클라우드와 엣지 시대의 AI 반도체 혁신과 지속가능성

서은정
삼성전자 마켓인텔리전스 그룹, Senior professional
e0ej.seo@samsung.com

AI 시대의 도래와 컴퓨팅 환경의 확장

인공지능(AI) 시대의 도래와 함께 AI 반도체는 디지털 인프라의 핵심으로 부상했다. AI 반도체의 정의는 조사기관마다 다소 차이가 있으나, 일반적으로 "AI 연산을 지원하는 훈련 및 추론에 사용되는 AI Processor 또는 AI Accelerator"로 규정된다 (김민식·오정숙, 2024). Gartner(2024)에 따르면 AI 반도체가 전체 반도체 시장에서 차지하는 비중은 2023년 10.1%에서 2028년 20.4%까지 확대될 것으로 전망된다. 반도체 시장에서 AI 반도체의 중요도는 갈수록 커지고 있다.

AI 산업의 발전은 하드웨어 자체의 향상뿐만 아니라 소프트웨어 생태계와의 동반 발전이 필수적이다(Allan, 2025). AI 칩 설계와 제조 기술의 진보는 새로운 소프트웨어 플랫폼과 알고리즘 발전을 이끌고, 반대로 혁신적인 AI 소프트웨어와 알고리즘의 등장은 새로운 반도체 아키텍처를 요구하며 하드웨어 발전을 견인

한다. 이러한 하드웨어-소프트웨어의 상호 작용적 진화는 클라우드 데이터센터부터 엣지 디바이스에 이르는 컴퓨팅 환경 전반에서 관찰된다. 예를 들어, 데이터센터에서는 NVIDIA GPU, Google TPU 등 전문 AI 가속기가 초대형 AI 워크로드를 처리하는 중심이 되며, 한편 엣지에서는 IoT, 자동차, 스마트폰 등의 NPU가 실시간 AI 연산을 가능케 한다.

본고에서는 클라우드 AI와 엣지 AI의 기능적 차이 및 특징을 정리하고, AI 반도체 산업의 최근 동향을 살펴본다. 또한 전력 효율과 지속 가능성 확보를 위한 발전 방향을 함께 정리하며, 클라우드 AI와 엣지 AI의 발전 과정과 미래 전망을 개관한다.

클라우드와 엣지: 대조적인 요구, 상호보완적 역할

클라우드(데이터센터) 환경과 엣지(말단 기기) 환경은 AI 활용 측면에서 서로 대조적인 요구 사항을 갖지만, 실무적으로는 상호 보완적으로 작동하여 전체 AI 생태계 발전을 이끌고 있다. 클라우드는 막대한 연산 자원을 바탕으로 초거대 AI 모델의 훈련과 대규모 추론을 담당하고, 엣지는 현장에서의 실시간 추론과 저전력 처리를 담당함으로써 두 환경 간에 효과적인 업무 분담이 이루어진다(KISDI, 2024). 실제로 엣지 AI는 데이터가 클라우드로 왕복 전달되며 발생하는 지연을 제거하지만, 그렇다고 클라우드의 역할이 사라지는 것은 아니다. 궁극적으로 엣지 AI와 클라우드 AI는 서로를 보완하며 함께 활용되는 기술로, 결합을 통해 데이터에 대한 더 높은 수준의 통찰을 제공한다(Statista, 2024). 이러한 분업 구조를 통해 양측은 각자의 강점을 극대화하면서 긴밀히 연계되어, 전체 AI 서비스의 효율과 사용자 경험을 향상시키고 있다.

예컨대 사용자 스마트폰이나 IoT 기기에서 AI 모델을 바로 현장에서 동작시키면 네트워크 지연과 프라이버시 문제를 크게 줄일 수 있다. 대형 모델 중심의 클라우드 AI 방식에서 한계점으로 지적되어 온 처리 지연과 개인정보 보안 문제가 엣지 AI를 통해 개선되고 있으며, 이러한 온-디바이스 AI 구현은 지연 없는 즉각적 응답과 데이터의 현장 처리를 가능케 하여 프라이버시를 보호한다는 이점이 있다(Google Cloud, 2023). 그 결과 자율주행차, 스마트폰 음성비서 등 실시간성이 중요한 다양한 분야에서 엣지 측 AI 처리가 적극 도입되고 있다. 요컨대, 클라우드와 엣지의 상호보완적 요구는 AI 칩과 AI 소프트웨어 양측면에서 전례 없는 혁신을 동시 촉발하고 있다고 볼 수 있다.

클라우드 AI의 역할과 초대형 연산 인프라

AI 데이터센터는 수천에서 수만 개에 이르는 병렬 컴퓨팅 노드로 구성된 인프라로, 복잡한 딥러닝 모델의

훈련과 추론을 수행하는 중추적인 역할을 한다. 수천, 수만 개의 GPU/TPU 등이 병렬로 연결된 초대형 데이터센터는 거대한 딥러닝 모델을 훈련시키고 서비스하는 중심지이다. NVIDIA H100이나 H200, Google TPU, AWS Trainium 같은 AI 가속기는 수천 개 이상이 병렬로 연결되어 LLM(Large Language Model) 학습에 최적화된다. 이러한 고성능 하드웨어를 효율적으로 활용하기 위해, CUDA, TensorRT, XLA와 같은 전용 소프트웨어 스택이 개발되어 각각 GPU와 TPU의 특수 기능을 최대한 끌어낸다. 예컨대 구글 TPU용 XLA 컴파일러는 TPU 하드웨어 구조에 맞춰 연산을 최적화하여 높은 성능을 이끌어 낸다(Nayak, 2025). 이처럼 클라우드 AI 인프라는 하드웨어와 소프트웨어의 정교한 결합으로 초대형 모델의 훈련과 추론을 가능하게 하여, 혁신적인 AI 서비스들의 탄생을 뒷받침하고 있다.

클라우드 사업자의 맞춤형 AI 칩 부상

AI 수요가 폭증하고 그 중요성이 커지면서 주요 클라우드 사업자들은 자체적인 맞춤형 AI 반도체 개발에 박차를 가하고 있다. 구글은 'Cloud Next 2025' 행사에서 7세대 TPU(Tensor Processing Unit) '아이언우드(Ironwood)'를 공개했다(Martin, 2025). 아이언우드는 이전 세대 TPU(v5p) 대비 10배 이상 성능이 향상된 고성능 칩이다. 고객 수요에 따라 256개 칩 구성부터 최대 9,216개 칩을 연결한 TPU 팟(Pod) 형태로 제공되며, 최상위 구성은 42.5 엑사플롭스(ExaFLOPS, 초당 10^18회 연산) 이상의 성능을 발휘한다. 아마존 웹서비스(AWS) 역시 Inferentia 및 Trainium 등 전용 칩을 개발하여 클라우드에 도입했다. 예를 들어 2세대 Trainium 칩을 탑재한 AWS Trn2 인스턴스는 16개 칩으로 20.8 페타플롭스(PetaFLOPS) 규모의 연산 성능을 제공하며, 동급 GPU 기반 인스턴스 대비 30~40% 높은 가격 대비 성능을 구현한다고 보고된다. 이처럼 클라우드용 맞춤형 AI 칩들은 특정 워크로드에 특화된 설계를 통해 범용 GPU 대비 뛰어난 성능/와트 및 비용 효율을 목표로 하며, 클라우드 소프트웨어와 긴밀히 통합되어 최적의 AI 서비스를 지원하고자 한다.

추론 수요 폭증과 '인퍼런스 시대'의 도래

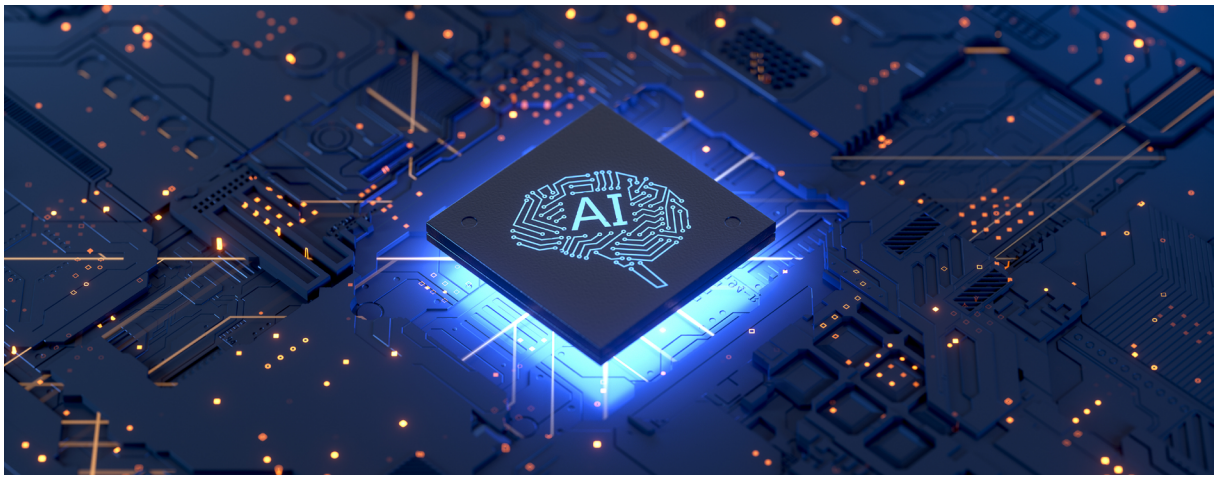
최근 AI 산업에서는 모델 훈련에서 추론 활용 중심으로 관심이 빠르게 이동하고 있다(O'Shea, 2025). ChatGPT와 같은 생성형 AI 서비스, 다양한 AI 애플리케이션과 AI 에이전트의 확산으로, 학습이 완료된 모델을 실제 서비스에 투입해 실시간으로 추론하는 작업량이 폭발적으로 늘어났기 때문이다. 오픈 AI의 샘 알트만(OpenAI CEO)은 이미지 생성 기능 추가 후 GPU 사용량 폭증을 두고 "GPU가 녹아내리고 있다"고 표현했는데, 이는 대중화된 AI 서비스의 막대한 추론 부하를 단적으로 보여준다.

이 전환을 가능하게 한 핵심은 추론 성능 향상과 비용 절감이다. 대표적으로, 엔비디아(NVIDIA)는 블랙웰 GPU가 기존의 Hopper 아키텍처 대비 4배 높은 추론 성능을 제공한다고 발표했으며, MLPerf 벤치마크에서도 성능 한계를 확장했다. 하드웨어의 성능 향상뿐만 아니라 소프트웨어 최적화를 통해 Hopper 플랫폼은 이전 세대 대비 추론 에너지 효율을 최대 15배 개선했다(O'Shea, 2025).

2025년 현재 업계는 AI 모델을 실제 제품화하고 대규모로 배치하는 단계에 진입했다. 이는 모델 개발 경쟁을 넘어, 이제는 얼마나 효율적으로 많은 사용자의 AI 요청을 처리하여 실질적인 가치를 창출하느냐가 관건인 '인퍼런스 시대'의 도래를 보여준다. 이에 따라 토큰당 비용 절감과 에너지 효율 향상이 AI 비즈니스의 핵심 과제로 부상했다.

엣지 AI의 역할과 실시간 처리

클라우드 AI와 대조적으로, 엣지 AI는 데이터 발생 현장에서 실시간으로 AI 연산을 수행하는 패러다임으로, 지연 감소와 프라이버시 보호를 위해 클라우드에 의존하지 않고 스마트폰, IoT 센서, 자동차 등 디바이스 자체에서 AI 모델을 실행한다. 이는 인터넷 연결 여부와 상관없이 밀리초 단위로 데이터 처리가 이루어져 지연을 최소화하며, 민감한 데이터가 밖으로 나가지 않으므로 프라이버시 보호에도 유리하다. 실제 온-디바이스 AI를 활용하면 저지연, 향상된 보안, 에너지 절감, 네트워크 독립성 등 다양한 이점을 얻을 수 있으며, 디바이스가 네트워크에 연결되지 않은 상황에서도 AI 기능을 활용할 수 있다(Gill 외, 2024). 특히 실시간성이 중요한 분야(자율주행, 제조 공정 등)에서는 클라우드 왕복 지연을 없애주는 엣지 AI가 필수적이다. 또한 대량의 센서 데이터 모두를 클라우드로 보내지 않고 현장에서 부분 처리하면 불필요한 대역폭 소모를 줄이고 비용을 절감할 수도 있다(Wang 외, 2025).



엣지 AI 반도체, 산업 전반으로 확산

스마트폰 산업은 엣지 AI 경쟁이 가장 빠르게 일어나는 분야이다. 삼성 엑시노스 2500이 3nm GAA 공정과 24K MAC NPU로 59 TOPS 성능을 달성, 전작 대비 39% 향상된 AI 처리 능력을 제공한다(SamMobile, 2025). 퀄컴 스냅드래곤 8 엘리트는 Hexagon NPU 성능을 45% 높이고 GPU 성능과 효율을 각 40% 개선했다(Qualcomm, 2024).

노트북과 PC에서도 AI 반도체 탑재가 본격화되고 있다. 인텔 Core Ultra가 통합 NPU로 AI 워크로드를 CPU 대비 효율적으로 처리하고(Intel Newsroom, 2024), HP의 OmniStudio X는 47 TOPS급 성능으로 AI 생산성을 강화했다(HP, 2025). 이러한 변화는 클라우드와 엣지를 아우르는 하이브리드 AI 컴퓨팅 환경을 확산시키고 있으며, 마이크로소프트·어도비 등 주요 소프트웨어 기업들도 NPU 기반 AI 가속 지원 범위를 빠르게 넓히고 있다.

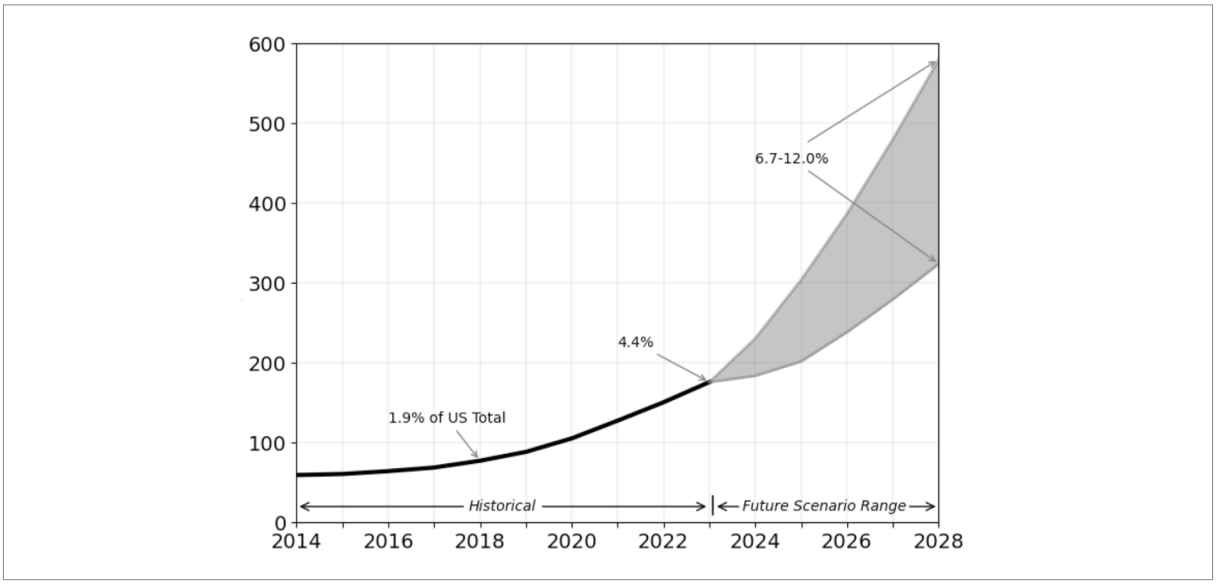
특히 자율주행차 시장은 엣지 AI의 필요성이 가장 두드러지는 산업이다. 엔비디아는 DRIVE Orin(최대 254 TOPS)과 차세대 DRIVE Thor(최대 2,000 TOPS)를 앞세워 시장을 선도하고 있다. Thor는 Orin 대비 4배 이상의 AI 연산 성능과 멀티 도메인 통합 설계를 통해, 인포테인먼트·안전·자율주행 기능을 단일 SoC에서 처리한다(NVIDIA, 2025).

2025년 현재 Li Auto, ZEEKR, Xiaomi, GWM 등이 Orin을, Volvo와 Lynk & Co는 Thor를 채택하여 소프트웨어 정의 차량(SDV) 개발에 나섰다(Volvo Cars, 2025). Volvo는 2020년대 후반 Thor 기반 차량 양산을 계획 중이며, 전문가들은 이를 '이동형 데이터센터' 시대의 개막으로 평가한다(TS2 Tech, 2025). 또한 테슬라는 자체 AI 반도체 전략을 재정비하면서 기존 Dojo 프로젝트를 중단하였다. 대신 AI5 및 AI6 inference 칩 개발에 집중하며 독자 개발 대신 외부 협업을 통한 신속한 상용화로 전략을 선회한 것이라는 분석이 나오고 있다(김흥식, 2025). 이렇듯 엣지 AI 반도체는 산업별 맞춤형 아키텍처와 고효율 설계를 통해, AI의 현장 활용도를 극대화하는 핵심 인프라로 자리매김하고 있다.

AI 인프라 확산의 그늘: 전력·효율성의 도전

클라우드와 엣지의 역할 분담이 본격화되면서 AI 인프라는 전례 없는 속도로 확산되고 있다. 그러나 그 이면에는 전력 소비와 효율성이라는 근본적인 제약이 자리하고 있다. 국제에너지기구(International Energy Agency; IEA, 2025)는 2030년까지 전 세계 데이터센터 전력 소비가 2022년 대비 두 배 이상 증가해 약 945TWh에 이를 것으로 전망하며, 특히 AI 특화 데이터센터가 이 증가분의 핵심 요인이 될 것이라고 분석한다.

그림 1 - 2014년부터 2028년까지 서버 스토리지, 네트워크 장비 및 인프라로 인한 총 전력 생산 대비 소비량 시나리오



출처: 2024 United States Data Center Energy Usage Report (2024)

미국과 중국이 주도하는 AI 데이터센터 확충 경쟁은 이러한 흐름을 더욱 가속화하고 있다. 미국 에너지부가 의뢰한 2024년 보고서에 따르면, 2023년 미국 내 데이터센터 전력 소비는 전체 발전량의 약 4.4%에 해당하는 176TWh였으며, 이는 2028년까지 325 ~ 580TWh로 증가할 것으로 예상된다. 이는 미국 전체 발전량의 약 6.712%에 해당하는 수치다(Shehabi 외, 2024).

민간 예측기관도 유사한 전망을 내놓고 있다. Goldman Sachs(2025)는 2023년 대비 2027년까지 데이터센터 전력 수요가 50% 이상 증가하고, 2030년까지 최대 165%까지 늘어날 수 있다고 분석했다. RAND 연구팀 역시 2027년 글로벌 AI 데이터센터 전력 수요가 약 68GW에 달할 수 있으며, 이는 미국 캘리포니아주의 전체 전력 설비 용량에 필적한다고 지적했다(Pilz 외, 2025).

문제는 이러한 전력 수요 폭증이 단순한 서버 증설이나 전력 인프라 확충만으로 해결되지 않는다는 점이다. Deloitte(2024)에 따르면, 평균적인 대규모 데이터센터에서 전력 소비의 약 40%가 냉각 시스템에서 발생한다. 특히 고성능 GPU와 AI 전용 가속기가 밀집된 환경에서는 열 밀도가 높아져 기존 공랭식 냉각 방식으로는 효율성을 유지하기 어렵다. GPU 세대가 발전할수록 와트당 연산 성능은 비약적으로 향상되었지만, 효율성이 높아질수록 오히려 전체 전력 사용량이 늘어나는 '제번스 역설(Jevons Paradox)'이 AI 산업에서도 나타나고 있다(Moulton, 2025).

때문에 전문가들은 칩 설계 단계에서 전력 관리 기능을 내재화하지 않으면, 전력망 확충 속도가 AI 수요 증가를 따라잡지 못할 것이라고 경고한다. 즉, 칩의 연산 성능 경쟁만으로는 AI 인프라 확장을 지속

가능하게 만들 수 없으며, 전력 효율성, 냉각 기술, 전력 인프라 혁신이 병행돼야 한다는 것이다.

이를 위해 산업계에서는 세 가지 주요 방향이 제시된다. 첫째, 칩 레벨 전력 최적화다. 파워 캐핑(power capping), 동적 전압·주파수 조절(DVFS), 연산 효율을 극대화하는 NPU·ASIC 설계가 대표적이다. 둘째, 냉각 기술 혁신이다. 액체 침지 냉각, 냉각수 재활용, 폐열 회수 시스템은 이미 일부 hyperscale 데이터센터에서 도입되고 있으며, 향후 AI 데이터센터 표준으로 확산될 가능성이 크다. 셋째, 그린 데이터센터 전환이다. PUE(Power Usage Effectiveness) 지표를 개선하고, 재생에너지 기반 전력 사용을 확대하며, 전력망과의 상호작용(Grid-interactive Data Center)을 강화하는 것이 핵심이다. 덧붙여 정책적 지원도 필수적이다. 데이터센터 인허가 절차 간소화, 재생에너지 인프라 투자 확대, 에너지 효율 장비 도입 인센티브 제공 등은 각국이 글로벌 AI 인프라 경쟁에서 우위를 확보하기 위해 검토해야 할 정책 수단이다.

맺음말

클라우드와 엣지의 보완적 역할은 이제 AI 산업 전반의 표준적 구조로 자리 잡았다. 클라우드는 초대형 AI 모델의 학습과 복잡한 분석을 수행하는 '지능의 심장' 역할을 맡고, 엣지는 사용자와 가장 가까운 지점에서 저지연·저전력의 실시간 추론을 구현하는 '촉수'로 기능한다. 이 둘은 물리적으로 떨어져 있으면서도 데이터·모델·컴퓨팅 자원을 끊임없이 주고받으며 하나의 통합 생태계를 이룬다. 다만, 전력 효율성·데이터 보안·네트워크 지연과 같은 현실적 제약은 여전히 양측의 성능 극대화를 가로막는 요인으로 남아 있다.

그럼에도 글로벌 AI 인프라는 앞으로의 AI 경쟁을 좌우할 핵심 기반으로 부상하고 있다. AI 칩과 소프트웨어가 클라우드와 엣지의 요구에 맞춰 함께 발전하며, 상호 보완적 구조 속에서 강점을 극대화할 때, AI는 확장성과 지속 가능성을 모두 갖춘 산업 인프라로 자리 잡을 것이다.

참고문헌

- 김민식·오정숙(2024.07.25.), 새로운 기회의 창으로 AI반도체 시장 현황과 전망, KISDI Perspectives, No. 1, 정보통신정책연구원, https://www.kisdi.re.kr/report/view.do?arrMasterId=4334696&artId=1776660&key=m2102058837181&masterId=4334696
- 김흥식(2025.08.08.), '도조' 해체한 테슬라, AI 전략 외부 협력으로 전환...삼성에 쏠리는 무게, 오토헤럴드, https://www.autoherald.co.kr/news/articleView.html?idxno=57418
- 박원익(2025.04.09.), 10배 강력한 TPU로 추론 AI 이끈다…구글의 3가지 필승 전략, 더밀크(The Miilk), https://themiilk.com/articles/a98f1b97a

- Allan, L.(2025.07.31.), For chip developers, HW/SW co-design key to data center efficiency, Semiconductor Engineering, https://semiengineering.com/for-chip-developers-hw-sw-co-design-key-to-data-center-efficiency/
- Deloitte(2024), Generative AI power consumption creates need for more sustainable data centers, Deloitte Insights, https://www.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2025/genai-power-consumption-creates-need-for-more-sustainable-data-centers.html
- Gill, S. S., Golec, M., Hu, J., Xu, M., Du, J., Wu, H., … Uhlig, S.(2024.10.18.), Edge AI: A taxonomy, systematic review and future directions, Cluster Computing, 28, Article 18, https://doi.org/10.1007/s10586-024-04686-y
- Goldman Sachs(2025), AI to drive 165% increase in data center power demand by 2030, Goldman Sachs Research, https://www.goldmansachs.com/insights/articles/ai-to-drive-165-increase-in-data-center-power-demand-by-2030
- International Energy Agency(2025), AI is set to drive surging electricity demand from data centres, IEA, https://www.iea.org/news/ai-is-set-to-drive-surging-electricity-demand-from-data-centres-while-offering-the-potential-to-transform-how-the-energy-sector-works
- Martin, D.(2025.06.20.), 7 new, cutting-edge AI chips from Nvidia and rivals in 2025, CRN, https://www.crn.com/news/components-peripherals/2025/7-new-cutting-edge-ai-chips-from-nvidia-and-rivals-in-2025
- Moulton, C.(2025.02.07.), What is Jevons paradox? And why it may (or may not) predict AI's future, Northeastern University News, https://news.northeastern.edu/2025/02/07/jevons-paradox-ai-energy-use/
- Nayak, S.(2025.04.22.), Google TPU Ironwood: Revolutionizing AI inference at scale, CloudOptimo, https://www.cloudoptimo.com/blog/google-tpu-ironwood-revolutionizing-ai-inference-at-scale/
- NVIDIA(2025.02.25.), Wave of EV makers choose NVIDIA DRIVE for automated driving, NVIDIA Newsroom, https://nvidianews.nvidia.com/news/wave-of-ev-makers-choose-nvidia-drive-for-automated-driving
- O'Shea, D.(2025.02.23.), Nvidia moves AI focus to inference cost, efficiency, Fierce Electronics, https://www.fierceelectronics.com/ai/nvidia-moves-ai-focus-inference-cost-efficiency
- Pilz, K. F., Mahmood, Y., & Heim, L. (2025, January 28). AI's power requirements under exponential growth: Extrapolating AI data center power demand and assessing its potential impact on U.S. competitiveness (RR-A3572-1). RAND Corporation
- Qualcomm(2024.10.21.), Snapdragon 8 elite mobile platform, Qualcomm, https://www.qualcomm.com/products/mobile/snapdragon/smartphones/snapdragon-8-series-mobile-platforms/snapdragon-8-elite-mobile-platform/
- SamMobile(2025.06.24.), Samsung Exynos 2500 is now official with 3nm tech and impressive specs!, SamMobile, https://www.sammobile.com/news/samsung-exynos-2500-official-launched-specs/
- Shehabi, A., Smith, S. J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M. A. B., Holecek, B., Koomey, J., Masanet, E., & Sartor, D.(2024.12.19.), 2024 United States Data Center Energy Usage Report, Lawrence Berkeley National Laboratory, https://eta.lbl.gov/publications/2024-lbnl-data-center-energy-usage-report
- Sperling, E.(2025.04.02.), Crisis ahead: Power consumption in AI data centers, Semiconductor Engineering, https://semiengineering.com/crisis-ahead-power-consumption-in-ai-data-centers/
- TS2 Tech(2025.03.05.), Self-driving supercomputer showdown: NVIDIA DRIVE Thor vs Tesla FSD Hardware 4 vs Qualcomm Snapdragon Ride Flex, TS2 Space, https://ts2.tech/en/self-driving-supercomputer-showdown-nvidia-drive-thor-vs-tesla-fsd-hardware-4-vs-qualcomm-snapdragon-ride-flex
- Volvo Cars(2025.02.19.), From car to cloud: Volvo Cars expands collaboration with NVIDIA, Volvo Cars Media, https://www.media.volvocars.com/global/en-gb/media/pressreleases/331848
- Wang, T., Guo, J., Zhang, B., Yang, G., & Li, D.(2025), Deploying AI on edge: Advancement and challenges in edge intelligence, Mathematics, 13(11), Article 1878, https://doi.org/10.3390/math13111878