SimPlatform

KOSDAQ
코스닥상장법인

기업 AI 인프라 운영에 최적화된 구축형 쿠버네티스

Enterprise Kubernetes Platform for Industrial AI Infrastructure

NUBISON Kubernetes

AX 시대 AI 인프라가 경쟁력을 결정합니다

수십 개의 AI모델이 동시에 실행되는 시대
인프라가 AI의 속도를 따라가지 못하면 현장 가치가 사라집니다.

NUBISON Kubernetes 배경 | AX의 속도는 곧 인프라의 속도

이기종 리소스의 분산 관리

GPU·CPU·NPU·스토리지가
개별적으로 관리되어 자원 활용
효율이 저하됩니다.

서비스 중단 위험

모델 배포, 패치, 장애 대응 시
서비스 중단이 발생하여
운영 연속성이 저하됩니다.

예측하기 어려운 AI 워크로드

학습과 추론 워크로드가 급변하여
수동 자원 운영만으로는
안정적인 대응이 어렵습니다.

NUBISON Kubernetes는 AI 인프라 운영을 위한 통합 제어면(Control Plane)을 제공합니다.

핵심 가치

- CPU·GPU·NPU를 하나의 플랫폼에서 통합 관리
- 무중단 운영으로 서비스 연속성 보장
- 기업 환경에 최적화된 구축형(On-Premises) AI 인프라 제공

핵심 가치 및 아키텍처

네 가지 핵심 역량으로 AI 인프라를 완성하고
다양한 AI 서비스가 하나의 Kubernetes 클러스터에서
독립적으로 격리되어 안정적으로 운영됩니다.

AI 인프라 운영의 핵심 가치

리소스는 더 효율적으로 서비스는 더 안정적으로 운영은 더 단순하게

01
리소스 통합 · 효율화
GPU·CPU·NPU·스토리지를 통합 관리하고
워크로드에 맞춰 자원을 최적으로 할당합니다.

02
무중단 · 자가 복구
AI 서비스의 지속적인 운영을 지원하는
고가용성 아키텍처를 제공합니다.

03
엔터프라이즈 제품화
13개 핵심 컴포넌트의 버전 호환성·의존 순서·운영
표준을 단일 벤더가 보증합니다.

04
구축형 운영 노하우
100여 건의 산업 AI 프로젝트를 통해
검증된 운영 노하우와 자동화 자산을 제공합니다.

아키텍처

하나의 클러스터, 수십 개의 AI 서비스, 완전한 운영 독립성

Kubernetes 핵심 기능

이기종 AI 리소스를 통합하고 최적으로 활용합니다.

기능01 UNITED RESOURCE | GPU·CPU·NPU·Storage를 하나의 리소스 풀로 통합

학습 · 대규모 추론
GPU 관리

- MIG / MPS 기반 GPU 분할 공유
- 모델별 GPU 쿼터 · 우선순위 스케줄링
- Fractional GPU로 가용률 극대화
- NVIDIA GPU Operator 통합 관리

온디바이스 · 엣지 가속
NPU 관리

- 이기종 NPU 벤더 통합 드라이버
- 모델 컴파일 · 배포 파이프라인 내재화
- 엣지 노드까지 확장되는 단일 제어면
- 디바이스 플러그인 자동 등록

전처리 · API · 오케스트레이션
CPU 관리

- 코어 수 기반 세밀한 리소스 쿼터
- NUMA · CPU Pinning 최적화
- AI 워크로드 외 일반 서비스 동거
- QoS 클래스별 우선 스케줄링

데이터셋 · 모델 아티팩트
Storage 관리

- 블록 · 오브젝트 · 파일 통합 프로비저닝
- 데이터 지역성(Locality) 기반 배치
- NUBISON Datalake와 직결
- 스냅샷 · PITR 기본 탑재

기능02 GPU INTELLIGENCE | GPU 리소스를 분할하고 공유하여 최적으로 배치

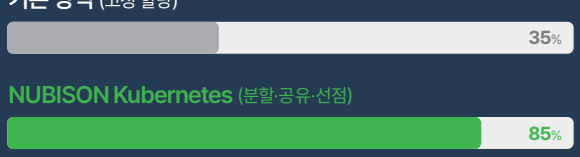
NUBISON Kubernetes는 GPU를 유연한 공유 자원으로 관리합니다. GPU를 여러 워크로드가 안전하게 공유할 수 있도록 분할하고, 우선순위에 따라 동적으로 할당하며, 유휴 자원은 즉시 회수하여 활용률을 극대화합니다.

Fractional GPU	1 GPU를 여러 추론 컨테이너가 공유
Quota & Priority	팀/프로젝트 단위 쿼터와 우선순위
Preemption	중요 학습 잡을 위해 유휴 잡 선점
Topology-Aware	NVLink · NUMA 위상 기반 최적 배치

GPU 활용률 비교

기존 방식 vs NUBISON Kubernetes

동일한 GPU 인프라에서 최대 2.4배
더 많은 AI 워크로드를 처리합니다.
GPU 증설 비용 절감 · TCO 최적화 · ROI 향상



기능03 SMART SCALING | AI 수요를 예측해 선제적으로 확장

STEP 01 | Observe

상시 관측

Pod · Node · GPU 지표,
모델 QPS, 큐 길이를 실시간 수집

STEP 02 | Predict

수요 예측

시계열 패턴 기반으로
향후 15~60분의 리소스 수요 예측

STEP 03 | Scale

선제 확장

HPA · VPA · Cluster Autoscaler를
조합해 수직·수평 확장

문제가 발생한 후 대응하는 것이 아니라 **발생하기 전에 준비하는 Predictive Kubernetes**

기능04 NON-STOP OPERATION | AI 서비스를 위한 4중 무중단 운영 체계

Rolling Update

무중단 롤링 업데이트

신규 모델 버전을 한 Pod씩 교체하며 배포
문제 발생 시 즉시 롤백

배포 다운타임 0초

Self-Healing

자가 복구 (Self-Healing)

Pod · 노드 장애를 자동 감지하여
건강한 노드로 즉시 재스케줄링

평균 복구 시간 < 30초

HA Control Plane

고가용 컨트롤 플레인

마스터 노드 다중화 · etcd 분산 구성으로
단일 장애점(SPOF) 제거

가용성 목표 99.95%

Traffic Mesh

안전한 트래픽 전환

Canary · Blue-Green 배포 전략으로
신·구 버전 트래픽을 비율로 제어

안전한 모델 검증

Kubernetes
차별점 및 기대효과

산업 AI의 경쟁력은 운영 환경과 노하우에서 시작되며
통합된 엔터프라이즈 플랫폼으로 도입 효과를 극대화합니다.

차별점01 현장에 최적화된 운영 환경 | 산업 AI는 데이터 주권, OT 연계, 예측 가능한 운영 비용이 핵심

구분	클라우드 매니지드 K8s	NUBISON Kubernetes (구축형)
데이터 주권	외부 클라우드로 데이터 전송 필요	기업 내부에서 안전하게 데이터 운영
현장 장비 연계	OT 장비 연계 시 네트워크 지연 발생 가능	NUBISON IoT와 초저지연 직접 연동
비용 구조	트래픽 · GPU 시간당 과금으로 예측 어려움	CAPEX 기반으로 예측가능한 총소유비용(TCO) 확보
커스터마이징	벤더 정책에 종속된 업그레이드 주기	기업 정책에 맞춘 유연한 운영
운영 지원	티켓 기반 표준 지원	전담 엔지니어의 구축·운영 지원

차별점02 운영 노하우 | 100여 건의 산업 AI 프로젝트에서 검증된 운영 경험을 제품에 내재화

Day-0 설치 자동화
표준 노드 이미지 · IaC
기반 1일 구축

Day-2 운영 대시보드
GPU · 모델 · 노드 상태를
통합 모니터링

보안 베이스라인
RBAC · 이미지 스캔
네트워크 정책 기본 탑재

장애 복구 플레이북
현장에서 검증된 RCA
Runbook 동봉

백업 · DR 표준
etcd · PV 스냅샷
리전 간 복제 자동화

운영 인력 교육
관리자 · 개발자 대상
맞춤 교육 프로그램

차별점03 통합된 엔터프라이즈 플랫폼 | 13개 핵심 컴포넌트를 하나의 검증된 플랫폼으로 제공

13

Core Components

사전 검증 · 통합 제공

단일 벤더 책임 · 일관된 운영 표준

배포 · 네트워크

- CD 환경
- 서비스 메시
- API 게이트웨이
- L2/L4 로드밸런서

데이터 · 저장소

- 분산 블록 스토리지
- MySQL 클러스터
- PostgreSQL
- 오브젝트 스토리지 (S3)
- 문서형 DB
- 인메모리 캐시

보안 · 관측 · 운영

- 시크릿 관리
- 통합 관측 플랫폼
- 내부 Git 저장소

도입 효과 | 리소스 효율, 서비스 안정성, 운영 생산성 전반에서 가시적인 성과를 제공

2.4X

행위 감사 추적

분할 공유
선점 스케줄링

99.95%

서비스 가용성

고가용 컨트롤 플레인
자가 복구

30초

장애 복구 시간

자동 재스케줄링
Pod 재시작

70%↓

운영 인력 공수

자동화된 Day-2
운영 플레이북

CAPEX 감축

GPU 증설 시점 지연으로
CAPEX 감축

시간 단축

AI 서비스
Time-to-Market 단축

수율/가동률 개선

현장 다운타임 최소화
수율/가동률 개선

SimPlatform
www.simplatform.com

(주)심플랫폼

(본사) 서울시 금천구 가산디지털1로 226 에이스하이엔드타워5차 2003호
(AX본부) 서울시 강남구 삼성로 555 알앤텍빌딩 4층
Tel. **02-6405-4338** | email **ax@simplatform.com**