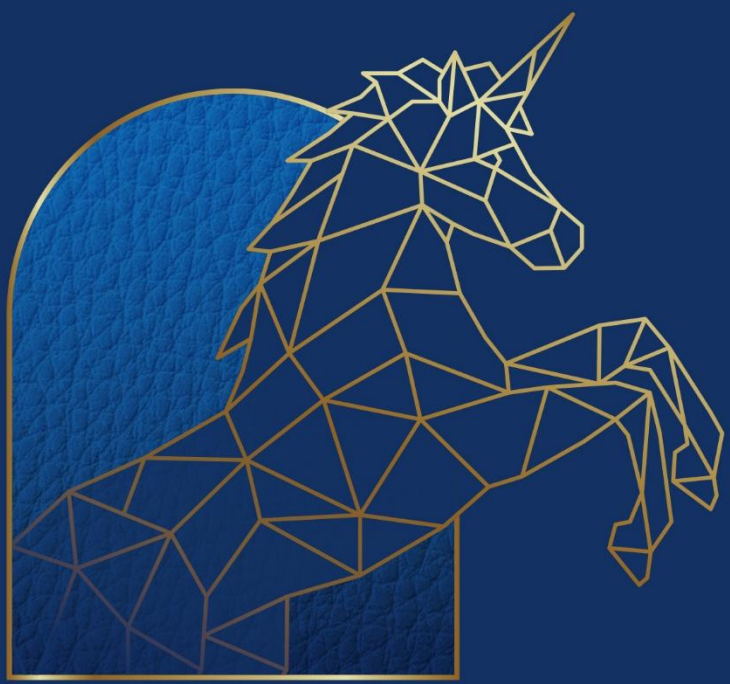
신한 Premier 리서치

Shinhan

망고부스트코리아

풀스택 인프라로 뚫는 에이전트 AI 시대의 병목



[IT팀]

김형태 수석연구원
✉ calebkim@shinhan.com

송혜수 연구원
✉ hyesus@shinhan.com

비상장/벤처 시리즈 #72

The Search for Unicorns
유니콘을 찾아서

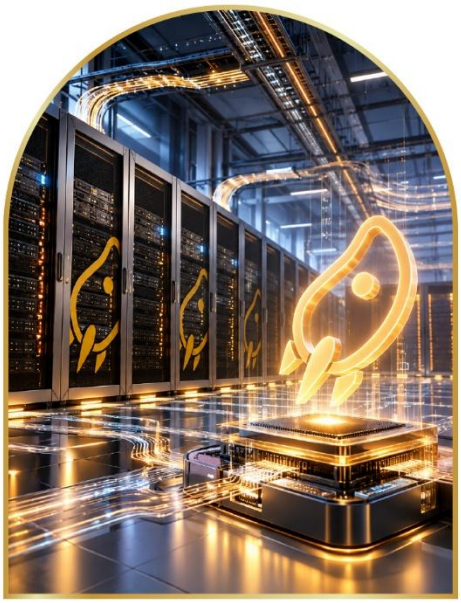
신한투자증권

비상장/벤처
시리즈 #72

The Search for Unicorns
유니콘을 찾아서



MANGOBOOST



* AI 생성 이미지

2021년
- 서울대학교 기반 DPU 핵심기술 연구 착수

2022년
- 망고부스트 설립 및 시드 투자 유치

2023년
- 시리즈 A 투자 유치 및 DPU 기술 PoC 시연

2024년
- UEC·UALink 컨소시엄 및 리눅스 재단 OPI 프로젝트 참여
- 400Gbps 고성능 DPU 샘플 출시

2025년
- MLPerf 5.0·5.1 성능 기록 수립 및 AMD·슈퍼마이크로와 공동 기술 백서 발표
- AI 추론 소프트웨어 Mango LLMBoost™ 출시

2026년
- 글로벌 CSP 대상 BoostX™ DPU 공급 확대 및 LLMBoost™ 기업 도입 확대
- MLPerf 6.0 세계 최초 멀티리전 이기종 추론 구현 및 에이전틱 AI 인프라 솔루션 출시

Business Highlights

Agentic AI 시대의 키워드 '풀스택 인프라'

망고부스트는 2022년 설립된 풀스택 인프라 기업으로 H/W와 S/W 공동 설계에 기반한 동적 인프라 생태계 구축에 앞장서고 있다. AI 인프라 스택의 모든 계층을 자체 보유한 기술력이 강점이다. 데이터 병목이 발생하는 여러 사례들에서 특화 칩의 중요도가 높아졌다. CPU 병목을 완화하는 DPU 개발에 그치지 않고 워크로드 특성 변화에 맞춘 H/W-S/W의 유기적 결합 솔루션을 지향하고 있다. 주요 파트너사들과의 활발한 기술 협력도 진행중이다. 본사는 미국 밸류에 위치해 있으며 산호세, 토론토, 도쿄, 인도네시아, 서울에 거점을 갖추고 있다.

기존 인프라의 한계를 개선할 세 가지 솔루션

① AI Agent OS(운영체제): H/W와 S/W 애플리케이션 사이 복잡한 오케스트레이션(진행 순서, 단계별 역할 분담, 검토, 승인 등 작업 전반 조율)을 담당한다. 분산 추론 엔진은 1) 멀티노드 확장, 2) 프리필(Prefill, 전체 프롬프트 인식, 입력 토큰에 대한 키/값 생성 후 KV캐시 저장)과 Decode(순차적으로 토큰을 생성)를 분리, 3) 도구 호출, 메모리 관리, Agent 간 경계 지정 등 부가 기능을 제공한다.

② 이기종 컴퓨팅(Heterogeneous Compute): 기존 인프라는 대부분의 단계를 동일 GPU에서 실행한다. 가벼운 워크로드까지 고가의 칩에서 구동하는 것은 자원 활용도 면에서 비효율적일 수밖에 없다. Agent의 비용 축소를 위해서는 단계별로 세분화된 칩 활용이 필요하다. 무거운 디코드는 플래그십 GPU에 할당하고, 프리필과 가벼운 디코드에 가성비 GPU 및 NPU를 활용한다. 도구 호출과 전반적인 작업 조정은 CPU에 일임하나 자체 DPU를 통해 연산 부하를 추가로 개선한다.

③ Agentic Memory Expansion: Agent AI는 방대한 양의 컨텍스트와 KV캐시를 유지해야 한다. HBM에 의존할 경우 용량 소진 속도가 너무 빠르다. 고가의 SSD 대신 HDD를 활용하는 하이브리드 인프라 솔루션도 관심이 높아지고 있다. 상대적으로 가격이 저렴한 HDD를 DPU로 가속하는 기술은 전체적인 스토리지 구축 비용 절감을 가능케 한다. 공급부족이 이어지고 있는 메모리 계층 전반의 다양한 연구개발은 풀스택 솔루션의 이점을 강화하는 요인으로 판단된다.

실적 및 주요 체크포인트

2025년 실적은 매출액 70만달러, 영업적자 2,270만달러를 기록했다. 기술 개발, 인프라 구축에 선제적 투자가 이뤄진 시기다. 상업화 이전 단계로 해석할 필요가 있다. 시장 개화가 본격화되는 2026년 이후부터 성장 가시성이 뚜렷해질 가능성이 높다. 보유 특허 50건 이상, 10년이 넘는 연구개발 이력, 누적 투자금 1억달러 초과 유치 등이 성장 잠재력과 시장 확장 기대감을 대변한다. 2026년 1~3분기 매출액은 190만달러 수준으로 이미 2025년 대비 2.7배 성장했으며 5백만달러 이상 수주잔고를 확보했다. 수주 기반 레퍼런스 강화가 동반될 2027년부터 동사의 풀스택 기술 경쟁력에 시장 관심이 높아질 것으로 예상한다.

기업 개요

'풀스택 인프라'를 제공하는 AI 병목 해결사

망고부스트는 Agentic AI 시대 개화 이후 가장 많이 언급된 '시스템 단위 병목'을 최적화하는 풀스택 솔루션 제공 업체다. H/W와 S/W 공동 설계에 기반한 풀스택 시스템으로 물리적 한계가 존재하는 단일 계층 최적화의 돌파구를 모색한다.

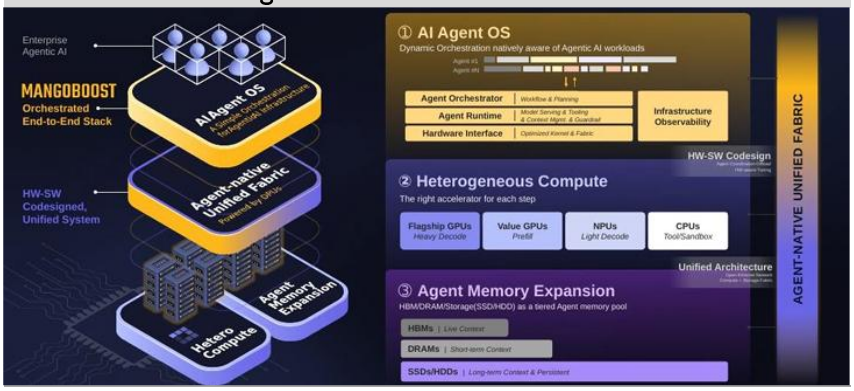
기존 인프라의 한계는 단일 모델 개발, 학습 위주로 구성된 시스템 구조에서 발생한다. 단순화된 시스템으로 모델 관리, 배포(Deploy), 운영이 비교적 쉽다는 장점이 있으나 Agentic AI 이후 여러 개의 모델을 묶어서 서비스하는 '멀티 모델 서빙'이 확산되고 있다.

AI 모델 확장(Model Scaling), 추론(Reasoning Scaling)이 주도한 AI의 발전은 또 한 번의 분기점에 도달했다. 초기 단계의 인프라 구성은 단일 모델이 자원을 독점하는 형태의 인프라 활용이 보편적이었다. 다수의 모델이 활용되는 현 시점은 사용자 중심의 시장 환경이 본격화되는 구간으로 볼 수 있다.

Agentic AI의 확산, 다수의 Agent가 협력하는 시스템 레벨의 확장에서는 이미 구축된 초기 단계 인프라의 구조 변화(전방위적인 데이터센터의 시스템 재설계)가 필연적이다. 추론 시장에서 요구하는 토큰 효율성(동일 GPU, 최대 가치창출) 개선을 위해서는 시스템 단위의 최적화가 동반돼야 한다.

동사는 자원 배분, 연산, 메모리를 3대 병목으로 지목하고 1) AI Agent 운영체제, 2) 이기종 연산 최적화 및 분배 체계 재구성, 3) 계층형 Memory 확장을 통한 시스템 재설계를 주력 사업으로 추진하고 있다. 'Agent AI Factory'로 언급된 사업 포트폴리오는 노드(Node), 랙(Rack), 데이터센터의 세 단계 구성을 제공한다.

망고부스트 주력 사업 'Agentic AI Infra 솔루션'



자료: 회사 자료, 신한투자증권

망고부스트 제품 포트폴리오 'Agent AI Factory'

제품명	제품 유형	판매 대상	핵심 차별점
"LLMBoost (→ Agent OS)"	소프트웨어	분산 추론 최적화 · 엔터프라이즈급 소프트웨어	자동 튜닝, 멀티노드 확장, 이기종 통합, 에이전트 특화 기능, 인프라 가시성
BoostX	DPU / IP	서버 간(East-West) 트래픽 서버-스토리지 간(North-South) 트래픽	완전 오프로드/가속 + 프로그래밍 가능성/적응성
Alphonso	AI 서버	턴키 컴퓨트 서버	플러그 앤 플레이, 멀티노드 확장, 이기종 확장, 네트워크 유휴시간 감소
Kesar	스토리지 서버	AI 가속기용 스토리지 서버	성능과 용량의 독립적 확장(분리형), 헤드리스 아키텍처(BOM 절감), KV 캐시 가속

자료: 회사 자료, 신한투자증권

사업 포트폴리오 단계별 구성

단계	활용 범위	대상 고객
노드(Node)	기존 인프라와 플러그 앤 플레이로 연결. 코딩 등 범용 에이전트 지원	모든 고객군
랙(Rack)	전문 분야별 멀티에이전트 서비스 운영	특화 연구 조직, 소버린 AI, 버티컬 AI
데이터센터	콜로케이션, API 기반 클라우드 배포 등 대규모 에이전트 집단 및 인프라 관리	대기업, AI 서비스 산업, 프론티어 AI

자료: 회사 자료, 신한투자증권

주요 경영진

성명	직책	주요 경력
김장우	CEO	카네기 멜런 전기공학, 컴퓨터공학 박사 / 썬 마이크로시스템즈
Eriko Nurvitadhi	CAIO	카네기 멜런 전기공학, 컴퓨터공학 박사 / 인텔 랩스
권동업	CTO	서울대 전기전자공학 박사 / 인텔 랩스
박준기	COO/CFO	서울대 경제학 박사 / 삼성경제연구소
박찬익	CCO	서울대 컴퓨터공학 박사 / 삼성전자
Tom Spencer	사업개발 VP	애리조나 전기공학 / 브로드컴, 인텔

자료: 신한투자증권

실적 및 주요 체크포인트

2022년 설립 직후 1,060만달러의 대규모 Seed 투자를 유치했다. 이후 Series A(2023년)에서는 5,520만달러 추가 조달에 성공했다. 2025년말~2026년 상반기 진행된 SAFE(Simple Agreement for Future Equity)에서는 3,470만달러를 조달하며 누적 조달 금액이 1억달러를 돌파했다.

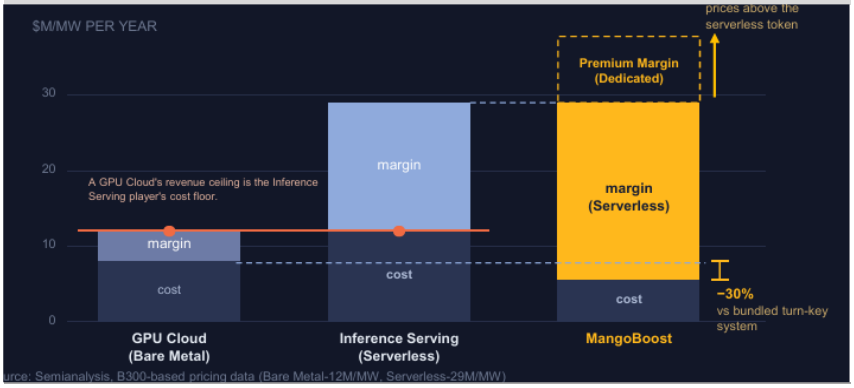
박사급을 포함한 140명 이상의 연구인력과 엔지니어들은 Agentic AI 시대를 위해 구축된 핵심 자원으로 평가된다. 연구개발만 10년 이상 진행하며 대형 프로세서 업체와의 협력도 도출해냈다. 현재 500만달러 이상의 수주잔고가 확보된 상태로 2027년부터 레퍼런스 강화가 본격화될 것으로 기대하고 있다. 동사의 사업성 및 경제성은 보편화된 데이터센터 임대 형태로 볼 수 있는 GPU 클라우드, 네오클라우드(추론 서빙), NPU 칩 적용과 직접 비교가 가능하다.

GPU 클라우드와 비교 시 1) 더 저렴한 GPU를 활용할 수 있는 워크로드에서 동등한 수준의 연산 성능 구현이 가능하다. 2) DPU·스위치·케이블을 통합해 프리미엄이 적용된 경쟁사들의 네트워크 시스템과 달리 비용 효율화를 지향한다. 3) 목적에 맞는 스토리지를 구분해 불필요한 범용 스토리지 용량을 최소화한다. 이를 통해 약 30%의 비용 절감 효과를 기대할 수 있다.

직관적인 차이점은 토큰 가격 대비 성능에서도 확인된다. 네오클라우드 업체의 GPU를 임대할 경우 임대료가 비용 하한선이 된다. 동등한 성능 구현이 가능하면서도 네오클라우드 대비 저렴한 비용이 수익성 제고 효과로 이어질 수 있다.

NPU 칩과의 비교에서는 구조적 차이가 존재한다. 1) NPU 칩이 경제적일 수 있으나 네트워크, 스토리지, S/W를 따로 구축해야 하는 경우 추가 비용이 발생한다. 2) 특정 칩에 맞춰 개발된 모델이나 애플리케이션에 따라 수익성이 크게 달라지는 경우가 많다. '풀스택 인프라'은 전체 시스템이 제공되고 H/W와 S/W의 최적화를 동시에 지원한다는 점이 강점으로 작용한다.

AI 인프라 사업 마진 형성 구조: H/W 내재화, S/W 수직계열화 → 수익성 차별화



자료: 회사 자료, 신한투자증권

투자포인트

Agentic AI 워크로드와 기존 시스템의 구조적 간극

학습에서 추론으로 연산의 중심 축이 이동하고 있다. 초기 AI 시장은 프론티어 모델을 만드는 것에 모든 자원이 집중됐다. 단일 모델을 주로 활용하는 생성형 AI와 Agentic AI의 가장 큰 차이점은 1) 모델의 수, 2) 요청 단계(Step), 3) 반복적인 도구 호출 구조에 있다. Agentic AI의 보편화를 주도한 Claude Opus 4.6 출시를 기점(2026년 2월)으로 토큰 소비량이 급격히 증가했다는 점에 주목한다.

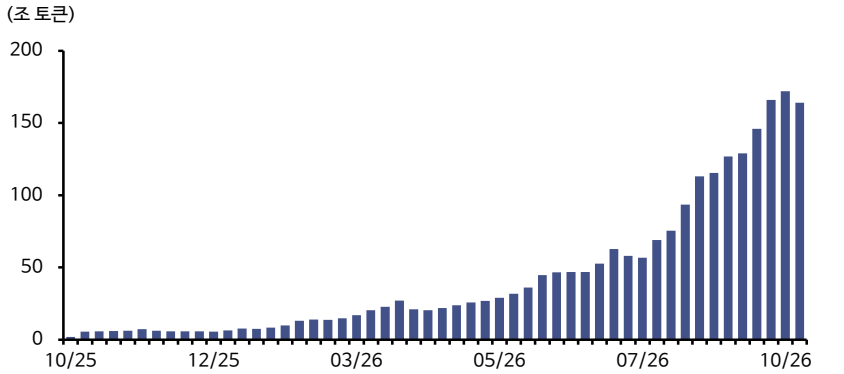
일반적인 생성형 AI의 프레임워크는 한 번의 요청을 단일 모델로 처리하며 한 가지의 답변을 내놓는다. Agentic AI는 결과물을 도출하는 과정에서 계획 수립, 도구 호출, 결과 검증, 반복 작업을 지정된 다수의 모델로 분산 처리할 수 있다.

결과를 재검토하는 과정에서는 반복적인 도구 호출이 이뤄진다. 기존 생성형AI 대비 토큰 소비량이 증가할 수밖에 없는 이유다. 글로벌 토큰 소비량은 2026년 초 35경에 도달했다. 2024년과 비교하면 150배가 넘는 수준이다. 추론형 AI 등장 이후 2025년 말부터 Agentic AI가 급부상했다는 점을 감안하면 성장 속도는 더욱 가속화될 것으로 보인다.

데이터센터 밸류체인 내 토큰 생성, 처리량을 보는 시선은 크게 세 가지로 분류할 수 있다. 1) 반도체 업체들은 이를 성능 지표로 활용한다. 2) 서비스 제공자는 AI 서비스 수요(Q)로 인식한다. 3) 가장 중요한 사용자(B2B/B2C)는 토큰 하나 하나가 모두 비용이다.

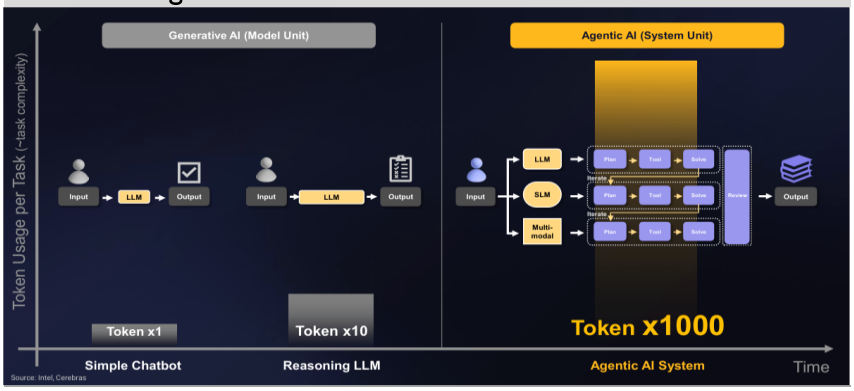
결국 AI 도입을 추진하는 최종 사용자들이 원하는 것은 토큰당 비용 절감으로 귀결된다. 서비스 사업자들이 토큰당 단가를 낮추기 위해서는 데이터 인프라의 자원 효율화가 필요하다는 점에서 시스템 레벨의 풀스택 솔루션이 갖는 중요도가 높아질 것으로 판단된다.

오픈라우터 주간 토큰 소비량 추이



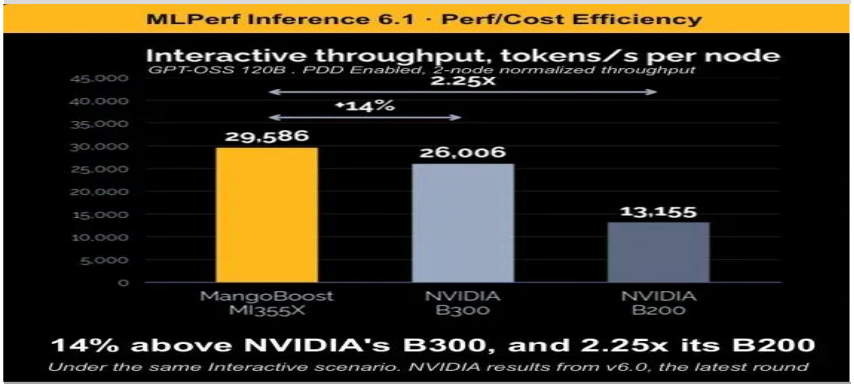
자료: OpenRouter, 신한투자증권

생성형 AI vs. Agentic AI 토큰 소비량



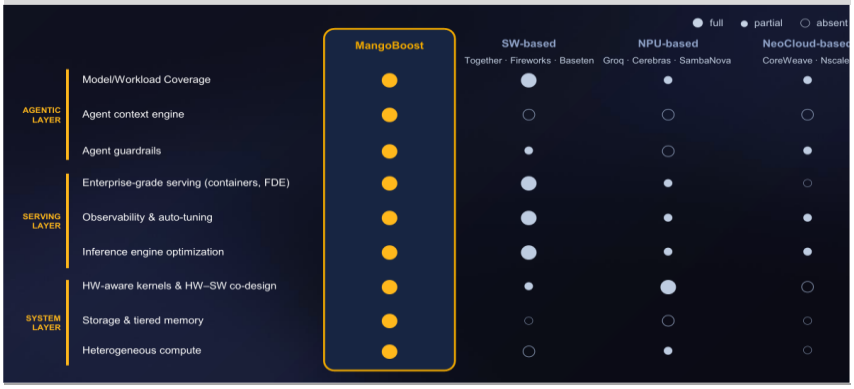
자료: 회사 자료, 신한투자증권

이종 연산 클러스터 벤치마크(성능 대비 비용 효율): MLPerf 6.1



자료: 회사 자료, 신한투자증권

풀스택 솔루션: 레이어 수직계열화 및 경쟁우위



자료: 회사 자료, 신한투자증권

인프라의 한계를 개선할 세 가지 핵심 솔루션

① AI Agent OS: LLM이 아닌 Agentic AI를 위한 인프라 운영체제(S/W)

AI Agent OS는 H/W와 S/W 애플리케이션 사이에서 복잡한 오케스트레이션(진행 순서, 단계별 역할 분담, 검토, 승인 등 작업 전반을 조율)을 담당한다.

분산 추론 엔진을 통해 1) 멀티노드 확장, 2) 프리필(Prefill, 전체 프롬프트를 인식하고 입력 토큰에 대한 키/값을 생성 후 KV캐시에 저장)과 Decode(순차적으로 토큰을 생성)를 분리, 3) 도구 호출, 메모리 관리, Agent 간 경계 지정 등 부가기능을 제공한다.

Agentic AI 구성에 특화된 강점으로 눈에 띄는 주요 기능은 'Dynamic Auto-tuned'다. 해당 기능은 1) 트래픽과 인프라 구동 상태를 실시간 모니터링, 2) 구성 단계의 병목 또는 작업 장애에 자동 대응을 지원한다.

모델 활용 빈도, Agent 활용 사례가 바뀐 상황에서도 매번 사람이 시스템을 재튜닝할 필요가 없도록 고도화하는 것이 해당 솔루션의 최종 목표다.

망고부스트 주력 사업: 'Agentic AI Infra 솔루션'



자료: 회사 자료, 신한투자증권

② 이기종 컴퓨팅(Heterogeneous Compute): 단계별 적합한 칩 활용

기존 인프라는 대부분의 단계를 동일 GPU에서 실행한다. 가벼운 워크로드까지 고가의 칩에서 구동하는 것은 자원 활용도 면에서 비효율적일 수밖에 없다. Agent의 비용 축소를 위해서는 단계별로 세분화된 칩 활용이 필요하다.

무거운 디코드는 플래그십 GPU에 할당하고, 프리필과 가벼운 디코드에 가성비 GPU와 NPU를 활용한다. 도구 호출과 전반적인 작업 조정은 CPU에 일임하는 구조를 반영하면 기존 시스템 대비 효율화가 가능하다.

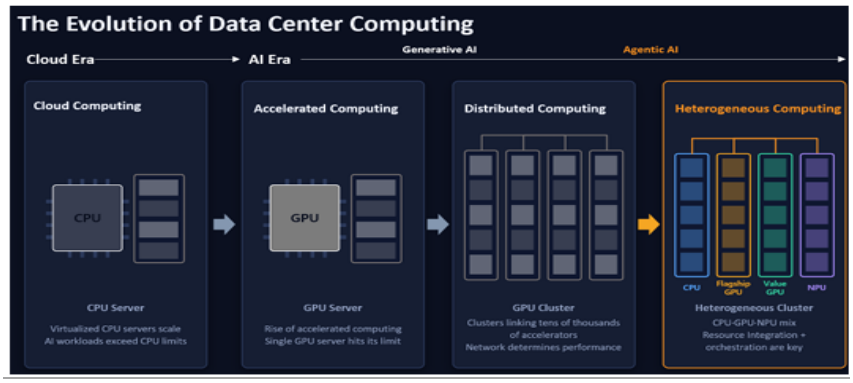
서로 다른 가속기를 하나의 패브릭으로 연결하는 RDMA 프로토콜과 집단통신(CCL) 최적화를 통해 연산 칩 조합을 자유롭게 배분할 수 있다. 표준 이더넷 기반의 연결 기술이 적용돼 데이터 이동간 발생하는 대기 시간도 일부 개선할 수 있을 것으로 보인다.

생성형 AI 모델이 비대해지면서 수백 개의 가속기를 연결하는 분산 컴퓨팅이 해결책으로 제시된 이후 대부분의 인프라가 같은 형식으로 구축되고 있다. Agentic AI 시대에서는 이기종 컴퓨팅을 통한 연산 최적화, 자원 분배 재구성이 토큰당 비용 효율화에 적합한 솔루션으로 부상할 가능성이 높아지고 있어 점진적인 시장 확장 기조가 확인될 것으로 예상된다.

이기종 컴퓨팅의 클러스터링 결과물도 주목할 만하다. 회사는 MLPerf(머신러닝 벤치마크) 추론 영역에서 서로 다른 세대의 AMD 가속기를 하나의 시스템으로 연결한 결과 비용 효율이 3배 개선됐다고 언급했다.

미국에 위치한 D사의 플래그십 GPU 플랫폼과 한국에 설치된 가성비 GPU 2종을 오케스트레이션한 결과 이기종 연산 클러스터링이 장거리 시스템에서도 충분히 구동 가능하다는 점을 입증했다. 총 4개 사이트, 2개 대륙에 걸쳐 GPU 32개를 연결한 멀티리전 클러스터 사례도 공유됐다. 확장 효율은 97%를 기록했는데 유휴 컴퓨팅 자원 발생 시 이를 최대로 활용할 수 있다는 점이 핵심이다.

데이터센터 컴퓨팅 기술 변화



자료: 회사 자료, 신한투자증권

③ Agentic Memory Expansion: 확장된 메모리 계층 적용

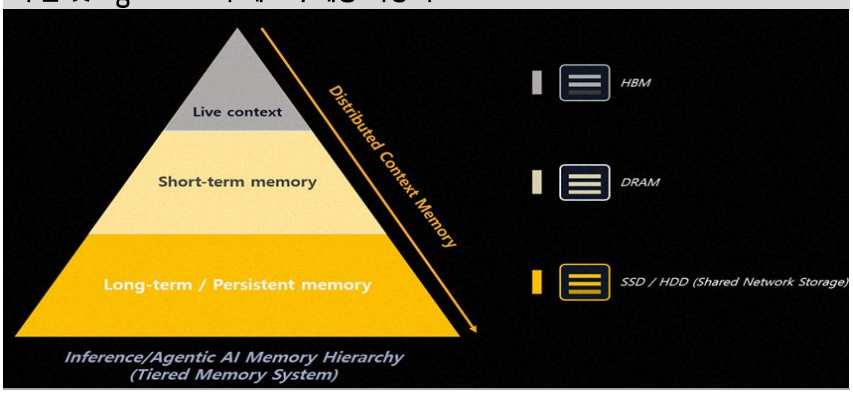
Agent AI는 방대한 양의 컨텍스트와 KV캐시를 유지해야 한다. HBM에 의존할 경우 용량 소진 속도가 너무 빠르다. 최근 전방 산업에서 나타나고 있는 메모리 계층 다각화 흐름이 이에 부합하는 기술 변화로 볼 수 있다.

사용 빈도가 높은 컨텍스트는 HBM과 DRAM에 할당하며, 장기 보관 데이터는 대용량 SSD에 배치한다. 망고 스토리지부스트(Mango StroageBoost) 솔루션도 주목할 만하다. 원격 저장장치도 NVMe-oF(NVMe over Fabrics, 원격 스토리지 자원을 로컬 자원으로 활용하는 기술) 네트워크를 통해 Near 스토리지와 유사한 활용도를 제공한다. 최근 화두가 되고 있는 GPU와 스토리지 간의 직접적인 데이터 전송(DMA)도 가능하다.

스토리지 계층의 기술적 유연성을 제공하는 것과 동시에 메모리 풀 압축 기술을 추가하면 활용도는 더욱 극대화된다. 단일 Agent와 달리 다수의 Agent를 활용하는 사례에서는 더 많은 컨텍스트가 공유될수록 작업 능력이 향상되는 것이 사실이다. 다만 토큰 비용이 기하급수적으로 상승해 이점이 상쇄되는 결과가 나타난다. 동사는 각 Agent 활용 정보 중 중요도가 높은 부분만 따로 전달하는 시스템을 구축해 데이터 인프라 자체가 이를 결정할 수 있도록 효율화를 지원한다.

고가의 SSD 대신 HDD를 활용하는 하이브리드 인프라 솔루션도 관심이 높아지고 있다. 상대적으로 가격이 저렴한 HDD를 DPU로 가속하는 기술로 전체적인 스토리지 구축 비용을 절감이 가능하다. 공급부족이 이어지고 있는 메모리 계층 전반의 다양한 연구개발은 풀스택 솔루션의 이점을 강화하는 요인으로 판단된다.

추론 및 Agentic AI의 메모리 계층 확장 구조



자료: 회사 자료, 신한투자증권

세 가지 솔루션과 앞으로의 기술 방향성

앞서 언급한 핵심 기능 세 가지를 활용하면 1) AI Agent OS가 워크로드를 분석하고, 2) 이기종 컴퓨팅 기능이 적절한 자원에 연산을 분배하며, 3) Agentic Memory Expansion이 계층화된 메모리 배치를 진행한다.

모든 연산 자원을 하나의 Pool로 통합해 패브릭을 구성한 이후에는 Agent 간 경계 지정, 컨텍스트 압축, 트래픽 경로 설정(라우팅) 등을 DPU에 분배한다. CPU의 연산 부하를 GPU가 감당했던 것과 유사한 흐름이다. CPU가 통제하는 네트워킹 및 경계 구성 등을 DPU가 대신 처리하는 구조다.

지속 가능한 AI 인프라는 1) 유휴 용량을 최소화하고, 2) 같은 자원 내에서 더 많은 토큰을 생성하며, 3) 통신 및 보안에 특화된 H/W를 적용해 안정성을 높이는 것이 기본 전제다. 시스템 전반의 효율화, 토큰 경제성 제고, 연결 기술의 융합이 풀스택 인프라 솔루션의 강점으로 작용할 것으로 보인다.

Agent-Native Unified Fabric: 하나의 Pool 안에서 작동하는 통합 시스템



자료: 회사 자료, 신한투자증권

Compliance Notice

- 이 자료에 게재된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 작성되었음을 확인합니다.(작성자: 김형태, 송혜수)
- 자료제공일 현재 조사분석 담당자는 상기 회사가 발행한 주식 및 주식관련사채에 대하여 규정상 고지하여야 할 재산적 이해관계가 없으며, 추천의견을 제시함에 있어 어떠한 금전적 보상과도 연계되어 있지 않습니다.
- 당 자료는 상기 회사 및 상기 회사의 유가증권에 대한 조사분석담당자의 의견을 정확히 반영하고 있으나 이는 자료제공일 현재 시점에서의 의견 및 추정치로서 실적치와 오차가 발생할 수 있으며, 투자를 유도할 목적이 아니라 투자자의 투자판단에 참고가 되는 정보제공을 목적으로 하고 있습니다. 따라서 종목의 선택이나 투자의 최종결정은 투자자 자신의 판단으로 하시기 바랍니다.
- 본 조사분석자료는 당사 고객에 한하여 배포되는 자료로 어떠한 경우에도 당사의 허락 없이 복사, 대여, 재배포 될 수 없습니다.

투자등급 (2017년 4월 1일부터 적용)

종목	◆ 매수 : 향후 6개월 수익률이 +10% 이상 ◆ Trading BUY : 향후 6개월 수익률이 -10% ~ +10% ◆ 중립 : 향후 6개월 수익률이 -10% ~ -20% ◆ 축소 : 향후 6개월 수익률이 -20% 이하	섹터	◆ 비중확대 : 업종내 커버리지 업체들의 투자의견이 시가총액 기준으로 매수 비중이 높을 경우 ◆ 중립 : 업종내 커버리지 업체들의 투자의견이 시가총액 기준으로 중립적일 경우 ◆ 축소 : 업종내 커버리지 업체들의 투자의견이 시가총액 기준으로 Reduce가 우세한 경우

신한투자증권 유니버스 투자등급 비율 (2026년 09월 21일 기준)

매수 (매수)	94.53%	Trading BUY (중립)	2.19%	중립 (중립)	3.28%	축소 (매도)	0.00%