엔비디아 생태계에 대한 업계·규제적 대응 현황

미래전략개발부 미래전략팀
정 희 채 (jhc0515@kdb.co.kr)

◆ AI 가속기 시장에서 CUDA* 종속 심화로 엔비디아의 시장지배력이 확대됨에 따라, 업계는 자체칩 개발 및 개방형 플랫폼 구축으로 대응하고 규제당국은 반독점 조사 중

* CUDA(Compute Unified Device Architecture): '07년 엔비디아가 발표, GPU 리소스를 사용할 수 있도록 각종 라이브러리 및 개발자 도구가 포함된 API로 기존의 다양한 프로그래밍 언어 지원

◆ 엔비디아의 기술적 우위와 미-중 기술패권의 핵심기업인 점 등 고려시 시장 내 지위는 당분간 유지할 전망이나, 규제당국의 향후 조치에 관심을 가질 필요

□ 엔비디아, CUDA 종속성을 기반으로 견고한 우위를 지속하며 독자적 생태계 확보

○ 연산·추론 성능이 개선된 AI 가속기(GPU) 지속출시 및 전용 CUDA 소프트웨어 업그레이드를 통해 사용자의 전환비용을 높이며 경제적 해자 구축

- 차세대 AI 가속기 '루빈 아키텍쳐'는 3나노 공정과 HBM4를 채택하여 연산성능을 5배 가량 향상시키고, 추론 비용을 절감하며 에너지 효율도 개선하였음
- 자사 제품에 최적화된 CUDA는 출시후 약 20년간 축적된 전용 라이브러리* 및 5백만명 이상의 개발자풀을 바탕으로 연구·산업계의 AI 표준으로 채택

* 복잡한 알고리즘이나 공통기능을 미리 구현해둔 코드의 집합으로 cuDNN(딥러닝) 등이 있음

○ 하드웨어와 소프트웨어의 수직계열화를 바탕으로 데이터센터 사업에서 높은 시장지배력을 유지하며 AMD, Intel 등 경쟁업체와 실적격차를 확대

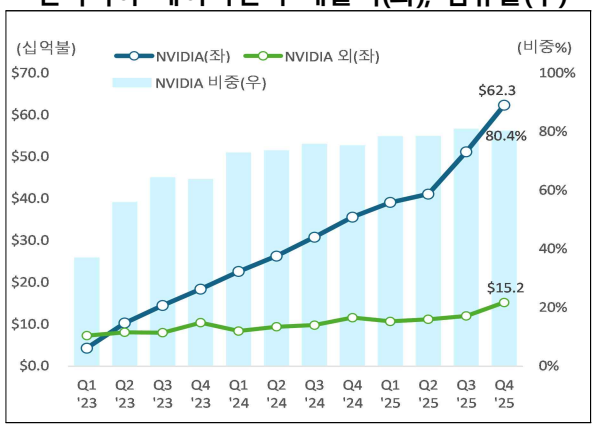
- '25년 매출(2,159억 달러) 중 데이터센터 비중 90%, 시장점유율은 80%를 기록

엔비디아 AI 가속기 세대별 사양 비교

구분	루빈 (Rubin)	블랙웰 (Blackwell)	호퍼 (Hopper)
출시·양산	**'26하**	'24~'25	'22~'23
주요모델	**R100 R200**	B200 B300	H100 H200
연산성능	**50 PFLOPS**	10 PFLOPS	4 PFLOPS
메모리 (대역폭)	**HBM4 (22TB/s)**	HBM3E (8TB/s)	HBM3E (3.4~4.8TB/s)
공 정	**3nm**	4nm	4nm/5nm

자료 : 엔비디아 홈페이지

엔비디아 데이터센터 매출액(좌), 점유율(우)



주 : "Nvidia 외"는 AMD·Intel·IBM 합산 기준
자료 : 각사 IR자료, Yahoo Finance 등

□ 엔비디아 의존성 완화를 위해 업계는 자체칩 개발 및 개방형 플랫폼 구축으로 대응

○ 구글을 비롯한 주요 빅테크 기업들은 자체칩(ASIC*) 개발을 통해 엔비디아의 영향력 확대에 대응하고 데이터센터 원가절감과 추론능력 향상에 주력

* ASIC(Application Specific Integrated Circuit): 범용성에 초점을 맞춘 엔비디아의 GPU와 달리, 특정 앱이나 기능을 위해 제작된 맞춤형 반도체로, 저렴하고 전력소모가 낮은 장점 보유

- 연산성능은 엔비디아 최신 GPU 대비 열위하나, 구글의 Gemini, MS의 Copilot 등 자사의 특정 AI모델 및 플랫폼 구동시 추론비용 절감에 기여

美 빅테크 기업들의 ASIC 및 특징

기 업	ASIC	공정	특 징
Google	TPU v7 (Ironwood)	4nm	Gemini 추론 및 학습에 최적화, 현재 양산 및 클라우드 배포중
Microsoft	Maia 200	3nm	Copilot 및 GPT-5 추론용으로, 데이터센터 구축 중
aws	Trainium 3	3nm	아마존 Bedrock 플랫폼의 LLM 학습
Meta	MTIA 300	3nm	Llama 추론 및 서비스 및 페이스북, 인스타그램 운영

자료 : 기업별 자료 자체 자료 참고 및 산업은행 재구성

○ 폐쇄적인 엔비디아 CUDA가 아닌 오픈소스 생태계 구축을 위해 '23.9월 주요 빅테크 기업 주도의 UXL(Unified Acceleration) 재단 출범

- 삼성전자, SK하이닉스, 마이크론, 퀄컴, 구글 등이 참여하여 표준화된 단일 코드로 GPU 및 AI 가속기에서 작동하는 "oneAPI" 개발
- 성능손실을 최소화하면서 기존 CUDA 기반 코드를 타사 ASIC으로 바꿀 수 있는 코드간 전환성 대폭 개선

□ 美·EU 규제당국은 엔비디아의 칩배정, CUDA 등을 중심으로 反독점 조사 진행중

○ '24.9월 美 법무부(DOJ)는 엔비디아의 반독점 위반 가능성 관련하여 기초조사 착수 후 현재 사업관행 및 기업 인수합병 적정여부 수사를 위해 소환장 발부

- 자사 GPU 구입시 관련 S/W나 장비를 함께 끼워팔거나(번들링), AMD, 인텔 등 경쟁사 칩 도입시 의도적으로 공급을 지연했는지(배타적 거래) 여부 조사
- '24년말 AI 인프라 관련 솔루션을 제공하는 Run:AI社* 인수를 통해 타사 GPU의 성능향상 가능성을 차단하는 등 시장 지배력 행사 여부

* 이스라엘 스타트업으로 GPU 성능 최적화를 통해 AI 가속기 효율을 향상하는 소프트웨어를 개발

- DOJ는 연방 반독점법*을 근거로 과거 구글, 마이크로소프트 등 빅테크기업과 AT&T, 스탠더드오일 등 전통산업 내 독점기업을 제소하고 규제조치를 시행

* 셔먼법(1890) : 기업의 가격담합 등 공모행위 및 시장 지배력을 가진 기업의 사업관행을 조사·규제

美 법무부(DOJ) 주도의 기업 반독점 대표적 규제 사례

구 분	구글(검색엔진)	마이크로소프트(OS/브라우저)	AT&T(통신)
시 기	2020~현재	1998~2001	1974~1982
쟁 점	애플·삼성 등 스마트폰 제조사앞 검색엔진(크롬) 독점 유지여부	윈도우(OS) 점유율을 이용, 인터넷 익스플로러(브라우저) 끼워팔기(tying) 여부	통신망과 장비시장 수직계열화를 통한 경쟁 배제 여부
규 제	스마트폰 기본 검색엔진 진입장벽 구축금지 및 관련계약 무효화	윈도우 OS 소스코드 공개 및 타사 브라우저 설치 방해 금지	반독점법 적용, 전화와 장거리 서비스 결합금지
결 과	기본 검색엔진 계약기간 제한(1년) 및 검색데이터 공개 (現 구글 항소 중)	사업관행 수정 및 API 공개	미국 내 7개 지역 전화 회사로 강제 분할

자료 : 산업은행

○ 유럽연합 집행위원회(EC) 및 EU 주요 회원국 규제당국은 엔비디아의 제품 끼워팔기, CUDA API의 폐쇄성 등 공정경쟁 저해여부 조사 중

- 프랑스는 '24.7월 엔비디아의 반경쟁 사업관행에 대한 공식 기소절차에 돌입하였고, 이탈리아는 Run:AI社 인수에 대해 EC앞 조사('24.9월)* 요청

* EC는 '24.12월 Run:AI社 인수를 최종 승인하였으나, 엔비디아는 이듬해 규제남용을 근거로 EC를 제소

□ 엔비디아 AI 가속기 및 CUDA의 기술적 우위와 전략적 위치 감안시 AI 시장 內 지위는 상당기간 유지될 전망

○ 주요 빅테크 기업들의 ASIC 개발과 개방형 표준플랫폼 구축 노력에도 불구, 엔비디아 생태계 종속성을 단기간 내 해소하는 데 한계 존재

- ASIC는 자사의 일부 특화된 애플리케이션 전용이나, 데이터 클라우드 이용 고객 대부분은 엔비디아 AI 가속기·CUDA 기반의 범용 개발플랫폼을 선호

○ 또한 엔비디아가 미-중 AI 기술패권 경쟁의 핵심기업인 점 감안시, 과거 AT&T의 강제분할 사례처럼 고강도 반독점 규제가 시행될 가능성은 낮음

- 中 화웨이社나 무어스레드社 등은 자체 AI 가속기·API 개발 또는 CUDA와 호환가능한 플랫폼을 구축하며 미국과 기술격차 해소를 위해 추격 중
- 다만, 과거 MS와 구글의 반독점 규제와 유사한 형태로 AI 가속기 공급망 투명성 강화, 차별적 배정금지 등에 이어 중·장기적으로 타사 AI 칩과의 CUDA 호환성 향상 등 점진적인 시장지배력 완화 조치가 시행될 가능성*

* '25.9월 美 법무부 반독점국 차관보(Gale Slater), 국제 반독점법·정책 컨퍼런스에서 AI시장의 경쟁력 확보를 위해 "AI 인프라자원의 배타적 행위 방지" 및 "오픈소스 모델"의 중요성 강조