글로벌전략 한눈에보는상상인자산전략

Analyst 김경태 | 글로벌전략/AI
02-3779-3427
kt.kim@sangsanginib.com

QR코드로 간편하게
상상인증권
투자전략팀 채널

AI Comment: 지연에 따른 CAPEX 사이클 하방 공포

Oracle 데이터센터 지연 이슈와 시장의 과민 반응: 결국 비우호적인 Macro 환경에 기인

전주 초 FOMC는 시장의 예상보다 훨씬 비둘기파적인 톤이 나왔고, 고용지표 등 또한 양호했으나 시장은 Oracle과 Broadcom의 실적에 따른 AI 수익성 논란에 집중하는 양상을 보였다. 11월 전반적인 조정 이후 다시 반등을 바라봤던 시장 입장에서는 다소 당황스러운 상황이다. 최근 Oracle이 OpenAI 관련 데이터센터 완공 시점을 기존 2027년에서 2028년으로 일부 조정할 가능성이 있다는 보도가 전해지며, AI 인프라 투자 전반에 대한 시장의 우려가 급격히 확대되었다. 실적 자체는 컨센서스에 대체로 부합했음에도 불구하고, 데이터센터 구축 일정 지연 가능성만으로 Oracle 주가는 이틀간 15% 이상 급락했으며, Nvidia, AMD, Micron 등 AI 하드웨어 관련주 역시 2~4%대 동반 하락을 기록했다.

Oracle 측은 이후 "OpenAI와의 계약 이행에 차질은 없으며, 데이터센터 프로젝트 지연은 사실이 아니다"라고 공식 부인했으나, 이미 시장 심리는 크게 훼손된 상태였다. 이는 단순한 개별 기업 이슈라기보다는 현 매크로 상황에서 여러 우려가 혼재된 상황에서 AI 인프라 투자에 내재된 구조적 리스크와 과거 투자 사이클의 기억이 동시에 자극된 결과로 해석된다.

시장에 만연한 AI 회의론의 근거는 과거 CAPEX 붕괴의 PTSD

이번 조정 국면에서 투자자들의 반응이 유독 민감했던 배경에는, 과거 신기술 인프라 투자 사이클에서 반복적으로 목격된 '주문 지연 → 재고 누적 → 신규 주문 취소/하락 →CAPEX 기반 투자 Narrative 붕괴'의 트라우마가 자리하고 있다. 특징적으로 수요 소멸 이전에 프로젝트 지연과 취소가 먼저 발생했다는 점은 과거의 트라우마를 불러일으키기에 충분했다.

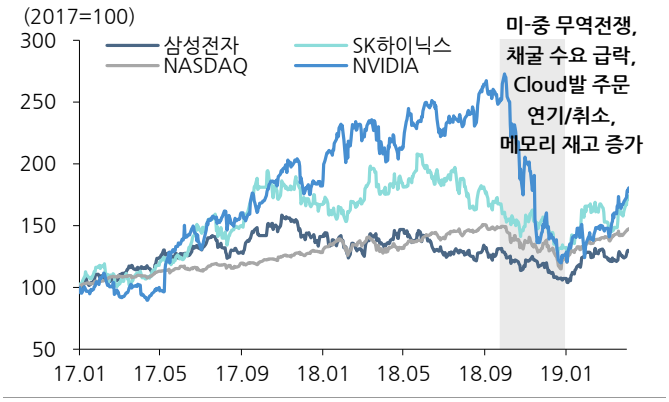
(1) 2018~2019년 GPU·메모리 사이클 하락 국면

2018년 하반기에는 비트코인/알트코인발 채굴 수요 둔화와 대형 클라우드 수요 조정이 맞물리며 CAPEX 사이클 하단이 나타났다. Nvidia는 당시 채널에 쌓인 구형 GPU 재고와 데이터센터 수요 둔화를 동시에 언급하며 가이던스를 대폭 하향했고, 이는 시장에 강한 충격을 주었다. 동시에 메모리 업황은 급격한 가격 하락 국면에 진입했다. 삼성전자와 SK하이닉스는 메모리 가격 하락과 재고 누적을 이유로 생산량을 축소했고, 밸류체인 내 반도체 소재/장비/부품 투자심리 역시 빠르게 위축되었다. 당시 시장은 이를 재고 소진 국면으로 해석했으며, 실제로 메모리 CAPEX 감소 → 장비 발주 축소 → 주가 급락이라는 전형적인 사이클이 재현되었다.

(2) 2022~2023년 데이터센터 재고 소진 국면

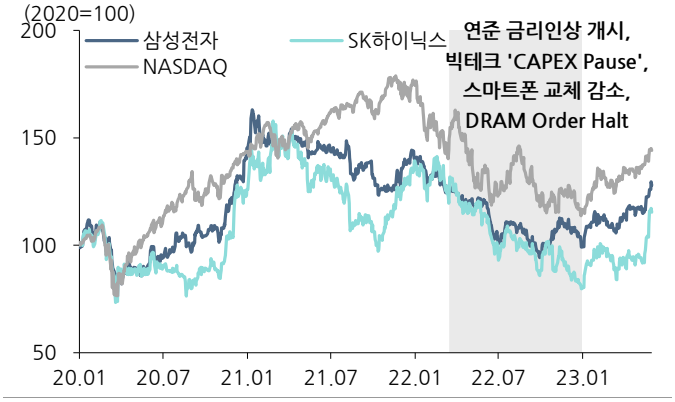
보다 최근인 2022~2023년에도 유사한 패턴이 반복되었다. PC·스마트폰 수요 둔화와 함께 데이터센터 고객들이 기존 재고를 우선 소진하기 시작했다는 점을 언급하였고, 그에 따라 메모리 가격은 급락했다. Micron과 SK하이닉스는 웨이퍼 투입 축소, 공급 성장률 하향 등 전형적인 CAPEX 방어 전략을 택했다.

그림 1. '17년~'19년 주요 종목/지수 상대주가 추이



자료: Bloomberg, Quantiwise, 상상인증권 리서치센터

그림 2. '20년~'23년 주요 종목/지수 상대주가 추이



자료: Bloomberg, Quantiwise, 상상인증권 리서치센터

현 국면의 본질적 차이: 수요 부족이 아닌 데이터센터 건설 관련 물리적 병목

다만 과거 사례들과 비교할 때 이번 국면에는 결정적인 차별점이 존재한다. 오라클 데이터센터 지연의 핵심 원인은 인공지능 수요 둔화가 아니라 전력, 인력, 계통 연결 등 물리적 인프라 병목에 가깝다. 실제로 다수 경영진은 그래픽처리장치 공급 자체보다 이를 수용할 데이터센터 건물, 전력 인프라, 계통 접속이 병목이라고 일관되게 언급해 왔다. 이는 과거처럼 주문 취소나 수요 소멸이 먼저 발생했던 국면과 질적으로 다른 양상이다.

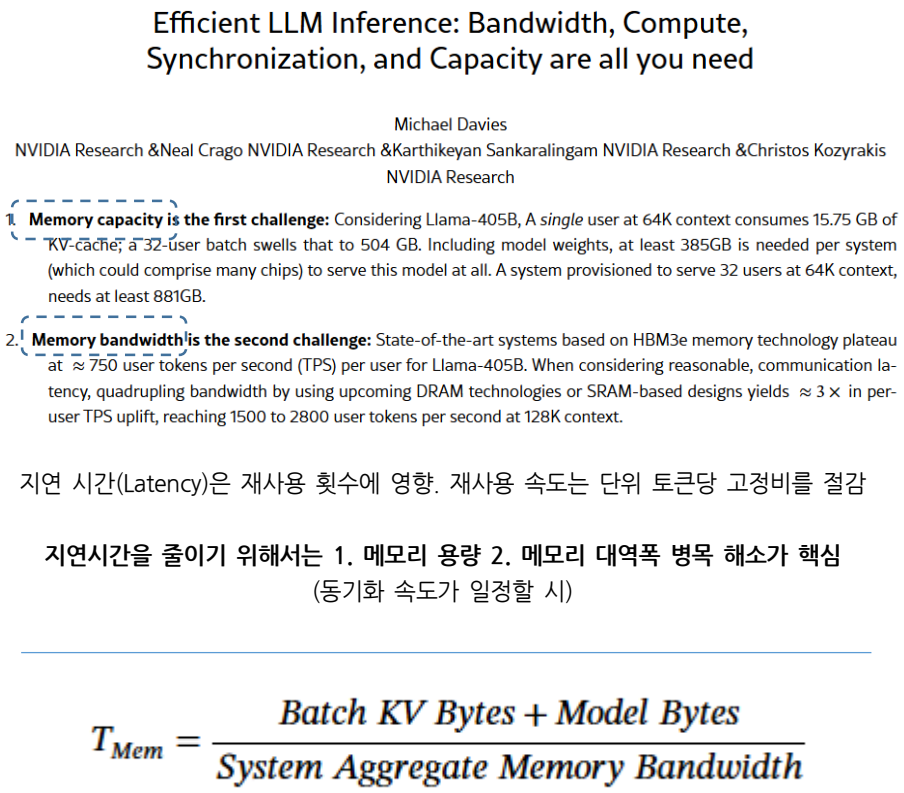
여기에 더해, 최근 대형언어모델 추론 비용 구조를 1차 원리로 분해한 연구는 디코딩 단계에서 메모리 대역폭과 메모리 용량, 그리고 다수 칩을 묶어 운영할 때의 동기화 지연이 처리량의 근본 제약임을 보여준다. 즉 인공지능 서비스 단가, 곧 인공지능 API 제공 비용은 단순히 연산 성능만이 아니라 메모리 대역폭이 얼마나 확보되는지, 대형 모델을 담을 충분한 메모리 용량이 있는지, 그리고 분산 시스템에서 동기화 지연이 낮아 고대역폭을 실제 처리량으로 전환할 수 있는지가 함께 결정한다. 이 관점에서 보면, 이번 이슈는 '수요 붕괴'의 조짐이라기보다 '물리적 병목'으로 인해 처리량 확장 경로가 일시적으로 비틀리며 투자자들이 비용과 수익성에 대한 민감도를 급격히 높인 사건에 가깝다.

또한 오라클을 포함한 주요 테크기업의 CDS 프리미엄이 가파르게 상승하고 있기는 하나, 지급 관련 문제가 실물로 나타났다고 보기는 어렵다. 위 기업들의 탄탄한 현금흐름과 유동성을 감안하면 향후에도 현실화될 가능성은 높지 않다.

이번 Oracle 이슈로 촉발된 AI 인프라 관련주 조정은 과거 CAPEX 붕괴의 실재했던 전례를 떠올리게 만든다는 점에서 시장의 반응 자체는 주가가 쉼 없이 상승을 반복했던 것의 반발심리가 작용한 것으로 보인다. 그러나 당시와 달리 현재 AI 인프라 시장은 수요 부족이 아닌 공급 제약 국면에 놓여 있으며, 이는 중장기 성장 스토리를 훼손할 요인은 아니다. AI 하드웨어와 전력·인프라 영역은 여전히 구조적 초과수요가 당분간 지속될 가능성이 높으며, 특히 증시 대부분 AI

관련 종목이 무차별적인 큰 하락을 시현한 만큼 이러한 상황에서 일부 재무상태가 건전하고 수익성이 탄탄한 기업들에 대해 재평가가 이루어 질 가능성이 농후하다. AI의 생산성 향상, 성능 증가에 대한 중장기 Conviction에 대해 훼손이 없는 한 AI 빅테크 및 관련 밸류체인, 인프라 기업에 대한 낙관론을 유지하며, 국내 반도체 펀더멘탈도 긍정적으로 전망한다.

그림 3. AI Token 제공 비용 & HBM의 관계식 분석 논문



Efficient LLM Inference: Bandwidth, Compute, Synchronization, and Capacity are all you need

Michael Davies
NVIDIA Research &Neal Crago NVIDIA Research &Karthikeyan Sankaralingam NVIDIA Research &Christos Kozyrakis NVIDIA Research

1. **Memory capacity is the first challenge:** Considering Llama-405B, A *single* user at 64K context consumes 15.75 GB of KV-cache, a 32-user batch swells that to 504 GB. Including model weights, at least 385GB is needed per system (which could comprise many chips) to serve this model at all. A system provisioned to serve 32 users at 64K context, needs at least 881GB.
2. **Memory bandwidth is the second challenge:** State-of-the-art systems based on HBM3e memory technology plateau at ≈ 750 user tokens per second (TPS) per user for Llama-405B. When considering reasonable, communication latency, quadrupling bandwidth by using upcoming DRAM technologies or SRAM-based designs yields ≈ 3 × in per-user TPS uplift, reaching 1500 to 2800 user tokens per second at 128K context.

지연 시간(Latency)은 재사용 횟수에 영향. 재사용 속도는 단위 토큰당 고정비를 절감

지연시간을 줄이기 위해서는 1. 메모리 용량 2. 메모리 대역폭 병목 해소가 핵심
(동기화 속도가 일정할 시)

$$T_{Mem} = \frac{Batch\ KV\ Bytes + Model\ Bytes}{System\ Aggregate\ Memory\ Bandwidth}$$

결국 시스템 레벨에서 LLM 비용 효율화에서 가장 중요한 것은 **고용량/고대역폭 HBM**

자료: NVidia Research, 상상인증권 리서치센터

Compliance Notice

- 본 자료에 기재된 내용들은 작성자 본인의 의견을 정확하게 반영하고 있으며 외부의 부당한 압력이나 간섭 없이 작성되었음을 확인합니다. (작성자: 김경태)
- 본 자료는 고객의 증권투자를 돕기 위한 정보제공을 목적으로 제작되었습니다. 본 자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료 및 정보를 바탕으로 작성한 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 참고자료로만 활용하시기 바라며 유가증권 투자 시 투자자 자신의 판단과 책임하에 최종결정을 하시기 바랍니다. 따라서 본 자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다.
- 본 자료는 당사의 저작물로서 모든 저작권은 당사에게 있으며 어떠한 경우에도 당사의 동의 없이 복제, 배포, 전송, 변형될 수 없습니다.
- 동 자료는 제공시점 현재 기관투자가 또는 제3자에게 사전 제공한 사실이 없습니다.
- 동 자료의 추천종목은 전일 기준 현재당사에서 1% 이상 보유하고 있지 않습니다.
- 동 자료의 추천종목은 전일 기준 현재 당사의 조사분석 담당자 및 그 배우자 등 관련자가 보유하고 있지 않습니다.
- 동 자료의 추천종목에 해당하는 회사는 당사와 계열회사 관계에 있지 않습니다.