| | | |
|---|---|---|
| | FP16 **312** TFLOPS (Tensor Core) | |
| **가격(2025 기준)** | A100 40GB: $12,000~13,000<br>A100 80GB: $15,000~16,000 | T4: $1,200~1,400 |
| **주요 용도** | 대규모 모델 훈련·추론, 대용량 벡터 연산, HPC 워크로드 | 경량 추론(특히 FP16 mixed precision), 엣지 AI, 소형 배포 환경 |
| **장점** | • 대규모 파라미터 모델(수백억~수조) 훈련 가능 - 높은 메모리 대역폭으로 대규모 배치 처리 - MIG(Matrix Instance GPU)로 자원 분할 기능 지원 | • 전력 효율이 뛰어나며, 소규모 모델 추론 비용 대비 성능 우수 - 저전력 데이터센터나 클라우드 환경에 적합 |
| **단점** | • 초기 투자 비용이 매우 높음 - 유지보수 및 전력 소비가 큼 | • 메모리 제한(16GB)으로 대형 모델 구동 불가 - 연산 성능이 대형 모델 훈련/추론에 부족 |

A100은 메모리 용량과 대역폭이 훨씬 크고 연산 성능이 뛰어나므로 대규모 LLM 추론과 다중 처리에 유리하다.

대량의 벡터 연산과 복잡한 RAG 연산을 원활히 수행하려면 대용량 메모리와 높은 연산량이 필수적이므로, DO 솔루션 최소 사양으로 A100 이상의 GPU가 요구된다.

### 9. RAG란 무엇인가? 구성과 기술 요소는?

RAG(Retrieval-Augmented Generation)는 외부 지식 기반 검색을 LLM 생성과 결합하는 기법이다.