TEN

회사소개서

Make AI accessible for all.

회사명	주식회사 텐 TEN Inc.
대표이사	오세진
설립일	2020.07.03
주소	서울특별시 강남구 테헤란로27길 18, 5층
주요 서비스	MLOps 솔루션 ‘AI Pub’ AI 인프라 컨설팅 ‘RA:X’
직원 수	27명
홈페이지 / SNS	https://ten1010.io/ https://www.youtube.com/@ten1300 https://www.linkedin.com/company/ten1010/

2020년 서비스를 시작하여
각종 사업 선정과 어워드 수상으로 경쟁력과 잠재력을 인정받고 있습니다.

2020

02 AI Pub 서비스로 비즈니스 시작
07 주식회사 텐 법인 설립

2021

03 NVIDIA Inception Program 체결
04 Kingsley Ventures Seed 투자 유치
06 벤처기업 인증 획득
07 TIPS 창업성장기술개발사업 선정
08 기업 부설 연구소 설립
08 대한민국 리딩기업대상 K-스타트업대상 'AI 서비스 부문' 수상
11 NVIDIA GTC³ Top Startup from Korea 선정

2022

03 AI 바우처 사업 선정
07 대한민국 리딩기업대상 K-스타트업대상 'MLOps 플랫폼 부문'
11 한국평가데이터 기술역량 우수기업 인증 (T4 등급)
12 Pre-series A 라운드 투자(30억 규모) 유치

2023

01 가상화 리소스 관리, 크기 추천 방법 등 특허 5개 등록
02 TEN 제1회 AI 워크샵 개최
07 NetApp Preferred Partnership 체결
08 IBM Partnership 체결
09 Redhat Partnership 체결
11 한국국방과학연구소 딥러닝 연산용 고밀도 GPU기반 클러스터 컴퓨팅 시스템 사업 구축 완료
11 TEN 제2회 AI워크샵 개최
12 AI Pub GS 인증 획득
12 CIO Review '2023년 가장 유망한 한국 테크기업' 선정

2024

01 성균관대 SAL・COMPASS LAB AI 인프라 산학협력 체결
04 코오롱 베니트 AI Alliance MOU 체결
05 제19회 디지털 이노베이션 대상 IT부문 수상

2025

05 Series A 라운드 투자(80억 규모) 유치
05 2025 디지털이노베이션 대상 선정
08 NVIDIA Partner Network – Solution Advisor 체결

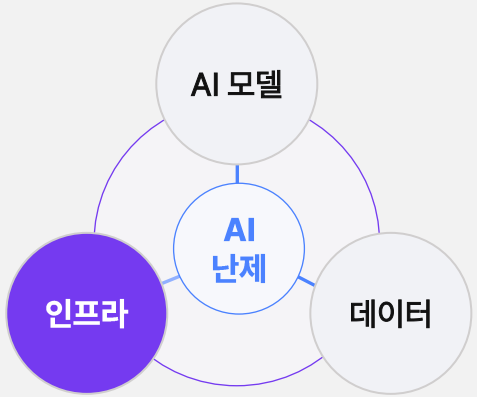
주식회사 텐은

Mission

세상을 널리 AI롭게 하기 위해

Vision 2024

기술의 장벽을 낮추는 소프트웨어 도구를 만듭니다.



특정 지식인이 아닌 모든 이가 AI를 만든다는 인력 보편화의 조건과
자신의 지식을 데이터로 구체화하면 AI를 만들 수 있다는 데이터 보편화의 조건은 점차 해결되고 있지만,
인프라는 철저하게 비용의 문제라 시간과 노력만으로는 해결하기 어렵습니다.

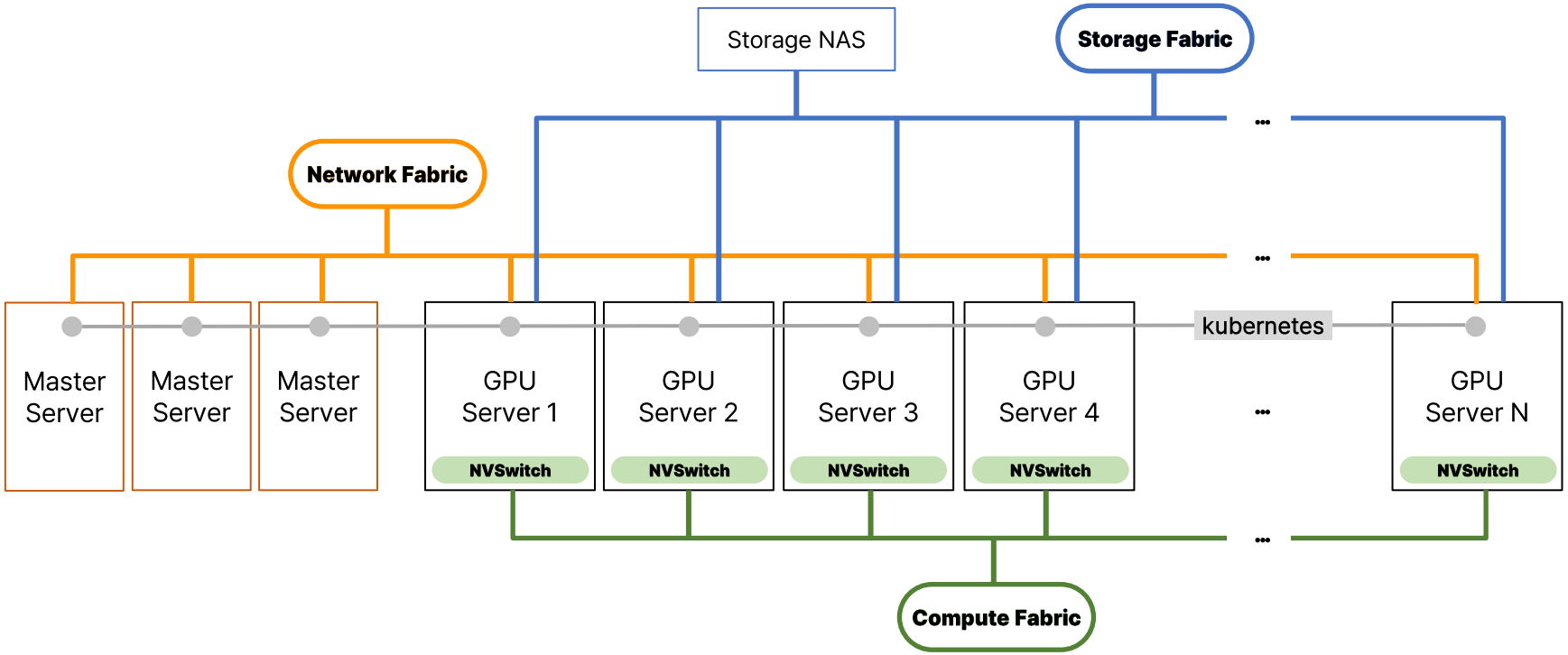
머신러닝 박사 학위가 없어도
AI을 만들 수 있는 시대

누구나 자신의 지식을
데이터로 구체화하고
이를 학습시킬 수 있음

인프라 = 비용

그러나 인프라 자원을 효율화하는 것은 어려운 일입니다.
AI 모델의 자원 수요가 지속적으로 증가함에 따라, 인프라 구성 역시 복잡해지기 때문입니다.

AI 모델 계산량 증가를 커버하기 위해, 다수의 GPU 머신을 도입할 수밖에 없고
이에 따른 AI 전용 인프라 구성은 더욱 복잡하고 어려워집니다.

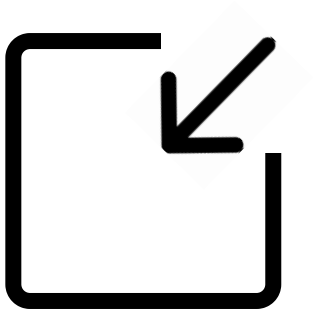


인프라 자원의 효율화는
AI 보편화의 핵심입니다.

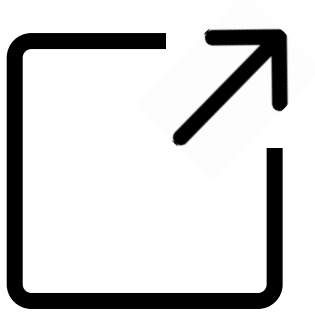
AI 구현은 얼마나 효율적으로 AI 인프라를 구축하고 쓸 수 있는 가에 좌우됨

AI의 등장은 기존 컴퓨팅 패러다임의 혁신적 변화를 초래했음
혁신적 변화의 핵심은 GPU 뿐 아니라 다른 HW 요소들 포함한 인프라의 통합적 성능 향상이 필수적이라는 것임

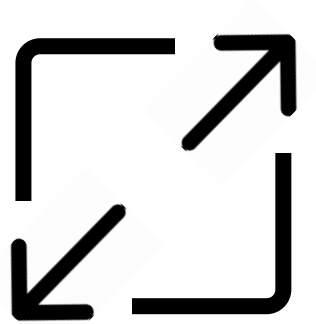
[AI로 인한 컴퓨팅 패러다임의 변혁]



서버 **내부**의 변혁



서버 **외부**의 변혁

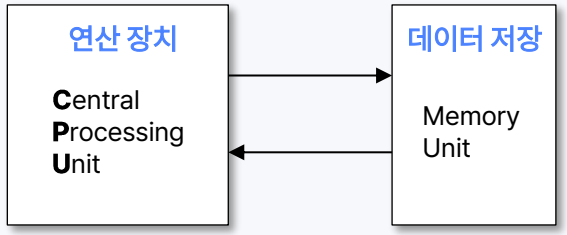


자원분배방식의 변혁

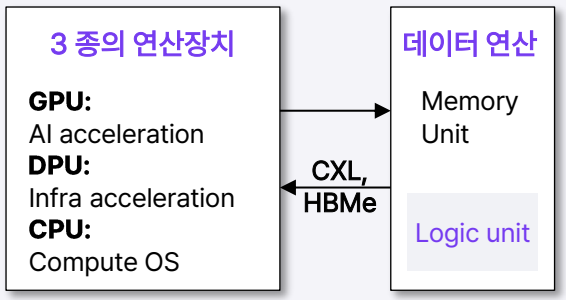
[AI로 인한 컴퓨팅 패러다임의 변혁 – 상세]

서버 내부의 변혁

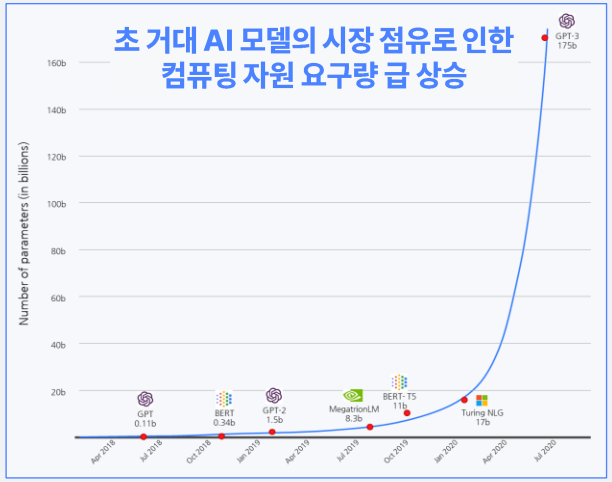
폰 노이만 구조
과거 컴퓨터 시스템의 기본 구조



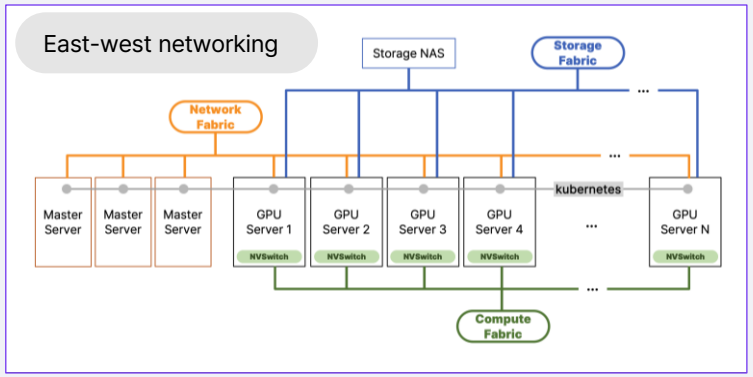
AI 데이터 센터 용 컴퓨터 시스템 구조
각기 다른 연산을 담당하는 3개의 처리 장치 &
연산용 프로세스를 더한 메모리 구조



서버 외부의 변혁



초 거대 AI 모델의 계산량을 커버하기 위해 다수의 서버를
클러스터로 구성하며, 그 사용 방법이 복잡해지고 있음

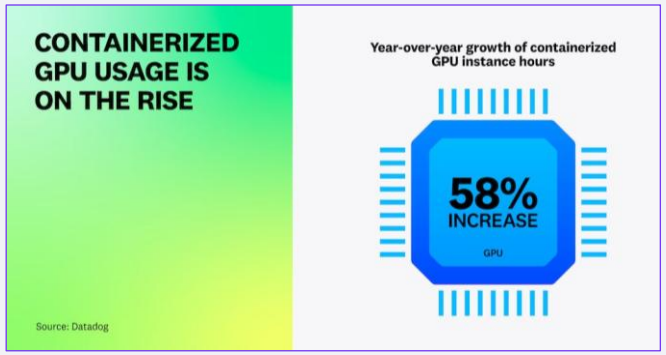


자원 분배 방식의 변혁

컨테이너 기술이 빠르게 확산함에 따라
컨테이너 플랫폼 쿠버네티스가
모던 인프라기술의 사실상 표준으로 부상

kubernetes

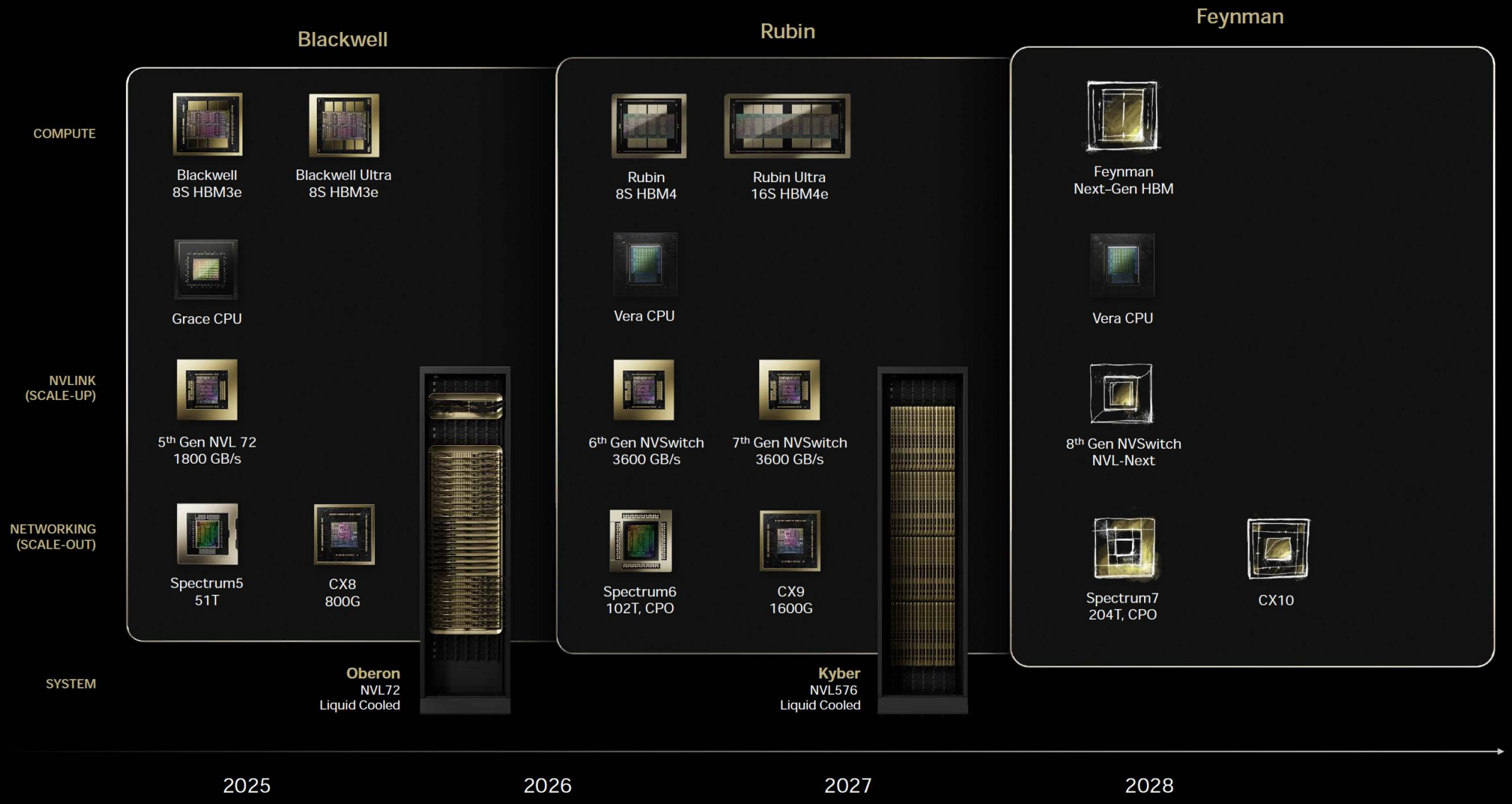
컨테이너화된 GPU 컴퓨팅이 YoY 58% 증가
AI의 방대한 데이터 처리 규모로 인해
컨테이너를 기반으로 한 GPU 컴퓨팅이 빠르게 성장 중

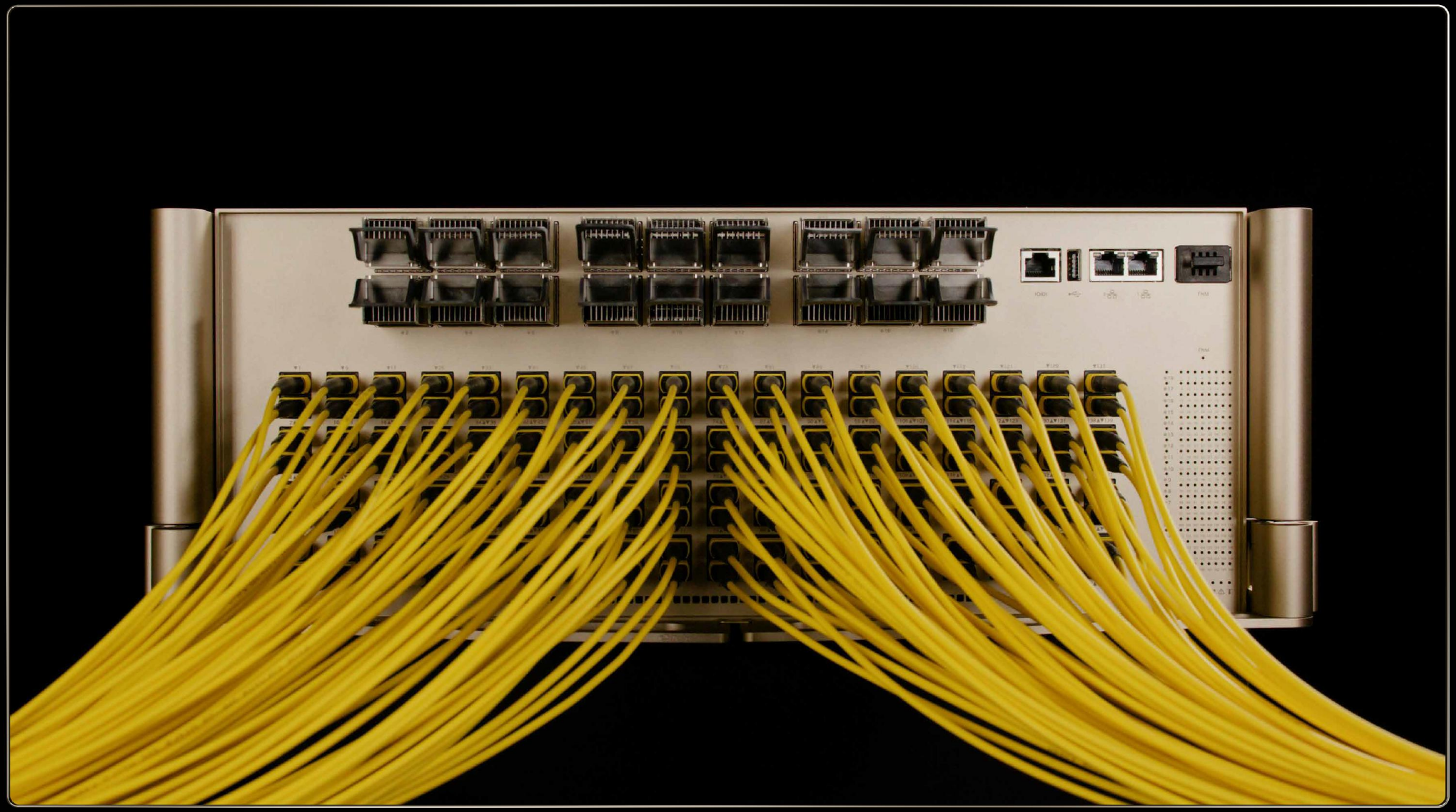


[출처] 10 Insights on real-world container use, Nov 2023 Datalog

NVIDIA Paves Road to Gigawatt AI Factories

One-Year Rhythm | Full-Stack | One Architecture | CUDA Everywhere





TEN은,
효율적인 솔루션을 제시합니다.

TEN은 AI를 잘 개발하고 운영할 수 있도록,
MLOps 소프트웨어와 AI 전용 인프라를 서비스하고 있습니다.

SW

Tools

MLOps

HW

Infrastructure

Server, GPU,
Network, Storage

Holistically served by TEN

Knowledge

분야전문가 & 데이터

AI 학습과 운영에서 인프라 활용 방식이 기존 IT 서비스와 다르다는 것이 AI 난제의 시작

모델 학습을 위한 GPU 인프라 문제

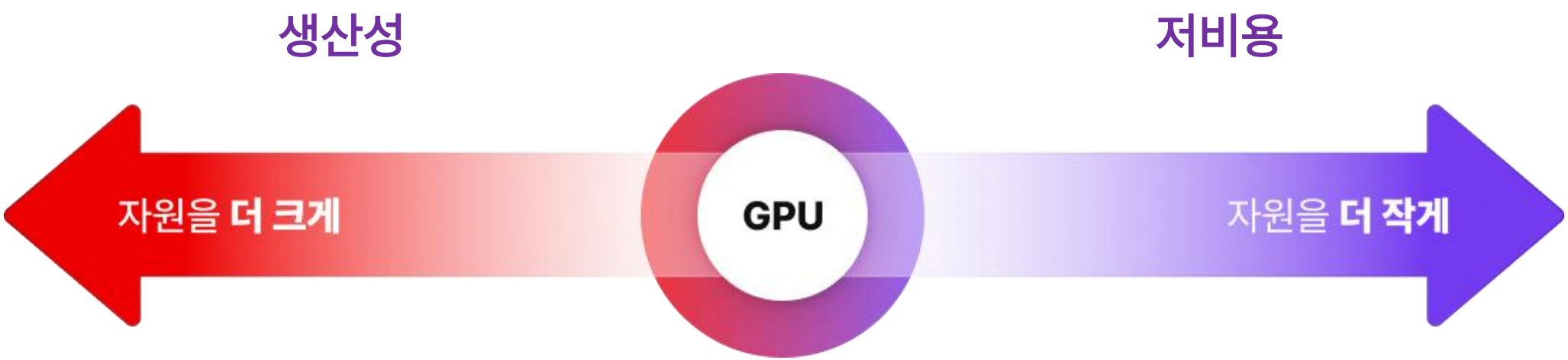
24 X 7, GPU 자원 가동률 극대화

- AI 학습의 계산량을 위해 다수의 GPU와 노드를 '묶어서' 사용 해야 함
- GPU와 노드를 '묶어' 쓰기 위해서는 네트워크와 스토리지 역할이 중요

모델 운영을 위한 GPU 인프라 문제

최소한의 GPU 사용으로 비용절감과 운영 품질 확보

- AI 추론의 계산량은 적은 GPU 자원으로도 처리 가능
- AI 비즈니스 전개 시 인프라 비용 최소화가 사업의 성패를 가름
  AI 서비스 원가로써 장기적으로 영향을 미침



차세대 AI 워크로드 오케스트레이션 플랫폼 AI Pub의 제품 라인업

K8s 관리를 위한

AI Pub K10s

for

Kubernetes에 능숙한 개발자
DevOps 및 MLOps 개발자

Kubernetes 오브젝트 관리

AI 개발을 위한

AI Pub Dev

AI 운영을 위한

AI Pub Ops

for

Kubernetes에 익숙하지 않은
AI 모델러 혹은 운영 관리자

인공지능 학습 오브젝트 관리

AI 인프라 구축을 위한
컨설팅 서비스

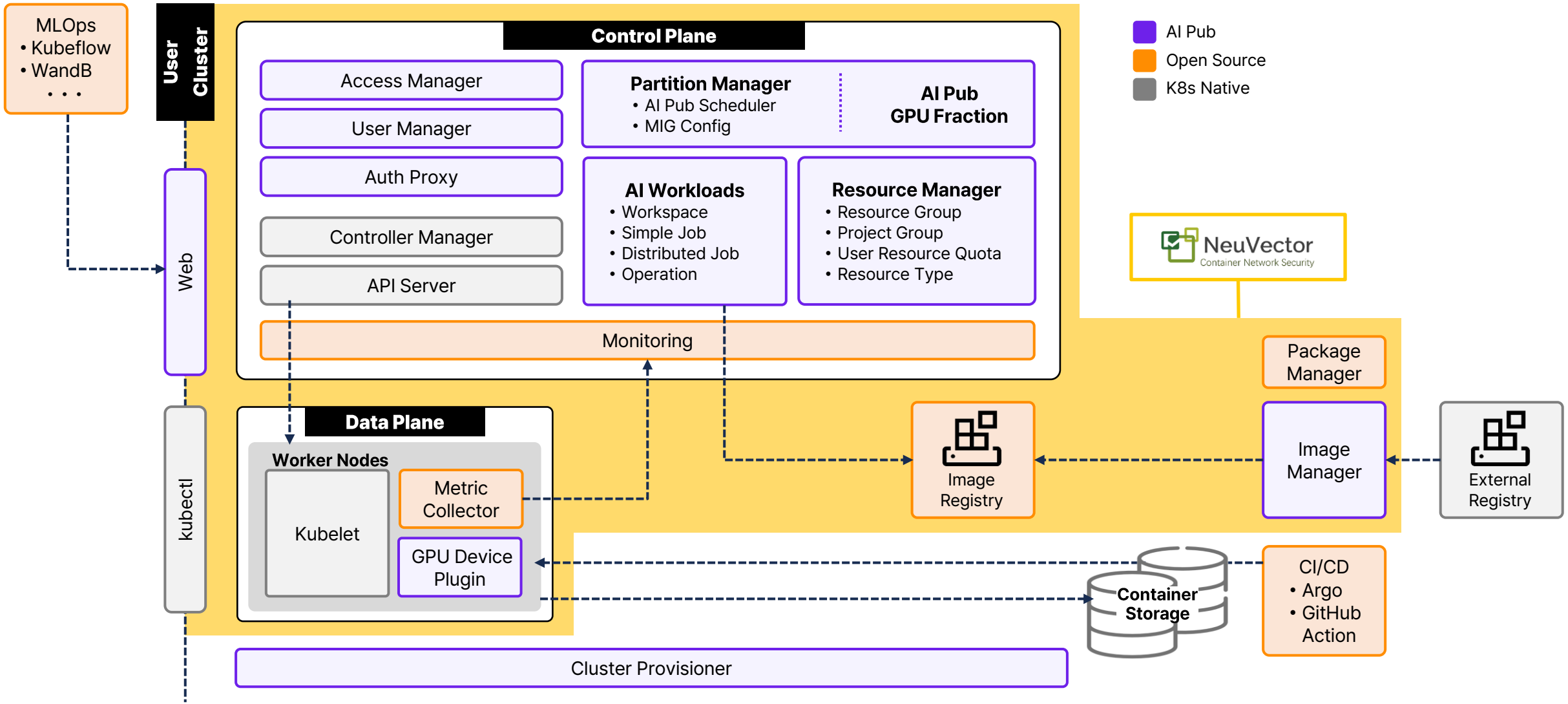
RA:X

for

AI인프라 신규 도입
또는 운영중인 기업

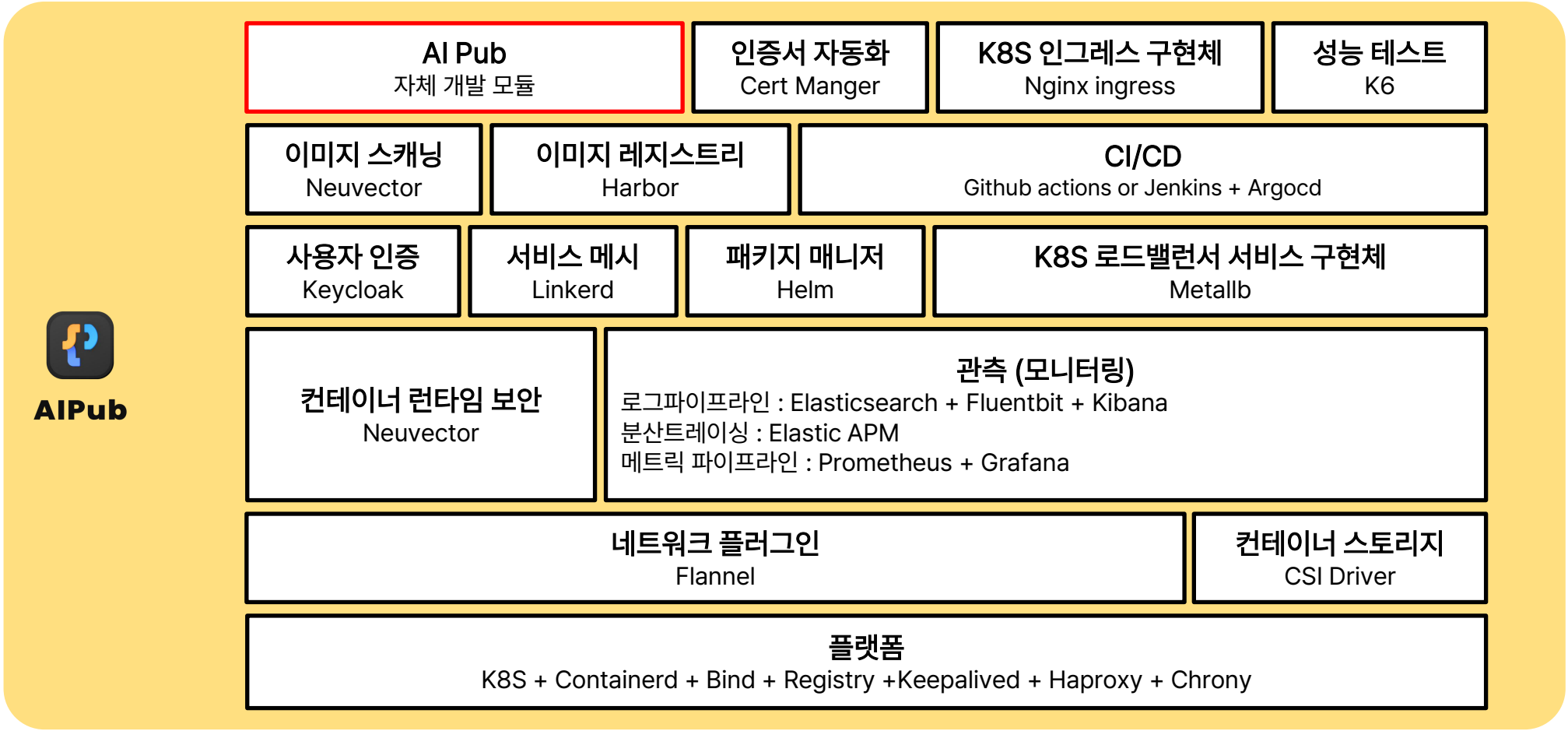
비용 효율적인 맞춤형 인프라 구성

Kubernetes 기반 AI Pub 플랫폼 구조



3. Kubernetes-Native Architecture

Kubernetes 에코 시스템의 검증된 도구들을 연동하여 서비스를 구현



GPU 활용률 극대화로 자원 낭비 최소화 및 비용 절감

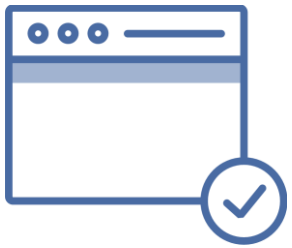
복잡한 환경 설정 없이 인프라 구축 시간을 획기적으로 단축하며, 최적의 AI 성능 및 비용 효율성 동시 확보

Ready-to-Accelerate

복잡한 환경 구성 없이 바로 사용 가능한 플랫폼으로 인프라 세팅과 AI 배포 기간을 획기적으로 단축

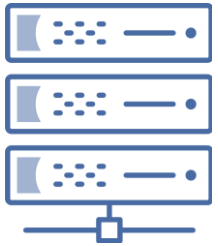
Maximized Efficiency

학습 단계에서는 GPU 자원을 최대한 활용하고, 운영 단계에서는 자원 사용을 최소화하여 비용은 줄이고 성능은 극대화



Faster to Market

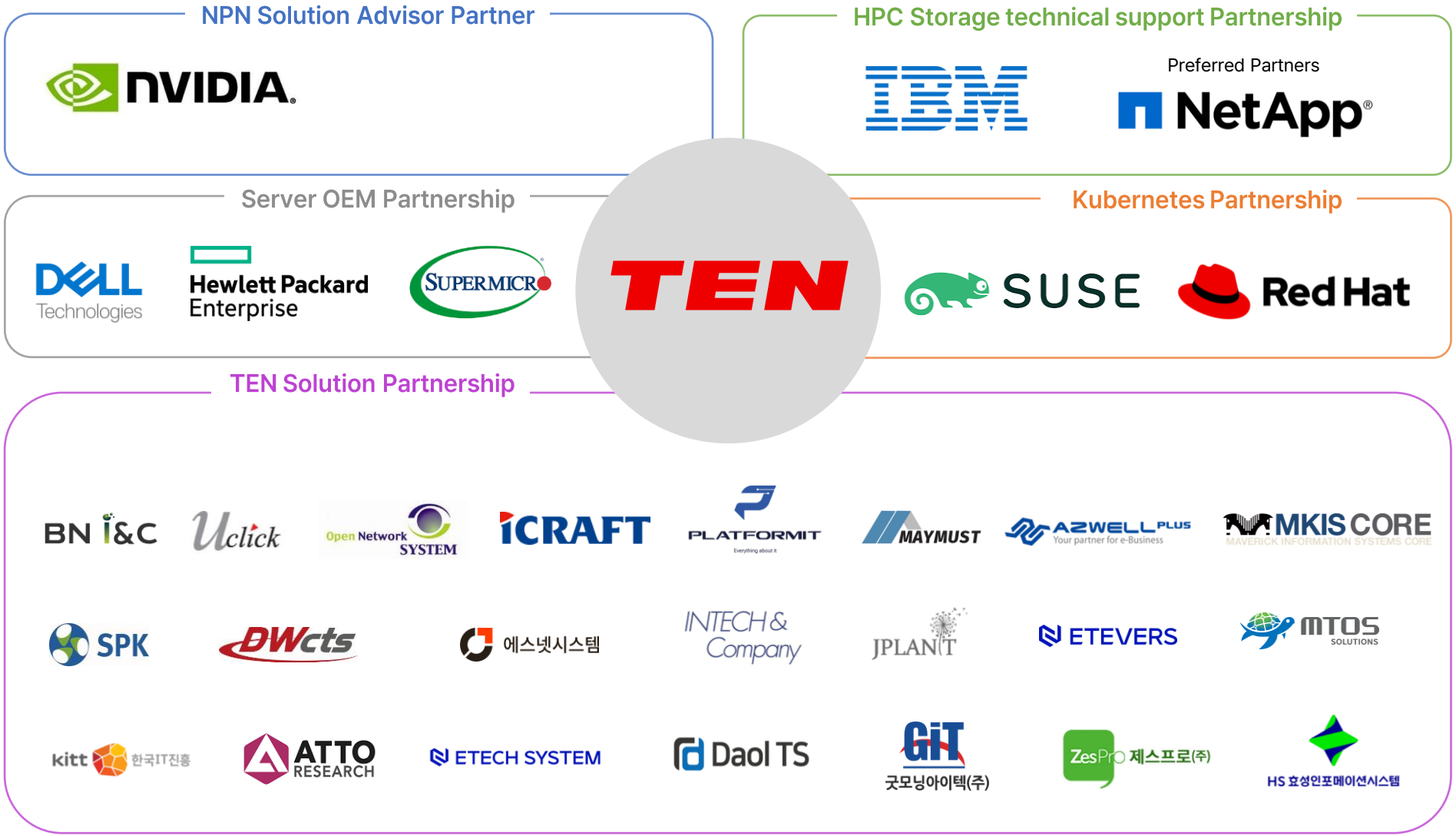
복수 워크로드, 사용자의 오케스트레이션
➔ 빠른 반복과 짧은 배포 주기를 실현



Control & Visibility

직관적인 사용자 인터페이스와 실시간 리소스 모니터링 기능을 통해, 인프라 상태를 파악하고 효율적으로 관리 가능

AI 전문 밴더들과 파트너십을 통해
AI 전용 서비스를 올인원으로 제공하고 있습니다.



다양한 분야의 기업 및 학교에서 TEN의 서비스를 이용하고 있습니다.

[Client] 정부 및 기관

국방과학연구소 AGENCY FOR DEFENSE DEVELOPMENT

세종특별자치시

KIMM 한국기계연구원 KOREA INSTITUTE OF MACHINERY & MATERIALS

[Client] 기업

SAMSUNG 삼성디스플레이

SAMSUNG SAMSUNG BIOLOGICS

SAMSUNG 삼성중공업

SAMSUNG SDI

HYUNDAI MOBIS

한화시스템

LG전자

LG디스플레이

NHN CLOUD

KB 국민카드

maum.ai

[Client] 교육

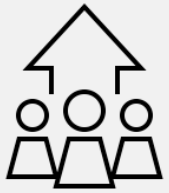
연세대학교 YONSEI UNIVERSITY

서강대학교 SOGANG UNIVERSITY

한양대학교 HANYANG UNIVERSITY

성균관대학교 SUNG KYUN KWAN UNIVERSITY

MLOps 관련 업체 중 국내 최초 Solution Advisor NVIDIA의 공식 파트너로 기술력과 서비스 역량을 인증 받았으며, NVIDIA의 다양한 솔루션을 통합하여 서비스를 제공할 수 있습니다.



Solution Advisor

Partners whose primary business model is to provide consultation services and expert advice to customers looking to implement NVIDIA products, NVIDIA-based solutions, and technologies.

NVIDIA Partner

Resource Management

GPU와 기타 컴퓨팅 자원을 얼마나 효과적으로 분배하고
제어할 수 있는지를 확인합니다.

1. GPU 분할 가상화
2. 자원 관리 기능

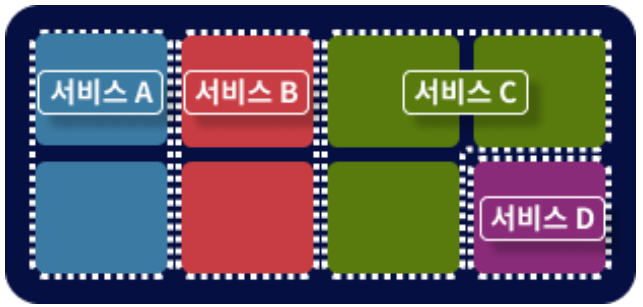
1. GPU 분할 가상화

- AI Pub은 Kubernetes-native 방식으로 GPU 분할 기능을 제공 (NVIDIA Volta 이상 GPU를 지원)
- Device Plug-in 및 K8s Extended Resource와 통합되어 자원 할당의 유연성과 가시성 제고

쿠버네티스 **기본 GPU 할당**

- GPU는 1개 단위로만 요청 가능
- 경량 워크로드도 전체 GPU 점유 → 자원 낭비
- 여러 서비스 간 공유 불가능
- 자원 충돌 제어 어려움

AI Pub의 **GPU 분할 할당 방식**



1개 GPU에 **여러 개의 서비스 운영**이 가능
서비스간 리소스 사용 침해를 막아 **안정성 확보**

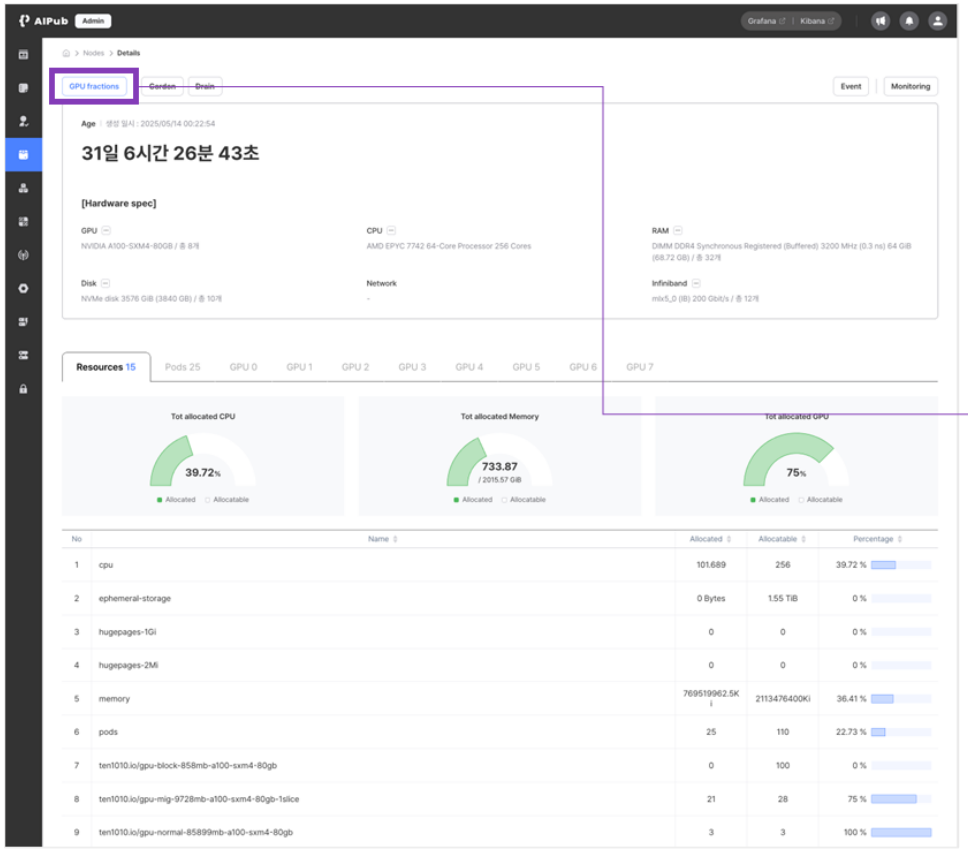
- GPU 1개의 코어와 메모리를 1% 단위로 분할, 100개 Block으로 분할
- 하나의 GPU를 여러 서비스가 동시에 사용
- Kubernetes 기반 리소스 스케줄링 가능
- 서비스 간 자원 침해 없이 안정적 운영

1. GPU 분할 가상화

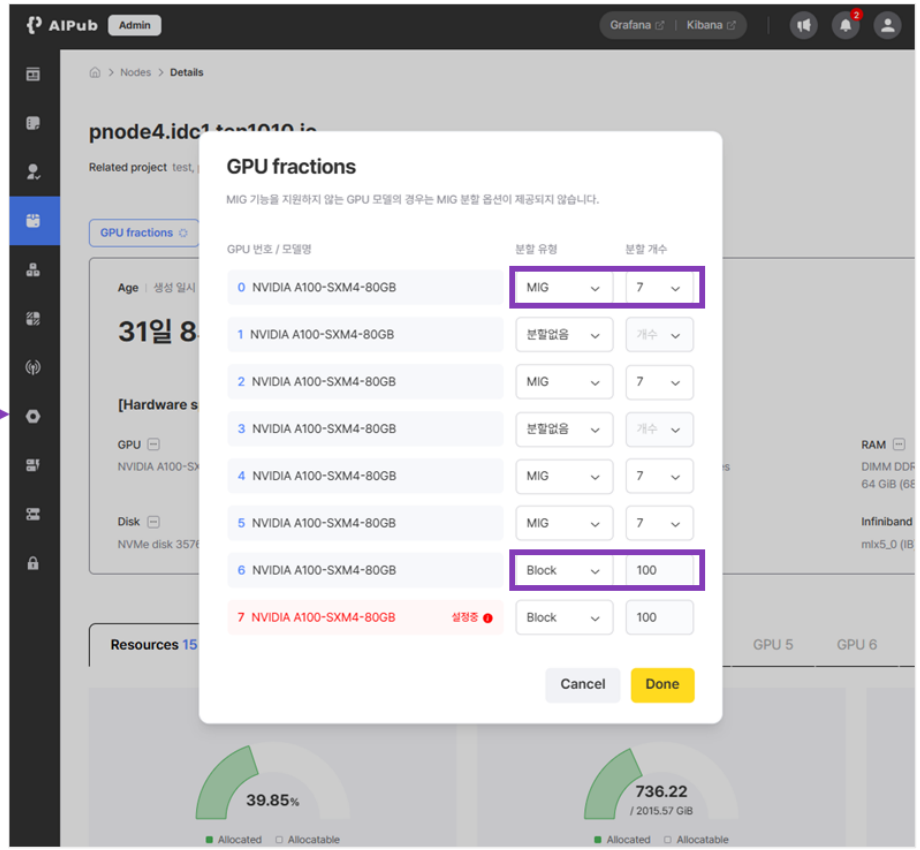
- **관리자 지정 GPU 분할 용이성**

- GPU 별로 UI를 통해 분할 설정 가능: 분할 타입 (분할 없음 / MIG 분할 / 블록 분할) 및 분할 개수 설정 가능
- MIG 분할 설정은 MIG 분할 가능한 GPU인 경우에 한하여 가능

[노드 상세 화면 > Resources 탭]

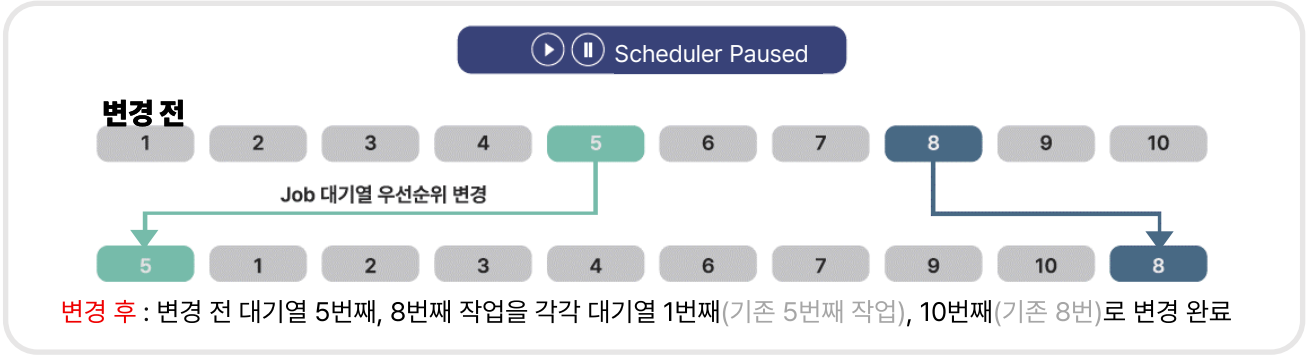


[GPU fractions 화면]

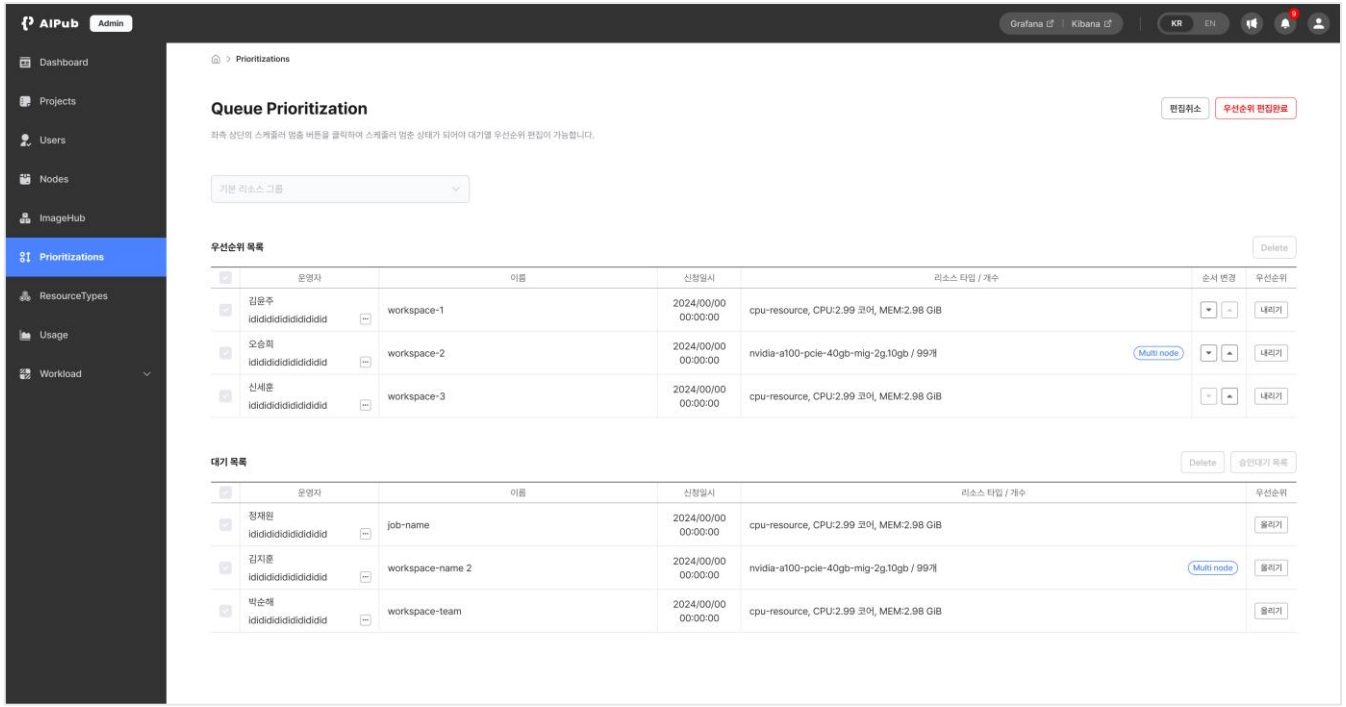


2. 자원관리 기능 - AI 워크로드에 최적화된 AI Pub 스케줄러

[워크로드 우선순위 조정 기능 예시]



변경 후 : 변경 전 대기열 5번째, 8번째 작업을 각각 대기열 1번째(기존 5번째 작업), 10번째(기존 8번)로 변경 완료



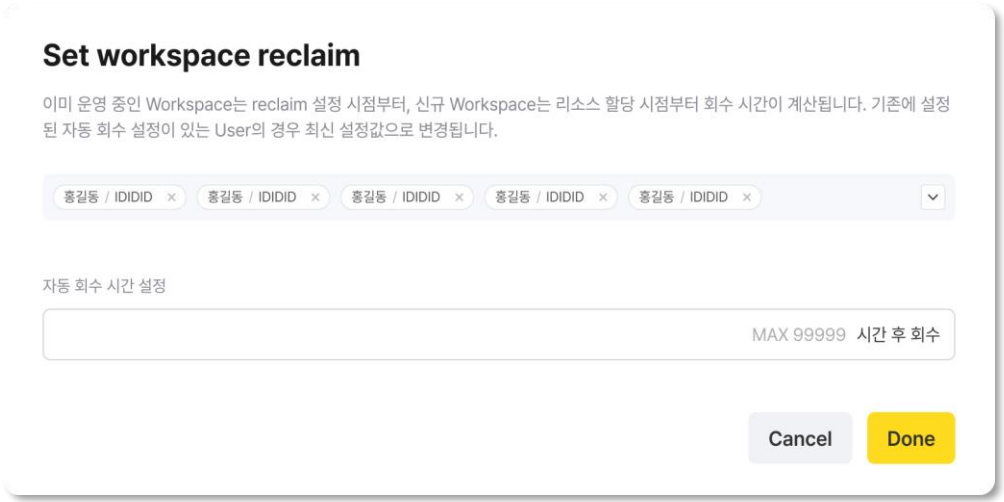
2. 자원관리 기능 - 정책 기반 리소스 관리

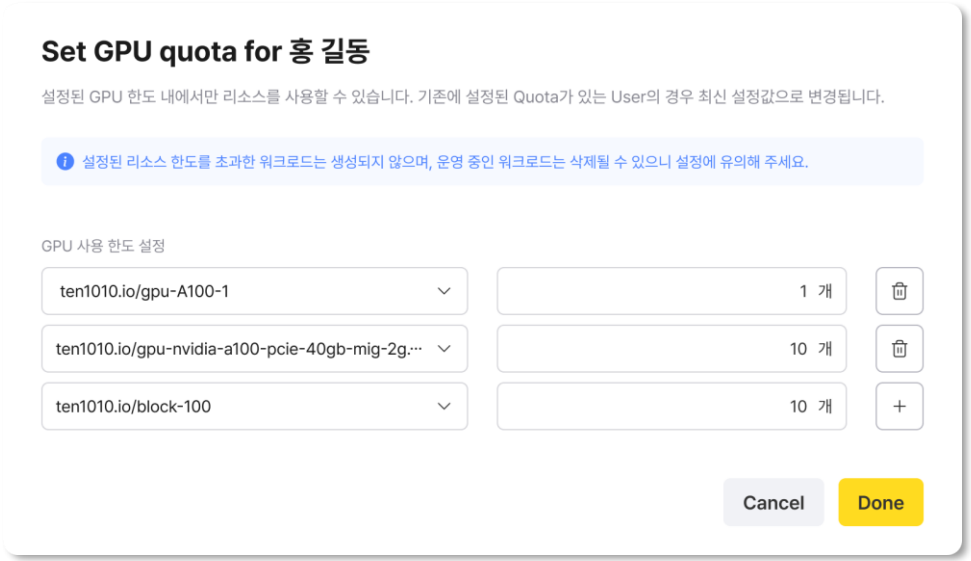
사용 정책에 따른 자원 할당 제어 및 자동 회수

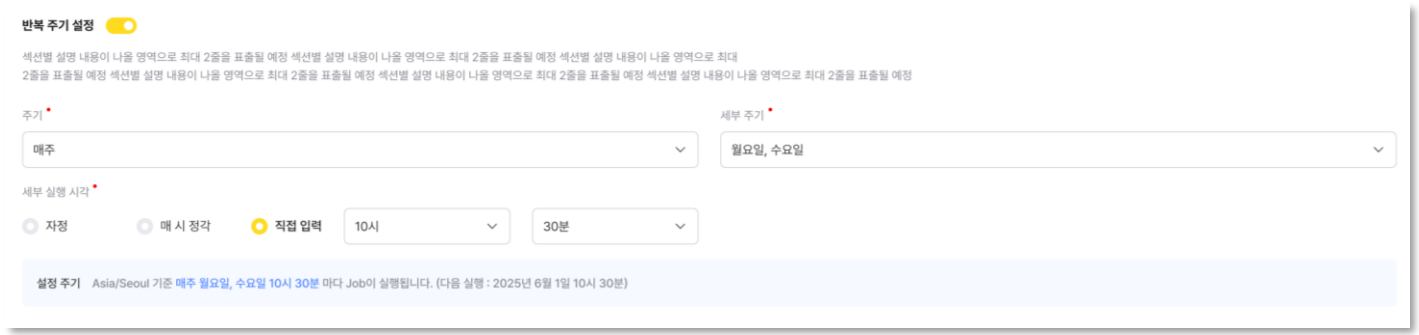
- **사용자 별, 프로젝트 별 자원 할당 제어**
  ResourceSet(GPU/CPU/RAM)에 대해 사용자 별, 프로젝트 별 사용한도 설정과 저장공간(Volume) 사용한도 설정 기능 제공

- **워크스페이스 회수 기능**
  - 관리자가 저 사용, 미사용 워크스페이스에 대한 강제 회수 조치 가능
  - 워크스페이스 생성 후 관리자가 지정한 시간 지나면 자동 회수 가능

- **배치 작업의 실행과 반복 회수 설정 가능**
  배치 작업을 예약 실행하고, 작업 종료 후 자원을 자동 회수함
  반복적인 연산 업무를 코드 없이 UI 로 손쉽게 자동화할 수 있음



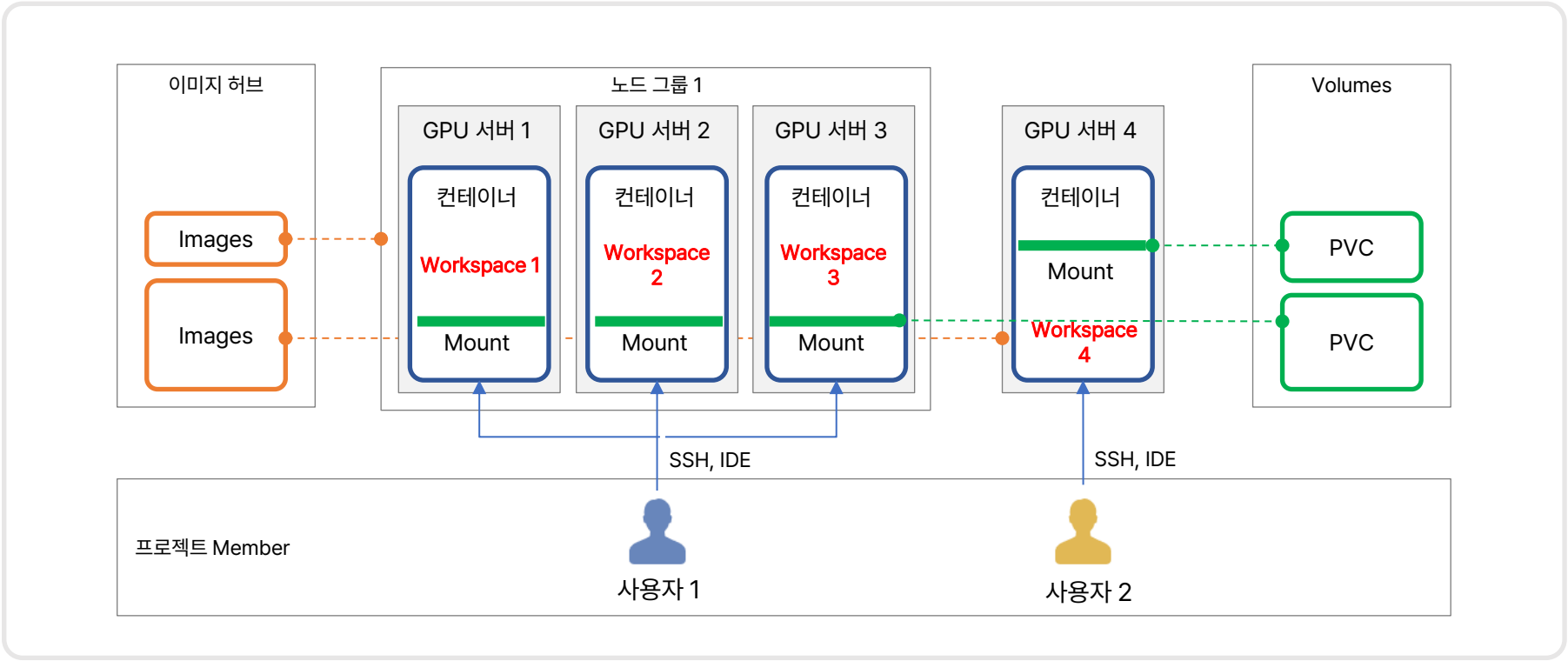




2. 자원관리 기능- 프로젝트 기반 자원과 역할 관리

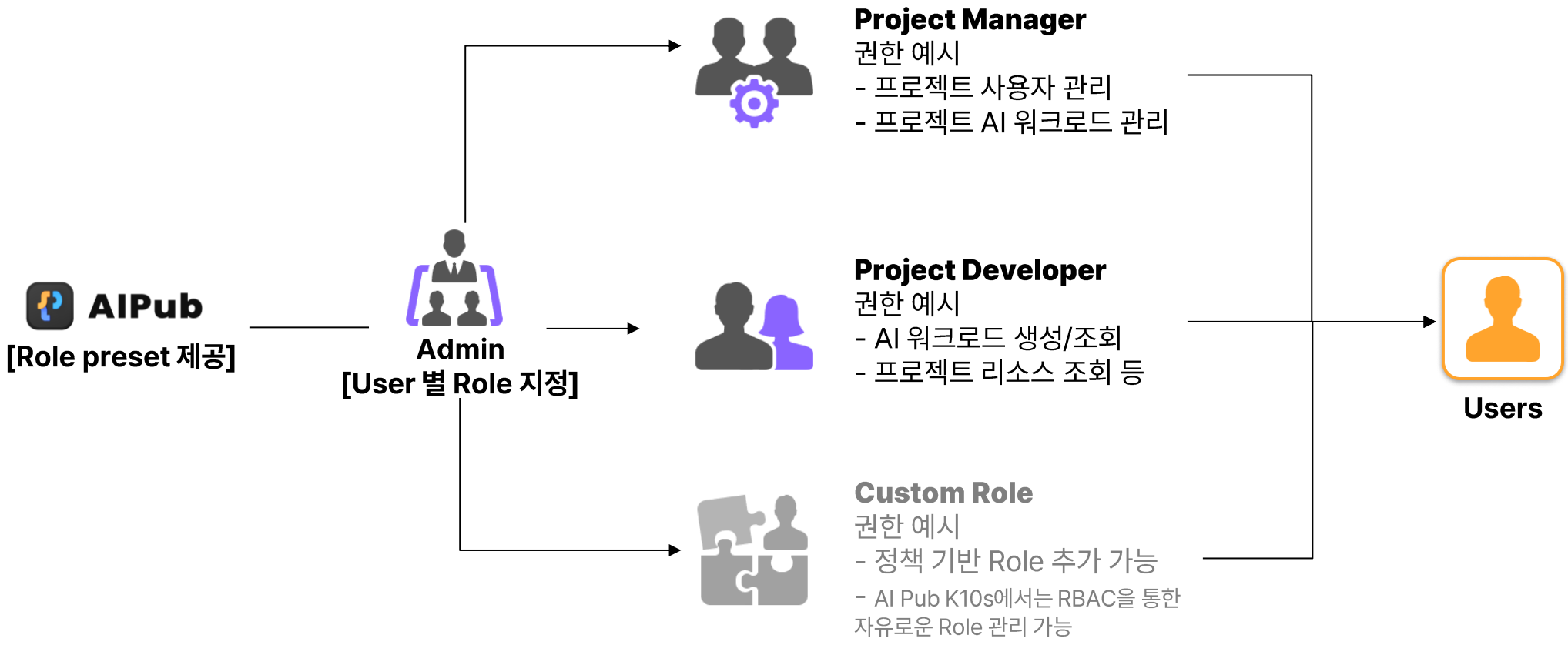
자체 개발한 '**프로젝트' 컨트롤러를 통해** 사용자와 자원을 유연하고 안전하게 매핑함
이를 통해 팀별, 목적별, 환경별로 **논리적으로 격리된 리소스 그룹을 구성**하고, 고도화된 자원 관리와 최적화를 지원

[프로젝트 별 리소스 접근 권한 관리]



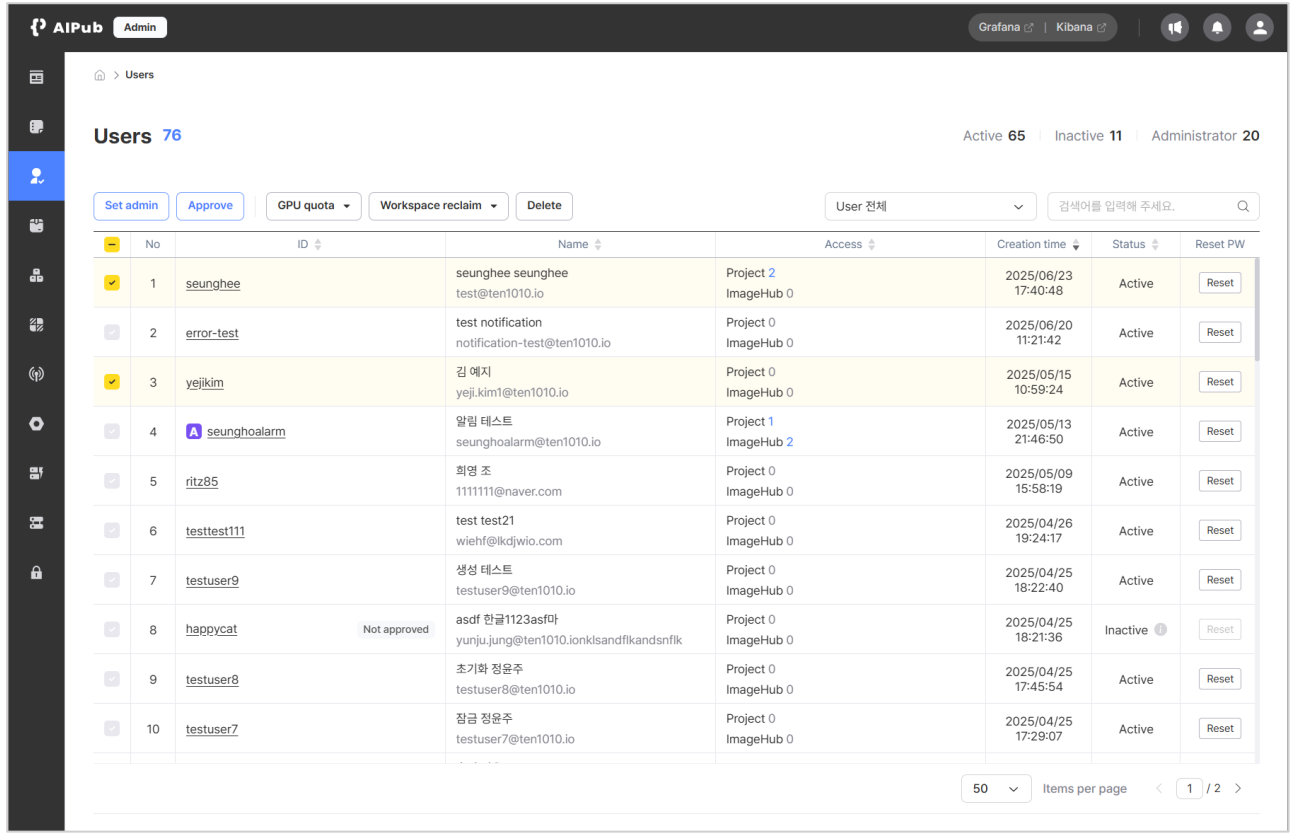
2. 자원관리 기능- 프로젝트 기반 자원과 역할 관리

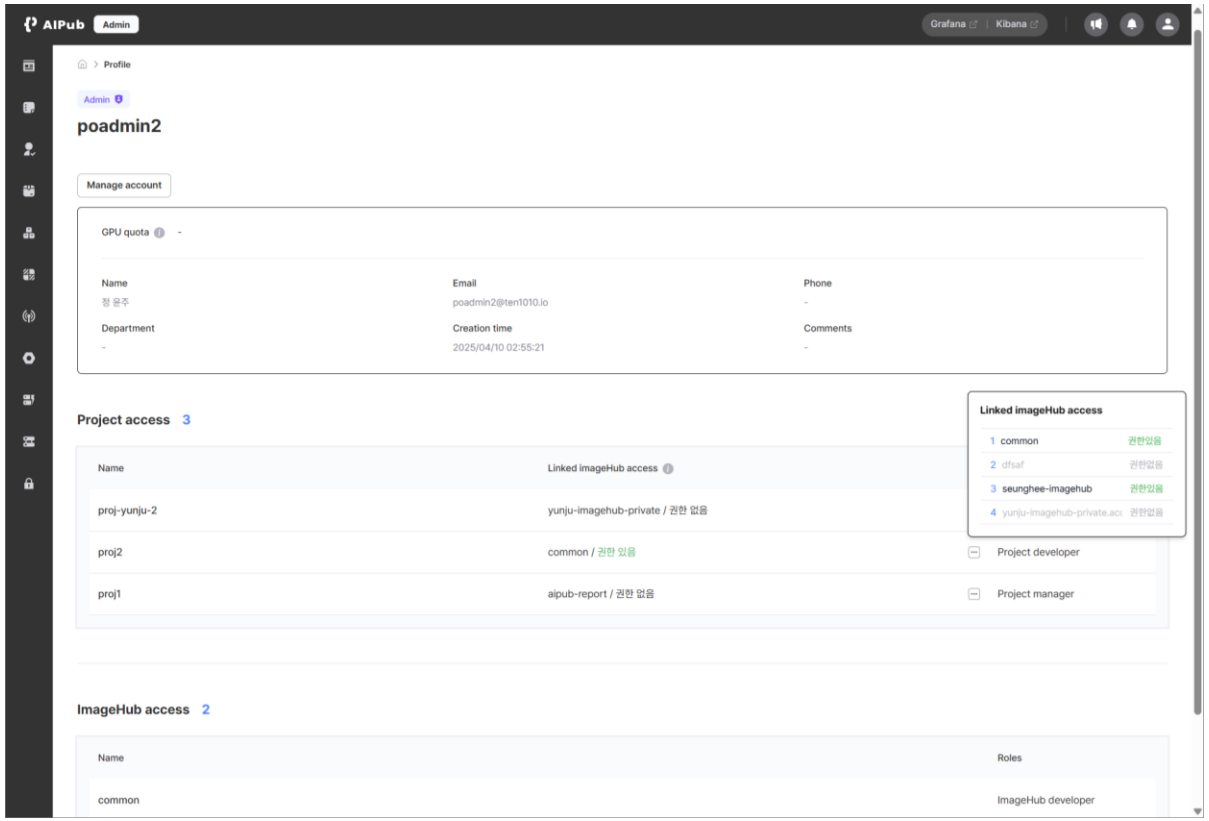
AI Pub은 역할 기반 사용자 권한 관리 기능을 제공함
쿠버네티스 RBAC 을 통해 사용자 정의 역할 및 권한 생성 가능하며, 주요 역할 프리셋 제공



User 관리

AI Pub은 사용자별 세밀한 리소스 및 권한 제어, GPU 및 워크스페이스 자원 관리, 직관적인 UI 기반의 통합 액션 기능을 통해 AI 클러스터 운영자가 안정적이고 효율적으로 사용자 관리를 수행할 수 있도록 지원합니다.

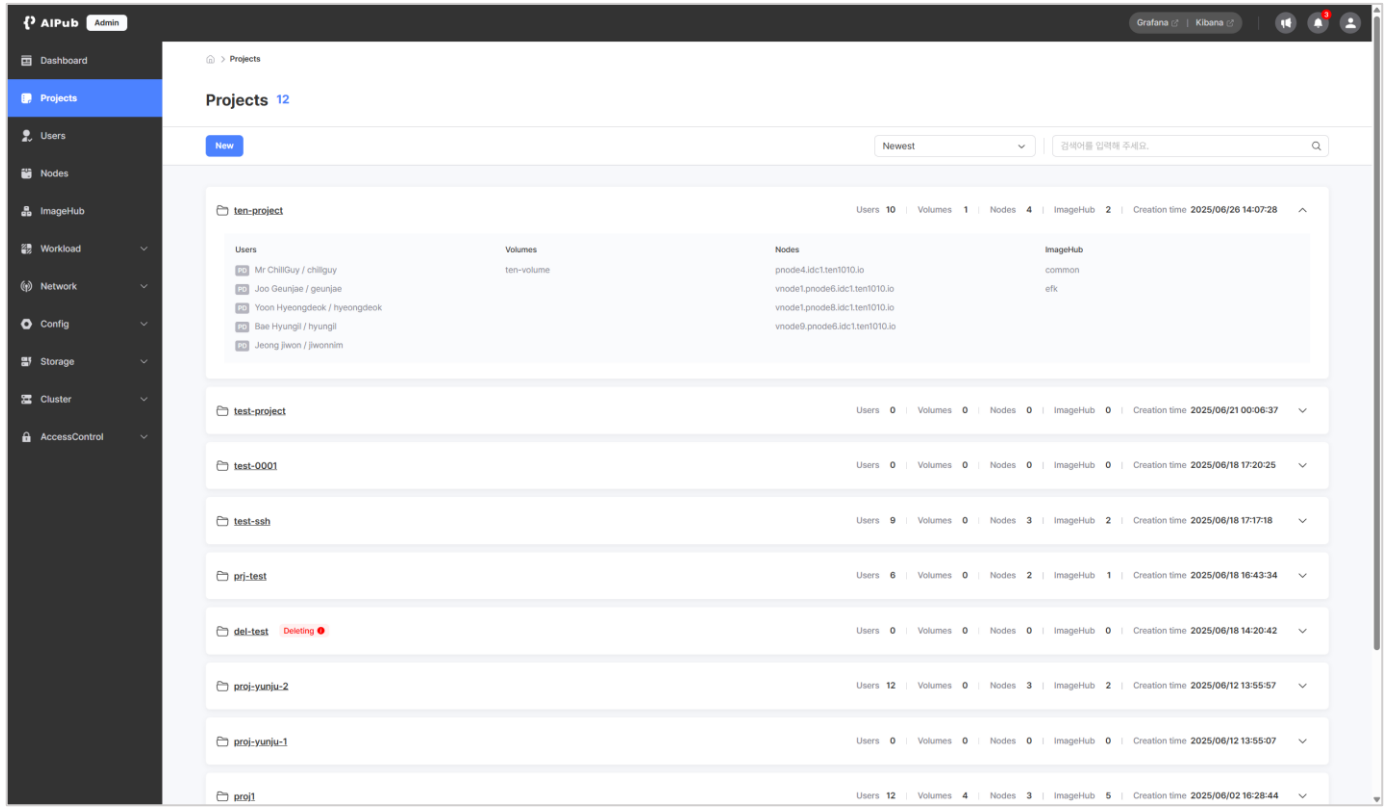




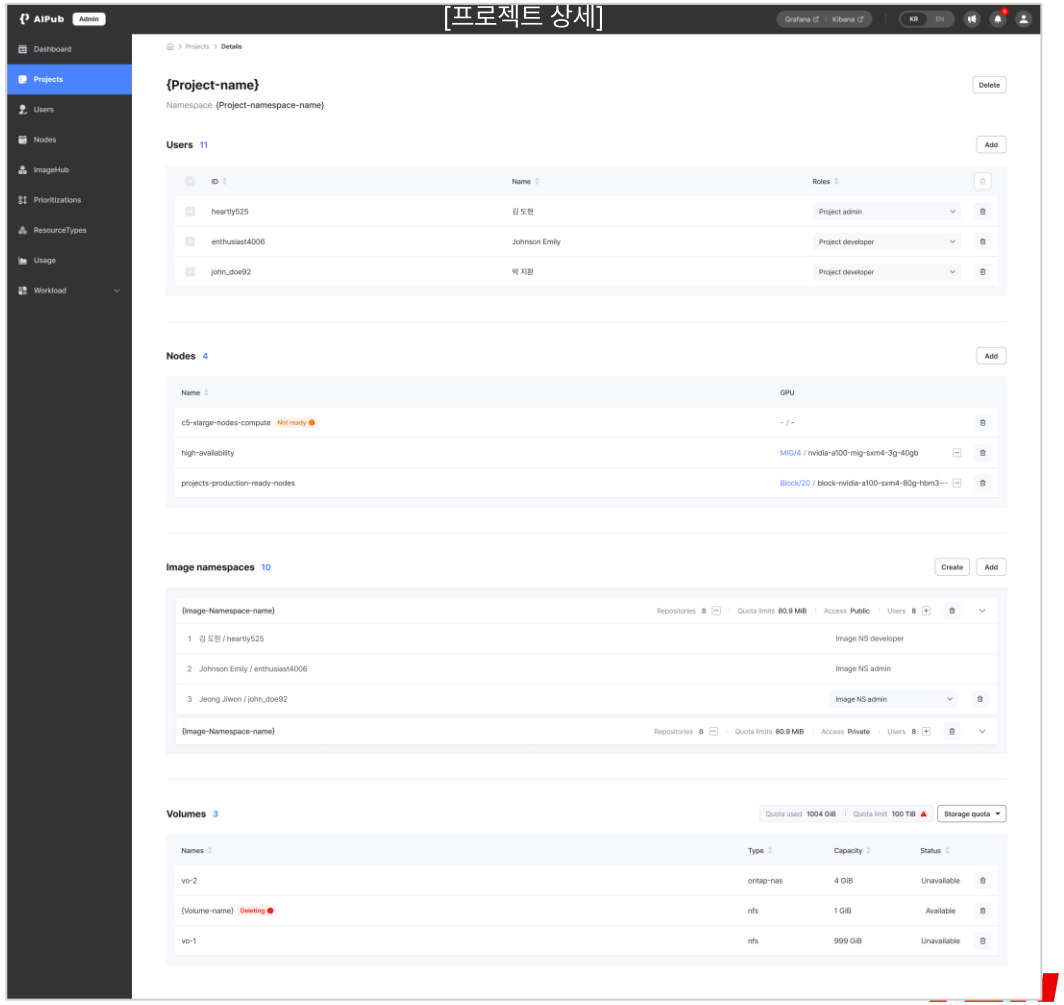
Project 관리

AI Pub의 프로젝트 관리 기능은 사용자, 볼륨, 노드, 이미지 저장소를 프로젝트 단위로 통합 관리하고 실시간 상태를 시각적으로 제공함으로써, 쿠버네티스 클러스터 운영자가 자원을 최적화하고 프로젝트별 협업을 안전하게 제어할 수 있도록 지원합니다.

[프로젝트 목록]



[프로젝트 상세]



Control and Visibility

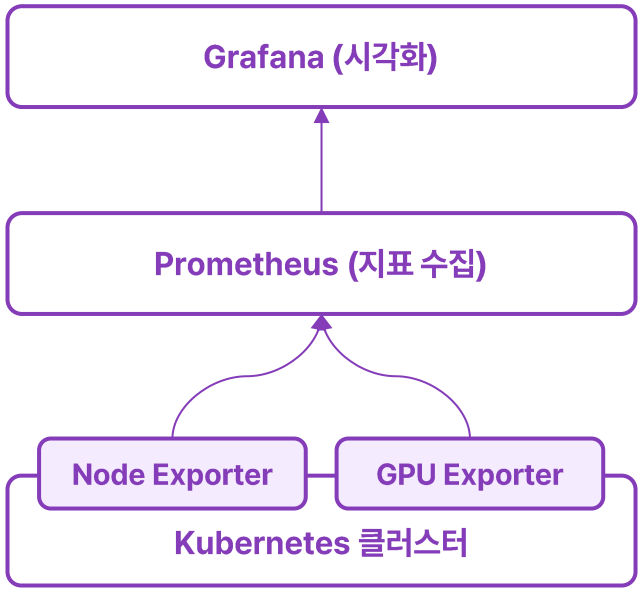
AI Pub의 서비스 별 UI와
모니터링 및 로그 기능을 소개합니다.

1. Monitoring – 메트릭 대시보드
2. Monitoring – 로그/이벤트 알림
3. 보고서 기능

1. AI Pub Monitoring: 메트릭 대시보드 개요

AI Pub은 리소스 사용량, 워크로드 상태, 성능 지표를 시각화한 대시보드를 통해 인프라와 서비스 상태를 한눈에 파악할 수 있음

[AI Pub의 메트릭 대시보드 스택 구성도]



1. 스택 개요

- AI Pub의 메트릭 대시보드 서비스는 Kube-Prometheus-Stack + Grafana 기반으로 구성되어 있음
- 인프라 및 워크로드 지표를 통합 수집하여 시각화 함

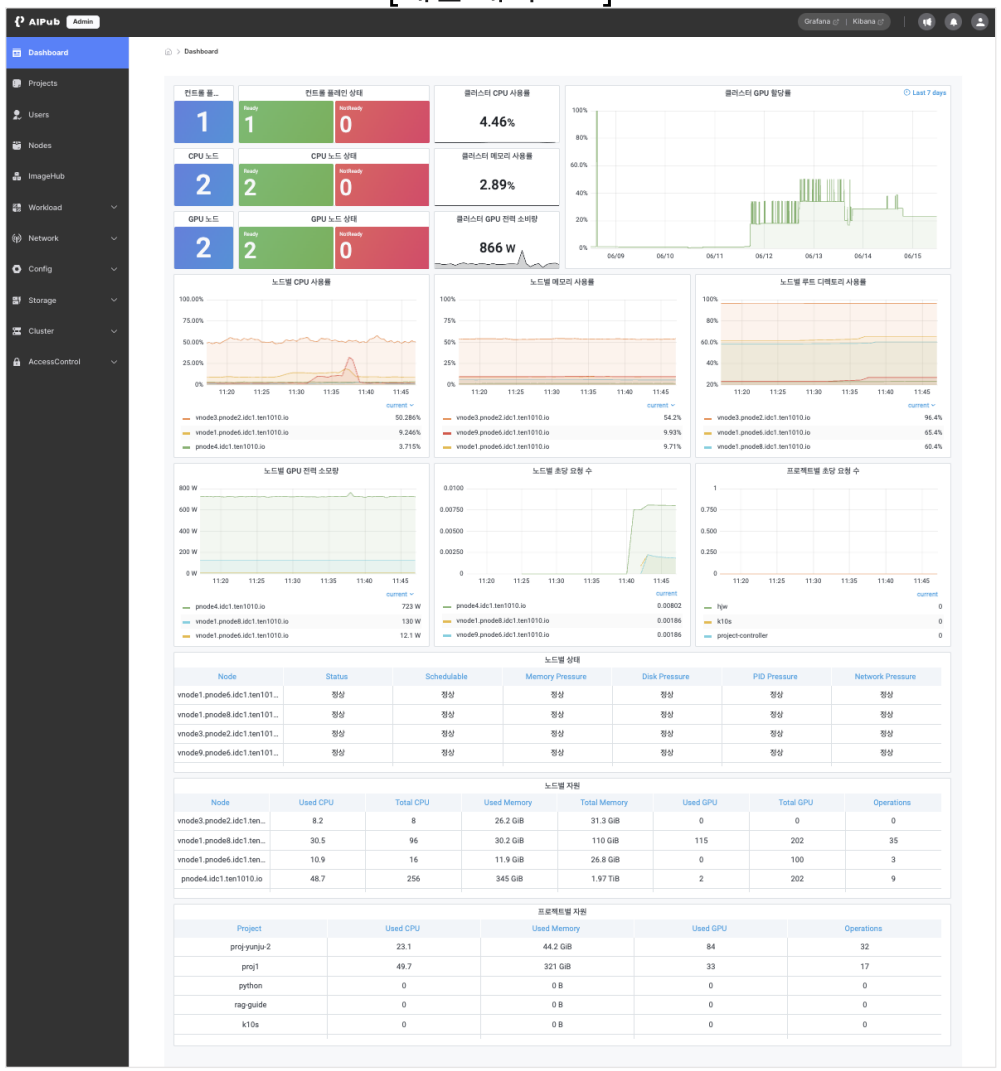
2. 커스텀 대시보드 구성 용이성

- Grafana 기반으로 고객 맞춤형 대시보드 구성이 자유로움
- Prometheus에서 수집되는 메트릭은 Kubernetes 오브젝트 단위로 구조화되어 있어 다양한 필터 및 정렬 가능
- 고객 요구에 맞는 지표 그룹핑 및 조건 기반 시각화(경고 임계값, 클러스터/네임스페이스 단위 등)가 가능
- 기존 템플릿 복제 및 변수 활용을 통해 비개발자도 손쉽게 대시보드를 편집가능

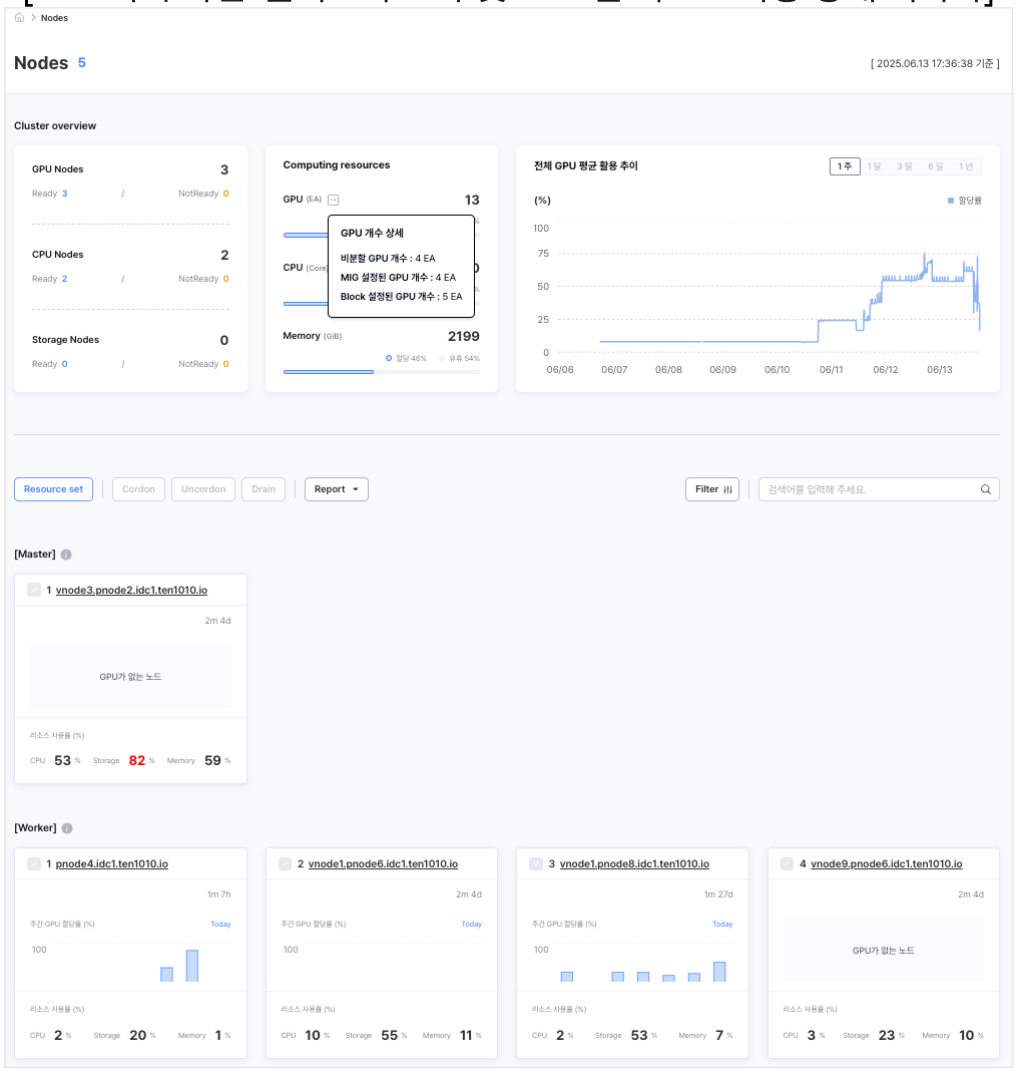
1. AI Pub Monitoring: 메트릭 대시보드

클러스터와 노드 모니터링: 클러스터 컴포넌트 별 통합/개별 지표 및 GPU의 통합/분할 설정 별 상세 지표 제공

[메인 대시보드]



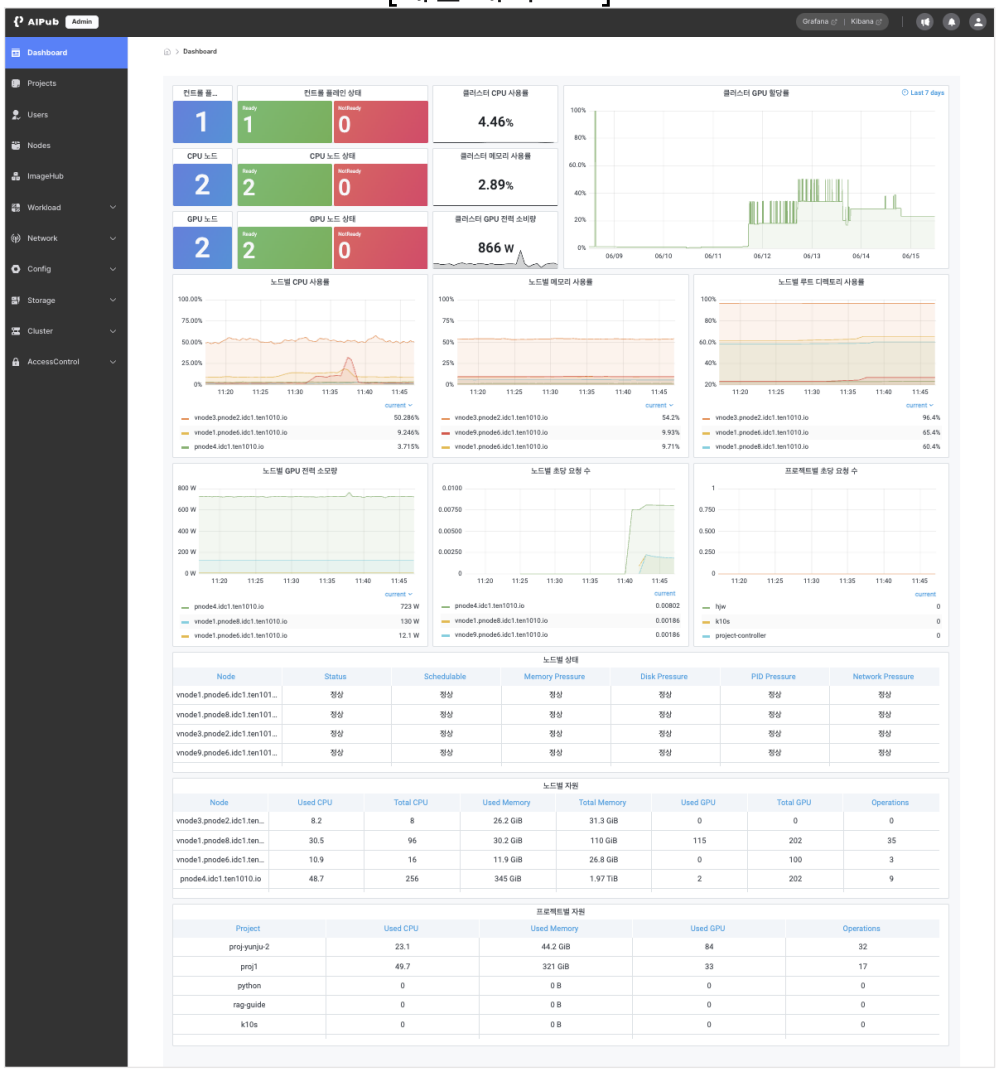
[노드 목록 화면-클러스터 요약 및 노드 별 리소스 사용 상태 시각화]



1. AI Pub Monitoring: 메트릭 대시보드

클러스터와 노드 모니터링: 클러스터 컴포넌트 별 통합/개별 지표 및 GPU의 통합/분할 설정 별 상세 지표 제공

[메인 대시보드]



- 요약

패널	내용
컨트롤 플레인/CPU/GPU 노드	각 컨트롤 플레인/CPU/GPU 의 노드 수
컨트롤 플레인/CPU/GPU 노드 상태	컨트롤 플레인/CPU/GPU 노드의 상태(Ready/Not Ready)
클러스터 CPU/메모리 사용률	클러스터 내 노드의 CPU/메모리 평균 사용률
클러스터 GPU 전력 소비량	현재 클러스터 내 모든 GPU의 전력 소비량 합
클러스터 GPU 할당률	일주일 동안의 클러스터 GPU 할당률

- 시간대 별 사용량 그래프

패널	내용
노드별 CPU 사용률	CPU 사용률 변화 그래프(%)
노드별 메모리 사용률	메모리 사용률 변화 그래프(%)
노드별 루트 디렉토리 사용률	노드 별 루트 디렉토리 사용량 변화 그래프(%)
노드별 GPU 전력 소모량	노드 별 GPU 전력 소모량 변화 그래프 (%)
노드별 초당 요청 수	Operation 자원에 대한 노드 별 초당 요청 수 변화 그래프(%)
프로젝트 별 초당 요청 수	Operation 자원에 대한 프로젝트 별 초당 요청 수 변화 그래프(%)

- 노드/프로젝트 현황 표

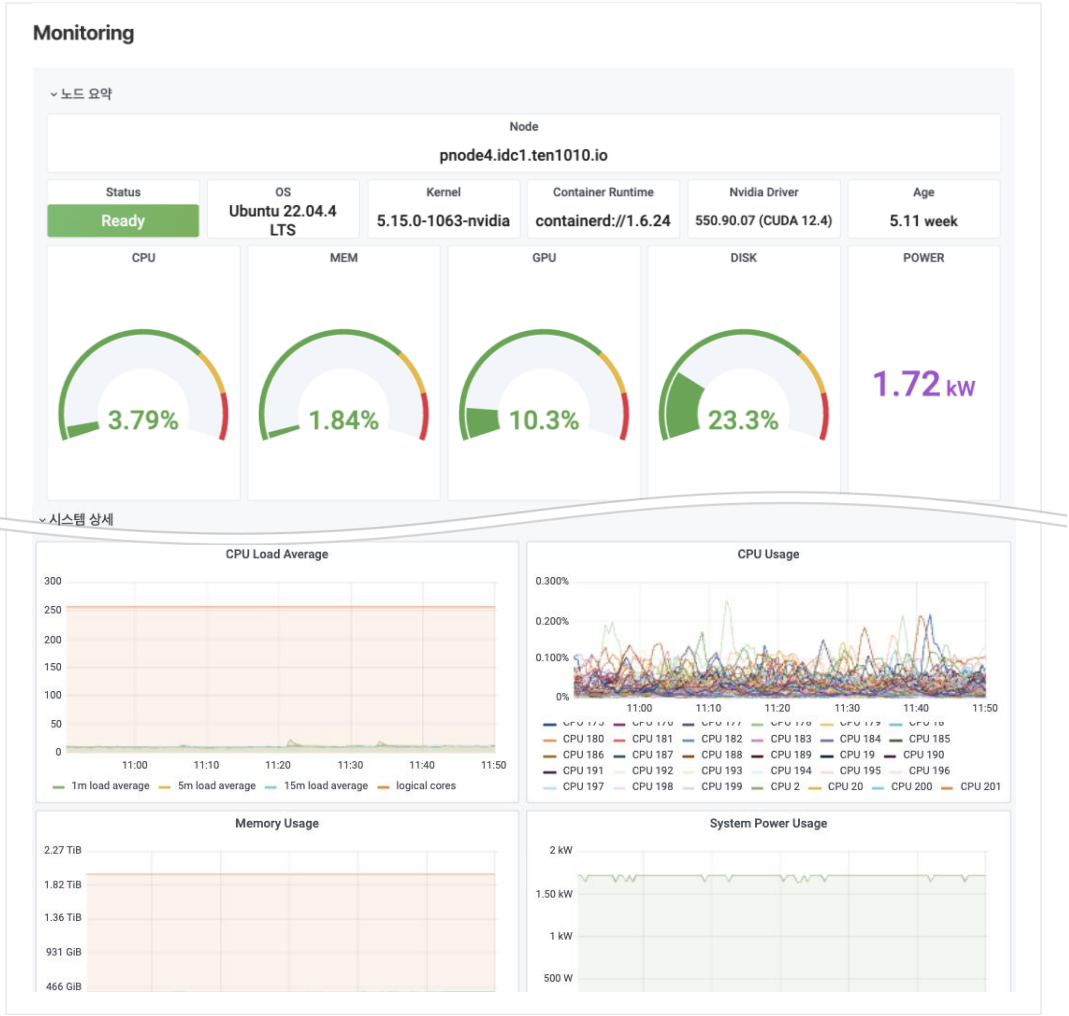
패널	내용
노드 별 상태	노드의 상태와 여러 Pressure(자원에 따른 정상작동 여부)를 정리한 표
노드 별 자원	노드 각 CPU/Memory/GPU 자원 전체와 현재 활용중인 자원 현황
프로젝트 별 자원	프로젝트 별 CPU/Memory/GPU 활용 자원 현황

EN

1. AI Pub Monitoring: 메트릭 대시보드

노드 별 모니터링: GPU 관련 다양한 메트릭은 노드별 모니터링 페이지에서 별도 대시보드로 구성 및 제공

[노드 별 모니터링 1/4 : 요약]



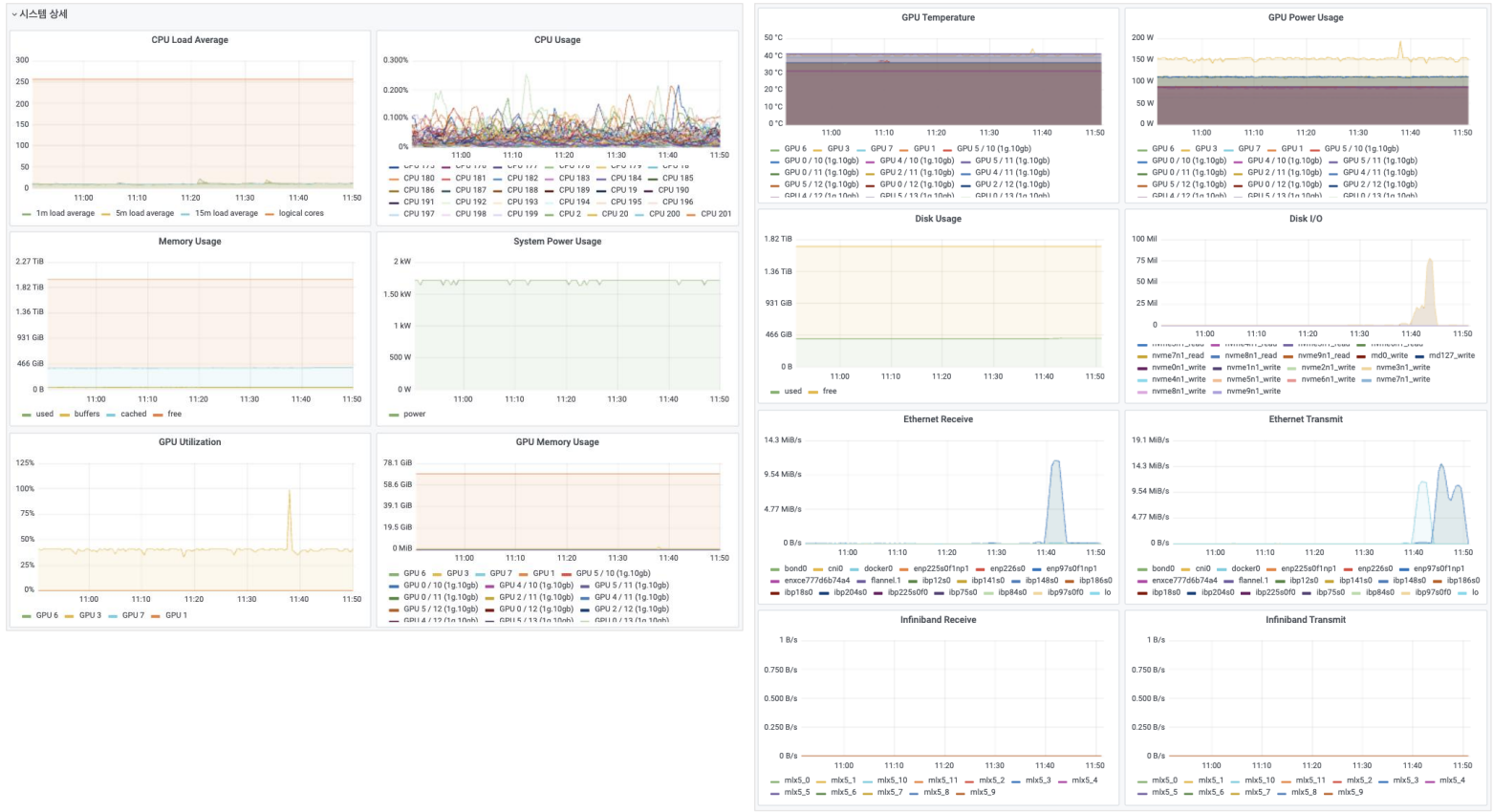
[노드 별 모니터링 1/4 : 요약 – 메트릭 상세]

패널	설명
Node	노드 이름
Status	노드 준비 상태 (Ready/Not Ready)
OS	운영체제 이름 및 버전
Kernel	커널 버전
Container Runtime	컨테이너 런타임 종류 및 버전
Age	생성 후 경과 시간
CPU	논리 코어별 사용률
MEM	전체 메모리 사용률
GPU	전체 GPU 사용률
DISK	디스크 사용률
POWER	전력 소비량

1. AI Pub Monitoring: 메트릭 대시보드

노드 별 모니터링: GPU 관련 다양한 메트릭은 노드별 모니터링 페이지에서 별도 대시보드로 구성 및 제공

[노드 별 모니터링 2/4 : 시스템 상세]



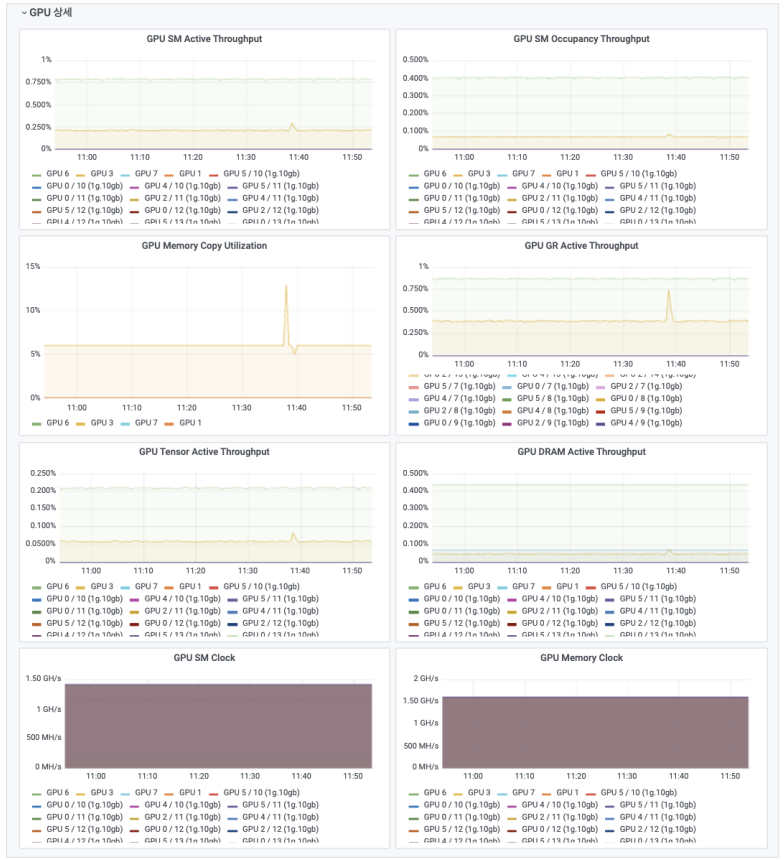
[노드 별 모니터링 2/4 : 시스템 상세 – 메트릭 상세]

패널	설명
CPU Load Average	1/5/15분 동안 평균사용한 논리 코어 수
CPU Usage	논리 CPU별 사용률
Memory Usage	메모리 사용량
System Power Usage	시스템 전체 전력 사용량
GPU Utilization	GPU 사용률
GPU Memory Usage	GPU 메모리 사용량
GPU Temperature	GPU 온도
GPU Power Usage	GPU 전력 소비량
Disk Usage	디스크 사용량
Disk I/O	디스크 읽기/쓰기 트래픽
Ethernet Receive	네트워크 수신 트래픽
Ethernet Transmit	네트워크 송신 트래픽
Infiniband Receive	인피니밴드 수신 트래픽
Infiniband Transmit	인피니밴드

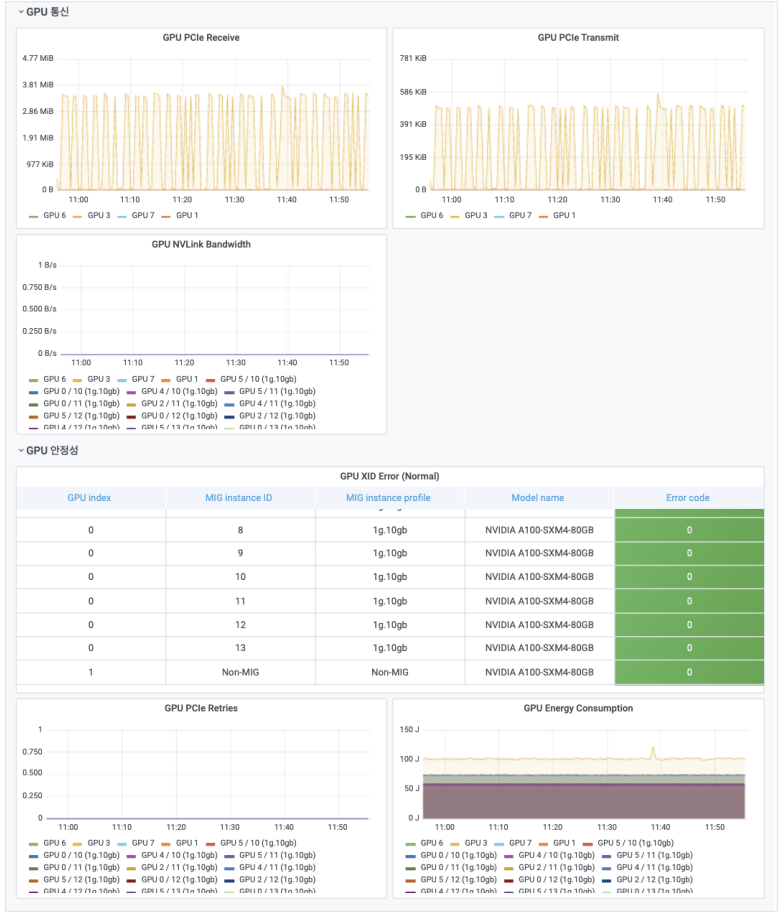
1. AI Pub Monitoring: 메트릭 대시보드

노드 별 모니터링: GPU 관련 다양한 메트릭은 노드별 모니터링 페이지에서 별도 대시보드로 구성 및 제공

[노드 별 모니터링 3/4 : GPU 상세]



[노드 별 모니터링 4/4 : GPU 통신 및 안정성]



[노드 별 모니터링 3/4 : GPU 상세 – 메트릭 상세]

패널	설명
GPU SM Active Throughput	SM 유닛의 실행률
GPU SM Occupancy Throughput	SM 유닛 점유율
GPU Memory Copy Utilization	메모리 복사 활용률
GPU GR Active Throughput	그래픽 유닛 실행률
GPU Tensor Active Throughput	텐서 코어 실행률
GPU DRAM Active Throughput	DRAM 활용률
GPU SM Clock	SM 유닛 클럭 속도
GPU Memory Clock	메모리 클럭 속도

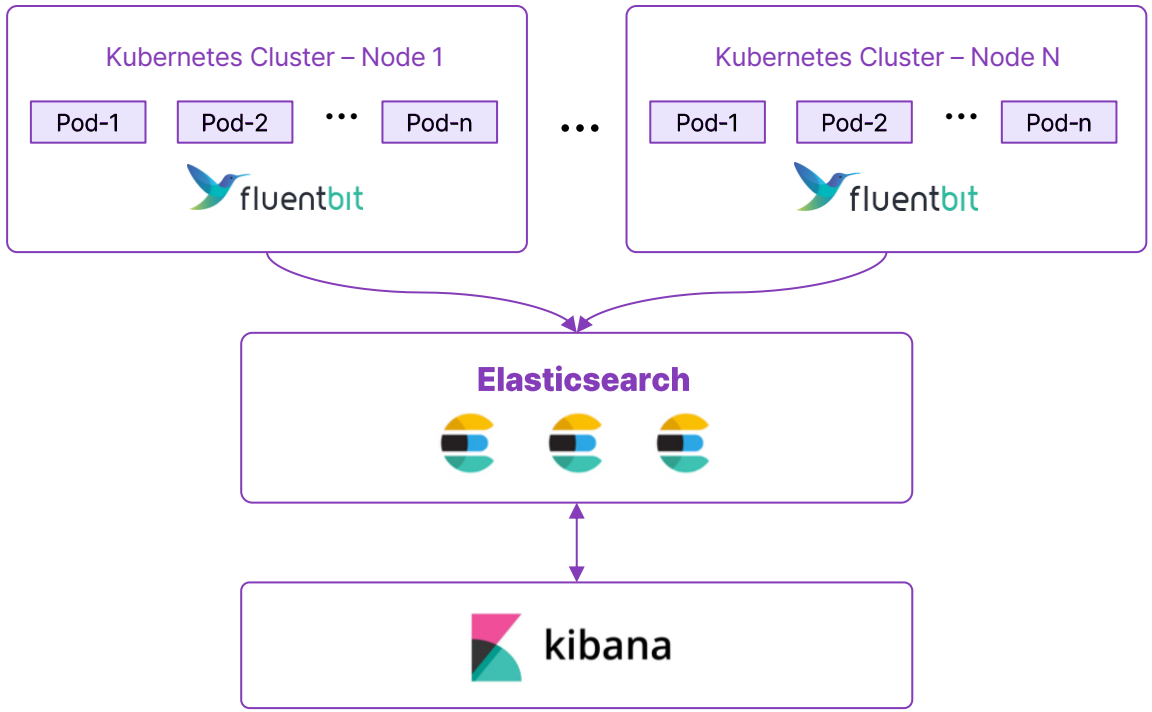
[노드 별 모니터링 4/4 : GPU 상세 – 메트릭 상세]

패널	설명
GPU PCIe Receive	GPU별 PCIe 수신 트래픽
GPU PCIe Transmit	GPU별 PCIe 송신 트래픽
GPU NVLink Bandwidth	GPU 간 NVLink 대역폭
패널	설명
GPU XID Error	GPU 오류 코드
GPU PCIe Retries	PCIe 재전송 발생률
GPU Energy Consumption	총 에너지 사용량 (joules)

2. AI Pub Monitoring: 로그/이벤트 알림 개요

에러, 이벤트, 사용자 행동 로그를 수집하여 실시간 알림과 함께 상세한 문제 원인 분석 기능을 제공

[AI Pub의 로그/이벤트 알림 스택 구성도]

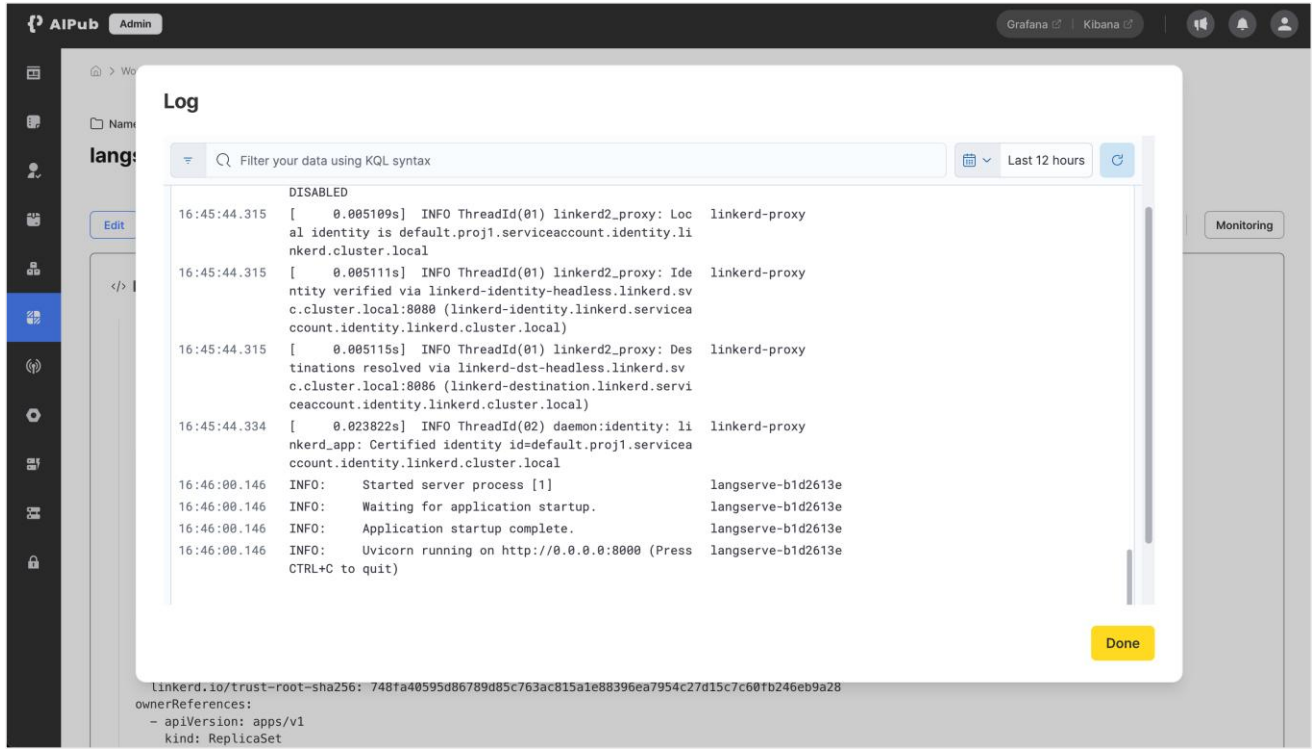


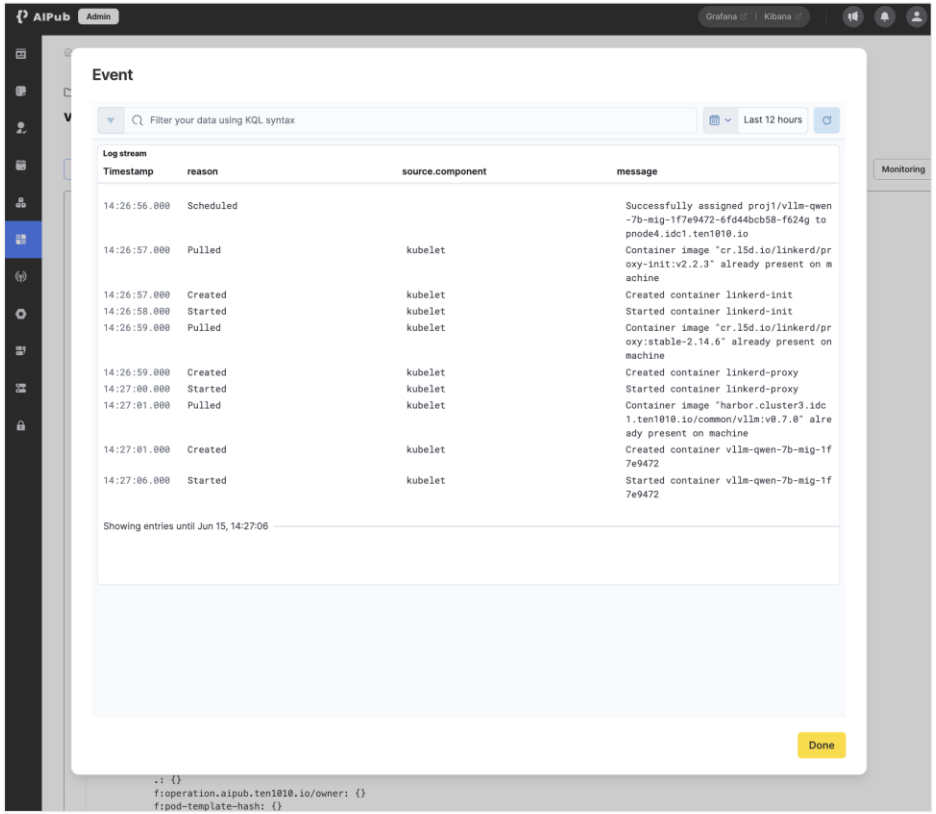
1. 스택 개요

- AI Pub의 로그/이벤트 알림 기능은 EFK(Elasticsearch, Fluent Bit, Kibana) 스택을 기반으로 구성되어 있음
- Kubernetes 기반 워크로드에서 발생하는 로그 및 이벤트 데이터를 수집/저장/시각화하는 구조로 구성
- 저장된 데이터는 검색 성능과 확장성을 보장하면서 Kibana를 통해 손쉽게 시각화 가능
- 수집된 데이터는 Elasticsearch을 통해 마스터 노드에 분산 저장하여 고가용성과 확장성을 동시에 확보

2. AI Pub Monitoring: 로그/이벤트 알림 개요

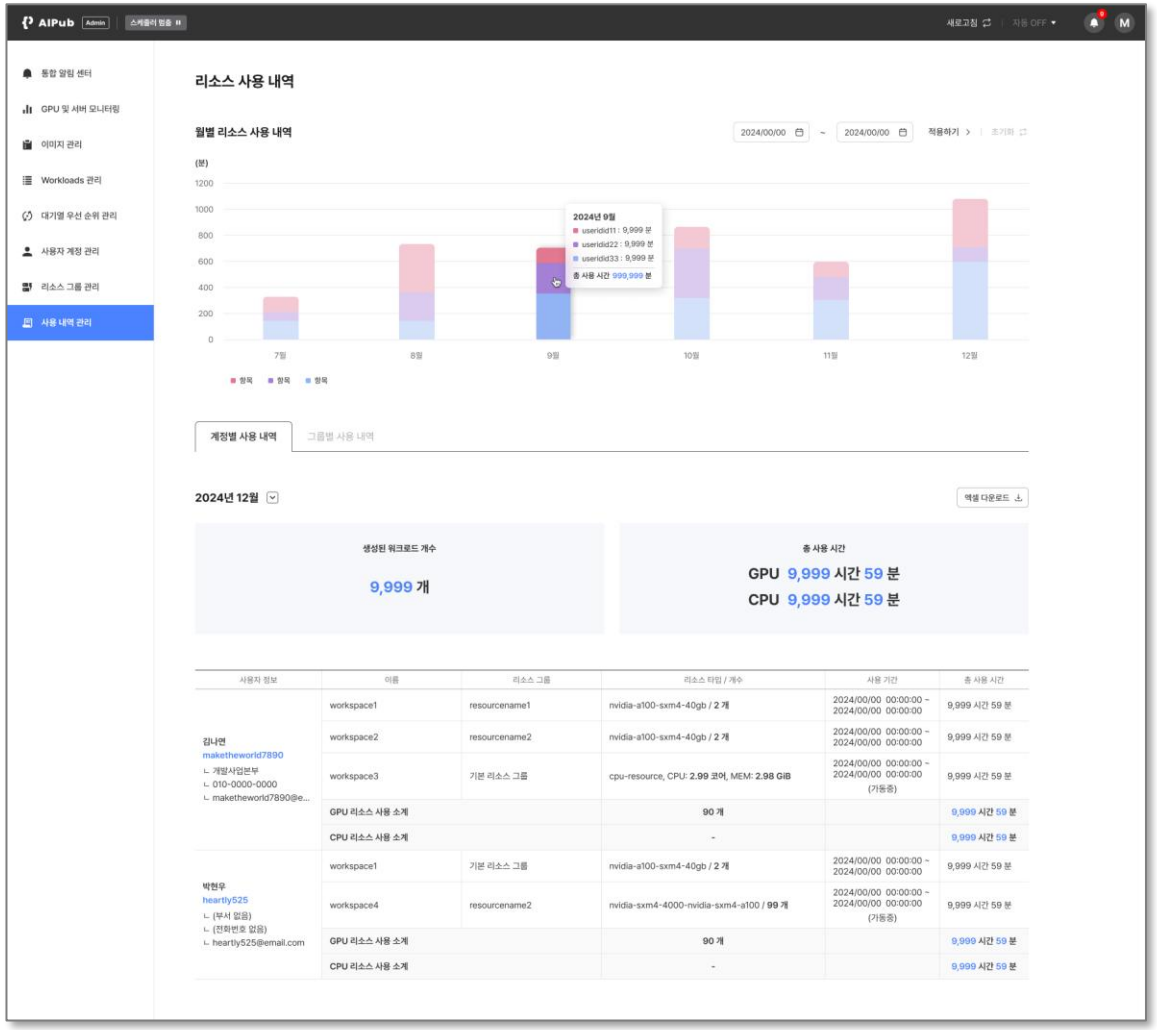
에러, 이벤트, 사용자 행동 로그를 수집하여 실시간 알림과 함께 상세한 문제 원인 분석 기능을 제공





3. AI Pub Monitoring: 보고서 기능

클러스터 현황에 대한 Daily/Weekly/Monthly PDF 리포트 제공 및 리소스 사용 내역 관리 기능

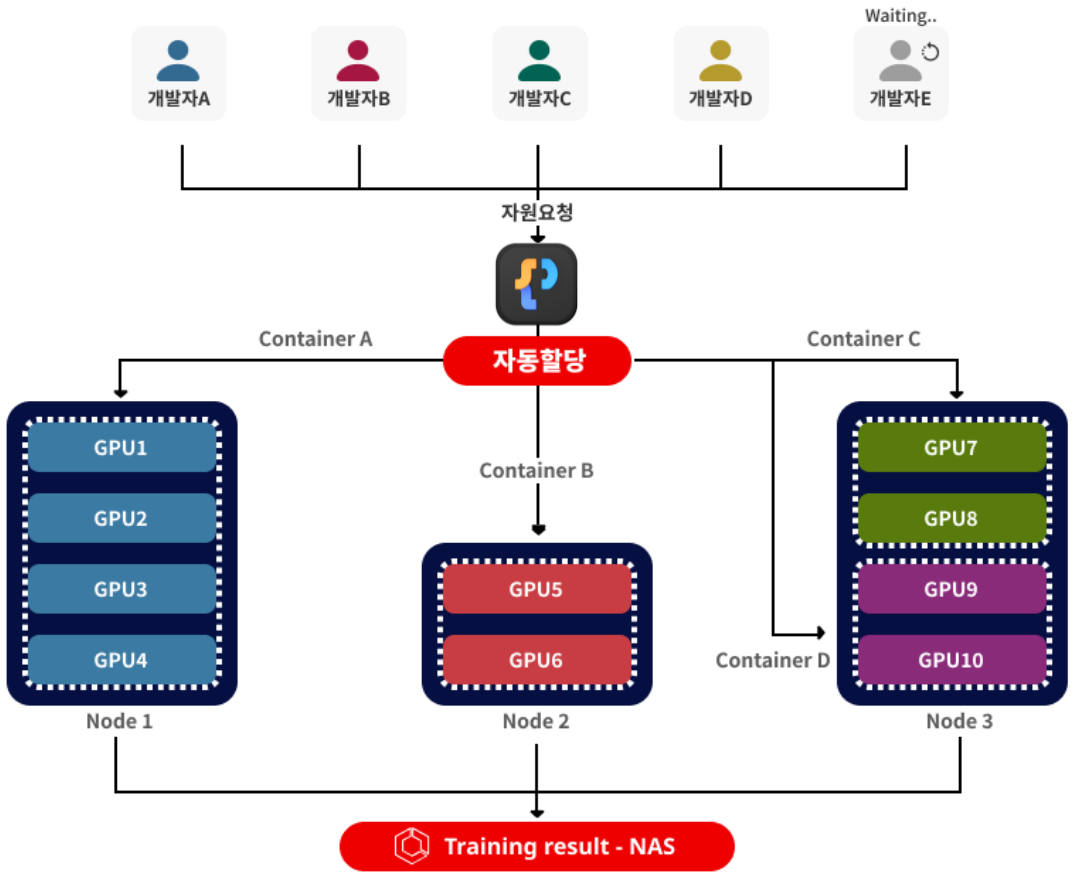


AIPub Dev

COASTER를 코어로 하여 AI 개발을 지원하는 완전 관리형 서비스

팀 혹은 사용자 단위로 AI 인프라를 할당할 수 있습니다.
팀 개발자들의 자원 사용량 및 팀 별 사용량을 집계할 수 있습니다.

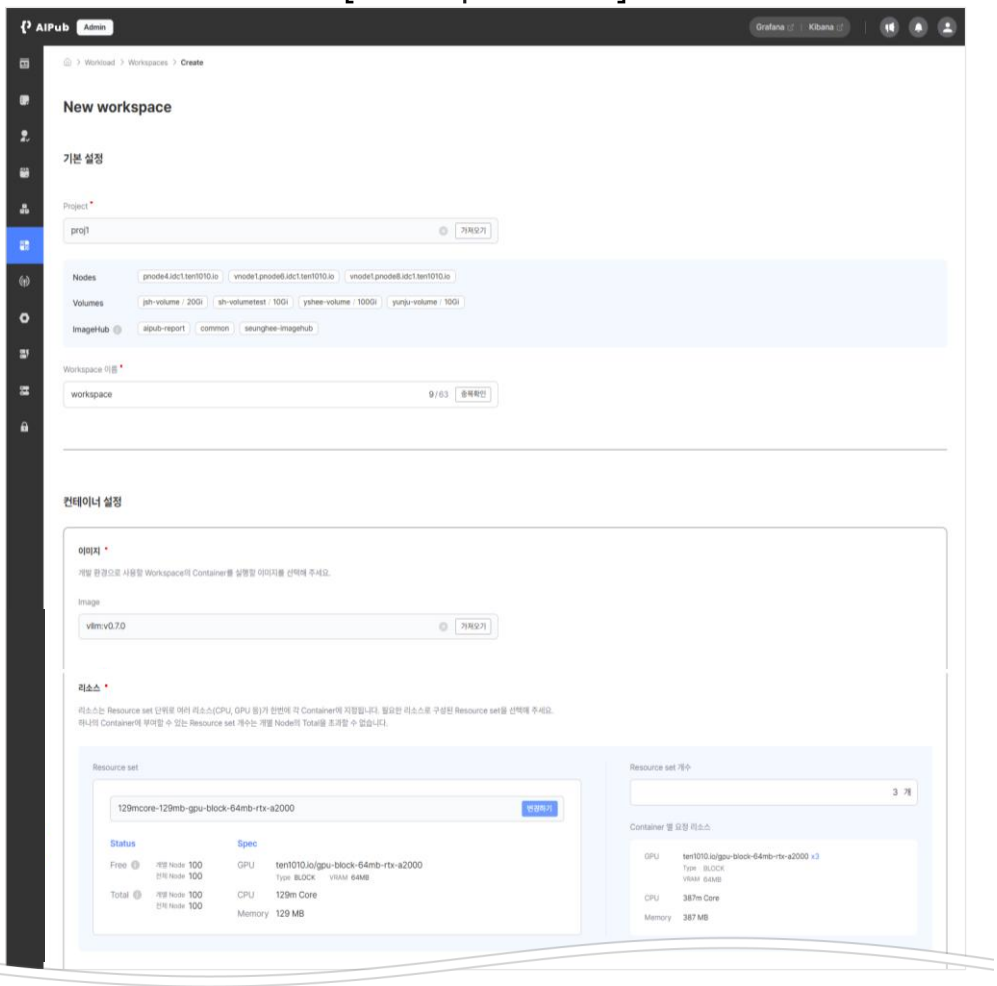
AI 개발용 인프라를 구축하여 중앙 관리할 때에도 자원 할당 및 관리가 가능하고, 각 단위 별 사용량을 측정할 수 있습니다.



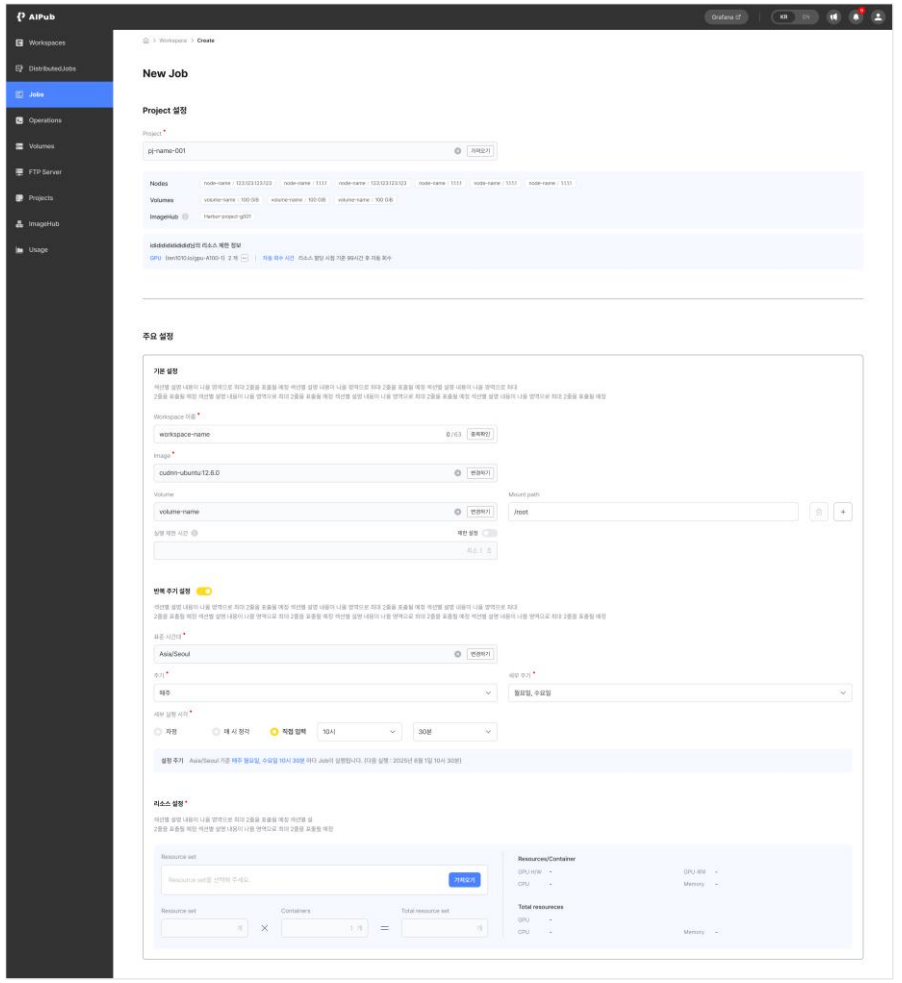
AI Pub Dev

AI Pub은 Kubernetes 기반이지만 직관적인 UI로 워크로드, 리소스, 볼륨을 통합 관리할 수 있어 비숙련자도 쉽게 접근 가능
실험 환경 구성, 실행, 결과 저장을 한 플랫폼 내에서 처리할 수 있어 전반적인 AI 연구 워크플로우가 단축됨

[Workspace 생성]



[Job 생성]

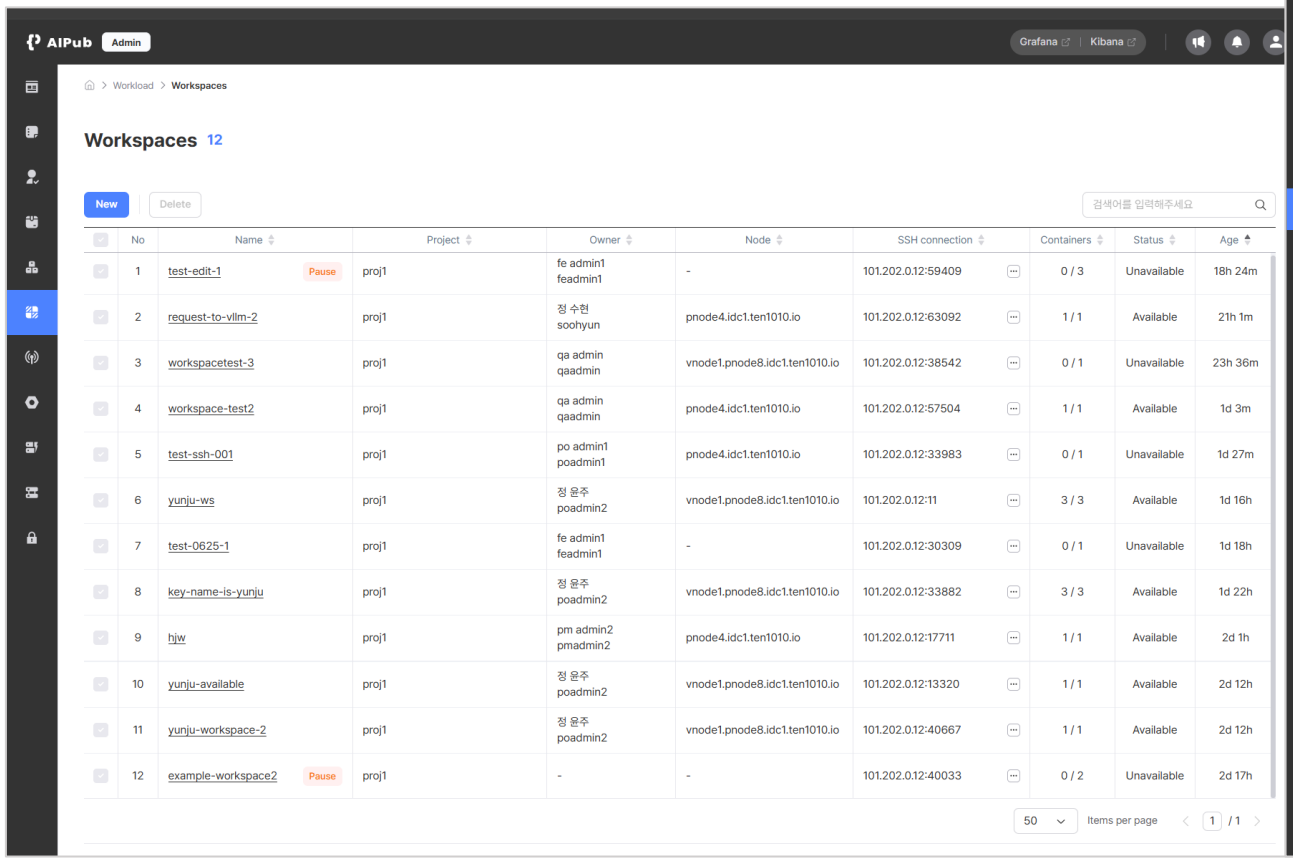


TEN

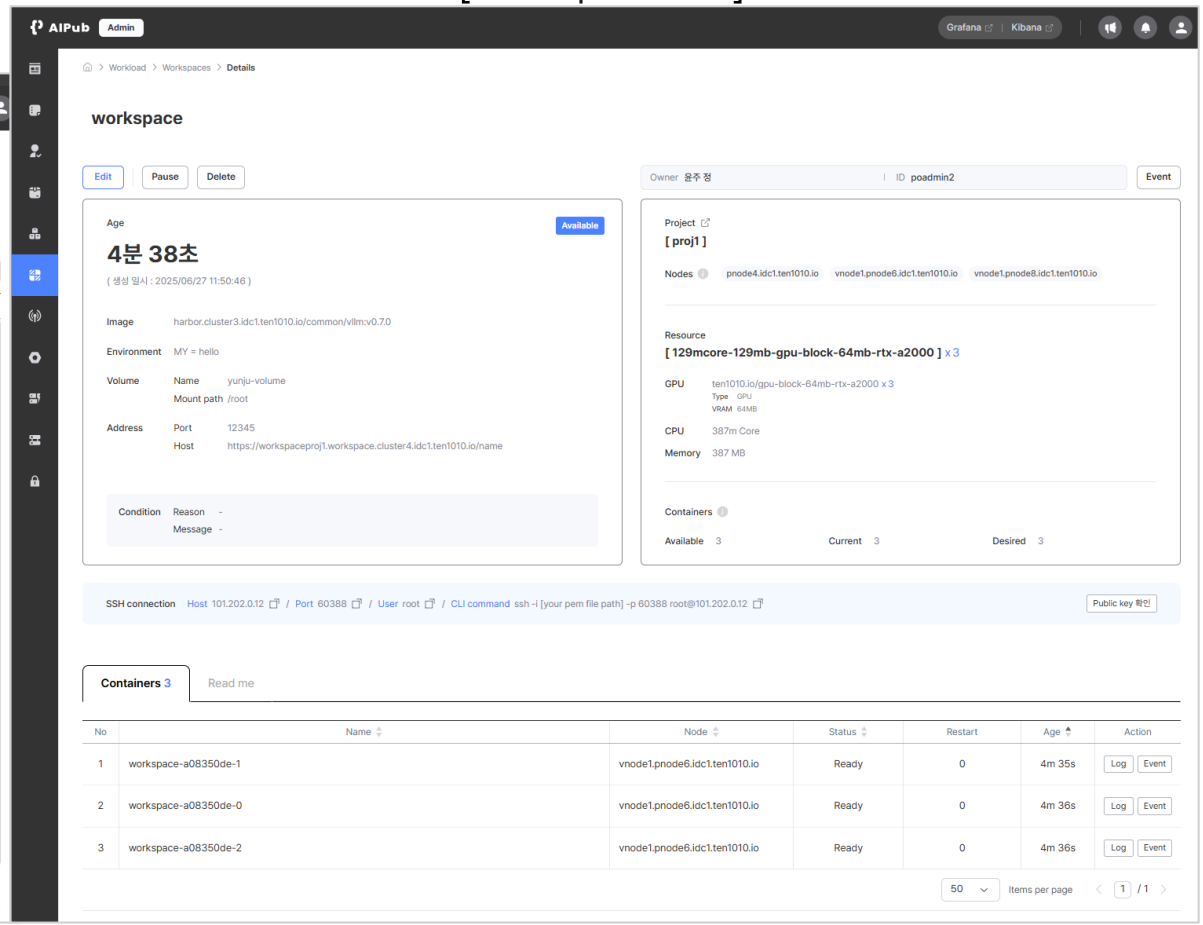
AI Pub Dev

AI Pub은 Kubernetes 기반이지만 직관적인 UI로 워크로드, 리소스, 볼륨을 통합 관리할 수 있어 비숙련자도 쉽게 접근 가능
실험 환경 구성, 실행, 결과 저장을 한 플랫폼 내에서 처리할 수 있어 전반적인 AI 연구 워크플로우가 단축됨

[Workspace 목록]



[Workspace 상세]



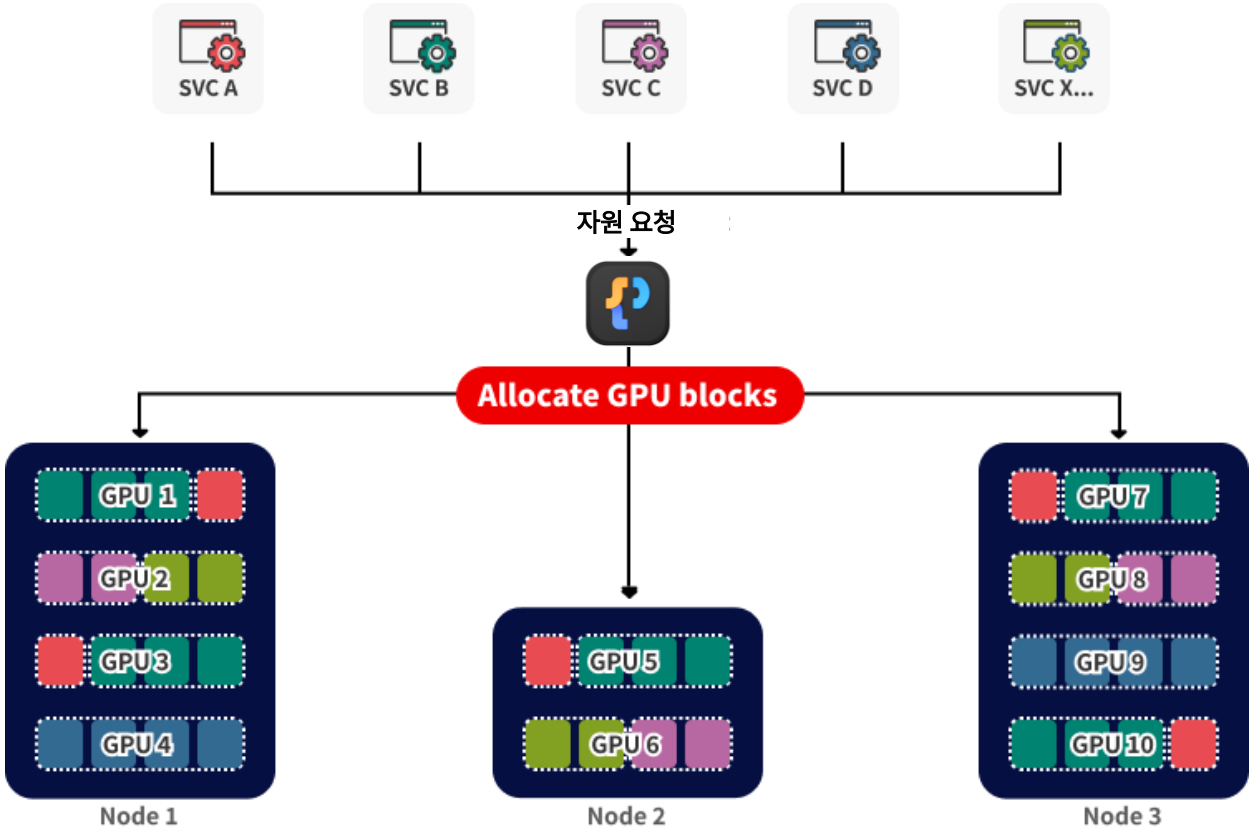
AIPub Ops

COASTER를 코어로 하여 AI 운영을 지원하는 완전 관리형 서비스

AIPub Ops 주요 기능 1.

사용자의 서비스를 이미지의 형태로 관리합니다.
GPU 1개를 100분할 하여 최소 단위로 운영해 비용 절감에 기여합니다.

비개발자도 서비스를 생성하거나 중지, 삭제할 수 있으며, 업데이트 및 롤백도 가능합니다.



AIPub Ops 주요 기능 2.

퍼블릭 클라우드의 서비스 운영 및 관리 기능을 제공하여 온프레미스 서버에서도 AI 서비스 운영을 돕는 기능을 지원합니다.

High Availability
이중화를 통해
Single point of failure 제거

Rolling Update
서비스 중단 없이
상시 업데이트 가능

Load Balancing
바쁘지 않은 서버로
서비스 요청을 자동 분배

Scale-out
서비스 요청에 따라
서버 수를 자동으로 늘려 트래픽 처리

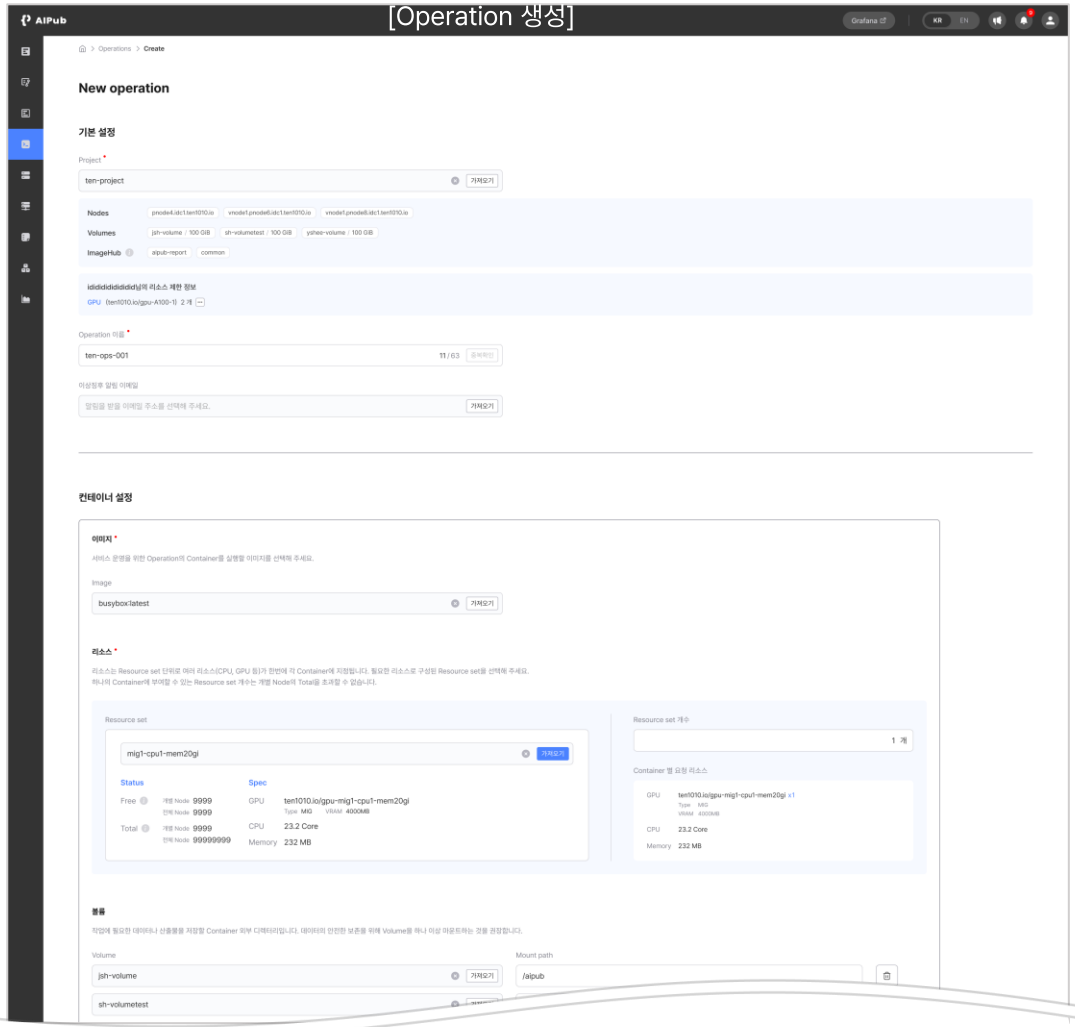
Fail over 대응
서비스 Fail over 시 탐지 기능 및
서비스를 새로 띄워 안정성 확보

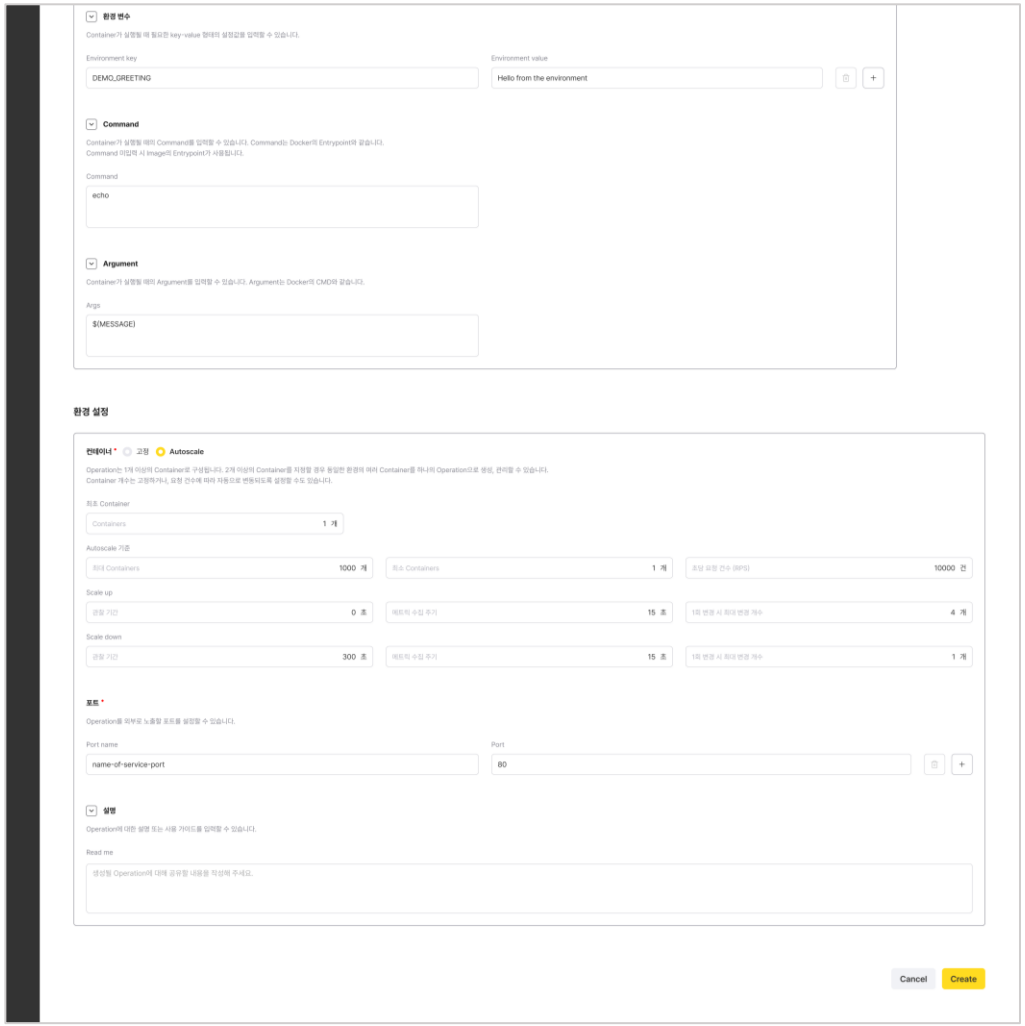
이상징후 알림
중요 운영 이벤트 발생 시
카톡, 슬랙으로 실시간 알림

AI Pub Ops

AI Pub은 Auto-scaling 서빙과 예약 실행 기반 배치 작업을 UI로 쉽게 구성할 수 있어, AI 모델 운영과 지속적인 개선을 효율적이고 안정적으로 지원합니다.

[Operation 생성]

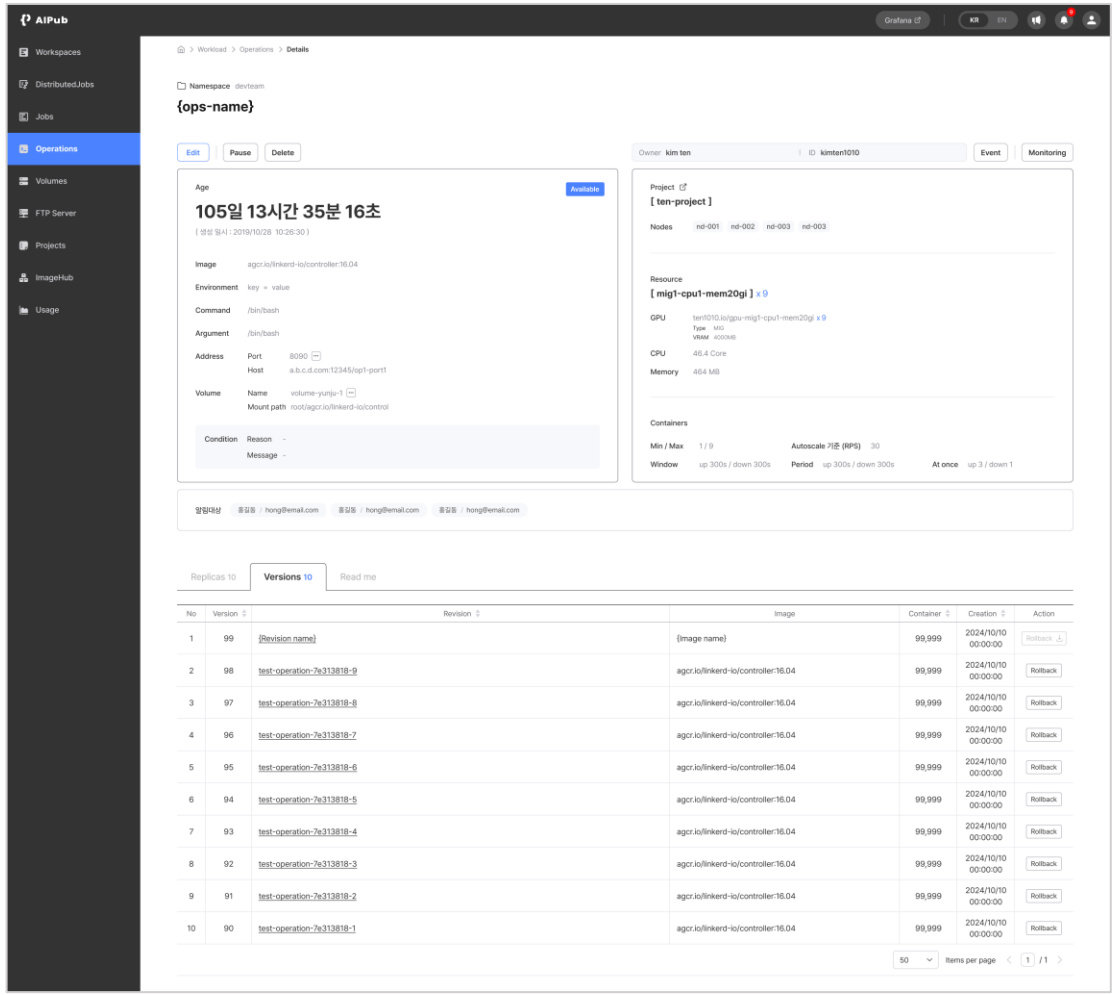




AI Pub Ops

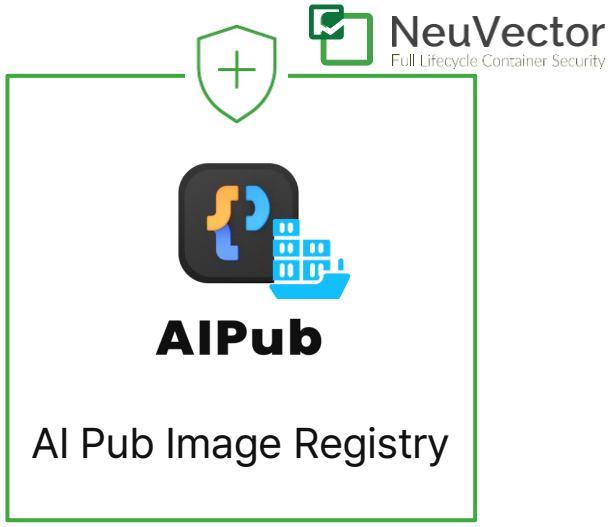
AI Pub은 Auto-scaling 서빙과 예약 실행 기반 배치 작업을 UI로 쉽게 구성할 수 있어, AI 모델 운영과 지속적인 개선을 효율적이고 안정적으로 지원합니다.

[Operation 상세]

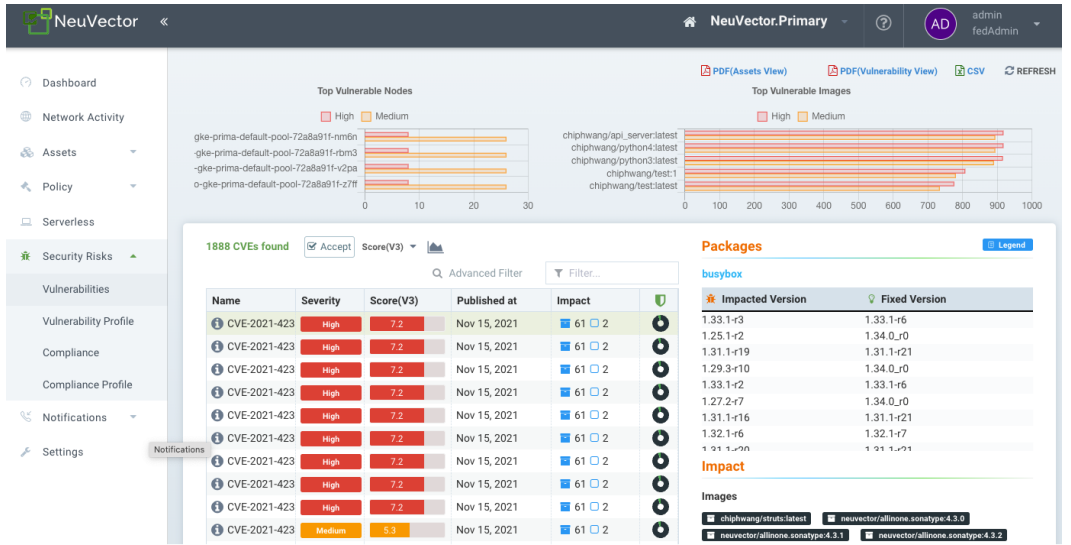


AI Pub Ops는 NeuVector의 스캐닝 기능을 사용하여 다양한 경로로 유입되는 도커 이미지에 대한 보안 필터링이 가능합니다.

컨테이너 이미지, 클러스터를 구성하는 노드, 동작 중인 컨테이너의 취약점을 검사할 수 있습니다.



COASTER X SUSE



CVE를 기준으로 스캐닝

CVE는 다양한 취약점을 통일해 고유한 넘버링을 부여한 것으로
장치, 시스템, 프로그램 해킹에 악용될 수 있는 컴퓨터 보안 취약성 및 시스템 결함에 코드를 부여하고 관리

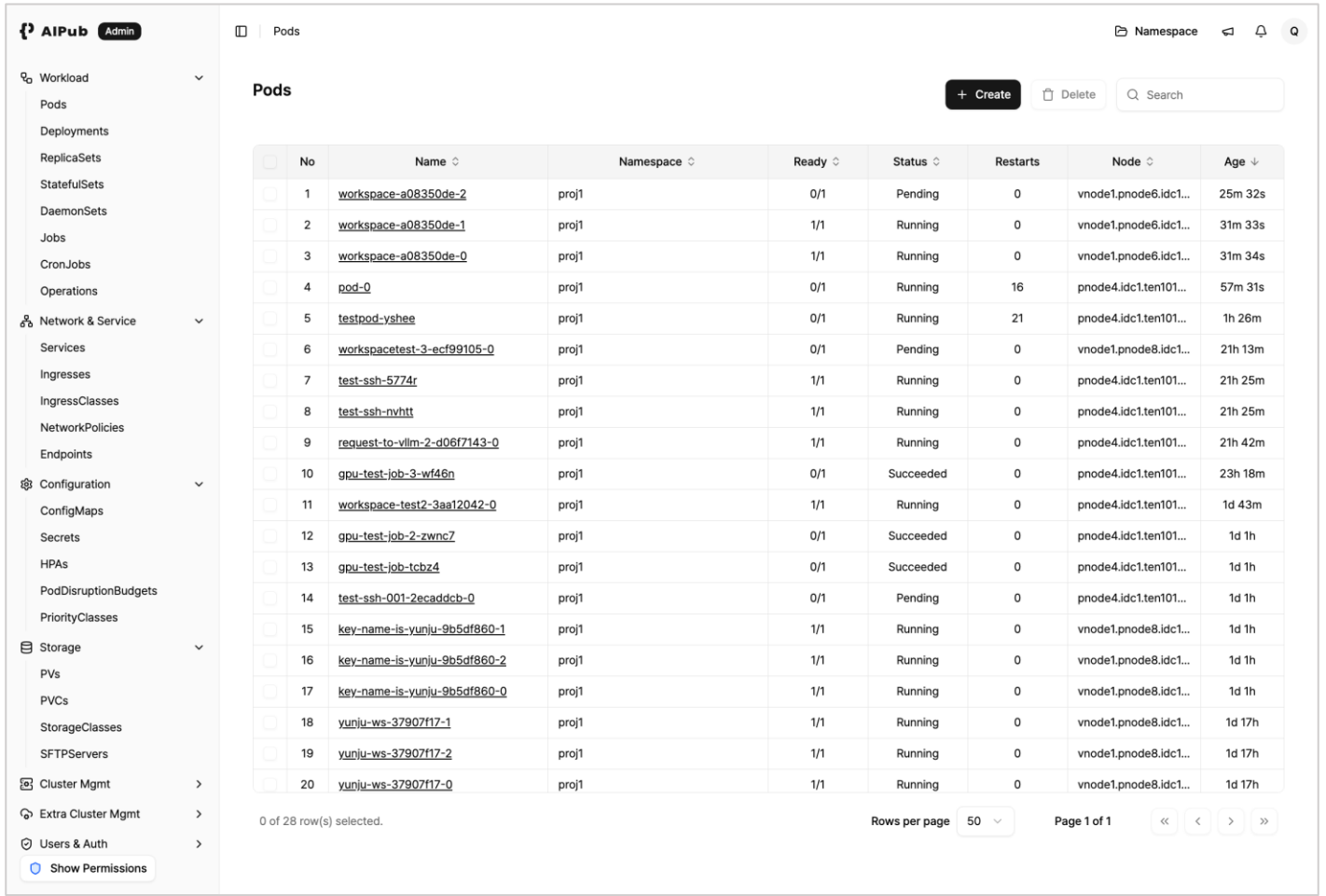
AI Pub K10s

K8s 숙련자를 위한 Kubernetes Native Platform

AI Pub K10s

AI Pub의 Kubernetes UI는 CLI만 사용하는 방식 대비 전체 리소스 상태를 한눈에 파악하고, 검색·관리·분석까지 직관적으로 수행할 수 있어 숙련자들의 운영 효율성을 극대화합니다.

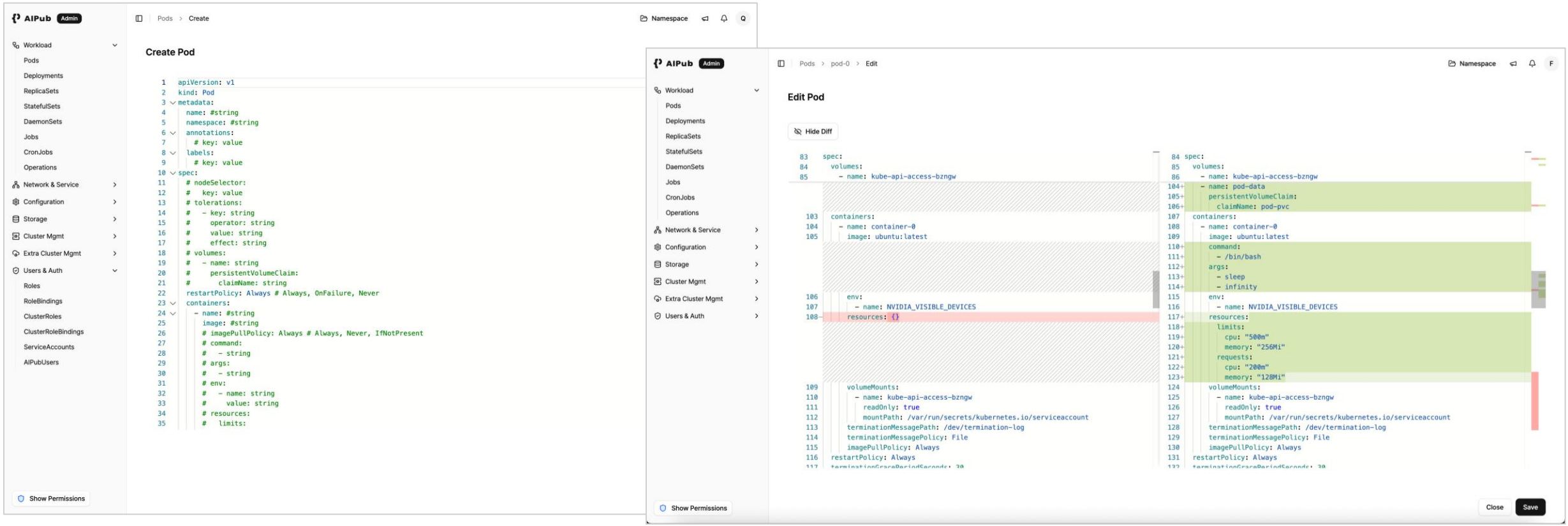
[Pod 목록]



AI Pub K10s

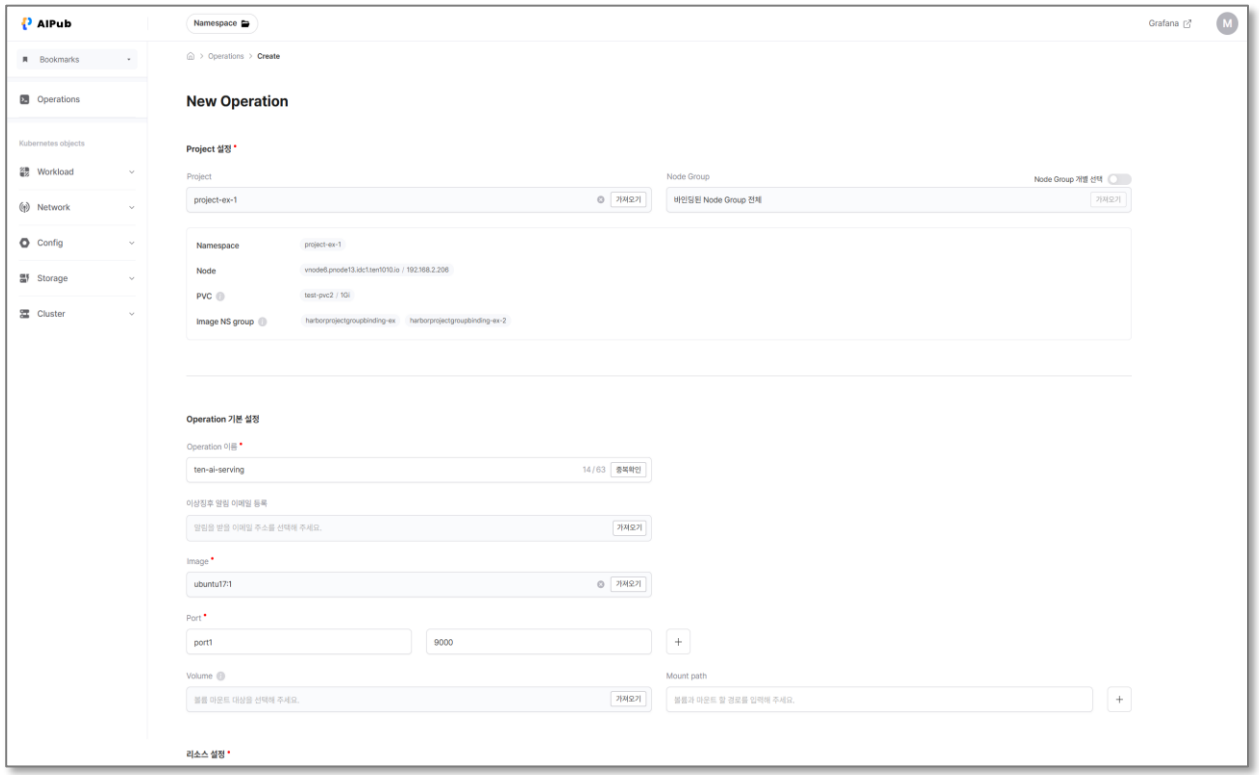
AI Pub의 Kubernetes UI는 CLI만 사용하는 방식 대비 전체 리소스 상태를 한눈에 파악하고, 검색·관리·분석까지 직관적으로 수행할 수 있어 숙련자들의 운영 효율성을 극대화합니다.

[Yaml Editor]



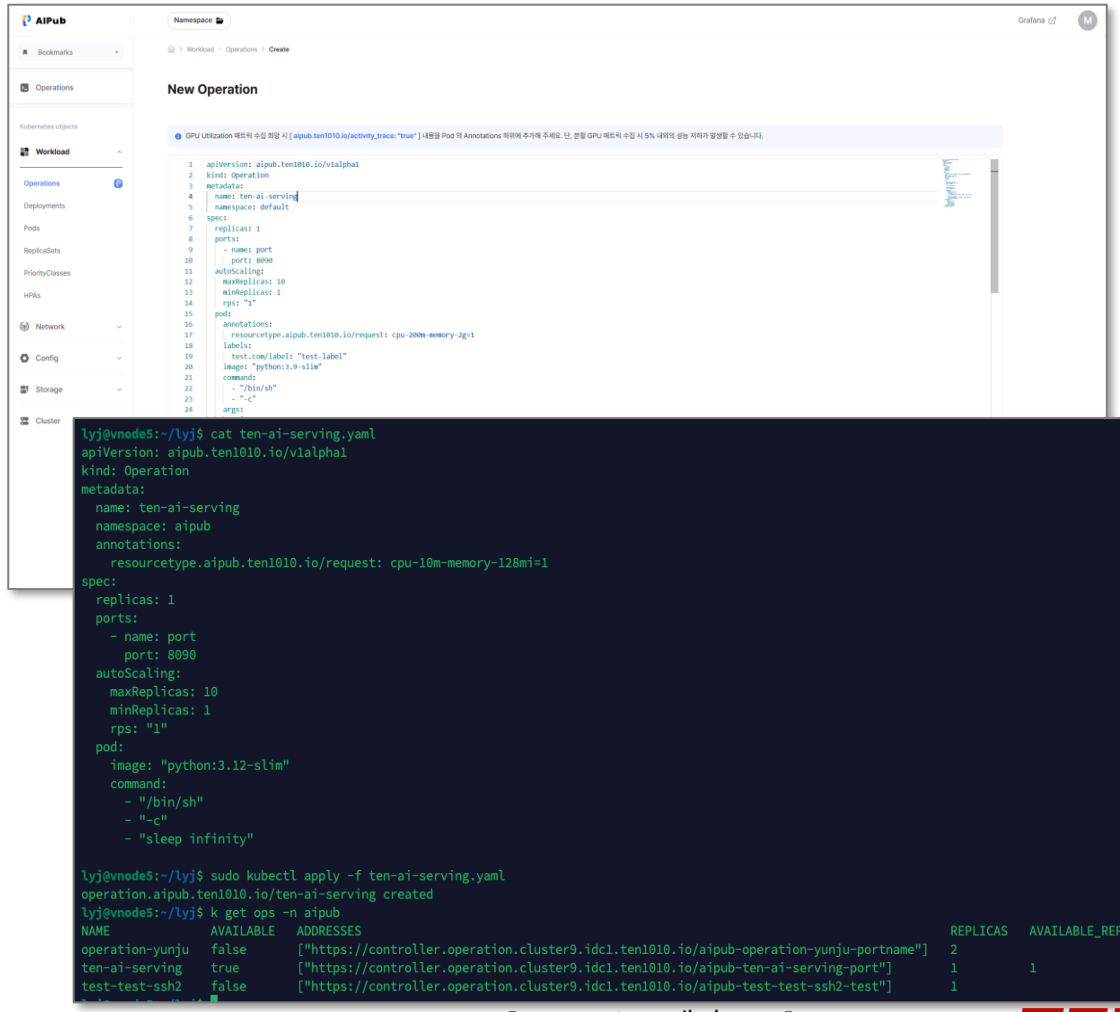
AI Pub K10s

다양한 사용자 환경(Web UI, YAML editor, CLI)에서 동일하게 생성 및 편집 가능하며 이를 하나의 GUI에서 관리



[Operation 예시: Form UI]

[Operation 예시: YAML editor]




```
lyj@vnode5:~/lyj$ cat ten-ai-serving.yaml
apiVersion: aipub.ten1010.io/v1alpha1
kind: Operation
metadata:
  name: ten-ai-serving
  namespace: aipub
  annotations:
    resourcetype.aipub.ten1010.io/request: cpu-10m-memory-128mi=1
spec:
  replicas: 1
  ports:
    - name: port
      port: 8090
  autoScaling:
    maxReplicas: 10
    minReplicas: 1
    rps: "1"
  pod:
    image: "python:3.12-slim"
    command:
      - "/bin/sh"
      - "-c"
      - "sleep infinity"

lyj@vnode5:~/lyj$ sudo kubectl apply -f ten-ai-serving.yaml
operation.aipub.ten1010.io/ten-ai-serving created
lyj@vnode5:~/lyj$ k get ops -n aipub
NAME             AVAILABLE   ADDRESSES                                                                                  REPLICAS   AVAILABLE_REPLICAS
operation-yunju  false       ["https://controller.operation.cluster9.idc1.ten1010.io/aipub-operation-yunju-portname"]   2
ten-ai-serving   true        ["https://controller.operation.cluster9.idc1.ten1010.io/aipub-ten-ai-serving-port"]        1          1
test-test-ssh2   false       ["https://controller.operation.cluster9.idc1.ten1010.io/aipub-test-test-ssh2-test"]        1
```

[Operation 예시: CLI]

RA:X

AI Reference Architecture of TEN

TEN의 노하우를 더한 테스트 기반의 AI 인프라 구축 컨설팅 서비스

하드웨어를 예산에 맞춰 각각 구입하도록 제안하지 않습니다.
인프라 전체의 가치를 볼 수 있게 하겠습니다.

“저희는 Llama 4를 학습할 계획입니다.”

고객

모델, 데이터 샘플

AI Pub &
AI 인프라 구성 제안

RA:X

성능 측정치

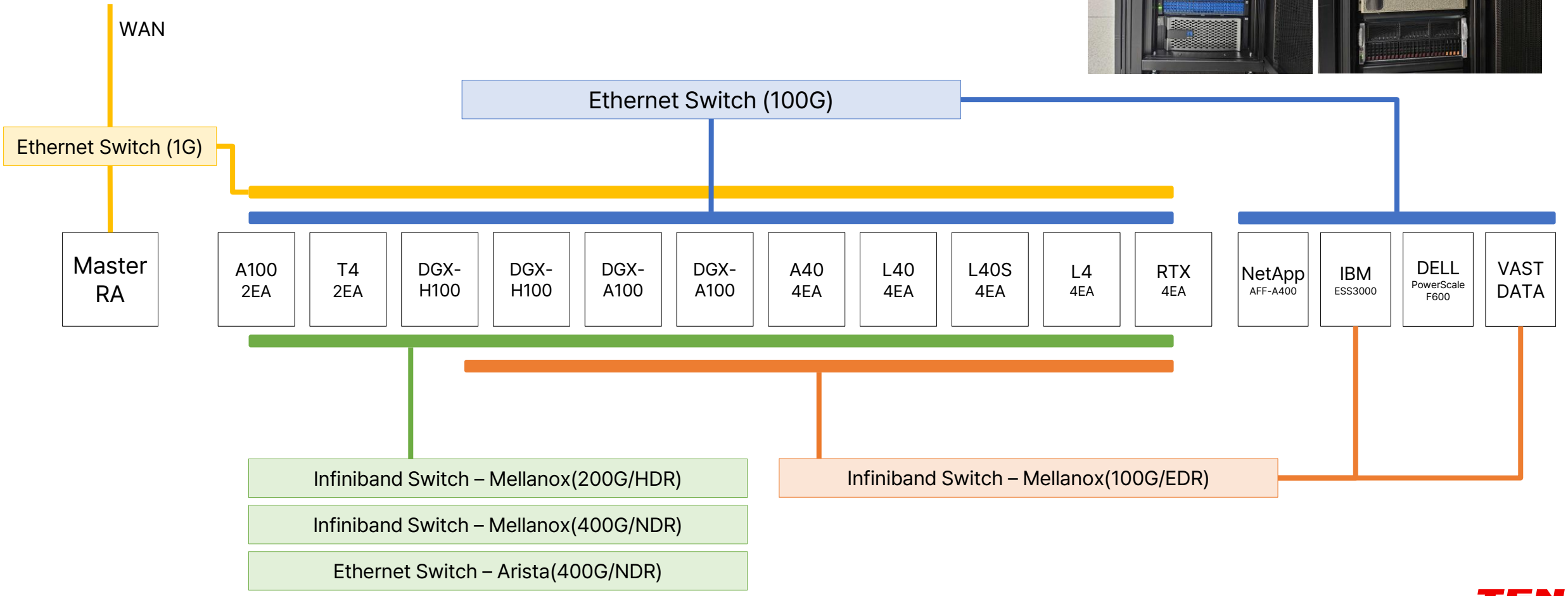
측정치에 맞는
HW 견적

“그 정도 수치면 이 HW 가 알맞습니다.”

하드웨어 벤더

TEN이 직접 구축하고 운영하는 Reference Architecture을 소개합니다.

최신 GPU 서버와, 스토리지, 네트워크로 구성한 레퍼런스 아키텍처를 구성하였습니다.
AI 모델 별 학습과 운영을 위한 컴퓨팅 리소스 측정이 가능합니다.



RA:X 를 통해 DeepSeek의 서비스 운영 인프라 구성을 비용 효율적으로 따져보기

DeepSeek-V3, R1

파라미터 개수: 671B
활성화 파라미터: 37B

8 NVIDIA H200 GPUs providing 1128 GB of GPU memory

Distilled Model Evaluation

Model	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	44.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

Inter-Token latency simulation

10ms

LLM 서비스를 제공하기 위해서는 초고속 연산 자원과 대용량 데이터 처리 능력이 필수적이어서 막대한 비용이 소요된다. 이로 인해 각 모델의 특성과 운영 환경에 맞는 최적의 인프라 구성이 절실하며, 효율성과 안정성을 동시에 갖춘 시스템 구축이 요구된다.

주식회사 Ten에서는 RAX 서비스를 통해 고객이 사용하려는 모델에 최적화된 효율적인 인프라를 구축할 수 있도록 지원하며, 비용 절감과 운영 효율성 극대화를 도모한다. 당사의 솔루션은 주어진 예산 내에서 최적의 성능을 발휘하여 LLM 서비스의 경쟁력을 강화하고, 고객의 비즈니스 성공을 견인하는 핵심 전략으로 자리매김할 것이다.

50ms

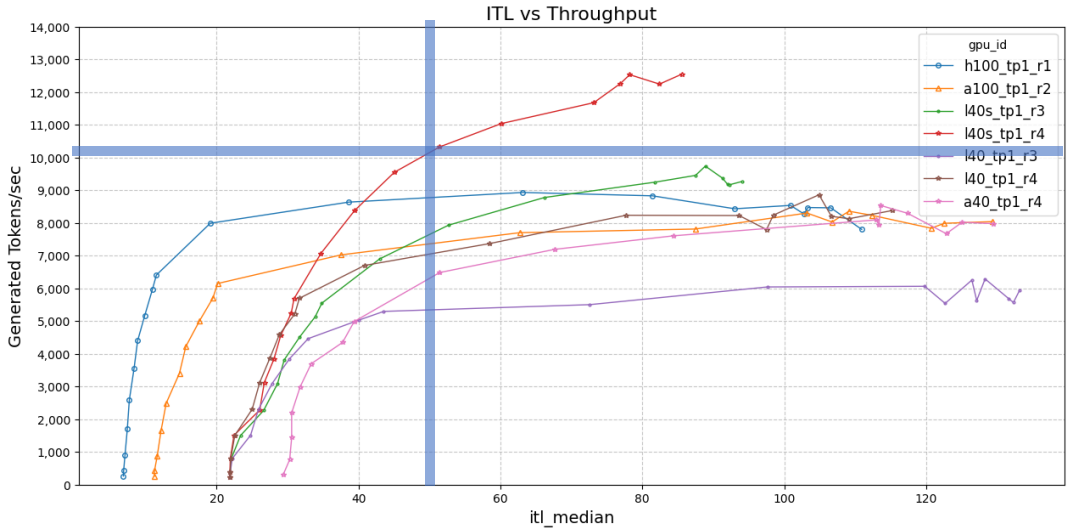
100ms

250ms

LLM 서비스를 제공하기 위해서는 초고속 연산 자원과 대용량 데이터 처리 능력이 필수적이어서 막대한 비용이 소요된다. 이로 인해 각 모델의 특성과 운영 환경에 맞는 최적의 인프라 구성이 절실하며, 효율성과 안정성을 동시에 갖춘 시스템 구축이 요구된다.

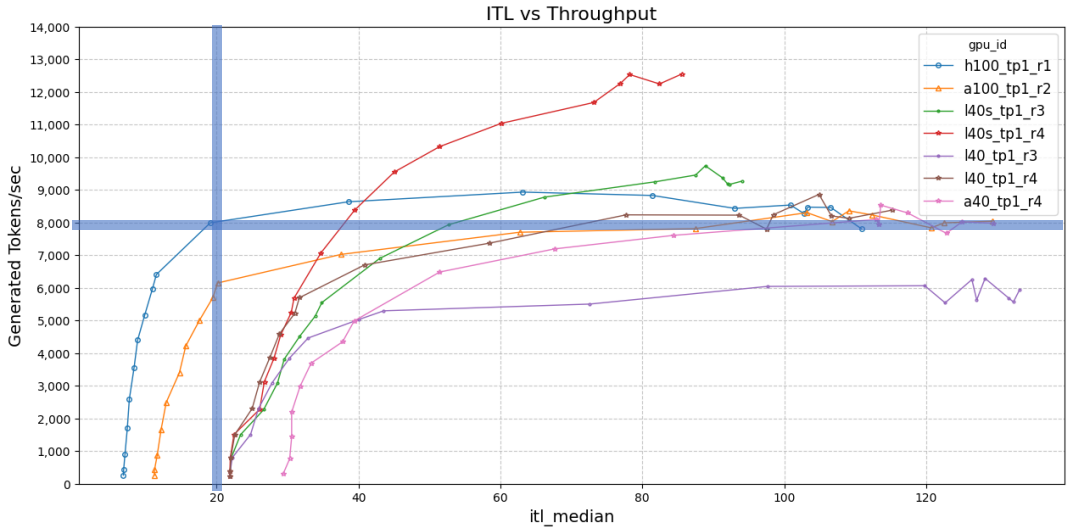
주식회사 Ten에서는 RAX 서비스를 통해 고객이 사용하려는 모델에 최적화된 효율적인 인프라를 구축할 수 있도록 지원하며, 비용 절감과 운영 효율성 극대화를 도모한다. 당사의 솔루션은 주어진 예산 내에서 최적의 성능을 발휘하여 LLM 서비스의 경쟁력을 강화하고, 고객의 비즈니스 성공을 견인하는 핵심 전략으로 자리매김할 것이다.

DeepSeek-R1-Distill-Qwen-7B 모델의 최적의 운영 GPU 자원 제안



- **Target Inter-Token Latency = 50ms**

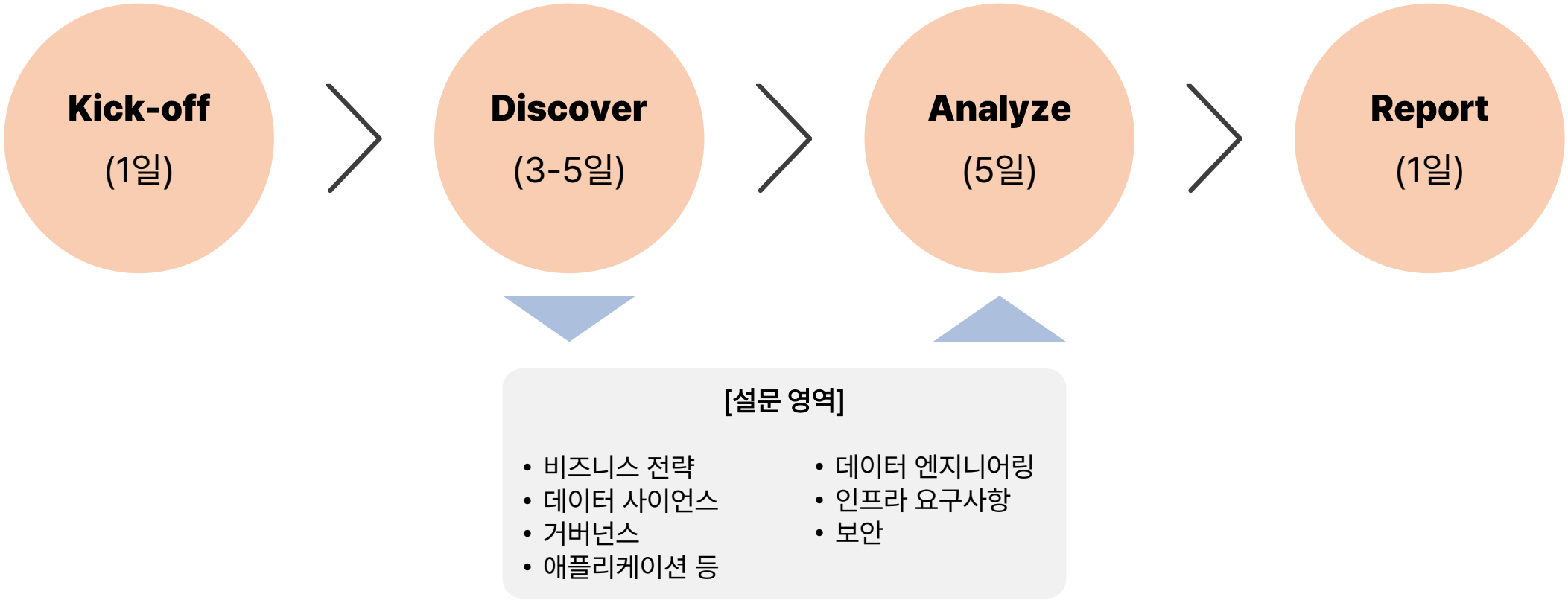
  ➔ L40s 4개



- **Target Inter-Token Latency = 20ms**

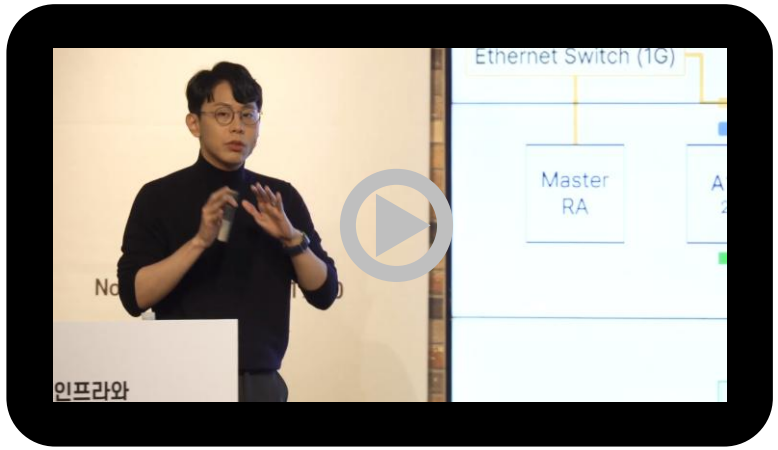
  ➔ H100 1개

RA:X 를 기반으로 한 인프라 구성 컨설팅은 아래의 프로세스로 진행됩니다.

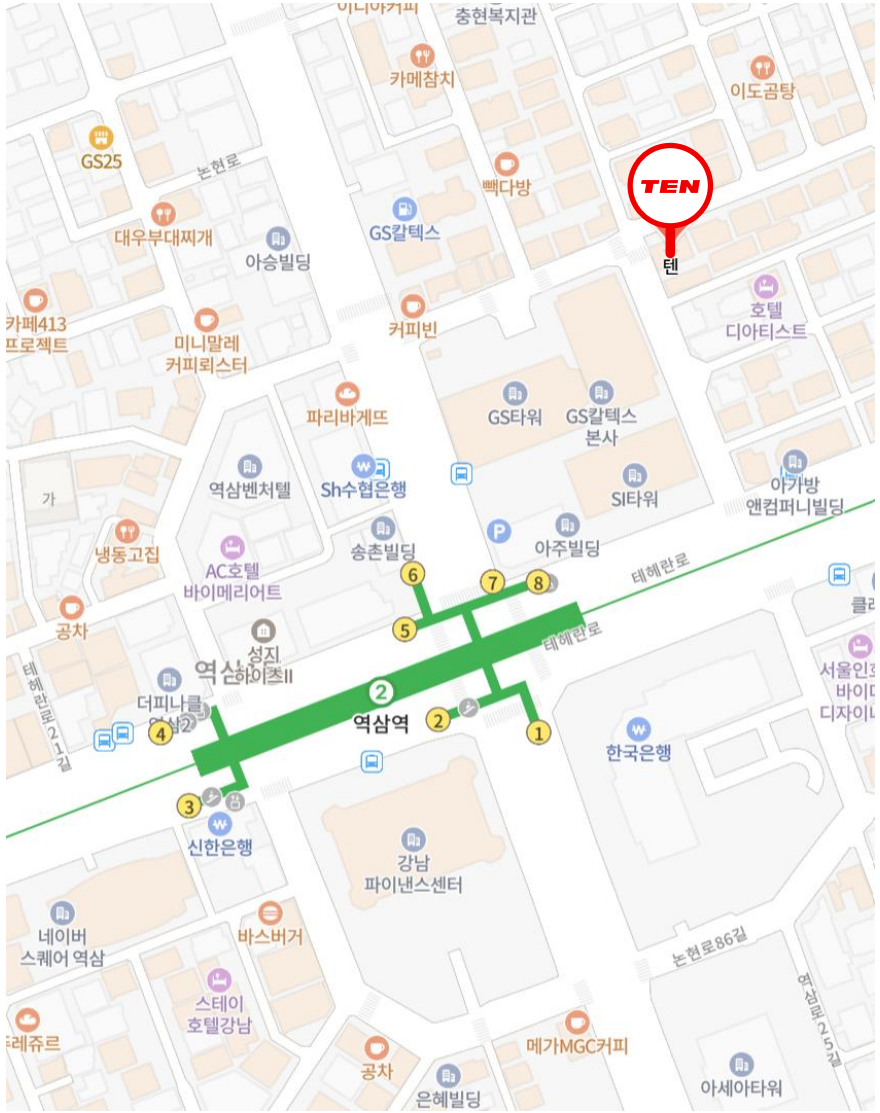


TEN은 AI 인프라 구성의 어려움과 해결 방안에 대한 세미나, 워크샵을 개최합니다.
인프라에 대한 고민에 공감하고, 해결을 돕기 위한 TEN의 목소리를 다양한 곳에 전합니다.

2023.11
AI 인프라와 NVIDIA 인증 솔루션
통합 구성 워크샵
Opening Session by TEN



2023.11
AI 인프라와 NVIDIA 인증 솔루션
통합 구성 워크샵
Session C, Closing Session by TEN



MORSE ME
WHEN YOU NEED
HELP WITH AI.

경쟁력 있는 AI를 위한 파트너가 필요하신가요?

고효율 솔루션과 신뢰도 높은 파트너십을 갖춘 주식회사 텐과의 협업으로
이전과 다른 AI 개발·운영을 경험하실 수 있습니다.

A 서울특별시 강남구 테헤란로27길 18, 5층

P +82-2-6956-1071

M helloten@ten1010.io

H https://ten1010.io/

AI로운
세상을 향해.

Make AI accessible for all

THANK YOU