lablup

AI Enterprise
AI Cloud
AI Open Source
AI MLOps

backend.ai

AI

세상 모든 AI를 위한

We'll Get You
Every Last Bit.

성장하는 AI 서비스,
가진 인프라를
충분히 쓰고 계신가요?

쏟아지는 LLM을
자체 인프라에서 빠르게 도입,
운영할 수 있는 방법은 없을까?

AI 모델 학습부터 서비스 배포까지
하나로 이어 가는 방법은?

비싼 GPU가 놀고 있거나,
작업 대기열이 막히는 문제를 어떻게 풀지?

NVDIA, AMD, Intel 등 다양한 AI 반도체를
한 번에 운영할 수 있는 방법은 없을까?

Backend.AI가
여러분의 해결사가
되어 드립니다.

원하는 모델을 추론 API로 띄우고 트래픽에 맞춰
자동 확장, 운영하는 관리형 모델 서빙

학습, 실험, 배포를 한 파이프라인으로
연결하는 MLOps 플랫폼 FastTrack 3

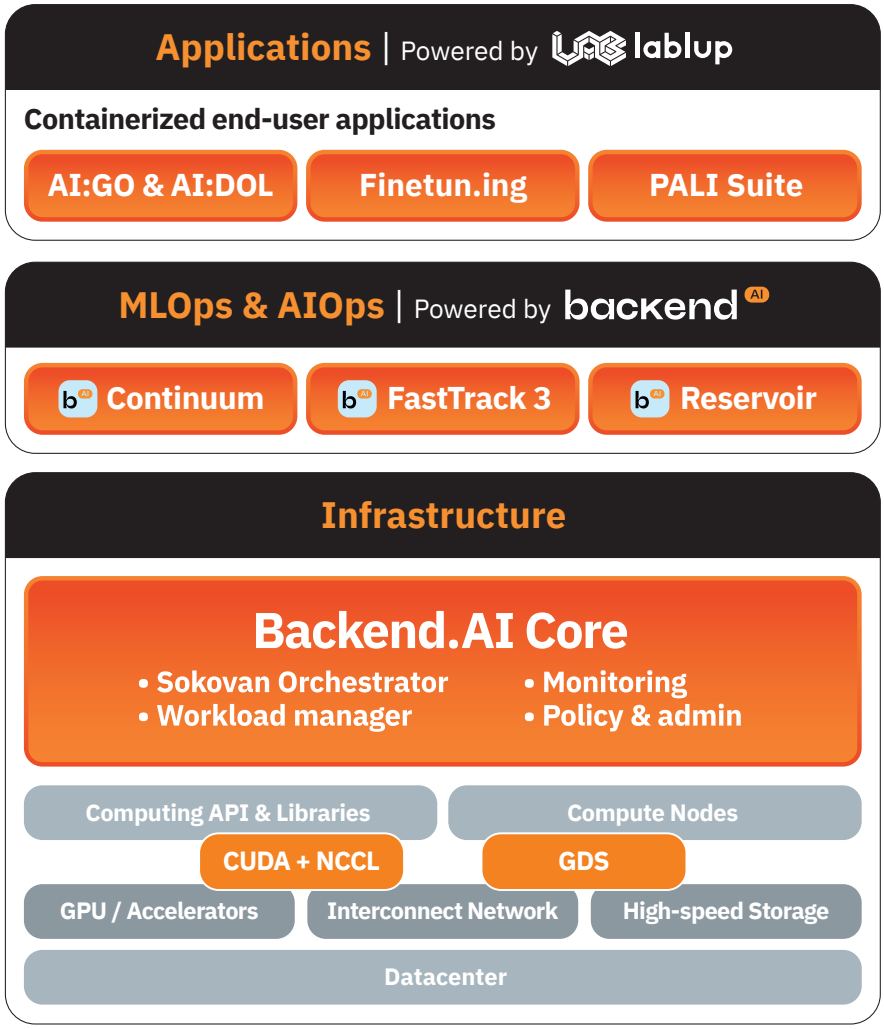
유휴 자원을 자동으로 거둬들이고
밀린 작업은 막힘 없이 배분하는
자원 오케스트레이터 Sokovan

서로 다른 AI 반도체를 한곳에 모아
운영하는 **12+종 이기종 가속기 지원**

backend AI

Project Map

시작부터 끝까지 AI를 위해 설계된
Backend.AI를 만나보세요.



GPU 관리부터 AI 서비스까지 하나로 연결하는 Backend.AI

모든 환경에 대응하는 배포 유연성
에어갭 온프레미스부터 클라우드까지 모든 환경 지원

업계 최고 수준의 하드웨어 호환성
NVIDIA DGX™-Ready 인증 및 11개 이상의 AI 가속기 지원

비교할 수 없는 독자 GPU 가상화 기술
컨테이너 수준 GPU 자원 관리로 효율성 극대화

한계까지 끌어올린 라인레이트급 데이터 처리
RDMA, GPUDirect Storage로 무지연 데이터 전송 구현

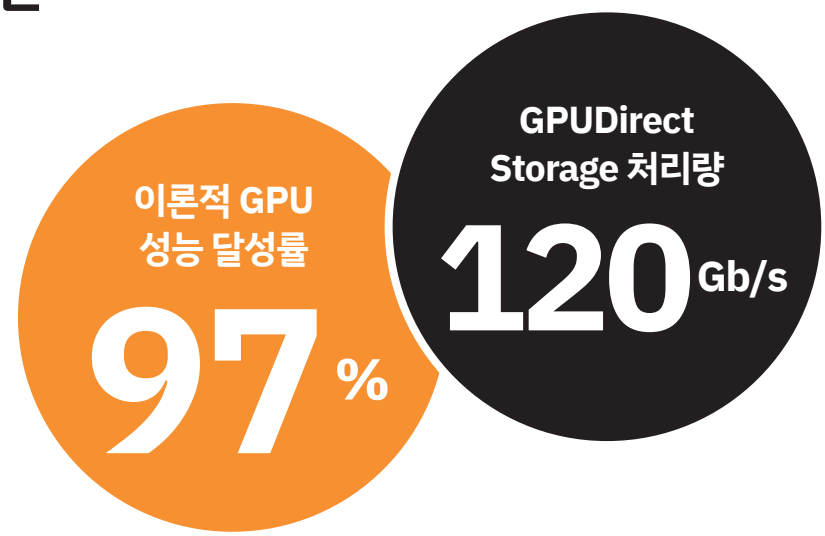
GPU 가동률의 차원을 높이는 오케스트레이션
Sokovan 오케스트레이터의 멀티테넌시와 멀티노드 지원을 통한 대규모 워크로드 운영

AI 애플리케이션을 위해 준비된 소프트웨어

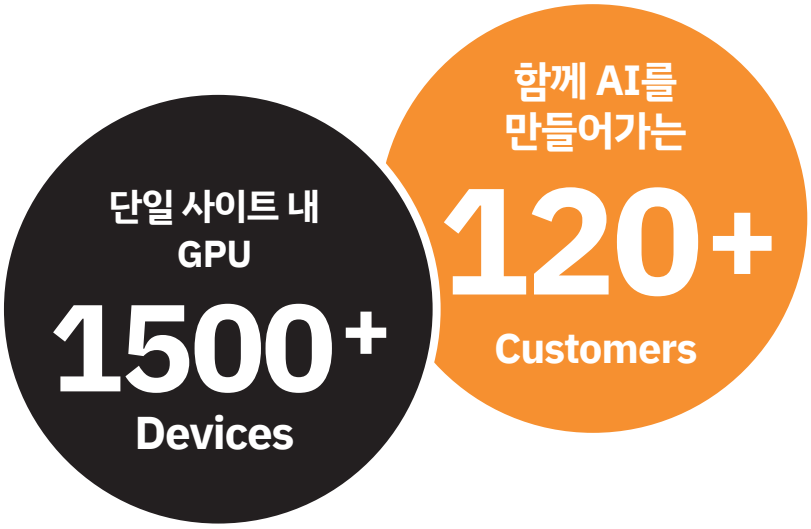
Backend.AI FastTrack 3 | 모델 개발부터 서비스까지 총괄하는 MLOps 플랫폼
AI:DOL | 바로 배포하고 활용할 수 있는 GenAI 옴니미디어 랩
PALI | 추론을 위한 고성능 AI 런처 서비스

기술력으로 업계를 선도하는
Backend.AI

극한의 퍼포먼스를 이끄는
Backend.AI



고객들이 믿고 사용하는
Backend.AI



Backend.AI 특장점

GPU 지원	컨테이너 수준 다중 GPU 할당 및 독자 개발 GPU 분할 가상화 / NVLink NVSwitch 최적화 다중 GPU 플러그인 아키텍처
스케일링	온프레미스 설치(실 서버/가상 서버) / 하이브리드 클라우드 운영(온프레미스+클라우드) 및 다중 클라우드 연동 / 분산 워크로드 실행 시 다중 네트워크 환경 자동 구성
스케줄링	GUI 어드민을 통한 통합 스케줄링 및 모니터링(CLI 지원) / 사용자 및 프로젝트 단위 자원 정책 / 다중 컨테이너 일괄 실행 및 제어 기능 제공 / 가용 슬롯 기반 스케줄링 / 확장 및 사용자화 가능한 배치 스케줄러 / 가상화폐 마이닝 검출 및 차단 / 유휴 자원 자동 수거를 위한 다양한 설정 제공
보안	다중 사용자 지원 / 하이퍼바이저 혹은 컨테이너를 통한 샌드박싱 / 프로그램 가능한 추가 샌드박싱 계층 / 관리자 모니터링 / SAML 2.0, OIDC, 커스텀 IdP 통합 지원
신뢰성	고가용성(HA) 구성 / 연산노드 핫플러깅 및 스케일링
UI/UX	사용자 애플리케이션(Windows 10/11, macOS, Linux Desktop) / 웹 기반 서비스 / 관제 콘솔
데이터 관리	공유 스토리지 기능을 통한 데이터 업&다운로드 및 공유 지원 / EFS, NFS, SMB 및 분산 파일 시스템 사용 / 사용자 및 프로젝트 단위 접근 제어 / 로컬 가속캐시(SSD, 메모리)
개발자 지원	범용프로그래밍 언어 지원(Python, C/C++, R, Java 등 17종) / 통합개발환경 플러그인(VS Code, IntelliJ, PyCharm) 제공 / 대화형 셸, 터미널 / 컨테이너 이미지빌드 GUI
AI 개발자 / 데이터 과학자 지원	웹 콘솔에서 JupyterLab, TensorBoard, VS Code 등 다양한 GUI 기반 개발도구 지원 / NGC(엔비디아 GPU 클라우드) 플랫폼 통합 / 주요 머신러닝 라이브러리 지원 : TensorFlow, PyTorch, MXNet, Jax 등 / 다양한 버전의 라이브러리 동시 사용 / 머신러닝 개발환경 자동 업데이트 / 서버리스 딥러닝 모델 서빙 / 사용자 작성 모델 서빙 및 버전 관리
폐쇄망 지원	Backend.AI Reservoir 를 통한 자체 패키지 저장소(PyPI, CRAN 및 Ubuntu 저장소) / 스토리지 프록시 기반의 스토리지 가속 플러그인 지원(Everpure, NetApp, Dell PowerScale, WEKA, VAST Data, CephFS, LustreFS, GPFS)
관리 및 제어	시스템 관리자 전용 대시보드 / 관리자 전용 컨트롤 패널 / 연산노드 설정 제어 / 연산노드 시스템 설정 변경 / 시스템 통계 수집 / 모니터링 솔루션 연동 / 무중단 노드 관리

backend.ai

AI

여러 벤더의 가속기도 하나로, 매끄럽게

Arm과 x86-64를 모두 지원하는 AI 인프라 플랫폼

NVIDIA · intel. · GRAPHCORE · rebellions_ · FURIOSA · tenstorrent · AMD · groq · sambanova · HYPER ACCEL

설계부터 고려된 고성능 스토리지 지원

I/O 병목 최소화를 통해 경험하는 전례 없는 속도

DELL Technologies · NetApp · VAST · WEKA · Everpure formerly Pure Storage · ddn · IBM Spectrum Scale

클라우드 고객을 위한 최적의 선택

자체 개발하는 것보다 구축 시간 최대 95% 단축

aws · kt cloud · NHN CLOUD · NAVER CLOUD PLATFORM · VULTR

다양한 산업 분야에서 사용하는 Backend.AI

- **NVIDIA DGX™-지원 소프트웨어**
- **클러스터단 GPUDirect Storage 세계 최초 구현 및 제공**

SAMSUNG · kt · LG · HYUNDAI MOBIS · 42dot · CJ · LIG Defense&Aerospace · 신한은행 · 한국은행 BANK OF KOREA · KOREAN AIR · 건강보험심사평가원 · SMC 삼성서울병원 · 충남대학교병원 · 대한민국해군 REPUBLIC OF KOREA NAVY · SAMSUNG 삼성웰스토리 · ETRI 한국전자통신연구원 · AICA · KISTI 한국과학기술정보연구원 · 경기도 · (재)경남테크노파크 · USC Advanced Research Computing · KMU · SUNGKYUNKWAN UNIVERSITY

각 기업 및 기관의 로고는 정보 제공의 목적으로만 사용됩니다.

AI Enterprise

AI Cloud

AI Open Source

AI MLOps

lablup

contact@lablup.com

https://www.backend.ai

https://github.com/lablup/backend.ai