

Tech & Semiconductor Insights: U.S. Market Brief July 2026

KSIA
한국반도체산업협회

COSAR
한국반도체연구조합

Korea AI & System IC
Innovation Center
한미 AI 반도체 혁신센터

본 뉴스 클리핑은 미국 시장 진출 및 관련 사업을 추진하는 기업들에게 미 정부 정책 및 글로벌 반도체 기업들의 주요 동향을 신속하고 정확하게 전달하기 위해 마련되었습니다.

한미 AI 반도체 혁신센터(K-ASIC)는 매일 미국 내 반도체 산업에서 주목해야 할 핵심 뉴스를 선정하고, 이를 심층 분석 및 번역하여 제공합니다. K-ASIC는 한국 반도체 기업들의 미국 시장 진출을 지원하는 핵심 허브로서, 기술 협력, 정책 분석, 비즈니스 구축 및 투자 유치를 종합적으로 지원합니다.

해당 자료는 한미 AI 반도체 혁신센터(K-ASIC)의 지식재산권에 의해 보호되며, 사전 승인 없이 무단 수정, 복사, 배포를 금합니다.

URL : www.kasicusa.com

Top 10 Headlines of the Month

1

OpenAI, Broadcom과 함께 맞춤형 AI 칩 디자인 공개

2

미 상무부, Anthropic Mythos 모델에 대한 규제 완화... 구체적인 사항은 미공개

3

Meta, 잔여 AI 컴퓨팅 판매 선언... 네오클라우드 기업 악영향

4

Nscale, 데이터 센터 9억 달러 리볼빙 신용 한도 확보... 과도한 부채에 의한 유동성 확보

5

Apple, Broadcom으로부터 300억 달러 규모 칩 구매 예정... 관세 혜택에 의한 국내 투자 확대

6

OpenAI, 새로운 모델 GPT-5.6 Sol 정식 출시... 미 정부의 사이버 보안 제재 조치 후 공개

7

Thinking Machines Lab, 첫 AI 모델 출시... 오픈 웨이트 모델을 통한 AI 독점 구조 탈피

8

Nvidia, 새로운 AI 모델 공개 및 일본 Physical AI 생태계 확장

9

AMD, 엔비디아에 대적할 첫 AI 랙 시스템 'Helios' 출시... 새로운 구매자 마이크로소프트 합류

10

AMD, Anthropic과 AI 칩을 위한 50억 달러 파트너십 체결

OpenAI, Broadcom과 함께 맞춤형 AI 칩 디자인 공개

2026. 06. 24

Content

OpenAI는 지난 10월 Broadcom과 함께 ChatGPT와 같은 AI 기술 구동을 위한 맞춤형 컴퓨터 칩을 제작할 것이라고 밝힘. 두 회사는 24일 수요일 Jalapeño(할라피뇨)라고 불리는 첫 번째 칩 디자인을 공개하며, 데이터 센터 전용 칩 기술을 구축할 계획임.

OpenAI는 Broadcom 협력 외에도 Nvidia, AMD, Cerebras와 계약을 체결하며 최종적으로 수백만 가구에 전력 공급이 가능한 10GW 용량 맞춤형 칩을 사용할 계획이라고 밝힘. 주요 하이퍼스케일러 기업(Amazon, Google, Microsoft, Meta)이 올해 약 7,000억 달러 지출 계획을 밝힌 가운데 OpenAI 또한 데이터 센터 건설에 수천억 달러를 지출할 계획임. 이미 텍사스 주 애빌린에 첫 번째 데이터센터를 건설 중이며, 유럽 및 중동에 추가 데이터 센터 건설 계획이 있음.

OpenAI의 하드웨어 부문을 이끄는 Richard Ho는 성명을 통해 “초기 테스트에 따르면, 할라피뇨는 하드웨어의 이론적 한계에 가깝게 우리의 가장 중요한 워크로드를 효율적으로 실행할 것”이라고 말함. OpenAI와 Broadcom은 이례적으로 짧은 기간인 9개월 만에 칩 설계를 완료했다고 밝힘. 회사는 이 칩이 언제 출시될 수 있는지에 대한 추가 정보는 제공하지 않음.

URL

<https://www.nytimes.com/2026/06/24/technology/openai-broadcom-chip-jalapeno.html>

미 상무부, Anthropic Mythos 모델에 대한 규제 완화... 구체적인 사항은 미공개

2026. 06. 26

Content

미 상무부는 Anthropic의 인공지능 모델 중 하나인 Mythos 5에 대한 일부 고객의 접근 권한을 복구할 수 있도록 허가했음. 2주 전 국가 보안상의 우려로 제한했던 모델의 제한을 풀고 다시 공개 사용 형태로 복구에 합의했음. 가장 강력한 모델인 Fable 5의 접근 권한 복구 협상도 여전히 진행 중이며, 이번 조치를 통해 트럼프 행정부와 Anthropic의 갈등이 완화되는 계기가 될 것으로 보임. 올해 초 Anthropic과 트럼프 행정부는 2억 달러 규모의 펜타곤 AI 계약을 두고 처음으로 충돌했음. 지난 2월 Pete Hegseth 국방부 장관은 Anthropic을 ‘공급망 위험 요소’로 지정했으며, 이에 Anthropic은 3월 정부를 상대로 소송을 제기했음. 이후 지난 6월 12일, Anthropic 측에 Mythos와 Fable 모델을 차단하는 데 단 90분의 시간만을 주며 갑작스럽게 통보한 바 있음. 행정부의 보안 우려는 Amazon의 연구원들이 발표한 논문에서 비롯되었으며, 해당 논문은 AI 모델이 소프트웨어 코드 조각의 결함을 찾아내도록 유도가 가능한 점에 있어 보안상 취약점을 지적했음. 이번 합의는 가장 영향력 있는 AI 기업 중 하나인 Anthropic과 트럼프 행정부 간의 반목을 완화하는 역할을 했음. 그러나 행정부가 새로운 AI 시스템 출시에 더 적극적인 역할을 하려 한다는 우려를 잠재우기에는 부족할 수 있음. 행정부가 Anthropic에 대한 제한을 전면 해제할 것인지, 아니면 다른 AI 기업들이 유사한 문제를 피하기 위해 따를 수 있는 더 공식적인 절차를 만들 것인지는 아직 불분명함.

URL

<https://www.nytimes.com/2026/06/26/technology/anthropic-mythos-government-restrictions.html>

글로벌 하이퍼스케일러에 이어 Meta, SoftBank 또한 잔여 AI 컴퓨팅 판매 선언... 네오클라우드 기업 악영향

2026. 07. 02

Content

Meta는 잔여 AI 컴퓨팅 자원을 외부 기업에 판매하는 클라우드 인프라 비즈니스 계획을 개발 중임. SoftBank 또한 AI 칩·클라우드 서비스 전문 법인 'SB Neo Inc'를 설립한다고 발표함. 막대한 자본력을 가진 두 기업들이 직접 클라우드 대여 시장에 진입함으로써, 기존 네오클라우드 기업 (CoreWeave, Nebius, etc)들에게 악영향이 갈 가능성이 있음.

Meta는 잔여 컴퓨팅 자원을 AI 모델 접근권과 함께 외부에 제공하는 구조를 검토 중이며, 자사는 수십만 개의 GPU로 구성된 자체 AI 인프라를 구축하고 있음. 이는 주요 하이퍼스케일러 기업인 Microsoft Azure·Google Cloud·AWS와 AI 전용 클라우드 기업 CoreWeave·Nebius와 직접 경쟁하는 사업 모델임. SB Neo Inc 또한 미국 기업 및 하이퍼스케일러 대상 AI 컴퓨팅 임대 사업을 추진하고, 2030년까지 데이터센터 규모를 10기가와트(GW)로 확대할 계획임.

Meta와 SoftBank의 동시 진입은 CoreWeave·Nebius 같은 기존 AI 전용 클라우드 기업들에 대한 경쟁 압력을 높이는 한편, CoreWeave는 "점점 더 다양해지는 고객층 전반에 엄청난 수요를 지속적으로 확인하고 있다"며, "이건 제로섬 게임이 아닌 빠르게 확장하는 시장을 더욱 확고히 해준다"라고 덧붙임. 주요 하이퍼스케일러 기업들은 Meta와 비슷한 방법을 통해 데이터센터 잉여 용량을 외부에 대여함으로써 투자 비용 회수가 가능한 반면, CoreWeave·Nebius는 담보로 막대한 대출을 감행한 부채 비율이 높아, 향후 대기업들이 고객에서 경쟁자가 되면, 연쇄적인 재정적 손실 리스크가 있음.

URL

<https://www.bloomberg.com/news/articles/2026-07-02/coreweave-junk-bonds-slide-further-as-investors-question-ai-boom?embedded-checkout=true>

영국 스타트업 Nscale, 데이터센터 확장 위해 9억 달러 리볼빙 신용 한도 확보... 과도한 부채에 의한 유동성 확보

2026. 07. 07

Content

영국 네오클라우드 스타트업 Nscale은 필요에 따라 자금을 상시 이동·상환이 가능한 '회전 거래 신용 한도'를 확보하며, 최대 대출금액이 9억 달러까지 가능하다고 발표함. 이 자금은 유럽, 미국, 아시아 전역의 데이터센터 확장 자금으로 쓴다고 밝힘. 해당 신용 한도는 J.P. Morgan, Goldman Sachs, Morgan Stanley, MUFG(미쓰비시), RBC Capital Markets, Bank of America, Cr dit Agricole CIB, Deutsche Bank, Mizuho, SMBC, TD Securities, KeyBank 등 12개 글로벌 투자은행이 대규모 자금 조달을 위한 연합 투자 방식으로 제공함. Nscale은 Nvidia·Dell Technologies·Nokia의 투자를 받은 회사로, 올해 초 기업가치 146억 달러를 돌파하고 20억 달러 투자금을 유치했음. 이번 리볼빙 신용 한도 확보는 Nscale이 지난 5월에 노르웨이 Narvik 데이터센터를 위해 7억 9천만 달러 대출에 이어, 추가 유동성을 확보하기 위함임. 또한 노르웨이 데이터센터에 최소 3만 개의 Nvidia Rubin 배치 계획을 발표했음. Nscale은 최근 AI 기술의 실질적 수익성에 대한 투자자의 의문과 과도한 부채 조달로 인한 'AI 거품론' 우려가 있음. 이에 대해 Nscale CEO Josh Payne은 "주요 글로벌 투자은행들과의 이번 회전거래 신용(RCF) 계약 체결은 당사의 플랫폼과 자본 구조, 그리고 팀에 대한 기관들의 강한 신뢰를 반영합니다. 이번 자원 덕분에 우리는 이를 신속하고 대규모로 수행할 수 있는 유연성을 더욱 확보하게 되었습니다."라고 발표함.

URL

<https://www.wsj.com/tech/ai/nvidia-backed-startup-nscale-locks-in-900-million-for-data-center-buildout-fd9820b3>

Apple, Broadcom으로부터 300억 달러 규모 칩 구매 예정... 관세 혜택에 의한 국내 투자 확대

2026. 07. 08

Content

Apple이 반도체 설계 기업 Broadcom과 300억 달러 이상의 장기 칩 공급 계약을 체결함. 양사는 다세대 Apple 제품을 위한 커스텀 주문형 반도체를 2031년까지 공동개발 및 공급에 합의함. 이번 계약은 Apple의 국내 제조 투자 기준 역대 최대 규모로, 총 150억 개 이상의 미국산 칩을 생산한다고 발표함.

Apple은 Broadcom의 셀룰러·Wi-Fi·블루투스 연결 부품들과 커스텀 주문형 반도체를 차세대 제품에 활용할 예정이라고 발표하며, 이 계약을 통해 Broadcom은 콜로라도주 포트 콜린스에 있는 제조 시설을 확장할 수 있을 것이라고 밝힘. 또한 Apple은 자국의 칩 공급망 강화를 위해 미국 내 제조 시설 구축 예정인 파운드리 TSMC, 칩 패키징 기업 Amkor Technology, 웨이퍼 기업 GlobalWafers부터 칩 구매를 약속함.

이번 계약은 Apple이 2025년 발표한 6,000억 달러 규모의 4개년 미국 내 투자 계획을 이행하는 조치로, 해외 부품에 대한 미국의 고율 관세 정책에 대응하여 부품 공급망과 첨단 제조시설을 미국 내로 확충하려는 전략적 움직임을 보여줌. 최근 Apple은 미국 정부가 지원하는 Intel Foundry와 칩 제조 계약을 체결한 바 있어, Broadcom 계약은 Apple이 미국 정부와 우호적인 관계를 강화하고 있는 것으로 해석됨.

URL

https://www.wsj.com/tech/apple-to-spend-30-billion-on-u-s-made-chips-from-broadcom-aaf23490?mod=tech_lead_pos4

OpenAI, 새로운 인공지능 모델 GPT-5.6 Sol 정식 출시 ... 미 정부의 사이버 보안 제재 조치 후 공개

2026. 07. 09

Content

OpenAI는 새로운 인공지능 모델 GPT-5.6 Sol을 공개하며 Anthropic과 AI 패권 경쟁을 더욱 가속화하고 있음. 트럼프 행정부는 기업들과 오랜 논의 끝에 OpenAI와 Anthropic 모두 새로운 기술을 대중에게 공개적으로 출시할 수 있도록 허용함.

6월 초 트럼프 대통령은 기업의 새로운 AI 모델 출시 전, 정부의 감독을 받도록 요구하는 AI 사이버 보안 관련 행정명령에 서명한 바 있음. 30일 가량 소요될 것으로 예상됐던 정부 검토 절차는 OpenAI가 기술 전문가팀을 정부에 즉시 파견하고 상무부와 밀착 협력하면서 검증 기간이 크게 단축됨. OpenAI는 미 정부의 안보 검토 체계를 최종 통과하고 현지 시간 7월 9일 글로벌 시장에 새로운 모델을 공식 출시함.

OpenAI GPT-5.6의 세 가지 변형 모델 '솔(Sol)', '루나(Luna)', '테라(Terra)'는 사용 목적과 예산에 맞춰 성능과 가격대를 차별화함. 최고 성능의 '솔(Sol)', 가성비를 극대화한 범용 '테라(Terra)', 접근성 중심의 경량화 모델 '루나(Luna)'로 선택의 폭을 넓히며 가격 다변화로 시장을 공략할 예정임.

URL

<https://www.nytimes.com/2026/07/09/technology/openai-sol-ai.html>

OpenAI, 첫 소비자 하드웨어 공개 예정 ... Apple은 영업비밀 유출을 주장하며 소송 제기

2026. 07. 14

Content

OpenAI의 관계자는 자사가 '인간 같은 가정용 AI 동반자 제품'을 개발 중이라고 밝힘. 가정용 기기 시장 진출을 위해 개발 중인 첫 소비자 하드웨어는 화면이 없는 이동식 스피커 형태로, 자사는 2027년까지 제품 출시를 계획했지만, Apple의 영업비밀 유출 고소로 계획에 차질이 생길 수 있음. 이번 스피커를 시작으로, 자사는 향후 스마트폰을 대체할 모바일 AI 기기, 웨어러블 기기, 가정용 로봇 공학 기술 등의 제품군을 검토 중이라고 밝힘.

이번 스마트 스피커는 단순 음성 응답을 넘어 다양한 서비스를 제공함. 카메라 및 센서를 통한 주변 환경 이해, 개인정보 학습을 통한 사용자 맞춤 서비스 제공, 최신 고도화 음성 모드 GPT-Live를 통한 자연스러운 대화 가능 및 충전식 배터리가 내장되어 집안 내 가전 제어, 미디어 재생, 업무 보조 등의 다양한 역할 수행이 가능함.

OpenAI는 Apple 출신 Jony Ive와 Tang Tan이 공동 창업한 하드웨어 스타트업 io Products를 65억 달러에 인수하고, Jony Ive의 디자인 스튜디오 LoveFrom 및 400명의 Apple 출신 인력(Evans Hankey, Paul Meade)을 대거 영입하며 자사의 스마트 스피커를 개발 중임. 이에 Apple은 OpenAI를 자사의 영업비밀 유출 및 제품 가속화로 고소함. OpenAI는 자사의 스피커는 Apple의 홈팟이나 기존 하드웨어와 완전히 설계가 다르다고 반박함.

URL

<https://www.bloomberg.com/news/articles/2026-07-14/openai-s-first-device-will-be-moveable-screenless-speaker-built-as-ai-companion>

Thinking Machines Lab, 첫 AI 모델 출시 ... 오픈 웨이트 모델을 통한 AI 독점 구조 탈피

2026. 07. 15

Content

OpenAI 전 CTO, 현 Thinking Machines Lab CEO Mira Murati는 Thinking Machines Lab의 첫 AI 모델 '잉클링(Inkling)'을 출시했다고 밝힘. 이 모델은 OpenAI와 Anthropic의 독점 구조를 막기 위한 '오픈 웨이트(open weights)' 형태로, 오픈 웨이트 형태는 다른 이들이 본인들만의 데이터로 모델 수정이 가능한 방법임. 자사의 오픈 웨이트 AI 모델 생태계 진출은 독점 기업들의 폐쇄형 모델 방식에 대한 업계 전반의 반발이 일어나는 가운데 나온 행보임. Thinking Machines Lab은 모델 사전 학습을 위해 텍스트, 이미지, 오디오, 비디오를 포함한 4,500만 개의 토큰을 활용하며 최첨단 Nvidia 하드웨어에서 훈련했다고 밝힘. Nvidia는 지난 3월 Thinking Machines Lab과 다년 파트너십을 발표한 바 있으며, 최소 1기가와트 규모의 칩을 제공하기로 합의함. Thinking Machines Lab은 다른 프런티어 연구소와 다르게 "비용과 성능의 균형"을 맞추도록 모델을 설계했다고 설명함. 잉클링이 가진 약 1조 개의 매개변수 중 '활성(Active)'상태인 것은 410억 개로, 질문이 주어졌을 때 AI 두뇌 일부분만 사용하여 비용이 저렴하고 속도가 빠름. 지난 금요일, 자사는 비전에 대한 첫 선언문을 발표함. CEO Murati는 어린 시절 고국 알바니아의 공산주의 몰락을 경험하고, 현재 폐쇄형 AI 패러다임을 '계획경제(central planning)'에 비유함. Murati는 "계획 경제가 실패하는 것은 지능 부족이 아닌, 생산적 지식의 본질이 묵시적, 지역적, 일시적이며 일을 통해 지식을 습득한 개인들이 사적으로 소유하고 있기 때문이다. 중앙 집중화된 인공지능이 사용하도록 지식을 모으는 시도 역시 동일한 한계에 직면하게 된다"라고 설명함.

URL

https://www.wsj.com/tech/ai/mira-muratis-ai-startup-releases-first-model-in-bid-to-loosen-ai-giants-grip-e042bb2b?mod=tech_lead_story

Nvidia, 새로운 AI 모델 공개 및 일본 Physical AI 생태계 확장

2026. 07. 16

Content

Nvidia는 로봇 및 비전 AI 에이전트를 위한 새로운 AI 모델 Cosmos 3 Edge를 공개함. 이 모델은 World model로서 시스템이 물리적 환경에서 실시간 인식 및 탐색을 돕기 위해 설계됨. Nvidia CEO Jensen Huang은 “AI의 다음은 Physical world이며, 이는 일본에 있어 엄청난 기회다”라며, “일본은 현 시대 제조업을 발명한 국가다. 이제 지능형 산업의 시대를 맞아 이를 재창조할 기회를 잡았다”라고 말함. Nvidia는 일본의 주요 대기업인 Fujitsu, Hitachi, Kawasaki 등 파트너십을 통해 일본 내 Physical AI 입지를 넓혀가고 있다고 전함. 이러한 파트너십은 Microsoft가 AI 인프라 구축 및 사이버 보안 강화를 위해 일본에 100억 달러 규모의 투자를 단행한 지 몇 달 만에 나온 것으로, 국제 무역 기구(ITA)에 따르면, 일본의 AI 시장은 2029년까지 약 279억 달러 규모에 달할 것으로 예상됨. Nvidia는 에이전트 AI 분야와 관련하여, 신약 개발 및 의료 로봇 공학 영역을 위해 Astellas Pharma, Daiichi Sankyo, Ono Pharmaceutical 등 주요 제약 기업들과 파트너십을 맺고 업무 프로세스 효율화를 위한 바이오 툴킷 활용을 확대하고 있다고 밝힘. Nvidia의 World model은 LLM(대규모 언어모델)에 비해 더 광범위한 입력을 학습할 수 있는 시스템임. 40억 개의 매개변수를 다룰 수 있으며, 실시간 추론을 통해 동급 대비 최고 처리량과 정확도를 갖춘 VLM (비전 언어모델) 역할이 가능하다고 설명함.

URL

<https://www.cnn.com/2026/07/16/nvidia-reveals-new-ai-model-and-expands-japan-physical-ai-ecosystem.html>

Meta, Anthropic에 컴퓨팅 파워 임대 논의 중 ... 2년간 최대 100억 달러 계약 규모

2026. 07. 17

Content

Meta는 AI 데이터 센터 컴퓨팅 파워를 Anthropic에 임대하기 위한 논의를 진행 중임. 이 계약은 2년간 최대 100억 달러에 달할 수 있으며, Anthropic은 지난 6월 Meta에 이 계약을 제안했고, Meta는 이 계약을 검토 중임. 이는 임대 사업을 통해 Meta AI 비즈니스를 향한 새로운 잠재적 방향이 될 수 있음을 시사함.

최근 Meta CEO Mark Zuckerberg는 올해 자사가 최대 1,450억 달러를 지출할 것이며, 그중 상당 부분이 AI에 투입될 것이라고 밝힘. 이는 지난해 지출한 720억 달러의 두 배가 넘는 금액이며, Meta의 컴퓨팅 파워 임대 사업이 투자의 일부 수익원이 될 수 있음을 암시함. 그러나 Anthropic이 Meta에 제안한 계약 규모는 지난 5월 SpaceX와 체결한 계약 규모의 약 3분의 1 수준이며, 당시 Anthropic · SpaceX 계약은 3년간 450억 달러 규모임.

Meta는 자체 시설 구축과 동시에 다른 데이터센터 제공업체와 계약을 체결함. 지난 3월 데이터센터 임대 기업 Nebius와 270억 달러 규모 계약을 체결했고, 4월 AI 컴퓨팅 파워 제공업체 CoreWeave와 210억 달러 규모 계약을 체결함. Bloomberg Intelligence Tech Analyst Mandeep Singh는 이러한 계약이 한정된 공급과 수요 증가로 인해 컴퓨팅 파워 가격의 상승을 이끌었으며, Meta가 장기적으로 자체 AI에 지속 투자 및 데이터 센터 임대 사업 기회도 생겼다고 덧붙임.

URL

<https://www.nytimes.com/2026/07/17/technology/meta-anthropic-ai-computing-power.html>

AI 추론 칩 스타트업 Etched 기업가치 50억 달러 돌파 ... 새로운 펀딩 라운드 통해 목표 기업가치는 200억 달러

2026. 07. 17

Content

AI 추론 칩 스타트업 Etched가 공식 스텔스 모드를 해제하면서 고객사와 체결된 10억 달러 초과분의 공급 계약을 발표함. 현재까지 8억 달러 자금 조달을 완료했고, 주요 투자 기관은 Jane Street, Hudson River Trading, Two Sigma, Venture Tech Alliance(TSMC Partner)등이 있음. 최근 기업가치 50억 달러 돌파 이후 Jane Street가 리드로 참여하는 새로운 펀딩 라운드에 기업 가치를 약 200억 달러 목표로 투자 유치 예정임. Etched는 향후 Nvidia의 잠재적 경쟁사로 주목을 받고 있음. Etched의 Sohu Chip은 트랜스포머(Transformer) 모델 아키텍처 기반 주문형 반도체(ASIC)로 범용 GPU 대비 AI 추론 비용 절감을 목표로 설계됨. 자사는 2026년 여름부터 고객사에 칩 출하를 시작할 계획이라고 밝힘. Sohu Chip은 HBM/SRAM 하이브리드 디자인과 Low Voltage Inference 기술로 메모리 용량, Memory-to-Memory latency, 칩 발열 등의 문제를 풀어내고 있으나 쓸 수 있는 모델 유형이 한정돼 있음. 순수 트랜스포머 어텐션 연산만을 지원하도록 설계되었기 때문에 비주얼 인코더, 이미지 생성, 영상 생성, 확산(diffusion) 등이 불가하다는 단점을 가지고 있음. 과거 NVIDIA GPU가 AI 시장을 장악했다면, AI 시대가 학습(training)에서 추론(inference)으로 넘어가면서 Google, Broadcom(OpenAI Jalapeño) 같은 주요 기업뿐만 아니라 Etched 같은 스타트업들이 AI Accelerator 시장에 진출하며 추론 전용 ASIC 진영을 넓히고 있음. 그중 Etched는 스타트업 최초 자체 추론 칩, 메모리, 서버 랙을 모두 디자인하는 점에서 주목받고 있음.

URL

<https://www.bloomberg.com/news/articles/2026-06-30/ai-chip-startup-etched-says-jane-street-tsmc-linked-vc-invested>

AMD, 엔비디아에 대적할 첫 AI 랙 시스템 'Helios' 출시 ... 새로운 구매자 마이크로소프트 합류 (2026.07.20)

2026. 07. 20

Content

AMD가 2025년 10월 Meta OCP에서 소개된 자사의 AI 랙 시스템 Helios를 정식 출시함. Meta, OpenAI, Oracle 등 주요 기업이 초기 고객으로 참여하며, Microsoft가 출시 발표와 함께 신규 고객으로 참여함. Helios 출시는 Nvidia의 Grace Blackwell, Vera Rubin 시스템의 경쟁 제품으로 주목받음.

AMD CEO Lisa Su는 "Helios는 Nvidia 랙 스케일 시스템에 비해 추론(Inference), 메모리 대역폭(Memory Bandwidth)과 메모리 기능 측면에 있어 상당한 이점이 있다."라고 언급하며 Helios의 성능을 강조함. 추가로 AMD 데이터센터 책임자 Forrest Norrod는 "우리는 최고의 총소유비용(TCO)과 토큰당 가장 낮은 비용을 종합적으로 제공하는 것에 집중한다."라고 말하며 "고객들은 우리가 그 목표를 달성하고 있다고 말해주고 있다"라고 덧붙임. Helios의 추정 가격은 500만 달러에서 550만 달러이며, 이는 Nvidia 2세대 랙 스케일 시스템 Vera Rubin의 추정 가격 350만 달러와 비교되는 수치임.

AMD는 Helios 출시로 데이터센터 CPU 및 GPU 시장의 입지를 넓히며, 2027년 매출 상승을 기대함. 그러나 Futurum Group CEO Daniel Newman는 "AMD가 기술적으로 우수해서 이기는 것인가, 아니면 단순히 공급 부족으로 인해 만들기만 하면 누군가 사주는 구조이기 때문인가?"라며, AMD 초기 도입 성과에 따라 시장 기회가 달라질 수 있음을 언급함.

URL

<https://www.cnbc.com/2026/07/20/amd-helios-microsoft-ai-nvidia.html>

영국 스타트업 CuspAI 기업가치 5배 상승 ... 신소재 개발을 위해 AI Material Foundry 연합 설립

2026. 07. 20

Content

영국 AI 스타트업 CuspAI가 시리즈 B 투자 유치에 성공하며 기업 가치를 5억 2000만 달러에서 26억 달러로 5배 상승함. 주요 투자자로 Kleiner Perkins · NEA가 리드하며 Jeff Bezos의 Bezos Expeditions 및 AMD Ventures가 후속 투자에 참여함. 또한 CuspAI는 AI Material Foundry 연합을 설립하며, Nvidia, Meta, 현대자동차그룹 등 48개의 기업 및 기관이 참여해 초고속 신소재 설계 및 개발을 진행할 계획임. CuspAI는 처음에 탄소 포집 및 수질 정화를 위한 신소재 발굴을 목표로 했으나, 반도체 제조 공급망의 강력한 수요를 발견하고 사업 방향을 전환했다고 밝힘. 공동 창업자이자 AI 연구자 Max Welling을 필두로, AI 기술 선구자 Yann LeCun과 Geoffrey Hinton이 자문위원회에 이름을 올림. 자사는 앞으로 반도체 제조의 주요 과제인 희귀 광물 의존도 및 에너지 소비 효율 문제를 해결할 계획임. 올해 자사 주요 계획 중 80%는 소재 발굴에 집중되어 있으며, 루테튬 혹은 이리듐처럼 공급망 리스크가 있는 희귀 금속 사용을 반도체 제조 공정 시 줄이거나 없애는 것이 목표임. CuspAI는 과거 프로젝트 중 하나인 기후 변화 대응 신소재 발굴 연구로 '금속-유기 골격체(MOF)' 내 300조 개 분자 구조 라이브러리를 구축했으며, 이후 AI 시스템을 통해 10개로 압축한 사례가 있음. 자사는 이러한 AI 학습 시스템을 통해 잠재적 신약 개발과 제조 및 재생 에너지 소재 설계 개발이 가능하다고 밝힘. CuspAI가 설립한 AI Material Foundry 연합은 참여기업들이 컴퓨팅 파워 및 과학 연구 데이터를 공유하고, 자사의 AI 플랫폼 고도화를 위한 기반을 제공함. 또한 신소재의 유효성을 싱가포르, 케임브리지, 샌프란시스코에서 검증하며, 참여 기업들은 이번 연합을 통해 초고속 신소재 설계 및 선점 기회 확보를 기대하고있음.

URL

<https://www.bloomberg.com/news/articles/2026-07-20/bezos-backs-startup-focused-on-material-discovery-for-chipmaking?srnd=phx-ai>

Google, 3개의 새로운 Gemini 모델 출시 ... AI 사이버 보안 시장 진입

2026. 07. 21

Content

Google이 새로운 인공지능 모델 Gemini 3.6 Flash, 3.5 Flash-Lite, 3.5 Flash Cyber를 출시함. Google은 이번 AI 주도권을 잡기 위해 Gemini 3.6 Flash 및 3.5 Flash-Lite 제공 비용을 낮추는 전략을 선택함. Gemini 3.5 Flash Cyber의 출시는 Anthropic의 Mythos와 OpenAI의 보안 모델 등이 선점한 사이버 보안 및 소프트웨어 취약점 탐지 분야에 대한 Google의 대응으로 해석되며, Gemini 모델 효율성을 위해 전용 칩 또한 개발 중임. Gemini 3.6 Flash는 코딩 및 금융 작업 특화 모델인 동시에 토큰 사용량을 17% 줄임. Gemini 3.5 Flash-Lite는 AI 에이전트 관리와 가벼운 업무 처리를 저렴한 비용으로 이용 가능하며, 앞선 두 모델 모두 대중 서비스로 공개함. 반면, Gemini 3.5 Flash Cyber는 사이버 보안을 위해 만든 모델로, 오픈소스 V8 자바스크립트 엔진 검사에서 다른 기존 모델들이 발견하지 못한 10개의 신규 취약점을 발견함. 사이버 보안 모델은 미국 정부와 일부 기업 한정으로 보안 강화를 위해 제공할 예정임. 최고 성능인 Gemini 3.5 Pro는 출시가 지연됐으며, Google은 “준비되는 대로 널리 이용 가능하게 할 것입니다”라고 밝힘. 또한, Google은 역대 최대 규모의 Gemini 4 사전 학습에 착수했으며, 많은 주목을 받고 있음.

URL

<https://www.nytimes.com/2026/07/21/technology/google-ai-cybersecurity-gemini.html>

AMD, Anthropic과 AI 칩을 위한 50억 달러 파트너십 체결

2026. 07. 22

Content

AMD와 Anthropic은 최대 50억 달러 규모의 AI 파트너십을 발표함. Anthropic은 AMD의 랙 스케일 솔루션 Helios를 활용해 최대 2GW 규모의 AMD MI450 시리즈 GPU를 배치할 계획임. 양사는 AMD 가속기에서 Claude AI 워크로드를 최적화하기 위해 공동 엔지니어링을 진행할 계획임.

이번 계약을 통해 Anthropic은 AI 서비스 확대에 필요한 컴퓨팅 자원을 안정적으로 확보하고 Nvidia 의존도를 낮추는 효과를 얻게됨. Anthropic은 컴퓨팅 용량을 늘리기 위해 지난 4월 Amazon, Google · Broadcom과 컴퓨팅 계약, 5월 SpaceX와 계약을 통해 용량을 확보한 바 있음. 또한, Meta로부터 컴퓨팅 파워를 임대하기 위한 예비 논의도 진행 중임. AMD 측은 1GW 규모의 AI 컴퓨팅 인프라 구축에 수백억 달러의 비용이 소모될 수 있다고 밝힘.

AMD AI 데이터센터 용량 중 1GW는 내년 상반기에 배포될 예정이며, 1GW는 미국 내 약 75만 가구에 전력을 공급할 수 있는 전력량임. AMD CEO Lisa Su는 “Anthropic과의 파트너십을 더욱 심화하고 AMD Helios를 배포할 수 있게 되어 매우 기쁩니다”라고 전함.

URL

<https://www.cnbc.com/2026/07/22/amd-anthropic-ai-chip-investment.html>