iM증권

Issue Comment

Overweight

2026.03.27

반도체 (Overweight)

Google TurboQuant의 영향

HBM 용량 증가율 하락 Vs NAND 수요 증가

TurboQuant에 따르면 기존 LLM 모델에서 HBM에 16비트 단위로 저장하던 데이터를 3비트로 압축해 AI가 대화 내용을 기억하는 단기 기억 저장소 (KV Cache)의 용량을 6배 이상 축소하는 것이 가능하다. 데이터를 압축하면 문맥에 대한 정확도가 떨어지나, 터보퀀트는 수학적 설계를 통해 정확도 저하 없이 압축에 성공했다.

두 가지 핵심 기술이 존재한다. ① PolarQuant는 데이터 벡터를 회전시켜 기하학적 구조를 단순화하고 이를 통해 복잡한 데이터를 예측 가능한 형태로 바꿔 효율적으로 압축한다. ② QJL (Quantized Johnson-Lindenstrauss)은 첫 단계에서 발생하는 미세한 오차를 단 1비트만 사용하여 보정한다. 이 과정이 수학적 에러 체크 역할을 하여 정확도를 완벽하게 보존하게 된다.

터보퀀트는 최근 AI 업계에서 뚜렷하게 나타나고 있는 HBM에 대한 의존도를 줄이고 SRAM, NAND를 더 많이 사용하려는 흐름을 가속화시키는 기술인 것으로 판단된다. 이미 Google은 터보퀀트를 제미나이 3.0에 일부 적용 중인 것으로 밝혔으며, 따라서 향후 현실화 가능성도 높은 것으로 보인다.

HBM의 현재 문제는 LLM이 긴 문맥을 처리할수록 KV Cache가 선형적으로 늘어나 HBM 용량을 상당량 차지하므로 한 번에 처리할 수 있는 사용자 수가 제한된다는 점이다. 터보퀀트를 이용하면, HBM에 저장되는 KV Cache의 점유율이 1/6 이하로 감소하므로 같은 HBM 용량으로 6배 더 많은 사용자를 동시 처리하거나, 6배 더 긴 문장을 입력할 수 있게 된다. 이는 HBM의 용량이 축소되거나 증가율이 크게 둔화될 수 있음을 의미한다.

GPU와 LPU 칩 내부에 위치한 SRAM은 용량이 극히 적어 데이터를 SRAM에 다 올리지 못하고, 매번 HBM에서 데이터를 불러와야 한다. 따라서 Memory 병목 현상이 발생해 계산 속도가 느려진다. 반면, 터보퀀트 활용 시 압축된 데이터의 크기가 작아 더 많은 양이 SRAM에 머무를 수 있고 HBM으로 데이터를 가지러 가는 횟수가 줄어들어 추론 속도가 최대 8배까지 빨라지고 전력 소모도 감소한다.

KV Cache를 NAND에 두면 추론할 때마다 데이터를 가져오는 속도가 너무 느려 AI의 답변 속도가 크게 저하된다. 터보퀀트를 활용하면 NAND에서 가져와야 할 데이터 양 자체가 1/6로 감소하므로 실질적으로는 6배 더 빠른 데이터 전송 효과를 내며 NAND 오프로딩 시 발생하는 속도 저하 문제를 상쇄할 수 있다. 따라서 굳이 고가의 HBM이 아니더라도 SSD에서 KV Cache를 관리할 수 있는 알고리즘 (ICMS) 설계가 가능하다.

터보퀀트의 데이터 6배 압축과 NAND의 저렴한 대용량이 만나면, 개인용 워크스테이션에서도 수백만 토큰의 대화를 기억하는 AI를 돌릴 수 있게 된다. 다시 말해, 메모리 가격 때문에 도입이 지연되어 왔던 온디바이스 AI의 상용화 정도가 강화될 수 있다. 한편, 빅테크 업체 입장에서는 고가의 GPU 서버를 늘리는 대신, 상대적으로 저렴한 NAND 기반 스토리지 서버로 고객의 문맥을 장기 보관할 수 있다.

터보퀀트의 사용이 단기적으로는 HBM의 용량 축소 또는 증가율 하락을 가져와 Memory 반도체 업계에 불리하게 작용할 수 있다. 단 장기적으로는 반대로 NAND 수요가 크게 증가하고 AI 사용량의 증가를 불러와 전반적인 Memory 반도체 수요에 긍정적인 영향을 미칠 가능성이 높은 것으로 판단된다.

[반도체]
송명섭
2122-9207
mssong@imfnsec.com

[IT RA]
박정하
2122-9195
jhpark@imfnsec.com

Compliance notice

당 보고서 공표일 기준으로 해당 기업과 관련하여,
- 회사는 해당 종목을 1%이상 보유하고 있지 않습니다.
- 금융투자분석사와 그 배우자는 해당 기업의 주식을 보유하고 있지 않습니다.
- 당 보고서는 기관투자가 및 제 3자에게 E-mail등을 통하여 사전에 배포된 사실이 없습니다.
- 회사는 6개월간 해당 기업의 유가증권 발행과 관련 주관사로 참여하지 않았습니다.
- 당 보고서에 게재된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 작성되었음을 확인합니다.

본 분석자료는 투자자의 증권투자를 돕기 위한 참고자료이며, 따라서, 본 자료에 의한 투자자의 투자결과에 대해 어떠한 목적의 증빙자료로도 사용될 수 없으며, 어떠한 경우에도 작성자 및 당사의 허가 없이 전재, 복사 또는 대여될 수 없습니다. 무단전재 등으로 인한 분쟁발생시 법적 책임이 있음을 주지하시기 바랍니다.

[투자의견]

종목추천 투자등급
종목투자의견은 향후 12개월간 추천일 종가대비 해당종목의 예상 목표수익률을 의미함.
- Buy(매수): 추천일 종가대비 +15%이상
- Hold(보유): 추천일 종가대비 -15% ~ 15% 내외 등락
- Sell(매도): 추천일 종가대비 -15%이상

산업추천 투자등급
시가총액기준 산업별 시장비중대비 보유비중의 변화를 추천하는 것임
- Overweight(비중확대)
- Neutral (중립)
- Underweight (비중축소)

[투자등급 비율 : 2025-12-31 기준)

매수	중립(보유)	매도
90.6%	9.4%	-