Market Insight | 2025.12.12

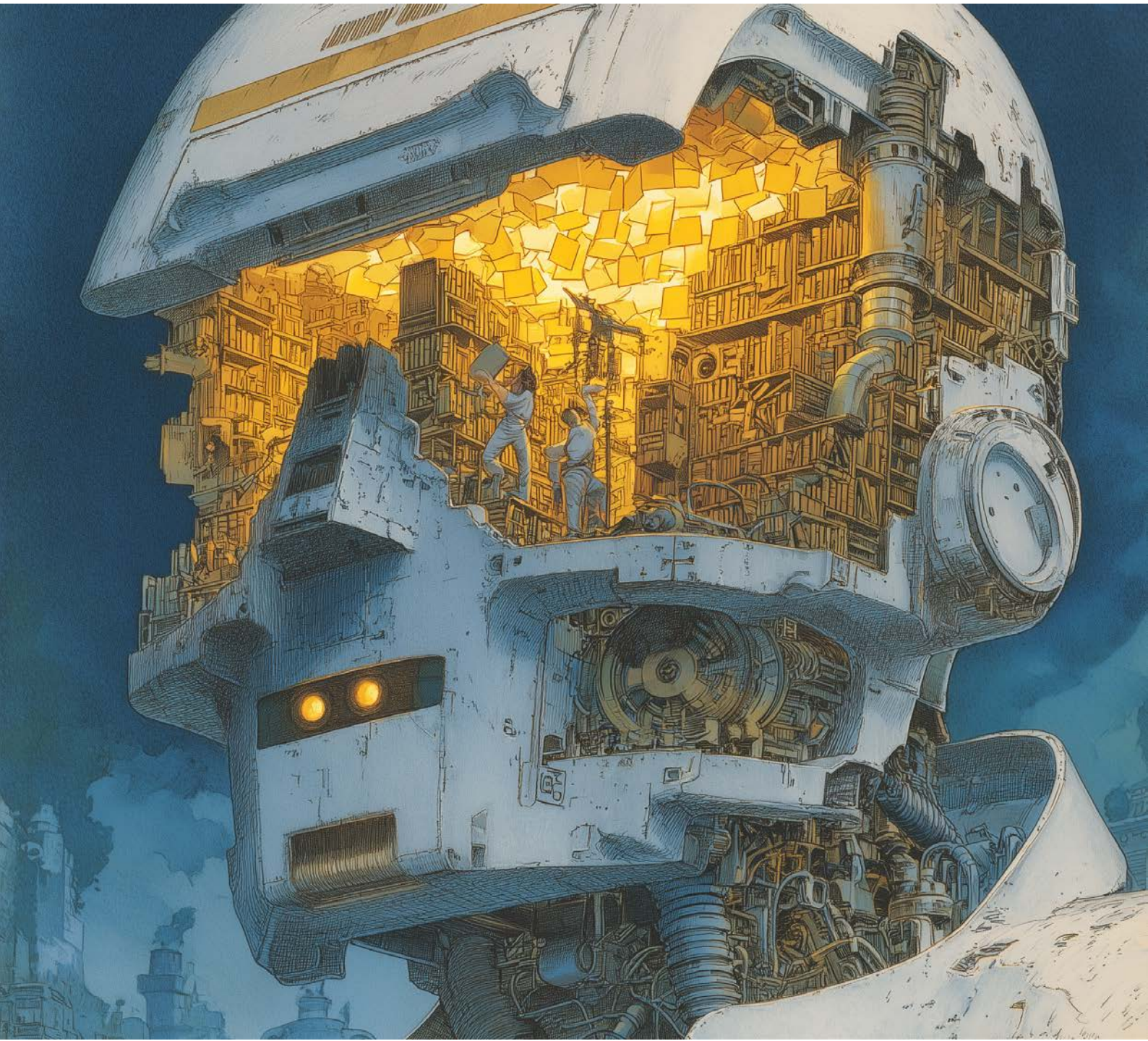
그림: Midjourney(포스트잇(단기기억)을 백과사전(장기기억)으로 옮겨 적고 있는 모습)

MIRAE ASSET 미래에셋증권

AI Hot Issue

구글의 지속적 학습과 메모리의 '질적' 슈퍼사이클

한종목
chongmok.han@miraeasset.com



CONTENTS

Executive Summary 3
구글의 AI 전략과 차세대 아키텍처, 이에 따른 반도체 시장의 변화 3

I. AI의 과학적 발견 4
1. 구글 딥마인드 CEO 데미스 하사비스가 제시하는 향후 1년 로드맵 4
2. 구글이 내비친 최근 역량: 선도적인 AI 연구(지속적 학습과 메모리 혁신) 9
3. 지속적 학습과 계층형 메모리 = 슈퍼사이클 19

Executive Summary

구글의 AI 전략과 차세대 아키텍처, 이에 따른 반도체 시장의 변화

구글 딥마인드의 CEO 데미스 하사비스는 노벨상 수상으로 입증된 과학적 통찰을 바탕으로, 향후 1년을 AI가 '실험실의 과학'에서 '일상의 제품'으로 완전히 전이되는 통합의 해로 선언했습니다. 그는 완전한 멀티모달 통합, 텍스트를 정교한 인포그래픽으로 변환하는 시각 지능, 그리고 가상 세계의 물리 법칙을 내재화한 월드 모델 등 5가지 핵심 트렌드를 제시하며, 챗봇을 넘어 스스로 계획하고 오류를 수정하는 '에이전트'가 1년 내 상용화될 것임을 예고했습니다. 이러한 고도화된 AI를 실현하기 위해 구글은 기존 모델의 치명적 약점인 '기억상실증(학습 후 지적 성장의 정체)'을 극복하는 '중첩 학습(Nested Learning)'과 'HOPE 모델'이라는 새로운 아키텍처를 제시했는데, 이는 정적인 모델에서 벗어나 환경과 상호작용하며 스스로 신경망을 재조립하는 동적인 유기체로 진화시키는 분기점이 될 것입니다.

이 혁신적인 아키텍처는 인간의 뇌가 단기 기억을 장기 기억으로 공고화하는 과정을 모방하여, 정보를 고주파(작업 기억), 중주파(맥락 기억), 저주파(장기 기억)의 세 가지 시간 척도로 나누어 처리합니다. 이를 통해 AI는 외부 정보를 단순히 검색해서 보여주는 RAG 방식의 한계를 넘어, 사용자와의 상호작용을 모델 내부의 가중치로 실시간 내재화하고, 새로운 정보를 배우면서도 기존 지식을 잊지 않는 진정한 의미의 '지속적 학습'을 구현하게 됩니다. 특히 중요한 점은 구글이 소프트웨어적으로 정의한 이 '주파수' 개념이 하드웨어의 물리적 '메모리 계층'과 잘 동기화된다는 것입니다. 즉, 알고리즘이 데이터의 중요도에 따라 저장 위치와 업데이트 속도를 결정하면, 하드웨어가 이에 맞춰 유기적으로 작동해야 하는 새로운 공학적 패러다임이 열릴 수 있습니다.

구체적으로 매 순간 가중치를 읽고 수정하는 작업이 빈번한 '고주파' 영역은 기존 메모리보다 훨씬 강력한 대역폭과 연산 기능이 포함된 맞춤형 HBM을 필요로 합니다. 이는 단순한 데이터 이동을 넘어 HBM4의 베이스 다이에 로직 공정이 도입되어 연산의 병목을 해소해야 함을 의미하며, HBM이 저장소를 넘어 가속기의 일부 영역으로 진입했음을 보여줍니다. 또한, 대화의 긴 맥락을 잇는 '중주파' 영역은 알고리즘의 예측에 따라 데이터를 미리 가져오는 프리페칭 기술이 적용된 CXL과 SOCAMM 메모리를 필수적으로 요구하게 되며, 방대한 지식을 보관하는 '저주파' 영역은 단순한 순차 읽기가 아닌, 모델이 필요한 정보만 핀포인트로 추출하고 벡터 검색을 수행하는 지능형 SSD(Computational Storage)로의 진화를 가속화할 것입니다.

결국 구글의 새로운 로드맵은 소프트웨어 알고리즘이 하드웨어의 물리적 특성을 직접 제어하고 활용하는 시대를 열었습니다. 이에 따라 메모리 반도체 기업은 단순 규격품을 대량 공급하는 데 그치지 않고, 구글과 같은 빅테크의 알고리즘에 최적화된 로직을 메모리에 심어주는 '시스템 솔루션 파트너'로 변모해야 합니다. "알고리즘을 모르면 메모리를 만들 수 없고, 하드웨어를 모르면 모델을 최적화할 수 없는" 이 필연적인 결합은 메모리 산업을 단순한 수급에 따른 가격 반등을 넘어, 반도체가 단순 부품에서 지능형 연산 장치로 격상되며 기술적 가치가 근본적으로 재평가되는 거대한 구조적 슈퍼사이클로 이끌 것입니다.

I. AI의 과학적 발견

1. 구글 딥마인드 CEO 데미스 하사비스가 제시하는 향후 1년 로드맵

(1) 데미스 하사비스의 말을 신뢰하는 이유

약 400일 전 노벨상을 수상하며 과학자로서 정점에 오른 구글 딥마인드의 데미스 하사비스(Demis Hassabis)는 이제 단순한 연구자가 아닌, 인류 역사상 가장 강력한 기술 변혁의 중심에 서 있는 설계자로서 향후 12개월의 기술 로드맵을 제시했다. 하사비스는 AGI 도달 시점을 향후 "5~10년 내"로 보수적으로 잡았지만, 그가 제시한 향후 12개월 동안 일어날 들을 감안하면, 그 시간이 예상보다 훨씬 빠르게 단축될 수 있음을 암시하여 눈길을 끌었다.

표 1. AGI 도달 시점 전망

그룹	예상 연도(25% 확률)
AI 기업 리더	2026
Metaculus 플랫폼 참가자들	2027
Samotsvety팀(예측 대회에서 활동하는 유명한 팀)	2029
AI 연구원들	2032
XPT(예측 토너먼트/프로그램에서 성과가 검증된 상위 예측가 집단)	2047
평균값	2032

자료: 80,000 Hours, 미래에셋증권 리서치센터

그의 말에 힘이 실리는 이유는, 데미스 하사비스의 예측은 지금까지 틀린 적이 없었기 때문이다. 하사비스는 원래 일론 머스크의 긍정적 전망보다는 훨씬 보수적인 톤으로 이야기하는 것에 가깝다. 그럼에도 이번에 그가 얘기한 것은 구글의 AI 역량에 대해 새삼 다시 생각해 보게끔 하는 자신감 있는 어투로 가득차있다.

하사비스가 정의한 향후 1년은 '실험실의 과학'이 '일상의 제품'으로 완전히 전이되는 시기이자, AI가 Narrow AI(좁은 의미의 AI)을 넘어 범용 인공지능(AGI)의 구성 요소들을 유기적으로 결합하는 '**통합의 해'가 될 것이다.**

표 2. 데미스 하사비스가 바라보는 2026년의 키워드: "통합"

통합 분야	세부 내용	기술적 의미	영향 범위
감각의 통합	텍스트, 이미지, 비디오, 오디오가 하나의 신경망에서 자유롭게 변환	멀티모달 AI의 완성 (모든 입출력 형식이 동일한 기반에서 처리)	모든 콘텐츠 형식 상호 호환
인지와 행동의 통합	세상을 이해하는 모델(월드 모델)과 행동하는 모델(에이전트)이 결합	AI가 수동적 이해에서 능동적 에이전트로 진화	자율 AI 시스템, 로봇공학, 자동화
과학과 제품의 통합	실험실의 최첨단 연구가 실시간으로 수십억 명의 사용자 제품에 반영	R&D와 상용화의 시간 간격 제거	기술 민주화, 대중의 AI 접근성 극대화

자료: 데미스 하사비스, 미래에셋증권 리서치센터

그는 현재의 AI 경쟁을 "기술 역사상 가장 치열한 전투"라고 묘사하면서도, 딥마인드가 가진 과학적 방법론(Scientific Method)과 엔지니어링의 결합이 구글에게 지속적인 우위를 제공할 것임을 확신했다.

하사비스가 언급한 5가지 핵심 기술 트렌드(완전한 멀티모달 통합, 시각 지능의 비약, 언어와 영상의 융합, 월드 모델의 대중화, 그리고 에이전트의 신뢰성 확보)다.

(2) 완전한 멀티모달 통합: 입력과 출력의 경계를 허무는 인지 혁명

하사비스는 Gemini 모델이 처음부터 멀티모달(Native Multimodal)로 설계되었음을 강조한다. 텍스트 모델에 이미지 인코더를 덧붙이는 식의 '덕테이프식' 접근이 아니라, 학습 초기 단계부터 텍스트, 이미지, 오디오, 비디오를 동일한 임베딩 공간에서 처리하도록 설계되었음을 의미한다.

2026년을 향해가는 지금, 이러한 설계 철학은 입력의 통합을 넘어 출력의 통합으로 진화하고 있다. 하사비스는 서로 다른 모달리티(Modality)의 결합이 단순한 기능의 합이 아니라, 지능의 질적 향상을 가져온다고 분석한다. 예를 들어, **텍스트 데이터만으로는 학습하기 어려운 물리적 인과관계나 공간적 추론 능력이 비디오 데이터를 통해 학습됨으로써, 모델의 전반적인 텍스트 추론 능력까지 향상되는 현상이다**.

향후 12개월 내에 Gemini와 같은 모델은 사용자가 어떤 형태로 정보를 던져주든(이미지를 보여주며 말로 질문하거나, 비디오를 입력하고 텍스트로 요약받는 등) 이를 지연 없이 처리할 뿐만 아니라, 그 결과물 역시 텍스트, 이미지, 영상, 오디오가 혼합된 형태로 즉각 생성해낼 것라고 했다. 멀티모달 통합의 핵심은 창의성과 추론의 결합이다.

하사비스는 **텍스트 기반의 논리적 추론이 이미지나 비디오 생성 과정에 개입함으로써, 기존 생성형 AI가 가졌던 환각(Hallucination) 문제를 줄이고 물리적 정합성이 높은 결과물을 만들어낼 것**이라고 예고했다. 이는 단순한 이미지 생성을 넘어, 상황에 맞는(Context-aware) 멀티미디어 생성이 가능해짐을 의미한다.

(3) 시각 지능의 비약적 발전: 'Nano Banana Pro'와 인포그래픽의 정복

하사비스는 구글의 최신 이미지 모델인 'Nano Banana Pro' 모델이 보여주는 시각적 이해 능력이 픽셀 단위의 패턴 매칭을 넘어선, 이미지 내의 의미론적(Semantic) 구조를 파악하는 능력을 의미한다고 했다.

기존 이미지 생성 모델들은 텍스트 렌더링이나 복잡한 도표 생성에 취약했다. 그러나 Nano Banana Pro는 데이터의 논리적 흐름을 시각적으로 구조화하는 '인포그래픽'을 매우 정확하게 생성해낸다. 이는 AI가 텍스트로 된 정보를 이해하고, 이를 다시 시각적 정보로 재구성하여 표현하는 '이중 변환(Double Translation)' 능력을 완벽하게 갖췄음을 보여준다.

특히 추상적인 개념을 구체적인 도식으로 시각화하는 데 있어 인간 전문가 수준에 근접했다. 이는 기업의 보고서 작성, 교육 자료 생성, 과학적 데이터 시각화 분야에서 즉각적인 생산성 혁명을 예고한다.

Nano Banana Pro는 미적인 완성도(Aesthetics)와 정보의 정확성(Accuracy)을 동시에 달성하며, 이는 AI가 디자이너와 데이터 분석가의 역할을 동시에 수행하는 하이브리드 툴로 진화하고 있음을 시사한다.

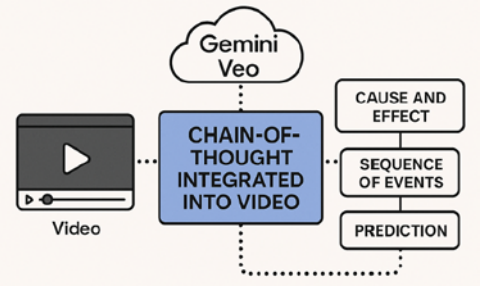
(4) 언어와 영상의 융합: 픽셀을 넘어 개념을 이해하다

하사비스는 비디오 이해(Video Understanding) 분야에서 거대언어모델(LLM)과 비디오 모델의 통합이 가져올 파괴적 혁신을 강조했다. Gemini에게 어떠한 영화의 한 장면을 보여주고 "주인공이 화면에서 하는 행동의 의미가 무엇인가?"라고 물었을 때, 단순히 비디오의 장면을 텍스트로 캡션(Captioning)하는 수준이 아니다.

이는 곧, Gemini(언어모델)와 Veo(비디오 모델)이 통합되어 탄생한 AI는 비디오 데이터에서 시각적 정보뿐만 아니라, 그 이면에 깔린 내러티브, 감정, 문화적 맥락까지 읽어낼 것이라 강조한 것이다.

이에 대해 가장 흥미로운 부분은, **하사비스가 유튜브 영상을 Gemini에 입력하여 처리하는 기능을 언급한 점이다**. 이는 전 세계에 존재하는 방대한 비디오 데이터를 AI가 학습하고, 검색하고, 분석하여 지식으로 변환할 수 있음을 의미하고, **유튜브야말로 구글이 갖고 있는 가장 큰 힘 중 하나라는 점을 방증한다.**

그림 1. Gemini Veo의 Chain-of-Thought
Gemini와 Veo가 통합될 때 탄생하는 AI는 그야말로 “Chan of Frame”의 능력을 갖출 것
시간 축에 따라 비디오 프레임과 생각의 흐름(Chain-of-Thought)이 겹쳐지는 모습을 시각화



자료: 미래에셋증권 리서치센터

그리고 이러한 "데이터 공장"을 활용해, 언어 모델의 추론 능력(Chain-of-Thought)이 비디오 처리에까지 적용하는 것을 구글은 준비하고 있다. CoT가 접목된 비디오 AI 모델은 영상의 인과관계를 단계별로 분석할 수 있게 된다.

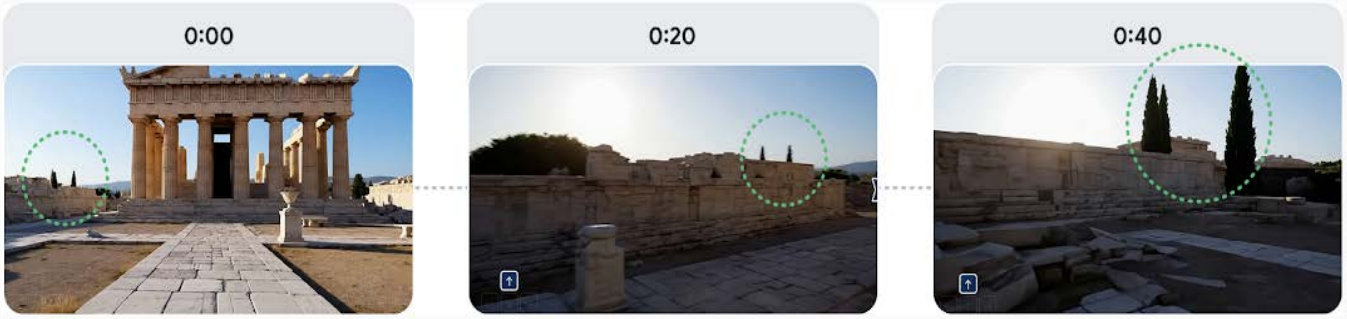
"왜, 어떻게 일어났으며 다음에는 무엇이 일어날 것인가"를 예측하는 능력으로 이어진다는 말이다. 이는 곧 소위 Chain-of-Frame(CoF)이라는 개념으로 생각할 수 있다. 그리고 우리가 살고 있는 현실의 세상도 엄밀하게 말하면 프레임의 연속인 점을 감안하면, 구글이 어떤 것을 준비하고 있는지 쉽게 알 수 있다.

(5) 월드 모델의 대중화: 에이전트는 내년이 더 기대되는 이유

xAI/테슬라의 일론 머스크도 항상 강조하는 개념이지만, 구글의 지향점도 결국 월드 모델이다. 하사비스는 현재 개발 중인 ‘Genie 3’ 시스템을 공개하며, 이를 ‘월드 모델(World Model)’의 본격적인 확산 사례로 꼽았다. Genie 3는 단순한 비디오 생성 모델이 아니다. 사용자가 생성된 비디오 내부로 들어가 마치 게임을 하듯 캐릭터를 조종하고 환경과 상호작용할 수 있는 시스템이다.

기존 비디오 생성 모델은 시간이 지나면 피사체가 일그러지거나 배경이 바뀌는 등 일관성 유지에 실패했다. 그러나 **Genie 3는 "최소 1분 이상" 시뮬레이션의 물리적, 시각적 일관성을 유지한다. 이는 AI가 단순히 다음 프레임을 예측하는 것이 아니라, 가상 세계의 물리 법칙과 공간 구조를 내재화(Internalize)하고 있음을 의미한다**.

그림 2. 텍스트만으로 생성된 구글 Genie 3의 3D 가상 세계 예시
0초, 20초, 40초로 이어지는 장면은 사용자가 가상 공간 내부를 실시간으로 이동하는 모습
시간 지나고 시점이 바뀌어도 나무와 지형(초록색 원)이 그대로 유지되는 것은 AI가 '월드 메모리'를 통해 장기 일관성을 구현했음을 의미



자료: 구글, 미래에셋증권 리서치센터

하사비스는 이를 통해 "몇 시간 만에 상용 수준의 게임을 코딩 없이 만들어내는(vibe-coding)" 미래를 예견했다. 이는 콘텐츠 제작의 장벽을 완전히 허무는 것으로, **텍스트 프롬프트만으로 플레이 가능한 가상 세계를 구축하는 메타버스(Metaverse)의 기술적 완성을 앞당길 것이다.**

우리는 "Everything Everywhere All at Once"라는 자료를 2023년 3월에 발간했던 적이 있는데, 그 당시 우리가 강조했던 것 중 하나는 "AI의 끝은 메타버스"라는 캐치프레이즈였다. **머스크도 하사비스도 만들려고 하는 것은 월드 모델이고, 그 월드 모델에 가장 근접한 게임을 AI로 정복하려고 하고 있는 것이다**. 월드 모델은 단순한 엔터테인먼트 도구가 아니라 AI가 현실 세계의 작동 방식을 이해하고 시뮬레이션할 수 있는 능력을 기르는 훈련장이 되기 때문이다.

그림 3. "Everything Everywhere All at Once" 보고서 표지

자료: 미래에셋증권 리서치센터

그림 4. AI는 메타버스를 밀어내는 것이 아니라 오히려 발전시킬 것

자료: 미래에셋증권 리서치센터

좀 더 사고를 확장하면, AI가 가상 환경에서 물리적 상호작용을 완벽하게 학습한다면, 이는 곧 현실 세계의 로봇 제어, 자율 주행, 과학적 실험 시뮬레이션으로 직결된다는 점을 알 수 있다. 즉, Genie 3의 대중화는 AGI가 '신체성(Embodiment)'을 갖추기 위한 필수적인 단계다.

또한, **데미스 하사비스는 월드 모델로서의 Gemini는 궁극적으로 '유니버설 어시스턴트'로 진화할 것이라고 전망했다**. 이는 사용자의 PC, 스마트폰, 스마트 글래스 등 모든 디바이스에 상주하며, 사용자의 일상과 업무를 능동적으로 보조하는 존재가 된다. **실제로 구글은 2026년에 Gemini 기반 AI 스마트 글라스를 처음 출시하겠다고 공식적으로 밝힌 상태다**. 참고로 하드웨어는 삼성, 젠틀 몬스터, 워비 파커와 협업해서 '패션+테크' 폼팩터로 간다는 전략으로써 메타의 그것과 유사하다.

그림 5. 메타(Meta) x 레이밴(Ray-Ban) 연합군을 잡기 위해 결성한 새로운 'AI 스마트 안경 동맹'
2025년 5월 Google I/O에서 공식화되었으며, 첫 제품은 2026년 출시가 유력

자료: 구글, 미래에셋증권 리서치센터

챗봇이 아니라, 복잡한 다단계 작업을 수행하는 '에이전트(Agent)'를 향후 1년 내에 목격하게 될 것이라고 타임라인을 정한 점은 앞으로의 12개월이 얼마나 흥미로울지 기대하게 하는 대목이다. 이와 관련해 하사비스는 "1년 뒤에는 에이전트가 전체 작업을 신뢰할 수 있는 수준으로 완수할 수 있게 될 것"이라고 단언했다. 이는 에이전트의 Planning 능력과 Self-correction 능력이 임계점을 넘어서고 있음을 시사한다.

하사비스가 이끄는 구글 딥마인드는 그 변화의 최전선에서, 과학적 엄밀함과 압도적인 인프라를 바탕으로 이 거대한 흐름을 주도할 준비를 마쳤다. 이제 우리는 "AI가 무엇을 할 수 있는가"를 묻는 단계를 지나, "AI와 함께 무엇을 정의할 것인가"를 고민해야 할 때다.

2. 구글이 내비친 최근 역량: 선도적인 AI 연구(지속적 학습과 메모리 혁신)

데미스 하사비스가 말하는 다섯 가지 트렌드는 공통적으로 모델이 더 많은 경험을 더 오래 기억하고, 상황에 맞게 즉시 수정하는 능력을 요구한다. 따라서 하사비스가 제시한 향후 12개월의 '제품 로드맵'은 동시에, 지속적 학습과 계층형 메모리를 전제로 한 '아키텍처 로드맵'이기도 하다고 생각한다.

(1) 문제의 본질: "기억상실증"에 걸린 거인들

NeurIPS 2025가 던진 화두: 더 큰 윈도우가 아니라 '지속성'이다

며칠 전 막을 내린 NeurIPS 2025의 현장은 뜨거웠지만, 그 열기의 방향은 작년과는 사뭇 달랐다. 참고로 NeurIPS(뉴립스라고 발음)는 AI 업계에서 가장 권위 있는 학회다. 지난 몇 년간 AI 업계는 "누가 더 긴 컨텍스트 윈도우(Context Window)를 가지는가"를 두고 치킨 게임을 벌여왔다. 100만 토큰, 200만 토큰, 심지어 무한 컨텍스트까지 된다는 모델까지 등장했다.

하지만 이번 **학회에서 전 세계 석학들이 입을 모아 지적한 'Next Big Thing'이자, 현재 AI가 직면한 가장 거대한 병목은 바로 '지속적 학습(Continual Learning)'의 부재였다**. 현재 AI 모델의 가장 큰 한계가 지속적 학습이 안된다는 것이다. 즉, ChatGPT 같은 모델이 사용자와 대화하면서 얻은 정보를 바탕으로 실시간으로 더 똑똑해지거나 성장하지 못하고, 매번 새로운 버전(예: GPT-5.5)으로 완전히 재학습되어야 하는 문제다.

우리가 사용하는 현재의 LLM(거대언어모델)은 아무리 파라미터가 거대해도 본질적으로 '선행성 기억상실증(Anterograde Amnesia)' 환자와 같다. 이와 관련해 뇌과학에서 유명한 사례는, 환자가 과거의 기억(장기 기억)은 온전하지만 방금 나눈 대화(단기 기억)를 뇌에 저장하지 못해 30초마다 세상이 리셋되는 것이다.

이와 똑같이 지금의 AI도 훈련이 끝나는 순간(지식 컷오프) 지적 성장이 멈춘다. 사용자와의 대화는 휘발되고, 어제의 실수를 오늘 교정하지 못하며, 새로운 정보를 주입하면 기존 지식이 파괴되는 '파국적 망각(Catastrophic Forgetting)'의 늪에 빠져 있다.

이것은 단순히 기술적 불편함의 문제가 아니다. 자율 AI 에이전트와 월드 모델이 실험실을 벗어나 현실 세계에 안착하는 데 있어 가장 치명적인 결함이다. 세상은 매초 변하는데 모델은 2025년의 어느 날에 박제되어 있다면, 그 괴리에서 오는 환각과 비효율은 에이전트의 상용화를 가로막는 거대한 장벽이 된다.

구글의 선언: 아키텍처는 곧 최적화(Optimization)다

이 시점에 공개된 **구글 리서치의 'Nested Learning(중첩 학습)'이라는 새로운 패러다임과 이를 구현한 'HOPE'라고 명명된 새로운 모델**은 논문 발표 그 이상의 의미를 가진다.

이때 가장 중요한 인사이트는, Nested Learning은 바로 **지속적 학습의 부재 문제를 해결하기 위해 구글 연구원들이 제시한 패러다임이라는 것이다**. **이는 곧 AI가 정체되지 않고 계속 발전하고 있다는 강력한 증거가 될 수 있다고 본다.**

그림 6. NeurIPS 현장에서 Nested Learning과 지속적 학습에 대해 설명 중인 구글의 알리 베루즈

자료: 알리 베루즈, 구글, Phil Trubey, 미래에셋증권 리서치센터

기존 모델(트랜스포머)에서 아키텍처(정보 처리)와 최적화(학습)는 분리되어 있었다. 입력으로 들어온 정보를 처리하는 아키텍처는 고정된 가중치만을 써야했고, 최적화 작업이 수반되는 사전훈련은 별도의 오프라인(실시간 아님)으로 이뤄져야 했다. 즉, 모델은 단순 계산기이고 더 잘하기 위한 공부는 바깥의 옵티마이저가 대신 시켜준다는 이분법이었다.

그런데, 지속적 학습을 하려면 추론 중에도(온라인으로) 가중치를 계속 수정해야 하는데, 기존 구조에서는 새로운 데이터 하나를 배우기 위해 전체 가중치까지 함께 업데이트 해야 하니 기존 지식이 와르르 무너지는 '파국적 망각'이 필연적이었다.

그런데 **Nested Learning 발표는, " 기억하는 것과 배우는 것은 사실 거의 같은 일이고, 신경망 구조(아키텍처) 자체가 계속 스스로 공부하는 장치가 될 수 있다"라는 기술적 선언문과 같다**. 다시 말해, 지난 수년간 딥러닝 씬을 지배해 온 "고정된 아키텍처 위에다가 최적화 알고리즘을 돌린다"는 이분법적 사고를 파괴하는 혁신이다.

연구진은 딥러닝의 모든 구성 요소를 '연상 기억' 문제로 환원했다. 여기서 연상 기억이라는 것은 사람의 얼굴을 보고 이름이 무엇인지 떠올리는 관계를 mapping 하는 메모리 시스템이다.

어텐션 레이어들로 구성되는 모델 아키텍처가 "텍스트 토큰 데이터를 학습하여 다음 단어를 예측하는 연상 기억 장치"라고 한다면, 이는 곧 '단어 패턴'을 기억하는 메모리다. 그리고 모델을 학습시킬 때 활용되는 옵티마이저(최적화 알고리즘)는 "그래디언트(오차가 어느 방향으로 큰지) 데이터를 보고, 다음에 파라미터를 어떻게 바꿀지 떠올리는 연상 기억 장치"로 볼 수 있다. 즉, '어떻게 공부하면 잘 느는지'를 기억하는 메모리가 옵티마이저다.

결국 둘 다 "어떤 입력 → 그에 맞는 출력"을 기억하는 똑같은 구조이고, 차이는 무엇을 기억하느냐(텍스트냐, 그래디언트냐)뿐이라는 관점이다. 구글이 간파한 이러한 수학적 본질은 지속적 학습에 매우 직접적이고 결정적인 의미를 갖게 된다. **학습(최적화)을 거창한 별도의 과정으로만 볼 게 아니라, 모델 아키텍처 내부의 레이어 하나하나가 사실은 실시간의 작은 최적화 엔진처럼 행동한다는 것이다.** 쉽게 말해, **추론(Inference)하는 매 순간이 곧 학습(Training)이 되어 별도의 재학습 없이, AI가 점 더 똑똑해지는 것이 가능해진다는 주장이다.**

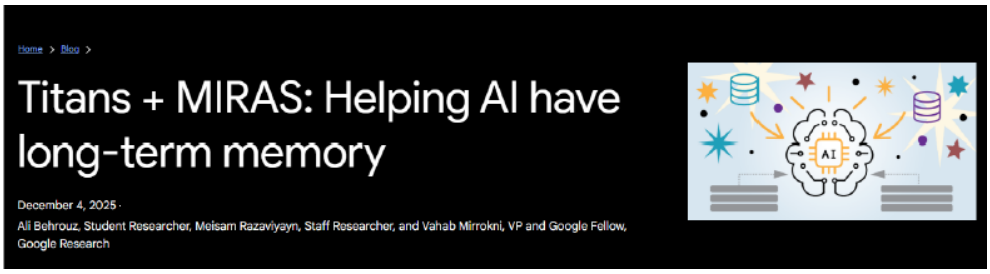
이런 철학을 기반으로 모델 안의 레이어들이 입력을 받을 때마다, 본인의 메모리를 조금씩 업데이트하며 실시간으로 배우게 만들 수 있게 된다. 구글의 이번 혁신은, 뇌가 새로운 정보를 배울 때 뇌 전체를 재부팅하지 않고 관련된 특정 시냅스 연결 강도만 즉각적으로 조절하는 것과 유사하다. **대부분의 기업들이 외부 데이터베이스를 검색하는 RAG(검색 증강 생성)라는 솔루션에 급급할 때, 구글은 인간 뇌의 구조를 모방하여 아키텍처 자체를 뜯어고치는 근원적인 접근을 택했다**. 이것이 구글이 가진 무서운 점이다. 딥러닝의 가장 밑바닥에 있는 수학적 원리를 재해석하는 역량은 여전히 압도적이기 때문이다.

(2) 핵심 엔진: Titans에서 HOPE로 이어지는 진화

Titans: 기억을 '버퍼'가 아닌 '시냅스'에 새기다

Nested Learning이라는 프레임워크를 이해하기 위해서는 먼저 몇 달 전 구글이 발표했던 Titans 논문을 짚고 넘어가야 한다. 이에 대해서는 AI Weekly에서도 소개한 바 있다. **Titans는 '중첩 학습' 연구의 핵심 엔진이자 선행 기술에 해당하고 이번 여러 연구들과 함께 더 다듬어진 것이라 볼 수 있다.** Titans가 던진 충격적인 제안은 "데이터를 메모리 버퍼(KV 캐시)에 띄워두지 말고, 실시간으로 학습해서 가중치(Weights)에 계속 녹여내자"는 것이었다.

그림 7. Titans의 제1저자인 알리 베루즈는 Nested Learning, HOPE 모델에도 기여



자료: 구글, 미래에셋증권 리서치센터

기존의 트랜스포머나 Mamba 같은 모델들은 정보를 'KV 캐시'라고 불리는 메모리 공간에 저장했다. 특히 트랜스포머는 컨텍스트(입출력 길이)가 길어질 수록 기하급수적으로 메모리 사용 요구량이 늘어난다. 따라서, 물리적 메모리(HBM)가 꽉 차면 오래된 정보는 밀려나는 구조적 한계를 갖는다. 이를 비유하자면, 책을 읽을 때 내용을 머리에 넣는 게 아니라, 책상을 넓혀서 책을 펼쳐두는 것과 같은 셈이다. 물론 책상이 꽉 차면, 오래된 책은 치워진다.

반면, **Titans는 추론 시간 학습(TTT: Test-Time Training)이라는 개념을 도입했다. 말 그대로 실시간으로 추론을 할 때마다 학습이 되는 방식**이다. 정보가 들어오면 모델 내부의 작은 신경망(MLP: 피드포워드 레이어)이 실시간으로 경사 하강법(예측의 손실을 줄이는 대표적 알고리즘)을 수행해 그 내용을 파라미터에 업데이트시킨다. 즉, 사용자로부터 받은 입력 시퀀스를 처리하는 과정 자체가 "모델 내부의 파라미터를 실시간으로 훈련시키는 과정"이 되는 것이다.

** 기본 개념: 사전 학습 때 피드포워드 신경망(또는 MLP)의 파라미터에는 세상의 온갖 사실, 문법 규칙, 논리적 관계, 개념 등이 패턴 형태로 새겨진다. 예를 들어 "대한민국의 수도는 서울"이라는 지식은 특정 뉴런들의 연결 강도(가중치)로 저장되는 것이다.*

이 과정에서 핵심 동력으로 작동하는 것이 바로 놀라움(Surprise)이라는 지표다. 인간이 평범한 출근길은 잊어버리고 교통사고가 날 뻔한 순간만 기억하듯, AI도 '정보의 충격량'을 스스로 판단하여 선별적으로 기억한다. 다만 여기서 놀라움은 문과적 표현이 아니라, "예측 오차에 의한 기울기의 크기(Gradient Norm)"를 의미한다. 쉽게 말하면, **모델이 예측하지 못한 정보(높은 Gradient)일수록 "이건 중요한 정보다!"라고 판단해 '깊게 새겨듣겠습니다'라고 판단하고 유의미하게 업데이트 하는 행동 양식이다**. 단계별로 말하면 다음과 같다.

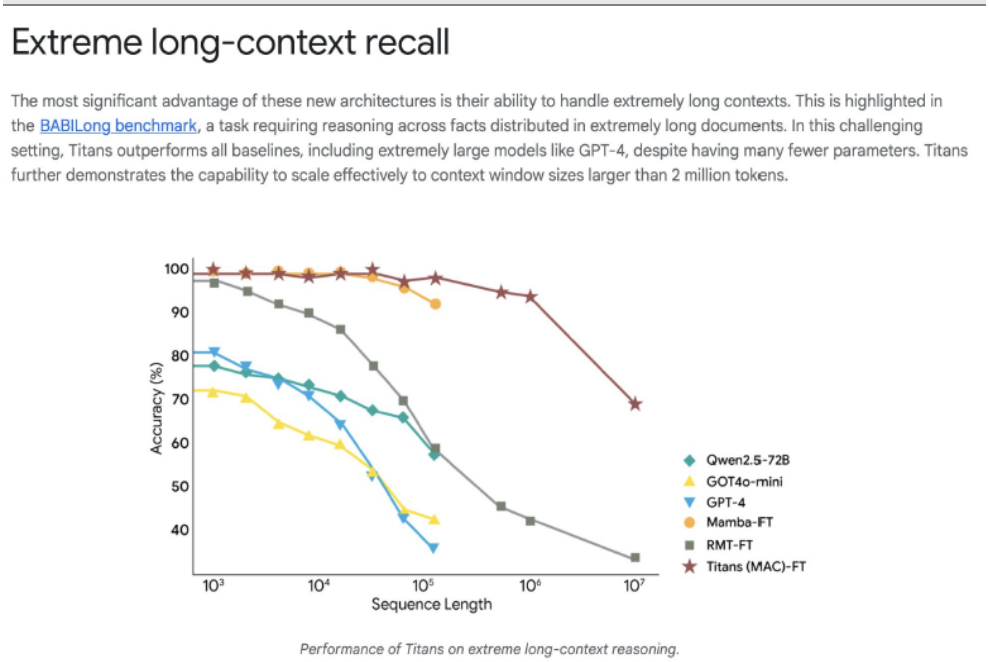
- 모델이 예측하지 못한 정보가 들어오면 손실(Loss)이 커지고, 급격한 그래디언트가 발생

- Titans는 이 높은 그래디언트를 "이것은 학습률(Learning Rate)을 높여야 할 중요한 새로운 정보다!"라는 신호로 해석

- 반면 뻔한 정보(Low Gradient)는 노이즈로 간주하여 가중치 업데이트를 최소화

결과적으로 Titans는 무려 1,000만 토큰이 넘는 방대한 문맥을 처리할 때도 메모리 버퍼를 늘리는 것이 아니라, 모델의 가중치를 정교하게 튜닝하며 정보를 압축 저장한다. 이는 메모리 효율을 극대화할 뿐만 아니라, **단순 검색(Retrieval)이 아닌 AI 모델의 내면화된 지식을 가능케 한다.**

그림 8. BABILong 벤치마크 (매우 긴 문서 곳곳에 흩어진 사실들을 연결하여 추론하는 능력 평가)
다른 모델들은 구조적 한계 혹은 메모리 용량의 한계로 더 긴 문맥을 처리하지 못하지만,
Titans 모델은 10만 토큰에도 정확도 100% 유지.
심지어 1,000만 토큰에 도달해서도 정확도 70%를 유지.
1,000만 토큰은 대략 해리포터 전권 시리즈를 10번 연속으로 읽은 분량.



자료: 구글, 미래에셋증권 리서치센터

중첩 학습과 HOPE: 주파수(Frequency)로 완성된 뇌의 구조

중첩 학습을 토대로 만들어진 Titans가 실시간으로 기억을 쌓는 뉴런이라면, 앞서 말했던 HOPE는 이러한 뉴런들을 조직화하여 만든 뇌의 설계도, 즉 완성된 모델로 이해하면 된다.

표 3. 구글이 내놓은 새로운 패러다임 '중첩 학습'을 이루는 4가지 계층

계층 명칭	분류	핵심 정의
Nested Learning	철학(물리 법칙)	AI 모델의 구조(Architecture)와 학습법(Optimization)은 별개가 아니라, 서로 꼬리를 물고 있는 중첩된 하나의 시스템
MIRAS	이론적 설계도(플랫폼)	Nested Learning 철학을 수학적으로 공식화한 것. 모든 AI 모델을 4가지 부품(기억 구조, 어텐션 편향, 유지 게이트, 학습 알고리즘)의 조립으로 설명
Titans	**핵심 기술(엔진)**	**MIRAS를 바탕으로 만들어진 구체적 아키텍처. 기존 모델들이 정보를 KV Cache에 일시 저장하는 것과 달리, 신경망(MLP) 가중치에 학습시켜 경험이 축적될 수 있게 기억하는 기술을 탑재**
HOPE	플래그십 모델(슈퍼카)	Titans 아키텍처를 기반으로 하되, MIRAS 프레임워크를 이용해 자기-참조+연속체 기억(CMS) 2가지 최고급 옵션을 추가한 모델

자료: 미래에셋증권 리서치센터

구글은 기존 LLM이 '지속적 학습'에 실패하는 원인을 '기억 업데이트 주기의 양극화'에서 찾았다. **기존 트랜스포머에는 단 두 가지 시간 척도만 존재했기 때문이다. 기존 AI 모델은 매 토큰마다 바뀌는 초고속 기억(어텐션 행렬)과, 훈련 후 절대 바뀌지 않는 초저속 기억(사전학습된 가중치)만 존재했다. 기억의 갱신 주기가 극단적**이었다는 말이다.

극단적인 두 상태 사이에 중간 지대가 비어 있었기에, 단기 기억(해마)이 장기 기억(대뇌피질)으로 넘어가는 기억의 공고화 프로세스가 원천적으로 불가능했다. 모델은 '현재'를 경험할 수 있지만, 그 경험을 '장기기억'으로 이관하지 못한다는 말이다. 인간은 수면 중에 이런 작업이 수행되지만 AI는 단기 기억과 장기 기억을 잇는 장치가 없다는 안드레이 카파시의 주장과 정확히 일치한다. 즉, **구글의 해당 연구는 카파시의 고민을 해결할 가능성이 있는 독자적인 연구 결과라서 더 의미가 깊다.**

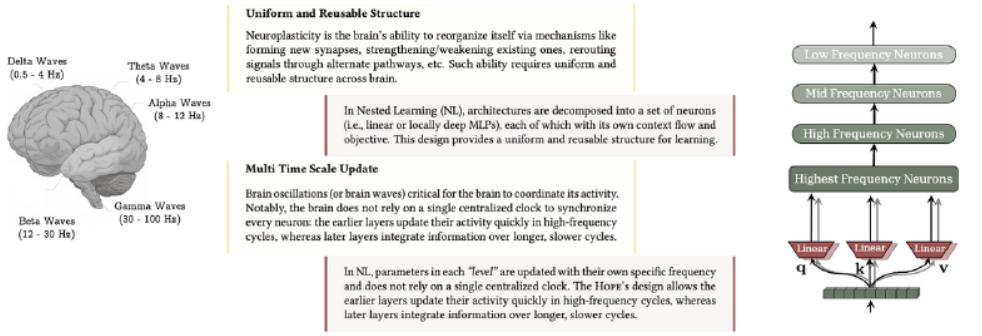
중첩 학습 아키텍처는, 이 문제를 '다중 주파수 업데이트(Multi-Frequency Update)' 개념이라는 다소 생소하게 들리는 방식으로 해결했다. 이를 연속체 기억 시스템(Continuum Memory System), 일명 CMS라 부른다. 초단기 정보는 고주파로 처리하고, 사용자와의 문맥은 중주파로 처리하며, 지속적인 지식은 점차 저주파로 서서히 침전시키는 방식이다. **단기기억이 곧바로 장기 기억으로 넘어가는 것이 아니라, 여러 단계의 주파수 대역을 거치며 서서히 느리게 '장기 기억화'되는 연속체(Continuum)를 만든 것이다. 뇌의 해마가 기억을 대뇌피질로 넘기는 과정을 공학적으로 구현한 셈이다**.

- 고주파 모듈(포스트잇에 기록): 작업 기억(Working Memory). 매 10 토큰마다 신경망을 업데이트. 방금 지나간 차의 색깔 같은 휘발성 정보를 처리. 뻔한 정보(Low Gradient)라서 빨리 처리하고 금방 버림.

- 중주파 모듈(업무 수첩에 기록): 일화적 기억(Episodic Memory). 100~1,000 토큰마다 업데이트. "프로젝트 마감일", "새로 사귄 친구 이름" 같은 맥락을 저장해 1달 정도는 보관.

- 저주파 모듈(비석에 기록): 의미론적 기억(Semantic Knowledge). 매 1,000,000 토큰마다 업데이트. 불변의 진실이나 사용자의 가치관을 저장. 특히 모델의 예측이 크게 빗나간 부분이 있다면(HIgh Gradient), 이 모듈에까지 업데이트.

그림 9. 뇌의 작동 원리와 중첩 학습의 설계도
왼쪽의 뇌는 다양한 주파수(속도)의 뇌파를 사용해 정보를 동시에 처리.
단 하나의 중앙 시계 없이, 여러 시간 단위로 학습이 일어남.
오른쪽의 중첩 학습 구조는 이를 직접적으로 모방한 AI 설계도.
가장 아래의 '최고 빈도 뉴런'은 즉각적인 데이터를 빠르게 처리하는 단기 기억 역할.
정보가 위로 올라갈수록 '저 빈도 뉴런'이 더 오랜 시간에 걸쳐 정보를 통합.



자료: 구글 논문, 미래에셋증권 리서치센터

이 구조가 혁명적인 이유는 '신호 간섭'의 제어에 있다. 예를 들어 기존 모델에서 새로운 학습을 시도하면 새로운 데이터의 그래디언트가 기존 가중치를 덮어써 버리는 '파국적 망각'이 발생했다.

반면, 중첩 학습에서는 새로운 정보의 쓰나미(High Frequency Signal)가 몰려와도, 이는 고주파 모듈에서만 처리되고 소멸된다. 심해에 있는 장기 기억(Low Frequency Module)은 물리적으로 격리되어 있어 영향을 받지 않는다. **기존 모델은 용량이 차면 덮어쓰기를 하지만, HOPE는 정보를 적절한 주파수 대역으로 격리(Isolation)시켜 보존한다**.

정보의 경중에 따라서, 저장되는 물리적 위치(어떤 MLP 블록인지)와 업데이트 속도가 다르기 때문에 서로 간섭하지 않고 독립적으로 존재할 수 있는 것이다. 즉, "새로운 정보를 받아들이면서도, 기존의 중요한 지식은 유지하는" 딜레마를 해결해냈다고 보면 된다. 이것이 HOPE가 다국어 실험에서 새로운 언어를 배워도 모국어를 잊지 않고 성능을 유지할 수 있었던 비결이다. 인간이 새로운 언어를 배운다고 모국어를 까먹지 않는 것과 유사한 메커니즘이 구현되었다!

게다가 HOPE 모델 내부에는 '학습을 조율하는 또 다른 신경망'이 있다. **이 신경망은 현재 들어온 정보가 얼마나 중요한지, 얼마나 오래 기억해야 할지를 판단하여 스스로 학습률(Learning Rate)과 망각률(Retention Gate)을 실시간으로 조절한다.**

기존 모델들은 학습률이나 가중치 감쇠(Weight Decay) 같은 하이퍼파라미터를 사람이 정해준 고정값으로 사용한다. 즉, HOPE는 외부 개발자가 "이건 이렇게 학습해"라고 정해주는 것이 아니라, 모델이 "이 정보는 중요하니 내 뇌 구조를 크게 바꿔서라도 깊이 새겨야겠다"라고 스스로 결정하는 메타인지 능력을 갖춘 것이다. **외부의 개입 없이 모델이 스스로 자신의 학습 규칙을 생성하는 자기 개선의 초기 단계를 보여주며, 이는 AI가 정체기와는 거리가 멀다는 것을 시사한다**.

그림 10. HOPE가 기존 트랜스포머와 근본적으로 어떻게 다른지를 보여주는 설계도
오른쪽의 트랜스포머는 모든 정보를 단일 속도로 처리하는 하나의 거대한 FFN(장기 기억)을 가짐.
왼쪽의 HOPE는 기억 장치가 '고/중/저 빈도'의 여러 계층으로 나뉨.
이것이 다양한 시간 척도로 정보를 공고화하는 '연속체 기억 시스템(CMS)'의 시각적 표현.
HOPE는 인간 뇌처럼 더 동적이고 다층적인, 완전히 새로운 종류의 학습 시스템.

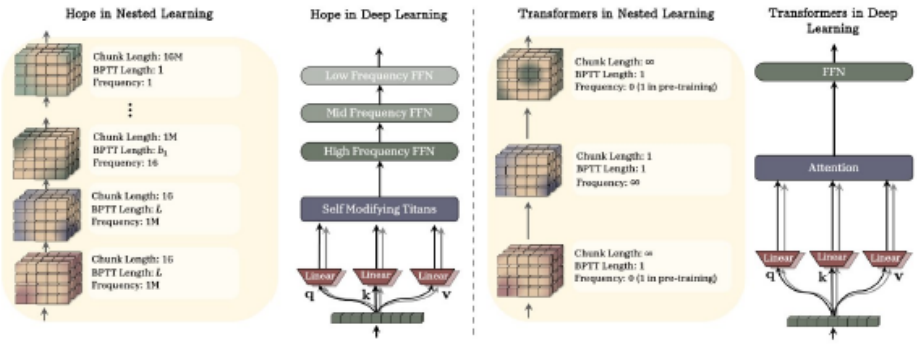


Figure 3: A comparison of Hope architectural backbone with Transformers (Normalization and potential data-dependent components are removed for the sake of clarity).

Table 1: Performance of HOPE and baselines on language modeling and common-sense reasoning tasks. Hybrid models are marked with *.

Model	Wiki. ppl ↓	LMB. ppl ↓	LMB. acc ↑	PIQA acc ↑	Hella. acc_n ↑	Wino. acc ↑	ARC-e acc ↑	ARC-c acc_n ↑	SIQA acc ↑	BoolQ acc ↑	Avg. ↑
HOPE (ours)	26.05	29.38	35.40	64.62	40.11	51.19	56.92	28.49	38.33	60.12	46.90
760M params / 30B tokens											
Transformer++	25.21	27.64	35.78	66.92	42.19	51.95	60.38	32.46	39.51	60.37	48.69
RetNet	26.08	24.45	34.51	67.19	41.63	52.09	63.17	32.78	38.36	57.92	48.46
DeltaNet	24.37	24.60	37.06	66.93	41.98	50.65	64.87	31.39	39.88	59.02	48.97
TTT	24.17	23.51	34.74	67.25	43.92	50.99	64.53	33.81	40.16	59.58	47.32
Samba*	20.63	22.71	39.72	69.19	47.35	52.01	66.92	33.20	38.98	61.24	51.08
Titans (LMM)	20.04	21.96	37.40	69.28	48.46	52.27	66.31	35.84	40.13	62.76	51.56
HOPE (ours)	20.53	20.47	39.02	70.13	49.21	52.70	66.89	36.05	40.71	63.29	52.26
1.3B params / 100B tokens											
Transformer++	18.53	18.32	42.60	70.02	50.23	53.51	68.83	35.10	40.66	57.09	52.25
RetNet	19.08	17.27	40.52	70.07	49.16	54.14	67.34	33.78	40.78	60.39	52.02
DeltaNet	17.71	16.88	42.46	70.72	50.93	53.35	68.47	35.66	40.22	55.29	52.14
Samba*	16.13	13.29	44.94	70.94	53.42	55.56	68.81	36.17	39.96	62.11	54.00
Titans (LMM)	15.60	11.41	49.14	73.09	56.31	59.81	72.43	40.82	42.05	60.97	56.82
HOPE (ours)	15.11	11.63	50.01	73.29	56.84	60.19	72.30	41.24	42.52	61.46	57.23

자료: 구글 논문, 미래에셋증권 리서치센터

표 4. 현재 주류 아키텍처인 트랜스포머와 구글이 내놓은 새로운 패러다임, "중첩학습 기반 모델 HOPE"

비교 항목	트랜스포머 (기존 LLM)	중첩 학습 / HOPE 모델 (새로운 패러다임)
기억 시스템	이분법적 & 단일 속도	연속적 & 다중 속도
세부 내용	**단기 기억 (어텐션): 사용자의 입력(문맥 창)만 처리. 대화가 끝나면 사라지는 휘발성 기억.** **장기 기억 (FFN): 사전 훈련으로 얻은 지식이 고정되어 있음. 모든 지식이 하나의 거대한 저장소에 동일한 방식으로 저장.**	**연속체 기억 시스템(CMS): 기억을 단기/장기로 나누지 않고, 빠른 속도(단기)부터 느린 속도(장기)까지 여러 계층으로 구성.** **정보의 중요도에 따라 각기 다른 속도의 기억 계층에 저장하고 공고화함.**
학습 방식	정적 (Static)	동적 (Dynamic) & 자기-참조 (Self-Referential)
학습 방식 세부	일회성 학습: 사전 훈련이 끝나면 모델의 핵심 지식(가중치)은 봉인됨. 추가학습(파인튜닝): 이때 기존 지식을 잊어버리는 파국적 망각 발생 가능.	지속적 학습 (Continual Learning): 새로운 경험(데이터)이 들어오면, 빠른 기억 계층은 즉시, 느린 기억 계층은 천천히 업데이트하며 평생 학습. 자기-참조: '어떻게 학습할 것인가' 자체를 스스로 학습하고 실시간 수정.
비유	**인쇄가 끝나면 내용 수정이 불가능한 백과사전.**	**새로운 경험을 통해 계속 회로를 바꾸고 성장하는 살아있는 뇌.**

자료: 미래에셋증권 리서치센터

(3) 전략적 함의: 구글이 그리는 'Parametric Memory'의 미래

RAG의 종말과 내면화된 지식의 시대

현재 엔터프라이즈용 AI 아키텍처는 벡터 데이터베이스(Vector DB)를 활용한 RAG(검색 증강 생성)가 사실상의 표준으로 자리 잡았다. 그러나 RAG 기반 벡터 데이터베이스의 최강자 중 한 곳인 Pinecone은 지난달 매각을 모색하고 있기도 했다. 우선, 기술적 관점에서 RAG는 본질적인 한계를 지닌 비모수적(Non-parametric) 메모리 솔루션이다.

RAG는 외부 데이터를 그때그때 필요할 때마다 단순히 검색만 해서 프롬프트에 이어 붙이는 방식이기에, 모델은 정보의 파편을 '참조'할 뿐 그 정보들 간의 복잡한 인과관계를 '학습'하지 못한다. 이는 **시험 중에 오픈북(Open-book)을 허용하는 것일 뿐, 학생 자체의 지능을 높이는 것이 아니라는 말과 같다**. 오픈북을 허용하더라도 학생들의 성적은 각자 천차만별이듯 중요한 것은 정보의 맥락을 정확히 이해하는 지능이지, 정보를 갖고만 오는 것이 아니다.

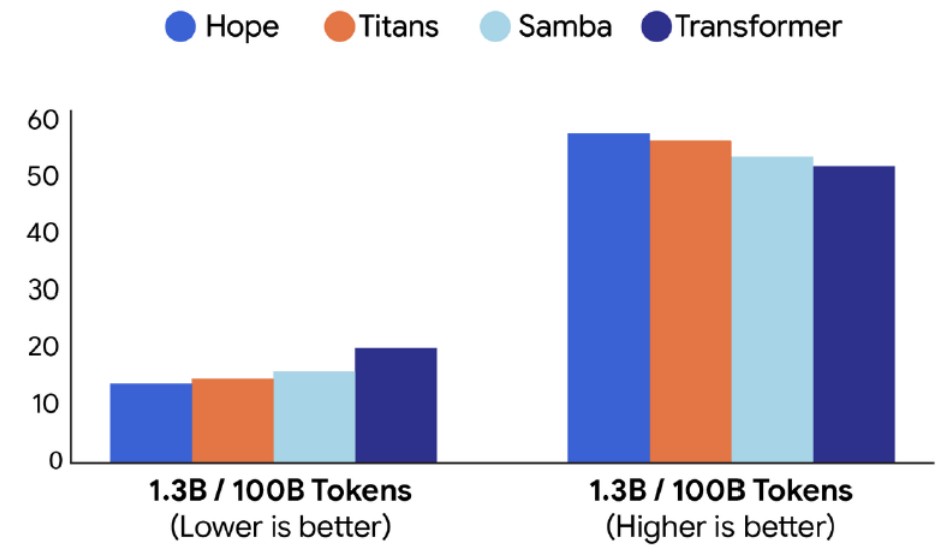
- RAG: "키워드 유사도(Cosine Similarity)에 기반해 문서 A, B를 가져온다." (단순 매칭)

- Nested Learning: "문서 A와 B를 읽고 난 후, 스스로 업데이트한 가중치 공간(AI의 배경 지식 공간)에서, 두 문서 간의 모순점을 추론(Reasoning)한다."

구글의 이번 연구는 외부의 지식을 모델 내부의 가중치(파라미터의 일종), 즉 '모수적(Parametric) 메모리'로 실시간 내재화하는 것이 가능함을 증명했다. 데이터 저장이 아닌 지식의 내면화를 의미하는 것이고, 검색(Retrieval)에서 회상(Recall)으로의 패러다임 전환을 의미한다. 1,000만 토큰 분량의 책을 읽었을 때, RAG는 책의 특정 문장을 찾아낼 뿐이라면 Nested Learning 모델은 방대한 텍스트 정보를 고차원 잠재 공간(Latent Space)에 압축하여 저장한다.

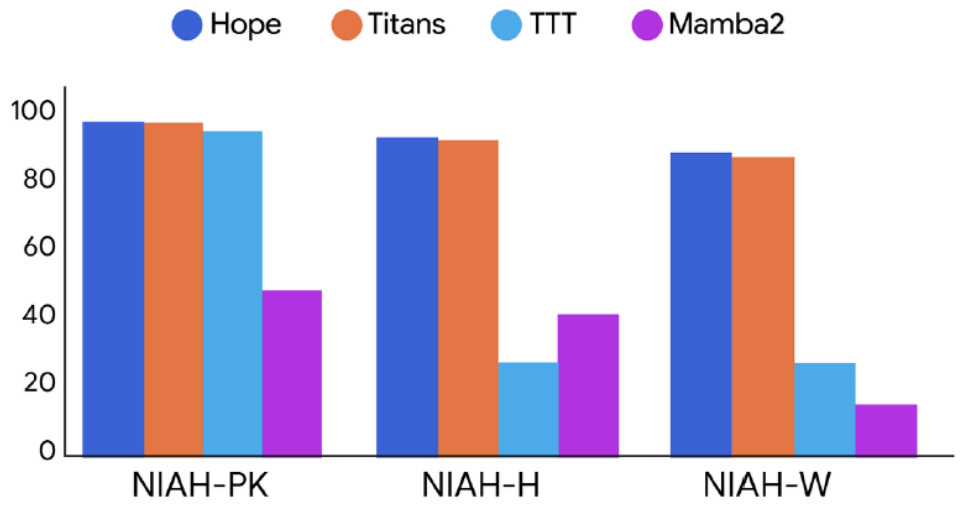
이를 통해 "이 책의 저자라면 이 상황에서 어떤 결정을 내렸을까?"와 같은, 흩어진 정보들을 합성해야만 가능한 복합 추론이 가능해진다. 데이터의 저장이 아니라 '체득화'다. 직관적으로 생각해보면, 휴머노이드 로보틱스에 AI의 이 체득화 특성이 접목될 때의 파급력이 엄청날 것임을 예고하기도 한다.

그림 11. 종합 지능 평가 (언어 이해력 및 상식 추론)
차트는 HOPE 모델의 종합적인 지능이 경쟁 모델들을 어떻게 능가하는지를 보여줌
왼쪽 그래프는 언어 자체의 구조적 이해도를 측정하는 언어 모델링 성능(Perplexity).
여기서 HOPE는 가장 낮은 점수를 기록하며, 가장 뛰어난 언어 예측 능력을 증명.
오른쪽 그래프는 다양한 상식 기반 문제 해결 능력을 평가하는 평균 정확도.
해당 지표에서 HOPE는 가장 높은 점수를 기록하여, 최고의 실용적 추론 능력을 입증.
두 영역 모두에서 최고 수준을 달성한 것은 HOPE가 균형 잡힌 고등 지능을 갖추었음을 의미.
중첩 학습이 단순 암기를 넘어, 깊은 이해와 응용 능력을 모두 향상시킨다는 증거.



자료: 구글, 미래에셋증권 리서치센터

그림 12. 장기 기억력 스트레스 테스트 (NIAH) – 방대한 정보 속에서 핵심을 기억하는 능력
AI 모델의 치명적인 약점인 장기 기억력 상실을 시험한 결과. 을 평가.
HOPE와 Titans는 거의 완벽한 기억력을 유지. 반면, 다른 모델들은 처참하게 실패.
이러한 성능 붕괴는 이들이 새로운 정보에 의해 과거의 중요 정보를 잊는 '파국적 망각' 때문.
HOPE의 압도적인 성공은 연속체 기억 시스템(CMS)의 위력을 보여주는 명백한 증거.



자료: 구글, 미래에셋증권 리서치센터

'나'를 기억하는 에이전트의 대중화

샘 알트만이나 마크 주커버그가 주창하는 1인 1 AI 에이전트 시대의 기술적 병목은 개인화(Personalization)에 있다. 진정한 개인화는 사용자의 발화 스타일, 과거의 실수, 변해가는 가치관 등 동적인 데이터를 실시간으로 반영해야 한다. 지금까지 이를 구현하려면 거대 모델 전체를 미세조정하거나 LoRA 같은 별도의 어댑터를 써야 했지만, 이는 또 다른 데이터 배치(Batch) 처리가 필요하고 비용 문제 때문에 실시간 서비스가 사실상 불가능했다.

하지만 Nested Learning이 적용된 에이전트는 '온라인 학습(Online Learning)'이 가능하다. 사용자와 대화하는 매 순간(Test-Time)이 곧 학습 과정이기 때문이다. 사용자의 데이터는 휘발되지 않고, 모델 내부의 중주파(Medium Frequency) 혹은 저주파(Low Frequency) 메모리 모듈의 가중치를 미세하게 조정한다. 이 과정에서 전체 네트워크는 망가뜨리지 않으면서 오직 사용자와 관련된 특정 시냅스 연결만을 강화한다.

즉, 시간이 지날수록 모델의 가중치 분포가 사용자의 뇌 구조와 동기화되는 사용자 특화 모델로 다듬어지게 된다. 별도의 재학습 파이프라인 없이도 "나를 닮아가고, 나보다 나를 더 잘 이해하는 AI"가 기술적으로, 그리고 비용 효율적으로 가능해짐을 뜻한다.

정적 도구에서 동적 유기체로

우리는 지금 딥러닝 역사에서 훈련(Training, 역전파)과 추론(Inference, 순전파)이라는 둘 간의 장벽이 무너지는 특이점을 목격하고 있다. 구글의 성과는 AI를 공장에서 찍어낸 후 동결시키는 고정된 제품에서, 환경과 상호작용하며 스스로 신경망을 재조립하는 동적인 유기체로 진화시키는 분기점이다.

최근 구글이 제품 상용화 속도(Go-to-Market)에서 경쟁사들에 밀리는 듯한 인상도 있기는 하지만, **Nested Learning과 같은 연구는 구글이 여전히 AI의 근원적 기술 주도권을 쥐고 있음을 방증한다**. 대부분의 빅테크가 이미 존재하는 트랜스포머 아키텍처를 대상으로 자본 경쟁에 몰두할 때, 구글은 "학습이란 무엇인가"라는 질문을 다시 던지며 다음 세대의 아키텍처를 설계하고 있다.

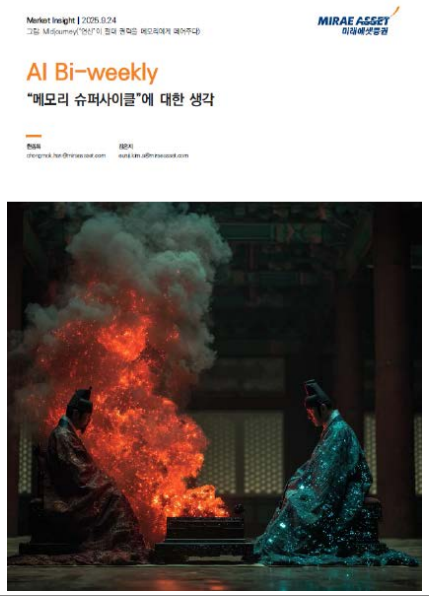
끊임없이 변화하는 리얼 월드를 이해하는 월드 모델(World Model), 그리고 진정한 의미의 AGI는, 과거의 데이터에 갇힌 정적 모델이 아니라, 오늘을 학습해 내일을 예측하는 이러한 구조적 혁신 위에서만 탄생할 수 있다.

3. 지속적 학습과 계층형 메모리 = 슈퍼사이클

(1) 구글의 메모리 주파수와 계층형 메모리의 데칼코마니

구글의 이번 발표는 우리가 지난 9월과 10월 리포트에서 다루었던 AI 반도체 시장의 변화(메모리 대역폭과 계층형 메모리로의 전환)가 참(True)임을 증명하는 수학적 증거라고 생각한다. 결론부터 말하자면, 구글이 소프트웨어적으로 정의한 주파수(Frequency)와 반도체 업계가 하드웨어적으로 구현하려는 계층(Tier)은 본질적으로 동의어라고도 볼 수 있다.

그림 13. "메모리 슈퍼 사이클"에 대한 생각 (지난 9/24 리포트)



자료: 미래에셋증권 리서치센터

그림 14. 에이전트 문제는 메모리가 해결한다 (지난 10/24 리포트)

자료: 미래에셋증권 리서치센터

엔지니어링 관점에서 중요한 통찰은, HOPE 모델이 내부적으로 생성하는 메타 파라미터인 학습률(에타)과 망각률(람다)이 더 이상 단순 수학적 상수가 아니라는 점이다. 이것은 시스템 레벨에서 데이터가 어디에 위치해야 하는지를 결정하는 강력한 제어 신호(Control Signal)로 해석되어야 한다.

즉, **알고리즘이 "이 정보는 중요해서 가중치를 크게 바꿔야 한다(높은 학습률)"고 신호를 보내면, 하드웨어는 이를 "이것은 Hot Data이니 당장 HBM에 둬야 한다"라고 해석하는 데칼코마니가 형성되는 것이다. 물론 모델이 판단할 때, "이 정보는 나중에 참조만 하면 된다." 정도로 해석되는 낮은 학습률이 필요한 데이터는 Cold Data (SSD 오프로딩)에 저장하면 되는 것이다.**

(2) HBM: 연산 밀도의 폭발 (고주파: 빠른 업데이트)

구글에서 말하는 고주파/중주파/저주파 모듈들은, 각자의 리듬대로 정보를 업데이트(실시간 학습)한다. 그런데 물리적인 연산(업데이트)은 오직 GPU와 HBM에서만 가능하다는 점을 명심해야 한다. 가중치를 업데이트하려면 미분(기울기 산출)을 계산하고 더하기/빼기를 해야 하는데, 이 고난도 수학 연산은 오직 GPU만 할 수 있다. GPU의 손은 HBM까지만 닿는다.

표 5. 구글의 중첩 학습에 기반해서 계층형 메모리의 역할을 나눠보았다

하드웨어 계층	소프트웨어 모듈	저장 데이터 유형	업데이트 주기
HBM (Hot Tier)	고주파 모듈(초단기 작업 기억)	활성 가중치: 실시간으로 변하는 뉴런 메타 파라미터: 학습률과 망각률 현재 입력: 처리 중인 토큰 활성 KV 캐시: 당장 연산에 필요한 직전 문맥	매 토큰 생성 시마다 실시간 수정 (RMW)
CXL DRAM (Warm Tier)	중주파 모듈(에피소드/맥락 기억)	맥락 가중치: 현재 대화 및 문서의 요약 정보 Warm KV 캐시: 윈도우를 벗어난 최근 대화 활성 LoRA 어댑터: 현재 작업에 필요한 전문가 모듈	문장이나 문단 단위로 간헐적 업데이트
NAND eSSD (Cold Tier)	저주파 모듈(장기 지식/"자아")	지식 가중치: 사전 훈련된 불변의 지식 아카이빙 된 메모리: 과거의 모든 대화 로그 벡터 DB: RAG를 위한 외부 지식 데이터셋	거의 안변하거나, 매우 드물게 대규모 갱신

자료: 미래에셋증권 리서치센터

먼저, **Nested Learning의 최전선인 고주파 모듈은 매 토큰마다 갱신된다**. 이곳에서는 입력 토큰을 실시간으로 분석하여 학습률(에타)과 망각률(람다)을 생성하는 초고속 신경망이 돌아간다. 이 모듈은 시스템 전체의 가장 좁은 병목이다. 이 계산이 끝나지 않으면 다음 토큰 생성이 불가능하기 때문이다.

여기서 하드웨어에 대한 요구사항은 명확하다. 현재의 PCIe 대역폭으로는 어림도 없다. 매 토큰마다 발생하는 가중치 읽기/쓰기(Read/Write) 트래픽을 감당하기 위해서는, **데이터가 반드시 프로세서(GPU)와 물리적으로 가장 가까운, TB/s급 대역폭을 가진 HBM에 상주해야 한다.**

특히 주목해야 할 기술적 변화는 트래픽의 성격이다. 기존 트랜스포머의 KV 캐시는 정보를 쌓아두고 참조(Read)만 하는 'Append-only' 방식이었다. 이미 저장된 정보나 모델의 가중치 자체를 수정(Update)하지는 않기 때문이다.

반면, **Titans와 HOPE의 단기 기억은 입력 데이터에 반응해 읽고, 계산하고, 다시 쓰는(Read-Modify-Write) 작업을 매 토큰마다 수행해야 한다. 이는 메모리 트래픽을 기존 대비 2배 이상 폭증시킨다**.

따라서 HBM의 대역폭 요구량은 계속해서 더 늘어난다. 향후 HBM 시장의 경쟁 포인트가 단순한 용량이 아니라, 이러한 연산 집약적 트래픽을 감당할 수 있는 '대역폭과 쓰기 속도(Latency)'로 이동하는 이유가 바로 여기에 있다. HBM이 속도 경쟁으로 가는 것은 선택이 아니라 필연이다.

(3) CXL DRAM(CMM과 CAMM): 지연 시간 은닉 (중주파: 낮은 레이턴시)

중주파 모듈은 청크(Chunk: 예를 들어 몇 백 ~ 몇 천 토큰 정도의 데이터 뭉치로 해석) 단위로 업데이트된다. 대화의 맥락이나 에피소드를 저장하는 이 기억들은 매 토큰마다 급하게 호출될 필요는 없지만, 수십 GB에서 수백 GB에 달하는 크기 때문에 단일 GPU의 HBM에 다 담기에는 부담스럽다. 여기서 **CXL(Compute Express Link) DRAM이 구원투수로 등판할 수 있다. 이 계층의 핵심 기술은 프리페칭(Prefetching)이다**.

구글에서 정의한 중주파 모듈은 매 토큰마다 호출되지 않으므로, 연산 시점에서 어느 정도 예측(Prediction)이 가능하다. 즉, 시스템은 "어, 청크가 거의 다 찼네? 50ms 뒤에 중주파 업데이트가 일어나겠군."이라고 미리 알 수 있는 것이다. 이 **예측 가능성 덕분에, 시스템은 GPU가 현재 데이터 청크를 처리하는 동안, 유휴 상태인 CXL 대역폭을 활용해 다음 데이터 뭉치를 DRAM(외부)에서 HBM(내부)으로 미리 복사해 둘 수 있다**.

** 예시: "512번째 토큰 처리가 끝나면, 반드시 중주파 기억(Medium Memory) 업데이트가 일어난다"는 것이 알고리즘(Nested Learning)에 정해져 있음. 시스템은 GPU가 400번째~512번째 토큰을 열심히 연산하는 동안, 놀고 있는 CXL 대역폭을 이용해 다음에 필요한 데이터를 DRAM(외부)에서 HBM(내부)으로 미리 Copy.*

다시 말해, Nested Learning 알고리즘이 '나 곧 이거 필요해!'라고 미리 신호를 보내주기 때문에, 하드웨어는 연산이 멈추지 않게 데이터를 미리미리 배달해 놓을 수 있는 것이다. 비록 CXL DRAM의 물리적 레이턴시(약 100ns)는 HBM보다 느리지만, 청크 단위의 연산 시간(수 ms) 뒤에 이를 숨길 수 있는 지연 시간 은닉(Latency Hiding)이 가능해진다. **CXL은 단순 메모리 확장이 아니라, 알고리즘의 호흡에 맞춰 데이터를 적시에 공급하는 '스마트 파이프라인'이 되는, 상급지로의 이동이 발생할 수 있다.**

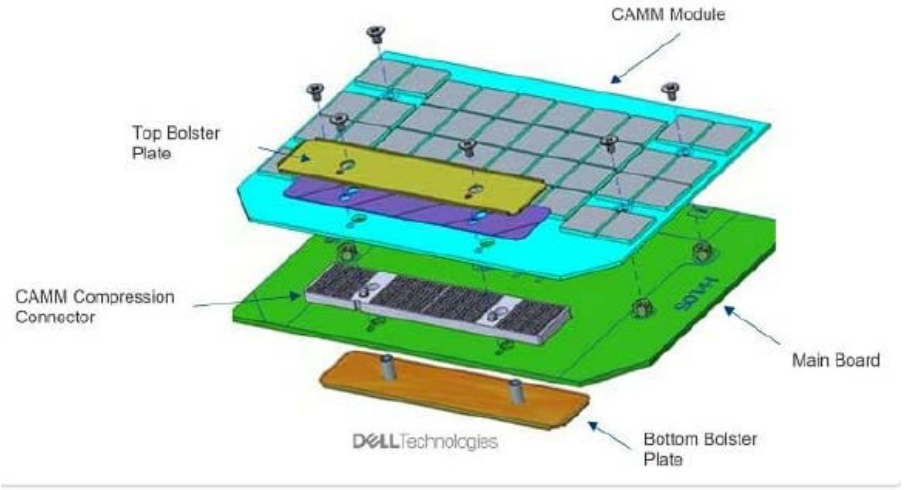
그렇다면 이 '스마트 파이프라인'을 구성하는 **물리적 실체는, 바로 SOCAMM(Small Outline Compression Attached Memory Module)일 가능성이 높다고 사료된다**.

그리고, 우리가 이론적으로 그린 '중주파 모듈용 대용량 버퍼'는 현실은, 지난 11월 SK하이닉스가 밝힌 'AI-D B (Breakthrough)'라는 이름의 제품군으로 구체화될 수도 있다고 본다. AI-D B 제품은 128GB SOCAMM2(LPCAMM2) 모듈을, 하나의 블레이드(서버)에 16개를 꽂아 2TB 용량을 구현한 CXL 메모리 시스템이다.

** SOCAMM: 기존 노트북 메모리(SO-DIMM)의 물리적 한계를 극복하기 위해 탄생한 차세대 메모리 모듈. JEDEC 표준 기준으로는 CAMM2/LPCAMM2으로 불림. 기존 메모리가 메인보드 슬롯에 수직으로 꽂혀 신호 경로가 길고 노이즈에 취약했던 것과 달리, SOCAMM은 압축 커넥터를 통해 메인보드에 수평으로 밀착 연결되어 프로세서와의 신호 거리를 극단적으로 단축. 이를 통해 LPDDR의 장점인 고속(9.6Gbps)과 저전력 효율을 유지하면서도, 기존 LPDDR의 치명적 단점인 '교체 불가' 문제를 해결해 용량 확장성을 확보. 고성능 AI PC의 필수 부품이자 전력 효율이 중요한 차세대 AI 서버의 핵심 메모리로 급부상 중*

Nested Learning이 요구하는 청크 단위의 대규모 데이터 이동을 감당하려면, 기존 DDR5의 전력 소모(약 25W)는 부담스러운데, SOCAMM2는 LPDDR 기반으로 전력은 절반(약 10W)이면서 속도는 2배(9.6 GT/s) 빠르다. 즉, **구글의 알고리즘이 "데이터 미리 보내줘!(프리페칭)"라고 신호를 보내면, 하드웨어 단에서는 저전력/고속의 SOCAMM 군단이 CXL 패브릭을 타고 데이터를 HBM 앞으로 배달하게 될 것 같다**. SOCAMM은 노트북용뿐만이 아닌, 전체 DRAM 수요의 중요 일부를 차지할 수 있는 이유도 여기에 있다.

그림 15. 슬롯을 없애고 압축 커넥터(Compression Connector)를 사용하는 CAMM의 구조도
기존은 슬롯에 끼워지는 방식, CAMM은 메인보드와 모듈 사이에 커넥터를 두고 나사로 조여 압착.
위아래의 금속 플레이트(Bolster Plate)가 모듈을 단단히 고정시켜, 안정적인 전기적 연결을 보장.
구조의 핵심은 CPU와 메모리 사이의 신호 경로를 수직으로 최단거리에 가깝게 연결할 수 있다는 점.

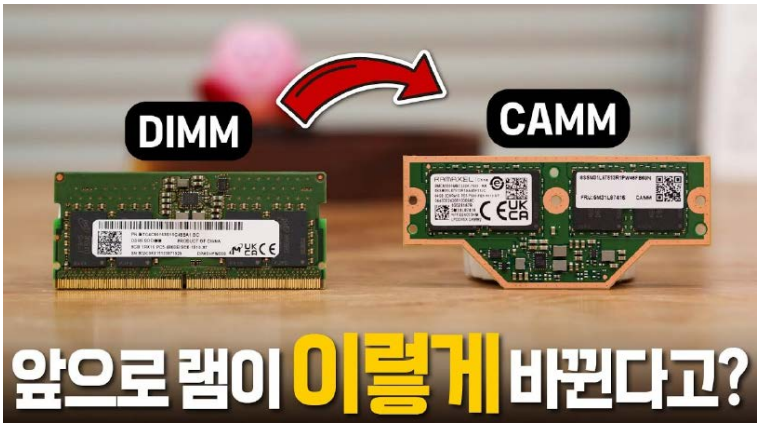


자료: Dell, 미래에셋증권 리서치센터

그림 16. 메인보드 상에서 SO-DIMM 대비 LPCAMM(SOCAMM)이 가져오는 극적인 공간 효율성
절약된 공간은 디바이스 전체의 설계를 바꿀 수 있는 여유 공간이 생겼음을 의미.

자료: 삼성전자, 미래에셋증권 리서치센터

그림 17. 20년 넘게 쓰인 표준인 SO-DIMM(왼쪽)과 AI 시대를 위한 새로운 표준 CAMM(오른쪽)
SO-DIMM은 슬롯에 세워서 꽂기 위한 길쭉한 막대 형태, CAMM은 넓고 얇은 판형
메인보드에 수평으로 밀착되기 위해 나사 구멍이 뚫려 있음. 땜질(Soldering) 없이 탈부착이 가능.



자료: 유튜브(@JN테크리뷰), 미래에셋증권 리서치센터

(4) SSD: 스토리지는 이제 어느정도 연산도 해야 돼 (저주파: 초거대 병렬 I/O)

저주파 모듈은 아주 가끔 대규모로 업데이트되는 만큼, 1PB(페타바이트) 급의 거대한 지식 저장소다. 위키피디아 전체나 과거의 모든 대화 로그가 여기에 해당한다. 이 거대한 데이터를 RAM에 올리는 것은 경제적으로/물리적으로 비현실적이다.

하지만 다 올릴 필요도 없다. 사용자들의 질문에 답하기 위해 필요한 건 그중 0.001%의 활성화된 가중치뿐이다. **업데이트는 드물게 일어나도 참조는 수시로 일어나기 때문에, 여기서 희소 접근(Sparse Access)이 핵심 키워드가 된다.**

따라서 페타바이트 데이터 전체를 로딩할 필요 없이, 필요한 블록만 NVMe SSD에서 읽어오면 된다. 최신 eSSD(PCIe Gen5)는 순차 읽기 속도가 14GB/s에 달하며, 수천 개의 병렬 I/O 요청(Queue Depth)을 처리할 수 있다. 물론 현재의 I/O(Random Read) 처리량으로는 아직 부족하고, SSD에서도 계속되는 혁신이 요구될 것으로 보인다.

여기서 **SSD의 역할이 더 중요해질 이유는, 데이터를 단순히 꺼내오는 게 아니라 SSD 컨트롤러 레벨에서 기본적인 필터링이나 벡터 검색을 수행한 뒤, 정말 필요한 데이터만 GPU로 보내야 하기 때문이다**. 따라서, Nested Learning의 저주파 모듈은 사실상 "거대한 벡터 DB가 내재화된 신경망"이다.

이로써 SSD는 단순 저장소가 아니라, 모델의 파라미터를 저장하고 서빙하는 실시간 추론 장치의 일부로 입지가 격상될 수 있다. GPU가 SSD에게 주소를 주고 '어디서부터 어디까지 있는 데이터 다 갖고 와'의 주소 기반 읽기 방식이 아니라 **질문(Query)을 던지면, SSD가 자체적으로 벡터 검색을 수행해 정답(Value)만 올려보내는 시대가 온 것이다**.

(5) 연산의 핵심 파트너로 격상되는 질적 변화

지금까지 시장은 메모리 반도체 산업을 용량(Capacity)과 가격(Price)의 함수로 주로 해석해왔다. HBM의 공급 부족이나 D램 가격의 등락에 일희일비했다는 뜻이다. 그러나 구글의 **Nested Learning이 시사하는 바는, 메모리가 확연히 Qualitative Shift를 경험할 수 있다는 가능성을 열었다. 이것이 바로 이번 메모리 슈퍼사이클이 과거와는 차원이 다른 이유다**.

기존 트랜스포머 추론은 메모리 대역폭을 '읽기'에만 썼다. 하지만 Nested Learning 기반의 HOPE 모델이 가진, 고주파 모듈(단기 기억)은 매 토큰마다 'Read-Modify-Write'를 수행한다. 현재 가중치를 읽고, 그래디언트를 계산한 뒤, 업데이트된 가중치를 다시 HBM에 써야 한다.

해당 작업으로 메모리 트래픽 부하가 기존 대비 2배 이상 폭증하며, 쓰기 작업(Write)이 읽기 작업을 블로킹(Blocking)하는 현상도 발생한다. 이 때문에 **HBM4부터 도입되는 베이스 다이(Base Die)의 로직 공정(5nm 등) 전환이 필수가 된다.**

단순 I/O가 아니라, HBM 내부에서 아주 **간단하더라도 어느정도의 가중치 업데이트 연산(Add/Decay)을 처리해줘야 한다. 이는 곧 PIM(Process In-Memory)의 역할인 셈이다**. 이렇게 해야 GPU와 HBM 사이의 병목을 줄일 수 있기 때문이다.

그림 18. SK하이닉스는 TSMC와 협력해 차세대 HBM4 베이스 다이를 설계
TSMC N5 공정의 1.55배 높은 로직 성능은 HBM 베이스 다이가 단순 버퍼링을 넘어, 메타 파라미터(학습률 등)연산을 수행할 수 있는 '연산 파트너'로 격상되었음을 의미. 면적(0.39x)과 전력(0.35x)의 획기적 절감은 HBM이라는 좁고 뜨거운 공간 안에 지능형 로직(ALU)을 심어도 발열 제어가 가능하다는 물리적 여유를 보여줌.

TSMC's Production Nodes for HBM4 Base Dies

null	N12FFC+	N5
Area	1X	0.39X
Logic GHz @ power	1X	1.55X
Power @ GHz	1X	0.35X

자료: Tom's Hardware, 미래에셋증권 리서치센터

구글이 자체 TPU에 맞춰 HBM 스펙을 커스터마이징하려는 이유도 여기에 있다고 사료된다. 이로써 **표준 JEDEC 스펙보다 쓰기(Write) 대역폭을 강화한 맞춤형 HBM 수요가 폭발할 가능성이 높다.** HBM은 단순한 데이터 탱크가 아니라, 실시간 학습 연산의 일부를 담당하는 가속기(Accelerator)의 영역으로 진입하고 있다.

그림 19. 구글 Custom silicon solutions를 위해 메모리 시스템 아키텍트를 채용한다는 경력직 공고

메모리 시스템 설계자, 실리콘

Google 미국 캘리포니아주 마운틴 뷰 중간

적용하다

최소 자격:

- 전기공학, 컴퓨터공학, 컴퓨터 과학 또는 관련 분야 학사 학위 또는 이에 상응하는 실무 경험.
- 마이크로아키텍처 및 ASIC 블록 설계 분야에서 8년의 경력을 보유하고 있습니다.
- 중앙 처리 장치(CPU), 그래픽 처리 장치(GPU), 캐시, 메모리 관리 장치(MMU) 또는 Coherent Fabrics 등 1개 이상의 블록에 대한 RTL(Register-Transfer Level)을 설계/구현한 경험이 있습니다.
- 마이크로 아키텍처/설계 성능 분석, 도구 및 시뮬레이터에 대한 경험.

선호하는 자격:

- 전기공학, 컴퓨터공학 또는 컴퓨터과학 분야의 석사 또는 박사 학위 소지자 (컴퓨터 아키텍처 전공).
- 구축 시간 내에 구성 가능한 설계를 구축한 경험이 있습니다.
- 코히런트 인터커넥트, 캐시, 메모리 시스템 중 하나 이상의 분야에 대한 지식.
- 시스템 버릴로그, 버릴로그와 같은 하드웨어 기술 언어(HDL)에 대한 지식.

자료: 구글, 미래에셋증권 리서치센터

그림 20. 마이크로소프트의 "Principal Memory Controller Architect" 공고.
LPDDR, HBM, CXL 등 여러 메모리 기술을 아우르는 메모리 컨트롤러 아키텍처 설계인력 찾는 중
대형 플랫폼 기업들이 메모리 컨트롤러/PHY 설계 인력을 내부에 확보하고 있음을 시사

주요 메모리 컨트롤러 설계자

마이크로소프트 - 약 1년 전에 게시됨 $ 161,600 - $ 286,200 /년

정규직 · 교장 툴리, 노스캐롤라이나주 출판 산업

직위에 관하여

수석 메모리 컨트롤러 아키텍트는 마이크로소프트 실리콘 엔지니어링 조직 내에서 메모리 컨트롤러 아키텍처를 정의하는 데 핵심적인 역할을 담당합니다. 이 직책은 다양한 팀과의 협업을 통해 마이크로소프트 클라우드 인프라를 위한 맞춤형 실리콘 설계에서 고성능, 효율성 및 안정성을 보장하는 업무를 포함합니다. 아키텍트는 다양한 메모리 기술에 맞는 메모리 컨트롤러를 개발하고 시스템 온 칩(SoC) 제품의 혁신을 주도할 책임이 있습니다.

책임

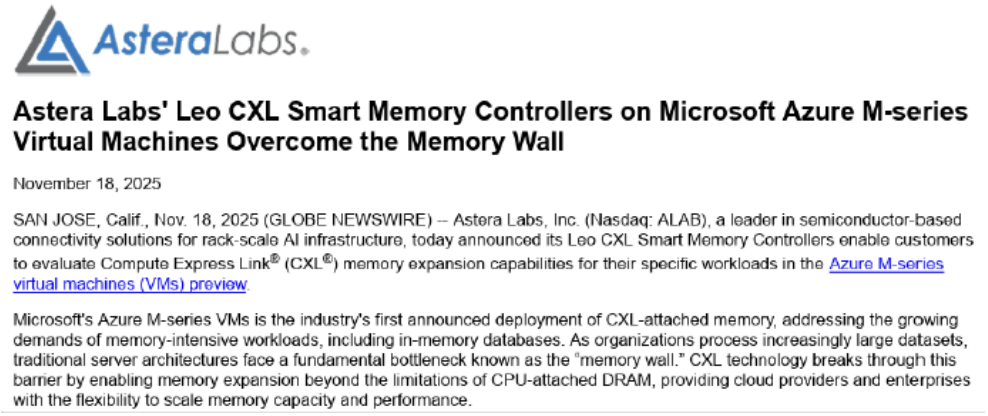
- 다양한 메모리 기술에 맞는 메모리 컨트롤러를 설계하고 개발합니다.
- DDR, LPDDR, HBM, Type 3 CXL 기반 메모리, RDIMM/MRDIMM 및 차세대 메모리 기술을 포함한 메모리 기술 로드맵을 검토하고 분석합니다.
- 메모리 컨트롤러 마이크로아키텍트 및 검증/확인 팀과 협력하여 새로운 기능을 제품화합니다.
- 벤더와 협력하여 지적 재산(IP)을 평가하고 통합 방안을 제안합니다.
- 성능 모델링 팀과 협력하여 메모리 컨트롤러에 대한 사이클 근사 모델을 개발하고, 특히 Azure 워크로드 결과를 활용하여 시스템 온 칩(SOC)/플랫폼 성능을 분석합니다.
- 전략적 계획, 아키텍처 및 내부 고객과 협력하여 워크로드와 사용 사례 요구 사항을 파악하고 메모리 제품 구조 내에서 풀스택 최적화 기회에 중점을 둡니다.

자료: 마이크로소프트, 미래에셋증권 리서치센터

한편, HBM은 너무 비싸고, SSD는 너무 느려서 격차를 메우기 위해 CXL(Compute Express Link)이 활용될 수 있지만, 그동안은 킬러 앱(Killer App) 부재에 시달렸다. Nested Learning의 중주파 모듈이 바로 그 킬러 앱이 될 수 있다.

위에서 설명한 Prefetching Buffer(미리 가져오기 버퍼) 역할을 CXL DRAM이 수행하면, 하드웨어는 현재 연산이 도는 동안 CXL 대역폭을 풀로 가동해 SSD나 원격 메모리에서 데이터를 HBM으로 미리 옮겨놓을 수 있다.

그림 21. 킬러 앱 부재에 따라 시장 성숙도에서 비판을 받았던 CXL... 최근 클라우드에서 활용 증대
최근 Astera Labs는 Azure의 가상 머신에서 CXL 메모리를 적용해 최대 2 TB의 메모리 확장 제공
CXL이 대형 데이터베이스·AI 추론·KV 캐시 등에 적합하다고 발표



AsteraLabs.

Astera Labs' Leo CXL Smart Memory Controllers on Microsoft Azure M-series Virtual Machines Overcome the Memory Wall

November 18, 2025

SAN JOSE, Calif., Nov. 18, 2025 (GLOBE NEWSWIRE) -- Astera Labs, Inc. (Nasdaq: ALAB), a leader in semiconductor-based connectivity solutions for rack-scale AI infrastructure, today announced its Leo CXL Smart Memory Controllers enable customers to evaluate Compute Express Link® (CXL®) memory expansion capabilities for their specific workloads in the Azure M-series virtual machines (VMs) preview.

Microsoft's Azure M-series VMs is the industry's first announced deployment of CXL-attached memory, addressing the growing demands of memory-intensive workloads, including in-memory databases. As organizations process increasingly large datasets, traditional server architectures face a fundamental bottleneck known as the "memory wall." CXL technology breaks through this barrier by enabling memory expansion beyond the limitations of CPU-attached DRAM, providing cloud providers and enterprises with the flexibility to scale memory capacity and performance.

자료: Astera Labs, 미래에셋증권 리서치센터

이러한 구조 하에서는 여러 GPU가 하나의 거대한 CXL 메모리 풀을 공유하며, 각자의 중기 기억을 동적으로 할당받는 Memory Pooling이 가능해져 TCO(총소유비용)를 낮출 수 있다. 또한 GPU가 HBM 내용을 수정했을 때 이것이 CXL 메모리에도 즉각 반영되어야 하므로, CXL 3.0의 캐시 일관성(Cache Coherency) 기능이 중요해진다. 즉, CXL DRAM이 Nested Learning의 리듬을 하드웨어적으로 지탱하는 필수 허리(Backbone)가 될 수도 있다.

마지막으로 SSD는 엔지니어들에게 그동안 '체크포인트 저장용' 혹은 '데이터셋 로딩용'에 불과했다. 느린 속도(Latency: ~100s) 때문에 연산의 핵심 경로(Critical Path)에 끼어들 수 없었기 때문이다. 하지만 Nested Learning은 이 패러다임을 뒤집는다. HOPE 모델의 저주파 모듈(장기 기억)은 페타바이트(PB)에 달한다. 중요한 것은 모델이 매 순간 이 모든 데이터를 로딩하지 않는다는 점이다.

기존에는 모델을 램에 올리기 위해 순차 읽기(Sequential Read) 성능이 중요했지만, 이제는 랜덤 읽기(Random Read) 성능이 핵심이 된다. 모델이 추론 중 "어? 예전에 봤던 그 논문 내용이 필요한데?"라고 판단하면, SSD의 특정 주소에 있는 파라미터 블록(Block)만을 핀포인트로 호출해야 하기 때문이다. 이는 SSD에게 데이터베이스(DB)와 유사한 높은 IOPS(초당 입출력 횟수)를 요구하게 된다.

이 지점에서 GPUDirect Storage(GDS)와 Computational Storage 기술이 필수가 된다. CPU를 거쳐서 데이터를 복사하면 레이턴시 때문에 실시간 처리가 불가능하다. 따라서 GPU가 PCIe 레인을 통해 NVMe SSD의 데이터를 직접 DMA(Direct Memory Access)로 꽂아 넣는 기술이 기본값이 된다. 더 나아가, **SSD 컨트롤러 단에서 "필요한 벡터(Key)와 유사한 데이터만 필터링해서 보내는" 전처리 기능이 요구된다. SSD는 GPU의 연산을 보좌하는 '가장 느리지만 가장 거대한 L4 캐시'로 승격되는 것이다.**

(6) 범용(General)의 종말, 전용(Custom)의 시대

컴퓨터 공학에는 수십 년간 지켜져 온 불문율이 하나 있었다. 바로 "소프트웨어는 하드웨어의 물리적 특성을 몰라도 된다"는 추상화(Abstraction)의 원칙이다. CUDA가 복잡한 메모리 주소를 가상화해주고, OS가 페이지 관리를 알아서 해줬기에 개발자들은 알고리즘에만 집중하면 됐다.

하지만 구글의 Nested Learning(HOPE/Titans)은 오랜 불문율을 깬다. 알고리즘이 직접 생성한 본인의 '학습률'과 '업데이트 빈도(주파수)'라는 메타 정보가, 물리적 메모리의 티어와 메모리 컨트롤러의 로직을 직접 제어하지 않으면 성능이 나오지 않을 것이기 때문이다.

소프트웨어 알고리즘이 "이 데이터는 중요하니 HBM에 넣자"라고 명령하고, 그러면 하드웨어가 "오케이. CXL에서 미리 땡겨놓겠다"라고 답하는 구조가 돼야 한다. 이는 곧 **구글과 같은 프로세서 제조업체와 메모리 업체들이 '칩의 설계 단계부터' 한 몸처럼 움직여야 하는 게 점점 더 필수가 되는 배경이다**.

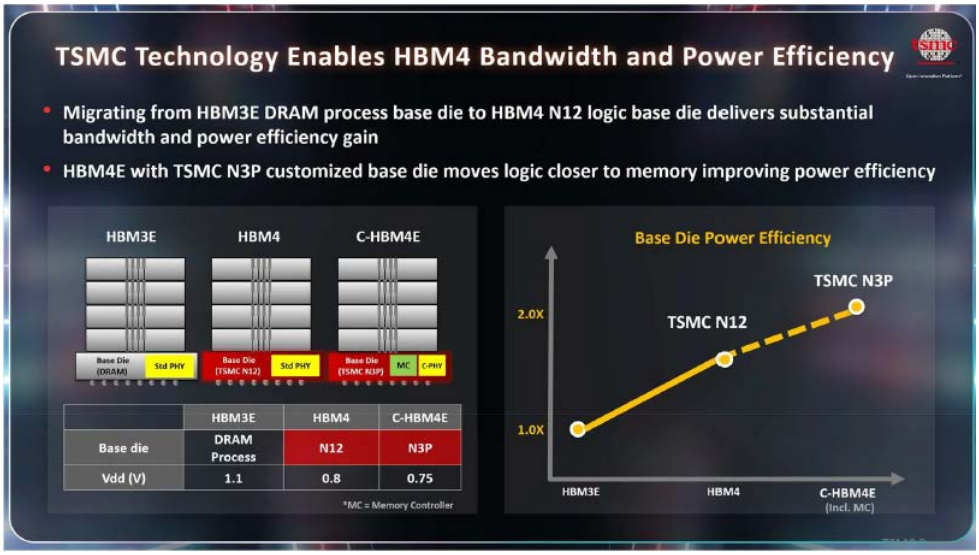
Nested Learning의 고주파 모듈은 매 토큰마다 '읽고-수정하고-쓰는(Read-Modify-Write)' 작업을 수행한다. 이는 기존의 표준 HBM(JEDEC)으로는 감당하기 힘든 '연산 집약적 메모리 접근'이다. 여기서 구글은 메모리 제조사에게 아주 까다로운 주문서를 보낼 것으로 사료된다. 이와 관련해 바로 HBM4의 로직 다이(Base Die)가 그만큼 중요한 이유라고 보인다.

메모리 제조사는 HBM의 가장 밑바닥(Base Die)에 구글이 설계하는 작은 Logic IP를 심는 작업이 수반될 수도 있다. 실제로 지난 11월 SK하이닉스는 "GPU의 특정 기능을 HBM 베이스 다이로 통합한다."라고 말했다. 구체적으로는 HBM 컨트롤러 기능을 GPU 쪽에서 HBM 베이스 다이 쪽으로 옮기는 것을 예시로 들었다. 이를 Nested Learning 관점에서 보면 구글과 같은 빅테크가 요구하는 여러 특수 연산 기능을 HBM에 심어주겠다는 의지가 깔려 있다고 사료된다.

이것은 곧 메모리 제조사가 더 이상 찍어내면 팔리는 규격품(Commodity)을 파는 곳이 절대 아니라는 말이다. 구글의 알고리즘에 최적화된 Custom ASIC HBM을 납품해야 하는, 사실상 파운드리(TSMC)의 성격을 띠게 된다는 것이다. 실제로 TSMC가 N3P 공정으로 커스텀 HBM4E 베이스 다이를 생산하고 Micron·SK하이닉스와 협력하여 2026~27년 양산을 목표로 하고 있는데, 이는 HBM 제조사가 고객 맞춤형 설계에 나서고 있음을 시사한다.

이런 흐름은 **메모리 업체들이 표준 DRAM만 공급하던 사업 구조에서 벗어나, "고객의 로직을 메모리에 박아주는 서비스"로 변모하고 있음을 보여준다**. 따라서, 최첨단의 HBM을 만드는 업체들의 밸류에이션 리레이팅을 정당화하는 기술적 요소로 점차 자리할 것으로 기대한다.

그림 22. HBM의 핵심 경쟁력이 미세공정에서 로직 공정(Base Die)의 전환으로 넘어갔음의 증거
HBM4부터는 TSMC의 12나노 공정을 채택하여 구동 전압(Vdd)을 낮추고, 전력 효율 비약적 상승.
특히 C-HBM4E(커스텀 HBM)는 TSMC N3P 공정을 사용하여 전력 효율이 HBM3E 대비 2배.
더 중요한 것은 C-HBM4E가 GPU에 있던 메모리 컨트롤러를 베이스 다이 내부로 통합했다는 점.



자료: TSMC, 미래에셋증권 리서치센터

중주파 모듈은 데이터(토큰) 문맥이 쌓이는 '청크(Chunk)' 단위로 관리된다. 이를 효율적으로 처리하려면 GPU 메모리와 외부 메모리 사이의 데이터 이동이 물 흐르듯 자연스러워야 한다. 예를 들어, 소프트웨어 알고리즘이 "지금 처리하는 이 데이터는 100ms 뒤면 중요도가 떨어져. 그때 알아서 CXL 메모리로 내려보내. CPU가 개입해서 스케줄링하면 너무 느리니까, 하드웨어가 알아서 해."라고 말할 수 있는데, 그렇다면 중요해지는 것은 메모리 컨트롤러가 된다.

HBM, CXL DRAM, SSD 사이의 데이터 이동작업을 하드웨어 레벨에서 자동화하는 것이다. 이는 **과거 OS가 하던 메모리 관리 기능을 하드웨어가 직접 수행하는 메모리 중심 컴퓨팅(Memory-Centric Computing)으로의 전환을 의미한다. 즉, AI 알고리즘의 일부를 수행하는 핵심으로서, 메모리 회사들의 또 다른 해자(Moat)가 된다**.

이런 모든 변화를 종합하면 결론이 나온다. 이제 알고리즘을 모르는 회사는 메모리를 만들 수 없고, 하드웨어를 모르는 회사는 모델을 최적화할 수 없다는 것이다. 메모리 업체들은 구글과 같은 ASIC 업체들과 숱한 미팅을 거치는 시스템 솔루션 프로바이더가 돼야 하고, 빅테크들은 본인들의 모델 역량을 십분 발휘하기 위해 메모리 컨트롤러의 동작 방식까지 정의한 커스텀 실리콘을 요구할 것이다. 이처럼 **메모리 슈퍼사이클은 가격(Price)의 상승 사이클만이 아니다. 메모리 반도체가 부품에서 지능형 연산 장치로 진화하며 발생하는 근본적 가치(Value)의 전환 사이클이다**.

Compliance Notice

- 당사는 자료 작성일 현재 조사분석 대상법인과 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트는 자료작성일 현재 조사분석 대상법인의 금융투자상품 및 권리를 보유하고 있지 않습니다.
- 본 자료는 외부의 부당한 압력이나 간섭없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.

본 조사분석자료는 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻은 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 종목 선택이나 투자시기에 대한 최종 결정을 하시기 바랍니다. 따라서 본 조사분석자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다. 본 조사분석자료의 지적재산권은 당사에 있으므로 당사의 허락 없이 무단 복제 및 배포할 수 없습니다.