

반도체

TurboQuant, DeepSeek 이후를 생각하면 결론은 명확

TurboQuant가 쏘아올린 큰 공

현지시각 3월 25일 구글은 성능의 저하 없이 KV Cache를 최대 6배까지 압축할 수 있다는 TurboQuant 알고리즘을 공식 블로그를 통해 공개했다. 시장은 이를 메모리 필요량의 감소로 해석해, 마이크론, 샌디스크는 물론 삼성전자, SK하이닉스를 포함한 메모리 반도체 기업 주가가 크게 하락했다. 그러나 이는 메모리 용량(Capacity)과 메모리 대역폭(Bandwidth)의 역할을 혼동한 데서 발생한 해석 오류이다. AI 추론의 병목은 메모리 용량 부족이 아니라, 메모리 접근 속도와 데이터 이동 효율에 의해 결정되는 문제이기 때문이다. TurboQuant는 이러한 병목을 일부 완화해 GPU 효율을 높임으로써, 동일한 GPU 자원으로 더 많은 토큰을 처리할 수 있게 해주는 기술로 이해해야 한다.

AI 추론의 구조: Prefill과 Decode

LLM 추론은 Prefill과 Decode, 두 단계로 구성된다. Prefill 단계는 GPU의 계산 능력이 성능을 제한하는 연산 중심 작업인 반면, Decode 단계는 데이터 이동의 속도가 성능을 좌우하는 메모리 집약적 작업이다. Decode에서는 새 토큰을 생성할 때마다 기존 KV cache를 반복적으로 참조해야 하므로, 구조적으로 메모리 대역폭(Bandwidth)과 접근 지연의 영향을 크게 받는다. 사용자가 체감하는 응답의 속도는 대부분 Decode 단계가 결정하므로, AI 추론 최적화의 핵심은 이 단계에 있다.

TurboQuant의 실제 작동 원리

TurboQuant가 KV Cache를 최대 6배 압축한다는 것은 필요 메모리 용량 자체를 줄인다는 의미라기보다, KV Cache가 차지하는 데이터 크기와 이에 따른 메모리 접근 부담을 크게 낮춘다는 의미에 가깝다. 이는 동일한 HBM 대역폭 내에서 처리해야 하는 데이터 양이 감소함을 의미하며, 결과적으로 메모리 접근 지연이 완화되고 GPU가 데이터를 기다리는 시간이 줄어들 수 있다. LLM 추론의 Decode 단계에서는 KV Cache 접근이 반복적으로 발생하기 때문에, 데이터 크기 축소는 메모리 병목 완화에 직접적인 영향을 미친다. 그 결과 GPU가 메모리 응답을 기다리는 비중이 줄어들고, 연산 자원이 보다 효율적으로 활용될 수 있다. 이 경우 GPU의 실제 가동률이 개선되며, 동일한 하드웨어 환경에서도 단위 시간당 처리 가능한 토큰 수(throughput)가 증가하는 효과가 나타난다. 이는 곧 토큰 당 비용이 낮아진다는 의미로도 해석할 수 있다.

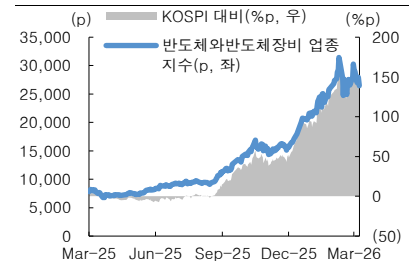
비중확대(유지)

종목	투자유언	목표주가
삼성전자	매수	270,000원
SK하이닉스	매수	1,500,000원

주요 종목 valuation

		2025A	2026F	2027F
SK하이닉스 (000660)	PER(x)	15.0	4.7	3.7
	PBR(x)	5.5	2.5	1.5
	EV/EBITDA(x)	10.7	3.3	2.3
	DY(%)	0.2%	0.2%	0.2%
삼성전자 (005930)	PER(x)	18.2	7.9	5.8
	PBR(x)	1.9	2.1	1.5
	EV/EBITDA(x)	7.6	4.0	2.9
	DY(%)	1.4%	0.9%	0.9%

12개월 업종 수익률



자료: FnGuide - WICS

채민숙

minsook@koreainvestment.com

김연준

fedkim@koreainvestment.com

AI 추론의 병목: 메모리 용량이 아닌 대역폭(Bandwidth) 문제

시장은 TurboQuant가 메모리 용량 사용을 줄여 HBM 수요를 감소시킬 수 있다는 방향으로 해석한 것으로 추정된다. 그러나 전술하였듯이, AI 추론에서의 핵심 병목은 메모리 용량 부족이 아니라, 메모리에서 데이터를 읽어 오는 속도, 즉 메모리 대역폭과 접근 지연(Latency)에 있다. GPU 연산 코어의 처리 속도는 HBM이 데이터를 공급하는 속도보다 훨씬 빠르기 때문에, 메모리로부터 데이터를 기다리는 동안 GPU는 대기 상태에 놓이게 된다. 업계 연구에 따르면 Decode 단계에서 Attention 연산 사이클의 50% 이상이 메모리 접근 지연으로 인한 대기 상태에 놓여있다. 즉, GPU가 이론 성능의 절반 이상을 메모리 응답을 기다리는 시간에 낭비하고 있는 것이다. Google DeepMind의 Xiaoyu Ma와 David Patterson이 2026년 1월 발표한 연구에 따르면 엔비디아 GPU의 64-bit FLOPS는 2012년~2022년 사이 약 80배 증가한 반면, 메모리 대역폭 증가는 약 17배에 그쳤으며, 이 격차는 향후에도 지속 확대될 것으로 전망된다. (arXiv:2601.05047, IEEE Computer 게재 예정) TurboQuant는 GPU 연산능력과 메모리 대역폭 사이의 Gap을 줄임으로써 동일한 하드웨어에서 단위 시간당 처리 가능한 토큰 수를 증가시킬 것이다. 이는 토큰 당 비용을 낮추어 AI 사용량 확대를 유도함으로써 더 많은 서비스와 사용자 유입을 촉진해 최종적으로 KV cache 사용량이 오히려 증가하는 결과로 이어질 것이다.

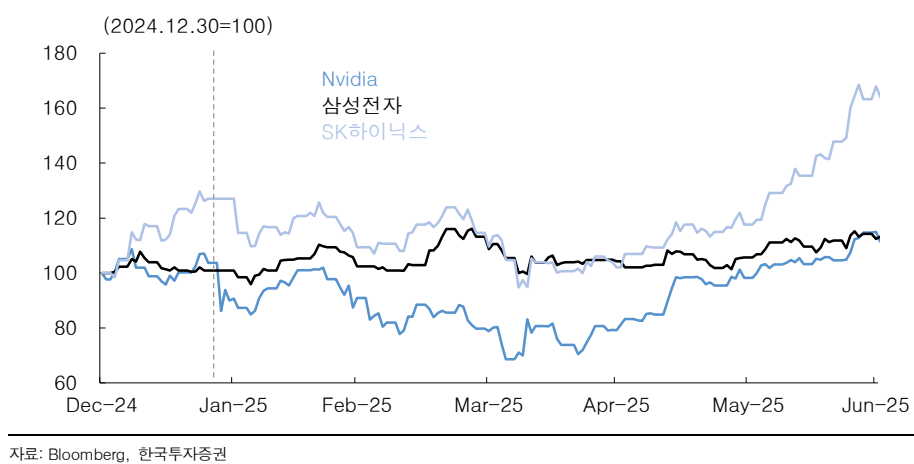
TurboQuant가 해결하지 못하는 병목: 칩 간 통신 Latency

AI 추론에는 TurboQuant가 전혀 해결하지 못하는 또 다른 병목이 있다. 대형 모델은 크기가 커서 단일 GPU 메모리 용량을 초과하기 때문에 여러 GPU에 분산되어 탑재된다. Decode 단계에서는 모델이 여러 GPU에 분산되어 있어, 각 토큰의 생성 과정마다 중간 계산 결과를 GPU 간에 교환해야 한다. 이 데이터들은 크기는 작지만 호출 빈도가 매우 높기 때문에 이로 인해 칩 간 통신 지연이 발생할 수 있다. 지연을 줄이기 위해서는 단일 GPU가 더 많은 파라미터와 KV cache를 처리할 수 있도록 해야 한다. 이는 결과적으로 GPU당 더 많은 HBM 용량을 요구한다. 모델 크기와 컨텍스트 길이가 지속적으로 확대되는 최근의 AI 환경에서는 이러한 HBM 용량 증가의 필요성이 더욱 강화될 것이다. 엔비디아가 세대별 GPU의 HBM 용량을 지속 확대해 온 것 역시 이와 무관하지 않을 것이다.

TurboQuant의 한계: 아직 초기 검증 단계

TurboQuant의 기술적 의미는 크지만, 적용 범위와 검증 수준에 한계가 존재한다는 점도 생각해 볼 문제다. 공개된 성능 검증은 LongBench, Needle-In-A-Haystack 등과 같은 단일 질문에 대한 정보 검색 중심의 비교적 단순한 테스트 환경에 국한되어 있다. 또한 실험 대상 역시 파라미터 규모 80억 수준의 비교적 작은 모델 위주로 진행돼, 실제 산업에서 사용되는 수백억 파라미터 급 대형 모델에서 동일한 효과가 재현될지 여부는 아직 검증되지 않았다. 더 중요한 점은, 최근 빠르게 확산되고 있는 Agentic AI 환경에서는 아직 검증이 이루어지지 않았다는 것이다. 이러한 환경에서는 모델이 여러 단계를 거쳐 반복적으로 판단을 수행하고, 더 긴 컨텍스트와 복잡한 KV Cache 구조를 사용하기 때문에, 단일 응답 기반 벤치마크와는 전혀 다른 메모리 사용 패턴이 나타날 수 있다.

[그림 1] Nvidia, 삼성전자, SK하이닉스 DeepSeek 발표 전후 주가 변동



<표 1> 커버리지 valuation

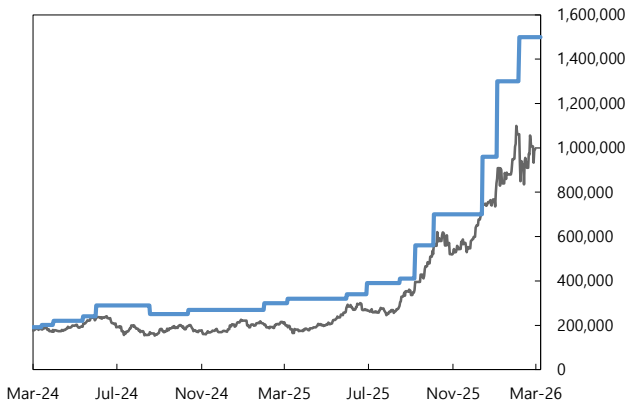
투자의견 및 목표주가				실적 및 Valuation									
종목				매출액 (십억원)	영업이익 (십억원)	순이익 (십억원)	EPS (원)	BPS (원)	PER (배)	PBR (배)	ROE (%)	EV/EBITDA (배)	DY (%)
삼성전자 (005930)	투자의견	매수	2024A	300,871	32,726	33,621	4,950	57,981	10.7	0.92	9.0	3.6	2.7
	목표주가(원)	270,000	2025A	333,606	43,601	44,261	6,605	63,997	18.2	1.87	10.8	7.6	1.4
	현재가 (3/26, 원)	180,100	2026F	515,821	202,112	151,943	22,706	86,446	7.9	2.08	30.5	4.0	0.9
	시가총액(십억원)	1,066,126	2027F	600,350	276,654	206,951	30,926	117,000	5.8	1.54	30.8	2.9	0.9
			2028F	660,385	293,571	221,103	33,040	149,696	5.5	1.20	25.1	2.3	0.9
SK하이닉스 (000660)	투자의견	매수	2023A	32,766	(7,730)	(9,112)	(13,244)	77,752	NM	1.82	(15.6)	21.1	0.8
	목표주가(원)	1,500,000	2024A	66,193	23,467	19,789	28,732	107,256	6.1	1.62	31.1	3.8	1.3
	현재가 (3/26, 원)	933,000	2025F	97,147	47,206	42,919	62,180	169,495	15.0	5.50	45.0	10.7	0.2
	시가총액(십억원)	664,951	2026F	242,499	168,378	136,047	197,088	366,604	4.7	2.54	73.5	3.3	0.2
			2027F	311,400	220,454	172,620	249,844	616,752	3.7	1.51	50.8	2.3	0.2

자료: 각 사, 한국투자증권

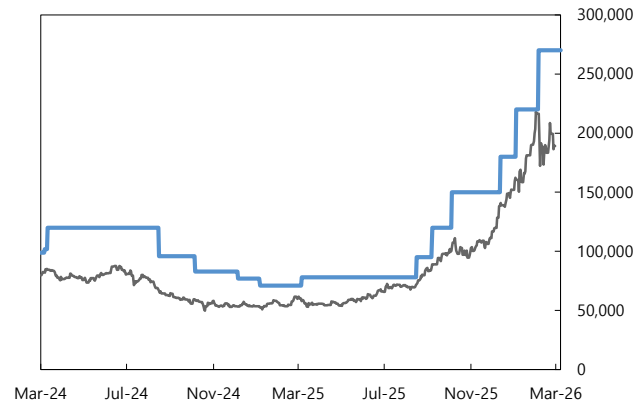
투자의견 및 목표주가 변경내역

종목(코드번호)	제시일자	투자의견	목표주가	과리율		종목(코드번호)	제시일자	투자의견	목표주가	과리율	
				평균주가 대비	최고(최저) 주가 대비					평균주가 대비	최고(최저) 주가 대비
SK하이닉스 (000660)	2024.02.26	매수	192,000원	-11.6	-2.1	삼성전자 (005930)	2026.03.02	매수	1,500,000원	-	-
	2024.04.08	매수	202,000원	-10.8	-6.7		2024.01.04	매수	99,000원	-24.8	-16.8
	2024.04.25	매수	220,000원	-14.5	-8.0		2024.04.01	매수	102,000원	-16.9	-16.4
	2024.06.07	매수	240,000원	-6.6	-1.0		2024.04.05	매수	120,000원	-34.2	-26.8
	2024.06.26	매수	290,000원	-32.5	-16.9		2024.09.10	매수	96,000원	-36.6	-30.9
	2024.09.12	매수	250,000원	-26.8	-19.6		2024.10.31	매수	83,000원	-33.8	-29.3
	2024.11.07	매수	270,000원	-30.2	-16.5		2024.12.31	매수	77,000원	-29.3	-25.6
	2025.02.25	매수	300,000원	-33.1	-28.2		2025.01.31	매수	71,000원	-20.5	-13.0
	2025.03.31	매수	320,000원	-36.7	-13.0		2025.03.31	매수	78,000원	-20.7	-6.9
	2025.06.25	매수	340,000원	-16.9	-11.8		2025.09.10	매수	95,000원	-14.4	-9.4
	2025.07.24	매수	390,000원	-32.2	-26.2		2025.10.02	매수	120,000원	-19.0	-15.0
	2025.09.10	매수	410,000원	-15.9	-12.0		2025.10.30	매수	150,000원	-28.7	-7.4
	2025.10.02	매수	560,000원	-16.4	-4.5		2026.01.07	매수	180,000원	-18.2	-9.8
	2025.10.29	매수	700,000원	-16.8	6.0		2026.01.29	매수	220,000원	-18.3	-0.9
	2026.01.08	매수	960,000원	-21.0	-12.4		2026.03.02	매수	270,000원	-	-
	2026.01.29	매수	1,300,000원	-29.1	-15.5						

SK하이닉스(000660)



삼성전자(005930)



■ Compliance notice

- 당사는 자료 공표일 현재 SK하이닉스, 삼성전자 종목의 발행주식을 1%이상 보유하고 있지 않습니다.
- 당사는 동 리포트의 내용 일부를 기관투자가 또는 제3자에게 사전에 제공한 사실이 없습니다.
- 동 리포트의 금융투자분석사와 배우자는 상기 발행주식을 보유하고 있지 않습니다.
- 당사는 SK하이닉스, 삼성전자 발행주식을 기초자산으로 하는 ELW(주식워런트증권)를 발행 중이며, 당해 ELW에 대한 유동성공급자(LP)입니다.

■ 기업 투자 의견은 향후 12개월간 시장 지수 대비 주가등락 기준임

- 매 수 : 시장 지수 대비 15%p 이상의 주가 상승 예상
- 중 립 : 시장 지수 대비 -15~15%p의 주가 등락 예상
- 비중축소 : 시장 지수 대비 15%p 이상의 주가 하락 예상
- 중립 및 비중축소 의견은 목표가 미제시

■ 투자등급 비율 (2025. 12. 31 기준)

매수	중립	비중축소(매도)
81%	19%	0%

※ 최근 1년간 공표한 유니버스 종목 기준

■ 업종 투자 의견은 향후 12개월간 해당 업종의 유가증권시장(코스닥) 시가총액 비중 대비 포트폴리오 구성 비중에 대한 의견임

- 비중확대 : 해당업종의 포트폴리오 구성비중을 유가증권시장(코스닥)시가총액 비중보다 높게 가져갈 것을 권함
- 중 립 : 해당업종의 포트폴리오 구성비중을 유가증권시장(코스닥)시가총액 비중과 같게 가져갈 것을 권함
- 비중축소 : 해당업종의 포트폴리오 구성비중을 유가증권시장(코스닥)시가총액 비중보다 낮게 가져갈 것을 권함

■ 본 리포트는 고객의 증권투자를 돕기 위하여 작성된 당사의 저작물로서 모든 저작권은 당사에게 있으며, 당사의 동의 없이 어떤 형태로든 복제, 배포, 전송, 변형할 수 없습니다.

■ 본 리포트는 당사 리서치본부에서 수집한 자료 및 정보를 기초로 작성된 것이나 당사가 그 자료 및 정보의 정확성이나 완전성을 보장할 수는 없으므로 당사는 본 리포트로써 고객의 투자 결과에 대한 어떠한 보장도 행하는 것이 아닙니다. 최종적 투자 결정은 고객의 판단에 기초한 것이며 본 리포트는 투자 결과와 관련한 법적 분쟁에서 증거로 사용될 수 없습니다.

■ 본 리포트에 제시된 종목들은 리서치본부에서 수집한 자료 및 정보 또는 계량화된 모델을 기초로 작성된 것이나, 당사의 공식적인 의견과는 다를 수 있습니다.

■ 이 리포트에 게재된 내용들은 작성자의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 작성되었음을 확인합니다.