Solution Brief

래블업 + 텐스토렌트

AI 인프라 관리 플랫폼 Backend.AI와 Tenstorrent Wormhole™ AI 가속기의 강력한 결합

Backend.AI®의 강력한 인프라 관리 능력에 Tenstorrent Wormhole™ AI 가속기의 성능을 더하다.

래블업 Backend.AI와 텐스토렌트 Wormhole이 함께 구축하는 AI 환경의 미래

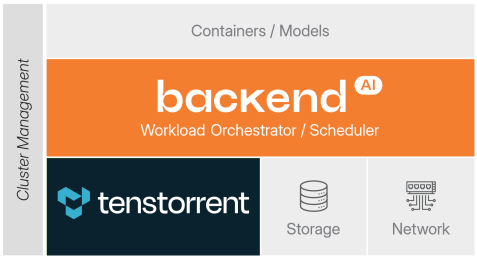
엔터프라이즈 생성형 AI 솔루션을 개발, 배포, 관리 및 확장하는 모든 과정을 그 어느 때보다 쉽고 효율적으로 수행할 수 있는 하드웨어와 소프트웨어의 조합, Lablup Backend.AI와 Tenstorrent Wormhole™을 소개합니다.

엔터프라이즈에 AI 전용 칩이 필요한 이유

현대의 AI 워크로드는 대규모 병렬 처리, 높은 메모리 대역폭, 낮은 지연 시간을 요구하며, 이러한 요구는 오직 설계 단계부터 AI 워크로드의 특성을 고려하여 설계된 전용 칩만이 충족할 수 있습니다. AI 전용 칩의 도입은 변화하는 AI 시장에서 경쟁력을 유지하고, 지속적으로 혁신하기 위한 필수 요소입니다.

비즈니스가 다루는 AI 워크로드의 특성과 규모, 예산 제약, 운영 환경 등에 따라 다양한 선택지가 존재할 수 있으며, 비즈니스 목표와 운영 요구에 부합하는 엔터프라이즈 AI 솔루션을 선택하는 것이 필수적입니다. 이러한 솔루션을 구성하는 토대를 Tenstorrent의 AI 전용 칩 위에서 시작해 보세요.

하드웨어를 뒷받침하는 강력한 소프트웨어의 중요성

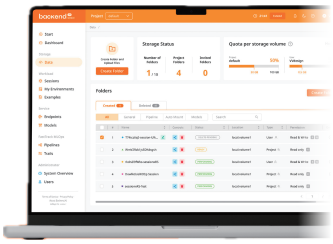


AI 가속기가 지니는 가치를 온전하게 활용하기 위해서는 지능적인 오케스트레이션이 필수적입니다. 지능형 오케스트레이터는 GPU 리소스를 동적으로 할당하여 활용도를 극대화하고, 리소스 경합을 방지하며, 여러 사용자와 프로젝트 간의 공정한 스케줄링을 보장합니다.

Backend.AI의 Sokovan 오케스트레이터는 엔터프라이즈 환경의 데이터 파이프라인 및 CI/CD 워크플로우와 통합되어, 대용량 데이터셋을 필요한 위치로 자동으로 이동시키고, 분산 학습이나 추론 환경에서 지연 시간을 최소화합니다. 이를 통해 조직이 인사이트를 효율적으로 도출하도록 지원하고, 모델 배포에 소요되는 시간을 단축할 뿐만 아니라, 운영의 복잡성과 잠재되어 있는 리스크를 줄여 조직의 혁신을 실현할 수 있도록 돕습니다.

Lablup Backend.AI®

Transforming GPU complexity into operational simplicity



Backend.AI는 자체 개발한 오케스트레이션 및 작업 스케줄러를 기반으로 한 이기종 가속 워크로드 호스팅 플랫폼으로, 온프레미스, 단절망 환경 뿐 아니라 클라우드 네이티브, 그리고 하이브리드 클러스터 환경에서도 운영할 수 있습니다. Backend.AI는 대규모 분산 모델 학습부터 추론까지 전체 워크플로우를 간소화하며, 멀티 노드와 멀티 테넌트를 아우르는 강력한 파이프라인 관리를 통해 AI를 위한 모든 과정을 안정적이고 효율적으로 처리합니다. Backend.AI는 AI 작업 워크플로우를 하나의 통합 소프트웨어 플랫폼 내에서 안전하고, 유연하고, 병렬적으로 실행할 수 있도록 보장합니다.

Backend.AI의 폭넓은 생태계를 통해 다양한 파트너 솔루션 및 플러그인과 긴밀하게 통합하고, 데이터 수집부터 모델 학습, 배포, 모니터링까지 모든 과정을 유기적으로 연결할 수 있습니다. 고객의 변화하는 요구에 지속적으로 대응할 수 있는 능력을 갖춘 Backend.AI와 함께 MLOps부터 실제 서비스 배포까지 함께하세요.

Tenstorrent Wormhole™

Flexible, scalable processors built with Tensix Cores™

Tenstorrent Wormhole™ ASIC은 전통적인 GPU 대비 비용 대비 성능 측면에서 우수한 효율을 제공하는 AI 가속기입니다. Wormhole™ PCIe 카드는 고성능 Tensix Core™ 기반의 유연하고 확장 가능한 프로세싱 솔루션을 제공합니다. 각 카드는 연산 유닛, 네트워크 온 칩(Network-on-Chip), 로컬 캐시, 그리고 "baby RISC-V" 코어를 통합하여 칩 내부에서 효율적이고 고속의 데이터 이동을 실현합니다.

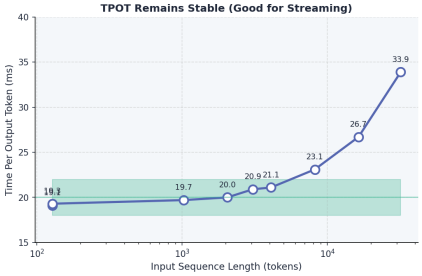
Wormhole™은 기존 GPU 대비 뛰어난 성능 대비 비용 효율성을 제공하며, 다양한 데이터 정밀도 포맷을 폭넓게 지원합니다. n150s와 n300s 두 가지 모델로 제공되는 Wormhole은 변화하는 엔터프라이즈 AI 워크로드 요구에 탁월하게 적응하고, 변하지 않는 가치를 조직에 제공합니다.

성능 벤치마크

실제 사용자가 경험할 수 있는 성능을 측정하기 위해 Wormhole LoudBox 환경에서 Llama-3.1-8B-Instruct 모델을 대상으로 성능 벤치마크를 수행했습니다. 측정 데이터는 동일 조건의 환경에서 수집되어 결과의 일관성과 재현성을 확보했습니다. Backend.AI를 Wormhole LoudBox에 설치하면 AI 가속기부터 스토리지, 네트워크 연결성까지 모든 하드웨어 스택을 유연하게 활용하는 최적의 운영 환경을 구성할 수 있습니다.

활용 시나리오 및 권장 구성

인터랙티브 채팅 · 실시간 어시스턴트

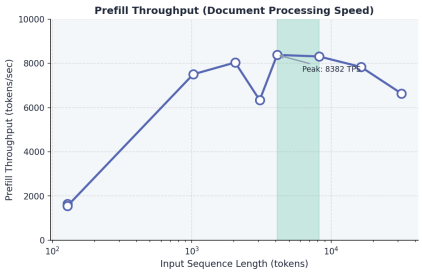


실시간 챗봇, 코파일럿, 고객 지원 어시스턴트와 같은 대화형 인터랙티브 환경에서는 사용자가 즉각적인 반응을 체감할 수 있는 낮은 TTFT(Time To First Token) 특징이 핵심입니다. 벤치마크 결과, 낮은 동시성(Concurrency 1~4) 구간에서 TTFT는 77~255ms로 빠르게 유지되었고, TPOT(Time Per Output Token) 역시 19~20ms 수준으로 안정적이었습니다. 이는 사용자가 첫 응답을 빠르게 확인하고, 이후에도 끊김 없이 자연스러운 스트리밍 응답을 경험할 수 있음을 의미합니다.

권장 구성:

- 낮은 동시성(Concurrency 낮게 유지) + 중간 컨텍스트(ISL ≤ 2K)
- 중간 길이 응답 (OSL 128~512)

RAG · 문서 기반 질의응답

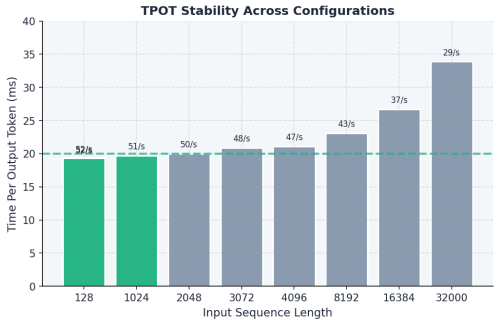


RAG 기반 문서 Q&A는 검색된 문서가 프롬프트에 포함되면서 입력 시퀀스 길이(ISL)가 길어지는 특성이 있어, 장문 컨텍스트 처리 능력과 대화형 응답성의 균형이 중요합니다. 일반적으로 이 유형의 워크로드에서는 ISL이 약 4K~8K 토큰까지 증가합니다. 벤치마크에서 동시성 조건 1 기준으로 TTFT는 489~986ms로 1초 이내를 유지했으며, Prefill 처리량은 8,300~8,400 TPS를 기록했습니다. 이는 긴 컨텍스트를 처리하는 Prefill 단계에서 텐스토렌트 하드웨어가 효율적으로 동작함을 보여줍니다.

권장 구성:

- 출력 길이를 128~256으로 제한하여 응답 지연 안정화
- 낮은 동시성에서 응답성을 확보한 후 캐싱/리트리벌 전략을 고도화

장문 콘텐츠 생성



출력 시퀀스 길이(OSL)가 1K 토큰을 초과하는 장문 생성 워크로드에서는, 초기 응답 속도(TTFT)보다 일관된 생성 속도와 안정성(TPOT)이 체감 품질을 좌우합니다. 벤치마크 결과 동시성(Concurrency) 1 조건에서 TTFT는 78~137ms로 빠르게 유지되었고, TPOT는 약 19~20ms 수준으로 안정적인 생성 성능을 보였습니다. 또한 1,024 토큰 생성에 소요되는 End-to-End 시간은 약 20초로 측정되어, 예측 가능한 처리 시간을 기반으로 운영 설계를 수립할 수 있습니다.

권장 구성:

- 낮은 동시성 유지로 TPOT 안정화(장문 생성 품질 및 일관성 확보)
- 최대 출력 길이 제한, 생성 중간 저장(Mid-generation saving) 등의 정책을 결합해 보다 예측 가능한 SLA 설계 지원

Recommendation matrix

사용 케이스	동시성	입력 시퀀스 길이	출력 시퀀스 길이	주요 메트릭
인터랙티브 채팅	1	≤ 2k	128-512	TTFT < 300ms
RAG / 질의응답	1	4k-8k	≤ 256	TTFT <1s
장문 콘텐츠 생성	1	≤ 1k	1k+	Stable TPOT

인터랙티브부터 배치까지 폭넓은 LLM 추론 호환성 확보

Tenstorrent Wormhole 아키텍처 기반의 LoudBox는 다양한 추론 워크로드 전반에서 예측 가능한 저지연과 높은 확장성을 동시에 제공할 수 있음을 보여줍니다. 실시간 채팅 및 에이전트 애플리케이션과 같은 인터랙티브 시나리오에서는 낮은 동시성 구간에서 TTFT 77~255ms, TPOT 19~20ms 수준을 유지하여 사용자에게 즉각적인 반응성과 자연스러운 스트리밍 경험을 제공합니다. 또한 RAG와 같은 장문 컨텍스트 중심 워크로드에서도 4K~8K 토큰 입력 구간에서 Prefill Throughput 8,300~8,400 TPS를 유지하면서 TTFT를 1초 이내로 제어해, 컨텍스트 처리와 생성 스트리밍이 모두 중요한 애플리케이션에 균형 잡힌 성능을 제공합니다.

Tenstorrent, 그리고 Backend.AI와 함께 여러분의 AI 서비스를 구축하기 위한 여정을 시작하세요.