학습된 메모리에서 선택적으로 조회하는 'Conditional Memory' 이다. Transformer가 고유명사, 관용 표현 등 반복되는 단어 조합('N-gram'이라고 명명)에 대해서도 계속해서 연산 하는 것을 비효율적으로 본 DeepSeek 연구진이 2026년 Engram을 대안으로 제시했다. MoE가 필요한 Expert만 활성화해 조건부 연산을 구현했다면, Engram은 N-gram을 Memory에서 조회하는 조건부 메모리의 개념이다.

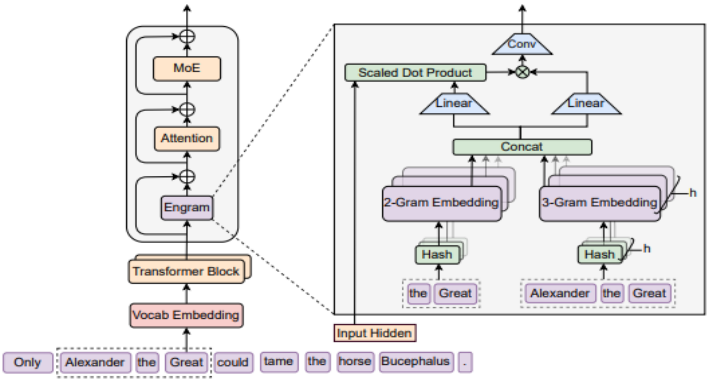
Engram은 쉽게 말해 '자주 사용하는 단어, 고유명사, 언어 패턴은 매번 다시 연산하지 말고 저장해놓고 조회하자'는 간단한 아이디어이며 저장되는 공간을 embedding table이라고 한다. 자주 사용하는 단어 조합이 입력되면 임베딩 테이블에 저장된 N-gram의 정보를 바로 불러오는 방식이며 이후 문맥과 불러온 정보가 잘 맞는지를 Gate가 판단(Apple처럼 한 단어에 여러 의미가 있을 수 있기 때문)하여 관련성이 높으면 반영하고 맞지 않으면 사용하지 않는 방식이다.

이러한 구조는 연산과 메모리의 역할을 나누는 것에 의미가 있다. 반복되는 단순 정보는 Memory 조회를 통해 빠르게 처리하고, Attention과 MoE는 복잡한 추론처럼 어려운 작업에 연산을 집중할 수 있다. 실제 Engram-27B는 파라미터와 토큰당 FLOPs가 동일한 MoE-27B보다 지식, 추론, 코딩과 수학 영역에서 높은 성능을 보였으며 이는 Memory와 Compute를 역할별로 분리하는 방식이 모델 효율을 높일 수 있음을 보여준다.

embedding table이 저장된 메모리, 스토리지 수혜

Engram이 확산되면 메모리 수요에 긍정적이다. Engram은 학습 가능한 embedding table이며, 메모리에 저장되어 있어야 한다는 점에서 메모리 수요와 직결된다. 또한 사용 빈도에 따라 HBM과 DRAM이 아니어도 SSD에 배치할 수 있어서 SSD도 수혜 범위 안에 있다.

그림19 DeepSeek 연구진이 제안한 Engram의 구조



자료: arXiv, DS투자증권 리서치센터

(2) 작업에 따라 동적 배분하는 방식

작업 난이도와 목적에 따라 자원을 효율적으로 배분하는 Model Routing

두 번째 발전 방향은 모든 요청과 작업 단계에 동일한 모델, 추론량, Tool, HW를 고정적으로 사용하는 구조에서 벗어나, 작업의 난이도와 목적에 따라 필요한 자원만 동적으로 배분하는 방식이다. 현재는 단순한 요약이나 분류에도 고성능 Model과 긴 Reasoning, 불필요한 Tool 호출과 범용 GPU가 함께 사용되는 경우가 많아 성능 대비 비용이 과도해질 수 있다. 앞으로는 최고 성능을 일률적으로 적용하기보다, Model, 추론량, Tool, HW를 작업별로 선택해 필요한 수준의 품질을 가장 낮은 비용으로 구현하는 시스템이 중요해질 가능성이 높다.

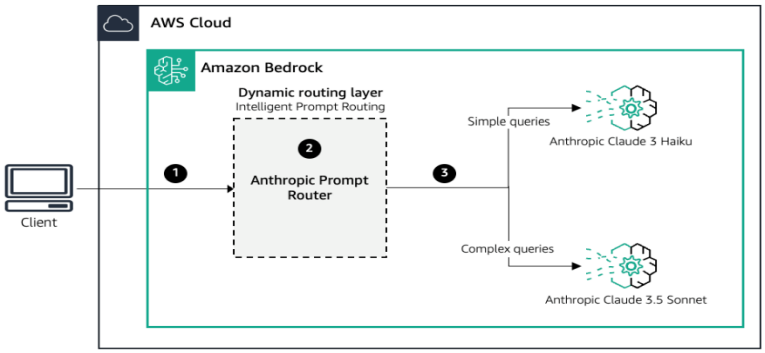
가장 먼저 확대될 방식은 Model Routing이다. 단순하고 반복적인 업무는 Small Model이나 기업이 직접 Fine-tuning한 오픈소스 Model이 처리하고, 복잡한 추론이나 높은 정확도가 필요한 요청만 Frontier Model로 전달하는 구조다. 이미 Amazon Bedrock은 요청별 예상 답변 품질을 비교해 같은 계열의 소형, 대형 모델 중 하나를 선택하는 Intelligent Prompt Routing을 제공하고 있어, Model Routing은 연구를 넘어 상용 AI Serving의 기능으로 발전한 상태이다.

동적 배분의 범위는 모델 선택을 넘어 Reasoning/Tool Routing으로 확대될 것이다. Reasoning Routing은 요청의 난이도에 따라 적절한 연산량을 배정하는 형태로 이미 OpenAI는 'reasoning.effort' 라는 설정값을 통해 모델이 답변 전 얼마나 많은 추론 연산을 사용할지 조절하고 있으며, Anthropic의 Adaptive Thinking은 요청별로 추론의 필요성과 투입량을 모델이 판단하는 방식을 적용하고 있다. Tool Routing는 모델이 외부 Tool을 호출할지, 언제 어떤 API와 인자를 사용할지를 직접 결정하는 방향으로 발전하고 있어 모델 내부 연산만으로 해결할 수 있는 작업에는 불필요한 검색, 코드 실행을 생략할 수 있다.

더 나아가 Workload에 따라 HW를 선택하는 HW Routing도 중요해질 수 있다. 현재 구현된 사례는 Prefill과 Decode를 서로 다른 GPU Pool에 배치하는 PD 분리로, 엔비디아는 Vera Rubin 플랫폼에 Groq 3 LPX를 결합해 GPU가 Prefill과 Decode Attention을 처리하고, LPX는 Decode FFN과 MoE 연산을 담당하도록 역할을 분리했다. 또한 OpenAI와 Broadcom이 개발한 Jalapeñ o처럼 추론의 특정 영역에 최적화된 ASIC도 등장하고 있다. 장기적으로는 범용성과 대규모 연산이 필요한 분야는 GPU, 반복적인 Decode는 특화 ASIC, Tool 실행과 데이터 처리는 CPU가 담당하는 이기종 컴퓨팅 구조가 확대될 가능성이 높다.

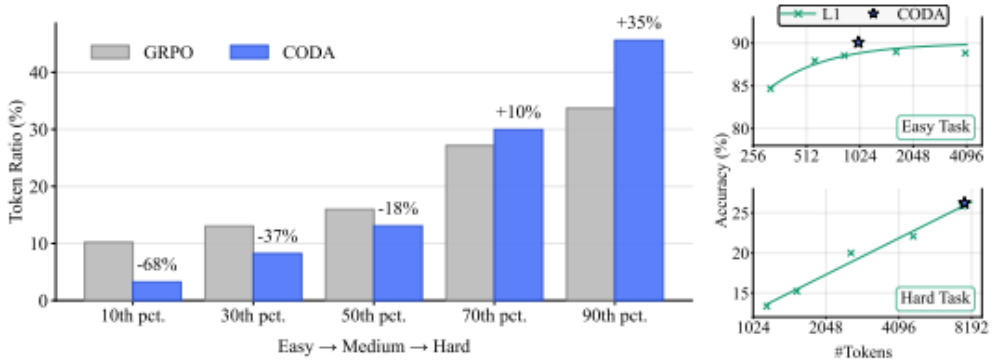
HW 별 수혜를 고민해보면, 모든 Workload를 Frontier Model, 범용 GPU에서 처리하던 구조가 약해지면서 GPU 수요는 약화될 환경이었으나 이기종 HW을 통해 엔비디아의 경쟁력은 한층 강화되는 그림이다. 또한 단순 컴퓨팅 제공에서 벗어나 Routing에서 역할을 확대할 수 있는 하이퍼스케일러와 반복적인 연산이 ASIC으로 이전되면서 ASIC 설계 및 개발 기업과 파운드리 기업에 수혜가 예상된다.

그림20 Amazon Bedrock의 Intelligent Prompt Routing



자료: AWS, DS투자증권 리서치센터

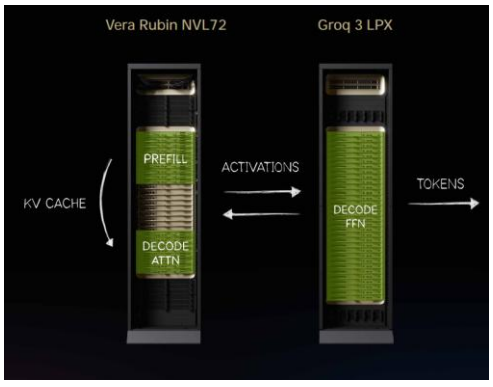
그림21 Reasoning Routing: 요청의 난이도에 따라 적절한 추론양 배정



(a) CODA allocates fewer tokens to easy tasks and more to hard tasks. (b) Accuracy vs. #Tokens.

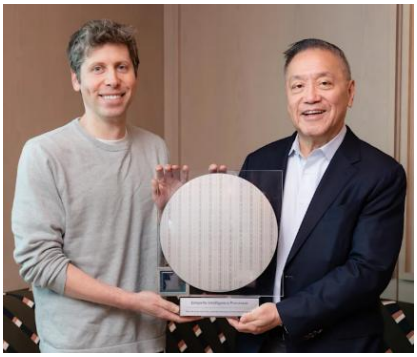
자료: CODA: Difficulty-Aware Compute Allocation for Adaptive Reasoning(Siye Wu), DS투자증권 리서치센터

그림22 Hardware Routing: 엔비디아의 이기종 컴퓨팅 구조



자료: NVIDIA, DS투자증권 리서치센터

그림23 OpenAI, Broadcom과 협업을 통해 추론 전용칩 개발



자료: OpenAI, DS투자증권 리서치센터

(3) KV Cache 재사용

비용 효율화의 핵심 기술: KV Cache 재사용

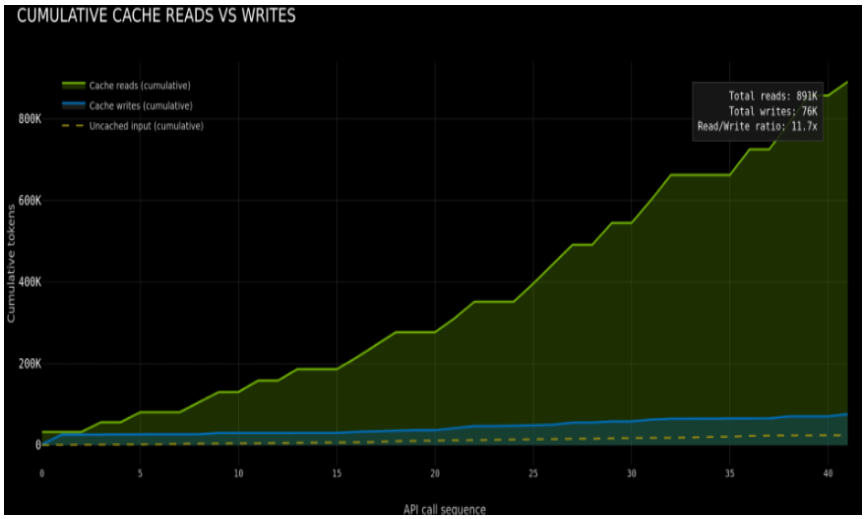
세 번째 발전 방향은 KV Cache 재사용을 확대하는 방향이다. Decode 단계에서 출력 토큰을 생성할 때마다 기존 컨텍스트의 K, V 값을 다시 계산하지 않고, 계산한 값을 KV Cache에 저장하여 이후 토큰 생성에 재사용하자는 내용이다. KV Cache 재사용은 Agentic System에 들어서 그 중요성이 높아지고 있다.

Coding Agent는 기존 코드, 작업 이력을 유지한 채 소량의 결과만 덧붙이기에 재사용 할 것이 많음

KV Cache 재사용은 새로운 요청의 앞부분이 과거 요청과 중복될 때 해당 구간에서 이미 계산한 K, V 값을 다시 불러와 Prefill 연산을 생략하는 것을 의미한다. Coding Agent의 활용이 확대되면서 그 중요성이 한층 높아졌는데 Coding Agent는 시스템 프롬프트와 도구 정의, 기존 코드와 작업 이력을 유지한 채 소량의 결과만 덧붙이며 수십~수백 회 호출되기 때문에 재사용 가능한 여력이 특히 높기 때문이다. DeepSeek의 'DualPath' 논문에 따르면 평균 32.7K Token의 기존 Context에 한 번의 Turn마다 429 Token만 추가돼 계산상 재사용 가능 비중이 98.7%에 달했다.

NVIDIA 역시 Coding Agent를 한 번 계산한 Context를 여러 번 읽는 'Write Once, Read Many' 구조로 설명하며 Claude Code를 첫 호출한 후 후속 호출에서 85~97%의 Cache 재사용률이 나타난다고 밝혔다. 주요 모델 개발사는 이러한 연산 절감 효과를 요금에도 반영하고 있으며, Anthropic은 Cache Read Token을 일반 입력 토큰 가격의 10%로 책정하고 있다. 향후 Agentic System이 확산될수록 KV Cache가 늘어날 개연성이 매우 높은 이유이며, Coinbase CEO는 토큰 사용량을 유지한 채 토큰 비용을 낮출 수 있었던 이유 중 하나로 KV Cache 재사용률을 5%에서 60%로 높인 것을 언급했다.

그림24 Coding Agent는 새로 계산하는 토큰(Cache Write)보다 기존 KV Cache를 재사용하는 토큰(Cache Read)이 훨씬 많음



자료: NVIDIA, DS투자증권 리서치센터

KV Cache 재사용이 늘어날수록 연산보다 저장이 병목이 됨

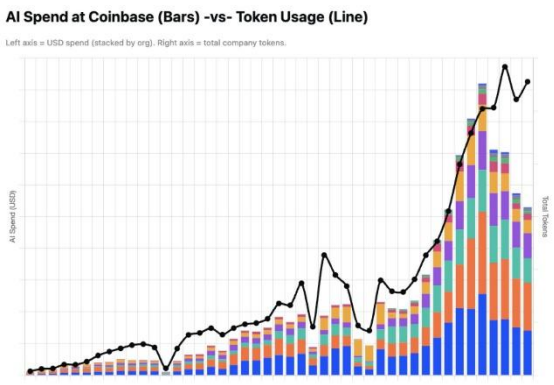
KV Cache 재사용이 늘어날수록 병목은 연산에서 KV Cache의 저장과 이동으로 옮겨가게 된다 (1) 필요한 Cache가 메모리, 스토리지에 저장되어 있지 않으면 긴 Context를 다시 Prefill해야 하는 비효율이 발생하고 (2) 저장되어 있어도 HBM으로 전송하는데 소요되는 시간이 길면 GPU가 대기하게 되면서 활용률이 낮아질 수 있기 때문이다. 특히 Agentic 워크로드에서는 새로 계산할 토큰은 적지만 불러와야 할 기존 KV Cache는 크기 때문에, 성능 병목이 GPU 연산량보다 저장소와 네트워크의 데이터 전송 속도로 옮겨가게 되는 것이고 이는 메모리 수요의 한 원인이기도 하다.

따라서 자주 사용하는 Cache는 HBM, 최근 사용한 Cache는 DRAM, 장기 보관할 Cache는 SSD에 배치하는 Memory Hierarchy(메모리 계층화)와, 이를 연결하는 PCIe, CXL, 네트워크 및 Cache가 위치한 GPU로 요청을 전달하는 Cache-aware Routing의 중요성이 높아질 가능성이 크다

[Non-prefix Caching] 더 나아가 Non-prefix Caching의 개념도 등장했다. 현재 일반적인 Caching(Prefix Caching)은 새 요청의 앞부분이 기존 요청과 토큰 단위로 정확히 같아야만 재사용할 수 있다는 한계가 있다. 예를 들어 기존 요청이 'A + B' 순서이고 이후 요청이 'B + A'이면, 같은 내용이라도 KV Cache를 활용하기 어렵다는 것이다. 이에 같은 문서나 코드가 입력의 다른 위치나 순서에 등장해도 KV Cache를 재사용하는 Non-prefix Reuse 연구가 진행 중이다.

Non-prefix Reuse 기술이 발전해 KV Cache의 재사용 범위가 확대되면, 반복 활용할 Cache를 더 오래 보관하기 위한 DRAM과 NAND SSD, 이를 메모리 계층에 연결하는 CXL 등 인터커넥트 수요에는 긍정적이다. 기존에는 요청 종료 후 폐기되던 KV Cache 중 일부가 여러 요청에서 반복적으로 활용되는 데이터로 전환되기 때문이다.

그림25 AI 지출(바 차트) 감축을 이뤄낸 코인베이스



자료: Coinbase CEO X 계정, DS투자증권 리서치센터

표8 코인베이스 CEO가 언급한 토큰 비용 감축의 5가지 비결

토큰 비용 감축 방법	설명
Better Default	- 저렴한 오픈소스 모델(GLM 5.2, Kimi 2.7 등)을 기본값으로 설정
Better Routing	- Prompt를 사전 처리한 뒤 Cache 재사용 여부와 모델 가격을 고려해 작업에 가장 적합한 모델
Better Caching	- Cache를 잘 인식하고 재사용할 수 있게 설계 - LibreChat에서 이를 제대로 구현한 뒤 Cache 재사용률이 5%에서 60%로 상승
Keep Context Lean	- 작업이 바뀌면 새 Session을 시작하고, 필요한 Context만 포함하는 등 간결한 Context 유지
Better Visibility	- 엔지니어의 Token 사용량을 제한하지 않되, 사용량을 투명하게 공개

자료: Coinbase CEO X 계정, DS투자증권 리서치센터

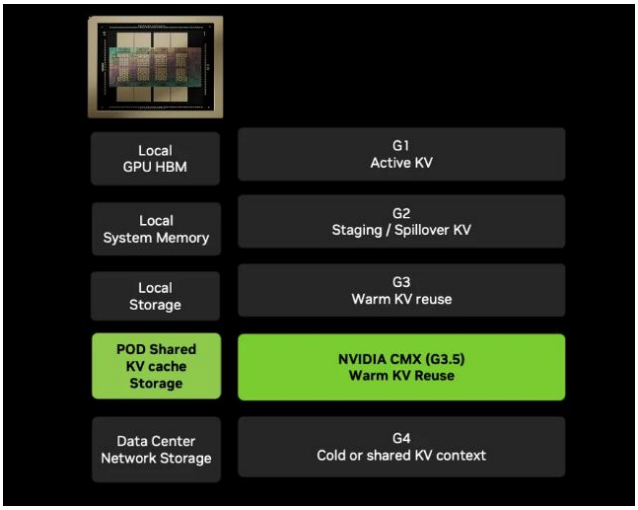
(참고 1) 엔비디아가 NAND를 Agentic AI 시대의 메모리로 점찍었다?

앞서 언급했듯 Coding Agent는 기존 코드의 작업 이력을 유지한 채 소량의 결과만 덧붙이며 수십~수백 회 호출되기에 KV Cache 재사용 여력이 높다. NVIDIA 역시 Coding Agent를 한 번 계산한 Context를 여러 번 읽는 'Write Once, Read Many' 구조로 설명하며 Cache 재사용률이 높게 나타난다고 언급한 이유이기도 하다.

엔비디아는 Vera Rubin POD를 구성하는 5종의 랙 스케일 시스템 중 하나로 'AI Native Storage'를 표방한 BlueField-4 STX 랙을 제시했다. STX는 SSD 기반인 CMX (Context Memory Storage)을 기반으로 BlueField-4 프로세서, Spectrum-X Ethernet 등으로 구성되었고, 핵심은 KV Cache 저장을 위한 POD 공용 SSD 계층을 추가한 것이다. 엔비디아 발표 기준으로 CMX는 GPU 당 약 16TB의 SSD 저장 용량을 지원한다. 여기에 소프트웨어인 Dynamo는 필요한 KV Cache를 GPU로 미리 전달하고 요청을 적절한 노드에 배치해, SSD의 낮은 전송속도를 보완하고 GPU 대기 시간을 줄이는 역할을 한다.

엔비디아가 KV Cache의 저장 계층으로 SSD를 활용한 배경에는 용량 대비 낮은 가격뿐 아니라 Agentic System의 워크로드에서는 SSD의 속도로도 활용 가능성이 있기 때문으로 해석된다. Coding Agent의 워크로드는 기존 코드를 반복하는 특성으로 추론 서비스보다 다음 호출에서 필요한 KV Cache를 예측하기 쉽다. 따라서 Agent 실행 흐름을 바탕으로 필요한 Cache를 HBM에 미리 올린다면 SSD의 낮은 전송속도에 따른 지연을 상당 부분 숨길 수 있다. 앞으로 Agentic AI의 시대에 NAND의 쓰임이 확장될 수 있다고 보는 주된 이유이다.

그림26 Vera Rubin POD 전체의 공용 SSD 계층인 CMX를 활용하여 KV Cache로 인한 메모리 용량 병목을 해소하려는 엔비디아



자료: NVIDIA, DS투자증권 리서치센터

(참고 2) Autoregressive 모델에서 벗어나려는 움직임: Diffusion 모델

AR모델은 토큰을 순차적으로 생성하며 문맥을 반영하는 과정에서 KV Cache를 누적

토큰을 순차적으로 생성하는 Autoregressive(이하 "AR") 모델은 앞서 생성된 토큰을 다음 토큰 생성에 반영해 문맥적으로 일관된 출력을 만들 수 있다는 장점이 있어 주류 LLM의 텍스트 생성 방식으로 자리 잡았다. 다만 순차적인 특성으로 지연 발생이 불가피한 점과 토큰 간의 관계를 반영하는 과정에서 KV Cache가 계속 증가한다는 단점도 있다. 이를 극복하려는 시도로 Diffusion LLM(이하 'dLLM')이 있다. 참고로 2026년 7월에 열린, 세계 3대 AI학회 중 하나인 ICML의 Outstanding Paper Award(최우수 논문)에 dLLM 관련 연구가 선정될 정도로 학계의 관심은 높다.

Diffusion은 병렬로 토큰을 예측하며 KV cache를 누적하는 구조도 아님

dLLM은 Token을 왼쪽에서 오른쪽으로 하나씩 생성하는 AR 방식과 달리, Masking된 여러 Token을 동시에 병렬 예측한 뒤 반복 수정하며 문장을 완성하는 구조다. 대표적인 모델인 LLaDA는 전체 Sequence의 일부를 가린 상태에서 여러 위치의 토큰을 복원하는 과정을 반복하며, 이론적으로는 한 번의 단계에서 여러 토큰을 함께 확정할 수 있어 순차적인 Decode 병목을 완화할 가능성이 있다. 다만 병렬 생성한 토큰 사이의 의존관계를 정확하게 반영하기 어렵고, 높은 품질을 얻기 위해 여러 차례의 Denoising이 필요하다는 한계도 있다.

메모리 수요에는 부정적이고 병렬 연산에 강한 GPU에는 긍정적인 방향

HW 관점에서는 dLLM은 연산과 메모리 간 병목 비중을 변화시킬 수 있다. 기존 dLLM은 AR 모델처럼 생성된 모든 Token의 KV Cache를 계속 누적하는 구조가 아니므로, KV Cache가 차지하는 메모리 용량 부담은 상대적으로 낮아질 수 있다. 반면 여러 Token을 병렬로 처리하고 전체 Sequence를 반복적으로 수정해야 하므로 연산 성능의 중요성은 높아질 수 있어 메모리 용량 수요에는 상대적인 하방 요인이 되는 반면, GPU와 같은 병렬 연산 가속기에는 새로운 수요 요인이 될 수 있다. 다만 최근에는 dLLM에도 KV Cache를 적용해 반복 연산을 줄이는 Fast-dLLM과 Block Diffusion 연구가 등장하고 있어 모든 dLLM이 KV Cache를 적게 사용한다고 일반화하기는 어렵다.

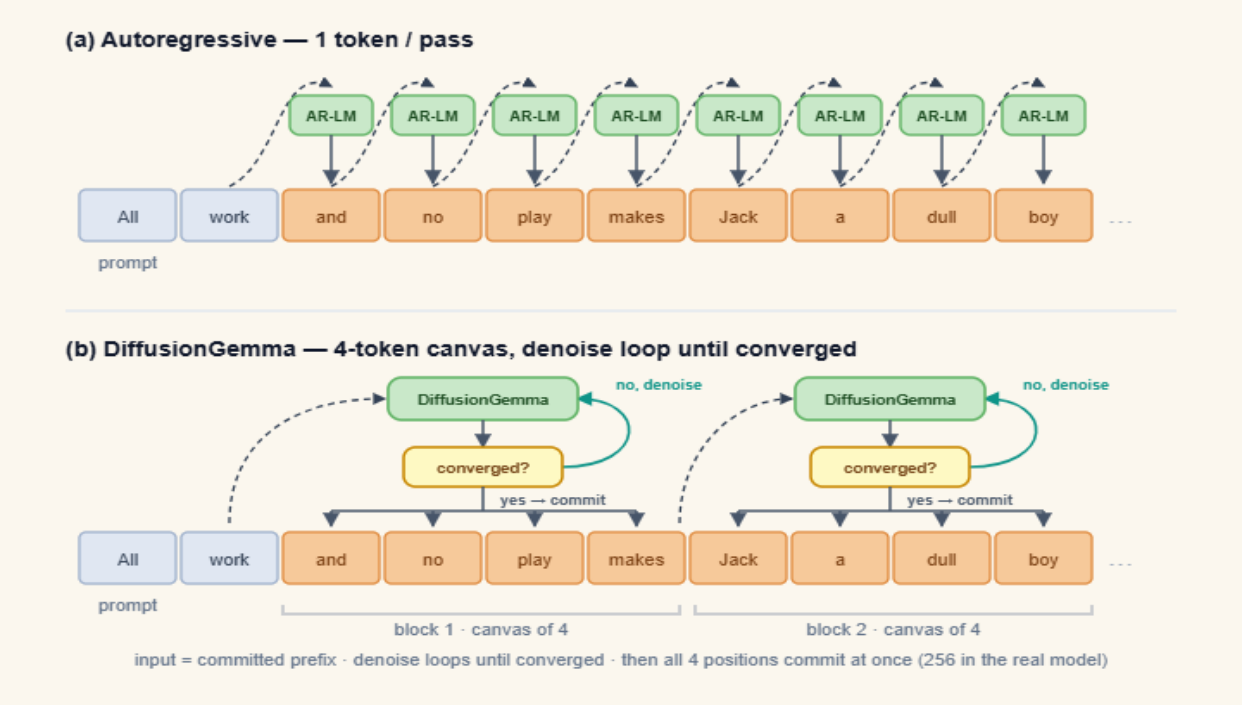
다만 현재 주류 LLM과 Serving Software, AI HW가 AR 구조에 맞춰 최적화됐다는 점은 dLLM의 확산의 제약이 될 수 있다. dLLM이 성능, Latency, 비용에서 뚜렷한 우위를 입증하지 못한다면, 생태계를 대체하기 어렵다는 의미이다.

AI HW가 AR Decode를 고도화하는 방향으로 가는 중

구체적으로 NVIDIA의 Vera Rubin POD은 Rubin GPU가 Prefill, Decode Attention을 처리하고 Groq 3 LPX가 Decode의 FFN과 MoE Expert 연산을 담당하도록 설계되었는데 이는 AR Decode를 고도화하는 방향이다. 모델 아키텍처가 HW의 방향을 결정하지만, 반대로 기존 HW, Software의 최적화 수준도 특정 모델이 경제성을 결정하기에 dLLM은 당분간 전체 AR방식을 대체하기보다 병렬 생성의 장점이 큰 일부 Workload에서 먼저 활용될 가능성 정도가 합리적이다. 다만 TurboQuant의 충격에서 경험했듯이 KV Cache로 인한 여러 병목에서 벗어나려는 움직임은 분명히 존재하는 만큼 관련된 내용을 인지해두는 것은 의미 있어 보인다.

당분간 dLLM이 LLM 대체는 어려울 것

그림27 AR모델(a)의 순차적인 연산과 Diffusion 모델(b)의 병렬적인 연산



자료: vLLM, DS투자증권 리서치센터

Agentic System 고도화

Agent AI은 이제 겨우 시작단계…

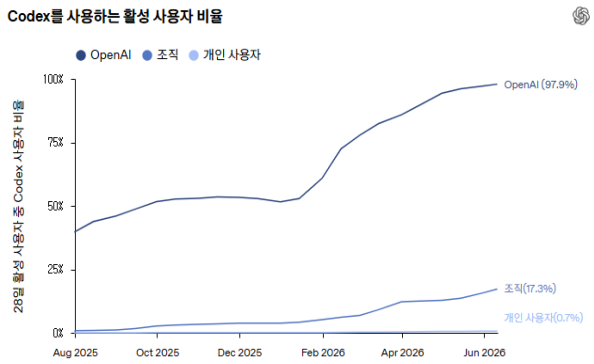
AI Agent는 AI 생산성을 유의미하게 향상시키면서 향후 빠르게 확산될 전망

Agentic System(편의상 "Agentic AI"로 표현)은 향후 수년간 AI 발전의 핵심축이 될 가능성이 높다. 질문과 답변 생성에 그친 기존 AI와 달리 Agent는 실제 업무를 완료할 수 있다는 점에서 AI의 생산성을 유의미하게 향상시킬 수 있을 것으로 전망한다. 이런 맥락 속에 Anthropic, OpenAI 등 주요 모델 개발사도 제품의 중심을 단일 답변에서 장기 작업과 기업 업무 수행으로 확대하고 있다.

6월 25일, OpenAI가 공개한 Codex의 사용 행태 분석도 이러한 가능성을 보여준다. 이 자료에서 주목한 부분은 (1) 기업, 개인 활성사용자 중 Codex를 사용 비중이 17.3%, 0.7%로 아직 낮지만 Codex가 전체 출력 토큰에서 차지하는 비중은 각각 16.5%, 63.3%에 달한 점(=Agentic AI의 높은 토큰 사용량을 증명) (2) 개인, 기업 Codex 활성사용자가 빠르게 증가하고 있다는 내용이다.

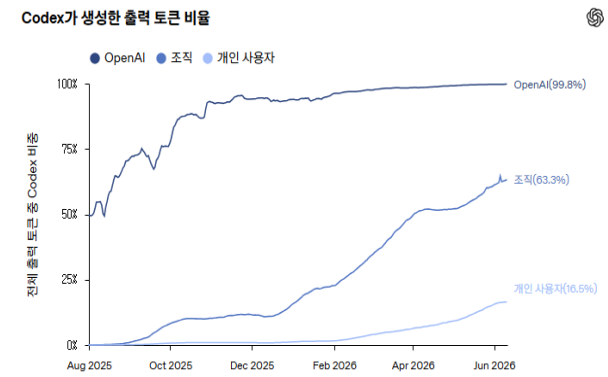
Agent AI는 기업과 사용자의 요구를 반영해서 다양한 방향으로 고도화될 것이다. 본 리포트에서는 (1) 장기 실행 Agent (2) Vertical Agent (3) Multi-Agent System 에 주목했다.

그림28 Codex를 사용하는 활성사용자: 기업17.3%, 개인 0.7%



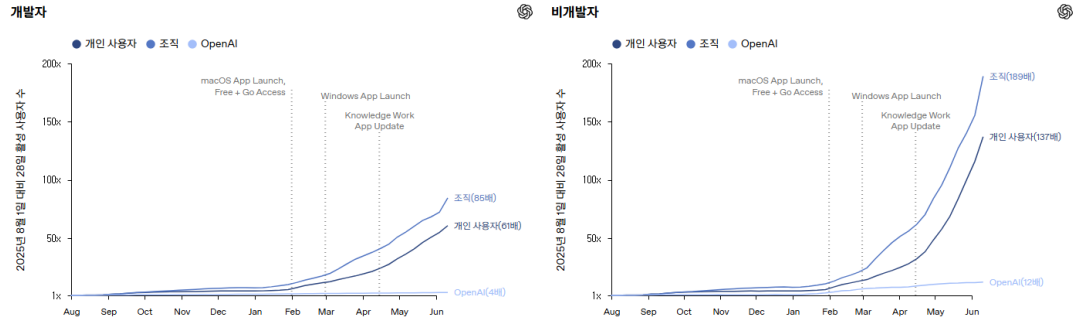
자료: OpenAI, DS투자증권 리서치센터

그림29 Codex 생성 출력 토큰 비율: 기업 63.3%, 개인 16.5%



자료: OpenAI, DS투자증권 리서치센터

그림30 개인, 기업을 중심으로 Codex 활성사용자는 빠르게 증가 중



자료: OpenAI, DS투자증권 리서치센터

Agentic AI의 발전 방향 (1) 장기 실행 Agent

장기 실행 Agent는 개별 업무를 보조하는 Copilot을 넘어 독립적인 실행 주체로 발전하는 방향

장기 실행 Agent가 단순히 Agent의 작업 수행 시간이 수시간 이상으로 늘어나는 변화에 불과해 보일 수 있으나 사용자 관점에서는 혁신에 가까운 효용이 생길 수 있다. 목표와 작업 상태를 유지한 채 복잡한 업무 과정을 장기간 자율적으로 이어갈 수 있게 되면, Agent는 개별 업무를 보조하는 Copilot을 넘어 업무를 위임받아 수행하는 독립적인 실행 주체로 발전한다. 사용자는 세부 단계마다 지시하는 대신 목표와 제약조건을 설정하고, Agent는 중간 결과와 예외 상황만 보고하며 업무를 End-to-End로 수행하게 된다.

특히 장기 실행 능력은 인간의 비근무 시간까지 생산 가능 시간으로 전환시킨다는 점에서 의미가 크다. 수면과 생활시간이 필요한 인간이 특성을 고려하면 Agent가 8시간 이상 자율적으로 업무를 수행할 수 있다는 것은 인간이 잠든 시간에도 업무가 지속될 수 있음을 의미한다. 또한 실행 시간의 확대는 단순히 더 오래 일한다는 의미에 그치지 않는다. 내부 추론 횟수를 늘려 성능을 향상시킨 Reasoning 모델의 사례처럼 업무 수행에 투입할 수 있는 시간과 연산이 늘어날수록 해결 가능한 과업의 복잡성과 범위도 함께 확대될 수 있다.

METR은 Agent가 수행할 수 있는 인간 기준 과업의 길이가 약 7개월마다 두 배씩 증가해 왔으며 현재 추세가 이어질 경우 향후 2~4년 내로 인간 기준 일주일 규모의 과업도 수행할 수 있을 것으로 전망했다. OpenAI 역시 GPT-5.1-Codex-Max가 Compaction(기존 작업 이력에서 핵심 상태만 압축해 새로운 Context Window로 넘기는 기술)을 활용해 여러 Context Window에 걸쳐 작업을 이어갈 수 있다고 설명했으며, 내부 평가에서는 24시간 이상 코드를 수정하고 테스트 오류를 해결한 사례를 제시했다.

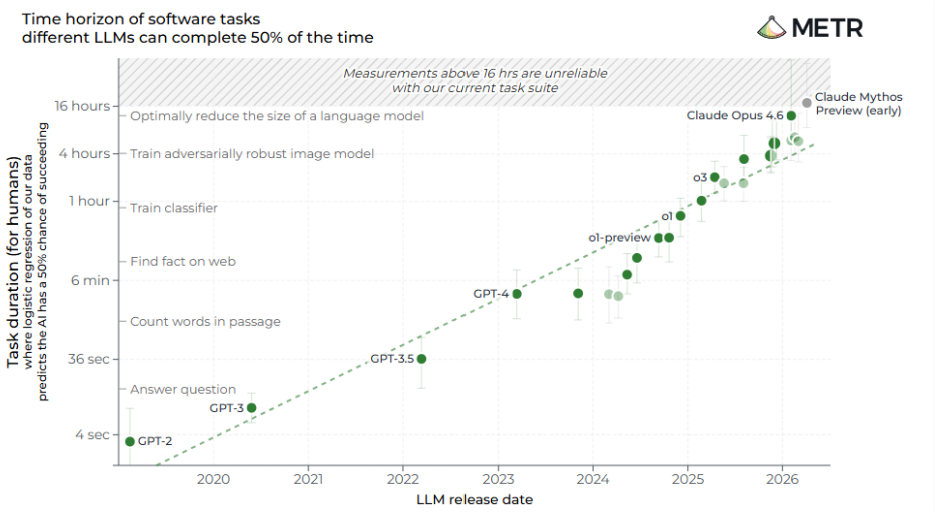
장기 실행 Agent는 컴퓨팅 수요와 AI인프라 전반의 수요에 긍정적

HW 측면에서도 장기 실행 Agent는 가장 근본적인 컴퓨팅 수요 확대 요인이다. 작업 시간과 작업 가능한 업무의 범위가 확대되면 모델 호출과 토큰 사용량이 늘어나고 Context와 KV Cache뿐 아니라 중간 결과, 체크 포인트, 오류 기록 등을 보존해야 한다.

장기 실행과 관련 있는 것은 메모리, 스토리지

장기 실행 Agent는 AI인프라 전반에도 긍정적이다. 반복 추론을 처리하는 GPU, 도구 실행을 담당하는 CPU, Worker 간 상태를 전달하는 네트워킹, 다양한 데이터의 저장을 위한 메모리, 스토리지 등 전반적인 수요가 증가할 것이다. 이 중 메모리, 스토리지의 수혜가 클 것으로 기대되는데 이는 메모리 용량 확대가 Agent의 기능적 확장으로 직접 이어지기 때문이다. GPU, CPU의 추가 투입은 연산 속도, 동시 처리량을 높이는 반면 메모리와 스토리지의 확대는 KV Cache, 작업 상태, 중간 산출물 등을 더 오래 보존해 Agent가 장기간 작업을 중단 없이 이어가고 더 넓은 범위의 업무를 처리할 수 있도록 하기 때문이다

그림31 (인간 기준) 장시간 소요되는 업무에도 50% 이상의 성공률을 보이는 LLM



자료: METR, DS투자증권 리서치센터

Agentic AI의 발전 방향 (2) Vertical Agent

Vertical Agent는 특정 산업, 직무에 특화돼 실제 과업을 완료할 수 있는 Agent

Vertical Agent는 특정 산업, 직무에 특화돼 기업의 데이터, 도구, 업무 프로세스와 연결되고 실제 과업을 완료하는 Agent를 의미한다. 기업은 다양한 업무를 준수하게 수행하는 범용 Agent보다 특정 업무를 전문적으로 수행해 생산성을 안정적으로 개선시키는 Vertical Agent에 비용을 지불할 가능성이 높다. 그리고 Agentic AI는 경제적 가치가 높고 자동화를 통해 생산성 향상이 가능한 분야에서 먼저 발전할 것으로 기대되는데, Anthropic의 기업 API 사용을 봐도 상위 10개 업무 비중이 2025년 8월 28%에서 11월 32%로 높아지는 등 소수 업무에 점차 집중되고 있다.

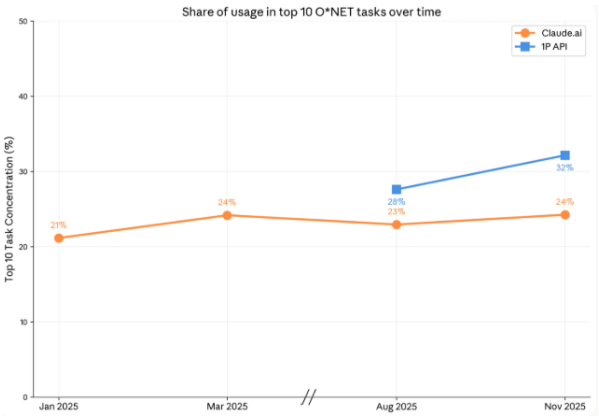
Vertical Agent는 단순히 특정 산업의 데이터로 조정된 모델이 아니라, 해당 산업의 규칙, 예외 절차 등을 이해하고 기업 데이터와 여러 업무 도구를 연결해서 특정 업무를 End-to-End로 수행하는 Agentic System이다. 범용 모델은 교체 가능하지만, 업무를 단계 별로 분해하고 단계에 맞는 도구, 권한, 검증 절차를 연결한 Workflow는 대체하기 어렵다. 따라서 경쟁력은 모델 성능뿐 아니라 업무 수행률, 결과의 신뢰성 및 추적 가능성, 기존 시스템과의 통합 수준에서 결정되며, 이는 모델 성능과 구분되는 새로운 경쟁의 축이다.

Agent가 기업의 실제 업무를 수행하는 실행 주체로 발전할수록 이를 관리하고 통제하는 운영, 보안 소프트웨어의 중요성도 높아질 것이다. 구체적으로 Agent가 권한을 받아 기업 시스템에 접속하고 실제 업무를 수행하기 시작하면 기업은 Agent가 어떤 정보에 접근했고 무엇을 실행했는지 관리해야 한다. 또한 중요 업무에는 사전 승인을 요구하고, 오류 발생 시 즉각 작업을 중단하거나 이전 상태로 되돌릴 수 있어야 한다. 이는 소프트웨어의 영역이므로 AI가 일부 기존 소프트웨어의 기능을 대체하더라도, Agent를 기업 시스템에 연결하고 안전하게 통제하는 소프트웨어의 역할은 오히려 확대될 가능성이 높다.

Agent가 기업 업무에 통합되면 클라우드의 리소스를 지속적으로 사용
→ 클라우드의 고부가 Software와 결합 시 수익성 개선

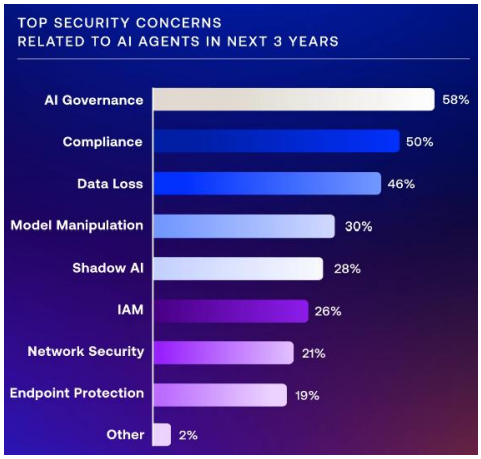
Vertical Agent가 기업의 핵심 업무로 통합되면 하이퍼스케일러에는 장기적이고 반복적인 매출로 이어질 수 있어 유리한 방향이다. Agent가 기업 데이터의 저장과 활용뿐 아니라 업무 실행, 보안, 운영 관리까지 클라우드 전반의 자원을 지속적으로 사용하기 때문이다. 이 과정에서 다른 플랫폼으로 이전하려면 데이터 연결과 업무 절차, 권한 체계를 다시 구축해야 하므로 전환비용도 함께 높아진다. 이에 더해 하이퍼스케일러가 그래왔듯이 클라우드 서비스에 고부가 소프트웨어를 결합하는 방향으로 수익성까지 향상시킬 개연성이 높다. AI CapEx의 핵심 투자 주체인 하이퍼스케일러가 이처럼 수익화 경로를 확보한다는 점은 AI 생태계 전반의 지속가능성에도 긍정적인 요소다.

그림32 Claude API 사용량은 상위 10개 업무에 집중되는 추세



자료: Anthropic, DS투자증권 리서치센터

그림33 향후 3년간 Agentic AI 보안의 중심은 거버넌스



자료: OpenAI, DS투자증권 리서치센터

표9 Amazon Bedrock AgentCore 요금제: Agentic AI의 시대에 수익화 경로를 확대 중인 하이퍼스케일러

버스/기능	리소스	요금
런타임	CPU	vCPU-시간당 0.0895 USD
	메모리	GB-시간당 0.00945 USD
브라우저 도구	CPU	vCPU-시간당 0.0895 USD
	메모리	GB-시간당 0.00945 USD
코드 인터프리터	CPU	vCPU-시간당 0.0895 USD
	메모리	GB-시간당 0.00945 USD
게이트웨이	API 간접 호출	간접 호출 1,000건당 0.005 USD
	검색 API	간접 호출 1,000건당 0.025 USD
	도구 인덱싱	인덱싱된 도구 100개당 월 0.02 USD
ID	비 AWS 리소스에 대한 토큰 또는 API 키 요청	에이전트가 요청한 토큰 또는 API 키 1,000개당 0.010 USD
메모리	단기 메모리	신규 이벤트 1,000건당 0.25 USD
	장기 메모리 스토리지	내장 메모리 전략 사용: 매월 저장된 메모리 1000개당 0.75 USD 내장 메모리 전략과 함께 오버라이드 또는 자체 관리형 메모리 전략 사용: 매월 저장된 메모리 레코드 1000개당 0.25 USD*
	장기 메모리 검색	1,000개의 메모리 레코드 검색당 0.50 USD
정책	권한 부여 요청	권한 부여 요청당 0.000025 USD
	입력 토큰 처리	입력 토큰 1,000개당 0.13 USD

자료: AWS, DS투자증권 리서치센터
주: 일부만 추출

Agentic AI의 발전 방향 (3) Multi-Agent System

Multi-Agent System의 핵심은 각자 독립된 컨텍스트에서 추론 → 관리자 Agent가 결과 통합

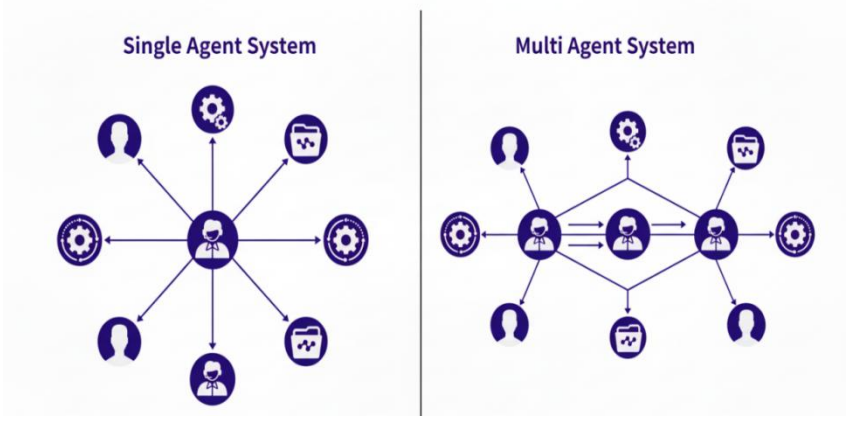
Multi-Agent System는 여러 Agent가 각자의 역할과 독립된 Context를 바탕으로 하나의 목표를 협업해 수행하는 구조이다. 단일 Agent가 하나의 Context 안에서 업무를 수행하던 방식과 달리, 대표적인 계층형 구조에서는 (1) 관리자 Agent가 목표를 하위 과업으로 나눠 전문 Agent에 배분하고 (2) 각 전문 Agent는 서로 다른 정보와 도구를 활용해 병렬로 탐색한다. 이후 결과를 통합하고 상호 검증하는 과정까지 더해지면서, AI 성능 향상의 축도 개별 모델을 넘어 역할 배분과 정보 전달, 검증 절차를 설계하는 시스템 영역으로 확장된다.

Multi-Agent System의 핵심은 각 Agent가 독립된 Context와 역할을 바탕으로 서로 다른 정보와 추론 경로를 탐색한다는 점이다. 이를 통해 하나의 Agent가 동일한 Context 안에서 모든 과정을 이어갈 때보다 다양한 접근을 병렬로 시도하고, 결과를 서로 비교 및 검증할 수 있게 된다. 더 나아가 계층형 구조에서는 관리자 Agent가 업무를 분배하고 결과를 취합해 다른 Agent에 다시 전달하면서, 하나의 정보가 여러 독립 Context에서 반복적으로 해석되고 보완된다. 따라서 Multi-Agent System은 단순히 모델 호출 횟수를 늘리는 것이 아니라, 독립적인 추론을 협업 구조로 결합해 성능을 높이는 방식이다.

각자 독립된 추론은 KV Cache의 총량을 늘림

이 구조는 우선 Agent 다수의 병렬 추론을 통해 시스템 전체의 지능 향상에 기여하지만 동시에 독립적으로 추론하기에 토큰과 메모리의 중복을 만든다. 실제로 올해 공개된 TokenDance 연구에서는 8개 Agent가 이전 라운드의 결과를 공유할 때 Agent별 KV Cache의 91~97%가 서로 유사한 것으로 나타났다. 또한 동일하게 250개의 하위 요청을 처리했을 때 Multi-Agent는 KV Cache 41.5GiB를 사용한 반면, 서로 독립적인 일반 요청은 24.8GiB를 사용했다.

표10 Multi-Agent System과 Single-Agent System 비교



자료: Codecademy, DS투자증권 리서치센터

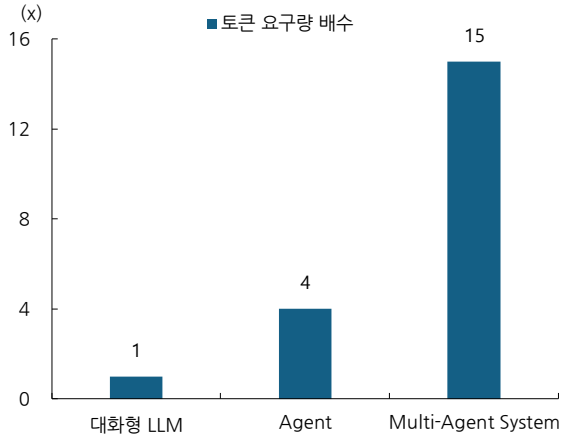
단순히 Agent의 수보다 공통 Context의 효율적 활용이 더욱 중요

이 연구는 (1) Multi-Agent의 연산 방식이 컴퓨팅은 물론 KV Cache를 증가시키는 방향이며 (2) 향후 Multi-Agent의 확장성을 결정하는 변수가 'Agent 수'가 아니라, '공통 컨텍스트를 얼마나 효율적으로 공유, 재사용할 수 있는가'임을 보여준다. 이러한 시스템이 잘 작동하지 않으면 Multi-Agent의 성능은 정체되거나 오히려 하락할 수 있음을 Google Research도 지적한 바 있다.

장기적으로 Multi-Agent는 개별 기업이나 플랫폼 내에서 미리 설정한 Agent 간 협업을 넘어서 필요에 따라 외부 전문 Agent와 협업 또는 업무를 위임하는 방향으로 발전할 가능성이 높다. Google의 Agent2Agent(A2A)는 이러한 변화를 보여주는 대표적인 사례로 서로 다른 기업의 Agent가 내부 메모리나 도구를 공개하지 않고도 자신의 기능을 알리고 업무와 진행 상태, 결과물을 주고받을 수 있게 한다.

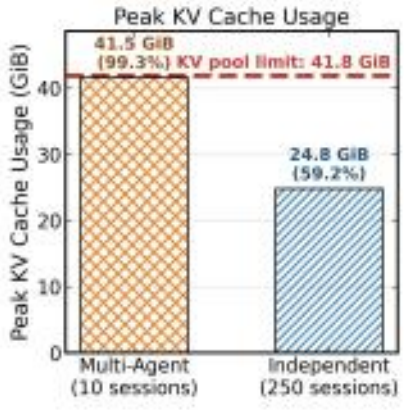
이러한 연결이 보편화되면 기업은 모든 기능별 Agent를 직접 개발하기보다 특정 산업과 업무에 특화된 전문 Agent를 선택적으로 활용할 가능성이 높기에 Vertical Agent 고도화와도 연결된다. 또한 기업 외부 Agent에 데이터, 실행 권한을 위임하기에 Agent의 신원/권한, 행동 이력을 관리하는 보안 인프라도 함께 중요해질 것이다.

그림34 Multi-Agent를 통해 토큰 생성량은 한 단계 증가



자료: Anthropic, DS투자증권 리서치센터

그림35 Multi-Agent는 KV Cache를 많이 사용하는 방식



자료: TokenDance: Scaling Multi-Agent LLM Serving via Collective KV Cache Sharing (Zhuohang Bian), DS투자증권 리서치센터

Multi-Agent System이 확산되면 (1) 메모리, 스토리지 (2) GPU, ASIC (3) Agent 보안 소프트웨어가 대표 수혜 영역이 될 수 있다. 이 중 메모리 계층의 수혜 논리는 가장 직관적이다.

KV Cache 총량 증가에 따른 메모리, 스토리지 수요 증가

[메모리, 스토리지] 앞서 기재했듯 Multi-Agent의 핵심은 Agent별로 독립된 Context이며 이를 통합, 검증하면서 높은 성능을 얻게 된다. 이는 각 Agent가 KV Cache를 별도로 저장해야 한다는 의미이며 KV Cache의 총량은 자연스럽게 늘어나 메모리, 스토리지의 수요를 강화시키는 방향이다. 그리고 메모리에 대한 수요는 서버 외부의 확장 DRAM을 풀링, 공유할 수 있는 CXL의 중요성까지 부각시킬 수 있다.

[CXL] Multi-Agent System에서는 여러 Agent가 동일한 시스템 프롬프트와 기업 문서 등 공통 정보를 반복적으로 참조하므로, 이를 서버마다 별도로 저장하면 메모리 중복이 커질 수 있다. 이때 공통 데이터와 관련 KV Cache를 CXL 기반 확장 메모리 풀에 저장하고 여러 서버가 공유하면 중복을 줄이는 동시에 필요한 메모리를 탄력적으로 배분할 수 있다. 따라서 Multi-Agent의 확산은 서버 외부 메모리를 풀링, 공유하는 CXL에게 역할을 부여할 수 있다.

복수 Agent의 병렬 작업으로 중요성이 높아질 GPU, ASIC

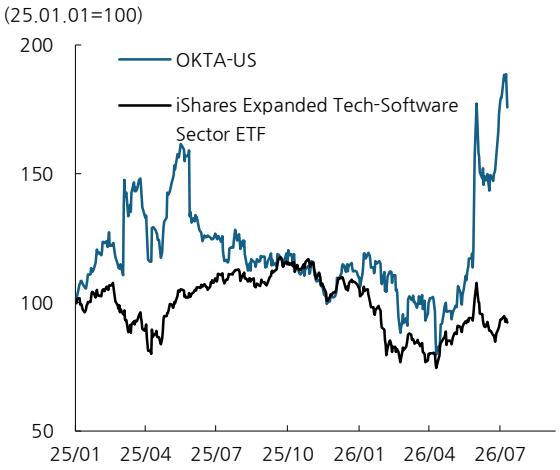
[GPU, ASIC] GPU, ASIC은 단일 요청을 다수의 Agent에 분배해 여러 병렬 작업으로 확장하는 과정에서 중요성이 높아질 수 있다. GPU는 범용성이 높고 대규모 병렬 연산에 특화돼 있어, 다수의 하위 Agent의 연산을 Batch로 묶어 병렬 처리하는 데 강점이 있기 때문이다. 이후 각 Agent의 추론 결과를 관리자, 검증 Agent가 통합하는 단계에서도 추가적인 Prefill과 Decode가 발생해 전체 가속기 수요가 늘어날 수 있다. 현재는 Multi-Agent 관련 기술과 업무 구조가 정립되는 초기 단계이므로 범용성이 높은 GPU가 시장을 주도하겠지만, 검색, 분류, 요약처럼 호출량이 많고 업무 패턴이 표준화된 하위 Agent의 작업은 비용과 전력 효율이 높은 추론 ASIC이 채택될 가능성이 있다.

Multi-Agent는 외부 Agent와도 협업하기에 신원, 권한 등을 통제하는 제3자 보안 플랫폼 수요 높아질 것

[Agent 보안 소프트웨어] Agent 보안 소프트웨어는 Multi-Agent가 사용자의 권한을 하위 Agent에 위임하고, Agent2Agent(A2A)를 통해 외부의 Agent까지 호출하게 되면 수요가 본격적으로 확대될 전망이다. 단일 Agent 환경에서는 하나의 실행 주체만 통제하면 되지만, Multi-Agent에서는 통제의 범위와 권한 위임 경로가 복잡해진다. (1) 어떤 권한을 어떤 Agent에 위임했는지 (2) 각 Agent가 어떤 데이터, 도구까지 접근했는지 (3) 최종 결과가 어느 Agent로부터 비롯됐는지 등을 지속적으로 추적해야 한다. 더 나아가 외부 Agent와 협업하면 기업의 신뢰 경계가 외부로 확장되므로 각 Agent를 독립된 비인간 신원으로 등록하고 업무별로 최소 권한과 단기 자격증명을 부여한 뒤 작업 종료 시 즉시 회수하는 등의 체계가 필수적이다.

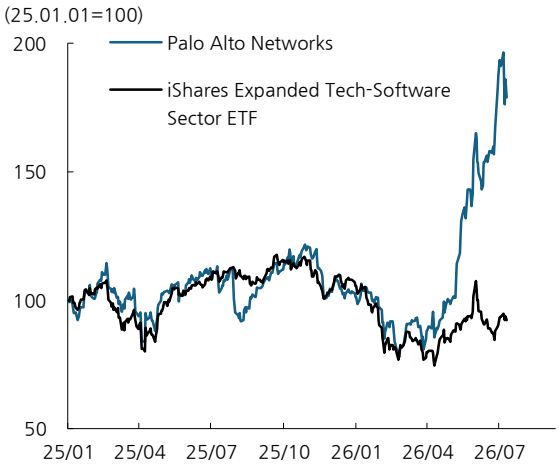
핵심은 외부 Agent까지 일관된 기준으로 통제해야 하는 만큼, 특정 클라우드나 AI 모델 내에 종속되지 않고 Agent의 신원, 권한, 행동 이력을 통합 관리하는 제3자 보안 플랫폼의 수요가 확대될 수 있다는 점이다.

그림36 최근 주가 급등한 Agent 보안 SW 관련주: Okta



자료: Bloomberg, DS투자증권 리서치센터

그림37 최근 주가 급등한 Agent 보안 SW 관련주: Palo-alto



자료: Bloomberg, DS투자증권 리서치센터

Enterprise AI가 온다

CapEx의 두번째 축, Enterprise AI

기업이 AI CapEx의 한 축으로 부상할 것

하이퍼스케일러에 뒤이은 AI CapEx의 또 다른 축은 기업이 담당하게 될 것으로 전망한다. McKinsey 조사에 따르면 기업의 88%가 특정 업무에 AI를 사용하지만 전사적으로 확장한 기업은 그 중 약 3분의 1에 불과하며 62%는 Agentic System을 도입하거나 실험하는 단계에 있다. 이는 점차 실제 수요로 반영될 예정이며 Agentic System이 고도화될수록 기업의 AI 도입도 가속화할 것이다. 실제로 Deloitte는 31개 이상의 AI서비스를 운영환경에 배치한 기업 비중이 2025년 44%에서 2028년 67%로 높아질 것으로 전망했다.

AI의 최전선에 있는 엔비디아는 지난 분기부터 DC 매출 분류를 Compute와 Networking에서 Hyperscale과 ACIE(AI Cloud, Industrial, Enterprise)로 변경했다. 이는 기존 핵심 시장인 하이퍼스케일러와 새로운 성장축인 ACIE를 구분하려는 시도로 해석할 수 있는데, 다수의 글로벌 IB는 ACIE 향 매출성장률이 하이퍼스케일러보다 높을 것으로 전망하고 있다.

보안이 중요한 기업은 온프레미스 혹은 프라이빗 클라우드를 직접 구축

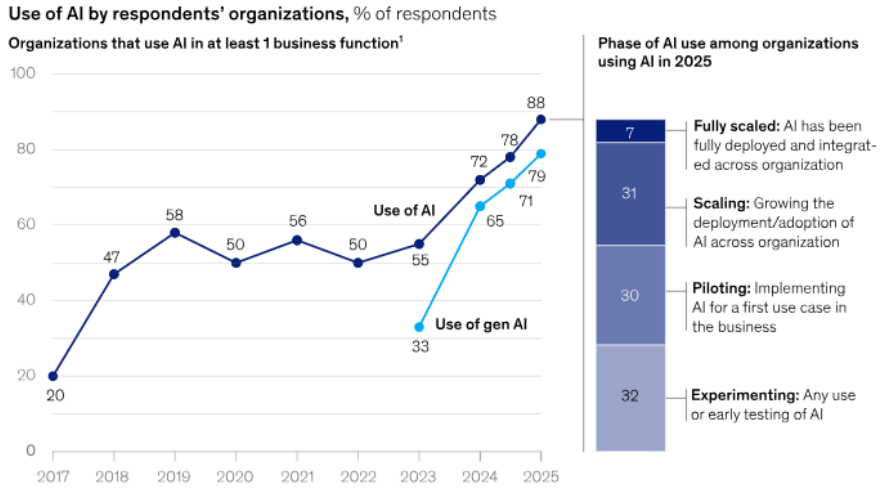
물론 기업이 모두 데이터센터를 짓는다는 의미는 아니다. IDC는 2028년 AI인프라 시장에서 클라우드 환경에 배치되는 서버가 82%를 차지할 것으로 전망했다. 이는 대부분의 AI 서비스가 여전히 하이퍼스케일러와 클라우드 사업자가 제공하는 인프라를 중심으로 운영될 가능성이 높다는 의미다. 다만 보안과 데이터 주권이 중요한 일부 글로벌 기업은 온프레미스나 프라이빗 클라우드를 직접 구축하고, 퍼블릭 클라우드와 병행하는 하이브리드 구조를 채택할 가능성도 높다.

최근 OpenAI와 Anthropic이 글로벌 컨설팅, SI 기업과 파트너십을 강화하는 현상은 Enterprise AI가 실제 구축 단계로 접어들고 있음을 보여준다. 기업이 Agentic System을 업무에 적용하려면 사내 데이터와 ERP, CRM 등 SW를 연결하고, 업무 프로세스와 접근 권한, 보안 정책까지 재설계해야 한다. 컨설팅 기업은 이 과정에서 시스템 구축과 조직 변화를 지원하는 역할을 담당한다.

보안에 철저한 삼성전자도 ChatGPT, Codex 도입

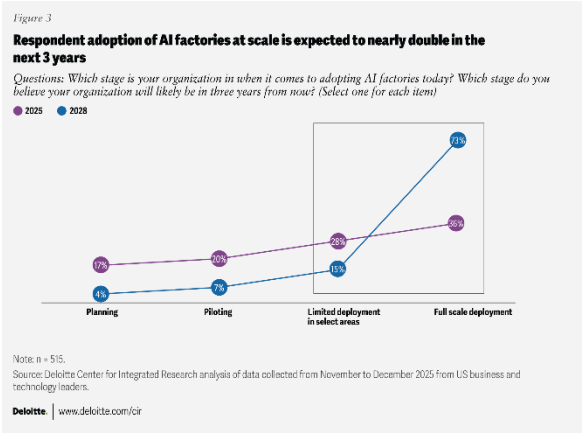
이에 더해 6월 22일, 삼성전자는 전사적 AI전환을 위해 OpenAI의 ChatGPT Enterprise와 Codex를 도입했다. 이는 OpenAI가 체결한 Enterprise AI 계약 중 최대 규모로 평가되는데, 그 규모보다 주목해야 하는 부분은 흐름이다. 2023년 보안을 이유로 ChatGPT 사용을 제한한 삼성전자가 대대적으로 전사에 AI를 도입할 수 있었던 배경에는 (1) 기업 수요에 맞춰 ChatGPT Enterprise가 보안을 강화했고 (2) 경쟁사(SK하이닉스, 마이크론)의 AI 도입 등이 언급된다. 보안에 대한 우려도 더 이상 AI도입을 지연시킬 수 없으며 한 기업이 AI도입은 경쟁사의 도입으로 점차 확산될 예정이다. Enterprise AI 확산은 멀지 않았다.

그림38 기업 88% 가 AI를 사용하지만, 전사적인 도입은 38%에 불과. 62%는 도입 중이거나 실험 단계



자료: McKinsey, DS투자증권 리서치센터

그림39 AI를 운영환경에 배치한 기업은 '25Y 44% → '28Y 67%



자료: Deloitte, DS투자증권 리서치센터

그림40 삼성전자: ChatGPT 제한 3년 만에 GPT Enterprise 도입

자료: OpenAI, DS투자증권 리서치센터

그림41 최근 OpenAI의 Enterprise AI 계약

발표일	계약 상대방	계약 내용
25년 2월	SoftBank	- ChatGPT Enterprise 및 맞춤형 기업 Agent 공급
25년 11월	Target	- ChatGPT Enterprise 공급 - OpenAI API를 고객, 매장 서비스에 적용
25년 12월	Accenture	- ChatGPT Enterprise 공급 - 기업 고객용 Agent 솔루션 공동 구축
25년 12월	Walt Disney	- ChatGPT 공급 - Sora, ChatGPT Images에 캐릭터 사용권 제공
2026년 1월	ServiceNow	- OpenAI 모델을 ServiceNow 업무 자동화 플랫폼에 통합
2026년 6월	삼성전자	- ChatGPT Enterprise와 Codex 공급

자료: OpenAI, DS투자증권 리서치센터

그림42 최근 Anthropic의 Enterprise AI 계약

발표일	계약 상대방	계약 내용
25년 10월	Salesforce	- 개발조직에 Claude Code를 공급 - Claude 모델을 Agentforce에 통합
25년 12월	Snowflake	- Claude 모델을 자체 플랫폼, Agents에 통합
25년 12월	Accenture	- 개발자에게 Claude Code를 공급 - 기업 고객용 Claude 솔루션을 공동 구축
26년 4월	NEC	- Claude, Code, Cowork 공급 - 산업별 솔루션에 Claude를 통합
26년 5월	PwC	- Claude Code, Cowork 공급 - 재무, 딜, 기업업무 솔루션을 공동 개발
26년 6월	삼성전자, 삼성SDS	- Claude, Cowork, Code를 공급해 지식업무와 소프트웨어 개발에 적용

자료: Anthropic, DS투자증권 리서치센터

발빠른 젠슨황은 이미 Enterprise AI 시장을 정조준

단순한 하드웨어 판매사가 아닌 AI 인프라 공급자를 지향하는 엔비디아

젠슨 황은 2022년 GTC에서 데이터센터를 전력과 데이터를 투입해 지능을 생산하는 'AI Factory'로 재정의한 뒤 이를 엔비디아의 핵심 사업 개념으로 강조해왔다. 이는 엔비디아가 개별 칩과 랙을 판매하는 회사를 넘어 전체 AI인프라 공급자를 지향한다는 의미로 해석된다.

이러한 전략은 기업 시장으로 확장되고 있다. 엔비디아는 2025년 기업용 AI Factory 검증 설계(Enterprise AI Factory Validated Design)를 공개해 기업이 검증된 표준 설계를 기반으로 자체 AI인프라를 구축할 수 있도록 했으며, 2026년에는 데이터센터 매출을 Hyperscale과 ACIE로 재분류했다. Enterprise AI를 새로운 AI인프라의 수요처로 판단하고 있음이 읽히는 대목이다.

Agentic AI는 단순히 질문에 답변하는 정도를 넘어서 실제 업무의 수행을 목표로 한다. 이는 Agentic AI가 고도화될수록 수행 가능한 업무의 영역 확대로 이어져 Enterprise AI 확산을 가속화시킬 것임을 의미하는데, 대부분의 기업은 클라우드 인프라를 임대하겠지만 보안, 데이터주권, 최적화를 목적으로 자체 AI Factory를 구축하려는 수요 또한 존재할 것이다. 특히 ASIC과 네트워킹 솔루션 등을 활용하려는 하이퍼스케일러보다 일반 기업이 오히려 엔비디아의 표준화된 'AI Factory'를 필요로할 가능성이 있다.

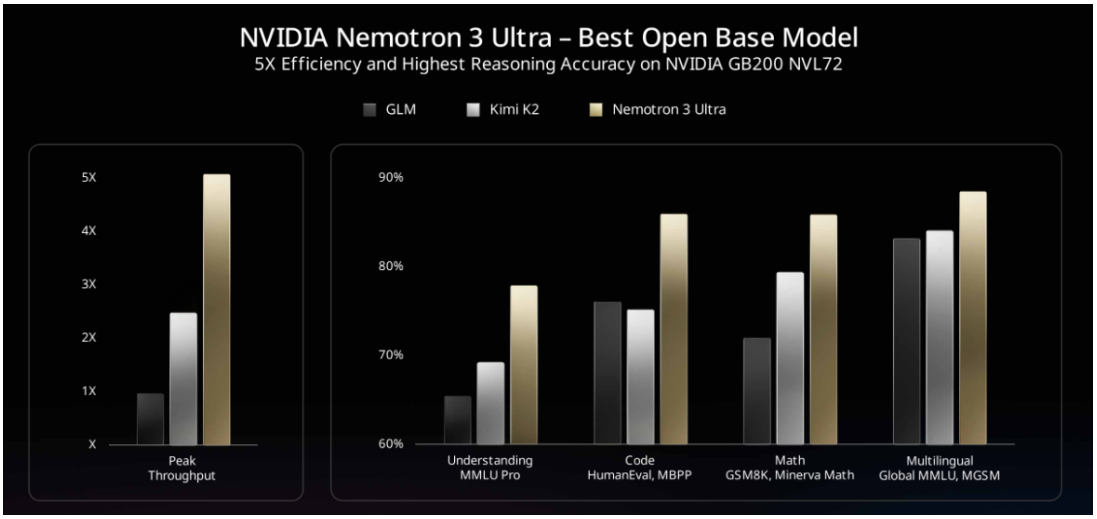
엔비디아의 자체 오픈모델 Nemotron도 Enterprise AI를 준비하는 포석일 수 있음

최근 엔비디아가 자체 오픈소스 모델 Nemotron의 성능을 강조하는 것도 기업용 AI Factory 전략의 일환으로 볼 수 있다. 기업이 자체 AI 인프라를 구축하려면 HW뿐 아니라 내부 데이터와 정책에 맞게 조정할 수 있는 모델도 필요하기 때문이다. Nemotron은 외부 모델 API에 데이터를 전적으로 의존하지 않고 자체 Agent를 구축할 수 있는 선택지를 제공하며, Agent Workload의 추론 효율과 처리량을 높여 운영비용을 낮추는 역할도 맡는다. 엔비디아는 강점을 가진 하드웨어에 모델과 운영 소프트웨어를 결합해 기업용 AI 인프라 전반을 선점하려는 것으로 보인다.

네오클라우드도 Enterprise AI의 수혜

네오클라우드도 자체 AI Factory를 구축할 CapEx나 데이터센터 운영 역량이 부족한 기업에게 대안이 될 수 있다는 점에서 Enterprise AI의 수혜를 받는다. 기업은 코어위브, 네비우스 등 네오클라우드에서 엔비디아 GPU 기반 컴퓨팅을 임대하고, Nemotron과 같은 오픈모델을 자사 데이터와 업무에 맞게 배포함으로써 대규모 선행투자 없이 Enterprise AI를 운영할 수 있다. 엔비디아는 네오클라우드 기업에 지분을 투자하는 동시에 최신 GPU를 제공하고 보증 등을 통해 자본 조달도 지원하면서 사실상 생태계를 만들어왔는데, 이는 네오클라우드를 엔비디아의 최신 HW와 SW 생태계를 시장에 확산하는 AI-native 클라우드 채널로 육성해온 것으로 보인다.

그림43 NVIDIA Nemotron3 Ultra - "Best Open Base Model"(GTC 2026)



자료: Nvidia, DS투자증권 리서치센터

그림44 Vera Rubin NVL72를 가장 먼저 배치한 코어위브

자료: Barron's, DS투자증권 리서치센터

표11 엔비디아의 코어위브, 네비우스 투자

투자 대상	발표일	투자금액
CoreWeave	2023-04-20	$100.0M
	2025-03-27	$250.0M
	2026-01-26	$2.0B
Nebius	2024-12-02	$25.0M
	2026-03-11	$2.0B

자료: SEC, DS투자증권 리서치센터
주: 2026년 3월 11일 Nebius 투자는 주식매수권 취득 계약

Enterprise AI의 수혜는? 하이퍼스케일러, 네오클라우드

Enterprise AI 본격화되면 하이퍼스케일러 수혜가 가장 직관적

Enterprise AI가 확산되면 기업의 AI활용이 늘어난다는 점에서 컴퓨팅 수요의 총량은 증가하게 된다. 이에 대응하기 위한 데이터센터 투자도 자연스럽게 증가할 것이며, AI반도체, 메모리, 네트워크, 전력 등 AI인프라 전반에 수혜가 있을 것이다. 더 나아가 기업이 Enterprise AI에 필요한 컴퓨팅을 어떤 방식으로 확보하는지에 따라 수혜 받는 기업이 달라질 수 있다.

기업의 컴퓨팅 확보 방식에는 크게 세 가지가 있다. (1) 하이퍼스케일러의 퍼블릭 클라우드를 이용하거나 (2) 네오클라우드의 GPU 컴퓨팅을 임대하거나 (3) 온프레미스 방식으로 자체 AI Factory를 구축할 수 있다. 네오클라우드는 전용 인프라를 구축할 수 있다는 측면에서, 온프레미스는 데이터 주권, 보안, 운영 통제의 측면에서 중장기적으로 수요가 확대될 개연성이 있다. 다만 높은 구축 비용과 아직 기술적 표준이 완전히 정립되지 않은 AI인프라의 상황을 고려하면 다수의 기업은 여전히 퍼블릭 클라우드를 주력으로 이용할 것이다. 그렇기에 Enterprise AI 초기에 가장 직관적인 수혜처는 하이퍼스케일러 기업이다.

하이퍼스케일러는 Enterprise AI를 계기로 클라우드 사업의 부활을 기대할 수 있다. 기존 클라우드는 범용 CPU 연산에 스토리지, 데이터베이스, 네트워크, 보안과 개발 환경을 통합하고 SaaS를 결합해 약 60~70%의 매출총이익률을 기록하는 사업이었다. 반면 AI 클라우드는 주요 사업이 GPU 기반 학습 및 추론 인프라 공급이기에 사업의 경쟁력이 고사양 GPU와 전력 확보로 이동했다. 오라클이 AI 클라우드의 매출총이익률을 30% 초반대로 언급한 점을 고려하면, 현재 AI클라우드에서 하이퍼스케일러는 본연의 경쟁력을 살리지 못할 뿐 아니라 수익성도 낮아진 셈이다.

Enterprise AI 시대에서는 AI 클라우드 경쟁의 기준도 풀스택 역량으로 확장 이는 하이퍼스케일러가 잘하는 영역

Enterprise AI가 확대되면 AI클라우드의 경쟁 기준은 최신 GPU 확보를 넘어 Agent 개발 및 배포 환경, 기업 데이터 연결, 보안 및 거버넌스를 포괄하는 풀스택 역량으로 넓어질 전망이다. 기업용 AI는 모델 성능뿐 아니라 기존 업무 시스템과의 통합, 안정적인 운영과 통제가 중요하기 때문이다. 그리고 이는 하이퍼스케일러가 그동안 축적해온 SW역량과 고객 기반을 통해 경쟁력을 발휘할 수 있는 영역이다. 더 나아가 고부가가치 SaaS를 함께 제공하면 고객당 매출과 수익성도 높일 수 있다. 구글과 마이크로소프트가 이미 기업용 Agent 플랫폼을 확대하는 것도 이러한 경쟁 변화에 대응하기 위한 움직임이다. 하이퍼스케일러의 시간이 돌아오고 있다.

네오클라우드의 장점

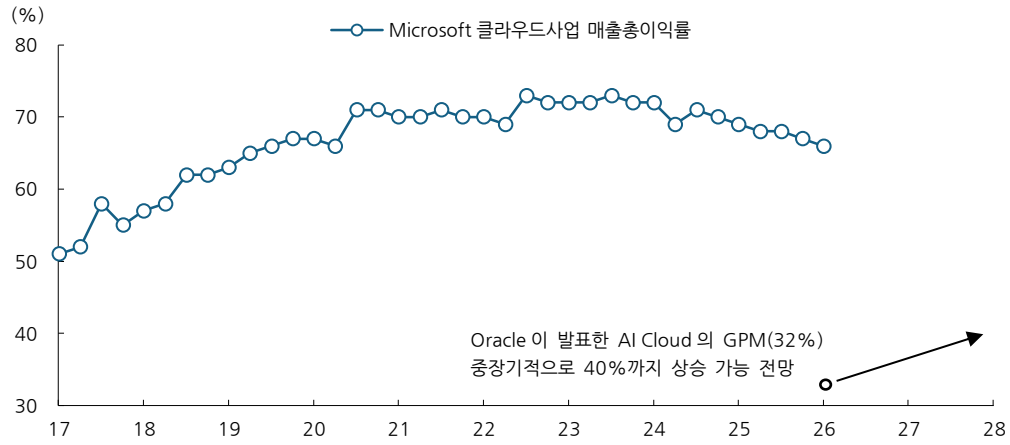
1) 최신 GPU 클러스터를 빠르게 확보 가능

2) 유연성과 가성비가 우수

네오클라우드 역시 수혜처가 될 수 있다. 네오클라우드는 (1) 기업이 대규모 투자 없이 연간 임대비용으로 최신 GPU 클러스터를 빠르게 확보하고 (2) 표준화된 하이퍼스케일러 클라우드 대비 특정 AI 워크로드에 맞춰 네트워크, 스토리지, 운영환경을 구성할 수 있다는 점이 강점이다. 또한 대규모 학습이나 고성능 추론처럼 GPU 활용률과 클러스터 성능이 중요한 워크로드에서는 가격 대비 성능이 우수할 수 있다. CoreWeave와 Nebius도 AI에 특화된 전용 인프라와 운영도구를 핵심 경쟁력으로 내세우고 있다.

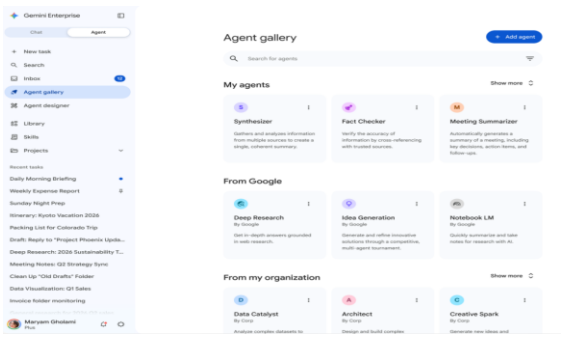
주요 고객은 자체 데이터센터를 구축하기 어려운 일반 기업에 한정되지 않는다. AI 모델 개발사와 소프트웨어 기업은 초기 CapEx와 인프라 운영 부담을 줄이기 위해 네오클라우드를 활용할 수 있고 자체 인프라를 보유한 하이퍼스케일러도 대규모 투자 없이 연간 임대 비용만으로 컴퓨팅 수요를 보완하거나 최신 GPU를 신속하게 확보하기 위해 이를 이용할 수 있다. Meta와 Nebius의 장기 인프라 계약은 네오클라우드가 단순히 일반 기업용 대체재가 아니라, 빅테크의 자체 인프라를 보완하는 탄력적 공급자로도 기능할 수 있음을 보여준다.

그림45 AI투자 및 AI클라우드의 상대적으로 낮은 마진으로 매출총이익률이 하락중인 Microsoft



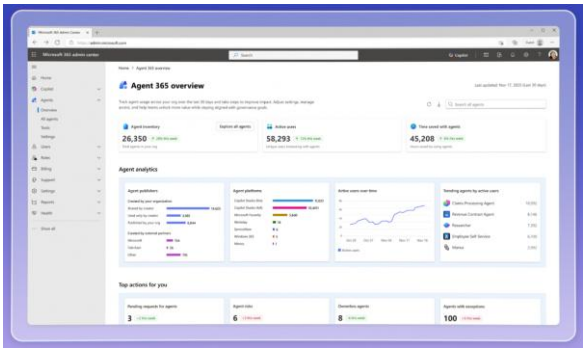
자료: Microsoft, Oracle, DS투자증권 리서치센터

그림46 기업용 AI Agent 시대를 준비해온 구글



자료: Google, DS투자증권 리서치센터

그림47 기업용 AI Agent 시대를 준비해온 마이크로소프트



자료: Microsoft, DS투자증권 리서치센터

그래서 뭐사요?

Show Me the Compute

컴퓨팅 없이는 AI도 없다

AI는 늘 컴퓨팅을 증가시키며 발전해옴

2022년 11월 ChatGPT의 등장을 AI산업 및 모델 발전이 본격적으로 가속화된 계기로 본다면, 지금과 같은 속도의 발전이 시작된 지는 아직 4년도 채 되지 않았다. AI 모델은 여전히 빠르게 진화하는 단계에 있으며, 성능, 비용, 전력효율 등 다양한 기준에서 최적점을 찾기 위해 아키텍처와 연산 방식이 무궁무진하게 변화하고 있다. 이번 장에서는 AI와 AI모델의 발전 방향에 근거해서 수혜가 전망되는 산업을 선정했다.

AI 모델의 성능을 높이는 방식은 여전히 더 많은 컴퓨팅을 투입하는 방향으로 전개되고 있다. 사전학습과 사후학습에 이어 더 많은 추론을 통해 답변의 정확도를 높이는 Test-time Scaling, 하나의 요청에 대해 계획, 도구 호출, 검증을 위한 복수의 모델 호출로 확장하는 Agentic System, 더 나아가 여러 Agent가 동시에 추론하고 결과를 통합하는 Multi-Agent까지 AI모델의 발전 방향을 보면 컴퓨팅 감소를 전망하기 어렵다. 이에 더해 Enterprise AI의 확산도 컴퓨팅을 크게 증가시킬 요인이다. 물론 연산 효율을 높이기 위한 기술 개발도 활발하지만, 이를 통해 절감된 비용이 더 복잡한 추론과 더 많은 업무 처리에 재투입되면서 총 컴퓨팅 수요를 낮추는 움직임은 아직 뚜렷하게 나타나지 않고 있다.

높은 컴퓨팅 수요의 단서 (1) GPU 임대 가격 상승

컴퓨팅에 대한 강한 수요는 AWS의 GPU 예약형 임대가격 인상에서도 확인된다. AWS는 2026년 7월 특정 기간의 GPU 용량을 미리 확보하는 계약의 임대 가격을 약 20% 인상했는데 인상 대상에는 B300, B200등 최신 제품 외에도 출시 후 상당 시간이 지난 H100(2022년 출시), A100(2020년 출시)도 포함됐다. 이는 일정 기간 대규모 GPU 용량을 보장받으려는 고객 수요가 공급을 웃돌고 있음을 보여준다.

높은 컴퓨팅 수요의 단서 (2) 최근 네오클라우드 컴퓨팅 임대 가격이 좋아짐

네오클라우드 사업을 확대 중인 SpaceX(xAI부문)의 컴퓨팅 계약에서도 강한 수요를 확인할 수 있다. SpaceX 체결한 전체 컴퓨팅 공급 계약은 완전 가동 기준 월 약 23.2억달러, 연환산 약 278억달러에 달한다. Colossus 1의 300MW와 Colossus 2의 1.2GW 목표용량을 합친 1.5GW를 적용하면 연간 계약금액은 1GW당 약 186억달러인데, 이는 2025년 9월에 발표된 Nebius-Microsoft 계약 금액(300MW 규모, 5년간 최대 194억달러)보다 약 44% 높은 수준이다.

물론 제공하는 데이터센터의 GPU 구성이 다르고, Colossus와 같은 고용량 데이터센터일수록 계약 금액이 큰 경향이 있어 단순 수치만 비교하기에 한계가 있다.

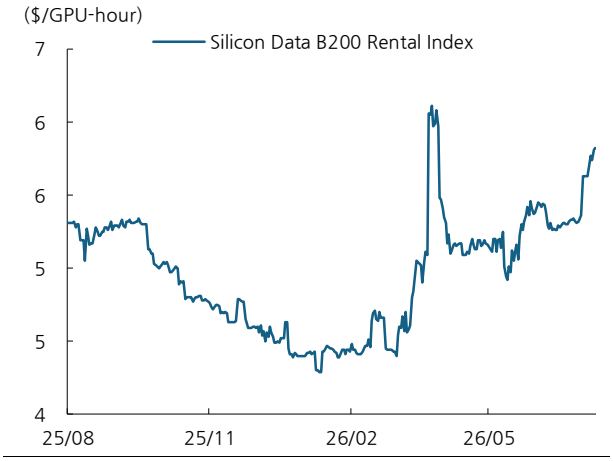
그럼에도 즉시 사용할 수 있는 대규모 최신 GPU 클러스터에는 상당한 프리미엄이 형성되고 있는 것으로 판단된다.

DC 건설 지연 이슈는 이미 구축된 데이터센터의 가치를 더욱 높이는 요소

컴퓨팅 수요가 강한 가운데 공급 확대 속도는 오히려 데이터센터 건설 지연에 제약받고 있다. 전력망 연결 및 전력기기 확보의 어려움, 전문 인력 부족, 지역사회의 반발 등의 이유로 데이터센터 건설 지연 소식은 꾸준히 생겨나고 있는데 이러한 상황에서 이미 전력과 냉각 인프라를 확보한 데이터센터의 희소성은 높아지고 있다.

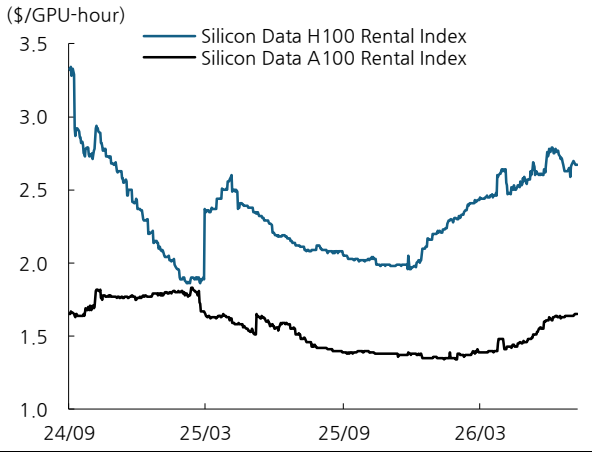
컴퓨팅에 대한 높은 수요는 하이퍼스케일러가 투자를 확대할 유인이 되기에 AI인프라 전반에 긍정적인데, 직접적인 수혜처에는 컴퓨팅을 직접 제공하는 네오클라우드가 있다. 이에 더해 엔비디아 차세대 AI가속기 Vera Rubin의 고객 배치가 하반기 본격화되면서 코어위브 등 이를 조기에 배치 받고 구동과 검증을 조기에 완료한 네오클라우드는 기존 계약잔고를 더 빠르게 매출로 전환하고, 희소한 최신 컴퓨팅에 프리미엄을 적용할 가능성이 있어 상대적인 수혜 강도가 강할 것으로 기대된다.

그림48 엔비디아 B200 임대가격 지수



자료: Bloomberg, DS투자증권 리서치센터
주: B200은 2024년 하반기 출시된 GPU

그림49 엔비디아 H100, A100 임대가격 지수



자료: Bloomberg, DS투자증권 리서치센터
주: A100/H100은 각각 2020년 상반기, 2022년 하반기 출시된 GPU

표12 스페이스X xAI부문 컴퓨팅 계약 단가: $18.56B/GW

항목	금액/용량	비고
월 계약금액 합계	$2.32B	- Anthropic: $1.25B - Google: $0.92B - Reflection AI: $0.15B
연 계약 금액 합계	**$27.84B**	
컴퓨팅 용량(GW)	1.5	- Colossus 1: 0.3 - Colossus 2: 1.2
1GW 당 컴퓨팅	**$18.56B**	

자료: 언론 종합, DS투자증권 리서치센터

표13 Nebius-Microsoft 컴퓨팅 계약 단가: $12.93B/GW

항목	계약금액	비고
5년 계약 금액	$17.4B	
5년 총 계약 금액 (옵션 포함)	$19.4B	
연 계약 금액 합계	$3.88B	
컴퓨팅 용량(GW)	0.3	
1GW 당 컴퓨팅	**$12.93B**	

자료: 언론 종합, DS투자증권 리서치센터
주: 2025년 9월 8일 발표된 계약

추론의 시대, 대장은 메모리

메모리를 덜 쓰기에는 너무나도 많은 KV Cache

Agent도 Reasoning 방식으로 연산

Reasoning모델의 핵심은 추론 단계에서 더 많은 연산과 토큰을 투입해 답변의 정확도를 높이는 'Test-time Scaling'이다. 또한 Agent가 복잡한 목표를 수행하려면 각 단계에서 판단과 재계획, 결과 검증을 반복해야 하므로 Agent와 Multi-Agent가 주류가 되어도 Reasoning은 지능과 업무 수행 성능을 결정하는 핵심 엔진으로 남을 가능성이 높다.

Reasoning 모델 → KV Cache 증가 → 메모리 용량과 대역폭 수요 증가

주류 LLM이 Reasoning 방식을 도입하면서 나타난 핵심 변화는 추론 단계에서 생성, 처리되는 토큰 수가 늘어나고, 길어진 컨텍스트를 유지하기 위한 KV Cache도 함께 증가한다는 점이다. 그리고 늘어난 KV Cache는 계속해서 기재했듯이 이를 저장할 메모리 수요로 연결된다. 더 나아가 비용 효율화라는 목표를 달성하기 위한 방법으로 KV Cache 재사용이 핵심 수단이 되었고, Coding Agent 등 Agentic AI에 들어서 반복되는 프롬프트, 코드가 늘어남에 따라 재사용의 이점이 증가하게 된다. 또한 Multi-Agent System은 에이전트 별 독립된 컨텍스트로 추론하기에 각각의 KV Cache를 저장해야 하기에 또 한번 KV Cache의 총량은 늘어나게 된다. 향후 모델이 더 긴 내부 추론과 반복 검증을 통해 성능을 높이는 방향이 이어진다면 KV Cache의 절대 규모와 이를 효율적으로 저장, 재사용하기 위한 메모리 수요의 감소를 예상하기 어렵다.

메모리 수요가 낮아지려면 AI모델이 KV Cache를 덜 쓰는 방향으로 변화해야 됨

아직 유의미한 변화는 감지되지 않음

이런 맥락에서 향후 메모리 수요를 예상하는 하나의 시각으로 AI모델의 변화를 추적하는 것도 중요하다. 메모리 수요가 구조적으로 둔화되려면 (1) Test-time Scaling을 통한 성능 향상의 한계가 뚜렷하거나 (2) TurboQuant와 같은 압축 기술이 KV Cache의 단위당 용량을 크게 낮추거나 (3) KV Cache 의존도가 낮은 dLLM, Delta Attention 등의 아키텍처가 주류로 자리 잡아야 한다. 반면 모델의 추론량, 컨텍스트와 Batch의 크기, KV Cache 재사용 수요, 에이전트의 수 등이 늘어나는 한 메모리 수요는 높은 수준을 유지할 가능성이 크고 현재의 상황이 그렇다. 이는 밸류에이션과 별개로 메모리 업체의 높은 이익을 뒷받침하는 구조적 수요 기반이 당분간 이어질 수 있음을 의미한다.

Cache를 DRAM, NAND로 이동시키는 계층화 확대 중

메모리 내에서도 계층별 수혜 강도는 차별화될 전망이다. 생성 중인 활성 KV Cache는 낮은 지연시간과 높은 대역폭이 필요해 HBM에 저장되지만, KV Cache의 총량이 빠르게 증가하면 이를 모두 HBM에 유지하기 어려워지면서 재사용 가능성이 낮거나 당장 사용하지 않는 Cache를 DRAM, NAND로 이동시키는 계층화가 확대될 수 있다.

HDD 역시 장기 데이터와 Agent의 결과물을 저장하는 용도로는 활용될 수 있지만, 현재의 대역폭과 지연시간을 고려하면 실시간 KV Cache 재사용 계층으로 사용하기에는 한계가 있다.

NAND는 AI 추론용 메모리 계층의 일부로서 재평가 받을 것

메모리 중에서는 DRAM, NAND 모두 수혜가 전망된다. 우선 NAND는 KV Cache의 증가로 AI추론용 메모리 계층의 일부로 편입 중이다. 이에 더해 NAND의 수혜 강도를 결정하는 핵심은 'Flash의 낮은 데이터 전송 속도를 고속 네트워크와 DPU, 데이터 사전 적재 기술로 얼마나 보완할 수 있는가'인데, 엔비디아의 CMX랙이 그 가능성을 보여줬다. CMX랙은 네트워킹과 소프트웨어를 통해 필요한 KV Cache를 미리 HBM으로 보내는 방법으로 낮은 속도를 보완했는데, 이 방법은 Coding Agent와 같이 반복 작업으로 필요한 KV Cache에 대한 예측 가능성이 높은 경우 더욱 효과적이다.

DRAM 역시 수혜 예상 다만 추가 마진 확대가 가능할 지 의문

DRAM 수요에 대해서도 긍정적으로 보고 있다. 우선 KV Cache 저장 구조에서 저비용 확장계층인 NAND SSD와 달리 DRAM은 HBM과 스토리지 사이에서 Cache를 빠르게 공급하는 핵심 중간 계층으로 추론 성능과의 연관성이 더 직접적이다. 또한 GPU에 HBM 탑재량이 증가하거나 HBM를 탑재한 ASIC 등 AI가속기 출하가 증가하면, HBM은 동일 용량의 DRAM보다 많은 웨이퍼를 소모하므로 DRAM의 공급을 제약해 가격의 하방을 방어하는 요인이 된다. 물론 DRAM 매출총이익률이 이미 80% 후반대로 추정되기에 추가 마진 확대가 가능할 지 의문이 있지만 KV Cache가 계속해서 증가하는 방향으로 AI모델이 발전한다면 현재 마진이 높더라도 급격한 하락을 예상하기는 어렵다.

앞서 설명한 Mamba, dLLM 그리고 Kimi K3의 KDA 방식까지 메모리 가격이 높다보니 메모리를 덜 사용하는 방향으로 발전해가려는 모델 단의 움직임은 분명 존재하고 이에 대한 노이즈는 계속 있을 것으로 생각한다. 그럼에도 여전히 주류 모델 아키텍처는 KV Cache를 누적해가는 방식이기에 메모리 수요의 급격한 둔화를 생각하기 어렵고, HBM 탑재량 증가가 DRAM의 공급을 제약한다는 점에서 여전히 메모리를 긍정적으로 본다.

CXL은 KV Cache 폭증의 대안이 될 수 있을까?

DRAM을 CPU의 제한된 메모리 채널에서 분리해 서버 외부로 확장하는 CXL에 관심 높아짐

늘어나는 KV Cache에 따른 메모리 수요를 해결하는 방식으로 CXL도 언급된다. 앞서 NAND가 사용 빈도가 낮은 KV Cache를 대용량 Storage로 이동시키는 접근이라면, CXL은 PCIe의 물리 인터페이스를 기반으로 CPU와 메모리, 가속기를 연결해 외부 DRAM을 일반 메모리에 가깝게 사용할 수 있도록 하는 기술이다. HBM과 로컬 DRAM만으로 늘어나는 KV Cache 저장에 필요한 용량을 감당하기 어려워지면서, CPU의 메모리 채널과 서버 내 메모리 슬롯 한계를 넘어 DRAM 용량을 확장하는 수단으로 주목받고 있다.

CXL의 핵심은 CPU에 직접 연결된 메모리의 물리적 한계를 완화하는 데 있다. CXL Controller는 서버 외부의 DRAM을 CPU의 메모리 주소 공간에 연결해 시스템 메모리의 한 계층으로 활용할 수 있도록 한다. 이를 통해 로컬 DRAM의 용량을 보완하고, AI 추론에 필요한 KV Cache와 모델 데이터를 저장하는 확장 메모리 계층을 구축할 수 있다. 나아가 CXL 규격이 고도화되면서 개별 서버의 Memory Expansion을 넘어, 여러 서버가 하나의 DRAM Pool에서 필요한 용량을 할당받는 Memory Pooling(CXL 2.0)과 동일한 메모리 영역에 접근하는 Memory Sharing (CXL 3.0에서 강화)으로 활용 범위가 확대되고 있다.

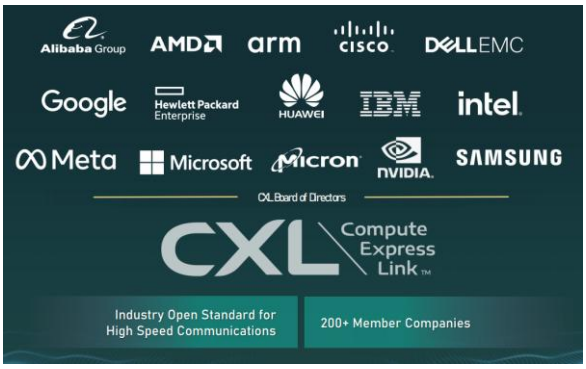
CXL의 발전 방향은 세대별 기능 변화를 살펴보면 더욱 선명하게 드러난다. CXL 1.1은 CPU에 직접 연결되지 않은 메모리를 추가하는 데 초점을 맞췄고, CXL 2.0은 Switch를 통해 메모리 자원을 여러 Host에 필요에 따라 할당하는 Memory Pooling을 도입했다. 다만 CXL 2.0은 하나의 스위치 아래에 Host와 메모리를 연결하는 구조에 머물러, 포트 수에 따라 연결 규모가 제한됐고 여러 스위치를 연결해 CPU, 가속기, 메모리를 유연하게 조합하기도 어려웠다.

CXL 3.0: 최대 4,096개의 CPU, AI가속기, 메모리 노드를 하나의 Fabric으로 연결

CXL 3.0은 이러한 한계를 크게 완화했다. 전송속도를 64GT/s로 높이는 동시에 스위치 아래에 다른 스위치를 연결할 수 있는 다단계 Switching, 비트리형(Non-Tree) 연결 구조, 정확한 데이터 전달을 위한 Port Based Routing을 도입해 최대 4,096개의 CPU, AI가속기, 메모리 Node를 하나의 Fabric으로 연결할 수 있도록 했다. 또한 CXL 2.0의 Pooling이 메모리 용량을 Host별로 나눠 할당하는 방식이었다면, CXL 3.0은 여러 Host가 동일한 메모리 영역에 접근하면서 데이터 일관성을 유지하는 Memory Sharing까지 지원한다. 이에 따라 CXL 3.0은 단순한 메모리 확장 규격을 넘어 CPU, 가속기, 메모리를 필요에 따라 조합하고 재배치하는 랙 단위 인프라의 기반으로 평가된다.

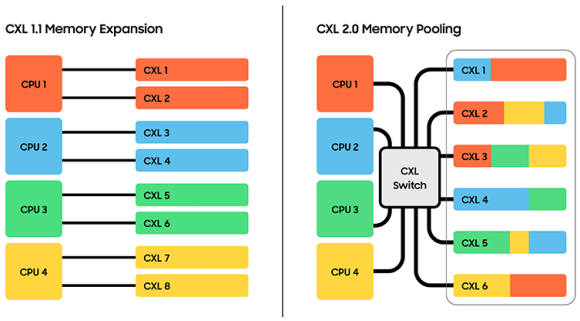
CXL DRAM은 새로운 종류의 메모리라기보다 CPU에 직접 연결된 로컬 DRAM의 용량 한계를 보완하는 외부 확장 계층으로 기능할 수 있으며, 메모리 계층상으로는 로컬 DRAM과 SSD 사이의 중간 계층에 위치한다. 접근 빈도가 높고 지연시간에 민감한 활성 데이터는 HBM과 로컬 DRAM에 유지하고, 당장 자주 사용되지는 않지만 SSD로 이동시키기에는 재호출 비용이 큰 데이터는 CXL 메모리에 배치하는 방식이다. NAND SSD가 대용량 데이터를 저장하는 Storage인 반면, CXL DRAM은 CPU의 메모리 주소 공간에 연결돼 일반 DRAM에 가까운 방식으로 활용할 수 있다.

그림50 CXL 컨소시엄 이사회 멤버(15개 기업)



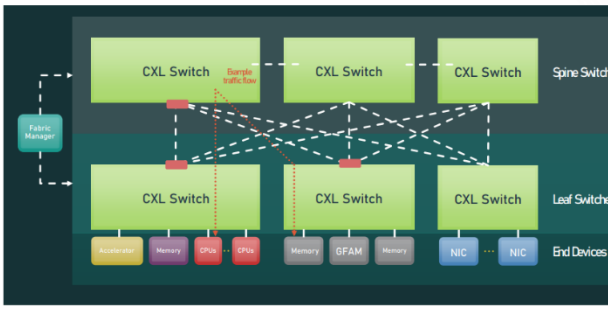
자료: CXL, DS투자증권 리서치센터

그림51 CXL 1.1 Memory Expansion / CXL 2.0 Memory Pooling



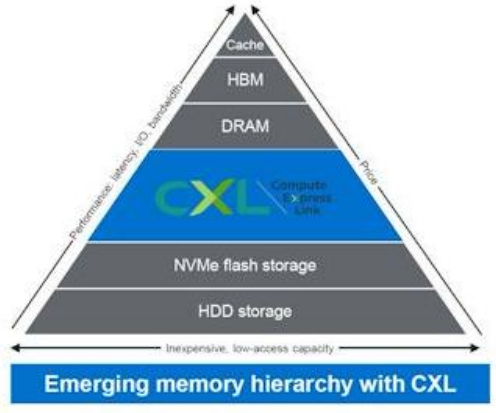
자료: 삼성전자, DS투자증권 리서치센터

그림52 CXL 3.0 구조도



자료: CXL, DS투자증권 리서치센터

그림53 CXL이 포함된 메모리 계층도



자료: Marvell Technology, DS투자증권 리서치센터

KV Cache의 저장소 혹은 호출 빈도가 낮은 Expert 저장에 활용하는 방안이 거론되는 중

AI 추론에서 CXL 메모리의 주요 활용처로는 KV Cache와 MoE 모델의 Expert 가중치가 거론된다. 긴 컨텍스트와 Multi-Agent 확산으로 KV Cache가 HBM과 로컬 DRAM의 용량을 넘어설 경우, 현재 생성에 필요한 활성 Cache는 HBM에 유지하고 접근 빈도가 낮아진 과거 토큰의 KV Cache는 CXL 메모리로 이동시켰다가 필요할 때 다시 불러오는 것이다. MoE 모델에서도 자주 호출되는 Expert는 HBM에 배치하고 호출 빈도가 낮은 Expert의 가중치는 CXL 메모리에 보관하는 계층화가 가능하다. 다만 CXL메모리는 HBM, 로컬DRAM보다 대역폭이 낮고 지연시간이 길기 때문에 실시간으로 자주 접근하는 데이터보다는 사용 빈도가 낮은 데이터를 보관하는 용도에 적합하다.

올해 7월, CXL기술이 활용된 사례가 있었는데 Meta는 자체 개발한 CXL 2.0 메모리 확장 ASIC 'Vistara'를 통해 신형 DDR5 서버에 퇴역 서버의 DDR4를 연결하는 구조를 실제 생산 환경에 적용했다. Meta 서버의 43.7%가 연산보다 메모리 용량에 제약 받고 있었으며, CXL로 서버당 256GB의 DRAM을 추가한 결과 일부 추론 작업에서는 필요한 서버 수를 최대 25% 줄일 수 있었다. 이는 CXL이 단순한 연구 기술을 넘어 메모리 부족으로 남는 컴퓨팅 자원을 되살리는 경제적 수단이 될 수 있음을 보여준 사례이다.

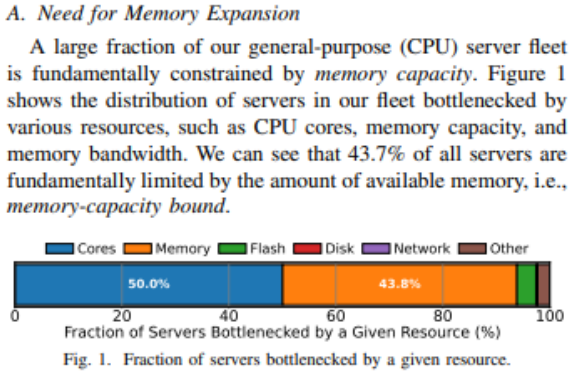
CXL 상용화를 대비해서 관련 기업들은 꾸준히 제품을 출시하고 있다. Astera Labs의 CXL 2.0 기반 Leo Controller는 마이크로소프트 Azure M-Series VM의 비공개 프리뷰에 적용됐으며, Controller당 최대 2TB의 메모리와 1.5배 이상의 서버 메모리 확장을 지원한다. Marvell은 현재 양산 중인 CXL 2.0 Switch에 이어 랙 단위로 CPU, GPU와 DRAM을 연결하는 CXL 3.0 Structera Switch를 2026년 3분기부터 고객사에 샘플링할 예정이다.

메모리 계층 상 KV Cache로 인한 NAND 수요는 일부 감소 가능

엔비디아도 Grace에서 Vera로 CPU 세대가 전환되면서 외부 I/O을 PCIe Gen5에서 PCIe Gen6 및 CXL 3.1로 확장했다. 이것 만으로 NVIDIA 시스템이 당장 CXL 메모리를 사용한다고 보기 어렵지만 향후 메모리 확장과 Pooling을 도입할 수 있는 기반을 마련한 것으로 해석할 수 있다.

KV Cache가 빠르게 증가하는 상황에서 CXL에 대한 관심도 높아지고 있다. 향후 CXL이 신규 DRAM 수요를 확대할지, Meta 사례처럼 기존 메모리의 재활용과 활용률 개선을 중심으로 확산되어 순수요를 감소시킬지 더 지켜볼 필요가 있다. 다만 메모리 계층 상 CXL메모리의 사용이 늘어나면 KV Cache가 NAND까지 내려가는 수요 일부는 감소할 수 있어 보인다.

그림54 Meta Vistara 기반 로컬 DRAM-CXL 확장 메모리 구조



자료: Meta, DS투자증권 리서치센터

주: Vistara: Making CXL Real—Full Path from ASIC Design and OS Support to Hyperscale Deployment

그림55 Meta Vistara 기반 로컬 DRAM-CXL 확장 메모리 구조

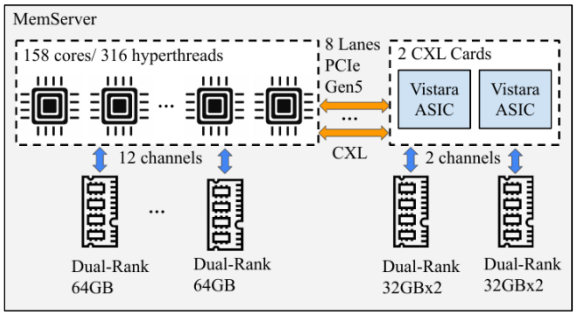


Fig. 6. High-level architecture of a *MemServer*.

자료: Meta, DS투자증권 리서치센터

주: Vistara: Making CXL Real—Full Path from ASIC Design and OS Support to Hyperscale Deployment

표14 NVIDIA Vera CPU와 Grace CPU 비교

Feature	Grace CPU	Vera CPU
Cores	72 Neoverse V2 cores	88 NVIDIA Custom Olympus cores
Threads	72	176 Spatial Multithreading
Memory bandwidth	Up to 512GB/s	Up to 1.2TB/s
Memory capacity	Up to 480GB LPDDR5X	Up to 1.5TB LPDDR5X
NVLINK-C2C	900 GB/s	1.8 TB/s
PCIe/CXL	Gen5	Gen6/CXL 3.1
Confidential compute		Supported

자료: Nvidia, DS투자증권 리서치센터

성능을 포기하지 않는 비용 효율화의 방법, ASIC

추론과 함께 한층 강화된 ASIC의 필요성

ASIC은 성능 향상과 비용 효율화를 동시에 달성하는 방법

LLM 서비스가 더 확산되기 위해서는 사용자, 모델 개발사, 하이퍼스케일러 모두의 비용 부담을 낮춰야 한다. 그와 함께 모델의 성능은 계속 높여야 하는 어려운 싸움이 AI산업에서 벌어지고 있다. 범용성과 개발 유연성이 높은 GPU는 여전히 핵심 가속기로 남겠지만 모든 연산을 고가의 GPU로 처리하는 방식만으로는 비용과 전력 부담을 낮추기 어렵다. 이에 따라 특정 목적에 맞춰 연산 구조와 데이터 이동을 최적화한 ASIC의 채택이 확대될 가능성이 높다. ASIC은 GPU를 전면 대체하기보다 표준화된 작업을 낮은 비용과 전력으로 처리해, 성능 향상과 비용 효율화를 동시에 달성에 활용될 것이다.

ASIC의 두 방향
(1) GPU 의존도를 낮춰 AI서비스의 경제성 개선

ASIC 확대는 크게 두 방향에서 나타나고 있다. 먼저 하이퍼스케일러가 GPU 의존도를 낮추고 AI서비스의 경제성을 개선하려는 시도이다. Google TPU와 AWS Trainium이 대표적이다. 이들은 자사 AI서비스와 클라우드 사업에 맞춰 연산 구조를 최적화해 비용, 성능을 개선하고, 자체 칩, 클라우드, 소프트웨어를 결합하여 AI 인프라의 수직계열화를 강화하려는 전략이다. 이들의 ASIC은 특정 영역에서 GPU 대비 높은 효율을 보이기도 하는데, AWS는 Trainium2가 동급 GPU보다 30~40% 우수한 가격 대비 성능을 제공한다고 설명했다.

(2) 특정 영역을 효율적으로 해결하기 위한 목적

다음으로 특정 추론 병목을 해결하기 위한 ASIC도 다수 출시되고 있다. 엔비디아가 사실상 인수한 Groq의 LPU, OpenAI와 Broadcom이 개발한 Jalapeño(할라피뇨), Cerebras의 Wafer-Scale Engine 등이 대표적인 사례이다. Jalapeño는 OpenAI가 모델과 서비스를 운영하며 축적한 추론 특성을 칩과 시스템 설계에 반영해 성능과 효율을 높이도록 설계됐다. 또한 Groq와 Cerebras는 대용량 온칩 SRAM과 높은 내부 대역폭을 활용해 모델 가중치와 중간 데이터의 외부 메모리 왕복을 줄이고, 추론 Latency와 처리량을 개선하는 데 초점을 둔다.

표15 주요 ASIC 프로젝트

고객사	프로세서	ASIC Partner(추정)	출시일(예정)	비고
Google	TPU v7 Ironwood	Broadcom	2025년 하반기	
	TPU 8t	Broadcom	2026년 하반기	
	TPU 8i	MediaTek	2026년 하반기	
Amazon	Trainium3	Marvell / Alchip	2025년 하반기	
	Trainium4	Marvell / Alchip	2027년 하반기	
Microsoft	Maia 200	Marvell	2026년 상반기	
	Maia 300	Marvell	2027년 예상	
Meta	MTIA 300	Broadcom	2026년 상반기	
	MTIA 400	Broadcom	2026년 하반기	
	MTIA 450	Broadcom	2027년 상반기	
OpenAI	Jalapeño	Broadcom	2026년 하반기	
Apple	Baltra	Broadcom	미정	언론보도, 미확정

자료: DS투자증권 리서치센터

그간 반복되었던 물량 축소와 출시 지연의 역사… 이번에는 다를까?

AI워크로드가 급격히 변화하는 환경에서는 범용성 높은 GPU가 유리

그동안 ASIC 채택이 예상보다 지연된 이유는 AI 워크로드가 급격히 변화하는 환경에서는 엔비디아의 GPU의 범용성과 CUDA 생태계가 대응에 유리하기 때문이다. 반면 ASIC은 특정 연산에 최적화될수록 효율이 높아지지만 개발 기간과 초기 비용이 있음에도 불구하고 설계가 완료될 무렵 모델의 아키텍처나 핵심 연산이 바뀌면 효율이 급감할 위험이 있어 개발은 계속되었지만 지연되거나 생산되는 물량이 예상치를 하회하는 흐름이 지속되었다.

표준화된 워크로드 증가로 ASIC의 쓰임이 늘어날 시점

AI 컴퓨팅의 중심이 사전학습에서 Reasoning으로 확장되고 반복적으로 처리되는 토큰과 표준화된 워크로드가 증가하면서 ASIC의 경제성도 높아질 전망이다. ASIC은 초기 개발비가 높지만, 대규모 추론 물량을 확보한 하이퍼스케일러와 모델 개발사는 이를 많은 요청에 분산해 요청당 비용을 낮출 수 있다. 또한 전력과 데이터센터 용량이 제한된 환경에서는 동일한 전력으로 더 많은 추론을 처리하는 전력 효율의 중요성도 커진다. 이에 따라 GPU는 최신 모델과 변화가 잦은 워크로드를, ASIC은 규모가 크고 패턴이 안정된 업무를 담당하는 이기종 컴퓨팅 구조가 강화될 가능성이 높다.

ASIC 시장 확대는 ASIC 설계사와 파운드리 기업의 수혜로 이어질 것이다. 하이퍼스케일러와 모델 개발사가 필요한 연산 구조와 시스템 사양을 정의하더라도 실제 ASIC을 구현하려면 Broadcom, Marvell과 같은 설계 업체와 CPU, 인터페이스, 메모리 관련 IP가 필요하기 때문이다. 또한 설계가 완료된 칩은 미세공정과 대형 패키징 역량을 갖춘 첨단 파운드리에서 생산되기에 파운드리는 GPU와 ASIC 간 경쟁 속에 수혜가 있을 것으로 예상한다. 특히 첨단 파운드리를 사실상 독점 중인 TSMC 외에도 삼성전자와 인텔의 파운드리 수주 확대를 전망되는데, 삼성전자의 Groq LPU 수주 등이 이러한 사례이다.

이 외에도 AI 연산 인프라가 광범위하게 성장함에 따라 CPU와 주변 반도체의 역할도 새롭게 정의되고 있다. 이 과정에서 CPU, DPU, 네트워크 Switch, CXL 등 다양한 목적의 ASIC 프로젝트도 늘어나고 있다.

표16 삼성전자, 인텔 프로젝트 수주 및 수주 기사화 현황

파운드리사	고객사	칩 프로젝트	목적	공식 발표 여부
삼성전자	NVIDIA(Groq)	Groq 3 LPU	추론 특화	공식 발표
	Tesla	AI6	자율주행, Optimus, AI데이터센터 연산용	공식 발표
	Meta	차세대 MTIA	학습 및 추론	
	Anthropic	자체 AI ASIC	추론 특화(추정)	
	ByteDance	자체 AI ASIC	추론 특화(추정)	
인텔 파운드리	Google	차세대 TPU	학습 및 추론	
	AWS	AI Fabric Chip	네트워킹	공식 발표

자료: 각사, DS투자증권 리서치센터

DS INVESTMENT & SECURITIES

Company
Analysis

DS INVESTMENT & SECURITIES

코어위브 [Coreweave]
CRWV US

컴퓨팅 쇼티지에 베팅하는 가장 좋은 방법

엔비디아가 밀어주는 네오클라우드 기업

코어위브는 대규모 GPU 클러스터를 기반으로 AI모델의 학습, 추론에 필요한 컴퓨팅을 제공하는 대표 네오클라우드 기업이다. 네오클라우드는 AI 시대의 클라우드로 AI워크로드에 특화된 GPU를 고속 네트워크로 대규모 연결하여 방대한 병렬 연산 컴퓨팅을 제공하는데 집중한다. 코어위브는 엔비디아의 지분투자를 받은 기업으로 컴퓨팅에 대한 높은 수요 속에 엔비디아의 차세대 GPU 조기 확보 능력을 기반으로 성장했으며 최근에는 AI 개발 플랫폼 Weights & Biases 인수를 통해 소프트웨어 역량 확보에도 주력 중이다. 중장기적으로 수익성 개선 기대.

최근 주가 부진의 이유는?

최근 코어위브의 주가는 네오클라우드 사업을 둘러싼 위험 인식이 높아진 영향으로 부진하다. 네오클라우드는 AI산업 혹은 컴퓨팅 수요에 대한 부정적 전망에 민감하게 반응하는 업종 중 하나이며, 코어위브는 오픈AI와 계약 비중이 높은 편이라 하락폭이 두드러지는 경향이 있다. 최근 '메타가 초과 컴퓨팅을 활용하여 클라우드 사업 진출을 검토 중'이라는 보도로 컴퓨팅 수요 둔화 및 경쟁 심화 우려 속에 급락했으나, 여러 정황 상 여전히 컴퓨팅 수요는 높고, 메타가 올해 들어 3건 이상 컴퓨팅 계약을 체결한 점과 높은 GPU가격을 고려했을 때 메타이 클라우드 사업 진출을 컴퓨팅 수요 둔화로 연결 짓기에는 모순되는 지점이 있다.

투자포인트

(1) 최근 감가상각이 상당부분 진행된 GPU(A100, H100) 가격도 반등 중이다. 향후 감가상각이 마무리되면 비용 없이 수익만 인식하게 되면서 수익성이 개선될 전망이다. 또한 컴퓨팅 수요에 따른 향후 컴퓨팅 임대 단가 상승을 기대한다. (2) 하반기 Vera Rubin NVL72가 본격 출하될 예정이다. 동사는 업계 최초로 해당 제품을 가동, 검증했으며 하반기에 배치가 본격화되면 단위 컴퓨팅 당 수익성 개선에 기여할 예정이다. (3) 자본조달 조건이 개선되고 있다. 올해 3월 $8.5B규모의 DDTL 4.0을 조달했으며 고정금리 기준 약 5.9%로 이전 조건보다 상당히 개선되었다. 동사는 차입을 통해 데이터센터를 구축하기에 낮아진 자본비용으로 인한 이익 개선이 기대된다.

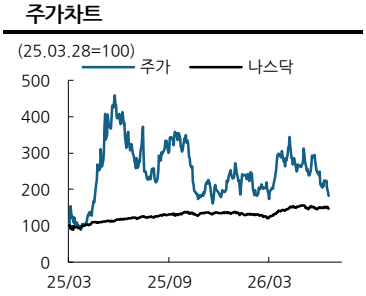
Key Data

국가	USA
거래소	NASDAQ
GICS(부문)	정보기술
GICS(산업)	IT 서비스
시가총액($ Bn)	39.78
시가총액(조원)	59.15
현재주가(07/17, $)	73.21
52주 최고가($)	153.20
52주 최저가($)	63.80
주요주주	
MAGNETAR FINANCIAL	11.71%
NVIDIA	10.55%

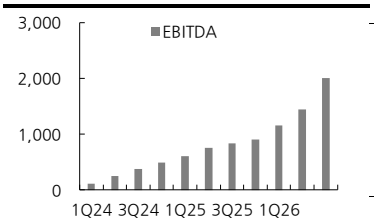
주가추이 및 상대강도

(%)	절대수익률	상대수익률
1M	-36.5	-34.5
3M	-37.3	-41.6
6M	-27.7	-36.2

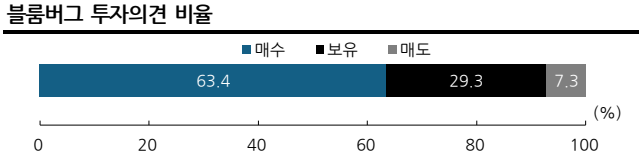
주가차트



EBITDA 추이($ Mn)



블룸버그 투자의견 비율



Financial Data

($ Mn)	FY2024	FY2025	FY2026F	FY2027F
매출액	1,915.4	5,131.0	12,683.1	25,146.1
영업이익	0.7	355.8	666.0	988.9
OPM(%)	0.3	18.6	13.0	7.8
순이익	-44.6	-64.9	-606.0	-1,850.9
EPS($)	-6.0	-1.4	-3.4	-1.0
증감률(%)	적자지속	적자지속	적자지속	적자지속
ROE(%)	-	-	-	-
PER(배)	-	-	-	-
PBR(배)	-	10.8	13.1	14.1
PSR(배)	-	10.2	6.0	3.6

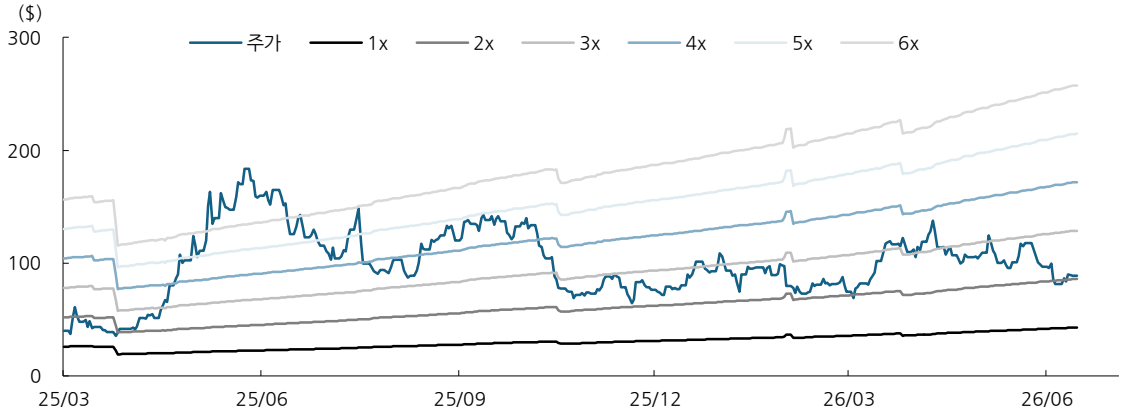
자료: Bloomberg, DS투자증권 리서치센터, Non-GAAP 기준, 비율은 GAAP 기준

표17 코어위브 실적 테이블

($ Mn, %)	FY1Q26	FY2Q26	FY3Q26	FY4Q26	FY1Q27	FY2Q27	FY3Q27F	FY25	FY26	FY27F
매출	**982**	**1,213**	**1,365**	**1,572**	**2,078**	**2,569**	**3,480**	**1,915**	**5,131**	**12,683**
%YoY	*420.3*	*207.0*	*133.7*	*110.2*	*111.7*	*111.8*	*155.0*		*99.7*	*264.5*
%QoQ	*31.3*	*23.5*	*12.5*	*15.2*	*32.2*	*23.6*	*35.5*		*167.9*	*147.2*
매출총이익	**719**	**587**	**996**	**1,063**	**646**	**1,714**	**2,349**	**1,422**	**3,366**	**8,615**
GPM(%)	73.3	48.4	73.0	67.6	31.1	66.7	67.5	74.2	65.6	67.9
%YoY	*455.6*	*105.0*	*125.9*	*88.1*	*-10.2*	*191.7*	*135.9*		*96.4*	*266.7*
EBITDA	**606**	**753**	**838**	**898**	**1,157**	**1,436**	**2,000**	**1,219**	**3,095**	**7,367**
EBITDA Margin(%)	61.7	62.1	61.4	57.1	55.7	55.9	57.5	63.7	60.3	58.1
%YoY	*479.8*	*201.3*	*121.3*	*84.8*	*90.9*	*90.7*	*138.6*		*115.5*	*268.5*
영업이익	**163**	**200**	**217**	**88**	**21**	**66**	**271**	**356**	**668**	**989**
OPM(%)	16.6	16.5	15.9	5.6	1.0	2.6	7.8	18.6	13.0	7.8
%YoY	*549.6*	*135.0*	*74.1*	*-27.3*	*-87.1*	*-67.0*	*25.0*		*912.0*	*264.3*
순이익	**-150**	**-131**	**-41**	**-284**	**-589**	**-559**	**-465**	**-65**	**-605**	**-1,851**
NPM(%)										
%YoY	*적자지속*	*적자지속*	*적자전환*	*적자지속*	*적자지속*	*적자지속*	*적자지속*		*적자지속*	*적자지속*
EPS	**-0.60**	**-0.27**	**-0.08**	**-0.56**	**-1.12**	**-1.06**	**-0.85**	**-3.48**	**-1.51**	**-3.42**
%YoY	*적자지속*	*적자지속*	*적자지속*	*적자지속*	*적자지속*	*적자지속*	*적자지속*		*적자지속*	*적자지속*

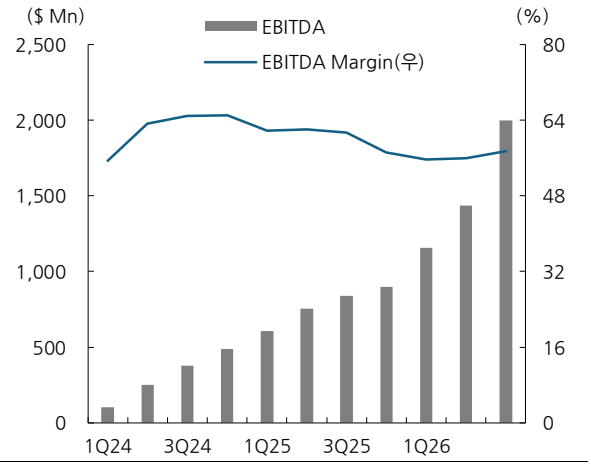
자료: Bloomberg, DS투자증권 리서치센터

그림56 코어위브12MF PSR 밴드 추이



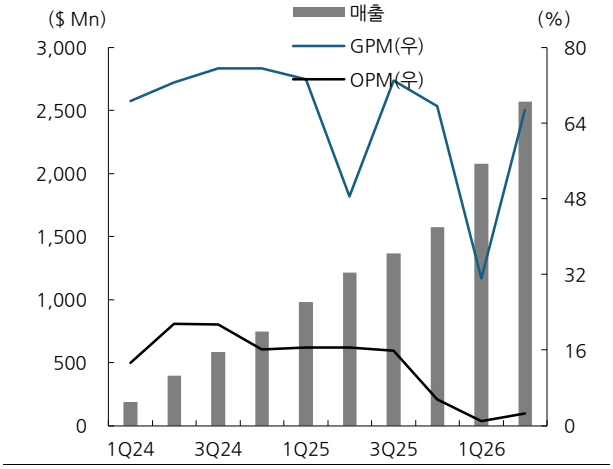
자료: Bloomberg, DS투자증권 리서치센터

그림57 코어위브 부문별 연간 EBITDA 추이



자료: Bloomberg, DS투자증권 리서치센터

그림58 코어위브 연간 매출액, 매출총이익률, 영업이익률



자료: Bloomberg, DS투자증권 리서치센터

투자포인트

컴퓨팅 수요
→ 컴퓨팅 임대 단가 상승

(1) 높은 컴퓨팅 수요는 네오클라우드의 가격 협상력과 가동률을 동시에 높일 수 있다. 특히 GPU클러스터 공급이 제한적인 상황이기에 컴퓨팅 수요는 신규 및 갱신 계약에서 단가 상승과 직결된다. 최근 자체 컴퓨팅을 외부에 임대하기 시작한 xAI의 계약은 GW당 계약가치가 작년 9월 체결된 유사 계약보다 약 44% 높은 것으로 추정된다. 계약별 HW와 제공 서비스가 달라 직접 비교에는 한계가 있지만, 이는 대규모 AI 컴퓨팅에 대한 고객의 지불의사가 여전히 높다는 점을 보여주며 코어위브의 가격 협상력에도 긍정적인 신호다.

컴퓨팅 수요
→ 구형 GPU 가치 상승

다음으로 높은 컴퓨팅 수요는 구형 GPU의 경제적 내용연수를 늘려 수익성 개선으로 이어진다. GPU는 약 6년간 감가상각되는데, 2020년 출시된 A100도 최근 임대가격이 반등하는 등 여전히 유효한 컴퓨팅 자원으로 활용되고 있다. 이는 최신 GPU가 필요한 최첨단 학습뿐 아니라 추론과 Fine-tuning 등 다양한 AI워크로드가 구형 GPU 수요를 지지하고 있음을 보여준다. 감가상각이 종료된 GPU가 계속 임대되면 매출은 유지되는 반면 감가상각비는 사라져 해당 자산의 수익성이 크게 높아질 수 있다.

표18 스페이스X xAI부문 컴퓨팅 계약 단가: $18.56B/GW

항목	금액/용량	비고
월 계약금액 합계	$2.32B	- Anthropic: $1.25B - Google: $0.92B - Reflection AI: $0.15B
연 계약 금액 합계	**$27.84B**	
컴퓨팅 용량(GW)	1.5	- Colossus 1: 0.3 - Colossus 2: 1.2
1GW 당 컴퓨팅	**$18.56B**	

자료: 언론 종합, DS투자증권 리서치센터

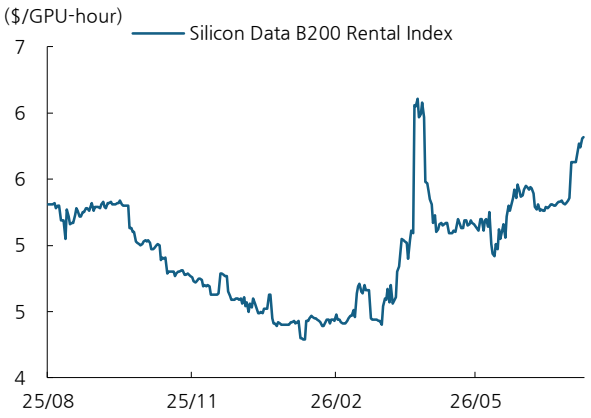
표19 Nebius-Microsoft 컴퓨팅 계약 단가: $12.93B/GW

항목	계약금액	비고
5년 계약 금액	$17.4B	
5년 총 계약 금액 (옵션 포함)	$19.4B	
연 계약 금액 합계	$3.88B	
컴퓨팅 용량(GW)	0.3	
1GW 당 컴퓨팅	**$12.93B**	

자료: 언론 종합, DS투자증권 리서치센터

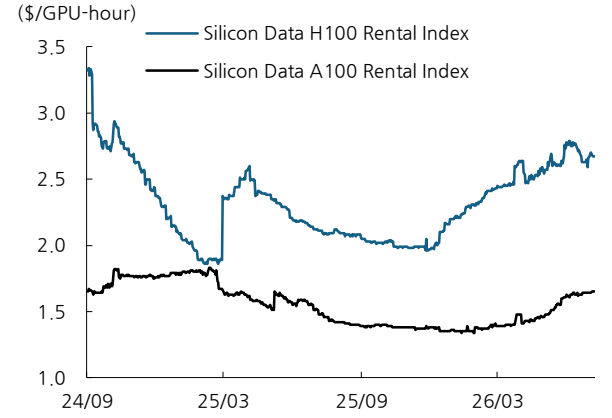
Silicon Data의 GPU 임대가격 지수를 보면 출시 후 약 6년이 지난 A100의 가격도 최근 반등하고 있다. AWS 역시 7월부터 GPU 가격을 약 20% 인상했는데, A100도 대상에 포함됐다. 이는 최신 GPU뿐 아니라 구형 GPU의 수요와 잔존가치도 유지되고 있음을 보여준다. 따라서 코어위브가 보유한 기존 GPU의 가동률과 임대단가가 유지되고 감가상각 부담이 점차 낮아진다면, 하반기 이후 수익성 개선에 기여할 수 있다.

그림59 엔비디아 B200 임대가격 지수



자료: Bloomberg, DS투자증권 리서치센터
주: B200은 2024년 하반기 출시된 GPU

그림60 엔비디아 H100, A100 임대가격 지수



자료: Bloomberg, DS투자증권 리서치센터
주: A100/H100은 각각 2020년 상반기, 2022년 하반기 출시된 GPU

Vera Rubin 배치 본격화되면 컴퓨팅 자산의 수익성 개선 기대

(2) 하반기 엔비디아 Vera Rubin NVL72의 출하와 상용 배치가 본격화되면서 코어위브의 수혜가 기대된다. 코어위브는 클라우드 사업자 중 최초로 Vera Rubin NVL72의 가동과 검증을 완료해 차세대 제품 도입 역량을 입증했다. Vera Rubin은 이전 세대보다 추론 성능과 전력 효율이 개선된 만큼, 높은 가동률과 임대단가가 유지된다면 토큰당 원가를 낮추고 컴퓨팅 자산의 수익성을 높일 수 있다. 또한 엔비디아의 지분 투자를 받은 전략적 파트너라는 점에서 차세대 제품을 경쟁사보다 조기에 도입, 상용화할 가능성이 높은 것도 강점이다.

자본 조달 여건도 개선되는 중

(3) 자본조달 여건도 개선되고 있다. CoreWeave는 올해 3월 $8.5B 규모의 DDTL 4.0을 확보했으며, 고정금리 Tranche는 약 5.9%, 변동금리 Tranche는 SOFR+2.25%로 책정됐다. 이는 고객 계약과 GPU 자산을 기반으로 이전보다 낮은 비용으로 대규모 CapEx를 조달할 수 있게 됐다는 의미다. 차입을 통해 데이터센터를 확장해온 동사에는 자본비용 하락이 신규 프로젝트의 수익성과 ROIC를 높이고, 장기적으로 이자비용 부담을 완화하는 요인이 될 수 있다.

네오클라우드 경쟁 심화

다만 메타에 이어 소프트뱅크가 'SB Neo'를 설립하고 네오클라우드 사업에 진출한 점은 향후 컴퓨팅 공급이 확대 요인이므로 동사에는 부정적이다. 소프트뱅크는 미국에 10GW 규모 인프라를 구축할 계획이다.

표20 코어위브가 체결한 DDTL 정리

체결일	구분	금리
2023년 7월	DDTL 1.0	SOFR+9.6196%
2024년 5월	DDTL 2.0	SOFR+6.0~13.0%
2025년 9월	DDTL 2.1	SOFR+4.25%
2025년 7월	DDTL 3.0	SOFR+4.00%
2026년 3월	DDTL 4.0	변동금리 SOFR+2.25%, 고정 약 5.9%
2026년 5월	DDTL 5.0	SOFR+4.50%

자료: DS투자증권 리서치센터

DS INVESTMENT & SECURITIES

옥타 [Okta]
Okta US

Agentic AI 시대의 사이버 수문장

신원 인증과 접근 권한 관리에 특화된 Identity 보안 기업

Okta는 기업 임직원의 신원 및 접근 권한을 관리하는 Workforce Identity와 Auth0를 기반으로 기업이 자사 고객의 로그인과 인증을 관리하도록 돕는 Customer Identity를 제공하는 클라우드 기반 Identity 보안 기업이다. 직원용 SSO(단일 로그인), MFA(다중 인증)에서 출발했지만 현재는 계정과 접근 권한의 생성, 변경, 회수부터 권한 승인 및 정기검토 등 관리자용 제품으로 확대하며 종합 Identity 보안 플랫폼으로 체질을 전환하고 있다. 다양한 산업의 2만개 이상 기업이 Okta 솔루션을 활용 중이며, 전체 매출의 98%가 구독 매출로 구성된다.

'신원과 권한'으로 사이버 보안의 흐름이 변해가는 중

기업의 IT 환경이 온프레미스에서 클라우드 중심으로 변화하면서 네트워크 경계와 방화벽만으로는 보안을 통제하기 어려워졌다. 이에 따라 보안의 중심도 접속 위치가 아니라, 사람, 애플리케이션, Agent의 신원을 확인하고 필요한 자원에 최소한의 권한만 부여하는 Identity 중심으로 이동 중이다. 특히 AI Agent가 이메일, CRM, 코드 저장소, 데이터베이스 등에 접근해 업무를 수행할수록 신원 및 권한의 부여/회수, 민감한 행동 승인 및 감사 등 Identity 보안의 역할은 커지게 된다. Okta는 기존 솔루션을 Agent Identity까지 확장 중이며 개발자를 위한 Auth0 for AI Agents와 사내 보안팀을 위한 Okta for AI Agents 등의 서비스를 출시했다. 다만 Microsoft가 신원, 접근 권한 관리 서비스인 Entra Agent ID를 MS365와 결합해 제공하는 등 경쟁 상황이 우호적이지만은 않다.

투자포인트

(1) Agentic AI의 확산은 기업이 관리해야 할 Identity와 접근 권한 검증 횟수를 구조적으로 늘린다는 점에서 동사의 타켓 시장이 유의미하게 확대될 전망이다. Agent는 업무 과정에서 데이터베이스 등 사내 여러 시스템에 연속적으로 접근하기에 고정된 장기 권한보다 업무에 필요한 최소 권한, 단기 인증정보만 부여하고 작업 종료 또는 이상행동 감지 시 즉시 회수하는 체계가 중요하다. 동사의 역할이 확대되는 그림이다. (2) 외부 전문 Agent와 협업하는 방식으로 Agentic AI가 고도화되는 점도 긍정적이다. 기업은 사내, 외부 Agent의 신뢰 수준을 구분하면서도 플랫폼 전반의 신원과 위임 권한을 일관되게 통제해야 하는데 특정 클라우드나 Agent 생태계에 종속되지 않은 당사 솔루션이 중립적인 통제 플랫폼으로 채택될 가능성이 있다 (3) 순유지율(net retention rate) 반등과 대형 고객(연간 10만달러 이상 지출)의 증가세도 긍정적이다. 최근 주가 급등은 다소 부담스럽지만 큰 그림은 좋다.

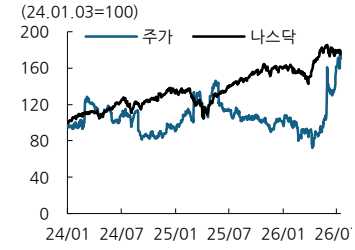
Key Data

국가	USA
거래소	NASDAQ
GICS(부문)	정보기술
GICS(산업)	IT 서비스
시가총액($ Bn)	25.66
시가총액(조원)	38.16
현재주가(07/17, $)	149.35
52주 최고가($)	157.00
52주 최저가($)	62.66
주요주주	
VANGUARD GROUP	10.87%
BLACKROCK	10.46%

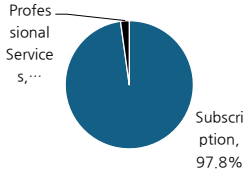
주가추이 및 상대강도

(%)	절대수익률	상대수익률
1M	32.1	34.1
3M	106.7	102.4
6M	66.8	58.3

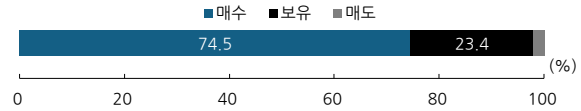
주가차트



사업부별 매출 비중



블룸버그 투자의견 비율



Financial Data

($ Mn)	FY2024	FY2025	FY2026F	FY2027F
매출액	2,263.0	2,610.0	2,919.0	3,199.2
영업이익	310.0	587.0	766.0	818.5
OPM(%)	13.7	22.5	26.2	25.6
순이익	286.0	510.0	646.0	708.6
EPS($)	1.6	2.8	3.5	3.8
증감률(%)	흑자전환	75.6	24.6	9.8
ROE(%)	-6.3	0.5	3.5	8.6
PER(배)	흑자전환	1,690.8	63.6	40.2
PBR(배)	2.4	2.6	2.1	3.7
PSR(배)	6.0	6.1	5.1	8.4

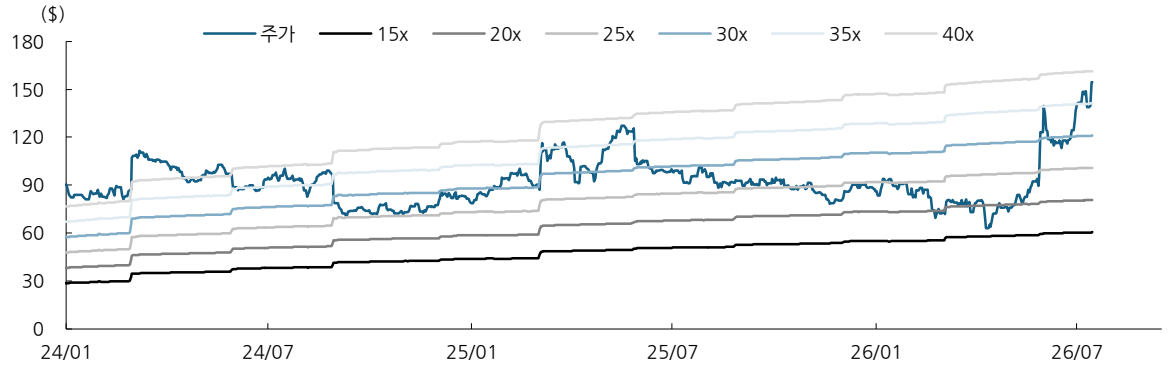
자료: Bloomberg, DS투자증권 리서치센터, Non-GAAP 기준, 비율은 GAAP 기준

표21 옥타 실적 테이블

($Mn, %)	FY1Q26	FY2Q26	FY3Q26	FY4Q26	FY1Q27	FY2Q27F	FY3Q27F	FY25	FY26	FY27F
매출	**688.0**	**728.0**	**742.0**	**761.0**	**765.0**	**793.3**	**808.9**	**2,610.0**	**2,919.0**	**3,129.2**
%YoY	*11.5*	*12.7*	*11.6*	*11.6*	*11.2*	*8.9*	*9.0*	*15.3*	*11.8*	*7.2*
%QoQ	*0.9*	*5.8*	*1.9*	*2.6*	*0.5*	*3.6*	*2.0*			
사업부별 매출										
Subscription	673.0	711.0	724.0	747.0	750.0	781.4	798.7	2,556.0	2,855.0	3,084.0
Professional Services	15.0	17.0	18.0	14.0	15.0	11.9	10.2	54.0	64.0	45.2
매출총이익	**563.0**	**594.0**	**604.0**	**624.0**	**624.0**	**649.9**	**649.9**	**2,130.0**	**2,385.0**	**2,625.0**
GPM(%)	81.8	81.6	81.4	82.0	81.6	81.9	80.3	81.6	81.7	83.9
%YoY	*11.9*	*12.5*	*11.6*	*11.8*	*10.8*	*9.4*	*9.8*	*17.1*	*12.0*	*10.1*
영업이익	**184.0**	**202.0**	**178.0**	**202.0**	**191.0**	**206.2**	**200.5**	**587.0**	**766.0**	**818.5**
OPM(%)	26.7	27.7	24.0	26.5	25.0	26.0	24.8	22.5	26.2	26.2
%YoY	*38.3*	*36.5*	*29.0*	*20.2*	*3.8*	*2.1*	*12.6*	*89.4*	*30.5*	*6.9*
순이익	**158.0**	**169.0**	**152.0**	**167.0**	**168.0**	**177.5**	**173.0**	**510.0**	**646.0**	**708.6**
NPM(%)	**23.0**	**23.2**	**20.5**	**21.9**	**22.0**	**22.4**	**21.4**	**19.5**	**22.1**	**22.6**
%YoY	*35.0*	*29.0*	*25.6*	*18.4*	*6.3*	*5.0*	*13.8*	*78.3*	*26.7*	*9.7*
EPS	**0.86**	**0.91**	**0.82**	**0.90**	**0.91**	**0.96**	**0.94**	**2.81**	**3.50**	**3.84**
%YoY	*32.3*	*26.4*	*22.4*	*15.4*	*5.8*	*5.8*	*14.4*	*75.6*	*24.6*	*9.8*

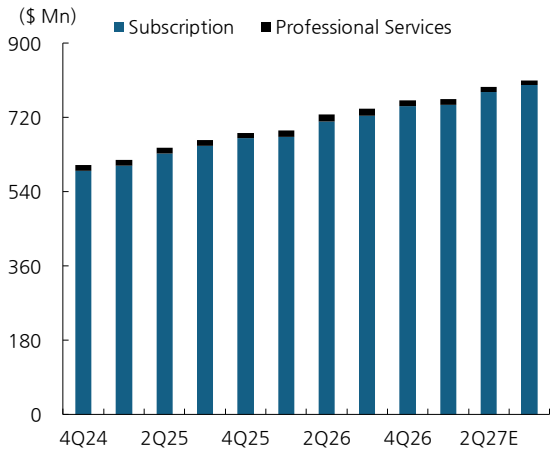
자료: Bloomberg, DS투자증권 리서치센터

그림61 옥타 12MF PER 밴드 추이



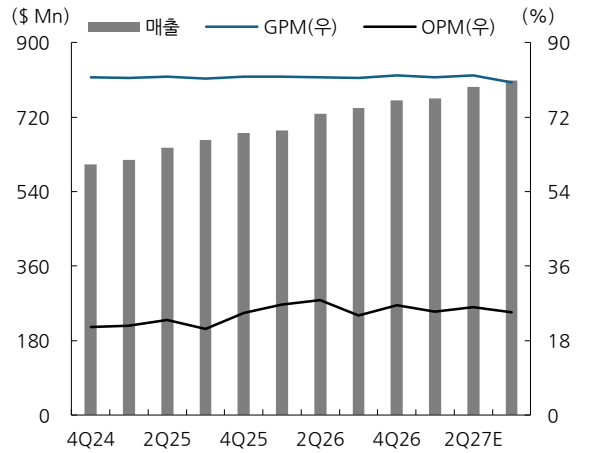
자료: Bloomberg, DS투자증권 리서치센터

그림62 옥타 부문별 연간 매출액 추이



자료: Bloomberg, DS투자증권 리서치센터

그림63 옥타 연간 매출액, 매출총이익률, 영업이익률



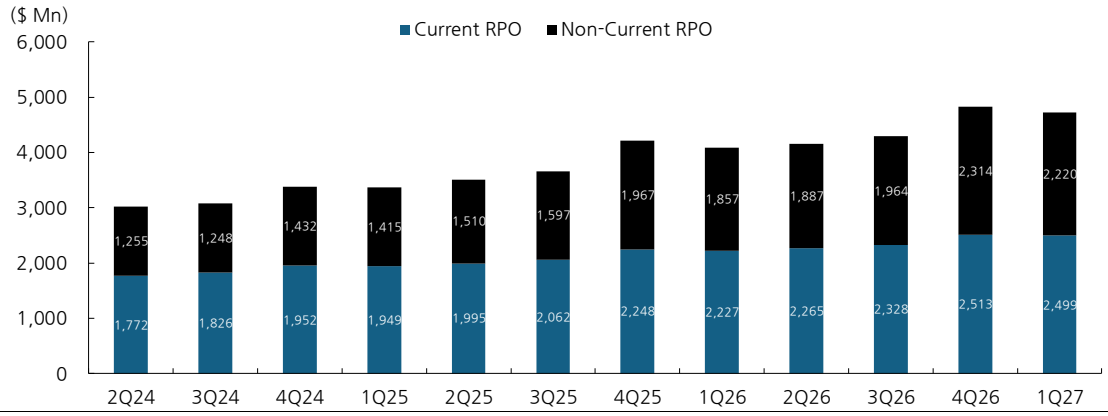
자료: Bloomberg, DS투자증권 리서치센터

표22 옥타의 주요 제품

제품	역할	대표 사용 사례
SSO(Single Sign On)	한 번의 로그인으로 여러 업무용 SW에 접속 권한을 제공(로그인 간소화)	직원이 Okta에 한 번 로그인하면, 기존에 권한을 받은 MS 365, Salesforce, Slack 등에 접근 가능
MFA(Multi-Factor Authentication)	접속 위치, 접속 기기, 위험도에 따라 추가 인증을 요구(로그인 안전성 강화)	의심스러운 접속에 생체 인증 혹은 보안키 요구
Universal Directory	사용자, 그룹, 기기와 관련된 권한 정보 등을 중앙에서 관리하는 디렉터리	인사시스템과 여러 업무용 SW에 분산 저장된 직원 정보를 통합
Lifecycle Management	입사, 부서 이동, 퇴사 직원 등의 계정과 접근권한을 자동 생성, 변경, 회수	신입 직원의 업무용 계정 자동 생성, 퇴사 시 일괄 차단
Okta Identity Governance	접근 권한 요청, 관리자의 승인 및 정기 검토, 회수, 감사 기록을 관리	ERP 접근 신청을 팀장이 승인하고 일정 기간 후 재검토
Auth0	기업이 자사 웹사이트, SW에 고객 로그인과 권한 관리 기능을 구축하도록 지원	쇼핑 사이트, 금융 앱 등의 회원가입, 소셜 로그인, MFA 구현
Okta for AI Agents	기업 내부 에이전트를 새로운 형태의 비인간 Identity이자 디지털 인력으로 관리	사내 에이전트에 담당자를 지정하고 업무 수행에 필요한 범위, 시간만큼 접근 권한을 부여
Auth0 for AI Agents	개발자가 고객용 에이전트에 사용자 인증, API 접근권한, 승인 절차를 구현하도록 지원	고객을 대신해서 Slack, Google 등에 접근하는 에이전트에 제한된 권한 부여

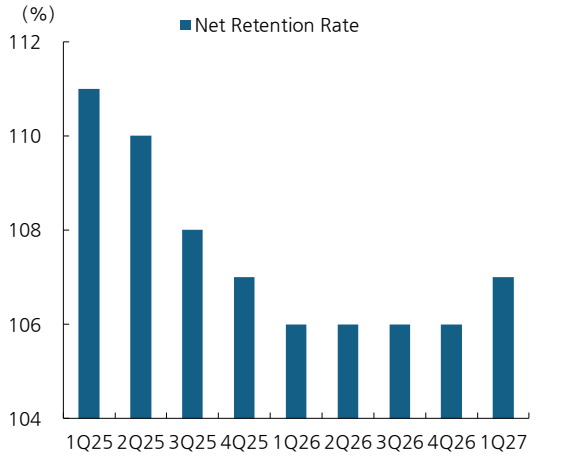
자료: DS투자증권 리서치센터

그림64 옥타 Total RPO, Current RPO 추이



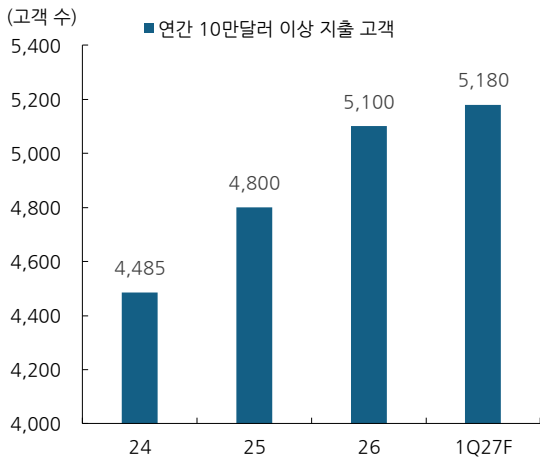
자료: Okta, DS투자증권 리서치센터, 주: Current RPO는 1년 내 매출로 전환될 것으로 예상되는 RPO

그림65 달러 기준 순유지율($ Net Retention Rate)



자료: Okta, DS투자증권 리서치센터
주: 동일고객의 현재 시점 계약 금액을 1년전 계약 금액과 비교

그림66 연간 10만달러 이상 지출하는 대형 고객 수



자료: Okta, DS투자증권 리서치센터

DS INVESTMENT & SECURITIES

엔비디아 [NVIDIA]
NVDA US

실적 성장이 고스란히 주가에 반영될 시간

본업에서 실망시킨 적은 없다. 단지 중요도가 떨어졌을 뿐…

AI가속기, 네트워킹, CPU, 소프트웨어 전반에서 기술 격차를 확대 중이며 자체 오픈소스 모델인 Nemotron3도 높은 성능을 입증하며 풀스택의 영역을 확장시켰다. 엔비디아의 핵심 경쟁력은 '모델 개발사와 긴밀히 협업해 차세대 모델의 구조와 추론 방식, 메모리 및 지연시간 병목을 조기에 파악하고, 이를 자사 제품에 반영하는 능력'이다. Prefill-Decode 분리를 Vera Rubin NVL72와 Groq 3 LPX를 결합한 이기종 컴퓨팅으로 구현한 사례, Agentic AI에 최적화된 Vera CPU 등이 이러한 전략의 연장선이다. AWS가 Trainium4에 NVLink Fusion과 MGX 랙 아키텍처를 채택했듯이 ASIC 시장의 확대 흐름속에서도 표준 설계자로서 경쟁력은 유지될 것이다.

왜 주가는 실적을 반영하지 못할까?

엔비디아가 한동안 탄력적으로 상승하지 못한 이유는 사업 경쟁력에 있지 않다. 높은 매출총이익률과 이익 성장률 그리고 제품에 대한 수요가 이를 뒷받침한다. 다만 AI모델이 학습에서 추론 중심으로 변화하면서 대규모 병렬 연산에 특화된 GPU보다 메모리, 네트워크, 스토리지, 전력 효율 등 시스템 병목을 해소하는 부품의 중요성이 높아졌기 때문이다. 특히 현재 추론 서비스에서 가장 중요한 비용 효율화를 달성하려면 Batch 크기를 키워야 하는데, 이는 KV Cache를 늘리고 메모리의 중요성을 강화하는 방향이다. 엔비디아가 주목받기 좋은 환경은 아니다.

투자포인트

엔비디아를 긍정적으로 보는 핵심 근거는 밸류에이션이다. CY26년과 CY27년 예상 PER은 각각 22배, 16배로 현재 밸류에이션의 절대 레벨은 높지 않고, EPS성장률과 수익성을 감안하면 낮다. 시장은 엔비디아의 이익 성장을 멀티플 압축으로 상쇄해왔지만, 지금 수준에서는 멀티플이 더 내려갈 수 있는 범위는 제한적이다. 하반기부터 EPS 성장의 상당부분이 주가에 반영될 것이다.

하반기 실적도 역시 좋을 것이다. (1) Vera Rubin 플랫폼이 7월 고객사 출하를 시작으로 대량 양산될 예정이며 (2) Vera CPU 단독 판매로만 연간 $20B 수준의 매출이 전망되기 때문이다. LPX, Vera CPU, CPX 등 HW에 더해 자체 오픈소스 모델 Nemotron 과 같은 SW 영역에서도 경쟁력을 확장시키고 있다. 27Y 기준 최근 3년 연평균 EPS성장률이 약 96%에 달하는 AI 풀스택 기업의 12MF PER이 20배 초반 수준이라면 분명히 좋은 기회이다.

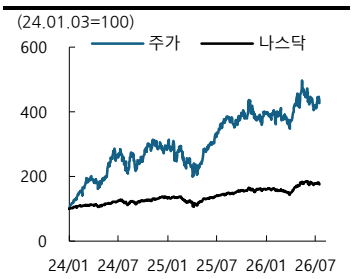
Key Data

국가	USA
거래소	NASDAQ
GICS(부문)	정보기술
GICS(산업)	반도체 & 반도체 장비
시가총액($ Tn)	4.90
시가총액(조원)	7,286.30
현재주가(07/17, $)	202.81
52주 최고가($)	236.54
52주 최저가($)	164.07
주요주주	
VANGUARD GROUP	9.48%
BLACKROCK	7.96%

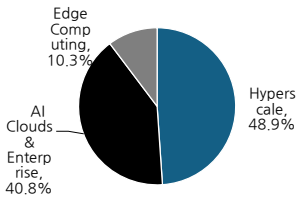
주가추이 및 상대강도

(%)	절대수익률	상대수익률
1M	-0.9	1.0
3M	0.6	-3.7
6M	8.9	0.4

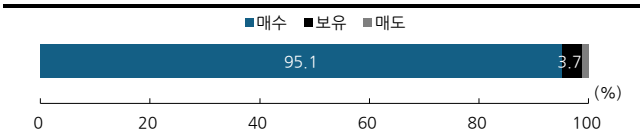
주가차트



사업부별 매출 비중



블룸버그 투자의견 비율



Financial Data

($ Bn)	FY2025	FY2026	FY2027F	FY2028F
매출액	130.5	215.9	392.9	561.5
영업이익	86.8	137.3	260.7	373.7
OPM(%)	66.5	63.6	66.4	66.6
순이익	74.3	117.0	219.4	309.2
EPS($)	3.0	4.8	9.0	12.9
증감률(%)	129.8	57.5	87.5	40.9
ROE(%)	119.2	101.5	91.8	76.0
PER(배)	48.6	43.0	23.5	16.3
PBR(배)	44.0	29.0	17.1	10.7
PSR(배)	26.8	21.2	13.0	9.1

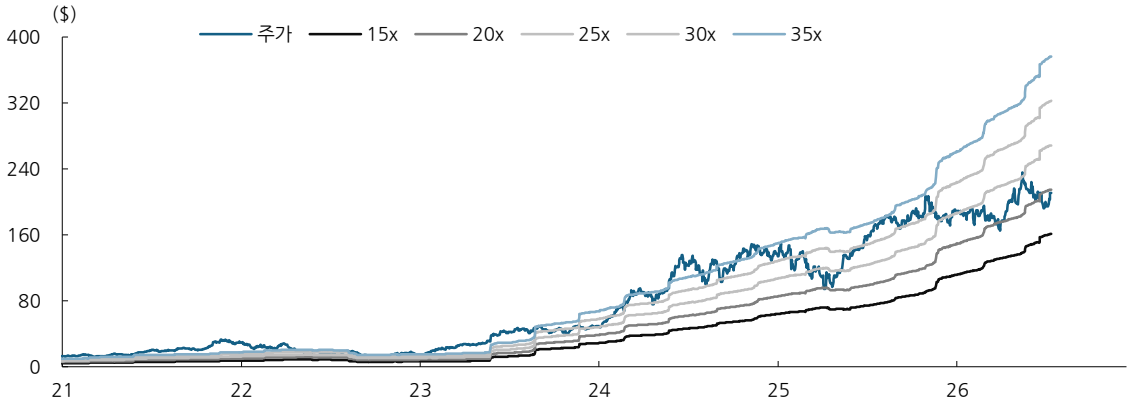
자료: Bloomberg, DS투자증권 리서치센터, Non-GAAP 기준, 비율은 GAAP 기준

표23 엔비디아 실적 테이블

($ Mn, %)	FY1Q26	FY2Q26	FY3Q26	FY4Q26	FY1Q27	FY2Q27F	FY3Q27F	FY25	FY26	FY27F
매출	**44,062**	**46,743**	**57,006**	**68,127**	**81,615**	**91,632**	**103,362**	**130,497**	**215,938**	**392,858**
%YoY	*69.2*	*55.6*	*62.5*	*73.2*	*85.2*	*96.0*	*81.3*	114.2	*65.5*	*81.9*
%QoQ	*12.0*	*6.1*	*22.0*	*19.5*	*19.8*	*12.3*	*12.8*			
사업부별 매출										
Datacenter	39,112	41,096	51,215	62,314	75,246	85,116	96,282	115,186	193,737	365,887
Hyperscale	17,599	23,883	30,340	33,814	37,869	43,420	49,204	53,796	105,636	184,777
ACIE	21,513	17,213	20,875	28,500	37,377	41,506	46,467	61,390	88,101	180,366
Edge Computing	4,950	5,647	5,791	5,813	6,369	6,607	6,807	15,311	22,201	26,725
매출총이익	**26,858**	**33,960**	**41,967**	**51,209**	**61,232**	**68,812**	**76,711**	**98,505**	**153,994**	**292,933**
GPM(%)	61.0	72.7	73.6	75.2	75.0	75.1	74.2	75.5	71.3	74.6
%YoY	30.6	49.4	59.4	77.2	128.0	102.6	82.8	219.1	56.3	90.2
영업이익	**23,275**	**30,165**	**37,752**	**46,107**	**53,783**	**60,822**	**68,703**	**86,788**	**137,299**	**260,671**
OPM(%)	52.8	64.5	66.2	67.7	65.9	66.4	66.5	66.5	63.6	66.4
%YoY	*28.9*	*51.3*	*62.2*	*80.7*	*131.1*	*101.6*	*82.0*	233.7	*58.2*	*89.9*
순이익	**23,635**	**25,783**	**31,767**	**39,552**	**45,548**	**50,922**	**56,953**	**74,266**	**120,737**	**219,407**
NPM(%)	53.6	55.2	55.7	58.1	55.8	55.6	55.1	56.9	55.9	55.8
%YoY	*55.1*	*52.1*	*58.8*	*79.2*	*92.7*	*97.5*	*79.3*	*129.8*	*62.6*	*81.7*
EPS	**0.81**	**1.05**	**1.30**	**1.62**	**1.87**	**2.09**	**2.35**	**2.99**	**4.78**	**8.99**
%YoY	*32.4*	*54.4*	*60.5*	*82.0*	*130.9*	*99.5*	*80.6*	*130.7*	*59.8*	*88.1*

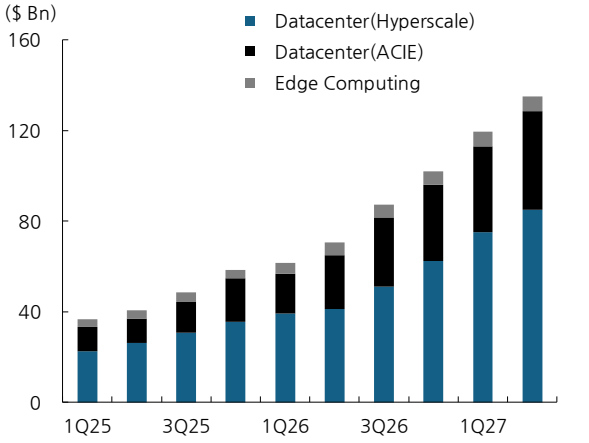
자료: Bloomberg, DS투자증권 리서치센터

그림67 엔비디아 12MF PER 밴드 추이



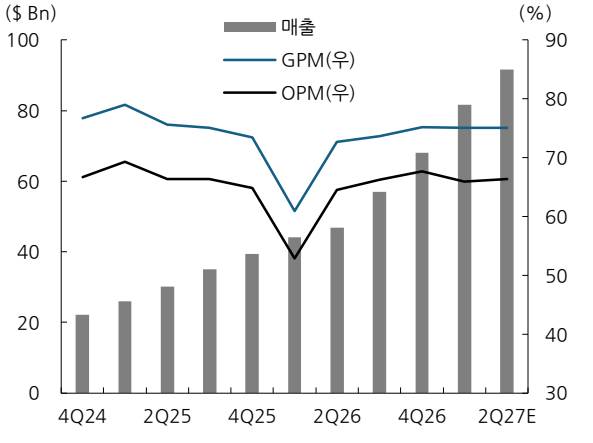
자료: Bloomberg, DS투자증권 리서치센터

그림68 엔비디아 부문별 연간 매출액 추이



자료: Bloomberg, DS투자증권 리서치센터

그림69 엔비디아 연간 매출액, 매출총이익률, 영업이익률



자료: Bloomberg, DS투자증권 리서치센터

DS INVESTMENT & SECURITIES

마벨 테크놀로지 [Marvel Technology]

MRVL US

AI네트워크, ASIC 그리고 CXL까지 좋은 건 다한다

AI 데이터센터용 네트워크 솔루션 기업

마벨 테크놀로지는 네트워크, ASIC 등 데이터센터 인프라에 활용되는 다양한 반도체를 설계하는 팹리스 기업이다. 광통신용 DSP와 스토리지 반도체에서 경쟁력을 쌓았으며 (1) Inphi 인수로 광통신 및 네트워킹 기술을, Innovium 인수를 통해 클라우드용 이더넷 스위치 역량을 확보하면서 AI 데이터센터용 네트워크 솔루션 기업으로 체질을 전환했다. 현재는 ASIC, 고속 스위치에서 경쟁력을 강화하는 동시에 CXL에서도 Structera 제품군을 통해 시장 개화에 선제적으로 대응하고 있다.

성장성이 높은 전방시장에서 경쟁력을 보여주는 중

전방사업의 성장성이 매우 높고 이로 인한 사업 기회가 무궁무진하다. (1) 광통신에서는 800G PAM4 수요가 견조한 가운데 1.6T 제품의 양산이 확대되고 있으며 Celestial AI와 Polariton 인수를 통해 Scale-up 광연결과 3.2T 이후의 기술 경쟁력도 강화했다. (2) ASIC에서는 AWS Trainium와 마이크로소프트 Maia의 파트너사로 알려졌는데 최근 구글과 차세대 TPU를 포함한 두 종류의 AI 칩 개발을 논의 중이라는 보도도 나왔다. 그 외에 최근 엔비디아로부터 $2B 지분 투자를 유치하고 NVLink Fusion, AI-RAN 등 분야에서 협력 계획을 발표했다. 엔비디아 GPU 생태계와 하이퍼스케일러 자체 ASIC에 모두 걸쳐 있는 기업으로 높은 경쟁력이 입증된다.

투자포인트

(1) AI네트워크와 ASIC 등 전방산업의 높은 성장성이 실적에도 본격적으로 반영될 예정이다. Marvell은 FY27 Interconnect 매출이 70% 이상 성장하고, ASIC 사업도 FY27 20% 이상 성장한 뒤 FY28에는 두 배 이상 확대될 것으로 전망하고 있다. AI 클러스터가 커지고 Agentic AI로 데이터 이동량과 연산이 늘어날수록 광통신, Switch, ASIC 수요가 함께 증가하는 구조이다. (2) KV Cache 폭증을 계기로 CXL 시장 개화 가능성도 기대할 수 있다. 동사는 CXL 2.0 기반 컨트롤러와 스위치, CXL 3.x 기반 Structera S 30260을 통해 서버의 메모리 확장부터 랙 단위의 메모리 Fabric까지 대응 가능하다. AI 인프라의 여러 고성장 영역에 동시에 노출돼 있다는 점을 Marvell의 가장 큰 투자 매력으로 평가한다.

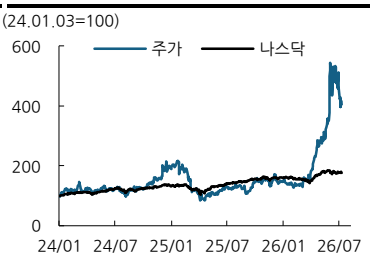
Key Data

국가	USA
거래소	NASDAQ
GICS(부문)	정보기술
GICS(산업)	반도체 & 반도체 장비
시가총액($ Bn)	165.06
시가총액(조원)	245.44
현재주가(07/17, $)	210.96
52주 최고가($)	329.88
52주 최저가($)	61.44
주요주주	
FMR LLC	9.64%
VANGUARD GROUP	9.15%

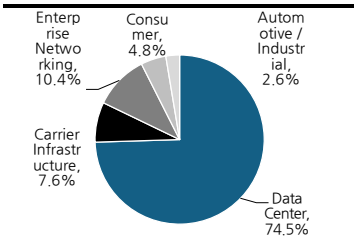
주가추이 및 상대강도

(%)	절대수익률	상대수익률
1M	-34.8	-32.9
3M	35.1	30.8
6M	134.5	126.0

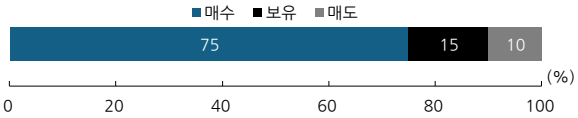
주가차트



사업부별 매출 비중



블룸버그 투자의견 비율



Financial Data

($ Bn)	FY2025	FY2026	FY2027F	FY2028F
매출액	57.7	81.9	114.0	166.7
영업이익	1.7	2.9	4.2	6.5
영업이익률(%)	28.9	35.3	36.8	38.8
순이익	1.4	2.5	3.7	5.7
EPS($)	1.6	2.8	4.0	6.9
증감률(%)	4.0	80.9	42.3	71.3
ROE(%)	-6.3	19.3	12.6	16.8
PER(배)	-	52.1	58.3	34.1
PBR(배)	7.3	4.7	11.6	9.8
PSR(배)	16.9	8.3	17.9	12.3

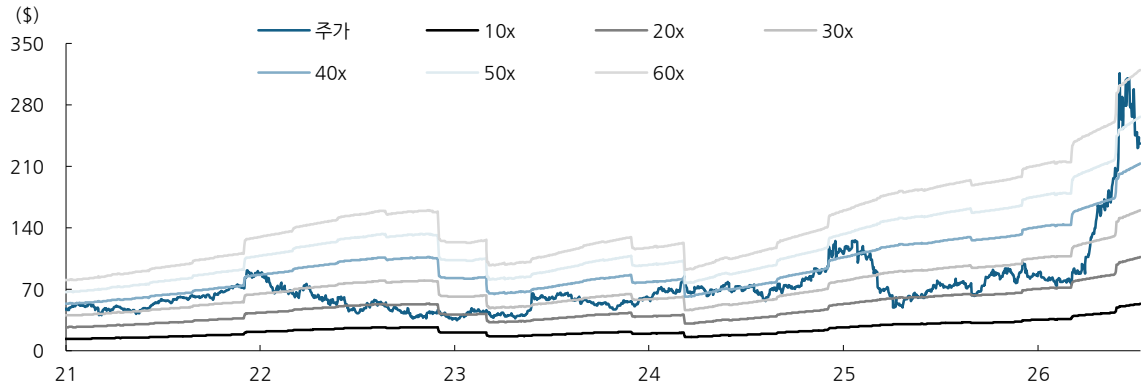
자료: Bloomberg, DS투자증권 리서치센터, Non-GAAP 기준, 비율은 GAAP 기준

표24 마벨 테크놀로지 실적 테이블

($ Mn, %)	FY1Q26	FY2Q26	FY3Q26	FY4Q26	FY1Q27	FY2Q27F	FY3Q27F	FY25	FY26	FY27F
매출	**1,895**	**2,006**	**2,075**	**2,220**	**2,418**	**2,699**	**3,015**	**5,767**	**8,196**	**11,554**
%YoY	*63.3*	*57.6*	*36.8*	*22.1*	*27.6*	*34.8*	*45.6*	*114.2*	*42.1*	*41.0*
%QoQ	*4.3*	*5.8*	*3.4*	*7.0*	*9.0*	*11.6*	*11.7*			
사업부별 매출										
Data Center	1,441	1,491	1,518	1,651	1,833	2,143	2,448	4,164	6,100	9,210
Infrastructure	138	130	168	193	205	194	195	338	629	756
Networking	178	194	237	251	256	247	255	626	859	997
Consumer	63	116	117	92	76	80	83	316	387	295
Automotive/Industrial	76	76	35	33	33	34	35	322	219	295
매출총이익	**952**	**1,011**	**1,070**	**1,148**	**1,261**	**1,406**	**1,564**	**2,382**	**4,181**	**6,023**
GPM(%)	50.3	50.4	51.6	51.7	52.1	52.1	51.9	41.3	51.0	52.1
%YoY	*80.4*	*72.0*	*206.2*	*25.1*	*32.4*	*39.2*	*46.2*	*3.9*	*75.5*	*44.1*
영업이익	**647**	**699**	**753**	**792**	**847**	**990**	**1,131**	**1,665**	2,891	4,189
OPM(%)	34.2	34.8	36.3	35.7	35.0	36.7	37.5	28.9	35.3	36.3
%YoY	*139.6*	*110.7*	*67.3*	*29.3*	*30.8*	*41.7*	*50.1*	*27.1*	*73.7*	*44.9*
순이익	540.0	585.5	655.0	685.1	718.0	848.4	978.8	**1,377.3**	**2,465.6**	***3,651.3***
NPM(%)	28.5	29.2	31.6	30.9	29.7	31.4	32.5	23.9	30.1	31.6
%YoY	*161.2*	*119.9*	*75.6*	*28.9*	*33.0*	*44.9*	*49.4*	*5.1*	*79.0*	*48.1*
EPS	**0.62**	**0.67**	**0.76**	**0.80**	**0.80**	**0.93**	**1.07**	**1.57**	2.85	*4.04*
%YoY	*158.3*	*123.3*	*76.7*	*33.3*	*29.0*	*39.0*	*40.7*	*4.0*	*81.5*	*41.8*

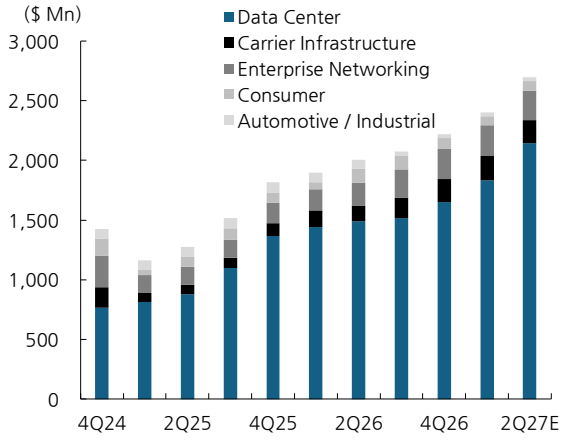
자료: Bloomberg, DS투자증권 리서치센터

그림70 마벨 테크놀로지 12MF PER 밴드 추이



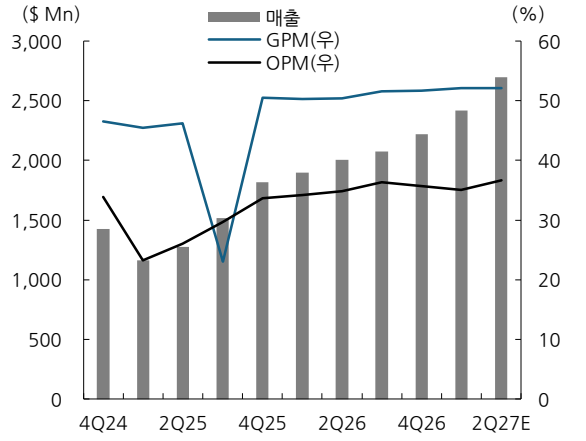
자료: Bloomberg, DS투자증권 리서치센터

그림71 마벨 테크놀로지 부문별 연간 매출액 추이



자료: Bloomberg, DS투자증권 리서치센터

그림72 마벨 테크놀로지 연간 매출액, 매출총이익률, 영업이익률



자료: Bloomberg, DS투자증권 리서치센터

마무리하며: 2026년 하반기 미국 주식시장 전망

투자의 시대. 'AI'에만 집중하자

AI투자가 견인하는 미국 경제

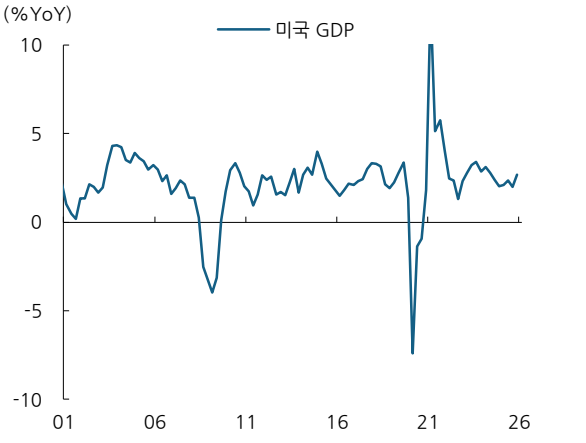
현재 미국 경기의 핵심 동력은 AI

미국 경기는 투자 사이클의 한복판에 위치해 있다. 그리고 투자는 미국 경제를 본격적으로 견인하기 시작했다. 2026년 1분기 미국의 실질 GDP는 전기 대비 연율 2.1% 성장하며 전분기 0.5%에서 반등했는데 민간투자가 1.35%p 기여한 반면 미국 경제의 근간인 민간소비의 기여도는 0.37%p에 그쳤다.

민간투자의 세부 항목을 보면 이번 투자 사이클의 성격은 더욱 명확해진다. 부동산 경기와 관련된 주거용 투자와 구조물 투자의 실질 GDP 성장률을 각각 0.3%p, 0.13%p 낮췄음에도 AI 투자 효과가 상당 부분 반영된 정보처리기기와 지적재산권은 각각 0.77%p, 0.74%p 기여했다. 두 항목의 합산 기여도가 1.5%p에 달했다는 점은 AI투자가 미국 경기를 결정하는 핵심 변수로 부상했음을 알 수 있다.

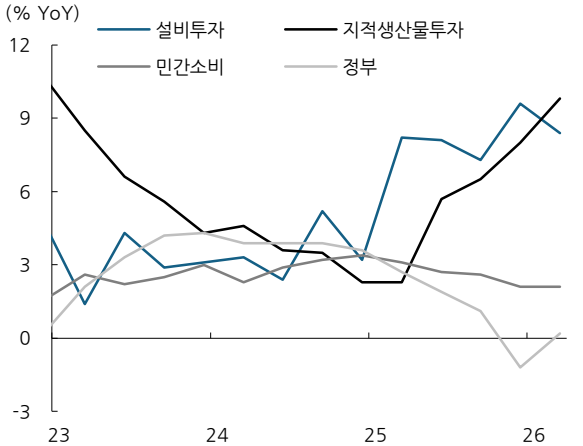
경기선행지표에서도 유사한 흐름이 관측된다. 5월 OECD 미국 경기선행지수는 101.0p를 기록하며 확장 국면을 이어갔다. 세부적으로는 내구재 신규주문과 ISM 제조업 PMI 등 투자 관련 지표가 강세를 보인 반면, 소비자기대지수와 주택착공은 부진했다. 선행지수, 동행지수에서 관측되는 지표 간 엇갈림은 오히려 일관된 메시지를 보낸다. 현재 미국 경기의 핵심 동력은 소비나 부동산이 아니라 AI투자이다.

그림73 1Q26 미국 GDP는 전기 대비 연율 2.1% 성장하며 반등



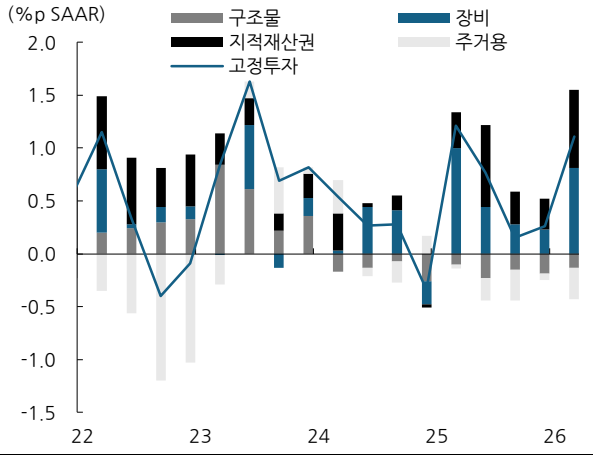
자료: Bloomberg, DS투자증권 리서치센터

그림74 부진한 소비를 대신해서 투자가 경제 성장을 주도



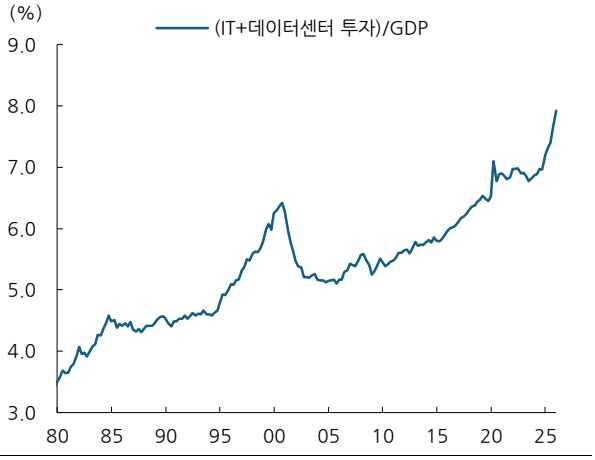
자료: Bloomberg, DS투자증권 리서치센터

그림75 투자 중에서도 AI 관련 투자가 핵심



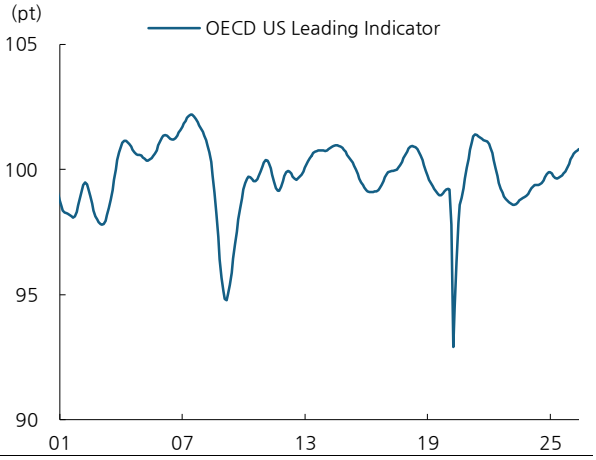
자료: Bloomberg, DS투자증권 리서치센터

그림76 과거 투자 사이클을 뛰어넘은 AI투자의 체급



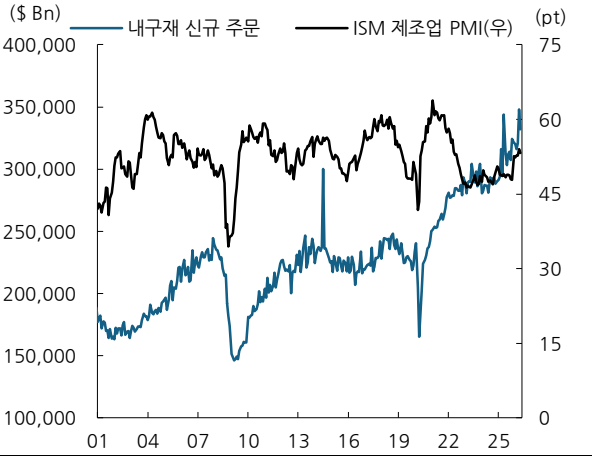
자료: Bloomberg, DS투자증권 리서치센터

그림77 OECD 미국 경기선행지수도 확장 국면에 위치



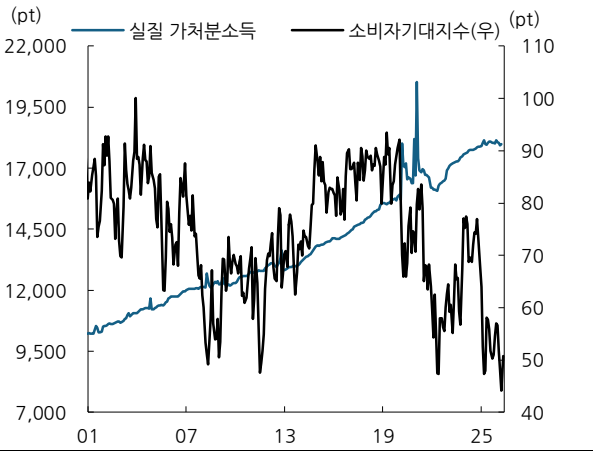
자료: Bloomberg, DS투자증권 리서치센터

그림78 투자 관련 지표는 성장세는 명확한데



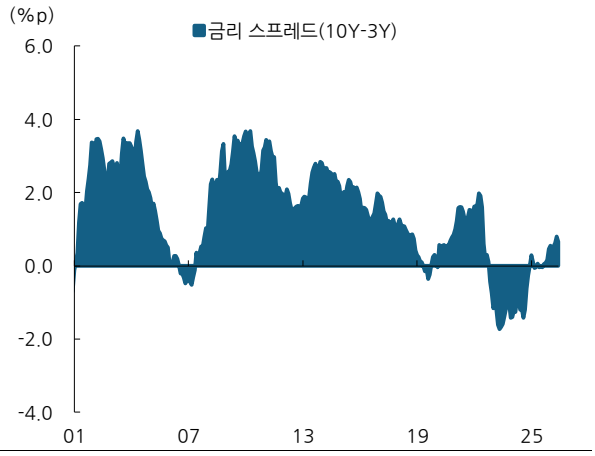
자료: Bloomberg, DS투자증권 리서치센터

그림79 소비의 선행지표인 소비 심리와 소득은 둔화 중



자료: Bloomberg, DS투자증권 리서치센터

그림80 금리 스프레드 정상화도 투자의 시대임을 방증



자료: Bloomberg, DS투자증권 리서치센터

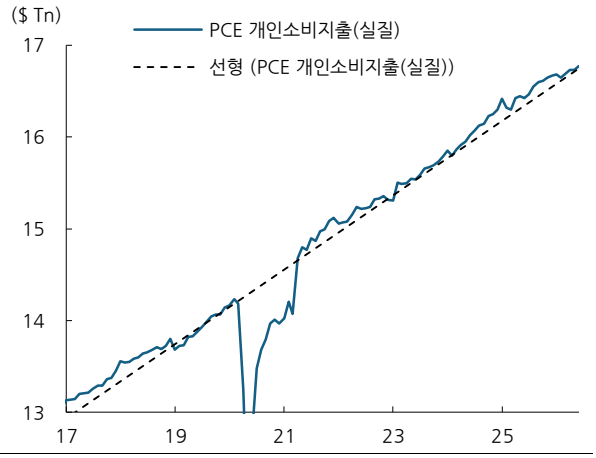
하반기도 AI투자 사이클에서 승부를 봐야한다

소비가 좋아지기 어렵다 하반기 역시 AI에서 승부 봐야함

하반기에도 AI투자 주도의 성장은 지속될 것이며 더욱 강화될 것이다. 밋밋한 고용, 정체된 임금상승률, 높은 금리 수준 등을 고려하면 소비는 한동안 장기 추세를 하회할 가능성이 높다. 반면 AI투자(정보처리장치, 소프트웨어, 데이터센터)는 올해 1분기 전년 동기 대비 16.6% 증가했는데, 이는 닷컴버블 당시 고점인 15.0%를 웃도는 수준이다. AI 투자 사이클의 규모를 과소평가해서는 안된다.

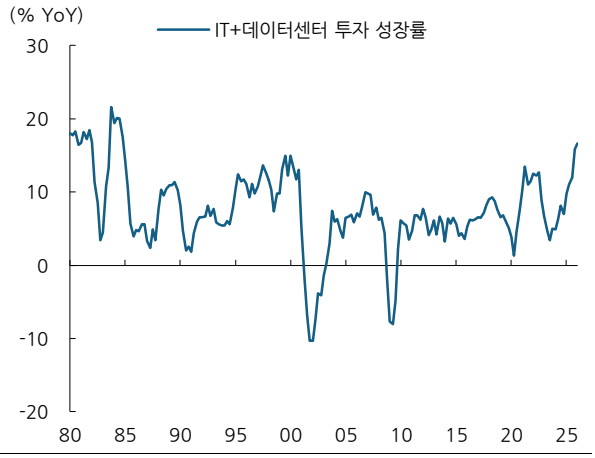
투자의 역사를 돌아보면 주식시장의 자금은 성장을 찾아 움직였으며, 성장이 불균형한 시기에 쏠림의 강도가 강했다. 현재처럼 (1) 실물투자와 이익 성장이 AI 관련 산업에 집중되고 (2) AI와 Non-AI의 실적 차별화가 지속되며 (3) AI 관련 기업의 밸류에이션이 과도하지 않다면 여전히 승부는 AI에서 봐야 할 것이다.

그림81 장기 추세선을 이탈한 소비지표



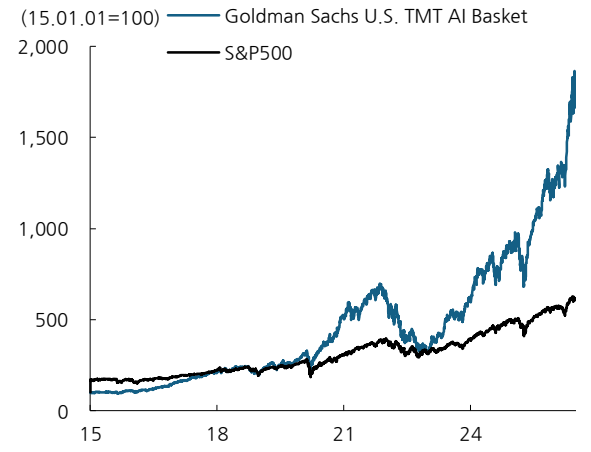
자료: Bloomberg, DS투자증권 리서치센터

그림82 닷컴 버블을 넘어선 AI투자의 성장 속도



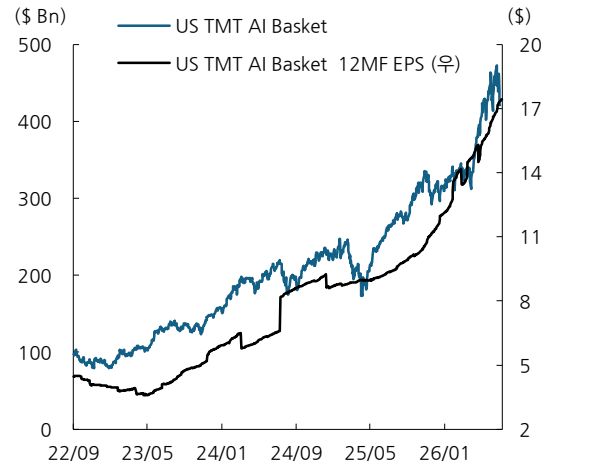
자료: Bloomberg, DS투자증권 리서치센터
주: IT = Information Processing Equipment + Software + R&D

그림83 AI 관련 기업의 가파른 주가 상승



자료: Bloomberg, DS투자증권 리서치센터
주: Goldman Sachs U.S. TMT AI Basket는 미국 상장된 AI 관련 기업 100여개를 지수화

그림84 주가 상승만큼 가파른 실적 성장



자료: Bloomberg, DS투자증권 리서치센터
주: Goldman Sachs U.S. TMT AI Basket는 미국 상장된 AI 관련 기업 100여개를 지수화

하반기 하이퍼스케일러 CapEx가 견고할 세 가지 이유

1. 돈이 되기 시작한 AI

1) AI 모델 ARR 급증
2) 클라우드 매출 성장
→ 하이퍼스케일러는 투자를 줄일 이유가 없다

AI 인프라 투자 사이클의 중심에는 하이퍼스케일러의 CapEx가 있다. 이들은 AI모델의 학습과 추론을 위한 AI데이터센터를 구축하며, 투자 수요는 AI가속기에서 메모리, 네트워크, 전력 인프라까지 밸류체인 전반으로 확산된다. 따라서 하이퍼스케일러의 CapEx는 AI인프라 투자 사이클의 지속성과 강도를 결정하는 핵심 변수다. 후술할 세 가지 근거를 바탕으로 주요 하이퍼스케일러의 공격적인 CapEx 기조가 지속될 것으로 전망한다.

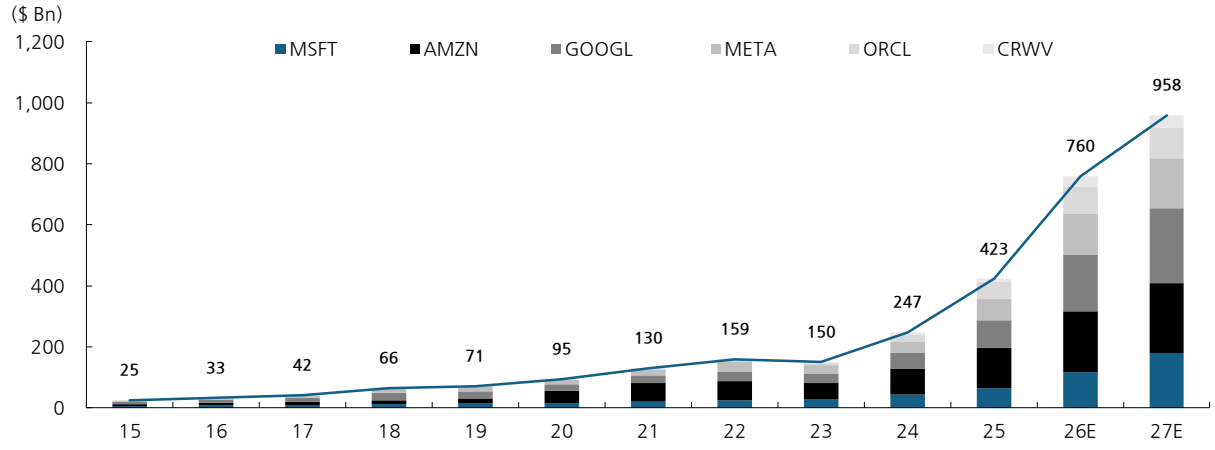
첫 번째 근거는 AI 투자의 수익화 가시성이 높아지고 있다는 점이다. 작년까지 하이퍼스케일러의 CapEx가 미래 수요와 기술 주도권을 선점하기 위한 선제적 투자 성격이 강했다면, 최근 (1) 전방기업인 AI모델 개발사의 ARR(Annual Recurring Revenue, 연간 반복 매출) 증가와 (2) AI 수요에 힘입은 클라우드 서비스의 매출 성장 가속화가 확인되고 있다. 이는 AI인프라 투자가 실제 매출로 전환되기 시작했으며, 하이퍼스케일러의 투자 회수 가능성도 높아지고 있음을 보여준다.

OpenAI, Anthropic의 ARR 증가가 이를 단적으로 보여준다. OpenAI와 Anthropic의 ARR은 '24년말 각각 $5.5B, $1B에서 '25년 $21B, $9B를 거쳐 '26년 5월 $25B 이상, $47B까지 증가한 것으로 전망되고 있다. 두 회사의 수치는 특정 시점이 매출을 연환산한 값이며, 산정 방식과 기준 시점도 달라 직접 비교에는 한계가 있지만 그럼에도 AI서비스의 매출이 빠르게 증가하고 있다는 방향성은 분명하다.

AI모델 개발사의 수익성 개선도 가시화되고 있다. WSJ에 따르면 Anthropic은 투자자들에게 2026년 2분기 $10.9B의 매출과 $559M의 영업이익을 전망했다. 아직 확정 실적은 아니지만, AI 모델 개발사의 가파른 매출 성장이 영업 레버리지로 이어질 가능성을 보여준다는 점에서 의미가 있다. 또한 전방기업인 이들의 수익성과 재무 여력이 개선되면 더 큰 규모의 컴퓨팅을 구매할 여력이 커지고, 이는 하이퍼스케일러의 클라우드 매출과 AI인프라 투자 회수 가시성을 높인다.

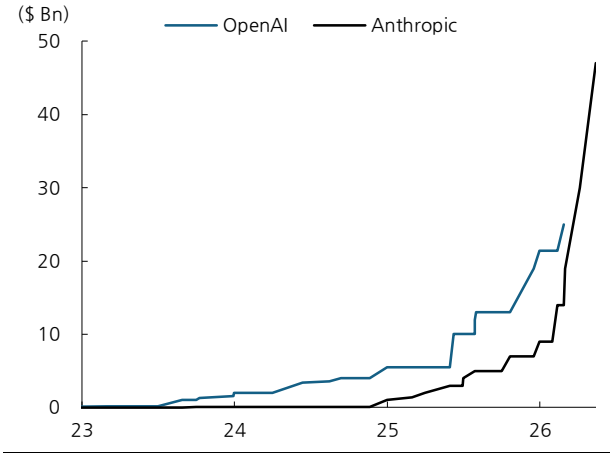
하이퍼스케일러의 높은 클라우드 매출 성장률도 AI 투자의 경제적 근거를 강화한다. 2026년 1분기 하이퍼클라우드 3사 Alphabet(GCP), Microsoft(Azure), Amazon(AWS)의 클라우드 매출은 전년 동기 대비 각각 63%, 40%, 28% 증가했다. 가장 낮은 성장률을 기록한 AWS조차 28% 성장했다는 점은 클라우드 수요가 전반적으로 강하다는 사실을 증명한다. 그리고 이러한 성장성은 공격적인 CapEx를 정당화한다.

그림85 AI투자 사이클의 지속성과 강도를 결정하는 하이퍼스케일러 CapEx 추이



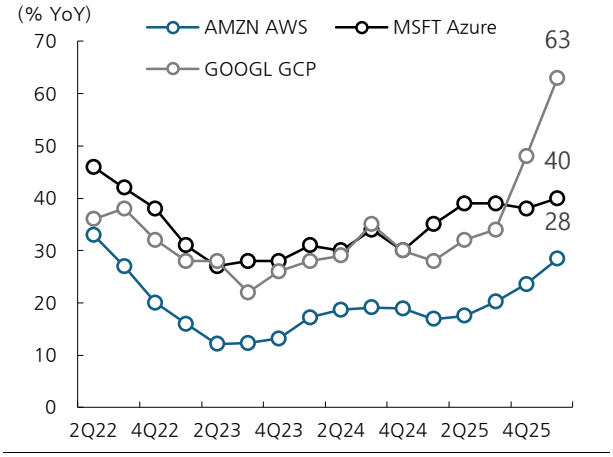
자료: Bloomberg, DS투자증권 리서치센터
주: 26년, 27년은 Bloomberg 컨센서스

그림86 상위 AI모델 개발사의 ARR 추이



자료: Epoch AI, DS투자증권 리서치센터

그림87 하이퍼스케일러 3사의 클라우드 매출 성장률



자료: 각사, DS투자증권 리서치센터

2. AI모델의 발전 방향이 만드는 구조적 컴퓨팅 수요

AI모델을 늘 컴퓨팅을 더 쓰는 방식으로 발전해옴

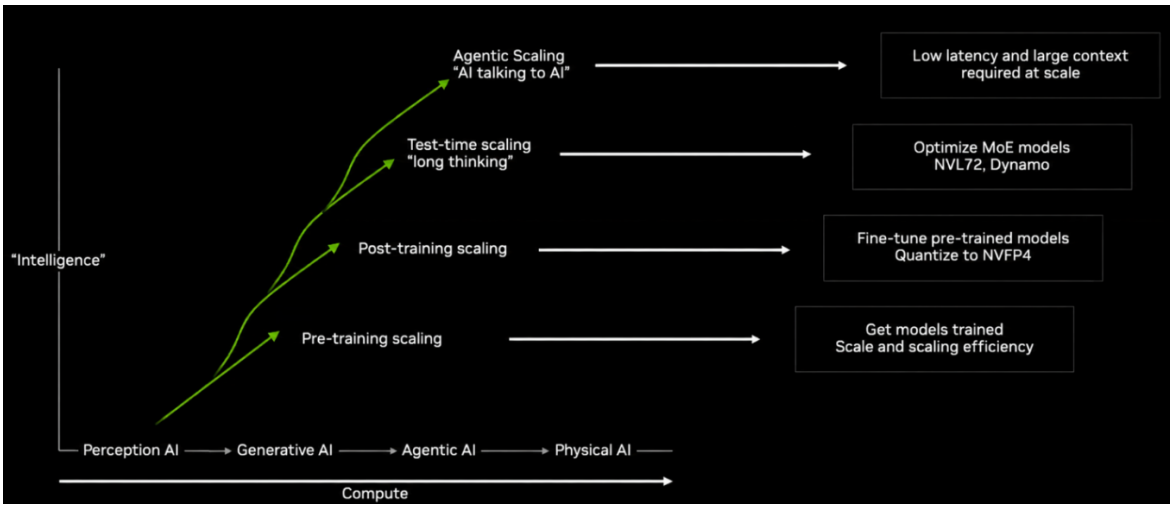
두 번째 근거는 앞서 설명했듯이 AI모델의 발전 방향 자체가 더 많은 컴퓨팅을 요구한다는 점이다. Agentic System에서 Multi-Agent System으로 발전해가는 과정에도 컴퓨팅의 증가는 필연적이다.

현재까지 AI 모델은 Pre-training, Post-training, Test-time Scaling을 거쳐 Agentic Scaling으로 확장되는 네 가지 성능 향상 축을 바탕으로 발전해왔다. 그리고 각 단계는 새로운 컴퓨팅 수요를 만들어냈다. 가장 먼저 확립된 Pre-training Scaling은 모델 파라미터, 학습 데이터, 훈련 컴퓨팅을 일정한 균형으로 확대할수록 모델이 성능이 개선된다는 법칙으로 AI모델이 더 높은 성능을 구현하려면 더 많은 컴퓨팅이 필요하다는 방향성이 명확하다.

두 번째 축인 Post-training Scaling은 사전학습 이후 Fine-tuning과 강화학습 등을 통해 모델의 유용성과 추론 능력을 높이는 과정이다. 과거에는 사전학습보다 필요한 컴퓨팅 규모가 작았지만, 최근에는 강화학습과 합성데이터 생성이 확대되면서 독립적인 컴퓨팅 수요로 부상하고 있다.

세 번째 축인 Test-time Scaling은 모델이 더 긴 추론 과정을 수행하거나 여러 해결 경로를 탐색, 검증할수록 성능이 개선된다는 법칙이다. 이 과정에서 하나의 요청에 응답하는 데 필요한 연산량이 크게 증가했으며, 추론 연산은 단순한 서비스 수행 비용을 넘어 새로운 성능 향상 축으로 자리 잡았다.

그림88 AI모델의 발전방향 정리 (1)



자료: NVDIA, DS투자증권 리서치센터

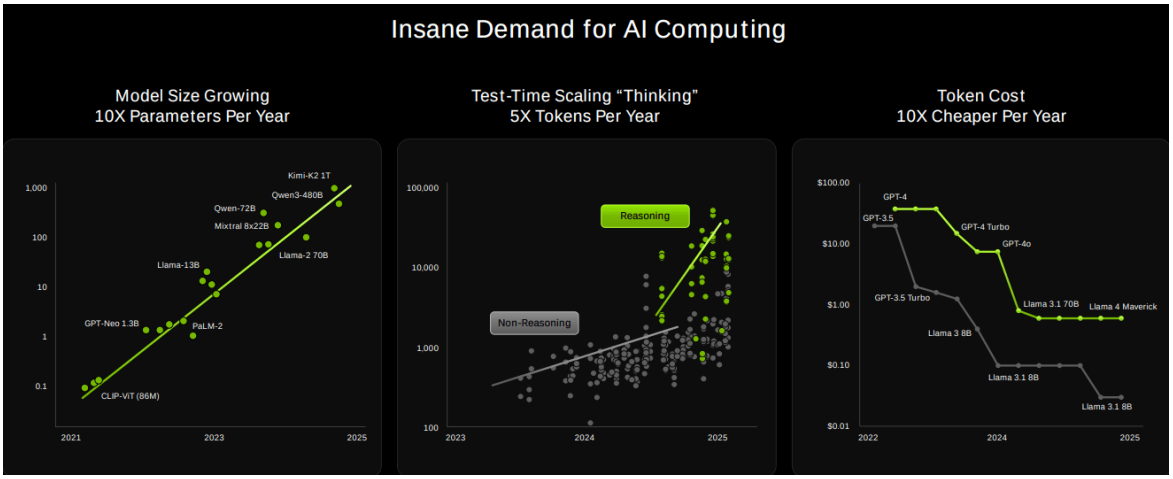
현재 시작된 Agentic AI도 컴퓨팅을 크게 늘리는 방향

네 번째 축인 Agentic Scaling은 개별 모델을 넘어, 모델이 도구를 활용하거나 복수의 에이전트가 반복적으로 협업하며 시스템 전체의 문제 해결 능력을 높이는 방식이다. 에이전트는 사용자 요청을 여러 단계로 나누고 계획 수립, 검색, 도구 호출, 코드 실행, 결과 검증과 수정을 반복한다. 이에 따라 작업당 토큰 사용량과 모델 호출 횟수, 실행시간이 늘어나면서 필연적으로 컴퓨팅 수요가 증가하게 된다. 더 나아가 복수의 에이전트가 협업하는 구조가 확산되면 작업당 컴퓨팅 사용량은 더욱 빠르게 증가할 수 있다.

AI 워크로드의 변화는 컴퓨팅 수요뿐 아니라 AI 인프라의 범위도 확장시켜왔다. 초기 AI CapEx가 GPU, HBM 등 모델 학습 인프라 확보에 집중됐다면, Reasoning과 Agentic System에서는 긴 컨텍스트와 KV Cache 증가, 반복 모델 호출과 데이터 이동으로 메모리, 스토리지, Scale-up 및 Scale-out 네트워크, CPU의 중요성이 함께 높아진다. 이는 투자 대상의 범위와 시스템당 투자액을 동시에 확대하며, CapEx에 지속적인 상방 압력을 가한다.

마지막으로 컴퓨팅 사용량을 늘리는 방식의 성능 향상은 LLM 사용 비용의 증가로 지속가능성에 대한 의문을 제기할 수 있다. 그러나 AI모델과 인프라는 성능 향상과 동시에 토큰 당, 작업당 추론 비용을 낮추는 방향으로 발전하고 있다. 일례로 Caching은 중복 연산을 줄이고, MoE와 저정밀 연산은 필요한 연산량과 처리 비용을 낮춰 동일 비용으로 더 많은 토큰을 처리하게 한다. 이러한 비용 효율화는 사용량 증가에 따른 비용 부담을 완화하고 AI 확산에 기여하여 오히려 총 컴퓨팅 수요와 CapEx 사이클의 지속성을 높이는 요인이 된다.

그림89 AI모델의 발전방향 정리 (2)



자료: NVDIA, DS투자증권 리서치센터

3. ROIC - 자본비용을 통해 확인한 CapEx의 경제적 정당성

세 번째 근거는 현재의 CapEx 투자가 경제적으로도 정당화된다는 점이다. 일각에서는 미국 국채 10년물 금리가 4.5% 안팎에 이르는 등 자본비용이 높아 대규모 투자가 지속되기 어렵다고 주장한다. 그러나 금리 레벨만을 고려하는 절대적 긴축은 투자의 성장성과 기대수익률을 반영하지 못한다.

중요한 것은 AI투자의 예상 ROIC가 기업의 자본비용을 비교한 '상대적 긴축' 정도이다. 투자수익률이 자본비용을 충분히 웃돈다면 높은 금리에도 투자를 축소할 유인은 크지 않다. 이를 기준으로 판단할 때 하이퍼스케일러가 현재 CapEx를 급격히 축소할 유인은 제한적이라고 본다.

아래 표에 근거하면 하이퍼스케일러의 AI CapEx에 대한 ROIC는 13.9% 로 계산된다. 이는 추정 자본비용 9.34%는 물론이고 3사의 5년만기 회사채 금리 평균인 4.7%를 넉넉히 상회한다는 점에서 충분한 수익성이 확인되는 것이다.

물론 아래 ROIC 계산식에서 (1) CapEx 와 RPO 간의 인과관계가 정확히 대응하지 않으며 (2) CapEx 중 향후 2년간 AI 매출 등 일부 항목이 가정, 추정치에 근거하는 등의 한계는 있다. 그럼에도 보수적인 수치를 적용한다면 투자의 경제성을 판단하는 지표로 활용될 수 있을 것이며, 향후 Cloud 부문의 OPM, 하이퍼스케일러 3사의 자본비용 등 변수가 변화를 반영하면 시의적절한 투자 판단을 내리는데 도움이 될 것으로 기대한다.

표25 하이퍼스케일러 3사의 ROIC 추정

#	Description	Value (USD Bn, %)	비고
1	2025 CapEx	287.8	- 3사 합산
2	2026 CapEx	501.6	- 3사 합산
3	2Y CapEx	789.5	= #1 + #2
4	CapEx 중 AI 비율	80%	- 80% - 근거 (1) MSFT: CapEx 중 80%를 클라우드/AI에 지출 - 근거 (2) GOOGL: 약 80%를 AI인프라에 지출
5	**2Y AI CapEx**	**631.6**	**= #3 * #4**
6	RPO	1,460.0	- 1Q26
7	AI 관련 RPO	893.4	- 3사 RPO 비례하여 가중평균
8	2Y AI 매출(RPO에서 발생분)	446.7	- 전체 AI 관련 RPO 중 50%가 2년 내 인식된다고 가정 - 근거 (1) MSFT "약 25%를 1년 안에 인식 예정" - 근거 (2) GOOGL "50%이상을 24개월 내 인식 예정"
9	2Y AI 매출(전체)	638.1	- RPO 외에서 발생하는 AI매출을 30%로 가정 - 근거: 기업 코멘트
10	2Y 영업이익	238.0	Cloud OPM: 37.3% - 3사 가중평균
11	2Y 세후 영업이익	195.8	- Federal Corporate Tax: 21%
12	2Y ROIC	29.8%	= #11 / #5
13	1Y ROIC	13.9%	- 연환산
14	1Y Cost of Capital	9.34%	- NYU Damodaran의 'Software (System & Application)' 업종 자본비용

자료: DS투자증권 리서치센터

연준의 하반기 금리 인상? 가성비가 너무 안나온다

Fed의 연내 정책금리 동결 전망

7월 17일 CME Fed Watch 기준으로 시장은 9월 금리 인상 가능성을 동결보다 높게 반영하고 있으며, 올해 말까지 금리를 동결할 가능성은 18%에 불과하다. 따라서 시장의 기본 시나리오는 연내 금리 인상이지만 당사는 아래 세 가지 이유로 정책 금리의 연내 동결을 전망한다. 금리 인상의 편익이 크지 않다는 생각이다.

(1) 물가 상승이 지속되기 어려운 환경

첫번째 이유는 인플레이션의 지속성을 확신하기 어렵다는 점이다. 과거 연준이 적극적으로 금리 인상을 단행했던 시기는 물가와 임금이 동반 상승했던 구간이다. 물가 상승이 임금 인상 요구로 이어지고, 다시 기업의 가격 인상을 자극하는 Wage-Price Spiral(임금-물가 악순환) 우려가 고조되던 상황에서 긴축적인 통화정책으로 대응을 했던 것이다. 하지만 현재 시간당 평균임금 상승률이 둔화되는 추세이며, 소비도 부진하기에 구조적인 물가 상승을 확신하기 어려워 보인다.

이에 더해 5월까지 WTI 선물 기준 100달러를 상회했던 유가가 전쟁 중단 합의가 발표되자 한달 만에 60달러 후반대까지 하락한 점도 시사하는 바가 크다. 셰일혁명의 영향으로 2019년부터 미국은 공식적으로 에너지 순수출국이 되었다. 과거 오일쇼크와 달리 에너지 가격의 통제 가능성이 높아진 상황이며, 에너지 가격이 인플레이션을 주도하는 과거의 국면이 반복되기 어려움을 시사한다.

(2) 금리 인상이 현재의 유가, 칩플레이션에 미치는 영향은 제한적

두 번째 이유는 금리 인상의 물가 억제 효과가 제한적일 수 있다는 점이다. 5월 PCE 세부 품목을 보면 상승률 상위 항목에는 '휘발유 및 기타 에너지 상품'(+40.5%), '금융 서비스 및 보험'(+7.8%), '운송 서비스'(+6.6%), '여가용 제품 및 자동차'(+6.4%), '기타 내구재'(+5.9%)가 있다. 이 중 금리 인상을 통해 총수요를 둔화시켜 물가를 억제할 수 있는 품목은 '여가용 제품 및 자동차', '기타 내구재' 정도 밖에 없다는 점에서 금리 인상의 효과가 크지 않을 수 있다고 생각한다.

또한 '여가용 제품 및 자동차 품목'에 PC, TV 등이 포함되고 '기타 내구재'에 스마트폰이 포함되는 등 두 항목에는 칩플레이션의 영향도 일부 반영되어 있다. 금리를 0.25% 혹은 0.5% 인상해도 반도체 수요에 미치는 영향이 크지 않을 것을 고려하면 금리 인상이 과거만큼 물가를 억제하지 못할 가능성이 있다.

(3) 민간이 체감하는 긴축은 시작됨. 앞으로도 계속 될 것

세 번째 이유는 정책금리 수준에 비해 이미 긴축적인 금융환경이다. 투자가 확대되는 시기에는 기업의 자금 수요가 늘면서 정책금리에 변화가 없어도 장기금리와 시장 조달금리에 상승 압력이 발생할 수 있다. 대표적으로 구글, 메타 등이 채권 발행에 나서는 흐름은 정책금리와 별개로 민간의 긴축 정도를 강화할 수 있다.

이는 국채금리를 보면 알 수 있다. 올해 초 미국 5년, 10년물 국채금리는 정책금리와 비슷한 수준이었는데 현재는 약 50bp, 80bp 이상 스프레드가 확대된 상황이다. 회사채 금리가 동일 만기의 국채금리에 크레딧 스프레드를 가산해서 계산되는 만큼, 기업이 느끼는 긴축 정도는 이미 기준금리 2~3회가 반영된 것으로 볼 수도 있다. 또한 이러한 긴축 강도는 모기지금리 등을 통해 가계에도 영향을 미칠 것이다. 물론 현재의 국채금리에는 금리 인상 전망이 이미 반영된 부분도 있지만, 실제 기준금리를 인상하지 않더라도 빅테크 등의 회사채 발행 등의 여파로 시중 금리는 높은 수준을 유지할 것으로 전망하며 기업이 연초대비 느끼는 긴축 정도는 이미 금리 인상을 1~2회 반영한 수준을 될 것으로 예상한다.

위 근거들을 종합하면 연준 입장에서 현재 정책금리를 인상해서 얻을 수 있는 편익이 크지 않을 것으로 판단한다. 특히 시장이 전망하는 하반기 금리 인상이 인상 사이클에서의 추가 인상이 아니라 금리 인상 사이클의 시작이라는 점에서 가성비가 매우 좋지 못한 선택으로 판단한다.

그림90 CME Fed Watch: 9월 금리 인상을 반영

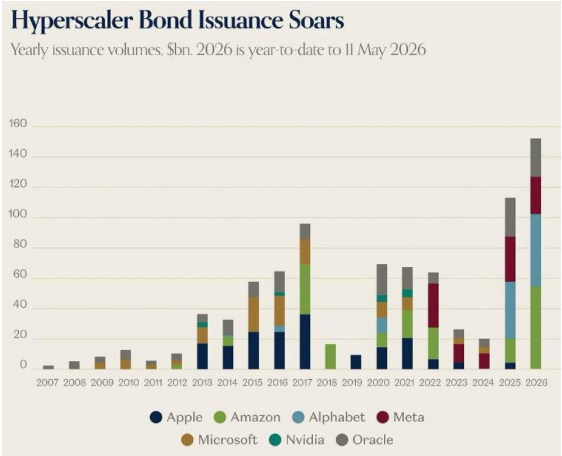
Meeting Date	3.25-3.50	350-375	375-400	400-425	425-450	450-475
2026-07-29	0.00%	85.60%	14.40%	0.00%	0.00%	0.00%
2026-09-16	0.00%	38.50%	53.50%	7.90%	0.00%	0.00%
2026-10-28	0.00%	30.40%	50.40%	17.50%	1.70%	0.00%
2026-12-09	0.00%	18.50%	42.60%	30.40%	7.90%	0.70%
2027-01-27	6.00%	26.20%	38.70%	23.10%	5.60%	0.40%
2027-03-17	1.80%	12.00%	29.90%	34.00%	17.90%	4.00%
2027-04-28	9.80%	26.00%	33.10%	21.40%	7.10%	1.10%
2027-06-09	3.40%	13.70%	27.80%	30.30%	17.90%	5.60%
2027-07-28	4.20%	14.80%	28.00%	29.30%	17.00%	5.20%
2027-09-15	5.40%	16.20%	28.10%	28.00%	15.70%	4.80%
2027-10-27	6.40%	17.30%	28.10%	26.90%	14.70%	4.40%
2027-12-08	7.50%	18.50%	28.00%	25.60%	13.60%	4.00%

자료: CME, DS투자증권 리서치센터

그림91 미국 시간당 평균 임금 상승률: 둔화 중

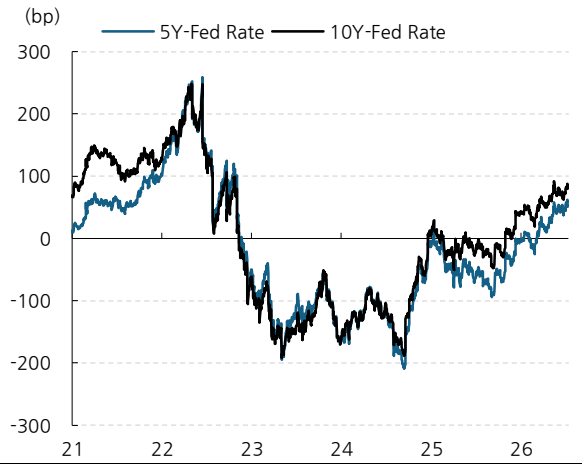
자료: Bloomberg, DS투자증권 리서치센터

그림92 하이퍼스케일러의 채권 발행 추이



자료: A16Z, DS투자증권 리서치센터
주: 2026년은 5월 11일까지의 채권발행액

그림93 미국채 5년, 10년 금리 - 정책금리 스프레드



자료: Bloomberg, DS투자증권 리서치센터

2026년 하반기 S&P500 밴드: 7,300~8,700pt

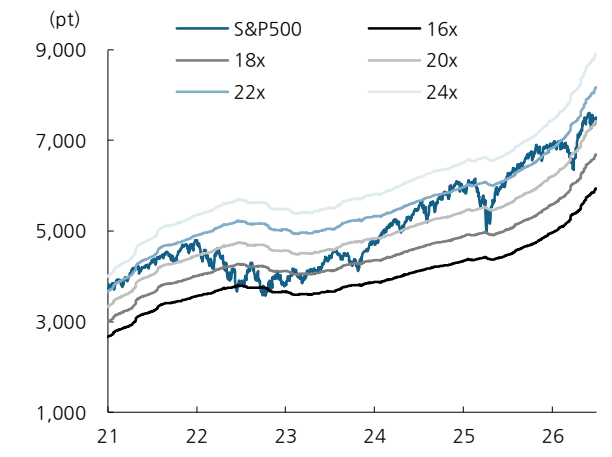
견고한 AI CapEx는 지수의 완만한 상승을 의미. AI인프라 종목 찾기 전략 유효

하반기 S&P500 7,300~8,700pt 전망

하반기 S&P500 예상 밴드로 7,300~8,700p로 전망한다. 기준 시나리오의 연말 목표치는 12개월 선행 EPS $395에 목표 PER 21배를 적용한 8,300p이다.

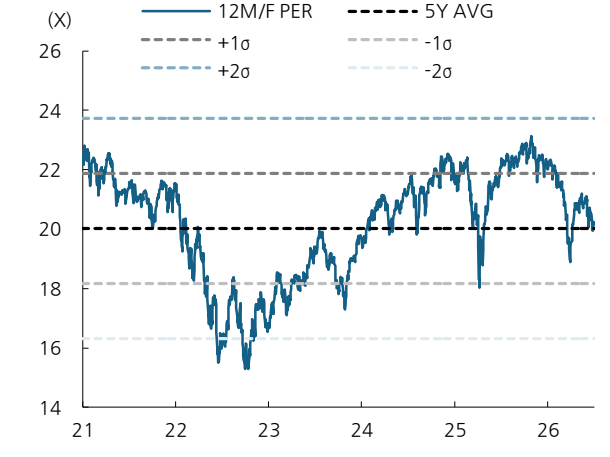
(1) Agentic AI가 확산될수록 클라우드 사업의 수익성이 개선될 것이고 (2) 전방 기업인 AI모델사가 돈을 벌기 시작했다. (3) 높은 컴퓨팅 수요가 여전하기에 하이퍼스케일러는 투자를 멈출 이유가 없다. 그리고 하이퍼스케일러의 CapEx는 중장기 성장 기반을 강화하는 대신 현재의 주주환원 여력을 제약하기에 빅테크 비중이 높은 S&P500은 완만한 속도의 상승을 전망한다. 반면 투자자금이 집중되는 AI인프라 기업은 직접적인 수혜를 받을 것이며, 하반기에도 AI인프라 종목 찾기 전략이 유효할 것이다.

그림94 S&P500 12MF PER 밴드



자료: Bloomberg, DS투자증권 리서치센터

그림95 S&P500 12MF PER 평균과 표준편차



자료: Bloomberg, DS투자증권 리서치센터

표26 S&P500 밴드 추정

		12MF PER								
		-1σ		-0.5σ		5년평균		+0.5σ		+1σ
		18	18.5	19	19.5	20	20.5	21	21.5	22
12MF EPS	365	6,570	6,753	6,935	7,118	7,300	7,483	7,665	7,848	8,030
현재	370	6,660	6,845	7,030	7,215	7,400	7,585	7,770	7,955	8,140
	375	6,750	6,938	7,125	7,313	7,500	7,688	7,875	8,063	8,250
	380	6,840	7,030	7,220	7,410	7,600	7,790	7,980	8,170	8,360
	385	6,930	7,123	7,315	7,508	7,700	7,893	8,085	8,278	8,470
	390	7,020	7,215	7,410	7,605	7,800	7,995	8,190	8,385	8,580
연말목표	395	7,110	7,308	7,505	7,703	7,900	8,098	8,295	8,493	8,690
	400	7,200	7,400	7,600	7,800	8,000	8,200	8,400	8,600	8,800
	405	7,290	7,493	7,695	7,898	8,100	8,303	8,505	8,708	8,910

자료: Bloomberg, DS투자증권 리서치센터

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)
- 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.