4 구글 '터보퀀트' 발표와 AI 메모리 패러다임 변화

구글 터보퀀트 발표, AI 추론 효율화의 새로운 접근

AI 서비스 확산으로 추론 단계의 KV 캐시·메모리 대역폭 병목 심화

- AI 서비스가 단순 질의응답에서 장문 처리·에이전트형 서비스 작업으로 확장되면서, 대형언어모델(LLM) 추론 시 KV 캐시에 쌓이는 데이터량이 급격히 증가하는 구조적 문제 발생
- 이러한 병목은 LLM 추론이 프리필(Prefill, 문맥처리)과 디코드(Decode, 답변생성)의 두 단계로 이뤄지는 구조에서 발생, 디코드에서는 누적된 KV 캐시 반복 조회 때문에 메모리 대역폭 제약이 두드러짐
  - (프리필 단계) 긴 문맥을 한꺼번에 처리하면서 KV 캐시 생성량과 연산 부담이 함께 증가해 GPU 메모리 사용량이 빠르게 확대
  - (디코드 단계) 토큰을 생성할 때마다 누적된 KV 캐시 전체를 반복 조회해야 해 메모리 대역폭 제약이 특히 크게 발생
- 결과적으로 KV 캐시 증가는 동시 처리 가능한 요청 수를 제한하고, 응답 지연과 인프라 비용 증가를 동시에 초래하면서 AI 서비스 품질과 운영 효율에 직접 영향
- 이에 따라 고가 GPU·HBM을 추가 구매하는 기존 대응 방식의 비용 부담이 확대되면서, SW 알고리즘을 통한 메모리 효율화가 대안으로 주목

구글, KV 캐시 병목 완화를 위한 '터보퀀트(TurboQuant)' 공개

- 이와 같이 KV 캐시가 추론 단계의 핵심 병목으로 부상하자, 구글은 2026년 3월 이를 3비트 수준으로 압축하는 알고리즘 '터보퀀트'를 공식 발표
- 엔비디아 H100 GPU 환경 테스트에서 메모리 사용량 최소 6배 절감, 연산 처리 속도 최대 8배 향상을 제시하며 추론 효율화 해법으로 주목
- 별도 재학습이나 추가 조정 없이 적용 가능한 방식이라는 점에서 기존 인프라의 활용도를 높일 수 있는 기술로 평가

AI 인프라 경쟁, GPU 확보에서 추론 메모리 효율화 경쟁으로 확대

- KV 캐시가 추론 단계의 핵심 병목으로 부상하면서, AI 인프라 경쟁도 GPU 수량 확보를 넘어 메모리 활용 효율을 높이는 방향으로 확대
- 구글은 터보퀀트를 통해 KV 캐시 자체를 더 작게 압축하는 알고리즘 해법을 제시, 엔비디아는 NVFP4 기반 저정밀화와 Dynamo 기반 오프로딩을 결합해 메모리 사용량과 대역폭 부담을 줄이는 해법 제시

- NVFP4는 KV 캐시를 더 적은 정밀도로 저장해 메모리를 최대 50% 줄이고, Dynamo는 GPU 메모리를 넘는 KV 캐시를 CPU 메모리·스토리지로 분산해 용량 한계를 우회하는 방식
- 즉, 구글이 기존 인프라에서도 적용 가능한 압축 알고리즘에 무게를 둔 반면, 엔비디아는 GPU 아키텍처와 메모리 계층 전체를 활용하는 방향으로 추론 효율화를 추진하고 있어, 접근 방식이 구분
- 결국 양사의 흐름은 단순한 신기술 경쟁이 아닐, AI 서비스 확산에 따른 추론 비용과 메모리 병목을 줄이기 위한 산업 전반의 구조적 대응으로 해석

터보퀀트의 압축 원리와 기대효과

- **(기술 원리)** 오버헤드 문제를 줄인 KV 캐시 압축 구현
  - 기존 양자화 기술은 데이터를 압축할 때 블록마다 정규화 상수와 같은 보조 정보를 별도로 저장해야 해, 이론상 압축률에 비해 실제 메모리 절감 효과가 줄어드는 한계 존재
  - 터보퀀트는 이 오버헤드 문제를 해결하기 위해 설계된 기술로, 폴라퀀트(PolarQuant)와 QJL을 단계적으로 결합해 오버헤드 없이 KV 캐시를 3비트 수준까지 압축하는 구조
  - 압축 과정에서 추가 학습이나 별도 데이터 없이 오차를 제거하는 구조로, 특정 모델이나 도메인에 관계없이 범용 적용이 가능한 점이 실용적 강점
- **(1단계)** PolarQuant, 좌표 변환으로 보조 정보 저장 부담 축소
  - 기존 방식이 벡터 압축에 필요한 보조 정보를 반복 계산·저장해야 했다면, PolarQuant는 좌표 변환(극좌표, 방향각·거리)을 통해 이를 줄여 압축에 더 유리한 구조를 형성
  - 이 방식은 정규화 상수 저장 부담을 줄이면서도 데이터의 핵심 방향성과 강도를 비교적 잘 유지하는데 초점, 1단계에서 압축의 대부분을 수행하고 이후 QJL이 남은 오차를 보정
- **(2단계)** QJL(Quantized Johnson-Lindenstrauss), 1비트 보정으로 잔여 오차를 축소
  - QJL은 고차원 벡터를 저차원으로 축소할 때 데이터 간의 핵심 거리·관계를 수학적으로 보존하는 변환 기법을 압축에 적용
  - 1단계 PolarQuant 압축 후 남은 미세 오차를 단 1비트만 사용해 제거, 추가 메모리 비용 없이 어텐션 스코어의 정확도를 유지하는 '제로 오버헤드' 구조를 완성

- 두 기법의 결합으로 터보퀀트는 다양한 장문 처리 벤치마크에서 기존 압축 기술 대비 최고 수준의 성능을 달성, 실제 서비스 환경에서의 신뢰성 확보

터보퀀트와 메모리 수요의 상관관계

● 추론용 메모리 효율화 확산, 단기적으로 AI 메모리 수요 둔화 우려를 자극

- 터보퀀트가 동일한 GPU 메모리로 더 많은 추론 작업을 처리할 수 있게 함에 따라, 단기적으로는 추론 인프라에 필요한 메모리 탑재 부담이 줄어들 수 있다는 우려 제기
- AI 시장이 학습 중심에서 추론 중심으로 전환되는 흐름 속에서, 추론 단계의 메모리 효율화가 HBM 등 고성능 메모리 수요 증가 속도를 일시적으로 완화할 수 있다는 시각도 존재
- 다만 터보퀀트는 아직 연구·검증 단계에 있어 실제 기업 서비스 인프라에 광범위하게 적용되기까지는 시간이 필요하며, 단기 수요 구조에 미치는 영향은 제한적일 가능성이 높음
- 아울러 KV 캐시는 추론 단계에서만 사용되는 임시 메모리로, AI 모델을 학습시키는 데 쓰이는 훈련용 HBM과는 구조적으로 분리되므로 전체 AI 메모리 수요 감소로 해석하기는 어려움

● 중장기적으로는 효율화가 오히려 전체 메모리 수요를 확대할 가능성도 존재

- 단기적 영향은 제한적일 수 있으나, 중장기적으로는 효율화가 더 긴 문맥 처리와 더 많은 서비스 수요를 자극해 전체 수요를 오히려 확대할 가능성 존재
- '25년 1월 딥시크(DeepSeek) 저비용 AI 모델 공개 당시에도 유사한 효율화 충격이 나타났으며, 이후 일부 하이퍼 스케일러들이 투자 계획을 유지하거나 오히려 상향 조정 사례 참고 가능
- 기업들은 효율화로 확보된 자원을 비용 절감보다 장문맥 처리, 동시 사용자 확대, 서비스 고도화에 재투입할 가능성이 높아, 총 메모리 수요의 방향을 단정하기 어려움
- 결국, 단기적으로는 효율화에 따른 수요 위축 우려가 제기될 수 있으나 AI 서비스 확산이 본격화될수록 효율 혁신이 수요를 제한하기보다 시장을 키우는 동력으로 작용할 가능성도 존재

출처 : ZDNet Korea 외(2026.3.)

https://zdnet.co.kr/view/?no=20260326163235
https://www.cnbc.com/2026/03/26/google-ai-turboquant-memory-chip-stocks-samsung-micron.html
https://www.computerworld.com/article/4150436/google-targets-ai-inference-bottlenecks-with-turboquant-2.html
https://techcrunch.com/2026/03/25/google-turboquant-ai-memory-compression-silicon-valley-pied-piper/
https://www.kmib.co.kr/article/view.asp?arcid=1774774167&code=11151100&cp=nv
https://www.chosun.com/economy/tech_it/2026/03/26/JHCQKRRSOBHXBHAXZCLUCJDMKE/
https://zdnet.co.kr/view/?no=20260326163235
https://www.aitimes.kr/news/articleView.html?idxno=39280
https://zdnet.co.kr/view/?no=20260327143604
https://zdnet.co.kr/view/?no=20260326192031
https://www.asteralabs.com/what-nvidia-gtc-2026-said-about-the-future-of-ai-connectivity/
https://www.marktechpost.com/2026/02/10/nvidia-researchers-introduce-kvtc-transform-coding-pipeline-to-compress-key-value-caches-by-20x-for-efficient-llm-serving/
https://venturebeat.com/orchestration/nvidia-shrinks-llm-memory-20x-without-changing-model-weights
https://quasarzone.com/bbs/qn_hardware/views/1986809
https://research.google/blog/turboquant-redefining-ai-efficiency-with-extreme-compression/