meritz Strategy Daily

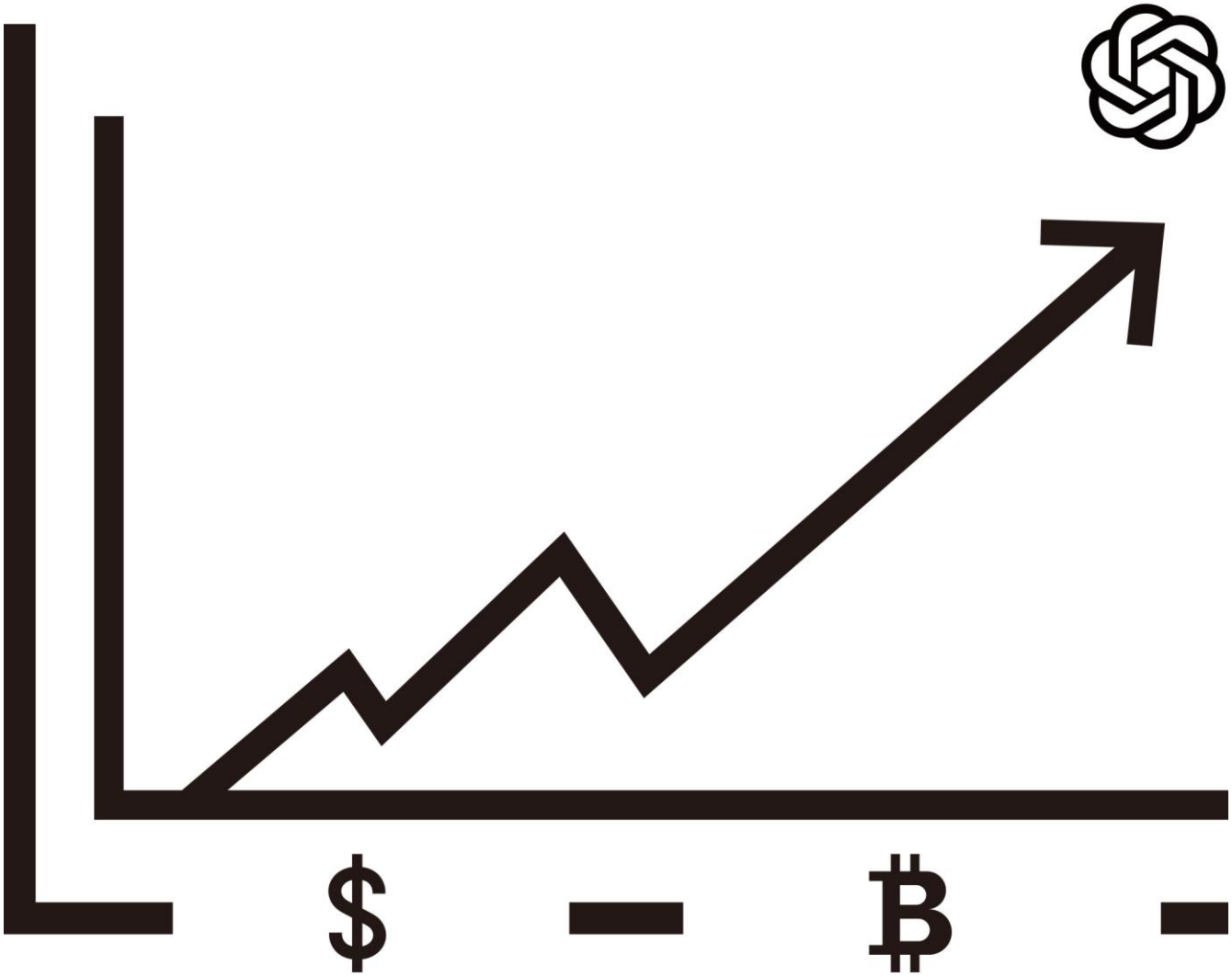
전략공감 2.0

2026. 6. 23 (화)

Strategy Idea
썰물(금리인상) 국면에서 더 중요해진 민낯(실적)

오늘의 차트
이익만큼 오른 KOSPI

칼럼의 재해석
GPU Rental Index: 오해와 진실



본 조사분석자료는 제3자에게 사전 제공된 사실이 없습니다. 당사는 자료작성일 현재 본 조사분석자료에 언급된 종목의 지분을 1% 이상 보유하고 있지 않습니다.
본 자료를 작성한 애널리스트는 자료작성일 현재 추천 종목과 재산적 이해관계가 없습니다.
본 자료에 게재된 내용은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 신의 성실하게 작성되었음을 확인합니다.

본 자료는 투자자들의 투자판단에 참고가 되는 정보제공을 목적으로 배포되는 자료입니다. 본 자료에 수록된 내용은 당사 리서치센터의 추정치로서 오차가 발생할 수 있으며 정확성이나 완벽성은 보장하지 않습니다. 본 자료를 이용하시는 분은 본 자료와 관련한 투자의 최종 결정은 자신의 판단으로 하시기 바랍니다. 따라서 어떠한 경우에도 본 자료는 투자 결과와 관련한 법적 책임소재의 증빙자료로 사용될 수 없습니다. 본 조사분석자료는 당사 고객에 한하여 배포되는 자료로 당사의 허락 없이 복사, 대여, 배포 될 수 없습니다.

meritz 메리츠증권

Strategy Idea

썰물(금리 상승) 국면에서 더 중요해진 민낯(실적)

투자전략
Analyst **황수욱**
soowook.hwang@meritz.co.kr

- ✓ 역사적으로 비교할 사례가 잘 안보이는, 원칙으로 돌아가 생각해야 하는 시장
- ✓ 썰물(금리 상승)에서 민낯(실적)이 두드러지는 국면. 이익에 근거한 쏠림
- ✓ 다만 펀더멘털이 양호한 종목까지 소외되는 불편한 쏠림, 완화 조건은?

쏠림에 대한 걱정, 원칙으로 돌아가 생각해야

KOSPI 9,000pt 시대가 개막했다. 1년 전에는 상상도 못하던, 2025년 순이익 200조를 벌던 시장이 2027년 1,000조를 벌 것으로 기대되는 시장이 되었다. 지금 주식 시장에서 보이는 숫자들은 대부분 누구나 처음 보는 숫자들일 것이다.

그런데 쏠림으로 모두에게 행복한 시장은 아닌 것 같다. 반도체만 오르는 시장이다. 대형주 두 종목에 쏠림이 심하다. 보통 주도주가 이끌면 시장 전반이 동반상승하는 그림이 일반적이었다. 그런데 이번 국면은 코스닥 등 중소형주는 하락하면서 대형주로의 쏠림이 심화되었다.

여러가지 시각이 있지만 이렇게 처음 보는 시장 양상이 전개될수록, 본질로 돌아가서 심플하게 봐야한다는 생각이다. 이하에서는 주가(P)를 이익(EPS)과 멀티플(PER)로 이분해서 접근하는 관점을 제시한다.

그림1 **코스피 내 삼성전자, SK하이닉스 시가총액 비중**

자료: FnGuide, 메리츠증권 리서치센터

그림2 **코스피 대비 코스닥 시가총액 비중**

자료: FnGuide, 메리츠증권 리서치센터

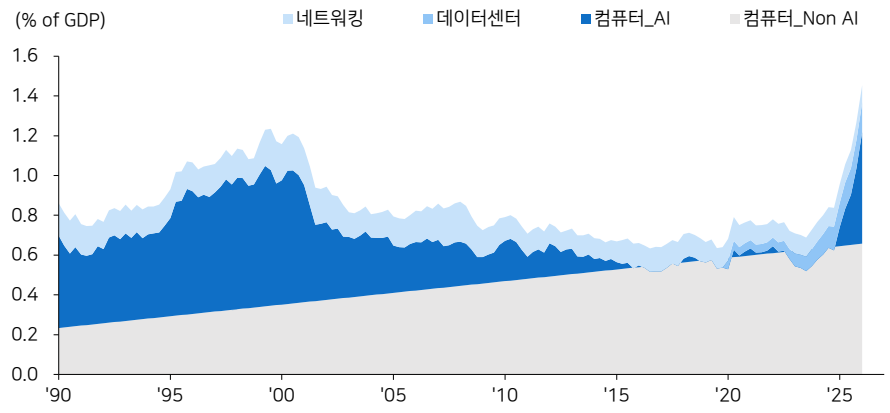
과거 사례와 비교할 수 있는 시기가 있을지

어떤 과거 사례와도 비교하기 어려운 이유는 지금 시장참여자들은 모두 처음 겪는 국면이기 때문이다. AI가 수반하는 대규모 투자 사이클은 90년대 PC/인터넷 사이클보다 크다고 본다(26.5.28 발간 Equity Strategy: Atom to Bit to Token 참고)

과거 인프라 사이클과 규모의 비교 기준을 GDP 대비 CapEx 사이즈로 본다면, AI 사이클은 보수적으로도 20년간 투자 사이클이 이어진 것으로 평가되는 Roaring 1920's의 전기화/대량생산 사이클과 필적한다. 여기에 젠슨 황 CEO 등의 AI 산업 내 등장하는 공격적인 전망은 50년의 투자 사이클이 이어졌던 19세기 철도/운송 사이클에 근접하는 규모다.

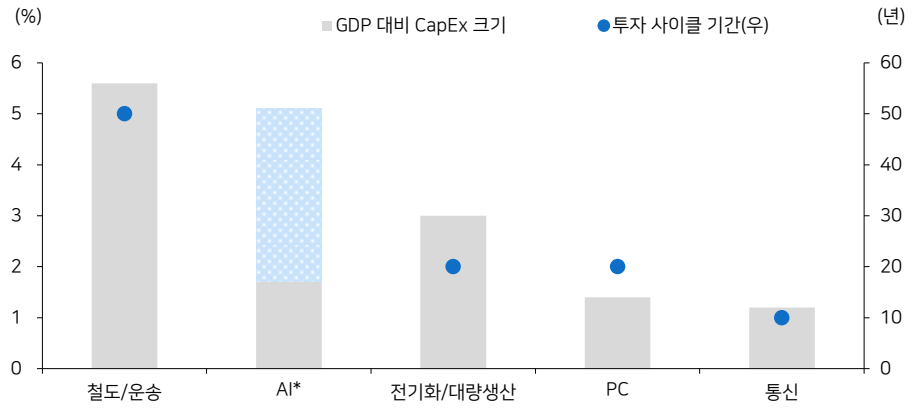
1990년대 이후 시계에서 IT 혁명과 산업 투자 사이클에 국한되었던 클라우드 등 소순환 투자 사이클의 주가 패턴의 틀에 갖혀 있기에는 지금이 훨씬 더 크다고 생각한다. 단순히 최근 30년 이내의 경험을 비교하기에는 차이가 있어 보인다.

그림3 **최근 30년 GDP 대비 IT CapEx 비중: 모든 시장 참여자가 처음 겪는 투자 사이클**



자료: US BEA, US Census Bureau, 메리츠증권 리서치센터

그림4 **각 기술 사이클 미국 GDP 대비 peak 연간 CapEx와 투자 사이클 길이**



*주: 2026년 IMF WEO, Dell'Oro 기준 1.7% 예상, 공격 전망인 엔비디아(젠슨 황) 전망 기반 2030년 5%까지 추정
자료: Association of American Railroads, NBER, AWS, Dell'Oro, IMF WEO, 메리츠증권 리서치센터

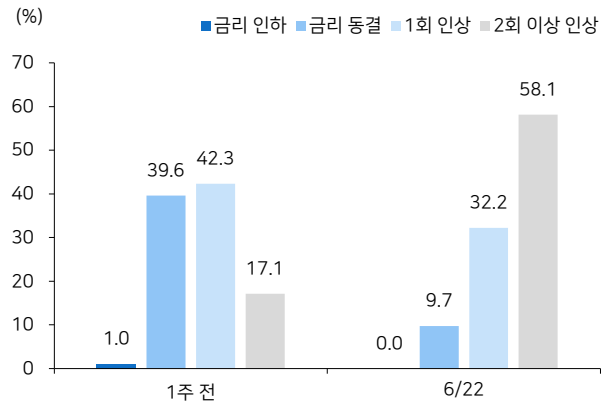
매파적으로 해석되는 FOMC, 기대 인플레와 별개로 우상향하는 실질금리

FOMC도 쏠림의 매개 중 하나로 본다. 6월 FOMC 이후 CME Fed Watch에 집계되는 연내 금리 인상 확률은 59.4%에서 90.3%까지 반영되기 시작했고, 이중에서 2회 인상 확률은 17%에서 58%까지 급등했다.

기준 금리에 대한 전망이 상향되는 가운데, 최근 TIPS로 대변되는 실질금리가 상승 중이다. 명목금리가 인플레 기대와 실질금리로 구성된다고 할 때, 연초 이후 전쟁으로 오르던 기대 인플레(BEI가 proxy)는 유가 안정과 함께 하락하며 전체 명목금리의 안정에 기여하고 있지만, 실질 금리는 계속 오르는 중이다.

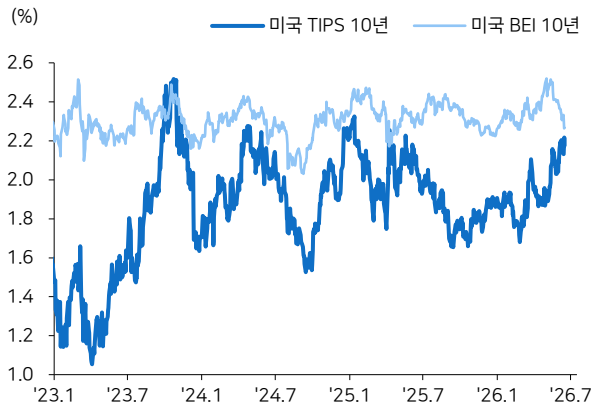
과거 산업혁명기를 보면 혁명 초기에는 장기금리가 우상향 추세를 그려 왔다. AI 진보가 노동시장이라는 연결고리로 경제 전체의 상당 부분을 차지하는 소비로의 낙수효과를 끊는다는 불안감은 있지만, 여전히 AI가 이끄는 견고한 미국 경기는 실질금리의 상승을 지지한다.

그림5 FOMC 이후 연내 금리 인상 기대 강화



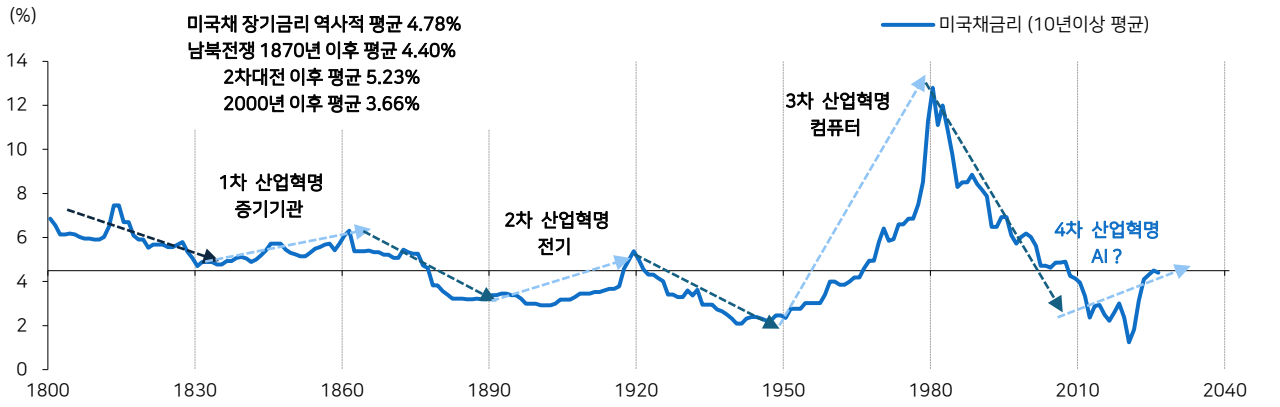
자료: CME Group, 메리츠증권 리서치센터

그림6 실질금리 중심으로 오르는 채권 금리



자료: Bloomberg, 메리츠증권 리서치센터

그림7 산업혁명과 미국 장기 금리



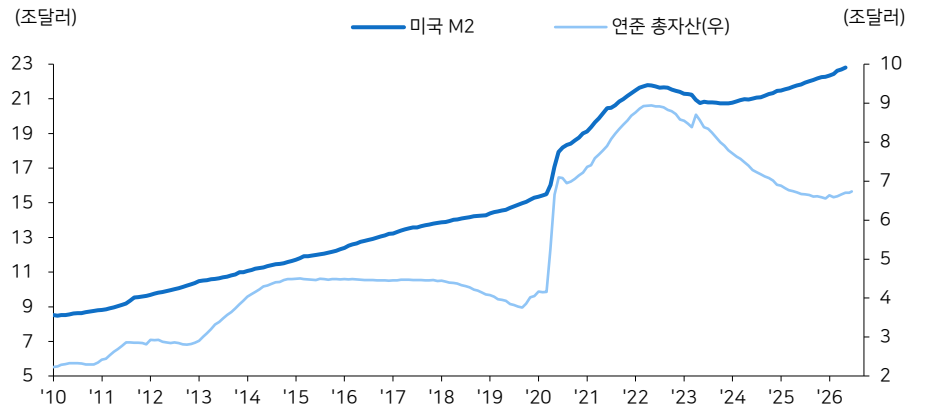
자료: NBER, 메리츠증권 리서치센터

정책 유동성과 시장 유동성의 구분: 지금은 유동성 장이 아니다

금리 상승은 유동성 환경의 위축을 의미한다. 지금은 유동성 장이 아니다. 시장 참여자들이 익숙한 유동성 장은 정책 유동성이 강하게 밀려들어와 자산 가격을 키워주는 국면이다. 제로금리와 QE가 수반되었던 2010년대 초, 2020년-2021년이 그 예이다. M2의 우상향을 근거로 유동성 장을 주장하는 시각도 있으나, M2의 대부분은 민간 유동성이다. 2023년 이후 연준 총자산 규모로 대변되는 정책 유동성은 QT 및 금리 인상과 함께 줄어왔다고 보는 것이 타당하다.

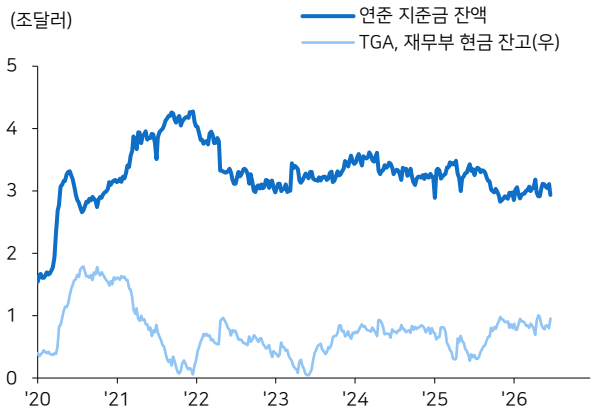
연초 이후 연준 총자산 규모가 다시 증가하는 점에 대해서도 올바르게 해석해야 한다. 26년 3월 10일자 전략공감에 상세하게 정리해두었으나, 최근 연준 총자산 규모 증가는 〈그림 10〉처럼 COVID 이후 가장 강하게 단기채 매입을 진행하기 때문이다. 연준은 이를 완화적 통화정책 맥락이 아니라 TGA 잔고 충당에서 비롯될 단기 유동성 시장 변동성을 줄이기 위한 예방적 공급의 일환으로 설명한다.

그림8 M2와 연준 총자산: 정책 유동성과 시장 유동성의 구분



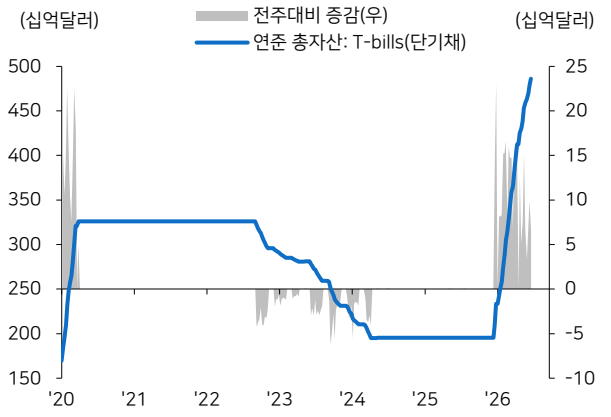
자료: Federal Reserve, 메리츠증권 리서치센터

그림9 TGA 잔고 증가, 연준 지준금 감소



자료: Federal Reserve, 메리츠증권 리서치센터

그림10 연준 단기채 매입: TGA 잔고 증가에 따른 우려 상쇄



자료: Federal Reserve, 메리츠증권 리서치센터

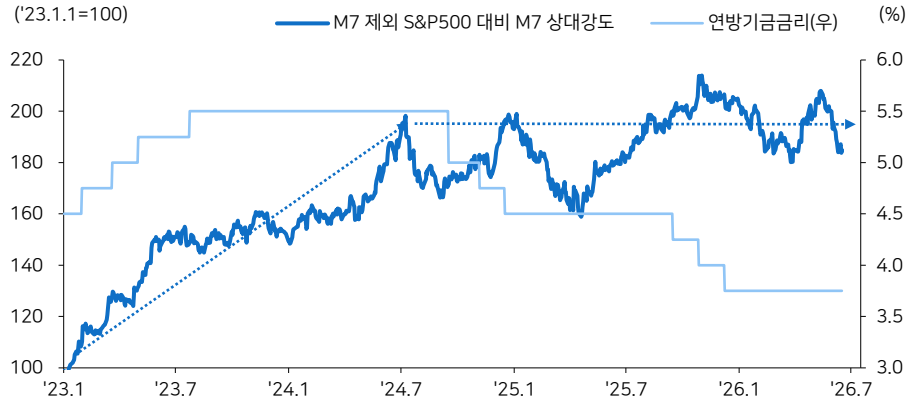
'23-'24년 미국도 금리 인상 가운데 대형 AI 실적주(magnificent7) 독주

필자는 2023년 이후 주식시장을 실적장이라고 주장해왔다. 그때는 미국 시장을 중점적으로 분석하던 시기였는데, 2023년 초기 Magnificent 7라는 개념이 처음 등장했다. 그리고 2024년 상반기까지는 미국에서도 이 주식들만 주가가 올랐다.

지금 복기해보면 이유는 더 선명해 보인다. 먼저 이익 모멘텀이 압도적이었다. Magnificent 7의 전년대비 순이익 성장률은 2023년 하반기에 60%에 근접했다. 동기간 나머지 S&P500의 이익은 여전히 역성장 중이었다. 2023년은 여전히 금리 인상 구간이었다. 연준은 2023년 하반기까지 기준금리 상단을 5.5%까지 인상하고 higher for longer를 주장했다.

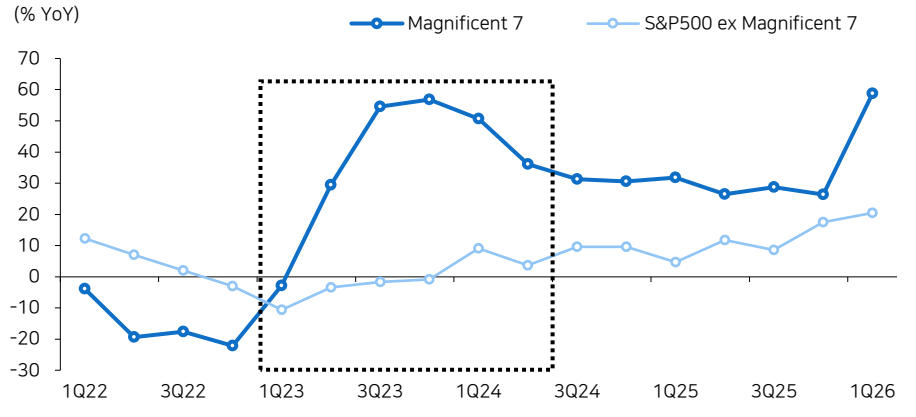
〈그림 11〉처럼 미국 Magnificent 7의 독주가 멈춘 것은 2024년 하반기였다. '23년까지 200%를 상회하던 엔비디아의 분기별 매출액 성장률이 눈에 띄게 피크아웃하기 시작했고, 9월 예상치 못했던 50bp 금리인하를 단행했다. 이때부터 M7 상대강도가 횡보하고, 팔란티어를 필두로 고밸류에이션 AI SW 주식이 아웃퍼폼했다.

그림11 M7 제외 S&P500 대비 Magnificent 7 상대강도: 금리 인상 국면에서 아웃퍼폼



자료: Bloomberg, 메리츠증권 리서치센터

그림12 Mangificent 7 vs M7 제외 S&P500 순이익의 YoY 성장률



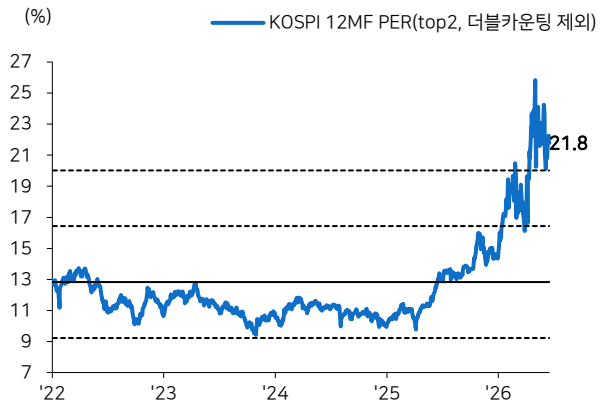
자료: Bloomberg, 메리츠증권 리서치센터

물이 빠질 때 기본은 펀더멘털에 집중

미국 사례에서 확인되는 점은, 금리가 상승하며 유동성 환경이 타이트해지는 구간에서는 실적의 민낯이 더 선명하게 드러난다는 것이다. 우호적인 유동성 환경이 모든 자산 가격 상승을 일으키는 국면과는 다르게 기본으로 돌아가 실적에 기반한 주가 차별화를 더 선명하게 드러나게 했다.

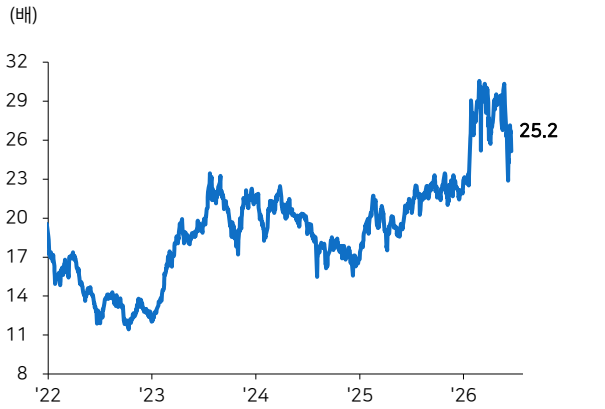
'Atom to Bit to Token' 하반기 전망에서 분석했듯, 삼성전자와 SK하이닉스를 제외한 코스피는 더블카운팅 이익을 제외한 실질적인 PER 수준을 고려하면 현재 금리 수준에 민감해질 수 있는 위치에 있다. 미국시장의 earnings yield(PER의 역수)가 장기금리에 근접한 이후 추가적인 밸류에이션 상단이 제한되고 있는데, 한국 시장의 멀티플도 유사한 수준으로 금리에 민감해질 수 있는 수준이라는 것이다. 삼성전자, SK하이닉스를 제외한 나머지 코스피뿐만 아니라 코스닥도 마찬가지다. 밸류에이션 플레이가 제한되는 금리 환경에서 펀더멘털에 집중되는 시장 환경이다.

그림13 Top2, 더블카운팅 이익 제외 코스피 12MF PER



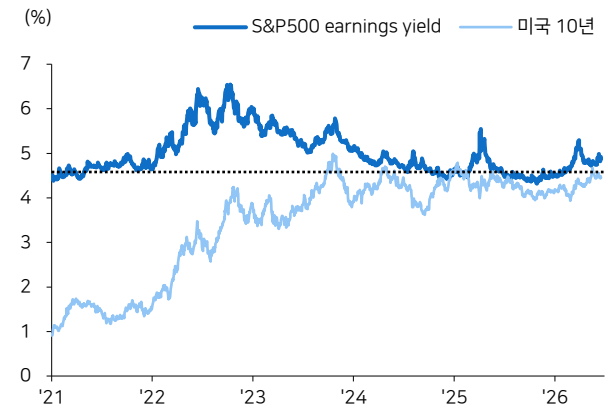
자료: FnGuide, 메리츠증권 리서치센터

그림14 코스닥 12MF PER



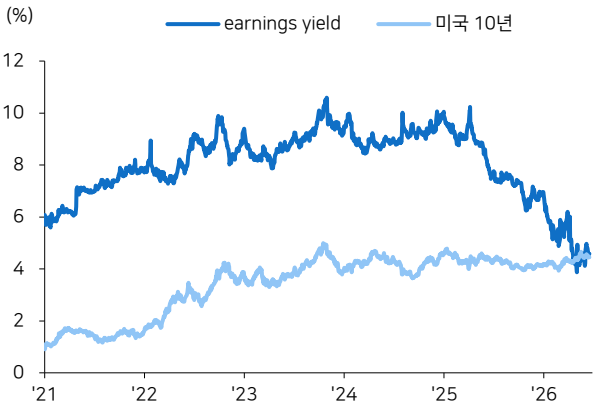
자료: FnGuide, 메리츠증권 리서치센터

그림15 S&P500 어닝스 일드 vs 미국채 10년



자료: Bloomberg, 메리츠증권 리서치센터

그림16 Top2, 더블카운팅 제외 코스피 어닝스 일드 vs 금리



자료: FnGuide, Bloomberg, 메리츠증권 리서치센터

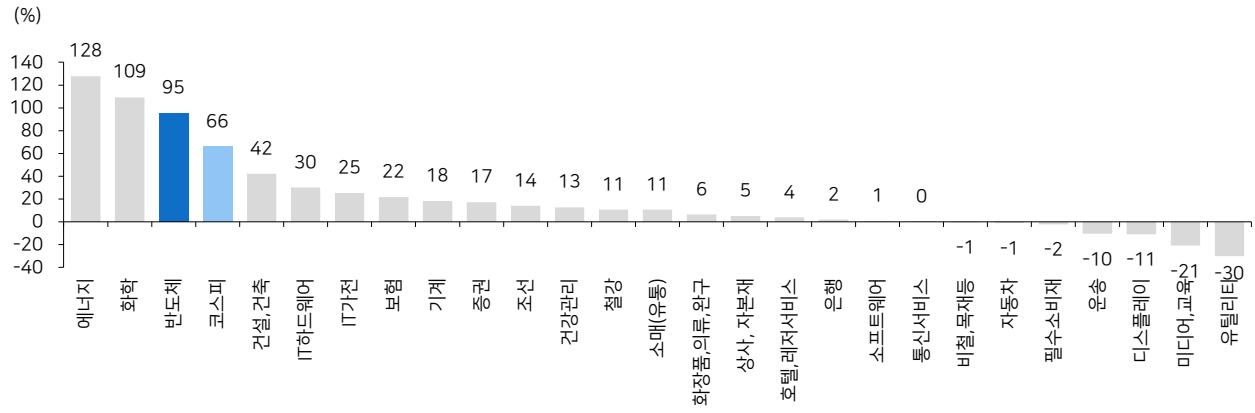
실적 모멘텀으로 시장을 이기는 업종이 실질적으로 반도체뿐

한국 시장은 유가 변동성으로 인해 지속 가능성이 높지 않고 저점에서 큰 폭으로 반등했던 에너지, 화학 업종의 실적 리레이팅을 제외한다면, 실질적으로 반도체보다 실적 추정치 상향 모멘텀이 강한 업종이 부재하다.

종목 단에서는 아쉽게 소외되는 종목들도 존재, 이런 분위기의 완화 조건은?

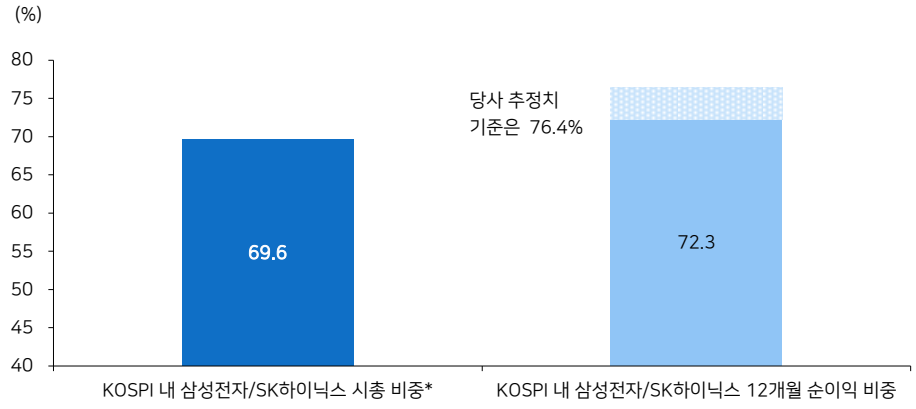
종목 단에서는 분명히 펀더멘털이 양호한 종목들이 손해를 보고 있는 것도 사실이라고 생각한다. Top2 일변도 분위기가 약해져야 이들도 제대로 된 평가를 받을 수 있을 것이다. 정확히 알 수는 없지만 기준을 한가지 생각해보면, 시장에 반영된 시가총액이 이익 대비 저평가 되어있다는 인식이 완화되는 조건을 고려해볼 수 있다. KOSPI 향후 12개월 순이익 중 Top2(삼성전자/SK하이닉스) 이익 비중은 당사 추정치 기준 76.4%다. 여기에 지주사 지분 가치까지 고려한 Top2 시총은 69.6%이다. 강한 반도체 실적 상향이 추가적으로 이어질 가능성이 높은 가운데, 이 갭이 줄어드는 것이 소외주 반등의 1차 조건일 수 있겠다.

그림17 **코스피 및 업종별 최근 3개월 이익 추정치 변화율: 유가 변수를 제외하면 코스피 업종 내 반도체를 아웃퍼폼 업종 부재**



자료: FnGuide, 메리츠증권 리서치센터

그림18 **KOSPI 내 삼성전자/SK하이닉스 비중: 시가총액 vs 향후 12개월 순이익 (6/22 기준)**

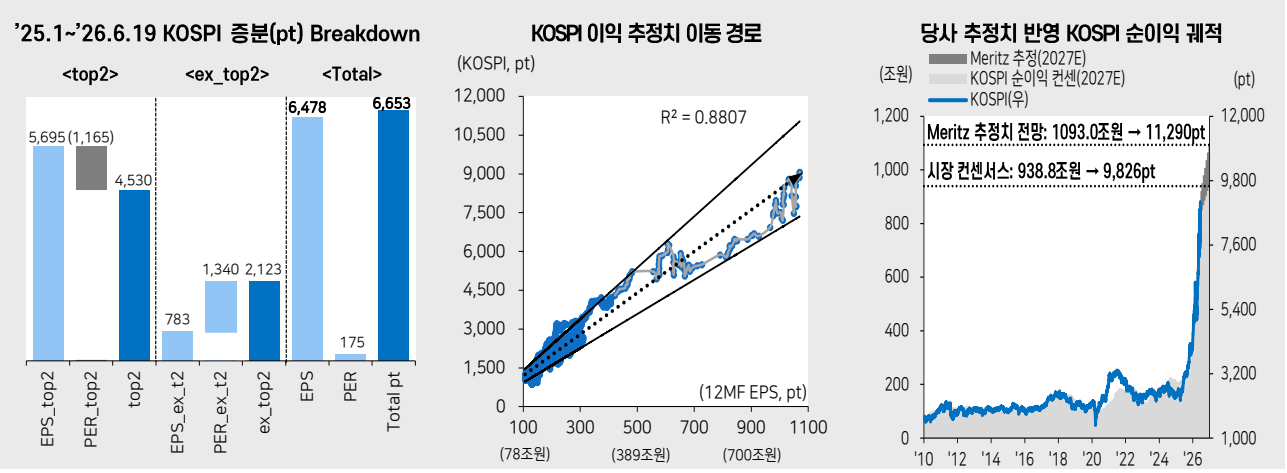


*주: 삼성물산, 삼성생명, 삼성화재, SK스퀘어의 시가총액에 반영된 삼성전자, SK하이닉스 지분 가치 포함
자료: FnGuide, 메리츠증권 리서치센터

오늘의 차트

염승준 연구원

이익만큼 오른 KOSPI



주1: 모든 데이터는 6월 19일 종가 기준으로 산출. '24.12.30 종가를 기준으로 일별 삼성전자, SK하이닉스 지수 기여 pt를 계산해 증분(pt) 기여도를 도출
주2: KOSPI 구성종목 중 이익 추정치가 존재하는 331개 사에서 삼성전자, SK하이닉스 이익 추정치를 제외한 값을 KOSPI 전체 순이익으로 간주해서 계산
주3: KOSPI 구성종목 중 이익 추정치가 존재하는 331개 사의 KOSPI 대비 시가총액 비중은 98.2%. Meritz 커버리지의 KOSPI 대비 시가총액 비중 85.3%
주4: Meritz 추정(2027E) 값은 당사 애널리스트 커버리지 추정치를 1순위, 시장 컨센서스를 2순위로 사용하여 Bottom-up 합산으로 도출
자료: FnGuide, 메리츠증권 리서치센터

'25.1~'26.6.19 기간 동안 KOSPI는 Δ6,652.93pt 증가했다(2,399.49 → 9,052.42). 해당 KOSPI 증분을 top2(SK하이닉스, 삼성전자)와 ex_top2 기여도로 나눠 계산해보면 top2가 증분의 68%(4,529.77pt)를 설명한다. 반도체 쏠림 현상은 시장이 모두 알고 있는 바 이번 차트에서 주목할 부분은 Total 관점이다.

동기간 KOSPI 12MF PER은 8.17배 → 8.46배, 12MF EPS는 293.7p → 1,070.0p로 상승했다. 주가 수익률, PER 변화율, EPS 증가율을 각각 로그 차분으로 계산하면 단순 변화율 기준에서 발생하는 교차항 효과를 배제하고 주가 변동을 EPS 요인과 PER 요인으로 가법적 분해할 수 있다. 이에 KOSPI 6,652.93pt 증분 중 6,478.16pt를 EPS로 설명할 수 있다.

KOSPI 지수는 이익 추정치 상향 정도를 가격에 온전히 반영하고 있다. 다만 이익 추정치를 가격에 반영하는 시차는 존재한다. KOSPI 이익 추정치 이동 경로에서 KOSPI는 실적 시즌을 소화하며 이익 추정치 상향 속도를 주가가 따라잡지 못하는 국면이 존재하지만 결국은 장기 추세를 회복하는 모습을 보인다.

시사점은 ① 이익 추정 정확도가 향후 지수 레벨을 판단함에 있어 중요한 변수로 작용할 것 ② 이익 추정치에 상응하는 지수 레벨을 Base, 추가적인 업사이드는 멀티플로 설명하는 것이 합리적으로 보인다. 당사는 2027E 순이익을 1,093.0조원으로 추정, 시장 컨센서스(938.8조원) 대비 154조원 업사이드가 있을 것으로 추정한다. 컨센서스와 괴리는 SK하이닉스, 삼성전자 이익 가정 차이에서 대부분 발생한다. 2026년말 KOSPI는 당사 이익 추정치 기준 이익 성장만을 반영해도 11,290pt를 달성할 수 있을 것으로 전망한다.

칼럼의 재해석

박보경 연구원

GPU Rental Index: 오해와 진실 (Silicon Data)

지난달 Silicon Data는 H100 하이퍼스케일러 지수를 "정체 상태의 벤치마크"로 설명했다. 실제로 4~5월 61일 동안 하이퍼스케일러 H100 가격은 매우 안정적이었던 반면, 같은 칩의 네오클라우드 가격은 더 낮지만 훨씬 큰 변동성을 보였다. 이는 AWS·Azure·Google Cloud의 공시 가격이 다년 예약계약과 이미 가격이 정해진 대형 계약 물량 위에 형성된 일종의 '관리되는 가격'이기 때문이다. 이 시장의 고객은 SLA, 데이터 이동 비용, 보안·컴플라이언스, 기존 운영환경 때문에 작은 가격 차이에 쉽게 공급자를 바꾸지 않는다.

반면 네오클라우드 H100 가격은 4월 초 2.63달러에서 4월 8일 2.47달러까지 약 6% 하락했다가, 5월 말에는 2.73달러까지 상승했다. 5월 한 달만 보면 약 7.1% 올랐고, 하루 최대 변동폭도 9센트에 달했다. 네오클라우드는 CoreWeave, Lambda, Crusoe, Nebius처럼 수주~수개월 단위로 용량을 재계약하는 경우가 많다. 따라서 신규 GPU가 들어오거나 기존 예약이 만료돼 가용 풀로 돌아오면, 해당 물량은 그 시점의 주간·월간 시장 가격에서 다시 거래된다. 또한 네오클라우드 고객은 베어메탈형 인프라를 쓰는 경우가 많아 전환비용이 상대적으로 낮고, 의미 있는 가격 차이가 나면 다른 공급자로 이동하기 쉽다. 그래서 수급 변화가 며칠 내 가격에 반영될 수 있다.

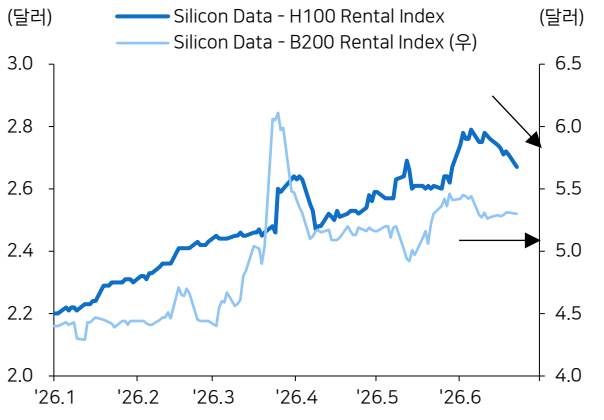
결국 하이퍼스케일러 지수는 기업 예산 편성과 장기 예약계약 협상의 기준점으로 적합하고, 네오클라우드 지수는 스팟·단기 조달 의사결정과 AI 인프라 수요 변화의 선행지표로 더 유용할 것이다.

GPU Rental Index 하락 = GPU 수요 둔화?

6월 이후 GPU Rental Index가 조정을 겪으며 AI 연산수요 둔화 우려를 자극하고 있다. 연초 이후 상승세를 지속해온 H100 Rental Index는 5월 말 대비 -2.9% 하락한 상태이다. B200는 연중 상승세가 제한되었음에도 불구하고 동 기간 약 -2.8% 하락했다. 최신 GPU를 중심으로 형성됐던 임대료 프리미엄이 약화되자, 단기 조정이 아닌, GPU 수요의 둔화 신호로 해석하려는 시각도 나타나고 있다.

최근 확산된 'TokenMaxxing' 흐름도 우려를 가중시키고 있다. 〈표 1〉과 같이 일부 기업에서 AI 사용량과 비용 효율을 재점검하는 움직임이 포착되면서 토큰 소비의 지속 가능성에 대한 의문이 제기되고 있다. 이는 곧 AI 연산수요가 예상보다 빠르게 비용 통제 국면으로 진입하고 있는 것 아닌 지에 대한 우려를 낳고 있다.

그림1 GPU Rental Index의 상승세 둔화,



자료: Silicon Data, 메리츠증권 리서치센터

표1 TokenMaxxing 우려까지 가중되는 상황

	토큰 정책	결과
Amazon	사용량 순위표 운영 개발자 사용 의무화	순위표 폐지, 사용 제한
Meta	사용량 순위표 운영, 1위 사용자 '토큰 레전드' 칭호 부여	30일간 토큰 60조개 사용, 순위표 삭제
Uber	사용량 순위표 운영, 사용 예산 수용	CEO, "토큰 소비와 기능 개선 연결 못찾아"
Shopify	가장 먼저 순위표 도입, 대량 사용자 옹호	순위표를 사용현황 대시보드로 변경

자료: 언론 종합, 메리츠증권 리서치센터

오해: 제한된 지수 산출 대상

해당 지수가 전체 GPU 시장의 수급을 반영하지 않는다는 점에 주목할 필요가 있다. GPU 임대 시장은 크게 장기계약, 온디맨드, 스팟 시장으로 구분 가능하다. 각각 중장기 및 반복적 운영수요, 유연성 확보 수요, 단기수요를 대변하는데, H100 및 B200 Rental index는 장기 전용계약 및 다년계약은 반영되어 있지 않다.

장기계약 및 온디맨드 계약이 AI 연산의 구조적인 수급을 더 잘 대변한다는 점에서 해당 지수를 통해 전체 GPU 수요를 조망하는 것의 한계가 더욱 분명해진다. 스팟 시장에서 거래되는 GPU 용량은 일반적으로 장기계약으로 이어지지 않은 유휴용량에 해당하는 경우가 많다. 따라서 신규 데이터센터 파트너의 편입, 특정 지역의 유휴 용량 증가, 기존 계약의 만료만으로도 스팟 가격은 하락할 수 있다.

수요 측면에서도 마찬가지이다. 고객 역시 중단 가능성이나 사용 지역의 제약을 감수할 수 있는, 비교적 초기 실험 단계의 AI 업무에 스팟 계약을 활용한다. 이처럼 유동적이고 가격 탄력성이 높은 수요 환경에서는, 공급이 소폭 증가해도 단기 가격이 빠르게 조정될 수 있다. Rental Index의 하락을 전체 GPU 수요의 둔화로 해석하기에는 한계가 있다.

표2 GPU 임대 시장 구분 및 관련 지표

계약 형태	구매 용량	특징	수요	관련 지표
장기계약	향후 수개월 - 수 년의 전용 GPU 용량	공급 보장, 옵션 및 일정 확실성	구조적 · 계획형 수요	-
온디맨드	필요할 때 바로 사용하는 안정적인 용량	가용성 및 서비스 편의성	유연성을 중시하는 운영 수요	Silicon Data B200 Rental Index : 네오클라우드 및 마켓플레이스의 온디맨드 B200 가격을 반영
스팟	유휴 GPU를 중단 가능한 조건으로 사용	당일 잔여 공급에 민감	가격에 민감한 단기 수요	Silicon Data H100 Index : 주로 네오클라우드 및 마켓플레이스의 단기 · 월 단위 조달 및 가격 민감형 수요 포착

자료: 메리츠증권 리서치센터 정리

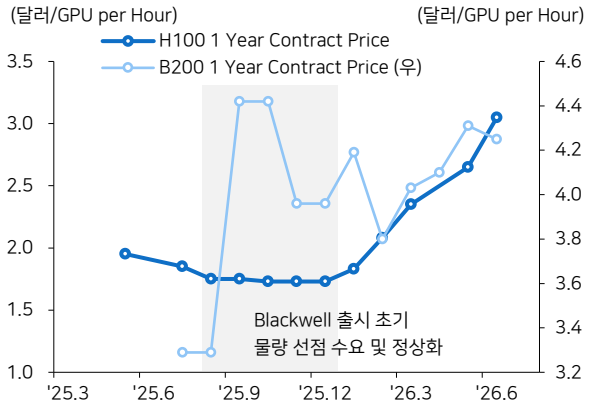
진실: 장기계약 가격은 견조, 안정성 조달의 필요성 반영

한편 중장기 GPU 임대시장에서는 초과수요의 지속이 확인할 수 있다. 대표적으로 H100의 1년 계약 임대료는 3-4월에 이어 상승 추세 지속되고 있다. B200 역시 출시 직후의 가격 급등 및 정상화 과정을 거친 이후, 지난 4월부터 상승세를 이어오고 있다. 네오클라우드 기업들의 백로그 및 장기계약 가치 급증도 장기 GPU 임대 수요가 확대되고 있음을 간접적으로 드러내는 요인이다.

그 외에도 중장기 GPU 임대료 인상의 증거들은 충분하다. 일부 AI 인프라 기업은 B200 임대 계약을 갱신할 때, 시간 당 사용료가 2.63달러에서 5.10달러로 두 배 수준 급등한 바 있다고 밝힌 바 있다. Nebius도 마찬가지로 6월 1일부터 전반적으로 시간당 GPU 임대료를 30% 인상한 바 있다.

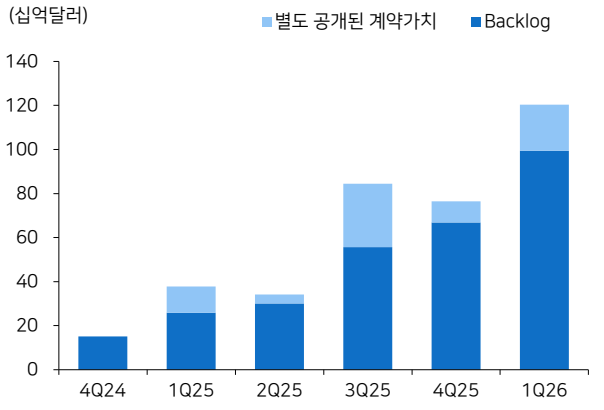
한편 장기계약 강세와 스팟 약세가 모순되어 보일 수 있으나, 두 현상을 동시에 설명할 수 있는 상황이 있다. 앞서 언급한 바 있듯이 스팟 물량은 주로 장기계약 용량을 배정한 뒤에 남는 비선호 지역, 단기, 중단 위험이 높은 물량이다. 이러한 스팟계약의 특성을 고려했을 때, 장기계약과 스팟 가격의 차별화는 수요의 방향이 즉각적인 GPU 확보에서 안정된 장기 공급으로 이동 중임을 시사한다. 이는 AI 워크로드가 단발성 실험에서 반복적·운영형 수요로의 전환과도 관련이 있다.

그림2 **1년 GPU 임대계약은 세대 불문 상승 추세 지속**



자료: GetDeploying, 메리츠증권 리서치센터

그림3 **네오클라우드 백로그도 장기계약 비중 확대를 시사**



자료: CoreWeave, Nebius, Iren, 메리츠증권 리서치센터

H100의 상대적 강세도 운영수요 확대를 반영하고 있을 가능성

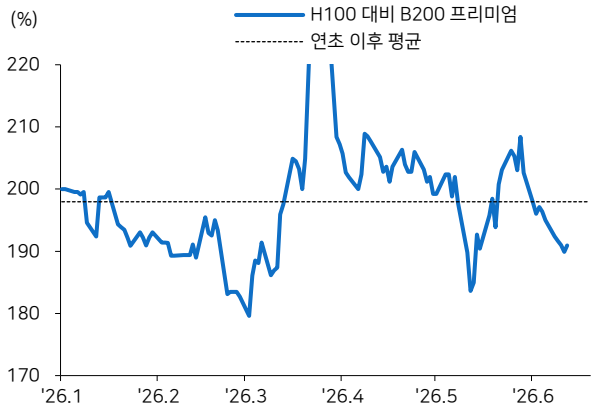
한편 시장이 주목하고 있는 또 하나의 변화는 B200 Spot 가격의 프리미엄 축소이다. 올해 4월까지만 해도 B200이 H100 대비 2.12배 더 높은 가격에서 거래되어 왔다. 다만 6월 19일 기준, B200의 프리미엄이 1.95배까지 하락했다.

한편 해당 현상은 B200의 수요 둔화보다는 H100 역할 재부각의 결과로 해석되고 있다. AI의 파일럿 단계에서는 최신의 GPU를 당장 확보하는 것이 더 중요하기 때문에 고사양 GPU 수요가 강할 수 있다. 다만 AI 워크로드가 실제 업무 운영 단계로 진입할수록 최소 비용으로 필요한 처리량을 안정적으로 맞추는 것이 중요해진다. 이에 따라 워크로드가 구형 GPU로 분산되는 흐름이 발생하는 것이다.

특히 B200의 성능우위는 높은 동시성, 대규모 배치, 긴 컨텍스트의 환경에서 토큰당 비용우위로 작용한다. 반면 중형 모델, 간헐적 수요 환경에서는 H100의 낮은 시간당 비용이 더 경제적일 수 있다. UC Berkley 역시 단일 종류의 고사양 GPU로만 워크로드를 수행하는 것보다, 낮은 사양의 GPU와 혼합할 경우 전체적인 비용이 더 낮아진다는 연구 결과를 제시한 바 있다 (그림 5).

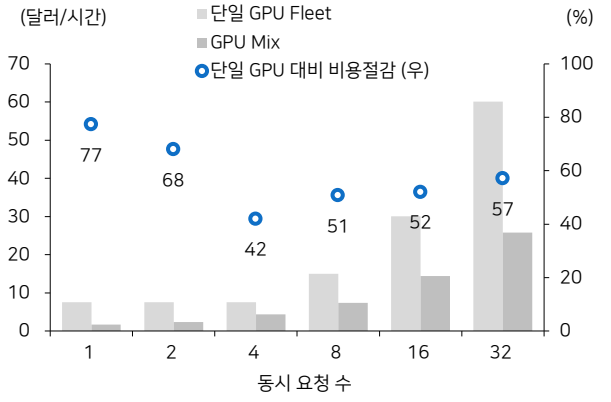
B200 프리미엄의 축소는 최신 GPU 수요 약화가 아닌, 워크로드별 비용 최적화를 대변한다고 보는 것이 더 적절하다. B200의 스팟 공급을 모두 흡수할 만큼 고성능 GPU의 수요가 넓지 않더라도, H100을 포함한 기존 세대 GPU의 반복적 수요가 유지될 수 있는 이유이다.

그림4 6월 이후 H100 대비 B200 프리미엄의 축소는



자료: Silicon Data, 메리츠증권 리서치센터

그림5 GPU 최적 배치 과정에서 발생했을 가능성



자료: UC Berkeley (2026), 메리츠증권 리서치센터

결론: GPU Rental Index 하락 ≠ GPU 수요 둔화

정리하자면 최근 스팟 가격의 조정은 GPU의 희소성 완화라기 보다는, 유휴용량의 가격이 정상화되는 과정에 가까워 보인다. 반면 장기계약 가격은 여전히 공급 제약을 가르키고 있다. 즉 스팟 가격이 약해져도, 배치 가능한 장기 용량의 희소성은 유지되고 있는 상황으로 해석할 수 있다. 게다가 GPU 세대 간 프리미엄 변화는 AI 수요가 실험단계를 넘어 반복적 운영으로 이동 중임을 시사한다.

종합했을 때, GPU 쇼티지가 일시적인 피크가 아닌, 더 지속적인 scarcity premium을 유도하는 요인으로 자리잡을 수 있는 환경이 조성될 수 있다는 판단이다. 물론 AI 투자 효용이 더 빈번하게 검증대에 올라섬에 따라 노이즈가 지속 발생할 수 있다. 다만 장기계약과 스팟계약 가격이 동시에 하락하지 않는 이상, scarcity-premium 논리를 훼손할 가능성은 제한적이라는 생각이다.

원문: *H100 Index: Same Chip, Two Tiers, Two Regimes (Silicon Data)*