2026·08

SK 주식회사 AX Ackerton Partners

AI 플랫폼 설계/구축

AI DC 핵심 요소와 추진 전략

글 | SK AX, Cloud EnableX본부 허민회 본부장
SK AX, Cloud EnableX본부 장문기 전문위원

Executive Summary

AI DC의 본질은 GPU 인프라가 아니라 토큰 생산 시스템이다.

- AI DC는 GPU를 많이 배치한 데이터센터가 아니라, 전력과 냉각이 허용하는 범위 안에서 GPU, 네트워크, 스토리지, 데이터, 모델, 운영 소프트웨어를 결합해 학습 Step, 추론 Token, Embedding, 자동화 결과를 지속적으로 생산하는 고밀도 컴퓨팅 시스템이다.
- 따라서 핵심 질문은 "AI DC를 보유했는가"가 아니라 "AI Workload를 안정적이고 경제적으로 생산·운영할 수 있는가"로 바뀌어야 한다.

AI DC는 전력·냉각과 Rack/POD 단위 플랫폼을 함께 설계해야 한다.

- AI 인프라는 단일 GPU 서버 중심에서 벗어나 Rack-Scale Platform, POD-Scale Architecture, AI Factory 운영 모델로 이동하고 있다.
- 이 변화는 GPU뿐 아니라 CPU, 메모리, 스토리지, 네트워크, 전력·냉각, 운영 소프트웨어를 함께 최적화해야 한다는 점을 보여준다.

Out-Rack은 AI DC가 확장될 수 있는 물리적 상한을 결정한다.

- Out-Rack은 전력 확보, 부지, 냉각 플랜트, 네트워크 회선, 물 사용, 보안, 설비 운영 검증처럼 AI DC가 어느 규모까지 안정적으로 확장될 수 있는지를 결정한다.
- 특정 공법보다 필요한 MW를 언제 확보할 수 있는지, 고밀도 랙의 열과 지속가능성 지표를 운영 기준으로 관리할 수 있는지가 중요하다.

In-Rack은 AI DC의 실제 성능과 경제성을 결정하는 핵심 영역이다.

- In-Rack은 GPU/HBM, Scale-Up Fabric, Scale-Out Network, Storage/Data Path, 랙 전력·냉각, Scheduler, Observability가 결합되어 실제 연산 효율과 토큰 생산성을 만드는 영역이다.
- 학습용 랙은 통신 패턴과 Checkpoint를, 추론용 랙은 p95 Latency와 토큰 경제성을, Agentic AI용 랙은 Context, Tool Call, 데이터 접근 경로를 함께 설명할 수 있어야 한다.

Executive Summary

GPU 인프라는 평균 사용률보다 유효 생산성 중심으로 운영해야 한다.

- 학습 인프라는 GPU 간 통신, 체크포인트 저장·복구, 느린 노드 발생, 스토리지 처리량이 전체 학습 시간을 좌우한다.
- GPU 운영 역량은 평균 사용률이 아니라 실제 GPU 가동 시간, 전력 대비 토큰 생산량, 작업 완료 시간, 장애 복구 시간을 함께 관리하는 능력으로 판단해야 한다.

기업 고객은 AI Factory를 위한 실행형 운영 모델이 필요하다.

- AI DC 투자는 GPU 구매가 아니라 Workload Portfolio 정의, Reference Architecture, 벤치마크, 운영 KPI 설계 순서로 추진해야 한다.
- AXgenticWire Core의 GPUaaS는 프로젝트 단위 GPU 요청·할당·관측·회수, 모델 서빙, 대시보드, Metering을 통해 AI DC 운영을 서비스형으로 구현하는 실행 접점이 될 수 있다.

01 AI DC를 어떻게 정의할 것인가

[그림 1] AI DC는 전력과 부지, 랙 내부 GPU fabric, 운영 자동화가 결합된 토큰 생산 시스템

AI DC는 "토큰 생산 시스템"이다.

전력, 냉각, GPU, 네트워크, 스토리지, 운영 자동화가 하나의 생산 라인으로 결합된다.

수요와 워크로드	Out-Rack 제약	In-Rack 엔진	운영과 사업 성과
Business Demand	**Site Envelope**	**GPU Fabric**	**AI Factory Ops**
● 학습, 튜닝, 추론 ● RAG, Agent, Batch ● 데이터 민감도와 SLA	● 전력 인입과 변전 ● 부지, 물, 광케이블 ● 열 제거와 지속가능성	● GPU, HBM4, Vera CPU ● Scale-up / Scale-out ● 스토리지와 데이터 경로	● 스케줄링과 용량관리 ● AIOps, FinOps, 보안 ● 고객별 서비스 품질
먼저 정의할 것: 모델, 데이터, 지연, 비용	**확장 상한:** MW, 랙밀도, 냉각 여유	**성능 단위:** Tokens/sec, Step Time	**운영 지표:** SLO, 비용, 활용률

AI DC 의사결정 핵심 지표

Time-to-GPU	Tokens/Watt	GPU Util.	Cost/Token	SLO	Resilience
용량 확보 리드타임	전력 대비 산출	실사용률과 유휴율	서비스 원가	지연과 가용성	장애 흡수 능력

1.1 AI DC는 기존 데이터센터의 고성능 버전이 아니다

기존 데이터센터는 범용 서버, 스토리지, 네트워크를 안정적으로 수용하는 CPU 중심 시설에 가까웠고, 가상화와 컨테이너, 스토리지 가용성, 네트워크 보안, 운영 자동화가 핵심 경쟁력이었다.

AI DC는 설계의 중심축이 다르다. 대규모 모델 학습과 추론은 수백에서 수만 개의 GPU 또는 AI 가속기가 하나의 계산 작업을 수행하는 구조를 요구한다. 단일 GPU 성능이 높아도 GPU 간 통신, HBM 용량, 스토리지 처리량, 네트워크 지연시간, 냉각 능력 중 **하나가 병목이면 전체 클러스터 생산성은 급격히 낮아진다.**

경제성도 다르다. AI DC는 GPU, 고속 네트워크, 액체냉각, 고밀도 전력 설비가 결합된 선투자 구조이므로 **GPU 유휴는 곧 손실로 직결된다.** 핵심 KPI도 CPU 사용률이 아니라 GPU 실 사용 시간, 전력당 토큰 수, Job 완료시간, p95 추론 레이턴시, 장애로 손실된 GPU-Hour가 된다.

따라서 AI DC는 단순히 "GPU를 넣은 데이터센터"가 아니다. 전력과 냉각이 제공하는 물리적 한계 안에서, GPU와 네트워크, 스토리지, 데이터, 모델, 운영 소프트웨어를 하나의 생산 시스템으로 묶는 아키텍처다. 고객의 질문도 "AI DC를 보유했는가"가 아니라 "**AI Workload를 안정적으로 생산·운영할 수 있는가**"로 바뀌어야 한다.

구분	기존 데이터센터	AI DC
설계 중심	CPU 서버, 가상화, 범용 스토리지, 네트워크 안정성	GPU/HBM, Scale-up Fabric, Scale-out Network, 고밀도 전력·냉각
성능 병목	CPU, 메모리, 스토리지 I/O, 네트워크 대역폭	GPU 간 통신, HBM, Checkpoint, KV Cache, Tail Latency
운영 KPI	서버 사용률, 장애 시간, 스토리지 가용성	유효 GPU 사용 시간, 전력당 토큰 생산량, Job 완료 시간, p95 지연시간
투자 리스크	용량 과소·과대 산정	GPU 유휴, 전력·냉각 한계, 네트워크·스토리지 병목

1.2 학습 중심에서 추론, Reasoning, Agentic AI로

AI 인프라 수요는 Foundation Model 학습에서 추론, Reasoning Model, Agentic AI, Long-Context Inference, Multimodal Generation으로 넓어지고 있다. 최근 주요 발표도 단일 칩보다 Rack-Scale 또는 POD-Scale 구성을 강조하므로, 기업 고객은 학습, 추론, RAG, Agentic AI가 서로 다른 병목을 가진다는 점을 전제로 인프라를 설계해야 한다.

기업 입장에서 중요한 점은 **모든 AI 업무가 같은 인프라를 필요로 하지 않는다는 것이다. 대규모 학습은 Scale-Out Network와 Checkpoint Storage가 중요**하고, **실시간 추론은 지연시간과 안정성이 중요**하다. **RAG와 에이전트는 GPU 성능보다 데이터 접근성, 권한, 검색 품질, 감사 추적이 더 큰 병목**이 될 수 있다.

1.3 AI DC의 본질은 토큰 생산 시스템

AI DC는 데이터를 투입해 모델을 만들고, 모델을 서비스로 배포하며, **토큰과 의사결정, 자동화 결과를 지속 생산하는 공장에 가깝다.** 생산성은 **컴퓨팅, 데이터, 운영**의 세 축으로 나눠 봐야 한다.

컴퓨팅 생산성은 동일 전력과 비용에서 **학습 Step, 추론 Token, Embedding**을 얼마나 많이 생산하는가를 의미한다. 데이터 생산성은 기업 데이터가 권한, 품질, Metadata, 검색 체계까지 포함해 AI-Ready 상태인지, 운영 생산성은 장애를 감지·격리하고 남은 자원으로 작업을 지속할 수 있는지를 뜻한다. AI DC 전략은 이 세 축을 하나의 생산 라인으로 연결하는 일이다.

더 많은 내용을 보시려면

파일 다운받기

버튼을 눌러주세요