

[GPU 관리솔루션]

AstraGo[®]

표준소개서

세상의 모든 GPU, AstraGo로 관리한다



Contents

목차

01. INTRO

02. 솔루션 소개

03. AstraGo 도입 여정

04. 도입 사례

05. Professional Service

06. Why Now?

I. INTRO

AI 영상분석 전문 기업 **씨이랩**은 인프라부터 AI 영상분석 · 디지털 트윈까지
산업군별 다양한 수요를 충족하는 **종합 Vision AI 역량** 보유

일반현황



업력 15년	특허 등록 건수 21건	연구 인력비율 70%
공공 데이터 사업 35개	출시 솔루션 06개	고객기업 176개

핵심 역량 및 기술

- 초고속 대용량 데이터 처리 기술 보유
 - 분산 영상 DB 기반 초고속 데이터 공급
 - 대용량 데이터 수집/저장/관리/분석
- 초고속 영상 데이터 가공 솔루션 보유
 - Auto Tracking 기반의 가공 기술
 - 소형 모델을 통한 데이터 자동 가공 기술

영상 데이터 기반
AI 분석 전문 기업

- 국내 최초 GPU 활용 솔루션 보유
 - 컨테이너 방식의 가상화
 - GPU 기반 딥러닝 개발 환경 경량화
- 국내 최초 실시간 영상분석 플랫폼 보유
 - 맞춤형 AI 모델 등록 및 분석
 - CH 별 분석 모델 스위칭 지원

주요 사업

AI Infra

AI 인프라 관리 솔루션
(NVIDIA GPU 내 당사 솔루션 탑재)

HPC, GPU서버, AI 데이터센터 등
GPU 관리효율 및 활용률 증대

데이터센터

연구기관

Vision AI

AI 영상분석 플랫폼
(Cloud/On-device)

CCTV, 미디어에서 생성되는
대용량 영상데이터를 AI로 분석

제조/바이오

미디어/엔터테인먼트

Digital Twin

디지털 트윈 플랫폼
(3D 시뮬레이션/합성데이터 생성)

단일 플랫폼에서 가상 시뮬레이션을
수행할 수 있는 디지털 트윈 기술

제조/건축

국방

대부분의 기업에서 GPU 활용률이 30% 미만, 프로젝트는 GPU 부족으로 3-6개월 지연 중, 고객은 더 많은 GPU가 아니라 더 똑똑한 관리를 원합니다.

자원이 급속히 늘고 있지만, 활용과 관리 효율성은 따라가지 못하는 실정



GPU 클러스터 자원 통합관리의 부재

평균 GPU 자원 활용률 약 20 ~ 30% 수준으로, 많은 GPU가 워크로드 없을 때 놀고 있음*
(최상위 기술기업의 GPU 활용률도 50% 미만)
Memory / Network 병목 때문에 전체 자원을 못 쓰기도 함



워크로드 대기 증가, 환경 세팅 복잡으로 인한 프로젝트 지연

84.7% 조직이 GPU 가용성 문제로 AI/ML 프로젝트 지연을 경험함*
전체 프로젝트 중 38.5%는 지연이 3-6개월 지속됨*



자원 낭비로 인한 비용 증가

under-utilized GPU 사용 → 연간 수십억원 손실 가능성



운영 복잡성 증대 및 DevOps 인력 부족

Memory/Network 병목, 다양한 워크로드 조정 필요
전문 인력 확보 어려움

*출처: Civo(2025). GPU-as-a-Service: The AI Tipping Point? Civo Press.



큐 관리에서 오케스트레이션으로, AstraGo가 AI 학습 인프라의 표준을 완성합니다

SLURM은 큐를 관리하고, Kubernetes는 환경을 표준화했습니다.
AstraGo는 그 위에서 GPU를 오케스트레이션하여 활용률 중심의 TCO 혁신을 완성합니다

01

명령어로 큐를 관리하던 HPC 시대

HPC 한정 # 수동 큐 관리 # 확장성 한계



- 연구자들이 서버에 명령어를 입력해 일일이 작업을 등록
- 모든 작업이 순서대로 대기하는 “줄서기 방식” 자원 관리
- 대형 슈퍼컴퓨터(HPC)에는 적합했지만, AI 개발엔 환경 설정이 어렵고 확장성이 떨어짐

02

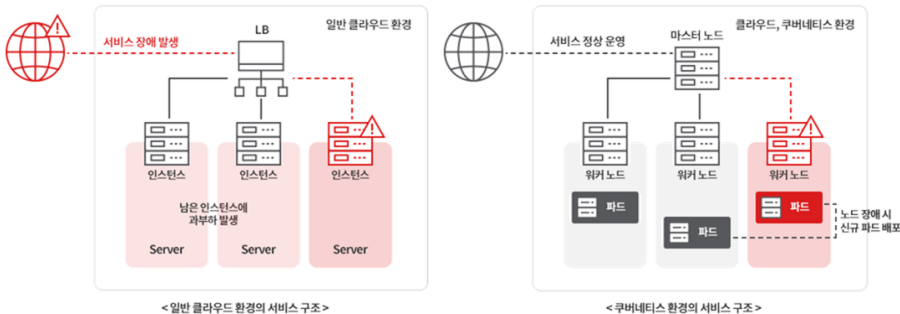
컨테이너 자동화를 통한 표준화의 시대

자동화 # 표준화 # GPU 관리 한계



- 여러 개발 환경을 컨테이너 단위로 자동 배포하고 관리
- 복잡한 시스템도 자동으로 복구(자가치유) 하고 확장 가능
- 하지만 GPU 자원을 세밀하게 관리하거나 AI 학습에 특화된 최적화는 지원이 제한적

일반 클라우드와 쿠버네티스 환경 비교



03

GPU 중심의 오케스트레이션 시대

#GPU 오케스트레이션 # 효율·안정성 #ROI 극대화



- GPU/CPU/RAM을 On-Demand분배하여 활용률을 극대화, Idle 자원 최소화
- 워크로드 변화에 따라 즉시 자원 확보, 초기 구축 부담 없이 빠른 프로젝트 착수 가능
- 자동화와 전문 엔지니어 지원으로 복잡한 관리 제거
- HW ~ SW 통합 유지보수 및 SLA 기반 복구체계

ML 프레임워크 라이브러리

AI 학습 지원 시스템

인프라/시스템 관리

컨테이너 오케스트레이션

H/W 인프라



워크로드 잡 선택

분산 학습

모델 허브 & 빌트인

학습시간 예측

모니터링 및 리포트

자원 최적화

잡 스케줄러

GPU 분할



GPU

CPU

RAM

네트워크

스토리지



서비스 서포트

단계별 Professional Service

Joint Delivery

클러스터 설계/최적화

무중단 운영지원&장애대응

단순 모니터링과 DevOps 중심의 접근으로는 GPU 문제를 해결할 수 없습니다.

A사 솔루션은 모니터링 중심, B사 솔루션은 DevOps 중심.
그러나 고객이 원하는 것은 **클러스터 자원의 효율적 오케스트레이션**입니다.

구분	A사 솔루션	B사 솔루션	AstraGo ⁺
핵심 포커스	GPU 자원 모니터링 및 스케줄링	- DevOps 워크플로우 자동화 - CICD 파이프라인 최적화	- GPU 오케스트레이션 서비스 지원 - 국내 고객환경 최적화 - On-site 기술 지원
강점	- GPU 사용 현황 파악 용이 - 글로벌 레퍼런스 다산	- DevOps 생산성 향상 - 워크플로우 효율화	- 효율성·생산성·확장성 동시 해결 - 고객 맞춤형 지원
한계	- 한국 현지화 부족으로 UI/UX 지원체계에서 불편 발생 - 특정 기능 위주로 고정되어 고객 환경 맞춤 적용이 어려움 - 높은 도입 및 운영비용으로 ROI 개선 효과 제한	- DevOps 팀 중심으로 관리자/사용자 관점 미흡 - GPU 활용률 최적화 한계 - End-to-End 관리 부재	-
지원범위	- 모니터링 중심 - 기능 제한적	DevOps 팀 중심의 톨	사용자 및 관리자 모두 지원 (서비스 엔지니어 포함)
차별 가치	가시성 제공 수준에 그침	DevOps 효율성에 한정	- 국내 고객 환경 최적화 - 경제적인 비용의 subscription 형태로 제공 - 전문 서비스 엔지니어 지원 - 높은 커스터마이징 유연성

AstraGo 도입 전과 후, 확연한 차이를 만듭니다

AstraGo는 고객의 문제를 직접 해결하는 GPU Orchestration 솔루션

AS-IS

GPU 자원 낭비

GPU 활용률 저조, Idle 자원 다수 → 투자 대비 ROI 저하

프로젝트 작업 대기시간 증가

GPU 가용성 부족으로 프로젝트 지연 및 일정 차질

환경 세팅·운영 복잡

DevOps 인력 의존 멀티워크로드 충돌로 운영 부담 확대

장애 대응 지연

대응 체계 부재 및 장애 근본원인 분석 부족으로 인한 동일 문제 반복 발생
외산 솔루션 현지 엔지니어 · 시차 · 커뮤니케이션 이슈로 대응 속도 저하

자체 교육 부족

운영팀 개별 학습 의존 및 표준화된 교육 부재

TO-BE

효율성 극대화

자동 최적화 및 GPU 분할(MPS/MIG)로 활용률 30~40% 향상, ROI 상승

프로젝트 운영 최적화

잡 스케줄링·쿼터 관리로 공정하고 빠른 자원 배분
프로젝트 착수속도 2배 향상

운영 단순화

Commit 이미지 & 빌트인 프레임워크로 운영 단순화
End-to-End 오케스트레이션으로 운영 관리 부담 최소화

신속 대응 & 안정성 보장

전문 엔지니어의 SLA 지원
장애 원인 분석 리포트 제공 및 재발 방지를 위한 구체적 대책 수립

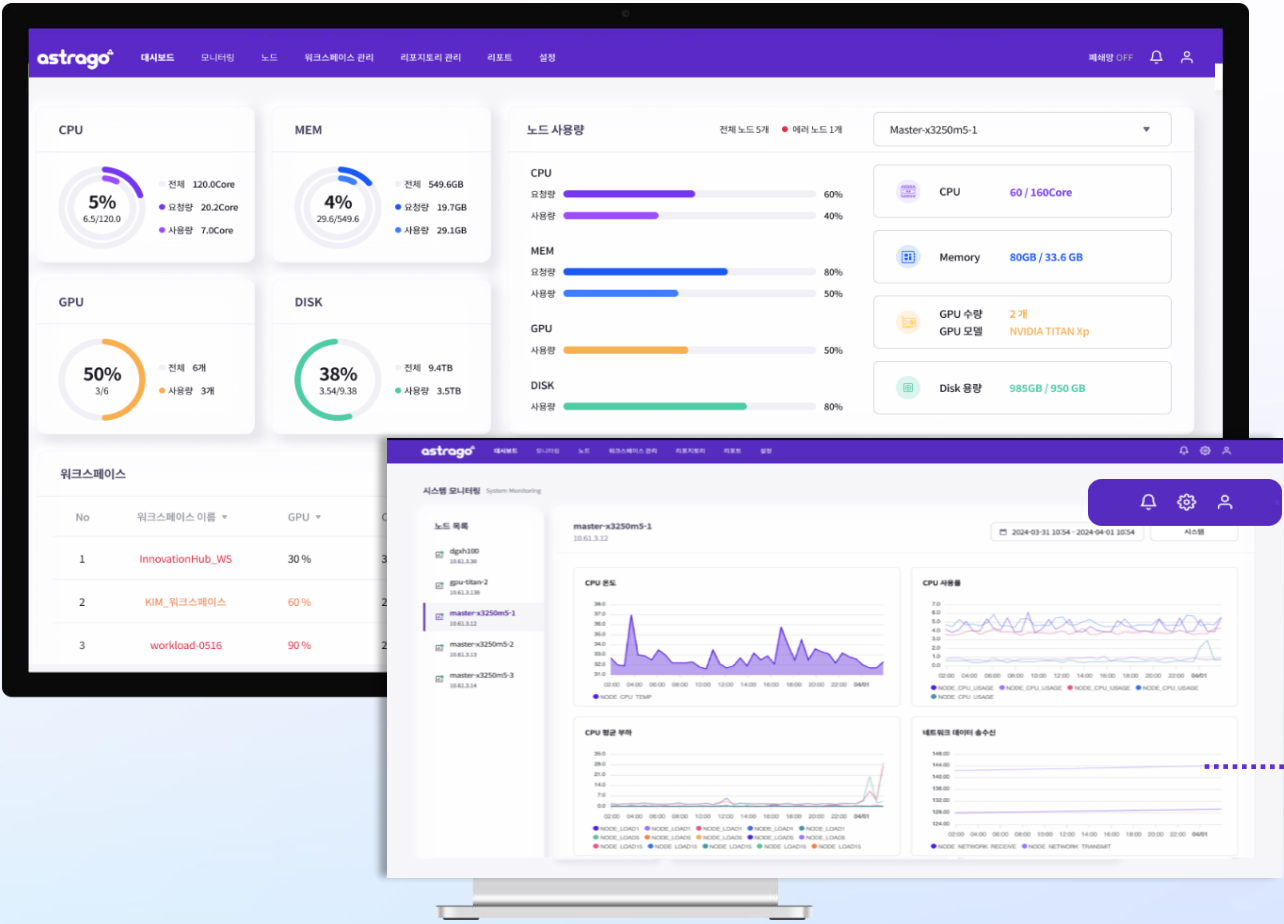
표준화된 교육 제공

AstraGo 무상 교육 & 맞춤형 현장 교육 제공

II. 솔루션 소개

모니터링을 넘어, 낭비없는 운영을 위한 가시성 확보.
실시간 모니터링& 리포트 + 데이터 기반 인사이트

실시간 모니터링 및 리포트



상단메뉴

노드별
모니터링

워크스페이스 관리

리포지토리 관리

리포트 및
설정

알람 표시

- 노드 장애, 사용자의 리소스 요청
- 최적화를 위한 리소스 회수
: 지정한 시간 내 지정한 수준 (ex. 10%)
- 미달 워크로드에 대한 알림 → 리소스 회수

Node 모니터링 정보

- GPU 사용률 및 메모리 용량
- GPU 온도, 전력 사용량, 에러
- CPU 사용률 및 평균부하, 온도
- 디스크 사용률 및 메모리 사용률
- 네트워크 송수신

GPU·CPU·메모리 사용 현황을 직관적으로 시각화하여, 자원의 Idle 상태와 병목을 빠르게 파악합니다.

낭비없는 GPU 활용, ROI를 극대화하다.

자원 최적화·GPU분할(MPS/MIG)·리소스 쿼터관리·Job 스케줄러로 GPU 효율성을 극대화합니다.

자원 최적화

워크스페이스 디폴트 리소스 설정

- 워크스페이스 당 배정할 최소 리소스 단위 설정

워크스페이스 기본 리소스 설정

워크스페이스 기본 리소스 설정은 워크로드 실행을 위한 초기 GPU/CPU/Memory 할당을 지정합니다.

GPU	2개	CPU	4Core	MEM	32GB	수정하기
-----	----	-----	-------	-----	------	------

정책 설정 Policy Settings

자원 최적화 설정

자원 사용률이 일정 시간 동안 설정 기준값을 넘지 못할 경우 자원이 수거됩니다.

Interactive Job	운영시간	24시간동안	GPU	10%	CPU	10%	MEM	10% 미만	사용유무	☐	수정하기
Batch Job	운영시간	24시간동안	GPU	10%	CPU	10%	MEM	10% 미만	사용유무	☒	수정하기

워크스페이스 기본 리소스 설정

워크스페이스 기본 리소스 설정은 워크로드 실행을 위한 초기 GPU/CPU/Memory 할당을 지정합니다.

GPU	2개	CPU	4Core	MEM	32GB	수정하기
-----	----	-----	-------	-----	------	------

자원 최적화 설정

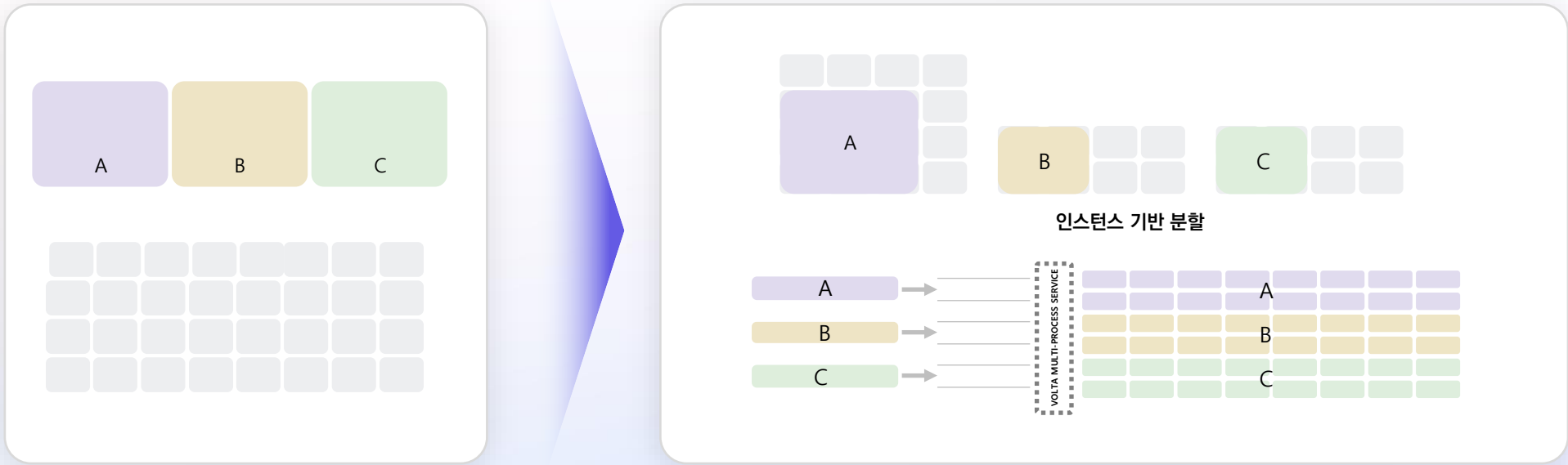
- 일정 수준의 사용량 미달 시 자동회수 정책 설정
- 자동 회수 작용 시 관리자 / 사용자에게 알림 전송

워크로드 상황에 맞춰 GPU/CPU를 자동 분배로 Idle 상태 최소화, 운영 효율을 극대화합니다.

낭비없는 GPU 활용, ROI를 극대화하다.

자원 최적화·GPU분할(MPS/MIG)·리소스 쿼터관리·Job 스케줄러로 GPU 효율성을 극대화합니다.

GPU 분할기술



학습 · 추론 공통으로 GPU 분할 기술 지원



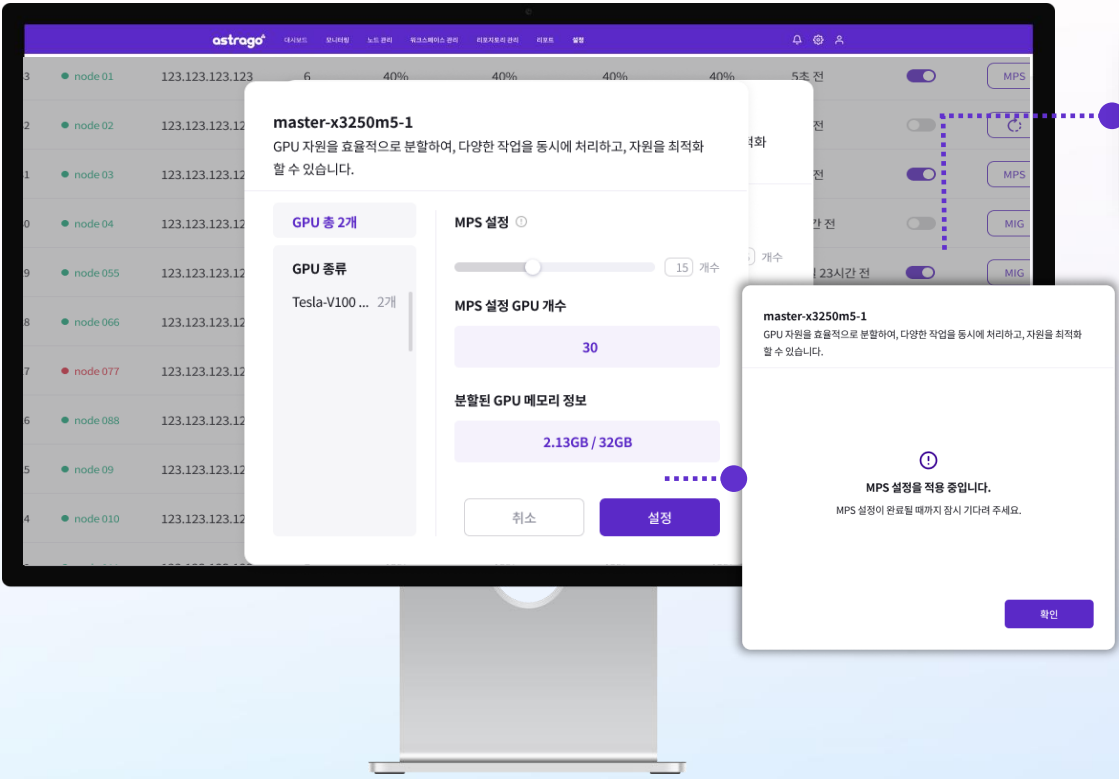
적은 리소스로도 수행될 수 있는 추론 등의 워크로드에 대해서는 GPU 분할 기술을 통한 자원 효율 극대화

씨이랩의 GPU 분할 기술로, 단일 GPU 내에서 추론 등 다수의 소형 워크로드 할당이 가능합니다.

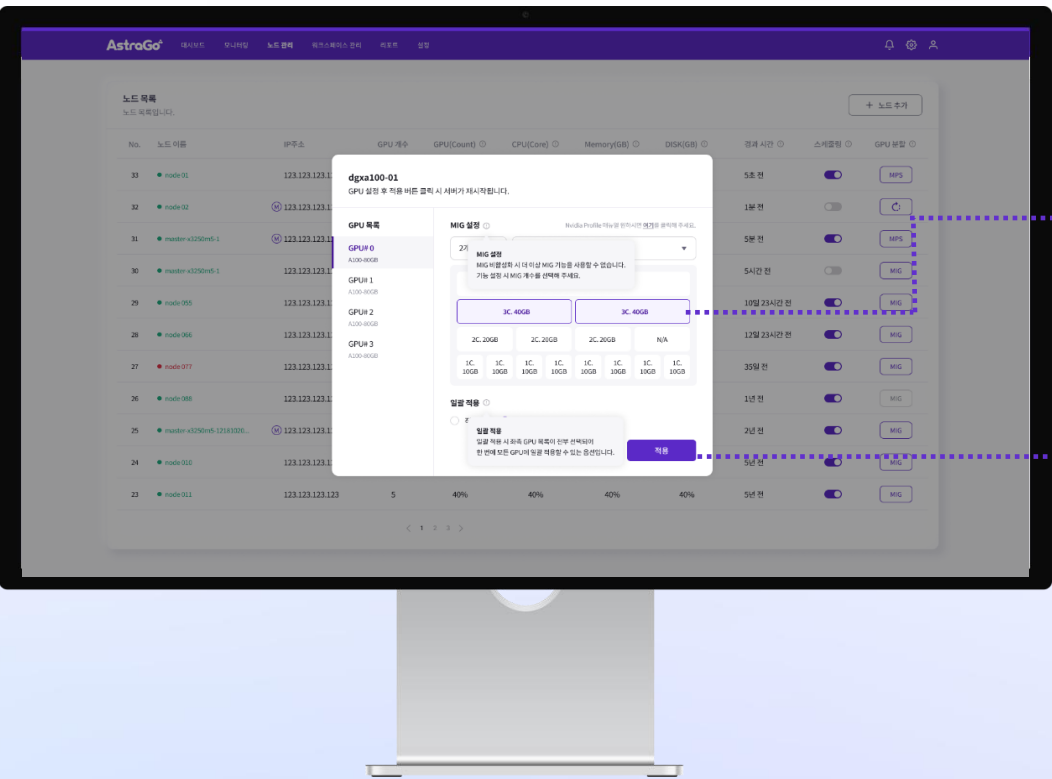
낭비없는 GPU 활용, ROI를 극대화하다.

자원 최적화·GPU분할(MPS/MIG)·리소스 쿼터관리·Job 스케줄러로 GPU 효율성을 극대화합니다.

GPU 분할(Multi-Process Service/Multi-Instance GPU)



- MPS
- MPS 설정 개수 선택 가능
 - 분할된 GPU 메모리 정보 제공



- MIG 활성화 개수 설정
- 관리자 권한으로 MIG 활성화 및 비활성화 개수 설정

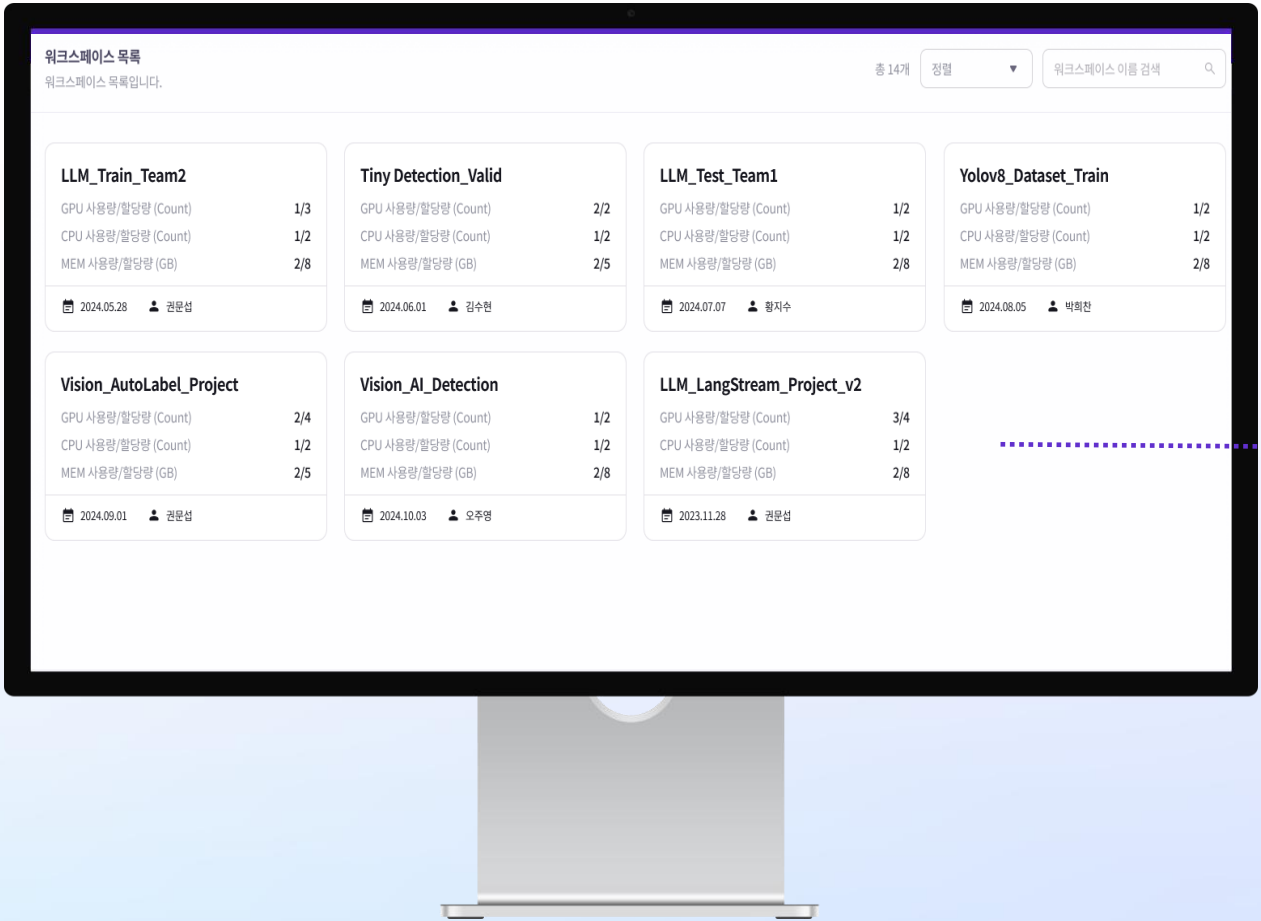
- GPU 자원 설정
- GPU 분할 개수 및 크기 설정
 - 좌측 GPU에 설정 사항 일괄 적용 가능

단일 GPU를 다수 사용자/Job에 할당하여 자원을 효율적으로 사용할 수 있습니다.

낭비없는 GPU 활용, ROI를 극대화하다.

자원 최적화·GPU분할(MPS/MIG)·리소스 쿼터관리·Job 스케줄러로 GPU 효율성을 극대화합니다.

워크스페이스 리소스 쿼터

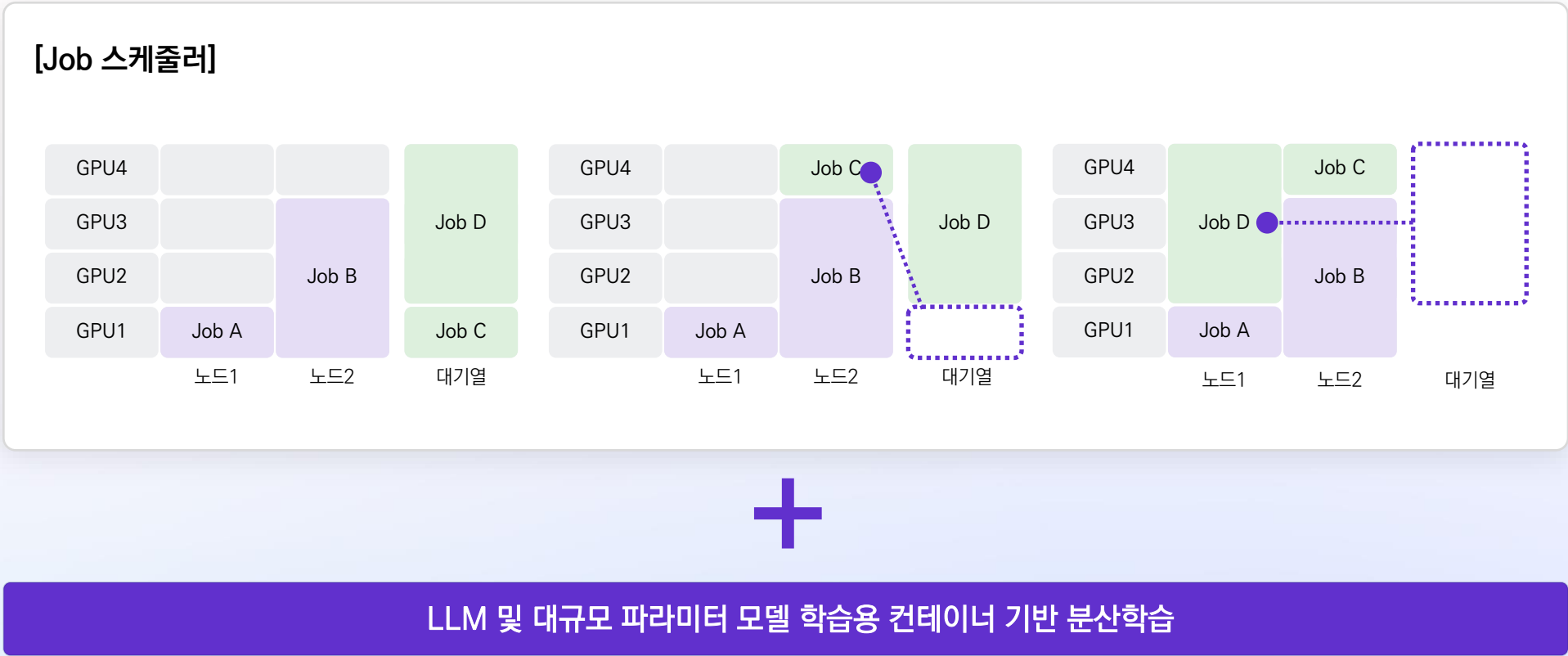


사용자/팀별 자원 사용 한도 설정으로 공정한 자원 분배와 안정성을 확보합니다.

낭비없는 GPU 활용, ROI를 극대화하다.

자원 최적화·GPU분할(MPS/MIG)·리소스 쿼터관리·Job 스케줄러로 GPU 효율성을 극대화합니다.

Job 스케줄러

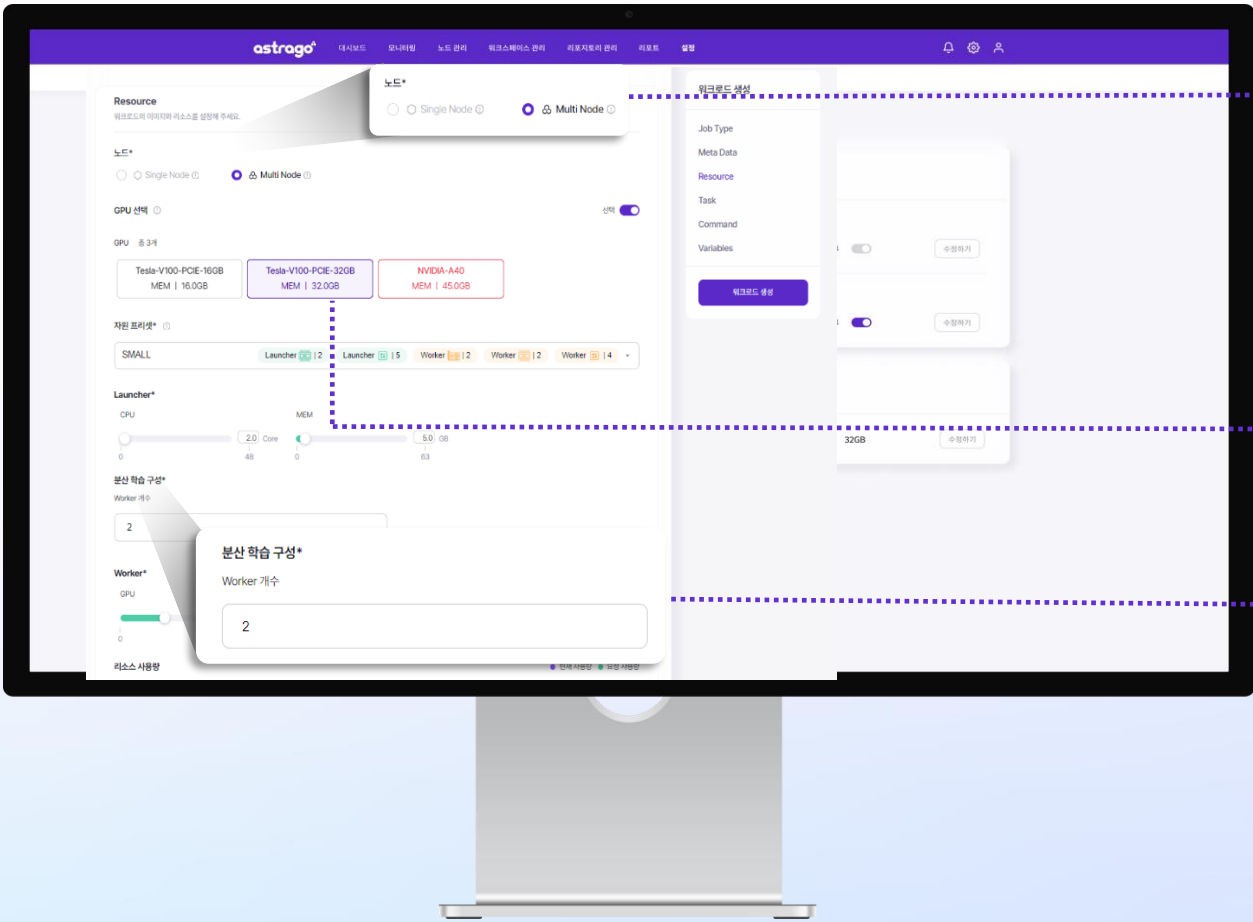


Job 배정 시 가용 자원이 적은 노드부터 배정하여 자원 활용률을 높입니다.

개발자는 개발에만 집중하세요. AstraGo가 나머지를 해결합니다.

분산학습 · 모델 허브로 연구개발 속도를 가속화합니다.

분산학습



분산 학습 시 GPU 선택

- 분산 학습 시, Multi-Node 선택
- 사용 희망하는 GPU 종류 선택 가능

자원 프리셋 선택

- 관리자가 설정한 Small / Medium / Large 중 선택 가능

분산 학습 워커 개수 설정

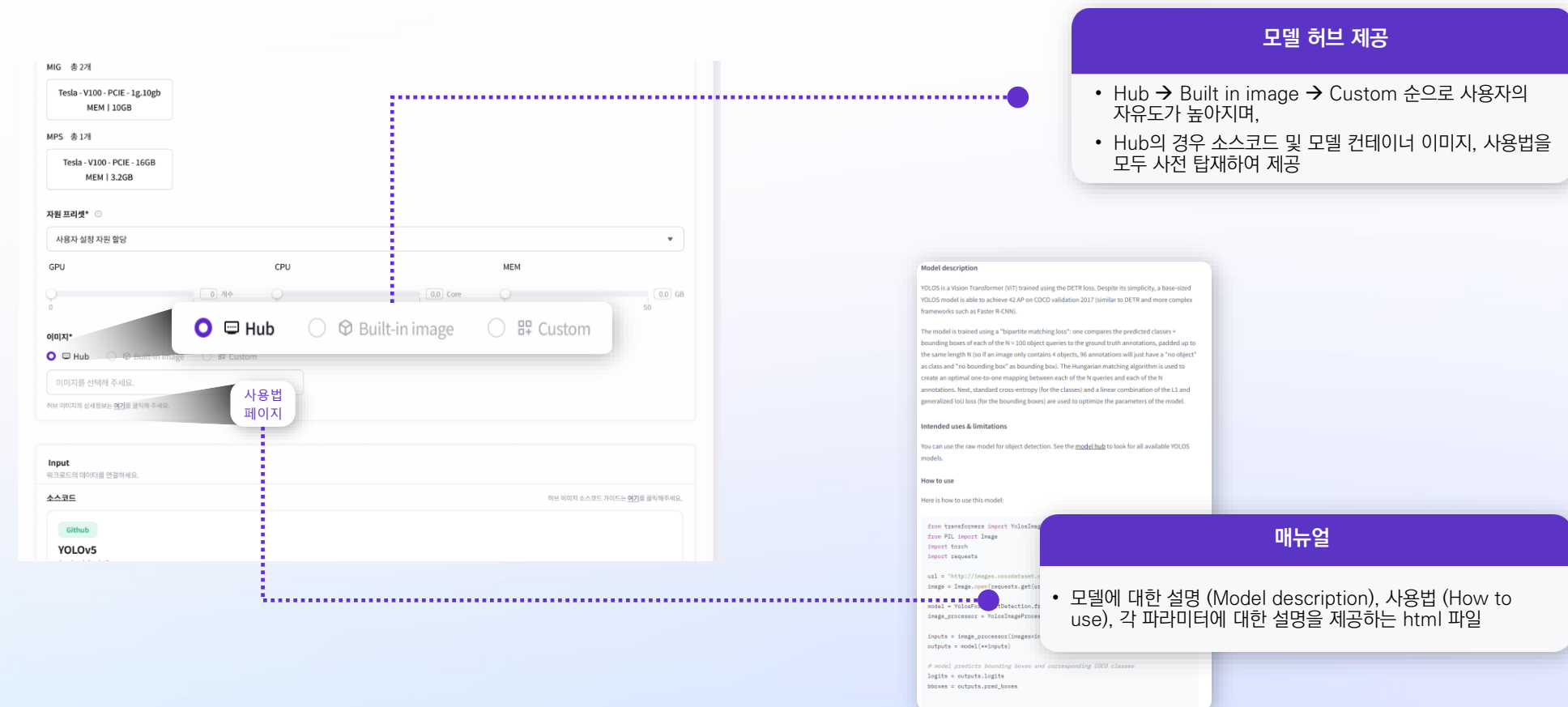
- 사용자가 분산 학습 진행할 워커 개수 기입

노드 학습의 분산학습 기능과 자동 명령어 생성 기능으로 세팅 부담을 최소화하여 학습 환경 준비시간을 단축할 수 있습니다.

개발자는 개발에만 집중하세요. AstraGo가 나머지를 해결합니다.

분산학습 · 모델 허브로 연구개발 속도를 가속화합니다.

모델 허브 & 빌트인 이미지

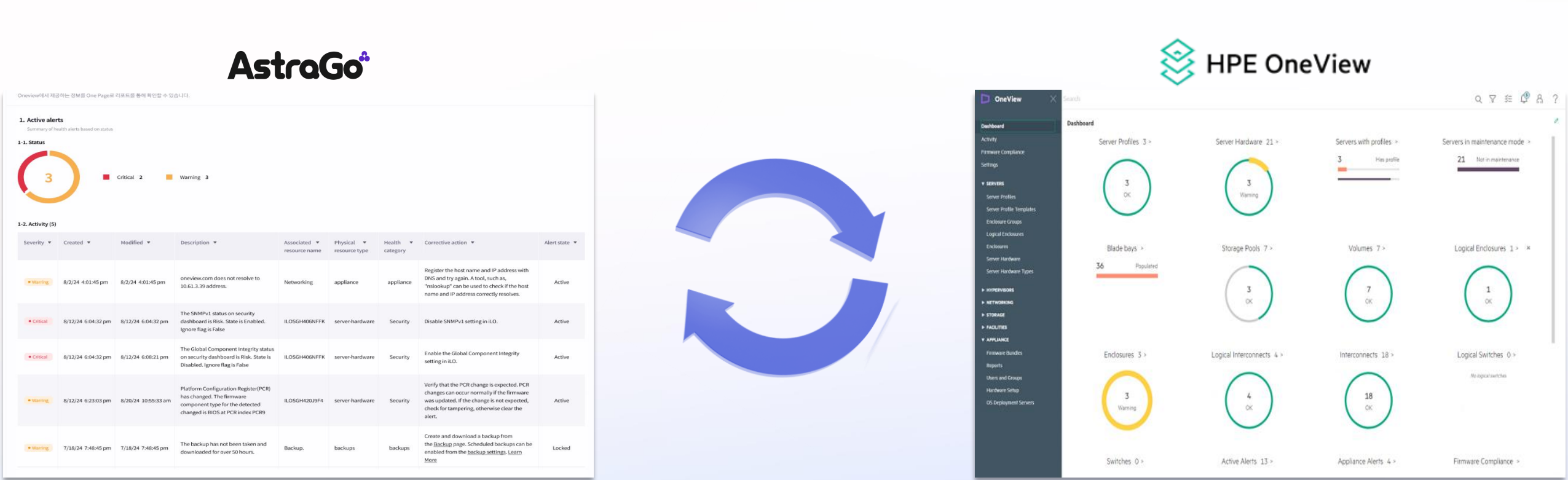


즉시 활용 가능한 모델과 환경 제공으로 복잡한 설치·세팅 없이 바로 학습 시작이 가능합니다.

늘어나는 AI 워크로드에도 흔들림 없는 확장성.

OneView 연동· Job 타입 선택지원·End-to-End 오케스트레이션으로 안정적 확장을 보장합니다.

OneView 연동

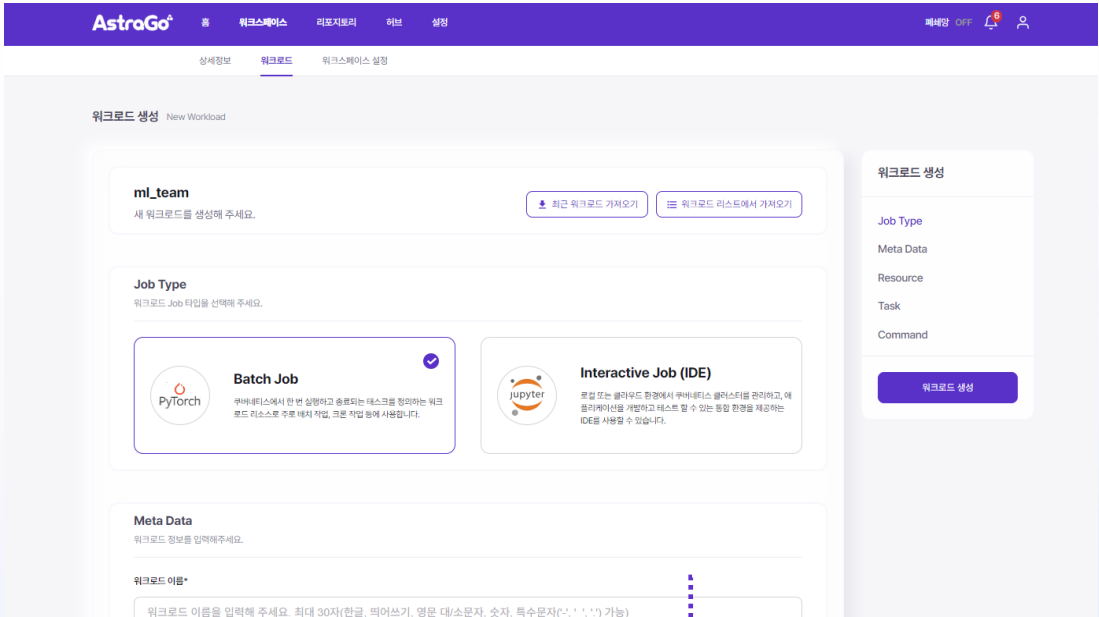


OneView와 연동하여 소프트웨어부터 하드웨어까지 단일 플랫폼에서 통합 관리가 가능합니다.
RedFish 기반의 모니터링 기능을 통해 HPE, Lenovo, NVIDIA 등 다양한 하드웨어 제품의 지표를 폭넓게 모니터링하고,
사용자는 GPU 클러스터 운영을 효율적이고 일원화된 방식으로 관리할 수 있습니다.

늘어나는 AI 워크로드에도 흔들림 없는 확장성.

OneView 연동· Job 타입 선택지원·End-to-End 오케스트레이션으로 안정적 확장을 보장합니다.

Job 타입 선택 지원



Batch Job



Interactive Job

워크로드 Job 타입 선택

- Batch Job : 일괄처리 및 반복 작업 시 사용
- IDE Job : 사용자와의 즉각적인 개입 및 피드백 필요

다양한 워크로드(개발·학습) 실행 지원을 통해 프로젝트 요구에 맞춘 유연한 운영이 가능합니다.

AstraGo 2.0 서버를 원활하게 구동하기 위한 Master Node 최소 사양

Master Node	
CPU 3.2 GHz, 8 Core 이상	Memory 128GB 이상
HDD 600GB 이상	NIC 10G 이상(네트워크)

Worker Node	
CPU 8 Core CPU 이상	Memory 64GB 이상
HDD 600GB 이상	NIC 1G 이상(네트워크)
GPU RTX 급 GPU 이상 (NVIDIA)	

*AstraGo 서버를 위한 최소 사양으로, Master Node는 3대 이상 권장합니다(Kubernetes 자체 권장 사항).
*Worker Node는 Geforce 계열의 GPU도 지원가능하나, 권장드리지 않습니다.

Ⅲ. AstraGo 도입 여정

GPU 클러스터 전 주기, AstraGo와 씨이랩이 책임집니다.

설계 → 배포 → 검증 → 운영지원 → 교육·인계까지 End-to-End 지원 프로세스를 제공합니다.



IV. 도입 사례

불필요하게 무거운 솔루션,
반복되는 세팅 공수,
HW/SW 단절

복잡한 세팅과 운영 부담을 단순화한
경량화 오케스트레이션 제공



Pain Point

- 컨테이너 환경을 준비하기 위해 **개발 및 사전 세팅에 많은 시간이 소요**되어 프로젝트 착수가 지연됨
- **비싸고 불필요한 기능이 많은 무거운 외산 솔루션**으로 인해 구매 및 운영 부담 발생
- 하드웨어와 소프트웨어를 아우르는 **통합 모니터링 체계 부재**로 운영 관리의 복잡성 증가.

Solution

- **빌트인 이미지 & 리포지토리 제공** → 개발자들이 별도 코드 없이 손쉽게 컨테이너 생성
- **필수 기능 중심의 경량화된 패키지** → 학습·추론에 필요한 기능만 제공, 경제적 비용 구조 확보
- **HPE OneView 통합 모니터링** → AstraGo 내에서 HW/SW 자원 상태를 일원화해 관리 효율화

Outcome

- 컨테이너 환경 준비 시간이 **약 40% 단축**, 개발 생산성 향상
- GPU 활용률이 **30% 이상 증가**, Idle 자원 감소로 ROI 개선
- 외산 대비 **총소유비용(TOC) 약 25% 절감**, 구매·운영 부담 완화
- HW/SW 통합 모니터링 도입으로 관리 업무 복잡도가 현저히 줄어듦

Pain Point

- GPU 활용률 저조, 일부 자원 Idle 상태 → ROI 저하
- 여러 프로젝트·연구자가 **동시에** 자원을 쓰며 **충돌** 및 **우선순위 혼선** 발생
- GPU/메모리/네트워크 모니터링 체계 부재
- GPU 인프라 비용 집행의 예측 불가능성

Solution

- **멀티 프로젝트 자원 격리/예약**: 연구자별 GPU 예약 관리, 충돌 방지 및 공정한 분배
- **프로젝트 이력 관리 기능**: AI 학습·추론 작업 기록을 체계적으로 관리, 연구 재현성 확보
- **Web IDE & 빌트인 환경 제공**: 프레임워크/라이브러리 사전 구성 → 연구자는 즉시 분석 착수 가능
- **경제적 GPUaaS 모델**: 불필요한 기능 제거, 연구 목적 중심의 경량화 기능만 제공 → 예산 효율화

Outcome

- GPU 활용률 30% ↑, Idle 최소화
- **멀티 사용자 충돌 0건**, 자원 배분 공정성 확보
- **프로젝트 준비 시간 40% 단축**, 연구 생산성 향상
- 운영 관리 업무 50% ↓, 연구원은 데이터 분석에 집중
- **TOC 비용 약 20% 절감**, 예산 집행 효율성 증대

GPU 자원 비효율, 멀티 사용자 충돌,
환경 세팅 지연, 운영 가시성 부족,
비용집행 불투명

다수 연구자의 병행 작업과
데이터 분석 수요를 위한
최적의 GPU 오케스트레이션

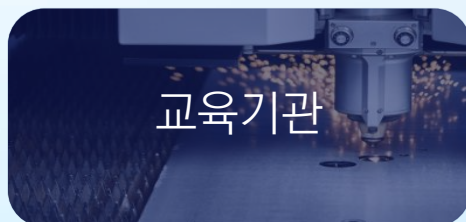
공공 연구기관

멀티 프로젝트 관리

예산 효율화

GPU 자원 부족과 자원배분 비효율,
환경설정·재현성 문제,
다수 과목·lab 동시 사용충돌

AI 교육과 초거대 모델 연구를 동시에
지원하는 대학 맞춤형 GPU 인프라



교육기관

GPU
하드웨어 구축

표준화된 학습 환경

Pain Point

- **공용 자원 배분 혼선**: 수업·연구 동시 운영으로 GPU 과점유·idle 발생
- **과목·랩마다 다른 프레임워크 버전, 결과 비교 어려움**
- **온보딩·지원 부담**: 다수 학생이 개별 환경을 세팅, 조교·관리자 과부하

Solution

- **GPU 클러스터 구축 & AstraGo 도입**: RTX A6000 서버 기반, 자원 자동 스케줄링·쿼터 관리
- **표준화된 개발 환경 제공**: 빌트인 이미지·Web IDE로 즉시 사용 가능, 수업·연구 일관성 확보
- **운영 가시성 강화 & 유지보수 지원**: 실시간 모니터링, 장애 대응, 중앙 거버넌스 체계 마련

Outcome

- GPU 활용률 40% ↑: 공용 자원 낭비 감소, 수업·연구 동시 운영 효율화
- 환경 세팅 시간 50% ↓: **표준 환경 기반으로 빠른 실습·연구 착수**
- **과제·실습 평가 일관성 확보**: 동일 환경 기반 성능 비교·재현 가능
- 운영 지원 부담 60% ↓: **조교·관리자 반복 업무 최소화, 학생은 학습에 집중**
- **초거대 모델 학습 가능**: GPU 클러스터 기반으로 대학 연구 역량 강화

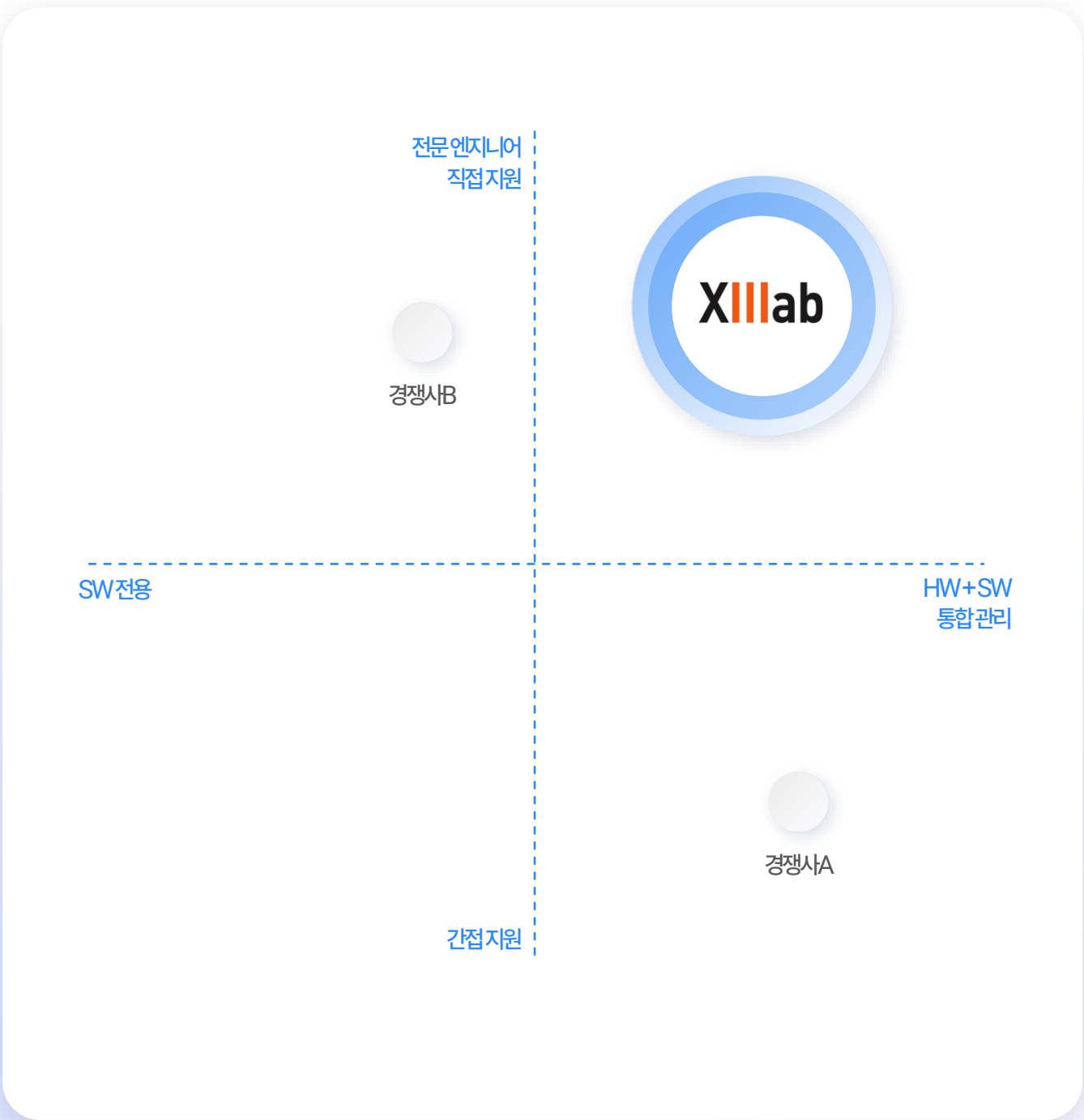
V . Professional Service

최신 NVIDIA SW 스택과 전문 엔지니어, 씨이랩이 직접 딜리버리합니다

4가지 핵심 차별화 요소

GPU 인프라의 설계부터 설치, 운영·유지보수까지 HW와 SW를 아우르는 원스톱 서비스를 제공합니다.

NVIDIA SW 파트너십 기반 최신 스택 즉시 제공	- NVIDIA AI Enterprise, CUDA, Container Toolkit 등 최신 SW 스택을 가장 빠르게 반영 - 고객은 버전 충돌·업데이트 지연 없이 곧바로 최신 환경에서 개발 가능
HW+SW 통합 유지보수 역량	- 서버·GPU·스토리지부터 Kubernetes까지 **단일 창구(One-stop)**에서 관리 - HW/SW 분리 지원의 한계를 넘어, 장애 시 원인 구분 없이 즉시 대응
전문 엔지니어 직접 지원	- 단순 설치 인력이 아닌, GPU 오케스트레이션·AI 워크로드 최적화 경험 보유 전문가 투입 - 인프라 설계, 자원 정책 수립, 보안 검증, 운영 자동화까지 End-to-End 컨설팅 제공
검증된 구축·운영 실적	- 대형 연구기관, 제조사, 대학 등 실제 GPU 클러스터 구축 레퍼런스 다수 - 현장 맞춤형 커스터마이징 경험으로, 해외 솔루션 대비 국내 최적화 보장



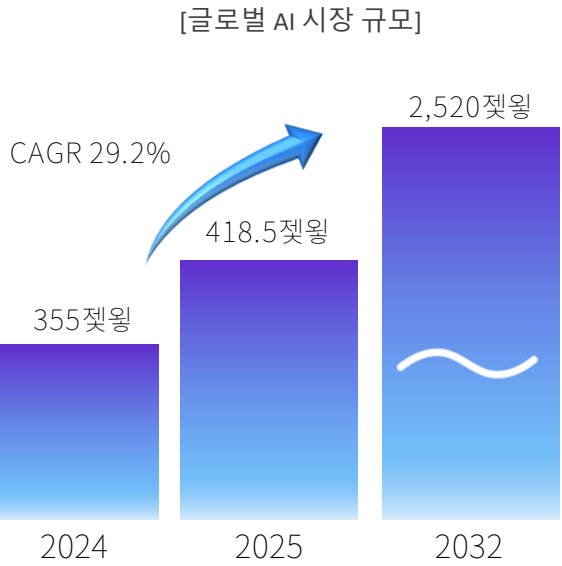
VI. Why Now?

GPU 인프라의 변곡점, 지금은 ‘효율’이 경쟁력입니다.

AI시장은 폭발적으로 성장하고 있지만, GPU 자원 수요는 공급과 예산을 압도하고 있습니다.
CapEx와 OpEx의 동반 상승은 이제 기업의 ROI를 직접 압박하고 있습니다.
지금 필요한 건 확장이 아닌, **효율 중심의 인프라 전환**입니다.

AI 수요 폭증

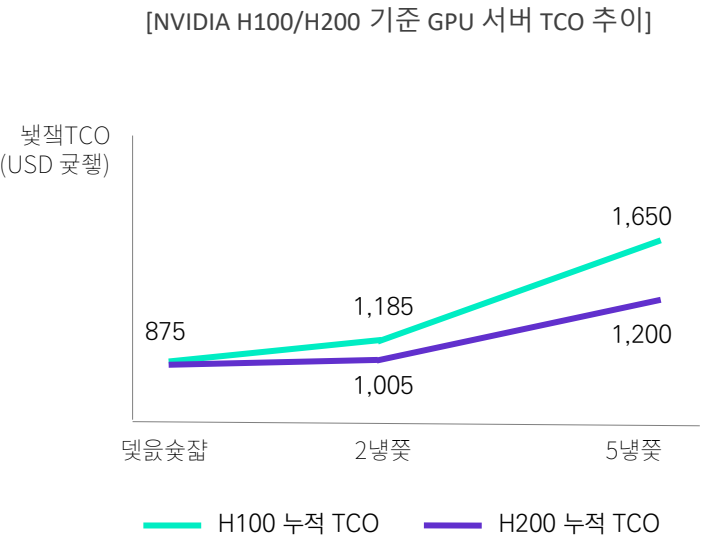
생성형 AI·퍼지컬AI의 확산으로
GPU 리소스 사용량 급등



* 출처: [Fortune Business Insight\(2025.09.29\)](#)

GPU TCO 급등

서버 단가·전력비·운영비 상승으로 인한
TCO(총소유비용) 부담 확대



* 출처: [Jarvis Labs](#), [TFG Datacenters](#), [Uvation](#), [Vast.ai](#), 씨이랩 자체 TCO 추정모델



해법은 ‘활용률 기반 TCO’

AstraGo는 GPU 클러스터의 모든 자원을 하나의 통합 관제 체계로 자동 조율합니다.
사용자는 모델 개발에만 집중하고, 시스템은 효율·안정성을 스스로 유지합니다.



AI 인프라의 새로운 기준, 지금 **AstroGo[®]**로 시작하십시오

씨이랩은 NVIDIA 파트너로서, GPU 인프라의 설계부터 운영까지 효율·안정·ROI 중심의 AI 인프라 혁신을 실현합니다.

➤ 문의 및 데모 요청: 씨이랩(Xillab) Sales팀 유진수 매니저(js.yoo@xillab.com)