2025. 2. 26 (목)

4Q25 엔비디아 실적발표

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

- 매출/가이던스 서프라이즈, 양호한 마진+피지컬AI 관련 매출 숫자까지 공개
- AI 에이전트로 연산량이 폭증하는 변곡점, 추론량=매출 → "단위 연산량당 비용 절감"이 2026 GTC 화두 중 하나일 수 있다는 생각
- 이 논의의 진전이 산업 전반에 걸친 AI 비용과 마진 우려를 덜어내는 업사이드 리스크로 작용할 수 있는 포인트일수도

엔비디아 4Q25 매출액 680억달러로 전년대비 73% 성장, 3분기 대비 성장률도 순차적으로 가속. GP 마진은 75%, 다음분기 가이던스로 780억달러 ±2%, GP 마진은 75% ±50bp, 연간으로는 70%대 중반 유지 전망

데이터센터 매출은 분기 중 110억 달러 성장, 클라우드 사업자, 하이퍼스케일러, AI 모델 개발사, 국가(소버린 AI) 고객군 다양하게 확대. 코렛 크레스 '에이전틱 및 피지컬 AI 어플리케이션이 성과에 기여하기 시작'했다는 컨콜 언급이 인상적. 연간 기준 데이터센터 매출은 1,940억달러로 전년대비 68% 성장, 2023회계연도 이후 13배 확장. 2026년 전망은 작년에 공유했던 5,000억 달러 규모의 블랙웰, 루빈 매출 기회 전망을 포함한 것보다 더 큼. 데이터센터들이 전력 제약을 받고 있다는 점도 언급

엔비디아 아키텍처의 강점은 토큰당 최저 비용을 제공, 경쟁 제품과 차별성. 클라우드에서는 Hopper뿐만 아니라, 6년된 Ampere 기반 제품 대부분도 매진된 상태라고 명시하며 작년 부각된 감가 상각 우려를 불식시킴. Scale up, Scale out, Scale across 관련 기술 수요 사상 최고치(네트워크 수요를 의미), 네트워크 사업은 멜라녹스를 인수한 2021년 대비 10배 이상 성장, 소버린 AI 사업도 전년대비 3배 성장, 장기적으로는 소버린 AI 기회가 AI 인프라 시장과 최소 동일한 속도로 가속할 것이라고 언급하며, 현재 엔비디아 핵심 성장 모두 고속성장 중인 가운데 향후 전망에 대해서도 청사진을 밝힘

피지컬 AI 관련 매출을 구체적으로 처음 공개. 2026 회계연도 엔비디아 매출에 60억 달러 이상 기여했다고 언급. 웨이모, 테슬라, 우버 등 상업용 로보택시 플릿 운행은 기하급수적으로 늘고 있고, 이는 수천억 달러 매출을 창출할 시장을 형성하게 된다고 언급. 이 확장에는 더 많은 컴퓨트가 필요하며 엔비디아 Cosmos와 Issac의 새로운 공개형 모델과 프레임 워크로 로보틱스 진전, 보스턴 다이나믹스, 캐터필러, LG전자 등 선도 기업들의 산업용 피지컬 AI 채택을 가속한다는 주장

젠슨황의 엔트로픽의 Cluade Cowork에 대한 언급, 혁신적이며 엔터프라이즈 AI 채택의 수문을 열었다고 평가, 거대한 AI(기존 AI보다 월등히 더 많은 연산량이 필요한 AI)의 ChatGPT 모먼트가 도래했다는 것. 엔비디아는 이런 프론티어 모델 성장 가운데 모든 성장 단계에서 파트너가 될 수 있는 독보적 위치

젠슨황은 최근 Groq과 저지연 추론 기술에 대한 비독점(non-exclusive) 라이선싱 계약을 체결했고, Mellanox 때와 마찬가지로 뛰어난 엔지니어 팀을 엔비디아로 맞이했다고 언급. Groq Innovations를 통해 엔비디아 아키텍처를 확장하여 AI 인프라의 성능과 가치에서 새로운 수준을 열어갈 것이며 다음 달 GTC에서 더 많은 내용을 공유하겠다고 밝힘. Groq 인수에 대한 기술적 시사점이 3월 GTC 주요 어젠다 중 하나일 수 있을 것이라는 것

표면적으로 Groq 인수는 TPU 등 ASIC이 강점이 있었던 초저지연 추론 영역에서 엔비디아의 경쟁력을 확보하기 위한 움직임으로 풀이되었음. HBM 타이트 국면에서 '추론 전용 아키텍처'의 경제성(지연/전력/비용)을 높이는 선택지라는 평가, AI 학습 영역뿐만 아니라, 추론 영역에서도 AI 종속성을 강화시키려는 전략의 일환

주요 Q&A

Q1: 2027년까지 성장 가시성, 다만 그 이후의 성장 속도는 올해만큼 어렵다는 게 시장의 중론. 여러 고객의 현금창출 능력까지 압박을 받는 중. 고객들의 CapEx를 계속 늘릴 수 있는 능력에 대해 어떻게 확신하는지? 고객 CapEx가 더 늘지 않으면 엔비디아는 그 한정된 총액 안에서 성장할 방법이 있을지?

A: 그들의 현금흐름이 성장할 것으로 확신. 그 이유는 간단함. 우리는 AI 에이전트의 유용성이 전세계와 모든 기업에서 변곡점에 들어섰다는 것을 확인. 이 때문에 엄청난 컴퓨팅 수요가 나타나고 있음. 이 세계에서는 컴퓨트가 곧 매출.

컴퓨트 없이는 토큰을 생성할 방법이 없고, 토큰 없이는 매출을 성장시킬 방법이 없음. 이 새로운 세계에서는 컴퓨트가 매출과 같음. 지금 시점에서 Codex의 생산적 사용, Claude Code, Claude Cowork에 대한 기대감, Open Claw에 대한 엄청난 열기, 그리고 그 엔터프라이즈 버전까지—이 모든 것들을 감안할 때, 모든 엔터프라이즈 ISV들이 자사 도구 플랫폼 위에 에이전틱 시스템을 구축하고 있다는 점을 고려하면, 우리가 변곡점에 도달했다고 확신

Q2: Anthropic, 잠재적으로 OpenAI, CoreWeave 등 여러 곳에 대한 전략적 투자에 대해 언급, Intel, Nokia, Synopsys 같은 파트너도 언급했음. 엔비디아는 분명히 모든 것의 중심에 있음. 이런 투자들의 역할과, 대차대조표를 엔비디아가 생태계에서의 입지를 키우고 그 성장에 참여하기 위한 도구로 어떻게 활용하는지?

A: 근본적으로 모든 것의 핵심은 엔비디아의 생태계. 우리는 전세계 엣지와 로보틱 시스템 전반에 걸쳐서도 존재함. 수천개의 AI 네이티브 기업들이 엔비디아 위에 구축되어 있음. 우리의 생태계는 과거보다 더 풍부. 과거에는 주로 GPU의 컴퓨팅 플랫폼이었지만, 이제 우리는 AI 인프라 회사이고, 컴퓨팅 플랫폼이 모든 측면에 존재. 컴퓨팅부터 AI 모델, 네트워킹, DPU까지, 그 위에는 컴퓨팅 스택 존재

Q3: 데이터센터 전체 프로필에서 네트워크 비중이 커지고 있음. 주문 흐름(Order book)으로 볼 때, 특히 곧 출시될 Spectrum XGS 및 102T Spectrum-6 스위칭 플랫폼을 고려하면 Spectrum의 런레이트는 현재 어디로 향하고 있는지?

A: 우리는 컴퓨터 '랙(rack)'을 출하. 그리고 그 NVLink 스위치 기반 스케일업 시스템은 Spectrum X와 InfiniBand로 scale-out 됩니다. 우리는 둘 다 지원합니다. 더 나아가, Spectrum scale-across를 통해 데이터센터 간 확장까지 함. 우리가 네트워킹을 생각하는 방식은 본질적으로 확장.

우리가 데이터센터에서 이더넷을 '인공지능 방식'으로 확장하는 기능을 만들었음. 이부분에 매우 강하고, Spectrum X의 성능이 이를 보여줌 100~200억 달러 규모의 AI 팩토리를 구축할 때, 10~20%의 차이-네트워크 효율과 처리량, 데이터센터 네트워크 활용율의 차이는 실제 돈으로 이어짐.

Q4: 대규모 컨텍스트 윈도우와 Groq이 코드에 특화된 솔루션을 추가할 가능성을 감안할 대, 향후 로드맵을 어떻게 봐야 할지? 워크로드별 고객별로 커스텀 실리콘을 더 강화하는 방향으로 봐야 할지? 특히 엔비디아가 분리형 아키텍처로 전환하면서 이런 방향성이 더 커진다고 봐야 할지?

A: 우리가 다이얼 인(dial-in)을 쓰지 않는 것은 아니지만, 원하는 방향은 가능한 이를 쓰지 않고 확장하는 것. 그 이유는 간단함, 다이얼-인을 할때마다 인터페이스를 하나 건너야 하는데, 이는 불필요한 지연과 전력을 추가시킴. GB 아키텍처와 Rubin 아키텍처를 보면, 레티클의 한계 수준의 거대한 다이 2개를 사용하고 이를 결합함. 이렇게하면 아키텍처 상에서 크로싱(경계/인터페이스를 넘는 것)의 양이 줄어들게 됨.

이 다이얼-인 tax는 경쟁사들의 아키텍처 효율성에서 드러남. 사람들이 이를 소프트웨어 우위라고 부르기도 하지만, 소프트웨어가 어디에서 시작되고 아키텍처가 어디서 시작/끝나는지는 구분하기 어려움

CUDA 아키텍처는 의심할 여지없이 더 효과적, 어떤 컴퓨팅 아키텍처보다도 flop 당, 와트당 성능이 뛰어남. 우리가 아키텍처를 설계하는 방식 때문. Groq과 저지연 디코더에 대해 어떻게 생각하냐는 질문에 대해서는, GTC에서 공유하고 싶은 좋은 아이디어들이 있음. 다만 핵심을 말하면, CUDA 덕분에 우리의 인프라는 매우 범용적, 우리는 그 방향을 계속 이어갈 것

GPU는 아키텍처적으로 호환됨. 즉, 오늘 Blackwell용으로 모델을 최적화하는 작업은 Hopper에도, Ampere에도 도움. A100이 지금도 '신선하게' 느껴지고, 세상에 배치된 지 수년이 지나도 성능을 유지하는 이유가 바로 이것때문. 아키텍처 호환성 덕분에, 우리는 소프트웨어 엔지니어링과 최적화에 막대한 투자를 하면서도, 클라우드·온프레미스·엣지 등 어디에 있든 여러 세대의 GPU 설치 기반 전체가 그 혜택을 받는다는 확신을 가질 수 있음

그래서 우리는 그 방향을 계속할 것이고, 유효 수명(useful life)을 늘리고 혁신·유연성·속도(velocity)를 확보할 수 있음. 이는 고객에게 성능으로, 그리고 무엇보다 중요한 성능/달러(performance per dollar)와 성능/와트(performance per watt)로 이어질 것

Q5: 이번 분기 데이터센터 매출이 전분기 대비 100억 달러 이상 증가했는데, 가이던스도 데이터센터에서 전분기 대비 증가분 100억 달러 대부분이 발생하는 것처럼 보임. 이 부분에 대한 연중 흐름은? 게이밍 산업에 대한 코멘트? 메모리 이슈 등은 이해하는데 2027 회계연도에서 전년대비 성장할 수 있다고 보는지?

A: 연간 전체를 생각해 보면, 우리는 Blackwell을 계속 판매하고 공급할 것이고, 동시에 Vera Rubin이 시장에 나오기 시작하는 모습도 보게 될 것. Vera Rubin의 초기 램프가 얼마나 될지에 대해서는 아직 판단하기 이름. 램프는 하반기에 시작될 것이나 수요와 관심이 강하다는 점에서는 혼선이 없음

사실상 모든 고객이 Vera Rubin을 구매할 것으로 예상함. 핵심은 '우리가 시장에 얼마나 빨리 내놓을 수 있는지', 그리고 고객들이 데이터센터에서 얼마나 빨리 이를 구축할 수 있느냐. 그게 첫 번째 질문에 대한 답

두 번째는 게이밍인데, 공급이 더 있으면 좋겠지만, 앞으로 몇 분기 동안은 공급이 매우 타이트할 것으로 보고 있음, 지금은 판단하기 어려움

Q6: AI 투자 수익의 더 많은 부분이 추론(inference) 워크로드에서 나오고 있는 상황에서, CUDA의 중요성에 대해 말씀해 줄 수 있을지?

A: CUDA가 없다면 우리는 추론을 어떻게 해야 할지조차 알기 어려웠을 것. 몇 년 전에 우리가 도입한 TensorRT-LLM부터 시작해 전체 스택이 있는데, 이건 여전히 세계에서 가장 성능이 좋은 추론 스택. 이를 NVLink에 최적화하려면, NVLink 72 전반의 총 대역폭을 활용할 수 있도록, CUDA 위에서 동작하는 새로운 병렬화 알고리즘을 발견하고 발명해야 함. 즉, CUDA 위에서 워크로드와 추론을 분산시키는 방법을 만들어야 함

NVLink 72 덕분에 우리는 세대 기준으로 와트당 성능을 50배 더 제공할 수 있었으며 이는 엄청난 리드. 스위칭 기술을 만들고, 스위치를 디스어그리게이션하고, 시스템 랙을 구축하고—우리는 그 모든 것을 '공개된 상태에서' 해왔고, 모두가 그것이 얼마나 어려운지 알고 있었음. 하지만 결과는 놀라움. 와트당 성능은 50배, 달러당 성능은 35배입니다. 그래서 추론에서의 도약이 엄청남

이제 추론은 고객에게 '곧 매출'이라는 점을 이해하는 것이 중요. 에이전트가 엄청나게 많은 토큰을 만들어내고, 그 결과가 매우 효과적이기 때문. 에이전트가 코딩을 할 때, 분 단위에서 시간 단위로 실행되면서 수천, 수만, 수십만 토큰을 생성합니다. 이런 에이전틱 시스템은 팀처럼 여러 에이전트를 동시에 생성해 협업시키기도 함. 생성되는 토큰 수는 정말 exponential하게 증가하고 있음

그래서 우리는 훨씬 더 높은 속도로 추론해야 함. 그리고 더 높은 속도로 추론하고, 각 토큰이 '달러화(dollarized)'될 때, 이는 곧바로 매출로 이어짐. 즉, 추론 성능 = 고객의 매출. 데이터센터 관점에서는 와트당 추론 토큰(inference tokens per watt)이 CSP의 매출로 직결됨

그 이유는 모두가 전력 제한(power limited)을 받기 때문. 데이터센터가 몇 개가 있든, 각 데이터센터가 100MW든 1GW든 전력 한계가 있음. 따라서 와트당 성능이 가장 좋은 아키텍처가 중요해짐. 와트당 토큰(=성능 토큰/와트)은 토큰이 달러화되므로 달러/와트로 전환되고, 이는 1GW 단위에서는 매출로 직결

그래서 모든 CSP와 하이퍼스케일러가 이제 이것을 이해함. CapEx는 컴퓨트로 이어지고, 올바른 아키텍처를 선택한 컴퓨트는 매출 극대화로 이어지며, 컴퓨트는 곧 매출. 오늘 용량(capacity)에 투자하지 않고, 컴퓨트에 투자하지 않으면, 매출 성장은 불가능합니다. 모두가 이를 이해

컴퓨트는 매출. 올바른 아키텍처를 고르는 것이 엄청나게 중요. 이제는 단순히 '전략적' 수준을 넘어서, 그들의 실적(earnings)에 직접 영향을 줌. 와트당 성능이 가장 좋은 아키텍처를 고르는 것이 말 그대로 전부

Q7: 총마진의 장기적으로 70%대 중반(mid-70s) 유지 가능성

A: 총마진의 가장 중요한 레버는 고객에게 '세대 도약(generational leaps)'을 제공하는 것. 그게 단연 가장 중요. 무어의 법칙(Moore's Law)이 제공할 수 있는 수준을 훨씬 뛰어넘는 와트당 성능을 세대 단위로 제공할 수 있고, 시스템 원가 상승 대비 가격보다 훨씬 더 큰 폭으로 달러당 성능을 제공할 수 있다면, 우리는 총마진을 유지할 수 있음

그 결과, 클라우드에 있는 6년 된 GPU마저 완전히 소진되고, 가격이 오르고 있음. 즉, '현대적 방식의 소프트웨어'에 필요한 컴퓨트가 지수적으로 늘어난다는 것을 우리는 알고 있음. 그래서 우리의 전략은 매년 '전체 AI 인프라'를 제공하는 것

Rubin 다음 세대도 여러 신규 칩을 제공할 것. 그리고 매 세대마다 우리는 와트당 성능과 달러당 성능에서 '여러 배(X factors)'의 개선을 제공하겠다는 약속을 지키려 함. 이런 속도와, 극한 공동 설계 능력 덕분에 우리는 고객에게 그 가치를 제공할 수 있고, 이것이 우리가 제공하는 가치의 핵심

Q8: 우주 데이터센터(space data centers), 현실적으로 얼마나 가능하다고 보는지? 타임라인(horizon)은 어느 정도이며, 경제성(economics)은 어떻고 시간이 지나며 어떻게 진화할지?

오늘 기준으로 경제성은 좋지 않지만, 시간이 지나며 개선될 것. 우주는 지상과 작동 방식이 근본적으로 다름. 에너지는 풍부. 태양광 패널은 크지만, 우주에는 공간이 많음

열 방출은—우주는 차갑습니다. 하지만 공기 흐름(airflow)이 없음. 그래서 열을 방출하는 유일한 방법은 전도 뿐임 필요한 라디에이터(radiators)는 꽤 큼

액체 냉각은 무겁고 얼어붙기 때문에 사실상 불가능. 따라서 지구에서 쓰는 방식과 우주에서 쓰는 방식은 다를 수밖에 없음. 하지만 우주에서 수행하고 싶은 컴퓨팅 문제는 많음. 그래서 엔비디아는 이미 "세계 최초의 우주 GPU"를 갖고 있고, Hopper는 우주에 있음.

우주에서 GPU의 가장 좋은 사용처 중 하나는 이미징(imaging). 광학과 AI를 통해 초고해상도 이미징을 하고, 다양한 각도의 데이터를 재투영(reprojection)하고, 업스케일(up-res)과 노이즈 제거(noise reduction)를 수행해, 매우 큰 스케일에서 아주 빠르게 초고해상도 이미지를 얻을 수 있음. 페타바이트급 이미징 데이터를 지구로 보내서 처리하는 건 어려움. 차라리 우주에서 처리하는 편이 더 쉬움

그리고 흥미로운 것을 발견할 때까지는 수집·처리한 데이터 대부분을 무시(폐기)할 수도 있음. 따라서 우주에서의 AI는 매우 흥미로운 응용을 갖게 될 것

Q9: 매출 다변화(revenue diversification)에 대한 스크립트 발언에 대한 질문. 하이퍼스케일러가 매출의 50%를 넘지만, 성장은 데이터센터의 나머지 고객들이 주도했다는 취지로 말씀하신 것으로 이해. 확인 차원에서, 비(非) 하이퍼스케일 고객이 더 빠르게 성장했다는 의미인지? 그렇다면 이들이 하이퍼스케일러와 무엇이 다른지? 이 트렌드가 지속될 것으로 보는지? 장기적으로 비하이퍼스케일러가 더 큰 비중을 차지하도록 고객 구성이 진화할지?

A; 우리가 말한 상위 5개 고객(CSP/하이퍼스케일러)은 현재 총매출의 약 50%를 차지. 그 외에 우리는 매우 다양한 고객군과 일하고 있음. AI 모델 개발사, 엔터프라이즈, 슈퍼컴퓨팅, 주권(국가) 고객 등이 포함. 여러 다른 팩터가 있지만, 말한 이해가 맞음

이 또한 매우 빠르게 성장하는 영역. 우리는 다양한 클라우드 제공자들 전반에서 플랫폼 포지션이 강하고, 동시에 전 세계적으로 매우 다양한 고객을 확보하고 있음. 이런 다변화는 우리가 전 영역을 서비스할 수 있게 해주며 큰 도움

CUDA 위에서 구축하면, 우리는 모든 클라우드에 존재하고, 모든 컴퓨터 제조사를 통해 제공되며, 엣지에서도 호환되는 유일한 가속 컴퓨팅 플랫폼. 우리는 통신도 키우고 있음. 미래의 무선 기지국은 모두 AI 기반 라디오가 될 것이고 미래 무선 네트워크도 컴퓨팅 플랫폼이 될 것

이는 이미 정해진 미래. 이를 가능하게 하는 기술을 발명해야 하는 누군가가 필요. 우리는 거의 모든 로봇, 모든 자율주행차에 들어가 있음. CUDA는 GPU 내부 텐서 코어라는 특수 프로세서 성능의 이점과 동시에, 언어·컴퓨터비전·로보틱스·생물학·물리 등 거의 모든 AI와 다양한 컴퓨팅 알고리즘을 풀 수 있는 유연성을 제공. 고객 기반의 다양성은 우리의 가장 큰 강점 중 하나

설령 프로세서가 프로그래머블하더라도, 우리가 생태계를 키우지 않았다면, 우리는 남의 생태계에서 디자인윈을 따낸 범위 이상으로 성장하기 어려움. 그래서 우리는 우리가 만든 플랫폼 덕분에 자연스럽게 생태계를 확장할 수 있음

마지막으로, OpenAI, Anthropic, xAI, Meta, 그리고 전 세계 오픈소스와의 파트너십이 매우 중요. Hugging Face에는 150만 개의 AI 모델이 있고, 그 모두가 NVIDIA CUDA에서 동작. 오픈소스 전체는 아마도 세계에서 두 번째로 큰 '모델 집합'일 것. OpenAI가 가장 크고, 두 번째는 모든 오픈소스의 집합

엔비디아는 그 모든 것을 돌릴 수 있기 때문에, 우리 플랫폼은 매우 대체 가능(fungible)하고, 매우 사용하기 쉽고, 투자하기에 안전. 그리고 이것이 고객의 다양성과 플랫폼의 다양성을 만들고, 모든 나라에서 사용 가능. 우리는 전 세계 생태계를 지원하기 때문

Q10: 플랫폼 극한 공동설계 관련 질문, 향후 아키텍처 진화에서 Vera가 갖는 중요성?

A; GTC에서 더 말씀드리겠습니다만, 큰 틀에서 말하면, 우리는 전 세계 다른 CPU들과는 근본적으로 다른 아키텍처 결정을 내렸습니다. Vera는 LPDDR5를 지원하는 유일한 데이터센터 CPU. 매우 높은 데이터 처리 능력에 초점을 맞춰 설계

우리가 관심 있는 컴퓨팅 문제의 대부분이 데이터 구동(data-driven)이기 때문. AI가 대표적. 단일 스레드 성능(single-threaded performance)과 대역폭 대비 성능의 조합은 압도. 훈련(training) 전에 데이터 처리(data processing)가 필요. 즉, 데이터 처리 → 프리트레이닝(pre-training) → 포스트트레이닝(post-training) 단계가 있고, 포스트트레이닝에서는 AI가 도구(tools)를 사용하는 법을 배우고 있음

도구 사용의 많은 부분은 CPU-only 환경에서 실행되거나 GPU+CPU 가속 환경에서 실행. Vera는 포스트트레이닝에 매우 뛰어난 CPU가 되도록 설계. AI 파이프라인 전체에서 CPU가 많이 쓰이는 유스케이스가 존재. 우리는 GPU뿐 아니라 CPU도 좋아함

알고리즘을 우리가 지금처럼 한계까지 가속하면, 아마달의 법칙(Amdahl's Law)에 따라 정말 빠른 단일 스레드 CPU가 필요. 그래서 우리는 Grace를 단일 스레드 성능에서 매우 뛰어나게 만들었고, Vera는 그보다도 훨씬 더 뛰어남

Q11: 자본 배분(capital deployment)에 대해? 그런데 실적이 아무리 좋아도 주가가 크게 오르지 않았음. 그렇다면 현재 주가는 자사주를 많이 매입하기에 꽤 좋은 가격이라고 느낄 법도 함

A: 우리는 자본 환원(capital return)을 매우 신중하게 봄. 그리고 우리가 할 수 있는 가장 중요한 일 중 하나는, 앞에 있는 '극한 생태계(extreme ecosystem)'를 지원하는 것이라고 믿음. 이는 공급업체 지원부터 시작해, 필요한 공급을 확보하고 공급업체의 캐파를 돕는 일, 그리고 우리 플랫폼 위에서 AI 솔루션을 개발하는 초기 개발자들을 지원하는 일까지 포괄

따라서 우리는 이를 프로세스의 매우 중요한 부분으로 유지하며, 전략적 투자도 계속할 것. 물론 우리는 자사주 매입도 계속하고 있고, 배당도 하고 있음. 그리고 연중 적절한 시점에 다양한 매입 기회를 계속 찾아갈 것

Q12: 이전에 2030년까지 데이터센터 CapEx가 3~4조 달러에 이를 수 있다는 잠재력을 언급했었음. 그 변곡점을 가장 가능성 높게 견인할 핵심 애플리케이션이 무엇이라고 보는지? 피지컬 AI, 에이전틱 AI, 혹은 다른 개념? 그리고 3~4조 달러라는 그 "범위(envelope)"에 대해서도 여전히 자신 있는지?

A: 미래의 소프트웨어는 AI를 사용해 만들어지는 'AI-활성화 소프트웨어'로 수행됩니다. AI는 토큰 구동(token-driven). 모두가 토크노믹스(tokenomics)를 얘기하고, 데이터센터가 토큰을 생성한다는 얘기를 함. 추론은 토큰을 생성하는 일이고, 우리는 토큰을 생성. 그래서 토큰 생성은 미래 소프트웨어와 컴퓨팅에 관한 거의 모든 것의 중심에 있음

과거의 컴퓨팅 사용 방식과 비교하면, 과거 소프트웨어가 요구하던 컴퓨트 수요는 미래에 필요한 것의 극히 일부에 불과. AI는 이미 왔고, 다시 뒤로 가지 않을 것. AI는 앞으로 더 좋아질 뿐. 따라서 과거에 세계가 전통적 컴퓨팅에 연간 3,000~4,000억 달러를 투자했다고 치면, 이제 AI가 등장했고 필요한 컴퓨트는 과거 방식 대비 천 배(1,000x) 수준으로 높음. 컴퓨트 수요가 훨씬 큼. 우리가 여전히 그 가치가 있다고 믿는다면, 세계는 그 토큰을 만들기 위해 투자할 것. 따라서 세계가 필요로 하는 토큰 생성 능력은 7,000억 달러보다 훨씬 큼

우리가 앞으로도 토큰을 계속 생성하고, 그에 필요한 컴퓨트 용량에 계속 투자할 것이라고 꽤 확신. 근본적으로 모든 기업은 소프트웨어에 의존하고, 모든 소프트웨어는 AI에 의존하게 될 것이기 때문. 그래서 모든 기업이 토큰을 생산할 것. 클라우드 데이터센터 기업이라면, 매출을 위해 토큰을 생성하는 AI 팩토리 갖게 됨

엔터프라이즈 소프트웨어 기업이라면, 자사 도구 위에 올라가는 에이전틱 시스템을 위해 토큰을 생성할 것. 로보틱스 공장이라면—자율주행차가 첫 사례—거대한 슈퍼컴퓨터(사실상 AI 팩토리)를 갖고 토큰을 생성해 차량으로 보내고, 그것이 차량의 AI가 됨. 그리고 차량 내부에도 지속적으로 토큰을 생성할 컴퓨터를 넣어야 함. 따라서 이것이 컴퓨팅의 미래라는 점에 대해 우리는 꽤 확신하고 있음

그렇다면 왜 이것이 컴퓨팅의 미래라고 그렇게 확신하느냐? 과거의 소프트웨어는 '사전 기록(pre-recorded)'이었음. 우리는 모든 것을 사전에 캡처다. 소프트웨어를 미리 컴파일했고, 콘텐츠를 미리 작성했고, 비디오를 미리 녹화. 하지만 이제 모든 것이 실시간으로 생성(generative in real time). 실시간으로 생성될 때는 개인, 상황, 질의, 의도까지 고려해서 결과를 생성할 수 있음. 이것이 우리가 말하는 새로운 소프트웨어, 즉 AI—에이전틱 AI. 따라서 필요한 컴퓨트 양은 사전 기록 방식보다 훨씬 큼. DVD 플레이어(사전 기록된 것)보다 컴퓨터가 훨씬 큰 컴퓨트 능력을 갖듯이, AI는 과거 방식의 소프트웨어보다 훨씬 더 많은 컴퓨트가 필요

이제 컴퓨트 지속성(sustainability)에 대한 첫 번째 레벨의 답은 컴퓨터 과학 관점에서 '컴퓨팅이 이렇게 수행될 것'이라는 것. 산업적 레벨에서는, 기업들은 결국 소프트웨어로 움직이고, 클라우드 기업도 소프트웨어로 움직임. 새 소프트웨어가 토큰 생성을 필요로 하고, 토큰이 수익화(monetized)된다면, 데이터센터 구축은 매출을 직접적으로 견인. 그래서 컴퓨트가 매출을 견인. 그들도 이를 이해하고 있고, 대중도 점점 이해하기 시작했다고 생각

마지막으로, AI가 세상에 제공하는 혜택은 결국 매출을 만들어내야 함. 우리는 바로 눈앞에서 그것이 개발되는 것을 보고 있음. 에이전틱 AI는 변곡점을 넘었고, 정말 지난 2~3개월 사이에 일어났음. 물론 업계 내부에서는 6개월 정도 전부터 보고 있었음. 하지만 이제 전 세계가 에이전틱 AI 변곡점을 인식하고 있음

에이전트는 매우 똑똑하고, 실제 문제를 해결하고 있음. 코딩은 이제 에이전틱 시스템이 지원. 엔비디아의 모든 코더는 Cloud Code나 OpenAI Codex를 매우 많이 사용. 종종 둘 다, 그리고 Cursor까지—유스케이스에 따라 세 가지 모두를 쓰기도. 그들은 문제를 풀기 위한 에이전트와 공동 설계 파트너, 엔지니어링 파트너를 갖고 있음

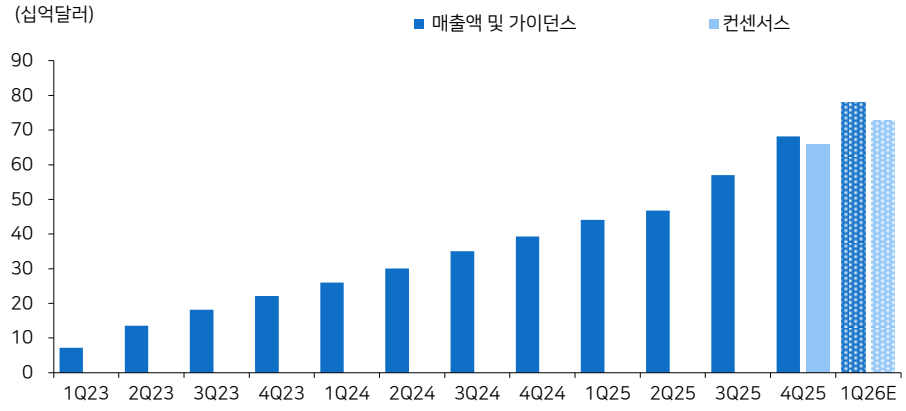
그들의 매출이 폭증하는 것도 볼 수 있음. Anthropic의 경우 매출이 1년 만에 10배가 된 것으로 알고 있고, 수요가 너무 커서 심각한 캐파 제약을 받고 있음. 토큰 수요는 엄청나고, 토큰 생성 속도는 지수적으로 증가

OpenAI도 마찬가지. 수요가 엄청남. 그래서 그들이 더 많은 컴퓨트를 온라인으로 올릴수록(stand online/bring online), 매출은 더 빨리 성장. 이는 '추론이 매출'이고 '컴퓨트가 매출'이라고 말한 것과 연결. 이것이 우리가 이를 새로운 산업혁명이라고 부르는 이유이기도 함

새로운 공장, 새로운 인프라가 건설되고 있고, 이런 새로운 컴퓨팅 방식은 되돌아가지 않을 것. 따라서 토큰 생산이 컴퓨팅의 미래라고 믿는 한, 우리는 이 시점부터 용량을 계속 구축하고, 여기서 더 확장해 나갈 것임

지금 우리가 보고 있는 파도는 에이전틱 AI 변곡점이고, 그 다음 변곡점은 피지컬 AI. 즉, AI와 에이전틱 시스템을 제조, 로보틱스 같은 물리적 애플리케이션으로 가져가는 것. 이것이 앞으로의 거대한 기회

그림1 **엔비디아 매출액 및 가이던스, 컨센서스 비교**



자료: Bloomberg, 메리츠증권 리서치센터

위 내용은 4Q25 엔비디아 실적발표 컨퍼런스 콜 내용을 번역 및 요약한 것임