경제

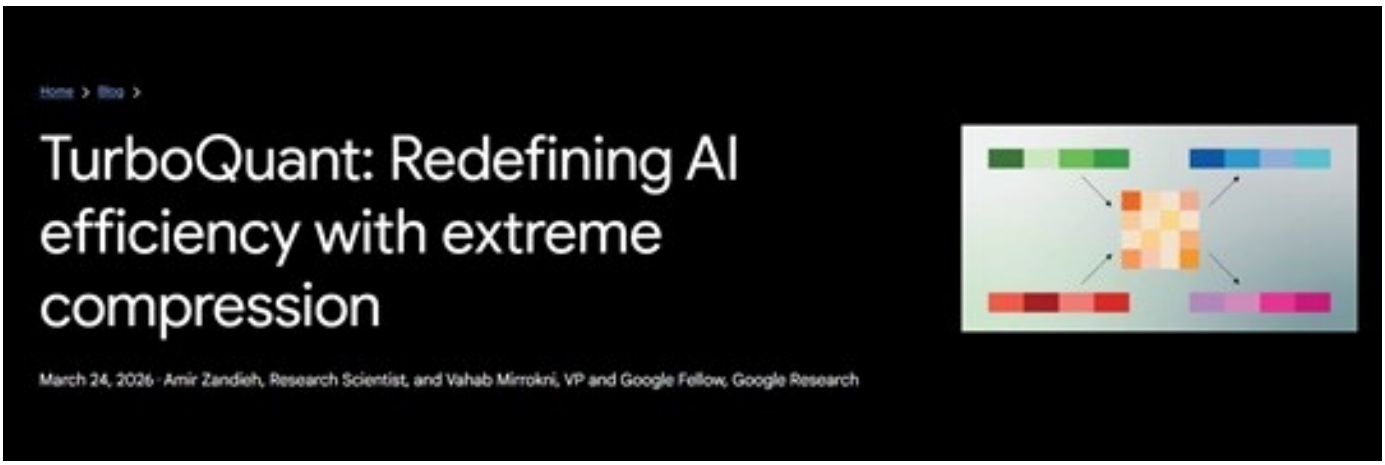
'게임 체인저'일까 '기우'일까…반도체 업계 뒤흔든 구글 '터보퀀트' 정체

[제1769호] | 26.04.03 17:06:51

KV캐시 압축해도 성능 유지, 범용성도 갖춰…HBM 대신 SRAM 등 대체 늘 수 있지만 전체 수요 확대 전망

[일요신문] 구글이 공개한 '터보퀀트(TurboQuant)'가 반도체 업계를 뒤흔들고 있다. 터보퀀트는 대규모 언어모델(LLM)의 임시 기억 장치인 '키밸류(KV) 캐시'를 성능 저하 없이 압축한 기술이다. 시장에선 소프트웨어 최적화로 메모리 반도체 의존도가 낮아져 수요가 줄어들 수 있다는 우려가 제기됐다. 하지만 글로벌 빅테크(대형 기술기업)들이 인공지능(AI) 모델 성능 경쟁에 뛰어든 상황에서, 오히려 메모리 수요가 늘어날 것이라는 전망이 지배적이다.

#KV캐시 압축했는데 성능은 유지



구글이 공개한 '터보퀀트(TurboQuant)'가 반도체 업계를 뒤흔들고 있다. 3월 24일(현지시각) 구글리서치 공식 블로그에 올라온 터보퀀트 소개 글. 사진=구글 리서치 블로그 캡처

"극강의 압축으로 AI 효율을 재정의한다." 지난 3월 24일(현지시각) 구글이 구글리서치 블로그를 통해 터보퀀트 논문을 소개했다. 구글은 KV캐시를 압축한 터보퀀트로 메모리 사용량을 기존의 6분의 1로 줄일 수 있다고 설명했다. 파장은 컸다. 사이버보안기업 클라우드플레어의 매슈 프린스 최고경영자(CEO)는 "구글의 딥시크"라는 평가를 내놨다. 2025년 1월 중국 딥시크는 챗GPT 20분의 1 비용으로 학습시킨 가성비 AI 모델 'R1'을 선보여 시장에 충격을 줬다. 딥시크와 구글 모두 소프트웨어 최적화 기술을 활용했다.

터보퀀트가 주목받은 건 AI 업계가 현재 직면한 기술적인 한계를 해결할 수도 있다는 점 때문이다. AI 업

터보퀀트의 핵심 기술은 '2중 양자화'다. 양자화는 소수점 단위의 데이터를 단순한 정수 형태로 바꾸는 것이다. 수백억 개의 파라미터로 구성되는 AI 모델은 보통 16비트의 소수점 형태다. 터보퀀트는 AI가 다루는 데이터를 직교좌표계(X·Y·Z)에서 크기와 방향 중심의 극좌표계로 바꾸는 '폴라퀀트' 기술을 활용해 1차 양자화를 한다. '동쪽으로 3칸, 북쪽으로 4칸 가라'는 지시를 '37도 각도로 5칸 가라'고 다루기 쉽게 바꾸는 식이다. 그다음 'QJL(양자화 존슨-린덴스 트라우스 변환)' 기술로 압축 과정에서 생기는 오차를 세밀하게 보정한다.

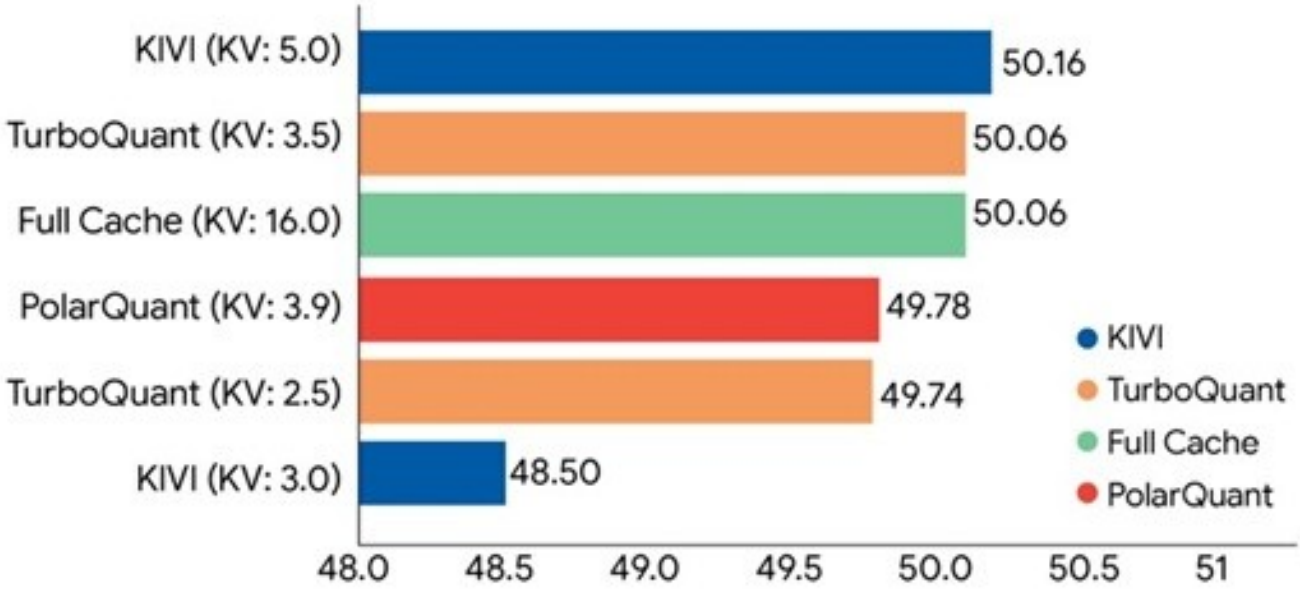
터보퀀트의 핵심 기술은 '2중 양자화'다. 미국 캘리포니아 구글 본사. 사진=AP/연합뉴스

구글이 KV캐시 압축을 시도한 유일한 기업은 아니다. 엔비디아는 2025년 11월 'KVTC'라는 기술을 공개했다. 사진과 영상을 압축하듯 KV캐시에서 중복되는 데이터를 압축하는 기술이다. 엔비디아는 KVTC를 통해 최대 20배까지 메모리를 절감할 수 있다고 밝혔다. 유회준 카이스트 AI반도체대학원 교수는 "양자화를 포함해 AI 모델 경량화는 꾸준히 연구되고 있는 연구 주제 중 하나"라고 말했다. 그간 KV캐시를 압축하면 다시 정보를 꺼내 쓸 때 AI 모델 답변 성능이 떨어지는 문제가 있었는데, 이를 구글이 개선했다는 분석이다.

한종목 미래에셋증권 연구원은 리포트를 통해 "데이터를 아주 작은 블록으로 쪼개 억지로 구겨 넣는 게 기

3~4비트까지 압축해도 성능 붕괴 없이 버티는 것"이라고 밝혔다.

터보퀀트는 범용성도 갖췄다. 터보퀀트는 구글의 AI 반도체인 'TPU'에서만 쓸 수 있는 게 아니다. 터보퀀트는 엔비디아의 GPU인 'H100'에서 연구가 수행됐다. 실험 결과 H100에서 기존 대비 연산 속도가 8배 빨라졌다. 터보퀀트의 폴라퀀트와 QJL 기술 연구에 참여한 한인수 카이스트 전기 및 전자공학부 교수는 3월 30일 온라인 기자간담회에서 "(터보퀀트는) 대규모 벡터 기반 검색 최적화에 범용적으로 활용될 수 있다"고 말했다.



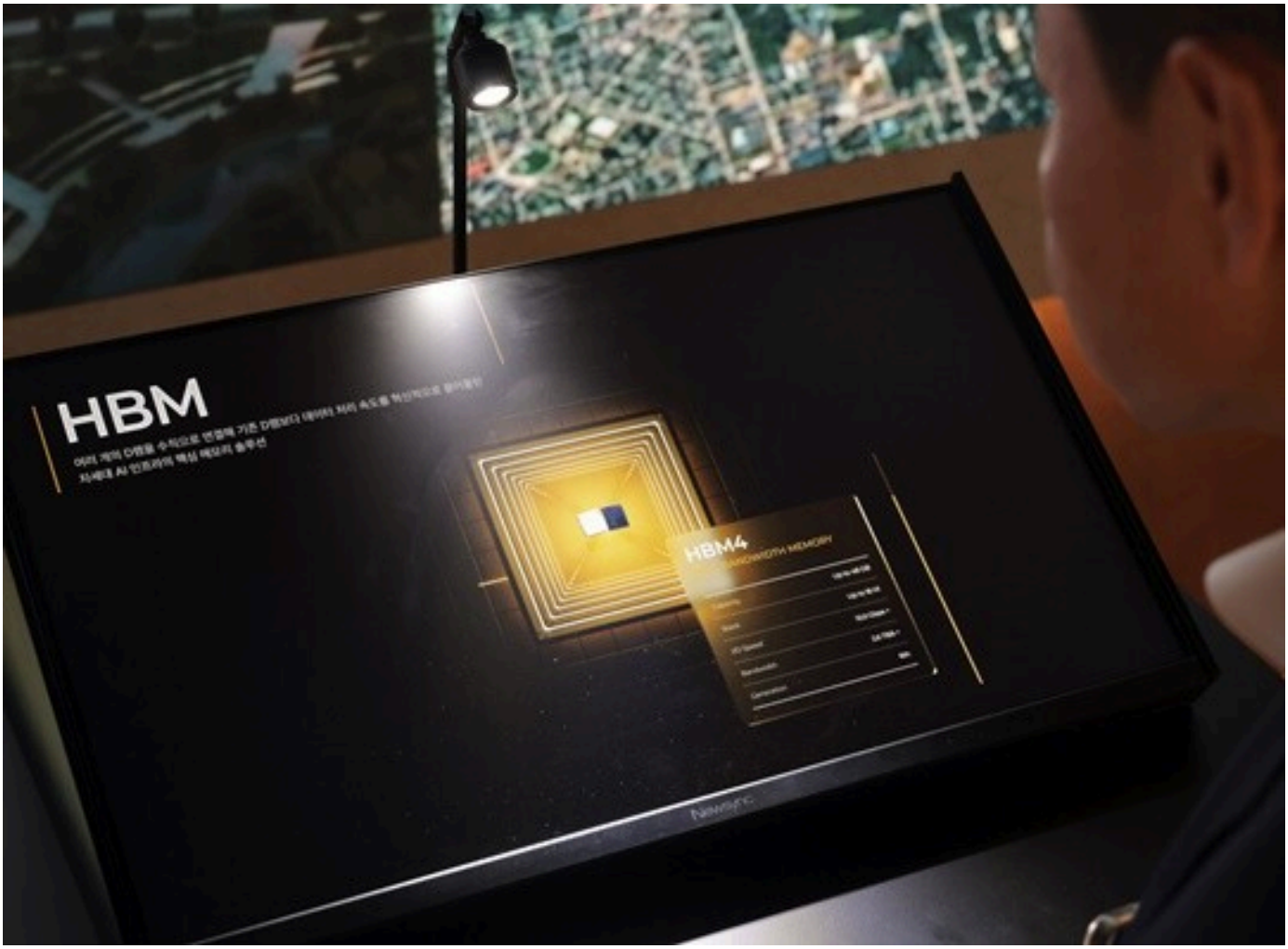
터보퀀트는 미국 대학 연구진이 2024년 개발한 KV캐시 양자화 압축 기술인 KIVI보다 더 낮은 비트수에서 비슷하거나 더 나은 성능을 냈다. 사진=구글리서치 블로그 캡처

#빅테크 성능 경쟁…메모리 수요 긍정적 전망도

시장에선 터보퀀트가 상용화되면 메모리 의존도가 낮아져 메모리 수요가 꺾이는 것 아니냐는 우려가 제기됐다. 최근 메모리 가격이 급등하면서 빅테크의 원가 부담이 커진 상황이다. 유재희 홍익대 전자전기공학부 교수(반도체공학회 부회장)는 "AI 산업이 진화하면서 요구되는 메모리양이 예상과 달리 급속히 늘어났다"며 "구글이 AI 산업에서 메모리 생산업체의 중요도를 줄이고 싶은 의도는 분명해 보인다"라고 분석했다.

터보퀀트가 단기적으로 HBM 수요에 영향을 줄 수 있다는 전망도 나온다. 비싼 HBM 대신 SRAM(정적램)이나 NAND(낸드)를 더 많이 쓰려는 움직임을 가속화할 수 있다는 이유에서다. 송명섭 iM증권 연구원은 리포트를 통해 "GPU(그래픽처리장치)나 LPU(언어처리장치) 칩 내부에 있는 SRAM은 용량이 극히 적어 데이터를 다 올리지 못하고 매번 HBM에서 데이터를 불러와야 한다. 터보퀀트를 활용하면 압축된 데이터

추론 단계에서 쓰이는 메모리인데, 현재 메모리 수요는 AI 학습 단계에서 대부분 발생한다. 이종환 상명대 시스템반도체공학과 교수는 "빅테크들은 AI 모델의 성능을 두고 경쟁 중"이라며 "AI 모델의 효율화 작업이 이뤄져도 대규모 학습을 통해 성능을 높이려는 욕구는 이어질 것"이라고 전망했다.



터보퀀트가 총 메모리 수요에는 큰 영향을 주지 못할 것이라는 의견이 전문가들의 중론이다. 2025년 10월 서울 강남구 코엑스에서 열린 제27회 반도체 대전(SEDEX 2025)을 찾은 관람객이 SK하이닉스 HBM4 실물을 살펴보고 있다. 사진=박정훈 기자

실제 최근 빅테크들은 메모리 업체들과 맺었던 기존 연간·분기 단위 계약을 3~5년 장기공급계약(LTA)으로 전환하기 위해 대규모 선수금까지 제시하고 있다. 강성철 한국반도체디스플레이학회 연구위원은 "최소 2027년까지는 메모리 수요 대비 공급이 부족할 것으로 본다"라고 "AI 중심축이 학습에서 추론으로 넘어가고 있고, 추론 분야에서 NPU(신경망공급장치) 등의 수요도 크기 때문에 메모리 수요를 염려할 단계는 아니다"라고 말했다.

터보퀀트로 인해 메모리 수요가 더 늘 수 있다는 전망도 있다. 19세기 경제학자 윌리엄 제본스의 이름을 딴 '제본스의 역설'이 거론되는 배경이다. 이는 기술 효율이 좋아지면 비용이 낮아져 오히려 수요가 증가한다는 경제 이론이다. 유재희 교수는 "영상 데이터를 다양한 방법으로 압축하고 있지만 해상도가 늘어나

서 "압축 기술은 이미 이전에도 존재했던 기술이다. 이 기술을 모든 업체가 사용하거나 보편화할 가능성 역시 미지수"라고 짚었다. 구글은 오는 4월 23~27일 브라질에서 열리는 '국제표현학습학회(ICLR) 2026'에서 터보퀀트 연구 성과를 정식 발표할 예정이다.

반도체 시장에 오히려 호재?

구글의 '터보퀀트' 발표 이후 반도체 대형주의 프리미엄에 흔들릴 수 있다는 시장의 평가가 나왔다. 그간 삼성전자와 SK하이닉스, 미국의 마이크론 같은 메모리 업체 주가 상승은 단순한 실적 개선이 아니라 '인공지능(AI) 인프라 확대로 고대역폭메모리(HBM)와 메모리 부족이 장기화할 것'이라는 기대가 동력이 됐기 때문이다. 기대에 비례해 우려의 폭도 컸다. 3월 26일 코스피가 3.22% 하락한 가운데 삼성전자는 4.71%, SK하이닉스는 6.23% 떨어졌다. '터보퀀트 쇼크'가 단순한 시장 조정 이상을 의미한다는 해석도 나왔다. 터보퀀트가 실제 상용 추론 환경에 안착하면 메모리 병목 현상이 생각보다 빨리 완화되고, 그동안 SK하이닉스와 삼성전자를 고평가했던 논리도 일부 약해질 수 있다는 우려가 나온 것이다.

구글의 '터보퀀트'가 알려지자 시장이 가장 먼저 흔든 것은 반도체 대형주의 프리미엄이었다. 3월 26일 오전 서울 중구 하나은행 딜링룸 현황판에 코스피 지수가 표시되고 있다. 사진=연합뉴스

다만 여기에는 과장된 해석도 섞여 있다는 지적이 나온다. 메모리 가격이 높고 AI 수요가 빠르게 늘어나는 상황에서 업계가 비용을 낮추고 효율을 높이기 위한 다양한 기술을 내놓는 것은 충분히 예상 가능한 범위였다는 것이다. 반도체업계 한 관계자는 "이런 시도가 계속 나오는 것이 자연스

준이 절대로 아니다. 앞으로 공장을 계속 지어도 감당하기 쉽지 않은 수준"이라며 "터보퀀트가 상용화된다 해도 어디까지나 효율을 높이는 기술일 뿐 기존의 수급 불균형을 단번에 해소할 게임체인저로 보기는 어렵다"고 말했다.

일론 머스크 테슬라 최고경영자(CEO) 역시 연간 1000억~2000억 개의 맞춤형 AI 및 메모리 반도체 생산을 목표로 지난 3월 21일 자체 반도체 생산 공장 '테라팹'을 출범하겠다고 밝힌 상황이다. 테라팹은 테슬라가 지난 1월 28일 실적 발표에서 처음 공식 확인한 프로젝트로 최소 추정 비용만 약 250억 달러에 달한다. 머스크는 당시 투자자들에게 3~4년 내 공급 부족이 현실화될 것으로 보고 이를 막으려면 자체 반도체 생산 시설이 반드시 필요하다고 설명했다. 실제로 뱅크오브아메리카(BofA)는 이번 매도세를 과도한 반응으로 평가하기도 했다.

최병호 고려대 휴먼인스파이어드 AI연구원 연구교수는 "효율성이 더 높아지면 기술은 다음 할 일을 찾아낸다"며 "가정용 로봇 같은 새로운 시장이 열리면 개별 기기마다 메모리반도체가 들어갈 수밖에 없다. AI 시장이 아직 충분히 개화하지도 않은 상황에서 효율화 기술만 보고 수요가 줄어든다고 단정하는 것은 무리"라고 설명했다.

과거 스토리지 시장도 비슷한 흐름을 보였다는 평가다. 저장장치 용량이 커지면 더 이상 추가 수요가 없을 것이라는 전망이 반복됐지만, 실제로는 더 큰 용량을 쓰게 만드는 애플리케이션이 등장하면서 시장이 오히려 확대됐다.

KB증권 리서치본부의 김동원 전무는 "터보퀀트의 등장은 반도체 업종에 중장기적으로는 오히려 호재에 가깝다. 최근 반도체주가 단기간에 급등한 상황에서 6개월 가까이 별다른 악재가 없었기 때문에 이번 이슈가 시장에 차익실현의 계기를 제공한 측면이 있다"고 덧붙였다.

김명선 기자 seon@ilyo.co.kr
김정민 기자 hurrymin@ilyo.co.kr