

규 격 / 사 양 서

1. 개요	품명	서버		
	수량	1	제조사	
2. 용도 및 특징	○ 용도 - 대규모 언어모델(LLM) 및 딥러닝 모델 학습·추론, 대용량 데이터 처리 등 인공지능 연구 수행을 위한 고성능 GPU 연산 서버임. - 멀티 GPU 병렬 연산 및 대용량 메모리 기반의 모델 학습·미세조정(fine-tuning) 실험 환경으로 활용하고자 함. ○ 특징 - 8개의 GPU 장착이 가능한 4U 랙마운트 새시에 96GB급 워크스테이션 GPU 2개를 기본 장착하여, 향후 GPU 추가 증설이 가능한 확장형 구조임. - 서버용 고성능 CPU 1개와 DDR5 ECC 메모리 672GB를 탑재하여 대용량 데이터 전처리 및 GPU 연산 지원에 적합함. - PCIe Gen5 기반 확장슬롯 및 PCIe 스위치를 통해 GPU 간 고속 데이터 전송을 지원하며, 이중화 전원공급장치로 안정적인 장시간 연산이 가능함. - 리눅스 서버 OS와 CUDA, 딥러닝 프레임워크, 컨테이너 환경이 설치·구축된 상태로 납품되어 즉시 연구에 활용 가능함.			
3. 기기구성 (세부내용, 수량 등)	- GPU 서버 본체 (4U 랙마운트, 8 GPU 확장형, 랙마운트 키트 포함) 1대 - 서버용 CPU (96코어/192스레드급) 1개 - 메모리 DDR5 ECC REG RDIMM 96GB 7개 (총 672GB) - OS용 SSD U.2 NVMe PCIe Gen5 1.92TB 1개 - GPU (96GB급 워크스테이션 GPU, Max-Q 300W급) 2개 - GPU 전원케이블 (12VHPWR용) 8개 - 전원공급장치 3000W Titanium 4개 (이중화, 총 12,000W) - 운영체제 및 AI 소프트웨어 환경 설치·구축 (OS, CUDA, Docker, 딥러닝 프레임워크, Jupyter, MPI 등) 1식			
4. 규격 및 사양	※ 본 규격은 최소 요구 성능을 정한 것이며, 동등 이상의 성능 및 품질을 갖춘 제품의 납품을 허용함. ○ 서버 본체 - 폼팩터: 4U 랙마운트, 1010×448×176mm(D×W×H) 동등, 랙마운트 키트 포함, 순중량 40kg 내외 - 메인보드: 듀얼 소켓 서버 보드, DIMM 슬롯 48개, DDR5 ECC REG RDIMM 최대 6TB 지원 - 스토리지 베이: 2.5" 핫스왑 베이 12개, M.2 2280/22110 슬롯 2개 - 스토리지 인터페이스: U.2 PCIe Gen5 x4 / SATA 포트 12개, M.2 PCIe Gen3 x4 2개 - 확장슬롯: PCIe Gen5 x16 더블폭 슬롯 8개, PCIe Gen5 x16 FHHL 슬롯 3개, PCIe Gen5 x16 LP 슬롯 3개 - PCIe 스위치: PCIe Gen5 스위치 칩셋 2개 - 네트워크: 1GbE RJ45 2포트(LAN1), 10GbE RJ45 2포트(LAN2) - 원격관리: IPMI 2.0 지원 BMC(10/100/1000Mbps 전용 관리포트) - 온보드 그래픽: 1920×1200@60Hz 이상 출력			

	- I/O 포트: D-sub VGA 1개, USB 3.2 2개, IPMI RJ45 1개, 1Gb LAN RJ45 2개 - 전원공급장치: 3000W 80PLUS Titanium 4개 이중화(Redundant), 최대 공급전력 12,000W ○ CPU - 서버용 x86 프로세서 1개: 96코어/192스레드 이상, 기본클럭 2.4GHz 이상, L3 캐시 384MB 이상, 프로세서당 12채널 메모리 지원, 최대 TDP 400W 지원 (AMD EPYC 9654 동등 이상) ○ 메모리 - DDR5 ECC REG RDIMM 96GB 5600MHz 7개 (총 672GB) ○ 스토리지 (OS용) - U.2 NVMe SSD PCIe Gen5 x4 1.92TB 1개: 순차읽기 12,000MB/s·랜덤읽기 1,700K IOPS 이상, 순차쓰기 3,500MB/s·랜덤쓰기 150K IOPS 이상, DWPD 1.0(5년), 데이터센터(엔터프라이즈)급 ○ GPU - 워크스테이션급 GPU 2개: GPU 메모리 96GB GDDR7 ECC 이상, CUDA 코어 24,000개 이상, PCIe Gen5 x16, 최대 소비전력 300W(Max-Q)급, 정품 리테일 박스 (NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition 동등 이상) - GPU 전원케이블 12VHPWR용 8개 ○ 소프트웨어 (설치 및 환경구축 포함) - 운영체제: Ubuntu 24.04 LTS Server - CUDA Toolkit, GPU 가속 라이브러리, 디버깅·최적화 도구, cuDNN, NCCL 설치 - Docker v29 이상 및 NVIDIA Container Toolkit(또는 Nvidia-docker) 설치, 사용자 Docker 이미지 구성 및 기술지원 - 딥러닝 프레임워크: PyTorch, TensorFlow, Keras, Theano, MXNet - Jupyter Notebook, JupyterLab 설치 및 환경구축 - 병렬처리 환경: Open MPI, MPICH, MPICH2, MVAPICH, MVAPICH2 및 AI 도구 커스터마이징 지원 - 소프트웨어 복구(recovery) 솔루션 제공
5. 설치, 교육 및 사후관리	- 장비 납품 시 GPU 포장 상태는 미개봉 상태로 봉인씰이 훼손되지 않아야 하며 검수 완료 후에 서버에 장착한다. - 업체는 사용자가 지정하는 장소에 설치하고 정상 작동을 확인하여 관리 담당자에게 안내해야 하며 필요한 교육이 있는 경우 실시한다. - 보증기간 내에 1~2회 교육 및 무상점검을 수행할 수 있어야 하며, 이때 발생하는 인건비 등은 업체 부담으로 한다.
6. 품질보증 기간 (Warranty)	3년 무상 방문하자보증
7. 납기일자	계약체결일로부터 60일 이내
8. 계약업체 제출서류	- 세부산출내역서 - 소프트웨어사업자 일반현황 관리확인서

9. 참가자격	- 소프트웨어사업자 신고를 마친 업체 - 서버 제조사(국내 공식 유통사)의 정품공급 및 기술지원 약속서를 제출할 수 있는 업체
10. 참가자격 제출서류	- 소프트웨어사업자 일반현황 관리확인서 - 서버 제조사(국내 공식 유통사)의 정품공급 및 기술지원 약속서
11. 기타사항	인공지능학과 / 김수경 교수님 / sookim@ewha.ac.kr