신설 [별지 제3호]

구매 규격서

COMMODITY DESCRIPTION

품목번호 Item No.	관세분류번호 HSK No.	정부물품분류번호(8자리) Korean Government Commodity Classification Code(eight-digit)	품 명 Description	단위 Unit	수량 Q'ty
1		43201401	그래픽또는비디오가속카드	EA	6

Ⅰ. 용도(End-user's Use)

본 장비는 본 과제의 핵심 목표인 초거대 AI 모델의 장기 기억 저장을 위한 벡터 DB 구축, 임베딩 모델 학습 및 대규모 데이터 기반 검색·추론 성능 고도화를 위한 고성능 GPU 가속 연산 장비임.

본 과제에서는 문서, 이미지, 로그, 사용자 질의, 대화 이력 등 다양한 데이터를 고차원 임베딩 벡터로 변환하고, 이를 기반으로 유사도 검색, 의미 기반 검색, RAG 검색 증강 및 사용자 질의응답 성능을 개선하고자 함. 이를 위해 대규모 임베딩 생성, 도메인 특화 임베딩 모델의 학습·파인튜닝·추론·평가 및 반복적인 벡터 검색 성능 개선 작업이 필요함.

기존 CPU 기반 연산 또는 저사양 GPU 환경에서는 대량의 텍스트·이미지·멀티모달 데이터 처리, 대규모 배치 추론, 임베딩 모델 학습 및 반복 실험에 과도한 시간이 소요됨. 따라서 본 장비는 벡터 DB 임베딩 모델 학습, 대규모 벡터 생성, Text RAG 및 MultiModal RAG 성능 평가, 검색·추론 파이프라인 최적화를 안정적으로 수행하기 위한 핵심 연구 장비로 활용하고자 함.

Ⅱ. 장비의 구성(Configurations of Goods)

1. 본체(Main body) AI 연산용 고성능 그래픽 또는 비디오 가속카드 6 EA
2. accessories
   - 서버 장착에 필요한 GPU 전원/연결 케이블 일체
   - 제품 장착 및 정상 구동에 필요한 기본 부속품 일체

Ⅲ. 성능 및 규격(Performance and Specification)

1. CUDA, CuDNN driver를 지원하고, GPU Driver 및 CUDA Toolkit 기반 개발·실행 환경 구성이 가능할 것
2. 180개 이상의 RT Core를 탑재하고 하드웨어 Ray Tracing 가속 기능을 지원할 것

3. 24,000개 이상의 CUDA Cores를 탑재할 것
4. GPU 메모리 96GB 이상, ECC 기능을 지원하는 GDDR7급 또는 동등 이상의 GPU 메모리를 탑재할 것
5. 메모리 대역폭 1,700 GB/s 이상 지원할 것
6. 워크스테이션 장착 인터페이스는 PCI Express 5.0 x16 이상을 지원할 것
8. GPU 1개당 최대 소비전력은 350W 이하일 것
9. AI 모델 학습·추론을 위한 행렬 연산 가속 기능을 지원할 것

Ⅳ. 기타 조건(Remarks)

1. 납기: 계약 후 발주 진행, 발주일로부터 60일 이내 납품 (단계/최종종료 2개월 전 납품 및 검수 완료)
2. 납품 장소: 서울특별시 관악구 관악로 1 서울대학교 942동 202호
3. 하자보증기간: 검수완료일로부터 1년 이상. 단, 제조사 또는 공급사의 기본 보증기간이 더 긴 경우 해당 보증기간을 적용함.
4. 규격담당자(연구원) 성함: 이호은 소속: 서울대학교 전기·정보공학부
   연락처: 010-2539-5107 이메일: hoeunlee@snu.ac.kr
5. 설치조건: 워크스테이션 또는 서버 내 장착 및 정상 인식 확인 포함, 장착에 필요한 GPU 전원/연결 케이블 및 기본 부속품을 포함하여야 함.
6. 시스템 호환성: 납품 장비는 PCI Express x16 이상 슬롯을 갖춘 서버 또는 워크스테이션 환경에서 정상 구동 가능하여야 하며, 장착 공간, 전원 공급 용량 및 냉각 조건을 충족하여야 함.
7. 드라이버 및 소프트웨어: NVIDIA GPU Driver, CUDA Toolkit 등 기본 GPU 연산 환경 구성 지원
8. 검수조건: 장비 장착 후 GPU 인식 여부, GPU 메모리 용량, 드라이버 정상 설치 여부 및 기본 연산 테스트 확인
9. 보증조건: 제조사 또는 공급사 기준 보증기간 적용
10. 유지보수: 초기 불량, 장착 오류, 드라이버 설치 및 정상 동작 여부 확인 지원
11. 기타: 연구용 워크스테이션 환경에서 안정적으로 구동 가능해야 하며, 딥러닝 프레임워크 기반 GPU 연산 실험에 활용 가능해야 함.