반도체

Overweight

Top Picks 및 관심종목

*CP 2026년 3월 26일

삼성전자(005930)
BUY | TP 300,000원
CP 180,100원

SK하이닉스(000660)
BUY | TP 1,450,000원
CP 933,000원

Google의 Turbo Quant와 AI생태계 및 메모리 시사점

Turbo Quant: 추론 병목현상 줄이기 위한 압축 기술

Google이 최근 공개한 Turbo Quant는 LLM 추론 과정에서 가장 큰 병목 현상 중 하나인 KV Cache 사용량을 획기적으로 줄이는 양자화(Quantization) 기술이다. 기존 모델(예: FP16, BF16)은 KV Cache를 저장할 때 토큰당 16비트를 사용하지만, Turbo Quant는 이를 4비트 이하로 압축하면서도 성능 저하를 최소화한다. KV Cache를 압축하기 위한 기술은 Nvidia의 NVFP4, Microsoft의 DeepSpeed 등도 이미 존재했다. 여타 압축기술과의 차별화 포인트는 압축 과정에서 발생할 수밖에 없는 정밀도 하향 및 오류 발생을 최소화했다는 점이다. 아울러 이를 오픈소스 형태로 공개하면서 여타 업체들도 활용할 수 있게 했다. 일례로 Google은 자사 모델인 Gemma뿐만 아니라 경쟁사 모델인 Mistral에도 이 기술을 적용해 성능을 증명했다. 오픈소스로 제공하는 이유는 하드웨어 제조사(Nvidia 등)가 자사 GPU 드라이버나 라이브러리에 Turbo Quant 연산을 가속하는 기능을 기본적으로 탑재할 가능성을 열어두었기 때문으로 보인다.

하이퍼스케일러 등의 고객들은 비용 효율화 가능

해당 기술은 추론 시스템에서 요구되는 메모리를 효율화시킬 수 있다. LLM 서비스 운영 시 GPU 메모리(HBM)는 크게 두 영역이 차지한다. 모델 가중치 (Weights: 모델 자체의 크기(고정값))와 KV Cache(사용자의 질문과 답변이 길어질수록 늘어나는 영역(가변값))이다. 기존 FP16 방식에서는 문맥이 길어질수록 KV Cache가 모델 가중치보다 더 많은 메모리를 점유하기도하면서, 이를 일반 메모리 및 스토리지로 오프로드했다. 이론적으로 이를 4비트로 줄이면, 동일한 메모리 용량에서 수용 가능한 컨텍스트 길이가 4배 늘어나거나, 동시 접속자(Batch Size)를 4배 더 받을 수 있게 된다. 이는 데이터 센터의 전력 소모 감소와 서버 구축 비용(CAPEX) 절감으로 직결될 수 있다. 이는 하이퍼스케일러들이 인프라 투자 효율(ROI)을 높이는 데 핵심적인 역할을 할 수 있게 된다.

메모리 업체 입장에서 진입장벽 낮아져 전체 수요 증대 가능성

시뮬레이션 결과가 실제 환경에서 어떻게 작용할지는 불확실하다. 해당 기술의 검증 및 확장 속도도 불명확하다. 만약에 압축 기술이 보편화되었을 때 일어날 변화를 HBM과 일반 DRAM, 스토리지로 구분해서 볼 필요가 있다.

Turbo Quant와 같은 기술은 압축된 데이터를 풀고(Dequantize) 연산하는 과정이 필요해 HBM 커스텀화가 가속화될 것으로 예상된다. 메모리 업체는 로직 다이에 가속 엔진(Hardware Accelerator)을 직접 심어, GPU로 데이터를 보내기 전에 메모리 단에서 압축/해제를 처리하는 'PIM(Processor-in-Memory)' 기술 고도화가 필요해질 것으로 예상된다. 아울러 용량보다는 대역폭, 즉 속도를 올리기 위한 경쟁이 중요해질 것으로 판단된다.

일반 DRAM의 경우, 워크로드되는 용량이 축소될 우려가 상존한다. 다만, 고객사 입장에서 당장 메모리 용량을 축소시키기 보다는 남는 공간에 4배 이상 많은 사용자 질문(Batch)을 집어넣을 수 있게 되기 때문에 운영 여력을 확보할 수 있게 된다. 또한, 데이터를 압축하고 푸는 과정에서 발생하는 연산을 빠르게 처리하려면, 일반 DDR5보다 대역폭이 높은

Analyst 김록호 roko.kim@hanafn.com
RA 김영규 kyg1019@hanafn.com
하나증권 리서치센터

MRDIMM이나 SOCAMM2 같은 고성능 규격의 수요가 보편화될 것으로 예상된다.

추론 워크로드에서 스토리지는 주로 모델의 가중치를 불러오거나 체크포인트를 저장하는 용도였다. 하지만 압축 기술은 스토리지를 확장된 메모리로 쓰게 만든다. 단순 저장용 저가 SSD보다는 읽기 속도가 극단적으로 빠른 NVMe Gen5/Gen6 SSD의 수요가 중요해진다. 즉, '단순 용량'보다는 '메모리처럼 작동하는 스토리지'로서의 비중이 커질 것으로 예상된다. HBF 상용화를 앞당길 가능성도 상존한다.

결론적으로 압축 기술은 이미 기존에도 존재했던 기술이고, 해당 기술을 모든 업체가 사용하거나 보편화할 가능성 역시 미지수다. 만약 해당 기술이 중장기적으로 보편화되어 비트당 수요가 효율화되면, 메모리 비용에 대한 진입장벽이 낮아지기 때문에 AI 서비스의 전체 사용량 및 DRAM과 스토리지의 절대적인 수요량을 감소시킬 가능성은 제한적이라 판단한다.

도표 1. 추론 압축 기술의 이해

기존 양자화 (KIVI 등)	TurboQuant
1) 블록 단위 분할 벡터를 소규모 블록(32~128개)으로 분할	**1) 랜덤 회전 + 극좌표 변환 (PolarQuant)** 데이터 기하 단순화 → 정규화/상수 저장 불필요
▼	▼
2) 블록별 양자화 (4비트) 각 블록에 스케일/제로포인트 계산 후 양자화	**2) 좌표별 스칼라 양자화 (b-1 비트)** 목표 비트폭에서 1을 뺀 만큼 MSE 최적 양자화
▼	▼
3) 양자화 상수 별도 저장 블록당 스케일, 제로포인트를 FP16으로 저장 → +1~2비트 메모리 오버헤드 발생	**3) QJL 잔차 보정 (1비트)** 잔여 오차를 1비트 부호 비트로 제거 → 무편향(unbiased) 내적 추정, 제로 오버헤드
▼	▼
실질 비트폭: **명목 + 오버헤드 1~2비트**	실질 비트폭: **명목 = 실질 (1~4비트, 오버헤드 0)**

자료: Google, TurboQuant: Online Vector Quantization with Near-optimal Distortion Rate, 하나증권

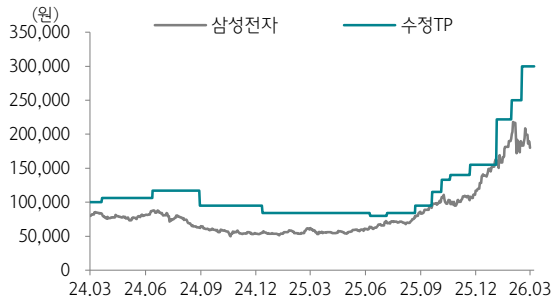
도표 2. 기존 기술들과의 비교

기술	개발사	비트 폭	압축률	정확도	특징
TurboQuant	Google	3비트	6배 이상	손실 0%	PolarQuant + QJL 2단계 압축, 데이터 비의존적, 오픈소스
KVTC	Nvidia	가변	최대 20배	1%p 미만	PCA 기반 변환 코딩, JPEG식 접근, Dynamo 프레임워크 통합
NVFP4	Nvidia	4비트	약 2배(FP8 대비)	1% 미만	Blackwell GPU 최적화, TensorRT 통합, HW-SW 공동 설계
KIVI	학계	2비트	약 2.6배	소폭 하락	비대칭 양자화 (Key: 채널별, Value: 토큰별), HuggingFace 통합
KVQuant	학계	2~4비트	최대 8배	양호	채널별 Key 양자화, A100 1장에 1M 토큰 처리 가능
DeepSpeed	Microsoft	4~8비트	2~4배	양호	분산 추론 프레임워크 내 양자화, 대규모 배포 최적화

자료: Google, Nvidia, 하나증권

투자의견 변동 내역 및 목표주가 괴리율

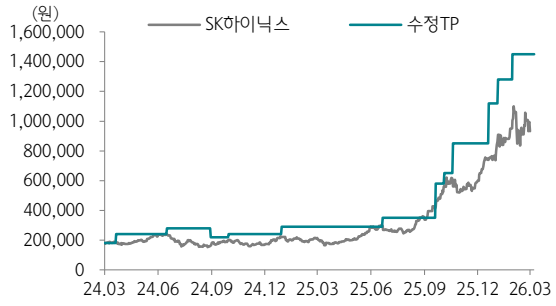
삼성전자



날짜	투자의견	목표주가	괴리율	
			평균	최고/최저
26.3.13	BUY	300,000		
26.2.24	BUY	250,000	-22.75%	-12.80%
26.1.30	BUY	222,000	-22.46%	-13.06%
25.12.17	BUY	155,000	-11.97%	4.77%
25.11.14	BUY	140,000	-26.46%	-21.79%
25.10.31	BUY	133,000	-22.83%	-16.47%
25.10.15	BUY	115,000	-13.74%	-9.48%
25.9.17	BUY	95,000	-9.35%	-0.63%
25.8.1	BUY	84,000	-15.30%	-5.48%
25.7.4	BUY	80,000	-17.84%	-9.25%

투자의견 변동 내역 및 목표주가 괴리율

SK하이닉스



날짜	투자의견	목표주가	괴리율	
			평균	최고/최저
26.2.24	BUY	1,450,000		
26.1.30	BUY	1,280,000	-30.87%	-25.70%
26.1.14	BUY	1,120,000	-30.91%	-23.13%
25.11.14	BUY	850,000	-29.80%	-11.06%
25.10.30	BUY	650,000	-8.14%	-4.62%
25.10.15	BUY	580,000	-14.37%	-3.79%
25.7.16	BUY	350,000	-15.26%	22.29%
25.1.24	BUY	290,000	-26.37%	3.45%
24.10.25	BUY	240,000	-23.33%	-6.04%
24.9.24	BUY	220,000	-16.06%	-9.91%

Compliance Notice

- 당사는 2026년 3월 27일 현재 해당회사의 지분을 1%이상 보유 하고 있지 않습니다
- 본 자료를 작성한 애널리스트(김록호)는 자료의 작성과 관련하여 외부의 압력이나 부당한 간섭을 받지 않았으며, 본인의 의견을 정확하게 반영하여 신의성실 하게 작성하였습니다.
- 본 자료는 기관투자가 등 제 3자에게 사전 제공한 사실이 없습니다
- 본자료를 작성한 애널리스트(김록호)는 2026년 3월 27일 현재 해당회사의 유가증권을 보유하고 있지 않습니다

본 조사자료는 고객의 투자에 정보를 제공할 목적으로 작성되었으며, 어떠한 경우에도 무단 복제 및 배포 될 수 없습니다. 또한 본 자료에 수록된 내용은 당사가 신뢰할 만한 자료 및 정보로 얻어진 것이나, 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 최종결정을 하시기 바랍니다. 따라서 어떠한 경우에도 본 자료는 고객의 주식투자의 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다.

투자등급 관련사항 및 투자의견 비율공시

- **투자의견의 유효기간은 추천일 이후 12개월을 기준으로 적용**
- **기업의 분류**
  BUY(매수)_목표주가가 현주가 대비 15% 이상 상승 여력
  Neutral(중립)_목표주가가 현주가 대비 -15%~15% 등락
  Reduce(비중축소)_목표주가가 현주가 대비 15% 이상 하락 가능
- **산업의 분류**
  Overweight(비중확대)_업종지수가 현재지수 대비 15% 이상 상승 여력
  Neutral(중립)_업종지수가 현재지수 대비 -15%~15% 등락
  Underweight(비중축소)_업종지수가 현재지수 대비 -15%~15% 등락

투자등급	BUY(매수)	Neutral(중립)	Reduce(매도)	합계
금융투자상품의 비율	96.76%	3.24%	0.00%	100%

* 기준일: 2026년 03월 24일