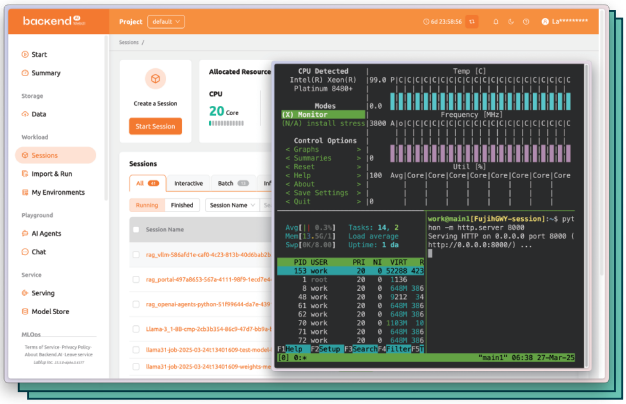
당신의 비즈니스를 새로운 차원으로 이끄는 혁신

Lablup® Backend.AI® meets Intel® Gaudi® 2 & 3 Platform

AI 가속기의 잠재력을 극대화하는 마법, Backend.AI®와 Gaudi®로 실현하는 AI 성능의 새로운 차원

고도화된 AI 비즈니스에는 강력한 AI 성능 뿐만 아니라 관리하기 쉬운 솔루션이 필수적입니다. 인텔®의 최신 가우디® 3 AI 가속기의 강력한 성능과 래블업의 Platform-as-a-Service 제품, Backend.AI가 만나 엔터프라이즈 AI 환경의 최대 잠재력을 보여드립니다.



똑똑한 스케줄러와 Sokovan™. 최고의 소프트웨어가 선사하는 최상의 하드웨어 성능

Backend.AI®는 고객 비즈니스 혁신을 위해 생성형 AI와 가속 컴퓨팅의 잠재력을 최대화하도록 설계되었습니다. 오로지 AI를 더 잘 해내기 위해 설계된 Sokovan™ 오케스트레이터와 똑똑한 스케줄러가 다양한 종류의 GPU와 AI 가속기의 무한한 가능성을 현실로 구현합니다.

소규모 모델부터 대규모 모델까지 원하는 규모의 모델을 개발, 학습, 서비스하는 동안 Backend.AI®는 하드웨어가 낼 수 있는 최대한의 성능을 제공합니다. 운영 비용과 복잡성을 크게 줄여주는 혁신적인 플랫폼, Backend.AI®는 빛의 속도로 AI 미래로 나아갑니다.

AI 가속기의 잠재력을 깨우는 완벽한 동행. 인텔® 가우디® 2/3 AI 가속기와 긴밀하게 통합된 Backend.AI®

엔비디아, 리벨리온, 퓨리오사AI, 그래프코어, AMD, 텐스토렌트를 포함해 다양한 제조사의 GPU와 AI 가속기를 지원하는 Backend.AI가 이젠 인텔을 품고 더욱 강력해졌습니다. 하드웨어 수준의 통합으로 긴밀하게 이어진 Backend.AI와 인텔 가우디 플랫폼의 강력한 조합을 소개합니다.

카드 단위 가속기 할당

Backend.AI는 인텔 가우디 2 및 인텔 가우디 3 AI 가속기 클러스터를 사용자들이 목적에 따라 활용할 수 있도록 지원합니다. 고객이 보유한 기존 플랫폼에서 모델을 학습한 후, 인텔® 가우디® 2 및 3 플랫폼에서 서비스를 배포하고 운영할 수 있습니다.

다양한 스토리지 솔루션 연동

Dell PowerScale, VAST Data, WekaFS, NetApp과 같은 초고속 스토리지 솔루션과의 완벽한 연동을 지원합니다. 클러스터 관리자나 사용자의 별도 설정 없이도, 연동한 솔루션이 제공하는 최고의 성능을 그대로 활용할 수 있습니다.

다양한 규모의 AI 워크로드 지원

소규모 모델을 구동하는 싱글 카드 AI 워크로드부터 초거대 사이즈의 모델을 구동하는 멀티 노드/멀티 카드 AI 워크로드까지, Backend.AI는 최상의 성능을 제공합니다. 다양한 환경의 AI워크로드도 Backend.AI와 함께라면 안심하고 사용할 수 있습니다.

추론 메트릭 통합 관리 기능

AI 프레임워크에서 제공하는 하드웨어와 소프트웨어의 실시간 지표를 모니터링하고 심층 분석할 수 있습니다. Backend.AI와 함께라면 복잡했던 추론 통계 관리가 간편해집니다.

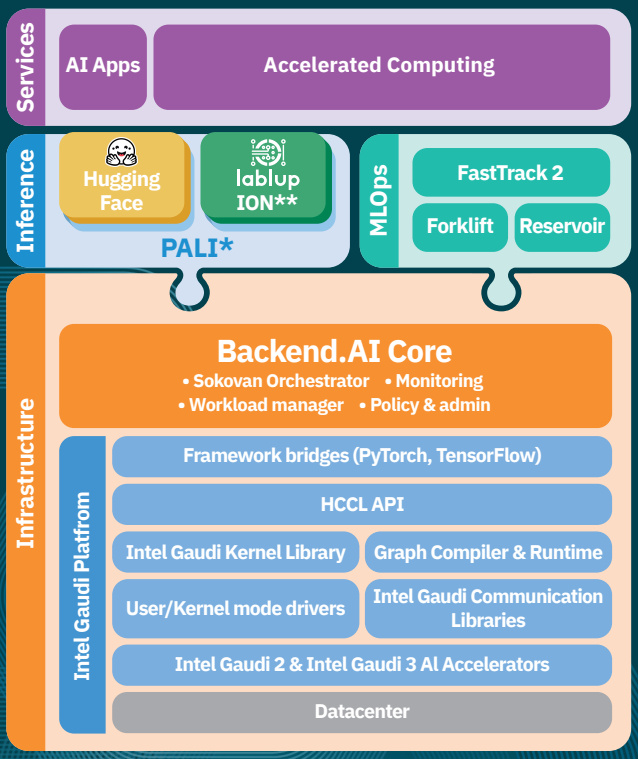
정책 기반 추론 워크로드 오토스케일링

자원 사용량을 스스로 최적화하는 스마트한 시스템이 관리자의 업무 부담을 줄여줍니다. Backend.AI는 하드웨어와 소프트웨어에서 수집된 다양한 정보를 분석하여 추론 워크로드를 효율적으로 관리합니다. 이를 통해 자원을 수동으로 모니터링하거나 제어할 필요 없이 시스템이 최적의 자원 할당을 수행합니다.

NUMA 최적화 자원 할당

Backend.AI는 여러 개의 CPU 소켓과 여러 개의 가속기로 구성된 시스템에서도 단일 노드 내에서 CPU 간 통신과 PCIe 버스 오버헤드를 제거함으로써 베어메탈 성능을 크게 향상시킵니다.

lablup | intel



*PALI : Performant AI Launcher for Inference **ION : Inference Objective Neuralnet

Lablup® Backend.AI®

베어메탈 급 성능을 확보하는 가속 컴퓨팅의 정점

- 독자 개발한 Sokovan™ 오케스트레이터로 구현된 뛰어난 Backend.AI Core 성능
- 유휴 리소스 재할당, 재사용을 통한 자원 활용도 극대화
- 소프트웨어와 하드웨어 정보를 결합하여 가속기 자원 최적화
- NUMA 최적화 자원 할당으로 CPU 간 오버헤드 및 PCIe 버스 오버헤드 제거
- RDMA 기반 스토리지 접근으로 고속, 저지연 데이터 처리
- 자동화된 멀티 노드 및 고속 연결 구성을 통한 성능 향상
- 대용량 페이지(Huge Pages) 할당을 통한 CPU 부담 최소화

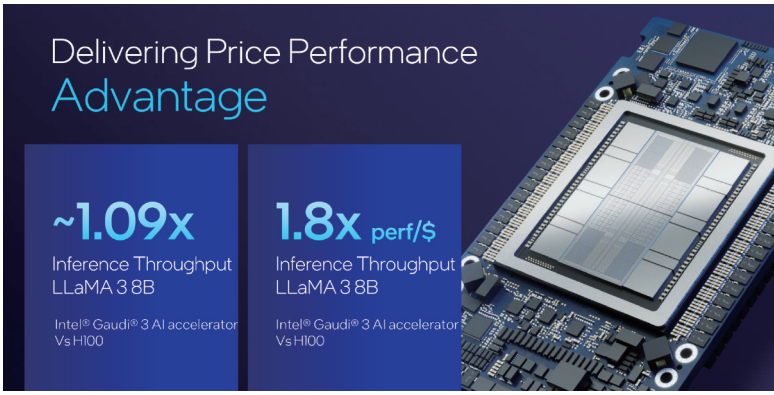
Intel® Gaudi® 3 AI Accelerators

나만의 방식으로 AI 워크로드를 처리하도록 구축된 새로운 고성능 옵션

인텔® 가우디® 3 AI 가속기는 대형 언어 모델 (LLM)과 스테이블 디퓨젼 (Stable Diffusion 등 이미지 생성)과 같은 생성 응용 프로그램부터 표준 객체 인식, 분류 및 음성 더빙에 이르기까지 모든 AI 워크로드에 대해 최첨단 데이터 센터 성능을 제공하도록 설계되었습니다.

인텔® 가우디® 3 AI 가속기에 대해 더 알아보고 싶으시다면,

intel.co.kr/gaudi3 을 방문하세요.



Source

인텔 측정 결과 vs H100 데이터 출처:
https://github.com/NVIDIA/TensorRTLLM/blob/main/docs/source/performance/perf-overview.md
테스트 조건: 가속기/GPU 1개당 128-2048tps 입출력 시퀀스.
인텔의 결과는 2024년 9월 9일에 측정되었음.
하드웨어: 인텔 가우디 3 AI 가속기 1개 (128 GB HBM) vs 엔비디아 H100 GPU 1개 (80 GB HBM); .
소프트웨어: 인텔 가우디 소프트웨어 릴리스 1.18.0.
엔비디아 H100에 대한 자료는 엔비디아에서 확인할 수 있습니다. 측정 결과는 환경에 따라 다르게 나타날 수 있습니다. 가격 추정치는 공개 자료 및 인텔 내부 분석에 기반합니다.

Disclaimers

인텔® 가우디® 3 AI 가속기에 대한 지원은 2025년 상반기에 예정되어 있습니다.
일정은 상황에 따라 변동될 수 있습니다.

제시된 데이터는 인텔에서 제공한 공식 수치이며, 래블업은 상기 데이터에 대한 일체의 보증을 하거나, 책임을 지지 않습니다.

인텔® 가우디® 3 AI 가속기가 Backend.AI 플랫폼에 통합될 경우 제시된 성능은 달라질 수 있습니다.

Make your AI Accelerator “manageable”

래블업은 과학자, 연구원, 개발자, 기업 및 AI 애호가들이 믿고 사용할 수 있도록 효율적이고, 확장 가능하며, 접근 가능한 AI 서비스를 만들고 있습니다. 래블업은 인텔과의 전략적 파트너십을 바탕으로 오늘날 인기를 끌고 있는 생성형 AI 및 딥 러닝 기반 서비스의 성공을 이끌어내고 있습니다.

검증된 기술력을 바탕으로, Backend.AI는 Intel® Gaudi® 2 및 3 플랫폼과 완벽하게 통합되어 고객들에게 최고의 성능을 제공합니다.

Backend.AI®에 대한 더 많은 정보는 backend.ai 웹사이트에서 확인할 수 있습니다.

lablup