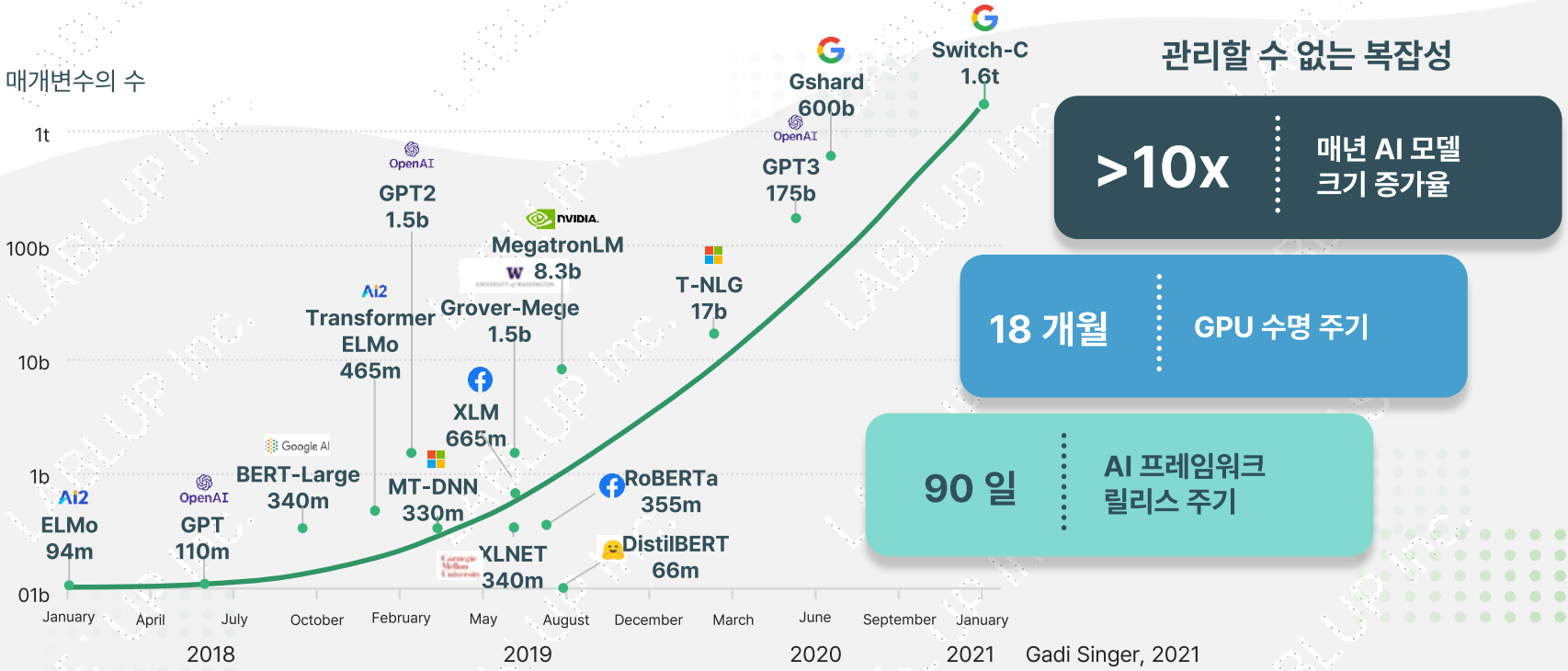
오픈소스 Backend.AI 플랫폼을 활용한 AI 트랜스포메이션

신정규
래블업 주식회사

오늘의 이야기

- **Backend.AI: 요약**
  - 개발 계기 / 목표
  - 핵심 요소 소개
  - 오픈소스 및 엔터프라이즈 버전
- **Backend.AI: AI 트랜스포메이션 사례**
  - AI 전환의 요구 사항
  - 시스템 구성 개요
  - 도입 케이스 요약
- **Backend.AI 클라우드: 꽃들에게 희망을**

1 초 거대 딥 러닝 모델의 시대



관리할 수 없는 복잡성

>10x 매년 AI 모델 크기 증가율

18 개월 GPU 수명 주기

90 일 AI 프레임워크 릴리스 주기

AI/HPC 워크로드: 실질적 문제들

파이프라인 단계별 이질적이고 복합적인 연산자원 요구사항

CPU 집약: 데이터 전처리, 분석, 특징 추출, ...

GPU 집약: 모델 훈련 및 검증, A/B 테스트, 저지연 추론, ...

I/O 집약: 데이터 조작, 배치, GPU 간 통신, ...

I/O 가속 기능들의 설정 복잡도 증가

노드 간 GPU–GPU peering

GPU Direct I/O

GPU Direct Storage

Resizable-BAR

계층화된 스토리지 캐시

블록 저장소와 파일시스템 조합

끝없는 호환성과의 싸움

소프트웨어

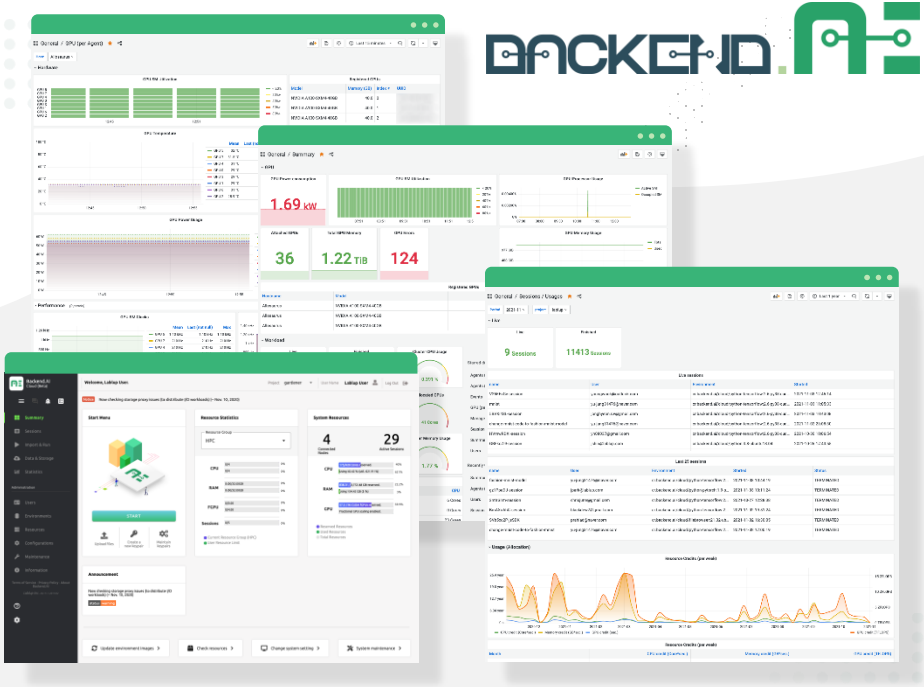
F77/F90 to Julia,
TensorFlow 1.X, 2.X,
PyTorch 1.X, JAX, Haiku...

하드웨어

CPU-only / SSE, AVX-based,
CUDA, ROCm,
Google TPU/Coral,
GraphCore IPU, Habana ...

Backend.AI

AI 개발 및 서비스를 위한
올인원 엔터프라이즈 운영 플랫폼



머신 러닝 모델 개발의 본질에만 집중하도록
AI 훈련 및 서비스 플랫폼에 필요한 관련 기술을
하나의 프레임워크로 제공하는
온프레미스 / 클라우드 통합 솔루션 및 플랫폼

Backend.AI 로 달성할 수 있는 목표

- 즉각적이고 자동화된 분산 컴퓨팅 환경
- 저비용 GPU AutoML (25% 미만의 비용만 소요)
- GPU 분할 가상화 기반의 경제적인 GPU AI 모델 서비스

PyTorch TensorFlow dmlc mxnet PaddlePaddle
DL4J DEEPLEARNING4J Microsoft Cognitive Toolkit Caffe torch Keras

7년 전...

연구자 및 교육자를 위한 계산 및 분석 플랫폼 설계삽질기

신정규 / 김준기 (+ 박종현)

Photo by © Jeongkyu Shin

https://www.slideshare.net/inureyes/pycon-kr-2015

6 현대 과학 연구와 개선

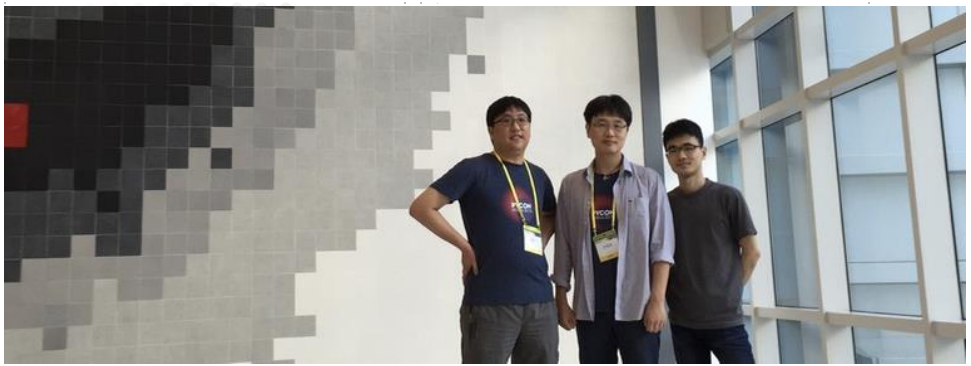
- **PyCon KR 2015 발표 요약**
  - 기술은 21세기, 시스템은 20세기, 학계는 19세기
  - 인간이 관여하는 분야의 발전 속도가 그렇지 않은 분야의 발전 속도를 따라가지 못하는 경향
  - 학계와 업계, 사회의 간극

- **연구말싸미 듕귁에 달아 코드로 서르 사맛띠 아니할쌔 이런 젼차로 어린 학도가 니르고져 홀빼이셔도 마참내 제 뜨들 시러펴디 못할 노미 하니라.**
- **내 이랄 위하야 어엿비 너겨 새로 연구 코드 나누미와 과학 공학 노리터를 맹가노니 사람마다 희여 수비니겨 날로 쑤메 뼈한크이 하고져 할따라미니라.**

  - 재현 가능한 데이터 연구 플랫폼
  - 논문용 코드로부터 실제 서비스까지 확장을 지원하는 클라우드 샌드박스 서비스

2015년 발표 후...

- **컨테이너 기반**의
- **고밀도**
- **분산처리** 플랫폼
- 및 **사용자 편의 기능** 구현

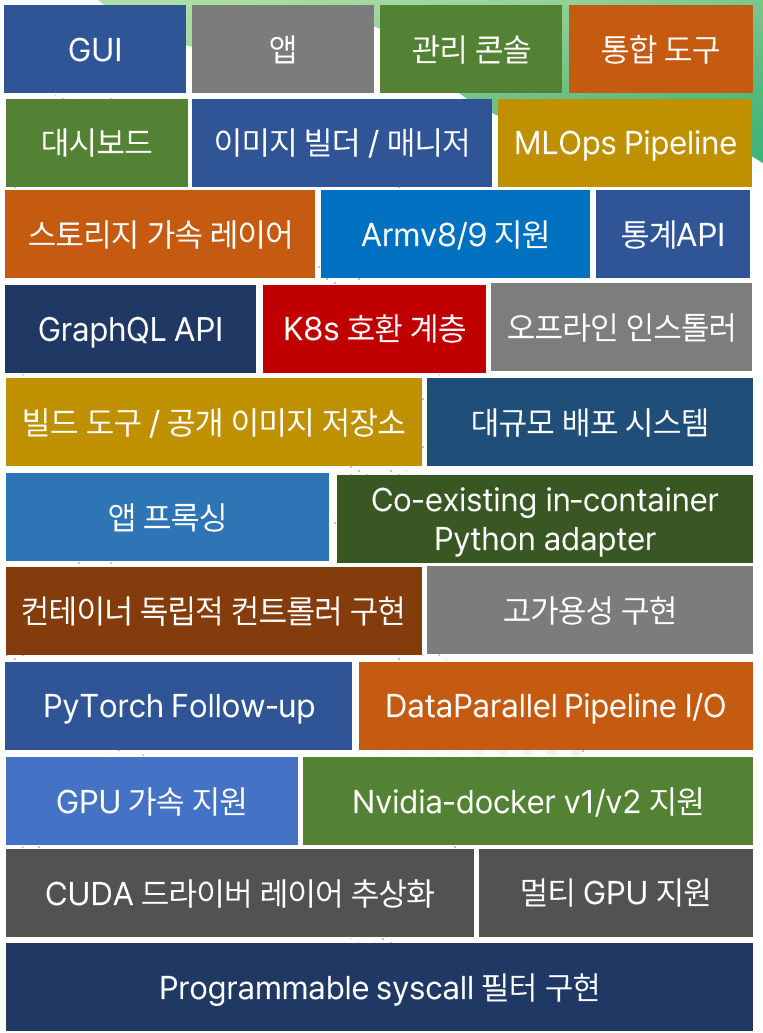


을 금방 할 줄 알았는데, **1.0까지 2년 반**, **엔터프라이즈까지 4년**이 걸렸습니다.

Photo by © Jeongkyu Shin

8 이후

- **2015년** 프로토타입 및 클라우드 서비스
- **2016년** 베타 테스트 / **오픈소스화**
- **2017년** 정식 버전 공개 / API 클라우드 서비스 시작
- **2018년 Fractional GPU 세계 최초 개발** / 분산훈련 기능
- **2019년 엔터프라이즈 버전** 발표 / MLOps 파이프라인 배포
- **2020년** 로컬 패키지 저장소 / **초고속 스토리지 가속** 기능
- **2021년** AI 가속기 추상화 / **ARM 아키텍처** 지원
- **2022년** 커스텀 컨테이너 이미지 빌더 / AI/MLOps 독립 기능



Backend.AI: 오픈소스

- **듀얼 라이선스 모델**
  - 비상업용: LGPLv3 (코어) + MIT (SDK/API/Apps)
  - 상업용: 별도 계약
  - 엔터프라이즈 기능들은 플러그인 또는 SDK를 이용한 별도 솔루션으로 제작하여 제공

- **API Server (backend.ai-manager + client + webserver + storage-proxy)**
  - 개인 PC 또는 서버에서 사용자가 스스로 Backend.AI 환경을 설치·사용 지원

- **API Client (backend.ai-client)**
  - 사용자가 직접 설치한 Backend.AI 환경 또는 공개 API 서비스에 접속하기 위한 클라이언트 라이브러리

- **Client SDK**
  - Python 3, JavaScript (ES6), Node.js 버전 제공
  - pip 및 npm / yarn 을 이용한 자동 라이브러리 설치

Backend.AI: 오픈소스

- **Web UI / WebServer (webserver + webUI + app)**
  - JavaScript SDK 기반의 GUI 인터페이스
  - 웹 서비스 / 데스크탑 앱 지원

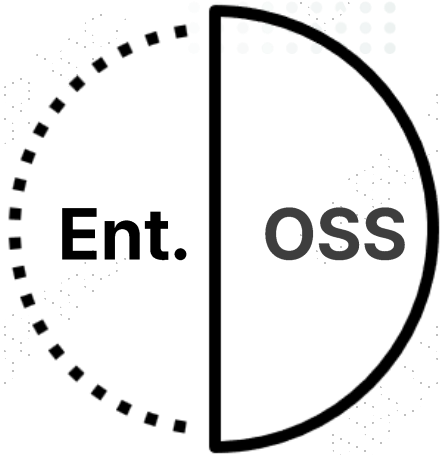
- **Environment**
  - Backend.AI 와 연동하여 사용하는 연산 환경 / 런타임 / 딥 러닝 환경 컨테이너 이미지
  - 17개 언어 및 5종의 딥 러닝 프레임워크를 docker hub 및 자체 컨테이너 레지스트리를 통해 제공
  - Backend.AI 없이도 사용 가능: **100만 다운로드** 이상 (2022년 초)

- **ForkLift 이미지 빌더**
  - 컨테이너 이미지를 템플릿 기반으로 쉽게 만들 수 있는 GUI 도구

Backend.AI: 오픈소스

조직 관리
대규모 시스템
비용절감기능
시스템 보증

Enterprise BACKEND.AI



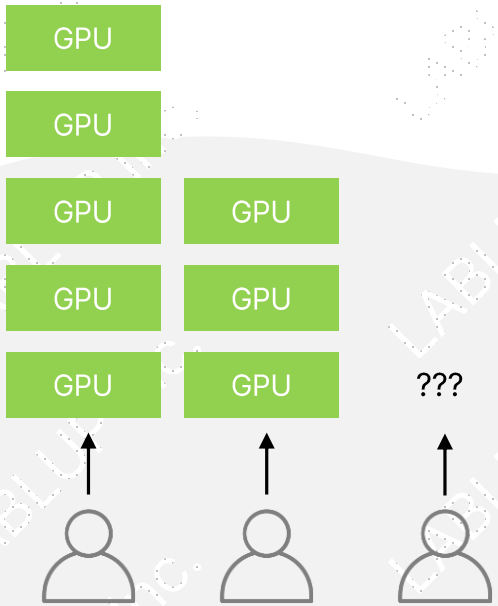
일반 기능
완전한 동작
소스코드

GPL MIT

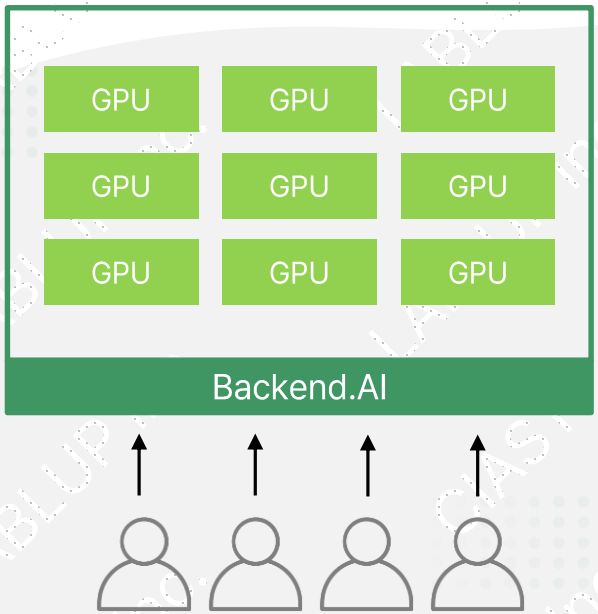
2


BACKEND.AI
E-mail로 로그인
다른 방식으로 로그인하기
IAM 사용
E-mail
비밀번호
엔드포인트
http://127.0.0.1:8090
로그인

Backend.AI: 개발 목표

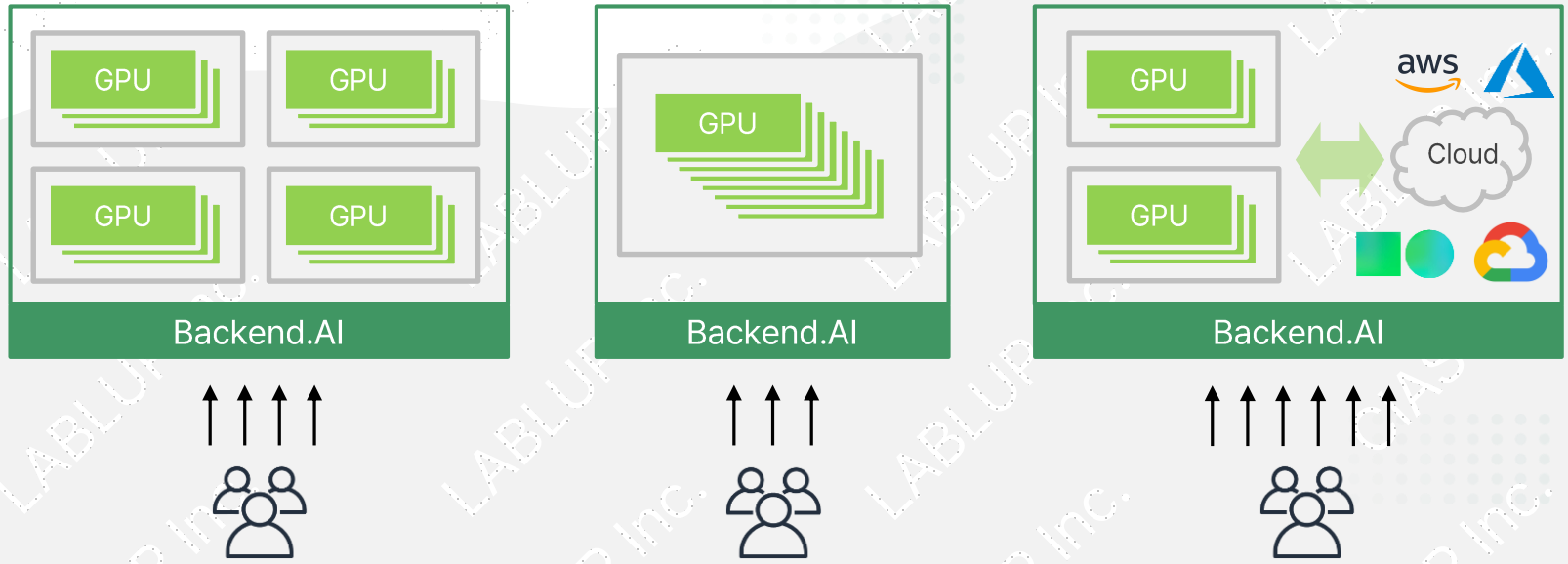


- 관리자가 사용자별로 GPU 할당
- 놀거나 부족한 자원 발생
- SW 유지관리 어려움



- Backend.AI가 GPU 통합 관리
- 동적으로 그때그때 필요한 만큼만 사용
- 컨테이너 기반으로 표준화된 SW 생태계

Backend.AI: 사용 시나리오



GPU 클러스터 구축

고성능 GPU 공유

클라우드로의 동적 용량 확장

Backend.AI: 개발 로드맵

Alpha | *Beta* | *Production-ready*

2015.8
프로젝트 공개 (PyCon KR)

2016.11
v0.9 릴리즈
- **GitHub을 통한 오픈소스 공개**
- LGPLv3/MIT 라이선스 적용

2017.10
v1.0 릴리즈
- REPL 기능 안정화
- 개발 매뉴얼 제공
- 가상폴더 기능 추가
- PyPI 공개 (pip install)

2018.1~3
v1.1 ~ v1.3 릴리즈
- 코드 안정화
- 설치프로그램 추가
- 플러그인 구조
- 브랜치 관리 규칙 적용

2018.9
v1.4 릴리즈
- GPU 부분공유 기능 첫 구현
- cloud.backend.ai 비공개 베타 시작

2018.12
v18.12 릴리즈
- 버전 번호 부여 정책 변경 (연.월)
- 공개 및 사설 Docker registry 연동
- Google TPU 지원 추가

2019.9
v19.09 (Enterprise R1) 릴리즈
- GPU 부분 공유 및 가상화 고도화
- 연산 자원 리소스 그룹 기능
- 이메일 기반의 사용자 관리 기능
- 도메인별 관리 기능 및 SSO 지원
- 엔터프라이즈용 control panel 기능
- 고가용성(HA) 지원
- Harbor Docker Registry v1 연동

시간

Backend.AI: 개발 로드맵 (2020-2021)

Enterprise R1 (20.03/09) | *Enterprise R2 (21.03/09/22.03)*

2020.6

v20.03 릴리즈

- Python 3.8 기반
- Callosum 보안 터널 연결 도입
- 표준화된 파이프라인 모듈 인터페이스
- LustreFS, GlusterFS 지원
- 리눅스 데스크탑 GUI 터널링 지원
- Harbor Docker Registry v2 연동
- DGX-A100 지원
- k8s pod 연동 지원 (베타)
- Google TPU 지원 (베타)
- AMD ROCm 지원 (베타)
- cloud.backend.ai 공개 베타 시작

2020.11

v20.09 릴리즈

- 멀티컨테이너 세션 지원
- 데이터 파이프라인
- XFS 파일시스템 지원
- PureStorage 통합
- DGX 통합 지원 (정식)
- AMD ROCm 지원 (정식)

2021.3

v21.03 릴리즈

- Python 3.9 기반
- SQLAlchemy v1.4 / aioredis v2 도입
- 대규모 클러스터 지원 안정화
- 스케줄러 HoL 회피기법 적용
- Watcher framework (베타)

시간

Backend.AI: 개발 로드맵 (2021-2022)

Enterprise R2 (21.03/09/22.03)

Enterprise R3 (22.09~)

2021.11

v21.09 릴리즈

- ARM64 (Apple Silicon, AWS Graviton, NVIDIA Jetson Nano) 지원
- RDMA 가속 지원
- NetApp 스토리지 통합
- 추론 워크로드를 위한 앱 스트리밍 최적화
- 실시간 통계 대시보드
- 파이프라인 스케줄러 통합 (베타)
- 템플릿 기반 세션 생성 (정식)
- Watcher framework (정식)

2022.3

v22.03 릴리즈

- Python 3.10 기반
- Storage proxy 파일브라우저 통합
- 동적 세션 자원 재할당(resizing/rescaling)
- 컨테이너 이미지 빌더 컴포넌트 (베타)
- Dell 스토리지 통합
- 파이프라인 스케줄러 통합 (정식)
- k8s pod 연동 지원 (정식)

2022.9

v22.09 릴리즈

- API 서비스에 특화된 AppProxy v3
- FastTrack AIOps UI 플랫폼 (베타)
- ForkLift 이미지 빌더 플랫폼 (정식)
- Weka.io 가속 기능 통합
- GraphCore IPU 지원 (베타)
- 사용자 관리 모듈 (베타)

2022.~

그 이후

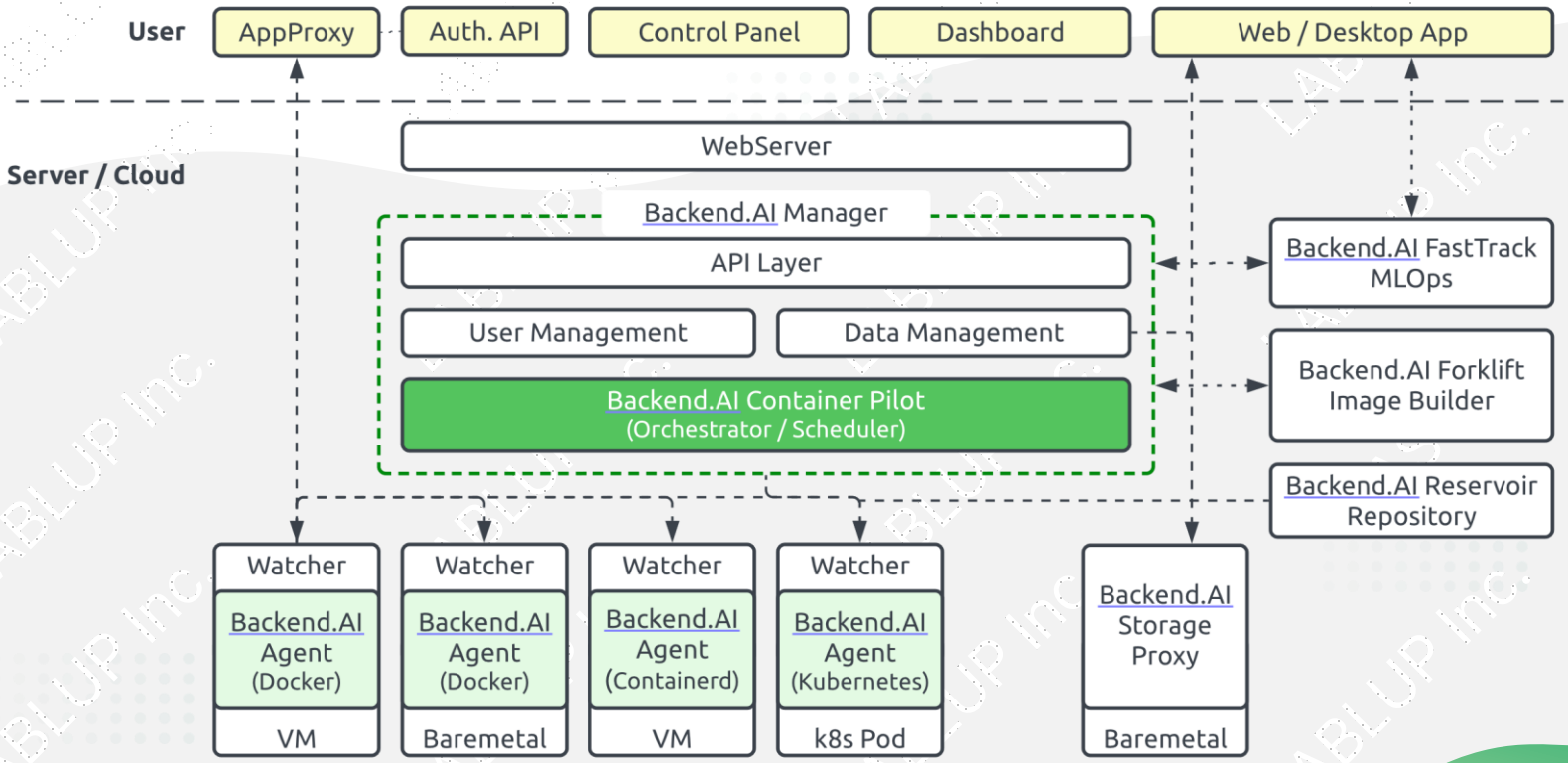
- cloud.backend.ai 정식 공개 서비스
- ARM64/x86 hybrid 환경 지원
- IoT/Edge 장치 환경 지원
- Federated Learning & Inference 통합
- World Console

시간

기여

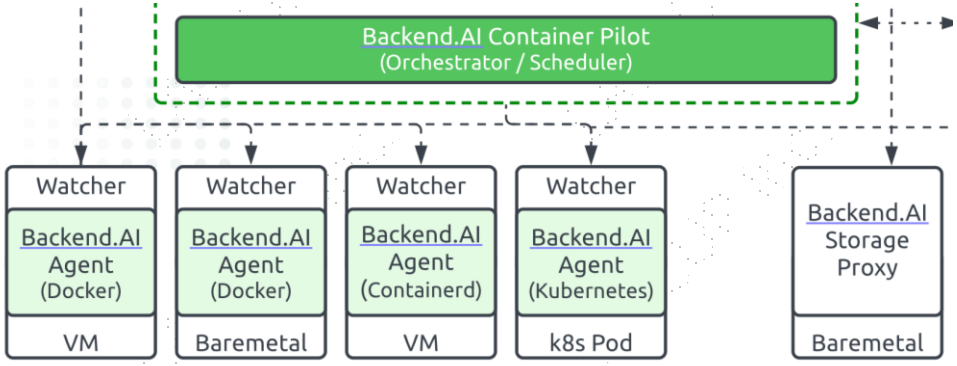
	2015 Alpha	2017 Beta	2019 Production	2020 Expansion
기술 진보	• 오픈 소스	• GPU 최적화 스케줄링 • 올 인 원 MLOps	• Fractional GPU sharing(GA) • 멀티 노드 세션 • 엔터프라이즈 기능	• 초대형 분산 컴퓨팅 실행을 위한 확장 • 최고의 분산 AI 성능 달성 • 다양한 파트너십
분야 기여		• 자체 ML/AI Platform 구축/실행 가능 • 하드웨어 지식 없이 높은 GPU 활용도 달성	• GPU 비용 < ½ 절감 • ML/AI 의 최근 규모 확장 경향을 백엔드 지식없이 지속적으로 따라가도록 지원 (엔터프라이즈)	• 관리 가능한 초대형 ML/AI 클러스터 제공 • AI 모델 추론 비용 대폭 절감

Backend.AI: 구조 요약



Backend.AI Container Pilot: 개요

- **워크로드의 배정 및 관리를 담당하는 계층**
- **Backend.AI Agent: 로컬 워크로드 관리**
  - Docker / Containerd 컨테이너
  - Kubernetes Pod
  - VMs
- **담당 기능**
  - 단일 / 멀티 컨테이너 세션 배치 및 상호 연결
  - 워크로드 메트릭 관리
  - 다중 가상 네트워크 설정 및 모니터링
  - 사용자 계층 / 데이터 트래픽 / GPU 네트워크 분리
  - 앱 프록시 계층으로의 연결 관리
  - 실시간 앱 이미지 조합을 통한 사용자 컨테이너 제공



Backend.AI Container Pilot: 필요성 및 강점

- **딥러닝 워크로드에 적합하지 않은 K8s**
  - 마이크로서비스 아키텍처에 특화된 철학, 구조, 구현
  - 컨텍스트 유지, 배치 작업과 정확히 반대되는 지향점을 향해 개발됨
  - 로드 밸런싱 정책의 차이: 분산 정책 / 집중 정책
  - 스토리지 관리 및 네트워크 계층 추상화 장벽

- **Backend.AI Container Pilot**
  - 딥러닝 / 빅데이터 분석등의 대규모 분산 및 자원 집중형 워크로드에 맞춘 설계
  - 근본 철학의 차이에 따른 완전히 다른 추상화 계층
    - ✓ 자원 그룹 개념, 네트워크 추상화 및 다중 네트워크, GPU 및 AI 가속기 지원 및 관리 영역 차이
  - 예: 멀티노드 GPU-GPU 직접 통신 연결을 통한 초거대 모델 개발 (컨테이너 기반으로 유일하게 지원함)
    - ✓ 노드간 가상 네트워크 구축으로 인한 속도 감소는 거의 없으면서도
    - ✓ 훈련 성능 비교 시 5배 이상의 차이 발생

Backend.AI: AI 소프트웨어 생태계 통합

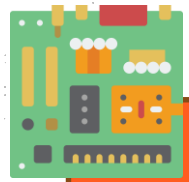
Technology		Apache AirFlow	Microsoft NNI	Databricks MLFlow	NVIDIA Triton	Weights& Biases	Backend.AI
Model training & Orchestration	Experiment tracking	✔	✔	✔		✔	✔
	Model Management	✔	✔	✔	✔	✔	✔
	Machine Orchestration	✔					✔
Data management	Pipeline automation	✔	✔		✔	✔	✔
	Data and pipeline versioning		✔	✔		✔	✔
Service production	Model deployment and monitoring				✔		✔
	Platform specific Deploy				✔		✔
User Interface	API / SDK		SDK	SDK		SDK	SDK
	Dashboard	✔	✔	✔		✔	✔
Backend.AI Integration		Serving as Backend.AI Environment	Serving as Backend.AI Environment	Serving as Backend.AI Environment	Serving as Backend.AI Environment	Serving as Backend.AI Environment	Serving as Backend.AI Environment

✔ : now in beta (Jan. 2021)

Backend.AI Resource Management: GPU부터 클러스터까지

Make GPUs
shareable & flexible

드라이버 레벨
GPU 분할 가상화



GPU-first scheduling /
Resource allocation

GPU-최우선
독자 스케줄러 / 오케스트레이터

Resource Group and
Resource Policy

자원 그룹과 **자원 정책** 기반
자동화 최적 자원 관리

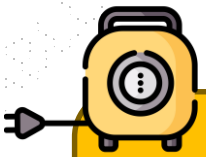
도입 연도	2018	2018	2018

Backend.AI MLOps: 클러스터부터 사용자까지

Pipeline Storage
for Distributed computing

분산처리, 재사용성 및
이식성에 특화한
파이프라인 설계

도입 연도 2020



App Proxy
for secure container access

다양한 앱을 **분산 환경** 및
보안 환경에서 실행하는
프록시 서버

2018

Storage Proxy
to optimize network I/O

데이터 입출력 부담을
분산하기 위한
스토리지 프록시

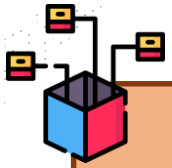
2020

Backend.AI Hyper-scale : 사용자부터 거대 모델까지

Cluster Session
to accelerate model training

여러 노드를 이용해
딥 러닝 모델을 분산훈련하는
클러스터 세션

도입 연도 2019



Domain
to run multiple Backend.AIs
with same cluster

여러 Backend.AI 서비스를
단일 클러스터에서 운영하는
도메인 기능

2019

Reservoir
To provide full-featured Air-gapped environment

완전 폐쇄망에서
패키지 서비스를 지원하는
자체 패키지 저장소

2021

BACKEND.AI

lablup

Backend.AI Future-ready : 미래에 대한 선제 대응

ARM64 HPC/AI Hybrid cluster

ARM64 CPU 아키텍처 지원 및 **멀티 아키텍처**로 구성된 하이브리드 클러스터 운영

Kubernetes Cluster as Backend.AI Agent

Kubernetes Pod를 Backend.AI의 연산 자원으로 **통합**

ASIC abstraction for Future Accelerators

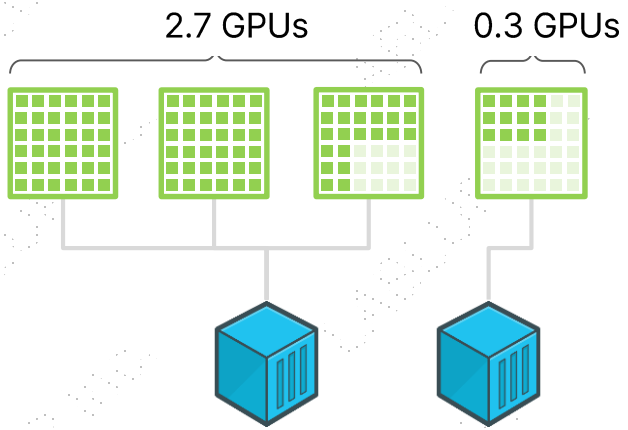
다양한 전용 가속기의 빠른 이식 및 통합을 지원하는 **가속기 추상화 레이어**

도입 연도	2021	2021	2021

Backend.AI: GPU 가상화 (fractional GPU™) 기반의 유연한 자원 관리

- **컨테이너 기반 GPU 스케일링**
  - 컨테이너별로 CUDA SMP 및 GPU RAM을 나눠줌
    - ✓ 예) 2.7 GPUs, 0.3 GPUs를 각 컨테이너에 할당
  - 단일 GPU 공유 : 교육 및 추론 워크로드에 적합
  - 다중 GPU 할당 : 모델 훈련 등 대규모 워크로드에 적합
  - 자체 개발한 CUDA 가상화 계층으로 구현 / 분할에 제한이 없음

    * 한국·미국·일본 등록 특허

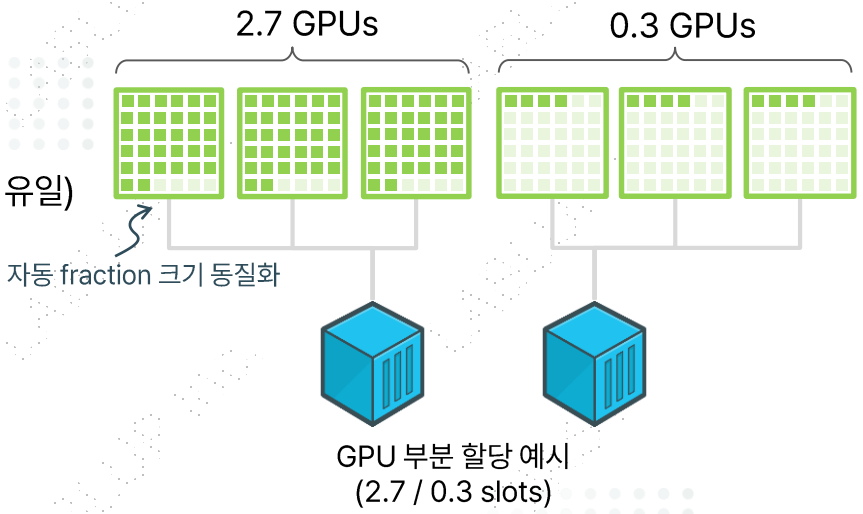


GPU 부분 할당 예시
(2.7 / 0.3 slots)

- **혜택**
  - 고가의 하드웨어인 GPU 의 사용률 향상을 통한 실질적 구입 비용 감소
  - 고가의 훈련용 GPU를 분할하여 추론 GPU로 운영
    - ✓ 수명주기에 따른 노후 장비 활용성 최대화 (훈련 → 추론 → 교육으로 용도 전환)
    - ✓ 감가상각을 고려하였을 때 GPU 클러스터 투자에 대한 의사결정 장벽 타파

Backend.AI: GPU 가상화 (fractional GPU™) 기반의 유연한 자원 관리

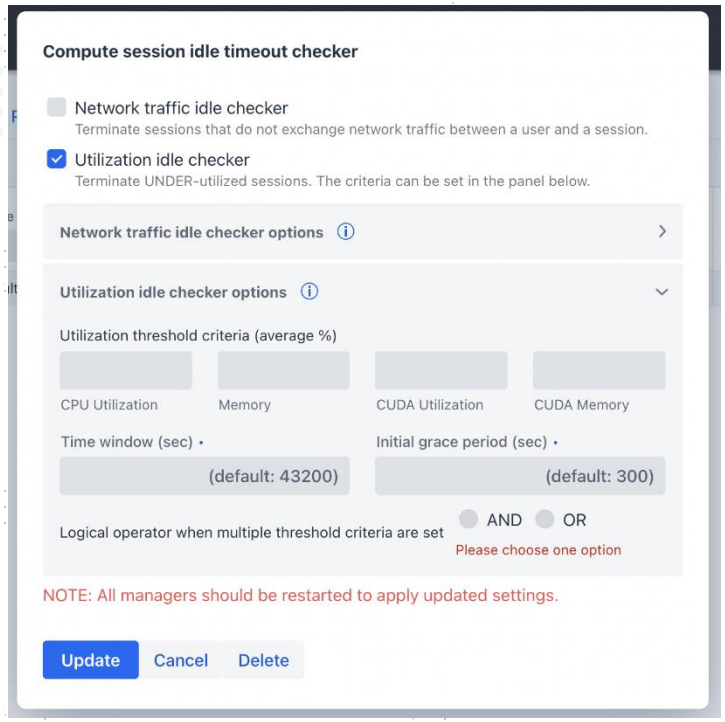
- **NVIDIA 플랫폼 통합**
  - NVIDIA **DGX-Ready** Software
    - ✓ 아태지역 최초 및 유일 (2021)
    - ✓ AI Accelerated Program 멤버 (GTC'22, AI플랫폼 아태 유일)
  - NGC (for DL / HPC) 이미지 그대로 실행



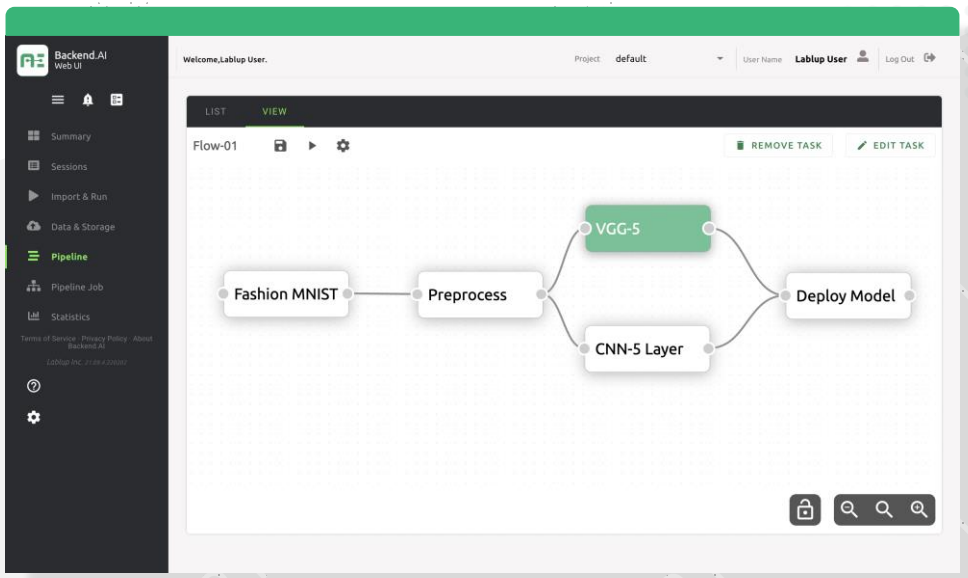
- **유사 기술과의 핵심 차별점**
  - CUDA 8 버전부터 11 버전까지 모든 GPU 모델 호환
    (데스크톱 / 워크스테이션 / 데이터센터)
  - 사용자 프로그램의 코드 변경 필요 없음
  - 딥러닝 프레임워크 수정 및 리빌드 필요 없음
    - ✓ TensorFlow/PyTorch뿐만 아니라 임의의 GPU 가속 워크로드 실행 가능
  - 여러 GPU로부터 할당한 fraction들을 하나의 컨테이너에 연결하는 multi-GPU 워크로드 지원
  - multi-GPU 환경에서 fraction 크기들을 동등하게 맞춰주는 기술 적용

Backend.AI: 정책 기반의 자동화된 자원 할당 및 회수

- **사용자·조직 단위의 자원 정책 설정**
  - API 키 / 사용자 / 프로젝트 / 도메인 별 최대 할당 허용량 설정 (CPU, RAM, GPU, Storage 등)
  - 최대 허용량 내에서는 별도 신청·허가 절차 없이 활용 가능
- **유휴 상태 검사 및 자동 세션 종료 정책**
  - 자원 점유율 기반
    - ✓ 예) 10분 동안 GPU 사용률 0% 및 CPU 사용률 5% 이하 유지되면 강제 종료
  - 상호작용 기반
    - ✓ 예) 1시간 동안 연산 세션과 사용자 사이에 상호작용(네트워크 트래픽) 없으면 자동 종료
  - 시간 기반
    - ✓ 예) 세션 시작 후 12시간 지나면 강제 종료
  - 사용자·프로젝트 및 전역 단위 설정 지원
- **배치 및 파이프라인 작업**
  - 메인 프로그램이 종료되면 연산 세션도 자동으로 종료
    - ✓ 예) 내일 아침 10시에 배치 연산 세션을 실행하고 작업이 완료되면 자동으로 종료 및 자원 회수



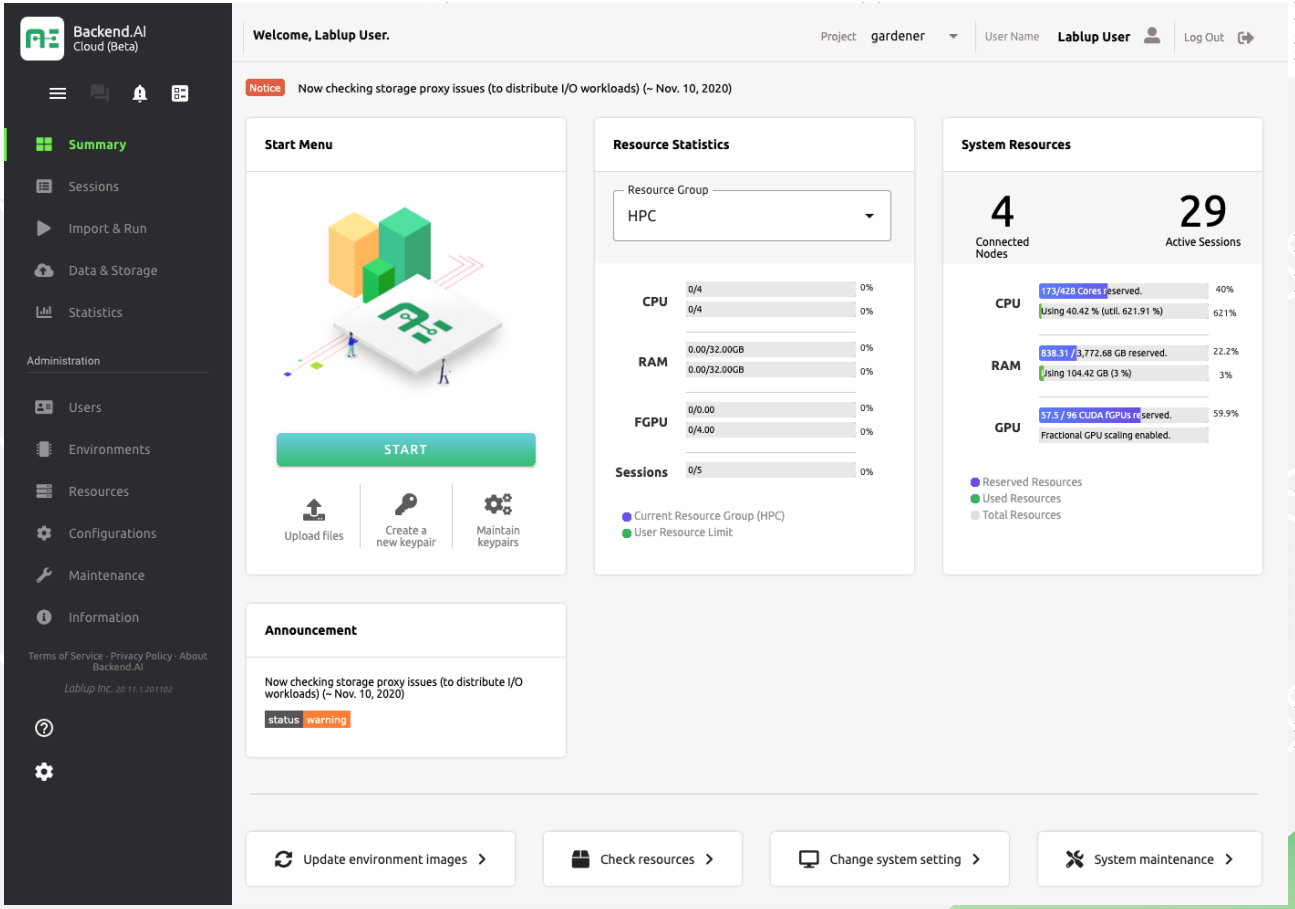
Backend.AI: FastTrack - MLOps 파이프라인 저작도구



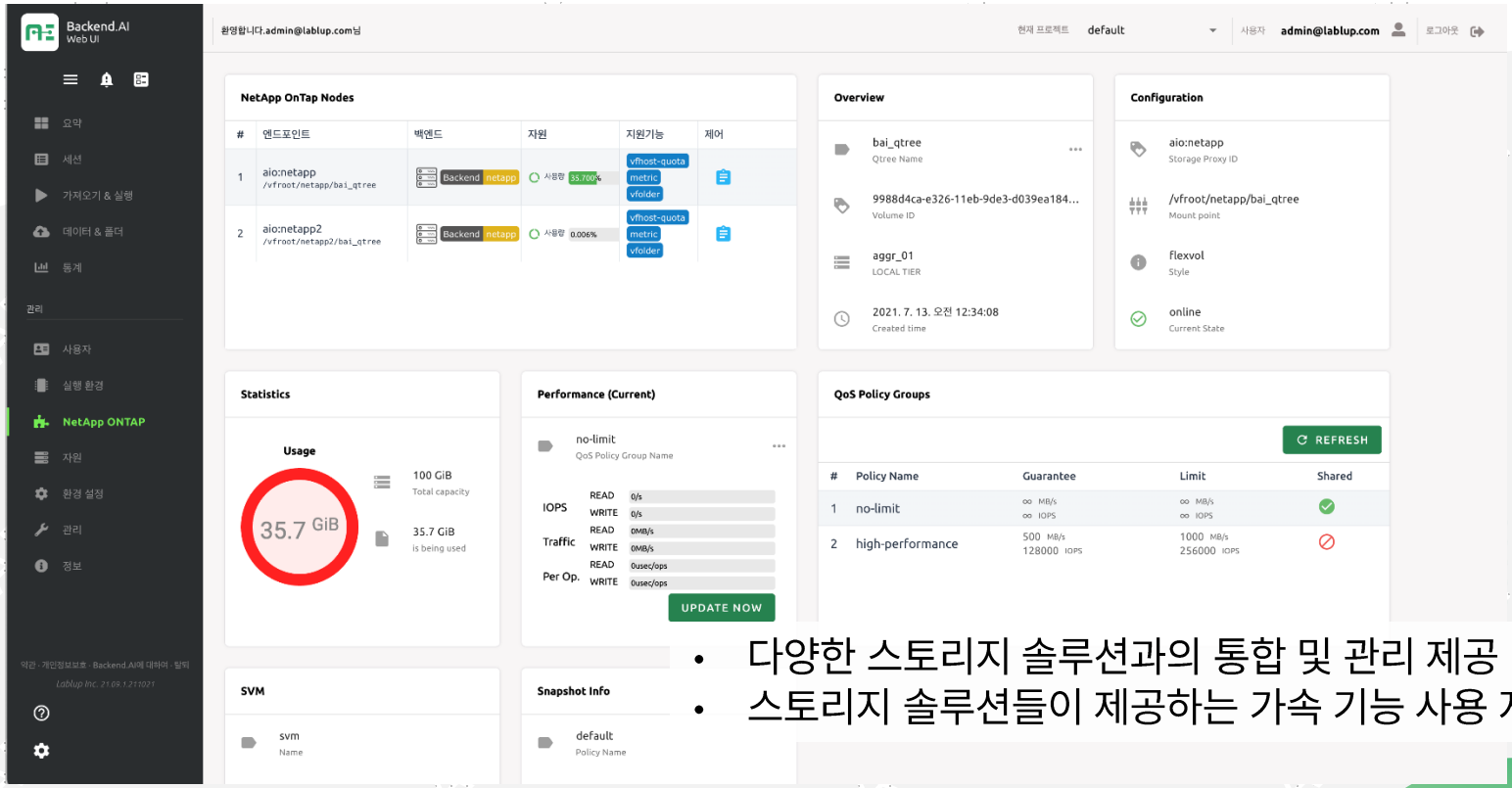
FastTrack MLOps 파이프라인 저작도구

- 파이프라인 템플릿 & 파이프라인 스토어
- GUI 기반 작업 그래프 작성
- YAML 기반 파이프라인 직렬화 형식
- 스토리지 폴더를 통한 메타데이터 및 입출력 연계
- 프로젝트 및 사내 배포 및 공유
- 모델 서빙을 위한 외부 보안 터널링

Backend.AI: 연산 자원 현황 확인

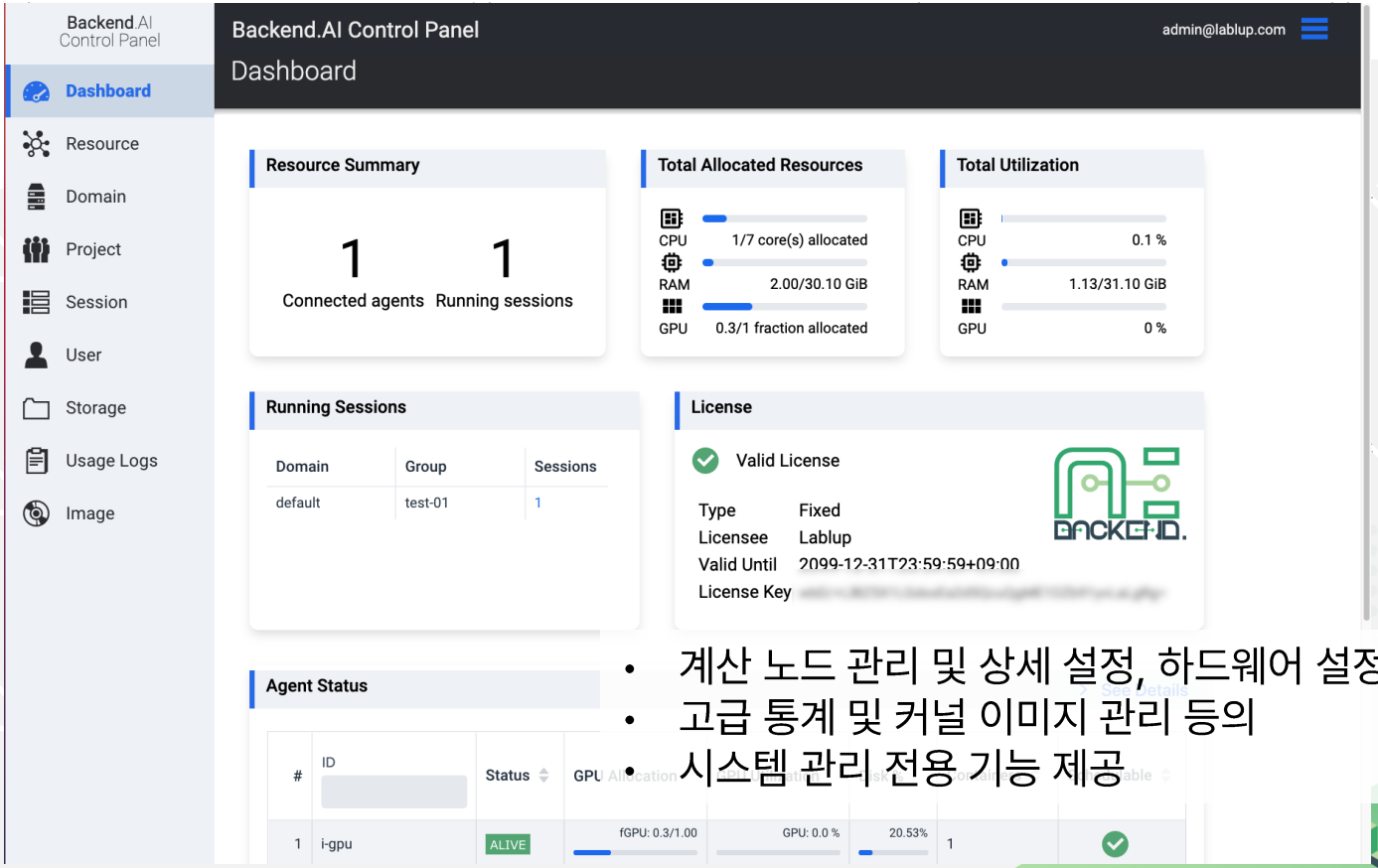


Backend.AI: 스토리지 솔루션 통합



- 다양한 스토리지 솔루션과의 통합 및 관리 제공
- 스토리지 솔루션들이 제공하는 가속 기능 사용 지원

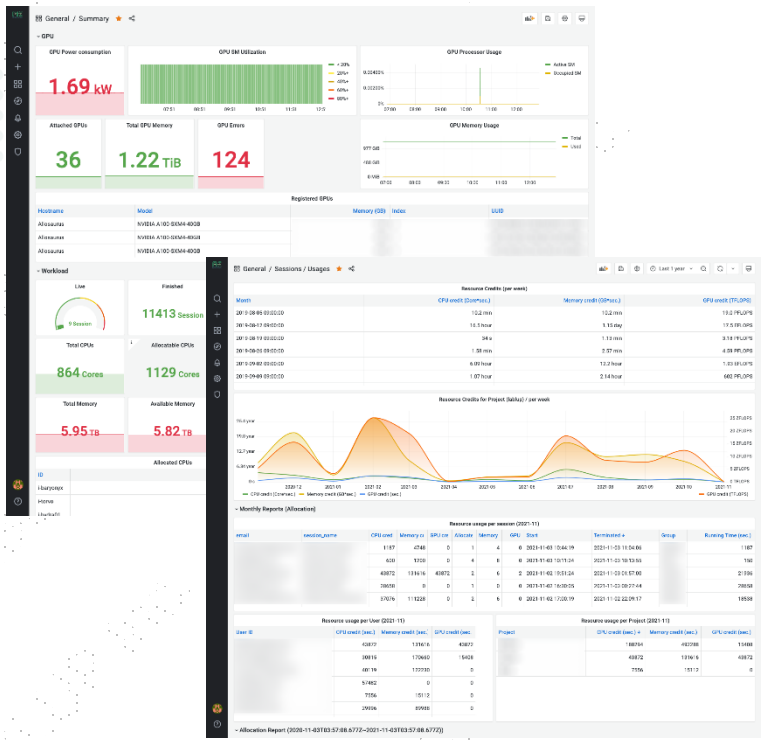
Backend.AI: 시스템 관리자 전용 컨트롤 패널



- 계산 노드 관리 및 상세 설정, 하드웨어 설정,
- 고급 통계 및 커널 이미지 관리 등의
- 시스템 관리 전용 기능 제공

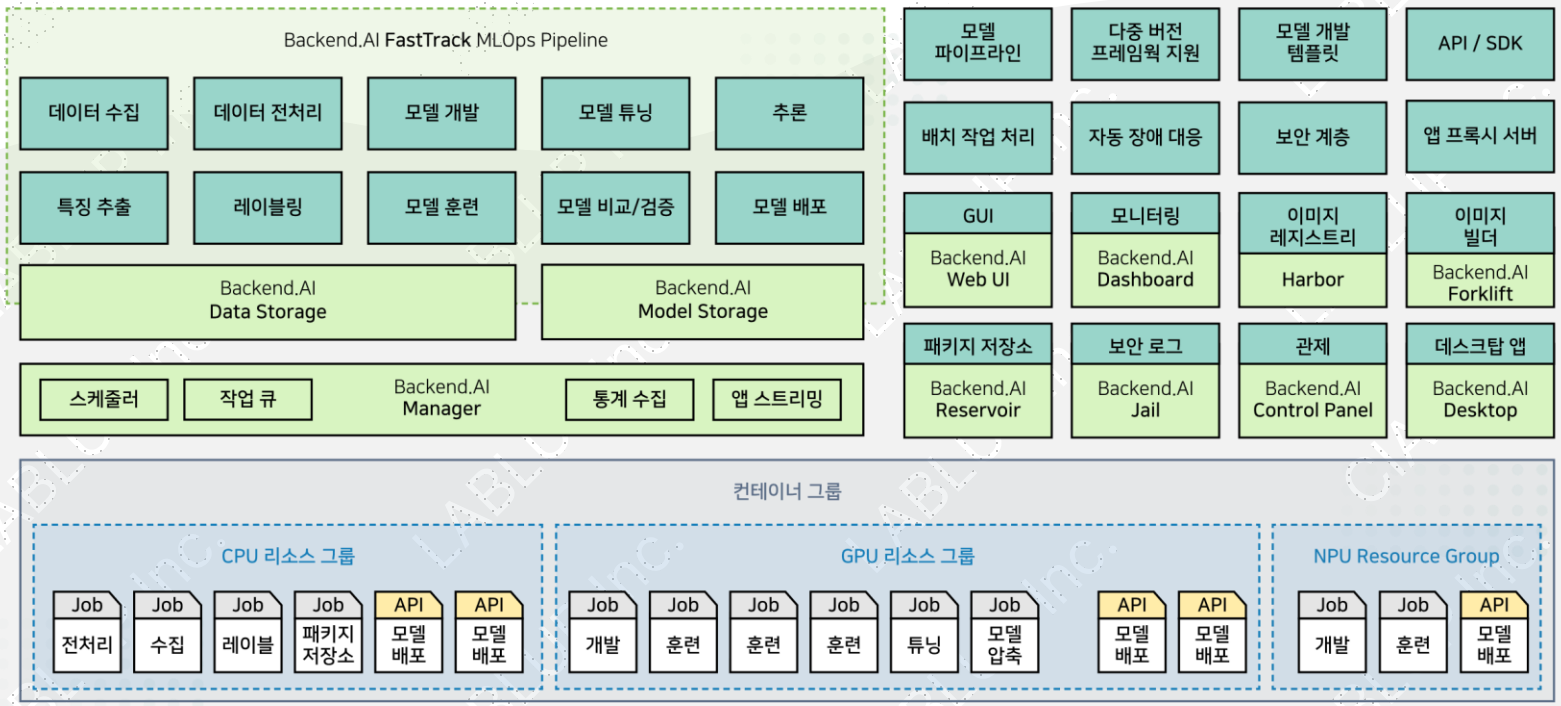
Backend.AI: 관제용 대시보드

- **대규모 시스템 구축시 통계 자동화 및 관제 기능 제공**
- **모니터링**
  - Backend.AI 구성 요소들의 상태 체크 및 보고
  - 실시간 모니터링, 자원 사용량 및 하드웨어 오류 보고
  - 연산 노드의 상태 및 자원 활용 정도
  - 연산 노드 내 GPU별 상세 통계 및 상태 제공
- **통계**
  - 사용자 / 그룹 / 조직별 누적 사용량 통계
  - 기간별 자원 사용률 변동 통계
  - CPU, RAM, GPU 등 자원 별 사용률 모니터와 기록 제공
  - 연산 자원 사용량의 FLOPS 변환 값 제공
  - 통계 데이터 리텐션 설정



Backend.AI: 프로덕션 시스템

: 시스템 개요



Backend.AI: 보급

기업 (국내 / 해외)

공공 기관 및 연구 기관

주요 대학

인증 / 파트너십 / 보급

- 아태지역 유일 NVIDIA DGX-Ready 소프트웨어 (전세계 13개 사 중 하나)
- NVIDIA, AWS, PureStorage, NetApp, Dell 등 다수 하드웨어 플랫폼 벤더들과 기술 파트너십
- 다수의 국내/국외 대기업 및 연구소, 대학에서 활용

AI 트랜스포메이션: 플랫폼 요구사항

- **자사 도메인과 AI의 결합**
  - 도메인 특화 기술 경쟁력을 AI 분야로 확장하는 과정의 어려움 해소
  - 데이터 과학 기반 의사 결정으로의 점진적 이전에 대한 이해 필요
- **투자 부담 경감**
  - 이미 보유하고 있는 AI 자원들을 통합하여 관리할 필요성
  - 비용 감소를 위한 자원 효율성 극대화 요구
- **쉬운 사용자 환경**
  - 기존 사용자들의 사용자 경험을 AI 자원에 구애받지 않고 유지
  - 자동화를 포함한 AI 특화 기능들
- **확장성 확보**
  - 지속적인 신규 하드웨어/소프트웨어 도입 과정의 복잡도 해소

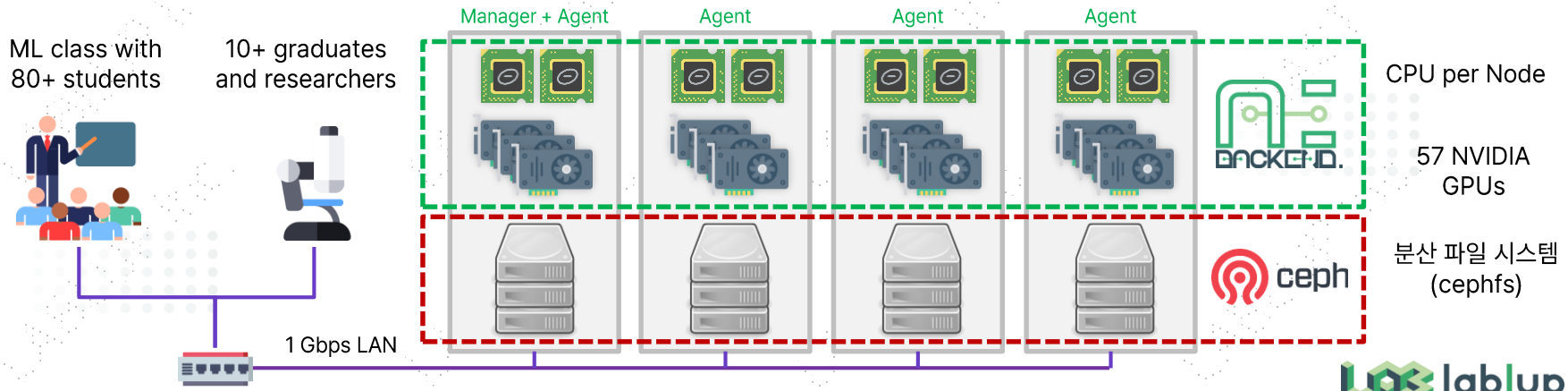
AI 트랜스포메이션: Backend.AI의 대응

- **자사 도메인과 AI의 결합**
  - 도메인별 모델 개발 파이프라인 템플릿 (지속적 추가)
  - 딥러닝 기반 시각화 도구 및 BI 도구들과의 유기적 결합
- **투자 부담 경감**
  - CPU 기반 K8s 자원의 통합 관리 및 모니터링
  - GPU 자원을 GPU 분할 가상화 및 Backend.AI Container Pilot 기반으로 최적화
  - 플러그인 기반의 LDAP 및 SSO 인증
- **쉬운 사용자 환경**
  - 다양한 딥 러닝 / 빅데이터 연구 소프트웨어 환경 제공 및 지속적 업데이트
  - AI/MLOps, 소프트웨어 환경 빌드 솔루션 개발 및 제공
- **확장성 확보**
  - 새로운 하드웨어 및 패러다임의 선도적 수용
  - 플러그인 형태의 이기종 AI 가속기 지원 / 다양한 스토리지 벤더의 가속 통합 제공

Backend.AI 사례 : 대학A (연구/실습 워크로드 통합 제공)

- **금융권 실무자 MBA 수강생 및 연구자들을 위한 GPU 서버팜 구축**
  - 초기: 2대 서버 16개의 GPU로 80명 이상의 수강생들이 동시에 실습
  - 다수 연구자들은 동일 시스템 내 별도 자원 그룹에서 모델링 수행

- **구성**
  - 7 노드 / 이종 GPU 조합 (총 57유닛 / 3종류)
  - 노드별 대용량 HDD를 LAN으로 묶어 18 TiB ceph 분산 파일시스템 구축 및 연동
  - 시기에 따라 순차적 확장 (2019년, 2021년, 2022년)

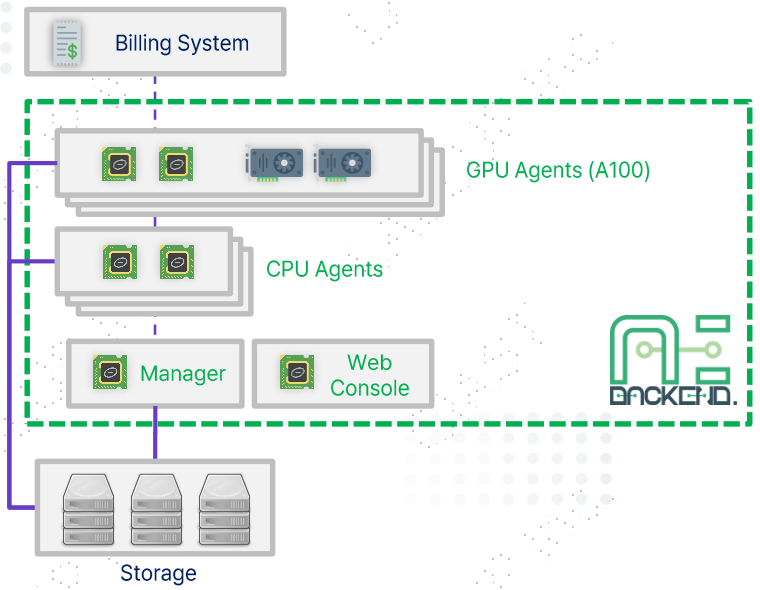


Backend.AI 사례 : 대학A (연구/실습 워크로드 통합 제공)

- **결과**
  - 제한된 갯수의 GPU로 훨씬 많은 인원 수의 사용자들이 동시에 사용
  - 수강생·연구자로 구분된 resource policy 적용
  - 학내 사용 가능한 Web GUI로 주요 기능을 제공하여 전담 관리자 없이도 최적 운영
  - 학기 중: AI 수업 및 연구 워크로드를 시간대에 따라 탄력 조정하며 사용
  - 방학 기간: 기업 AI 위탁 교육 플랫폼으로 이용

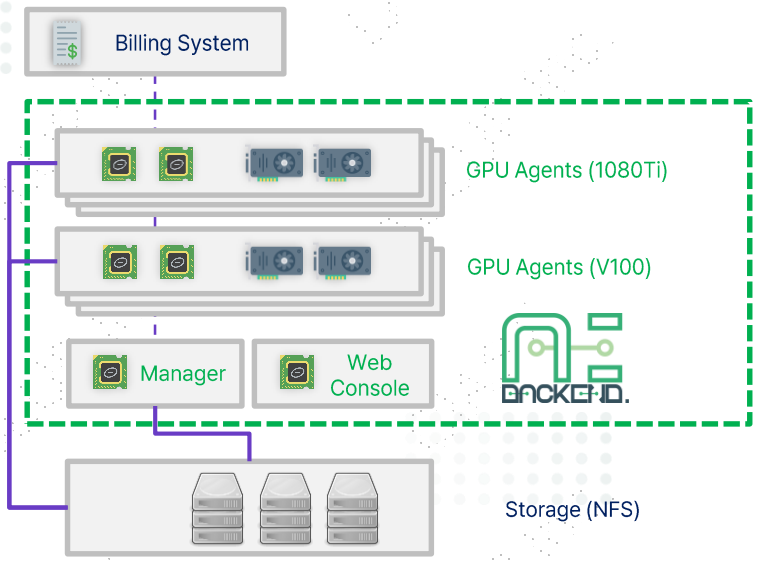
Backend.AI 사례 : 대학B (대학 구성원들을 위한 슈퍼컴퓨팅센터 구축)

- **학내 AI 클라우드 플랫폼으로 전 구성원 사용**
  - https://www.skku.edu/skku/campus/skk_comm/popup_news.do?mode=view&articleNo=96929
- **구성**
  - A100 GPU 40유닛 이상
- **혜택**
  - 수업, 교육, 연구등의 다양한 용도에 따른 학내 AI/ML 자원의 효율적 사용
  - 사용자 및 프로젝트 별 자원량 할당 기반의 수업 및 연구 과제, 일반 연구 목적으로 탄력적 운영
  - 할당 및 실사용 메트릭 기반의 학내 과금 시스템과의 통합
    - ✓ 자체 AI 클라우드 운영 및 과금을 학내에서 가능하도록 함



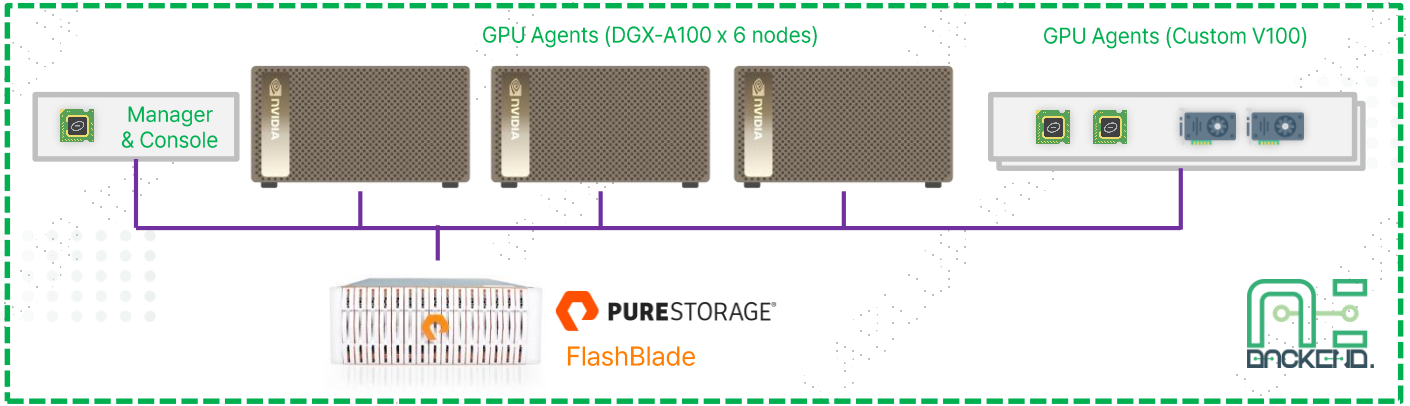
Backend.AI 사례 : 그룹C (사내 AI플랫폼 / 내외부 리소스 제공)

- **사내 AI 플랫폼 구축**
- **구성**
  - 소규모 구성으로 시작 / 현재 GPU 100유닛 이상
  - 연산 자원 규모 지속적으로 확장 중
- **혜택**
  - 대규모 GPU 클러스터의 활용성 증대
  - 비교적 구매 시점이 지난 GPU (3년 이상) 리소스와 현세대 GPU 리소스의 통합 관리
  - 사용자 환경의 변화 없는 시스템 확장
  - 사용 메트릭 기반의 리소스 재배분 및 사내 과금 시스템 연동



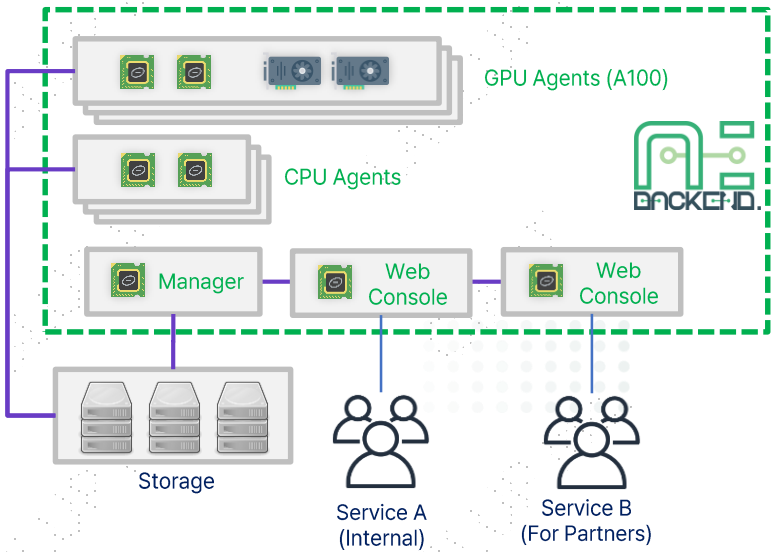
Backend.AI 사례 : 그룹D (사내 AI 연구 개발 플랫폼)

- **최신 GPU의 활용성 증대 및 기존 리소스와의 통합 관리**
- **구성**
  - 도입 규모 : 현재 GPU 48유닛 / 연산 자원 규모 지속적으로 확장 중
- **혜택**
  - 가장 최신의 GPU 노드들을 기존 사용 노드들과 통합 관리하여, 연산 자원 사용 효율을 대폭 상승
  - 기존 분석/모델링 환경을 계속 유지하면서 새로운 하드웨어의 혜택을 누릴 수 있음



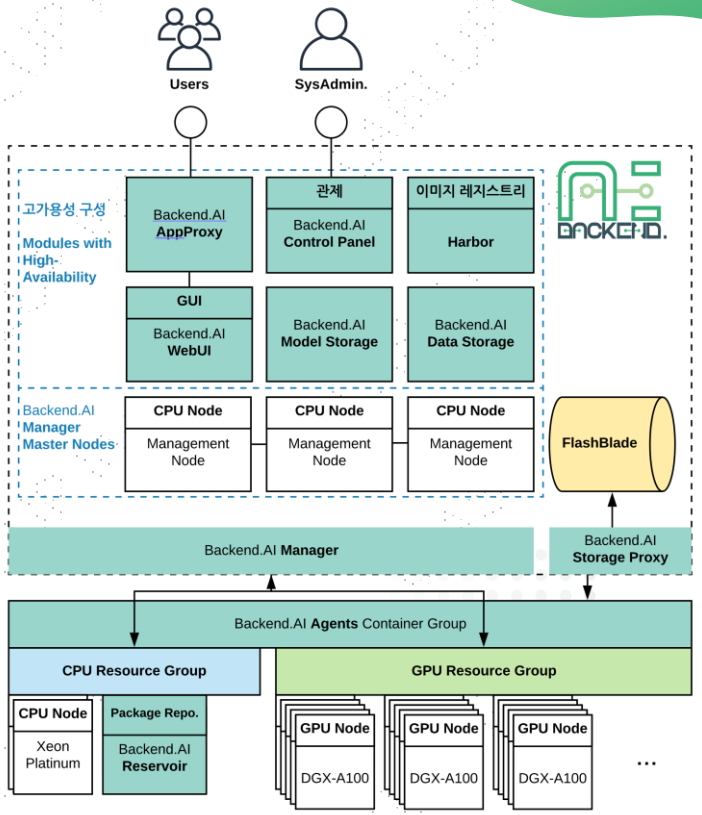
Backend.AI 사례 : 그룹E (사내 AI플랫폼 / 내외부 리소스 제공)

- **사내/사외 AI 협업 클라우드 플랫폼**
- **구성**
  - A100 GPU 300유닛 이상
  - 빅데이터 처리용 CPU 노드 통합 관리 / 운영계 삼중화 구성
  - 데이터, 컨트롤, GPU간 통신용 네트워크를 별도 구축 및 사용
- **혜택**
  - 분산 훈련 기반의 초대규모 딥 러닝 모델 개발 및 서비스 제공
  - 사용 메트릭 기반의 개인 및 팀별 사용 메트릭 측정 및 정기적 자원 재분배
  - 공동 작업등을 위하여 외부 사용자에게 서비스 제공시 동일한 하드웨어 팜으로 보안 격리된 독립 사용자 경험 제공 (Backend.AI 도메인 기능)



Backend.AI 사례 : 해외A (대단위 개방형 AI 연구개발 플랫폼)

- **멀티 리전, 다중 조직 사용자를 위한 AI 개발 클러스터**
- **구성**
  - (GPU 워크로드용) A100 GPU 200유닛
    (빅데이터 분석 및 데이터 전처리용) 고성능 CPU 노드 다수
  - 주기적 업데이트 및 사용자별 패키지 설치를 포함한 **완전 Air-gapped 환경 구축**
- **고객 혜택**
  - 대단위 팜 구성 설계 제공 및 SLA 극대화 고가용성 구성
  - 멀티노드 분산 훈련 및 GPU간 직접 네트워크 기반의 대규모 초고속 딥러닝 훈련
  - 로컬 패키지 저장소 솔루션인 Backend.AI Reservoir와 결합하여 PyPI 및 Ubuntu 저장소를 완전 폐쇄망 내에서 자유롭게 사용 지원
  - 기관 내외부 동시 서비스 시 시스템/데이터 보안을 위한 격리 도메인 구성



cloud.Backend.AI: Backend.AI 를 모두에게!

- **설치의 어려움, GPU 자원 구입에 드는 비싼 비용, 모델 서비스시의 어려움을 해결하기 위한 해법**

Personal Cloud[BETA]

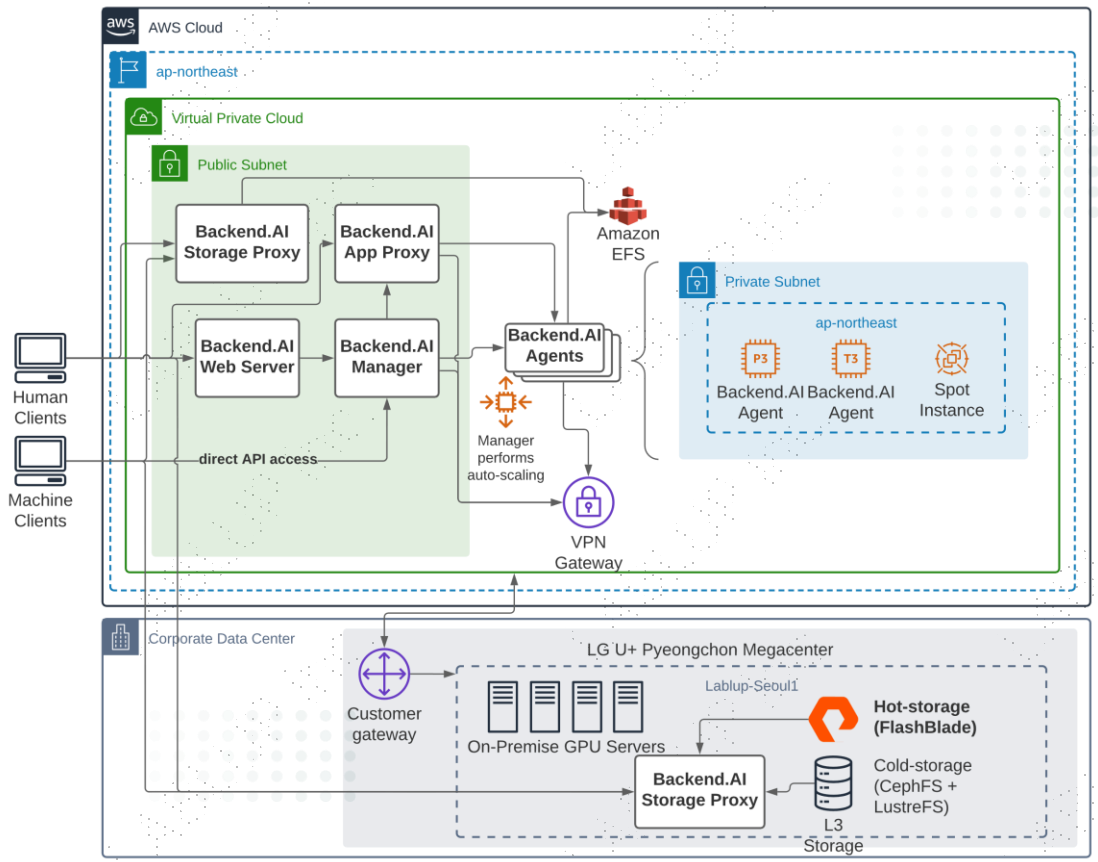
- **관리의 부담 없이** 바로 시작
- 모델을 개발하고 데이터를 공유하는 과정을 **최대한 간단하게**
- **원하는 만큼만** 자원을 사용하고, 필요하면 늘려 사용
- 베타 테스트 (2020~) / 2022년 중 정식 오픈 예정

Team Cloud[ALPHA*]

- 누구나 쉽게 만드는 자신만의 AI 팀
- 팀 개발 + 커뮤니케이션 + 공유 + AI서비스를 하나의 서비스로
- **World Console:** 내 PC들을 연결해서 클라우드에서 사용
- 작은 규모로 시작하여 초 대규모 AI 개발까지 쉽게 확장

*협업 그룹 대상 우선 테스트

Cloud.Backend.AI: 멀티클라우드+하이브리드 클라우드



- **구성**
  - AWS + Azure + GCP 통합 (평균 20 노드 / 가변 150 노드)
- **GPU / TPU**
  - IDC내 GPU 리소스 기본 사용
  - 로드 초과 예상시 AWS로 오토스케일
  - Google TPUv3 지원
- **데이터 저장**
  - AWS EFS (Elastic File System) + Lablup L3 On-prem 스토리지 연동 (Hot + Cold)
  - 분리된 자원 그룹간 데이터 캐싱 (Backend.AI 내장 기능)

lablup

마치며

- **Backend.AI: 요약**
  - 개발 계기 / 목표
  - 핵심 요소 소개
  - 오픈소스 및 엔터프라이즈 버전
- **Backend.AI: AI 트랜스포메이션 사례**
  - AI 전환의 요구 사항
  - 시스템 구성 개요
  - 도입 케이스 요약
- **Backend.AI 클라우드: 꽃들에게 희망을**

오픈소스부터 사람을 위한 AI까지

- **왜 이 일을 하고 있는가?**

기업용 AI 개발 플랫폼
세계 최고 성능의 엔드투엔드 AI 소프트웨어

Backend.AI Enterprise : 토탈 AI/ML 솔루션

초거대 AI 모델 빌드 / 개발 플랫폼
교육, 프로토타이핑부터 프로덕션 AI/ML 모델까지
데이터를 모델로, 모델을 서비스로 연결하는 **통합 플랫폼**

AI 서비스 플랫폼 / 클라우드
전세계인 모두를 위한 AI 서비스

Backend.AI Cloud SaaS : AI를 위한 빌딩 블록

모든 IT 솔루션들을 위한 **AI 빌딩 블록**
데이터 + 클라우드 + 모델
어디에나 결합 가능한 범용적 지능 서비스 제공

AI 를 어디에서나, 누구나 사용할 수 있도록 합니다.

감사합니다.

contact@lablup.com

https://www.facebook.com/lablupInc

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