2026.03.27

Memory Watch

박재환_jaehwan124@
이승우_swlee6591@

AI Agent 시대에 새롭게 주목받는 CPU

- **AI Agent 워크로드:** 초기 생성형 AI는 단순 응답형 구조이기에, 연산의 핵심 병목은 GPU 연산 구간에 집중되었음. 그러나, OpenClaw와 같은 AI Agent의 워크로드는 DB 접근, 외부 툴 호출 및 실행, 재추론 등의 다층 레이어로 이루어져 있어 작업의 스케줄링과 오케스트레이션이 중요해지고 있음. CPU는 직접 AI 연산을 진행하지는 않지만, Agent AI 워크로드를 관리하고, DB, 툴, 스케줄링, 오케스트레이션 등 대부분의 Non-GPU 작업을 담당. 따라서 AI Agent 시장이 본격적으로 개화함에 따라, GPU와 CPU의 적절한 조화가 클러스터의 효율을 좌우하는 방향으로 진화.

- **엔비디아 Vera Rubin Pod:** 엔비디아는 가장 효율적인 데이터센터 클러스터 랙시스템으로 Vera Rubin Pod을 제시하고 있음. VR Pod은 NVL72 랙 16대, Vera CPU Rack(CPU 256개로 구성) 2대가 탑재된 구조. 이를 프로세서 단에서 보면 Pod은 GPU 총 1,152개(72x16), CPU 총 1,088개(36x16 + 256x2)로 구성됨. 따라서, NVL72 랙 기준 33%였던 CPU의 비중은 VR Pod 기준으로는 49%로 확대됨. Vera CPU Rack을 별도로 출시한 것은 AI Agent 시대의 CPU 중요도 증가를 암시하는 변화로 해석됨.

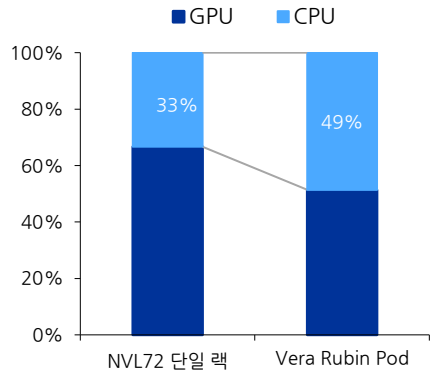
- **ARM AGI CPU:** IP/라이센스 업체인 ARM도 AGI CPU를 개발하고 판매한다는 계획을 발표. AGI CPU 또한 랙시스템으로 제공되며, 랙 당 최대 336개의 CPU와 최대 1PB 용량의 DRAM을 탑재함. ARM은 AI Agent 시장의 개화로 DC 향 CPU TAM이 2030년까지 1,000억달러에 달할 것으로 전망. 또한 향후 5년 이내에 AGI CPU 매출은 연간 150억 달러를 달성할 것이라고 언급. (ARM 2025년 연간 매출 46.7억달러) 엔비디아의 Vera CPU와 하이퍼스케일러들의 자체 CPU가 ARM 기반인 점 역시 기존 사업부인 IP/라이센스 매출에 긍정적인 드라이브가 될 전망.

- **So What:** AI 사이클 초기부터 연산 성능 확보를 위한 GPU, 추론 시장의 개화 이후 급증하는 KV Cache로 인한 메모리, 그리고 고성능 네트워크 확장을 위한 광 네트워킹이 순차적인 병목 요인으로 부상한 데 이어 AI Agent 시장으로의 급격한 전환에 따라, CPU까지 새로운 병목 요인으로 부각될 전망. GPU, 메모리, 실리콘 포토닉스에 이어 데이터센터 CPU 밸류체인에 대한 주목이 필요.

ARM AGI CPU

자료: ARM, 유진투자증권

NVL72와 VR Pod의 GPU/CPU 비중



자료: 유진투자증권

CPU 랙 스펙 비교

구분	Vera CPU	AGI CPU
CPU 탑재량	256 개	336 개
메모리	LPDDR5X	DDR5
메모리 모듈	SOCAMM	DIMM
메모리 용량	384TB	1PB

자료: 유진투자증권

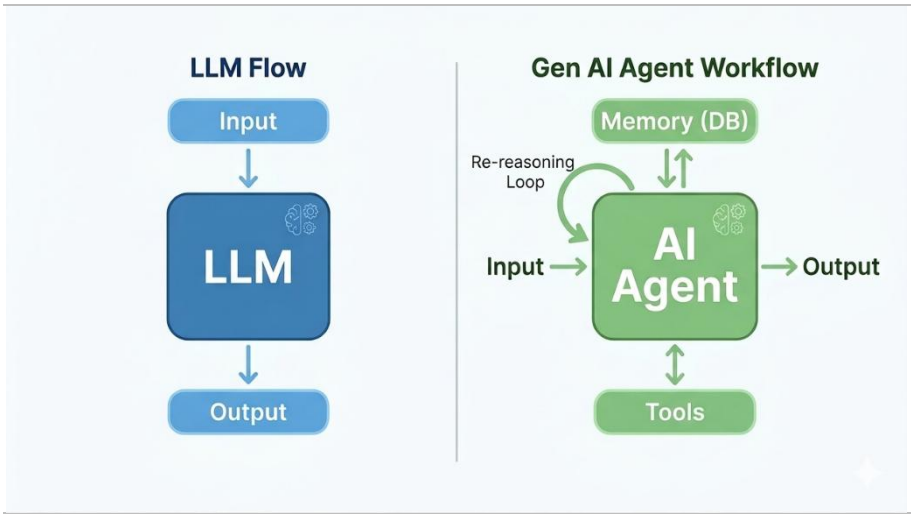
유진투자증권

AI Agent 워크로드

초기 생성형 AI 의 워크로드는 사용자의 입력에 대해 답변을 생성하는 단순한 응답형 구조. 이 경우 연산의 핵심 병목은 대부분 모델 내부의 행렬 연산, 메모리 대역폭, KV Cache 처리 효율 등 GPU 중심의 연산 구간에 집중되었음. 이에 반해 CPU 는 AI 연산을 직접적으로 처리하기에는 한계가 있었기에, 네트워크 관리, 스케줄링, 스토리지 관리 등 시스템을 제어하고 GPU 의 연산을 보조하는 역할을 수행.

다만 AI Agent 시장이 개화하며 추론의 워크로드 양상은 새로운 국면에 진입하고 있음. Open Claw 와 같이 직접 행동하는 AI Agent 가 출시되며, AI Agent 의 추론은 단순한 답변 생성이 아니라 사용자의 요청을 해석하고, 외부 툴과 데이터를 호출 및 실행하고, 그 결과를 다시 반영해 추가적인 의사결정을 반복하는 다층 실행 구조로 변화 중. 또한 워크로드의 레이어가 늘어나며 작업의 스케줄링, 오케스트레이션 과정 역시 고도화되어 Non-GPU 연산 비중이 급증하고 있음. CPU 는 Agent 추론 과정에서 요청 해석, 작업 스케줄링 등의 오케스트레이션, 데이터베이스 접근, 툴 호출 및 실행, 세션 및 메모리 관리 등 사실상 AI 워크로드 전반의 제어와 외부 툴 실행을 담당. 과거에는 GPU 가 AI 클러스터 성능의 핵심이었다면, **AI Agent 환경에서는 GPU 와 CPU 의 적절한 조화가 클러스터의 효율을 좌우하는 방향으로 변화**하고 있음.

도표 1. LLM 과 AI Agent 의 차이



자료: 유진투자증권

엔비디아의 Vera CPU Rack

올해 2 월 메타는 엔비디아의 Grace CPU 와 Vera CPU 를 대량 구매하는 공급 계약을 체결함. 이에 더해 엔비디아는 GTC 2026 에서 CPU 의 개별 판매를 공식화함. 엔비디아가 인텔, AMD와 달리 CPU를 개별적으로 판매하지 않았다는 점을 감안하면, 이는 매우 이례적인 행보로 판단됨.

아울러 엔비디아는 Vera CPU Rack 을 공개함. Vera CPU Rack 은 CPU 만으로 구성된 랙으로, 노드당 8 개의 Vera CPU, 랙당 총 256 개의 CPU 가 탑재됨. Vera CPU 는 프로세서당 약 1.5TB 의 LPDDR SOCAMM 을 탑재하기에, Vera CPU Rack 의 총 디램 용량은 384TB 에 달할 예정. 엔비디아가 CPU 의 개별 판매를 결정한 배경에는 AI Agent 시대로 접어들며 CPU 가 AI 인프라 성능을 좌우하는 핵심 요인으로 부상할 것이라는 판단이 반영된 것으로 추정됨. **또한 CPU 병목이 지속된다면 GPU 에 대한 투자 역시 둔화될 수 있다는 점에서, CPU 물량을 시장에 추가적으로 공급하려는 의도도 일부 반영된 것으로 추정 가능.**

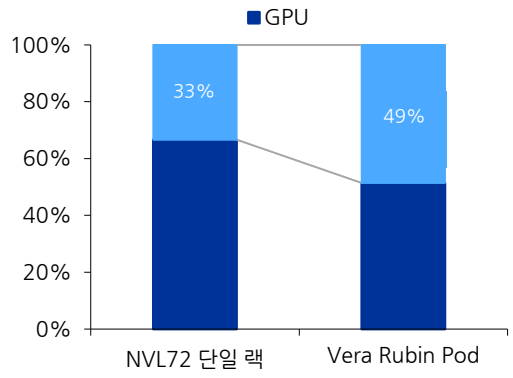
이에 더해 엔비디아는 40 개의 랙으로 구성된 Vera Rubin Pod 을 공개. Vera Rubin Pod 은 AI Agent 시대에 엔비디아가 제시하는 AI 데이터센터 클러스터의 랙 최적 비율로 해석할 수 있음. 엔비디아가 제시한 Pod 에서는 컴퓨팅을 담당하는 NVL72 랙이 16 대, Vera CPU Rack 이 2 대 탑재됨. 이를 프로세서 단에서 보면 Pod 당 GPU 는 총 1,152 개(NVL72 Rack 72 개 x 16), CPU 는 총 1,088 개(NVL72 Rack 36 개 x 16 + Vera CPU Rack 256 개 x 2)가 탑재됨. 이는 기존 단일 NVL72 랙에서 GPU 와 CPU 가 2:1 비율로 고정되어 있던 것과 달리, 클러스터 단위에서는 약 1:1 비율로 **CPU 탑재량이 두 배 수준으로 확대될 수 있다는 점을 암시.**

도표 2. Vera Rubin Pod 구성

랙 종류	개수(대)	용도
VR NVL72	16	AI 연산
Vera CPU Rack	2	AI 워크로드 관리
Groq 3 LPX	10	AI 연산 (Decode)
Spectrum-6 SPX	10	스케일아웃 네트워킹
Bluefield-4 STX	2	KV Cache 용 SSD 스토리지

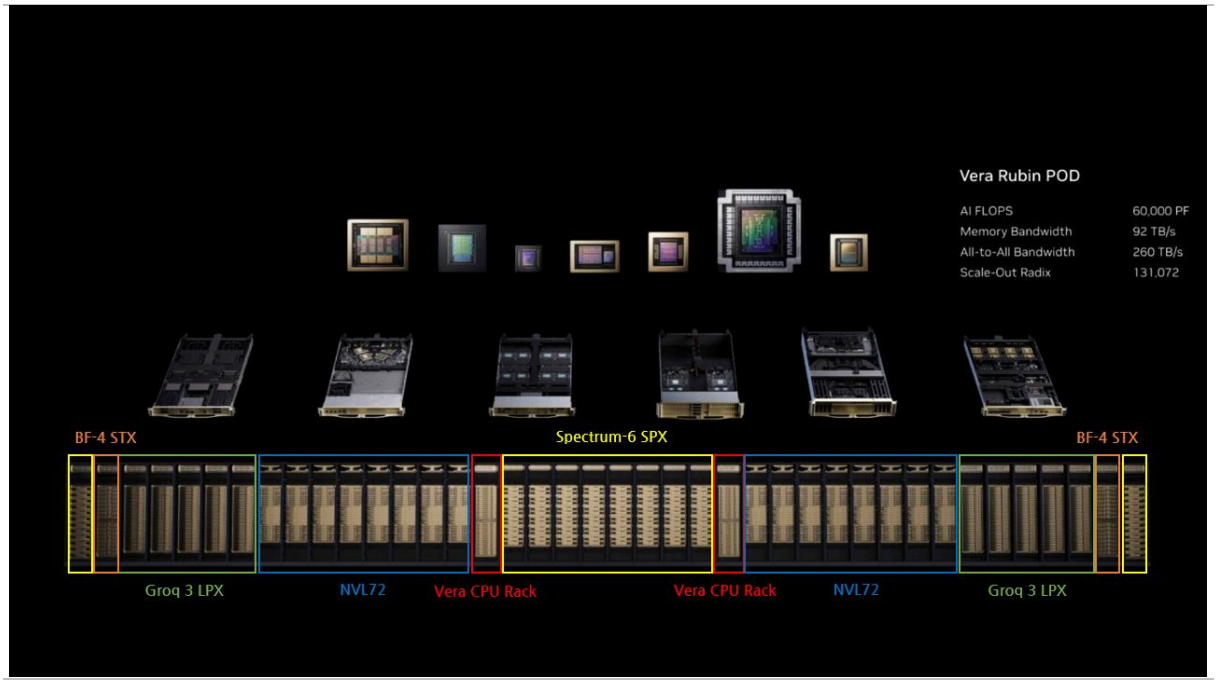
자료: 유진투자증권

도표 3. 프로세서 기준 GPU&CPU 비중



자료: 유진투자증권

도표 4. Vera Rubin Pod



자료: Nvidia, 유진투자증권

도표 5. Blackwell & Rubin 랙시스템 스펙

구분	Blackwell		Rubin		
랙 시스템 명	GB200 NVL72	GB300 NVL72	VR200 NVL72	Vera CPU Rack	Groq 3 LPX
가속기 종류	B200 GPU	B300 GPU	V200 GPU	-	LP30 LPU
가속기 메모리 종류	HBM3E 8hi	HBM3E 12hi	HBM4	-	SRAM
CPU 종류	Grace CPU	Grace CPU	Vera CPU	Vera CPU	x86 CPU
CPU 메모리 종류	LPDDR5X	LPDDR5X	LPDDR5X	LPDDR5X	DDR5
랙 당 가속기 탑재량	72	72	72	-	256
랙 당 CPU 탑재량	36	36	36	256	32
랙 당 가속기 메모리 용량	13.8TB	20.7TB	20.7TB	-	128GB
랙 당 CPU 메모리 용량	17.3TB	17.3TB	54TB	384TB	32.8TB

자료: 유진투자증권

ARM AGI CPU

ARM 역시 ARM Everywhere 2026 행사에서 데이터센터 AGI CPU 를 공개함. 기존 ARM 은 반도체 IP 에 대한 라이선스와 로열티 수익이 주 매출원으로, 1990 년 창립 이래 최초의 칩 판매라는 점에서 매우 이례적. AGI CPU 는 메타와 공동 개발되었으며, TSMC 의 3nm 공정으로 제조됨. AGI CPU 는 엔비디아의 Vera CPU Rack 과 유사하게 랙 시스템으로 출시될 예정. 랙시스템은 공랭식과 수랭식 두 가지 형태로 제공되며, 공랭식은 랙당 CPU 60 개, 수랭식은 랙당 CPU 336 개가 탑재됨. **AGI CPU 는 칩당 12 개의 DDR5 DIMM 이 탑재될 예정이며, 공랭식 기준 180TB, 수랭식 기준 최대 1PB 의 DRAM 이 탑재될 예정.** 주류 상용 AI 랙 중에서는 최초로 페타바이트 단위의 DRAM 이 탑재된다는 점이 특징적. AGI CPU Rack 의 첫 대규모 고객사는 메타가 될 전망.

ARM 은 AI Agent 시장의 개화로 데이터센터 CPU TAM 이 2030 년까지 1,000 억달러에 달할 것이며, 5 년 내에 AGI CPU 는 연간 150 억달러의 매출을 창출할 것이라고 언급함. **2025 년 ARM 의 연간 매출이 46.7 억달러인 점을 고려하면 칩 매출 150 억달러 증분은 매우 긍정적인 가이던스임을 알 수 있음.**

엔비디아와 ARM 의 CPU 판매는, 데이터센터의 병목이 가속기 외부로 이동하고 있으며, 더 이상 데이터센터의 성능은 단순히 GPU 의 개수만으로 결정되지 않는다는 점을 시사. AI 사이클의 초기에는 연산 성능 확보를 위한 GPU, 추론 시장의 개화 이후 급증하는 KV Cache 로 인한 메모리, 그리고 고성능 네트워크 확장을 위한 광 네트워킹이 순차적인 병목 요인으로 부상하였음. 더 나아가 현 시점 AI 산업은 AI Agent 시장으로의 급격한 전환기에 위치함. 이에 따라 Non-GPU 워크로드 비중이 빠르게 확대되고 있고, **CPU 가 새로운 병목 요인으로 부각되는 국면이라고 판단.**

도표 6. AGI CPU 랙 스펙

구분	공랭식	수랭식
CPU 종류	AGI CPU	AGI CPU
CPU 탑재량	60	336
메모리 종류	DDR5 DIMM	DDR5 DIMM
메모리 탑재량	180TB	1PB

자료: ARM, 유진투자증권

도표 7. AGI CPU 랙

자료: ARM 유진투자증권

Compliance Notice

당사는 자료 작성일 기준으로 지난 3 개월 간 해당종목에 대해서 유가증권 발행에 참여한 적이 없습니다
당사는 본 자료 발간일을 기준으로 해당종목의 주식을 1% 이상 보유하고 있지 않습니다
당사는 동 자료를 기관투자가 또는 제 3 자에게 사전 제공한 사실이 없습니다
조사분석담당자는 자료작성일 현재 동 종목과 관련하여 재산적 이해관계가 없습니다
동 자료에 게재된 내용들은 조사분석담당자 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 작성되었음을 확인합니다
동 자료는 당사의 제작물로서 모든 저작권은 당사에게 있습니다
동 자료는 당사의 동의 없이 어떠한 경우에도 어떠한 형태로든 복제, 배포, 전송, 변형, 대여할 수 없습니다
동 자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료 및 정보로부터 얻어진 것이나, 당사는 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 어떠한 경우에도 자료는 고객의 주식투자의 결과에 대한 법적 책임소재에 대한 증빙자료로 사용될 수 없습니다

투자기간 및 투자등급/투자의견 비율

종목추천 및 업종추천 투자기간: 12 개월 (추천기준일 종가대비 추천종목의 예상 목표수익률을 의미함)		당사 투자의견 비율(%)
· STRONG BUY(매수)	추천기준일 종가대비 +50%이상	0%
· BUY(매수)	추천기준일 종가대비 +15%이상 ~ +50%미만	98%
· HOLD(중립)	추천기준일 종가대비 -10%이상 ~ +15%미만	2%
· REDUCE(매도)	추천기준일 종가대비 -10%미만	0%

(2025.12.31 기준)