ETRI Insight

ICT전략연구 2026-01

추론의 시대, AI 구조 전환과 기술패권 전망

이승민

ETRI 한국전자통신연구원
Electronics and Telecommunications Research Institute

본 보고서는 ETRI ICT전략연구소 기본사업인
"ICT 기술전략 및 기술정책 연구"를 통해 작성된 결과물입니다.

목 차

표 목차

그림 목차

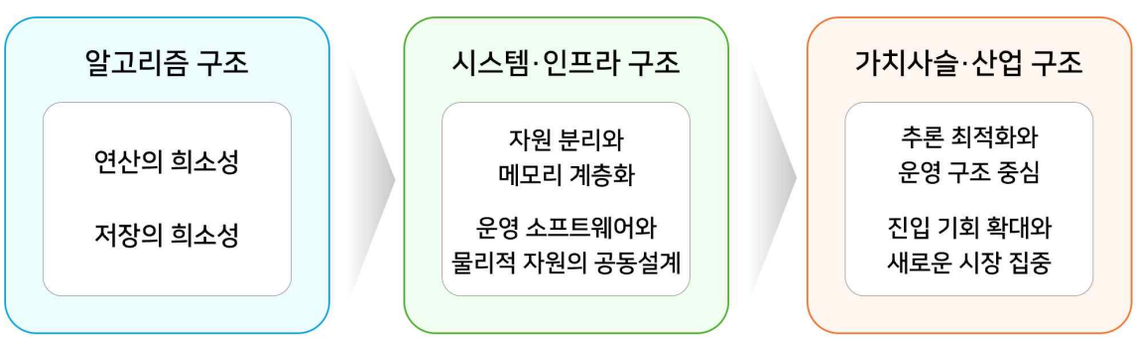
Executive Summary

AI 경쟁은 이제 학습 중심의 확장 국면을 넘어 **추론 중심의 운영 경쟁이 본격화되는 변곡점**에 진입하고 있다. 기술 경쟁의 핵심도 더 큰 모델의 선제적 확보에서, 대규모 추론을 더 낮은 비용과 더 높은 효율로 안정적으로 운영하는 구조의 설계로 빠르게 이동하고 있다. 이 과정에서 알고리즘 구조는 연산, 저장, 데이터 이동 방식을 규정함으로써 시스템·인프라·산업구조 전반을 좌우하는 핵심 변수로 부상하고 있다. 본 보고서는 이러한 구조 전환의 의미를 기술·산업·정책 관점에서 종합적으로 검토하고자 한다.

추론 중심 경쟁이 촉발한 구조적 전환은 알고리즘 구조 변화에서 출발해 가치사슬과 산업구조 재편으로 확산되는 과정으로 나타난다. 추론 중심 AI로의 전환과 알고리즘 구조 변화는 알고리즘·시스템·인프라 공동설계를 가속하고, 이는 다시 AI 가치사슬 재편과 운영 구조 경쟁을 거쳐 산업구조 변화와 기술패권 경쟁 재편으로 이어진다. 이러한 흐름에 따라 1장에서는 기술 경쟁의 전환, 2장에서는 알고리즘 구조 혁신, 3장에서는 시스템·인프라 구조 전환, 4장에서는 기술패권 경쟁 재편의 산업·정책적 함의와 전략을 살펴본다.

〈 추론 중심 경쟁이 촉발한 구조적 전환 〉

(학습) 모델 규모 중심 : 성능 ⇒ (추론) 운영 구조 중심 : 비용·처리량·지연시간



1. 추론 중심 AI 경쟁의 본격화

추론의 시대 진입과 함께 **AI 경쟁의 기준도 구조적으로 재편**되고 있다. 대규모 데이터와 연산 자원을 투입하여 성능을 고도화하는 방식은 여전히 중요하나, 실제 공공·산업 현장에서는 모델의 절대적 성능보다 안정적 운영, 비용, 지연시간, 통제 가능성이 직접적인 제약 요인으로 작용하고 있다. 특히 모델 규모가 커질수록 추론 비용과 시스템 병목이 부각되면서, 단순한 규모 경쟁만으로는 지속 가능한 우위를 확보하기 어려운 상황이 전개되고 있다. 이에 따라 AI 경쟁은 학습 능력 중심에서 운영 능력까지 포괄하는 총체적 역량 중심으로 재편되고 있으며, 기술 최적화와 투자 또한 훈련용 대형 데이터센터에서 대규모 추론 운영 서비스로 이동하고 있다.

이러한 전환은 단순한 기술 흐름의 변화가 아니라 기술패권 경쟁의 무대 자체가 바뀌고 있음을 의미한다. 추론 단계에서는 연산 효율보다 메모리 용량과 대역폭, 데이터 이동, 상태 관리가 성능과 경제성을 좌우하는 핵심 변수로 작용한다. 따라서 AI 경쟁력은 더 이상 우수한 모델의 보유 여부만으로 설명되기 어렵고, 서비스 운영·통제 구조에 의해 좌우된다. 나아가 이는 국가 차원에서 필요한 인프라 형태, 산업 생태계 구성, 기술 주권 범위까지 다시 정의하게 만드는 구조적 변화라 할 수 있다.

2. AI 알고리즘 구조 혁신

본 장에서는 현재 AI의 표준적 기반이 된 **트랜스포머 구조가 추론 중심의 서비스 확산 국면에서 직면한 한계**를 중심으로 혁신 방향을 살펴본다. 첫째, 모든 파라미터 활성화와 연산 비효율의 한계다. 입력 토큰이 매 계층에서 동일한 경로를 거치며 모든 파라미터가 활성화되는 구조에서는 모델이 커질수록 연산량과 메모리 접근 비용이 급증한다. 학습 단계에서는 병렬 처리로 일정 부분 이를 흡수할 수 있었으나, 추론 단계에서는 시간 응답, 동시 요청 처리, 비용 제약이 직접적인 확장성 문제로 이어진다. 둘째, 파라미터·상태 접근 비용 증가와 메모리 병목의 한계다. 긴 문맥과 대규모 동시 요청 환경에서는 상태 정보(KV cache)와 파라미터의 반복적 접근·저장·이동이 요구되며, 실제 병목은 연산 자체보다 상태 접근과 메모리 용량·대역폭, 데이터 이동 비용으로 이동한다. 이로 인해 AI 기술 혁신의 방향은 데이터와 모델 크기 경쟁에서, 연산 수행 방식과 상태 저장·관리 구조의 설계 문제로 전환되고 있다.

이러한 혁신의 핵심에는 희소성이라는 설계 원리가 있다. 희소성은 단순히 모델 크기를 줄이는 기법이 아니라, 제한된 연산·메모리·대역폭·지연 안에서 필요한 부분만 선택적으로 활용하도록 구조를 바꾸는 방식이다. 첫 번째 축은 **연산의 희소성**이다. 이는 전체 파라미터를 항상 모두 사용하는 대신, 필요한 일부만 선택적으로 활성화함으로써 모델 규모의 확장과 실제 연산 비용을 분리하려는 접근이다. 이를 통해 전체 모델 규모는 확대하지만, 토큰당 연산 부담은 통제할 수 있기 때문에 초거대 모델의 경제성을 개선할 수 있다. 다만 연산 효율의 개선만으로는 추론 환경의 병목이 충분히 해소되지 않는다. 상태 정보의 유지와 이동 비용이 여전히 크게 작용하기 때문에, 혁신의 범위는 결국 저장의 희소성으로 확장된다.

저장의 희소성은 기억 대상과 저장 위치·형태의 선택 문제를 AI 구조 설계의 중심에 둔다. 이는 단순한 연산량 절감만으로 해결할 수 없는 메모리 중심 병목을 다루기 위한 접근으로, 상태 관리 효율화, 어텐션 구조 재설계, 외부 메모리 조회를 통한 지식 접근 방식의 전환 등으로 구체화된다. 핵심은 모든 상태를 동일한 방식으로 저장하지 않고 필요한 정보만 효율적으로 유지하며, 저장 대상 자체를 축소하고, 나아가 일부 지식 처리를 계산이 아니라 조회 중심 방식으로 전환하는 데 있다. 이는 성능 향상의 원천이 더 이상 단순한 규모 확대에 있지 않고, 계산·저장·이동 방식을 재조합하는 구조 혁신에 있음을 보여준다. 동시에 AI 알고리즘의 경쟁력은 벤치마크 성능을 넘어 실제 운영 환경에서 데이터 이동 비용과 메모리 사용량까지 함께 절감할 수 있는 역량에 의해 평가되어야 함을 시사한다.

3. 희소성이 강제하는 시스템·인프라 구조 전환

본 장에서는 알고리즘 구조 혁신이 시스템과 인프라를 어떻게 재설계하도록 만드는지를 다룬다. 알고리즘의 희소성 확산은 연산과 저장을 별도로 최적화하는 수준을 넘어, 두 요소를 하나의 설계 문제로 통합하도록 요구하고 있다. 연산을 줄이더라도 상태 유지 비용이 크면 전체 효율은 제한되며, 상태를 줄이더라도 토큰마다 과도한 계산이 수행되면 비용 문제는 다시 확대된다. 이 때문에 **알고리즘은 더 이상 하드웨어와 시스템 환경을 고려하지 않은 채 독립적으로 설계될 수 없으며**, 어떤 메모리 계층을 활용할 것인지, 어떤 연산 자원이 필요한지, 어떤 운영 구조가 효율적인지를 함께 고려하는 방향으로 이동하고 있다. 이는 AI 경쟁의 축이 학습에서 효율적 추론으로 이동하였음을 보여주는 결정적 변화이기도 하다.

추론 중심 AI에서는 성능의 척도도 근본적으로 달라진다. 과거에는 빠른 학습 능력이 핵심 경쟁력이었다면, 이제는 얼마나 낮은 지연시간과 비용으로 얼마나 많은 요청을 안정적으로 처리할 수 있는지가 핵심 기준이 된다. 생성형 AI의 추론은 입력 전체를 처리하는 단계와 토큰을 순차적으로 생성하는 단계로 나뉘며, 두 단계는 요구하는 자원과 병목의 성격이 다르다. 전자는 연산 집약적이고 후자는 메모리 및 상태 접근 집약적이기 때문에, 동일 자원에서 함께 처리할 경우 부하 간섭이 발생할 수 있다. 이에 따라 자원을 논리적·공간적으로 분리하고 각 단계에 맞는 역할을 부여하는 구조가 중요해지고 있다.

이와 함께 **메모리 구조 또한 단일 고대역폭 메모리에 의존하는 방식에서 다층적 계층 구조로 전환**되고 있다. 추론 과정에서 상태의 저장·이동·재사용이 핵심이 되면서, 메모리는 단순한 저장 장치가 아니라 지연시간, 처리량, 비용을 동시에 결정하는 전략 변수로 작용한다. 빠른 처리가 필요한 데이터와 활성 상태는 최상위 메모리 계층에 두고, 대용량이면서 재사용 가능한 상태 데이터는 보다 저렴한 계층으로 분산하는 구조가 현실적 대안으로 부상하고 있다. 결국 이러한 흐름은 **알고리즘·시스템·인프라 공동설계**의 필요성으로 이어지며, 앞으로의 AI 경쟁력은 단일 기술의 우수성보다 각 계층을 반복적으로 조정해 전체 운영 효율을 극대화하는 능력에서 결정될 것이다.

4. 기술패권 경쟁의 재편과 전략적 선택

본 장에서는 알고리즘, 시스템, 인프라 전반의 구조 변화가 산업구조와 국가 전략에 미치는 영향을 종합적으로 살펴본다. 첫째, **AI 가치사슬의 중심**이 모델 규모 자체보다 운영 구조로 이동하고 있다. 서비스 확산 단계에서는 지연시간, 처리량, 비용이 실질 경쟁력을 좌우하며, 이에 따라 기존의 모델 규모 중심 가치사슬은 알고리즘 구조, 시스템 설계, 인프라 표준이 결합된 운영 구조 중심 가치사슬로 재편되고 있다.

둘째, **산업구조**는 진입 기회의 확대와 새로운 집중이 동시에 나타나는 방향으로 바뀌고 있다. 구조적 최적화와 추론 효율화는 상대적으로 적은 자원으로도 경쟁 가능한 영역을 넓히지만, 실제로는 AI 서비스를 운영·배포하고 운영 기준을 설계하는 단계에 새로운 지배력이 집중되고 있다.

셋째, **반도체 산업의 경쟁 구도** 또한 범용 학습 가속기 중심에서 추론 특화 구조 중심으로 이동하고 있다. 추론 중심 시대에는 단순 연산 성능보다 메모리 접근, 데이터 이동, 계층화된 자원 구조와의 결합이 더 중요해지며, 반도체 경쟁도 단일 칩의 최고 성능보다 시스템 전체의 운영 효율과 공동설계 역량에 의해 좌우되는 방향으로 전환되고 있다.

넷째, **국가 AI 주권의 의미** 또한 모델 보유 중심에서 운영 주도권과 통제력 중심으로 확장되고 있다. 공공 행정, 산업 자동화, 국방 의사결정과 같이 상시적 서비스가 요구되는 영역에서는 모델 자체의 우수성보다 지속적 운영 가능성, 장애 대응, 정책 반영, 비용 통제 능력이 더 직접적인 경쟁 기준이 되기 때문이다. 추론 중심 AI 시대에는 데이터센터가 지능형 토큰을 생산하는 핵심 설비로 변화하면서, **AI 기술패권의 중심**도 모델 보유에서 지능의 생산·배포·통제 능력으로 이동하고 있다. 결국 실질적 AI 주권은 소버린 모델만으로 완성되지 않으며, 소버린 운영과 소버린 통제가 결합될 때 비로소 확보될 수 있다.

따라서 향후 국가 전략은 개별 기술 확보나 단일 산업 지원을 넘어, AI 구조 전환이 제기하는 핵심 전략 요구를 식별하고 이를 기술 전략, 산업 전략, 정책 방향으로 연계하는 방식으로 설계될 필요가 있다. 본 보고서는 이러한 관점에서 AI 추론 시대에 대응하기 위한 10대 전략 과제(안)을 다음과 같이 제시한다.

《 AI 추론 시대 대응 10대 전략 과제(안) 》

10대 전략 과제(안)		주요 내용
기술	① 추론 효율형 알고리즘의 선도 기술 확보	연산·저장 희소성 기반 구조 혁신을 통해 토큰당 비용, 지연시간, 처리량을 개선하는 추론 특화 알고리즘 기술을 선도적으로 확보
	② 국산 추론 운영 체계의 공통 기반 확보	국산 추론 운영 체계의 공통 기반을 마련해 국내 기술이 실제 서비스 환경에서 작동할 수 있는 기반 구축
	③ 메모리 중심 추론 인프라의 공동설계 역량 확보	HBM·DRAM·SSD·CXL 등 다층 메모리 구조와 가속기·시스템 소프트웨어를 통합 최적화하는 공동설계 및 실증 역량 확보
	④ 국산 NPU의 서비스형 통합 역량 확보	개별 칩 성능 경쟁을 넘어 오픈소스 추론 스택과 결합된 서비스형 통합 구조를 구축하고, 공공·제조·통신 등에서 실증 가능한 운영 표준 형성
산업	⑤ AI 반도체-운영 소프트웨어 결합 산업 육성	운영 계층 전문기업 육성과 함께 국산 AI 반도체와 운영 스택이 결합된 솔루션·플랫폼형 사업모델을 발굴하여 운영 계층 산업 기반 강화
	⑥ 수요산업 중심의 고신뢰 추론 시장 선점	제조, 통신, 공공, 국방 등 한국의 강점 수요산업에서 고신뢰·저지연·고효율 추론 서비스를 조기 확산해 실증 기반 시장 선점
	⑦ 개방형 상호운용 생태계 구축	오픈소스 추론 스택, 표준 인터페이스 등을 바탕으로 특정 플랫폼 종속을 줄이고 국내 중소·전문기업 참여가 가능한 개방형 생태계 조성
정책	⑧ 소버린 AI의 국가 운영체계 정립	소버린 AI의 범위를 단순 모델 보유에서 운영 가능성·통제력·지속가능성으로 확장, 분산된 AI 정책을 범정부 차원의 국가 운영체계로 구조화
	⑨ 국가 AI 컴퓨팅 정책의 중심을 운영 인프라로 전환	훈련용 대형 컴퓨팅 중심 정책에서 벗어나 추론 서비스 운영, 메모리 계층화, 시스템 효율을 중심으로 국가 AI 인프라 정책 재편
	⑩ 운영 거버넌스의 선제적 구축	공공·산업 현장에서 요구되는 안정성, 보안, 통제 가능성, 표준·인증, 책임체계 등을 포함하는 AI 운영 거버넌스를 선제적으로 정립

Ⅰ 추론 중심 AI 경쟁의 본격화

1 학습 규모 경쟁에서 추론 운영 경쟁으로

AI 경쟁 기준의 변곡점

- 2026년 3월 GTC 2026은 인공지능 경쟁의 무게중심이 학습 중심의 확장 경쟁에서 **추론 중심의 운영 경쟁으로 이동**하고 있음을 산업 차원에서 분명히 보여준 계기[1)]
  - 2022년 11월 ChatGPT 등장 이후 본격화된 초거대 모델 경쟁은, 2024년 9월 OpenAI의 추론형 모델 'o1'을 계기로 경쟁의 기준이 바뀌기 시작
  - 이후 2025년 Claude Code와 Claude 4 계열, 그리고 2026년 초에 부각된 자율 에이전트 및 OpenClaw 사례는 AI 경쟁이 단순 생성에서 지속적 추론과 작업 수행 중심으로 확대되는 흐름을 제시
  - 특히, GTC 2026에서 제시된 'Agentic Scaling'은 인간이 아닌 에이전트형 AI가 주도하는 추론 수요가 확대되면서, AI 경쟁이 모델 학습 능력뿐 아니라 대규모 추론을 안정적으로 운영하는 능력으로 재편되고 있음을 보여준 산업적 계기

※ AI 스케일링 4단계: Pre-training scaling → Post-training scaling → Test-time scaling → Agentic scaling

- 추론 시대의 AI 산업은 학습 중심의 규모 경쟁에서, 제한된 자원 안에서 토큰당 생산 단가를 낮추는 **'토큰 공장' 효율화 경쟁**으로 빠르게 재편
  - 이는 제한된 자원 내에서 더 높은 품질의 추론 결과를 더 낮은 지연과 비용으로 산출하는 능력, 즉 '지능의 밀도'[2)] 제고 문제로 요약될 수 있으며, 결과적으로 AI 토큰 경제학의 중요성을 부각

※ SemiAnalysis의 'AI Token Factory Economics'는 데이터센터를 단순히 데이터를 저장하고 처리하는 곳이 아니라 지능이라는 결과물인 토큰을 생산하는 현대적 공장으로 정의하고 수익 구조를 분석[3)]

  - 이러한 변화는 결국 AI 경쟁의 초점을 모델 규모 자체보다 추론 비용 절감과 운영 구조 효율화 중심으로 이동

※ SemiAnalysis 벤치마크 결과, 최신 NVIDIA 플랫폼은 타사 대비 운영 비용 최대 35배 절감 발표[1)]

AI 기술 경쟁의 현주소와 중심 이동

- 최근 AI 경쟁은 더 큰 모델의 확보 자체보다, 실제 서비스 환경에서의 효율적·안정적 운영 역량의 문제로 빠르게 이동 중

1) NVIDIA(2026), Watch Jensen Huang's GTC 2026 Keynote: On Demand.
2) NVIDIA GTC 2026 Keynote에서 젠슨 황은 '지능의 밀도(Intelligence Density)'를 주어진 물리적 제약 내에서 추출 가능한 지능의 총량으로 설명하며, 이를 추론 시대의 핵심 경쟁력으로 제시
3) SemiAnalysis, Tokenomics Model, https://semianalysis.com/tokenomics-model/.

- 이 과정에서 알고리즘 구조는 계산 대상, 저장 대상, 데이터 이동 방식을 결정하는 핵심 원리로 부상
- 따라서 **알고리즘 구조**는 단순한 모델 내부 설계를 넘어 시스템 설계, 인프라 구성, 운영 효율을 좌우하는 핵심 변수로 작용

○ ChatGPT 등장 이후 AI 경쟁은 대규모 데이터와 막대한 연산 자원을 투입하면 성능이 개선된다는 경험의 법칙이 산업 전반에 확산하고 있으나 동시에 두 가지 현실적 한계에 직면

- 첫째, 학습 중심의 경쟁 방식은 막대한 선투자 비용과 공급망 제약을 동반하여, 스케일링 자체가 지속되더라도 그 성과가 서비스 가치로 전환되는 속도와 효율은 별개의 문제

※ 공공·산업 현장에서 모델의 성능보다 안정적 운영, 비용, 지연시간, 보안·통제 가능성이 더 직접적인 제약으로 작용

- 둘째, 모델 규모 확대는 필연적으로 추론 비용과 시스템 병목을 표면화하여, 단순히 파라미터 수를 늘리는 전략만으로는 경쟁 우위 확보에 제한적

※ 사용자 체감 품질은 결국 신속성, 경제성, 그리고 대규모 요청에 대한 서비스 안정성 문제로 귀결

- 이로 인해 경쟁의 초점은 모델 규모 자체보다, 추론 비용과 시스템 병목을 줄일 수 있는 구조 설계 역량으로 빠르게 이동

○ AI 경쟁의 중심이 학습에서 추론으로 이동하면서, 데이터센터 컴퓨팅 수요·투자·기술 최적화의 초점도 '훈련용 초대형 데이터센터'에서 **'대규모 추론 운영 서비스'**로 전환

- 최근 Google은 대규모 사전학습용 TPU 8t와 추론용 TPU 8i 분리 구조를 제시하는 등[4] 주요 클라우드 사업자의 AI 인프라 중심축은 범용 학습 가속기에서 추론 특화 구조로 이동 중
- 이와 함께 추론 특화 칩, 운영 구조, 메모리 계층화 전략이 빠르게 부상하고 있으며, 이는 기술패권 경쟁의 무대가 모델 학습을 넘어 모델 운영과 인프라 설계로 확장되고 있음을 시사
- 즉, 초거대 모델 경쟁이 계속되더라도, 국가·기업의 실질적 우위는 학습 능력만이 아니라 운영 능력까지 포함하는 총체적 역량으로 재정의

➡ 추론 중심 경쟁의 기술적 함의

○ 추론은 학습과 달리 사용자 패턴에 따른 가변성이 높고, 토큰을 순차적으로 생성하는 특성으로 인해 연산 효율보다 메모리 및 데이터 이동의 병목 현상이 지배적으로 발생

- 추론에서는 입력 프롬프트를 처리하는 프리필(Prefill) 단계와 토큰을 생성하는 디코드(Decode) 단계를 구분하여 자원을 분리 운영하는 설계가 지연시간과 처리 효율을 좌우

※ Prefill은 사용자가 입력한 프롬프트(컨텍스트)를 병렬적으로 처리하며 내부 상태(KV cache 등)를 생성하는 단계, Decode는 그 상태를 읽어가며 토큰을 순차적으로 생성하는 단계

4) Google(2026.4.23.), Inside the eighth-generation TPU: An architecture deep dive.

- 이 구분이 중요한 이유는 추론 단계에서의 병목이 연산량(FLOPs) 자체보다 메모리 용량·대역폭·데이터 이동에서 발생하기 때문
- 추론 메커니즘의 특징은 알고리즘, 추론 시스템, 인프라가 더 이상 분리된 계층으로 최적화되기 어렵고, 메모리 계층, 데이터 이동, 통신 대역폭, 서빙 구조[5]의 제약 속에서 함께 설계되어야 한다는 점에서 구조적

〈 추론 비용의 경제적 장벽 〉

- 스케일링 법칙은 AI 모델의 지능 수준이 투입된 자원의 지수적 증가에 비례한다는 것을 증명하였으나, 그 이면에는 추론 단계에서의 비효율성이라는 치명적 약점 노출
- 즉, LLM은 단일 토큰 생성에도 대규모 모델 가중치와 이전 문맥의 KV cache를 반복적으로 읽고 저장·참조해야 하므로 연산 비용과 메모리 부담이 함께 누적되는 구조적 한계를 보임
- 학습 비용은 일회성 비용인 반면, 추론 비용은 사용자 수에 비례하여 발생하는 반복적 비용이라는 점에서 큰 차이
- 기술 발전에 따라 단위당 생산 원가가 절감되는 경우와 달리, LLM은 모델 크기를 키울수록 추론에 필요한 고대역폭 메모리(HBM)의 요구량과 GPU 유지 비용이 기하급수적으로 증가

→ 토큰당 단가를 낮추기 어려운 **'경제적 임계점'**을 만들면서, 이는 자원 효율화를 위한 **알고리즘의 구조 혁신**을 가속화

- 추론 중심 경쟁은 단순 수요 확대를 넘어 모델 최적화 설계 및 효율적 운영 역량이 국가 AI 경쟁력의 핵심 척도로 부상함을 시사
  - 학습 중심 경쟁이 자본과 GPU 확보(양적)에 치중했다면, 추론 중심 경쟁은 시스템 소프트웨어 스택과 인프라 운영 능력(질적)에 따라 서비스 주도권이 결정
  - 이러한 기술적 전환은 국가 AI 인프라 투자, 공공 서비스 도입, 산업 생태계 구축의 기준을 단순 하드웨어 성능이 아니라 전체 운영 생태계의 효율성 중심으로 재편하는 변수로 작용

AI 알고리즘 변화와 운영 구조 전환

- 초거대 AI의 병목은 연산량 자체보다 파라미터와 KV cache[6]를 읽고 저장하는 과정에서 발생하며, 이를 효율적으로 제어하기 위한 알고리즘·시스템 차원의 설계가 필수적
  - 동시 요청과 문장 길이에 따라 급격히 증가하는 메모리 점유 문제를 해결하기 위해, 캐시 관리, 상태 압축, 메모리 계층화, 자원 분리 등 운영 구조 전반의 최적화 기술이 고도화
  - 단순히 빠른 하드웨어를 도입하기보다, **알고리즘·시스템·인프라를 함께 재설계해 데이터 이동과 운영 비용을 최소화하는 방향으로 진화 중**

5) 서빙 구조(Serving Architecture)는 학습된 AI 모델이나 알고리즘을 실제 사용자 서비스에 연결하여, 요청 처리·자원 배치·추론 실행·응답 제공이 이루어지도록 하는 실행·운영 체계를 의미
6) 추론 시 이전 단계에서 계산된 Key와 Value 벡터를 메모리에 저장해 재연산을 방지하는 기법

표 1 알고리즘·시스템·인프라 계층 구분

계층	의미	기능(예)
알고리즘	계산·저장 방식의 설계로, 모델이 결과를 생성하는 내부 구조	어텐션 메커니즘, KV cache 저장·압축·재사용 방식, 긴 문맥에서의 메모리 낭비 최소화
시스템	요청·상태·자원을 배치·제어하는 운영 소프트웨어 계층	배칭, 스케줄링, Prefill·Decode 분리, 라우팅, KV cache 이동·공유·오프로딩
인프라	운영 소프트웨어가 실행되는 물리적 자원과 네트워크 계층	GPU/NPU/TPU, CPU, HBM·DRAM·SSD·CXL, 초고속 인터커넥트, 랙·데이터센터 구조

※ 출처: 저자 작성.

- 연산의 희소성[7]과 저장의 희소성을 두 축으로 하는 구조적 최적화가 모델의 크기를 키우는 것보다 실질적인 성능 및 경제성 향상을 주도
  - 연산의 희소성은 모든 파라미터를 사용하지 않고 필요한 부분만 활성화하는 조건부 연산으로 대규모 모델의 성능을 유지하면서 토큰당 연산 비용을 획기적으로 낮추는 구조가 핵심
  - 저장의 희소성은 추론 단계의 상태 정보를 효율적으로 관리하거나 압축하여 메모리 점유를 최소화하고 데이터 이동 경로를 단축함으로써 시스템의 동시 처리 용량을 극대화
- 이러한 흐름에서 주목할 점은 '알고리즘의 구조 변화'가 이론적 논의를 넘어, 모델의 연산 경로를 조건부로 만들고(연산의 희소성), 상태를 줄이거나 계층화하며 데이터를 통제하는(저장의 희소성) 방향으로 상용화되고 있다는 사실
  - 즉, 성능 향상이 더 이상 모델의 규모 확장만의 결과가 아니라 구조적 최적화의 결과가 되고 있으며, 이러한 구조 변화는 시스템·인프라와 산업 경쟁의 규칙까지 재편
  - 이러한 재편은 우리나라처럼 메모리·패키징·제조 기반이 강한 국가에 기회 요인으로 작용할 수도 있으나, 동시에 소프트웨어 스택·운영 표준을 선점하는 주체에게 새로운 지배력이 집중될 가능성도 있기에 면밀한 주의와 선제 대응이 요구

표 2 AI 경쟁의 패러다임 전환

구분	학습 중심 경쟁	추론 중심 경쟁
주요 목표	절대적 모델 성능 고도화	운영 효율(지연시간, 처리량, 비용)
연산 구조	밀집(Dense) 연산 구조	희소성(Sparse) 연산 구조
주요 병목	연산 한계(Compute Bound)	메모리 용량·대역폭(Memory Bound)
인프라 형태	중앙집중 훈련용 대형 인프라	사용자 인접형 대규모 분산 인프라

※ 출처: 저자 작성.

7) AI 분야에서 '희소성(Sparsity)'은 '전체 중 아주 일부만 활성화된다'는 의미로, 촘촘하게 가득 찬(Dense) 데이터 대신, 대부분이 0(Zero)이고 소수의 유효한 값만 드문드문(Sparse) 존재하는 상태

2 AI 구조 전환과 기술패권의 전략적 의미

AI 기술 경쟁의 구조 전환과 대응 체계 재정립 필요

- 글로벌 AI 경쟁의 축이 단순 모델 성능과 파라미터 규모 중심에서 알고리즘 구조, 시스템 설계, 인프라 결합 역량 등 복합적인 운영 효율성 중심으로 빠르게 이동
  - 초거대 모델의 학습·운영 비용 급증에 따라 AI 활용의 지향점이 대규모 학습 단계에서 실질적인 서비스 확산을 위한 추론 및 최적화 단계로 전이되는 양상
  - 국가별 AI 역량을 단일 성능 지표가 아닌 시스템 설계 능력과 인프라 통합 역량 등 다각적인 구조적 대응력 관점에서 재평가하고 전략을 수정해야 할 시점
- 한국의 현재 위치는 단순 선도·후발 구도로 규정하기 어려우며, 독자적 AI 모델 확보라는 필수 과제와 구조적 활용이라는 전략적 선택지를 동시에 보유한 상태
  - 선도국 대비 자본력의 한계는 존재하나, 소버린 AI 정책을 통해 기술 주권의 핵심인 독자 모델 개발 역량을 국가 전략의 필수 기반으로 설정하고 추진 중
  - 단순 개발 성과에 머물지 않고 확보된 모델 역량이 산업 전반에 지속적으로 활용될 수 있는 최적의 구조적 설계안을 마련하는 것이 차세대 AI 경쟁력의 핵심
  - 모든 모델을 자체 개발하는 단편적 접근 대신, 연산·저장·데이터 이동의 최적화를 통해 **국가 전략 목표에 맞춘 현실적인 AI 역량 확보 방안을 마련할 필요**
  - 즉, 독자 모델 역량은 자체 기술 확보의 기반이며, 구조적 활용 역량은 이를 실제 산업·공공 현장에서 경쟁력으로 전환하는 힘이라는 점에서 두 과제는 병행적으로 추진될 필요

기술패권 경쟁의 본질 변화와 구조 전환의 전략적 의미

- 기술패권은 단일 기술의 우열을 넘어 가치사슬, 표준, 인프라, 제도를 포괄하는 장기 경쟁이며, 추론 중심 AI로의 구조 전환은 국가 전략을 결정짓는 핵심 변수로 작용
  - 첫째, AI 가치사슬은 모델 자체의 개발·보유에서 운영 가능한 시스템의 구축·확산으로 중심이 이동하며, 국가가 구축해야 할 인프라와 생태계의 형태 및 판단 기준을 결정

※ OECD 등 국제기구의 사례에서 보듯, 정부 기능 전반의 AI 확산은 기술적 성능을 넘어 신뢰성, 비용 효율성, 운영 역량이 뒷받침될 때 비로소 지속 가능

  - 둘째, AI 주권의 범위도 독자 모델 보유를 넘어 인프라 설계와 운영 관리 영역으로 확장되며, 국가의 자율적 배포·업데이트·제어 능력을 실질적으로 강화하는 토대

※ AI 주권은 데이터・모델・인프라・인재에 대한 자율성과 통제력을 포함하는 개념으로, 국가가 자국의 전략 목표에 맞춰 AI를 독립적으로 개발・배포할 수 있는 역량을 의미

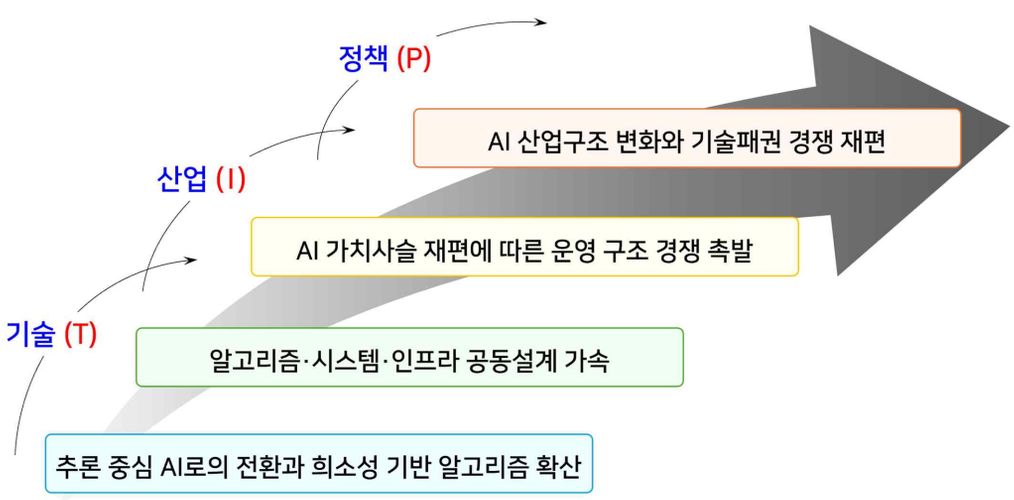
- 셋째, 알고리즘·시스템·인프라의 결합 방식은 반도체와 클라우드, 운영 소프트웨어, 표준·거버넌스 경쟁의 기준을 바꾸고 있으며, 이는 기술패권 경쟁이 단일 모델 성능을 넘어 운영 생태계 전반의 통제력 경쟁으로 확대되고 있음을 의미

※ AI 확산에 따른 기술패권은 거대 AI 모델 보유를 넘어 책임 있는 운영(캐시 정책, 서빙 구조, 분리 운영 등)과 신뢰 체계 구축 여부에 의해 결정되며, 이는 글로벌 표준 경쟁의 핵심 요소

- 본 보고서에서 다루는 **'구조(적) 전환'**은 알고리즘과 시스템 차원의 기술 변화에 그치지 않고, AI 가치사슬 재편, 반도체 경쟁 구도 변화, 운영 주도권과 국가 AI 주권의 결합, 그리고 국가 전략과 정책 방향의 재정립까지 포괄

- 이는 AI 가치사슬 재편, 산업구조 변화, 서비스·운영 주도권과 AI 주권의 결합, 컴퓨팅 인프라(반도체 등)의 전략적 위상 변화, 국제 표준·규범 경쟁 등의 전략적 변화를 수반
- 특히 반도체는 이러한 전환의 물적 기반으로서 가장 직접적인 영향을 받는 영역이지만, 동시에 알고리즘 구조 변화가 촉발하는 경쟁은 반도체 단독이 아니라 운영 스택과 생태계, 국가의 인프라 전략까지 동시에 재편한다는 점에서 더 큰 전략적 함의를 내포
- 본 보고서는 이러한 전환을 기술–산업–정책 관점에서 체계적으로 분석함으로써, 기술개발 기획과 국가 정책 설계를 위한 분석 프레임과 시사점을 제시하고자 함

그림 1 AI 구조 전환이 가져온 기술패권 재편의 확산 경로



※ 출처: 저자 작성.

Ⅱ AI 알고리즘 구조 혁신

1 트랜스포머 구조의 한계와 진화 방향

➡ 트랜스포머의 두 가지 구조적 한계

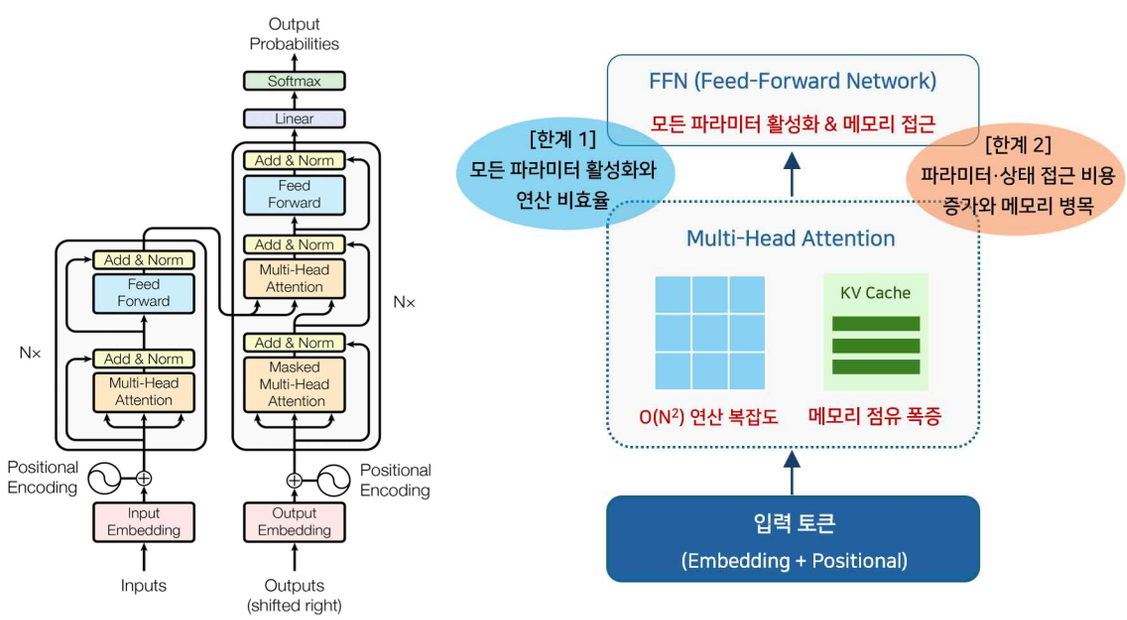
- 트랜스포머는 지난 수년간 AI 분야에서 사실상의 표준 아키텍처로 자리 잡으며 언어 처리, 비전, 로봇 제어 등 다양한 영역으로 빠르게 확산하고 있으나 근본적인 한계에 직면
  - **[한계 1: 모든 파라미터 활성화(Dense)와 연산 비효율]** 입력 토큰이 모든 레이어에서 동일한 연산 경로를 거치며 대부분의 파라미터가 활성화되므로, 모델 규모가 커질수록 연산량과 메모리 접근 비용 증가

※ 학습 단계에서는 대규모 병렬 처리를 통해 이러한 비용을 어느 정도 흡수할 수 있었으나, 추론 단계에서는 시간 응답, 대규모 동시 요청, 비용 제약 등이 확장성 제약으로 작용

  - **[한계 2: 파라미터·상태 접근 비용 증가와 메모리 병목]** 긴 문맥과 대규모 동시 요청 환경에서는 파라미터와 상태 정보(KV cache)의 반복적 접근·저장·이동이 요구되므로, 메모리 용량·대역폭·데이터 이동 비용이 핵심 병목으로 부상

※ 특히 실제 대규모 서비스 환경에서 추론 병목의 상당 부분이 연산 자체보다 KV cache 접근과 메모리 대역폭에서 발생

그림 2 트랜스포머 구조와 병목



※ 출처: (좌) 트랜스포머 구조: Ashish Vaswani et al.(2017), (우) 트랜스포머 병목 지점: 저자 작성

○ 알고리즘의 구조적 한계는 **'데이터와 모델 크기 경쟁'**에서 **'연산 수행 방법, 상태 저장 및 관리 구조'**로 경쟁의 방향을 전환

- 트랜스포머의 한계는 단순 최적화가 아니라, 알고리즘 구조 자체를 재검토하게 만드는 압력으로 작용
- 이러한 문제를 해결하기 위해, 알고리즘 설계의 핵심 쟁점으로 '희소성(Sparsity)' 개념과 관련 연구가 부상

〈 병목 지점의 이동 〉

- 트랜스포머 기반 모델은 임베딩-어텐션-FFN(Feed-Forward Network) 구조가 레이어 단위로 반복되며, 이 구조에서 학습과 추론 모두 가중치(파라미터)를 읽고 행렬곱을 수행
- 그러나 추론에서는 여기에 KV cache라는 상태가 추가되며, 이로 인해 병목은 크게 ① 파라미터(정적) 접근 비용과 ② KV cache(동적) 유지·접근 비용으로 구분
- 전자는 모델이 커질수록 고정적으로 매번 읽어야 하는 비용이고, 후자는 컨텍스트 길이와 동시성에 따라 계속 커지는 비용
- 이를 해결하기 위한 **최신 최적화 흐름**은 단순히 GPU의 성능을 높이는 물리적 해결책을 넘어, 모델 구조 자체가 연산의 종류와 저장 방식, 데이터의 이동 경로를 설계 단계부터 최적화하여 **전체적인 운영 효율성을 극대화하는 방향**으로 이동

→ 대표적 연구 방향이 연산의 희소성과 저장의 희소성

➡ 희소성의 2대 축

○ 희소성은 제한된 자원 예산(연산·메모리·대역폭·지연) 하에서 모델이 필요한 일부만 선택적으로 사용하도록 만드는 구조적 설계 원리를 의미

- 희소성은 단순히 모델 크기를 줄이는 기술이 아니라 모델이 작동하는 방식(연산 경로와 상태 관리)을 바꾸는 전략
- 즉, 계산할 대상을 줄여 데이터 이동과 저장할 상태를 줄이거나, 저장과 이동의 방식을 바꾸어 병목을 재배치하여 전체 효율을 개선하려는 기술

○ 결과적으로, 트랜스포머의 병목이 연산에서 메모리 용량과 대역폭, 데이터 이동으로 확대되면서, '연산의 희소성'과 '저장의 희소성'이라는 두 가지 축이 등장

- 연산의 희소성은 모든 파라미터와 모든 연산을 토큰마다 수행하지 않고, 전체 파라미터의 일부만 선택적으로 사용하여 GPU 연산 부하를 줄이고 성능을 높이려는 원리
- 저장의 희소성은 모든 상태를 동일한 형태로 동일한 위치에 저장하지 않고, 데이터 자체를 압축하거나 불필요한 정보를 제거해 메모리 낭비를 줄이고 전송 효율을 높이려는 원리

표 3 트랜스포머의 한계 및 해결 방향

구조적 한계	해결 방향
[한계 1] 모든 파라미터 활성화와 연산 비효율 입력과 관계없이 대부분의 파라미터가 동일한 연산 경로를 따라 활성화되므로, 모델이 커질수록 불필요한 연산 부하와 연산을 위한 메모리 접근 비용이 함께 증가	**연산의 희소성(Compute Sparsity)** 전체 파라미터 중 일부만 선택적으로 활성화하여 불필요한 연산 부하를 줄이고 효율 제고 (사례: FFN/MoE 경로의 선택적 활성화)
[한계 2] 파라미터·상태 접근 비용 증가 긴 문맥과 대규모 동시 요청 환경에서는 모델 파라미터와 상태 정보(KV cache)를 반복적으로 읽고 저장·이동해야 하므로, 메모리 용량·대역폭·데이터 이동 비용이 핵심 병목으로 부상	**저장의 희소성(Storage Sparsity)** 파라미터와 KV cache를 압축·계층화하여 메모리 점유율과 데이터 이동 비용을 줄이고 전송 효율 제고 (사례: KV cache 관리, 파라미터 압축, 메모리 계층화)

※ 주: 연산의 희소성과 저장의 희소성은 특정 블록(어텐션, FFN)과 1:1로 대응하는 구분이 아니라, 트랜스포머 전반의 병목을 설명하는 두 개의 상위 설계 축(FFN/MoE는 연산 희소성의 대표 사례로, KV cache는 저장 희소성의 대표 사례로 자주 논의되나, 실제 추론 병목은 파라미터와 상태 접근 비용이 결합된 형태로 발생)

2 연산의 희소성

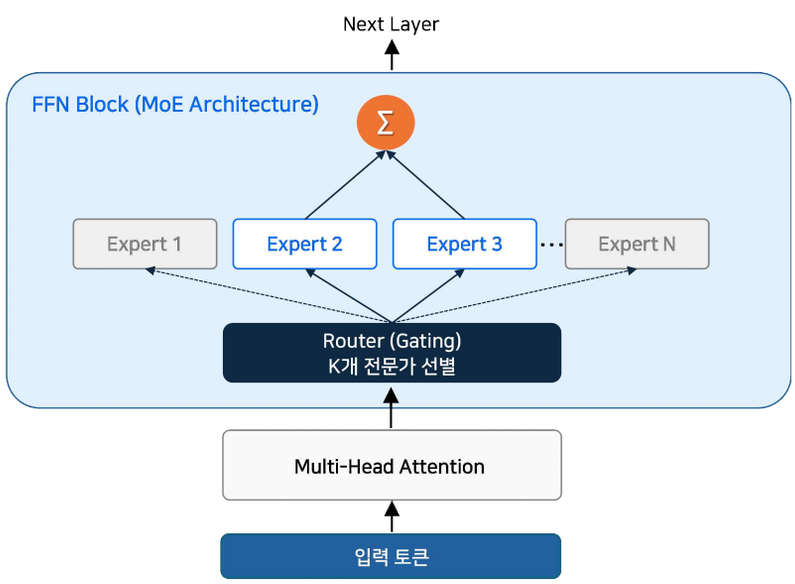
➡ 조건부 계산과 선택적 활성화

- 연산의 희소성은 모델의 전체 파라미터 수를 줄이기보다는 활성화되는 파라미터의 비율을 제어하는 조건부 계산(conditional computation) 설계 개념
  - 즉, '모델 크기 확장'과 '연산 비용 통제'를 분리하는 전략으로 Mixture of Experts(MoE) 계열 알고리즘이 대표적
  - MoE 구조에서는 입력 토큰이 라우터(router)를 통해 일부 전문가 모듈로만 전달되며, 선택된 전문가만 활성화되어 연산 수행

※ MoE는 트랜스포머의 FFN 레이어를 여러 개의 작은 FFN(전문가)과 라우터로 대체

  - 이를 통해 모델 전체 파라미터 수는 크게 늘릴 수 있지만, 토큰당 연산량(FLOPs)은 제한된 수준으로 유지 가능

그림 3 MoE 구조



※ 출처: 저자 작성.

➡ MoE 계열 알고리즘 발전 및 대표 사례

- (GShard[8]) 조건부 연산을 트랜스포머에 도입하여 성능 손실 없이 모델 규모를 확장하는 기반을 마련하고, 대규모 분산 학습의 기술적 가능성을 확인

8) Lepikhin, D., et al.(2020), GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding, arXiv preprint arXiv:2006.16668.

- 자동화된 샤딩(Sharding)[9] 기법을 통해 수천 개의 가속기에서 거대 모델을 학습시키는 표준 라이브러리를 제시하며 고효율 분산 학습 체계의 기반을 구축
- 6천억 개 이상의 파라미터를 가진 모델을 안정적으로 구현함으로써, 학습 자원의 한계를 기술적으로 극복할 수 있는 희소 아키텍처의 연산 효용성을 입증

- (Switch Transformer[10]) GShard의 복잡한 라우팅을 **'단일 전문가'** 방식으로 과감히 단순화하여, 연산 오버헤드를 줄이고 모델 규모를 1.6조 개까지 확장
  - 복잡했던 하이퍼파라미터 튜닝을 안정화하고 통신 비용을 획기적으로 낮추어, 현대적 초거대 MoE 모델이 실질적으로 작동할 수 있는 최적화 기준을 제시
  - LLM 경쟁의 축을 단순 밀집(Dense) 모델에서 효율적인 희소(Sparse) 모델로 전환한 촉매제 역할을 제공해, 이후 다양한 MoE 변종 모델의 탄생을 견인

- (Mixtral 8x7B[11]) 8개의 **전문가 중 2개만 활성화**하여 실제 연산량을 12B 수준으로 억제하면서도 성능은 수배 큰 모델을 압도, 오픈소스 MoE 모델의 상용화 가능성을 제고
  - 적은 연산 자원으로 최첨단 성능을 달성하여 기업용 고효율 AI 모델의 실질적 표준을 제시하고 추론 비용 최적화와 성능 유지라는 상충하는 두 과제를 동시에 해결
  - 오픈소스 생태계에 고성능 MoE 모델을 배포함으로써 중소규모 기업과 연구소에서도 최적화된 희소 아키텍처 기반의 독자적인 서비스 구축 환경을 조성

- (DeepSeek-V3[12]) **세분화된 전문가** 구조와 공유 전문가 시스템을 도입해 지식 습득 효율을 극대화함으로써 적은 비용으로도 최상위권 모델과 대등한 성능 달성
  - 특히 MLA(Multi-head Latent Attention)와 MoE의 결합은 데이터 전송과 연산 과정에서 발생하는 병목 현상을 해결하는 핵심으로, 이러한 희소 아키텍처는 하드웨어 성능을 알고리즘에 최적화하여 구현 효율을 극대화
  - 추가적인 제약 조건 없이 전문가 간의 작업량을 고르게 분산하여 학습의 효율과 안정성을 동시에 확보하여 비용 효율적 AI 개발 전략을 구체화

➡ 연산 희소성에서 저장 희소성으로 확장

- 연산의 희소성은 연산 대비 성능 효율은 달성하였으나, 토큰별 상태 정보인 **KV cache**의 선형적 증가는 여전히 추론 시스템의 물리적 확장성을 저해하는 **핵심 병목**으로 작용

9) 샤딩(sharding)이란 하나의 거대한 데이터베이스나 네트워크 시스템을 여러 개의 작은 조각으로 나누어 분산 저장하여 관리하는 기술
10) Fedus, W., et al.(2022), Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, Journal of Machine Learning Research.
11) Jiang, A. Q., et al.(2024), Mixtral of Experts, arXiv preprint arXiv:2401.04088.
12) DeepSeek-AI(2024), DeepSeek-V3 Technical Report.

- 라우팅 과정의 통신 오버헤드와 메모리 대역폭 정체가 맞물리며 실질적인 서비스 지연 시간 개선 효과로 이어지지 못하는 구조적 한계 노출

※ 통신 오버헤드와 시스템 복잡성 증가는 특히 멀티 GPU · 멀티 노드 환경에서는 expert 간 부하 불균형이 성능 저하로 직결

- 무엇보다 추론 단계에서 병목이 연산보다 KV cache 접근, 메모리 이동, 상태 관리에 좌우되면서[13] 연산 중심의 구조 변화는 결과적으로 저장의 희소성으로 확장

표 4 연산 희소성의 특징과 한계

구분	내용
핵심 개념	계산 대상의 선택
기본 원리	조건부 연산, 선택적 파라미터 활성화
대표 사례	MoE 계열 알고리즘(GShard, Switch Transformer 등)
주요 장점	연산 비용 통제, 모델 규모 확장
구조적 한계	상태(KV cache), 메모리 병목 미해결

※ 출처: 저자 작성.

13) Kwon, W. et al.(2023), Efficient Memory Management for Large Language Model Serving with PagedAttention, arXiv:2309.06180.

3 저장의 희소성

➡ 저장 희소성 부상과 다차원적 접근

○ 초거대 모델의 연산 비용 통제를 넘어 추론 단계의 실질적 병목인 상태 유지와 메모리 용량·대역폭 문제를 해결하기 위한 저장 희소성의 기술 중요성 급부상

- 이는 기억 대상과 저장 위치·형태의 선택 문제를 모델 구조와 시스템 설계의 중심에 두어 데이터 이동 효율성을 극대화하는 구조적 재설계
- 단순 연산량(FLOPs) 저감만으로는 해결하기 어려운 메모리 중심 워크로드의 한계를 극복하기 위해, 상태 관리와 저장 방식에 대한 알고리즘 차원의 재설계가 필요

〈 기억이 병목이 되는 시대 〉

- 트랜스포머 기반 생성 모델은 추론 과정에서 **과거 토큰의 Key/Value(KV cache)**를 저장해 다음 토큰 생성에 재사용
- KV cache는 컨텍스트 길이가 늘어날수록 선형적으로 증가하고, 동시 요청이 늘어날수록 요청 수만큼 복제되어 증가하며, 디코딩 단계에서 모든 토큰마다 반복 참조
- 결과적으로, 추론 과정에서는 GPU 연산 능력뿐 아니라 HBM 용량, 메모리 대역폭, 그리고 캐시 관리 방식이 성능과 비용을 결정하는 핵심 요인

→ 저장의 희소성을 위한 다차원적 접근 부상

○ 저장의 희소성은 다양한 기술을 통해 구현될 수 있으며, 서로 다른 메모리 병목 지점을 겨냥한 세 가지 접근으로 구분

- **① 상태 관리·표현 최적화**: KV cache 구조는 유지하되 낭비와 파편화를 최소화하여 물리적 배치를 효율화하고, 필요시 KV 데이터 자체를 압축해 저장 비용과 전송 부담을 함께 줄이는 접근
- **② 어텐션 구조 재설계**: KV cache의 형태와 크기 자체를 변형하여 상태 저장 필요량 자체를 줄임으로써 메모리 점유의 물리적 한계치를 낮추는 아키텍처 혁신 병행
- **③ 조건부 메모리**: 정적 메모리 호출(Lookup) 기반의 외부 저장소를 활용하여 모델의 내부 상태 유지 부담을 줄이고 지식 접근의 효율성과 정확성을 동시 제고
- 요약하면, 저장 희소성은 소프트웨어 운영의 혁신에서 시작하여, 재사용 가능한 상태의 압축 저장과 활성 상태의 표현 압축으로 확장되며, 나아가 어텐션 구조 재설계와 lookup 방식 내장으로 발전

표 5 저장 희소성의 세 가지 접근 비교

구분	해결 과제	접근 방법	특징 및 장점
① 상태 관리·표현 최적화	KV cache 파편화·낭비	블록·페이지 단위 관리, KV cache 양자화·압축	메모리 낭비 대폭 감소, 적용 용이
② 어텐션 구조 재설계	KV cache 규모	Latent KV, 선형· 하이브리드 어텐션	긴 컨텍스트에서 캐시·대역폭 부담 감소
③ 조건부 메모리	지식 조회 계산 낭비	Lookup 기반 정적 메모리	지식 접근을 계산에서 저장/조회로 부분 이전

※ 출처: 저자 작성.

접근 1: 상태 관리·표현 최적화

- 추론 단계의 병목이 연산(FLOPs)에서 메모리 및 데이터 이동으로 전이됨에 따라, GPU 메모리의 상당 부분을 점유하는 KV cache의 효율적 관리가 필수적
  - 요청마다 길이가 다르고 동적으로 증가하는 KV cache의 특성상, 기존의 연속 메모리 할당 방식은 심각한 메모리 파편화와 자원 낭비를 초래
  - 특히 대규모 동시 요청 처리가 요구되는 서빙(serving) 환경에서 불필요한 메모리 점유는 시스템 전체의 확장성을 제약하고 운영 비용을 상승시키는 핵심 요인으로 작용
  - 상태 관리·표현 최적화는 이를 위해 같은 KV cache를 저장해도 낭비 없이 더 잘 관리하는 운영 최적화와, 같은 KV 구조를 더 작은 표현으로 저장하는 압축 최적화를 모두 포함하는 접근으로 이해하는 것이 바람직
- (PagedAttention[14]) 운영체제의 가상 메모리 및 페이징 개념을 AI 추론에 도입하여, KV cache를 연속된 공간이 아닌 작은 블록 단위로 나눠 비연속적으로 관리
  - 기존 방식은 KV 캐시의 크기가 요청별로 다르고 동적으로 변화함에 따라 메모리 파편화와 중복 점유를 초래해 동시 작업 규모와 처리량을 제한하는 병목으로 작용
  - PagedAttention은 이러한 문제를 해결하기 위해 KV cache 메모리 낭비를 거의 제로 수준으로 줄이고, 요청 간 KV cache 공유를 통해 메모리 사용을 추가로 절감하도록 설계됨
  - 논문은 동일 지연시간 기준에서 기존 시스템 대비 처리량을 2~4배 높였다고 보고하며, 특히 긴 문맥과 복잡한 디코딩일수록 효과가 커진다고 설명
  - 주목할 점은 **어텐션 구조의 수학적 변경 없이도**, 상태 저장과 관리 방식의 혁신만으로 저장 희소성 문제를 크게 완화했다는 데 있음

14) Kwon, W. et al.(2023), Efficient Memory Management for Large Language Model Serving with PagedAttention, arXiv:2309.06180.

- (KVTC[15]) NVIDIA가 제안한 KV Cache Transform Coding은, 이미 생성된 KV cache를 더 작게 저장해 재사용 가능한 상태의 보관·복원 비용을 줄이는 방식
  - NVIDIA는 연속 대화나 반복적인 코드 편집처럼 재사용 지시문이 많은 환경에서, 장기 잔류 KV 캐시가 GPU 메모리를 과도하게 점유하는 현상을 해결 과제로 설정
  - 이를 위해 KVTC는 고전적 미디어 압축 방식과 유사하게 PCA 기반 특징 분리, 적응형 양자화, 엔트로피 코딩을 결합해 KV cache를 GPU 내외부에 분리해 압축 저장
  - KVTC는 최대 20배 압축, 특정 경우에는 40배 이상 압축을 달성하면서도 추론과 긴 문맥 정확도를 유지했고, TTFT(Time-To-First-Token)를 최대 8배까지 개선
  - 이는 저장 희소성 기술이 단순히 현재 사용 중인 캐시의 효율적 분할 관리에 그치지 않고, 향후 **재사용될 캐시의 저장 용량을 최소화**하는 방향으로 고도화되고 있음을 시사
- (TurboQuant[16]) 구글은 기존 어텐션 구조를 유지한 채, KV cache를 획기적으로 압축해 메모리 사용량과 데이터 이동량을 동시에 줄이는 표현 최적화 기법을 제안
  - TurboQuant는 KV cache 압축과 벡터 검색을 동시에 지원하는 압축 방법으로, 고차원 벡터 압축 과정에서 문제였던 추가 메모리 오버헤드를 최소화
  - 학습이나 추가 파인튜닝 없이 KV cache를 3비트 수준까지 압축하면서도 정확도 저하 없이 동작할 수 있고, 일부 실험에서 **KV 메모리 크기는 최소 6배** 줄이고, H100 GPU 기준 어텐션(Attention) **연산 속도는 최대 8배**까지 높일 수 있음을 확인
  - 이는 저장 희소성 기술이 단순한 메모리 낭비 제거를 넘어, 데이터 자체를 가볍고 빠르게 이동할 수 있는 최적화된 형태로 재설계하는 방향으로 진화하고 있음을 시사

〈 TurboQuant: KV cache 표현 압축을 통한 저장 희소성 혁신 〉

- 구조는 크게 벡터 정보를 압축하기 쉬운 형태로 바꾼 뒤, 남는 작은 오차를 별도로 보정해 압축률은 높이면서도 정확도 저하는 최소화하는 두 단계로 구성
- 1단계: 고차원 벡터를 무작위 회전한 뒤 각 부분을 개별적으로 압축하는 PolarQuant 관련 아이디어를 통해 대부분의 압축 효과를 확보
- 2단계: 1차 압축 이후 남는 작은 오차에 대해서는 QJL(Quantized Johnson-Lindenstrauss) 을 1비트 수준으로 적용하여, attention score 계산의 편향을 줄이고 정확도를 유지
- 이는 저장 희소성이 단순한 캐시 관리 최적화를 넘어, KV 상태를 더 작고 계산 친화적인 표현으로 재구성하는 방향으로 확장되고 있음을 보여줌

→ TurboQuant: 활성 KV cache의 압축·계산 효율 개선에 중점 vs. KVTC: 재활용 가능한 KV cache의 압축 저장과 복원 효율 개선에 중점

15) Konrad Staniszewski et al.(2025), KV Cache Transform Coding for Compact Storage in LLM Inference, arXiv:2511.01815v2.
16) Amir Zandieh et al.(2026.3.24.), TurboQuant: Redefining AI efficiency with extreme compression, Google Blog.

접근 2: 어텐션 구조 재설계

- 상태 관리 최적화가 KV cache를 효율적으로 운영하는 기술이라면, 구조 재설계는 긴 문맥과 대량 추론을 위해 저장 부피 자체를 근본적으로 줄이는 방식
  - 즉, 어텐션 메커니즘 자체를 재설계함으로써 모델 크기 확대에 따른 캐시 폭증 문제를 해결하고, 하드웨어의 메모리 대역폭 한계를 돌파할 수 있는 기술적 경로 확보
  - 구조 재설계를 위해, 고정밀 연산이 필요한 범용 모델의 저장 공간 최적화 기법으로 MLA(Multi-head Latent Attention)/TransMLA와, 실시간 응답이 강조되는 장문 서비스 최적화 기법으로 Kimi Linear가 대표적
- (MLA[17]) KV 레이어에서 저랭크(low-rank)[18] 구조를 활용해 명시적 KV cache를 고도로 압축된 잠재 상태(Latent state)로 변환하여 저장의 효율화 추구
  - 즉, 전통적인 멀티헤드 어텐션 대비 모델의 추론 정확도는 유지하면서 KV cache의 물리적 크기를 대폭 축소하여, 메모리 대역폭 병목 문제를 근본적으로 해결하려는 시도
  - 또한, TransMLA[19]에서는 대규모 신규 학습 없이도 MLA를 기존 모델(GQA[20] 등)에 빠르게 이식하거나 변환하는 방법을 제시
- (Kimi Linear[21]) Moonshot AI는 KV cache 의존도를 낮추고 거대 분량의 문서 처리 속도를 높일 수 있는 KDA(Kimi Delta Attention)이라는 정보 요약 기술을 제안
  - 모든 데이터를 개별 대조하는 대신, 정보를 요약된 상태로 누적 업데이트하는 선형 연산을 수행하여, 100만 토큰 이상의 **초장문 환경에서도 캐시 사용량을 최대 75% 절감**
  - 알고리즘 혁신을 모델 구동 플랫폼(서빙 엔진[22]) 등 기존 서비스 인프라에 교체(Drop-in[23]) 가능한 형태로 배포하여, 실제 서비스 현장에 신속하게 적용 가능
  - 이는 저장 희소성 문제가 알고리즘 영역을 넘어 알고리즘×소프트웨어×운영 생태계 영역으로 확장되어야 함을 시사

17) DeepSeek-AI(2025), Multi-head Latent Attention: Scaling KV Cache for Efficient Inference, arXiv:2502.07864v2.
18) 행렬의 데이터 간 상관관계를 활용하여 원래의 고차원 데이터를 훨씬 낮은 차원의 행렬 곱으로 분해·압축함으로써, 정보 손실을 최소화하면서 연산량과 저장 공간을 줄이는 수학적 기법
19) Meng, F. et al.(2025), TransMLA: Multi-Head Latent Attention Is All You Need, arXiv:2502.07864.
20) GQA(Grouped-Query Attention)은 여러 개의 쿼리(Query) 헤드가 하나의 Key & Value(KV) 헤드를 공유하도록 그룹화하는 기술로, 기존 멀티헤드 어텐션(MHA)의 성능과 멀티쿼리 어텐션(MQA)의 효율성 사이의 절충안으로 널리 사용되는 방식
21) Moonshot AI(2025), Kimi Linear: An Expressive, Efficient Attention Architecture, arXiv:2510.26692v2.
22) 학습된 AI 모델을 실제 서비스 환경에서 구동시키고 사용자의 요청에 실시간으로 답변을 생성·제공하는 핵심 소프트웨어
23) 기존 시스템의 전체 설계를 바꿀 필요 없이, 특정 부품을 갈아 끼우듯 즉시 교체하여 바로 사용할 수 있는 방식

〈 Kimi Linear: 장문 추론에 특화된 어텐션 구조 혁신[24] 〉

- Kimi Linear 논문은 단순히 속도를 높인 것을 넘어, 가장 큰 모델이 아니라 가장 효율적인 모델이 우위를 결정한다는 주장으로, AI 모델 설계 방식을 바꿀 수 있음에 주목
- 두 가지 핵심 기술은 중요한 정보만 선별해 기억하는 KDA(Kimi Delta Attention)과, KDA와 Full Attention을 3:1 비율로 구성하는 하이브리드 결합 구조
- KDA는 채널별 망각 게이트(channel-wise gating)를 도입해 **정보의 중요도에 따라 기억 비율을 다르게 처리**
- 하이브리드 구조는 델타 어텐션 레이어 3개가 정보를 처리한 후, 주기적인 풀 어텐션 레이어가 전체 문맥을 조망하는 설계를 통해 처리 속도 향상과 정교한 문맥 파악

- (DeepSeek V4[25]) DeepSeek는 100만 토큰 문맥 처리를 위해 압축 희소 어텐션(CSA)과 고압축 어텐션(HCA)을 결합한 하이브리드 어텐션 구조를 제안
  - 긴 문맥에서 모든 정보를 동일하게 유지하는 대신 오래된 정보는 압축하고 현재 처리에 중요한 정보에 선택적으로 집중하는 방식으로 연산·저장 부담을 동시에 완화
  - 즉, 100만 토큰 문맥 기준 V4-Pro는 이전 V3.2 대비 단일 토큰 추론 FLOPs를 27%, KV cache를 10% 수준으로 낮추고, V4-Flash는 각각 10%, 7% 수준까지 효율화
  - 이는 장문맥 처리의 경쟁력이 긴 입력 중 어떤 정보를 압축하고 어떤 정보를 보존할 것인지 결정하는 어텐션 구조 설계에 의해 좌우됨을 시사

접근 3: 조건부 메모리

- 상태 관리·표현 최적화와 어텐션 구조 재설계가 KV cache의 파편화 제거 및 크기 축소에 주력했다면, 조건부 메모리는 모델의 지식 보유 및 활용 구조 자체를 재설계하는 데 집중
  - 앞의 두 가지 접근이 KV cache의 운영 및 저장 효율화에 관한 것이라면, 조건부 메모리는 지식 처리 자체를 연산 중심에서 조회 중심으로 전환하려는 접근
  - DeepSeek의 Engram이 대표적이며, 이는 저장의 희소성을 단지 KV cache 절감으로 축소하지 않고 모델의 지식 보유와 접근 구조 자체를 재구성하는 방향으로 확장
- (Engram[26]) 연산 희소성에서 MoE가 어떤 전문가를 쓸지 조건부로 결정하듯이, Engram은 조건부 메모리 개념을 도입해 어떤 지식을 메모리에서 조회할지 결정

24) MIT Technology Review(2025), 트랜스포머 이후 가장 중요한 논문이 나왔다, 2025.11.18.
25) DeepSeek-AI(2026), DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence.
26) Xin Cheng et al.(2026), Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models, arXiv: 2601.07372v1.

- 기존 트랜스포머는 단순히 기억에서 꺼낼 수 있는 정보조차 복잡한 연산 과정을 거쳐 KV cache를 반복적으로 생성·저장해야 하므로, 이 과정에서 계산 자원과 메모리 공간을 비효율적으로 낭비하는 구조적 한계가 존재
- Engram은 연산과 기억을 구조적으로 분리해, 연산 중에 필요한 '활성 데이터(KV Cache 등)'는 최소화하고, 방대한 지식은 필요할 때만 인덱싱하여 불러오는 방식
- 조건부 메모리 개념에 기반한 Engram 등장은 첫째, 계산과 기억의 역할을 명확히 분리하여, 지식 접근을 반드시 계산으로 해결해야 한다는 연산 중심 구조를 벗어나는 계기 마련
- 둘째, 저장 희소성을 단지 GPU 메모리(HBM) 내부 최적화가 아니라, 외부 메모리 계층(DRAM 등)과의 유기적 결합으로 시스템 전체 효율을 극대화하는 방향으로 확장

〈 Engram: 조건부 메모리를 통한 기억 구조 혁신 〉

- DeepSeek-V3[27]/R1[28]이 MoE 기반 조건부 연산과 MLA 기반 KV cache 절감을 결합해 연산·저장 효율화를 동시에 보여준 1차 쇼크였다면, Engram은 조건부 기억이라는 새로운 희소성 축을 도입해 **연산과 기억을 분리하는 모델 구조 혁신을 제시한 2차 쇼크로 평가**
- 첫째, Engram은 O(1)[29] 속도의 결정적 조회[30]를 통해 정적인 지식 인출을 담당함으로써, 트랜스포머가 복잡한 추론에만 집중할 수 있도록 연산 효율성을 극대화
- 둘째, 전체 모델 자원(파라미터)에서 MoE와 Engram 사이의 최적 균형점을 찾는 희소성 배분 문제를 공식화(formulation)하여 'U자형 스케일링 법칙' 발견

※ 전체 희소 파라미터[31] 중 20~25%를 Engram(메모리)에, 나머지 75~80%를 MoE(연산)에 할당할 때 모델 성능 최대(손실 최소)

※ Engram에는 주로 '임베딩[32]' 형태의 파라미터를 할당하여 방대한 지식을 저장하게 하고, MoE에는 주로 '연산 가중치' 관련 파라미터를 할당하여 논리적인 사고를 담당

→ Engram: HBM뿐 아니라 대규모 지식 데이터를 상대적으로 저렴한 고용량 메모리(DRAM/SSD)과 분산 배치하는 메모리 계층화를 강화

27) DeepSeek-AI(2024), DeepSeek-V3 Technical Report, arXiv:2412.19437v1.
28) DeepSeek-AI(2025), DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, arXiv:2501.12948v2.
29) 데이터의 규모와 상관없이 지식 인출 소요 시간이 일정(상수시간)함을 의미
30) 복잡한 연산 없이 입력된 단어 조합(N-gram)에 따라 정해진 메모리 주소를 즉시 찾아가는 방식
31) 모델의 전체 파라미터 중에서 특정 입력값을 처리할 때 일부만 사용하여 연산하는 파라미터를 의미
32) 방대한 지식을 컴퓨터가 이해할 수 있는 숫자 형태의 벡터로 변환하여 저장한 값

Ⅲ 희소성이 강제하는 시스템·인프라 구조 전환

1 희소성이 만드는 설계 변화

➡ 설계 변화의 방향과 범위

- (방향) 트랜스포머 기반 AI 알고리즘의 구조적 병목을 해결하기 위해 등장한 **두 가지 희소성** 개념은 서로 다른 병목을 겨냥하지만, 추론 운영 관점에서는 **함께 최적화**되어야 하는 설계 축으로 접근할 필요
  - 연산의 희소성은 주로 토큰당 활성화되는 연산 경로와 파라미터를 줄이는 접근이며, 저장의 희소성은 문맥 길이와 동시 요청 증가에 따라 커지는 상태 정보의 저장·이동·접근 부담을 줄이는 접근
  - 즉, 두 축은 동일한 문제를 해결하는 대체 관계라기보다, 추론 서비스의 전체 비용·지연시간·처리량을 결정하는 서로 다른 병목을 분담

※ 연산을 줄이더라도 KV cache 등 상태 유지 비용이 크면 HBM 점유와 데이터 이동 병목으로 전체 처리량이 제한되고, 반대로 상태를 줄이더라도 토큰마다 과도한 연산이 수행되면 토큰당 비용과 지연시간 문제가 다시 발생하기 때문

  - 따라서 최근 알고리즘 구조 변화의 핵심은 연산 희소성에 기반한 선택적 활성화와 저장 희소성에 기반한 상태 관리·표현 최적화를 별개의 기술로 나열하는 데 있지 않고, 계산·저장·이동 비용을 하나의 추론 운영 문제로 통합해 줄이는 방향으로 이동

※ 특히 Engram 연구는 일부 지식 처리를 매번 계산하는 방식에서 필요한 정보를 조회하는 방식으로 전환할 수 있음을 보여주며, 연산·저장·조회 간 기능 배분이 모델 성능과 운영 효율을 좌우할 수 있음을 제시

  - 이러한 흐름은 트랜스포머 구조가 더 이상 단일한 형태로 고정되지 않고, 연산·저장·조회 방식이 조합되는 방향으로 진화하고 있음을 시사

- (범위) 희소성은 단순한 국소적 최적화 기법이 아니라 AI 알고리즘 구조 전반을 재편하는 설계 원리로 작용하며, 이제는 **알고리즘 차원을 넘어 시스템 설계와 인프라 구성에까지 영향력의 범위를 확장**
  - 희소성 확산은 알고리즘이 더 이상 하드웨어와 시스템 환경을 추상화한 채 독립적으로 설계될 수 없음을 의미
  - 연산과 저장에 대한 선택은 메모리 계층의 활용 방식, 연산 자원의 구성 형태, 운영 구조의 효율성을 정의하는 핵심 결정 요인
  - 다시 말해, 희소성은 알고리즘이 시스템과 인프라 환경을 가정하는 수준을 넘어, 시스템 구조를 전제로 설계되도록 강제하는 요인으로 작용

추론 중심 AI 전환의 의미

- 알고리즘이 시스템 구조를 전제로 설계된다는 점은 단순한 구현 방식의 문제를 넘어, AI 경쟁의 축이 학습에서 효율적 추론으로 완전히 이동했음을 보여주는 결정적 지표
  - 학습에서 추론으로 AI 경쟁 축의 이동은 시스템과 인프라가 더 이상 보조적 요소가 아니라, 알고리즘 성능과 경제성을 결정하는 핵심 변수로 작용
  - AI가 연구 단계를 넘어 검색·행정 등 상시 가동되는 서비스 체계로 자리 잡으면서, 데이터센터 수요 구조 또한 실시간 추론 대응 중심으로 빠르게 전환
  - 이에 따라 AI 성능 평가 기준이 최고 정확도나 벤치마크 점수보다, **지연시간(latency), 처리량(throughput), 비용(cost)**과 같은 운영 지표로 재정의

※ 과거의 경쟁력이 신속한 모델 학습 역량에 있었다면, 현재는 대규모 요청에 대한 저지연·저비용 기반의 안정적 처리 능력이 기업의 생존을 결정

- 추론 중심 AI 전환은 성능의 척도가 단순 연산 속도(FLOPs) 중심에서 지연시간(Latency), 처리량(Throughput), 비용(Cost) 간의 균형을 최적화하는 운영 지표[33]로 이동
  - 생성형 AI의 추론은 토큰별 반복 연산과 상태 접근이 필수적이므로, 단순 계산 성능보다 메모리 접근 및 스케줄링 효율이 실질적 성능을 좌우
  - 결국 메모리 대역폭은 지연시간을, 용량은 처리량을, 상태 관리 효율은 비용 경쟁력을 결정짓는 핵심적인 통제 변수로 작용

※ 대역폭이 좁으면 지연시간이 늘고, 용량이 부족하면 처리량이 제한되며, 상태 관리가 부실하면 운영 비용이 상승하는 구조적 인과관계 형성

  - 이로 인해, 추론 중심 AI는 단순히 서버를 증설하는 문제를 넘어 자원의 역할 분리와 메모리의 구조적 전환을 요구하는 설계 문제로 확장

33) Latency: 단일 요청이 응답을 받기까지 걸리는 시간, Throughput: 단위 시간당 처리 가능한 요청 또는 토큰 수, Cost: 요청당 또는 토큰당 발생하는 계산·메모리·에너지 비용

2 자원 분리와 메모리 계층화

Prefill(입력 처리 단계)과 Decode(출력 생성 단계) 자원 분리

- 생성형 AI 추론은 프롬프트를 처리하는 Prefill과 토큰을 생성하는 Decode 단계로 구분할 수 있고, 각 단계는 요구되는 자원 특성과 병목 지점이 근본적으로 상이
  - Prefill은 연산 집약적(Compute-bound) 작업으로 TTFT(Time To First Token)를 결정하며, Decode는 메모리/상태 접근(Memory-bound) 위주로 TPOT(Time Per Output Token)에 영향

    ※ TTFT: 첫 토큰 생성 시간, TPOT: 토큰 간 생성 시간
  - Prefill과 Decode를 동일 자원에서 처리할 경우, 한 단계의 부하 변화가 다른 단계의 지연을 악화시키는 구조적 문제 발생

표 6 Prefill과 Decode 비교

구분	Prefill	Decode
주요 기능	사용자가 입력한 문맥 전체를 한 번에 읽고 답변을 위한 KV cache 생성	이전의 모든 내용을 참조하여 다음 단어를 예측하고 문장을 완성
처리 방식	**대규모 병렬(Parallel) 처리**	**순차(Sequential) 처리**
병목 현상	매우 많은 양의 행렬 연산이 성능 좌우 → 연산 중심(Compute-bound)[34]	메모리에서 KV캐시를 읽어오는 속도가 성능 좌우 → 메모리 중심(Memory-bound)
대표 HW	NVIDIA Rubin CPX, 화웨이 Ascend 950PR → 높은 연산 성능에 집중, 비용 효율적 메모리 탑재	NVIDIA Rubin R200, 화웨이 Ascend 950DT → 극도로 높은 메모리 대역폭을 가진 HBM 탑재
성능 지표	요청 후 첫 글자가 나올 때까지 걸리는 시간 → TTFT(Time-To-First-Token)	첫 글자 이후, 나머지 토큰들이 생성되는 평균 시간 → TPOT(Time-Per-Output-Token)
비용 구조	**연산(FLOPS) 비용 중심**	**메모리(대역폭, 용량) 비용 중심**
역할 특화	Prefill 전용 가속기로 연산 극대화	KV cache 오프로딩(Offloading)[35] 최적화

※ 출처: 저자 작성.

- 동일 자원 처리의 구조적 문제를 해결하기 위해 자원을 논리적·공간적으로 나누는 **자원 분리**와 각 자원에 최적화된 임무를 부여하는 **역할 특화**가 핵심으로 부상
  - 분산 추론(Distributed Inference)은 단일 장치의 저장과 연산 한계를 극복하기 위해 모델과 워크로드(Workload)를 여러 장치에 분산하여 처리하는 기술로, Prefill과 Decode 구조 분리가 대표적인 고도화 사례

34) 전통적으로 Prefill은 대규모 병렬 연산 특성상 연산 중심(Compute-bound)으로 분류되나, 에이전틱 AI 확산에 따른 컨텍스트의 극단적 거대화로 인해, 최근 프리필 단계에서도 대규모 입력 데이터 및 KV 캐시를 지연 없이 공급하기 위한 '메모리 대역폭 및 캐시 관리'가 새로운 핵심 병목으로 부상 중임

35) GPU 메모리가 부족할 때, 당장 필요하지 않은 모델 가중치나 KV cache를 DRAM 또는 SSD로 옮겨두는 것

- 자원 분리는 prefill과 decode를 서로 다른 자원(실행 엔진)에서 처리함으로써, 단계 간 간섭을 제거하고 각 단계의 성능을 독립적으로 관리할 수 있도록 하는 설계 방식
- 역할 특화는 단순히 자원을 나누는 것을 넘어, 각 단계가 수행하는 기능적 목적에 맞춰 하드웨어 사양과 소프트웨어 최적화 전략을 차별화하는 설계 방식

※ 분산 추론 아키텍처를 실제로 구현한 사례로 Prefill 단계 특화 칩(NVIDIA Rubin CPX, 화웨이 Ascend 950PR)과 Decode 단계 특화 칩(Rubin R200, Ascend 950DT)이 대표적

HBM 단독 구조에서 SRAM-HBM-DRAM-SSD 계층 구조로

○ 추론 단계에서 발생한 지연·처리량·비용을 보다 세밀하게 관리하기 위해 등장한 분리 구조(Prefill & Decode)가 확산될수록 시스템은 또 하나의 근본적인 메모리 제약에 직면

- 추론 과정에서 발생하는 상태(State), 특히 KV cache의 저장·이동·재사용이 추론 성능의 상한을 결정하며 추론에서 메모리가 가장 큰 병목으로 등장
- Prefill & Decode 분리 구조의 확산은 KV cache가 더 이상 단순한 로컬 메모리 객체가 아니라, 저장되고 이동되며 재사용되어야 하는 상태가 되었음을 의미

※ vLLM의 PagedAttention 연구[36]는 요청별 KV cache가 동적으로 생성·확장되는 환경에서 추론 성능이 단순한 연산 속도가 아니라, 메모리 관리 방식에 의해 좌우될 수 있음을 입증

○ 메모리 관리 방식 관점에서, 고대역폭 메모리(HBM) 중심의 전통적 추론 구조는 거대 모델의 긴 문맥 처리와 대규모 동시 요청 환경에서 용량 및 비용의 임계치에 도달

- 긴 문맥 처리 시 KV cache가 급증하여 고가의 HBM 자원을 우선 소진하고 연산 효율을 저하시키는 병목 발생
- 모든 상태 데이터를 HBM에만 유지하는 방식은 높은 단가와 공급 제약으로 인해 AI 서비스의 경제적 지속성 확보에 한계 노출
- 따라서 최근의 인프라 설계는 데이터를 활성도에 따라 최적의 메모리 계층에 배치하는 **'계층형 메모리 아키텍처'**로 진화 중
- 즉, HBM을 유일한 저장 공간으로 사용하는 방식에서 벗어나, 온칩(On-chip) SRAM은 즉시 참조가 필요한 활성 상태를, HBM은 핵심 실행 메모리를, DRAM/GDDR과 SSD는 재사용 가능한 대용량 상태를 담당하는 방향으로 역할 분화가 진행

※ 자원 분리와 역할 특화가 추론 구조를 재편했다면, 메모리 계층화는 그 구조가 실제로 작동하도록 만드는 인프라 차원의 필수 조건

36) Kwon, W. et al.(2023), Efficient Memory Management for Large Language Model Serving with PagedAttention, arXiv:2309.06180.

- 추론 중심 AI 인프라는 데이터 활성 특성에 따라 메모리 계층을 분리하고, 각 계층의 역할을 명확히 정의하여 시스템 효율을 극대화
  - (온칩 SRAM) 연산 장치 내부 또는 바로 인접한 위치에 배치되는 초저지연 메모리로, 즉시 참조가 필요한 최신 활성 KV cache나 작업 집합(working set)을 칩 내부에 상주시켜 데이터 이동을 최소화하고 디코딩 병목을 완화하는 역할을 담당
  - (HBM) 초고속 대역폭을 통해 실시간 토큰 생성에 필요한 모델 가중치와 최신 활성 KV cache를 배치하여 지연시간 최소화에 집중
  - (DRAM/GDDR) HBM 대비 높은 용량과 경제성을 바탕으로, 즉각적인 연산에는 포함되지 않으나 재사용 빈도가 높은 상태 데이터를 보관하는 보조 캐시로 활용

※ CXL(Compute Express Link): DRAM 계층의 물리적 한계를 극복하기 위해 도입된 차세대 인터페이스로, 장치 간 메모리 공유와 확장을 가능하게 하는 핵심 기술

  - (SSD) 초거대 용량을 활용하여 실시간 범위를 벗어난 긴 문맥이나 비활성 KV cache 데이터를 저장함으로써 시스템 전체의 용량 한계 극복

분산 추론 서빙 엔진

- 분산 추론 서빙 시스템은 대규모 모델의 추론 효율 극대화를 위해 **알고리즘 최적화와 메모리 계층화를 연결**하는 핵심 기술
  - 이는 실시간 자원 조율과 캐시 전송 최적화로 지연을 최소화하고 GPU 가동률을 높이는 지능형 오케스트레이션 역할
  - 분산 추론 서빙 엔진은 알고리즘과 인프라(GPU, 메모리 등) 계층 사이에 위치하는 시스템 기술로 알고리즘-시스템-인프라 공동설계를 위한 성능 병목을 해결하고 하드웨어 제약을 소프트웨어로 극복하는 실질적인 통합 실행 환경
  - 주요 기능으로 Prefill & Decode 자원 분리, KV cache 지능형 라우팅, KV cache 오프로딩 및 프리페칭, GPU와 메모리 간 데이터 이동 최적화, 통합 메모리 관리 등

※ DistServe 연구, NVIDIA의 Dynamo, Moonshot AI의 Mooncake 등이 분산 추론 서빙의 대표 기술

〈 오프로딩 & 프리페칭 〉

- 오프로딩(Offloading): 주 메모리(HBM)의 용량 제한이나 병목 현상을 해결하기 위해, 상대적으로 느리지만 용량이 큰 하위 메모리(DRAM, SSD)로 데이터를 이동시키는 기술
- 프리페칭(Prefetching): 오프로딩된 데이터를 연산에 사용하기 직전, 하위 메모리(DRAM, SSD)에서 주 메모리(HBM)로 미리 가져와 전송 지연과 연산 대기 시간을 최소화하는 기술

→ NVIDIA Dynamo는 이러한 오프로딩과 프리페칭을 동적으로 제어하여 메모리 병목 문제를 소프트웨어적으로 해결하는 핵심 기능 제공

3 알고리즘-시스템-인프라 공동설계

➡ 공동설계(Co-design) 필요성 및 개념

- 추론 과정에서는 **알고리즘, 시스템, 인프라**가 서로 강하게 얽혀 있어, 어느 한 부분만 개선해서는 전체 성능을 끌어올리기 어려운 구조
  - 첫째, Prefill과 Decode 단계는 성격이 달라 두 단계를 같은 자원에서 같은 방식으로 처리하면 한 부분을 개선하더라도 다른 부분에서 병목이 발생하기 때문
  - 둘째, KV cache가 저장, 공유, 전송, 오프로딩, 재사용의 대상이 되어 모델 파라미터뿐 아니라 상태(KV cache) 관리가 추론 성능의 핵심 기능이 되기 때문
  - 셋째, 실제 현장에서는 최고 속도보다 공동설계를 통해 운영 기준(GPU 가동률, 초당 처리 토큰 수, 서비스 수준 목표 등)에 맞춘 안정적 서비스 제공이 중요하기 때문
  - 넷째, 추론 AI가 긴 문맥과 다단계 작업을 기본 전제로 발전하고 있어, 앞으로 단계별 역할에 특화된 인프라 공동설계가 필요하기 때문
- 추론 중심 AI에서 공동설계는 알고리즘, 시스템, 인프라를 개별 최적화하지 않고 전체를 최적화하는 방식
  - 알고리즘은 계산과 상태 관리 방식을 바꾸고, 시스템은 요청과 자원을 배치·제어하며, 인프라는 시스템이 실제로 실행할 물리 자원을 제공
  - 따라서 공동설계의 핵심은 어느 한 층의 절대 성능이 아니라, 여러 층을 관통하는 병목을 얼마나 줄일 수 있는가에 있음

※ 알고리즘: '계산·저장 방식의 설계'에 대한 것으로 모델이 결과를 생성하는 내부 방식(어텐션 메커니즘, KV cache 저장 및 재사용 방식, 긴 문맥에서 메모리 낭비 최소화) 결정

※ 시스템: '요청·상태 전달 방식의 설계'에 대한 것으로, 실제 서비스 운용(batching: 요청 처리 방식, prefill & decode 분리, routing, KV cache 이동 및 재사용) 지원

※ 인프라: '실행 자원의 구상과 배치'에 대한 것으로, 시스템이 실행되는 물리 자원(연산 자원: GPU·CPU, 메모리 자원: 온칩 SRAM·HBM·DRAM·SSD, 연결 자원: NVLink·CXL·Ethernet) 제공

➡ 공동설계 방향 및 주요 기술

- 공동설계는 알고리즘이 상태를 잘 만들고, 시스템이 그 상태와 요청을 잘 운영하며, 인프라가 시스템 운영 방식을 비용 효율적으로 제공하는 구조를 기반으로 통합 최적화
  - (알고리즘) 긴 문맥 추론에서는 모델 파라미터보다도 KV cache 부담이 빠르게 증가하므로 연산량뿐 아니라 상태량을 줄이는 방향이 중요
  - (시스템) 요청에 따라 Prefill과 Decode를 분리하고, 요청 흐름의 단계별 KV cache 이동과 재사용 등 실행 제어 방식이 중요

– (인프라) 단일 GPU 성능 경쟁보다, 단계 특화 GPU와 메모리 계층, 네트워크를 함께 설계하는 방향이 중요

표 7 공동설계 주요 기술

구분	주요 내용
알고리즘	• KVTC[37] : 재사용 KV cache의 저장 용량을 최대 20배 압축해 TTFT(Time-To-First-Token)를 최대 8배까지 개선 • TurboQuant[38] : KV cache 압축과 벡터 검색을 동시에 지원하는 압축 방법으로 KV 메모리 크기는 최소 6배 줄이고, 어텐션(Attention) 연산 속도는 최대 8배까지 향상 • Kimi Linear[39] : 중요한 정보만 선별해 기억하는 방식으로, 1M 토큰 문맥 기준으로 기존 MLA 대비 KV cache 사용량 최대 75% 감소, 디코딩 처리량(TPOT 기준) 최대 6배 개선 • Engram[40] : 지식 조회(Lookup) 기반 조건부 메모리 개념을 도입해 순수 MoE 구조보다 MoE와 메모리를 함께 배분하는 방식이 더 효율적임을 제시 • MoE[41] : DeepSeek-V3는 6,710억 개 파라미터를 가진 언어 모델임에도 불구하고, 실제 추론 시에는 토큰당 약 370억 개의 파라미터만 활성화해 비용 효율을 극대화
시스템	• vLLM[42] : PagedAttention을 통해 KV cache를 블록 단위로 관리해, 같은 지연 수준에서 기존 대비 2~4배 처리량 향상 및 메모리 낭비를 거의 제로 수준으로 감소 • Splitwise[43] : Prefill과 Decode를 서로 다른 하드웨어를 선택하는 구조로, 기존 대비 최대 1.4배 높은 처리량을 20% 낮은 비용으로 달성 • DistServe[44] : Prefill과 Decode를 서로 다른 GPU에 분리 배치해, 기존 대비 최대 7.4배 높은 처리량, 12.6배 더 엄격한 서비스 목표 수준 달성, 전체 요청의 90% 이상을 지연 기준 내 처리 • Mooncake[45] : Prefill & Decode 분리와 CPU·DRAM·SSD 자원을 활용해 분산 KV cache를 운영해, 유효 요청 수용 능력을 59~498% 향상, Prefill 계산 시간을 최대 48% 감소 • NVIDIA Dynamo[46] : Prefill & Decode 분리와 CPU·DRAM·SSD 다계층 메모리 구조를 논문 수준을 넘어 실제 운영 플랫폼 수준으로 발전시켜 상용화
인프라	• NVIDIA Vera Rubin CPX[47] : Prefill 특화 GPU로 HBM 대신 비용 효율적인 GDDR7 메모리를 사용해 GB당 메모리 비용을 50% 이상 절감 • HBM4[48] : 대규모 추론에서 필요한 메모리 대역폭과 전력 효율을 동시에 높이는 핵심 인프라 기술로 SK hynix는 HBM3E 대비 대역폭 2배, 전력 효율 40% 이상 개선 제시 • CXL 4.0[49] : 대역폭을 64GT/s에서 128GT/s로 높이고 CPU·GPU·메모리를 더 유연하게 연결해 메모리 확장성과 신뢰성 기능을 강화한 차세대 인터커넥트 기술

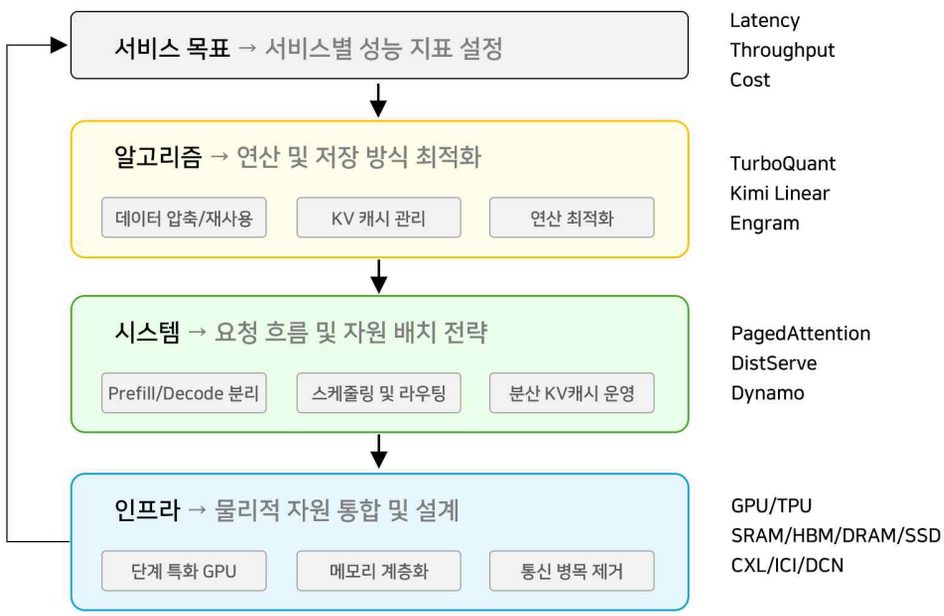
※ 출처: 저자 작성.

37) Konrad Staniszewski et al.(2025), KV Cache Transform Coding for Compact Storage in LLM Inference, arXiv:2511.01815v2.
38) Amir Zandieh et al.(2026.3.24.), TurboQuant: Redefining AI efficiency with extreme compression, Google Blog.
39) Moonshot AI(2025), Kimi Linear: An Expressive, Efficient Attention Architecture, arXiv:2510.26692v2.
40) Xin Cheng et al(2026), Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models, arXiv: 2601.07372v1.
41) DeepSeek-AI(2024), DeepSeek-V3 Technical Report, arXiv:2412.19437v1.
42) Kwon, W., et al.(2023), Efficient Memory Management for Large Language Model Serving with PagedAttention, arXiv:2309.06180v1.

공동설계 구조

○ 공동설계는 서비스 목표 → 알고리즘 → 시스템 → 인프라 흐름으로 이를 최적화하고 지속적으로 개선하는 구조

그림 4 공동설계 구조



※ 출처: 저자 작성.

- 실제 공동설계는 각 층이 독립적으로 순차 설계되는 방식이 아니라 서로의 제약과 성능 특성을 반영하며 함께 조정
- 즉, **서비스 목표가 알고리즘·시스템·인프라를 결정**하는 것처럼 보이나, 반대로 사용 가능한 **메모리 구조, 통신 방식, 가속기 특성도 알고리즘 선택과 시스템 설계에 영향**

※ 서비스 목표: 지연, 처리량, 비용, 안정성 기준을 정하는 것으로 서비스별 기준이 상이

43) Patel, P., et al.(2024), Splitwise: Efficient Generative LLM Inference Using Phase Separation, arXiv:2311.18677v2.
44) Yinmin Zhong et al.(2024), DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving, arXiv:2401.09670v3.
45) Ruoyu Qin et al.(2025), MOONCAKE: Trading More Storage for Less Computation - A KVCache-centric Disaggregated Architecture for Serving LLM Chatbot, FAST 2025.
46) NVIDIA Document Dynamo, https://docs.nvidia.com/dynamo/latest/user-guides/kv-cache-aware-routing.
47) SemiAnalysis(2025.9.11.), Another Giant Leap: The Rubin CPX Specialized Accelerator & Rack.
48) SK hynix(2025), Completes World's First HBM4 Development and Readies Mass Production.
49) Compute Express Link, https://computeexpresslink.org/wp-content/uploads/2025/11/CXL_4.0-Specification-Release_FINAL_Website-Copy.pdf

- 실제 최근 **추론 중심 AI 알고리즘**은 더 이상 독립 변수처럼 다루기 어렵고, 메모리 계층, 데이터 이동 비용, 통신 대역폭, 서빙 방식 등 **시스템·인프라 조건**과 함께 설계되는 방향으로 이동 중
- 따라서 공동설계는 일방향 최적화가 아니라, 여러 층이 상호작용하며 반복적으로 조정되는 전체 최적화 과정으로 이해함이 바람직

공동설계 효과

- 공동설계의 효과는 단순히 모델 구동 속도 개선을 넘어 지연시간(Latency)의 안정적 관리와 처리량(Throughput) 극대화, 시스템 최적화를 통한 운영 비용(Cost) 절감 지향

표 8 공동설계 효과 분석

구분	핵심 효과	관련 내용
지연시간 (Latency)	TTFT와 TPOT를 분리 관리해 지연 목표를 안정적으로 달성	- DistServe: 90% 이상 지연 목표 충족 - Splitwise: 두 번째 토큰 지연 오버헤드 16.5% 감소 (기존 하나의 GPU 처리 64% 대비)
처리량 (Throughput)	동일 인프라에서 더 많은 요청을 처리	- vLLM: 동일 지연시간에서 기존 대비 2~4배 향상 - Splitwise: 기존 대비 20% 낮은 비용으로 최대 1.4배 처리 또는 동일 비용·전력에서 2.35배 처리 - DistServe: 기존 대비 최대 7.4배 처리(동시에 90% 이상 지연 목표 충족) - Mooncake: 유효 요청 처리 59~498% 향상
비용 (Cost)	질문당 비용 및 메모리·GPU 자원 낭비 감소	- Splitwise: 20% 낮은 비용으로 1.4배 높게 처리 - DistServe: 최대 4.48배 낮은 질의당 비용 - Mooncake: 캐시 히트율 최대 2.36배 향상 및 Prefill 연산 시간 최대 48% 절감으로 HBM·GPU 의존 감소

※ 출처: 저자 작성.

① 지연시간(Latency)

- DistServe[50]는 Prefill과 Decode를 서로 다른 GPU 자원에 분리 배치해 전체 요청의 90% 이상을 지연 목표 충족
- Splitwise[51]는 Prefill & Decode 단계 분리를 통해 기존 하나의 GPU에서 순차적 처리 대비, 두 번째 토큰 지연 증가를 64%에서 16.5% 수준으로 감소

50) Yinmin Zhong et al.(2024), DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving, arXiv:2401.09670v3.
51) Patel, P., et al.(2024), Splitwise: Efficient Generative LLM Inference Using Phase Separation, arXiv:2311.18677v2.

② 처리량(Throughput)

- vLLM[52]은 PagedAttention과 KV cache 블록 관리 방식을 통해 같은 수준의 지연 시간에서 기존 최신 시스템 대비 2~4배 처리량 향상을 제시, 알고리즘·시스템 공동설계가 배치 크기와 실제 요청 처리량을 직접 바꿀 수 있음을 입증
- Splitwise는 Prefill과 Decode 특화 GPU를 선택해 기존 대비 최대 1.4배 높은 처리량을 20% 낮은 비용으로 달성(또는 같은 비용·전력 예산에서 2.35배 더 높은 처리)
- DistServe는 90% 이상 지연 목표를 만족하면서 기존 대비 최대 7.4배 더 많은 요청을 처리(또는 12.6배 더 엄격한 서비스 목표 수준 달성)
- Mooncake[53]는 실제 서비스 환경의 운영 데이터 기준으로 유효 요청 수용 능력을 59~498% 향상 확인

③ 비용(Cost)

- Splitwise는 1.4배 높은 처리량을 20% 낮은 비용으로 달성하거나, 같은 비용·전력 예산에서 2.35배 높은 처리량을 제시
- DistServe는 사용자 체감 품질 기준 하에 GPU 한 대가 처리하는 양을 늘려 질의당 소요 비용을 최대 4.48배 감소
- Mooncake: 캐시 히트율 최대 2.36배 향상하고 Prefill 연산 시간을 최대 48% 절감해 HBM과 GPU 의존도를 감소

〈 NVIDIA 공동설계 혁신 사례 〉

- GTC 2026 Keynote[54]에서 젠슨 황(Jensen Huang)은 'NVIDIA Extreme Co-Design Revolutionized Token Cost'를 소개하며 자사의 혁신적인 공동설계 기술을 강조
- NVIDIA가 제시한 공동설계는 칩 단위를 넘어 알고리즘부터 인프라까지 전체 스택을 동시에 설계함으로써 토큰 생산성을 극대화
- 즉, NVFP4, Multi-Token Prediction, Dynamo, Rubin/LPU 등 알고리즘·런타임·네트워크·칩·메모리·랙 설계까지 동시에 최적화
- 특히, Rubin/LPU 간 네트워크 병목을 해소하기 위해 Spectrum-X라는 CPO(Co-Packaged Optics) 광학 기술을 통해 기존 플러그형 광 모듈 대비 3.5배 전력 효율 향상

→ ① NVIDIA는 공동설계를 통해 실제 서비스 운영 비용(**토큰 비용**)을 경쟁사 대비 **35배 절감**
② **전력당 생산되는 토큰의 수**는 경쟁사 및 이전 세대(H200) 대비 최대 **50배 성능 향상**

52) Kwon, W. et al.(2023), Efficient Memory Management for Large Language Model Serving with PagedAttention, arXiv:2309.06180.
53) Ruoyu Qin et al.(2025), MOONCAKE: Trading More Storage for Less Computation - A KVCache-centric Disaggregated Architecture for Serving LLM Chatbot, FAST 2025.
54) NVIDIA(2026), Watch Jensen Huang's GTC 2026 Keynote: On Demand.

Ⅳ 기술패권 경쟁의 재편과 전략적 선택

1 AI 가치사슬 재편과 산업구조 변화

AI 가치사슬 이동: 모델 규모에서 운영 구조로

- 추론 중심 AI로의 전환은 모델 파라미터 수와 학습 데이터 규모를 앞세운 모델 규모 중심의 AI 경쟁의 중심축을 운영 구조로 빠르게 이동
  - 모델 규모가 커질수록 학습 비용뿐 아니라 추론 비용이 기하급수적으로 증가하며, 서비스 확산 단계에서는 이 비용이 가장 큰 제약 요인으로 작용
  - 특히 대규모 서비스 환경에서는 모델 성능의 미세한 향상보다, 지연시간(latency), 처리량(throughput), 비용(cost)이 실제 경쟁력을 좌우
  - 이는 모델 규모 확대가 더 이상 일방적인 경쟁 우위를 보장하지 않으며, 일정 수준 이후에는 오히려 운영 부담을 증폭시킬 수 있음을 의미
- 기존 AI 가치사슬이 모델의 '아키텍처·대규모 학습·연산 자원 확보'에 집중된 모델 규모 중심이라면, 새로운 가치사슬은 '알고리즘 구조·시스템 설계·인프라 운영 표준'이 결합된 운영 구조 중심
  - AI 가치사슬 재편은 기술패권의 개념 자체를 변화시켜, 추론 효율을 극대화하는 알고리즘과 시스템 설계 능력, 인프라 운영 생태계와 사실상 표준(de facto standard)의 영향력이 패권 경쟁의 핵심 요소로 부상
  - 이는 초대형 모델을 가장 먼저 개발하거나, 최대 규모의 학습 클러스터를 보유한 국가가 반드시 AI 주도국이 된다고 어려운 구조
  - 즉 모델 규모 경쟁에서 불리하더라도, 운영 구조·시스템 공동설계·표준 대응 역량을 확보할 경우, 글로벌 AI 가치사슬에서 핵심적 위치를 차지할 수 있는 전략적 공간이 존재
  - 결과적으로 거대 모델 중심의 경쟁 구도가 약화되는 현상은 단순한 기술적 흐름이 아니라, 향후 기술패권의 판도를 재편하는 구조적 변곡점으로 작용
- 요약하면, 학습 중심 시대는 연산 자원 확보 → 대규모 학습 → 모델 성능 확보 → 배포로 이어진다면, 추론 중심 시대는 **알고리즘 구조 최적화 → 시스템·인프라 공동설계 → 추론 실행·운영 → 서비스 통합·표준화의 가치사슬로 재편**
  - 학습 중심 시대의 AI 가치사슬이 대규모 연산 자원 확보, 모델 아키텍처 설계, 대규모 학습 수행을 중심으로 형성되었다면, 추론 중심 시대의 AI 가치사슬은 알고리즘 구조 최적화, 시스템·인프라 공동설계, 추론 실행·운영, 서비스 통합과 사실상 표준 형성 단계로 중심축이 이동

- 즉 과거에는 더 큰 모델을 만들 수 있는 역량이 가치사슬의 핵심이었다면, 앞으로는 모델을 얼마나 효율적으로 실행하고 운영할 수 있는지가 가치사슬의 중심으로 부상
- 이는 AI 가치가 모델 개발 단계보다 실행·운영·통합 단계에서 더욱 크게 실현되는 방향으로 재편되고 있음을 의미

표 9 AI 가치사슬 재편 비교

단계	학습 중심 시대	추론 중심 시대
[1] 핵심 인프라 기반 구축 단계	- 대규모 GPU 클러스터 및 학습 데이터 중심 - 연산 및 학습 자원 확보가 핵심 인프라	- 추론 특화 칩, 메모리 용량·대역폭, 메모리 계층화 - 추론 운영 인프라가 핵심 인프라
[2] 모델 생산 단계	- 모델 아키텍처 설계, 파라미터 규모 확대, 대규모 학습 수행 - 벤치마크 성능 향상과 모델 규모 확대가 주요 경쟁 기준	- 알고리즘 구조 최적화, 시스템·인프라 공동설계 - 토큰당 비용, 지연시간, 처리량을 고려한 구조 설계가 주요 경쟁 기준
[3] 실행 및 운영 단계	- 대규모 학습 수행과 모델 성능 확보가 중심 - 배포와 운영은 학습 이후의 후속 단계로 인식	- 추론 실행·서빙 구조가 가치사슬의 핵심 단계로 부상 - 동시 요청 처리, 긴 문맥 처리, 서비스 안정성이 운영 경쟁력 좌우
[4] 가치 창출 및 시장 지배 단계	- 선도 모델 개발 여부와 모델 보유 자체가 가치 창출의 중심 - 모델 성능, API 제공, 클라우드 연동을 통해 시장 영향력 확보	- 실제 서비스 운영·배포·통합·표준화 역량이 가치 창출의 중심 - 운영 구조와 사실상 표준 형성 역량이 시장 지배력과 기술패권의 핵심 변수

※ 출처: 저자 작성.

AI 산업구조 변화: 독점 완화와 새로운 집중

- AI 알고리즘의 구조적 변화는 산업 진입 기회를 넓히는 동시에, 시장 지배력은 다른 단계로 이동시키는 양면적 변화를 초래
  - 초대규모 학습 능력이 절대적 진입장벽이던 시기와 달리, 최근에는 구조적 최적화와 추론 효율을 통해 상대적으로 작은 자원으로도 경쟁 가능한 영역이 확대
  - 그러나 진입장벽 완화가 곧 산업 전반의 분산으로 이어지는 것은 아니며, 실제로는 AI 서비스를 직접 배포하는 단계와 운영 기준을 설계하는 단계를 중심으로 새로운 집중 구조가 형성
  - 즉, 모델 개발 자체의 장벽은 일부 낮아질 수 있지만, 대규모 서비스 운영에는 클라우드, 추론 엔진, 서비스 배포 체계가 필요하며, 동시에 이러한 요소를 하나의 효율적 구조로 결합해 사실상 표준을 제시하는 기업이 새로운 시장 지배력의 핵심 주체로 부상
  - 과거의 독점이 초대형 모델과 학습 인프라에 기반했다면, 앞으로의 집중은 AI 서비스를 직접 운영·배포하는 기업과 그 운영 방식의 기준을 설계하는 주체를 중심으로 강화될 전망

- AI 운영 방식(추론 엔진·실행 환경·API 등)에 대한 집중은 사실상 표준(de facto standard) 선점을 통해 모델 자체가 아닌 **운영 환경에 대한 통제권**을 가진 주체를 중심으로 시장 지배력을 재편할 전망
  - NVIDIA는 GTC 2026[55]에서 AI 반도체, BlueField-4, NVLink, Dynamo, Open Models까지 함께 제시하며, 알고리즘-시스템-인프라 공동설계를 실제 산업 전략으로 구체화
  - 이는 단순히 GPU뿐 아니라 고객이 사용하는 AI 서비스가 개발·배포·운영되는 전 과정을 NVIDIA 방식에 맞추도록 유도함으로써 사실상 산업 표준을 선점하려는 전략으로 해석
    ※ 국내에서는 아직 초기 단계이지만, 운영 방식의 기준을 국내 생태계 안에서 만들기 위한 Rebellions의 vLLM 기반 NPU 추론 시도가 대표적
  - 이러한 흐름은 오픈소스 진영에서도 나타나며, 소수의 실행 방식이 사실상 표준 경로로 자리 잡고 있음을 시사
    ※ 추론 서빙 엔진인 vLLM은 2026년 3월 기준 GitHub에서 Star 7.3만 개, Fork 1.4만 개 이상을 기록하며[56] 특정 운영 방식이 빠르게 산업 표준으로 확산

표 10 AI 산업구조 변화의 주요 특징

구분	주요 변화	의미
진입 장벽	- 초대규모 학습 능력 중심에서 구조적 최적화·추론 효율 중심으로 이동	- 전문·중소 AI 기업, 도메인 특화 기업의 진입 가능성 확대
시장 집중	- 모델 개발 단계에서 서비스 운영·배포와 운영 환경 기준 설계 단계로 이동	- 서비스를 직접 운영하는 주체와 운영 방식의 기준을 설계하는 주체를 중심으로 시장 지배력 재편
사실상 표준	- 추론 엔진, 실행 환경, API 등 운영 방식의 사실상 표준화	- 특정 운영 방식이 산업의 기준 경로로 자리 잡으며 운영 환경 통제 주체의 영향력 확대
정책 방향	- 경쟁 촉진과 집중의 동시 대응 필요성	- 모델 개발 지원뿐 아니라 운영·배포 단계의 집중 구조까지 함께 고려할 필요

※ 출처: 저자 작성.

55) NVIDIA(2026), Watch Jensen Huang's GTC 2026 Keynote: On Demand.
56) Star는 관심·인지도, Fork는 복제·수정·활용 확산의 정도를 보여주는 지표

2 반도체 산업의 영향과 새로운 경쟁 국면

➡ 반도체 경쟁의 중심 이동: 범용 학습 가속기에서 추론 특화 구조로

- 추론 AI의 확산은 반도체 경쟁의 초점을 범용 GPU 확보에서 **추론 단계에 특화된 구조 경쟁**으로 이동 가속화
  - 추론 중심 시대에는 지연시간(latency), 처리량(throughput), 토큰당 비용(cost)이 더 직접적인 경쟁 변수로 부상
  - 이에 따라 반도체도 연산의 성능뿐 아니라, Prefill & Decode 분리, 대규모 동시성, KV cache 접근, 긴 문맥 처리에 얼마나 적합한가를 기준으로 재평가되는 구조로 전환
  - 결국 추론 시대의 반도체 경쟁은 범용성보다 작업 특성 적합성, 단일 칩 성능보다 시스템 전체의 운영 효율, 그리고 칩 자체보다 공동설계 능력에 의해 좌우되는 방향으로 이동
- 이러한 변화는 추론 전용 칩 경쟁을 본격화하며 GPU 일변도의 시장 구도에 새로운 변수로 작용
  - NVIDIA가 GTC 2026에서 공개한 Groq 3 LPU(Language Processing Unit)는 AI 추론 성능을 극대화하기 위해 설계된 전용 가속기의 대표 사례[57]
  - AWS와 Cerebras는 2026년 3월 공식 발표에서 추론을 Prefill과 Decode로 분리하고, Trainium은 Prefill, CS-3는 Decode를 담당하는 구조를 제시[58]
  - Microsoft의 'Maia 200'는 하이퍼스케일러 자체 설계형 추론 칩 사례로서, 추론 칩 경쟁이 클라우드 사업자로 확대[59]
  - AMD는 2026년 1월, MI355X가 MoE 모델 추론에서 유의미한 성능 우위를 확보했음을 제시하며, 범용 가속기 시장의 중심축을 추론 최적화로 빠르게 이동[60]
  - Google은 Google Cloud Next 2026에서 학습과 추론의 요구 조건이 더 이상 동일하지 않음을 전제로, TPU 8t를 대규모 사전학습용, TPU 8i를 사후학습 및 추론용으로 구분해 제시[61]

➡ 메모리 중심 시대: HBM의 전략적 가치와 메모리 계층화

- 메모리 중심 시대에 HBM은 핵심 전략 자산으로 메모리 대역폭 경쟁이 반도체 산업의 중심축이나, HBM의 중요성 유지가 곧 **HBM 단독 체제의 지속을 의미하지는 않음**에 유의

57) NVIDIA(2026.3.16.), Inside NVIDIA Groq 3 LPX: The Low-Latency Inference Accelerator for the NVIDIA Vera Rubin Platform.
58) Amazon(2026.3.13.), AWS and Cerebras Collaboration Aims to Set a New Standard for AI Inference Speed and Performance in the Cloud.
59) Microsoft(2026.1.26.), Maia 200: The AI accelerator built for inference.
60) AMD(2026.1.6.), Single Node and Distributed Inference Performance on AMD Instinct MI355X GPU.
61) Google(2026.4.23.), Inside the eighth-generation TPU: An architecture deep dive.

- 대규모 추론에서는 모델 가중치와 KV cache를 빠르게 읽고 갱신해야 하므로, 고대역폭 메모리와 이를 연산 칩에 밀착시킨 적층형 패키징 구조가 여전히 결정적
- 따라서 앞으로 HBM의 전략적 가치는 줄어드는 것이 아니라, 오히려 추론 효율을 지탱하는 가장 비싼 핵심 자원으로서 그 중요성이 유지되는 구조로 보는 것이 타당
- 그러나, 추론 서비스는 학습보다 훨씬 더 긴 시간 동안 상시 운영되는 과정에서 모든 상태를 항상 고가의 HBM에만 유지하는 구조는 비용 측면에서 비효율적
- 메모리 중심 시대의 경쟁은 HBM 단독 확장이 아니라, 온칩 SRAM-HBM-DRAM·SSD-외부 메모리로 이어지는 다층 구조를 통해 지연시간과 비용을 함께 최적화하는 방향으로 전개

○ 향후 반도체 산업의 핵심 경쟁은 메모리 계층을 설계하고 결합하는 경쟁으로 확대되고, 이러한 변화는 메모리 산업 내부에서도 새로운 분화를 촉진

- 추론 시대의 메모리 전략은 HBM 고도화와 함께 DRAM, SSD, CXL, HBF를 포함한 다층 조합 경쟁으로 확장
- HBM은 즉시성이 중요한 데이터와 활성 상태를 담당하고, 상대적으로 저렴한 DRAM과 SSD는 대용량 상태 저장과 재사용 계층을 담당하는 다층 구조가 현실적 대안으로 부상
- 이때 CXL은 단순한 인터페이스가 아니라, 메모리 공유·풀링을 통해 GPU 중심 구조의 제약을 완화하는 핵심적인 시스템 확장 기술
- 이와 함께 Sandisk와 SK hynix는 2026년 2월 HBF(High Bandwidth Flash) 표준화 추진을 발표하면서, 이를 AI 추론 시대를 위한 새로운 메모리 계층으로 규정[62]

※ HBF는 아직 표준화 초기 단계이므로 장기적 관점에서 HBM과 SSD 사이의 새로운 계층을 만들려는 시도라는 점에서 새로운 구조 변화의 신호로 주목할 필요

➔ 메모리 중심 시대의 역설: 메모리와 비메모리의 경계 약화

○ 메모리 중심 시대는 한편으로 저장 계층을 늘리지만, 다른 한편으로는 연산과 메모리의 물리적 거리 축소를 동시에 추구하는 역설적 흐름이 등장

- Cerebras는 CS-3의 decode 최적화와 함께 온칩 SRAM 기반 효율을 강조하고 있으며, 이는 데이터 이동을 줄여 추론 효율을 높이려는 대표 사례
- NVIDIA는 GTC 2026에서 공개한 Groq 3 LPU에는 HBM 대신 SRAM을 핵심 메모리로 탑재하며, AI 메모리 공급망의 판도 변화를 예고

※ NVIDIA는 LPU 역할 분담을 통해 파라미터(매개변수)가 조 단위인 최고급 AI 모델의 추론 처리량을 35배 향상할 수 있다고 보도[63]

62) Sandisk(2026.2.25.), Sandisk and SK hynix Begin Global Standardization of Next-Generation Memory Solution, High Bandwidth Flash (HBF™).
63) NVIDIA(2026.3.16.), Inside NVIDIA Groq 3 LPX: The Low-Latency Inference Accelerator for the NVIDIA Vera Rubin Platform.

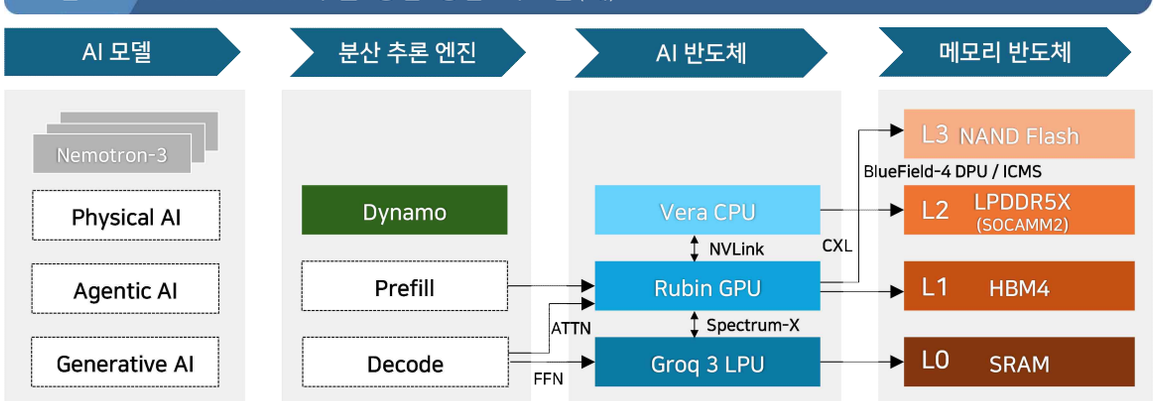
- Google 역시 HBM-온칩 SRAM-호스트 메모리의 다층 메모리 구조를 명시하고, TPU 8i에서는 KV cache 확장을 위해 온칩 SRAM을 크게 확대한 점을 강조
- 이러한 흐름은 고가의 대용량 저장 구조를 확대하기보다, 연산 장치 바로 옆의 초고속 메모리와 외부 계층 메모리를 함께 사용하는 방향으로 설계가 이동하고 있고, **장기적으로는 메모리와 비메모리의 경계를 약화하는 구조로 발전할 전망**

※ HBM4부터 GPU 특정 기능이 베이스 로직 다이로 통합되고 있으며, 장기적으로는 PIM(processing-in-memory)과 같이 메모리 내부 연산 기술도 메모리와 비메모리 산업의 전통적 구분이 점차 흐려지는 방향과 일치

새로운 경쟁 국면

- 추론 전용 칩 경쟁, HBM의 전략적 가치 지속, 메모리 계층화, 메모리와 비메모리의 경계 약화 흐름은 반도체 경쟁이 더 이상 칩 단품 성능 경쟁이 아님을 시사
  - 앞으로의 우위는 개별 칩의 최대 성능보다, 실제 추론 서비스의 병목을 기준으로 **어떤 메모리 계층과 어떤 실행 구조를 결합하느냐에 따라 결정될 가능성**
  - NVIDIA가 GTC 2026에서 LPU, 인프라, 소프트웨어, 모델을 통합 제시한 것은 단순 칩 공급을 넘어 AI 운영 구조 자체를 설계하려는 전략 변화의 대표적 사례
  - Google 역시 Google Cloud Next 2026에서 온칩 SRAM·CAE·Boardfly 통합 구조를 통해 추론 시대의 경쟁이 연산 성능 자체보다 메모리·네트워크·서빙 구조의 결합으로 이동하고 있음을 강조
  - 이는 향후 경쟁의 초점이 개별 반도체의 성능 우위보다, 모델-실행 엔진-가속기-메모리 계층이 통합된 추론 인프라의 표준 구조 설계 역량으로 이동하고 있음을 시사
  - 따라서 추론 중심 AI 시대의 반도체 패권은 칩의 단순 성능보다 추론 인프라의 통합 설계와 운영 구조 주도권을 중심으로 재편될 전망

그림 5 NVIDIA AI 추론 중심 통합 시스템(예)[64]



※ 출처: 저자 작성.

64) NVIDIA가 GTC 2026에서 제시한 내용을 바탕으로 AI 모델-분산 추론 엔진-AI 반도체-메모리 반도체가 하나의 추론 체계로 결합한 구조를 도식화한 것

3 AI 운영 주도권과 국가 AI 주권의 결합

AI 주권 기준의 변화

- 추론 중심 AI로의 전환은 국가 AI 주권의 실체를 단순한 모델 개발 능력에서 서비스 운영 주도권으로 확장
  - 과거에는 자체 AI 모델의 확보 여부가 주권 논의의 핵심 지표로 인식되었으나, AI가 공공·산업·국방 영역으로 확산되면서 AI 주권 개념이 운영 역량과 통제 체제로 재정의
  - 이는 AI 주권이 기술 국산화 여부만으로 완결되지 않으며, 추론 엔진·실행 환경·배포 체계·운영 구조까지 포함하는 운영·통제 가능한 구조라는 의미
  - 이러한 변화는 AI 경쟁의 무게중심이 학습 중심의 기술 확보에서 추론 중심의 운영 체계 설계로 이동한 구조적 변화와 직접 관련
- 이제 국가 AI 경쟁력은 AI 모델 개발 능력과 보유 여부를 넘어, **AI 모델의 운영·배포·통제 능력**에 의해 결정
  - 공공 행정, 산업 자동화, 국방 의사결정과 같이 상시적 서비스가 필요한 영역에서는 모델 자체보다 지속적 운영 가능성, 보안 및 장애 대응 능력이 더 직접적인 경쟁 기준
  - 따라서 국가 AI 경쟁력은 모델 자체의 우수성보다, 이를 안정적으로 운영·배포·통제할 수 있는 역량에 의해 좌우
  - 이는 AI 서비스 경쟁의 중심이 모델 소유에서 운영 구조의 설계와 배포 경로의 통제 능력으로 이동하고 있음을 의미

운영 주도권과 국가 통제력의 결합

- AI 운영 주도권은 단순한 기술 경쟁이 아니라 사실상 표준(de facto standard)을 선점하는 운영 주도권 경쟁
  - 특정 추론 엔진, 분리형 추론 구조, KV cache 관리 방식, 운영 구조가 산업 전반에 확산될수록 후발 주자는 모델 성능보다도 기존 운영 표준과의 호환성에 더 크게 제약
  - AI 운영 방식에 대한 집중은 모델 자체가 아니라 운영 환경 통제 주체 중심으로 시장 지배력을 재편하는 양상
  - 결국 운영 주도권은 모델 개발 능력의 문제를 넘어, 누가 AI 운영 방식의 기준을 만들고 확산시키는가의 문제로 수렴
- AI 주권의 핵심은 모델의 국내 개발 여부만이 아니라, 운영 경로와 의사결정 권한의 통제 범위 문제

- 외형상 동일한 모델이라도 특정 해외 클라우드, 외부 추론 구조, API 등에 의존할 경우, 실제 운영 권한과 업데이트 권한, 장애 대응 권한은 외부 주체에 집중될 가능성
- 이는 AI 주권이 단순한 기술 자립 개념이 아니라, 법적 관할·운영 통제·위험 관리까지 포괄하는 구조적 통제력의 개념임을 시사
- 따라서 **국가 AI 주권**은 소버린 모델만으로 완결되기 어렵고, **소버린 운영과 소버린 통제**가 결합될 때 비로소 실질적 의미를 확보

○ 중요한 점은 AI 주권이 단순히 운영 주체의 문제가 아니라, 알고리즘 구조 선택 문제와도 밀접히 연결

- 희소성 기반 알고리즘과 추론 중심 설계는 동일한 성능을 더 적은 자원으로 구현할 가능성을 높임으로써, 대규모 학습 인프라가 없는 국가에도 자체 운영·배포 역량을 확대할 현실적 기회를 제공
- 이는 AI 주권이 단순한 모델 보유의 문제를 넘어, 운영 가능성과 통제 가능성을 확보할 수 있는 구조 선택의 문제라는 점을 보여주며, 따라서 알고리즘 구조 선택과 운영 설계는 국가 통제력의 범위를 결정하는 핵심 요소

➡ 비가시적 종속과 새로운 경쟁 축

○ 운영 주도권은 AI 시스템이 겉으로는 동일해 보여도, 실제로는 운영 환경에 따라 전혀 다른 통제 구조를 갖게 된다는 점에서 더욱 중요

- 특정 해외 플랫폼이나 운영 스택에 대한 의존이 심화될 경우, 국가는 모델을 보유하더라도 서비스 중단, 비용 증가, 정책 반영 지연, 보안 통제 제약에 직면할 가능성
- 특히 이러한 종속은 하드웨어 의존처럼 명시적으로 드러나지 않고, 실행 환경·운영 소프트웨어·데이터 이동 경로·관리 체계에 포함된 형태의 비가시적 종속으로 나타나는 특징
- 따라서 AI 주권의 핵심 질문은 국산 모델 개발 여부가 아니라, 운영 경로와 통제 지점의 국내 통제 가능성 문제로 이동

○ 한편, 추론 중심 전환과 운영 구조 경쟁의 심화는 후발국에도 글로벌 AI 가치사슬에서 핵심적 위치를 차지할 수 있는 전략적 공간을 제공

- 모델 규모 경쟁에서는 불리한 국가라도, 공공·산업 분야에서 필요한 운영 구조와 서비스 통제 체계를 설계할 수 있다면 AI 가치사슬에서 전략적 위치를 확보할 가능성
- 특히 **추론 효율화, 자원 분리, 메모리 계층화, 운영 스택 설계 역량**은 대규모 학습 인프라 보유 여부와는 다른 경쟁 축이라는 점에서 의미
- 따라서 앞으로의 국가 경쟁은 대규모 AI 모델 확보 여부보다, 자국의 목적에 맞는 AI 서비스의 운영·통제 구조에 의해 더 크게 좌우

토큰 생산 체제와 AI 기술패권 재정의

- 추론 중심 AI로의 전환은 데이터센터의 성격을 단순 저장·처리 인프라에서 지능형 토큰을 지속적으로 생산하는 고부가가치 생산 설비로 변화시키고 있으며, 이는 AI 운영 주권이 기술패권 경쟁에서 갖는 전략적 의미를 새롭게 재정의

  - 과거 데이터센터가 데이터를 저장하고 서비스를 제공하는 디지털 기반시설에 가까웠다면, 추론 중심 AI 시대의 데이터센터는 모델, 가속기, 메모리, 운영 소프트웨어를 결합해 **토큰이라는 지능 산출물을 생산하는 'AI 공장'으로 진화**
  - 최근 미국 주요 하이퍼스케일러의 AI 데이터센터 및 컴퓨팅 인프라 투자가 수천억 달러 규모로 확대되고 있는 흐름은 AI 경쟁이 모델 개발을 넘어 지능 생산 설비와 운영 인프라를 선점하는 경쟁으로 전환되고 있음을 시사
  - 최근 미국 주요 하이퍼스케일러의 AI 데이터센터 및 컴퓨팅 인프라 투자가 7,000억 달러를 상회하는 규모로 확대되는 흐름은 AI 경쟁이 모델 개발을 넘어 지능 생산 설비와 운영 인프라를 선점하는 경쟁으로 전환되고 있음을 시사

  ※ 2026년 1분기 실적 발표 이후 집계된 분석은 Big 5 하이퍼스케일러(Amazon, Alphabet, Meta, Microsoft, Oracle 등)의 2026년 AI 인프라 관련 투자 규모를 약 7,250억 달러로 추정[65]

- 이러한 관점에서 과거 산업혁명이 **'제품 생산 능력'**을 중심으로 기술패권을 형성했다면, AI 혁명 시대의 기술패권은 **'토큰 생산 능력'**과 이를 배포·과금·통제하는 운영 플랫폼 지배력을 중심으로 재편될 가능성

  - 산업혁명기의 공장이 원료, 기계, 노동을 결합해 제품을 생산했다면, AI 혁명 시대의 데이터센터는 모델, 반도체, 메모리, 운영 소프트웨어를 결합해 지능형 토큰을 생산
  - 이에 따라 AI 기술패권의 핵심은 더 큰 모델을 보유하는 능력만이 아니라, 지능을 얼마나 낮은 비용과 지연시간으로 생산하고, 어떤 플랫폼과 과금 체계를 통해 배포하며, 어느 주체가 그 운영 경로와 통제 지점을 장악하는가의 문제로 이동
  - 글로벌 AI 이용국이 자체 운영 체계와 통제력을 확보하지 못할 경우, 모델을 보유하더라도 해외 추론 인프라, 운영 스택, 배포 경로, 과금 구조에 종속될 가능성
  - 따라서 AI 주권은 소버린 모델 확보를 넘어 소버린 운영과 소버린 통제를 포함하는 개념으로 확장되어야 하며, 이는 향후 국가 AI 전략이 추론 인프라, 운영 소프트웨어, 배포 체계, 거버넌스를 함께 설계해야 하는 이유

65) A.L. Capital Advisory(2026.5.22.), The AI Capex Cycle: $725B Hyperscaler Buildout and the Five High-Conviction Positions.

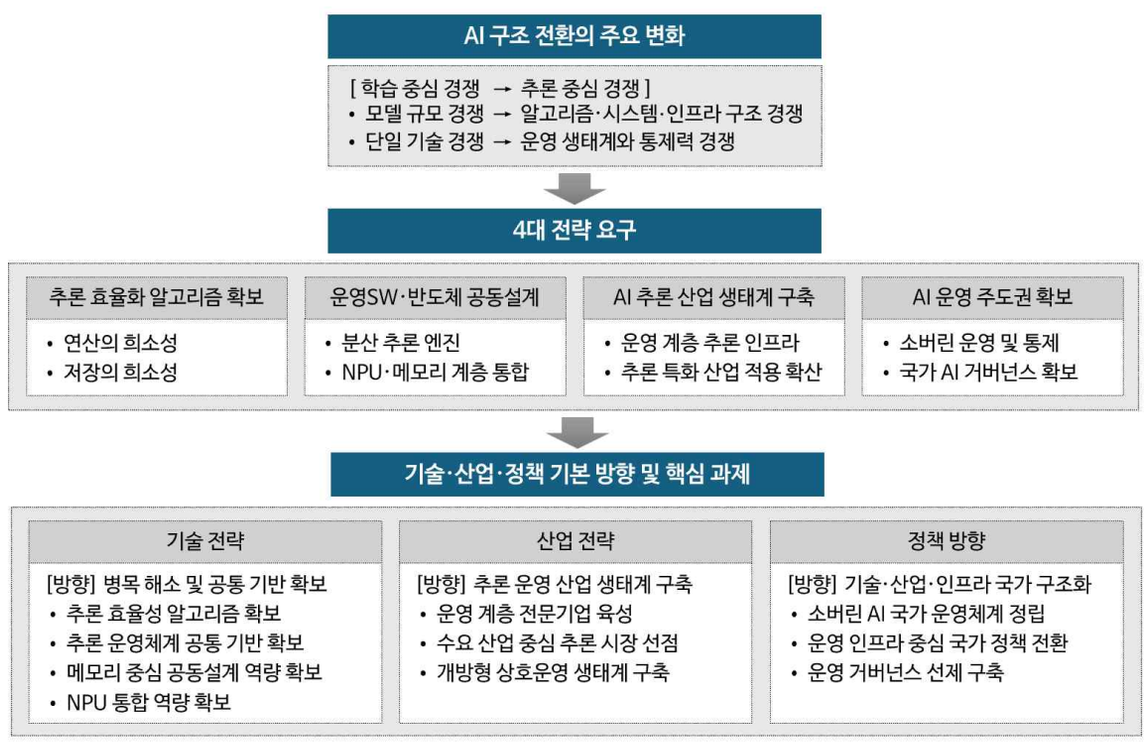
4 전략적 선택과 정책 방향

AI 추론의 시대, 대응 과제 도출

- 추론 중심 AI로의 전환은 단순히 모델 성능 경쟁의 변화에 그치지 않고 알고리즘 구조, 시스템 운영 방식, 인프라 설계, 산업 가치사슬, 국가 AI 주권의 재편으로 확산
  - 이에 따라 향후 국가 전략은 개별 기술 확보나 단일 산업 지원을 넘어, **AI 구조 전환**이 제기하는 **핵심 전략 요구**를 식별하고 이를 기술 전략, 산업 전략, 정책 방향으로 연계하는 방식으로 설계될 필요
  - AI 구조 전환이 제기하는 핵심 전략 요구는 추론 효율화 알고리즘 확보, 운영 소프트웨어·반도체 공동설계, AI 추론 운영 산업 생태계 구축, AI 운영 주도권 확보 등 네 가지로 구분
  - 이러한 4대 전략 요구는 각각 기술 전략, 산업 전략, 정책 방향의 기본 방향으로 수렴되어, 추론 중심 AI 시대에 대응하기 위한 10대 전략 과제 도출로 연결

※ 기술 전략은 국가 R&D 차원에서 해결해야 할 구조적 병목과 공통 기반 확보 방향을 제시하고, 산업 전략은 이를 시장과 수요 산업에 연결하는 확산 경로를 제시하며, 정책 방향은 기술·산업·인프라를 국가 차원에서 통합·구조화하는 역할을 담당

그림 6 AI 구조 전환에 따른 10대 전략 과제(안) 도출 흐름



※ 출처: 저자 작성.

(1) 기술 전략

➡ 기본 방향: 구조적 병목 해소와 공통 기반 확보

- 기술 전략의 기본 방향은 더 큰 모델 자체의 개발보다, 추론 단계의 구조적 병목을 줄이고 국내 기술이 실제 서비스 환경에서 작동할 수 있는 공통 기반을 확보하는 데 초점을 둘 필요
  - 초거대 모델 경쟁이 계속되더라도 후발국이 현실적으로 공략할 수 있는 지점은 최고 성능 경쟁 그 자체보다, 동일 성능을 더 낮은 자원과 비용으로 구현하는 구조 기술에 있음
  - 특히 한국은 메모리, 패키징, 제조 기반의 강점을 보유하고 있는 만큼, 이를 알고리즘·시스템·인프라 공동설계와 연결하는 방향이 전략적으로 중요
  - 기술 전략은 개별 요소기술의 성과를 넘어, 알고리즘 구조–운영 소프트웨어–가속기–메모리 계층이 결합되는 기술 체계를 구축하는 방향으로 설계할 필요

➡ 기술 전략 1: 추론 효율형 알고리즘의 선도 기술 확보

- 국가 R&D는 초거대 모델(Dense) 개발 경쟁 자체보다, 추론 단계의 연산·저장·데이터 이동 병목을 줄일 수 있는 알고리즘(Sparse) 구조 확보에 집중
  - 연산의 희소성과 저장의 희소성은 추론 비용을 낮추고 동시 처리 효율을 높이는 핵심 구조 원리라는 점에서 전략적 가치가 큼

※ DeepSeek-V3는 MoE 구조와 MLA를 채택해 효율적 추론과 비용 효율적 학습을 지향했고, 이는 연산 희소성과 저장 희소성을 함께 겨냥한 대표 사례

  - 한국이 공략해야 할 방향은 최고 성능 모델 자체보다, 긴 문맥 처리, 실시간 응답, 온디바이스·엣지 배치까지 고려한 서비스 친화적 알고리즘 구조

※ Kimi Linear는 하이브리드 선형 어텐션 구조를 도입하여 긴 문맥 추론에서 효율적인 정보 기억 및 계산 구조 설계가 핵심 경쟁력임을 증명

  - 결과적으로 알고리즘 분야의 목표는 더 큰 모델이 아니라, 더 적은 자원으로 배포 가능한 모델 구조와 추론 친화적 설계 원리를 확보하는 데 둘 필요

➡ 기술 전략 2: 국산 추론 운영 체계[66]의 공통 기반 확보

- 추론 시대에는 칩 자체의 성능뿐 아니라, 모델을 실제 서비스 환경에서 실행·관리하는 운영 소프트웨어의 완성도가 경쟁력을 좌우하므로, 공통 운영 계층 확보가 핵심 과제

66) 여기서 운영 체계는 모델 배치, 자원 배분, 메모리 관리, 실행 제어를 담당하는 시스템 소프트웨어를 의미

- NVIDIA의 경쟁력은 GPU를 넘어, Dynamo와 같은 분산 추론 프레임워크로 저지연·고처리량으로 전체 스택을 제공하는 데 있으며, 이는 운영 주도권이 칩보다 런타임, 스케줄러, 메모리 관리 등 소프트웨어 계층에서 결정됨을 의미
- 반면 한국은 칩이나 모델의 개별 성과에 비해, 추론 단계의 자원 분리, 상태 관리, 실행 제어, 성능 최적화를 담당하는 공통 운영 계층이 상대적으로 취약한 편

※ 그러나 최근 Rebellions와 Red Hat이 Rebellions NPU, vLLM, OpenShift AI를 결합한 풀스택 엔터프라이즈 AI 플랫폼을 제시한 것은 국내 NPU도 운영 스택과 결합해 국산 추론 운영 체계의 기반 확보 가능성을 보여준 대표 사례[67]

- 따라서 국가 R&D는 특정 기업 제품 개발에 한정되기보다, 표준 인터페이스, 런타임, 벤치마크 환경 등 운영 체계의 공통 기반을 확보하는 방향으로 추진될 필요

기술 전략 3: 메모리 중심 추론 인프라의 공동설계 역량 강화

- 한국의 강점인 메모리와 패키징 역량을 추론 인프라 경쟁력으로 연결하기 위해서는, 연산 장치와 메모리 계층을 함께 최적화하는 공동설계 역량을 국가 핵심 기술로 육성할 필요

- 추론 단계에서는 연산량 자체보다 상태 저장, 이동, 재사용의 효율이 성능과 비용을 좌우하므로, 메모리 중심의 접근이 필수

※ NVIDIA는 Groq 3 LPX와 Vera Rubin 구조에서 저지연 추론을 위해 연산, 저장, 통신 기술을 긴밀히 결합한 랙 스케일의 구조를 제시

- 고대역폭 메모리(HBM)뿐 아니라 DRAM, SSD, CXL 기반 메모리 확장 구조까지 포함하는 다층 메모리 체계를 실전형 인프라 설계의 핵심 축으로 다룰 필요

※ CXL은 CPU・GPU・메모리 간 자원 확장과 공유를 지원하는 고속 인터커넥트 기술로, 추론 인프라의 유연성과 확장성을 높이는 핵심 기반이라는 점에서 전략적 중요성이 큼

- 이에 따라 국가 차원에서는 알고리즘, 시스템 소프트웨어, 가속기, 메모리 계층을 통합적으로 검증할 수 있는 공동설계 역량과 실증 기반을 확보할 필요

기술 전략 4: 국산 NPU의 서비스형 통합 역량 확보

- 국산 NPU 전략은 개별 칩 성능 향상에 머무르기보다, 오픈소스 추론 스택과 결합해, 실제 서비스 환경에서 활용 가능한 통합 구조를 만드는 방향으로 전환될 필요

- 국산 NPU의 경쟁력은 칩 자체의 이론 성능보다, 어떤 운영 소프트웨어와 결합되어 어떤 수요 환경에서 안정적으로 작동하는가에 의해 결정

67) Rebellions(2025.12.11.), Rebellions and Red Hat Introduce Red Hat OpenShift AI Powered by Rebellions NPUs to Fuel Choice and Flexibility in Enterprise AI.

※ 국산 NPU가 있어도 서빙 인터페이스, 런타임, 라이브러리 등의 소프트웨어 생태계가 뒷받침되지 않으면, 실제 운영 주도권 없이 단순 벤치마크 성과에 머무를 가능성

- 특히 제조·통신·공공·국방과 같이 한국이 실증 가능한 분야에서 동일 작업을 반복 검증할 수 있어야, 국산 NPU가 단품이 아니라 운영 표준의 일부로 자리 잡을 수 있음
- 따라서 국가 R&D의 목표는 국산 NPU 개발을 넘어 국산 NPU가 공통 운영 체계 위에서 활용되는 구조 확립

(2) 산업 전략

➡ 기본 방향: 추론 운영 산업 생태계 구축

○ 산업 전략의 기본 방향은 추론 중심 AI를 실제 산업 현장에 배치·운영·확산시키는 운영 산업 생태계 구축에 둘 필요

- 최근에는 선도 모델 간 성능 수렴과 오픈웨이트 모델의 경쟁력 제고가 동시에 나타나면서, AI 산업의 경쟁 우위가 모델 자체보다 운영·배포·통합 역량으로 이동하는 흐름

※ AI Index 2026은 2026년 3월 기준 Arena 상위권 모델들이 사실상 구별하기 어려운 수준으로 수렴했다고 평가하며, 오픈웨이트 모델의 경쟁력도 크게 높아졌다고 지적. 이에 따라 AI 시장에서는 모델 성능 자체보다 비용, 지연시간, 신뢰성, 실제 운용성이 경쟁력의 핵심 변수로 부상

- 반면 서비스 운영과 배포를 담당하는 플랫폼과 실행 환경의 중요성은 오히려 커지고 있어, 시장의 중심은 모델 자체보다 운영 구조와 확산 역량으로 이동하는 양상
- 이에 따라 한국의 산업 전략은 프런티어 모델 경쟁에만 집중하기보다, 운영 계층, 추론 인프라, 산업 적용 솔루션이 결합된 생태계를 구축하는 방향으로 재설계할 필요

➡ 산업 전략 1: AI 반도체-운영 소프트웨어 결합 산업 육성

○ 운영 계층 전문기업 육성과 AI 반도체-운영 스택 결합형 사업모델 구축을 하나의 전략 축으로 통합 추진할 필요

- AI 추론 시대에는 모델 자체보다 이를 실제 서비스로 전환하는 서빙 엔진, 스케줄러, KV cache 관리, 보안, 비용 최적화, 배포 자동화와 같은 운영 계층[68]이 경쟁력을 좌우

※ 오픈소스 추론 엔진 vLLM이 빠르게 확산되고 있다는 사실은, 운영 스택이 사실상 표준의 출발점이 될 수 있음을 보여주는 대표 사례[69]

- 따라서 국내 산업 전략은 운영 중간층을 독자 산업 영역으로 인식하고, 추론 실행 환경과 서비스 운영 체계를 담당하는 전문기업을 체계적으로 육성하는 방향으로 전환할 필요

68) 운영 계층은 추론 엔진, 자원관리, 배포 자동화, 보안 등 모델과 인프라 사이를 연결하는 영역
69) vLLM은 2026년 3월 기준 GitHub에서 Star 7.3만 개, Fork 1.4만 개 이상을 기록하며 특정 운영 방식이 빠르게 산업 표준으로 확산

- 동시에 국산 AI 반도체가 실제 시장에서 활용되기 위해서는 운영 스택과 결합된 플랫폼형·서비스형 사업모델 발굴이 병행되어야 함
- 따라서 국산 AI 반도체 산업은 칩 설계 기업, 시스템 소프트웨어 기업, 클라우드·플랫폼 사업자, 수요기업이 결합하는 구조를 구축해야 하며, 이는 한국의 반도체 강점을 운영 가능한 AI 인프라 경쟁력으로 연결하는 핵심 경로라는 점에서 의미가 큼

산업 전략 2: 수요산업 중심의 고신뢰 추론 시장 선점

- 한국의 산업 전략은 범용 AI 서비스 경쟁보다 제조, 통신, 공공, 국방 등 반복 수요와 높은 신뢰성이 요구되는 분야에서 고신뢰 추론 시장을 선점하는 데 초점을 둘 필요
  - 이러한 영역에서는 최고 성능 모델보다 낮은 지연시간, 비용 예측 가능성, 보안, 안정적 운영이 더 중요한 경쟁 기준으로 작용
  - 한국은 대규모 소비자 플랫폼 경쟁에서는 불리할 수 있으나, 수요산업과 현장 실증 기반에서는 비교우위를 가질 수 있음
  - 따라서 산업 전략은 국내 강점 산업과 연계된 특정 시장에서 반복 적용 가능한 추론 서비스 모델을 조기에 확보하는 방향으로 추진할 필요

산업 전략 3: 개방형 상호운용 생태계 구축

- 산업 전략의 중요한 원칙은 폐쇄형 단일사업자 종속보다, 오픈소스 추론 스택·표준 인터페이스·이기종 가속기 호환성을 기반으로 한 개방형 상호운용 생태계의 구축
  - 운영 구조 경쟁이 심화될수록 특정 플랫폼에 대한 종속 위험도 커질 수 있으므로, 중장기적으로는 개방형 생태계 기반이 중요
  - 개방형 상호운용 구조는 국내 중소·전문기업의 참여 공간을 확대하고, 국산 반도체와 소프트웨어의 결합 가능성을 높이는 효과
  - 이에 따라 산업 전략은 폐쇄형 단일 생태계 구축보다, 표준 인터페이스와 상호운용성을 바탕으로 다양한 기업이 연결될 수 있는 구조를 지향할 필요

(3) 정책 방향

➡ 기본 방향: 기술·산업·인프라의 구조화

- 정부 정책의 핵심 역할은 개별 기술이나 기업에 대한 단순 지원을 넘어, AI 경쟁의 중심축 이동(모델 개발 → 운영 구조)에 따른 기술·산업·인프라의 국가적 구조화
  - AI 정책은 모델 개발과 GPU 확보를 포함해, 운영 가능한 국가 AI 체계의 조성 여부가 중요한 판단 기준으로 부상
  - 특히 공공·산업·국방 영역에서는 성능 자체보다 지속적 운영 가능성, 책임성, 보안, 비용 통제 능력이 더욱 중요
  - 이에 따라 정책 방향도 인프라 확충, 운영 거버넌스, 부처 간 역할 조정, 평가 체계까지 포함하는 방향으로 확장될 필요

➡ 정책 방향 1: 소버린 AI의 국가 운영체계 정립

- 소버린 AI 정책의 범위를 모델 보유 중심에서 운영 체계 중심으로 확장하고, 이를 범정부 차원의 국가 AI 운영 체계로 정립할 필요
  - 실질적 소버린 AI는 소버린 모델만으로 완성되지 않으며, 소버린 운영과 소버린 통제가 결합될 때 비로소 확보 가능
  - 이에 따라 AI 정책은 데이터·모델·인프라·운영 소프트웨어·실증 체계를 포괄하는 통합적 운영 체계 관점에서 설계할 필요
  - 또한 현재 AI 정책은 개발, 인프라, 산업 확산, 공공 활용 등 여러 영역으로 분산되어 있으나, 추론 중심 시대에는 이들이 실제 운영 구조 안에서 함께 작동해야 함
  - 따라서 정책의 목표는 개별 사업의 병렬 추진이 아니라, 국가 AI 운영 인프라를 체계적으로 구축하고 공공과 산업 현장에 일관되게 적용 가능한 구조를 만드는 데 둘 필요

➡ 정책 방향 2: 국가 AI 컴퓨팅 정책의 중심을 운영 인프라로 전환

- 국가 AI 컴퓨팅 정책은 단순한 연산 자원 확보를 넘어, 추론 서비스의 지연시간, 처리량, 비용, 보안, 장애 대응 능력을 검증할 수 있는 운영 인프라 중심으로 재설계 필요
  - 추론 중심 AI의 병목은 단순 연산량보다 메모리 계층, 상태 관리, 데이터 이동, 운영 스택에 더 크게 좌우되며, 이에 따라 AI 경쟁의 무대는 모델이 아니라 운영 구조로 이동
  - 이에 따라 국가 AI 컴퓨팅 정책은 학습 중심의 자원 논리에서 벗어나, 추론 실행 환경과 서비스 운영 체계를 실증할 수 있는 방향으로 확장되어야 함

- 공공 인프라 역시 국산 모델, 국산 반도체, 국내 통제 가능한 운영 스택이 결합될 수 있는 시험·실증 기반으로 기능할 필요

정책 방향 3: 운영 거버넌스의 선제적 구축

- 공공·산업 현장에서 AI가 상시적으로 작동하는 환경에 대비하여, 성능관리, 장애 대응, 책임성을 포괄하는 운영 거버넌스를 선제적으로 구축할 필요
  - 추론 중심 AI 시대에는 기술 도입 자체보다, 운영 과정에서 발생할 수 있는 사고, 중단, 책임 소재, 안전성 문제에 대한 관리 체계가 중요
  - 따라서 정책은 규범과 안전 원칙을 선언하는 수준을 넘어, 실제 운영 과정에서 적용 가능한 관리 기준과 점검 체계를 마련하는 방향으로 진화할 필요
  - 이는 공공 서비스 도입뿐 아니라 산업 현장의 AI 확산을 안정적으로 뒷받침하는 핵심적인 인프라로 기능

표 11 AI 추론 시대 대응 10대 전략 과제(안)

10대 전략 과제(안)		주요 내용
기술	① 추론 효율형 알고리즘의 선도 기술 확보	연산·저장 희소성 기반 구조 혁신을 통해 토큰당 비용, 지연시간, 처리량을 개선하는 추론 특화 알고리즘 기술을 선도적으로 확보
	② 국산 추론 운영 체계의 공통 기반 확보	국산 추론 운영 체계의 공통 기반을 마련해 국내 기술이 실제 서비스 환경에서 작동할 수 있는 기반 구축
	③ 메모리 중심 추론 인프라의 공동설계 역량 확보	HBM·DRAM·SSD·CXL 등 다층 메모리 구조와 가속기·시스템 소프트웨어를 통합 최적화하는 공동설계 및 실증 역량 확보
	④ 국산 NPU의 서비스형 통합 역량 확보	개별 칩 성능 경쟁을 넘어 오픈소스 추론 스택과 결합된 서비스형 통합 구조를 구축하고, 공공·제조·통신 등에서 실증 가능한 운영 표준 형성
산업	⑤ AI 반도체-운영 소프트웨어 결합 산업 육성	운영 계층 전문기업 육성과 함께 국산 AI 반도체와 운영 스택이 결합된 솔루션·플랫폼형 사업모델을 발굴하여 운영 계층 산업 기반 강화
	⑥ 수요산업 중심의 고신뢰 추론 시장 선점	제조, 통신, 공공, 국방 등 한국의 강점 수요산업에서 고신뢰·저지연·고효율 추론 서비스를 조기 확산해 실증 기반 시장 선점
	⑦ 개방형 상호운용 생태계 구축	오픈소스 추론 스택, 표준 인터페이스 등을 바탕으로 특정 플랫폼 종속을 줄이고 국내 중소·전문기업 참여가 가능한 개방형 생태계 조성
정책	⑧ 소버린 AI의 국가 운영체계 정립	소버린 AI의 범위를 단순 모델 보유에서 운영 가능성·통제력·지속가능성으로 확장, 분산된 AI 정책을 범정부 차원의 국가 운영체계로 구조화
	⑨ 국가 AI 컴퓨팅 정책의 중심을 운영 인프라로 전환	훈련용 대형 컴퓨팅 중심 정책에서 벗어나 추론 서비스 운영, 메모리 계층화, 시스템 효율을 중심으로 국가 AI 인프라 정책 재편
	⑩ 운영 거버넌스의 선제적 구축	공공·산업 현장에서 요구되는 안정성, 보안, 통제 가능성, 표준·인증, 책임체계 등을 포함하는 AI 운영 거버넌스를 선제적으로 정립

※ 출처: 저자 작성.

참고문헌

A.L. Capital Advisory(2026.5.22.), The AI Capex Cycle: $725B Hyperscaler Buildout and the Five High-Conviction Positions.

AMD(2026.1.6.), Single Node and Distributed Inference Performance on AMD Instinct MI355X GPU.

Amazon(2026.3.13.), AWS and Cerebras Collaboration Aims to Set a New Standard for AI Inference Speed and Performance in the Cloud.

Ashish Vaswani et al.(2017), Attention Is All You Need, arXiv:1706.03762v7.

Compute Express Link, CXL 4.0 Specification Release, https://computeexpresslink.org/wp-content/uploads/2025/11/CXL_4.0-Specification-Release_FINAL_Website-Copy.pdf

CSET(2023.8.), Assessing South Korea's AI Ecosystem.

DeepSeek-AI(2024), DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model, arXiv:2405.04434v5.

DeepSeek-AI(2024), DeepSeek-V3 Technical Report, arXiv:2412.19437v1.

DeepSeek-AI(2025), DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, arXiv:2501.12948v2.

DeepSeek-AI(2025), Multi-head Latent Attention: Scaling KV Cache for Efficient Inference, arXiv:2502.07864v2.

DeepSeek-AI(2026), DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence.

Fedus, W., et al.(2022), Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. Journal of Machine Learning Research.

Google(2026.4.23.), Inside the eighth-generation TPU: An architecture deep dive.

Jiang, A. Q., et al.(2024), Mixtral of Experts, arXiv:2401.04088.

Konrad Staniszewski et al.(2025), KV Cache Transform Coding for Compact Storage in LLM Inference, arXiv:2511.01815v2.

Kwon, W., et al.(2023), Efficient Memory Management for Large Language Model Serving with PagedAttention, arXiv:2309.06180v1.

Lepikhin, D., et al.(2020), GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding, arXiv:2006.16668.

Meng, F., et al.(2025), TransMLA: Multi-Head Latent Attention Is All You Need, arXiv:2502.07864.

Microsoft(2026.1.26.), Maia 200: The AI accelerator built for inference.

MIT Technology Review(2025.11.18.), 트랜스포머 이후 가장 중요한 논문이 나왔다.

Moonshot AI(2025), Kimi Linear: An Expressive, Efficient Attention Architecture, arXiv:2510.26692v2.

NVIDIA, Dynamo Document, https://docs.nvidia.com/dynamo/latest/user-guides/kv-cache-aware-routing.

NVIDIA(2026.3.16.), Inside NVIDIA Groq 3 LPX: The Low-Latency Inference Accelerator for the NVIDIA Vera Rubin Platform.

NVIDIA(2026), Watch Jensen Huang's GTC 2026 Keynote: On Demand; NVIDIA CEO Jensen Huang GTC 2026 Full Keynote, https://www.youtube.com/watch?v=jw_o0xr8MWU&t=3662s.

Patel, P., et al.(2024), Splitwise: Efficient Generative LLM Inference Using Phase Separation, arXiv:2311.18677v2.

Rebellions(2025.12.11.), Rebellions and Red Hat Introduce Red Hat OpenShift AI Powered by Rebellions NPUs to Fuel Choice and Flexibility in Enterprise AI.

Ruoyu Qin et al.(2025), MOONCAKE: Trading More Storage for Less Computation - A KVCache-centric Disaggregated Architecture for Serving LLM Chatbot. FAST 2025.

Sandisk and SK hynix(2026.2.25.), Sandisk and SK hynix Begin Global Standardization of Next-Generation Memory Solution, High Bandwidth Flash (HBF™).

SemiAnalysis, Tokenomics Model, https://semianalysis.com/tokenomics-model/.

SemiAnalysis(2025.9.11.), Another Giant Leap: The Rubin CPX Specialized Accelerator & Rack.

SK hynix(2025), Completes World's First HBM4 Development and Readies Mass Production.

Stanford HAI(2026.4.), AI Index Report 2026.

Synergy Research Group(2025.2.6.), Cloud Market Jumped to $330 billion in 2024.

Xin Cheng et al.(2026), Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models, arXiv:2601.07372v1.

Yinmin Zhong et al.(2024), DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving, arXiv:2401.09670v3.

저자소개

이승민 ETRI ICT전략연구소 기술정책연구본부 기술경제연구실 책임연구원
과학기술연합대학원대학교(UST) ETRI 스쿨 과학기술경영정책 교수
e-mail: todtom@etri.re.kr Tel. 042-860-1775

추론의 시대, AI 구조 전환과 기술패권 전망

발행인 한 성 수
발행처 한국전자통신연구원 ICT전략연구소
발행일 2026년 6월 30일

www.etri.re.kr

ETRI Electronics and Telecommunications Research Institute

34129 대전광역시 유성구 가정로 218
TEL.(042) 860-6114 FAX.(042) 860-6504