2026.03.18

Memory Watch

이승우_swlee6591@
박재환_jaehwan124@

엔비디아 GTC 2026 Keynote

- **Groq LPU:** 3세대 Groq LPU인 LP30이 공개됨. LPU는 Decode의 FFN 단계를 전담하는 프로세서로 Latency가 극히 낮은 SRAM 500MB로 구성됨. LPX 랙 시스템은 256개의 LP30으로 구성되어 있어 총 128GB의 SRAM을 탑재하게 됨. (참고로 Vera Rubin NVL72의 20.7TB HBM이 탑재됨). FFN은 행렬 연산이 주를 이루는 연산 집약적 단계이기에, 메모리의 절대적인 용량보다는 레이턴시가 중요. LPX는 추론 과정을 분담하여, 고객은 단일 VR NVL72 랙을 가동할 때 보다 최대 2배 수준의 높은 아웃풋을 창출할 수 있을 전망. **LP30은 삼성 파운드리 4nm로 제조될 예정.**

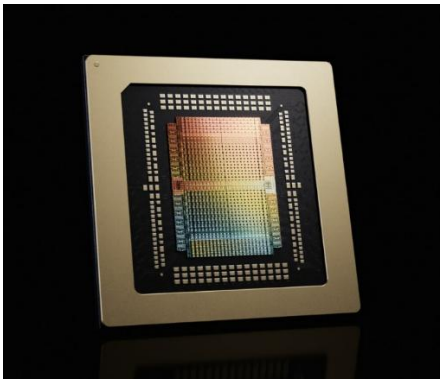
- **Vera 시스템:** 엔비디아는 지난 2월 메타와의 Grace+Vera CPU 개별 공급 계약을 발표한 바 있음. 이번 GTC에서 젠슨황은 Vera CPU의 개별 판매 의지를 다시금 강조함. 특히 Vera CPU 8개가 탑재되어 있는 트레이 형태의 Vera CPU 시스템을 추가로 공개한 점이 특징적.

- **Rubin Ultra 업데이트:** 금번 GTC에서 Rubin Ultra Kyber 랙의 컴퓨트 블레이드 실물이 공개됨. 컴퓨트 블레이드는 Oberon의 트레이 방식과 달리 세로로 삽입되는 방식이며, 각 블레이드 당 4개의 Rubin Ultra GPU, 2개의 Vera CPU가 탑재될 전망. 또한 Kyber 랙 아키텍처부터 NVLink 스위치는 랙 후면부에 탑재, 중앙의 Midplane PCB가 컴퓨트와 네트워킹 블레이드를 연결해주는 방식일 것으로 예상됨.

- **Feynman 업데이트:** Feynman GPU에는 **Die Stacking 기술이 적용**될 가능성이 있으며, 이는 로직 다이를 수직으로 적층하는 기술. Die Stacking은 데이터 전송 구간이 짧아지고 연산 집약도가 상승한다는 장점이 있으나, 발열·수율 관리가 핵심 과제가 될 것으로 판단됨. 이에 더해 Feynman 플랫폼부터 **NVLink Switch에도 CPO가 적용될 가능성**이 부각됨. 이는 올 하반기 스케일아웃에만 제한적으로 적용될 예정인 CPO가 예상보다 빠르게 광범위한 영역으로 확장될 수 있다는 의미. 다만 젠슨황은 Feynman 플랫폼이 구리·광을 모두 사용할 것을 언급해, 기존 구리 케이블과 LPO, CPO가 공존할 것임을 시사.

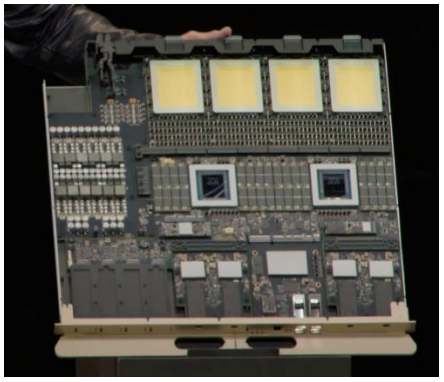
- **2027년 누적 매출 가이던스:** 젠슨황은 2027년까지 Blackwell·Rubin 플랫폼의 누적 매출이 1조달러에 달할 것이라고 언급함. 이전 가이던스인 '2026년까지 누적 데이터센터 매출 5,000억달러 이상'을 충족한다고 가정할 시, 이번 가이던스는 2027년 데이터센터 매출 약 4,500억~5,000억 달러 수준을 시사함. 엔비디아는 GTC를 통해 LPU, CPO의 도입과 같이 아키텍처의 지속적인 고도화를 통해 중장기 AI 인프라 로드맵에 대한 주도권을 지속 강화하고 있음을 재확인.

엔비디아 LP30



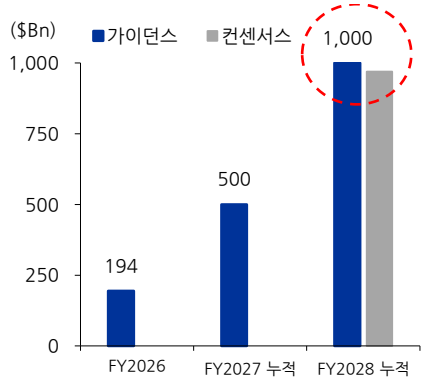
자료: Nvidia, 유진투자증권

엔비디아 Kyber Compute Blade



자료: Nvidia, 유진투자증권

엔비디아 데이터센터 매출 가이던스



자료: 유진투자증권

유진투자증권

I. 추론의 분리 - Groq LPU 3

LP30 LPX 랙의 메모리 용량은 128GB 뿐

젠슨 황은 엔비디아 Rubin 아키텍처에 신규 편입되는 Groq의 3세대 LPU, LP30을 공개함. LP30 은 삼성 파운드리의 4nm 공정으로 제조되며, 칩당 500MB 의 SRAM 용량과 150TB/s의 SRAM 대역폭을 제공하는 것이 특징.

LP30은 LPX라는 별도의 랙 시스템으로 제공될 예정. LPX 트레이 1 개에는 8 개의 LP30과 CPU, BlueField-4 DPU가 각각 1 개씩 탑재되며, LPX 랙 1 개에는 총 32 개의 LPX 트레이, 즉 256 개의 LP30이 집적되는 구조. LPX 랙에는 SRAM이 총 128GB 탑재될 예정.

Decode FFN 연산에 최적화된 LPU

LP30 의 핵심 포인트는 Decode 구간에서 FFN(Feed Forward Network) 연산을 처리하는 데 최적화되어 있다는 점. Decode 단계는 크게 AATN(Attention)과 FFN 두 단계로 구분됨. AATN 단계는 현재 토큰이 과거 KV Cache 중 어떤 정보를 참고할지를 결정하는 구간으로 대규모 KV Cache를 빠르게 읽고 써야 하기에 메모리 용량 및 대역폭 부담이 큼. 반면 FFN 단계는 AATN 단계에서 집계된 정보를 바탕으로 비선형 연산을 수행하는 단계로, 연산 집약적 단계의 특성상 상대적으로 메모리 용량 부담은 낮지만 연산 지연(latency) 최소화가 중요.

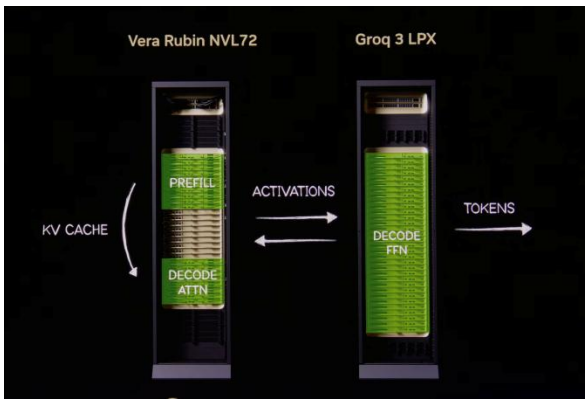
이러한 특성 차이를 고려해 엔비디아는 Decode 단계 내에서 한번 더 역할을 분리한 것으로 해석됨. 즉, Rubin NVL72가 추론의 Prefill 단계와 Decode AATN 단계의 연산을 담당하고, 별도의 LPX 랙에서는 LPU가 Decode FFN 단계의 연산을 담당하는 구조. GPU는 SRAM 대비 높은 메모리 용량을 바탕으로 KV Cache를 다루는 데 유리하고, LPU는 SRAM 기반의 초저지연(Low Latency) 구조를 통해 FFN과 같은 연산 집약적 워크로드를 빠르게 처리하는 데 적합하기 때문.

도표 1. NVL72 와 Groq LPX 스팩표

구분	NVL72	Groq 3 LPX
역할	Prefill+Decode ATTN	Decode FFN
노드 당 탑재량	4	8
랙 당 탑재량	72	256
프로세서 메모리	HBM	SRAM
메모리 용량	20.7TB	128GB
메모리 대역폭	260TB/s	640TB/s

자료: 유진투자증권

도표 2. GPU 와 LPU 의 추론 분담



자료: Nvidia, 유진투자증권

도표 3. Decode 의 2 단계

구분	AATN(Attention Layer)	FFN(Feed Forward Layer)
단계	토큰이 과거 KV Cache 중 무엇을 참고할 지 연산	Attention 결과를 받아 토큰의 비선형 연산
연산	KV Cache 조회/반영	MLP/행렬곱
핵심 병목	메모리, KV Cache	컴퓨트
메모리 부담	매우 큼	낮음
프로세서	GPU	LPU

자료: 유진투자증권

엔비디아는 추론을 3 단계로 분리하는 아키텍처를 채택

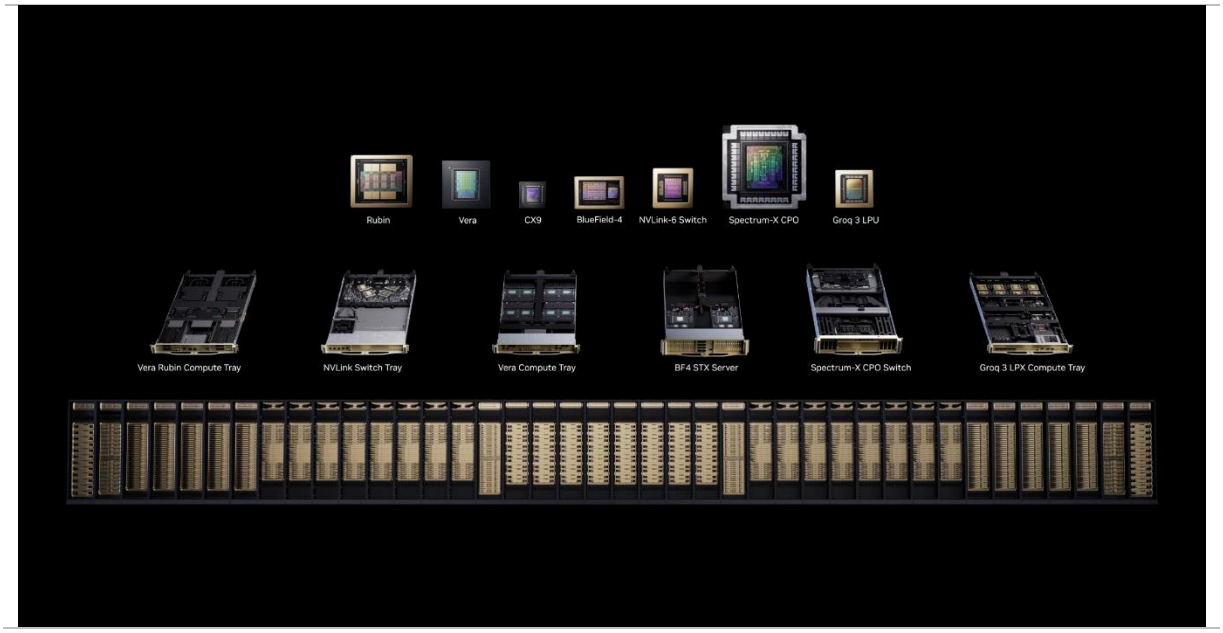
결과적으로 해당 구조는 추론 과정을 Prefill, Decode AATN, Decode FFN 3 단계로 세분화한 뒤, 각 단계의 병목 특성에 맞는 하드웨어를 배치한 사례로 볼 수 있음. 이는 단순히 새로운 가속기를 추가한 것이 아니라, 추론 시스템 아키텍처를 메모리 병목 구간과 연산 병목 구간으로 분리해 전체적인 추론 성능을 최적화하려는 시도로 해석됨. Rubin NVL72 와 LPX 랙을 통합하면, 프론티어급 AI 에이전트 모델에서 Blackwell 대비 최대 35 배 높은 토큰 처리량과 10 배 많은 매출 창출 기회를 제시함. 이는 매출 기준 단일 Rubin NVL72 플랫폼 대비 약 2 배 수준에 해당. 엔비디아는 LPX 랙 시스템을 Rubin 플랫폼과 함께 2026 년 하반기 출시할 계획.

도표 4. 엔비디아 칩 로드맵

구분	Blackwell	Blackwell Ultra	Rubin	Groq LPU	Rubin Ultra
FP4 Dense	10 PFLOPS	15 PFLOPS	50 PFLOPS	1.2 PFLOPS (FP8)	-
프로세스 노드	4nm	4nm	3nm	4nm	3nm
트랜지스터 수	2,080 억 개	2,080 억 개	3,360 억 개	980 억 개	-
메모리	HBM3E 8hi	HBM3E 12hi	HBM4	SRAM	HBM4E
메모리 패키지 수	8	8	8	-	16
메모리 용량	192GB	288GB	288GB	500MB	1,024GB
메모리 대역폭	8TB/s	8TB/s	22TB/s	150TB/s (SRAM)	-

자료: 유진투자증권

도표 5. 엔비디아 Vera Rubin 플랫폼



자료: Nvidia, 유진투자증권

Ⅱ. 파인만(Feynman) 플랫폼 구체화

엔비디아의 로드맵 슬라이드에서 2028 년 출시될 Feynman 플랫폼의 구성을 일부 파악할 수 있었음. Feynman 플랫폼의 7 개의 신규 칩으로 구성될 전망.

Feynman GPU, 3D Die Stacking 적용 전망

Rubin GPU 까지 적용되었던 2.5D 패키징 방식인 CoWoS 와 달리, Feynman GPU 에는 3D Die Stacking 이 적용될 가능성이 있으며, 이는 로직 다이를 HBM 과 유사하게 위아래로 직접 적층해 연결하는 패키징 방식. 3D 다이 적층은 동일 면적 내에서 더 높은 수준의 연산 자원 집적이 가능하고, 더 짧은 die-to-die 연결을 통해 더 작은 폼팩터에서 더 높은 대역폭과 더 낮은 전력 소모를 구현할 수 있다는 장점이 있음. 다만 DRAM 다이보다 발열이 큰 로직 다이의 특성상 수직으로 적층될 시 패키지 내부 발열이 극심해질 수 있으며, 적층 구조상 다이 하나의 문제가 패키지 전체의 문제로 이어질 수 있어 수율 저하 측면의 단점 또한 존재. 따라서 열 관리와 파운드리·패키징 측면에서의 혁신이 핵심 과제가 될 전망. 현재 파운드리 업체들은 3D 패키징 기술을 활발히 개발 중이며, 대표적으로 TSMC 의 SoIC, Intel 의 Foveros, 삼성전자의 X-Cube 등이 있음.

CPO 의 기회는 스케일업까지 확대

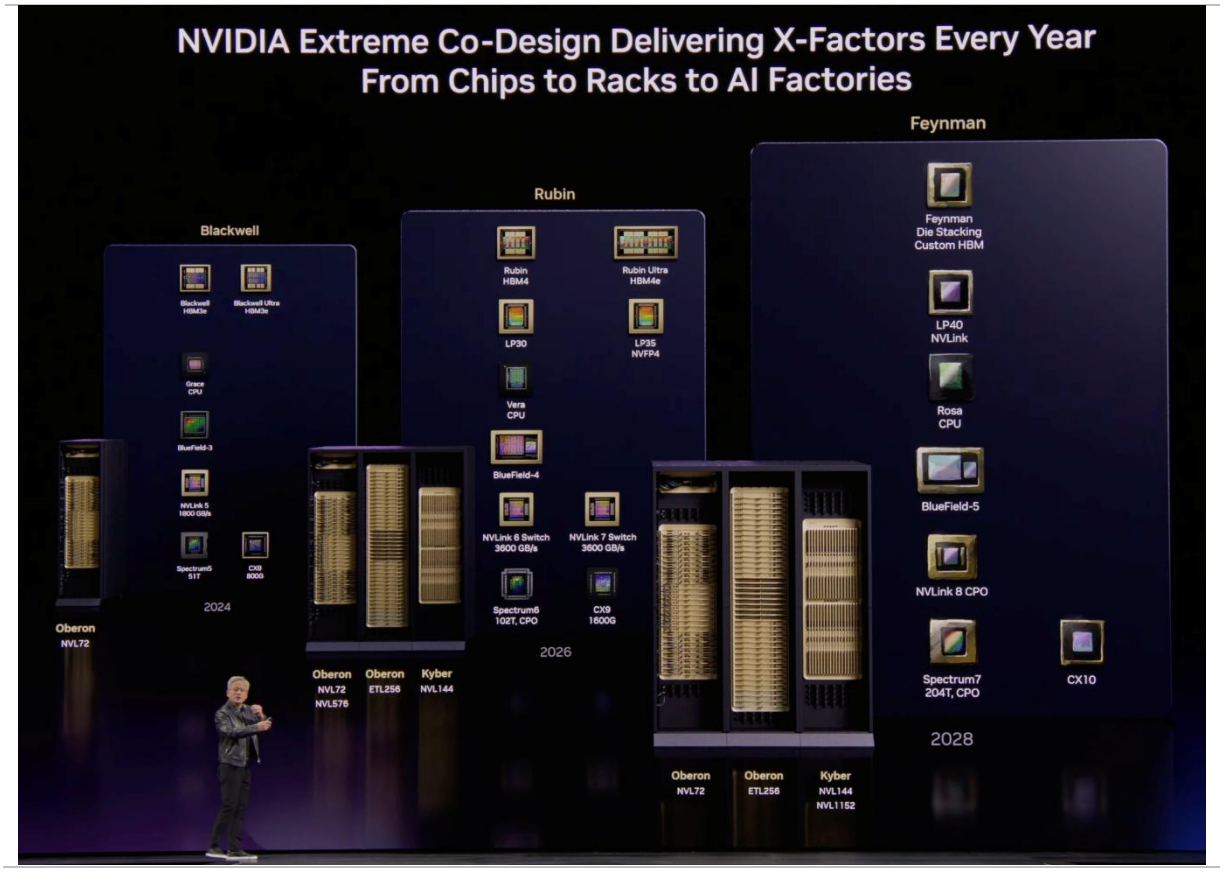
또한 특징적인 부분은 Feynman 플랫폼부터 NVLink Switch 에 CPO 가 적용될 수 있다는 점. NVLink Switch 는 Spectrum-X 와 달리 NVL 랙 내부 GPU 들을 연결하는 스케일업 네트워킹 스위치로, 2028 년부터 CPO 가 본격적으로 스케일업까지 TAM 을 확장할 수 있음을 시사. 다만 젠슨 황은 추가적으로 구리 인터커넥트 역시 여전히 유효하며, Feynman 플랫폼에서도 구리와 광 연결이 모두 사용될 것이라고 언급함. 이는 향후 CPO 가 기존 구리 기반 연결을 전면 대체하는 방식보다는, 대역폭·전력효율·배선 밀도 측면에서 광 연결의 필요성이 높아지는 구간을 중심으로 CPO 가 점진적으로 확대해 나갈 가능성이 높다고 해석됨.

도표 6. Feynman 플랫폼 신규 칩 리스트

구분	용도	비고
Feynman GPU	AI 연산	Die Stacking, Custom HBM 적용
Rosa CPU	워크로드 관리	-
LP40 NVLink	추론 특화 연산	NVFP4, NVLink 지원
Bluefield-5	스토리지 가속	-
NVLink 8 CPO	스케일업 네트워킹	스케일업에 구리, 광(CPO) 동시 적용
Spectrum7 CPO	스케일아웃 네트워킹	이전 세대 대비 2 배 빠른 204Tb/s 네트워킹 지원
CX10	스케일아웃 네트워킹	-

자료: 유진투자증권

도표 7. 엔비디아 플랫폼 로드맵



자료: Nvidia, 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 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)