우송대학교 소프트웨어학부 GPU서버 도입과 효율적 활용

㈜스타셀
www.starcell.co.kr

Starcell

Deepcell

Agenda

- 스타셀 소개
- 도입현황
- Kubeflow
- DeepcellKube User Manager
- GPU 분할
- 활용방안
- 시스템 확장
- 데모

스타셀 소개

회사 개요

2004년 IT 전문 기업으로 출발한 ㈜스타셀을 기술 발전에 발맞추어 다양한 분야의 소프트웨어 솔루션을 개발·공급해 왔으며, 현재는 인공지능(AI) 관련 솔루션 개발과 서비스를 핵심 사업 영역으로 삼아 전문성을 강화하고 있습니다.

주식회사 스타셀

2004년 5월 설립

S/W 솔루션 개발 회사로 시작하여 현재 AI 분야 사업 수행

주소 : 서울시 강서구 마곡중앙6로 11 보타닉파크타워3 415호

대표 : 박노헌

웹페이지 : www.starcell.co.kr

연락처 : npark@starcell.co.kr, ☎02-540-7853, (fax)02-540-7761

스타셀 소개

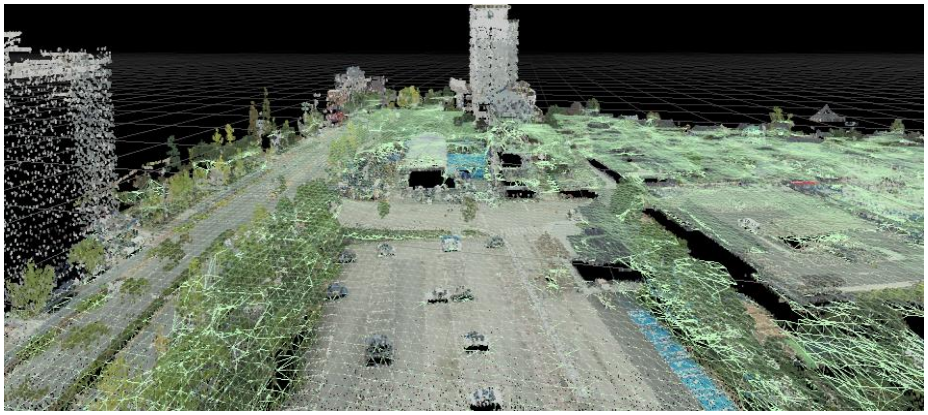
주요 연혁

- 2004 : IT 전문가들에 의해 기업용 솔루션 개발업체로 창립
- 2004 ~ : 자체 솔루션 ITMon과 JPA 국내 500여 기관에 공급
- 2015 : 엔비디아 솔루션 파트너 -> AI 전문 기업으로 전환 시작
- 2019 : 인텔 기술 파트너 -> AI 교육솔루션 공동 개발
- 2021 : AI 학습용 데이터 구축사업 컨설팅 : 한국도시3차원영상데이터
- 2022 : 경기도 기술개발과제 개발 : 건축물 외관 진단용 딥러닝 AI 시스템
- 2023 : AI 학습용 데이터 구축사업 : 중노년층 방언 음성인식 모델링
- 2024 : DeepcellKube AI Toolkit 솔루션 개발

스타셀 소개

스타셀 AI 관련 사업 수행 사례

2021	• AI 학습용 데이터 구축사업 컨설팅 : 한국도시3차원영상데이터(타임게이트컨소시엄)
2022	• 경기도 기술개발과제 AI 분석시스템 개발 : 건축물 외관 진단용 딥러닝 AI시스템(레인보우테크)



스타셀 소개

스타셀 AI 관련 사업 수행 사례

2022	• AI 학습용 데이터 구축사업 : 중노년층한국어방언데이터(음성인식모델링)

개요

GPU Server 운영 인프라가 필요한 이유

적절한 GPU Server 운영 인프라가 없이는 고가의 GPU 서버를 도입해도 유휴 시간이 많고 제대로 활용하기 어려움

개요

GPU 서버 활용을 위해 필요한 방안

공용 GPU 서버 도입 : 특정 개인이나 조직이 전용으로 도입하는 경우 유휴 시간에 다른 부서에서 사용하기 어려우므로 사용율이 낮을 가능성이 높음

편리한 접근성 필요 : 원격지(서버실, 데이터 센터)에 설치된 GPU 서버를 사용자들이 쉽게 접근할 수 있는 사용자 인터페이스가 제공되어야 여러 사용자들이 편리하게 이용할 수 있음

가상화 환경 사용 : 사용자 별로 가상화 시스템을 이용하여 사용 환경을 독립적으로 구성해야 상호 간에 간섭 없이 사용할 수 있음

구체적 활용 계획의 필요 : GPU Server 운영 인프라의 특성에 맞는 적절한 활용 전략과 계획이 있어야 시행착오를 최소화 하며 운영 효율을 높일 수 있음

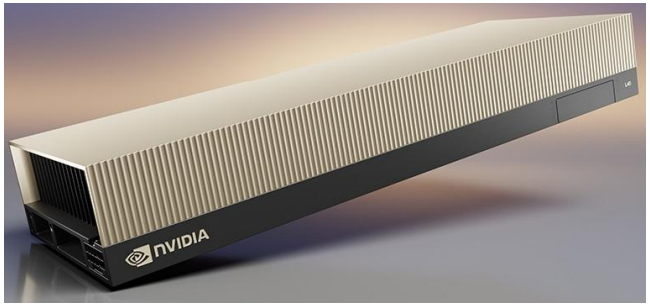
도입 현황

하드웨어 시스템

Single Node
GPU Server

Mater/Worker node(GPU node)

1. Gigabyte GPU Server G493-SB1-AAP1
2. CPU : XEON(R) GOLD 6530
3. CPU cores : 2 Socker, 32C/64T -> 128 cores
4. Memory : 512 GB
5. Storage : SSD 1TB(RAID1, OS), SSD 7TB(data)



GPU : 4개 장착 됨

1. NVIDIA L40S
2. GPU 아키텍처 : NVIDIA Ada Lovelace Architecture
3. CUDA 병렬 처리 코어 : 18,176
4. 단정도 성능(FP32) : 91.6테라플롭스
5. GPU 메모리 : ECC 포함 48GB GDDR6
6. 전력 소비량 : 최대 350W(구성 가능)

도입 현황

GPU Server 운영 인프라

Kubeflow

Kubernetes 기반의 AI Platform

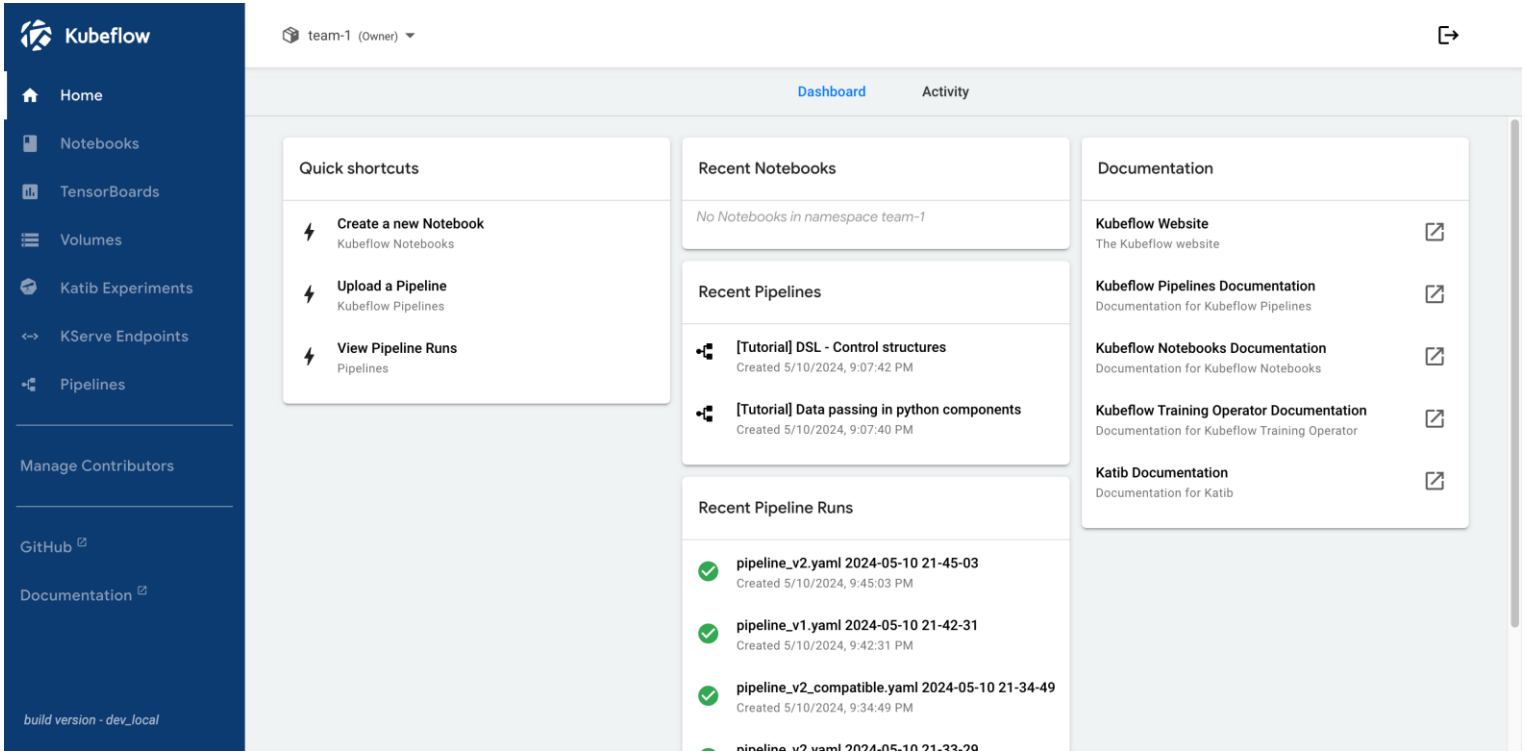
- **Kubernetes** : 멀티 노드 클러스터 인프라
- **Kubernetes 기반의 Kubeflow** : AI 연구 개발을 위한 사용자들에게 Container(가상 시스템)를 제공
- **AI Platform** : AI, ML, DL를 위한 개발환경부터 서비스까지 전체 워크플로를 관리하고 운영할 수 있는 통합 솔루션을 제공

Kubeflow

Kubeflow의 주요 프로젝트

Central Dashboard

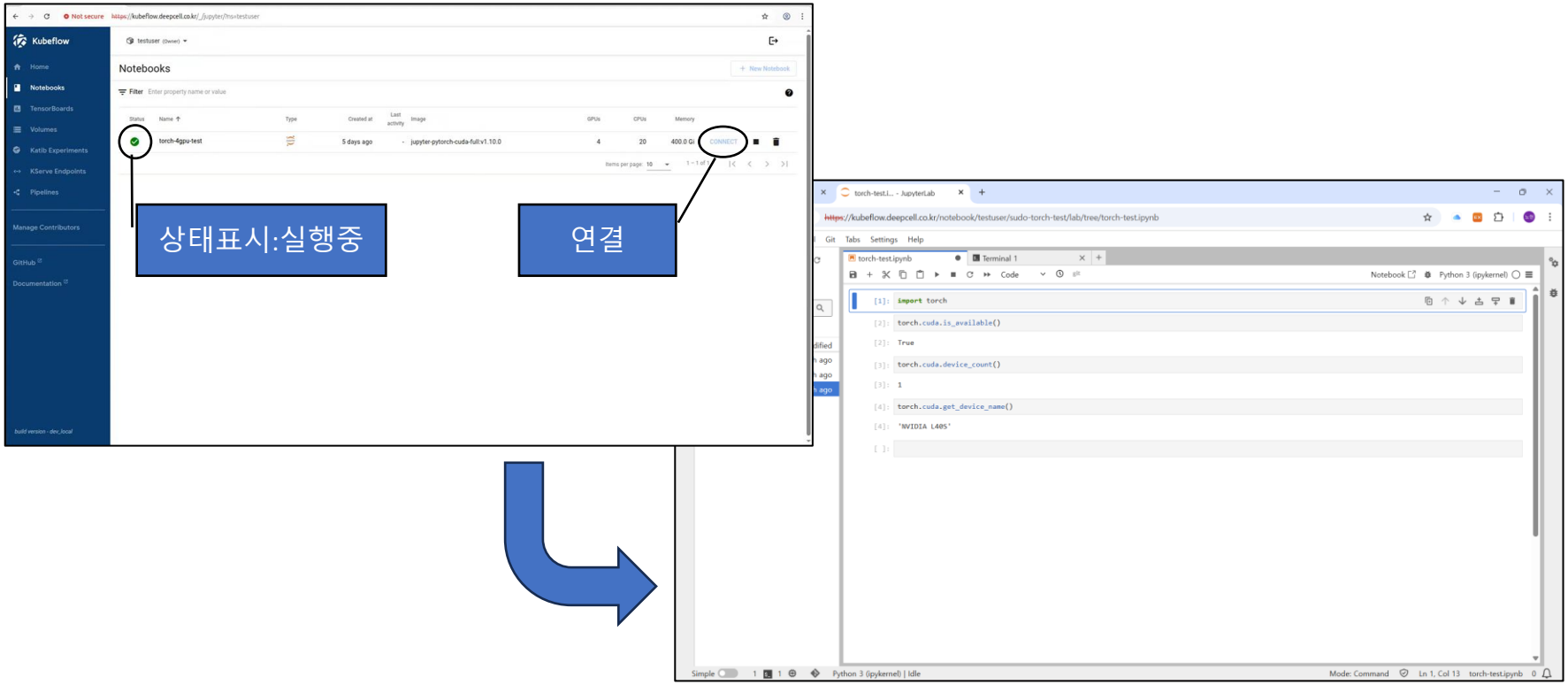
- Kubeflow 사용을 위한 기본 Web UI
- 사용자 인증 후 사용 가능
- Kubeflow 컴포넌트 UI 제공



Kubeflow

Notebooks

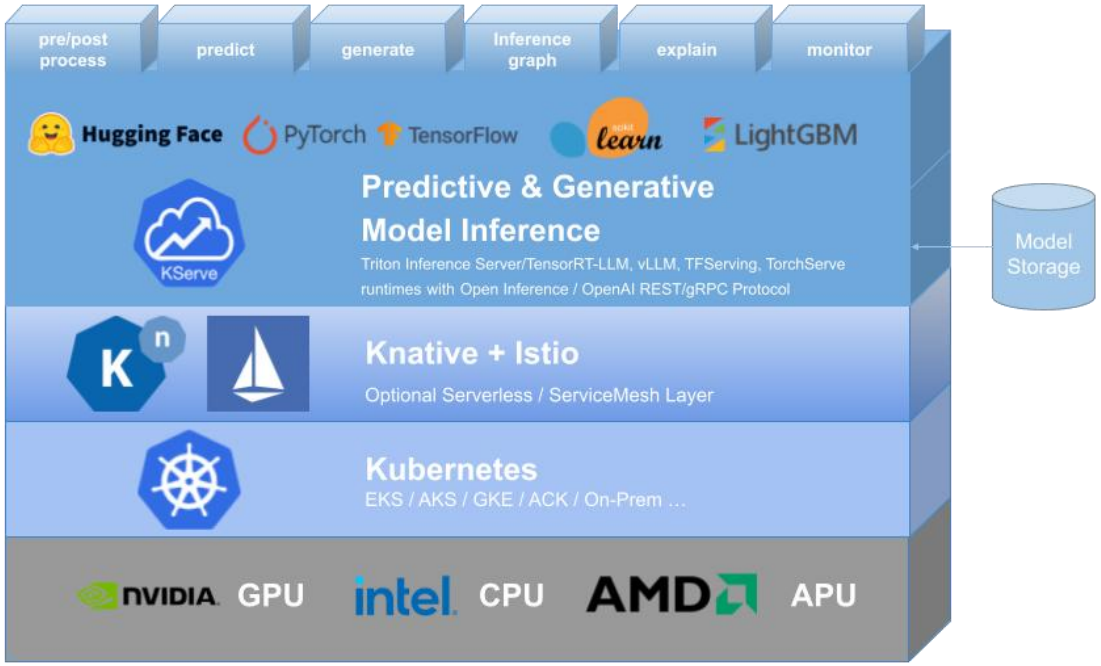
- 웹기반 개발 환경
- JupyterLab, RStudio, and Visual Studio Code (code-server) 지원
- 개발 환경을 클러스터에 컨테이너로 생성



Kubeflow

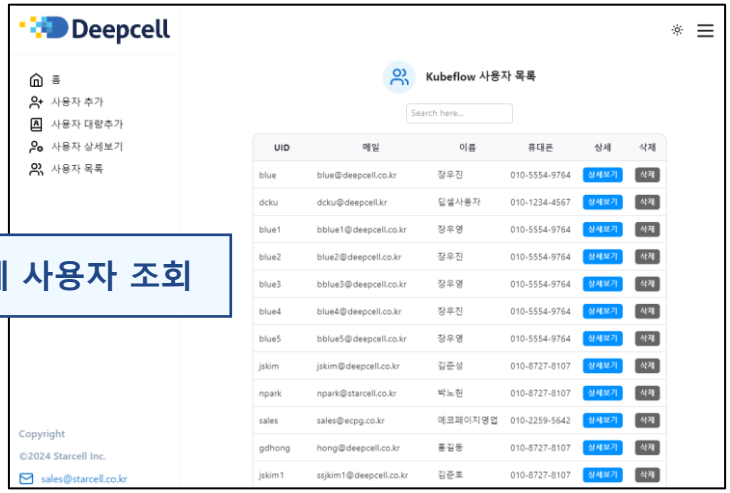
KServe

- AI Inference Platform
- 일반 딥러닝, 머신러닝 Inference Service 구축 지원
- 생성형 AI, LLM Inference Service 구축 지원
- 다양한 AI 프레임워크 지원

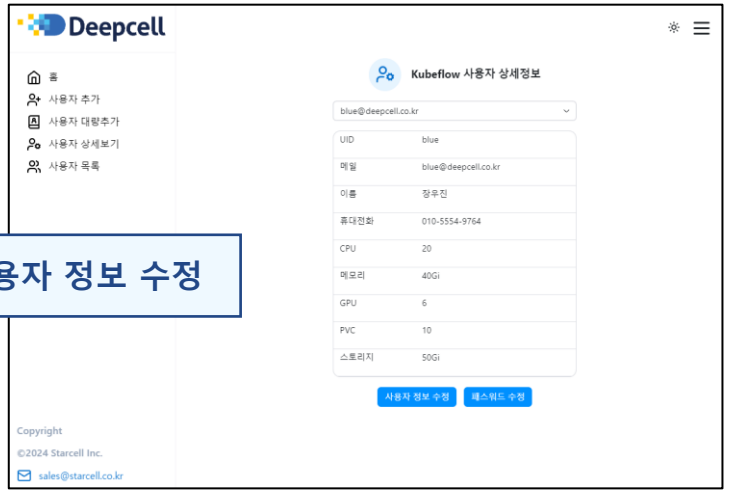


DeepcellKube User Manager

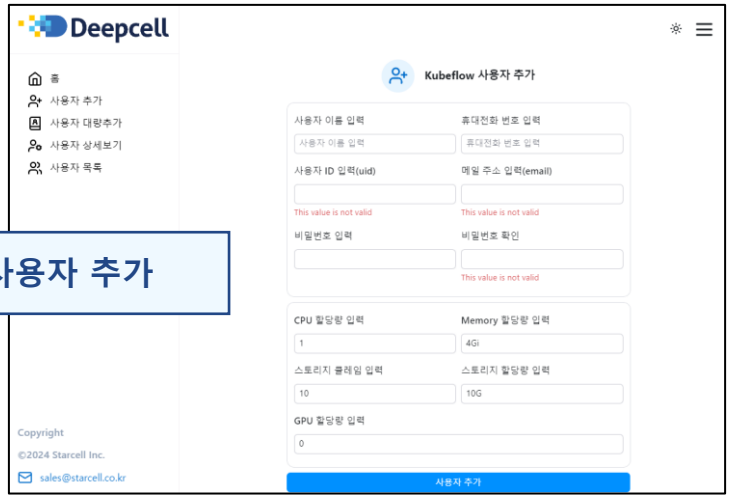
Kubeflow를 위한 사용자 관리 솔루션



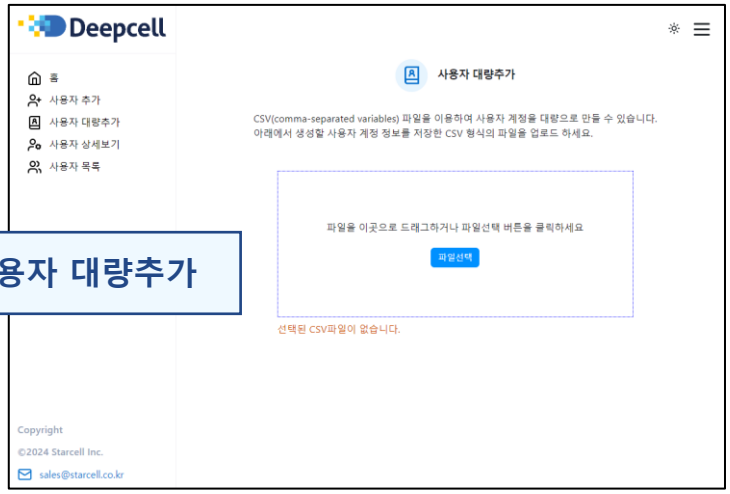
전체 사용자 조회



사용자 정보 수정



사용자 추가



사용자 대량추가

GPU 분할 사용

NVIDIA GPU의 분할 사용 방법

NVIDIA GPU는 하나의 GPU를 그대로 사용하는 방법과 물리적으로 분할해서 사용하는 방법 그리고 논리적으로 공유해서 사용하는 방법들을 제공

분할방법	방법 설명	제공 시스템	장점	단점
MIG(Multi GPU Instance)	GPU에서 제공하는 기능으로 GPU 자원(GPU, GPU Memory)을 물리적으로 분할하여 사용	A100, H100, H200, Pro 6000 등 최신 아키텍처 또는 Datacenter용 GPU에서 지원됨	물리적으로 분할 되어서 분할 영역이 고립됨, 따라서 분할된 영역이 영향을 주고 받지 않으므로 안정적으로 사용 가능	특정 GPU에서만 사용 가능 정해진 규격(프로파일)으로만 분할 가능 최대 7개로 분할 가능(GPU 모델에 따라 분할 가능 인스턴스 수가 다름) 미리 분할해 놓고 사용해야 함
MPS(Multi Process Service)	하나의 GPU를 여러 개의 프로그램이 사용할 수 있도록 소프트웨어적으로 제어하는 방법	Volta 아키텍처 이후의 모든 NVIDIA GPU에서 지원됨	특별한 하드웨어 요구사항 없이 소프트웨어적인 설정으로 GPU를 분할하여 사용 컨트롤 프로세스에 의해 GPU 오류를 감지하여 오류의 확산을 방어함	GPU를 물리적으로 완전히 고립시키지 못함
Time-Slicing	하나의 GPU를 여러 개의 프로그램이 시용할 수 있도록 시분할 방식을 제공	Pascal 아키텍처 이후의 모든 NVIDIA GPU에서 지원됨	가장 간단하게 설정하여 GPU를 분할 사용할 수 있는 방법	Context Switching이 빈번하게 발생하여 성능 저하가 큼 메모리 보호 기능이 가장 약하여 오류 발생 시 같은 GPU를 사용하는 다른 프로그램에 영향을 많이 미침

GPU Server 활용 방안

Notebooks를 이용한 개발 환경 활용

Jupyter Notebook을 이용

- 코딩 실습 교육
- 머신 러닝 실습 교육
- 딥러닝 실습 교육
- API 프로그램 교육

Terminal을 이용

- Linux OS 실습
- Shell Script 실습

VS Code 이용

- 기타 개발 실습

R-studio 이용

- 데이터 분석

GPU 분할 사용을 이용한 딥러닝 실습

MPS를 이용한 GPU 분할

- 4개의 GPU를 32(4 x 8 = 32)개로 분할 하여 딥러닝 실습 가능
- 필요에 따라 분할 설정하여 사용
- 하나의 노드에 있는 4개의 GPU는 동일한 분할 정책이 적용됨

GPU Server 활용 방안

Kserve를 이용한 LLM Service 구축 활용

LLM 서비스 구축

- qwen 등 경량 LLM 서비스 구축

OpenAI-Compatible APIs 지원

- 챗GPT와 같은 AI 서비스 개발 실습
- 챗봇 개발 실습

향 후 확장 방안

시스템 확장

향후 필요에 따라 시스템을 확장할 수 있으며 시스템 확장은 다음과 같은 범위에서 단계별로 확장 가능

- GPU 추가
- 노드 추가
- 스토리지 추가
- 네트워크 고도화

향 후 확장 방안

GPU 추가

현재 서버에 GPU 추가 장착

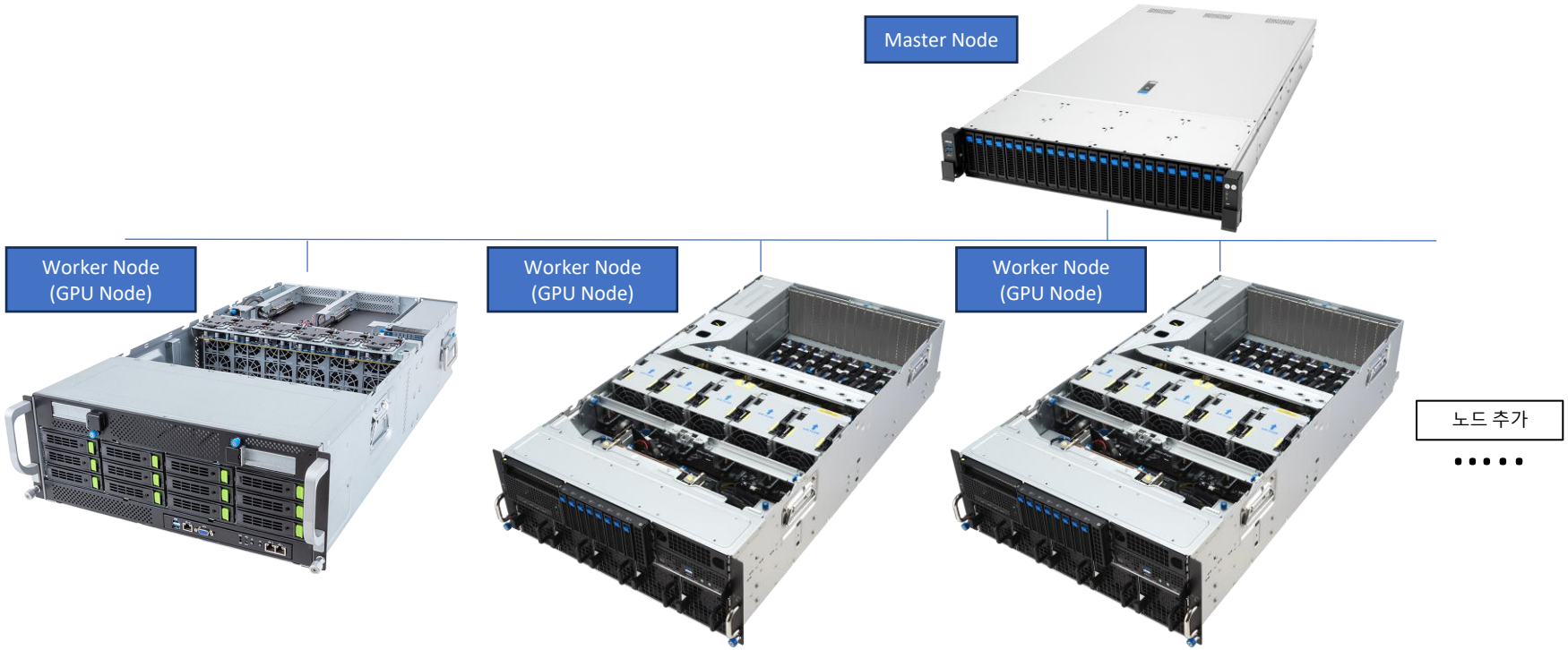
- 4개의 GPU 추가 가능
- 추가 가능 GPU
  - L40s
  - RTX 6000 Ada
  - H100 NVL(94GB)
  - RTX Pro 6000 Max-Q

향 후 확장 방안

Node 추가

Node를 추가하여 멀티 노드 환경 구축

- Master Node 분리가 먼저 필요
- 유연한 확장으로 Scale Out 가능



향 후 확장 방안

Storage 추가 및 네트워크 고도화

필요에 따라 스토리지를 추가하고 클러스터 네트워크를 고도화

- NAS 장비 추가
- 10GB 이상 네트워크 구성

DEMO

질의

Question