lablup

Backend.AI Partner Intro Session

복잡한 GPU 운영을 단순화하는
AI 인프라 운영 플랫폼

April 2026

lablup

창립일	2015.04
대표자	신 정 규
직원 수	46

제품 / 서비스

- **Backend.AI**: **온프레미스** 및 **하이브리드 클라우드 환경**에서 **AI 개발, 서비스, 슈퍼컴퓨팅을 관리**할 수 있는 엔터프라이즈 AI 운영체제(AI OS)
- **Backend.AI FastTrack 3**: 작업 파이프라인을 자동화, 통합, 관리할 수 있도록 하여 AI 개발을 간소화하며, 원활한 작업을 위해 재사용 가능한 템플릿을 제공하는
- **Backend.AI:DOL:** 복잡하고 파편화된 LLM API를 제어 가능한 게이트웨이로 통합하는 Continuum Router를 기반으로 하는 오픈 모델 기반 GenAI 서비스 환경
- **Backend.AI:GO:** 추론을 위해 클라우드에 연결할 필요 없이 로컬 PC에서 sLM부터 LLM까지 실행하는 데스크톱 애플리케이션

핵심 기술

- **Container-level GPU Virtualization** (컨테이너 수준 GPU 가상화) 기술
- 분산형 딥러닝 모델 학습 및 **초대형 모델 개발** 플랫폼
- GPU를 기반으로 하는 **저비용 / 초저지연 AI 모델 서비스**
- AI 오퍼레이션 파이프라인을 통한 **다양한 종류의 AI 가속기/반도체 지원**
- **포터블한 하이퍼스케일 AI 플랫폼**

사업 개요

- 대한민국 내 가장 많은 수의 GPU를 관리하고 있음
- **NVIDIA® DGX™ 지원 소프트웨어 인증 획득**
- 대기업, 금융 기관, 의료 기관, 국내외 연구 기관 및 대학교에 이르기까지 약 110여개 이상의 고객사 및 파트너와 함께 AI 혁신을 위한 협력관계 구축

플랫폼 파트너

NVIDIA® DGX™ 지원 소프트웨어 파트너

DELL AI for Telecom Certification Program Partner

Tech. Alliance Partner

고객사

110+ and Growing!

backend AI

래블업에 대하여

래블업은 AI를 쉽게 운영하고 확장할 수 있는 환경을 만들고 있습니다.

창립일 **2015.04** | 대표자 **신정규** | 직원수 **48**

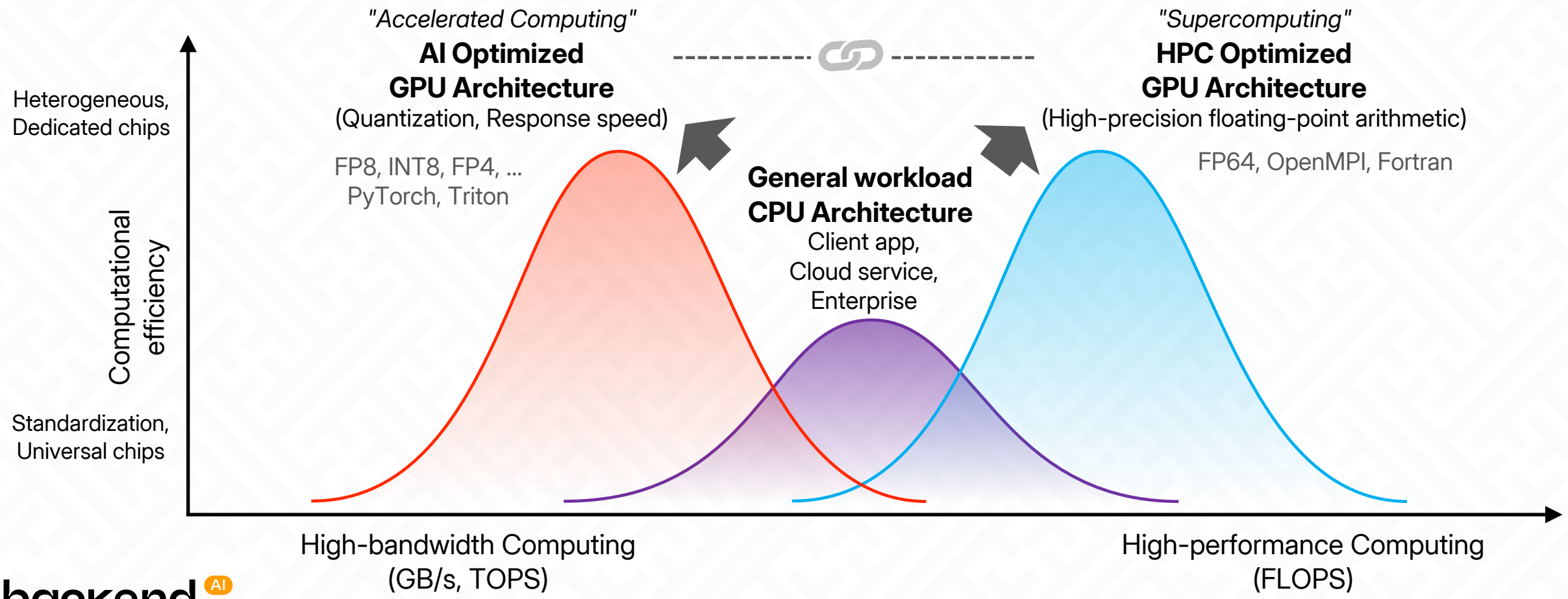
NVIDIA® DGX™-지원 소프트웨어 Backend.AI 개발

16,000대 이상의 GPU 운영 및 관리

110개 이상의 엔터프라이즈 고객

backend AI

HPC에서 AI로, 범용 컴퓨팅에서 가속 컴퓨팅으로



컴퓨팅의 발전 방향

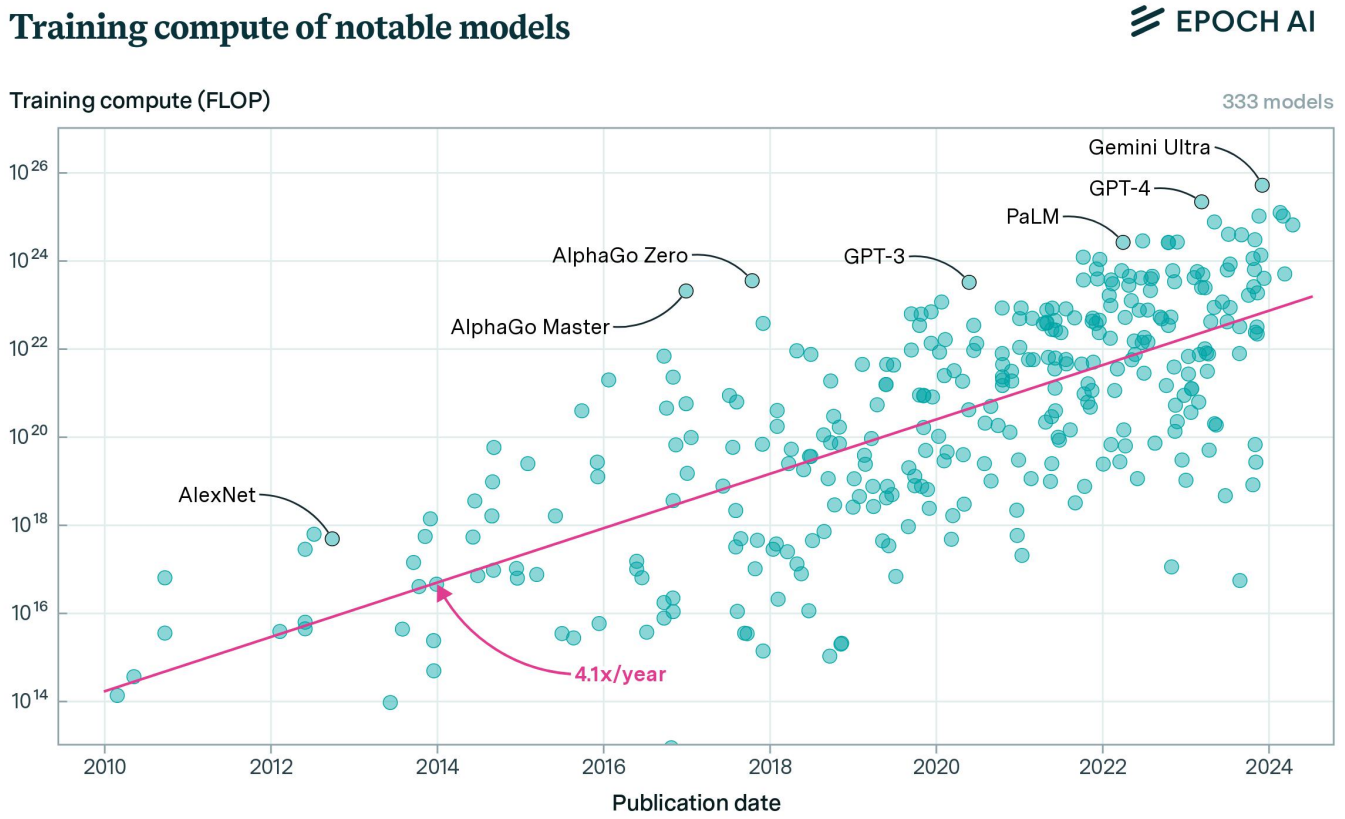
AI 서비스를 움직이는 숨은 뼈대,
지능형 미래를 만드는 인프라의 능력

[1] RSC of Meta

증가하는 컴퓨팅 수요

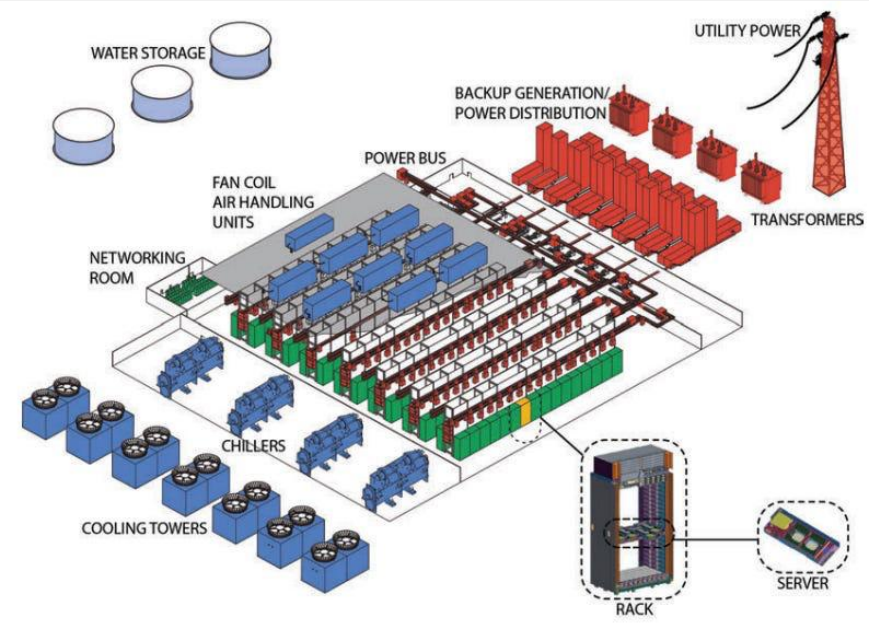
4x/ 연간

(Approx.)



지능 인프라를 향한 도전

성장 속도를 따라잡지 못하는 HW·SW 인프라



[1] https://www.construction-physics.com/p/how-to-build-an-ai-data-center

상호 연결 병목
Interconnect Bottleneck

내부 지연으로 인한
전체 시스템 속도 저하

냉각 인프라
Cooling Infrastructure

냉각이 부족한 경우
성능 저하 발생

활용 효율성
Utilization Efficiency

AI 리소스 이용에 따르는
활용 부족과 자원 불균형

스케줄러 확장성
Scheduler Scalability

복잡한 작업 부하에 필요한
확장 가능한 스케줄링

가시성 / 텔레메트리
Observability / Telemetry

적절한 시스템 모니터링
및 진단의 중요성

데이터 / 모델 보안
Data / Model Security

진화하는 보안 위협과
점점 복잡해지는 시스템

AI/HPC 워크로드 운영에서 발생하는 실질적인 문제들

비용 문제

- 값비싼 GPU 하드웨어
- GPU의 엄청난 데이터 소비량
- AI 데이터 소비 속도의 병목

전력 문제

- 랙당 4kW to 250kW까지 요구되는 어마어마한 전력량
- 기존에 존재하지 않았던 전력 소비 패턴의 등장
- 단위 면적 당 유례 없는 전력 집중

복잡도 문제

- 매일 출시되는 새로운 AI 모델들 (HuggingFace에 100만개 이상의 AI 모델 등록됨)
- 2주 단위로 업데이트되는 새로운 소프트웨어들 (TensorFlow, PyTorch, vLLM, TensorRT-LLM...)
- 빠르게 성장하는 AI 가속기들 (NVIDIA CUDA, AMD ROCm, ATOM+, RNGD, Graphcore IPU, Intel Gaudi...)

래블업의 솔루션

backend AI

복잡한 GPU 운영을 단순화하는
AI 인프라 운영 플랫폼

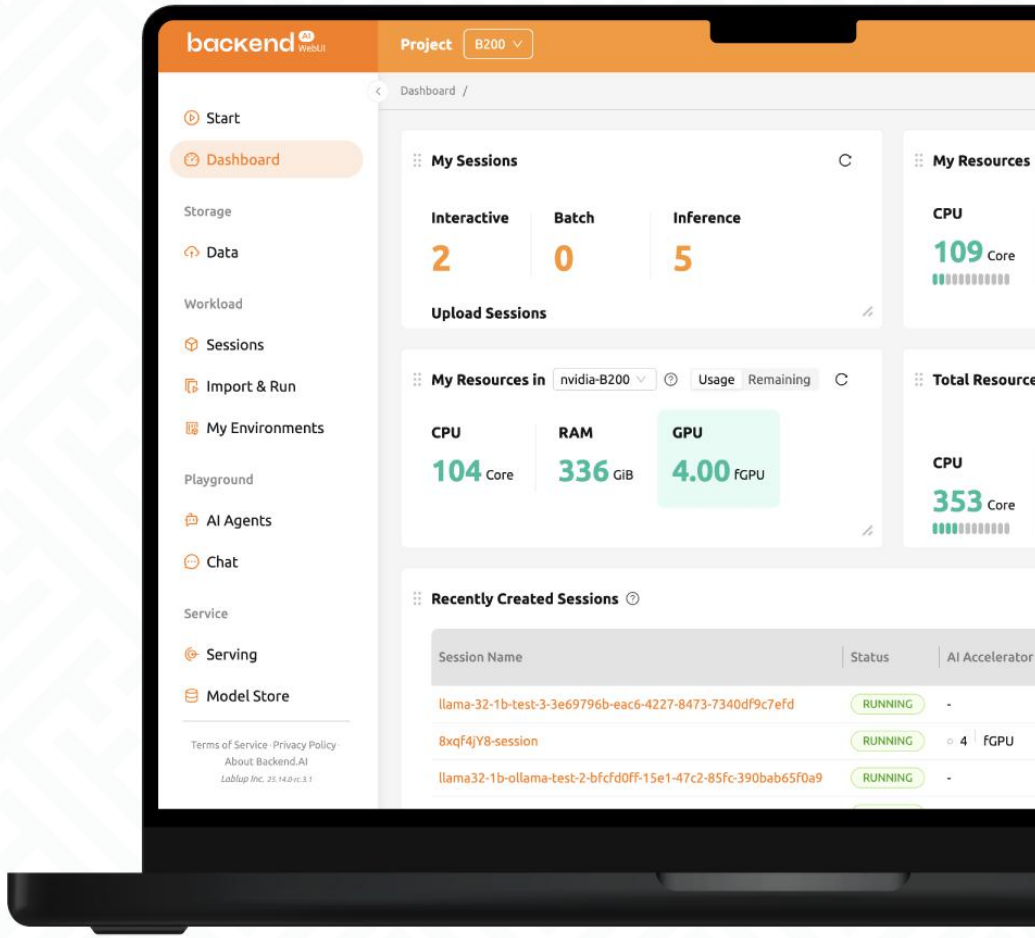
원하는 방식의 GPU 효율화 실현
컨테이너 수준 GPU 분할가상화 지원

데이터 및 스토리지 가속
RDMA / NVIDIA GPUDirect Storage 지원

이종 아키텍처 및 이기종 AI 가속기 하드웨어 지원
x86-64 / Arm64 환경 지원

모든 종류의 환경에서 안정적인 배포 지원
에어갭, 온프레미스, 클라우드, 하이브리드 환경에서도 운영 가능

대규모 AI 워크로드 오케스트레이션
Sokovan™ 오케스트레이터로 **멀티 테넌트, 멀티 노드** 워크로드 관리



backend AI

래블업의 솔루션

엔터프라이즈 GPU 가동률을 새로운 차원으로 끌어올리는 Backend.AI

- **가장 효과적인 방법으로** 고급 GPU 리소스를 사용하는 엔지니어링
- **병목 현상 최소화**를 통한 베어 메탈 수준의 고성능 달성

Performance Overview

<3%
이론상 GPU 오버헤드
Theoretical GPU Overheads
베어메탈 수준의 성능을 발휘하도록 병목 현상 최소화[1]

4x
가동률 향상
Utilization Enhancement
정밀하고 신속하게 AI 작업을 오케스트레이션하는 시스템[2]

110%
파이프라인 퍼포먼스
Pipeline Performance
개발 과정을 효율적으로 관리하기 위한 AI 파이프라인 최적화[3]

120Gb/s
NVIDIA GPUDirect Storage 처리량
스토리지와 GPU 메모리 사이 직접 경로 활성화[4]

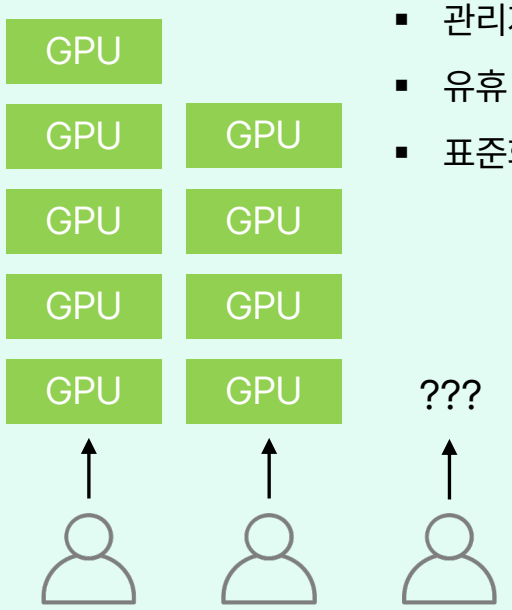
backend AI

[1] 대규모 언어 모델 트레이닝 성능 벤치마크, Deepspeed에서 7.5B GPT-2 모델을 훈련할 때 달성한 FLOPS 기준 (2022.11)
[2] BMT를 위한 18개 AI 워크로드 벤치마크 평균 기준. (2023.10)
[3] 데이터 전처리 워크로드, GPU 기반 학습 워크로드, 레퍼런스 테스트 세트를 모두 합쳐 실행한 전체 실행시간 기반 벤치마크
[4] Magnum IO에 기반한 GPUDirect Storage 1MB 블록 쓰기 기준, Weka 13 Node Cluster Configuration 적용 (2022.12)

lablup

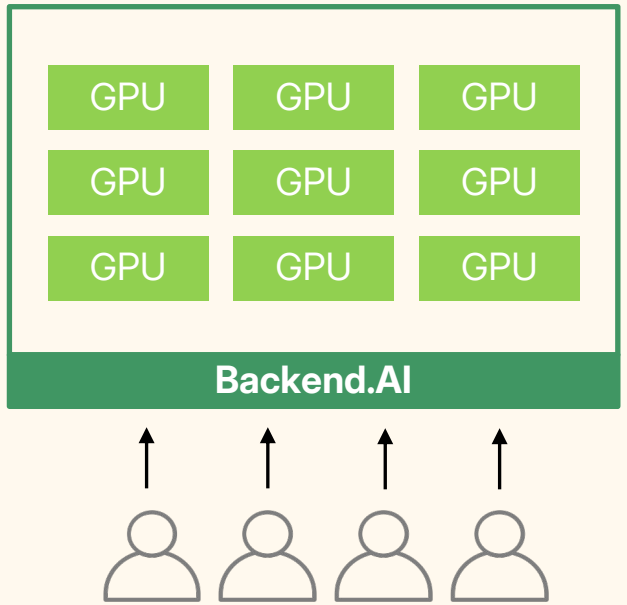
AI에 의한, AI를 위한 플랫폼.
엔터프라이즈 GPU의 관리 혁신

기존 관리 방법



- 관리자에 의한 사용자별 GPU 수동 할당
- 유휴 자원 혹은 부족 자원의 발생
- 표준화되지 않아 어려운 SW 유지관리

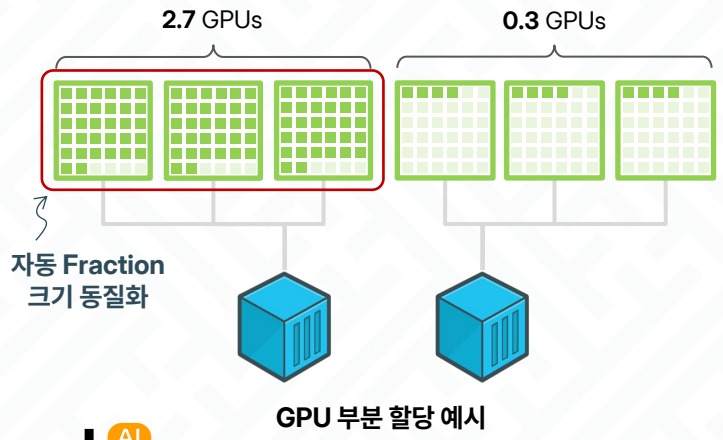
Backend.AI 방식의 관리 방법



- Backend.AI에 의한 GPU 통합 관리
- GPU를 동적으로 할당, 사용자는 필요한 만큼만 사용
- 컨테이너 기반으로 표준화된 SW 생태계 구축

GPU 분할가상화 (Fractional GPU) 기반 유연한 자원 관리

- **컨테이너 기반 GPU 스케일링**
  - 컨테이너별 CUDA SMP, GPU RAM 분배
  - 단일 GPU 공유: 교육, 추론 워크로드에 적합
  - 다중 GPU 할당: 학습 등 대규모 워크로드에 적합
  - 자체 개발 CUDA 가상화 계층으로 구현 (한/미/일 특허)



GPU 부분 할당 예시

- **고객 혜택**
  - 고가의 하드웨어인 GPU의 사용률 향상을 통한 **장비 구입 비용 감소 효과**
  - 기존 학습용 GPU를 분할하여 추론용 GPU로 운영, **노후 장비 활용성 극대화**
  - GPU 생애주기를 총괄하는 효율적인 장비 운영

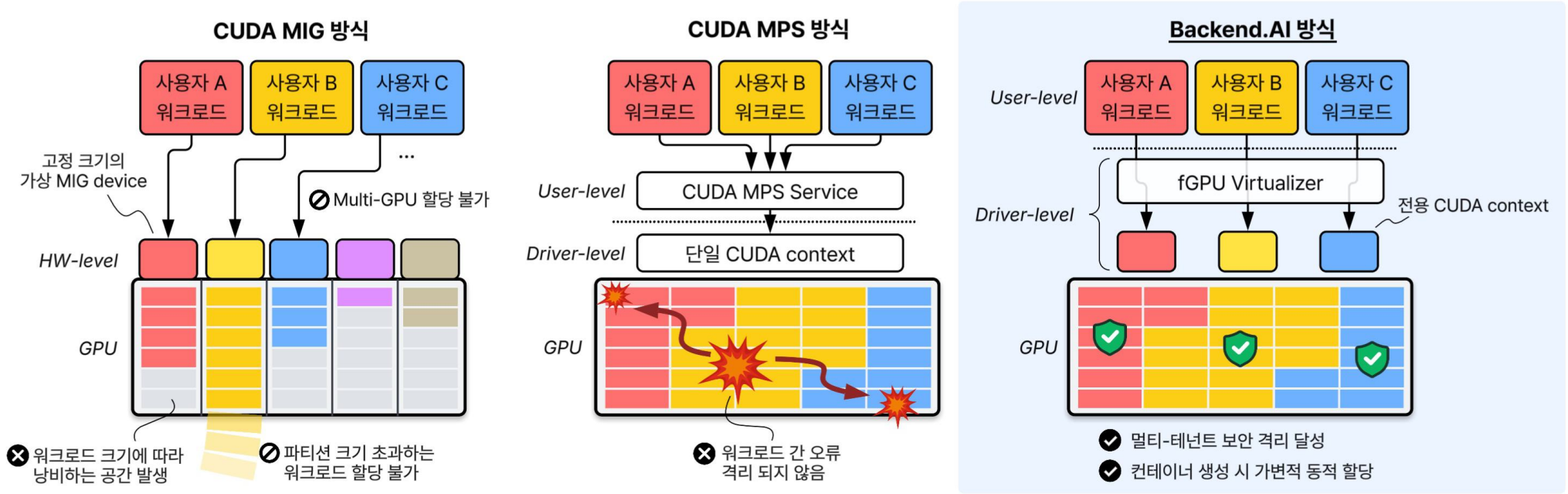
유사 기술과의 핵심 차별점

- **폭넓은 CUDA 버전의 GPU 모델 호환** (CUDA 8~13 / 데스크톱, 워크스테이션, 데이터센터)
- 사용자 프로그램 코드 변경 불필요
- 딥러닝 프레임워크 수정, 리빌드 불필요 (TensorFlow/PyTorch 외 임의 GPU 가속 워크로드 실행 가능)
- 여러 GPU에서 할당된 Fraction을 하나의 컨테이너에 연결하는 멀티GPU 워크로드 지원
- Multi-GPU 환경 Fraction의 크기 동질화 기술 적용

lablup

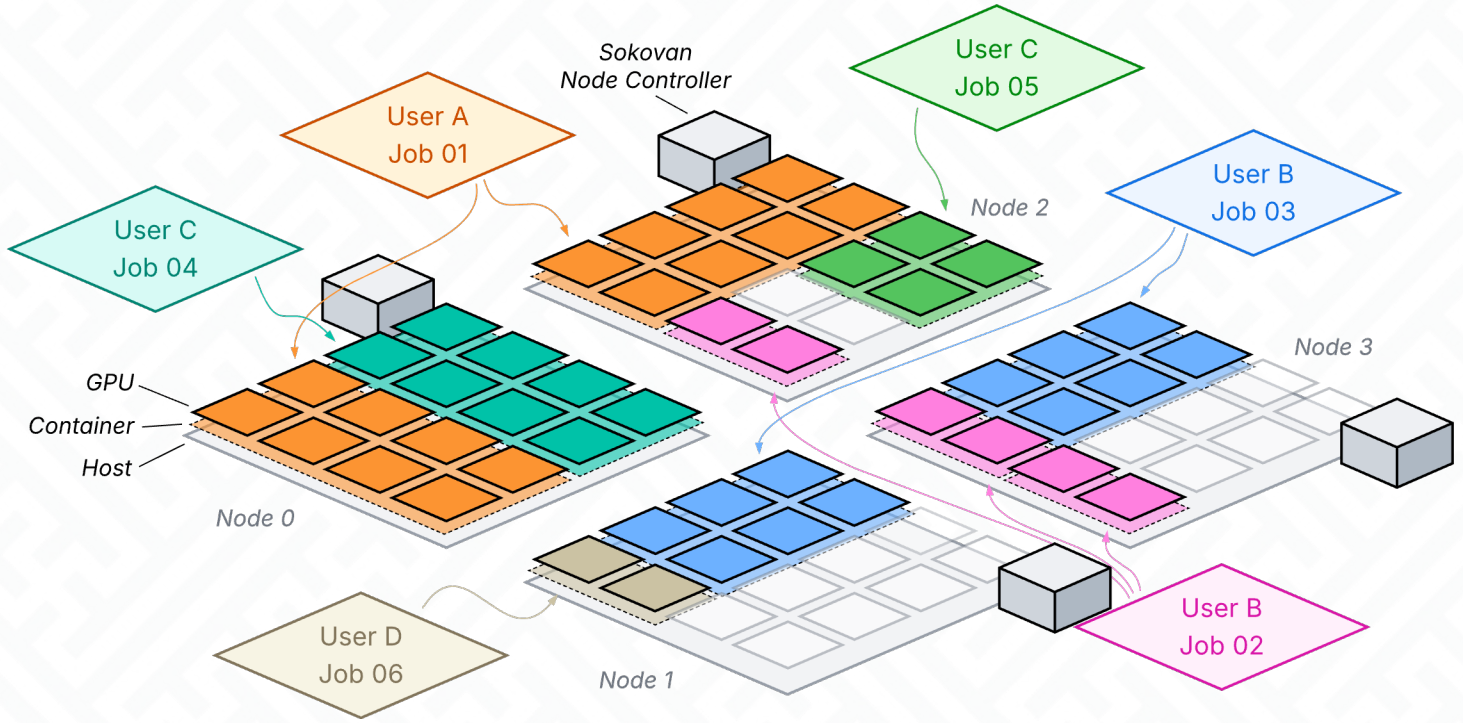
유연성과 멀티테넌시 지원 특화 GPU 가상화 제공

- Backend.AI fGPU 가상화 적용 시 GPU 최적의 공유 성능 특성 및 오버헤드 가능
- CUDA context 분리로 CUDA kernel 수준의 오류 전파 격리



backend AI

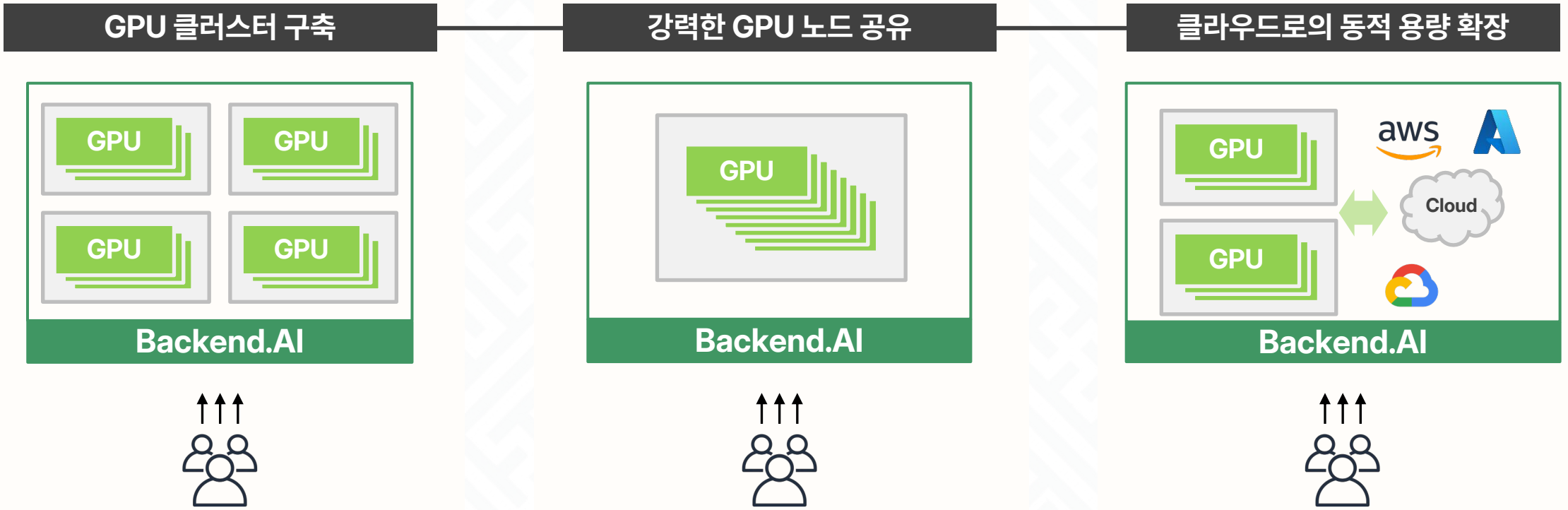
멀티테넌시 네이티브 구조, 멀티노드 확장 지원



- **Sokovan 오케스트레이션 엔진**
  - 컨테이너 환경에 최적화된 독자 GPU 분할가상화 기술(fGPU) 및 MIG 지원
- **멀티테넌시 네이티브 구조**
  - 사용자별, 프로젝트별 리소스 할당량 및 역할 기반 접근 제어(RBAC)
  - 다수 사용자가 노드를 공유하여 자원 통합 효율성 향상
- **멀티 노드 확장 지원**
  - 하나의 작업을 여러 GPU, 여러 컨테이너, 여러 노드에 걸쳐 실행
  - 자동 오버레이 및 RDMA 네트워크 구성
  - 자동 크로스 컨테이너 SSH 인증

lablup

인프라를 구축, 공유, 확장하기 위해 설계된 Backend.AI



lablup

최대 성능, 최대 효율.
기술의 최전선에서 키워드로 살펴보는 Backend.AI

AI 인프라 CAPEX 효율성 제고 (성능 극대화 측면)

- 멀티 레벨 스케줄러
- NUMA-고려 자원 매핑
- 이종 에이전트 백엔드
- 자원 그룹 및 네임 스페이싱
- 멀티 노드, 멀티 클러스터 클러스터링
- I/O 가속 플레인

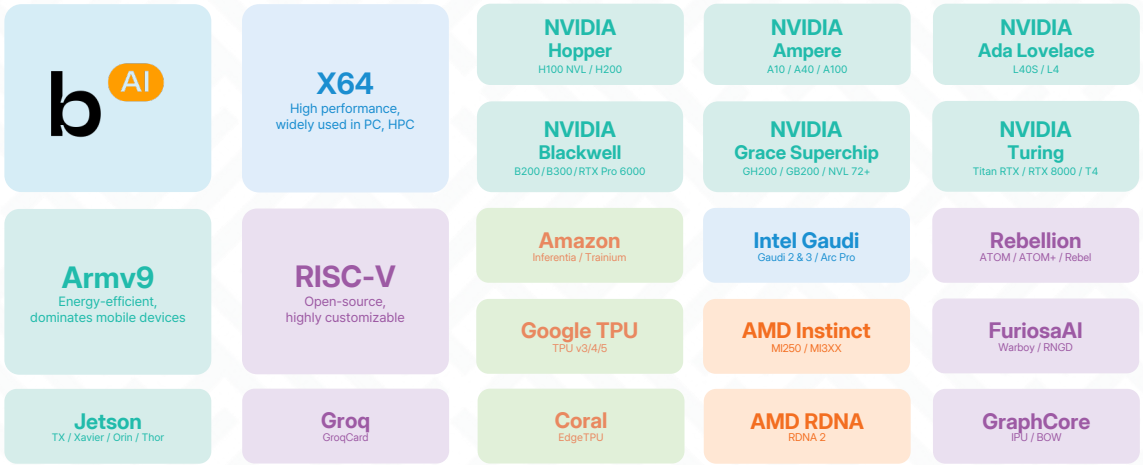
AI 인프라 OPEX 최적화 (전력 소비 절감 측면)

- 컨테이너 수준 GPU 분할가상화
- 동적 자원 할당
- GPU/NPU 추상화
- MEC / IoT를 위한 마이크로 Backend.AI
- 멀티 아키텍처 & Arm 지원
- 전력 소비 기반 스케줄링

lablup

11종류 이상의 AI 가속기를 아우르는 광범위한 지원

- 다양한 AI 반도체 지원을 통한 AI 워크로드의 효율성 극대화
  - NVIDIA, Intel, Graphcore 등 다양한 AI 반도체 제품(GPU, NPU, TPU 등) 지원
  - NVIDIA Jetson, Google Coral 등 AI IoT 가속기 지원
  - 워크로드 특성에 맞춘 전력 대 성능비 향상
- Arm 및 x86 상호 운영 가능한 멀티 아키텍처 플랫폼
  - NVIDIA Grace™ Hopper™ 및 Grace™ Blackwell을 완벽하게 지원
  - 다양한 환경에서의 유연성과 최적의 성능을 보장하는 멀티 아키텍처 대응 설계



AI Accelerators Works with Backend.AI

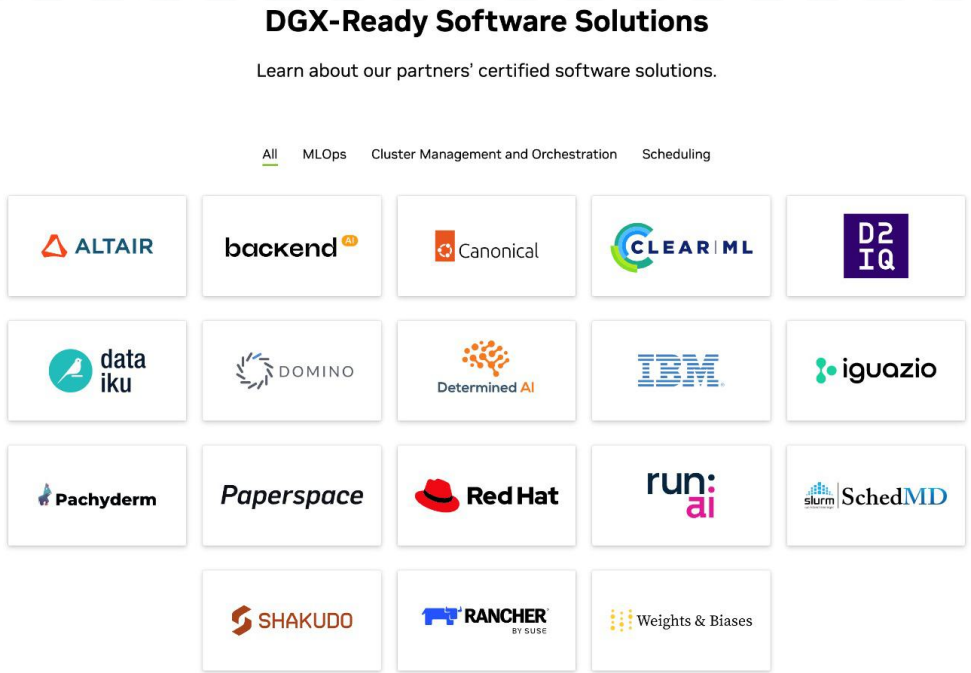
NVIDIA intel GRAPHCORE rebellions_ FURIOSA tenstorrent Google AMD

가장 최신의 NVIDIA 플랫폼까지 폭넓게 지원

NVIDIA Hardware Platform	Operating System	CPU Architecture
DGX (B300, B200, H200)	DGX OS	x86-64
DGX (GB300, GB200, GH200)	DGX OS	Arm64
HGX	Ubuntu / RHEL-compatible	x86-64
DGX Spark (GB10)	(Ubuntu-based)	Arm64
Jetson (NX, Xavier, Orin)	JetPack (Ubuntu-based)	Arm64

Backend.AI + DGX™: Backend.AI DGX™ 클러스터

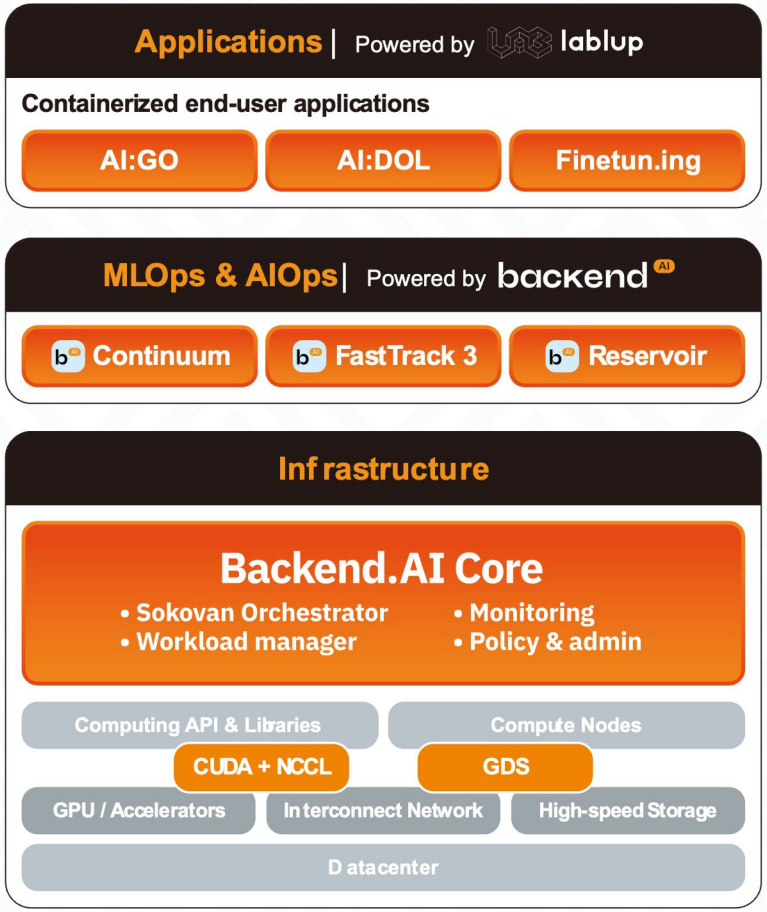
- **NVIDIA DGX™ series: From DGX™ station to DGX™ B300/B200**
  - 단일 노드로 가장 강력한 다중 GPU 계산 시스템
    - ✓ Ubuntu / RHEL 기반 운영체제
    - ✓ NVLink / NVSwitch를 통한 고속 GPU-GPU 네트워킹
    - ✓ 다양한 크기의 GPU 워크로드 실행
- **Backend.AI와 DGX™ 패밀리의 통합**
  - NVIDIA가 제공하는 컨테이너 런타임에서 부족한 기능 제공
    - ✓ 다중 사용자를 위한 GPU 공유 및 할당
    - ✓ 기계학습 파이프라인을 위한 기본 구성요소 제공
    - ✓ CPU/GPU 토폴로지를 고려한 스케줄링



https://www.nvidia.com/en-us/data-center/dgx-ready-software/

AI 인프라 수직 통합

인프라부터 서비스까지 Backend.AI로 완전 정복



- **Applications**
  - AI:GO 및 AI:DOL을 통해 로컬 환경에서 클라우드까지 아우르는 다양한 레이어 제공
  - 소규모 서비스부터 대규모 서비스까지 유연한 확장성
- **MLOps (FastTrack 3)**
  - 전처리, 학습, 추론, 검증, 배포까지의 모든 단계를 자동화 및 하나의 플랫폼으로 통합
  - GUI와 CLI 듀얼 모드 인터페이스로 모든 MLOps 단계의 유연한 지원
- **DevOps (Reservoir)**
  - 폐쇄망 환경 및 프라이빗 클라우드를 위한 로컬 패키지 저장소
  - 내부 네트워크 환경에서 오픈소스 패키지 및 운영체제의 안전한 제공
- **Infrastrucure (Backend.AI)**
  - 컨테이너 수준 GPU 분할가상화 및 다중 사용자 동시 리소스 활용 설계 (Multi-node, Multi-tenant)
  - 온프레미스, 클라우드, 하이브리드 및 폐쇄망에서의 일관된 운영 경험 제공
  - 이기종 AI 가속기, 고성능(초고속) 스토리지, 네트워크 스택의 효율적 통합 관리

제품 포트폴리오

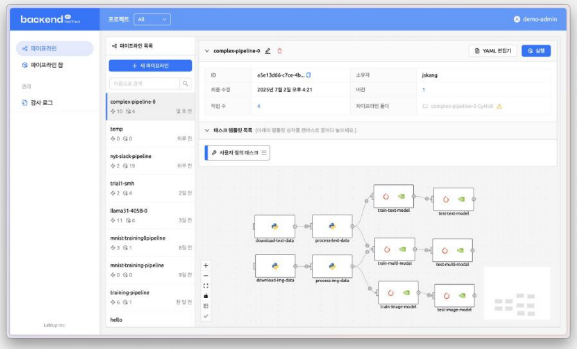
lablup

더욱 저렴하고, 편리한 AI 서비스를 위한 래블업의 레이어 확장

FastTrack 3

노드 단위 커스터마이징이 가능한
자동화된 ML 워크플로우 플랫폼

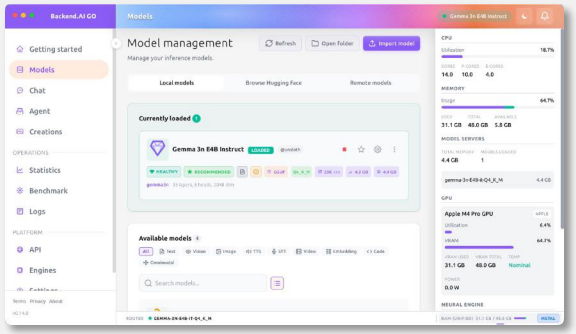
- 멀티노드 자동 설정으로 거대 모델 학습 용이
- 간편하게 배치 작업 파이프라인 구축
- 재사용성이 높은 파이프라인 템플릿 제공



AI:GO

sLM부터 LLM까지 로컬 PC에서
실행하는 데스크톱 애플리케이션

- 폭넓은 사용자 AI 가속 하드웨어를 지원
- 여러 대의 Go를 연결하여 사용하거나, Backend.AI와 연결하여 사용



Reservoir

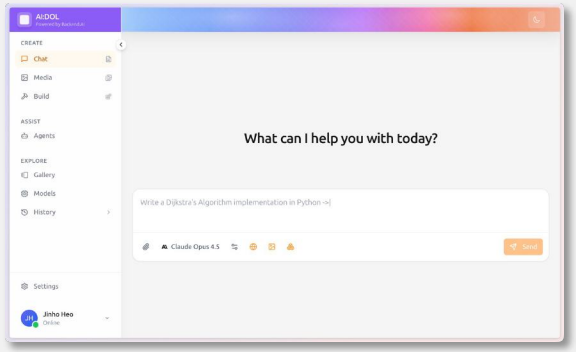
에어갭 엔터프라이즈 환경을 위한
자체 패키지 저장소

- 프라이빗 클라우드 및 폐쇄망에 최신 개발환경 제공
- 래블업 저장소 허브 서비스와 로컬 패키지 서버로 구성되어 보안 검사 거친 패키지만 제공

AI:DOL (Beta)

AI 네이티브 시대를 위한
아티팩트 생성 플랫폼

- 오픈모델 기반 GenAI 경험 제공
  (챗봇/Agent 개발, 이미지 생성, 코딩 등)
- 고객 시스템과 클라우드 AI의 통합



backend AI

FastTrack 3

노드 단위 커스터마이징이 가능한, 자동화된 ML 워크플로우 플랫폼

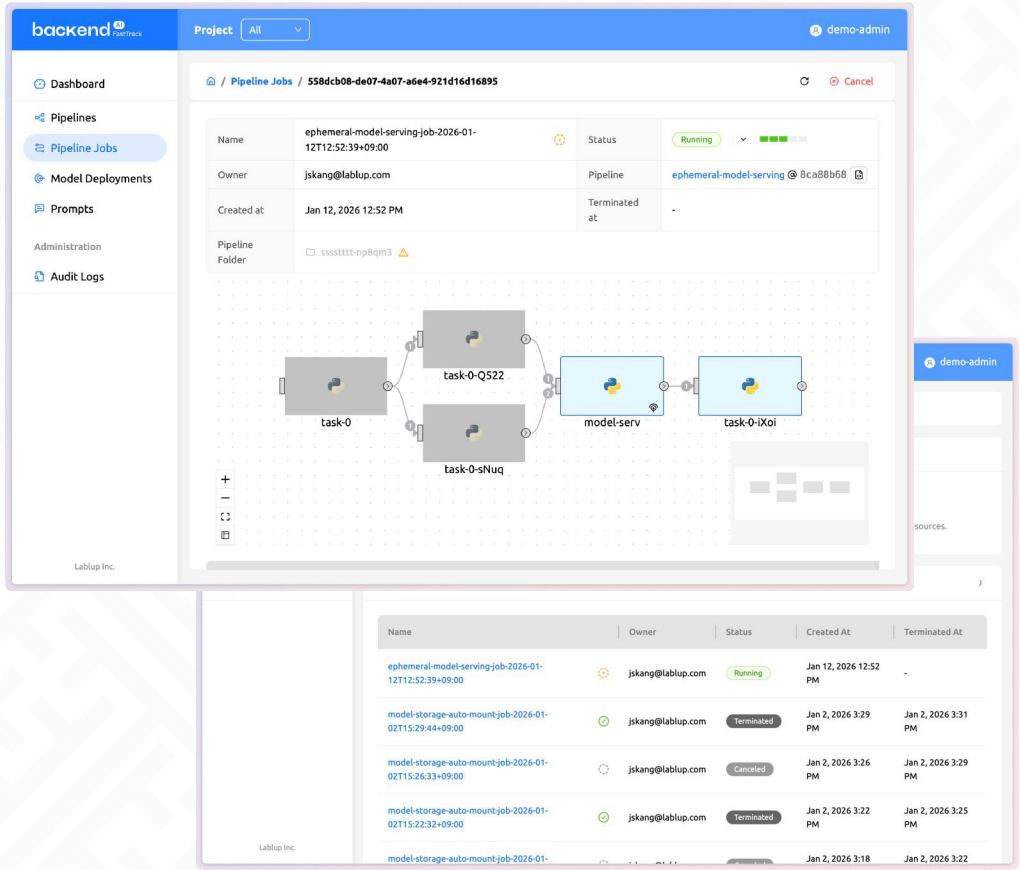
직관적인 UI로 쉽게 구축하는 Batch Job 파이프라인
마우스 드래그와 클릭만으로 파이프라인을 구성하고, 각각의 태스크를 눌러 자원 할당, 의존성 편집

대규모 모델 학습을 위한 자동화된 다중 노드 환경 설정
손쉬운 다중 노드 설정을 통해 대규모 모델 학습도 어려움 없이, 최대한의 리소스를 활용

데이터 준비부터 배포까지 하나의 파이프라인으로 구성
데이터 전처리, 학습, 검증, 모델 서빙까지 하나로 연결된 엔드투엔드 MLOps 플랫폼

환경 변수 맞춤 설정을 통한 정밀 제어 제공
공통 변수, 작업별 변수, 그리고 일회성 변수까지 다양한 형태의 환경변수로 실험 환경 정밀 조정

파이프라인 템플릿을 통해 높은 재사용성 확보
자주 사용하는 템플릿으로부터 시작하고, 예제로부터 불러와 필요한 파이프라인을 구성



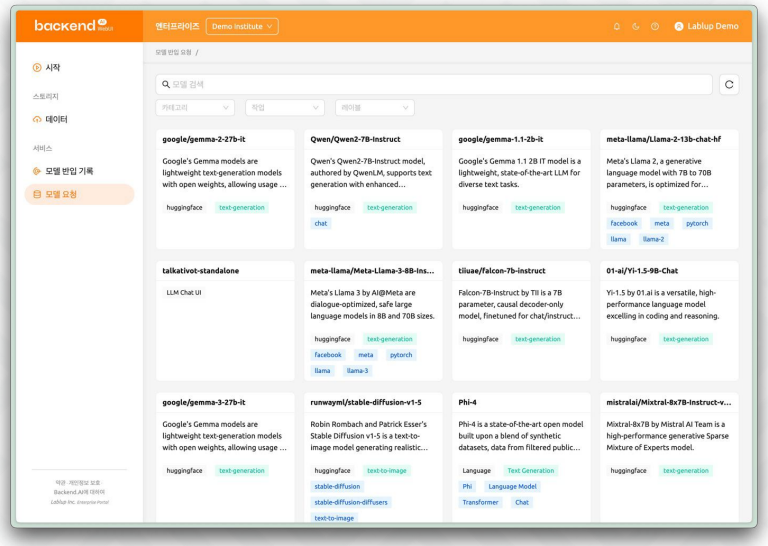
backend AI

제품 포트폴리오

lablup

Reservoir & Reservoir AI

Backend.AI 용의 로컬 패키지 / 모델 저장소 서비스



프라이빗 클라우드, 폐쇄 환경, 오프라인 사이트에 최적 솔루션
Lablup 저장소 허브 서비스와 로컬 패키지 서버로 구성

모델 반입 시 Backend.AI 클러스터의 Model Store에 등록
필요한 경우 해당 모델을 원활하게 실행할 수 있는 runtime 대응 컨테이너 이미지도 포함

Reservoir AI
AI 모델 및 런타임용 컨테이너 이미지

- ✓ Hugging Face
- ✓ PALI Model Store

Reservoir
프로그래밍 언어별 라이브러리

- ✓ PyPI (Python Package Index)
- ✓ CRAN®

Reservoir
리눅스 배포판 패키지

- ✓ Ubuntu (20.04, 22.04, 24.04)
- ✓ DGX OS (5, 6) / AlmaLinux 9

backend AI

lablup

AI:GO Generative On-device

sLM부터 LLM까지 로컬 PC에서 실행하는 데스크톱 애플리케이션

사용자 프라이버시를 우선하는 설계
로컬 추론을 위해 클라우드에 연결할 필요 없이 사용자 PC 안에서 모든 데이터 관리

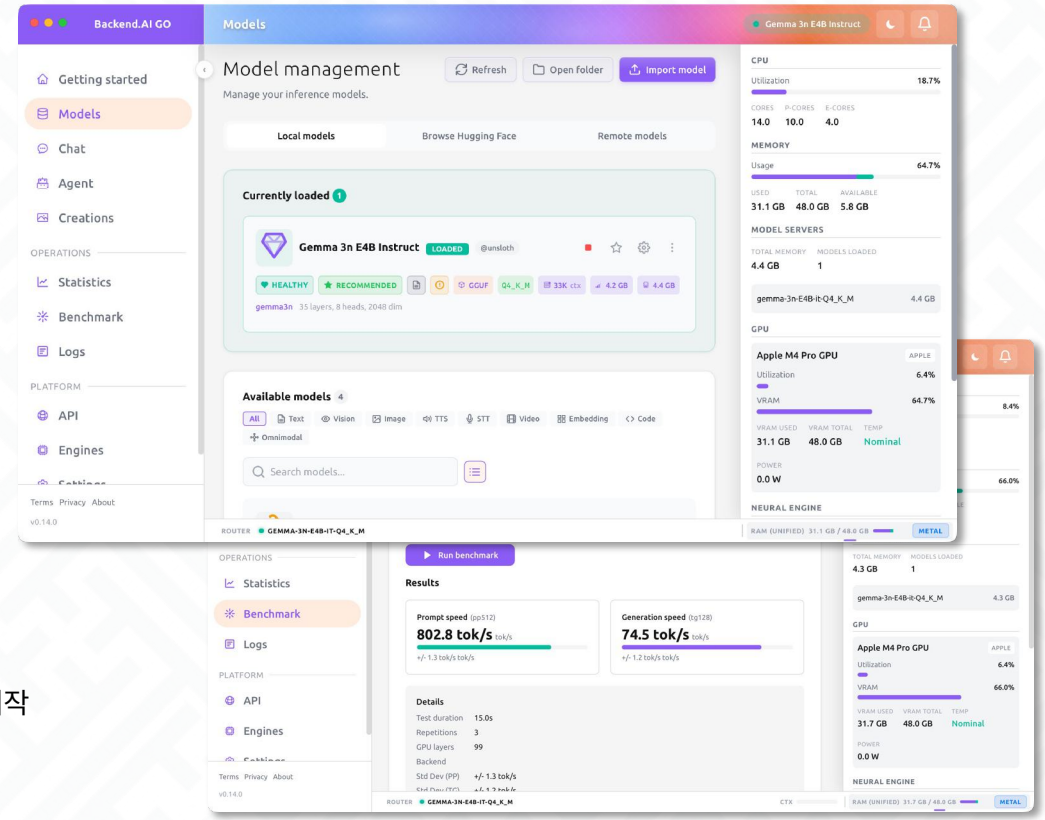
사용자 하드웨어에 최적화된 기반 엔진
macOS (Metal/MLX), Windows & Linux (CUDA, ROCm, oneAPI) 및 표준 CPU에서 높은 성능 발휘

오픈소스를 우선하는 애플리케이션
오픈소스 애플리케이션인 AI:GO 내부에서 Hugging Face의 모델을 직접 검색하고 다운로드

표준 API를 통해 널리 확장할 수 있는 백엔드 기반
OpenAI 호환 API를 통해 즐겨 사용하는 AI 도구의 백엔드로 AI:GO를 활용

여러 대의 Backend.AI:GO를 연결하여 사용
고사양 PC에서 대형 모델을 구동하고, Mesh 네트워크를 통해 Edge 디바이스에서 연결해서 사용

Backend.AI와 함께, Backend.AI:GO로 더욱 넓게
Backend.AI 클러스터에 이미 만들어져 있는 인퍼런스를 사용하거나, Backend.AI에서 새로운 인퍼런스 세션 시작



backend AI

lablup

AI:DOL Deployable Omnimedia Lab

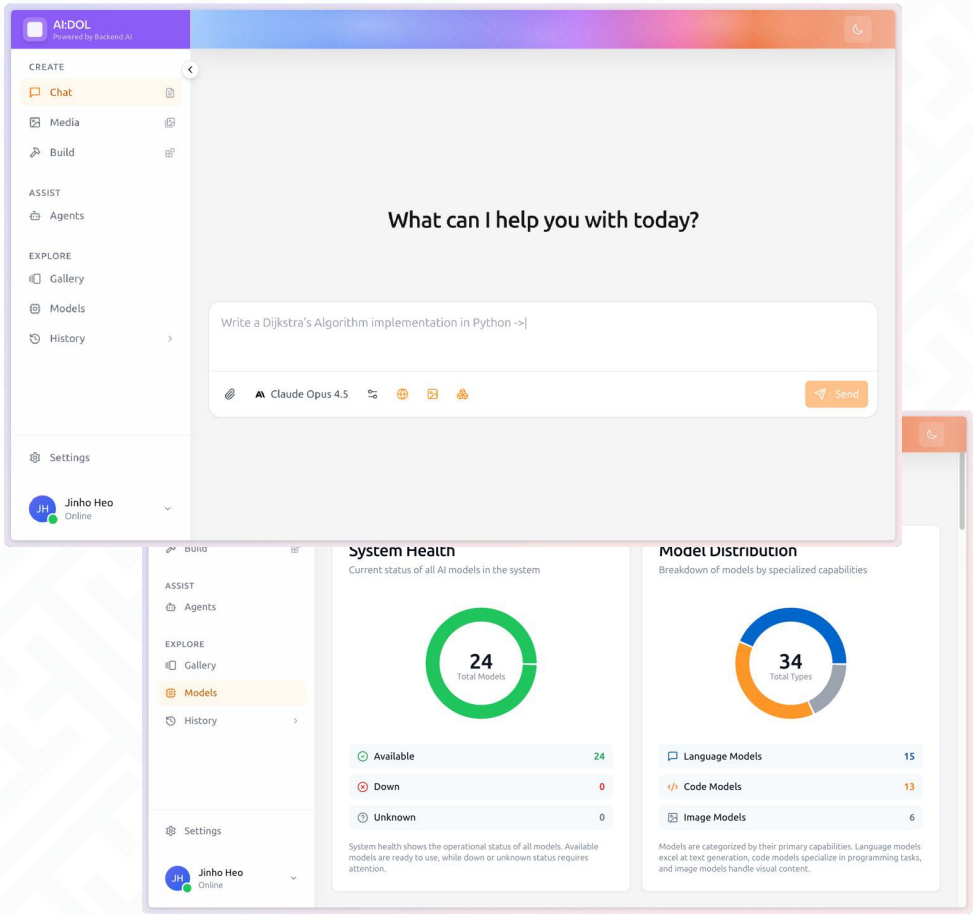
AI 네이티브 시대를 위한 아티팩트 생성 플랫폼

채팅부터 코딩, 멀티모달 창작까지 – 오픈 모델 기반 생성형 AI 서비스 환경 제공
다양한 오픈 모델을 클라우드로 연결하여 채팅, 코드, 이미지, 비디오 등의 아티팩트를 생성할 수 있는 플랫폼

AI로 이룰 수 있는 모든 것을 브라우저 환경에서 손쉽게 사용
별도의 명령줄 인터페이스 없이 자연어 기반으로 브라우저 안에서 멀티모달 창작

사용자의 시스템과 클라우드까지 수직으로 통합하여 최고 성능 달성
수직 통합을 통해 기존 오픈소스 및 다양한 도구 묶음과 비교하여 효율적인 동작 수행

어떤 상황에서도 사용자의 명령을 최우선으로 수행하는 플랫폼
Backend.AI Continuum Router를 백엔드로 이용하여 안정적인 모델 라우팅 및 장애 복구 지원



backend AI

lablup

Continuum Router

AI:DOL의 기반이 되는 엔터프라이즈 수준 Multi-LLM 게이트웨이

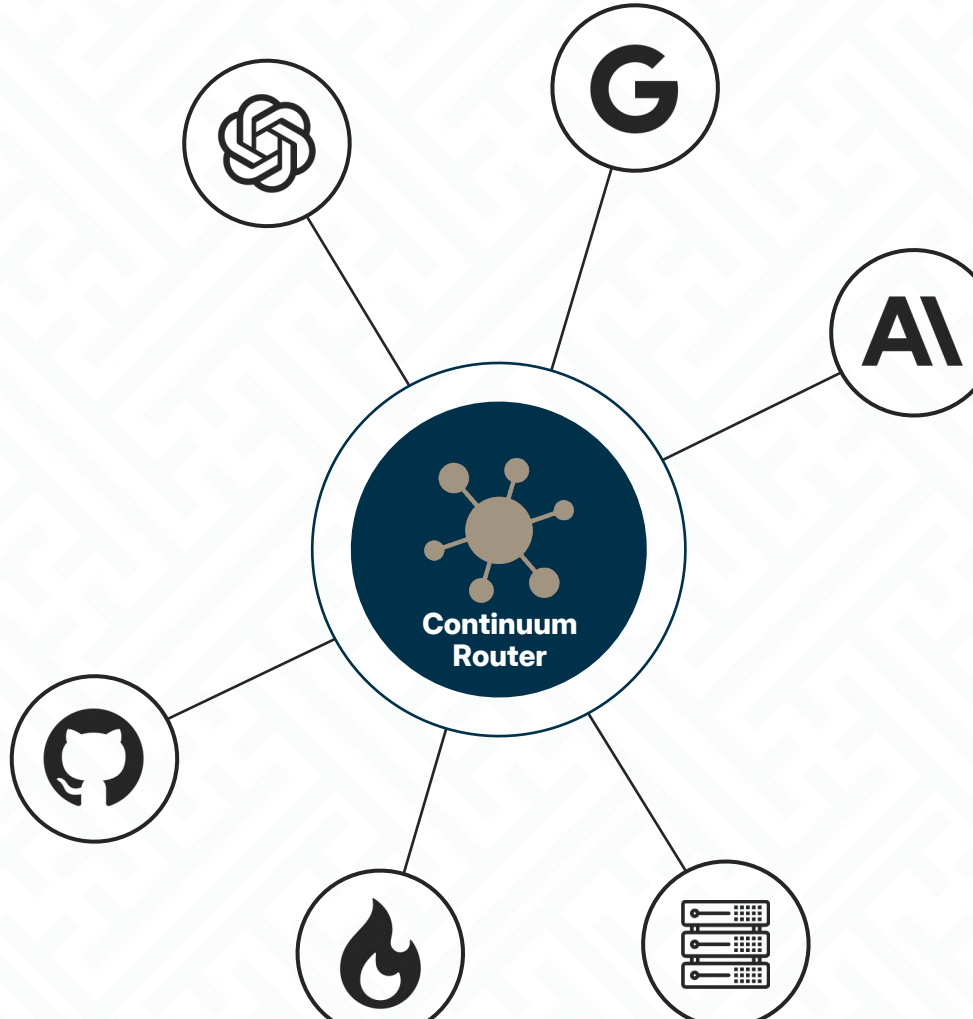
다양한 LLM API를 직접 연결하는 것 보다 효율적인 통합
OpenAI 호환성을 갖추고 있는 Continuum Router로 기존 생태계와 코드 변경 없이 즉시 백엔드 통합 가능

프로덕션에 적합한 엔터프라이즈급 안정성 제공
헬스체크(장애 감지) 자동화, 서킷브레이커, 모델 폴백 기능으로 장애 격리 및 서비스 안정성 보장

비교할 수 없는 최고의 성능
재시작 없는 실시간 설정 변경 (Hot-reload), 고급 라우팅 전략 설정, 5ms 미만의 오버헤드로 최적화된 성능

프록시 그 이상의 온전한 API 생태계 구현
Files API, 자동 콘텐츠 주입, 자동 모델 탐색, SSE 스트리밍 등 LLM 애플리케이션에 필수적인 기능 통합 제공

Admin REST API를 통한 운영자 친화적인 기능 제어
스크립트를 통해 라우터의 모든 측면을 프로그래밍 방식으로 관리할 수 있는 인터페이스 제공



backend AI

제품 포트폴리오

All-SMI

엔터프라이즈 AI 인프라를 위한 통합 GPU/NPU 모니터링 솔루션

데이터센터 전체를 효율적으로 관제할 수 있는 통합 솔루션
각 노드 및 섀시의 발열 정보 모니터링을 통해 대규모 데이터센터의 정밀한 관제 가능

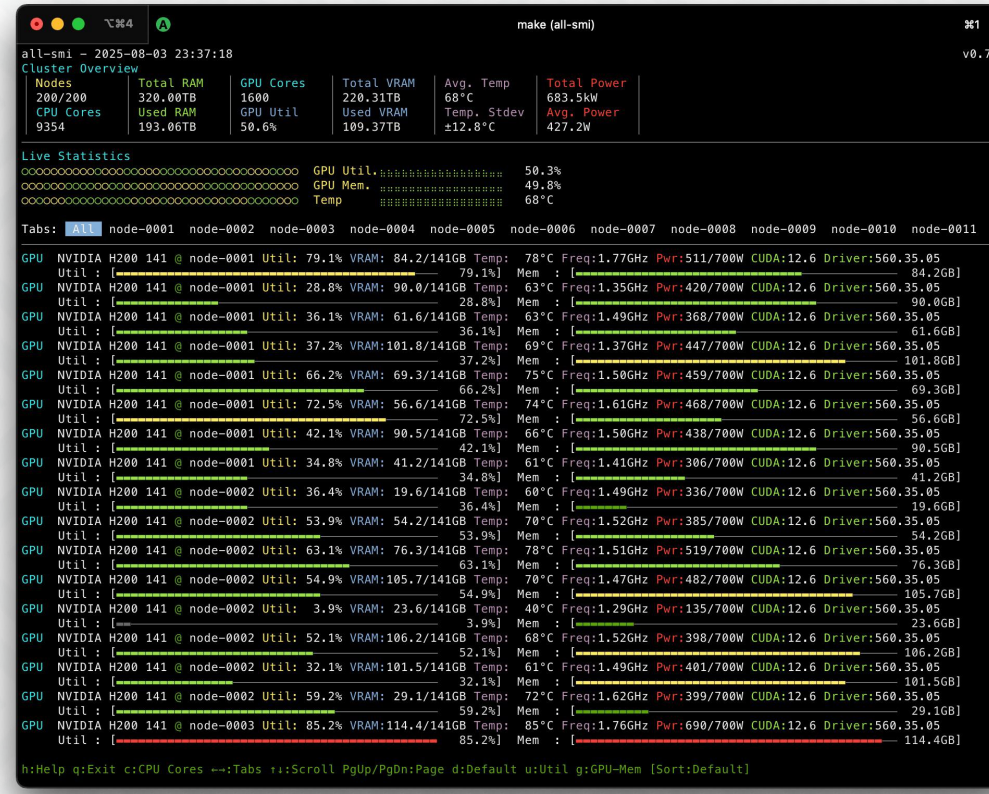
통일된 메트릭으로 다양한 AI 가속기 플랫폼 동시 운용에 적합
NVIDIA, AMD, Apple Silicon, Intel Gaudi, Google TPU 등 9+종의 플랫폼을 단일 UI로 관리, 이기종 운영 복잡도 해소

시스템에서 점유되고 있는 리소스 가시성 확보
256+개의 원격 시스템 동시 모니터링, 총 노드 수, GPU 활용률, 메모리 사용량, 온도 및 전력 소비량 실시간 제공

엔터프라이즈급 안정적인 원격 모니터링 아키텍처
연결 풀링, 동시 연결 제한, 자동 재시도 및 TCP Keep-alive로 대규모 분산 환경에서도 안정적인 모니터링 제공

상세한 프로세스 추적으로 리소스 낭비 원인 파악
GPU 메모리 사용량, CPU 사용률, 프로세스 상태를 추적, 성능 병목 현상을 신속하게 진단

빠른 문제 대응을 위한 직관적 인터랙티브 UI 제공
색상 상태 표시(Green ≤60%, Yellow 60-80%, Red >80%), 실시간 그래프로 운영팀의 신속한 의사결정 지원



고객

lablup

모든 규모의 비즈니스를 뒷받침하는
고객 성공 플랫폼 Backend.AI

래블업이 관리하는 GPU
16,000 대 이상

단일 사이트 내 GPU
2,000 대 이상

래블업을 신뢰하는 고객사
110 곳 이상

Backend.AI 레퍼런스

SAMSUNG · kt · LG · HYUNDAI MOBIS · 42dot · Samsung Research · LIG Nex1 · CJ · SMC · 삼성서울병원 · 대한민국해군 REPUBLIC OF KOREA NAVY · 대한항공

삼성웰스토리 · 한국은행 BANK OF KOREA · 신한은행 · HIRA HEALTH INSURANCE REVIEW & ASSESSMENT SERVICE · ETRI 한국전자통신연구원 · TTA · KAERI Since 1959 · KICT · KIST · AICA 인공지능산업융합사업단

kt cloud · NHN CLOUD · 가천대학교 · 단국대학교 · kit · GIST · KMU · 성균관대학교 · KENTECH 한국에너지공과대학교 Korea Institute of Energy Technology · 한양대학교 1939 · 충남연구원 ChungNam Institute · KU MEDICINE · UIPA 울산정보산업진흥원 · *110+ and Growing!*

backend AI

Backend.AI와 함께 구축하는 고객 사례

USC | Center for Advanced Research Computing
Enabling scientific breakthroughs at scale

김병도 교수 (부연구정보책임자)
USC Center for Advanced Research Computing

> 학교 차원에서 제공된 리소스가 학생들의 실질적인 활용으로 이어지지 못한다면, 해당 자원의 가치는 퇴색됩니다. Backend.AI는 기관의 컴퓨팅 리소스에 대한 관리 편의성을 높이고 학생들의 손쉬운 접근성을 보장하여, HPC 인프라가 본연의 목적을 다할 수 있게 하는 AI 운영 플랫폼입니다.

upstage

이승윤 기술이사
Upstage

> 업스테이지는 인프라 구축과 트러블슈팅에 시간을 낭비하는 대신, Backend.AI에 모든 운영을 맡기고 '독자 AI 파운데이션 모델 프로젝트'의 핵심인 차세대 Solar 모델 개발에만 몰두할 수 있었습니다. 이는 Backend.AI가 복잡한 인프라 뒷단을 모두 책임져 준 덕분에 가능했던 결과입니다.

KIST Korea Institute of Science and Technology

이정민, David S. Yang, 정철현* 박사
CJLab, 화학생명융합연구센터
* 교신저자

> 래블업의 클라우드 리소스를 도입한 후 연구 운영이 물 흐르듯 매끄러워졌습니다.
> 무엇보다 빠르고 편리한 데다 어디서든 접속할 수 있는 Backend.AI 플랫폼 덕분에,
> 팀원들이 어디에 있든 생산성 저하 없이 연구를 이어가고 있습니다.

국내 독자 파운데이션 모델

- **구성**
  - B200 GPU x 63 nodes
  - 고속 스토리지 VAST E-Box, 2 PB
- **고객사 요구사항**
  - 대규모 클러스터를 안정적으로 운영할 수 있는 인프라 운영 플랫폼
  - 장애 발생 시 유연하게 복구되고, 그 과정에서 사람의 개입을 줄여주는 소프트웨어
- **Backend.AI 핵심 요약**
  - 래블업 자체 개발 GPU 및 서버 모니터링 소프트웨어 및 NVIDIA DCGM을 통한 실시간 장애 감지, 메신저 알림
  - 워크로드 중단 유발 장애 감지 시 자동 노드 격리, 예비 노드 핫 스왑, 세션 재시작
  - 최신 체크포인트로부터 학습을 재개하는 무인 작업 복구

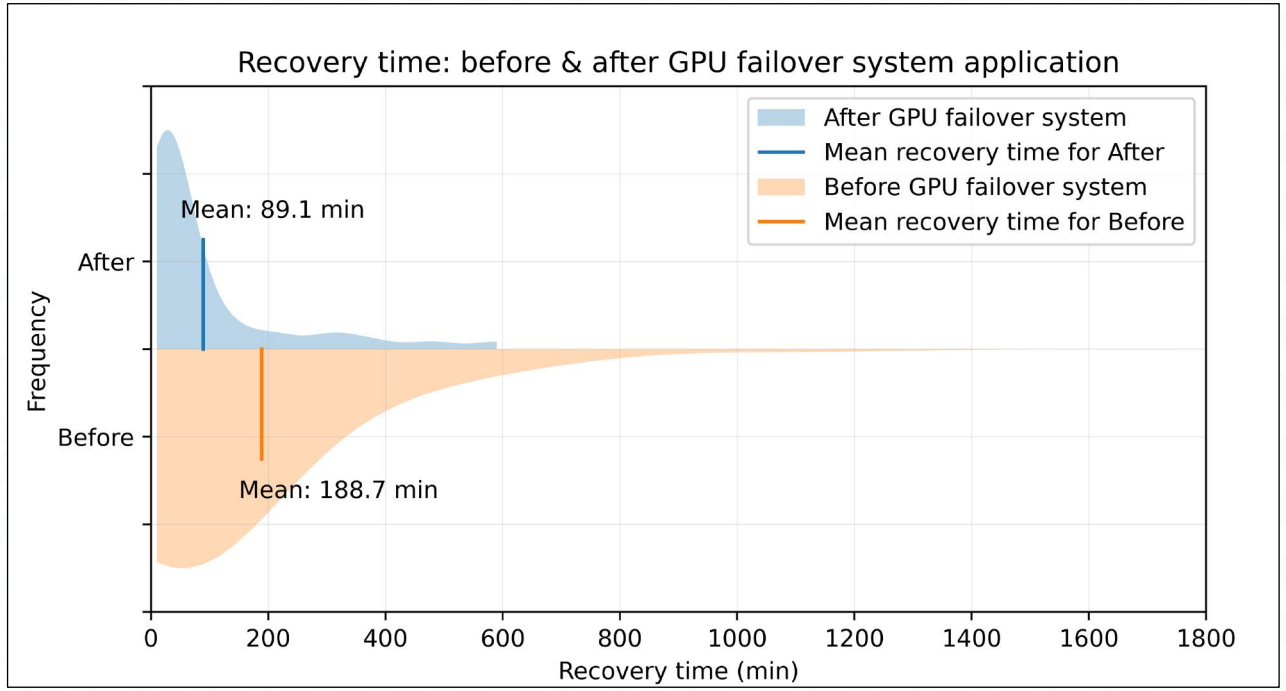
작업 및 클러스터 사용현황 대시보드

SKT AIDC '해인 클러스터'

국내 독자 파운데이션 모델

- **기능 구성**
  - 래블업 자체 개발 GPU/서버 모니터링 소프트웨어 (all-smi) 및 NVIDIA DCGM을 통한 실시간 장애 감지
  - 장애 감지 후 예비 노드 핫 스왑 (Hot-swap) 실행
  - 멀티노드 워크로드를 중단된 체크포인트로부터 재개

- **고객 혜택**
  - 장애 발생 후 학습 재개까지 소요되는 평균 시간 188.7분에서 89.1분으로 약 47% 단축
  - 차세대 Solar 모델 (Solar Open 100B) 학습 시 동일 타임프레임 기준 장애 대응시간 감소를 통한 전체 학습 시간 효율화
  - 인프라 유지보수 소요 감축



학습 중단 후 재개까지의 소요시간 (Backend.AI 적용 전후 비교)

labup

"Make AI Accessible"

AI를 모두에게

backend AI

"Make AI Scalable"

AI를 널리널리

backend AI

Democratizing intelligence for humanity.

래블업은 지능을 공급하는 회사가 되고자 합니다.

MAIL	contact@lablup.com
Lablup Inc.	https://www.lablup.com
Backend.AI	https://www.backend.ai
GitHub	https://github.com/lablup/backend.ai
Cloud	https://cloud.backend.ai