INDUSTRY ANALYSIS REPORT

AI 기술 어디까지 왔나
- 산업, 에너지, 데이터센터

수요는 **추론 · Agent · On-device**로 변하고,

공급은 **반도체 · 에너지 · 네트워크 · 데이터**에서 막힌다

SK AI 위원회 김지현
ioojoo@gmail.com

한 문장 요약

> AI 데이터센터는 “돈 먹는 하마”에서
> “토큰을 생산하는 공장”으로 바뀌고 있다.

2025: RLM 추론의 시대

생성형 AI에서 생각하는 AI로의 전환

2026: Agent 실행의 시대

자율적으로 실행하는 AI로 대전환

핵심 : AI 수요가 바뀌면서, AI 연산을 처리하기 위한 최소 단위인 토큰 비용이 중요해짐

달라진 AI로 인해 infra에 대한 요구가 달라짐

구분	의미	핵심 변화
수요 Side	AI 워크로드	학습 → 추론 → Agent → On-device → Physical
공급 Side	AI 데이터센터·반도체·전력·망·데이터	컴퓨팅 파워 → 토큰 처리 효율 → 분산 실행
바틀넥	공급이 따라가지 못하는 지점	Chip / Energy / Network / Data

관점 전환

‘얼마나 큰 모델을 만들까?’가 아니라
‘어떻게 토큰을 낮은 ‘비용·지연·전력’으로 실행할까?’가 경쟁의 중심이 된다.

PART 01

AI 수요의 변화

AI 워크로드는 **학습 중심**에서 **실행 중심**으로 이동한다

1-1.학습에서 추론으로

1-2.Agent로의 고도화

1-3.On-device AI의 등장

1-1. 학습에서 추론으로: KPI가 달라진다

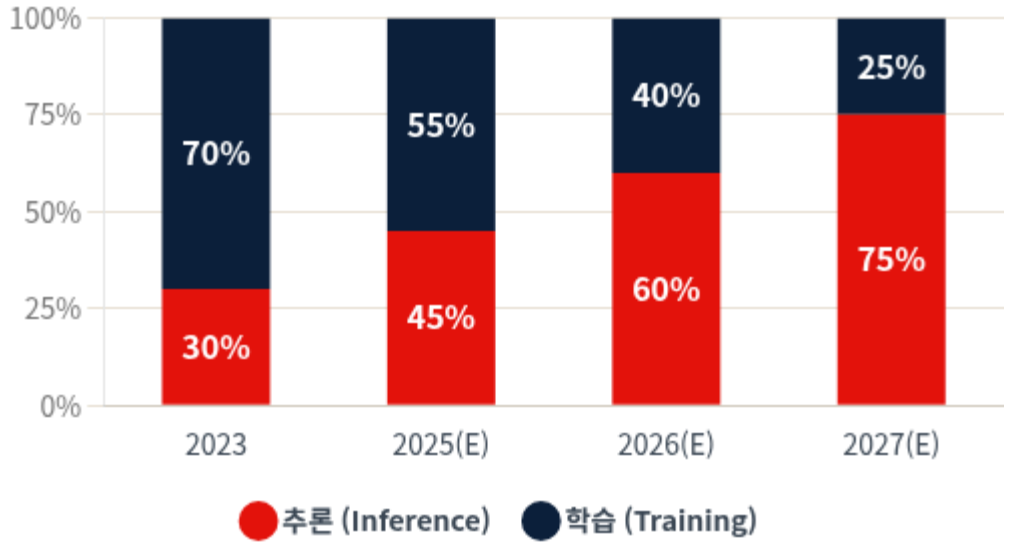
클러스터 유형	핵심 KPI (Key Performance Indicator)
학습 (Training)	최대 처리량 (Throughput) · 모델 크기 · 학습 완료 시간
추론 (Inference)	Latency · Cost per Token · Cache Hit Rate · SLA · 가동률
에이전트 (Agent)	Trajectory Cost · Tool Call 성공률 · Context 재사용률 · 상태 지속성

결론

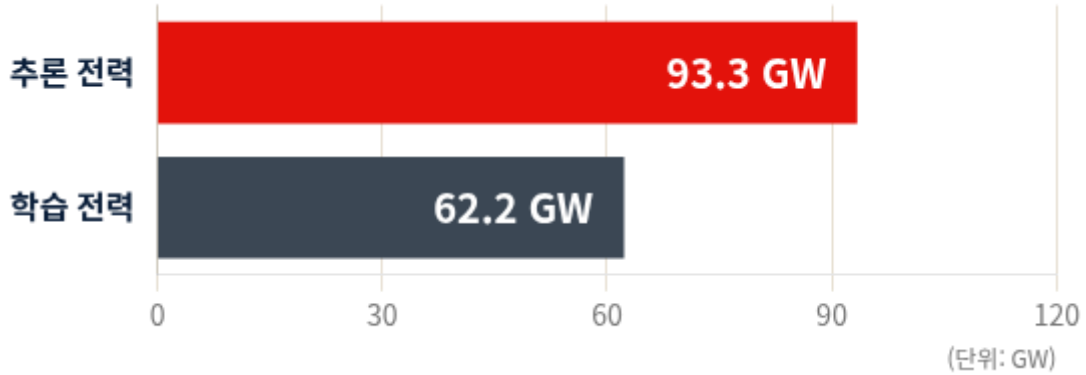
추론과 Agent 시대의 데이터센터는 '큰 계산기'가 아니라 "실시간 토큰 공장"이 된다.

※비중과 전력: 추론 중심의 재편

AI 워크로드 비중 전망 (2023-2027)



AI 데이터센터 전력 수요 전망 (2030년)



2030년 전망에서 추론 전력 수요가 학습 수요를 크게 앞지르는 구조적 역전이 발생한다.

핵심 의미

추론은 한 번 학습하고 끝나는 작업이 아니라, 사용자가 호출할 때마다 반복되는 생산 활동이다.

1-2. LLM → RLM → LAM: 토큰 폭증

단계 / 모델	핵심 능력 (진화 도식)	토큰 및 비용 예시 (폭증)
1. LLM / LMM	단순 생성 (Parameter Scale 중심)	Inference: 약 2,000 Tokens / **$0.02**
2. RLM	Reasoning (계획, 검증, 자기수정)	Reasoning: 약 14,000 Tokens / **$0.23**
3. LAM(Agent)	Action (도구 호출, 화면 조작, 업무 실행)	Agent: 약 650,000 Tokens / $4.00

왜 토큰이 폭증하는가?

Agent는 단발성이 아니라 Planning → Tool Call → Observation → Reflection → Retry의 루프를 반복하며 인프라 자원을 지속적으로 소비한다.

1-3. On-device AI : Hybrid AI로 이원화

실행 계층별 역할

위치	역할	핵심 가치
Device	Context Capture	개인·현실 맥락 즉시 이해
Edge	Low-latency Inference	가까운 곳에서 빠르게 처리
Cloud	Heavy Reasoning	복잡한 계획·장문 생성
Agent	Orchestration	최적 실행 위치 판단/호출

3개 실행 입구 (Front Door)

1. AI PC

업무 자동화 Agent
(문서, 메일, 코드, 사내 시스템)

2. AI Phone

개인 생활 Agent
(위치, 일정, 결제, 건강)

3. AI Glass / MR

Spatial Agent
(시야, 공간, 음성, 현실 환경)

핵심 메시지

설치형 Agent는 클라우드 수요를 대체하지 않고 하이브리드 수요를 키운다.

PART 02

AI 공급의 바틀넥 4가지

Chip · Energy · Network · Data

2-1. 반도체　2-2. 에너지　2-3. 네트워크　2-4. 데이터

2-1. 반도체: GPU에서 xPU 포트폴리오로

진영	대표 플레이어	강점	한계
Silicon Giants	NVIDIA, AMD, Intel	SW 생태계, Full Stack	비용·전력 부담
Hyperscalers	Google TPU, AWS Inferentia, MS Maia, Meta MTIA	자사 워크로드 최적화	외부 고객 범용성 한계
Challengers	Groq, Rebellions, FuriosaAI 등	특정 추론 워크로드 전성비	생태계·개발자 기반 한계

핵심 메시지

학습은 GPU가 강하지만, 추론·Agent·Edge는 NPU·ASIC·CPU·메모리 조합이 중요해진다.

2-1. 메모리 계층과 KV Cache

워크로드별 메모리 요구 매트릭스

대형 모델 · 범용 추론

HBM

141GB 용량 / 4.8TB/s 대역폭

긴 컨텍스트 추론

KV Cache

저장/공유를 통한 재사용 최적화

NPU 특화 추론

SRAM / INT8

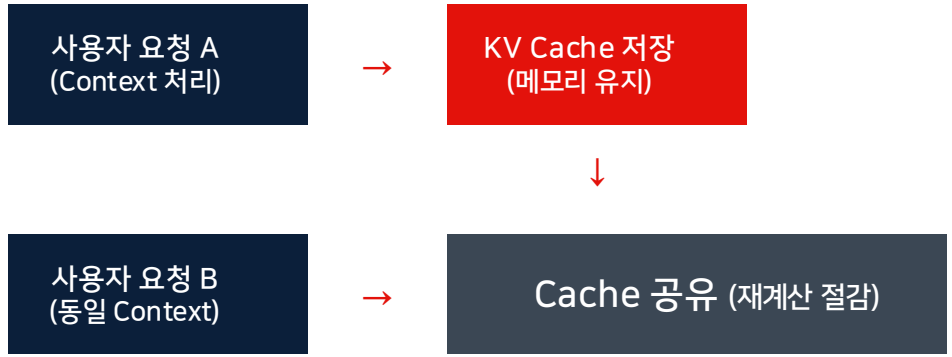
절대 용량보다 전성비 중심

On-device & Edge

LPDDR5X

초저지연 및 저전력 배터리 관리

KV Cache 재사용 메커니즘



※ 구글 터보퀀트 : KV 캐시 메모리를 6배 압축, 처리 속도 8배 향상

핵심 의미 & KPI

HBM 경쟁을 넘어 SRAM·LPDDR·CXL까지 포함한 메모리 계층 설계가 핵심이 된다.

주요 평가 지표: Cache Hit Rate · Memory Bandwidth Efficiency · Cost per Token

2-2. 에너지: DC 유형별 전력·냉각 요구

데이터센터 유형	특성 및 요구사항	설계 방향
GPU 학습 DC	고밀도 · 고발열 · 대규모 전력	대형 전력 계약, 액침/액랭, 고집적 랙
추론 DC	지속 가동 · 저지연 · 비용 효율	캐시·라우팅·전력 효율 최적화
Edge AI DC	30~50kW/Rack 수준의 분산형	Prefab, 모듈형, 근접 배치
On-device · Edge	저전력 · 저발열	NPU, LPDDR, Local SLM

핵심 메시지

AI 데이터센터의 한계는 '칩'이 아니라 '전력'이 병목이다.
기존 그리드에 연결해 전력을 공급할 수 없어, DC 자체적인 전력 생산을 필요로 한다.

2-2. 에너지 바틀넥: 3층 해법과 사회적 용납

1. 덜 쓰는 설계

- 저전력 NPU 최적화
- 캐시(KV Cache) 재사용
- 데이터 이동 거리 최소화
- 실리콘 포토닉스 도입

2. 잘 식히는 설계

- 액랭/공랭 하이브리드 냉각
- 랙(Rack) 밀도 최적화
- 고효율 모듈형 DC 구조

3. 더 확보하는 설계

- 재생에너지 연계 (PPA)
- SMR (소형모듈원전) 검토
- 스마트그리드 연동
- 송배전 인프라 최적화

숨은 바틀넥: 사회적 용납 (Social Acceptance)

전력·부지·소음·냉각수 이슈로 지역사회 갈등이 커지면서,

지상 물리적 제약을 우회하는 보조 레이어로 위성(우주), 선박 데이터센터가 적극 검토되고 있다.

2-3. 네트워크: Pipe → Control Tower

> “네트워크는 연결이 아니라
> 실행 위치와 경로를 결정하는 플랫폼이 된다.”

문제 01
동일 문맥 반복 계산

문제 02
KV Cache 재사용 부족

문제 03
서버 간 데이터 이동 지연

문제 04
DC 간 경로 비효율

핵심 메시지: 토큰 처리 비용을 줄이려면 네트워크가 경로·상태·캐시·배치를 함께 제어해야 한다.

2-3. 네트워크: 구간별 요구사항

구간	핵심 기술 및 요소	목표 지연(Latency)
서버 내부	NVLink, NVSwitch, HBM, CXL	< 1μs
서버 ⇔ 서버	RoCEv2, InfiniBand, RDMA, Silicon Photonics	< 10μs
DC ⇔ DC	DCI, DWDM, OCS, Submarine Cable, Path Pinning	< 5ms
기지국 ⇔ Edge	AI-RAN, MEC, Network Slicing, 5G/6G	< 10ms

핵심 메시지

실리콘 포토닉스는 단순 통신 기술이 아닌 AI 인프라의 지속가능성과 확장성을 보장하는 물리 계층이다.

2-4. 데이터: AI 단계별 데이터 진화

AI 단계	필요한 데이터	병목
Generative AI	인터넷 텍스트·이미지·코드	공개 데이터 품질·저작권
Agentic AI	업무 프로세스, 도구 호출, 실행 로그	**권한·보안·실패 데이터 부족**
On-device AI	개인 맥락, 위치, 앱 사용, 인증·승인	프라이버시와 단말 내 처리
Physical AI	센서, 공간, 동작, 로봇·차량 데이터	**현실 데이터 수집 비용과 안전성**

핵심 메시지

다음 데이터 병목은 “실행 데이터와 현실 데이터 부족”이다.

2-4. 데이터: Agentic·Physical 데이터 확보

"AI의 원유는 텍스트가 아닌 실행 로그와 현실 맥락이다."

Agentic 데이터

- 워크플로 패턴 및 승인/거절 이력
- Tool 호출 성공 및 실패 로그
- 기업 권한 경계 및 MCP/API 결과

Physical / 합성 데이터

- 센서, 공간, 동작 등 물리적 환경 정보
- 디지털 트윈 및 롱테일 시뮬레이션
- On-device 사용자 개인 맥락 채널

비즈니스 기회: 컴퓨팅 공급자에서 데이터·런타임·거버넌스 운영자로 확장

❍ DC 설계 기준의 변화

AS-IS | GPU 집적 중심

- 더 많은 GPU
- 더 넓은 상면
- 더 큰 전력
- 더 빠른 학습

→

TO-BE | 토큰 실행 효율 중심

- ✓ Cost per Token
- ✓ Latency SLA
- ✓ Cache Hit Rate
- ✓ Memory Sharing
- ✓ Agent Runtime Control
- ✓ Energy per Token

핵심 메시지

AI DC의 경쟁력은 '얼마나 크게 짓는가'이 아닌 '얼마나 싸고 빠르게 실행시키는가'로 평가된다.

❍ 왜 지금 Neocloud가 주목받는가: AI 전용 인프라 공급자의 등장

① AI 산업 구조 변화

- Generative → Agentic → Physical AI로 진화
- 학습에서 추론으로 이동
- AI 서비스가 기업 업무·서비스·디바이스에 내재화

② 기존 Cloud의 한계

- CPU 기반 범용 Cloud 구조
- 멀티테넌트 구조로 인한 GPU Fragmentation
- 대규모 AI 학습·추론에 비효율 및 GPU 공급 지연

③ Neocloud의 등장 배경

- GPU를 대규모 확보해 AI 특화 인프라 제공
- BMaaS(Bare Metal as a Service) 및 GPUaaS 제공
- AI 학습/추론에 최적화된 클러스터 운영

④ 핵심 가치

- GPU 즉시 공급 (Time-to-GPU)
- 대규모 단일 클러스터 제공
- Hyperscaler 대비 저렴한 비용 및 워크로드 최적화

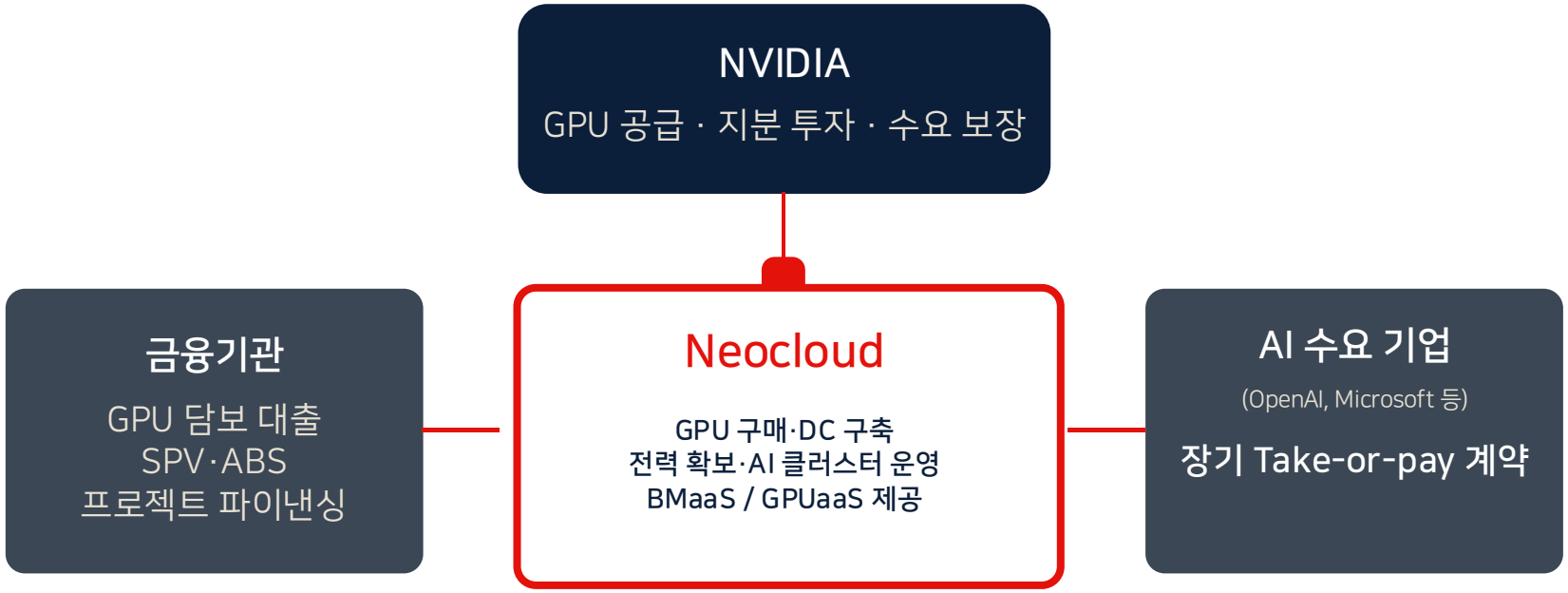
Neocloud는 단순 Cloud가 아니라 "AI Factory를 운영하는 AI 인프라 사업자"

AI 수요 증가 → GPU 부족 → AI 특화 Infra 사업 등장 → Neocloud 성장

• 대표 Neocloud 기업의 사업 구조: AI 인프라 생태계

대표 기업

- CoreWeave
- Nebius
- Crusoe
- Lambda
- Nscale
- IREN



핵심 메시지

Neocloud는 단순 GPU 임대가 아니라 GPU + 계약 + 금융이 결합된 인프라 금융 사업이다.

• Neocloud의 진짜 경쟁: AI Platform화

⚠ Bare-metal / GPUaaS의 한계

- ↓ GPU 가격 지속 하락
- 📈 감가상각 부담
- 👥 경쟁 심화 (Commodity화)
- 단순 임대 시 낮은 EBIT

★ 고부가가치 영역 (AI Platform)

AI Orchestration | Distributed Inference | MLOps | AI 개발도구 | AI Workflow | Agent Runtime | AI 운영 자동화

기존: GPU 공급 경쟁 > 현재: Token 생산 원가 경쟁 > 미래: AI Platform 경쟁

경쟁 상대	Hyperscaler (AWS·Azure·GCP·Oracle)	Neocloud
핵심 강점 비교	• 압도적 **고객 기반** 및 풍부한 **자본력** • **자체 AI 칩** 개발 통한 원가 절감 • 방대한 기존 **SW Ecosystem** 연계	• **AI workload 최적화** 인프라 설계 • **빠른 GPU 공급** (Time-to-GPU) • **대형 단일 클러스터** 기반의 높은 유연성

핵심 메시지: Neocloud의 미래 경쟁력은 "GPU 수"가 아니라 "AI 운영 플랫폼 역량"에 있다.

Neocloud 사업의 5대 성장 허들

① 막대한 초기 CAPEX

· GPU 구매 /DC 구축
· 전력·냉각·네트워크 투자

→ 초기 수조원 필요

② 금융 리스크

· 고금리 차입 구조
· 금리 상승 시 부담 확대
· GPU 가치 하락→담보 위험

→ 레버리지 산업 한계

③ GPU 가격 하락

· 신규 GPU 지속 출시
· 기존 GPU 가치 급락
· GPUaaS 가격 하락

→ 감가상각 압박 심화

④ AI 수요 변화 리스크

· 학습 → 추론 전환 가속
· On-device AI 시장 확산
· MoE·Distillation 효율화

→ GPU 수요 구조 변화

⑤ Hyperscaler 경쟁

· AWS·Azure AI Infra 강화
· 자체 특화 AI칩 확대
· 클라우드 고객 Lock-in

→ 장기적 시장 압박 가능성

살아남는 기업의 필수 역량

대규모 장기 계약 확보 · 저금리 자금 조달 · 전력 및 DC 인프라 선점 · GPU 활용률 극대화 · AI Platform 구축 역량

Scale + 금융 + AI Platform을 동시에 확보하지 못하면 장기 생존이 어려운 산업

기업 전략 시사점

1. GPUaaS만으로는 부족

Agent 는 단순 렌탈을 넘어 토큰 흐름을 최적화하는 통합 인프라 스택 필요

2. 메모리·네트워크·데이터 차별화

KV Cache, CXL, 포토닉스, Edge Routing, 데이터 거버넌스가 핵심 경쟁력

3. 하이브리드 실행 오케스트레이션

Cloud의 심층 추론과 Device의 즉시 실행을 연결하는 끊김 없는 환경 구축

4. KPI의 전환

모델 성능 중심에서 Cost per Token, Latency 토큰 단위 운영지표로 평가 기준 이동

✓ Key Takeaways & Q&A

① **AI 수요의 이동:** 학습 중심에서 추론·Agent·On-device로 실질적 이동

② **공급 바틀넥 극복:** 반도체·에너지·네트워크·데이터의 동시 다발적 재설계

③ **네트워크 & 메모리:** '더 빠른 전송'을 넘어 '더 적은 재계산과 더 높은 재사용'이 핵심

④ **에너지 3층 해법:** 물리적 전력·냉각의 최적화를 넘어 사회적 수용성 확보까지 관리

⇒ **코어 메시지: 반도체 집적 공장에서 "토큰을 싸고 빠르게 실행하는 공급망"으로**

Q & A