엔비디아의 토탈 솔루션 전략과 탈 엔비디아 동향

NVIDIA's Total Solution Strategy and the De-NVIDIA Trend

신강선 (K. Shin, sin3401@etri.re.kr) 기술전략연구센터 박사후연수연구원
현종민 (J. Hyun, jmzxc1024@naver.com) 충북대학교 경영정보학과
박정렬 (J. Park, jrpark16@etri.re.kr) 미래전략연구실 연구원

ABSTRACT

As generative AI gained prominence, the global demand for GPUs tailored for this technology surged. NVIDIA, a leading manufacturer of these GPUs, became the first semiconductor company to surpass a market capitalization of $1 trillion. The company implements a total solution strategy that offers everything from dedicated GPU cards for computation to the entire infrastructure required for cloud services. Through this approach, it aims to secure a competitive advantage in the market, effectively countering its competitors, and transforming itself into a comprehensive semiconductor company that not only produces AI chips but also builds and provides the entire computing environment. However, NVIDIA's dominant market share has led to supply shortages and price increases, prompting major tech companies that rely on these products to seek ways to reduce their dependence—a trend referred to as "De-NVIDIA." This study analyzes NVIDIA's total solution strategy, which can be seen as a key growth factor, alongside the global trend of De-NVIDIA, to draw meaningful insights.

KEYWORDS GPU, AI반도체, 생성형 AI, 엔비디아, 토탈 솔루션

Ⅰ. 서론

인공지능(이하 AI) 기술을 활용한 부가가치 창출이 활발해지면서 AI의 성능을 증가시키는 반도체(이하 AI반도체)에 대한 수요와 관심이 증가하고 있다. OMDIA에 따르면 AI반도체 시장은 '21년 368억 달러에서 '27년 1,313억 달러로 연평균 23.6% 성장이 예상된다[1]. AI반도체는 수집 · 가공된 데이터를 학습하고 서비스로 제공하기 위한 핵심 인프라로 AI 확산에 중추적 역할을 담당하고 있다. 특히, ChatGPT와 같은 생성형 AI의 등장으로 그 중요성이 더욱 부각되었다. AI반도체 시장은 크게 데이터센터와 엣지 시장으로 구분되며, 초기에는 스마트폰, 가전 등 소비자 디바이스와 데이터센터 서버 분

* DOI: https://doi.org/10.22648/ETRI.2024.J.390608
* 이 논문은 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음[24ZF1130, ICT 국가기술전략 정책연구].

야에서 주도적 성장을 보였다. 향후에는 자동차와 사물인터넷(IoT) 분야로의 확장이 예상된다.

현재 호황을 맞고 있는 여러 반도체 기업 중 GPU(그래픽 처리 장치) 기반 클라우드 인프라 서비스를 제공하는 NVIDIA(이하 엔비디아)는 '22년 AI반도체 시장의 92%를 점유하였고, 다음해 5월, 애플, 마이크로소프트, 구글, 아마존에 이어 시가총액 1조 달러를 돌파하는 최초의 반도체 기업이 되었다[2]. 이를 가능하게 한 원동력은 엔비디아가 제공하는 '토탈 솔루션'이라고 볼 수 있다. 이는 소프트웨어 개발 툴, 하드웨어, 통신 네트워크 인터페이스, 반도체 칩 간 통신 등 모든 인프라 서비스를 통합 제공하는 방식이다[3]. 엔비디아는 이를 통해 시장에서의 경쟁우위를 확보하고, 경쟁자들을 효과적으로 견제하면서 AI반도체 생산을 넘어 전체 컴퓨팅 환경을 구축 및 제공하는 종합 반도체 기업이 되고자 한다.

그러나 높은 시장 점유율을 차지할 수 있게 하는 엔비디아의 토탈 솔루션 전략 이면에는 제품 생산 및 공급의 한계로 인한 문제점도 드러나고 있다[4]. 일부 파트너사에 대한 원활한 공급이 어려워지고, 제품의 품귀현상이 지속되면서 일부 빅테크 기업은 독자적인 AI반도체 개발에 나서기 시작했다. 이는 '탈(脫) 엔비디아'로 이어지고 있으며, 장기적으로 엔비디아는 경쟁사들의 제품 개발 경쟁과 시장 경쟁력 감소라는 도전에 직면할 것으로 예상된다[5,6].

이러한 현상은 AI반도체 시장에서 엔비디아의 지배적 위치에 대한 산업계의 대응 양상을 보여주는 것이라 인식되며, 이는 시장 구조의 잠재적 변화 가능성을 시사한다. 따라서 이러한 변화가 우리나라의 반도체 산업과 AI 기술 발전에 미칠 수 있는 영향을 고려할 때, 이에 대한 체계적인 조사와 분석이 필요하다.

이러한 배경에서 본고는 AI반도체 시장에서 엔비디아가 지배적 위치를 차지하게 된 핵심 요인인 토탈 솔루션 전략을 살펴보았다. 또한 엔비디아와 수요 기업들이 직면하고 있는 도전과제들을 살펴보고, 이에 따른 탈 엔비디아 동향을 분석하여 결론과 시사점을 제시하였다.

II. 엔비디아의 토탈 솔루션 전략

1. 엔비디아 소개

엔비디아는 '93년 설립된 미국의 반도체 기업으로 초기에는 PC, 노트북, 콘솔 게임기용 GPU 설계에 주력하였다. '04년부터는 GPU의 고성능 병렬 연산 능력이 다양한 분야에 활용될 수 있음을 인지하면서[7], AI 컴퓨팅 학습용 반도체 제조로 사업 영역을 확장하였다[8]. CEO 젠슨황의 뛰어난 리더십과 비전을 바탕으로 AI와 IoT, 자율주행 기술의 핵심인 반도체를 생산하면서 전 세계 AI반도체 시장의 80% 이상을 점유하고 있다[9].

엔비디아가 이처럼 높은 시장 점유율을 확보하고 시가총액 1조 달러를 돌파할 수 있었던 핵심 전략 중 하나는 '토탈 솔루션'이라 볼 수 있다. 이 전략은 고객들이 엔비디아 제품을 지속적으로 사용하도록 유도하는 데 중요한 역할을 하였으며, 결과적으로 엔비디아가 반도체 시장을 석권하게 된 결정적 요인이 되었다[10].

2. 토탈 솔루션 전략

토탈 솔루션은 AI 연산에 최적화된 전체 컴퓨팅 아키텍처를 의미하며, 소프트웨어 개발 툴, 하드웨어, 통신 네트워크 인터페이스 및 반도체 간 통신 등을 포함한다[3].

엔비디아는 이러한 아키텍처를 기반으로 그래픽 처리와 연산을 위한 전용 GPU부터 클라우드 서비스까지 모든 과정을 통합 제공한다[11]. 대표적인

예로, 엔비디아 DGX 클라우드는 AI 모델 구축 및 학습을 위한 컴퓨팅 자원과 소프트웨어 시스템 접근권을 제공한다. 더 나아가, 사전 학습된 모델, 생성형 AI 모델, 맞춤형 대규모 언어 모델의 구축 · 개선 및 운영을 위한 종합적인 솔루션을 제공한다[12].

이러한 토탈 솔루션 공급을 통해 엔비디아는 모든 컴퓨팅 영역에서 사용하기 쉬운 플랫폼을 구축해 고객 가치를 창출하고, 강력한 락인효과를 통해 수익 극대화를 추구하기 위한 전략을 구사하고 있다[9]. 이를 가능케 하는 토탈 솔루션의 핵심 요소는 소프트웨어 개발 플랫폼인 '쿠다', 하드웨어의 범용성을 높이는 'ASSP', 그리고 자체 개발하는 'CPU'라고 볼 수 있다.

가. 소프트웨어 개발 플랫폼(쿠다)

쿠다(CUDA)는 그래픽 처리 장치에서 수행하는 알고리즘을 C++, 파이썬 등 다양한 프로그래밍 언어로 활용할 수 있게 하는 GPGPU 기술이다[10]. 쿠다의 주요 특징은 빠른 명령어 사용, 쿠다 커널 분할, 데이터 배열 단순화, 루프 풀기 및 텍스처 메모리 사용 등이 있다. 이러한 특징들은 GPU 최적화를 위해 다각도로 개발되어 GPU 구현 속도를 높인다.

쿠다는 지속적인 업그레이드를 통해 GPU 코어의 다양한 명령어를 개발자들이 쉽게 사용할 수 있도록 일종의 '번역기' 역할을 수행하면서, 동시에 더 많은 기능을 지원하도록 발전해왔다[3]. 엔비디아는 모든 개발자에게 쿠다를 무료로 공개하면서도 이를 자사 GPU에서만 사용 가능하도록 제한했다. 이를 통해 시장 입지를 강화하고 경쟁사들을 효과적으로 견제하고 있다.

나. GPU 범용성

엔비디아의 GPU를 포함한 서버용 GPU는 ASSP(Application Specific Standard Product)로 분류된다. ASSP는 특정 용도로 설계된 ASIC(Application Specific IC)가 여러 고객에 의해 사용되면서 발전한 형태다. '24년 기준, 엔비디아는 서버용 GPU 시장에서 98%의 점유율로 독점적 지위를 차지하고 있으며, 경쟁사인 AMD와 인텔의 제품도 ASSP 카테고리에 속한다.

ASSP는 ASIC에 비해 최대 AI 연산 성능이 다소 낮고 전력 소모가 클 수 있지만, 다양한 AI 연산에서 안정적인 성능을 발휘할 수 있는 범용성이 큰 장점이다. 반면 ASIC는 CNN(Convolutional Neural Networks) 연산과 같은 특정 목적에 최적화되어 있어 다양한 AI 연산에 활용하기에는 제한적이다. 그러나 엔비디아의 GPU는 GAN(Generative Adversarial Networks), RNN(Recurrent Neural Networks) 등 다양한 인공신경망 연산을 지원하며, 안정적인 성능을 제공한다. 이러한 범용성을 갖춘 GPU를 토탈 솔루션에 포함시킴으로써 엔비디아는 폭넓은 고객층을 확보하고 있다.

다. CPU 개발

AI 연산에 사용되는 데이터는 대체로 크기가 방대한 반면, AI반도체에 탑재할 수 있는 메모리 용량은 상대적으로 매우 작다. 이로 인해 데이터 처리 시 DRAM으로부터 여러 차례 데이터를 불러와야 하는 상황이 발생한다. 그러나 CPU가 지원하는 메모리와 가속기 간 통신 방식인 PCI Express의 데이터 전송 속도가 가속기의 처리 속도를 따라가지 못해 컴퓨터 성능 저하가 발생한다[3]. 이를 개선하기 위해 엔비디아는 타사의 CPU를 사용하지 않고 자사의 GPU에 최적화된 CPU 'Grace'를 직접 개발 및 생산하고 있다. CPU의 직접 개발은 성능 저하 현상을 해결하는 것뿐만 아닌 AI에 최적화된 컴퓨팅 아키텍처를 CPU 단계부터 구축해 이를 사용하는 고객들에게 최상의 AI 컴퓨팅 환경을 모두 제공하겠다는 엔비디아의 토탈 솔루션 전략 중 하나이다[3].

Ⅲ. 엔비디아, 성공 이면의 도전과제

전 세계 AI반도체 시장의 80% 이상을 점유하고 있는 엔비디아는 토탈 솔루션 전략을 통해 급속한 성장을 이루었다. 그러나 이러한 성공 이면에는 양면적인 도전과제가 존재한다. 엔비디아 입장에서는 특정 부문에 대한 높은 의존도로 인한 리스크가 있으며, 제품 사용 고객 입장에서는 고전력 소비 및 높은 비용 등의 문제에 직면해 있다. 이러한 이중적 문제는 엔비디아의 지속 가능한 성장과 고객사들의 효율적인 AI 인프라 구축 모두에 잠재적인 걸림돌로 작용하고 있다.

1. 공급 부족 문제

엔비디아는 주로 반도체 설계에 특화된 기업으로 실제 제품 생산을 위해서는 반도체 위탁 생산(이하 파운드리) 업체와 긴밀히 협력해야 한다. 이와 같은 구조는 전체 공급망에 걸쳐 효율적인 관리와 조정이 필요할 뿐만 아니라 엔비디아를 공급망 문제에 더욱 취약하게 만든다. 즉, 외부 요인으로 인한 생산 차질이나 공급 지연이 엔비디아의 전반적인 사업 운영에 직접적인 영향을 미칠 수 있는 구조인 것이다.

실제로 엔비디아는 자사의 모든 반도체 제조를 대만 TSMC에 과도하게 의존하고 있다[13]. TSMC는 자동차 제조업체부터 스마트폰 제조업체에 이르는 많은 글로벌 고객이 필요로 하는 AI반도체를 생산하는 업체이다. 그러나 TSMC의 생산 능력도 한계가 있기 때문에 향후 AI반도체의 급격한 수요 증가, 대만과 중국 간의 안보 문제 등 글로벌 이슈에 따른 구조적인 공급망 문제가 발생할 경우, 엔비디아의 제품 생산에도 부정적 영향을 미칠 수 있다[13]. 이러한 상황은 엔비디아의 성장 전략에 잠재적인 위험 요소로 작용할 수 있어, 장기적인 안정성과 지속 가능한 성장을 위해 해결해야 할 중요한 과제라 볼 수 있다.

2. 고전력 문제

엔비디아는 아마존, 마이크로소프트와 같은 주요 클라우드 서비스 제공업체에 GPU를 공급하고 있다. 이들 클라우드 업체의 핵심 사업인 데이터센터에서 GPU는 워크로드 가속을 통해 작업처리 속도와 용량을 크게 향상시킨다[14,15]. 코로나19 팬데믹 이후 가속화된 디지털 전환으로 데이터 양이 폭증하면서 데이터센터와 GPU의 중요성은 더욱 커지고 있다.

그러나, 엔비디아의 GPU는 고전력 소비라는 문제를 안고 있다. 예컨대, 엔비디아의 GPU 제품인 'H100'의 최대 소비전력은 700W에 달한다. 이 제품을 연간 61% 사용률로 가동할 경우, 그 전력 소비량은 미국 평균 한 가구(2.51명 기준)가 사용하는 연간 전력량과 비슷하다. 슈나이더 일렉트릭의 최근 보고서에 따르면 '24년 엔비디아 'H100'의 연간 총 전력 소비량은 조지아, 리투아니아, 과테말라와 같은 국가들의 연간 전력 소비량과 비슷한 수준이다[16]. 이러한 고전력 소비 문제는 엔비디아의 비즈니스에 직접적인 위협이 되고 있다.

이러한 고전력 소비 문제의 심각성은 엔비디아 내부에서도 인식되고 있다. '23년 12월에 열린 임원회의에서는 GPU 수요가 높은 데이터센터 관련 문제가 주요 안건으로 다뤄졌는데, 특히 GPU의 고전력 문제로 인해 자사의 매출이 정체될 수 있다는 점이 가장 큰 우려사항으로 지적되었다[17]. 이러한 상황은 엔비디아가 직면한 도전 과제를 명확히 보여주며, 지속 가능한 성장을 위해 에너지 효율성 개선이 시급함을 시사한다.

3. 비용 및 최적화 문제

엔비디아 GPU의 또 다른 문제점은 높은 가격이다. 토탈 솔루션 전략을 통해 확보한 시장 우위로 인해 엔비디아 제품은 주요 빅테크 기업들의 필수 구매 품목이 되었다. 특히, ChatGPT 등장 이후 생성형 AI 기술 경쟁이 가열되면서 엔비디아 GPU에 대한 수요는 더욱 커지고 있다[18]. 이러한 독보적 위치는 엔비디아에게 프리미엄 가격 정책을 가능케 했지만, 동시에 고객사들에게 상당한 비용 부담을 안겨주고 있다.

대표 제품인 'H100'은 개당 가격이 4만 달러를 상회할 뿐만 아니라 제품 공급에도 상당한 시일이 소요되는 실정이다. 이러한 상황에 대해 오픈AI의 샘 올트먼도 엔비디아 제품의 높은 가격과 공급 부족 문제를 지적한 바 있다[19].

또한, 엔비디아 GPU는 구조적 한계도 지니고 있다. 이들 제품은 온전한 AI 딥러닝용 신경망처리장치(NPU)가 아닌, GPU에 ASIC가 더해진 형태다. 이러한 구조는 범용성을 높이는 장점이 있지만, 동시에 무겁고 전력 효율이 상대적으로 떨어진다는 단점도 있다. 로이터 통신에 따르면 '24년 2월 엔비디아는 이를 위기로 인식하고 완제품 판매뿐만 아니라 주문형 칩 사업부를 신설하여 사업 영역을 확장하려 하고 있다[20].

이와 같은 문제들로 인해 많은 빅테크 기업은 자체 AI반도체 개발의 필요성을 절감하고 있으며, 이는 '탈 엔비디아'로 이어지고 있다. 이러한 추세는 엔비디아의 시장 지배력에 대한 잠재적 위협이 될 수 있으며, 장기적으로 AI반도체 시장의 구도 변화를 초래할 가능성이 있다.

Ⅳ. 탈 엔비디아

엔비디아의 GPU는 고성능을 기반으로 '06년 소프트웨어 '쿠다' 출시와 함께 토탈 솔루션을 제공하면서 AI반도체 시장에서 독점에 가까운 영향력을 보이고 있다. 그림 1을 보면, '23년 3월 오픈AI의 GPT-4가 출시된 이후, 엔비디아 제품에 대한 수요가 급증하여 '23년 7월 시가총액 1위를 달성한 것을 확인할 수 있다. 그러나 엔비디아 제품은 공급 부족, 고전력 그리고 비용 및 최적화 문제 등의 한계를 안고 있다. 이로 인해 제품을 사용하는 빅테크 기업들은 엔비디아에 대한 과도한 의존도를 줄이고, 다양한 하드웨어 솔루션을 활용한 기술적 유연성 확보를 위해 노력하고 있다. 또한, 자체 AI반도체 개발을 통해 시장 차별화를 추구하고 있다. 구글의 TPU, 아마존의 트레이니엄 등이 대표적 사례라 볼 수 있다. 이처럼 AI반도체 시장에서 엔비디아의 독점에 대응하는 산업계의 다양한 변화가 관찰되고 있다(표 1 참고).

먼저 연합체 구성이 주목할 만하다. 여기에는 엔

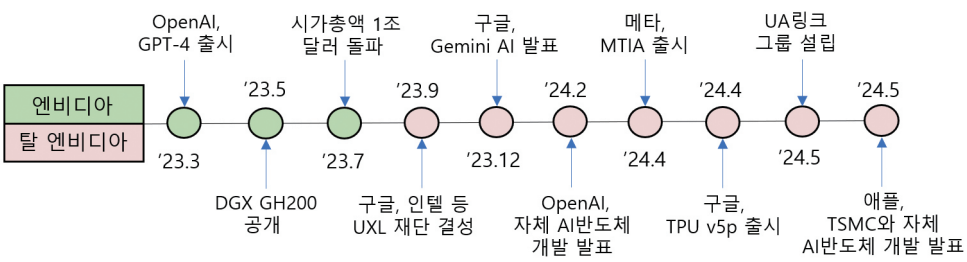

그림 1 AI반도체 생태계 다변화 추이

표 1 주요 기업별 탈 엔비디아 전략

기업명	HW 개발	SW 개발	협력 구축		기업 인수(시도)	대체 기술 연구
			투자	대체품 탐색		
구글	• 자체 AI 전용 반도체 '트릴리움' 출시_'24.05 • UA 링크 프로모터 그룹 연합	• 오픈소스sLM모델 '젬마' 출시_'24.02 • UXL 재단 소속	• 말레이시아데이터센터 건설 등 20억 달러 투자_'24.05	• 데이터 센터용 ARM 기반 CPU '엑시온' 공개_'24.04	• 사이버 보안 스타트업 '위즈' 인수 시도_'24.07	• 양자 컴퓨터 '시커모어' 개발_'19.10
애플	• 추론용 AI반도체 개발('ACDC 프로젝트')_'24.05	• 온디바이스 AI용 오픈소스 공개_'24.06	• 공급망 다각화를 위해 베트남 공급업체 투자 확대 발표_'24.04	• '애플인텔리전스' 구글 TPU 탑재_'24.07	• 온디바이스 AI 구축을 위해 '다윈AI' 인수_'24.03	-
마이크로소프트	• 자체 개발 GPU '마이아 100', CPU '코발트 100'_'23.11 • UA 링크 프로모터 그룹 연합	• SW 개발 환경 'MS 데브 박스' 공개_'22.05	• UAE AI 기업 'G42' 약 15억 달러 투자 발표_'24.04	• '애저 클라우드 컴퓨팅 서비스'에 AMD社 GPU 탑재_'24.05	• 오픈소스 공유 플랫폼 '깃허브' 인수_'18.06	• 양자 슈퍼컴퓨터 개발 로드맵 발표_'23.06
오픈 AI	• AI반도체 사내 전담팀 구성, 브로드컴과 협력 논의_'24.07	• GPT-4o 무료 토큰 제공_'24.08	• 중동지역 투자금 유치_'23.11 • 콘텐츠 기업 '콘데나스트' 파트너십 체결_'24.08	• MS와 '스타게이트 프로젝트' 여러 AI반도체 탑재_'24.03	• SW개발 업무 플랫폼 스타트업 '멀티' 인수_'24.06	-
아마존	• 자체 개발 AI반도체 '그래비톤 4', '트레이니움 2' 공개_'23.11	• 'Amazon Q Developer' 서비스 출시_'24.08	• AI 스타트업 '엔트로픽' 총 40억 달러 투자_'24.04	• '엔트로픽' 모델 구축에 '트레이니움', '인퍼런시아' 사용_'24.04	-	• 양자컴퓨팅 인프라 '브라켓' 제공_'20.08
메타	• 자체 개발 AI반도체 'MSVP', 'MITA'_'24.04 • UA 링크 프로모터 그룹 연합	• AI 모델 '라마 3.1' 오픈소스 공개_'24.07	• AI 분야 400억 달러 투자 발표_'24.03	• 엔비디아 제품 외 25만 개 AI반도체 확보 발표_'24.04	-	-
네이버	-	• vLLM v1 오픈소스 공개 예정_'24.04	• 프랑스 AI 유니콘 기업 '미스트랄 AI' 투자_'24.07	• '네이버 플레이스' 인텔 CPU로 변경_'23.10	• 네이버웹툰, 컴퓨터 비전 AI 스타트업 'V.DO' 인수_'20.01	-
인텔	• AI반도체 '가우디 3' 개발 및 서버용 CPU '제온 6' 공개_'24.04 • UA 링크 프로모터 그룹 연합	• 온디바이스 'AI PC 가속 프로그램' 발표_'23.10 • UXL 재단 소속	• '뉴 올버니 프로젝트' 포함 5년간 1,000억 달러 자국 투자 발표_'24.03	• 네이버, KAIST와 공동연구센터 설립_'24.03	• 이스라엘 파운드리 '타워세미컨덕터' 인수 시도_'23.08	• 양자컴퓨팅 기술 '터널 폴스' 공개_'23.09 • 뉴로모픽 시스템 '할라 포인트' 발표_'24.04
AMD	• AI반도체 'MI 시리즈' 발표_'24.06 • 2세대 버설 AI 엣지 시리즈 제품 출시_'24.04 • UA 링크 프로모터 그룹 연합	• AI용 오픈소스소프트웨어 'ROCm' 운영	• AI 스타트업 '코히어' 투자 참여	• MS, 메타 등 빅테크 기업 포함 100곳 이상 고객사 확보_'24.06	• 오픈소스 SW 스타트업 '노드닷AI' 인수_'23.10 • AI 스타트업 '사일로AI' 인수_'24.07 • 서버 제조업체 'ZT시스템' 인수_'24.08	• SIMD 컨트롤러 기반 양자컴퓨팅 기술 특허 공개_'21.09
삼성전자	• AI반도체 '마하-1' 개발 및 '25년 초 공개 예정 발표_'24.02	• 오픈 소스 전략 회사 OSPO 운영 • UXL 재단 소속	• AI반도체 스타트업 '그로크' 투자 참여_'24.08	• 네이버와 LLM 용 ASIC 개발 협력_'23.12	• 지식 그래프 기술 전문 기업 '옥스퍼드 시멘틱 테크놀로지' 인수_'24.07	• 차세대 메모리 대안 'CXL' 양산 발표_'24.07 • 뉴로모픽 연구 및 개발

비디아 토탈 솔루션의 핵심이라 볼 수 있는 소프트웨어 개발 플랫폼 '쿠다'에 대한 대응으로 인텔, 구글, 퀄컴 그리고 삼성이 주축이 된 'UXL 재단'이 있다. '23년 9월에 출범한 이 연합은 회원 기업과 외부 기관으로부터 기술을 기부받기 시작하여, '24년 3월 엔비디아 제품 외의 다양한 하드웨어와 소프트웨어에서 AI 개발을 가능케 하는 오픈 소스 프로젝트를 진행하고 있다[21].

또한, 엔비디아가 GPU와 CPU 간 데이터 전송을 최적화하는 NV링크 기술로 시장을 주도하는 가운데, 이에 대응하기 위해 '24년 5월 AMD, 구글, 마이크로소프트 및 인텔 등 주요 기업들이 '울트라 가속기(UA) 링크 프로모터 그룹'을 결성했다. 이 연합체는 AI 가속기 연결을 위해 개방형 표준을 개발함으로써 NV링크에 대한 대안을 제시하고, 시장 경쟁력을 확보하고자 한다.

한편, 주요 기업들 중에서는 구글과 애플이 가장 두드러진 움직임을 보이고 있다. 구글은 '24년 4월 '구글 클라우드 넥스트 2024'를 개최하여 생성형 AI 기반의 혁신적인 AI 생태계 비전을 제시하며, 다양한 신제품과 서비스를 공개하였다. 특히 데이터센터용 ARM 기반 CPU인 '엑시온'과 자체 AI반도체 'TPU v5p'가 주목받았다. 또한, '24년 5월 '2024 Google I/O'를 통해 6세대 TPU인 '트릴리움(Trillium)'을 공개하며 자체 하드웨어 제품 개발을 통한 탈 엔비디아 전략을 강화하였다. TPU는 구글이 개발한 ASIC로서 TensorFlow를 통해 머신러닝 워크로드를 처리하는 AI반도체이다[22]. 또한, 구글은 자사의 생성형 AI '제미나이' 개발에 사용된 핵심 기술과 연구를 기반으로 제작된 오픈소스 모델 '젬마'를 공개하며 AI 개발 생태계에서의 영향력 확대를 꾀하고 있다[23].

애플은 '24년 5월 반도체 생산 기업 TSMC와 협업하여 엔비디아의 GPU를 대체할 추론용 AI 개발 프로젝트 'ACDC'를 추진 중이다. 또한 엔비디아 의존도 감소를 위해 '애플 인텔리전스 파운데이션 언어 모델(AFM)'에 구글의 TPU를 적용했다[24]. 그리고 온디바이스 AI를 개발하기 위해 '24년 3월 '다윈 AI'를 인수하고, 같은 해 6월 온디바이스 AI용 오픈소스를 공개하는 등 적극적인 행보를 보이고 있다[25,26].

마이크로소프트는 엔비디아 의존도 감소를 위한 다각적 전략을 펼치고 있다. '23년 11월 자체 AI반도체 '마이아(Maia)100'과 고성능 컴퓨팅 CPU '코발트(Cobalt)100'을 공개하였다. '24년 5월에는 미국 반도체 회사인 AMD의 GPU를 탑재한 '애저 클라우드 컴퓨팅 서비스' 상용화 계획을 발표하였다[27]. 더불어 '24년 4월 UAE의 AI기업 'G42'에 약 15억 달러를 투자하며 중동 시장 진출을 모색하고 있다[28]. 또한, '23년 6월에는 양자 슈퍼컴퓨터 개발 로드맵을 발표하였는데, 이는 대체 기술 개발에 대한 마이크로소프트의 의지로 볼 수 있다[29].

한편, 네이버는 실용적 접근으로 '탈 엔비디아' 행보를 보이고 있다. '네이버 플레이스' 서비스의 AI 추론용 서버를 엔비디아 제품에서 인텔 CPU로 전환했으며, 인텔과의 협업을 통한 실제 서비스 환경 실험 후, vLLM 버전 1의 오픈소스 공개 계획을 발표하였다[30].

AI반도체를 주력 사업으로 하는 엔비디아의 경쟁업체들 또한 '탈 엔비디아'에 가세하고 있다. 이들은 주로 혁신적인 신제품 개발에 주력하여 빅테크 기업들에게 엔비디아 제품의 대안을 제시하고 있으며, 자사의 경쟁력 강화를 위한 다양한 전략을 추진하여 시장 점유율 확대를 꾀하고 있다.

인텔은 '인텔 비전 2024'에서 네이버를 핵심 제휴 파트너로 소개하며 자체 개발한 가우디 소프트웨어 생태계의 확장 전략을 발표하였다[31]. 또한 카이스트와 '공동 연구센터'를 설립하여 AI반도체 생

태계에서의 경쟁력 강화에 주력하고 있다[32]. 그리고 '23년 9월, 양자 컴퓨팅 기술 '터널 폴스'와 '24년 4월, 뉴로모픽 시스템 '할라 포인트'를 공개하는 등 대체 기술개발에도 박차를 가하고 있다[33,34].

AMD는 기업 인수를 통한 역량 강화에 주력하고 있다. '23년 10월 오픈소스 소프트웨어 스타트업 '노드닷 AI', '24년 7월 핀란드 AI 스타트업 '사일로 AI'를 인수하였다. 그리고 '24년 8월 서버 제조업체 'ZT 시스템'을 인수하며 기술력과 생산 능력을 확대하고 있다[35]. 또한, 'MI 시리즈'와 '2세대 버설 AI 엣지 시리즈' 등 자체 AI반도체를 개발 · 출시하고 있으며[36], AI 오픈소스 소프트웨어 'ROCm' 운영을 통해 엔비디아의 '쿠다'에 대응하는 소프트웨어 생태계 구축에도 힘쓰고 있다[37].

삼성전자는 자체 AI반도체 개발과 오픈소스 전략을 중심으로 대응하고 있다. '24년 2월 추론에 특화된 AI반도체 '마하-1'의 개발을 발표하며 엔비디아 의존도 감소 방안을 제시하였다[38]. 또한, 오픈 소스를 AI 시장 주도권 확보를 위한 핵심 전략으로 인식하고 'OSPO' 전담 조직을 운영하여 '21년 ISO/IEC 국제 표준 인증을 획득하였다[39]. 더불어 삼성 첨단기술연구소에서는 차세대 AI반도체로 주목받는 뉴로모픽 연구를 진행 중이며, 해당 분야의 권위자인 함돈희 하버드대학교 교수를 연구소 부원장으로 영입하는 등 기술 경쟁력 강화에 박차를 가하고 있다[40].

Ⅴ. 결론 및 시사점

엔비디아는 소프트웨어 개발 도구, 하드웨어, 통신 네트워크 인터페이스, 반도체 칩 간의 통신 등 모든 인프라 서비스를 통합한 토탈 솔루션 제공을 통해 시장에서의 경쟁우위를 확보하고, AI반도체 생산을 넘어 전체 컴퓨팅 환경을 구축 및 제공하는 종합 반도체 기업으로 도약을 꾀하고 있다.

그러나 엔비디아 제품을 사용하는 기업들은 여러 가지 어려움에 직면하고 있다. GPU의 높은 가격으로 인한 재정적 부담, 제품 공급의 불안정성, 그리고 높은 소비전력 문제가 주요 원인이다. 이러한 문제들이 지속되면서 기업들은 새로운 대안을 모색하기 시작했고, 이는 '탈 엔비디아'로 이어지고 있다.

탈 엔비디아를 보이는 기업들의 활동을 분석한 결과, 크게 다섯 가지로 구분된다. ① GPU를 대체하는 기업 맞춤형 NPU 개발, ② AI를 개발하는 사람들의 필수 소프트웨어인 쿠다를 대체할 오픈 소스 소프트웨어(OSS) 개발, ③ 기업 투자와 대체 공급업체 모색을 통한 협력 구축, ④ 반도체 스타트업 기업 인수, ⑤ 뉴로모픽, 양자컴퓨팅 등 대체기술 연구로 구분할 수 있다. 본고에서는 이러한 분석 결과를 바탕으로 국내 AI반도체 산업과 관련 기업들을 위한 시사점을 제시하였다.

첫째, 엔비디아의 토탈 솔루션 전략에서 핵심은 세 가지로 요약된다. ① 다양한 프로그래밍 언어를 지원하는 소프트웨어 '쿠다'를 통한 개발자 친화적 환경 구축, ② 다양한 AI 연산 환경에 적용 가능한 GPU의 범용성 확보, ③ GPU에 최적화된 자체 CPU 개발을 통한 종합적 AI 컴퓨팅 환경 제공이다. 국내 기업들도 이러한 사용자 중심의 제품 생태계를 구축하기 위해 차별화된 연구 · 개발 전략이 필요하다.

둘째, 엔비디아는 토탈 솔루션 공급 과정에서 다양한 문제가 발생하고 있다. ① 엔비디아 주요 제품인 GPU 생산을 위한 웨이퍼, 실리콘, 일부 기판 및 부품 등의 공급 부족, ② 빅테크 기업들의 자체 AI반도체 개발로 인한 시장 경쟁 심화, ③ TSMC에 대한 과도한 의존으로 인한 잠재적 공급 제약이라는 문제점을 가지고 있다. 국내 기업들도 공급사슬상에서 발생할 수 있는 문제점들을 효과적으로 관리

하기 위한 방안을 강구할 필요가 있다. 국내 팹리스 기업들의 경우 파운드리를 특정 기업에게 집중하면 엔비디아와 같은 공급망 약점이 발생할 수 있기 때문에 공급망 안정성 확보도 향후 중요한 과제이다. 이를 대비하기 위해서는 공급사슬 전반에 걸친 효율적인 관리와 조율이 필요하며, 다양한 파운드리 기업과의 협력적 관계를 구축할 필요가 있다.

셋째, 글로벌 기업들의 '탈 엔비디아' 전략을 참고하여 국내 AI반도체 기업들도 다양한 반도체 기술과 기업에 투자해 해외 제품의 의존도를 줄일 수 있는 생태계 마련이 필요하다. 특히 국내 기업들의 강점인 메모리 반도체를 바탕으로 PIM 인공지능 반도체(차세대 AI반도체) 경쟁력을 확보하고, 이를 바탕으로 글로벌 AI반도체 시장에서의 경쟁력을 확보하는 것이 중요하다. 또한 GPU, NPU 이후의 신기술 확보를 위해 뉴로모픽 반도체에 대한 체계적이고 지속적인 투자도 함께 진행되어야 할 것으로 보인다.

끝으로 정부는 AI반도체 산업 육성을 위해 관련 기업 지원, 협력 강화 그리고 전문인력 양성을 위한 노력을 기울일 필요가 있다. AI반도체 초격차 기술 선점을 위해 미국의 '국가 반도체 기술센터(NSTC)'와 같은 민 · 관 합동 연구센터 설립을 고려할 수 있다. 또한, 글로벌 반도체 인력 부족 문제 해결을 위해 고급 인력뿐만 아니라 초급 인력 양성에도 관심을 기울여야 하며, 각국의 경쟁적 지원에 대응하여 반도체 공장(팹) 건설 보조금 및 다양한 금융 · 세제 혜택 제공을 검토해야 할 것이다.

AI반도체는 글로벌 디지털 혁신을 주도하는 핵심 기술로, 국제 사회에서 그 중요성이 날로 증대되고 있다. 이러한 국제적 흐름에 발 맞추어, 우리나라도 산 · 학 · 연 협력을 통해 이러한 비전을 실현하기 위한 구체적인 전략과 정책을 수립하고 실행해 나가는 것이 앞으로의 과제일 것이다.

용어해설

AI반도체 데이터 연산의 성능, 비용, 전력 소모 등을 최적화해 인공지능 용도로 특화한 시스템 반도체

NV Link 엔비디아의 고속 GPU 상호 연결 기술로, 다중 GPU 시스템에서 GPU 간 데이터 전송 속도를 높이는 기술

DXG Cloud 엔비디아의 AI 소프트웨어와 DXG AI 슈퍼컴퓨팅 전용 클러스터를 결합하여 제공하는 플랫폼 서비스

약어 정리

AP	Application Processor
ASIC	Application Specific Integrated Circuit
ASSP	Application Specific Standard Product
CNN	Convolutional Neural Networks
GPU	Graphics Processing Unit
OSPO	Open Source Program Office
RNN	Recurrent Neural Network
TPU	Tensor Processing Unit

참고문헌

[1] Gartner, "Forecast: AI Semiconductors, Worldwide," Gartner, 2023 2Q.

[2] 박찬, "엔비디아, 시가총액 '1조 달러' 돌파…반도체 기업 최초," AI타임스, 2024. 10. 16.

[3] 신창환, 김영우, "AI반도체 생태계분석," Digital Insight, 한국지능정보사회진흥원, 2022.

[4] 박찬, "엔비디아, 연내 'H100' GPU 55만 개 공급 예정," AI타임스, 2023. 8. 16.

[5] 박두호, "글로벌 빅테크, 자체 'AI 반도체' 개발 경쟁 치열," 전자신문, 2024. 2. 4.

[6] C. Metz, K, Weise, and M. Isaan, "Nvidia's Big Tech Rivals Put Their Own A.I. Chips on the Table," The New York Times, 2024. 1. 29.

[7] 조명현, "'소비자가 쓰기 쉬운 신기술'에 초점 선두업체 밀어낸 비결은 新생태계 구축," DBR, 2020. 12.

[8] 위키백과. https://ko.wikipedia.org/

[9] 최진석, "엔비디아 '폭발적 성장'…젠슨 황 '기술 중심' 리더십 빛났다," 한경, 2023. 8. 24.

[10] CUDA, 위키백과. https://ko.wikipedia.org/wiki/CUDA

[11] NVIDIA. https://www.nvidia.com/

[12] T.W. Kim, "엔비디아는 어떻게 성장했는가: How They Grow 시리즈," Brunch story, 2023. 11. 13.

[13] 준윤, "[기고] 엔비디아의 최대 약점은 '대만 의존도 100%'…공급망 다변화도 어려워," 한국무역협회, 2023. 9. 4.

[14] https://aws.amazon.com/ko/what-is/data-center/

[15] Intel, "데이터 센터 GPU가 혁신에 필수인 이유," https://www.intel.co.kr/content/www/kr/ko/products/docs/discrete-gpus/data-center-gpu/what-is-data-center-gpu.html

[16] 박찬, "GPU 연간 소비전력량이 소규모 국가 수준에 근접," AI타임스, 2023. 12. 27.

[17] 오로라, "'세계 시총 1위' 찍은 엔비디아, '시스코 리스크' 없을까," 조선경제, 2024. 6. 19.

[18] 최진석, "엔비디아 'H100' 뭐길래…대당 6000만 원까지 치솟았다," 한경, 2023. 4. 20.

[19] 이인준, "오픈AI, 반도체 직접 만들려는 이유…"비싸도 너무 비싸, 수천만원 웃돈"," BLOXKMEDIA, 2024. 1. 24.

[20] 최창원, "달라진 반도체 환경…주목해볼 3가지 변화 [COVER STORY]," 매경ECONOMY, 2024. 2. 29.

[21] 김정아, "인텔·구글·퀄컴, '反엔비디아' AI 오픈소스 SW 프로젝트 추진," 한경, 2024. 3. 25.

[22] 아민 바닷, "[I/O 2024] 구글 클라우드 TPU 6세대, 트릴리움(Trillium)을 소개합니다," 구글 코리아 블로그, 2024. 5. 14.

[23] 박찬, "구글, 온디바이스 AI용 오픈 소스 sLM '젬마' 출시," AI타임스, 2024. 2. 22.

[24] 강광우, "[팩플] 엔비디아 AI 칩 시장 독점 균열 생기나…구글 칩 택한 애플," 중앙일보, 2024. 7. 30.

[25] 최진석, "생성AI 뒤쳐진 애플의 반격…"캐나다 스타트업 '다윈AI' 인수"," 한경, 2024. 3. 15.

[26] 박찬, "애플, 온디바이스 AI용 모델·데이터셋 오픈 소스로 대거 출시," AI타임스, 2024. 6. 18.

[27] 우예진, "마이크로소프트, AI 칩 '코발트 100' 다음 주 출시…탈 엔비디아 가속," 베타뉴스, 2024. 5. 19.

[28] S. Sharma, "Microsoft expands UAE presence, inks $1.5B deal with AI giant G42," VentureBeat, 2024. 4. 16.

[29] 김태종, "양자컴퓨터 개발 경쟁 '삼국지'…IBM·구글 주도에 MS 도전장," 매일경제, 2023. 6. 23.

[30] 석대건, "네이버 "올해 내 생성형AI vLLM 오픈소스 공개"," Digital Today, 2024. 6. 6.

[31] R.r. Oh, S.-h. Ahn, and M.-G. Kim, "Intel teams up with Naver, challenging Nvidia's dominance," The Chosun Daily, 2024. 2. 12.

[32] 전미준, "네이버-KAIST-인텔, 인공지능 반도체 신 생태계 조성…'NAVER-Intel-KAIST AI 공동연구센터' 설립," 인공지능신문, 2024. 4. 30.

[33] S. Shankland, "Intel Plans a Quantum Computing Approach to Leapfrog Rivals," CNET, 2023. 9. 21.

[34] Intel, "인텔, 세계 최대규모 뉴로모픽 시스템 공개." https://www.intel.co.kr/content/www/kr/ko/newsroom/news/intel-builds-worlds-largest-neuromorphic-system.html

[35] 오로라, "엔비디아 정조준한 AMD, 서버 제조업체 ZT시스템 인수," 조선일보, 2024. 8. 20.

[36] AI리포터, "AMD, 새로운 AI가속기 'MI325X' 공개…용량·속도 개선 '눈길'," Digital Today, 2024. 6. 4.

[37] ADM, "AI용 AMD ROCm™ 소프트웨어," https://www.amd.com/ko/products/software/rocm/ai.html

[38] 배태용, "엔비디아 독주 끝나나…AMD 선택한 MS·삼성-네이버 연합 [소부장반차장]," 디지털데일리, 2024. 5. 20.

[39] 양대규, "[현장] 박수홍 삼성 오픈소스 그룹장 "오픈소스, 기업 경쟁력의 핵심"," SMARTTODAY, 2023. 11. 14.

[40] 임유진, "삼성도 찜했다…'인간 뇌 닮은꼴' 뉴로모픽 반도체," 뉴스토마토, 2024. 5. 27.