Lablup | VAST

Solution Brief

Backend.AI + VAST Data 스토리지에서 KV 캐시 오프로딩: 에이전트 코딩 TTFT를 절반으로 단축하는 방법

Backend.AI와 VAST Data로 고부하 에이전트 환경에서
라인 레이트 수준 KV 캐시 오프로딩 제공

상태를 유지하는 방향으로 진화하는 추론 작업

초기 LLM 서빙은 단일 프롬프트를 처리하는 Stateless 패턴에 최적화되어 설계되었지만, 에이전트 워크로드는 여러 턴의 대화에 걸쳐 동일한 작업 맥락을 다시 불러와야 하는 Stateful 패턴을 형성합니다. 특히 프로그래밍을 위해 사용하는 코딩 에이전트는 방대한 코드베이스, 도구 호출 맥락, 이전 추론 결과를 포함한 긴 Prefix를 반복적으로 재사용하므로, 사용자 체감 성능은 단순 토큰 처리량(Throughput)보다 동일한 컨텍스트에 얼마나 빠르게 재진입할 수 있는지에 더 크게 좌우됩니다.

GPU에 KV 캐시를 두는 접근만으로 부족한 이유

에이전트 코딩 세션은 세션별로 긴 Prefix를 형성하고 이를 여러 턴에 걸쳐 재사용합니다. 각 턴에서 추론 엔진은 먼저 입력 전체를 처리하며 KV 캐시를 생성하고(Prefill), 이후 해당 캐시를 재사용해 토큰을 순차 생성합니다(Decode). 문제는 여러 세션이 하나의 인퍼런스 서버를 번갈아 점유하면 GPU 메모리 내 KV 캐시 Working Set이 빠르게 커지고, 이전 턴에서 사용된 KV 블록은 다른 세션에 의해 축출되어 다음 턴에서 Prefill을 다시 수행하게 됩니다.

KV 캐시 오프로딩: GPU 바깥으로 KV 캐시를 이동하는 전략

KV 캐시 오프로딩은 GPU가 현재 활성 처리에 사용하지 않는 KV 블록을 외부 저장소로 이동하고, 동일한 컨텍스트가 다시 필요해질 때 이를 재계산하는 대신 외부 저장소로부터 다시 로드하는 방식입니다. 이 구조에서 GPU는 즉시 다음 연산에 필요한 컨텍스트(Hot context)를 유지하고, 당장 사용되지 않지만 나중에 다시 필요할 수 있는 컨텍스트(Cold context)는 외부 계층으로 이동하게 됩니다. 이 방식을 적용하면 GPU 메모리가 모든 컨텍스트를 떠안아야 하는 부담에서 벗어납니다.

왜 KV 캐시 오프로딩이 중요한가요?

더 낮은 TTFT

반복되는 턴의 비용을 70% 가까이 절약 가능

더 높은 GPU 활용율

반복 Prefill 작업을 회피, 중요한 작업에 GPU를 더 많이 할당

라인 레이트 성능

호스트 RAM을 거치지 않고 GPU와 스토리지 간 KV 블록 직접 이동

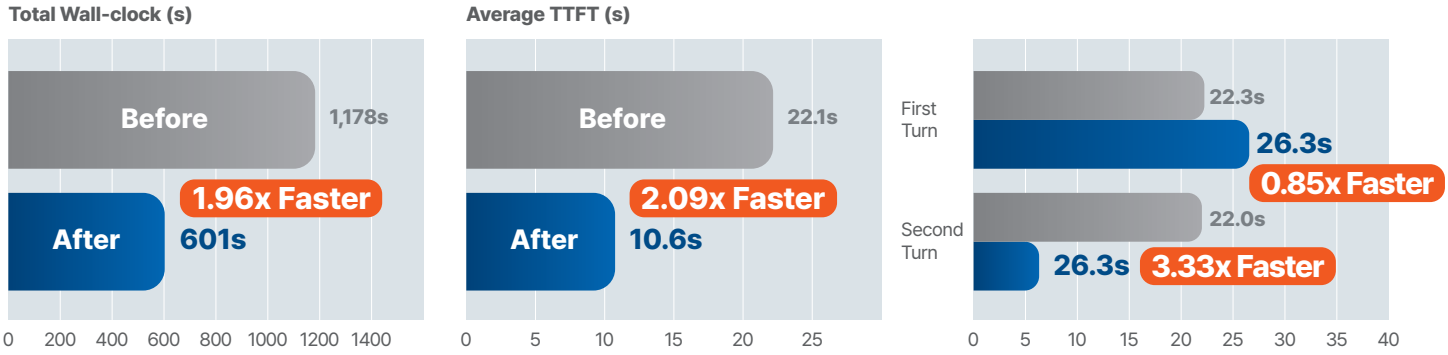
Backend.AI + VAST Data 솔루션

Backend.AI는 AI 개발과 서비스 운영을 아우르는 래블업의 엔터프라이즈 AI 인프라 운영 플랫폼입니다. GPU, NPU, TPU 등 다양한 형태의 이기종 AI 가속기를 하나의 운영 체계 안에서 다룰 수 있도록 설계되었으며, 다양한 스토리지 시스템과 체계적으로 연동하여 스토리지 본연의 성능을 최대한 끌어낼 수 있도록 설계되었습니다. VAST Data 스토리지 및 VAST Data AI OS는 초고속 AI, 머신러닝, 고성능 컴퓨팅(HPC) 워크로드를 위해 설계된 차세대 올플래시(All-Flash) 데이터 플랫폼으로, Backend.AI와 결합하여 라인 레이트 KV 이동을 가능하게 하는 스토리지 계층을 구성합니다.

Backend.AI에서 모든 연산 작업은 '세션'이라는 단위로 표현되며, 세션은 Storage Proxy를 통해 VAST KV 캐시 VFolder를 마운트합니다. 이 세션 내부에서 vLLM과 LMCache가 실행되며, LMCache의 gds_path가 VAST 마운트를 직접 참조합니다. Backend.AI와 VAST Data의 스토리지는 NVIDIA Magnum IO GPUDirect Storage를 활용, KV 블록을 Host RAM 등 다른 경로를 거치지 않고 GPU 메모리에서 스토리지로 직접 전송합니다.

벤치마크 결과

2배 개선된 전체 소요 시간 및 TTFT, 두 번째 턴부터 드러나는 명확한 우위



- Without offload
- With offload
- Accelerated drainage

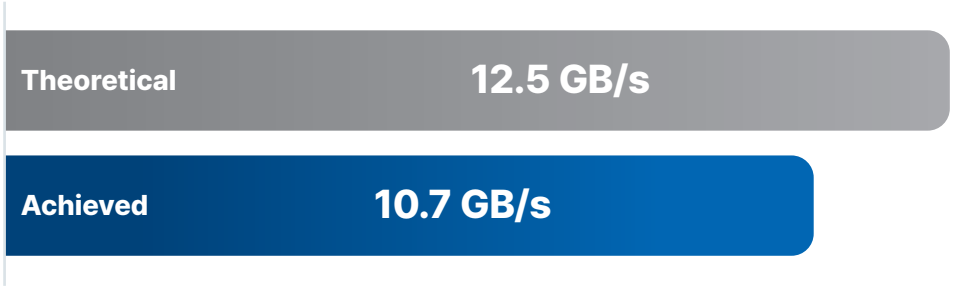
전체 소요 시간(wall-clock)과 평균 TTFT(Time-To-First-Token)는 모두 약 2배 개선되었습니다. 이 성능 향상은 주로 첫 번째 토큰을 더 빨리 가져오는 데서 비롯됩니다. 평균 TTFT는 첫 방문 턴과 동일한 컨텍스트를 다시 사용하는 이후 턴을 구분해 볼 때 특히 의미가 있습니다. 첫 번째 턴에서는 프리필과 스토리지 쓰기가 함께 발생하므로 약 4초의 오프로딩 오버헤드가 발생합니다. * 그러나 이후 턴에서는 재계산 대신 캐시된 블록을 다시 읽어오므로 TTFT가 기준치의 약 3분의 1 수준으로 줄어듭니다. 그 결과 사용자는 토큰 생성이 22초가 아니라 약 6초 후에 시작되는 것을 체감합니다.

참고: 2026년 7월 현재 기준 KV cache offloading을 활성화하면 vLLM 세션의 첫 번째 턴에서 수 초 수준의 추가 지연이 발생합니다. 래블업과 VAST Data의 공동 실험 결과 이러한 지연 경향이 저장 매체와 무관하게 일관되게 관찰되었으므로, 원인은 하드웨어 구성보다 LMCache 구현에 있는 것으로 추정되고 있습니다.

대규모 서빙에서 경험하는 라인 레이트 처리량

VAST는 100Gbps 패브릭에서 최대 10,693 MB/s와 10,441 ops/s 성능을 기록했으며, 평균 round-trip time은 8.265ms, retransmit은 0건이었습니다. 이 수준의 처리량은 장문 컨텍스트 KV 캐시를 VAST와 Backend.AI 환경에서 다시 불러오는 것이 GPU에서 재계산하는 것보다 빠르다는 점을 보여줍니다. 고속 네트워크의 한계치에 근접한 데이터 전달은 대규모 멀티테넌트 워크로드에서도 스토리지 성능 저하를 최소화합니다.

VAST Reads Throughput



Backend.AI와 VAST Data AI OS의 결합을 통해 스토리지 계층에서 병목을 만들지 않으면서 워크로드를 수평으로 확장할 수 있습니다. 이러한 환경은 여러 세션 컨텍스트가 여러 개의 테넌트로부터 동시에 유입되는 프로덕션 LLM 서빙 환경에서 필수적입니다. 법률 문서 처리, 다중 사용자 멀티턴 엔터프라이즈 챗봇 등 다양한 프로덕션 상황에서 Backend.AI와 VAST Data는 KV 캐시를 GPU 메모리에만 의존하는 구조를 벗어나, 스토리지 계층까지 활용하여 실질적인 성능 이득을 경험할 수 있습니다.

벤치마크 조건

벤치마크는 Backend.AI 26.4 환경과 VAST AI OS v5.4 위에서 수행했으며, 8x H100 GPU에서 Mistral Medium 3.5 128B를 사용했습니다. 단일 추론 서버는 140K 토큰 컨텍스트 길이를 기준으로 10개의 서로 다른 Agent context를 5턴씩 순환 처리했습니다. 동일한 컨텍스트를 다시 처리하기 전에는 다른 9개의 장문 컨텍스트 세션이 GPU를 점유했습니다.

Backend.AI 소개

Backend.AI는 래블업이 개발한 AI 운영체제입니다. 어떤 가속기, 어떤 클라우드, 어떤 데이터센터에서도 워크로드를 오케스트레이션할 수 있도록 설계되어, 팀이 인프라 관리보다 모델 개발에 더 많은 시간을 쓸 수 있게 합니다. Python, R, C/C++, Java, Rust 등 다양한 프로그래밍 언어와 PyTorch, Megatron-LM, vLLM 같은 다양한 AI 라이브러리를 지원하는 등 특정 프레임워크에 종속되지 않는 범용 실행 환경을 지향하며, 다양한 설치환경을 지원하여 온프레미스, 클라우드, 그리고 하이브리드 환경에서도 동일한 경험을 목표로 Backend.AI를 사용할 수 있습니다. 래블업은 AI 실행 환경을 소프트웨어를 통해 간소화해 왔으며, 업계가 AI 인프라에 본격적으로 주목하기 시작한 시점보다 이른 2015년부터 AI 인프라 오케스트레이션을 다뤄온 전문가 집단입니다. 여러분의 AI 스택을 재구축하지 않고 모든 가속기에서 AI를 실행하고 싶으신가요? contact@lablup.com으로 문의하세요.

VAST Data 소개

VAST Data는 AI Operating System 기업으로, AI의 잠재력을 최대한 끌어내기 위해 설계된 통합 소프트웨어 인프라 스택을 제공합니다.
VAST AI OS는 데이터 서비스, 컴퓨트 서비스, 그리고 에이전트 실행 기능을 하나의 확장 가능한 플랫폼으로 통합하여, 조직이 AI 에이전트 간 배포와 상호 통신을 지원하며, 실시간 데이터에 대한 추론을 수행하고, 복잡한 워크플로를 글로벌 규모에서 자동화할 수 있도록 지원합니다. VAST는 자체 DASE 아키텍처를 기반으로 합니다. DASE는 성능, 확장성, 단순성, 복원성 사이의 트레이드오프를 제거하도록 설계된 세계 최초의 진정한 병렬 분산 시스템 아키텍처입니다. 이를 바탕으로 VAST는 현대적 인프라를 추론형 AI를 위한 글로벌 패브릭으로 발전시켰습니다. KV 캐시를 사용하여 추론 속도를 높이는 방법을 알아보고 계신가요? hello@vastdata.com으로 VAST 솔루션 엔지니어에게 문의하세요.