IBK기업은행 금융그룹 IBK투자증권

Industry Comment

Analyst 강민구
02) 6915-5473
kmg@ibks.com

비중확대 (유지)

새로운 병목으로 부상한 광(光)네트워크

- Computex 기조연설 이후 Marvell의 주가가 급등
- 데이터센터는 AI Factory로 진화 중이며, 연결성이 새로운 병목으로 부상
- 메인 병목으로 여겨지던 칩 공급은 여전히 중요한 문제
- Scale Across, Out, Up, 패키징 등 계층별 네트워크 밸류체인에 대한 관심 필요

AI 트렌드를 대변하는 Marvell의 주가

최근 Marvell의 주가 상승은 AI 시장의 병목이 칩 생산에서 연결로 이동하고 있음을 보여준다. 생성형 AI 시장에서는 GPU, HBM, CoWoS 등 첨단 칩의 성능과 생산 능력이 병목으로 작용해왔다. 최근 광네트워크 기업의 급등은 AI Agent 시대의 병목이 연산 자원 확보를 넘어, 효율적 연결로 이동했음을 나타낸다.

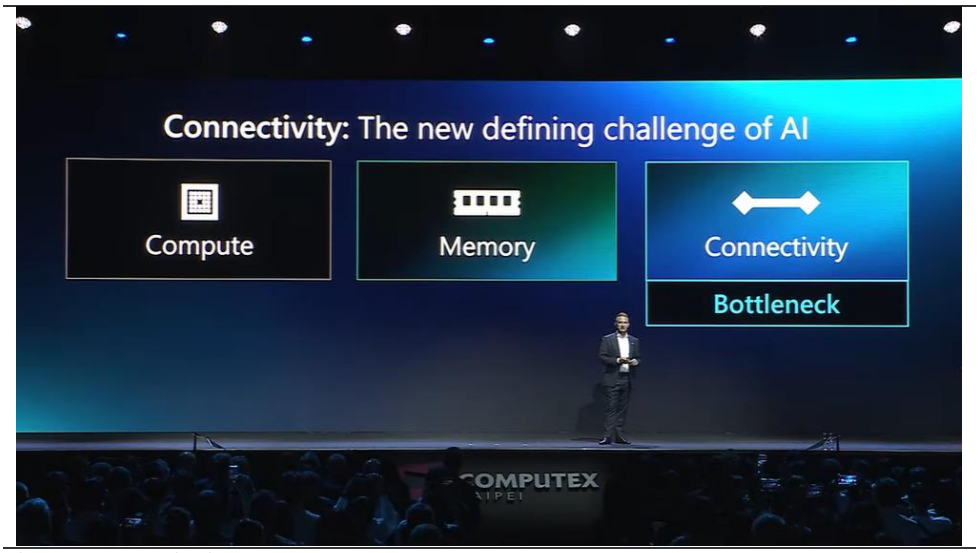
Marvell도 COMPUTEX 2026에서 〈연결성〉이 AI Factory의 새로운 병목임을 강조했다. 데이터센터가 공장화되기 위해서는 단일 프로세서의 성능보다 수백만 개 칩의 유기적 연결이 효율적이기 때문이다. 이를 지지하듯 젠슨황은 Marvell이 차세대 1조 달러의 기업이 될 것이라고 언급했고, 연설 이후 주가는 32.5% 급등한 $290.7로 마감했다.

Marvell은 AI 데이터센터용 Custom silicon, Ethernet 스위치, 광통신용 DSP (Digital Signal Processor), SerDes, PHY 등의 반도체를 설계한다. Keynote에서도 custom XPU, scale-up 네트워크, 광통신, 고속 인터커넥트 기술이 데이터센터의 미래 과제임을 강조했다. 특히 800Gbps·1.6Tbps 대역폭을 지원하기 위해서는 광 네트워크가 데이터센터 간(Scale Across), 데이터센터 내부(Scale Out) 통신을 넘어 랙 내부(Scale Up)에도 적용이 되어야 함을 언급했다.

Marvell의 최근 실적과 NVIDIA의 투자 및 협력은 연결 시장의 잠재력을 보여준다. FY1Q27 매출액은 전년동기 대비 27.5% 증가한 2.4B 달러를 기록했고, 데이터센터향 비중이 매출의 76%를 차지했다. NVIDIA는 Marvell에 20억 달러를 투자하고, NVLink Fusion 파트너십을 체결해 custom XPU와 네트워킹 기술을 생태계에 연결했다. Marvell은 custom silicon, Ethernet switch, optical DSP, SerDes 등을 공급하고, NVIDIA는 NVLink, Spectrum-X, ConnectX, BlueField 등 네트워크 플랫폼을 제공한다.

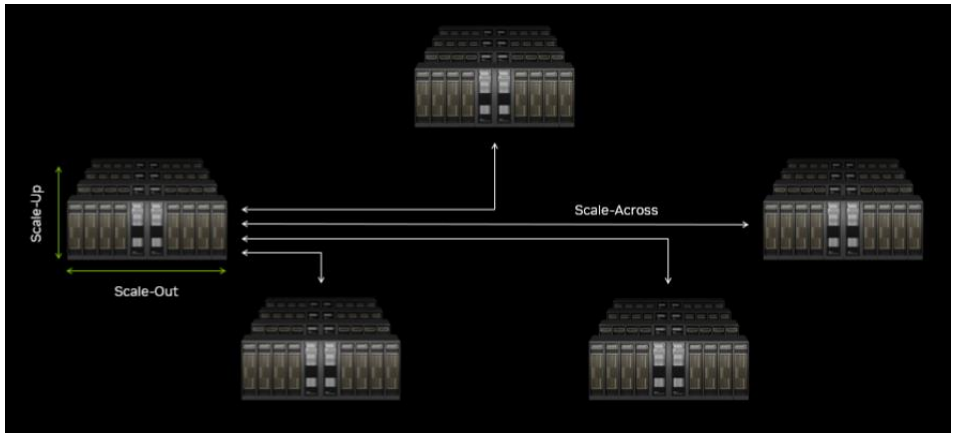
본 조사분석자료는 당사 리서치본부에서 신뢰할 만한 자료 및 정보를 바탕으로 작성한 것이나 당사는 그 정확성이나 완전성을 보장할 수 없으며, 과거의 자료를 기초로 한 투자참고 자료로서 향후 주가 움직임은 과거의 패턴과 다를 수 있습니다. 고객께서는 자신의 판단과 책임 하에 종목 선택이나 투자시기에 대해 최종 결정하시기 바라며, 본 자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다.

그림 1. Marvell은 Keynote에서 연결성을 새로운 병목으로 지목



자료: Marvell, IBK투자증권

그림 2. AI Factory 계층 별 네트워크 유형

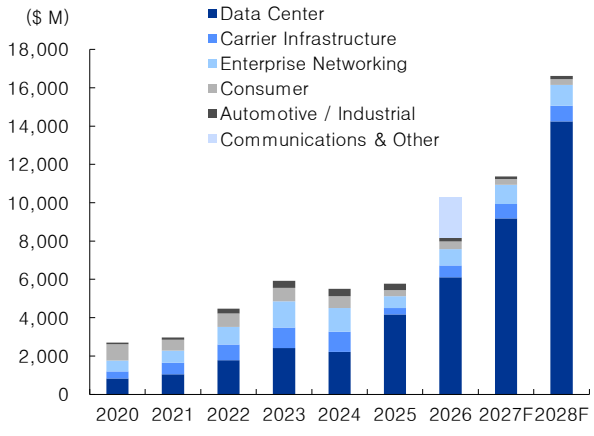


자료: NVIDIA, IBK투자증권

그림 3. Marvell 주가 추이

자료: Bloomberg, IBK투자증권

그림 4. 회계연도 기준 Marvell 매출액 추이 및 전망



자료: Bloomberg, IBK투자증권

Chip Supply는 여전히 중요한 문제

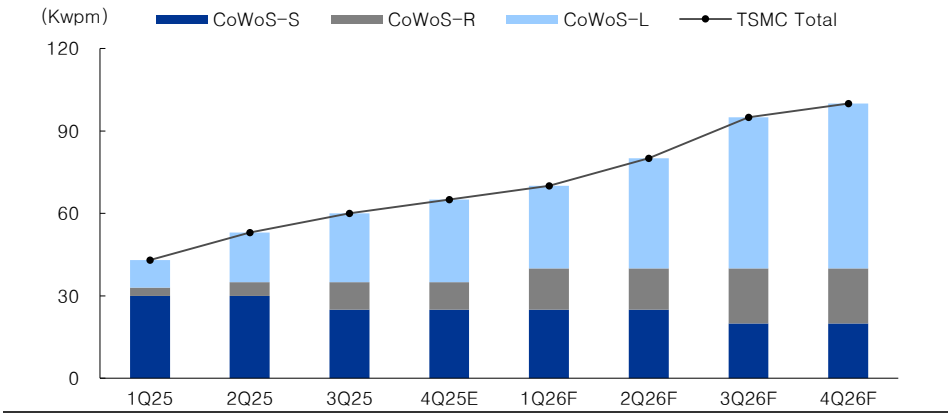
생성형 AI 시대의 가장 큰 병목은 고성능 칩 생산과 첨단 패키징이었다. LLM의 학습 및 추론 수요가 폭증하면서 선단 공정, 첨단 패키징, HBM이 주요 병목으로 지목되었고, 관련 기업들의 실적과 주가도 재평가되었다. NVIDIA는 금년 1분기 Earnings Call에서 기존 병목이 아직 해소되지 않았음을 언급한다. Blackwell GPU가 역사상 가장 빠르게 램프업된 제품이라고 설명했고, 차세대 Vera Rubin은 더 큰 수요가 기대된다고 밝혔다.

첨단 칩의 공급을 제약하는 첫 번째 병목은 TSMC 선단 공정 CAPA이다. 2025년 말부터 양산이 본격화된 N2 공정은 140k/m(연초 50K/m) CAPA를 목표로 증설이 진행중이다. 3nm 공정 역시 추가 증설을 통해 올해 말까지 160k/m 수준까지 CAPA를 확대한다. TSMC는 1분기 실적발표에서 2026년 Capex 가이던스를 52~56B 달러로 제시해 전년 대비 약 30% 증가할 것을 예고했다.

두 번째는 CoWoS로 대표되는 첨단 패키징 공정이다. TSMC는 CoWoS CAPA를 2026년 말까지 100~130k/월 이상으로 확대할 계획이지만, 패키징 대형화가 증설 효과를 상쇄할 것으로 전망된다. Hopper가 약 3배의 레티클을 쓰던 것과 달리 Rubin 이후 AI 칩은 9~12배 면적을 사용해, 동일 라인에서 처리하는 패키지 수량은 크게 감소한다.

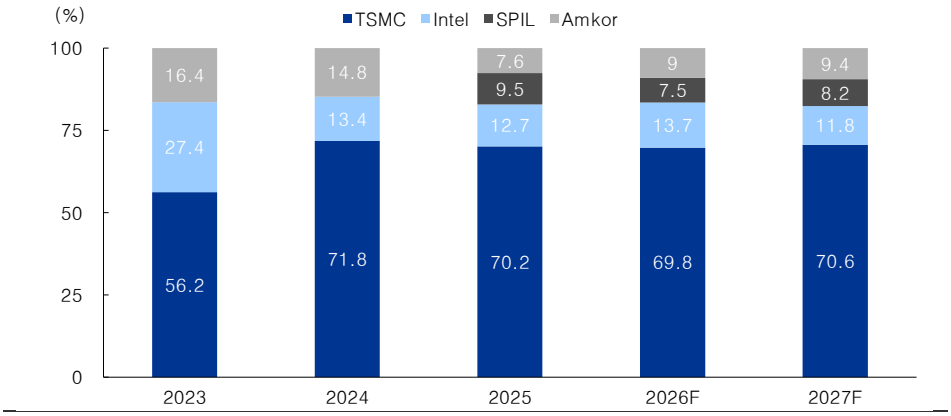
메모리와 HBM 역시 AI 인프라 공급을 제약하는 병목이다. 주요 IDM사는 HBM4가 HBM3E 대비 50% 수준의 가격 프리미엄에도 2026년 물량이 매진되었음을 밝혔다. 2025년 SK하이닉스의 CAPEX는 30.2조 원이었지만 금년 큰 폭의 증액이 예상되며, 마이크론도 직전 가이던스 대비 약 25% 증가한 25B 달러 규모의 CAPEX를 제시했다. COMPUTEX 2026에서 젠슨 황은 SK하이닉스 HBM4E 웨이퍼에 "Please Make More"라는 문구를 적어 칩 공급 부족이 현재 진행형임을 나타냈다.

그림 5. TSMC CoWoS CAPA 추이



자료: TrendForce, IBK투자증권

그림 6. 글로벌 2.5D Packaging Capacity 점유율 전망



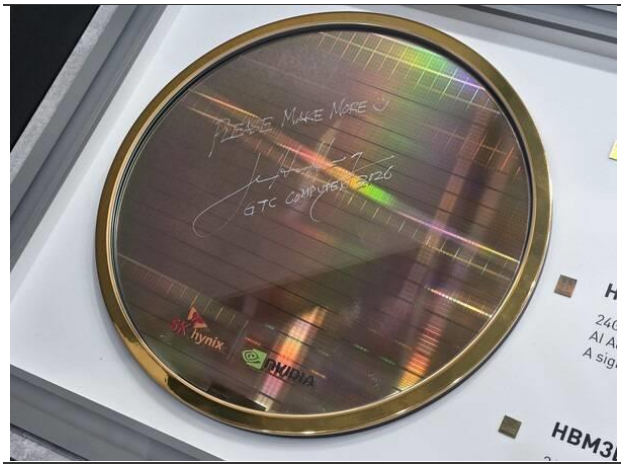
자료: TrendForce, IBK투자증권

표 1. TSMC CoWoS 제품별 특징

구분	TSMC		
	CoWoS-S	CoWoS-R	CoWoS-L
Interposer	Silicon	RDL	RDL, LSI
Silicon Bridge	×	×	LSI embedded in Interposer
Reticle Size	3.3x	9x (2027)	9x (2027) ~ 12x
Price	12,000	11,000	15,000
Current Product	Hopper, TPU, MTIA, Maia…etc.	Trainium	Blackwell /Rubin
Difference	– Higher Price – Size limitation		

자료: TrendForce, IBK투자증권

그림 7. SK하이닉스 HBM4E 웨이퍼



자료: 언론보도, IBK투자증권

연결성이 새로운 병목인 이유

1. Agent를 위한 새로운 Computing의 등장

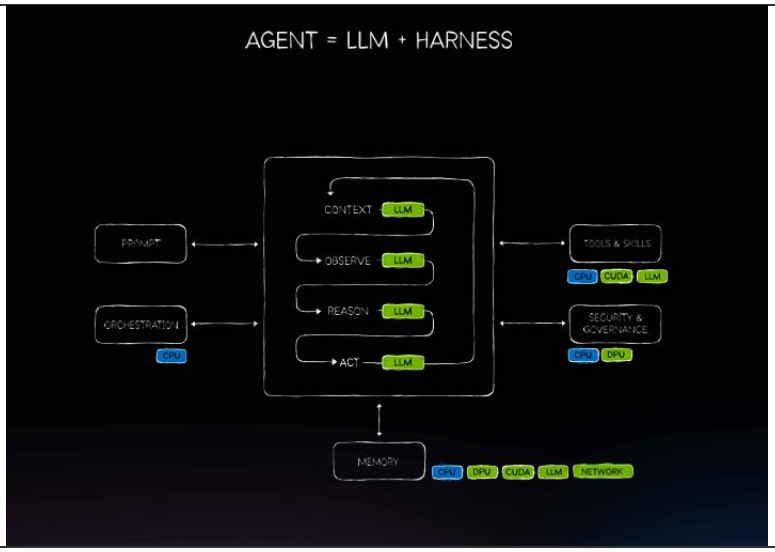
GTC Taipei 2026에서 젠슨 황은 〈Agent AI〉와 〈새로운 Computing〉의 시대가 도래했음을 선언했다. Agent는 AI가 스스로 사용자를 위해 필요한 도구를 사용해 작업을 수행하는 것을 의미한다. 기존의 Computing은 사용자가 어플리케이션을 열고, 키보드로 입력하면 Code가 운영체제 위에서 실행되는 구조였다. 새로운 Computing은 사용 주체가 사람에서 Agent로 바뀌면서, 명령 수행을 위한 입력과 스킬 사용을 Agent가 대신한다.

Agent는 모델-하네스-도구/스킬-런타임이 결합한 시스템이다. 젠슨 황은 모델을 뇌, 하네스를 몸, 런타임을 연장이 놓인 작업장에 비유하며, 단일 모델의 성능이 아닌 적합한 스킬·데이터베이스로의 접속을 조율하는 하네스의 역할을 강조했다. 에이전트가 하나의 LLM에 머무르지 않고 작업에 따라 여러 모델을 사용하기 때문에, 조율 소프트웨어 계층인 하네스에 따라 결과물이 달라진다.

Agent를 위한 Computing은 하나의 작업을 여러 프로세서가 분담하는 이종(heterogeneous) 구조로 Vera Rubin 플랫폼 역시 이를 반영해 설계했다. 사고 및 추론은 Rubin GPU가, 도구/스킬의 실행은 GPU와 CPU가 나눠 맡는다. 보안 하네스에는 Vera CPU와 BlueField DPU가 동원된다. Agent Computing에서는 Context의 장기화로 스토리지 시스템이 KV 캐시 관리에도 참여할 것으로 보인다.

작업을 다양한 하드웨어로 분산할수록 연산보다 연결성이 성능을 결정짓는 변수가 된다. GPU가 빠르게 연산해도 CPU·DPU·스토리지를 오가는 데이터가 지연되면 전체 처리량이 감소하기 때문이다. 소프트웨어에서는 하네스를, 하드웨어에서는 CPU·DPU 및 네트워킹에 집중해야 하는 이유이다.

그림 8. Agent 구조도



자료: NVIDIA, IBK투자증권

2. AI Factory의 생산성을 좌우하는 연결성

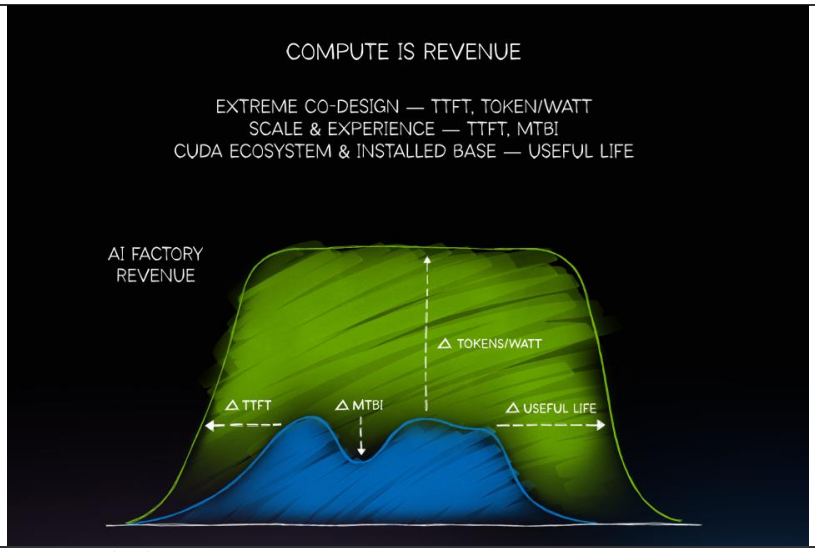
데이터센터의 기본 단위인 1GW 당 증설 비용이 급격히 증가하고 있다. 2020년대 말까지 글로벌 데이터센터 규모는 1,000GW를 상회할 전망이며, 자본투자 역시 급증할 것으로 보인다. 과거 20B 달러 수준에 불과했던 1GW 당 증설 비용이 현재는 50B 달러까지 증가했으며, 향후 100B 달러를 상회할 것으로 전망한다.

젠슨 황은 데이터센터를 토큰 기반의 수익성을 만들어내는 AI Factory로 정의한다. 공장이라는 표현처럼 와트당 토큰 산출량(Token/Watt), TTFT(Time to First Token) 등 핵심 수익 지표를 제시했다. 베라루빈 플랫폼은 수익성 개선을 위해 극단적인 통합 설계(extreme co-design)되었으며, compute, networking, storage, power, cooling을 최적화했다고 언급했다.

AI Factory의 경제성은 GPU가 쉬지 않고 토큰을 생산하는 시간에 달려 있다. GPU·CPU·스토리지·네트워크 등 계층 간 데이터 이동이 지연되면 GPU는 대기 상태가 된다. Blackwell은 72개 GPU를 하나의 추론 엔진으로 통합했다면, Vera Rubin은 연결 계층을 CPU와 스토리지까지 확장했다. Agent에서는 계획 수립, 도구 실행, 코드 처리, 메모리 접근 등을 조율하는 Orchestration 부담이 커지는데, Vera CPU는 해당 작업을 위해 설계되어 NVLink-C2C 기반 CPU-GPU 연결로 GPU 유휴 시간을 줄인다.

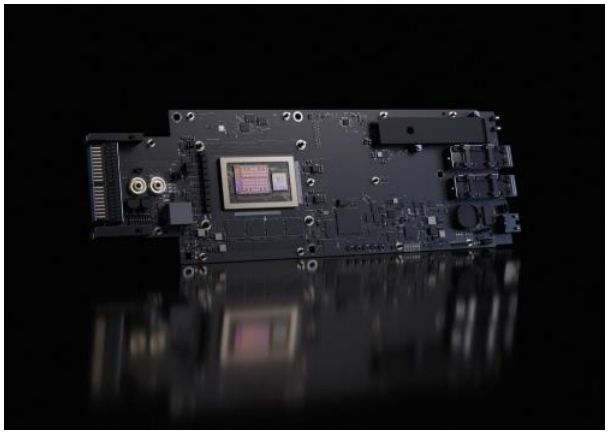
역설적으로 TCO(Total Cost of Ownership)는 낮아질 것으로 전망한다. 1GW당 증설 비용은 상승하지만, Vera Rubin NVL72는 GB200 NVL72 대비 토큰당 비용을 1/10 수준으로 낮추는 것을 목표로 하기 때문이다. 하드웨어 차원의 핵심은 BlueField-4 DPU와 STX이다. BlueField-4 기반 context memory storage 계층은 KV cache와 장기 컨텍스트 데이터를 HBM 밖에서 관리해 GPU 유휴 시간을 줄인다. Agent에서 컨텍스트와 메모리 호출이 늘어날수록, 네트워크와 스토리지 병목을 줄이는 하드웨어가 와트당 토큰 생산성과 TTFT를 좌우한다.

그림 9. NVIDIA Computing 효율 개선 예시



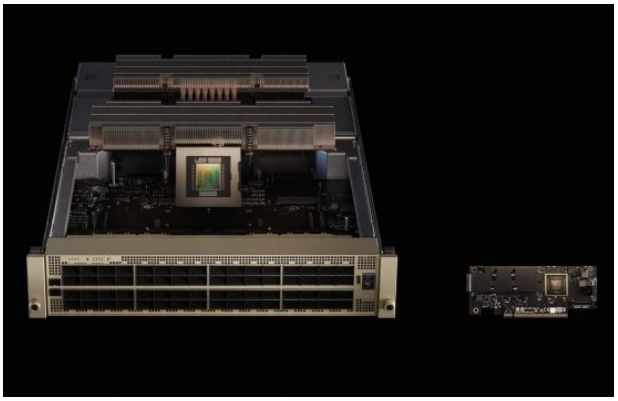
자료: NVIDIA, IBK투자증권

그림 10. NVIDIA BlueField -4 DPU



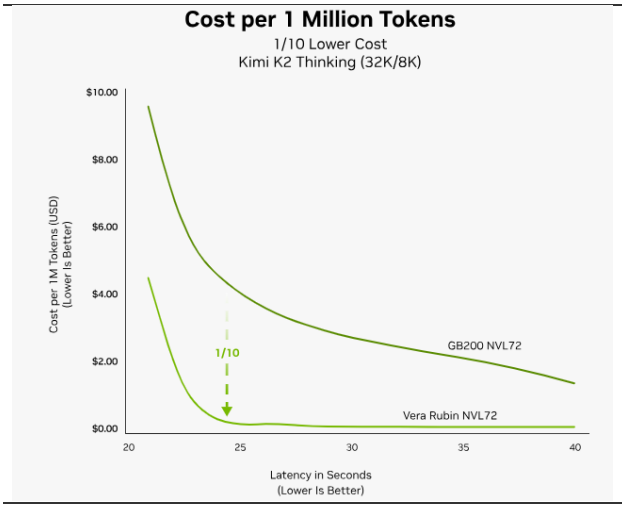
자료: NVIDIA, IBK투자증권

그림 11. NVIDIA Spectrum-X 이더넷 플랫폼



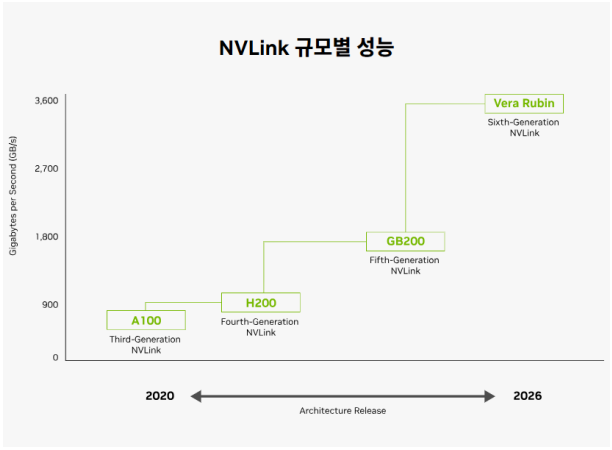
자료: NVIDIA, IBK투자증권

그림 12. Vera Rubin 1M 토큰당 비용 추이



자료: NVIDIA, IBK투자증권

그림 13. NVLink 세대별 대역폭 추이



자료: NVIDIA, IBK투자증권

광 네트워크, Scale Up 시작

연결성을 위한 가장 유력한 해답은 광 네트워크 기술이 될 것으로 전망한다. Marvell은 Computex 2026에서 연결성이 새로운 병목임을 선언했다. AI Factory 구축을 위해서는 프로세서 간 연결이 Computing Power의 구축만큼 중요해질 것으로 전망한다. GPU와 HBM 공급 부족이 NVIDIA와 메모리 3사의 밸류에이션을 재평가로 연결되었듯이, 다음 리레이팅 후보는 연결성 밸류체인에서 나올 가능성이 높다.

광 네트워크는 전송 거리에 따라 Scale across / out / up / 패키징으로 나뉜다. Scale-across는 데이터센터 사이를 연결하며, Scale-out은 데이터센터 내부에서 서버, 랙, 스위치를 묶는다. Scale-up은 랙 내부에서 다수의 GPU를 연결하고, Chiplet 패키징 내부의 연결성 개선도 논의 중이다.

표 2. AI Factory 단계별 네트워크 구분

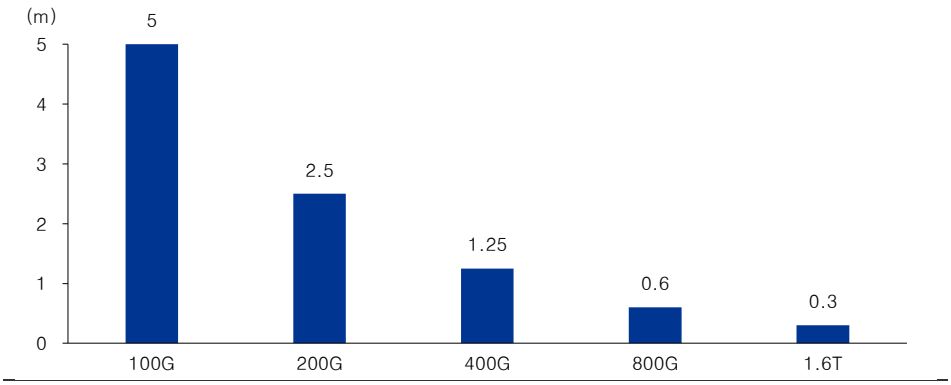
영역	대략적 거리	주 소재	연결 구간	핵심 부품 · 방식	핵심 기업	Vera Rubin 예시
Scale-across	수백 km 이상	광섬유	DC-DC	코히런트 변조, 코히런트 DSP, 장거리 광모듈, DCI, 실리콘 포토닉스	Marvell · Cisco(코히런트 DSP), Coherent · Lumentum (레이저 · 광부품), Ciena(장거리 전송 시스템)	-
Scale-out	수백 m	광섬유 중심	서버-스위치, 랙-스파인-코어	PAM4, PAM4 DSP, 800G · 1.6T 광모듈, TIA · 레이저 드라이버, SuperNIC, 이더넷, InfiniBand 스위치	NVIDIA(Spectrum-X · Quantum-X), Broadcom(Tomahawk), Marvell (PAM4 DSP · 스위치), Lumentum · Coherent (레이저), Corning(광섬유)	ConnectX-9 SuperNIC, Quantum-X800 InfiniBand, Spectrum-X Ethernet (Spectrum-6 = 800G CPO)
Scale-up	수 m	구리	GPU-GPU, CPU-GPU (any-to-any)	SerDes(200G→400G), NVLink · NVSwitch, retimer / (장기) CPO · 실리콘 포토닉스	NVIDIA(NVLink · NVSwitch), Broadcom(CPO 스위치), Marvell(SerDes · XConn 스케일업 스위치, 광 패브릭), Ayar Labs(optical I/O)	NVLink 6 Switch, NVLink-C2C (Vera CPU-Rubin GPU), Vera-Rubin 울트라 NVL576
패키징 / chiplet	mm ~ cm	실리콘 인터포저 · 유리기판	die-die, chiplet-chiplet (2.5D/3D)	die-to-die SerDes, UCIe, 인터포저, CoWoS, optical I/O(광 정렬 · 테스트)	TSMC(CoWoS 등 첨단 패키징), ASE(OSAT), Broadcom · Marvell (die-to-die SerDes/IP), Ayar Labs (co-packaged optical I/O)	NVLink-C2C 기반 superchip

자료: 각 사, 언론보도, IBK투자증권

광 네트워크의 scale-up(랙 내부) 영역 침투가 본격화될 것으로 전망한다. 구리는 대역폭이 커질수록 신호 손실이 커지고, 이를 보상하는 SerDes의 소비 전력도 급증한다. 레인당 속도가 200Gbps를 넘어 400Gbps로 향하면 한계 거리가 1m 안팎까지 좁혀져, 단일 랙을 넘어서면 광 연결은 불가피해진다. 이미 scale-out은 800Gbps에서 1.6Tbps로 넘어가는 과도기에 있으며, scale-up에서의 광네트워크 수요는 Rubin Ultra(NVL576) 플랫폼이 등장하는 2027년부터 본격화될 것으로 전망한다.

네트워크 계층별로 주목해야 할 주요 기술과 부품도 상이하다. Scale-across에서는 장거리 전송을 위한 coherent optics와 coherent DSP가 핵심이고, scale-out에서는 데이터센터 내부 서버·랙·스위치를 연결하는 800Gbps·1.6Tbps 광모듈과 PAM4 DSP가 단기적인 병목으로 예상되며, Scale-up에서는 GPU·XPU 연결을 지원하는 CPO와 silicon photonics 기업에 주목할 필요가 있다.

그림 14. Bandwidth별 구리 케이블 길이



자료: Marvell, IBK투자증권

표 3. NVIDIA의 NVL72 세대별 스펙 비교

구분	Blackwell GB200/GB300 NVL72	Vera Rubin NVL72	변화의 의미
GPU 패키지	72개 GPU	72개 GPU	–
GPU-GPU	NVLink 5, 130TB/s aggregate	NVLink 6, 260TB/s aggregate	랙 내부 통신 대역폭 2배
GPU당 NVLink	1.8TB/s	3.6TB/s	MoE·long-context all-to-all 통신 병목 완화
CPU-GPU	Grace CPU · 900GB/s NVLink-C2C	Vera CPU · 1.8TB/s NVLink-C2C†	CPU 오케스트레이션·메모리 접근 강화
Scale-out	Spectrum-X Ethernet / Quantum InfiniBand	ConnectX-9 SuperNIC (GPU당 800G×2), Spectrum-X, Quantum-X Photonics(CPO)[2]	랙 밖 네트워크가 플랫폼 핵심으로 부상
DPU	BlueField-3	BlueField-4	보안·격리·오프로드 중요도 상승
Storage	외부 스토리지 연동 중심	BlueField-4 STX storage processor	KV cache·long-term memory 병목 대응
보안	Confidential Computing(TEE) 지원	저장·전송·연산 구간 암호화 확장	agent AI·기업 데이터 보호
랙 설계	NVL72 rack-scale (Oberon, 구리 스파인)	cable-free 모듈러 랙(케이블·호스·팬 제거)[3]	조립 2시간→5분

자료: NVIDIA, 언론보도, IBK투자증권

Compliance Notice

동 자료에 게재된 내용들은 외부의 압력이나 부당한 간섭없이 본인의 의견을 정확하게 반영하여 작성되었음을 확인합니다.
동 자료는 기관투자가 또는 제3자에게 사전 제공한 사실이 없습니다.
동 자료는 조사분석자료 작성에 참여한 외부인(계열회사 및 그 임직원등)이 없습니다.
조사분석 담당자 및 배우자는 해당종목과 재산적 이해관계가 없습니다.
동 자료에 언급된 종목의 지분율 1%이상 보유하고 있지 않습니다.
당사는 상기 명시한 사항 외 고지해야 하는 특별한 이해관계가 없습니다.