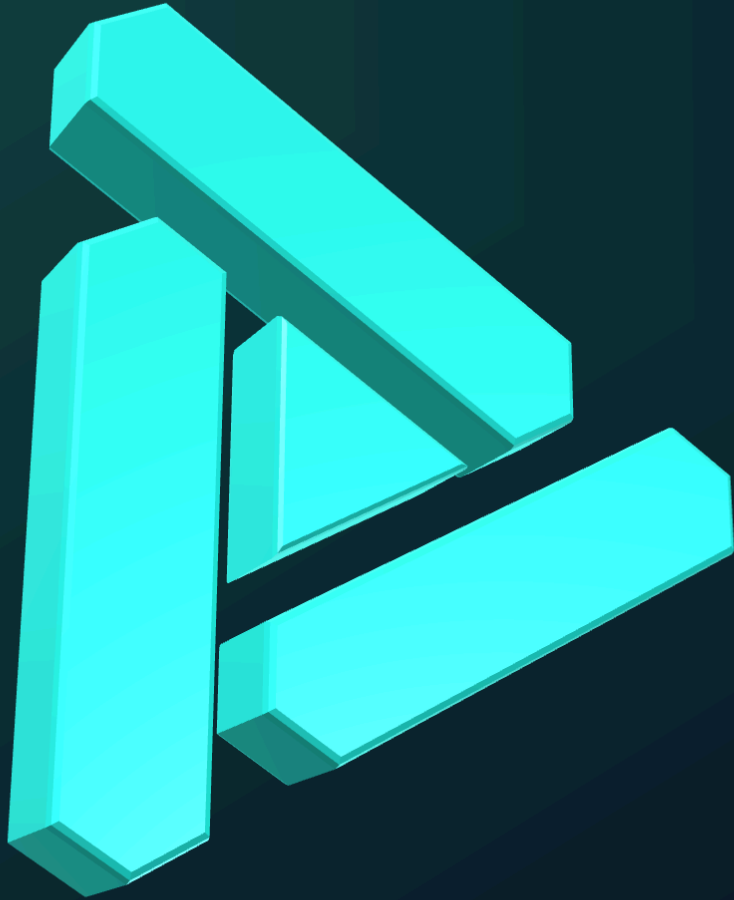




AI Inference Platform

Metis



Contents

1.	개요	02
2.	비즈니스 임팩트	04
3.	Metis 제품 구성 Metis (AI Inference)	06
4.	관측성·보안·운영	10
5.	Works with	12

1. 개요

흩어진 xPU 자원을 하나로 묶어 inference를 돌립니다.

GPU·NPU가 여러 클러스터에 나뉘어 있어도 한 곳에서 다룹니다. Kueue 기반 스케줄링이 우선순위와 할당량에 따라 자원을 배분하고, 서비스는 쓴 만큼 내는 방식과 전용 자원을 잡아 두는 방식 중에 골라 운영합니다.

xPU 이기종 오케스트레이션

다양한 가속기(GPU·NPU)를 효율적으로 조율하는 클라우드 네이티브 inference 플랫폼

복잡성 추상화

쿠버네티스, 멀티클러스터 스케줄링, xPU 관리 복잡성 제거로 개발 경험 향상

스케줄링과 자원 배분

Kueue 기반 대기열과 우선순위로 GPU를 배분하고, 쓰지 않는 자원은 회수합니다. 여러 팀이 같은 자원을 나눠 써도 순서가 엉키지 않습니다.

두 가지 운영 방식

쓴 만큼 내는 방식과 전용 자원을 잡아 두는 방식 중에 고릅니다. 트래픽이 들쭉날쭉하면 앞을, 성능이 일정해야 하면 뒤를 씁니다.

모델은 어디서 오나

회사가 검증해 둔 모델을 목록에서 골라 그대로 띄웁니다. 파일을 내려받아 다시 올리는 과정이 없습니다.

2.

비즈니스 임팩트



AI 인프라 투자를 실질적인 비즈니스 성과로 전환합니다.

Metis는 Kueue 스케줄링과 워크로드 실행, serverless inference를 하나의 플랫폼에서 제공합니다.

01 출시 속도 가속화

준비된 inference 엔진으로 endpoint를 띄우는 원클릭 배포

02 단가 및 비용 효율성

Kueue 기반 대기열·우선순위 스케줄링과 유휴 자원 회수

03 운영 안정성 및 우수성

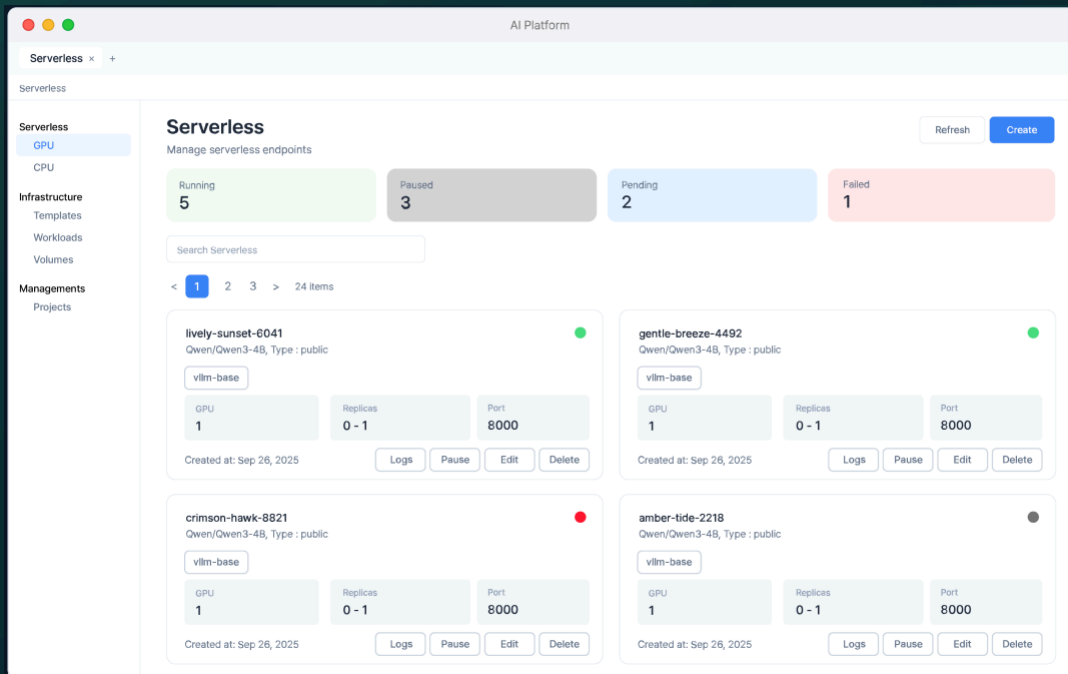
쿠버네티스 네이티브 멀티클러스터 운영과 통합 관측성

3.

Metis 제품 구성

사용량에 따라 자동으로 확장되는 완전관리형 inference 레이어

인프라를 직접 준비하지 않고 endpoint를 띄웁니다. 요청량에 따라 자동으로 늘고 줄며, 쓰지 않을 때는 자원을 최소로 내려 비용을 아낍니다.



Serverless Inference: 완전관리형 사용량 기반 inference

인프라 관리 없이 사용량에 따라 비용
지불하는 완전관리형 inference 레이어

OpenAI 호환 API

폐쇄형 제공자로부터 쉬운
마이그레이션을 위한 OpenAI 호환 API

오픈소스·멀티모달 모델

오픈소스 및 멀티모달 모델 지원

자동 스케일링

초당 토큰 처리량 및 요청 볼륨 기반으로
자동 확장

vLLM 기반 엔진

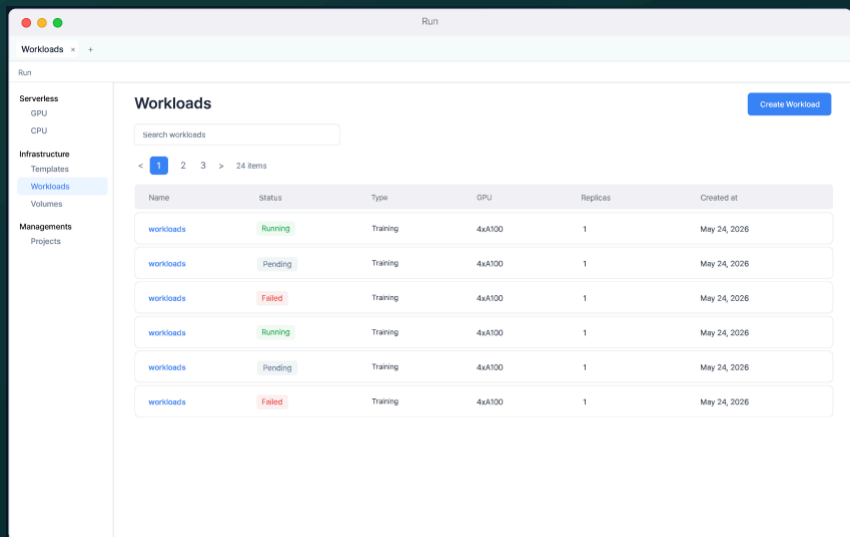
높은 처리량, 낮은 지연시간, 효율적인
KV 캐시 활용

수명주기

쓰지 않을 때는 잠시 멈춰 비용을
아끼고 필요할 때 바로 되살립니다.
지우는 것만 한 번 더 확인을 거칩니다.

전용 자원으로 일관된 성능을 보장하는 워크로드 실행 환경

성능이 일정해야 하는 서비스는 자원을 미리 잡아 두고 띄웁니다. 배포부터 종료까지 한 화면에서 다루고, 지금 할 수 없는 일은 버튼이 열리지 않습니다.



전용 자원으로 직접 띄우기

성능이 일정해야 하는 서비스는 자원을 잡아 두고 운영합니다. 트래픽이 몰려도 응답 속도가 흔들리지 않습니다.

실수를 막는 운영

배포부터 종료까지 한 화면에서 하되, 지금 할 수 없는 일은 아예 눌리지 않습니다. 운영에 들어간 서비스는 잠가 두면 지우거나 멈출 수 없습니다.

전용 자원 · 독립 환경

보안이 중요한 서비스는 다른 고객과 섞이지 않는 전용 자원에 올립니다. 네트워크도 분리됩니다.

SLA

업타임, 지연 시간, 용량을 보장하는 엔터프라이즈급 서비스 수준 협약

예측 가능한 성능·비용

안정적인 프로덕션 환경에서 일관된 성능과 명확한 비용 구조

멀티 클러스터 배포

Virtual Cloud를 통해 여러 클러스터에 동시 배포



검증된 자산과 구성을 반복 활용하는 재사용 기반

저장 공간과 템플릿, 회사가 승인해 둔 모델을 매년 새로 만들지 않고 재사용합니다. 한 번 정해 둔 구성이 다음 배포에서 그대로 섭니다.

저장 공간

데이터를 담은 공간을 만들고 필요할 때 늘립니다. 한 번 정한 저장 방식은 나중에 바뀌지 않아 운영 중에 설정이 흔들릴 일이 없습니다. 시점별로 백업해 두었다가 되돌릴 수도 있습니다.

자산 카탈로그 연동

회사가 승인해 둔 모델을 파일로 내려받아 다시 올리지 않고 목록에서 골라 바로 띄웁니다.

템플릿

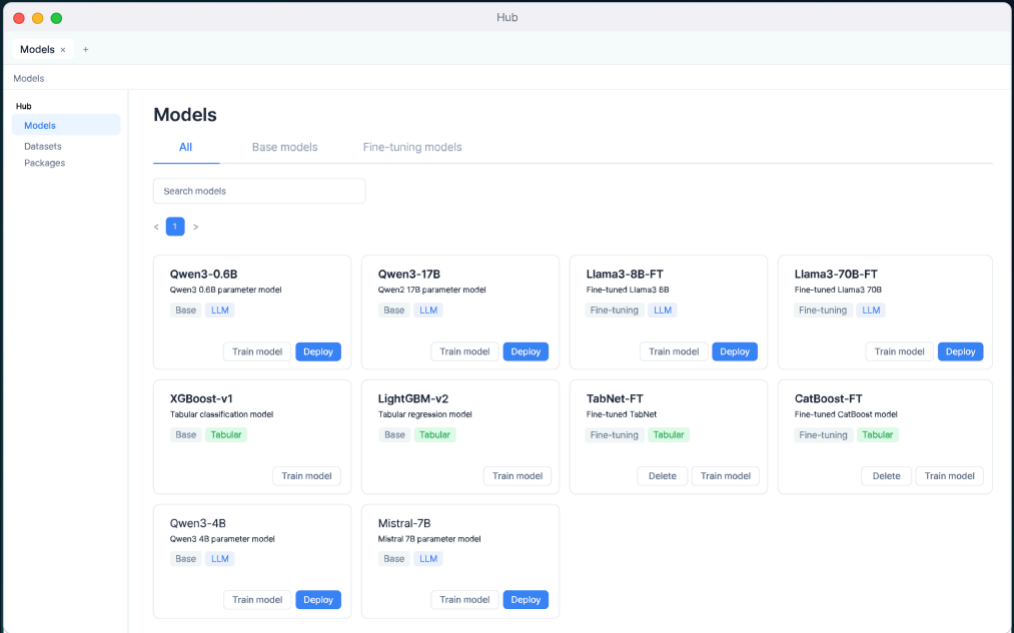
자주 쓰는 구성은 템플릿으로 저장해 다음에 클릭 몇 번으로 띄웁니다. 용도별로 분류해 두고 팀 전체에 공유할지 나만 쓸지 고를 수 있습니다.

Maxis에서 넘어온 모델

우리 데이터로 다듬은 모델도 같은 목록에서 골라 띄웁니다. training을 마친 뒤 옮기거나 다시 준비하는 단계가 없습니다.

프로젝트 격리

내 프로젝트 것만 보입니다. 다른 팀이 만든 자원은 목록에 뜨지 않아, 부서나 고객사별로 나눠 쓰는 환경에서 서로의 데이터를 볼 수 없습니다.



4.

관측성·보안·운영

엔터프라이즈 AI 운영에 필요한 가시성, 보안, 운영 통제를 제공합니다.

AI 모델과 워크로드, GPU/xPU 인프라 전반의 운영 상태를 통합적으로 관찰하고 관리할 수 있는 운영 기반을 제공합니다. 성능 지표, 로그, 접근 제어, 네트워크 격리, 배포 자동화 기능을 통해 기업이 요구하는 안정성, 보안성, 운영 일관성을 확보할 수 있습니다.

엔터프라이즈 AI 운영 기능

xPU 리소스 지표 모니터링과 자원 배분까지 안정적 AI 운영에 필요한 것을 플랫폼에 내재화

관측성

자원 사용률·처리량·응답 시간·오류율을 한 화면에서 봅니다. 여유가 없어지면 색으로 알려줘 미리 대응할 수 있습니다.

관리자 화면

관리자는 별도 화면에서 누가 얼마나 쓰는지 보고 자원 배분을 조정합니다. GPU가 모자라면 요청이 줄을 서고 자리가 나면 시작되며 급한 것을 먼저 보내도록 순위를 정할 수 있습니다.

격리 & 규정 준수

네임스페이스 격리, 쿠버네티스 네트워크 정책, 암호화된 시크릿 관리

5.

Works with



쌓인 자산이 그대로 서비스가 됩니다.

자산 허브에 등록된 모델을 골라 띄우고, 성능을 재 보고, 프롬프트로 응답을 확인한 뒤 서비스로 씁니다. 쓴 만큼은 토큰으로 집계됩니다. Maxis에서 다듬은 모델도 같은 목록에 있습니다.

01 자산 선택

자산 허브에 등록된 모델을 목록에서 고릅니다. 파일로 내려받아 다시 올리는 과정이 없습니다.

02 Endpoint 배포

고른 모델을 inference endpoint로 올립니다. 준비된 엔진을 쓰는 길과 직접 만든 이미지를 쓰는 길이 있습니다.

03 성능 측정

띄운 endpoint의 성능을 그 자리에서 잹니다. 후보가 여럿이면 견줘 보고 고릅니다.

04 프롬프트 확인

프롬프트를 넣어 응답을 바로 확인합니다. 애플리케이션을 붙이지 않아도 됩니다.

05 토큰 사용량 집계

inference에 쓴 토큰 사용량을 집계해 보여줍니다. 어느 서비스가 얼마나 썼는지 파악할 수 있습니다.

함께 쓰는 것

Signum 자산 허브와 접근·권한·감사의 공통 기반 · Capsis 쿠버네티스 실행 기반 · Maxis 모델 training

CTA

도입 문의 · 데모 요청