2026. 06. 15

금융으로 만나는 새로운세상

NVIDIA GTC Taipei / COMPUTEX 2026 참관기

젠슨 황이 한국에서 GTC를 개최하겠다는 이유

IT/반도체 **김 운 호**
02) 6915-5656
unokim88@ibks.com

*AI로 제작된 이미지 입니다.

IBK기업은행 금융그룹
IBK투자증권

IBK투자증권에서 반도체 담당하는 김운호입니다.

6월 초 대만 타이페이에서 개최된 Computex 2026, GTC Taipei 참관 이후 작성한 보고서입니다. 보고서 제목을 "젠슨 황이 한국에서 GTC를 개최하겠다는 이유"로 정했습니다. Computex 다녀와서 작성한 보고서 제목이 생뚱맞다는 생각이 드실 수도 있겠습니다.

Computex 주제가 NVIDIA GTC와 같은 맥락이고 Agent AI라는 Keyword를 공유하고 있습니다. Agent AI의 핵심은 Computing 개선, Token 급증, 이를 해결하기 위한 Distributed AI, Agent AI On-Device인데 모든 것이 Memory 수요 폭발로 귀결됩니다.

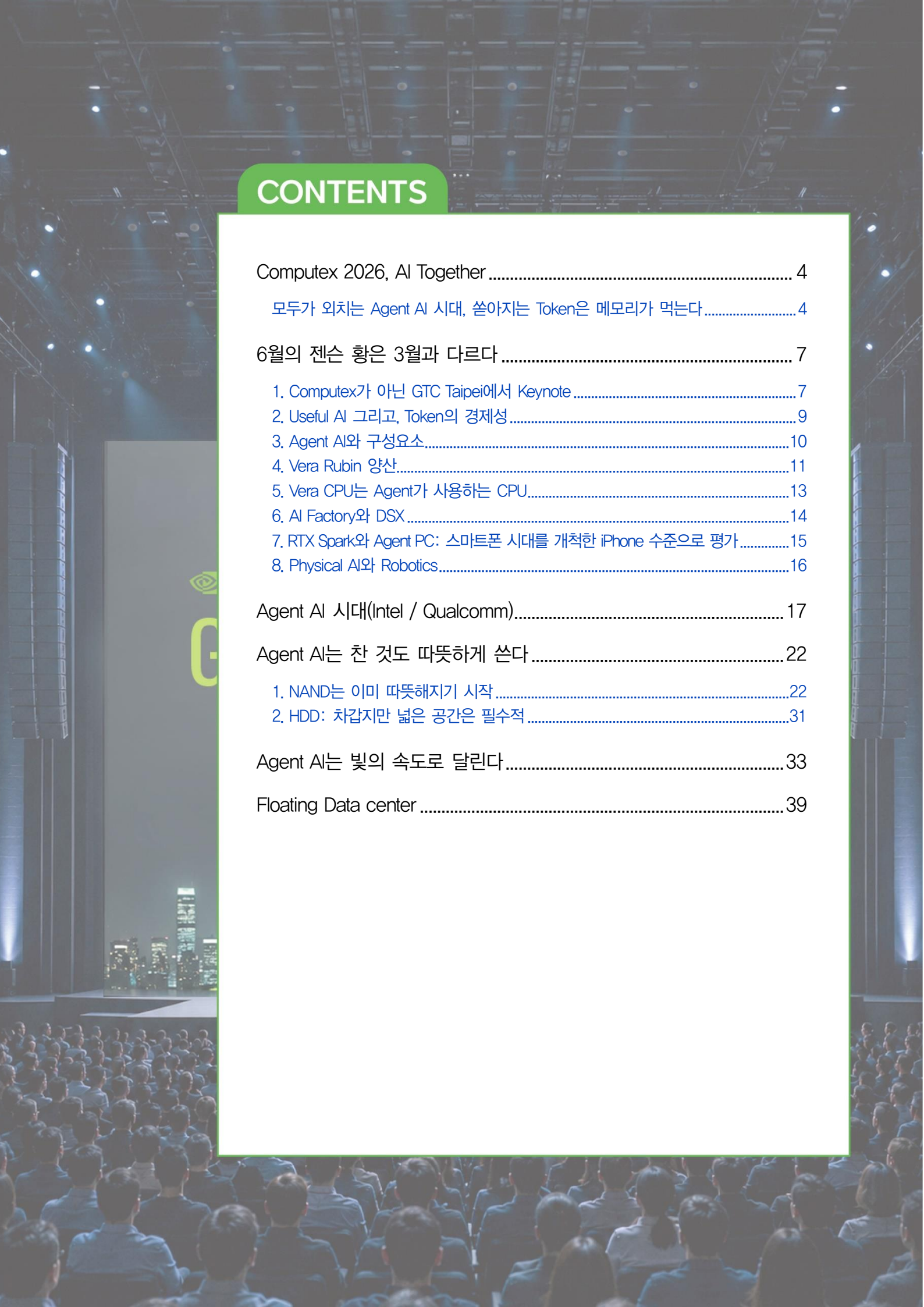
NVIDIA는 미국 San Jose에서만 개최했던 GTC를 2025년부터 대만에서 개최하고 있습니다. AI 서버에 공급되는 부품 생산이 가장 많은 국가이기 때문입니다. 행사 성격도 San Jose는 개발자 중심이고, 대만은 하드웨어 중심입니다. 그러면 한국은 어떨까요? Memory, Foundry, 패키지 기판, MLCC 뿐만 아니라 Cloud, Humanoid, 자율주행 자동차, 전력 공급 시스템 등 AI 관련 완벽한 생태계를 갖추고 있습니다. 젠슨 황은 대만 타이페이에서 GTC Korea 개최 의향을 밝혔고 지난 주 한국 방문에서 배경훈 과기부 부총리와의 긍정적인 의견 교환을 발표했습니다. 대한민국은 Agent AI 시대에 NVIDIA와 동반 성장할 최적의 생태계(Ecosytem)를 보유하고 있음을 확인하는 순간입니다.

한편, 이번 Computex 2026은 Agent AI라는 화두를 주요 IT 회사의 CEO들이 어떻게 생각하고, 어떤 전략을 펼칠 지에 대한 토론의 장이었습니다. AI 서버 시장을 주도했던 NVIDIA, 다소 소외되었던 Intel, Qualcomm도 Agent AI 시대를 대응하기 위해서 PC용 제품을 앞다투어 출시했습니다. 그 맥락에서, GPU와 Cloud 중심에서 CPU와 Edge로 확산되고 있음을 확인했습니다.

"AI is an Ecosystem Game"이라는 화두는 지금의 Agnet AI 시대를 정의하고 있습니다. 독불장군처럼 혼자서 모든 걸 처리할 수 없을 정도로 시장의 확산 속도가 빠르기 때문입니다. NVIDIA는 LPU 칩 Groq을 도입했고, GPU를 위해서 HBM 업체들과 협업하고 있고, Network를 위해서 Marvell, Corning에 직접 투자하고 있습니다. 신제품 PC용 AI Agent 칩인 RTX Spark는 대만 MediaTek과 공동 개발했습니다. Intel도 미국 팹리스 SamvaNova의 RDU를 NVIDIA의 LPU를 대체하기 위해서 시스템에 도입하고 있습니다. Cloud에서 Edge로 확장하고 있는 AI 산업 환경이 Memory / Storage 시장의 성장 동력입니다.

지난 해부터 Las Vegas CES, Barcelona MWC, Taipei Computex, Berlin IFA, San Jose GTC 등 다양한 곳에서 AI의 발전을 현장에서 확인하고 있습니다. 3월 미국 실리콘밸리의 NVIDIA GTC와 6월 대만 타이페이의 Computex와 GTC의 변화는 괄목할 만한 수준이었습니다. NVIDIA는 Agent AI를 PC로 구현하고, Agent AI 지원 플랫폼인 DSX를 개발했습니다. 점점 더 완성되어가는 NVIDIA의 비전을 확인할 수 있었습니다. 또한 NAND 업체들의 시장 성장에 대한 자신감, Intel과 Qualcomm의 Agent AI 출사표도 흥미로운 대목이었습니다. 이러한 흐름은 GTC Korea 개최가 현실로 바짝 다가오고 있음을 느끼게 했습니다.

주식 시장은 AI 성장에 대한 끊임없는 의구심을 제시하지만, 타이페이 현장에서 AI가 폭발적으로 성장하는 시대에 어떻게 대응할 지에 대한 치열한 전략을 보고 들었습니다. AI 정점까지는 아직 3부 능선도 채 넘지 않은 구간이라 생각합니다.



CONTENTS

Computex 2026, AI Together

모두가 외치는 Agent AI 시대, 쏟아지는 Token은 메모리가 먹는다

2026년 Agent AI Boom up

젠슨 황은 AI가 Generative AI, Reasoning AI, Agent AI의 단계로 발전한다고 제시한 바 있다. 그리고 2026년이 바로 그 시점이라고 Computex 2026 참가 기업들이 이구동성으로 발표했다. 바야흐로 Agent AI 전성시대가 되었다.

Memory Full Stack

Agent AI에서 가장 큰 변화는 Memory가 Full Stacking된다는 점이다. GTC 2026, Memory Dynamics에서 언급했듯이 SRAM, HBM, DRAM, NAND까지 아우르는 시스템으로 AI 서버는 진화하고 있다.

NVIDIA, Intel, Qualcomm, Marvell CEO의 목소리

보고서 내용은 젠슨 황의 GTC Taipei Keynote, Agent AI에 출사표를 던진 Intel, Qualcomm CEO의 Keynote, KV cache 및 Bulk Storage 수요 확대에 따른 SSD / HDD 전망, 지난 아모텍 보고서에서 언급한 빛으로 가는 AI에 대한 비전을 제시한 Marvell CEO의 Keynote, 마지막으로 Data center 병목을 해소할 수 있는 솔루션 중에 하나인 Floating Data center 순으로 구성되어 있다.

6월 젠슨 황은 달랐다

젠슨 황은 1년에 Keynote만 3번(CES, GTC, Computex) 하고 있다. 지난 San Jose GTC 이후 2개월 조금 넘게 지난 시점에서 진행된 GTC Taipei Keynote에서는 RTX Spark, DSX와 같은 Agent AI에 특화된 솔루션을 공개한 점이 새롭다.

Intel, Qualcomm도 Agent AI 출사표

Agent AI 시대에 Intel, Qualcomm도 각자의 사업 전략과 Agent AI에 대응하는 PC 솔루션을 공개했다. 서버에서만 움직였던 AI가 PC에서도 필요하게 되었기 때문이다. 전체 생태계를 아우르는 솔루션까지 제공한 업체는 NVIDIA이지만, Intel은 Core Ultra Series 3, Qaulcomm은 Snapdragon X를 출시했다. 이를 통해 Agent AI 확산 속도는 빨라질 것으로 예상한다.

Agent AI는 SSD / HDD를 키운다

Agent AI는 Memory와 Storage에도 영향을 주는데, KV cache 필요성과 대용량 저장 공간의 필요성을 Solidigm, KIOXIA, WD(Western Digital)가 설명했다. Agent AI는 많은 Token을 생성하기 때문에 저장 공간의 확대는 필수적이다. Context Storage 필요성은 Solidigm과 KIOXIA가 공통적으로 제시했고, KIOXIA의 독창적 솔루션은 GPU Direct Storage이다. GPU와 SSD로만 구성되어 있는 서버이다. HBM의 용량을 극복할 수 있는 솔루션으로 제시했다.

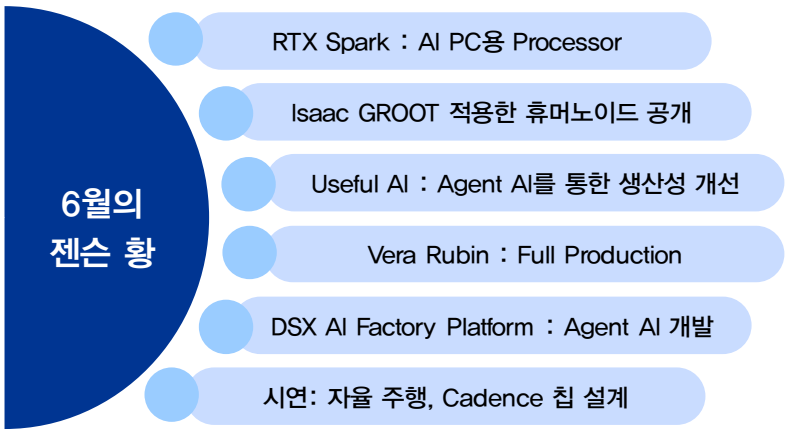
구리로는 한계, 광으로 가야 한다

전송 솔루션에 대한 변화 필요성을 Marvell이 제시했다. 구리선의 한계가 명확하고 데이터 전송 속도가 Agent AI 병목이 될 가능성이 높을 것으로 설명했다. 이를 해결하기 위해서는 광 전송의 도입이 필수적이라고 한다. CPO(Co-Packaged Optics)도 공개했다. NVIDIA Vera Rubin 시스템과 공조할 것으로 전망한다.

Agent AI, 메모리 전성시대

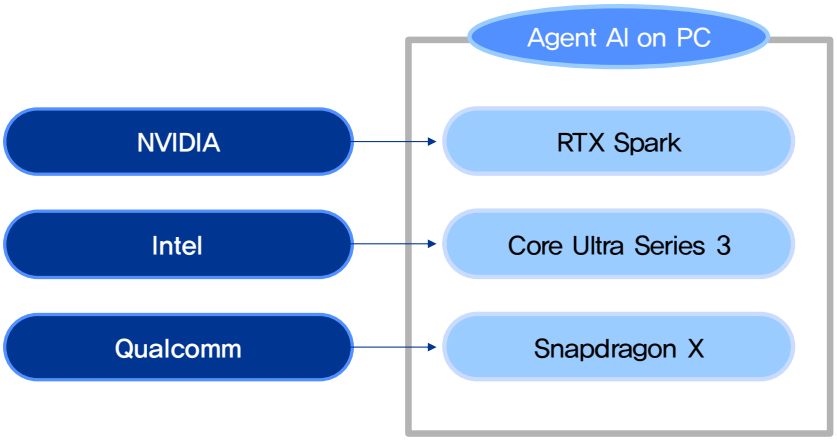
Agent AI는 메모리 시장에 강력한 성장 동력이 될 것으로 기대한다. 서버뿐만 아니라 PC로의 확장, 급증하는 Token과 Data를 저장할 장치 용량도 동반해서 성장할 수 밖에 없다.

그림 1. 6월의 젠슨 황



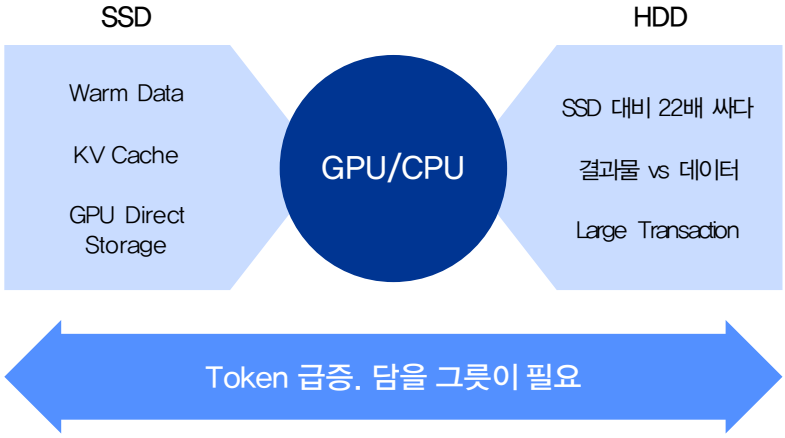
자료: IBK투자증권

그림 2. PC에서 불 붙은 Agent AI



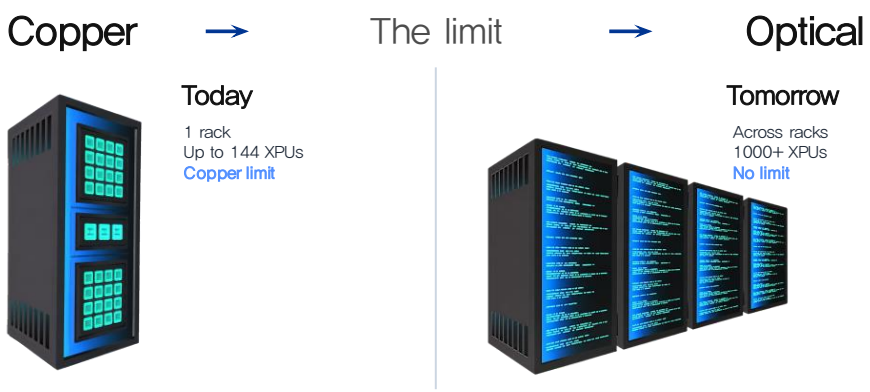
자료: IBK투자증권

그림 3. Agent AI는 메모리를 데운다



자료: IBK투자증권

그림 4. Agent AI는 빛의 속도로 달린다

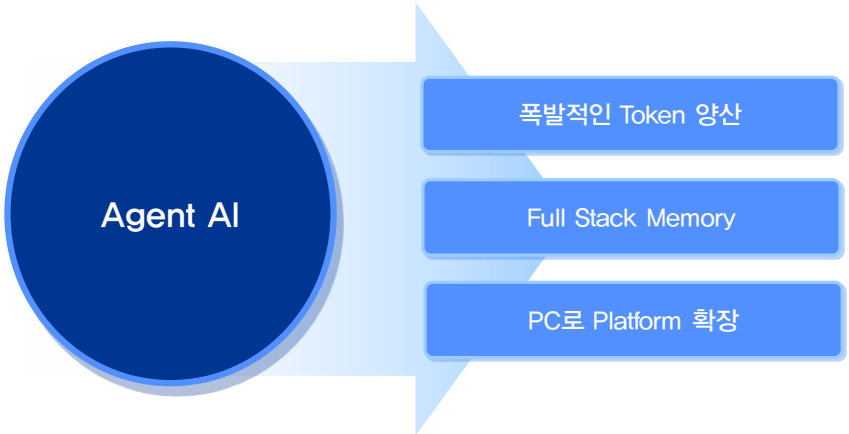


자료: Marvell, IBK투자증권

그림 5. New Solution : Floating Data center

자료: 삼성중공업, IBK투자증권

그림 6. Agent AI는 Token 공장



자료: IBK투자증권

6월의 젠슨 황은 3월과 다르다

1. Computex가 아닌 GTC Taipei에서 Keynote

Computex가 아닌 GTC에서 진행

NVIDIA의 CEO인 젠슨 황은 지난 해와는 달리 Keynote 발표를 Computex와는 별도의 행사장인 NVIDIA GTC Taipei에서 진행했다. NVIDIA GTC는 2009년부터 미국 캘리포니아 San Jose에서 개최되는 행사이다. 2025년부터 GTC Taipei를 Computex와 연계해서 진행하고 있다. 지난 해에는 젠슨 황이 Computex에서 NVIDIA Keynote를 발표했다.

2개월의 차이: 새로운 것과 구체화

Taipei에서 발표한 내용은 San Jose의 내용과 크게 다르지 않지만 새로운 내용도 포함되어 있고, 개념적 비전에서 현실적 구체화로 표현된 것이 특징으로 판단된다. 2개월이 조금 지난 이후의 발표에서 신제품과 여러 아이디어가 구체화 되었다는 것은 AI 진화 속도가 얼마나 빠른지 체감하게 한다.

Agent AI 시대 NVIDIA는 플랫폼 회사

큰 맥락에서는 Computex 2026 전반에 녹아 있는 Agent AI 시대를 전제로 하고 있다는 점과 NVIDIA는 System / Platform 회사라는 점을 강조하고 있다.

GTC Taipei의 Keynote는 아래 7가지로 요약할 수 있다.

1. Useful AI 그리고, Token의 경제성

2. Agent AI와 새로운 컴퓨팅 모델

3. Vera Rubin 양산

4. Vera CPU

5. AI Factory와 DSX

6. RTX Spark와 Agent PC

7. Physical AI와 Robotics

표 1. 젠슨 황 Keynote San Jose & Taipei 비교

구분	내용
신규	• RTX Spark 발표 : Agent PC용 Processor • Isaac GR00T Reference 휴머노이드 로봇 실물 공개 • GitHub commit 수, 개발자 생산성 $9조 등 "유용한 AI"를 정량화하는 데이터 신규 제시 • Cadence 칩 설계 슈퍼 에이전트 실제 시연 신규 공개 • Alpamayo 자율주행 실시간 주행 시연(차량이 행동을 음성으로 설명) 신규 공개, Hyperion 탑재 브랜드가 전 세계 자동차 제조사 80%, 이동 서비스 97% 연결 수치 공개 • Cosmos: GTC에서 Cosmos 1·2 수준 → Taipei에서 Cosmos 3 발표. 픽셀·액션·사운드·언어 스트림 통합 하이브리드 아키텍처로 진화
공통	• Agent = model + harness + tool / kit + runtime • NVIDIA가 GPU 회사 → AI 인프라 회사로 변신 • 컴퓨팅 = 수익이라는 토큰 경제 논리 • Vera Rubin이 추론을 넘어 에이전트를 위해 설계됐다는 포지셔닝 • CUDA-X 라이브러리가 에이전트 도구로 활용된다는 개념 • 수직 통합·수평 개방 전략
구체화	• Vera Rubin : GTC에서 "2026년 출시 예정" → Taipei에서 "전면 양산 진입" 선언. 공급망 규모 Grace Blackwell 두 배, Rack 조립 시간 2시간 → 5분으로 단축 • Vera CPU 성능 : GTC에서 "와트당 2배, x86 대비 우수" 수준 → Taipei에서 IPC 세계 1위, 코어 간 6TB/s 대역폭, SQL 3배, NYSE 스트림 처리 6배 등 실측 수치 공개 • DSX AI Factory Platform 개념 : NemoClaw의 상용화

자료: NVIDIA, IBK투자증권

2. Useful AI 그리고, Token의 경제성

돈이 되는 AI 시대

젠슨 황은 "유용한(useful) AI가 도착했다"고 Keynote를 시작했다. 지금까지 AI는 유용하지 않았다는 의미인가 반문할 필요는 없어 보이지만 실생활에 사용할 수 있는 그리고, 돈이 되는 AI 시대가 되었다는 의미로 해석할 수 있다.

실행하는 AI, Agent AI

AI가 단순 답변 생성(Generative AI) 혹은 좀 더 오랜 시간에 걸쳐 고민한 대답을 생성(Reasoning AI)했던 것에서 벗어나 실제 업무 수행과 생산성 향상 단계(Agent AI)로 진입했다고 밝혔다.

GitHub commit는 개발 코드를 공유 플랫폼에 저장한다는 의미

GitHub commit 증가 사례를 통해 AI가 소프트웨어 엔지니어링 생산성을 크게 높이고 있다고 설명했는데 GitHub는 전 세계 개발자들이 코드를 저장하고 협업하는 플랫폼이다. 자세한 설명은 아래와 같다.

Git : 코드 변경 이력을 관리하는 시스템

GitHub : Git로 관리한 코드를 인터넷에 올려 공유하는 서비스

Commit : 코드 변경 기록 저장

GitHub가 AI에서 의미가 있는 것은 오픈 소스 AI 모델, 프레임워크(vLLM, TensorRT-LLM, Hugging Face 관련 코드)가 GitHub에 올라오고, NVIDIA, AMD, Google, Meta 등이 샘플 코드나 SDK(Software Development Kit)를 공개한다. 이로 인해서 AI Agent가 GitHub를 연결하면 코드를 읽고, 수정하고, 버그를 찾고, PR(Pull Request)까지 만들 수 있다.

GitHub commit 폭증. Agent AI 수요 증가로 해석

GitHub commit 수는 2023년 3억 건 → 2024년 4억 건 → 2025년 5억 건 → 2026년 초 약 3배 증가한 15억건으로 폭증하고 있다. AI Agent로 인해서 생산성이 크게 증가했다는 것을 의미한다. 또한, 이에 필요한 Computing을 위해서 데이터센터에 더 많은 투자를 해야한다는 것을 의미한다. 소프트웨어 개발자 3,000만 명이 $3조 임금으로 $9조의 생산성을 창출하고 있다는 것도 이러한 의미이다.

이제는 Token으로 돈 버는 시대

Token은 이제 단순 연산 단위가 아니라 수익을 창출하는 단위로 변화하고 있다. 젠슨 황은 Token을 새로운 Currency로 정의한다. AI 시대에는 주요 지표가 이전 인터넷 시대처럼 클릭, 조회수, 노출이 아니라 Token 사용량으로 바뀌었다는 의미이다.

더 많은 Token을 생산하려 하고, 생산이 곧 compute demand 확대로 이어지고, 이는 AI Factory 구축 수요를 견인하게 된다.

3. Agent AI와 구성요소

Agent AI 시대라고 모두가 말하다

젠슨 황은 AI의 중심이 LLM 단독에서 Agent AI로 이동하고 있다고 언급했다. 이는 Computex 2026에 참가한 모든 기업들이 공통적으로 언급하는 내용이기도 하다.

Agent AI는 Token 먹는 하마

Agent AI는 이전 Generative AI, Reasoning AI처럼 질문에 대한 대답만 하는 것이 아니라 요구 사항을 직접 수행하는 AI를 의미한다. 이로 인해서 처리되는 Token은 Generative AI 대비 100배가 요구 된다고 NVIDIA는 얘기하고 Intel은 1,000배라고까지 언급했다.

Agent는 LLM, harness, tools, runtime으로 구성되어 있다.

LLM은 사고 / 추론
Harness는 작업 지휘

LLM은 판단하고, 계획하고, 언어를 이해하고, 다음 행동을 결정하는 두뇌 역할을, harness는 AI Agent를 작동시키기 위해서 여러 도구와 절차를 통합해서 실제 작업을 수행하도록 제어하는 소프트웨어 구조이다. Middle ware의 성격이 강하다.

Tool:
실제 수행하는 기능

Tool은 LLM이 실제 일을 수행하기 위해 사용하는 외부 기능이다. 검색 / 계산 / 코딩 / DB 조회 / API(Application Programming Interface) 호출 등이 모두 Tool에 해당한다.

Runtime:
소프트웨어 환경

Runtime은 Agent AI의 Tool과 코드가 실제로 수행되는 소프트웨어 환경을 의미한다.

표 2. Agent AI 구성 요소

구성 요소	뜻	역할	NVIDIA식 대응
LLM / Model	두뇌	이해, 추론, 계획, 코드/텍스트 생성	Nemotron, Llama, DeepSeek, OpenAI/Anthropic 모델 등
Harness / Orchestration	작업 지휘 체계	모델에게 목표를 주고, 단계별로 일을 나누고, 도구 호출을 관리	NeMo, NIM, Agent Toolkit, LangChain/LlamaIndex류
Tools / Skills	손과 발	검색, 코드 실행, 이메일, DB 조회, CAD, 로봇 제어 등 실제 작업 수행	NVIDIA Agent Skills, Omniverse, Cosmos, Metropolis, enterprise apps
Runtime	실행 환경 / 안전 울타리	권한, 보안, 정책, 메모리, 로그, 실행 격리, 배포 관리	NVIDIA AI Enterprise, NIM runtime, Agent Toolkit secure runtime

자료: IBK투자증권

4. Vera Rubin 양산

Vera Rubin 양산 진입. HBM4도 양산

Vera Rubin은 2026년에 발표한 신규 CPU+GPU 시스템이다. Vera가 CPU이고 Rubin이 GPU이다. 지난 Grace Blackwell을 잇는 신규 CPU / GPU 시스템이다. 지난 San Jose GTC에서는 발표만 있었고 이번 Keynote에서는 생산에 진입했다고 언급했다. 정확히는 Full production이라고 말했다. 이는 HBM4 역시 양산에 진입했다는 의미이다.

Vera Rubin 시스템은 7개 칩으로 구성

Vera Rubin 시스템을 구성하는 7개 칩은 CPU Vera, GPU Rubin, Rack 내부 GPU를 연결하는 NVLink 6 Switch, Rack / 서버를 연결하는 ConnextX-9 SuperNIC (Network Interface Card), Network, Storage, 보안, 인프라 오프로딩을 담당하는 BlueField-4 DPU, 데이터센터 / AI Factory간 연결을 위한 Spectrum-6 Ethernet Switch, GPU에 필요한 HBM4를 언급한다.

Agent AI를 위한 시스템

Grace Blackwell이 AI Inference를 위한 시스템이었다면, Vera Rubin은 Agent를 실행하기 위한 시스템이다. 더 빠르고 많은 Token 처리능력이 있다는 의미이다.

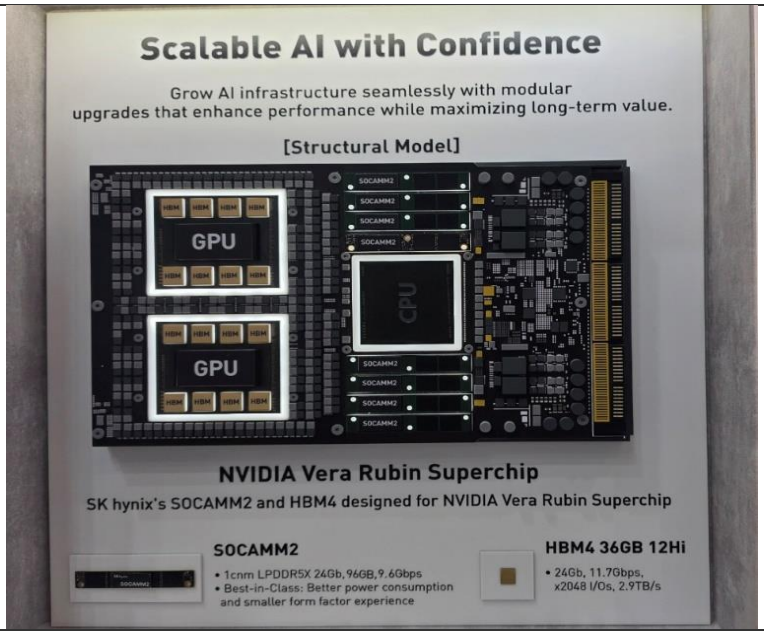
Grace Blackwell 대비 2배 SCM 확보

젠슨 황은 Vera Rubin 공급망이 Grace Blackwell 대비 2배 규모로 구축되어 있다고 언급했다. 시장 규모가 훨씬 커질 것으로 전망하고 있다는 의미이다.

조립 편의성 극대화

또한 신규 시스템의 특징으로 연결을 위한 Cable이 없고 방열을 위한 Fan이 없다고도 언급했다. 이로 인해 조립 시간이 Grace Blackwell은 2시간이었던 것에 비해서 Vera Rubin은 5분으로 단축되었다고 한다.

그림 7. Vera Rubin

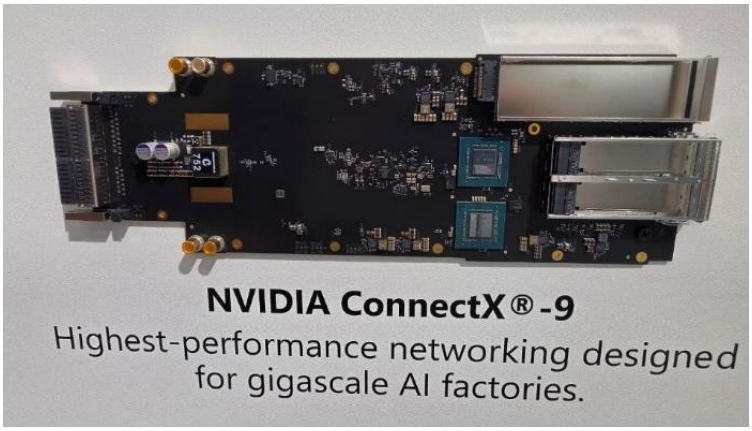


자료: Computex 2026, IBK투자증권

그림 8. NVIDIA NVLink Switch Tray

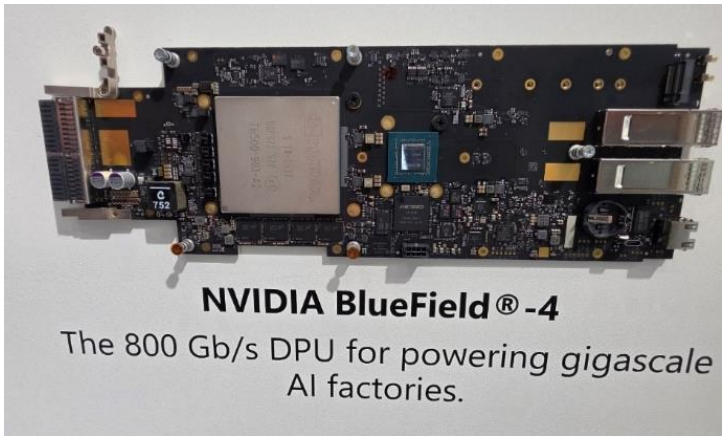
자료: Computex 2026, IBK투자증권

그림 9. NVIDIA ConnectX-9



자료: Computex 2026, IBK투자증권

그림 10. NVIDIA BlueField-4(Vera / ConnectX-9)



자료: Computex 2026, IBK투자증권

5. Vera CPU는 Agent가 사용하는 CPU

서버용 최초 CPU는 Grace

NVIDIA는 GPU Hopper 세대부터 AI 서버용 독자 CPU Grace를 개발 적용하고 있다. Grace는 NVIDIA의 GH200 시스템으로 공급되기도 하고, CPU Superchip으로 CPU 2개를 NVLink로 연결한 서버를 공급한 바가 있다. Meta는 최근 Grace CPU를 대규모 도입하기로 했다.

Vera는 Agent가 사용. 그래서 빠르다

NVIDIA는 Vera CPU를 기존 CPU 대체가 아니라 Agent용 CPU라는 신규 시장으로 제시하고 있다. Vera CPU는 Vera Rubin 내부에서 GPU 조율, KV cache 관리, software execution, tool-use orchestration을 담당하며 latency를 낮춘다.

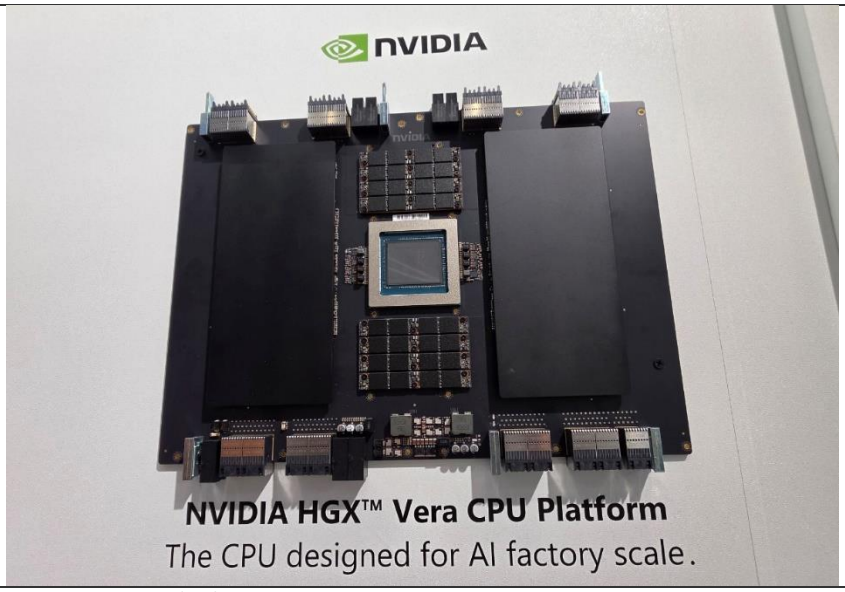
젠슨 황은 Vera CPU가 사람을 위한 제품이 아니라 Agent를 위한 CPU라고 했다. 여기에 latency라는 의미가 부각되는데 사람보다 기계(Agent)는 데이터 속도에 훨씬 더 민감하다는 의미이다. Agent는 나노 초 단위 세계에 살기 때문에 latency를 가능한 낮춰야 한다.

Vera CPU Rack이 별도로 존재

Vera CPU는 Vera Rubin을 지원하는 제품이지만 NVL72에는 Vera Rubin Rack과 별도로 Vera Rack이 있다. Vera Rubin Rack을 지원하는 역할이다. 최근 AI에서 컴퓨팅 속도가 그만큼 더 중요해지고 있다는 의미로 해석한다.

Agent AI 4가지 구성에서 Tool이 실제 작업을 하고 이와 관련된 내용을 LLM으로 전달하는 과정이 반복적으로 이루어지는데 이를 지원하기 위해서 CPU 성능이 강화될 필요가 있다.

그림 11. Vera CPU



자료: Computex 2026, IBK투자증권

6. AI Factory와 DSX

Agent AI를 만들기 위해서 AI Factory가 필요

젠슨 황 대표는 "Compute is revenue, Compute is profit"이라고 강조했다. Agent AI는 모든 기업이 사용할 것으로 전망한다. 가용 자원을 효율적으로 사용할 수 있게 지원하기 때문이다. 각 사업부별 특성에 맞는 Agent AI가 필요하고 이를 위해서 AI Factory를 구축해야 한다.

외부에서 필요한 Agent를 구입할 수도 있겠지만, 내부 자료의 외부 유출, 사업의 특성 등 각 기업만의 특성을 모두 적용하기 위해서는 독자 개발이 맞는 방향일 것으로 판단된다.

AI Factory를 만들기 위한 매뉴얼이 DSX

이를 지원하기 위해서 NVIDIA가 제공하는 시스템이 DSX이다. DSX는 AI Factory를 만들기 위한 complete playbook(완전한 실행 지침서)이라고 설명한다. 오픈소스 / 모듈형 소프트웨어 라이브러리, API, 레퍼런스 디자인, NVIDIA 가속 컴퓨팅 플랫폼(CPU / GPU 시스템), 파트너 기술을 하나로 묶는 플랫폼이다. DSX는 토큰을 낮은 비용, 높은 효율로 생산하기 위한 AI Factory 설계 / 운영을 지원한다.

표 3. DSX 구성

구성	설명
Reference design	AI Factory 설계 표준안
APIs	전력·냉각·네트워크·운영 시스템 연결
Software libraries	설계·운영 자동화용 소프트웨어
Simulation	AI Factory를 짓기 전 디지털 트윈으로 검증
Operations	구축 후 운영, 상태 모니터링, 자동화
Partner technologies	Schneider Electric, Siemens 등 인프라 파트너 생태계와 연결

자료: IBK투자증권

표 4. 문제별 DSX 역할

문제	DSX의 역할
GPU를 얼마나 넣을 것인가	rack / 클러스터 설계
전력 한도 안에서 어떻게 최적화할 것인가	전력·냉각 최적화
GPU 간 통신 병목은 어떻게 줄일 것인가	NVLink / Ethernet / InfiniBand 설계
토큰당 비용을 어떻게 낮출 것인가	MaxLPS, token performance per megawatt 최적화
AI Factory를 어떻게 운영할 것인가	모니터링·자동화·운영 소프트웨어

자료: IBK투자증권

7. RTX Spark와 Agent PC : 스마트폰 시대를 개척한 iPhone 수준으로 평가

새로운 생태계 창조하는 NVIDIA

이번 Keynote에서 가장 관심있게 들었던 부분이다. 지난 San Jose GTC에서 NVIDIA는 AI 생태계를 만들어 가고 있다고 보고 느꼈었다. NemoClaw를 이용한 Open Model 개발 솔루션을 제공하고 이를 통해서 다양한 Agent AI 모델이 구축될 것으로 기대하고 있다.

결국 이러한 수요는 AI Factory 수요로 이어지고, 이로 인해 CPU / GPU 뿐만 아니라 관련 네트워크 솔루션, 저장장치 솔루션의 확대를 모색하고 있다.

Agent AI를 PC로 개발. RTX Spark로 지원

여기에 더 나아가 Agent AI 모델 개발이 서버 레벨에서만 진행되는 것이 아니라 PC 레벨에서 가능할 것으로 보고 개발한 제품이 RTX Spark이다. Microsoft와 NVIDIA는 PC를 Agent AI 시대에 맞게 재정의했다.

RTX Spark는 iPhone수준의 혁명

젠슨 황은 RTX Spark를 탑재한 노트북을 스마트폰 시대를 열었던 iPhone에 비유했다.

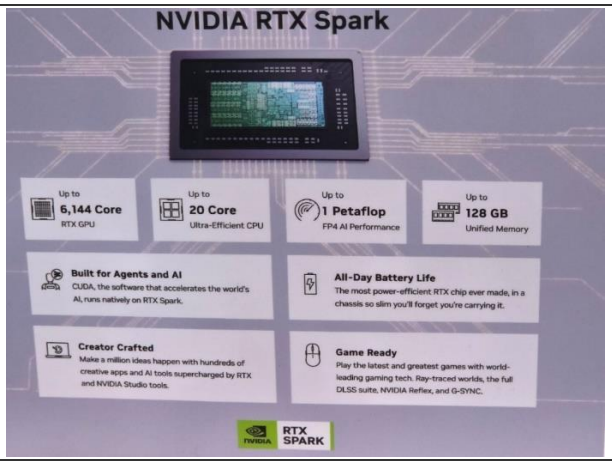
GPU+CPU 통합 칩

RTX Spark는 Blackwell RTX GPU와 MediaTek과 공동 개발한 N1X(custom 20-core Grace CPU)를 NVLink로 결합한 Agent PC 플랫폼이다. 6,144개 CUDA core, 1 PFLOPS AI 성능, 128GB unified memory를 제공하며, TSMC 3나노 공정 기반 약 700억 개 트랜지스터를 집적한 구조이다.

서버에서 PC로

PC가 더 이상 클릭·타이핑 중심 도구가 아니라 개인 Agent를 상시 실행하는 AI computer로 진화했다는 기념비적인 제품이라 평가한다. Desktop, laptop, workstation 모두 Agent AI를 실행하는 Windows machine으로 확장하는 제품이고, AI compute가 데이터센터뿐 아니라 client device로 내려오는 흐름을 제시한 첫 제품이다.

그림 12. RTX Spark



자료: Computex 2026, IBK투자증권

8. Physical AI와 Robotics

Physical AI는
Agent AI의 확산

Agent AI는 디지털 로봇이며, Agentic model이 물리 세계로 확장되는 것을 Physical AI라고 한다. 로봇의 정의가 스스로 판단하고 작업을 수행하는 기계라고 보면 Agent AI도 광의 로봇에는 포함된다고 할 수 있다. 다만 물리적 기능이 없어서 디지털 로봇으로 볼 수 있다. NVIDIA는 Physical AI를 현실 세계의 물리 법칙을 이해하는 AI라고 설명하고 있다. Physical AI는 cloud, PC, vehicle, robot, factory, base station, edge 전반으로 확장될 것으로 전망한다.

Keynote에서 Physical AI 관련 Cosmos, Alpamayo, GR00T 3가지 플랫폼을 제시했다.

Cosmos가
중추 브레인 역할

1. Cosmos는 로봇이나 자율주행차가 현실 세계를 이해하도록 물리 세계를 생성 / 예측 / 시뮬레이션 하는 AI 모델이다. NVIDIA는 World Foundation Models, tokenizers, guardrails, accelerated data processing pipelines를 포함하는 Physical AI 개발 플랫폼으로 설명한다.

Alpamayo는
자율 주행 특화 모델

2. Alpamayo는 자율주행차용 Physical AI 모델 / 도구이다. Reasoning 기반의 자율주행 개발을 가속하기 위한 오픈 AI 모델 / 시뮬레이션 도구 / 데이터셋 패밀리로 정의한다. Cosmos가 가상 세계와 학습 데이터를 만들면 Alpamayo는 그 세계에서 학습해 자동차의 운전 솔루션을 제공하는 개념으로 이해하면 편하다.

Isaac GROOT는
휴머노이드용 모델

3. Isaac GR00T reference humanoid robot은 휴머노이드 개발을 위한 오픈 플랫폼이다. Isaac는 시뮬레이션, 로봇 학습 프레임워크, CUDA 가속 라이브러리, AI 모델, reference workflow를 포함해 AMR, 로봇 팔, 매니퓰레이터, 휴머노이드를 만드는 오픈 Roboitcs 플랫폼이다.

Isaac GR00T 플랫폼 기반 휴머노이드를 공개한 것도 이번 Keynote 행사의 특징이다. 25 자유도, 손당 31 자유도, 키 6피트, 무게 150파운드. Jetson Thor 칩 및 전체 소프트웨어 스택을 NVIDIA가 제공했다.

그림 13. Isaac GR00T로 개발한 Humanoid

자료: NVIDIA, IBK투자증권

Agent AI 시대 (Intel / Qualcomm)

모두가 인정하는 Agent AI

Computex 2026을 관통하는 Keyword는 Agent AI이다. 지난 해가 개념의 정의, 방향성 제시였다면 올해는 좀 더 구체화되고 성장한 Agent AI이다. Agent AI를 구현하는 수단은 서버와 PC로 나누어진다.

Agent AI를 PC에서

Generative AI, Reasoning AI는 서버 자산이 절대적으로 중요했지만 Agent AI는 PC 레벨에서도 가능하다고 판단한다. Agent AI에도 서버는 절대적으로 중요한 자원이기 때문에 없어서는 안된다.

NVIDIA: RTX Spark

NVIDIA는 Vera Rubin을 Agent AI를 위한 시스템으로 소개하고, Agent의 컴퓨팅 속도에 대응하기 위해서 Vera CPU를 따로 강조했다. 그리고 Agent AI 개발이 PC 레벨에서 가능해야 한다는 생각으로 RTX Spark를 공개했다.

Intel, Qualcomm도 Agent AI 시대에 맞게 적당한 Chip Solution을 제공할 계획이다. Training AI, Inference AI 시기에 핵심 Chip은 GPU였다. 연산에서 사용되는 비중이 GPU가 7, CPU는 1 수준이었다. Agent AI를 실행할 때에는 Intel이 Keynote에서 시연했던 것처럼 CPU와 GPU의 비중은 1대 1 수준이다. NVIDIA가 GPU 중심에서 RTX Spark를 생산한 이유이기도 하다.

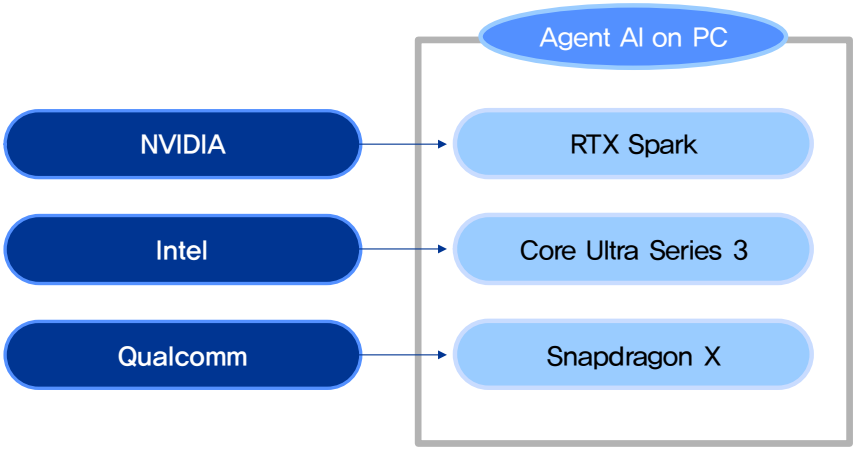
Intel: Core Ultra Series 3

Intel은 CPU가 주력 제품이다. Agent AI용 CPU를 개발하기 가장 좋은 입지이다. Agent AI 개발용 PC를 위해서 Core Ultra Series 3를 발표했다. 코드명은 Panther Lake이다. GPU와 NPU가 포함되어 있다. Intel은 상대적으로 약한 GPU를 극복하기 위해서 Hybrid Agentic Interface를 시연했다. 민감한 데이터는 PC에서, 외부 리서치와 대형 모델은 클라우드 서버에서 구동하는 방식이다.

Qualcomm: Snapdraon X

Qualcomm도 Agent AI 시대에 동참했는데 서버와 PC에 국한하지 않고, 스마트 글래스, 자동차, 로봇까지 Agent AI가 적용될 것으로 전망하고 있다. 노트북용으로 Snapdragon X를 제공한다.

그림 14. Agent AI 쏟아지는 출사표



자료: 각 사, IBK투자증권

1. Intel / Lip-Bu Tan : AI-driven computing, ecosystem collaboration

Agent AI는 Intel에 기회

Intel은 Agent AI를 새로운 기회로 판단하고 있다. Generative AI, Reasoning AI로도 CPU를 꾸준히 공급했지만 GPU만큼은 아니었고, AI 학습에서 핵심적인 부품은 아니었다. Agent AI를 Intel은 4가지 카테고리로 대응할 전략이다.

1) PC·Edge : 모든 PC를 Agentic Platform으로 전환

PC는 CPU, GPU, NPU 성능을 강화한 Core Ultra Series3로 대응

Intel은 Agent AI를 위해서 Core Ultra Series 3를 출시했다. 프리미엄 모바일 성능, 배터리 수명, CPU·GPU·NPU 통합 AI 성능이 이전 세대 제품에 비해서 강화되었다. Core Ultra Series 3는 300개 이상 디자인, Core Series 3까지 합산 시 총 400개에 가까운 디자인을 확보하며 PC 생태계 내 빠른 속도로 성장하고 있다.

Edge 영역에도 Core Ultra Series 3 제품군을 투입해 제조, Robotics, 리테일 등으로 확장하고, Physical AI를 장기 성장 시장으로 제시했다.

2) On-device AI와 Hybrid Agentic Inference

Hybrid Agentic Inferface : Local + Cloud

Intel은 Agent AI 운용 방식으로 Hybrid 방식을 적용했다. 이를 적용한 AI업체는 Perplexity이고 행사장에서 CEO가 동반 등장해서 시연했다. Hybrid agentic inference는 민감한 파일과 기밀 데이터는 Core Ultra 기반 로컬 모델이 처리하고, 외부 리서치와 대형 모델 추론은 클라우드 모델이 처리하는 구조이다.

Intel은 미래 컴퓨팅이 데이터센터에 더 많은 compute, 로컬 머신에도 더 많은 compute가 함께 존재하는 hybrid architecture로 진화한다고 강조했다.

3) 데이터센터 : Agent AI가 CPU 수요를 재부각

CPU 필요성 재부각: Agent AI는 CPU GPU 사용 비율이 반반

기존 LLM inference는 GPU-heavy 구조였지만, Agent AI는 도구 사용, 파일 읽기·쓰기, linting(코드에서 문법적 오류, 스타일 문제, 잠재적 버그를 자동으로 찾아주는 검사 과정), web fetch, compile, unit test 등 CPU 중심 작업이 증가하게 된다.

시연 기준 기존 inference는 GPU:CPU 비중이 약 7:1에 가까웠지만, agentic pipeline에서는 CPU와 GPU 비중이 거의 동등하거나 CPU-heavy로 전환되었다.

Intel은 x86이 약 50년간 데이터센터를 구동해온 범용 컴퓨팅 아키텍처이며, 2030년까지 설치 서버 10대 중 8대가 x86 기반일 것으로 전망한다. Xeon 6 Plus를 Computex 2026에서 공식 출시했다. 288개 E-core, 576MB L3 cache, Intel 18A 기반 고밀도·고효율 CPU로 소개했다. Xeon 6 Plus는 32U rack 기준 36,000개 이상 코어, 최대 15만 개 agent 구동 가능성을 제시했다.

4) Rack Scale과 이종 추론 (CPU, RDU, GPU)

Socket에서 Rack

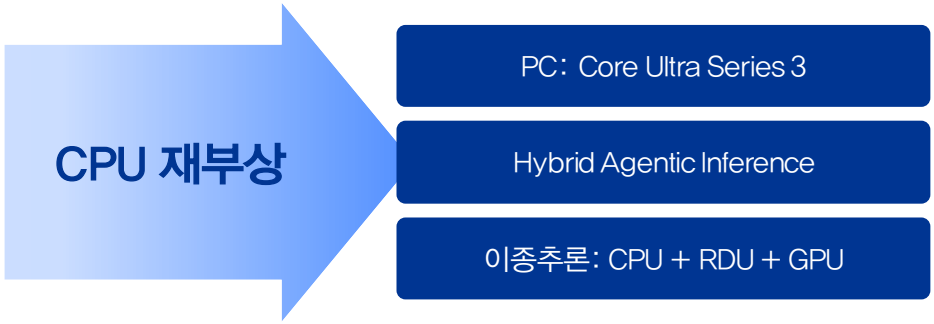
Intel은 소켓 단위 compute만으로는 Agent AI workload를 감당하기 어렵다고 보고 Rack Scale 청사진을 발표했다. Foxconn과 Xeon 기반 rack scale AI infrastructure 개발·통합·상용화 협력을 확대하기로 했다.

AI Server : CPU+GPU+RDU

Xeon CPU, SambaNova RDU, NVIDIA GPU를 결합한 heterogeneous disaggregated inference를 시연했는데 이는 NVIDIA의 NVL 시스템과 유사한 구조이다. GPU를 제외하면 CPU는 Intel, LPU를 대신해서 SambaNova의 RDU(Reconfigurable Dataflow Unit)를 적용했다.

역할 분담은 Xeon이 tool execution, RDU가 decode와 token generation, GPU가 prompt caching과 prefill을 담당하는 구조이다. 자체 테스트 기준 GPU 단독 대비 2~3배 빠른 inference 성능을 주장하고 있다.

그림 15. Intel의 Agent AI 대응 솔루션



자료: IBK투자증권

2. Qualcomm / Cristiano Amon : Year of Agents

2026년은 Agent의 해, Agent AI와는 다른 의미

Qualcomm은 2026년을 '에이전트의 해'라고 정의했다. NVIDIA, Intel과 같은 맥락의 흐름이다. 하지만 비즈니스 모델의 차이에서 나오는 접근 방법과 솔루션 구현 방법에서는 다소 차이를 보인다. AI 서버에 가장 근접한 순서는 NVIDIA, Intel, Qualcomm이고 반대로 Device에 가장 근접한 순서는 Qualcomm, Intel, NVIDIA이다.

Agent는 디바이스

Qualcomm의 Agent AI 시대에 대응하는 전략은 Device 중심이고 Cloud보다는 Edge 기능을 좀 더 강조하고 있다.

Qualcomm은 AI가 발전하며 개인용 컴퓨팅 디바이스의 아키텍처가 변화하기 시작했다고 보고 있고, AI가 단순 도구(tool)에서 자율 실행 에이전트(agent)로 전환하기 시작했고, Agent AI는 사용자 개입 없이 다중 작업을 독립적으로 처리하는 방식이라고 설명한다. Qualcomm은 이 전환이 PC·모바일·자동차·로봇 전 카테고리에 걸친 대규모 업그레이드 사이클을 촉발한다고 전망한다.

토큰 수요 증가 속도는 상상초월

Qualcomm은 토큰 수요 증가에 대해서 NVIDIA, Intel과 유사한 견해를 갖고 있다. 레벨 1(단순 대화) ~1만 → 레벨 3(자율 에이전트 AI) 수백만 토큰이 소요되고, 두 단계 만에 토큰 사용량이 수백 배 증가하고 있다고 추정한다.

그 근거로 일반 사용자는 하루 약 5,000개 token, power user는 하루 최대 2.5만 개 token을 생성하고, 대기업은 이미 하루 약 80억 개 token을 생성하고 있으며, 이는 단순 채팅이 아니라 실제 업무 프로세스를 수행하는 agent 활동에 가까운 수준이다. AI가 실험 단계를 지나 기업 업무를 조용히 구동하는 실제 인프라가 되고 있음을 의미한다.

2026년 현재 전 세계 10초 내 토큰 수요 약 317억 개, 2030년은 추정이 가능할 지 모르겠지만 401경 4,810조 개로 증가할 것으로 보고 있다.

분산 AI로 효율성 제고

증가하는 token 수요에 대응 방식은 Intel과 유사하게 분산 AI를 제시했다. 분산 AI로 코딩 워크로드하면 약 140만 토큰을 절약할 수 있고, 비용도 60% 절감 가능하다.

상호 보완적 관계: Cloud vs Edge 동반 성장

Cloud와 Edge는 대체재가 아닌 상호보완 관계임을 강조한다. Cloud에서 실행되어야 하는 것은 Cloud에서 실행될 것이고, Edge에서 실행되어야 하는 것은 Edge에서 실행될 것이라고 언급했다. 중요한 것은 Cloud-Device가 함께 작동한다는 의미이다.

디바이스가 Agent로 진화

현재 상용되는 Device에도 Agent AI가 적용될 것으로 보고 있다. 과거에는 phone, watch, XR glasses, PC를 각각 분리된 기기로 봤지만, 이제는 이들 기기가 서로 context를 공유하는 구조로 변화하고 있다.

Agent로 연결된 디바이스

AI는 사용자의 phone, app, wearable, PC 안에 자연스럽게 들어가며 일상 활동을 이해하는 방향으로 발전하고, Agent는 app, device, workplace를 오가며 사용자가 phone에서 시작한 작업을 PC에서 이어가고, calendar·wearable·car data까지 연결해 작업을 완성하는 구조로 발전할 것으로 전망한다.

PC가 Agent AI의 강력한 디바이스

Qualcomm도 PC가 Agent AI의 중심으로 이동할 것으로 보고 있다. Agent가 scale 있게 작동하려면 항상 켜져 있고, 즉각 반응하며, local에서 실행 가능한 compute가 필요한데 PC는 power, efficiency, 사용자의 daily workflow 이해도를 동시에 갖춘 기기이기 때문이다. 이에 따라 AI agent 실행의 중심축이 PC로 이동하고 있으며, PC는 단순 작업 기기가 아니라 여러 앱과 디바이스를 연결하는 agent orchestration hub로 변화할 것으로 예상한다.

Snapdrgon X series를 Agent AI PC 솔루션으로 제시

Qualcomm은 Snapdragon X series를 PC에서 full agent orchestration을 실행할 수 있는 플랫폼으로 제시했다. Agent는 여러 app과 device를 동시에 이해하고, 실시간으로 사용자의 작업 맥락을 파악해 직접 action을 수행하는데 Snapdragon X는 이를 충분히 지원한다고 발표했다.

Cursor, Claude Desktop, OpenAI Codex 등 Agent AI workflow가 Snapdragon 기반 PC에서 native로 구동되고 있고, 35억 개 이상 device reach를 기반으로 device, agent, platform을 연결하는 AI 생태계 중심이 될 것으로 기대한다.

6G로 연결성 강조. Qualcomm만의 차별화 포인트

6G 통신을 강조하는 것도 역시 Qualcomm의 비즈니스 모델을 반영한 결과이다. AI를 위한 6G 통신 시스템 구축은 연결성·컴퓨팅·센싱이 중요한 변수이다. 연결성은 약한 신호 환경에서도 고속 유지, 자율 네트워크 운영이 가능해야 하고, 컴퓨팅은 디바이스에서 Cloud까지 지속적으로 제공되어야 한다. 이를 위해서 분산 AI가 기지국· 인프라 전반에 내재화되어 있어야 한다. 통신사가 근본적으로 다른 회사로 변모해야 한다. 센싱(최대 변화)은 무선 연결에서 발생하는 데이터를 통해 차량·보행자 등 움직이는 모든 객체를 실시간 식별 → Agent의 실시간 컨텍스트 데이터로 활용할 수 있어야 한다.

디바이스 솔루션에 강점. Agent 구동 가능하게 하는 컴퓨팅 엔진 필요

Qualcomm은 가장 작은 wearable부터 Agent와 연결되는 모든 디바이스까지 전 계층 커버하는 플랫폼을 구축하고 있고, 모든 디바이스의 모든 컴퓨팅 엔진이 Agent를 실행하는 데 있어서 절대적으로 중요한 변수임을 강조했다.

그림 16. Qualcomm은 Agent AI가 아닌 Agent를 강조

자료: IBK투자증권

Agent AI는 찬 것도 따뜻하게 쓴다

1. NAND는 이미 따뜻해지기 시작

Agent AI 확대로 SSD 수요 확장

Agent AI가 빠르게 확산되면서 SSD 역할이 커지고 있다. HBM, DRAM만큼은 아니지만 이제는 충분히 AI 연산에 도움이 되면서 Cold data에서 Warm Data로 이동 중이다.

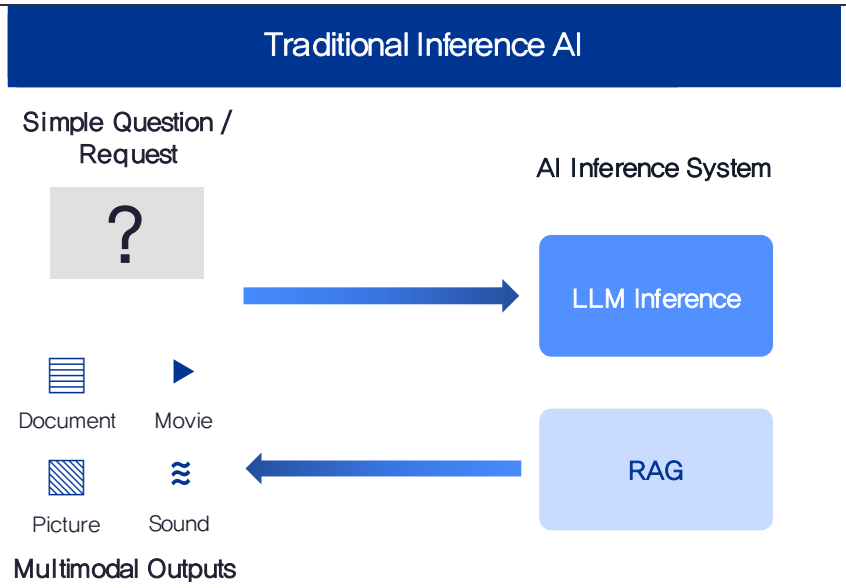
Agent AI는 Inference AI보다 Token 소요량이 급증

AI가 단순 질의응답 중심의 traditional inference에서 계획, 실행, 관찰, 개선을 반복하는 Agent AI 구조로 진화하고 있다. Agent AI는 LLM inference와 RAG(Retrieval-Augmented Generation)를 기반으로 복잡한 목표, 데이터, 외부 시스템을 연결해 구체적인 결과와 자동 실행으로 이어지는 구조이다. 이 과정에서 inference context, token 사용량, 지식 활용 범위, inference application이 모두 확대되며 AI token 사용량은 2030년까지 빠르게 증가할 것으로 전망한다.

25개 Token이 발생하는 질문에 답하기 위해서 13.1GB KV cache 사용

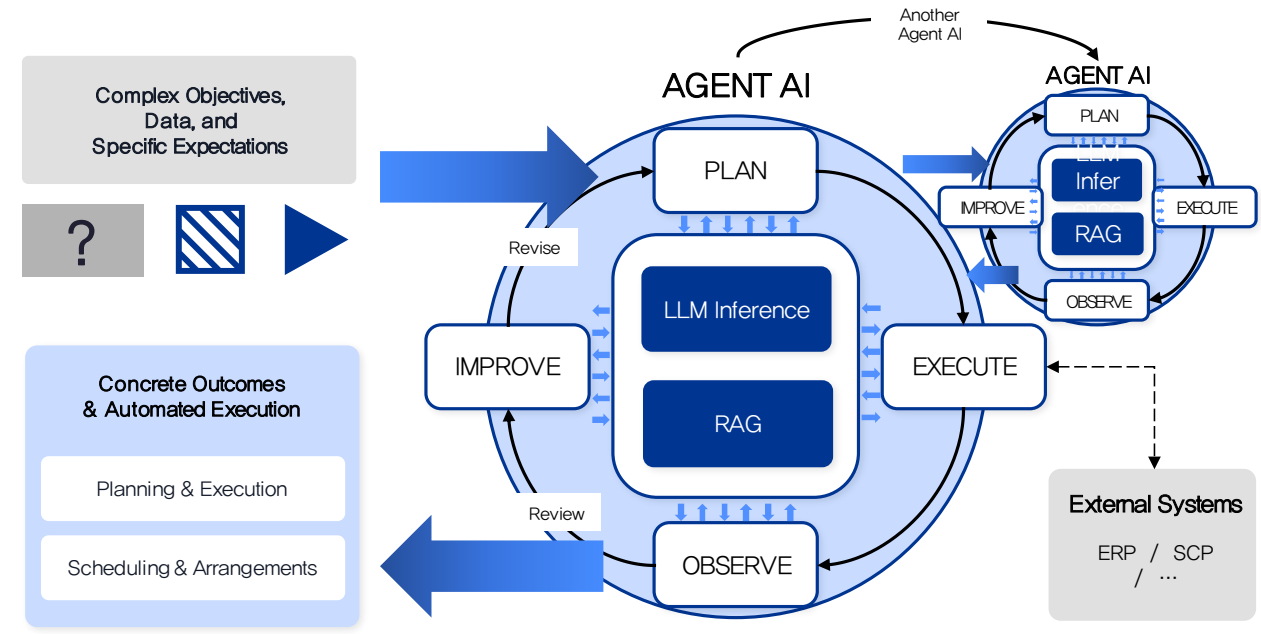
Solidigm은 16개 단어, 25개 token 수준의 단순 질문(ex. Tell me about the top 5 seafood restaurants I should try when I'm in Taipei)도 실제 처리 과정에서는 42,025 token이 발생하고 이는 13.1GB KV cache 용량을 차지한다고 설명했다. 메모리 수요가 급격히 증가하는 이유이다.

그림 17. Inference AI 서버



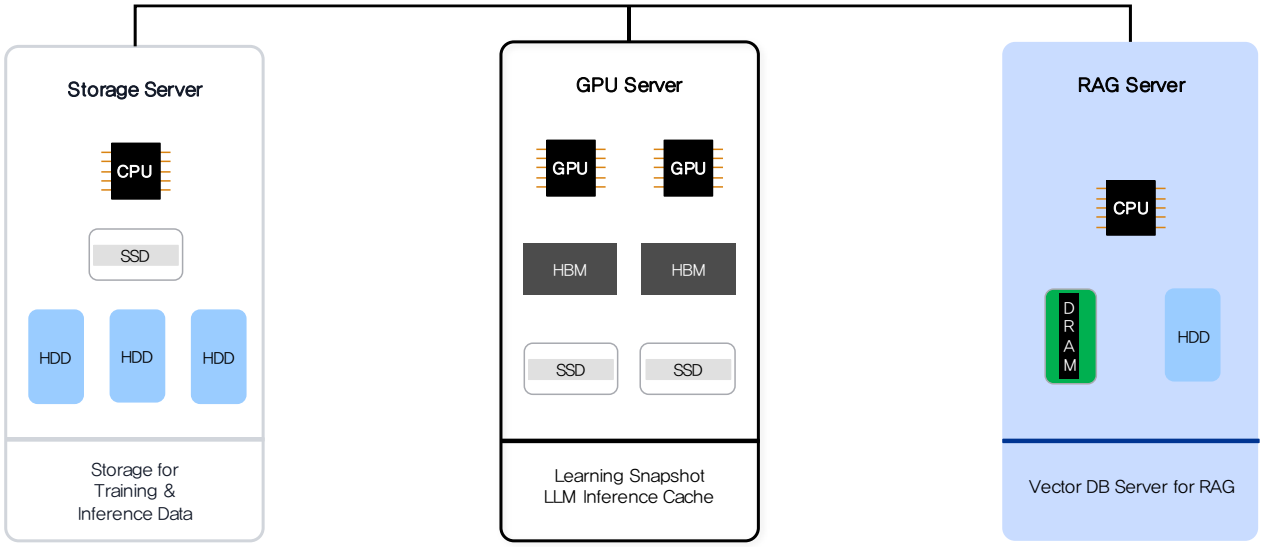
자료: KIOXIA, IBK투자증권

그림 18. Agent AI 서버



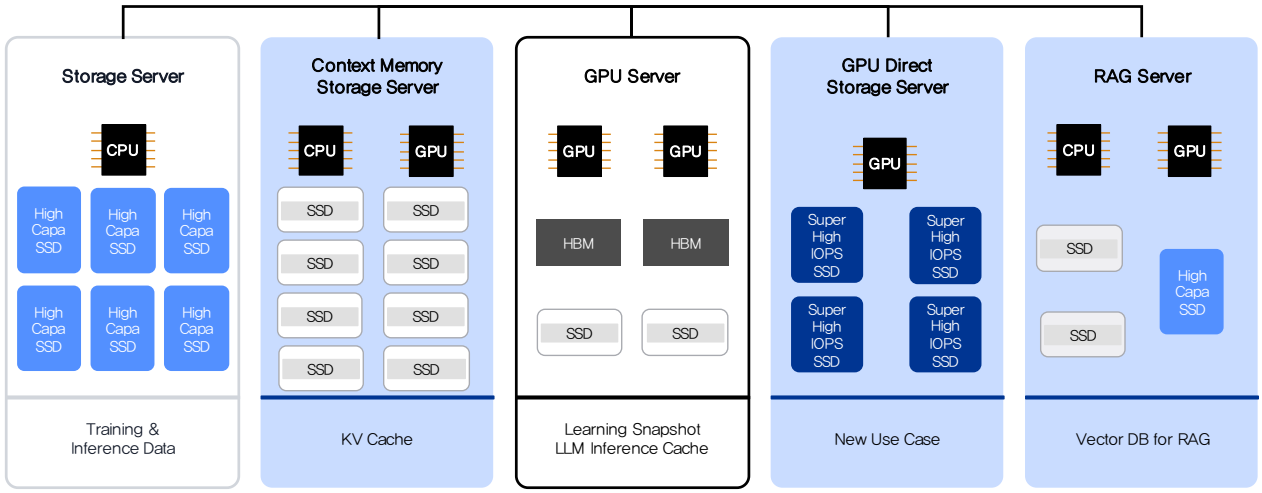
자료: KIOXIA, IBK투자증권

그림 19. Inference AI Server System–2025



자료: KIOXIA, IBK투자증권

그림 20. Agent AI Server System – 미래



자료: KIOXIA, IBK투자증권

AI Data는 3개 축으로 성장

AI Data 증가에 대한 여러 가정이 있지만 Solidigm은 Linear하게 증가하는 것이 아니라 3가지 축으로 성장한다고 분석했다. 3가지 축은 Data Richness, Inference Complexity, Operational Persistence이다.

- Data Richness : Size of each data object
- Inference Complexty : Data generated per task
- Operational Persistence : What survives, how long

2년 만에 토큰은 330배 증가

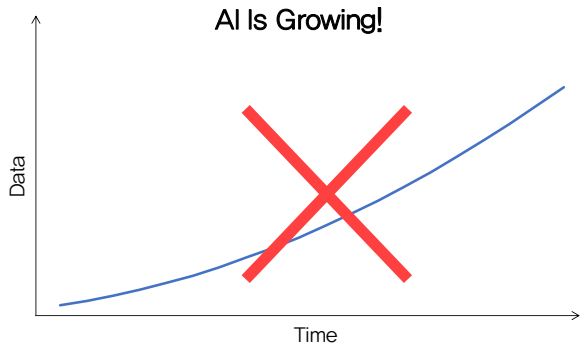
이 축들은 이용할 때마다 증가하고, 이용하는 사람들이 증가할수록 더 빠르게 배속되고 있다. 구글에서 보여준 2년 누적 효과는 330배 커졌다. 2024년에는 9.7조에서 2026년에는 3,200조까지 토큰이 증가했다.

표 5. AI Data 성장은 3차원

Vector	Data Richness	Inference Complexity	Operational Persistence
What is it?	Size of each data object	Data generated per task	What survives, how long
Evolution	Text → Multimodal → Sensor streams	Single-shot → Reasoning chains → Multi-step workflows	Ephemeral → Stored-by-default → Continuous logging
Key Indicator	In 2 years, Llama model training corpus grew by 15x ▪ 2T tokens (Llama 2) → 30T (Llama 4)	In 2 years, DeepSeek model context window grew by 31x ▪ 32K (V2) → 1M (R1)	OpenAI API defaults ▪ Assistants : 2.5TB project storage ▪ Responses : stored by default
Vector	Data Richness	Inference Complexity	Operational Persistence

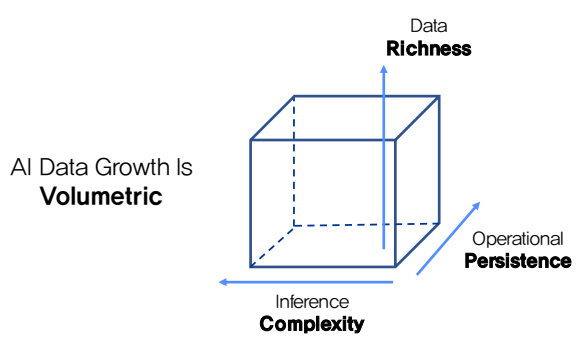
자료: Solidigm, IBK투자증권

그림 21. AI Data 선형 증가 모델



자료: Solidigm, IBK투자증권

그림 22. AI Data는 3차원으로 증가



자료: Solidigm, IBK투자증권

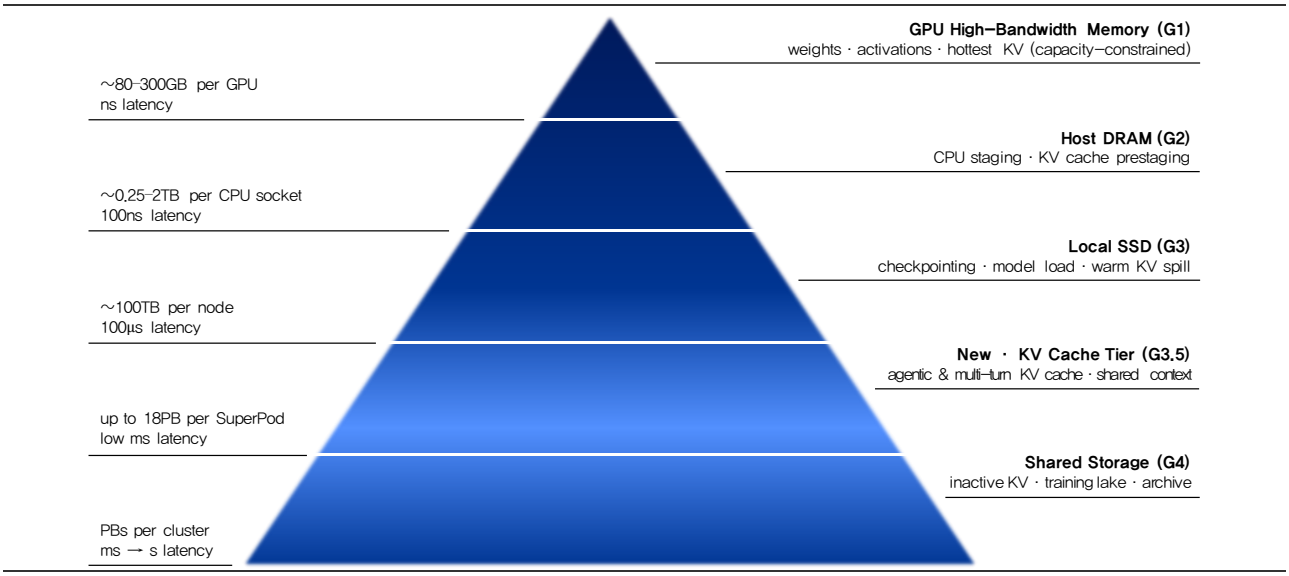
AI memory hierarchy 재편성 중

Agent AI로 발전하면서 AI memory / storage hierarchy가 재편되고 있다. AI 인프라는 GPU HBM, Host DRAM, Local SSD, KV cache Tier, Shared Storage로 이어지는 계층 구조로 Solidigm은 분류하고 있다.

Solidigm은 G1에서 G4로 온도차이로 구분

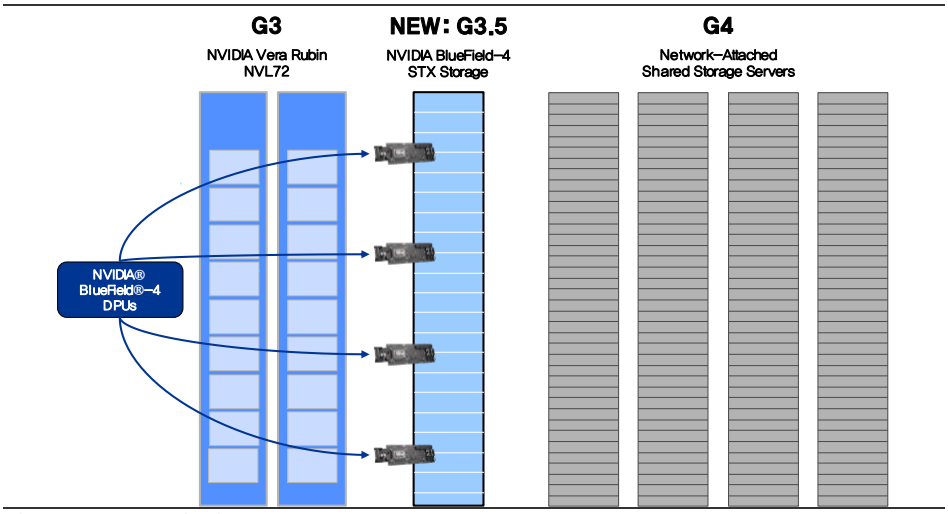
GPU HBM은 G1으로 weights, activations, 가장 뜨거운 KV cache를 담당하고, Host DRAM은 G2로 CPU staging과 KV cache prestaging을 담당하고, Local SSD는 G3인데 checkpointing, model load, warm KV spill에 사용되며, 새롭게 부상하는 KV cache Tier는 3.5G로 구분하고 agentic·multi-turn KV cache와 shared context를 처리한다. NVIDIA의 Blue-Field4 STX를 의미한다. Shared Storage는 G4이고 inactive KV, training lake, archive 등 장기 보관 데이터 역할을 담당한다. KIOXIA는 Network Storage라고 부른다.

그림 23. Storage Hierarchy



자료: Solidigm, IBK투자증권

그림 24. Storage Hierarchy: 3.5G는 NVIDIA BlueField 4 STX



자료: Solidigm, IBK투자증권

Local Memory 한계 극복하기 위한 Context Storage

AI Inference가 복잡해질수록 local storage만으로는 context data와 KV cache를 효율적으로 관리하기 어려워진다. Solidigm은 NVIDIA G3.5 계층에 해당하는 context storage를 통해 cache node와 KV cache offload를 지원할 수 있다고 본다.

Agent AI 대응에 필요한 솔루션

KIOXIA도 Context Memory Storage를 Agent AI를 대응하기 위한 솔루션으로 제시했다. SSD를 활용한 Context Memory Storage를 통해 long-context AI Inference를 가속할 수 있다고 설명한다. Context Mermoy Storage에는 CPU, GPU, SSD가 내장된다.

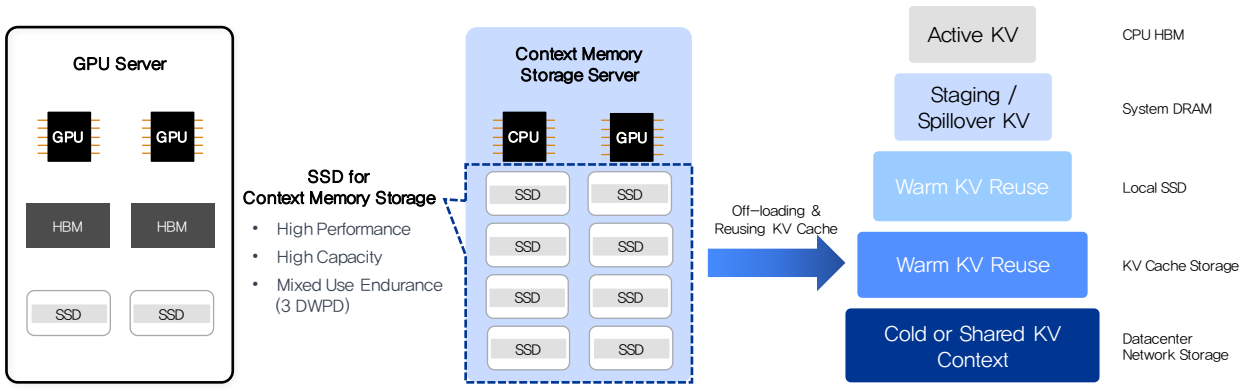
TFTT 27배 RAG Data 65% 절감

핵심은 KV cache를 SSD로 offload하고 재사용해 HBM과 DRAM의 용량·비용 제약을 보완하는 것인데 Solidigm은 Context Storage 단계에서는 recomputing context 보다 TTFT(Time To First Token)가 27배 빠르고, RAG Data는 DRAM 사용량을 줄임으로써 65% 절감할 수 있다고 분석했다. Shared Storage에서는 SSD 솔루션이 HDD를 장착한 서버 대비 1/9로 줄일 수 있고, 동일한 전력 기준으로는 GPU 서버를 50% 더 사용할 수 있는 것으로 분석했다.

AI Inference Data에 대해서 Solidigm은 저장공간 계층별로 역할을 자세히 분석했다. AI Infernce는 단순히 요구한 것을 서버에서 찾아서 대답하는 간단한 구조가 아니라고 강조한다.

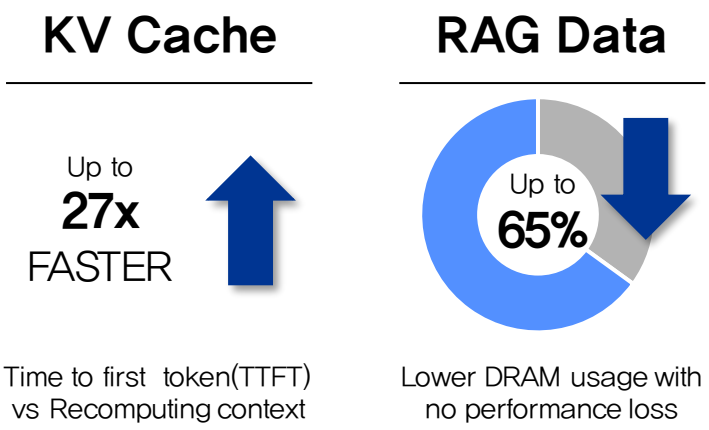
Request(Gateway and Policy / Schedular) 〉〉 Retrieval 〉〉 Context assembly 〉〉 Perfill + KV cache(Decode) 〉〉 Toll / action loop(Guardrails) 〉〉 Response or acton(Traces, evals, and logs)의 단계를 거치는 것이 AI Inference이다.

그림 25. Context Memory is accelerating long-context AI Inference



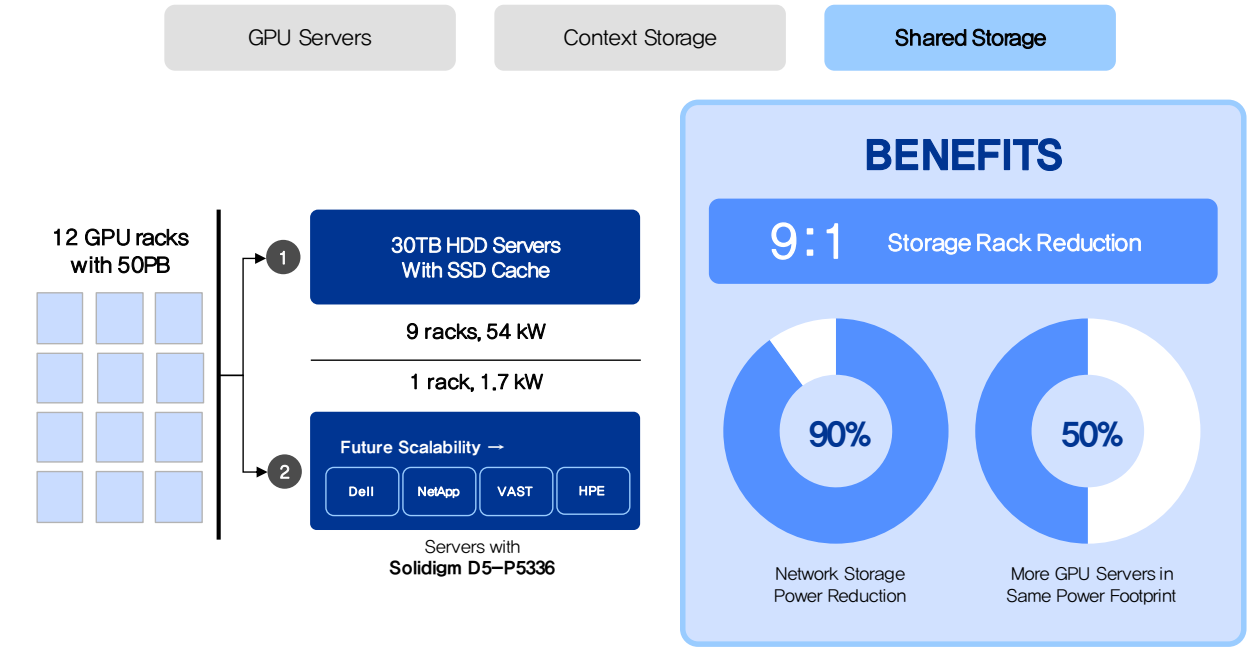
자료: KIOXIA, IBK투자증권

그림 26. Context Memory 효과



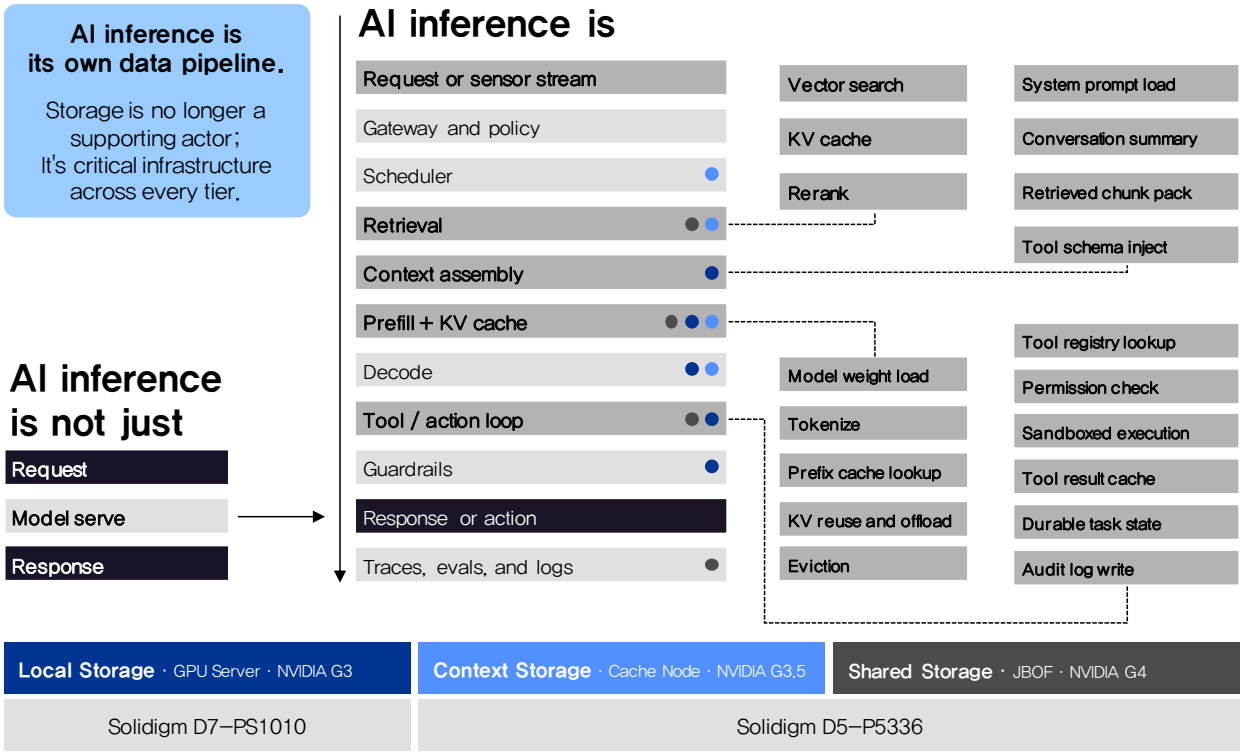
자료: Solidigm, IBK투자증권

그림 27. Shared Storage



자료: Solidigm, IBK투자증권

그림 28. AI Inference 단계별 사용되는 메모리 영역



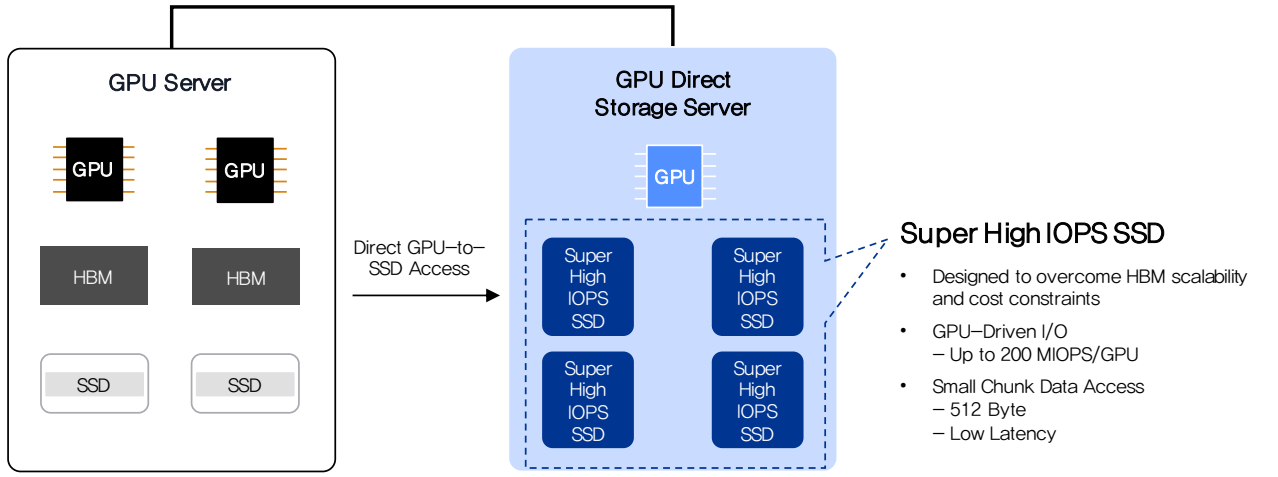
자료: Solidigm, IBK투자증권

GPU와 SSD로 구성된 서버

Computex 2026에서 가장 인상 깊은 제품 중 하나는 GPU Direct Storage이다. HBF는 GPU 옆에 DRAM이 아닌 NAND가 그 역할을 하는 개념인 것에 비해서 GPU Direct Storage는 서버 내에 GPU와 SSD만으로 구성되어 있다. GPU Server가 GPU와 HBM 그리고 SSD로 구성된 것과도 차별화되는 방식이다.

GPU Direct Storage는 GPU가 SSD에 직접 접근해 HBM의 확장성과 비용 제약을 보완하는 구조이다. 주로 graph neural network용 대규모 데이터 처리와 GPU 기반 vector DB indexing 가속에 사용된다.

그림 29. GPU Direct Storage



자료: KIOXIA, IBK투자증권

2. HDD : 차갑지만 넓은 공간은 필수적

대용량 저장장치도
AI 컴퓨팅에 필수

AI가 성장 동력이 되면서 HBM, DRAM, NAND까지는 거론되지만 HDD에 대한 필요성을 거론한 자료는 Computex 2026에서 처음 접했다. Data의 온도로 따지면 Cold data이다. 하지만 처리해야 하는 데이터가 계속 쌓이게 된다는 것을 고려하면 필요성에 대해서는 굳이 고민할 필요는 없다.

HDD 관련 내용은 Western Digital(이하 WD)의 의견으로 정리하였다.

AI 핵심 인프라는
Data system

WD는 AI 인프라는 compute보다 data system이라고 주장한다. AI 인프라는 단순 compute 경쟁이 아니라 데이터가 계속 생성·저장·재사용되는 data system으로 보고 있다.

반복되는 연산 이후
데이터는 쌓여 있다

연산은 반복 실행되지만 데이터는 시간이 갈수록 누적되는 특징을 갖고 있다. AI scale이 커질수록 storage는 보조 인프라가 아니라 AI 비용 구조와 시스템 설계를 결정하는 핵심 요소가 될 수 밖에 없다.

경제성을 무시할 수 없다

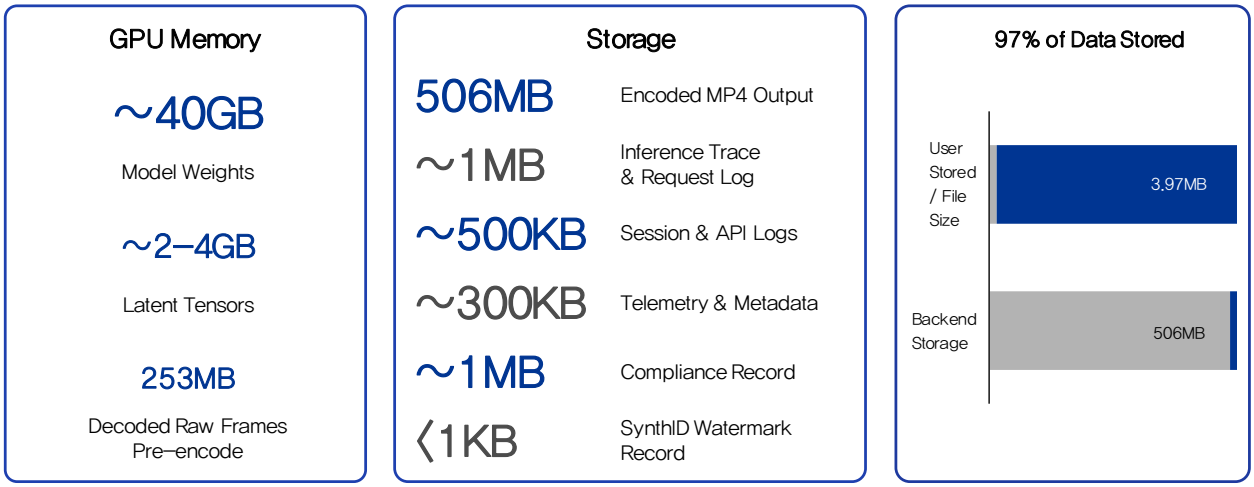
AI 규모가 작을 때는 storage 비용 차이가 크게 보이지 않지만, EB·ZB 단위로 커지면 작은 단가 차이가 막대한 비용 차이로 확대된다. $0.01/GB 차이가 1TB에서는 $10, 1PB에서는 $1만, 1EB에서는 $1,000만, 100EB에서는 $10억 차이로 커진다. WD는 HDD 30TB, QLC SSD 30TB를 비교했을 때 비용은 22배 차이 난다고 분석했다.

AI 인프라가 대규모화되면 storage 선택은 단순 부품 선택이 아니라 전체 architecture와 경제성을 결정하는 문제로 봐야 한다.

결과물은 3.97MB,
서버 데이터는 506MB

AI에게 여러 가지 설정을 한 이후 동영상을 제작하라고 지시한 후 8초 후에 720p 3.97MB의 비디오 파일을 결과를 얻었다. 여기에 사용된 GPU Memory 253MB를 포함한 Storage의 총 합은 506MB이고 이는 Storage에 저장되어 있다.

그림 30. 97% Data는 저장

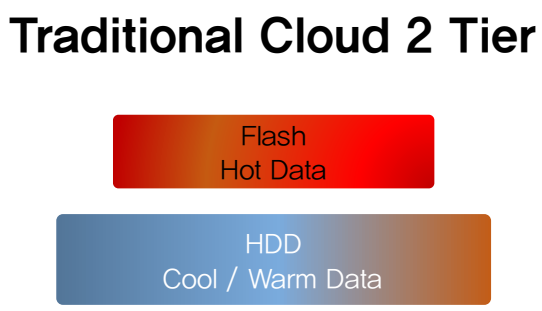


자료: Western Digital, IBK투자증권

Storage는 2 Tier에서 4 Tier

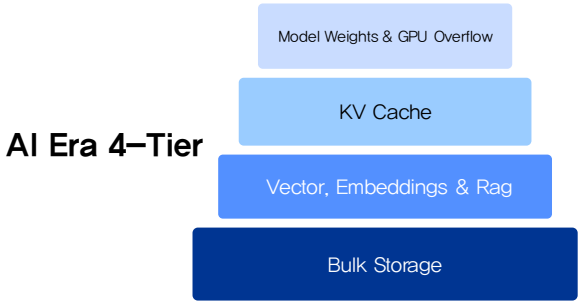
AI storage architecture의 다층화가 진행 중이다. 앞서 언급한 HBM, DRAM, NAND보다 상위 구분이다. 기존 cloud storage는 flash가 hot data, HDD가 cool·warm data를 담당하는 2-tier 구조가 중심이었고, AI 저장장치는 model weights·GPU overflow, KV cache, vectors·embeddings·RAG, bulk storage로 나뉘는 4-tier 구조이다.

그림 31. 전통적 Cloud 저장장치



자료: Western Digital, IBK투자증권

그림 32. AI Cloud는 4 Tier

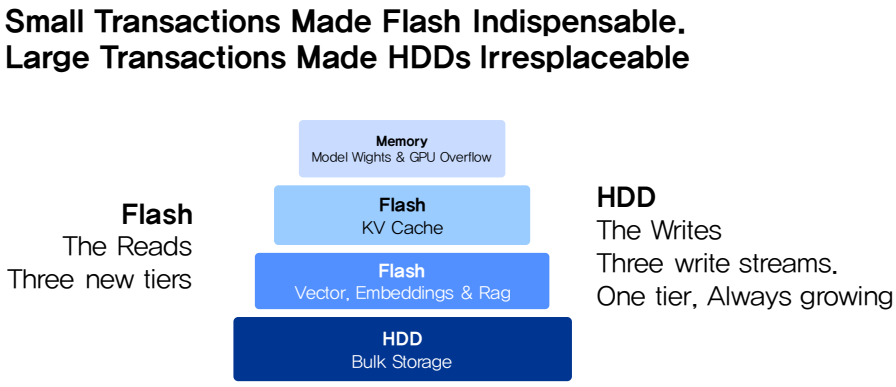


자료: Western Digital, IBK투자증권

Large transaction에서 HDD는 필수품

Small transaction 증가가 Flash를 필수 인프라로 만들었고, large transaction 증가는 HDD를 대체 불가능한 bulk storage로 만들었다. Flash는 빠른 read 중심 workload와 KV cache, vector·embedding, RAG 처리에 적합하고, HDD는 대규모 write stream, archive, synthetic data, training data, 장기 보관 데이터에 적합한 것으로 보고 있다. 결론적으로 AI storage는 Flash와 HDD 중 하나를 선택하는 문제가 아니라, workload별로 역할을 나누는 multi-tier 구조가 핵심이다.

그림 33. Large Transactions on HDD



자료: Western Digital, IBK투자증권

Agent AI는 빛의 속도로 달린다

컴퓨팅에서 전송이 병목이 될 것

Marvell은 새로운 관점으로 AI 인프라의 병목을 언급했다. 컴퓨팅 파워보다 데이터 전송 시스템 병목이 더 중요해 질 것으로 전망하고 있다. NVIDIA가 CPO(Co-Packaged Optics)를 통한 새로운 스위치를 개발한 것과 같은 맥락이다.

AI 성능을 더 이상 GPU / XPU 성능, 공정 노드, HBM 대역폭만으로 설명하기 어려워진다고 한다. 초기 AI 인프라 병목은 compute였고, 이후 Memory bandwidth가 핵심 과제로 부각됐으나, 이제는 connectivity가 다음 병목이 될 것으로 전망한다.

고속 / 저지연성 해결해야 할 문제

reasoning model, MoE(Model of Expert), Agent AI 확산으로 데이터 이동량이 급증하고, 인프라 전반에서 더 높은 대역폭과 낮은 지연시간을 요구하고 있다. 수만~수백만 개 프로세서가 하나의 거대한 컴퓨팅 엔진처럼 작동하려면 고속·저지연 연결성이 핵심이다.

세계 최대 하이퍼스케일러들도 AI 인프라 확장의 본질을 연결성 문제로 보고 네트워크 아키텍처를 전면 재설계 중에 있다고 한다.

표 6. 주요 기업 연결성 관련 발언

구분	설명
Meta	"The requirements have grown beyond what can fit in an existing data center campus."
Google	"Transforming globally distributed infrastructure into one seamless supercomputer."
AWS	"The way networks had been built wasn't going to scale into the future and that something fundamentally different had to happen."
Microsoft	"The speed of light is now a key bottleneck."

자료: Marvell, IBK투자증권

데이터간 이동 거리는 다양

AI 데이터센터 연결은 데이터센터 간 수백~수천 km, 데이터센터 내부 수백 m, Rack 내부 수 m, 패키지 내부 mm 단위까지 다양한 거리 구간으로 구성되어 있다. 각 거리마다 필요한 기술, 엔지니어링 역량, 공급망이 다르기 때문에 단일 솔루션으로 전체 인프라를 커버하기 어렵다.

거리별로 광모듈, DSP, PAM4 등 다양한 솔루션

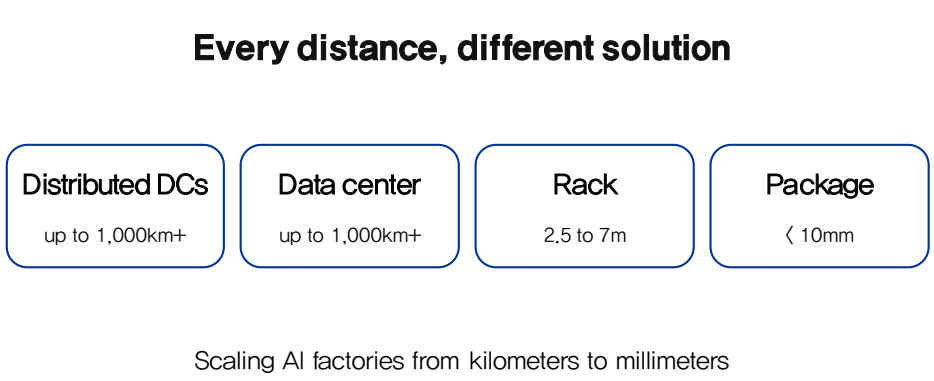
데이터센터 간 Scale Across 영역에서는 Coherent DSP(Digital Signal Processor) 기반 장거리 광모듈이 핵심이며, 데이터센터 내부 Scale Out 영역에서는 PAM4(Pulse Amplitude Modulation 4-level) DSP, TIA(Transimpedance Amplifier), 레이저 드라이버, 이더넷 스위치가 핵심이다.

Rack 내부는 아직 구리선

Rack 내부는 모든 프로세서가 다른 모든 프로세서와 직접 통신해야 한다. Scale Up 영역은 현재 구리 기반 고속 신호 전송 중심이다.

패키지 내부에서는 die-to-die interface와 Advanced Packing이 중요하며, chiplet 간 초고속 short-reach 연결이 핵심 과제로 부상하고 있다.

그림 34. Every Distance, different solution



자료: Marvell, IBK투자증권

구리선은 물리적 한계

AI 데이터센터 확장의 핵심 변곡점으로 Marvell은 Copper Wall을 제시했다. 구리는 단순하고 비용이 낮아 가능한 오래 사용해야 하지만, 대역폭이 높아질수록 전송 가능 거리가 짧아지는 물리적 한계가 존재한다.

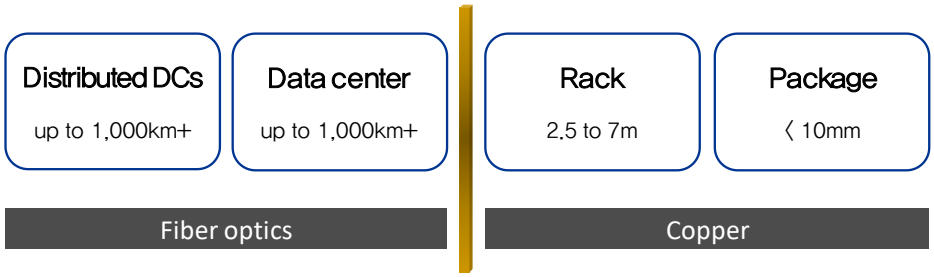
200Gbps는 2.5m까지만 대응

대역폭이 2배 늘어날 때마다 구리 케이블의 유효 전송 거리는 절반 수준으로 감소한다. 현재 가장 빠른 200Gbps/lane에서 구리 케이블 길이가 약 2.5m 수준까지 제한되며, Rack 높이와 내부 배선 경로를 고려하면 이미 한계에 근접한 것으로 보고 있다.

400Gbps는 Rack 내부도 광

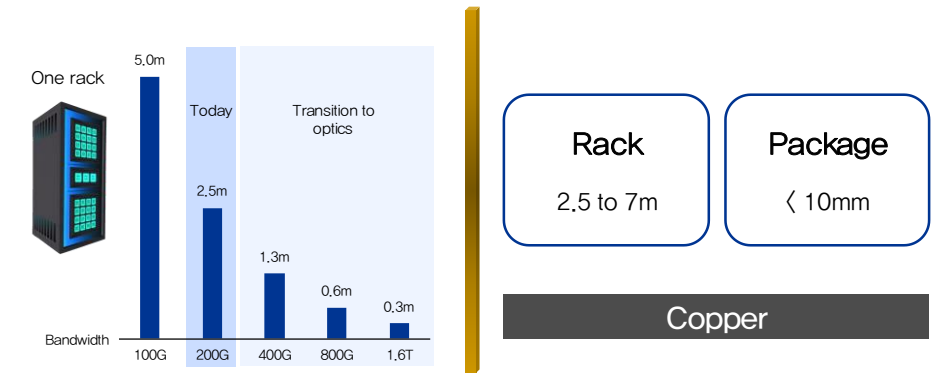
400Gbps/lane 시대로 넘어가면 Rack 내부 전체를 구리로 완전히 연결하기 어려워지고, 결국 Rack 내부 연결도 광으로 전환될 수밖에 없을 것으로 전망하고 있다. 20년 전 데이터센터 내부 연결이 구리에서 광으로 전환됐던 것처럼, 이번에는 전환 범위가 Rack 내부까지 확장되는 국면이다.

그림 35. Every Distance, different solution



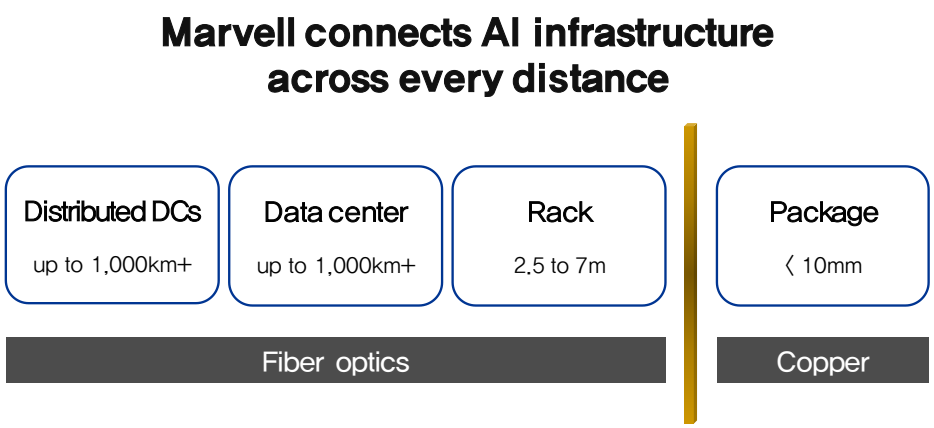
자료: Marvell, IBK투자증권

그림 36. 현재는 Copper



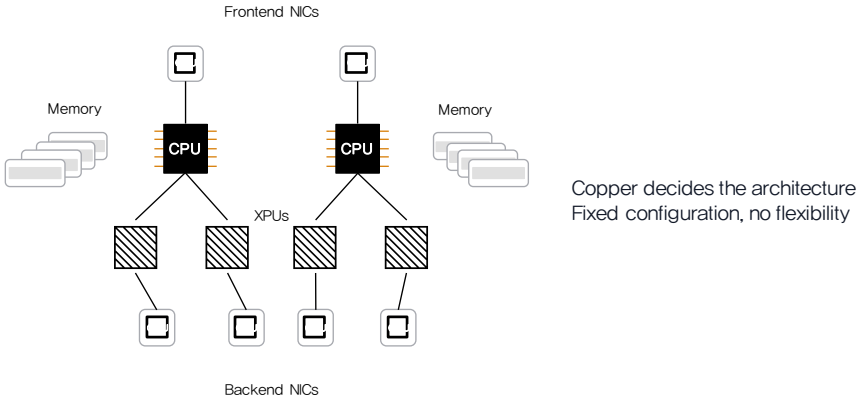
자료: Marvell, IBK투자증권

그림 37. Rack까지 Optics



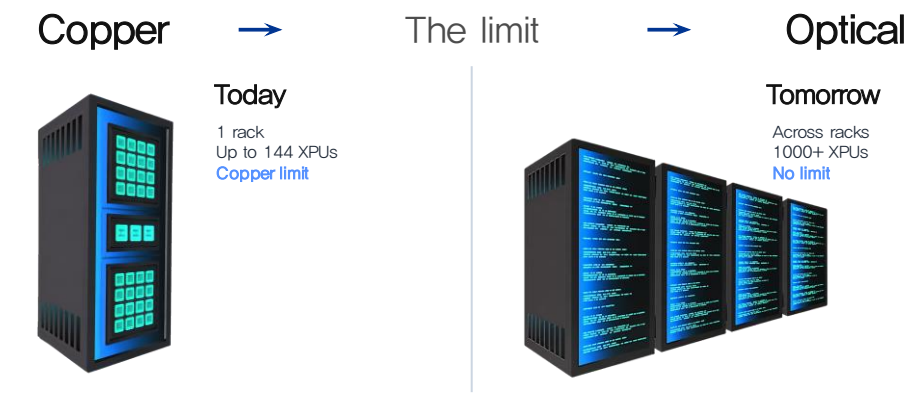
자료: Marvell, IBK투자증권

그림 38. Copper는 유연성이 낮다



자료: Marvell, IBK투자증권

그림 39. Optical이 미래다



자료: Marvell, IBK투자증권

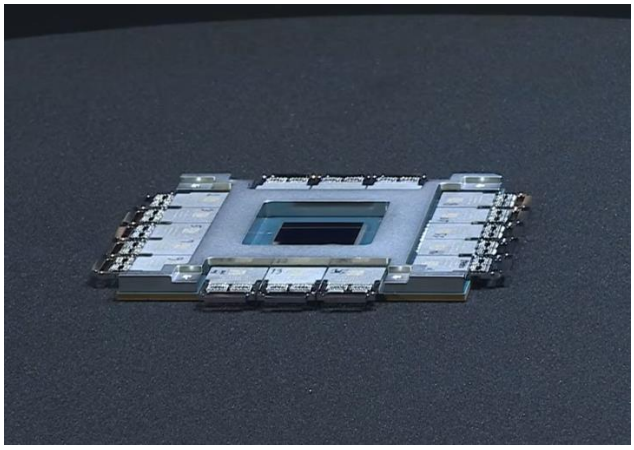
CPO가 Rack 내부 전송의 솔루션

CPO는 Rack 내부 Optical 전환의 핵심 기술이다. Vera Rubin 시스템에 NVIDIA가 적용하고 있다. 기존 방식은 스위치 실리콘에서 PCB 구리 trace를 거쳐 전면 패널의 광모듈로 신호를 보내는 구조인데 비해서 CPO는 optical engine을 스위치 또는 compute package 바로 옆에 배치해 구리 trace를 줄이고, 광섬유를 package 가까이까지 끌어오는 방식이다.

Rack 내부 연결 수는 Rack 간 연결보다 약 10배 많기 때문에 기존 pluggable optical module 방식으로는 전력·공간·밀도 문제를 해결하기 어렵다. CPO는 이를 지원한다.

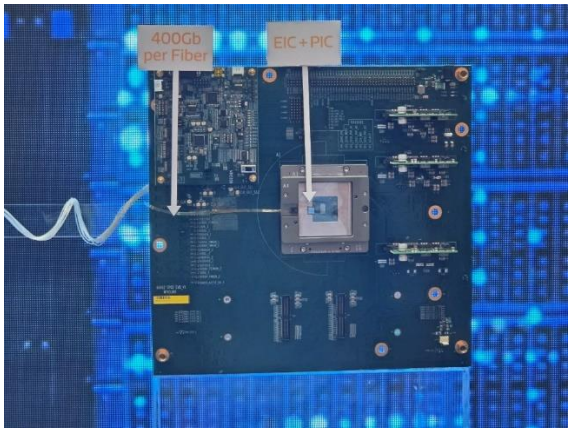
Computex 2026에서 MediaTek도 관련 제품을 공개했다.

그림 40. CPO(Co-Packaged Optic)



자료: Marvell, IBK투자증권

그림 41. MediaTek CPO



자료: MediaTek, IBK투자증권

거리 제약 없는 데이터센터

광 네트워크의 확산은 거리의 한계가 없어진 데이터센터를 의미한다. 광 연결이 Rack 내부와 서버 내부까지 확장되면 CPU, XPU / GPU, Memory, network interface를 물리적으로 한 시스템 안에 고정할 필요가 없어진다.

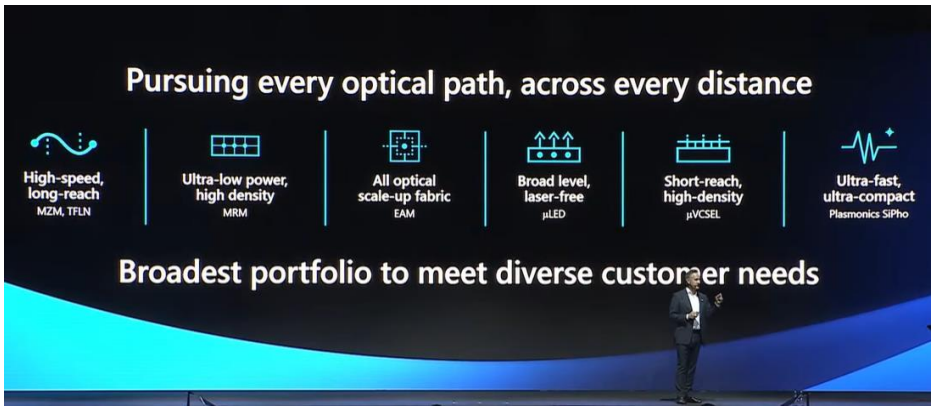
Compute, Memory, Network이 따로 분리 되어 배치도 가능

현재는 메모리 대역폭과 지연 때문에 CPU·XPU·Memory가 보드 위에서 가까이 배치되어야 하지만, optical 연결이 확대되면 자원을 별도 pool로 분리할 수 있을 것으로 전망한다. compute pool, Memory pool, networking pool을 workload에 따라 동적으로 조합하면 고정된 CPU:GPU:Memory 비율에서 발생하는 자원 낭비를 줄일 수 있다.

AI인프라는 연결성이 아니라 모델 / workload가 변수

AI workload는 scale-up cluster 크기에 맞춰 쪼개지는 것이 아니라, 모델의 필요에 따라 더 큰 단위로 실행 가능하게 되고 궁극적으로 수백만 개의 Compute, Memory, Network resource가 하나처럼 작동하는 구조가 가능해진다. 결론은 AI 인프라 아키텍처가 연결성의 한계가 아니라 모델과 workload의 필요에 의해 정의되는 시대로 이동할 것으로 전망한다.

그림 42. 고객별 다양한 솔루션



자료: Marvell, IBK투자증권

Floating Data center

부지, 전력 부족을 극복하기 위한 대안

Super Micro 발표 세션에 삼성중공업이 찬조 출연하면서 Floating Data center에 대한 발표가 있었다. Data center가 절대적으로 부족하지만 전력, 부지 확보, 냉각 시스템 등으로 건축 속도가 수요를 못 따라오고 있기 때문에 고안된 대안으로 파악된다.

Floating Data center는 육상 건물이 아니라 수상에 바지선, 해상 플랫폼, 부유식 구조물 위에 서버, 전력 설비, 냉각시스템을 올려 운영하는 방식이다.

부지 확보에 강점이 있다. Keppel은 싱가포르에서 25MW 규모로 진행 중이고 2028년 가동을 목표로 하고 있다. 확장성에서도 강점이 있을 것으로 삼성중공업이 제시하고 있다.

공기 단축, 확장성이 강점

지상에서 구축하는 것 보다 50% 빠른 공정 속도로 건축할 수 있다는 장점이 있다. 삼성중공업은 모듈화 된 블록을 해상 구조물에 쌓는 방식을 제안하고 있다.

그림 43. Floating Data Center

자료: 삼성중공업, IBK투자증권

그림 44. Floating Data Center– 모듈화로 높은 확장성

자료: Super Micro Computer, IBK투자증권

외부 환경을 이용한 냉각

바다나 강의 물을 열교환에 활용할 수 있다는 장점이 있다. Nautilus Data Technologies는 캘리포니아 Stockton에서 수냉 기반 인프라를 사용하고 있다.

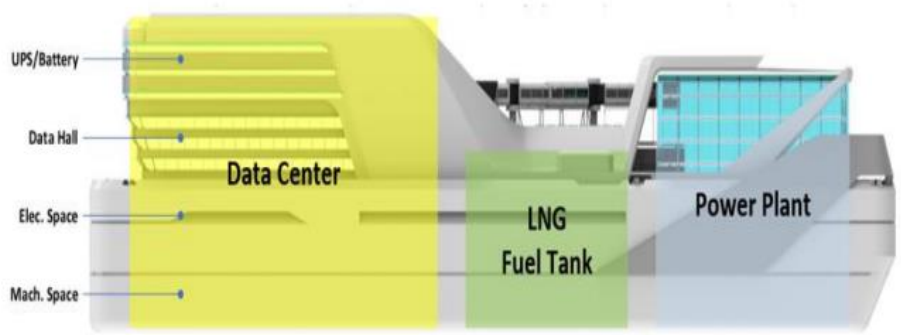
전력 확보 용이

독립적인 전력 공급 시스템 확보도 가능한데 LNG를 이용한 발전, 해상 풍력을 이용할 수 있다.

Risk도 고려 사항

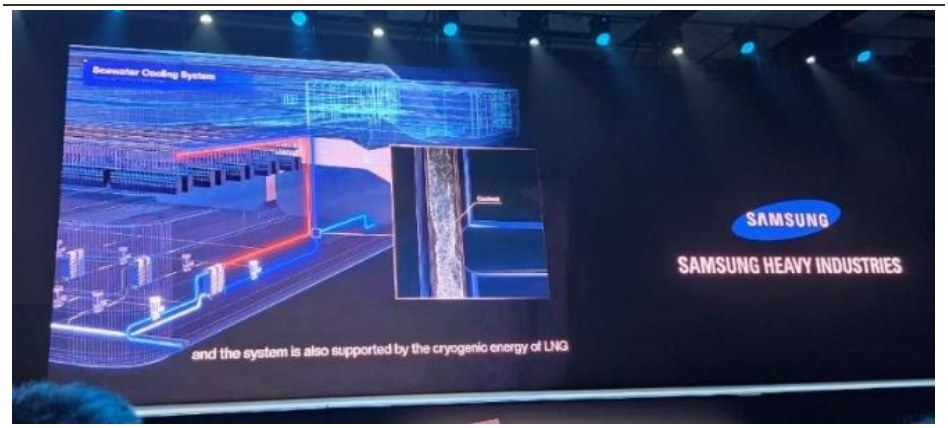
지상 건물에 비해서 Risk도 있는데 염분, 파도, 규제, 정비 난이도, 환경 문제가 발생할 여지가 있다.

그림 45. Floating Data Center – 독립 전략 시스템 확보



자료: 삼성중공업, IBK투자증권

그림 46. Floating Data Center – LNG를 이용한 자체 냉각



자료: Super Micro Computer, IBK투자증권

IBKS Research Center

성명	직급	담당업종	전화	이메일
용대인	전무(부문장)	총괄	6915-5400	daeinyong@ibks.com
이승훈	상무대우(본부장)	AI/인터넷/게임	6915-5680	dozed@ibks.com
투자분석부				
변준호	연구위원	Strategy	6915-5670	ymaezono@ibks.com
정용택	수석 Economist	Economy	6915-5701	ytjeong0815@ibks.com
김인식	연구위원	자산배분/ETF	6915-5472	kds4539@ibks.com
정형주	연구위원	채권/크레딧	6915-5654	hj.jeong@ibks.com
기간산업분석부				
이동욱	연구위원	에너지/소재	6915-5671	treestump@ibks.com
남성현	연구위원	유통·식자재/지주	6915-5672	rockrole@ibks.com
이현욱	연구원	자동차/2차전지	6915-5659	hwle1125@ibks.com
오지훈	연구원	조선/기계	6915-5662	jihoonoh@ibks.com
혁신기업분석부				
김운호	연구위원	IT/반도체	6915-5656	unokim88@ibks.com
김태현	연구위원	음식료/유틸리티/통신	6915-5658	kith0923@ibks.com
조경진	연구위원	해외주식	6915-5464	ckjins@ibks.com
조정현	연구원	건설/부동산	6915-5660	controlh@ibks.com
코스닥리서치센터				
이건재	연구위원	소재·부품·장비/스몰캡	6915-5676	geonjaelee83@ibks.com
정이수	연구위원	제약/바이오	6915-5677	ysjeong306@ibks.com
강민구	연구원	IT/디스플레이/미드·스몰캡	6915-5473	kmg@ibks.com
김혜빈	연구원	로봇	6915-5669	hyebhinkim@ibks.com
유창근	연구원	헬스케어	6915-5686	ucck0726@ibks.com

“국민과 중소기업에 필요한 참 좋은 IBK투자증권”

IBK기업은행 금융그룹
IBK투자증권

서울특별시 영등포구 여의도동 국제금융로 6길 11
대표번호 02-6915-5000
고객지원부 1588-0030, 1544-0050

IBKS Family Office	02) 536-4070	IBK WM센터 대구	053) 752-3535
영업부	02) 6915-2626	IBK WM센터 광주	062) 382-6611
강남센터	02) 2051-5858	IBK WM센터 일산	031) 904-3450
강남역 금융센터	02) 532-0210	IBK WM센터 판교	031) 724-2630
분당센터	031) 705-3600	IBK WM센터 평촌	031) 476-1020
IBK WM센터 강남센트럴	02) 556-4999	IBK WM센터 천안	041) 569-8130
IBK WM센터 목동	02) 2062-3002	IBK WM센터 부산	051) 741-8810
IBK WM센터 도곡	02) 2057-9300	IBK WM센터 창원	055) 282-1650
IBK WM센터 한남동	02) 796-8500	IBK WM센터 울산	052) 271-3050
IBK WM센터 중계동	02) 948-0270	IBK WM센터 시화공단	031) 498-7900
IBK WM센터 반포자이	02) 3481-6900	IBK WM센터 남동산단	032) 822-6200
IBK WM센터 동부이촌동	02) 798-1030		