AI Enterprise

AI Cloud

AI Open Source

AI MLOps

lablup

backend AI

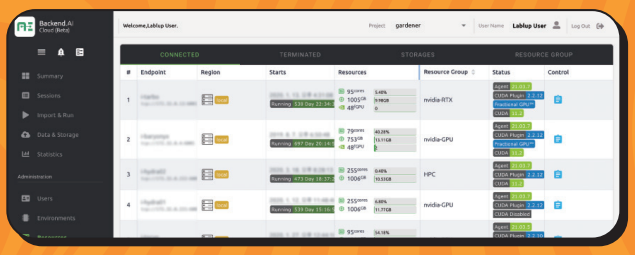
We'll Get You Every Last Bit.

Backend.AI는
아태지역 최초 NVIDIA DGX-Ready Software로 검증된
Hyper-Scalable 인공지능 연구&개발 플랫폼입니다.

GPU 클러스터 활용도 극대화

Backend.AI는 모든 유형의 딥 러닝 워크로드에 대하여 성능과 비용 모두를 만족시키는 최적화를 제공합니다. 많은 연산 자원이 필요한 모델 훈련을 고성능 GPU를 초고속 네트워크로 묶어 수행하고, 동시 추론 및 교육 워크로드에 대응하여 독자적인 GPU 분할 가상화 기능을 제공합니다. 빠르게 변하는 GPU 시장에 맞추어, 고가의 GPU를 항상 최적으로 활용할 수 있도록 AI 모델 훈련부터 추론 서비스, 머신러닝 교육으로 이어지는 전체 수명 주기 사용 계획을 제공합니다.

- 독자 기술 기반 컨테이너 수준 GPU 분할 가상화 ™ 지원
- GPU간 초고속 통신 기반 멀티노드 모델 훈련
- NVIDIA 다중 인스턴스 GPU (MIG) 지원



직관적인 관리 및 사용자 경험

클러스터에서 여러 사용자와 작업을 관리하고 모든 시스템을 완전히 활용하는 것은 어려울 수 있습니다. Backend.AI는 단일 노드부터 대규모 다중 노드 클러스터에 이르기까지 간단하고 일관된 사용자 및 관리 경험을 제공합니다.

- 웹 UI / 데스크톱 앱
- GUI 기반 MLOps 파이프 라인 / 배치
- 모니터링 솔루션과 통합되는 상세 로그 및 통계
- 자동화 및 통합을 위한 CLI / API / SDK

AI 및 HPC(고성능컴퓨팅) 최적화

Backend.AI 는 AI와 고성능 컴퓨팅에 특화된 독자적인 GPU 중심 오케스트레이터 및 스케줄러를 통해 딥러닝 친화적인 리소스 배치, 분산 처리용 다중 노드 워크로드, 데이터 I/O 병렬화를 지원하는 스토리지 프록시를 통해 연산자원을 자동으로 최적 관리하여 최대한의 잠재력을 발휘합니다.

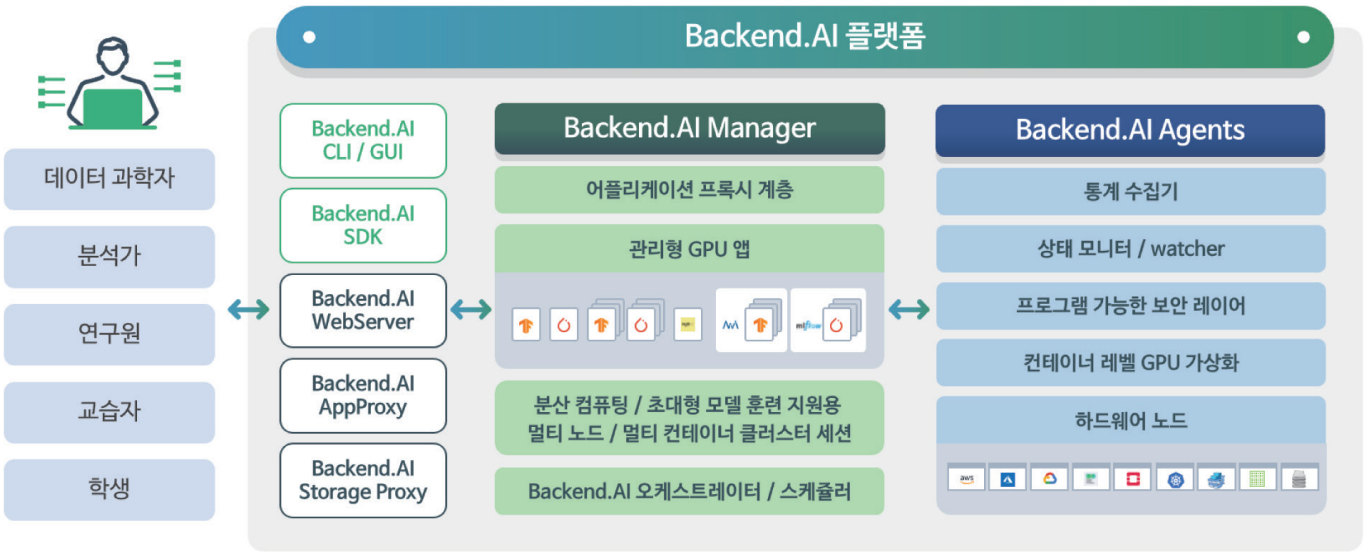
- 최적의 연산 자원 배치를 구현하는 독자적인 GPU 중심 오케스트레이터
- Air-gapped 클러스터를 위한 로컬 PyPI / CRAN / APT / Yum 저장소
- 사용률 기반 자동 자원 회수
- 배치 / 파이프라인 연산 스케줄링

쉬운 워크로드 규모 확장

분할 가상화된 GPU를 통해 작은 자원으로 딥 러닝 개발을 시작하십시오. 준비가 되면 Backend.AI 가 간단하게 AI 모델 훈련 규모를 효율적으로 확장해 드립니다.

- 자동 분산 교육 설정이 포함된 다중 노드 / 다중 컨테이너 세션
- 모델 훈련 및 데이터 I/O 파이프 라인 분리
- CephFS, PureStorage, NetApp, Dell, Weka.io 등의 분산 / 초고속 스토리지 솔루션에 대한 파일 입출력 지원 가속

Backend.AI 플랫폼



GPU 지원 컨테이너 수준 다중 GPU 할당 및 GPU 분할 공유 / NVLink 최적화 다중 GPU 플러그인 아키텍처

스케일링 온-프레미스 설치(실 서버/가상 서버) / 하이브리드 클라우드 운영(온-프레미스+클라우드) 및 다중 클라우드(이종)연동 / 워크로드 분산 처리시 멀티 네트워크 환경 자동 지원

스케줄링 GUI 어드민을 통한 통합 스케줄링 및 모니터링(CLI 지원) / 사용자 및 사용자 그룹별 자원 사용 / 다중 컨테이너 일괄실행 및 제어 기능 제공 / 가용 슬롯 기반 스케줄링 / 확장 및 사용자화 가능한 배치 스케줄러 / 가상화폐 마이닝 검출 및 차단 / 다양한 설정의 유휴 자원 자동 수거

보안 다중 사용자 지원 / 하이퍼바이저 혹은 컨테이너를 통한 샌드박싱 / 프로그래머블 샌드박싱 / 시스템콜 수준 로깅 / 관리자 모니터링

신뢰성 고 가용성 (HA) 구성 / 연산노드 라이브 추가 및 제거

UI/UX 사용자 어플리케이션(Windows10, MacOS 10.12~) / 웹 기반 서비스 지원 / 관제 콘솔 지원

데이터 관리 공유 스토리지 기능을 통한 데이터 업&다운로드 및 공유 지원 / EFS, NFS, SMB 및 분산 파일 시스템 사용 / 사용자 & 그룹 별 접근 제어 지원 / 로컬 가속 캐시(SSD, 메모리)

개발자 관리 범용프로그래밍 언어 지원(Python, C/C++, R, Java 등 17종) / 통합개발환경 플러그인(VS Code, IntelliJ, PyCharm) 제공 / 대화형 셸, 터미널 지원 / 컨테이너 이미지빌드 GUI

AI 개발자 / 데이터 과학자 지원 사용자 어플리케이션 내 Jupyter, TensorBoard 등 GUI 기반 도구 지원 / NGC(엔비디아 GPU 클라우드) 플랫폼 통합 / 주요 머신러닝 라이브러리 지원 : TensorFlow, PyTorch, CNTK, Mxnet 등 / 라이브러리 버전별 동시 지원(예:TensorFlow 1.0~2.10) / 웹 콘솔 내 Jupyter, TensorBoard, DevOps / MLOps 등 GUI 기반 도구 지원 / 머신러닝 라이브러리 자동 업데이트 / 함수화 딥러닝 모델 / 사용자 작성 모델 서빙 / 서빙모델 버전 관리

폐쇄망 지원 Backend.AI Reservoir 를 통한 자체 패키지 저장소 (PyPi, CRAN 및 Ubuntu 저장소 대상) / 스토리지 프록시 기반의 스토리지 가속 플러그인 지원 (CephFS, PureStorage, NetApp, Dell, Weka.io)

관리 및 제어 시스템 관리자 전용 대시보드 / 관리자 전용 컨트롤 패널 / 연산 노드 설정 제어 / 연산 노드 시스템 설정 변경 / 시스템 통계 수집 / 모니터링 솔루션 연동

Backend.AI 엔터프라이즈 패키지

Essential

엔터프라이즈 환경을 위한
딥러닝 연구 플랫폼

교육 기관 및 비영리단체

- ML/AI 개발환경
- 웹 UI 및 데스크탑 앱
- 개발환경 허브
- 관리자 전용 컨트롤 패널
- GPU 분할 가상화™
- 하이브리드 클라우드 구성

Pro

서비스 운영 및 프로덕션 모델 개발용
토탈 솔루션 및 컨설팅 서비스

기업 및 공공기관, 연구소

- Essential 포함
- AutoML 라이브러리 지원
- 모델 서비스
- MLOps 파이프라인
- 모델/데이터 저장소
- 관제용 대시보드

Reservoir

완전 폐쇄 환경 운영을 위한
Backend.AI와 통합된 패키지 저장소

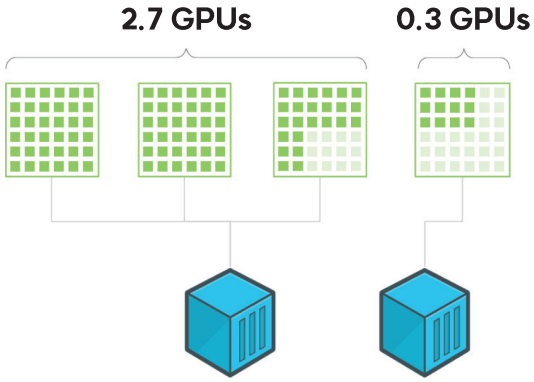
폐쇄망 운영 기업 / 기관

- PyPI / CRAN 저장소
- APT (Ubuntu, Debian)
- RPM / Yum (CentOS)
- 패키지 보안 체크
- 패키지 동기화 서비스
- 대시보드 통합

독자 기술 GPU 분할 가상화™

컨테이너 기반 GPU 스케일링

- CUDA MIG 및 CUDA MPS 여부에 구애받지 않음
- 단일 GPU 공유 : 교육 및 추론 워크로드에 적합
- 다중 GPU 할당 : 모델 훈련 등 대규모 워크로드에 적합
- 자체 개발한 CUDA 가상화 계층으로 구현
  [대한민국, 미국, 일본 등록 특허]
- GPU간 초고속 네트워킹 기반의 멀티노드 분산 GPU 훈련 지원



고성능 최신 하드웨어 기술 통합으로
AI 워크로드 성능과 효율 극대화

- 초고속 스토리지 솔루션 전용 I/O 가속 계층 제공(NetApp, PureStorage, Dell, Weka 등)
- RDMA 및 GPUDirect Storage를 통한 거대 모델 훈련 가속
- 다양한 AI 반도체 지원으로 워크로드 특성에 맞춘 전력 대 성능비 향상
- 최고 성능과 최저 비용을 함께 달성하는 GPU-NPU 통합 파이프라인 구성
- 저전력 고성능 컴퓨팅용 아키텍처 지원 (ARM & x86)

Backend.AI Reservoir

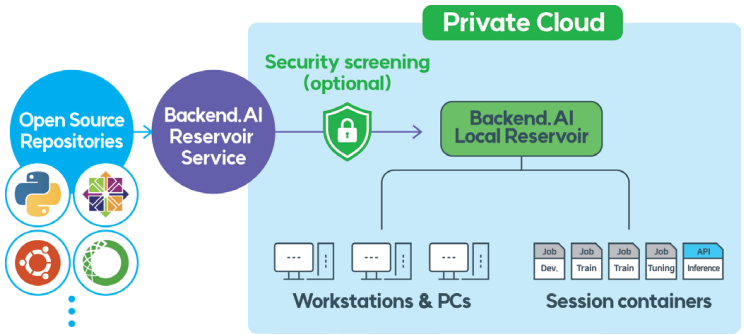
Backend.AI 클러스터 내부 / 외부 (옵션) 를 대상으로 한 독립 패키지 저장소 + 래블업의 업데이트 서비스

Backend.AI 기반 로컬 패키지 저장소 서비스

- Lablup 저장소 허브 서비스 + 로컬 패키지 서버
- Backend.AI Enterprise 의 추가 컴포넌트

고객 혜택

- 격리된 네트워크 내의 올인원 오프라인 서비스 구축
- 패키지 저장소 캐시 서비스로 네트워크 트래픽 절약



Backend.AI Forklift

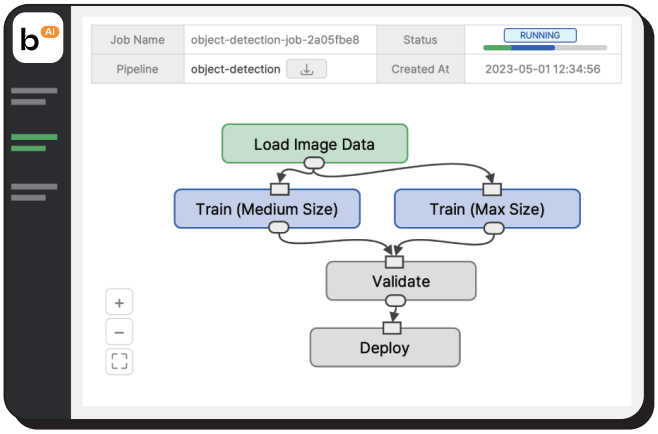
GUI로 누구나 쉽고 빠르게 사용할 수 있는 컨테이너 이미지 제작 도구 몇 번의 클릭만으로 개발과 연구 환경을 간편하게 구축하여 시간과 노력을 절약

- 빠르고 쉬운 Backend.AI용 컨테이너 이미지 구성
- 다양한 사전 설치된 소프트웨어 패키지 및 디폴트 설정
- 용이한 유지보수 및 업그레이드
- 만든 이미지를 Backend.AI에서 바로 사용

Backend.AI FastTrack

AI 엔지니어링 효율 극대화를 위한 MLOps 플랫폼, AI 개발의 모든 과정을 빠르고 효율적으로 관리

- 다양한 AI 파이프라인 템플릿을 제공
- Backend.AI와의 완벽한 호환성을 고려해 설계되어 최적 성능과 비용 효율 제공
- GUI 드래그 앤 드랍 기반 노코드 AI 파이프라인 엔지니어링 제공
- 데이터와 코드를 패키지로 관리하여 이전, 관리와 공유 용이
- AI 모델 서비스를 자동 터널링을 통해 보안 우려 없이 제공



Backend.AI 왜 도입해야 하는가?

AI/ML/HPC를 이용한 데이터 분석 및 모델 개발 R&D 부터 Business Service, AI 서비스 추론까지 하나의 일관된 플랫폼인 Backend.AI 를 통해 누구나 시간과 비용의 제약 없이 인공지능을 만들고 효과적으로 관리/서비스 할 수 있습니다.

Q. 수천 개의 시뮬레이션을 실행하고 단시간에 수백만 개의 데이터를 분석할 수 있게 최적화 된 초거대 팜을 구성하고 싶습니다.

A. Backend.AI는 AI, ML, HPC, 수치해석 등 연구 개발 환경에 최적화되어 있습니다. 멀티노드 분산 훈련 및 GPU간 네트워크 기반의 대규모 분산처리가 최적화되어 있으며, 고성능 컴퓨팅에 특화된 다양한 배치 자리, 자원 할당 및 병목 제거 구현을 통해 수백만 개의 데이터를 단시간에 분석할 수 있습니다.

Q. 머신러닝 교육 및 개발 클라우드 서비스 도입 시 최소의 비용과 인력만으로 개발 환경을 구성하여 효율적으로 관리하고 싶습니다.

A. Backend.AI는 ML / HPC 전문가들이 직접 만든 플랫폼으로 GPU 분할가상화 (Fractional GPU™) 를 통한 고가 GPU의 활용성 증대 및 고가용성을 달성할 수 있습니다. 더 적은 하드웨어(GPU)로 동일한 성능, 동일한 교육환경 제공으로 비용 절감은 물론 강력한 장애 대응(연속성 Fail Over, 쉬운 장애 원인분석 및 로그 API / 로그 솔루션 통합) 시스템 관리 컨트롤 패널을 통한 상세한 관리자 제어 기능을 제공하여 적은 인원으로 효율적으로 관리할 수 있습니다.

Q. 갑자기 많은 자원이 필요할 때 즉시 퍼블릭 클라우드 인스턴스를 추가하여 하이브리드 클라우드로 구성하여 사용하고 싶습니다.

A. Backend.AI는 Public Cloud 물론 On-Prem에서 Hybrid 클라우드까지 빠르고 쉽게 확장할 수 있습니다. 또한 다양한 GPU 및 머신러닝 가속 H/W 지원과 완전 문서화된 API 및 SDK(Python, Node.js) 제공으로 단시간 구성이 가능합니다.

Q. 언제 어디서나 내 머신러닝 개발환경에 접속해서 개발만 하고 싶습니다.

A. Backend.AI는 웹에서 접속만하면 개발자가 직접 설정해 둔 개발 환경, 연산 자원 등 변경없이 언제 어디서든 실행이 가능합니다. 네트워크를 통해 사용자가 개발, 운영하는 제품에 연동할 수 있는 공용 API 및 SDK를 제공하며, 사용자 유형별 커스터마이징 할 수 있습니다.

Customer story

국민대학교 경영대학원
조윤호 경영대학원장

Backend.AI를 활용하여 세 대의 GPU서버를 가지고 80여명의 수강생들이 동시에 모델링 실습과 과제를 수행할 수 있었습니다.

전담 관리자 없이 여러 서버의 개발 환경을 구성하고 자원을 효율적으로 관리할 수 있다는 점 그리고, 대부분의 기능을 편리한 웹 GUI로 제공하는 점도 인공지능 교육에 있어 아주 매력적이었습니다.

SIA
전태균 CEO/설립자

Backend.AI의 GPU 가상화 기술을 활용하여 DGX-1,2와 같은 다중 GPU 장비를 활용하는 대규모 딥러닝 모델 훈련 시 I/O bottleneck 등을 회피하여 GPU의 활용률을 최적화 할 수 있었습니다. 일시적으로 많은 자원이 필요할 때는 복잡한 작업 없이 퍼블릭 클라우드 인스턴스를 즉시 추가하여 하이브리드 클라우드를 구성하여 사용할 수 있어 효율적이었습니다.

Backend.AI Success story

초거대 AI 인프라에 백엔드닷에이아이(Backend.AI)와 NVIDIA DGX 클러스터를 결합한 슈퍼컴퓨터가 수천 개의 시뮬레이션을 실행하고 단시간에 수백만 개의 데이터를 분석하고 있습니다.

구성

- 대륙간, Multi-organization 사용자용 머신러닝 최적화 클러스터 팜 솔루션
- A100 GPU 수백대, 추가 고성능 CPU nodes(데이터 분석용)
- 완전 폐쇄환경에서의 운영을 위한 Backend.AI와 통합된 패키지 저장소인 "Backend.AI Reservoir" 추가 도입으로 완전 Air-gapped 환경을 구축

고객혜택

- 대단위 팜 구성 설계 제공 및 SLA 극대화를 위한 고가용성 구성
- 멀티노드 분산 훈련 지원 기능 및 GPU간 네트워크 기반의 대규모 초고속 딥러닝 훈련
- Backend.AI Reservoir를 통해 PyPI 및 Ubuntu 저장소를 완전 폐쇄망내에서 자유롭게 사용 지원
- 기관 내외부 동시 서비스 시 시스템/데이터 보안을 위한 격리 도메인 구성

대표 고객사

기업

SAMSUNG ELECTRONICS, kt, LG, DSME, LOTTE, CJ

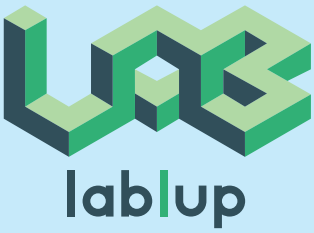
공공 기관 및 연구 기관

한국은행 BANK OF KOREA, 건강보험심사평가원, KISTI, TTA, KAERI, KIST

주요대학

기술 / 플랫폼 파트너

NVIDIA, PURESTORAGE, NetApp, DELL, Microsoft, aws, NAVER Cloud



contact@lablup.com
https://www.backend.ai
https://github.com/lablup/backend.ai