SECTOR
UPDATE

2025. 1. 31

삼성증권

Tech팀

이종욱
팀장
jwstar.lee@samsung.com

오동환
Senior Analyst
dh1.oh@samsung.com

문준호
Senior Analyst
joonho.moon@samsung.com

류형근
Analyst
hyungkeun.ryu@samsung.com

글로벌주식팀

김중한
Senior Analyst
joonghan1.kim@samsung.com

이영진, CFA
Analyst
youngjin91.lee@samsung.com

Tech (OVERWEIGHT)

Seeking Deeper: DeepSeek 사건의 오해와 본질

- 투자자들은 AI 내러티브 훼손에 대한 공포. 그러나 특별하지 않은 기술 방향의 단면
- 쏠림 현상 해소하며 주가 조정 불가피. 그러나 곧 명확한 AI 투자 방향성을 재확인하게 될 것

WHAT'S THE STORY?

주가 급락의 본질: 투자자들 공포의 본질은 다음과 같다. "중국 AI 스타트업이 미국의 반도체 견제에도 알고리즘 혁신을 통해 오픈AI를 따라잡았다. 최신 엔비디아 반도체를 쓰지 않고도 성능은 비슷하지만 비용은 매우 작다" 이 문장들이 맞다면 미국의 대중 재제는 효과가 없었으며 빅테크 기업들의 AI 해자는 무너졌고, 엔비디아 반도체 구매는 돈낭비였다.

오해가 많은 이벤트: 그러나 우리는 DeepSeek의 많은 부분들에 오해가 있었고, 본질이 왜곡된 채 전파되었다고 생각한다. 위기처럼 보이지만 실상 AI 기술 전개의 방향성이다.

- **DeepSeek, 기술적으로 특별하지 않다:** 우리는 LLM의 진입 장벽이 깨졌다고 보지 않는다. DeepSeek 모델에 사용된 경량화 기술은 특별하지 않으며 경쟁 모델 대비 비용을 크게 줄인 것도 아니다. AI 기술 방향이 추론으로 확대되기 위한 자연스런 효율화 작업이다.
- **클라우드 Capex 상향에 의심 금지:** 추론 비용은 지난 3년간 1,000배 이상 하락했지만 CSP 3사의 클라우드 이익은 지속적으로 상향 조정되고 있다. 추론 비용의 하락이 수요 증가의 동기였기 때문이다. DeepSeek은 모델 훈련의 효율화를 이야기하지만 메타의 AI 책임자 얀 르쿤은 대규모 투자가 훈련이 아닌 추론 인프라 때문이라 언급했다. MS CEO 사티아 나델라는 AI 모델의 비용이 하락하면서 AI 수요가 지수 함수로 확대된다고 언급했다.
- **DeepSeek이 엔비디아의 두번째 성장 뚜껑을 열었다:** DeepSeek-V3의 모델 훈련에 블랙웰을 사용했다면 더 낮은 비용으로 구현 가능했을 것이다. DeepSeek-R1을 블렉웰로 추론한다면 시간을 더 아낄 수 있다. DeepSeek이 제시한 모델의 방향성은 사실 엔비디아의 기술 로드맵과 일치한다. DeepSeek이 열게 될 추론 영역으로의 시장 확장은 엔비디아의 두번째 스케일링-AI Agent의 시작을 의미한다.
- **HBM의 수요는 데이터 이동 수요를 의미:** 비효율적인 반도체의 대명사인 HBM 수요에 미칠 영향은 제한적이다. 여전히 언어모델의 메모리 대역폭 부족 문제는 해소되지 않았다. 오히려 모델 경량화의 몸부림은 메모리 대역폭이 부족함을 보여주는 증거다. 엔비디아뿐만 아니라 추론용 ASIC에서조차 HBM 탑재량은 급증하고 있다.

오해의 원인은 미중 정치 문제: 이번 DeepSeek 사태는 미중 경쟁 때문에 부각되었으며 미중 경쟁에 불을 붙이는 계기가 될 것이다. 전개 양상은 1) 미국 투자의 증가와 2) 중국향 AI 재제 강화의 두 가지 방향성일 것이다. 정치적 관점에서 보면 DeepSeek은 현재의 방향성에 대한 Unwinding trigger가 아니라 현재의 방향성을 고착화시키는 Accelerator이다.

투자 아이디어- 길지 않은 조정과 새로운 기회: 우리는 한결같은 AI 관심도와 클라우드 수요/Capex 사이클을 바탕으로 이 이벤트가 단기 조정에 불과할 것이라고 결론지었다. 공포심 해소에는 시간이 필요하지만 클라우드의 투자상향과 엔비디아 실적, 그리고 미국 AI 모델의 경량화 성과, AI 비용 절감에 따른 시장 진입자 소식들이 중요한 반등 트리거이다.

엔비디아와 브로드컴 위주의 반도체, 클라우드 업체의 투자 상향이 여전히 긍정적이다. 한국 메모리 업체는 조정장 속에서 시장을 아웃퍼폼할 수 있다. AI 비용 하락은 이를 활용하는 엔터프라이즈 소프트웨어나 어플라이케이션 기업에게는 호재다. 한국 인터넷업체들에게도 재입장의 기회가 주어졌다. AI 효율화는 On device AI의 내러티브와도 맞닿아 있다.

Deepseek은 사실 특별하지 않다.

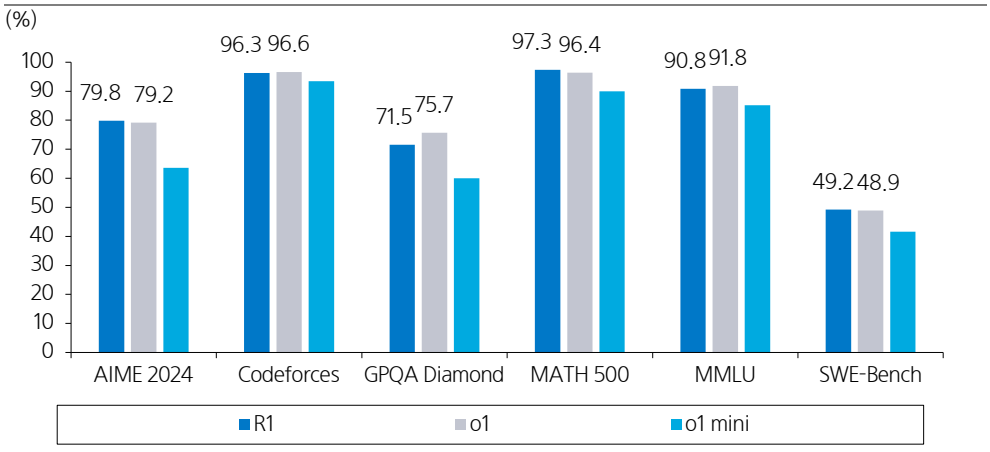
DeepSeek은 좋은 모델이지만 특별한 모델은 아니다

DeepSeek(딥시크)발 조정을 이해하기 위해서는 DeepSeek AI 모델의 특별함에 대한 분석이 선행되어야 한다. DeepSeeek V3와 R1 모델은 좋은 모델이다. 하지만 엄청나게 특별하지는 않다.

먼저 기술적 측면에서 1) 강화학습(RL)을 통한 R1 Zero의 재귀적 성능 개선 달성, 2) 콜드 스타트 미세조정을 포함한 다단계 보상 기반 강화학습 전략 활용, 3) GRPO(그룹 상대 정책 최적화) 적용을 통한 효율성 추구 등은 주목해야하는 부분이다.

하지만 반도체 제한이라는 한계를 돌파하기 위해 온몸을 비틀어 돌파구를 모색했을 뿐 우위를 점한 것이 아니다. DeepSeek 모델에 적용된 증류(Distillation), MoE(Mixture of Expert) 구조, MLA(Multi-head latent attention), 양자화(Quantization)등의 경량화 및 효율성 도구는 과거에서도 지속되던 방향성이다. 중국이 숨겨놓은 비기나 게임 체인저라고 하기에는 미국 기업들도 활용하던 전략이며, Reasoning 스케일업 목적으로 중요성과 주목도가 더욱 높아지던 분야다.

오픈AI o1과 유사한 수준의 벤치마크를 달성한 딥시크의 R1



자료: Deepseek, 삼성증권

DeepSeek의 비용은 호도되었다

비용적 측면에서는 V3의 낮은 학습 비용과 R1의 낮은 추론 비용이 주목을 받았다. V3 모델의 학습 소요 비용으로 알려진 규모는 557만 달러다. 이는 2,048개의 H800 GPU 인프라를 280만 시간 활용함에 따라 시간 당 임대 비용($2)을 고려해 산출한 것이다. 하지만 논문에도 명시된 것처럼 이전 연구 및 실험, 아키텍처, 알고리즘, 데이터 구축에 소요된 비용은 반영되지 않은 숫자다. 또한 V3 모델 기반으로 R1을 개발하는데도 상당한 비용이 소요되었을 것이다. 강화학습의 연산 비용이 지도학습 대비 파격적으로 저렴한 것도 아닐 뿐 더러 R1 구축에는 강화학습과 지도학습이 모두 활용되었다.

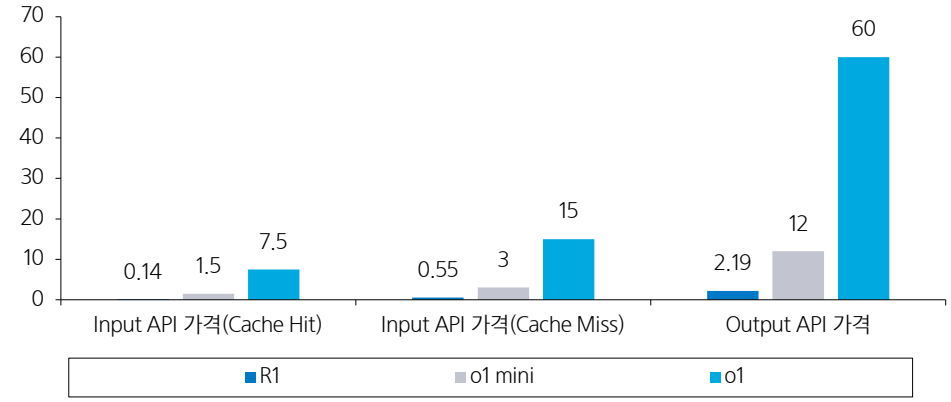
R1은 o1보다 저렴한 API 가격을 표방하고 있다. 하지만 추론 서비스 소요 비용은 API 가격 만으로 판단할 수 없다. 기업마다 수익화 전략을 어떻게 설정하는 지는가 다르기 때문이다. 오히려 구글의 Gemini 2.0 Flash Thinking 모델 API 가격이 1.5 Flash와 동일하게 적용될 경우 R1 대비 가성비에서 우위다. 또한 파라미터 일부만 활성화되는 MoE 구조는 모델 구동 시 효율성을 추구하지만, 실제 추론에서는 복합적 리퀘스트 처리가 필요하며, GPU의 분산 구동에 따라 상대적으로 비효율적인 구조를 만든다.

DeepSeek이 GPU Poor를 표방하고 있지만 Hopper GPU 5만 장(H20, H800 위주에 일부 H100 혼합)을 보유하고 있다는 주장이 다양한 AI 산업 관계자로부터 제기되고 있다. 보유 자원을 활용하지 않을 이유가 없다. SemiAnalysis 딜런 파텔은 딥시크가 GPU에 5억 달러 이상을 지출했다고 추정하고 있다.

또한 DeepSeek이 기술 격차를 빠른 속도로 줄일 수 있었던 이유로 오픈AI 포함 프론티어 모델을 자사 모델 구축에 활용했다는 이야기가 나오고 있다. 마이크로소프트와 오픈AI도 관련 조사를 진행 중이다. 선두의 결과물이 팔로워의 길잡이로 활용되는 것을 피할 수는 없다. 다만 이 경우 기초가 된 파운데이션 모델의 성능을 뛰어넘는 것에는 한계가 존재한다.

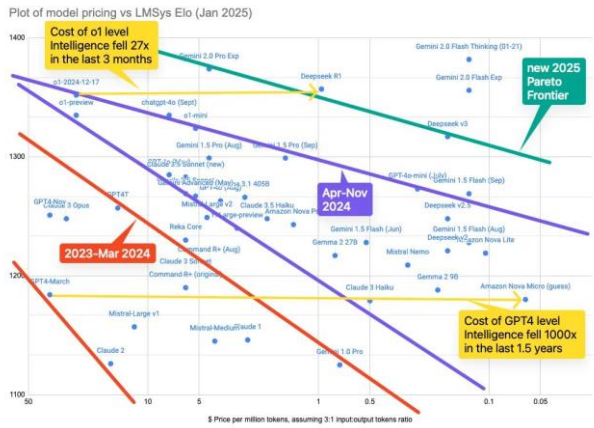
o1 대비 API 가격은 95% 이상 낮은 수준. 가성비를 추구하는 딥시크 R1

(달러/백만토큰)



자료: Deepseek, 삼성증권

모델 가격의 인하 트렌드는 과거에도 계속되었다



자료: X, 삼성증권

Gemini 2.0 Flash Thinking과 Deepseek R1 비교

	Gemini 2.0 Flash Thinking	DeepSeek R1
Input Token Price (/M)	**$0.075** (1.5 Flash Price)	$0.55
Output Token Price (/M)	**$0.30** (1.5 Flash Price)	$2.19
Context Length	**1M**	128k
AIME 2024 (Math)	**73.3%**	71.0%
GPQA Diamond (Science)	**74.2%**	71.5%
MMMU	**75.4%**	n/a
MMLU Pro	76.4% (not thinking)	**84%**

자료: X, 삼성증권

투자자들은 미국 Big tech의 AI 독점 전략을 의심한다.

우려의 이유는 기술이 아닌 AI 해자의 의심 때문

DeepSeek발 조정의 본질이 기술적 이슈가 아니라는 점은 타임라인에서도 드러난다. V3는 지난해 12월 26일, R1은 올해 1월 20일 공개되었다. 시장의 급락과는 시차가 있다. 주식시장에 딥시크발 공포가 드리운 것은 기술적 이슈를 정치적 이슈로 확대 해석한 언론보도가 시작점이다. 단순히 기술 혁신 및 역전 문제가 아니라 미국의 해자 정책이 작동하고 있는지에 대한 우려가 극심하게 반영된 것으로 판단한다.

"중국 AI 스타트업이 미국의 반도체 견제에도 불구하고, 알고리즘 혁신을 통해 오픈AI를 따라잡았으며, 성능은 비슷한데 비용은 훨씬 작다."는 틀린 문장이 아니지만, 추가적 의미 부여는 경계해야한다. 반대로 중국 AI 기업들의 돌파구를 미국 프론티어 AI 기업이 활용해 격차를 더욱 벌릴 수 있다. 추론 모델을 기반으로 차세대 모델을 학습시키기 위한 데이터셋을 만들고, 구축된 차세대 모델을 기반으로 또 다시 진보된 추론 모델을 만드는 선순환 구조의 작동이 확인되고 있다. 현재보다 발전된 프론티어 모델을 구축하기 위한 컴퓨팅 자원의 필요성은 오히려 높아진다. 중국의 반격으로 미국 빅테크 AI 독점 전략의 공고함에 약간의 의심이 생겨났지만 AI 개발 경쟁은 o1 수준의 모델을 구축에서 멈추는 것이 아니라 초지능(ASI)를 바라보고 있다.

미국 AI의 진입장벽과 반도체 투자 하향을 우려

1) LLM 진입장벽의 의심: 미국 AI의 해자가 의심된다. 더 큰 스케일로, 더 많은 반도체를 투입하여 자본집약적으로 모델을 훈련하는 것이 진입 장벽이라고 생각했는데, 후발 업체는 더 낮은 자본으로 누구나 참여할 수 있는 것일까.

2) 클라우드 업체들의 Capex 하향 우려: 반도체 성장성이 의심된다. 클라우드 업체들의 투자는 모델의 스케일업에 맞춰져 돈낭비를 했던 것이 아닐까. 향후 효율성을 따지기 시작하면서 엔비디아와 반도체 주문량이 감소하지 않을까.

3) 중국 AI의 부상: 우리는 미국의 AI 독점을 믿어서 미국 관련 주식 위주로 투자했는데, 더욱 효율적인 중국 AI가 승리하는 것은 아닐까.

반면에 현재의 주가 조정을 중국 기업이 미국 big tech들이나 떠오르는 스타트업들을 압도할 만한 기술 혁신이 발생했다거나, 아니면 중국의 LLM 기술이 미국을 역전하기 시작했다는 등의 기술 관점에서의 접근은 적절해 보이지 않는다. 키워드는 LLM의 효율성이다.

이러한 리스크의 결론만 요약하면 다음과 같다. 첫째, 우리는 LLM의 진입장벽이 깨졌다고 보지 않는다. DeepSeek 모델에 사용된 경량화 및 효율성 도구는 미국 기업도 사용한 산업의 흐름이었으며 DeepSeek이 경쟁 모델 대비 크게 비용을 줄인 것도 아니다. 둘째, 이번 사건은 AI의 관심이 학습에서 추론으로 확대되는 계기를 만들며 오히려 빅테크의 Capex가 확대되는 계기가 될 것으로 믿는다. 셋째, 중국 AI의 부상으로 규제는 증가하겠지만, 이와 동시에 투자 경쟁에도 불이 붙을 것으로 본다. 결론적으로 겉모습은 위험해 보였지만, 생각할수록 AI 성장의 단면을 보여준 단편적 이벤트에 불과하다.

비용 하락은 수요 증가의 어머니

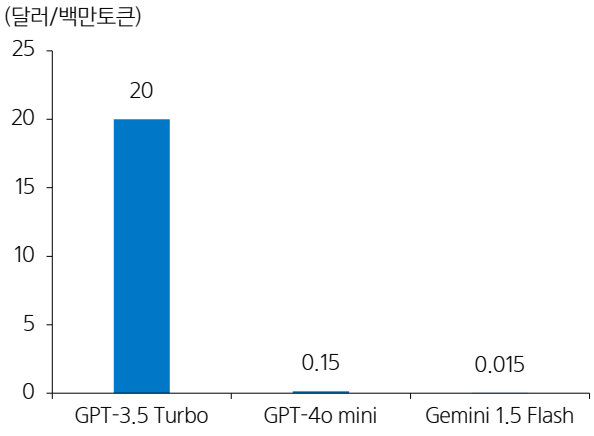
API 가격 하락이 클라우드 인프라 수요 감소가 아닌 이유

추론 코스트는 지난 3년간 약 1,000배 이상 하락했다. P의 하락에 따른 Q의 증가가 AI 서비스시장 개화의 본질이다. DeepSeek은 이런 트렌드를 가속화했을 뿐이다. 클라우드 기업에게는 당연히 호재다. API 사용량 증가(Q)가 API 가격(P)의 하락을 능가하기 때문이다.

실제로 하이퍼스케일러 3사의 클라우드 부문 영업 마진은 AI사이클이 시작된 이후 지속적으로 확대되고 있다. 특히 추론은 훈련대비 토큰마진 측면에서 유리한 것으로 알려져 있다(50~70% 추정). 돈을 잘 번다면 투자를 망설일 이유가 없다. 마이크로소프트(FY25년 800억 달러 투자 예상), 메타 플랫폼스(FY25년 600~650억달러 투자 예상)의 CAPEX 가이던스가 지속적으로 상향되는 이유다.

메타 플랫폼스의 AI 수장인 얀 르쿤 또한 딥식 사태를 겨냥해 자사의 대규모 인프라 투자가 트레이닝이 아닌 추론에 집중하고 있다는 점을 명확히 했다. 마이크로소프트가 초기 스타게이트 이니셔티브를 주도했던 입장에서 한발 물러나 기술적인 서포트를 말하는 이유도 추론 인프라에 집중하려는 의도다. 상대적으로 리스크가 크고 트레이닝 인프라가 중요한 AGI 달성은 오픈AI에게 위임하겠다는 의미다.

'22년과 '24년 주요 LLM의 API 가격 비교(1,000배 이상 하락)



자료: 삼성증권 정리

트레이닝 vs 추론에 대한 주요 기업 입장

Meta
- 자사의 대규모 인프라 투자(FY25 $600~650억)는 트레이닝이 아닌 **추론에 집중**하기 위한 투자임 밝힘

OpenAI
- **트레이닝에 대한 투자**를 통해 AGI를 우선적으로 달성하는 것을 중점적 목표로 삼음

Microsoft
- 초기 스타게이트 이니셔티브 주도 → 기술적 서포트 입장 전환
- FY25 $800억 투자로 **추론 인프라 확충**

자료: 삼성증권 정리

오픈AI에게는 악재일까?

그렇다면 DeepSeek의 등장은 오픈AI에게 악재일까? 좋은 소식은 아니지만 그렇다고 대단한 악재도 아니다. 경량화 기술이 발달할수록 어차피 고성능의 대형 베이스 모델과 이를 가능하게 하는 인프라를 지닌 기업들이 유리하기 때문이다. 새로운 알고리즘 방법론의 도입 또한 다양한 시행착오를 감내할 수 있어야 한다는 측면에서 안정된 인프라를 보유한 기존의 강자들이 유리하다.

결국 DeepSeek 또한 기존 기술의 기반 위에서 이루어진 결과물이다. 만일 오픈AI의 최신 모델 o3를 상회하는 탑티어 모델이 등장한다면 위기라고 볼 수 있다. 그러나 단순히 가성비 모델의 출현을 근거로 탑티어 기업의 위기를 결론짓는 것은 성급하다. 오히려 그 사이에 있는 애매한 위치의 세컨티어 기업들, 혹은 메타처럼 동일한 오픈소스 전략으로 생태계를 확장하려는 기업의 입장에서 위기감이 커질 수 있다는 판단이다. 실제로 DeepSeek 등장 이후 메타 플랫폼스가 무려 4개의 워 룸을 운영하고 있다는 뉴스 또한 이러한 맥락에서 비롯된 것으로 판단된다.

DeepSeek 충격은 역설적으로 수요 증가의 계기가 될 것.

DeepSeek의 충격은 저비용으로 실행할 수 있는 AI 모델들에 대한 관심을 증가시키는 계기가 되었다. 이제 곧 증류와 양자화로 경량화된 AI 모델들이 집중적으로 증가하기 시작할 것이다. 클라우드 업체들의 capex의 감소를 걱정하기보다는 새로운 AI 수요들로 클라우드의 성장 속도가 가속화될 것이라고 전망한다. 경량화된 AI모델들 조차도 추론 과정을 진행하거나(scaling for reasoning) 실제 서비스 사용량이 늘어날 때(scaling for inference), AI 데이터센터의 규모의 경제가 큰 진입 장벽이 될 것이다. 일반 기업들의 경우에는 높은 비용때문에 주저했던 AI 모델 도입의 문턱이 낮아졌으며, 소프트웨어 업체들은 더 낮은 반도체 비용으로 비슷한 성과를 낼 수 있는 AI 개발이 가속화될 것이다.

마이크로소프트 CEO인 사티아 나델라는 DeepSeek의 영향에 대해 자신의 X계정에 '제본스 역설(Jevons Paradox)'을 언급했다. 제본스 역설은 19세기 증기기관의 효율성이 개선되면서 석탄을 덜 쓰게 될 것이라 예상했지만, 실제로는 증기기관이 널리 퍼지는 계기가 되면서 석탄을 더 쓰게 되었다는 제본스의 연구에서 유래되었다. 이처럼 AI 모델의 효율성 증가가 AI 모델과 반도체 수요의 확대를 불러오게 될 것이라는 의미이다.

우리도 AI도 효율성의 증가가 수요의 증가로 이어진다는 예상에 동의한다. 제본스의 역설이 작동하려면 세 가지 전제조건이 필요하며, 이번 케이스는 전제조건에 부합한다고 생각한다. 첫째, 수요의 가격민감도가 커야 한다. 우리는 1월 14일 CES2025 리포트에서 AI 서비스를 비즈니스로 실행하기에 투자금액이 너무 높다고 언급한 바 있다. 이는 현재 AI의 비용이 수요 증가의 가장 큰 허들이라는 의미이다. 둘째, TAM(Target available market)이 커야 한다. 현재 AI에는 Reasoning scaling과 Physical AI scaling이라는 TAM이 존재한다. 가격 하락은 잠재수요를 이끌 트리거가 될 것이다. 셋째, 공급 확대가 가능해야 한다. 마치 석탄 광산처럼, AI도 많은 투자가 수요의 전제 조건이다.

한국 인터넷에 주어진 두번째 기회

DeepSeek-R1 등 오픈소스 모델의 등장은 제한적인 Capex를 가지고 AI 사업을 진행해야 하는 국내 인터넷 기업들에게 새로운 기회를 제공해줄 전망이다. 네이버와 카카오는 24년부터 AI전략을 수정하여 글로벌 프론티어 AI 모델과 직접 경쟁하기보다는 '비용 효율적'으로 AI기술을 기존 서비스에 접목시키는데 주력하고 있다. 그러나 하이퍼클로바 등 기존 자체 모델과 글로벌 업체들의 AI 모델과의 격차가 점차 벌어지며 서비스 내재화에도 어려움을 겪던 상황이다. 이러한 상황 속에서 DeepSeek의 모델들은 설계 최적화를 통해 훈련과 추론에 필요한 비용을 크게 낮추고, 이를 오픈소스로 공개하여 다른 기업들도 기술을 활용할 수 있게 하였다. 네이버의 하이퍼클로바나 카카오의 카나나도 오픈소스로 공개된 모델들을 학습하며 개발되었던 만큼 저비용 오픈소스 모델을 학습하여 성능을 크게 개선시킬 수 있을 전망이다.

이에 네이버는 자체 검색과 쇼핑, 광고, 컨텐츠에 생성형 언어 모델을 접목하여 서비스 경쟁력을 높이고, 카카오는 대화형 AI 챗봇을 도입하여 채팅 서비스의 기능을 강화하려는 전략이 힘을 받을 것으로 예상된다.

한편, 미국과 중국 기업들의 글로벌 프론티어 모델 경쟁에서 국내 기업들이 소외되는 모습을 보임에 따라 지난 수년간의 기술 개발의 성과에 대한 비판적인 시각이 제기될 수 있다. 그러나 소수의 미국 빅테크들의 기술 독점 상황보다는 다각 체제의 경쟁 체제가 국내 기업들에게는 추격의 발판을 마련할 수 있다는 점에서 긍정적인 효과가 더 크다고 판단된다.

AI 반도체는 큰 위기를 맞이했는가.

AI 가속기 사이클은 끝나지 않았다

DeepSeek가 현저히 낮은 투자 비용 및 저렴한 가격의 서비스 제공으로 인해 AI 인프라, 특히 AI 반도체의 수요와 효용에 대한 의구심에 제기되고 있지만, 우리는 우려는 과장되었고 투자자 반응 또한 과도했다고 판단한다.

AI 모델의 성능은 학습 데이터, 모델의 매개변수, Computing Power에 의해 결정된다. 학습 데이터 측면에서 보면, 기존 Text 중심에서 이미지, 동영상 등으로 학습의 범위가 넓어지고 있고, 이를 효율적으로 Support하기 위한 Computing Power 확대 노력도 지속 전개되고 있다. 이는 더 많은 고품질의 AI 반도체가 필요하다는 점을 시사한다. 고품질은 최대 성능과 효율성을 모두 포함하는 개념이다.

효율성을 위해서는 또다시 엔비디아의 기술력이 더없이 중요한 순간이 올 것이다. DeepSeek V3의 경우, FP(Floating Point)16을 활용한 Open AI와 달리, FP8로 학습이 진행되었다. 낮은 Floating Point는 장단점이 존재한다. 압축을 통해 데이터를 단순화하고, 이를 통해 연산 속도를 늘릴 수 있다는 장점이 있으나, 정밀도가 떨어진다는 한계가 있다. 그러나 최근에는 FP8 학습의 일부 과정에서는 정밀도 손실보다 연산 비용의 절감이 더욱 뛰어난 순간이 있어 FP8 학습 비중을 늘리는 것이 트렌드이다. 효율성을 위해서 더욱 정확도가 떨어지는 FP4까지 학습에 이용되는 시대가 온다. 성능의 희생을 최소화하면서 얻게 되는 연산 비용의 절감이 DeepSeek 효율성의 핵심이다. 그런데 역설적으로 FP8과 FP4의 학습은 신형 GPU에서 가장 효율적이다. 근래 엔비디아 제품 로드맵이 AI 모델 개발의 트렌드와 궤를 같이하기 때문이다. 엔비디아는 Hopper를 기점으로 Transformer Engine을 탑재하여 소프트웨어에서 FP8/16간 동적 변환과 최적화를 지원하기 시작했고, Blackwell 들어서는 FP8과 FP4를 하드웨어에서 네이티브로 최적화하기 시작했다. Blackwell FP4는 Hopper FP8 대비 5배 이상의 연산 성능 (Blackwell FP4 20,000TFLOPS vs Hopper FP8 4,000TFLOPS)을 구현하는 데 성공했다. 그리고 차세대 Rubin은 Blackwell 보다도 더욱 강력해진 성능을 보여줄 것이다.

또 하나 중요하게 보아야 할 관점은, 모델이 효율적이라는 것이 AI 계산이 줄어든다는 것을 의미하지 않는다는 점이다. AI 모델의 효율적으로 변화하면서 AI 서비스의 경쟁력이 모델에서 강화학습과 추론 기술로 확장되고 있다. 많은 강화학습의 기술이 발전되어 왔지만 DeepSeek를 계기로 강화학습으로의 자원 투입이 더욱 확장될 것이다. 그런데 강화학습과 추론기술이 엔비디아가 주장하는 두번째 스케일링인 AI Agent 사이클의 시작을 의미한다. 모델을 여러번 계산하고, 다양한 모델로 확장하고, 데이터를 바탕으로 강화학습하는 등 AI Agent가 제시하는 방법론을 현실화시키기 위해서는 모델의 효율화(경량화)가 필수적이다. 성능은 사전 학습으로 높이고, 증류를 통해 가벼운 모델로 변환하지 않으면, 추론의 과정 속에 나타나는 계산 시간을 감당할 수 없다. 엔비디아는 AI agnet가 모델링하는 방식을 세트메뉴로 구성하고, text 형식으로 대화할 수 있는 AI Blueprint 도구를 제공하여 서비스 사업자들의 추론 모델 활용성을 높였다.

여전히 AI반도체의와 AI 데이터센터는 여전히 AI 서비스의 확실한 해자이다. 모델 경량화는 연산 비용과 속도를 최적화하기 위한 과정이며 AI 반도체 비용으로부터의 독립을 의미하는 것이 아니다. DeepSeek v3 모델이 블랙웰을 사용하여 훈련했다면 훨씬 더 낮은 비용으로 성능 개선이 가능 (FP8의 가속 효과, NVLink를 사용하여 훈련 속도 개선, 메모리 밴드위스 개선)했을 것이다. DeepSeek R1 모델을 블랙웰을 통해 사용(inference)한다면 더욱 비용과 시간을 아낄 수 있다. Reasoning은 더 많은 컴퓨팅 파워를 소구한다. 이 말은 AI Agent 시대가 올 수록 AI반도체의 능력이 중요해진다는 의미이다. AI 데이터센터의 규모는 AI agent와 AI 서비스의 확장력을 의미한다. 모델을 서비스하기 위해 얼마나 최신의 AI 반도체가 탑재된 AI 데이터센터를 얼만큼의 규모로 확보할 수 있는지는 굉장히 중요한 경쟁력이다. 결국 자본 집약적인 AI 데이터센터는 여전히 AI 서비스의 해자 중 하나라 볼 수 있다.

HBM 메모리의 사이클은 끝나지 않았다.

HBM은 언어모델을 훈련하거나 추론할 때 메모리 대역폭이 너무 모자라기 때문에 발생한 고가 메모리 반도체이다. 신형 GPU 내 HBM의 세대 진화와 탑재량 증가가 지속되고 있다는 점이 시장에서 HBM 수요의 고성장을 예견해왔던 배경이다. 하지만 AI 모델 훈련 비용이 높은 이유는 비싼 GPU와 비싼 HBM을 사용했기 때문이며 AI가 효율성을 위해 몸부리치는 상황이 온다는 것은 비싼 GPU와 HBM을 우회하는 기술을 개발하고 있다는 뜻이다. GPU의 스펙 Upgrade 사이클이 둔화된다면, HBM 수요 성장에 대한 회의론이 부각될 것이며, 이에 대한 진단이 필요하다. 정말 최근의 AI 기술 동향이 메모리 대역폭을 덜 요구하게 되는 것일까?

MoE(전문가혼합, Mixture of Expert)구조와 양자화(Quantization)는 이론적으로 메모리 수요를 떨어뜨릴 것이라는 우려가 있으나 실제 영향은 미미하다고 생각한다. MoE는 모델의 모든 파라메터를 활성화시키는 대신에 각 입력 토큰에 대해 가장 적합한 전문가를 선택한 뒤 활성화하는 방식으로 효율성을 높이는 전략이다. DeepSeek V3 모델은 6,710억개의 파라메터로 훈련했으나 토큰 처리시 최소 370억개의 파라메터만 활성화된다. 이론적으로 보면 이는 메모리의 대역폭이 일시적으로 6%만 필요한 상황이 될 수도 있다는 것을 의미한다. 그러나 MoE는 훈련(training)할 때의 효율성을 의미할 수 있어도 추론(inference)할 때의 효율성을 의미하는 것은 아니다. 추론할 때에는 각 전문가 조각들이 언제 쓰일지 모르는 상황 속에서 프롬프트에 따라 모든 엑스퍼트가 동시에 활성화되어야 하며 전문가 조각끼리의 데이터 이동을 생각하면 메모리 사용량은 더욱 늘어날 가능성이 있다. AI 서비스가 성장한다는 것은 추론 계산 중심으로 메모리가 사용된다는 이야기이기 때문에 결국 메모리의 잠재 시장은 떨어지지 않는다는 것을 뜻한다. DeepSeek에서 사용된 양자화는 정확히 이야기하면 FP32나 FP16으로 데이터를 학습하지 않고 FP8이나 FP4로 학습하는 저비트 학습(Low-bit Quantization)을 의미한다. FP16에서 FP8로 학습한다면 이론적으로 디램의 대역폭이 절반만 필요하다. 그러나 현실적으로는 데이터 계산 시 메모리 오버헤드가 클 수밖에 없다. 블랙웰처럼 하드웨어에서 FP8을 네이티브로 계산하는 AI반도체가 아니라면 일부 FP8을 이용해 데이터가 입력되었더라도 실제 계산 시 FP16에 맞추면서 메모리 사용량이 줄지 않는다. Optimizer state, scale factor, zero point 등의 데이터를 추가적으로 저장하는 과정 속에서도 메모리 오버헤드가 발생한다. DeepSeek에서는 모든 데이터를 FP8로 학습하지 않기 때문에 추론 과정에서 FP8/FP16 캐스팅이 발생한다면 메모리 사용량이 더욱 늘어날 수 있다.

추론용 ASIC을 사용하면 메모리 대역폭을 더 줄일 수 있을까? 물론 추론용 ASIC은 효율적인 계산을 위한 것이며 메모리 대역폭을 줄일 수 있는 다양한 기술들을 도입하고 있다. 특히 연산 패턴이 고정적이고 일정하다면 메모리의 탑재량을 줄이기에 효과적이다. 하지만 기본적을 LLM을 돌릴 정도의 ASIC은 파라메터의 수가 너무 많아서 HBM이 필수이다. 심지어 메모리 대역폭이 병목 현상이 나타날 가능성도 높다. 강화훈련(RL)의 경우에는 사후훈련임에도 불구하고 많은 데이터를 처리하고 매번 새로운 데이터가 필요하므로 HBM의 사용량이 더욱 늘어날 수 있다. 속도 향상과 계산의 효율성을 위해 한번에 여러 토큰을 처리할 때에도 커다란 메모리 대역폭이 필요하다.

현재 가장 유력한 ASIC인 구글 TPU와 아마존의 Tranium이 제품 업그레이드가 계속될수록 HBM 탑재량이 증가하는 것은 여전히 ASIC에서도 HBM의 수요가 부족하다는 결정적 증거이다. 아마존은 24년 메모리 탑재량이 작은 Inferencia의 향후 개발을 중단하고 추론시에도 Tranium 반도체를 사용하기로 결정했다.

AI 반도체의 HBM 탑재량 증가

			Launch	Tech	Content (GB)	Density (GB)	HBM/p (unit/package)
GPU	Nvidia	A100	3Q20	HBM2e	80	16	5
		H100	4Q22	HBM3	80	16	5
		H200	2Q24	HBM3	144	24	6
		B200/GB200	1Q25	HBM3e 8hi	192	24	8
		B300/GB300	3Q25	HBM3e 12hi	288	36	8
		GB300A	3Q25	HBM3e 12hi	144	36	4
		R100	1Q26	HBM4 12hi	384	48	8
	AMD	MI200	4Q21	HBM2e	128	16	8
		MI300	4Q23	HBM3	192	24	8
		MI325	4Q24	HBM3e 12hi	288	36	8
		MI355	4Q25	HBM3e 12hi	288	36	8
		MI400	1Q26	HBM4 12hi	384	48	8
	Intel	Gaudi2	2Q22	HBM2e	96	16	6
		Gaudi3	4Q24	HBM2e	128	16	8
		Falcon Shore	n/a	n/a			
ASIC	Google	TPU v5e	2Q23	HBM2	16	16	1
		TPU v5p	4Q23	HBM2e	96	16	6
		TPU v6e	3Q24	HBM3e 8hi	48	24	2
		TPU v6p	2Q25	HBM3e 8hi	192	24	8
		TPU v7e	3Q26	HBM4	216	36	6
		TPU v7p	1Q26	HBM4	288	36	8
	Amazon	Inferentia 2	4Q22	HBM3	32	16	2
		Inferentia 2.5	2Q24	HBM3	32	16	2
		Tranium 2	4Q23	HBM3e 8hi	96	24	4
		Tranium 2.5	2Q25	HBM3e 12hi	144	36	4
		Tranium 3	4Q25	HBM3e 12hi	144	36	4
	Tesla	Dozo	3Q23	HBM2e, 3	160	32	5
	MS	Maia	3Q25	HBM3	64	16	4
		Maia 2	n/a	n/a			

자료: 각 사, 삼성증권 추정

꺾이지 않은 HBM의 전성비, 새로운 메모리 수요 자극

전성비 - 전력 효율성은 고객들이 비싼 돈을 주고 HBM을 고수하는 배경 중의 하나이다. 기술적으로 보면, 통신 속도 최적화 및 동일 용량 기준 전력 소모 축소 등의 측면에서 HBM이 여전히 강점을 가지고 있다. AI 솔루션이 당면한 문제가 있다면, 전력이다. GPU도 전력을 많이 소모하고, GPU를 Support하는 메모리반도체에도 많은 전력이 필요하다. 그렇다면, 전력 효율화를 메모리에서도 가져가야 하는데 이에 적합한 솔루션이 DRAM에서는 HBM, NAND에서는 고용량 eSSD와 HAMR (HDD 진영)이다. 현 기술 구조 상, 동일 용량 기준 HBM이 기존 DRAM 대비 전력을 덜 소비하는 것으로 추정되며, 병렬 구조의 데이터 처리 구조 상, I/O를 늘리는 것이 Pin당 데이터 처리 속도를 늘리는 것 대비 AI에 유리한 솔루션인 것으로 판단된다.

오히려 효율적인 AI 모델의 증가와, AI 서비스의 확대는 새로운 메모리 수요를 자극하게 될 것이라고 생각한다. 엔비디아는 25년 로봇이나 자율주행에 사용할 수 있는 추론용 보드 Thor를 업그레이드했다. 128GB 정도의 경량 디바이스에서도 증류 기술을 통한 고사양 모델이 돌아가기 시작한다면, 로봇, 자율주행, 엣지디바이스의 수요를 자극할 것이라고 기대한다.

물론 미리 예고하건대, 현재의 비싼 CoWoS와 HBM구조를 대체할 혁신적인 저비용(저전력) 고사양 반도체의 출현은 필연적이다. AI 모델과 소프트웨어 관점에서 효율적인 AI 훈련과 추론 기법들이 나타나기 시작하는 이유와 같다. 하지만 경량화 AI 모델이 새로운 수요를 창출하는데 기여한 것처럼 AI반도체의 혁신도 총 반도체 수요를 확장하는데 기여할 것이다. AI 반도체 시장은 AI가속기와 HBM 조합이 AI 데이터센터에 탑재되는 고사양 하이엔드 시장과 SOC와 저전력 AI 메모리의 조합이 디바이스와 Edge AI에 탑재되는 고효율 시장으로 양분화될 것이다. 이때에도 주식 시장은 위기를 이야기하겠지만, 곧 시장 성장 속도가 가속화됨을 감지하게 될 것이다.

수급으로 보면, 현시점에서 바라보는 2025년 HBM 수요는 240억 Gb 내외인 것으로 판단된다. 최근 부각되는 AI Capex 상향 기조와 CoWoS Capa의 Bottleneck 해소 등을 감안 시 수요는 260억 Gb까지도 Stretch될 수 있을 것이라 생각한다. 반면, 공급은 여전히 제한적이다. 적정 수율 확보가 어려운 고단 제품 중심으로 생산 Mix가 변화하고 있으며, 하반기에는 Net Die 축소를 추가 유발하는 HBM4가 등장한다. 그만큼 생산을 늘리는 작업은 시장 예상 대비 순탄하지 않을 것이라 생각한다. 그렇다면, 수급상의 Mismatch가 올해 나타날 가능성은 제한될 것이라는 판단이다.

미중 경쟁의 도화선

이번 DeepSeek 사태는 미중 경쟁에 불을 붙이는 계기가 될 것이다. 이 계기는 1) 미국 투자의 증가와 2) 중국향 AI 재제 강화의 두 가지 방향성으로 전개될 것이라 믿는다.

미국 내 AI 데이터센터 투자의 자극

모두가 알다시피 AI와 반도체는 경제의 논리와 정치적 논리가 결합되었다. 정치적 논리는 경제적 숫자를 변질시킨다. 예를 들어 AI SW의 경우 미국 진영은 큰 자본력과 반도체 독점이 필요함을 부각하고, 중국 진영은 작은 자본력과 비용 절감을 강조한다. AI반도체의 경우 미국 진영은 낮은 자본 투하와 공급 부족, 차별적 공급을 강조하고, 중국 진영은 기술 격차 해소와 높은 자본 투하를 강조한다. 이러한 정치적 경향성은 투자자들이 현실과 방향을 객관적으로 파악하기 어렵게 만든다. 투자자들이 이번 Deepseek 쇼크를 현실보다 큰 충격으로 느낀 이유는 미국은 비용을 부풀리고, 중국은 비용을 축소했기 때문이다. 정치적 논리의 결과는 항상 공포심과 투자의 정당성을 불러 일으킨다.

미국 업체들로서는 투자를 더 확대할 여지가 크다고 판단한다. 오히려 강화되어 온 규제 속에서도 중국 업체들이 다양한 기법 또는 혁신을 통해 우수한 성능의 모델을 개발했다는 것은 미국 투자를 자극하게 만든다. 이미 오픈 소스로 풀린 모델들은 모두가 활용 가능하다. 결국 경쟁사들 대비 더 뛰어난 파운데이션 모델을 보유해야 이로부터 파생되는 소형 모델들의 성능도 경쟁력을 가질 수 있을 것이다. DeepSeek 모델 공개를 전후로 스타게이트 프로젝트가 발표되고 또 메타가 CAPEX 계획을 상향한 것은 우연이 아닐 것이다. 해당 과정에서 더 우수한 성능의 반도체는 미국 업체들이 가진 상대적 이점이 아닐 수 없다. DeepSeek가 더 적은 수의 저렴한 GPU들로 우수한 모델을 개발했다는 것에만 주목하지 말고, 만약 더 많은 수의 최신 GPU들로 중무장했을 때 얼마나 뛰어난 모델을 개발했을 수도 있었을 지를 두려워해 볼 필요도 있을 것이다. 어떤 상황에서든, 아무리 효율적인 AI가 나오더라도 AI의 성장 과정 속에서 AI데이터센터의 규모는 높은 해자를 만들 수 있다.

중국향 규제는 강화될 것

물론 이번 DeepSeek 노이즈가 가져올 리스크도 공존한다. DeepSeek를 통해 중국 AI의 가능성을 확인한 이상, GPU 및 HBM에 대한 규제 범위가 확대될 가능성이 높다고 생각한다. 최근 언론 보도와 같이 H20과 같은 중국산 반도체에 대한 수출 통제가 예상 가능한 변화이며, Nvidia에서도 이러한 Risk를 감안하여 H20에 대한 빌드업을 적극적으로 하고 있지 않은 것으로 추정된다.

규제는 강해질 것

향후 대중 규제는 강화될 가능성이 높다. 알고리즘 단의 방법론이 상위 업체를 중심으로 빠르게 퍼지는 상황에서(커머디티화) 이를 서포트할 수 있는 인프라가 오히려 경쟁력 차별화의 키포인트가 될 것이기 때문이다. 인프라 축소에 대한 시장의 오해가 풀려가는 와중에도 엔비디아 주가가 빠른 회복을 못하는 이유 또한 결국 규제 강화에 따른 실적 불확실성 때문인 것으로 판단된다.

미국의 스타게이트(4년간 5,000억 달러) 뿐만 아니라 중국 정부 또한 5년간 1조 위안(1,370억 달러) 투자를 예고하며 대대적인 지원을 예고했다. 그야말로 민관합동 국가 총력전의 양상이다. 알고리즘이 인프라를 완전하게 극복했다면 결코 나오기 어려운 숫자들이다. 실제로 DeepSeek CEO 량원펑은 미국의 GPU 수출 통제를 주요 제약 요인으로 언급했다. 완벽한 통제는 못해도 상대방에게 최대한의 데미지를 주겠다는 것이 AI 패권경쟁의 기본 노선이 될 가능성이 높다.

미국 규제 관련 주요 주체의 입장

주체	입장
하워드 러트닉 (상무부 장관 지명자)	중국의 첨단기술 개발을 견제하기 위한 수출 통제를 강화할 것
엔비디아	AI 규제와 관련하여 미 행정부와 협력할 준비가 되어 있음
량원펑 (DeepSeek CEO)	미국의 반도체 수출 통제가 “가장 큰 도전”이 될 것

자료: 삼성증권 정리

중국이라는 대형 시장이 잠식될 Risk 가 있다는 점은 잠재 수요에 있어 부정적이지만, 이를 통해 확인할 수 있는 것은 AI 에 대한 미국의 의지이다. AI 가 미래 생태계에 미칠 파급력이 큰 만큼, 기존의 반도체 부양책 CHIPS Act 와 스타게이트 프로젝트를 필두로 정부 차원에서의 육성 정책이 지속 강화될 것이며, 기업 단에서도 Capex 의 상향이 연이어 나타나고 있다. 그렇다면, AI 반도체와 이에 동행하는 HBM 과 같은 AI 메모리 수요에도 작년과 같은 Upside 가 발생할 가능성을 배제하기 어렵다는 판단이다.

투자 전략 - 조정 후 새로운 기회 탐색

조정기는 불가피하나 길지 않을 것

최근의 주가 급등락은 단순히 DeepSeek의 기술적 충격 때문이 아니다. 투자자들의 특정 섹터 쏠림 현상, 미중 AI 전쟁의 불확실성, AI가 아닌 매크로 우려까지 복합적으로 작용하였다고 생각한다. 현상 발생과 주가 반영 사이의 시차로 볼때 DeepSeek 이벤트는 오히려 단순한 트리거에 불과했을 가능성이 높다. 바꿔 이야기하면 DeepSeek의 진실이 무엇이냐와 상관없이 쏠림 현상과 공포심 해소에는 시간이 필요하다. 당분간 주가는 높은 변동성 또는 조정기를 거칠 가능성이 높다.

짧은 조정기간: 하지만 우리는 조정기간이 길 것이라 생각하지 않는다. 현재는 AI의 관심도가 떨어지지 않았다. 즉, AI 섹터에서 다른 섹터로 투자자 수요가 넘어간 것은 관찰되지 않는다. 그리고 클라우드 업체들의 투자 확대 사이클이 지속되는 데다가 연달아 있을 실적 발표에서 상황이 변하지 않았다는 것을 투자자들이 인지하게 될 것이다. 무엇보다 이미 주가 급락의 과정에서 투자자들이 각자 나름의 리스크 헷징을 끝낸 뒤라고 생각하기 때문이다.

무엇을 사야 할까? 첫째, 우리는 반도체, 클라우드, 소프트웨어 업체들의 주가 회복 혹은 추가 상승을 전망한다. DeepSeek이 성장의 단면을 보여줬지만 리스크로 오인한 것이 기회를 제공한다고 생각한다. 둘째, 우리는 조정기의 과정에서 상대적으로 쏠림이 덜했던 한국 테크 주식이 당분간 아웃퍼폼할 것이라고 생각한다. 셋째, AI의 소외주로 여겨졌던 한국 인터넷, On device AI와 디바이스 업체들에서 주가 상승의 기회를 모색할 것이다. 반면 Pre-training scaling의 관심도가 떨어지면서 데이터센터의 주변 인프라의 관심 하락이 전망된다.

Catalyst가 무엇일까? 클라우드의 capex와 엔비디아의 실적이다. AI 네러티브에서 가장 큰 줄기의 논리가 굳건하다는 증거가 가장 중요하다. 미국 AI 모델들의 경량화 성과들도 중요한 반등의 트리거이다. API 가격 하락과 AI Agent 서비스의 등장은 AI의 성장 과정에서 당연히 나타날 호재이다. 높은 가격때문에 할 수 없었던 많은 AI 관련 비즈니스가 언급되고 자본집약적이라고 주저했던 잠재적 경쟁자들의 참여 소식도 좋은 catalyst가 될 것이다.

진정한 리스크는? 클라우드의 수요 감소를 목격하는 것이 결정적 peak-out의 신호이다. 이 관점에서 DeepSeek은 오히려 클라우드의 수요 촉진을 알리는 뉴스이다. peak-out의 계기는 AI에서 되는 서비스와 안되는 서비스가 갈리기 시작하는 서비스 시작 단계에서 드러날 가능성이 높다. 지금은 현실성과 상관없이 모든 상상 속 AI 서비스들을 잠재적인 클라우드 토큰 수요로 잡고 있기 때문이다.

굳건한 엔비디아와 ASIC 1등 브로드컴

AI 반도체에 대한 긍정적 시각은 여전히 유효하다. 특히 기술 개발 경쟁이 더욱 심화되고 있는 국면 속 가장 우수한 반도체를 많이 확보하려는 노력은 올해도 지속될 것으로 전망한다. 2025년 들어서도 연일 빅테크 CAPEX 상향 움직임이 지속되고 있듯 말이다.

엔비디아가 올해도 대장주라고 믿는 이유다. 다만 경쟁 심화 속 업체들은 차별화 요인을 지속해서 찾아 나갈 것이다. ASIC은 차별화 방법 중 하나이며, 또 추론 수요가 급증하는 환경에서 개발 움직임이 더욱 가속화 될 전망이기에 역시나 긍정적 시각이 유효하다. 이미 세 개의 프로젝트를 진행 중이고, 또 새로이 AI ASIC 개발에 참전하는 오픈AI와 애플을 고객사로 확보한 것으로 기정사실화 된 상태의 브로드컴에 함께 주목하는 이유다.

GPU와 ASIC이 경쟁 선상에 있어 보일 수 있지만, 당사는 AI 인프라가 성장하는 환경 속 모두 성장을 이어갈 수 있다고 생각한다. 단, 이같이 경쟁이 치열한 산업의 성장의 결실은 단연 각 영역의 1위 업체들에게 몰릴 수밖에 없다는 판단이다. GPU 1등 엔비디아와 ASIC 1등 브로드컴을 함께 긍정적으로 보는 배경이다.

메모리의 오해는 곧 풀릴 것

범용 반도체에 대한 의심이 산재한 상황에서 AI에 대한 의구심이 피어나고 있다는 점은 메모리반도체와 Supply Chain의 단기 주가에 있어 부정적으로 작용할 수 있다. 하지만, 이러한 의심은 반등의 재차 변곡점을 맞이할 가능성이 더 높다고 생각한다. 신제품 로드맵에서 HBM의 탑재량 증가는, 여전히 AI 생태계 속에서 메모리 대역폭의 중요도가 떨어지지 않고 있음을 보여주는 결정적 증거다. AI 비용 축소가 구체화될 때마다, 시장은 생태계의 추가 확산 가능성에 주목할 것이다. Edge Device로의 AI 확산 가능성은 그간 침체되어 있던 범용 수요에 새로운 기대감을 안겨다 줄 것이라는 판단이다. 특히 쏠림 현상이 덜했던 한국 메모리 업체는 조정장 속에서도 시장을 아웃퍼폼할 가능성이 높다고 생각한다.

클라우드는 여전히 좋다

DeepSeek의 등장이 클라우드 기업에게 나쁠 것은 없다. 오히려 좋다고 판단한다. 학습과 추론 모두에서 클라우드 수요는 확대될 것이기 때문이다. 딥시크의 논문에도 나온 것처럼 증류 모델을 구축하기 위해서는 '좋은' 파운데이션 모델이 필요하다. R1의 증류 버전이 R1 zero의 증류 버전 보다 성능이 좋은 것이 분명한 근거다. 차세대 모델 구축을 위한 컴퓨팅 자원과 이를 강화하고 증류하기 위한 컴퓨팅 자원의 필요는 계속될 것이다. 또한 가성비를 추구하는 오픈 모델의 등장은 당연하게도 AI 서비스 즉 추론 성장으로 이어진다.

ROI에 대한 챌린징이 계속되겠지만, 클라우드 기업 입장에서 Capex 투자를 망설일 이유는 없다. 전일 실적을 발표한 마이크로소프트도 공격적 자본 지출 의지를 유지했다. "AI가 더 효율적이고 접근 가능해질수록 수요가 지수적으로 성장한다"는 샤티아 나델라 CEO의 코멘트가 현재의 상황을 요약한다. 일련의 산업 변화는 빅테크 클라우드 사업의 호조 전망을 강화하는 재료다.

OpenAI의 입장은?

DeepSeek의 발전을 바라보는 오픈AI의 기분은 개운하지 않을 것이다. 자신들이 구축한 모델 그리고 연구 방향성이 참고 자료처럼 쓰였으니 말이다. 하지만 DeepSeek의 발전은 프론티어 모델 시장을 교란하기보다 오픈 소스 및 미들 마켓을 교란한다고 보는 편이 맞다. 오히려 API를 통한 수익화에 초점을 맞췄던 앤스로픽이나 워 룸을 가동시켰다는 이야기가 들려오는 메타의 전략에 주목해야한다.

DeepSeek앱이 챗GPT를 제치고 앱 스토어 1등을 달성하기도 했지만, 이는 일시적일 가능성이 높다. 유저의 활성 사용량, 유지율 및 사용 시간 등 지표의 추가 확인이 필요하다. 관심이 과도하게 쏠리자 서비스 장애를 겪기도 했고, 현재 중국 로컬 번호를 보유해야 가입이 가능한 상태로 바뀌었다. 추론 서비스를 제공에도 컴퓨팅 파워가 당연히 필요하다. 감춰 놓은 반도체로 수요를 대응하는 것에도 한계가 존재할 것이다.

DeepSeek 모델의 단점인 편향성, 안전성, 개인정보 측면을 부차적으로 생각하더라도 보유한 데이터와 컴퓨팅 자원 그리고 구축한 생태계 규모에서 오픈AI가 경쟁에서 밀릴 이유는 없다. 오픈AI는 o1 대비로도 추가 개선을 달성한 o3로 이미 앞서있다. 샘 알트먼의 코멘트처럼 오히려 경쟁은 일종의 자극제가 될 것이다. 유저 입장에서 오픈AI가 압도적인 힘을 보여주기 위해 차세대 모델을 더 빠르게, 더 저렴하게 내놓을 가능성이 생겼다는 것은 오히려 반가운 일이다.

소프트웨어 기업에게는 기회의 장

AI 인프라 기업의 주가 조정과 달리 AI 소프트웨어 기업의 주가는 오히려 호조를 보였다. 단순하게 생각해보면 모델 레이어에서 알고리즘 효율성이 향상되고 비용이 감소하면 이를 활용하는 엔터프라이즈 소프트웨어에게는 긍정적이다. 특히 R1의 증류 모델의 높은 성과는 소형 및 도메인 맞춤형 모델 구축의 길을 제시하고 있다. 엔드 유저 활용과 다양한 유즈 케이스 확대를 기대할 수 있다.

2025년 AI 서비스가 본격화되는 원년으로 예상하는 기반에는 똑똑한 추론 모델의 등장도 있지만, API 가격이 급격하게 하락한 것도 있다. 파운데이션 모델 레이어의 경쟁 심화는 어플리케이션 레이어에게는 나쁠 것이 없다. 하지만 수급적 요인이 작용했다는 점도 분명히 생각해야한다. 서두에 확인한 것처럼 경량화와 효율성을 추구하는 모델의 발전 방향은 갑자기 나타난 것이 아니다. 엔드 솔루션을 제공하는 SW 기업은 Agentic AI 산업 개화의 기회를 제대로 포착할 수 있는지, 미들웨어 SW 입장에서는 추론 수요 상승에 따른 맞춤형 솔루션을 제공하며 실제적 수혜를 누릴 수 있는 지가 중요하다.

한국형 AI, 재입장하세요

AI의 비용 하락과 오픈소스의 발전은 한국 인터넷 기업들에게 제 2의 기회를 제공할 것이다. 네이버의 하이퍼클로바나 카카오의 카나나도 오픈소스로 공개된 모델들을 학습하며 개발되었던 만큼 저비용 오픈소스 모델은 이들에게 추가 성능 개선의 호재가 될 수 있다. 네이버는 자체 검색과 쇼핑, 광고, 컨텐츠에 생성형 언어 모델을 접목하여 서비스 경쟁력을 높이고, 카카오는 대화형 AI 챗봇을 도입하여 채팅 서비스의 기능을 강화하려는 전략이 힘을 받을 것으로 예상된다.

다시 빛을 보는 On device AI 내러티브

우리가 주장하는 On device AI는 철저하게 비용과 효율성 관점에서 접근하는 논리이다. 클라우드에서 AI 모델을 계산할 때의 성능이 매우 좋지만 비쌀뿐더러 Inference 과정에서 폭발적인 서비스 증가를 감당하지 못할 가능성이 있다는 단점이 있다. 디바이스에서 AI 모델을 계산할 때 비용은 드디어 무료에 가까워지지만 전력 소모와 디바이스 크기의 한계 때문에 아직까지 AI 계산 성능과 속도가 형편없다는 것이 시장이 열리지 않고 있다는 이유이다. 이번 DeepSeek 사태에서 투자자들은 증류와 양자화가 결합된 경량 모델이 얼마나 좋은 성능을 내는지 깨닫는 계기가 되었다. 물론 DeepSeek V3 모델 자체는 671b 파라메터의 크기로 디바이스에서 사용하기에는 매우 무겁다. 현재의 반도체 사양이라면 13b정도로 가벼워야 한다. 하지만 선생님 모델의 성능이 좋다면 이를 증류한 학생 모델의 성능이 좋을 것이라는 믿음이 생겼으며, 추론(Reasoning) 과정 속에서 효율성을 위해 일부는 작은 모델을 사용할 수 있다는 것도 공감하게 되었을 것이다.

이제 스마트폰, PC, 또는 일부 산업용 기기와 IoT에서 추론용 반도체 수요가 증가하게 될 것이라는 것이 부각될 가능성이 높다. 언어 모델의 계산은 항상 메모리 바운드가 상존하며 메모리 탑재량의 증가 스토리가 동반된다. AI 기능이 탑재된 디바이스와 탑재되지 않은 디바이스 사이의 성능 격차가 체감되는 것은 덤이다. 스마트폰과 PC 수요, HBM을 제외한 디램 탑재량 증가, 파운드리와 파운드리에서 생산하는 가벼운 AI 반도체들이 모두 수혜를 받게 된다.

Compliance notice

- 본 조사분석자료의 애널리스트는 2025년 1월 24일 현재 위 조사분석자료에 언급된 종목의 지분을 보유하고 있지 않습니다.
- 당사는 2025년 1월 24일 현재 위 조사분석자료에 언급된 종목의 지분을 1% 이상 보유하고 있지 않습니다.
- 본 조사분석자료에는 외부의 부당한 압력이나 간섭 없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.
- 본 조사분석자료는 당사의 저작물로서 모든 저작권은 당사에게 있습니다.
- 본 조사분석자료는 당사의 동의 없이 어떠한 경우에도 어떠한 형태로든 복제, 배포, 전송, 변형, 대여할 수 없습니다.
- 본 조사분석자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료 및 정보로부터 얻어진 것이나, 당사는 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 어떠한 경우에도 본 자료는 고객의 주식투자의 결과에 대한 법적 책임소재에 대한 증빙자료로 사용될 수 없습니다.
- 본 조사분석자료는 기관투자가 등 제3자에게 사전 제공된 사실이 없습니다.

삼성증권

삼성증권주식회사
서울특별시 서초구 서초대로74길 11(삼성전자빌딩)
Tel: 02 2020 8000 / www.samsungpop.com

삼성증권 Family Center: 1588 2323

고객 불편사항 접수: 080 911 0900