meritz DM
메리츠 선진증시(DM) Direct Message
2025. 2. 27 (목)

글로벌 투자전략-선진국
Analyst 황수욱
02. 6454-4896
soowook.hwang@meritz.co.kr

# 엔비디아 4Q24 실적발표 및 Q&A

엔비디아 FY4Q24 실적발표, 매출액은 393억달러로 전년동기대비 78% 성장하며 컨센서스(381억달러) 상회. 블랙웰 시리즈 초기물량 시작, H200 시리즈의 지속적 분기 성장세를 보인 결과. FY2024 연간 매출액은 전년대비 142% 성장한 1,152억 달러 기록. GP 마진은 73%로 전분기대비 2%p 하락. 다음 분기 매출액 가이던스, GP 마진 가이던스를 각각 430억 달러, 71%로 제시, 매출액 가이던스는 컨센서스를 10억달러 상회. 핵심 동력인 데이터센터(컴퓨팅, 네트워크) 매출액이 전년동기대비 93.3% 성장. 엔비디아는 전일 3.7% 상승마감, 마진 하락에도 실적발표 이후 시간외에서 2% 상승 중

5,000억 달러 규모 스타게이트 프로젝트에 엔비디아가 핵심 기술 파트너로 참여하게 되었다고 발표. 글로벌 CSP들이 급증하는 AI 수요 충족을 위해 전세계 클라우드 region에 GB200 시스템 도입한다고 밝힘

로보틱스, 자율주행, 비전 AI 등 물리적 AI 애플리케이션으로의 엔비디아 Omniverse 통합을 확장하기 위해 생성형 AI 모델과 서비스 청사진 발표. 도요타가 엔비디아 드라이버 OS 탑재, 현대자동차가 엔비디아 AI와 Omniverse를 활용해 제조 고도화 및 첨단 로보틱스 도입하겠다고 발표. Physical AI 학습 플랫폼인 Cosmos 공개 이후 로보틱스 및 자동차 분야 선도 기업인 1X, Agile Robots, Waabi, Uber가 도입

젠슨황은 FY4Q24에 Blackwell 아키텍처 매출액은 110억 달러로 엔비디아 역사상 가장 빠른 제품 확장 속도를 기록한 것이라고 평가. Blackwell에 대한 수요는 엄청난데, 추론 AI는 또 하나의 확장 법칙을 추가하며, 훈련을 위한 컴퓨팅이 증가할수록 모델이 더 똑똑해지고(Scaling Law), 장기적인 사고를 위한 컴퓨팅이 증가할수록 더 나은 답을 제공(Test Time Computing, Time Scaling Law)한다고 언급. AI 학습뿐만 아니라 추론에도 막대한 컴퓨팅 수요 시사

학습 후 미세조정과 모델 맞춤화는 사전 학습보다도 훨씬 더 방대한 컴퓨팅 리소스를 필요로 할 수 있으며, 이는 토큰(token)을 대규모로 생성해 추가로 학습에 활용하기 때문. OpenAI의 o3, DeepSeek-R1, Grok 3 등 새로운 추론 모델은 "long-thinking reasoning AI"라 불리는 기술을 통해 100배 이상의 추가 컴퓨팅을 필요. Blackwell은 이런 추론용 모델을 위해 설계, H100 대비 최대 25배 높은 토큰 처리량과 20배 낮은 비용으로 Reasoning AI를 구현. GB200 초도물량 중 상당 부분이 추론용으로 할당된 것은 이번이 처음

NVIDIA는 지난 2년간 추론 비용을 200배 절감하는 혁신을 이룸. 4분기 데이터센터 매출의 절반가량은 대형 CSP(클라우드 서비스 제공업체)로부터 발생했으며, 전년 동기 대비 약 2배 성장했습니다. 이들은 Blackwell을 가장 먼저 구축했고, Azure, GCP, AWS, OCI가 전 세계 여러 클라우드 리전에 GB200 시스템을 도입해 AI 수요에 대응

AI에는 새로운 유형의 네트워킹이 필요. Spectrum X는 AI 컴퓨팅을 위해 이더넷을 최적화한 기술. 엔비디아는 이더넷 환경을 위한 Spectrum X 제공. 스타게이트(Stargate) 데이터센터 역시 Spectrum X를 사용할 예정

**이하는 컨콜 주요 Q&A 내용**

**Q:** 학습과 추론 사이의 경계가 모호해지며 추론 전용 클러스터의 축소 가능성?

**젠슨 황:** 사전학습이든, 사후학습이든 데이터센터를 유연하게 활용할 수 있어야 하는데, 엔비디아의 아키텍처는 이를 쉽게 구현함. 사전학습은 계속 확장될 예정, 멀티모달 데이터를 활용할 것이며 추론 데이터도 포함

사후 학습(파인튜닝, 강화학습, 지식 증류) 등 단계는 이미 사전학습보다 많은 컴퓨팅 자원을 요구. 모델이 생성한 토큰을 활용하기 때문에 여기서 생성되는 토큰 수가 폭발적으로 늘어날 수 있음. 테스트 시점에서의 추론, Long Thinking reasoning AI에서는 기존 일회성 추론 대비 100배 이상의 연산을 소모. 미래는 이런 추론 방식을 통해 수천 배, 수만 배의 연산이 필요할 수 있으며 블랙웰은 이런 방향을 염두에 두고 설계

**Q:** CES에서 언급된 랙단위 GB200 아키텍처의 복잡성 난관에서 병목 현상이 아직 있는지, NGL72 플랫폼에 대한 기대는 여전히 유효한지?

**젠슨 황:** CES 때보다 지금 기대가 더 큼. 이유는 단순한데 CES 이후로 훨씬 더 많은 수량을 출하했기 때문. 엔비디아는 GB의 대규모로 성공적으로 양산하며 지난 분기에만 110억 달러 매출 올림. 앞으로도 수요가 높아 제조 능력을 지속적으로 확장해야 함

**Q:** 마진 관련, GP 마진이 저점이라고 확언할 수 있는지. 내년까지 강력한 수요가 이어질 수 있다고 확신하는 근거가 무엇인지. DeepSeek이 내놓은 혁신이 그 전망을 바꿨는지?

**젠슨 황:** Blackwell 초도물량 생산에서 매출 총이익률은 70%대 초반을 형성할 것. 현재 엔비디아의 초점은 생산 속도를 높이는 것. 최대한 빨리 고객에게 제품을 제공해야 하기 때문. 그런데 블랙웰 생산이 완전히 안착되면, 원가를 절감해 마진을 개선할 수 있을 것으로 예상. 그래서 올해 하반기에는 70%대 중반으로 회복할 것으로 예상

전망에 대해서는 단기, 중기, 장기 모두 긍정적으로 전망. 단기적으로는 실제 발주나 파트너사들의 수요 전망, 중기로는 인프라 및 자본지출 확대 규모를 과거와 비교해볼 수 있음. 장기적으로는 기존의 CPU 기반 직접 코딩에서 머신러닝/AI 기반으로 완전히 전환되는 추세. 향후 모든 소프트웨어와 서비스는 결국 AI를 사용하게 될 것. Agentic AI, Physical AI를 만들기 위해 아직도 수많은 혁신 스타트업이 등장 중. 신생 기업들도 모두 상당한 규모의 컴퓨팅 인프라가 필요

**Q:** 블랙웰 울트라 하반기 출시 목표인데, 수요를 어떻게 전망하는지?

**젠슨 황:** 첫 번째 Blackwell 출시 과정에서 몇 달 정도 지연을 겪기도 했지만, 현재는 완전히 회복. Blackwell의 양산 생산이 순조롭게 진행 중, 다음 세대 제품 개발 일정 역시 그대로 유지하고 있음. '호퍼(Hopper)에서 블랙웰(Blackwell)로' 넘어갈 때처럼 큰 폭의 변경이 있진 않음. 이후 세대인 '베라 루빈(Vera Rubin)'도 이미 파트너사들에게 미리 공유하고 있으며, 해당 전환 과정도 준비 중. GTC에서 블랙웰 울트라, 베라 루빈, 그 이후 이어질 제품에 대해 자세히 이야기할 예정

**Q:** GPU 상용 제품(merchant GPU)과 맞춤형 ASIC이 어떻게 균형을 이룰지?

**젠슨 황:** 전반적으로 완전히 다른 영역. NVIDIA 아키텍처는 범용성(general)을 지향, 데이터 전처리부터 학습, 그리고 강화학습을 통한 사후 학습, 최종적으로 테스트 시점 추론까지 전 과정을 가속할 수 있음. 그리고 성능과 혁신 주기가 매우 빠름. 성능이 2배, 4배, 8배씩 늘어나면, 동일 전력에서 처리 가능한 작업량이 그만큼 늘어남. AI 팩토리는 생성되는 토큰에서 직접 매출이 나오기 때문에, 토큰 처리량이 4~8배 늘어나면 해당 데이터센터에서 나오는 매출도 그만큼 늘어남

ASIC을 만드는 것 자체가 간단하지 않을뿐더러, NVIDIA가 제공하는 생태계와 SW 스택이 2년 전보다 10배는 복잡해짐. AI가 급속도로 발전하고, 전 세계 개발자들이 엔비디아의 아키텍처에 맞춰 소프트웨어를 계속 쌓아 올리고 있기 때문. 이걸 다른 칩셋 위에서 다시 구현한다는 건 엄청난 난관

**Q:** 규제로 인한 다른 지역(예: 중국) 매출이 제한될 경우 미국 시장이 그 공백을 충분히 메울 수 있을지?

**젠슨 황:** 중국 매출 비중은 이전과 비슷한 수준. 수출 규제와 비교하면 절반 정도 되는 수준. 지역별 매출을 볼 때 가장 중요한 점은 AI가 소프트웨어라는 점. AI는 매우 혁신적이고 현대적인 소프트웨어 형태이며, 이미 대중화(mainstream)가 되었음. 미래의 컴퓨팅은 가속 컴퓨팅을 중심으로 구축될 것, 지금은 그 전환의 초기 단계. 규모 측면에서도 엄청난 잠재력이 있어 갈 길이 멂(질문에 대한 직접적인 답을 피하고 산업의 큰 전망으로 질문의 답을 갈음함)

**Q:** 현재 하이퍼스케일러가 AI 인프라를 사들이는 목적이 내부(자체 사용)와 외부(CSP 사업)용 중 어느 쪽이 더 큰지 말해줄 수 있는지?

**젠슨 황:** 하이퍼스케일러들은 자체적으로 대규모 언어 모델을 학습하거나 추론을 수행하기도 하지만, 동시에 엔터프라이즈가 클라우드를 이용해 AI를 돌릴 수 있도록 인프라도 제공. 어느 쪽이든 지금 둘 다 매우 빠른 속도로 성장하고 있음

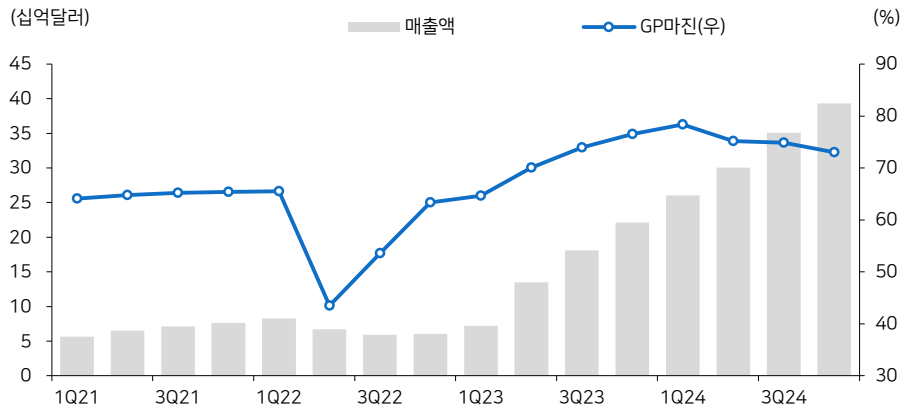
**Q:** 앞으로의 로드맵을 보았을 때, 이미 구축된 인프라의 교체 주기(Replacement Cycle)에 대해 어떻게 생각하는지?

**젠슨 황:** 우선, 아직도 볼타(Volta), 파스칼(Pascal), 암페어(Ampere) GPU가 실제로 활용되고 있음. 비교적 덜 무거운 연산들은 과거 세대인 암페어(Ampere) GPU로 충분히 처리할 수 있음. 반면, 실제 모델 학습은 호퍼(Hopper) 시스템에서 진행. 결국, 모든 세대의 GPU가 유용하게 쓰이고 있음

**Q:** 하반기에 마진이 70% 중반대로 가려면 분기별로 200bp 정도씩 상승해야 하지 않은지? 반도체 업계 전체로 볼 때 관세 문제도 어떻게 작용할지 아직 확실치 않음. 올해 후반기에 그렇게 마진을 개선할 수 있다는 자신감의 근거는?

젠슨 황: 장기적으로 마진 개선을 위한 기회가 여럿 존재. 원가 구조를 효율화, 고객들에게 제품을 신속히 제공하기 위해 ramp up에 집중하고 있지만, 가능한 한 빨리 마진 개선 작업도 병행. 관세 문제는 아직 불확실한 부분이 많음. 미국 정부의 정책(시기, 적용 범위, 세율 등)이 어떻게 나올지 기다려봐야 알 수 있기 때문. 하지만 항상 수출 규제나 관세 등 관련 규정을 철저히 준수할 것

그림1 엔비디아 매출액과 GP 마진




자료: Bloomberg, 메리츠증권 리서치센터