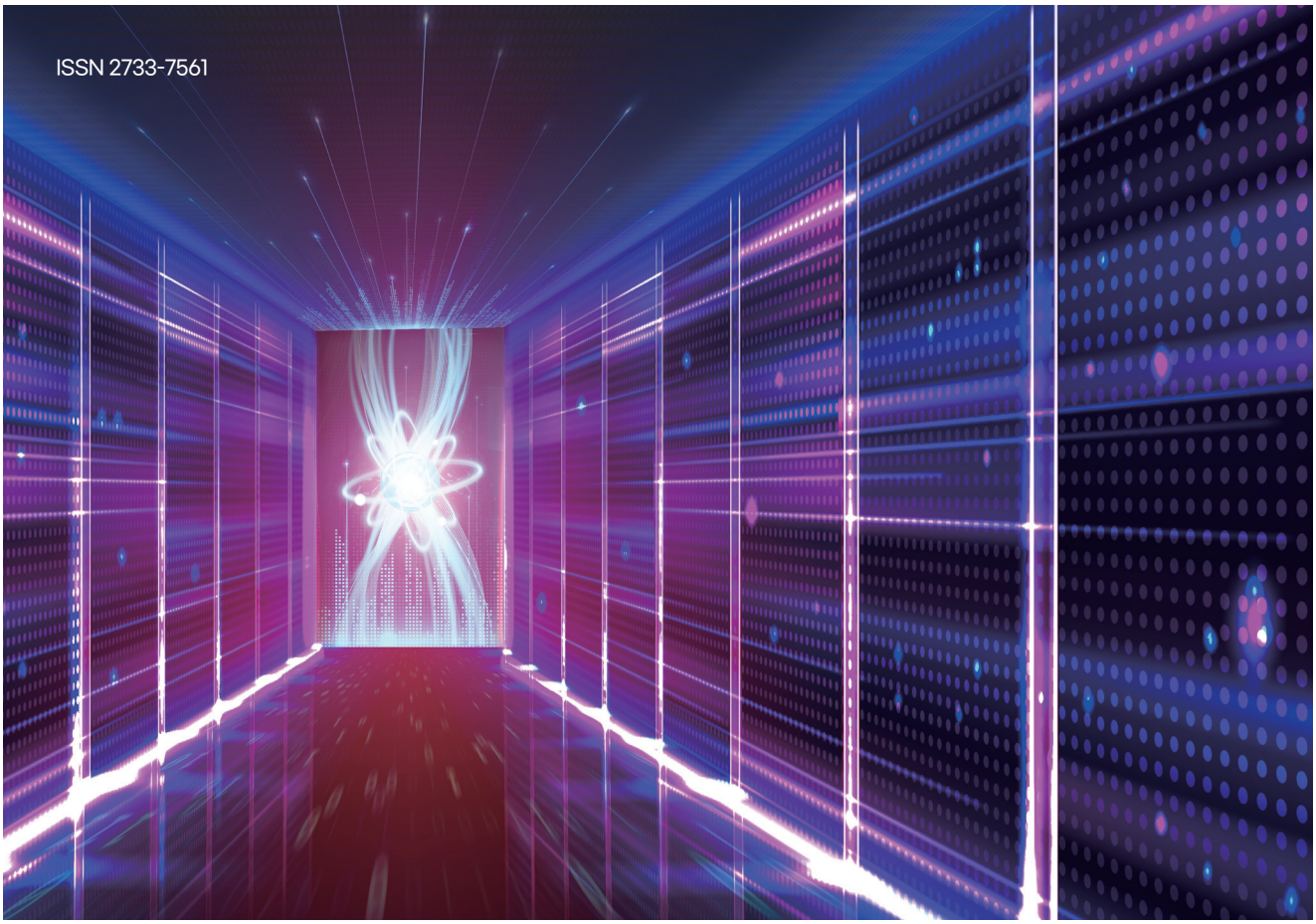
ISSN 2733-7561



GLOBAL HPC HORIZON

국가슈퍼컴퓨팅본부 초고성능컴퓨팅정책센터

허영주, 고동건, 권민우, 김성준, 안준언, 우준, 박형우, 김남규,
박소영, 김재욱, 윤태호, 이승희, 최선희, 곽재혁

vol. 1
2025 SUMMER

KiSTi www.kisti.re.kr 한국과학기술정보연구원 Korea Institute of Science and Technology Information

CONTENTS

목 차

CHAPTER 01

초고성능컴퓨팅 시장 및 서비스 동향

01 인프라부터 에이전트까지… 'SK표 AI패키지' CES 간다 6
02 AI PC에서 서버까지: 퀄컴과 AMD, Intel의 시장 지배에 도전 7
03 2025년 데이터 센터를 혁신하는 4가지 주요 트렌드 8
04 SK하이닉스, CES 2025에서 '풀스택 AI 메모리 공급업체' 비전 공개 9
05 펭귄 솔루션스, SK 텔레콤 및 SK 하이닉스와 AI 데이터센터 협약 체결 10
06 인공지능(AI) 칩 시장, 2025-2029년 동안 902.65억 달러 성장 11
07 Miltiverse Computing, Kinesis와 파트너십으로 AI 최적화 및 에너지 소비 절감 추진 12
08 Supermicro, NVIDIA RTX PRO 6000 Blackwell Server Edition GPU 지원 발표 13
09 Softbank 그룹, Ampere 컴퓨팅 인수 14
10 젠슨 황, NVIDIA의 인공지능 기반 미래를 설계하다 15
11 AI 기반 지식 관리로 엔지니어링 워크플로우 혁신 16
12 IBM, AI 추론 및 시스템 통합 확장을 위한 z17 공개 17

CHAPTER 02

초고성능컴퓨팅 인프라 구축 동향

제1절 산업계 동향 18

01 NVIDIA의 블랙웰, 인공지능의 미래를 보여줌. 당장은 수냉식 18
02 AWS, 오하이오 주 데이터 센터 확장을 위해 100억 달러 투자 19
03 NVIDIA의 Blackwell, AI의 미래가 수냉식임을 시사 20
04 Eni, 자원 탐사 및 탈탄소화를 위한 1억 유로 HPC6 슈퍼컴퓨터 공개 21
05 Infortrend, HP와 미디어 및 엔터테인먼트용 스토리지 공개 22
06 Seagate, 데이터 센터 구축을 위한 36TB HAMR 하드 드라이브 준비 완료 23
07 Spectra Logic, 테이프 연결 확장을 위한 24G 광 SAS 스위치 도입 24
08 VDURA, Rice 대학교 에너지 HPC 컨퍼런스에서 차세대 데이터 플랫폼 공개 25
09 DOE, 차세대 주요 슈퍼컴퓨터 세부 정보 공개 - E Capitan의 동반 시스템 26
10 Fluidstack, 프랑스에 1GW 인공지능 슈퍼컴퓨터 구축 27
11 Tier 0 스토리지가 GPU 스토리지의 판도를 바꾸는 이유 28
12 CoolIT, AI 및 HPC를 위한 고용량 row(행) 기반 냉각 솔루션 공개 29
13 Oracle, 30,000개의 AMD MI355X 가속기로 AI 클러스터 구축 30

제2절 학계, 연구계 동향 31

01 HLRS: 동적 전력 캠핑으로 HPC에서 더 나은 에너지 효율성 실현 31
02 Cineca, 이탈리아에서 가장 강력한 양자 컴퓨터 IQM Radiance 54 도입 32
03 Pasqal, EuroHPC 조달 계약에 따라 EuroQCS-Italy 양자 시스템 제공 33
04 QpiAI, 인도 국가 양자 미션의 일환으로 25큐비트 초전도 양자 시스템 출시 34

CHAPTER 03

초고성능컴퓨팅 기술개발 동향

제1절 산업계 동향 36

01 Meta, 2030년까지 루이지애나 북동부에 100억 달러 규모의 AI 데이터 센터 건설 36
02 AWS, HPC 서버용 대규모 인메모리 데이터베이스 EC2 U7inh 인스턴스 출시 37
03 OpenAI, 2024년 12월 20일 새로운 AI모델 'o3'와 'o3-mini' 발표 38
04 Microsoft, Hugging Face에 Phi-4 언어 모델 출시 39
05 SandboxAQ가 과학과 의학에서 조용한 혁명을 주도하는 방법 40
06 Altair, 인공지능 기반 스케줄링이 가능한 클라우드 플랫폼 HPC 웍스 업그레이드 출시 41
07 세계 최초로 100% AI 생성 논문이 세계 탑 AI 학술대회 워크샵 동료평가 통과, Sakana ai「AI Scientist」가 달성 42
08 대형 언어 모델(LLM) 추론 최적화 43
09 Fujitsu, 오픈소스 기반 양자 컴퓨터 운영 소프트웨어 출시 44
10 UALInk 컨소시엄, Ultra Accelerator LInk 200G 1.0 사양 발표 45
11 젠스파크, 마누스보다 뛰어난 '슈퍼 에이전트' 출시 - "진정한 첫 범용 에이전트" 46
12 AMD, TSMC-2nm 공정 기반 첫 HPC 제품 'Venice' EPYC CPU출시 47
13 OpenAI, AI 에이전트 검사를 위한 벤치마크 BrowseComp 발표 48

제2절 학계, 연구계 동향 49

01 ANL, AI에 대비하여 새로운 세대의 연구자를 양성하기 위한 교육 시리즈 운영 49
02 Sandia 국립 연구소, 에너지 효율적 AI 및 컴퓨팅 기술 개발을 위해 연구소들과 협력 50
03 BSC, MareNostrum 5에 통합할 새로운 양자 시스템 출시 51
04 중국, 양자 컴퓨터로 10억 개 매개변수 AI 모델 파인튜닝 완료 52

CHAPTER
04 초고성능컴퓨팅 응용 및 활용 동향

01 슈퍼컴퓨터를 활용한 연쇄 지진 활동 연구 54
02 화학 소재분야를 위한 확장 가능한 기계학습 모델 개발 55
03 슈퍼컴퓨터를 활용한 해양 풍력 터빈 설계를 위한 대와류 시뮬레이션 연구 56
04 혈관 내 단백질의 기계적 힘에 대한 반응 연구 57
05 ITER 토카막에서의 전자 행동 연구를 위한 슈퍼컴퓨터 시뮬레이션 58
06 슈퍼컴퓨터를 활용한 유전 데이터 분석 가속화 59
07 PSC: Bridges-2 시뮬레이션, 로켓 배출수가 달의 얼음 저장소에 미치는 영향 분석 60
08 슈퍼컴퓨터를 활용한 맞춤형 암 치료법 연구 61
09 슈퍼컴퓨터를 활용한 배터리용 고체전해질 설계 62
10 슈퍼컴퓨터를 활용한 소규모 해양과정의 폭풍 발달 영향 확인 63
11 ESA의 Space HPC를 통한 태양 폭풍 및 우주 기상 모델링 가속화 64
12 Ansys와 Baker Hughes, Frontier를 활용하여 획기적으로 CFD 시뮬레이션 시간 단축 66
13 TACC Frontera를 활용한 미세소관 말단에서의 새로운 행동 확보 67
14 LLNL Sierra 슈퍼컴퓨터를 활용한 둔감 고폭탄에서의 핫스팟 형성 시뮬레이션 69

CHAPTER
05 초고성능컴퓨팅 정책 동향

01 유럽의 AI 모델 개발을 발전시키기 위한 EuroHPC 자금 지원 MINERVA 프로젝트 70
02 EuroHPC, 유럽의 독자적인 HPC 및 AI 개발을 위한 DARE 프로젝트 지원 71
03 DOE, 잠재적 AI 중심 데이터 센터 개발을 위한 16개 연방 부지 선정 72
04 EuroHPC, 국가 AI 생태계 지원을 위한 AI 팩토리 안테나 제안 개시 73

출처 74

요약

Executive Summary

인공지능 기술이 사회 전반을 근본적으로 재편하면서, 이를 실현할 수 있는 기반 기술로서 초고성능컴퓨팅(High Performance Computing, HPC)의 중요성이 더욱 대두되고 있다. 생성형 AI, 대규모 언어 모델(LLM), 정밀 과학 시뮬레이션 등 고복잡도 작업을 수행하기 위해서는 막대한 연산 능력과 효율적인 데이터 처리가 필수적이며, 이에 따라 고성능 서버, 메모리, 스토리지, 냉각 시스템 등 HPC 인프라 전반에 대한 수요가 빠르게 확대되고 있다. 동시에 글로벌 기업과 주요국들은 양자컴퓨팅 기술을 미래 전략 기술로 인식하고 적극적인 투자와 협력에 나서고 있다. 특히, 유럽의 EuroHPC 프로젝트, 이탈리아의 양자 시스템 도입, 중국과 인도의 양자 미션과 같은 사례는 HPC와 양자 기술의 융합이 향수 기술 주권과 경쟁력 확보의 핵심축임을 보여준다.

Global HPC Horizon 2025년 제1호는 이처럼 빠르게 진화하고 있는 HPC 분야의 시장 동향, 인프라 구축, 기술 개발, 응용 사례, 각국의 정책 흐름을 종합적으로 분석, AI 시대를 견인할 기술 기반의 현재와 미래를 조망하고자 한다.

1장에서는 AI 패키지, 차세대 AI 칩, 데이터 센터 트렌드, 글로벌 기업들의 전략 등 HPC와 밀접한 시장 및 서비스 동향을 소개한다. 특히 SK하이닉스, 퀄컴, AMD, NVIDIA 등 주요 기업들의 움직임을 중심으로 기술 진보와 산업 재편의 흐름을 조망한다.

2장에서는 수냉식 서버, 대용량 스토리지, 차세대 슈퍼컴퓨터 등 하드웨어 기반 기술과 각국의 전략적 투자 사례를 포함해서 산업계와 학계에서 진행중인 HPC 인프라 구축 동향을 다룬다.

3장에서는 HPC 기술의 발전과 AI 통합 흐름을 짚어본다. OpenAI, Microsoft, Fujitsu 등 기술 선도 기업들이 선보인 최신 기술 및 AI 에이전트, LLM 최적화, 양자 컴퓨팅 등의 기술 개발 동향을 다룬다.

4장에서는 슈퍼컴퓨터가 실제로 적용되고 있는 다양한 응용 사례를 제시한다. 기후 예측, 생명과학, 에너지, 우주과학 등 다양한 분야에서 HPC가 이끌어낸 혁신적 결과를 통해 HPC의 실질적 가치를 확인할 수 있다.

마지막으로 5장에서는 각국의 정책 방향과 HPC 생태계 조성을 위한 국제적 협력 노력을 조망하며, 유럽의 EuroHPC 프로젝트, 미국의 DOE 정책 등 글로벌 차원의 정책적 대응과 전략을 소개한다.

01 초고성능컴퓨팅 시장 및 서비스 동향

01

인프라부터 에이전트까지… ‘SK표 AI패키지’ CES 간다

개요

SKT, CES 2025에서 AI 데이터센터 기술력 소개 예

- SK텔레콤, ‘CES 2025’에서 인공지능(AI) 기술 및 서비스 소개 예정
- SKT는 SK하이닉스, SKC, SK엔무브 등과 함께 ‘혁신적인 AI 기술로 지속가능한 미래를 만든다’는 주제로 공동 전시관을 운영
- 센트럴 홀 내 1950㎡ (590평) 규모 전시 공간에서 AI 데이터 센터(DC) 기술과 각종 AI 서비스, 파트너 협업 내용을 소개 예정

AI를 통한 지속가능한 미래 지향 AI 인프라 모습 제시

- SKT는 지난달 열린 ‘SK AI 서밋’에서 AI 데이터센터와 GPU 클라우드 서비스(GPUaas), 에지(Edge) AI 기술을 근간으로 한 전국 단위의 ‘AI 인프라 슈퍼 하이웨이’ 전략 발표
- SK 그룹 전시관의 핵심 소재는 AI 데이터 센터로, SKT는 AI DC 부스 중앙에 AI 데이터센터의 데이터 흐름을 표현한 6m 높이 대형 LED 기둥을 설치하고 이를 중심으로 SK그룹이 보유한 네 가지 AI DC솔루션(에너지·AI·운영·보안) 등 총 21개 아이템을 소개할 예정
- AI DC 내 분산 발전원을 설치, 안정적·효율적으로 전력을 공급하는 기술, 액체를 활용한 발열 관리 등 에너지 관련 솔루션 전시 예정
- 신경망처리장치(NPU) 기반 리벨리온의 AI 가속기도 선보일 예정으로 SK 하이닉스가 개발중인 현존 D램 최고 솔루션 ‘HBM3E 16단’ 등 AI 데이터 센터를 구성하는 다양한 AI 반도체와 반도체 공정의 ‘게임체인저’로 불리는 SKC(앱솔릭스)의 유리기판 기술 등도 소개 예정

결론 및 시사점

- SK는 미래 HPC 인프라에서 핵심이 될 AI 데이터 센터 인프라 관련 첨단 기술들을 CES 2025 에서 전세계를 대상으로 주도적으로 선도할 계획을 밝힘

02

AI PC에서 서버까지: 퀄컴과 AMD, Intel의 시장 지배에 도전

개요 **AMD와 퀄컴, 인공지능과 고성능 컴퓨팅 분야에서 Intel의 시장 지배에 도전**

퀄컴, 스냅드래곤(Snapdragon) X 시리즈 프로세서를 통해 인텔과 AMD에 직접적인 도전

- 퀄컴은 2024년 투자자 행사에서 QCT[1] 부문 매출을 2024년 83억 달러에서 2029년 220억 달러로 성장시키겠다는 계획을 발표

AMD는 데이터 센터 시장에서 Intel에 도전

- Epyc 프로세서를 통해 엔터프라이즈, 중견, 중소기업 시장으로 확장 추진
- AMD는 2022년 4세대 Epyc 프로세서 Genoa 출시 이후 x86 서버 CPU 시장 점유율을 24.2%까지 끌어올렸으며, 이는 2006년 기록한 26.2%에 근접한 수치
- 2024년 5세대 Epyc 프로세서 Turin을 출시하며 성능, 에너지 효율성, 총소유비용 측면에서 우위를 강조

AI 도입 증가는 GPU와 가속기 수요를 촉진하고 있으며, 퀄컴과 AMD는 각각 강점을 보유

- 퀄컴은 스냅드래곤 X 시리즈에 AI 기능을 통합
- AMD는 Instinct 제품군과 Epyc 제품으로 HPC와 AI 클러스터에 적용
- 특히, AMD의 CPU는 NVIDA GPU 시스템 보완 측면에서 상당한 수요 존재

결론 및 시사점

- 퀄컴과 AMD는 각사가 보유한 기술적 장점을 바탕으로 시장 점유율을 확대하고 있으며 이를 통해 Intel의 오랜 시장 지배력에 변화를 가져올 것으로 전망

1) QTC(Qualcomm CDMA Technologies): 퀄컴의 핵심 사업 부문으로 주로 반도체 설계 및 제조를 담당

03

2025년 데이터 센터를 혁신하는 4가지 주요 트렌드

개요

미래의 데이터 센터 관리는 과거와 같이 단순하지 않음

• 데이터 센터 업계는 지금까지는 증가하는 수요를 효과적으로 관리하고 지속 가능한 성장을 보장할 수 있었지만 향후 지속적인 수요 보장을 위해서는 기존의 고립된 관점을 넘어 전체적인 관점을 가지고 접근하는 방식으로 다목적의 최적화 로드맵을 수립해야 함

2025년 데이터 센터를 혁신할 4가지 주요 트렌드

- 에너지: 2022년 데이터 센터와 암호화폐는 전 세계 전력 수요의 약 2%를 차지하는 약 460TWh(10^{12} Wh)를 소비했으나 2026년까지 1,000TWh를 초과할 것으로 예상
- 컴퓨팅 집약적인 워크로드가 급증하면서 새로운 전환점을 만들고 있음: AI 전용 데이터 센터는 전력 밀도는 기존 데이터 센터에 비해 훨씬 높지만 워크로드와 애플리케이션이 기하급수적으로 증가함에 따라 시설 규모도 증대함. 2026년까지 AI 전용 데이터 센터는 100~300TWh를 소비할 것으로 예상
- 더 높은 전력 요구 사항을 충족하기 위한 새로운 접근 방식의 발전: 액체 냉각 방식 채택 증가 등
- 데이터 센터의 지속 가능성과 효율성에 대한 노력 예상: Google은 AI 모델 학습에 필요한 에너지를 크게 줄이는 방안을 추진. 예를 들어 6세대 텐서 처리 장치(TPU)인 트릴리움은 이전 세대인 TPU v5e보다 에너지 효율이 67% 이상 높음

결론 및 시사점

- 미래 HPC 인프라의 핵심 요소는 AI와 데이터 컴퓨팅이며 이를 지원하기 위한 에너지 관리 기술이 될 것으로 예측됨

04

SK하이닉스, CES 2025에서 '풀스택 AI 메모리 공급업체' 비전 공개

개요

SK하이닉스, CES205에서 인공지능 시대를 선도할 다양한 메모리 신기술을 적용한 신제품 전시를 통해 풀스택 AI 메모리 공급업체로서의 비전 공개

SK하이닉스가 미국 라스베이거스에서 열리는 CES 2025에 참가하여 혁신적인 인공지능(AI) 메모리 기술력을 선보임

- SK하이닉스, 2025년 1월 7일부터 10일까지 미국 라스베이거스에서 열린 CES 2025에 참가하여 AI 메모리 기술을 선보임
- 곽노정 CEO, 김주선 AI 인프라 사장, 안현 개발총괄 사장 등 SK하이닉스 C-level 경영진이 행사에 참석
- HBM, eSSD, 온디바이스 AI 솔루션 등 AI 메모리 기술을 선보이며 「풀 스택 AI 메모리 프로바이더」로서의 기술 경쟁력 강조

HBM3E 16단 제품, 122TB 및 61TB 고용량 eSSD, LPDDR5X 기반 LPCAMM2, CXL과 PIM 기반 CMM-Ax 등 다양한 신제품 전시

- 세계 최초 5세대 HBM (HBM3E) 12단 제품을 양산한 SK하이닉스는 MR-MUF 공정을 적용하여 16단 구현 및 칩 휨 현상 제어, 방열 성능 향상한 제품 샘플을 전시
- 자회사인 솔리다임(Solidigm)이 작년 11월 개발한 현존 최대 용량 「D5-P5336」 122TB(테라바이트) 제품을 포함하여 고용량, 고성능 기업용 SK하이닉스의 61TB QLC 기반 SSD(eSSD, enterprise SSD) 제품도 전시
- 온디바이스 AI용 메모리인 LPCAMM2는 LPDDR5X 기반 모듈로 DDR5 SODIMM 2개를 LPCAMM2 1개로 대체
- 차세대 데이터센터 인프라인 CXL과 PIM((Processing in Memory) 기반 CMM(CXL Memory Module)-Ax, AiMX(AiM Accelerator) 전시

결론 및 시사점

- SK하이닉스는 6세대 HBM(HBM4) 양산계획을 통해 맞춤형 HBM 시장 선도하고 AI 기반 기술 혁신을 통해 AI 시대의 선도자로서 역할 강화

05

펭귄 솔루션스, SK텔레콤 및 SK하이닉스와 AI 데이터센터 협약 체결

개요

펭귄 솔루션스와 SK, AIDC 관련 전략적 협약 체결

- **펭귄 솔루션스는 SK텔레콤(이하 "SKT") 및 SK하이닉스와 포괄적인 AI 데이터센터(이하 "AIDC") 솔루션 개발 및 제공을 위한 전략적 협력 계약을 체결**
- **이번 계약은 SKT가 펭귄 솔루션스에 2억 달러를 전략적으로 투자하기로 한 12월의 결정에 따라 CES 2025 기간에 체결됨**

AI 데이터 센터를 위한 지능형 대규모 GPU 관리 기능이 강화될 것으로 기대

- 펭귄 솔루션스의 Scyld ClusterWare[1] 지능형 관리 소프트웨어를 기반으로 한 OriginAI[2] 솔루션 아키텍처를 통해 AI 컴퓨팅 클러스터의 활용률을 더욱 높임
- 펭귄 솔루션스의 Syld ClusterWare 소프트웨어와 SKT의 AI 인프라 관리 소프트웨어의 결합을 통해 펭귄 솔루션의 전문 AI 서비스 팀이 엔드투엔드로 관리할 수 있는 풀스택 AI 배포의 설치, 배포 및 최적화 기능을 확장할 수 있음
- SK하이닉스와 펭귄 솔루션스의 제품 브랜드인 스마트 모듈러 테크놀로지스는 가속화된 컴퓨팅 인프라 환경에서 효율성과 성능을 향상시키는 혁신적인 메모리 솔루션 개발을 위해 협력할 계획

결론 및 시사점

- 미래 HPC 인프라의 방향이 AI 데이터 센터로서의 기능을 강화하는 방향으로 나아가고 있음을 보여줌

1) Scyld ClusterWare: 노드 프로비저닝, 이미지 커스터마이징, 클러스터 모니터링을 포함한 지능형 관리 소프트웨어

2) OriginAI: 수백 개에서 16,000개 이상의 GPU 클러스터로 확장되는 검증된 사전 정의된 AI 아키텍처를 기반으로 구축된 AI 팩토리 인프라 솔루션

06

인공지능(AI) 칩 시장, 2025-2029년 동안 902.65억 달러 성장

개요

인공지능(AI) 칩 시장이 2025년부터 2029년까지 미화 9,026억 5천만 달러 성장할 것이라고 예측

- **성장은 AI 칩 기술의 혁신, 산업 전반에 걸쳐 AI 기반 애플리케이션에 대한 수요 증가 등 여러 요인에 의해 주도될 예정**

시장 성장 및 예측

- 전 세계 AI 칩 시장은 크게 성장하여 2029년까지 9,026억 5천만 달러에 이를 것으로 예상
- 이러한 성장은 주로 스마트폰, 자동차, 소비자 전자제품과 같은 부문에서 AI 기술 채택이 증가함에 따라 주도
- Technavio는 예측 기간 동안 약 38%의 CAGR(연간 복합 성장률)을 예측

AI 칩 혁신

- 시장 성장을 촉진하는 주요 요인 중 하나는 AI 칩의 지속적인 혁신
- AI 칩은 머신러닝, 컴퓨터 비전, 자연어 처리와 같은 작업에 점점 더 특화
- ASIC(Application-Specific Integrated Circuit), GPU(그래픽 처리 장치), TPU(텐서 처리 장치)의 개발은 특히 AI 애플리케이션의 성능과 효율성을 향상

자동차 및 엣지 컴퓨팅

- 자동차 산업은 AI 칩 수요에 또 다른 중요한 기여
- 자율주행 차량이 등장하면서 AI 칩은 센서 융합, 실시간 처리, V2X(Vehicle-to-Everything) 통신에 매우 중요
- AI 칩은 클라우드가 아닌 장치 수준에서 처리가 발생하여 더 짧은 대기 시간으로 더 빠른 데이터 처리를 제공하는 에지 컴퓨팅 애플리케이션에 필수적

결론 및 시사점

- AI 칩 시장은 칩 기술의 혁신, 스마트폰과 같은 가전제품의 AI 수요 증가, 향후 몇 년 동안 엄청난 성장을 경험할 것으로 예상되며 이로 인해 이 시장은 AI의 미래 발전을 위한 중추적인 영역이 되고 있음

07

Multiverse Computing, Kinesis와 파트너십으로 AI 최적화 및 에너지 소비 절감 추진

개요

스페인의 Multiverse Computing과 미국의 Kinesis Network Inc.가 AI 성능은 최적화하면서 자원 소비는 줄이는 혁신적인 파트너십을 체결

- **Multiverse Computing은 양자 영감 알고리즘을 활용해 대규모 언어 모델(LLM)을 최적화하는 기술을 보유하고 있으며, 특히 자사의 압축 소프트웨어인 CompactifAI는 양자 영감 텐서 네트워크를 사용하여 AI 모델의 효율성 및 성능을 향상시켰고 이는 모델 크기를 줄이는 데에 일조하여 필요한 컴퓨팅 파워를 감소시켜 결과적으로 운영 비용을 절감하고 있음**
- **Kinesis는 분산된 미사용 컴퓨팅 자원을 통합하여 효율성을 극대화하고 낭비를 최소화하는 컴퓨팅 최적화 전문 기업으로, 이들의 플랫폼은 컴퓨팅 자원을 지능적으로 할당하여 인프라 비용을 절감하고 탄소 배출량을 감소시키고 있음**
- **두 회사의 협력으로 AI 분야의 주요 과제인 고성능 AI 구현과 환경적, 재정적 비용 절감을 동시에 해결할 수 있게 되었는데, Multiverse Computing의 LLM 최적화 기술과 Kinesis의 컴퓨팅 자원 최적화 기술을 결합함으로써 기업들은 혁신을 추진하면서도 비용과 지속가능성을 모두 고려할 수 있게 됨**
- **이번 파트너십은 AI 기술의 환경 영향에 대한 산업계의 관심이 높아지는 시점에 이루어졌다는 점에서 의미가 있으며, 두 회사는 에너지 소비를 줄이고 자원 사용을 최적화함으로써 운영 효율성을 높이는 동시에 지속가능한 AI 실천의 새로운 기준을 제시하고 있음**
- **결론 및 시사점**
  - Multiverse Computing과 Kinesis가 AI 성능을 최적화하여 자원 소비를 줄이고자 하는 파트너십을 체결하였으며, 이들은 AI 워크로드가 현재 인프라가 처리할 수 있는 한계를 넘어섰다고 판단하여 이를 최적화하고 비용이나 지속가능성을 손상시키지 않은 상태로 고객에게 서비스를 지속하기 위해 노력할 것이라 밝힘

08

Supermicro, NVIDIA RTX PRO 6000 Blackwell Server Edition GPU 지원 발표

개요

Supermicro, 최신 NVIDIA RTX PRO 6000 Blackwell Server Edition GPU를 자사의 다양한 GPU 서버 및 워크스테이션 라인업에서 공식 지원한다고 발표

- 이 GPU는 AI 추론 및 파인튜닝, 생성형 AI, 그래픽 처리, 가상화 등을 위한 워크로드에 최적화됨
- Supermicro 시스템은 대부분 NVIDIA 인증을 받아 NVIDIA AI Enterprise와 호환 가능
- Supermicro는 NVIDIA H200 NVL의 2-way 및 4-way NVLink 구성도 지원하여 대규모 AI 추론 성능을 극대화할 수 있음

RTX PRO 6000 Blackwell GPU의 주요 특징

- 차세대 GDDR7 메모리 탑재, 이전 세대(L40S) 대비 2배 메모리 용량
- PCIe 5.0 인터페이스 지원: GPU와 CPU간 통신속도 향상
- MIG(Multi-Instant GPU) 기능: 하나의 GPU를 최대 4개의 독립 인스턴스로 분할 사용 가능

Supermicro에서 지원하는 NVIDIA RTX PRO 6000 블랙웰 서버 에디션

- 5U PCIe GPU 서버: 최대 10개 GPU 탑재, AI 추론 및 시뮬레이션에 적합
- MGX 모듈형 시스템: 2U에서 최대 4개의 GPU 또는 4U에서 최대 8개 GPU 지원, 산업 자동화, 과학적 모델링, HPC 및 AI 추론 애플리케이션 지원
- 3U Edge 최적화 PCIe GPU: 엣지 데이터 센터 구축용
- SuperBlade: 고밀도, 에너지 효율적인 다중 노드 아키텔처로 랙당 최대 120GPU 지원
- 랙마운트 워크스테이션: 중앙 집중식 리소스를 활용하고자 하는 조직에 적합

결론 및 시사점

- Supermicro는 최신 NVIDIA Blackwell 아키텍처를 빠르게 지원함으로써 AI 및 고성능 그래픽 시장에서의 경쟁력을 강화하고 있음

엔터프라이즈 AI 용 Supermicro GPU

출처: Supermicro

09

SoftBank 그룹, Ampere 컴퓨팅 인수

개요

SoftBank 그룹은 선도적인 실리콘 설계 회사인 Ampere 컴퓨팅을 65억 달러 규모로 인수한다고 발표

- SoftBank 그룹이 Cristal Intelligence, Stargate 등의 벤처기업에 대한 AI 인프라 투자를 확대함에 따라 이번 인수는 SoftBank 그룹의 핵심 분야에서 역량을 강화하고 성장 이니셔티브를 가속화하는데 도움을 줄 것으로 예상

Ampere 컴퓨팅은 2018년 실리콘밸리에서 설립되어 처음에는 클라우드 네이티브 컴퓨팅에 집중했지만 그 이후로 지속 가능한 AI 컴퓨팅으로 확장

- 엣지에서 클라우드 데이터 센터에 이르기까지 다양한 클라우드 워크로드를 위한 여러 제품을 보유

Ampere의 설립자 겸 CEO인 Renee James는 AI를 발전시킨다는 공통의 비전을 가지고 선도적인 기술 기업 포트폴리오에 협력하고 고성능 Arm 프로세서와 AI를 위한 AmpereOne 로드맵을 추진하게 되어 기대된다는 입장을 발표

결론 및 시사점

- SoftBank는 반도체 설계 기업인 Arm에 이어 Ampere도 인수함으로서 Arm 기반의 인공지능 생태계 확장을 주도할 것으로 예상

10

젠슨 황, NVIDIA의 인공지능 기반 미래를 설계하다

개요

젠슨 황, GTC 2025 기조연설에서 NVIDIA의 AI 관련 최신 성과와 향후 몇 년간 업계가 어떻게 발전할 것인지에 대한 전망 발표

- **캘리포니아 산호세에서 열린 GTC 2025는 AI 혁신의 급속한 가속화를 보여 주었을 뿐만 아니라, NVIDIA가 어떻게 기술 혁신의 선도자로 자리매김하고 있는지를 보여주는 행사였음**

NVIDIA의 차세대 그래픽 아키텍처인 블랙웰 울트라(Blackwell Ultra)와 베라 루빈(Vera Rubin)이 공개됨

- 올해 말 출시 예정인 블랙웰 울트라 칩을 이용해 랙 하나에 1 엑사플롭의 연산 성능(GB300 NVL72, FP4 기준)을 갖춘 시스템을 구축할 수 있음
- 2026년 말 출시 예정인 베라 루빈 칩은 블랙웰 GPU의 성능을 2배 이상 뛰어넘을 것으로 예상되며 2027년 하반기에는 루빈 울트라 칩이 출시될 예정

NVIDIA는 기존의 전통적인 데이터 센터를 AI Factory로 전환하는 것을 목표함

- AI Factory는 AI 학습과 추론을 위한 초고성능 컴퓨팅 환경으로 전통적인 데이터 센터의 다음 단계 개념
- Omniverse Blueprint를 사용하여 1GW 규모의 AI Factory를 설계 및 시뮬레이션하는 방법을 시연

결론 및 시사점

- AI 혁신 가속화를 위해 NVIDIA는 신규 GPU 출시 및 AI Factory를 설계 및 시뮬레이션하는 기술을 발전시킴
- 국내 AI 연구 환경 혁신을 가속화할 수 있는 NVIDIA GPU 기반 슈퍼컴퓨터 6호기의 중요성을 시사

11

AI 기반 지식 관리로 엔지니어링 워크플로우 혁신

개요

AI 기반 엔지니어링 제조 지식 관리의 필요성

- **데이터 폭증 문제 해결: 엔지니어링 데이터와 기술 문서의 양이 급증하여 체계적인 관리 필요**
- **중복 작업 방지: 동일 또는 유사 프로젝트 간 데이터 활용 부족으로 시간과 자원 낭비 발생**
- **협업 강화: 엔지니어 간 지식 공유와 협력이 어려워 혁신 속도 저하**

엔지니어링 제조 분야 AI 지식 관리의 주요 역할

- (데이터 중앙화 및 접근성 향상) 모든 데이터와 문서를 중앙 허브로 모아 자동 색인 및 분류를 통해 데이터 검색 속도 향상하고 실시간 데이터 통합으로 프로젝트 상태 파악 용이해짐
- (지능형 데이터 분석 및 인사이트 도출) AI가 수집한 데이터를 자동 분석하여 프로젝트 상태, 문제점, 개선 방향을 파악하고 과거 데이터를 활용하여 유사 프로젝트 위험 요소를 예측하는 등 데이터 기반 의사결정 지원
- (자동화된 규정 준수 관리) 안전 규제나 법적 요구사항을 자동 모니터링하여 위반 리스크를 최소화하여 엔지니어들이 최신 규정을 준수할 수 있도록 지원

실제 적용 사례

- 자동차 산업: 제조 공정 데이터를 AI가 실시간 모니터링하여 결함 탐지를 자동화하고 데이터 기반 결함 원인 분석으로 품질 관리 효율성 극대화
- 항공우주 산업: 수천 개 부품과 기술 문서를 중앙에서 관리하여 조립 오류를 방지하고 모델 기반 시뮬레이션을 통해 설계 오류 조기 발견
- 방위 산업: 다양한 무기 시스템 데이터 관리 및 분석을 통해 유지보수 예측 모델을 개발하고 AI를 활용해 프로토타입 성능 검증 속도 향상
- 엔지니어링 제조 관련 주요 AI 도구와 플랫폼으로는 IBM Watson Discovery, Microsoft Project Cortex, Mindbreeze Insight 등이 있음

기대 효과와 전망

- 생산성 향상: 중복 작업 방지 및 실시간 협업으로 업무 효율성 증대
- 비용 절감: 데이터 관리 자동화로 인적 자원 낭비 감소
- 의사결정 고도화: 실시간 인사이트로 전략적 의사결정 지원
- AI 기반 엔지니어링 제조 플랫폼의 발전으로 더욱 강력한 지식 자동화 가능

결론 및 시사점

- 엔지니어링 제조 현장에 존재하는 문서 기반의 지식을 AI를 적용하여 활용가능하게 됨
- 피지컬 AI와 결합하여 제조 AI 플랫폼을 통한 생산성 확대 및 효율적인 공정관리로 인간을 지원함

12

IBM, AI 추론 및 시스템 통합 확장을 위한 z17 공개

개요

IBM은 최신 메인프레임 제품인 z17을 공개하며, 전통적인 메인프레임 환경에 AI 인퍼런스(추론) 시스템의 통합 기능을 대폭 확장하는 전략 발표

- 새로운 IBM Telum II 프로세서로 구성된 z17은 이전 모델인 z16 대비 50% 더 많은 AI 추론 작업을 처리할 수 있도록 설계
- z17은 멀티 모델 AI 기능, 향상된 데이터 보호를 위한 새로운 보안 기능, AI를 활용하여 시스템 사용성과 관리를 개선하는 도구 등을 제공
- AI 추론 기능을 위해 IBM Telum II 프로세서는 내장된 2세대 온칩 AI 가속기를 탑재하고 있으며, 캐시용량이 40% 증가하여 하루 4,500억건 이상의 추론 연산 수행 및 1밀리초 이내 응답속도 제공

IBM Spyre 가속기[1]는 Telum II 프로세서를 보완하여 추가 AI 연산 능력 제공, 두 가속기를 함께 사용하면 AI 다중 모델 지원을 위한 최적 환경 구축 가능

z/OS의 차세대 버전인 z/OS 3.2[2]는 AI 어시스턴트와 에이전트를 활용, 개발자와 운영팀의 역량과 효율성을 강화하도록 설계됨

IBM Z Operation Unite[3]는 주요 성능 지표 및 로그를 OpenTelemetry[4] 형식으로 통합하여, 시스템 운영 효율성 향상 및 이상 징후 감지 시간을 단축을 통한 신속한 문제 해결 기반 제공

결론 및 시사점

- IBM z17은 메인프레임 시장에 AI 기술을 적용을 가속화하는 중요한 전환점이 될 전망
- 전통적인 메인프레임의 안정성과 보안을 유지하면서 첨단 AI 추론 기능을 통합함으로써 기업들이 보다 혁신적이면서 실시간 데이터 분석 기반의 의사 결정을 할 수 있는 환경 제공
- 특히 금융, 제조, 공공 분야 등 미션 크리티컬 시스템이 요구되는 분야에서 경쟁력 강화와 비용 효율성 증대에 크게 기여할 것으로 예상

1) IBM Spyre 가속기: 2025년 4분기 출시 예정인 PCIe 형태의 가속기
2) z/OS 3.2: 2025년 3분기 출시 예정인 IBM 메인프레임용 운영체제
3) Z Opearation Unite: 2025년 5월 출시 예정인 시스템 성능 지표 관리 솔루션
4) OpenTelemetry: 로그, 메트릭 등을 수집, 처리 및 전송하는 표준화된 도구 및 라이브러리

02 초고성능컴퓨팅 인프라 구축 동향

제1절 산업계 동향

01
NVIDIA의 블랙웰, 인공지능의 미래를 보여줌. 당장은 수냉식

개요

블랙웰[1] 프로세서의 고밀도와 열 문제로 수냉식 데이터 센터가 필수이며, AI 확산에 따라 온수 냉각 방식이 미래의 핵심 기술로 예약

NVIDIA의 블랙웰 프로세서는 매우 밀집도가 높은 GPU로 이로 인해 많은 열이 발생하는 문제해결 필요
- 높은 밀도와 열 발생으로 인해 공냉식 한계를 초과하여 NVIDIA는 이를 위해 수냉식 사양의 랙을 출시
- Dell과 Lenovo 등 주요 서버 공급업체가 수냉식 기술을 빠르게 채택 중

Lenovo는 IBM의 x86 서버 사업 인수하면서 IBM의 첨단 수냉 기술을 가지게 되어 수냉식 기술에 경쟁 우위를 확보
- IBM의 첨단 수냉 기술 인수 후, 수냉식 서버 분야 선도
- Blackwell의 Neptune 수냉 시스템을 적극 활용
- 수냉식의 전자 기기 열처리 기술에 높은 전문성 보유

집적도가 높아져 발열이 문제인 블랙웰 GPU를 설치한 서버랙에 대해 수냉식 데이터 센터 필요성 증대
- 공냉식의 한계와 블랙웰의 열 발생량 증가로 수냉식 필요성 확대
- 수냉식은 효율적 냉각, 전력 절감, 환경친화적 이점 제공
- 온수 냉각 기술 발전으로 서버 유지 비용 절감 및 부품 수명 연장

GPU 직접도가 높은 데이터 센터 냉각에 대한 미래 전망과 계획
- AI 확산에 따라 온수 냉각 데이터 센터 수요 증가 예상
- 경험 있는 설치팀 확보와 장기적 계획 수립 권장
- 미래 데이터 센터는 온수 냉각 방식 중심으로 전환될 가능성이 증가

결론 및 시사점
- NVIDIA 블랙웰 프로세서는 인공지능 기술 발전과 데이터센터의 고밀도화에 대응하여 수냉식 냉각 시스템은 필수적이며, 이는 AI 인프라의 효율성과 지속 가능성을 위해 온수 냉각 기술을 포함한 첨단 냉각 솔루션 채택의 필요성이 증대함을 보임

1) Blackwell: NVIDIA가 2024년 3월 18일에 열린 'GTC(GPU Technology Conference) 2024'에서 공개한 반도체. 2,080억 개의 트랜지스터를 탑재하고 있으며 TSMC 4NP 프로세서로 제조된 GPU 프로세서

02

AWS, 오하이오 주 데이터 센터 확장을 위해 100억 달러 투자

개요

AWS[1]는 오하이오 전역에 데이터 센터 인프라를 확장하기 위한 약 100억 달러의 추가 자금을 투자 계획 발표

- 2030년 말까지 수백 개의 새로운 저임금 일자리를 창출하고 주요 기술 허브로서의 주 역할을 강화해나갈 계획

새로운 데이터 센터에는 컴퓨터 서버, 데이터 저장 드라이브, 네트워킹 장비 및 인공지능(AI) 및 머신 러닝을 포함한 클라우드 컴퓨팅을 구동하는 데 사용되는 기타 형태의 기술 인프라가 포함될 예정

- 새로운 투자는 AWS가 작년에 발표한 78억 달러의 투자 계획을 기반으로 하며, 2015년 이후 오하이오에 대한 투자 계획은 2030년 말까지 230억 달러가 넘는 금액임
- Husted 부지사는 "인공지능과 데이터 센터는 혁신을 주도하고, 첨단 산업을 지원하며, 모든 부문에서 생산성을 향상시키고, 글로벌 경쟁력에 필수적인 방대한 데이터를 분석하고 관리할 수 있게 해주기 때문에 미국의 경제적 우위에 필수적"이라 언급

결론 및 시사점

- HPC 인프라에서 앞으로는 세상이 디지털 서비스에 더 의존적으로 될 것이기에 미래에는 데이터 센터가 인공지능과 함께 경제 및 산업 발전에 중요한 역할을 할 것이라 예측

1) AWS(아마존 웹서비스): 클라우드 컴퓨터 분야에서 세계 1위의 점유율을 차지하고 있는 아마존닷컴의 클라우드 컴퓨팅 서비스 기업

03

NVIDIA의 Blackwell, AI의 미래가 수냉식임을 시사

개요

NVIDIA, Blackwell[1]이 랙당 72개의 프로세서가 들어감에 따라 급증한 열밀도로 인해 공랭식으로는 처리가 불가능해짐에 따라 수냉식 사양의 랙 출시

- Dell은 Blackwell 전용 서버를 발빠르게 출시중
- 전통적인 수냉식의 강자인 Lenovo는 데이터센터의 수냉식 전환에 적극적이며 자신의 넵튠[2] 수냉식을 사용하여 현재 Blackwell과 관련한 수냉기술에 우위에 있음

Blackwell은 AI 산업의 발달로 엄청난 인기를 끌고 있으며 NVIDIA는 수요량을 충족하기 어려운 상황

- NVIDIA는 AI에 대한 비젼을 확신하고 지속적인 대규모 투자를 단행한 결과 가장 가치 있는 회사의 대표가 됨
- 프로세서의 발전에 따라 프로세서는 더욱 높은 발열량으로 뜨거워지고 있음

Blackwell의 발열량 과다로 인한 성능 저하, 공랭식의 경우 공기의 빠른 유속으로 인한 소음과 고온 환경으로 근무자들의 건강에 유해

- 앞으로 출시될 프로세서의 발열량 증대로 인해 수냉식 도입이 필수
- 최근 다수의 제조사가 온수 냉각(warm water cooling)[3]을 활용함으로써 서버의 전력 사용 및 물 사용 비용을 줄여 환경친화적 접근 방식 시도
- 수냉식은 메모리, 프로세서 뿐만 아니라 전원공급장치(PSU)등과 같은 기타 서버 부품으로 확대되고 있으므로 작업자의 근무환경과 서버의 내구 연한 향상에도 기여할 것으로 기대

결론 및 시사점

- Blackwell의 등장은 수냉식 도입이 필수임을 확인시켰으며, 에너지 효율적이며 환경친화적인 온수냉각(warm water cooling)이 향후 데이터센터의 중요한 설계 대안으로 고려되고 있음

1) Blackwell: NVIDIA가 2024년 3월 18일에 열린 'GTC(GPU Technology Conference) 2024'에서 공개한 반도체

2) 넵튠(Neptune): Lenovo社의 자체 수냉식 하드웨어 및 소프트웨어 냉각 기술

3) 온수냉각(Warm water cooling): 압축기를 사용하지 않고(compressorless) 냉각탑이나 dry cooler 등의 냉각장비에서 생성되는 물의 온도를 활용한 냉각

04

Eni, 자원 탐사 및 탈탄소화를 위한 1억 유로 HPC6 슈퍼컴퓨터 공개

개요

이탈리아 에너지 회사인 Eni는 석유 및 가스 탐사 기술을 확장하고 탈탄소화 및 청정 에너지 전략을 지원하는데 사용될 슈퍼컴퓨터인 HPC6를 공식 발표

- HPC6는 477.9페타플롭스 성능(Rmax)으로 TOP500 목록에서 5위에 등재
- AMD 3세대 EPYC CPU와 AMD MI250X GPU (대략 14,000개의 GPU)로 구동되는 HPE Cray EX235a이며 HPE의 Slingshot-11 패브릭을 활용
- 시추 작업, 지진 탐사 및 저수지 시뮬레이션을 통해서 데이터를 확보하고 석유 및 가스 매장 위치 및 규모, 시추 전략을 파악하는데 도움을 줄 것으로 기대

Eni는 이미 슈퍼컴퓨팅 기술을 사용하여 탄소 저장을 위한 유체 역학 및 지질학적 연구 개선, 산업 플랜트 운영 개선, 더 나은 배터리 생산, 바이오연로 공급망 효율화를 달성하였으며 HPC6를 사용하여 재생 에너지 자원의 생산 효율성을 개선할 것으로 예상

결론 및 시사점

- 자원 탐사 및 탈탄소화 산업 분야에 슈퍼컴퓨터를 활용한 사례로 보이며 고효율 에너지 확보를 위한 기반 도구로서 슈퍼컴퓨터의 활용 가능성에 대해서 주목할 필요가 있음

05

Infortrend, HP와 미디어 및 엔터테인먼트용 스토리지 공개

개요

Infortrend, HPC를 위한 통합 스토리지 발표

- Infortrend[1]는 고밀도 4U[2] 90-베이[3] HDD(하드 디스크) 솔루션인 통합 스토리지 EonStor GS 5090과 확장 인클로저 JB 4090을 소개
- 이 솔루션의 대용량 및 고처리량 성능은 고성능 컴퓨팅, 미디어 및 엔터테인먼트와 같은 애플리케이션을 위해 설계됨

EonStor GS 5090은 45GB/s 읽기, 20GB/s 쓰기, 200GbE 네트워크를 지원

- 이중 컨트롤러 설계로 중단 없는 운영을 보장
- Intel Xeon 확장가능 프로세서 사용으로 45GB/s 읽기와 20GB/s 쓰기를 달성하며, 초고속 200GbE 연결을 도입하여 데이터 전송 속도에 대한 새로운 벤치마크를 달성함
- SAS 12G[4]에 비해 처리량을 두 배로 늘린 SAS 24G 지원
- 캐시용 4개의 전용 U.2 NVMe SSD[5] 슬롯을 탑재, 90 베이 스토리지의 넉넉한 용량을 최대한 활용하면서 빠른 데이터 액세스를 제공하여 스토리지 시스템과 JBOD[6] 간의 병목 현상을 효과적으로 제거

결론 및 시사점

- HPC 인프라에서 AI와 ML 서비스가 중심이 되어가면서 대규모 데이터를 지원하기 위한 고성능 스토리지의 필요성이 증대하고 있음

1) Infotrend: 대만의 스토리지 전문 기업
2) U: 전자 장비 높이의 단위. 1U = 44.45mm(1.750 inch), 한 랙의 높이는 42U임
3) 베이(Bay): 하드디스크 장착 단위. 1베이면 하드디스크 1개를 장착할 수 있음
4) SAS 12G: 12Gbps의 전송속도를 지원하는 직렬 통신 방식의 스토리지(SAS: Serial Attached Storage)
5) U.2 NVMe SSD: U.2 sms 물리적 커넥터 표준. 데이터 센터 환경에서 고속 및 고밀도 연결 지원. NVMe SSD는 비휘발성 반도체 메모리를 나타냄
6) JBOD(Just a Buch of Disks): 단순히 디스크를 묶어 사용하는 것으로, 별다른 성능 향상 없이 단지 단위 볼륨처럼 하나로 묶어 사용하는 것

06

Seagate, 데이터 센터 구축을 위한 36TB HAMR 하드 드라이브 준비 완료

개요

Seagate, 업계 최대 용량의 HDD 발표

- Seagate[1]는 세계 최고 용량인 36테라바이트(TB)의 Exos M 하드 드라이브 샘플을 출하하여 일부 고객에게 제공
- Seagate의 열 보조 자기 기록(HAMR)[2]기술 플랫폼인 Mozaic 3+를 기반으로 하는 Exos M은 대규모 데이터 센터 배포를 위한 세계 최대 스토리지를 제공

HAMR 기술은 단위 공간 기준으로 300% 더 많은 데이터 저장 능력을 저비용으로 지원

- Exos M 스토리지는 동일한 데이터 센터 공간 내에서 300% 더 많은 저장 용량, 테라바이트당 25% 비용 절감, 테라바이트당 전력 소비 60% 감소 등 데이터 센터 운영자에게 상당한 규모, 총 소유 비용(TCO), 지속 가능성 이점을 제공

결론 및 시사점

- HPC 인프라에서 AI와 ML 사용이 증대하면서 대규모 AI/ML 데이터를 장기적으로 보관할 수 있는 대용량의 스토리지 인프라의 도입 필요성이 증대하고 있음

1) Seagate: 미국의 하드디스크 생산 전문업체

2) HAMR(Heat-assisted magnetic recording, 열 보조 자기 기록(HAMR)(발음:"해머")): 데이터 write 중에 디스크 재료를 일시적으로 가열하여 HDD와 같은 자기 장치에 저장할 수 있는 데이터의 양을 크게 늘리기 위한 자기 저장 기술.

07

Spectra Logic, 테이프 연결 확장을 위한 24G 광 SAS 스위치 도입

개요

Spectra Logic, 능동 광케이블로 테이프 저장장치 연결 거리 확장

- Spectra Logic[1], 데이터 센터 테이프 스토리지 연결성을 개선한 Spectra OSW-2400 광 SAS 스위치 출시
- OSW-2400 스위치는 능동 광케이블로 최대 100미터의 연결 거리를 지원하며 이를 통해 SAS[2] 패브릭은 데이터 센터의 공간을 최대 10,000m²까지 커버 가능

최대 1.08Tb/s 대역폭을 지원하는 가성비 높은 대용량 테이프 저장장치를 소개

- 스위치드 SAS 패브릭 채택함으로 인랙 또는 랙(Rack) 투 랙 케이블 거리를 넘어 배치된 테이프 스토리지에 대한 접근이 필요한 서버들의 광 SAS 연결이 가능해짐
- 이를 통해 저비용 SAS 테이프 드라이브를 사용할 수 있게 되었고, 더 비싼 파이버 채널 네트워크 인프라의 필요성을 줄이거나 없앨 수 있게 됨
- 직렬 연결 SCSI 4.0(SAS-4) 표준에 따르면, OSW-2400 스위치는 최대 48개의 24Gb/s 레인을 22.5Gb/s로 작동하며, 1.08Tb/s의 대역폭 또는 108GB/s의 총 데이터 전송 속도를 제공

결론 및 시사점

- 대용량의 데이터를 요구하는 AI/ML 중심의 HPC 인프라에서 가성비 높은 대규모 데이터 저장 시스템 구축을 위한 초고속 대용량의 테이프 저장 시스템이 개발되고 있음

1) Spectra Logic: 백업 및 아카이브를 위한 테이프 라이브러리 전문 기업

2) SAS(Serial Attached SCSI): 데이터 스토리지 장비와 송수신할 수 있는 직렬 프로토콜로 SCSI의 발전형

08

VDURA, Rice 대학교 에너지 HPC 컨퍼런스에서 차세대 데이터 플랫폼 공개

개요

VDURA[1] 데이터 플랫폼, 에너지 탐사 및 생산 워크플로를 통합, 가속화하는 하이브리드 아키텍처 제공

- **에너지 기업들은 에너지 비축량을 식별하고 재생 가능한 솔루션을 최적화하며 지속 가능성을 높이기 위해 대규모 데이터셋을 관리 및 분석해야 하는 과제에 직면함**
- **VDURA의 차세대 데이터 플랫폼은 혁신적인 기능을 통해 더 빠른 결과, 감소된 운영 복잡성, 타의 추종을 불허하는 데이터 내구성을 보장**

HDD의 비용 효율성과 SSD의 고성능을 동시에 지원하는 하이브리드 모델 형태의 데이터 플랫폼이 대두

- VDURA의 하이브리드 모델은 HDD의 비용 효율성과 SSD의 고성능을 결합하여 에너지 회사들이 운영 비용과 작업 속도를 최적화함
- VDURA의 고급 알고리즘은 가장 효율적이고 비용 효율적인 저장 매체에 데이터를 자동으로 투명하게 배치
- 동시성이 높은 워크로드와 방대한 데이터 볼륨을 처리하도록 설계된 VDURA 데이터 플랫폼은 온-프레미스[2], 클라우드 및 하이브리드 환경 전반에 걸쳐 원활한 통합을 제공
- 이 플랫폼은 데이터 사일로를 제거하고 정보를 통합된 글로벌 네임스페이스로 통합하여 팀 간의 접근성, 협업 및 의사 결정을 개선함

결론 및 시사점

- 대용량의 데이터를 요구하는 AI/ML 중심의 HPC 인프라에서 에너지 소비가 증가하면서 에너지 효율성과 고성능 데이터 처리를 동시에 지원할 수 있는 하이브리드 형 데이터 플랫폼이 소개됨

1) VDURA: HPC 인프라와 병렬 파일 시스템 전문 기업

2) On-Premis: 기업이 자체적으로 IT 인프라를 소유, 관리 및 운영하는 경우를 지칭

09

DOE, 차세대 주요 슈퍼컴퓨터 세부 정보 공개 - El Capitan의 동반 시스템

개요

미국 에너지부(DOE)는 차세대 슈퍼컴퓨터 ATS-5를 2027년 로스앨러모스 국립연구소(LANL)에 설치할 계획으로 이는 현재 운영 중인 Crossroads(ATS-3)의 후속 모델로, 기존 30페타플롭스(PF)급 성능을 크게 뛰어넘을 것으로 기대

- **2026년 말 시스템 납품, 2027년 2~3분기 벤치마크 테스트 진행 후 본격적인 운영 예정**
- **ATS-5는 미국 국가핵안보국(NNSA)의 핵무기 비축 유지 임무를 지원하는 대규모 3D 시뮬레이션 작업을 수행할 계획임**

ATS-5의 주요 특징

- (성능 및 에너지 소비)최대 20MW 이하의 전력을 소모하도록 요구하고 있으며, 이는 21MW를 소모하는 Frontier(세계 최초 엑사스케일 컴퓨터) 및 30MW의 El Capitan보다 낮은 수준임
- (AI 및 HPC 혼합 시스템) 인공지능(AI)과 고성능 컴퓨팅(HPC) 작업을 모두 수행할 수 있도록 설계 (CPU 전용 아키텍처도 허용)
- (모듈형 설계) 모듈형 아키텍처를 채택하여, 새로운 프로세서, 가속기, 메모리 기술을 시스템 수명 주기 동안 유연하게 추가할 수 있도록 설계
- (네트워크 및 메모리 대역폭) 이더넷 및 인피니밴드(Infiniband) 혼합 구성으로 네트워크 대역폭은 방향당 100GB/s ~ 300GB/s 수준을 목표로 하며 메모리는 9페비바이트(10.1PB) 이상의 용량을 요구함

결론 및 시사점

- 2024년 11월 세계에서 가장 빠른 슈퍼컴퓨터인 El Capitan의 동반 시스템 도입을 통해 DOE의 컴퓨팅 능력을 강화하고, 복잡한 과학 및 안보 문제를 해결하는 데 기여할 것으로 기대

10

Fluidstack, 프랑스에 1GW 인공지능 슈퍼컴퓨터 구축

개요

프랑스는 Fluidstack[1]과 함께 100억 유로를 투자해 2026년부터 1기가와트 규모의 인공지능 전용 컴퓨팅 인프라를 구축하는 MoU를 체결

프랑스 정부는 100억 유로를 투자해 2026년부터 1기가와트 규모의 인공지능 전용 컴퓨팅 인프라를 구축하고 2028년까지 이를 확장

- 최대 1기가와트 규모의 인공지능 전용 컴퓨팅 자원을 제공하여 유럽 최대의 인공지능 인프라 구축하고 2028년까지 단계적으로 확장하여 차세대 인공지능 모델 개발을 위한 독보적인 컴퓨팅 용량 확보
- 초기 투자금으로 100억 유로(약 103억 6천만 달러)를 투입하여 2026년 상업 운영 시작 예정
- 프랑스의 인공지능 인프라, 에너지 안보, 디지털 주권 분야에서의 리더십 강화를 위한 전략적 투자

원자력 에너지를 활용하여 50만 개의 AI 전용 칩을 탑재한 친환경 AGI 개발 인프라를 구축하고, 폐열 회수 시스템으로 에너지 효율을 극대화할 계획

- 저탄소 원자력 에너지를 활용하여 100% 탈탄소화된 운영 체계 구축하고 폐열 회수 시스템을 통한 에너지 효율성 극대화 및 환경 영향 최소화
- 프랑스 국영 전력망 기업 RTE와의 전략적 파트너십을 통해 안정적인 전력 공급 및 효율적인 그리드 연결 보장
- 1단계 시설에 약 50만 개의 최신 AI 전용 칩을 설치하여 최고 수준의 컴퓨팅 파워를 확보하여 AGI(인공 일반 지능) 개발 인프라를 구축

의료, 우주, 국방 등 다양한 분야의 인공지능 혁신을 주도하며, 미국, 중국과 함께 세계 3대 AI 허브로 도약하여 유럽의 디지털 주권을 확립

- 프랑스를 미국, 중국과 함께 세계 3대 인공지능 허브로 발전시키는 핵심 인프라 역할을 수행하며 유럽 내 AI 기술 주권 확보 및 글로벌 AI 연구 개발의 중심지로 도약
- 수천 개의 고숙련 AI 연구직 및 인프라 관련 일자리 창출로 지역 경제 활성화
- 의료, 우주, 국방, 대규모 언어 모델 등 다양한 분야의 AI 혁신 가속화
- 프랑스의 에너지 주권과 디지털 주권을 동시에 강화하는 시너지 효과 창출

결론 및 시사점

- 프랑스의 대규모 인공지능 인프라 구축 프로젝트는 원자력 에너지를 활용한 친환경 컴퓨팅 시설을 통해 유럽의 인공지능 주권을 확립
- 미국과 중국에 대응하는 제3의 글로벌 인공지능 허브로 도약하려는 프랑스의 전략적 비전을 보여주는 획기적인 시도

1) Fluidstack: 2017년 설립된 AI 클라우드 제공 업체로 미스트랄(Mistral), 캐릭터닷AI(Character.AI), 풀사이드, 블랙포레스트 랩스 등을 지원

11

Tier 0 스토리지가 GPU 스토리지의 판도를 바꾸는 이유

개요

GPU 스토리지로 Tier 0 스토리지가 대두

- **기술 분야에선 첨단 기술을 사용할 것을 강요받거나 서비스 트래픽 버틀넥 현상에 갇혀있거나 둘 중의 하나의 상태에 있게 되는데, Tier 0 스토리지[1]는 데이터 전송에 있어서 단순히 도로를 정리하는 것이 아니라 아우토반을 구축**
- **비효율성을 없애고 병목 현상을 해소함으로써 GPU의 진정한 힘을 발휘하게 함**
- **MLPerf1.0 벤치마크를 통해 Tier 0 스토리지는 점진적인 개선이 아니라 고속 혁명임을 증명**

Tier0 스토리지는 CPU 오버헤드와 네트워크 제약을 해결

- (CPU 오버헤드 제로) GPU 서버는 스토리지 비효율성으로 악명이 높은데 티어 0 스토리지는 리눅스 커널만 사용하여 스토리지 서비스에 대한 프로세서 이용률을 거의 0으로 줄임으로써 이 문제를 해결
- (네트워크 제약 프리) 네트워크는 대역폭 집약적인 워크로드에서 GPU 컴퓨팅의 핵심 요소인데, Tier 0 스토리지는 네트워크 의존성을 완전히 제거함
- 로컬 NVMe[2] 스토리지를 사용하면 데이터가 네트워크 파이프를 통해 크롤링될 때까지 기다리지 않고도 GPU를 최대 속도로 실행할 수 있음

결론 및 시사점

- HPC를 위한 GPU 스토리지에서 Tier 0 스토리지가 고성능 스토리지 인프라의 솔루션으로 제기되고 있음

1) Tier 0 스토리지: 다른 어떤 수준보다 빠르게 처리되는 데이터 스토리지 수준

2) NVMe: 컴퓨터의 고속 PCIe 버스를 통해 SSD와 같은 플래시 메모리 저장 장치에 있는 데이터에 빠르게 액세스할 수 있게 해주는 전송 프로토콜

12

CoolIT, AI 및 HPC를 위한 고용량 row(행) 기반 냉각 솔루션 공개

개요

북미의 액체냉각 솔류션 전문기업인 CoolIT[1]는 AI 및 HPC 시장의 엄격한 냉각 요구사항을 만족하도록 설계한 최신 CDU[2] 장비인 CHx1500 출시

- **주요 프로세서 제조업체 및 하이퍼스케일 기업과 긴밀한 협력을 통해 최대 Approach Temperature Difference(ATD)[3]가 4℃기준에서 1,360kW의 냉각부하 용량을 확보함으로써 고밀도 AI 칩 및 서버에도 충분한 성능을 제공함과 동시에 액체 냉각 솔루션 업계를 선도**

CoolIT의 CDU는 AI 및 HPC 시장에서 요구되는 고밀도 열부하를 안정적으로 처리할 수 있는 고성능 기능을 보유

- CHx1500은 최대 9개의 NVIDIA GB200 NVL72구성된 랙 냉각이 가능
- ATD 5℃ 기준으로 1,500kW의 냉각 열량 공급
- 최근 액체 냉각 솔루션이 고력되는 상황에서 냉각 용량, 유량, 압력 및 공간 측면에서 신뢰성 있는 데이터 제공

현재 액체냉각 솔루션 시장은 CoolIT를 포함한 여러 글로벌 기업들이 경쟁중

결론 및 시사점

- AI 및 HPC 환경에서 CDU를 이용한 액체냉각 기술이 핵심인프라 요소로 자리매김 할 것이며 향후 더 개선된 냉각 기술이 지속적으로 개발될 것으로 예상
- GPU기반의 AI 및 HPC 인프라는 전력 효율성, 지속가능성, 워크로드의 안정성을 유지하기 위한 액체냉각 솔루션을 필요로 하며, CoolIT를 비롯한 글로벌 기업이 다수가 경쟁하는 상황
- 액체냉각 솔루션의 확대를 통해 시설인프라의 운용 효율성 향상과 에너지 절감 효과도 기대

1) CoolIT: 2001년 캐나다에서 설립된 기업으로 데이터 센터와 데스크탑 컴퓨터 시장을 위한 고급 액체 냉각 솔루션을 제공

2) CDU: 액체 냉각(Liquid Cooling) 시스템에서 냉각수를 순환시키고 온도를 제어하며 냉각 성능을 최적화하는 장치

3) ATD: 열 교환기에서 냉각제(예: 물 또는 공기)와 냉각 대상(예: 서버, 프로세서 등) 사이의 온도 차이. 냉각수가 시스템에서 열을 제거한 수 얼마나 효과적으로 냉각됐는지를 평가하는 지표로 ATD가 작을수록 열 교환 효율이 높음

13

Oracle, 30,000개의 AMD MI355X 가속기로 AI 클러스터 구축

개요

Oracle은 최근 분기 실적 발표에서 AMD와 수십억 달러 규모의 계약을 체결하여 AMD의 차기 MI355X GPU 30,000개를 탑재한 클러스터를 구축한다고 발표

- MI355X GPU는 TSMC의 3nm 공정과 AMD의 CDNA 아키텍처를 기반으로 이번 여름에 출시될 예정으로 288GB의 HBM3E 메모리, 최대 8TB/s의 대역폭, FP6 및 FP4 저정밀 컴퓨팅 지원을 특징으로 엔비디아의 블랙웰 B100 및 B200 GPU의 경쟁자로 자리매김할 예정

Oracle 회장 겸 CTO인 래리 엘리슨은 실적발표 컨퍼런스 콜에서 5,000억 달러 규모의 AI 인프라 투자에 해당하는 스타게이트 프로젝트에는 AI 훈련을 위한 64,000개의 GPU 액체 냉각 NVIDIA GB200 클러스터가 포함될 예정이라고 언급

- 스타게이트 프로젝트는 1월말 백악관에서 트럼프 대통령, Oracle 회장 겸 CTO인 래리 앨리슨, OpenAI CEO인 샘 알트만, SoftBank 창립자 겸 CEO인 마사요시 손이 발표한 대규모 AI 데이터 센터 프로젝트

결론 및 시사점

- 대규모의 GPU로 구성된 AI 클러스터 구축이 전세계적으로 진행되고 있고 AI 기술을 둘러싼 국가 간 경쟁이 더욱 심화할 것으로 예상

제2절 학계, 연구계 동향

01

HLRS: 동적 전력 캡핑으로 HPC에서 더 나은 에너지 효율성 실현

개요

HLRS[1]의 Hawk 슈퍼컴퓨터에 동적 전력 캡핑을 적용하여 애플리케이션의 성능 저하 없이 전체 전력 소비를 20% 줄일 수 있음을 확인

- 동적 전력 캡핑(Dynamic Power Capping)은 시스템의 전력 소비를 실시간으로 조절하여 특정 수준 이하로 유지하는 기술이며, 이를 통해 불필요한 에너지 소비를 줄이고 성능 저하를 최소화
- 전력 절감액은 약 1,500채의 단독 주택의 연간 전력 소비량과 유사함

HPE가 HLRS와 협력하여 개발한 지능형 전력 관리 솔루션은 전력 예산 내에서 시스템 성능을 최적화 하기 위해 슈퍼컴퓨터 전체에 전력 분배를 조절

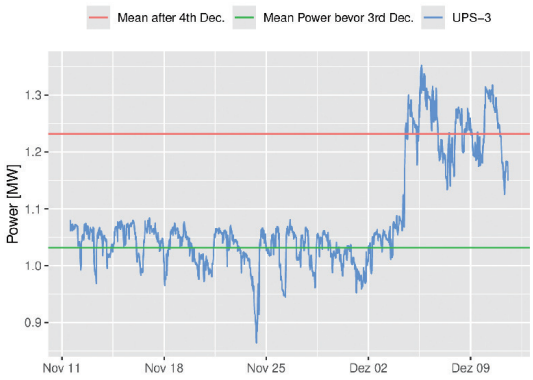
- 컴퓨팅 바운드 코드와 메모리 바운드 코드 간에 사용 가능한 전력을 균형 있게 조정하면 전체 시스템 전력 사용량의 갑작스러운 급증과 감소도 줄어들어 HLRS의 전력 소비 목표를 준수하는 일관되고 안정적인 전력 수준을 유지할 수 있음
- '24년 10월 HLRS는 "혁신" 부문에서 데이터센터 전략상을 수상

HLRS와 HPE는 동적 전력 캡핑을 HLRS의 차세대 GPU 기반 슈퍼컴퓨터인 Hunter와 Herder로 확장하기를 기대

- '27년 도입 예정인 엑사스케일 슈퍼컴퓨터 Herder의 경우 운영 비용이 많이 들기 때문에 동적 전력 제한을 완벽하게 구현하는 것이 특히 중요함

결론 및 시사점

- 동적 전력 캡핑 기술은 에너지 효율성 향상을 위한 혁신적인 접근 방식으로 주목받기 시작했으며, 엑사스케일 슈퍼컴퓨터에서 급증하는 전력 비용의 절감을 통해 지속 가능한 슈퍼컴퓨터 운영을 가능하게 할 것으로 예상됨



동적 전력 제한을 통해 달성한 약 20%의 에너지 절감 효과
(빨간선) 비제한 모드에서의 편균 전력 사용량, 녹색선)
동적 전력 제한 적용한 평균 전력 사용량

출처: HPE/HLRS

1) HLRS(HIgh-Performance Computing Center Stuttgart): 독일 슈트트가르트 대학교에서 운영하는 슈퍼컴퓨팅센터

02

Cineca, 이탈리아에서 가장 강력한 양자 컴퓨터 IQM Radiance 54 도입

개요

이탈리아 슈퍼컴퓨팅센터인 Cineca는 IQM Quantum Computers와의 협약 체결을 통해 IQM Radiance 54 양자 컴퓨터 설치 계획을 발표

- 54큐비트 풀 스택 초전도 양자 컴퓨터인 IQM Radiance는 이탈리아 볼로냐에 위치한 세계에서 가장 빠른 슈퍼컴퓨터 중 하나인 Leonardo에 통합될 예정
- IQM Radiance는 2025년 4분기에 배송 및 설치될 예정이며, Cineca에서 설치되는 첫 번째 양자 컴퓨터
- IQM은 이 시스템을 통해 이탈리아의 연구자들에게 최신 양자 컴퓨터 플랫폼과 도구를 제공하며, 복잡한 과학적 문제를 해결할 수 있도록 지원
- Cineca는 이 시스템을 양자 응용 최적화, 양자 암호화, 양자 통신 및 인공지능 양자 알고리즘 연구에 활용할 계획임

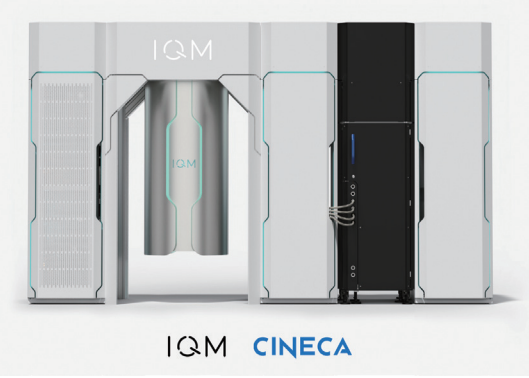
Leonardo 시스템은 EuroHPC JU1)의 자금 지원을 받아 개발되었으며, 2022년 11월부터 운영이 시작됨

- 이 시스템은 현재 세계에서 TOP500 리스트에서 9위에 랭크

IQM은 2018년에 설립된 핀란드의 초전도 방식의 양자 컴퓨터 기업으로, HPC와 연구소, 대학, 기업에 풀 스택 양자 컴퓨터와 응용 프로그램을 제공

결론 및 시사점

- 이탈리아의 Cineca와 IQM의 협약은 양자 컴퓨팅 분야에서 이탈리아의 입지를 확고히 하고, 전 세계 양자 기술 생태계에서 중요한 발전을 의미
- IQM Radiance 양자 컴퓨터는 고전 컴퓨터로는 불가능한 문제들을 해결할 수 있는 잠재력을 지니고 있어, 연구자들에게 새로운 가능성을 열어줄 수 있음



출처: IQM

1) EuroHPC JU: 유럽 내에서 슈퍼컴퓨터를 구축, 운영함과 동시에 HPC 기술연구 및 혁신을 촉진하기 위해 설립된 유럽연합 공동 기구

03

Pasqal, EuroHPC 조달 계약에 따라 EuroQCS-Italy 양자 시스템 제공

개요

EuroHPC JU(유럽 고성능 컴퓨팅 조인트 사업)와 프랑스 양자 컴퓨팅 기업 Pasqal이 EuroQCS-Italy 양자 컴퓨터 조달 계약을 체결

- 총 1,300만 유로(약 200억 원) 규모의 프로젝트로, EuroHPC JU와 이탈리아 정부가 각각 50%씩 자금 부담
- EuroHPC JU는 2023년 체코, 독일, 스페인, 프랑스, 이탈리아, 폴란드 등 유럽 6개국에 양자 컴퓨터를 구축하는 계약을 체결한 바 있음

EuroQCS-Italy는 중성 원자 기반의 양자 시뮬레이터로, 이탈리아 볼로냐의 CINECA에서 호스팅 및 운영될 예정

- 양자컴퓨터 설치는 2025년부터 시작되어 초기에는 최소 140 큐비트를 지원하며 아날로그 모드로 작동함
- 알고리즘 활용사례 확대를 위해 2027년 하이브리드 아날로그/디지털 모드로 업그레이드될 예정
- CINECA에 호스팅될 이 시스템은 Leonardo(TOP500 9위) 슈퍼컴퓨터와 양자-고전 하이브리드 시스템으로 통합될 예정
- 이 시스템은 과학계, 산업계, 공공 부문 등 유럽 사용자에게 광범위하게 제공될 예정이며 양자 다체계(many-body) 물리학, 최적화 문제, 머신러닝 등의 연구를 지원할 것임

결론 및 시사점

- EuroHPC의 사전 엑사스케일 시스템 “Leonardo”와 통합되어 양자-고전 하이브리드 컴퓨팅 기술 발전에도 기여할 전망
- 유럽은 다양한 양자 기술 및 하이브리드 구조를 도입하며 미국, 중국과 함께 양자 컴퓨팅 선도국으로 도약하는 전략을 추진중에 있음
- 이번 프로젝트를 통해 유럽 내 양자 기술의 상업적·학문적 활용이 본격화되며, 연구자 및 기업이 실질적인 성과를 도출할 수 있는 환경이 마련됨

04

QpiAI, 인도 국가 양자 미션의 일환으로 25큐비트 초전도 양자 시스템 출시

개요

인도 테크기업인 QpiAI는 25큐비트 초전도 양자컴퓨터 'QpiAI-Indus'를 출시하였으며, 이는 인도 내에서 가장 강력한 양자 시스템 중 하나로 평가됨

- QpiAI-Indus는 양자 프로세서, 고성능 컴퓨팅(HPC) 연동 소프트웨어 스택, 그리고 AI 기반 하이브리드 최적화 도구 등 3가지 주요 요소로 구성됨
- T1과 T2 시간은 각각 30μs, 25μs이며, 향후 100μs 이상을 목표로 하고 있음. 단일 및 이중 큐비트 게이트 정확도는 각각 99.7%, 96%에 도달
- QpiAI는 향후 64 큐비트 시스템으로 확장 예정이며, 장기적으로는 오류보정 기반 FTQC(Fault-Tolerant Quantum Computing) 시스템을 목표로 하고 있음
- 응용 분야는 물류 최적화, 신약 개발, 소재 설계, 기후 모델링 등이며, 이와 관련된 제품군으로 QpiAI-Logistics, QpiAI-Pharma, QpiAI-Matter 등을 운영

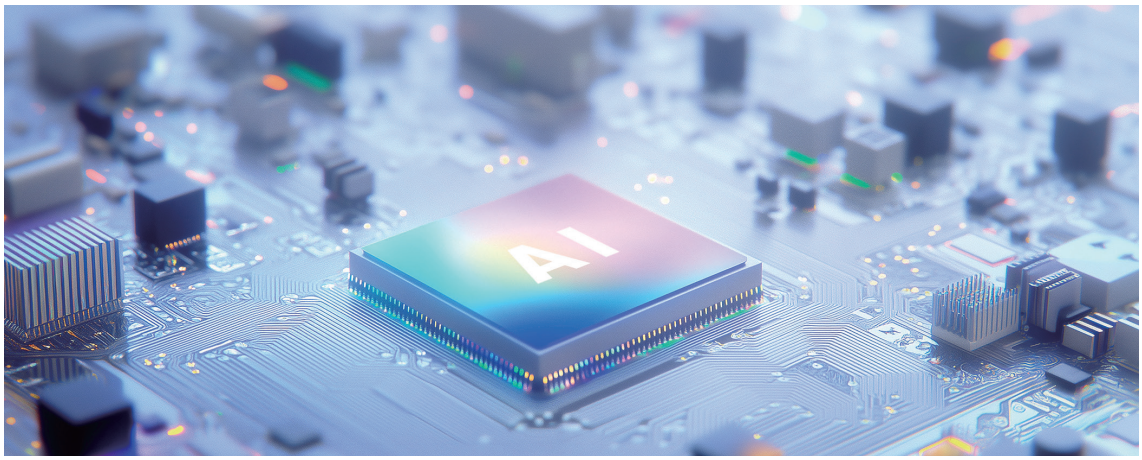
이번 발표는 2025년 4월 14일 '세계 양자의 날'에 공식 공개되었으며, 인도의 국가 양자 미션(National Quantum Mission, NQM)의 일환으로 진행됨

- NQM은 2023~2031년 동안 약 7억 4천만 달러 규모로 운영되며, 인도 내 양자 컴퓨팅, 통신, 센서, 측정, 소재 분야의 기술 자립과 인프라 구축을 목표로 함
- NQM은 벵갈루루, 마드라스, 뭄바이, 델리의 4개 테마 허브(T-Hub)를 통해 연구와 산업 간 협업 생태계도 구축 중임

결론 및 시사점

- QpiAI의 25큐비트 양자컴퓨터 출시는 인도의 국가 양자 미션(NQM) 전략에 따라 양자 기술 국산화를 실현하는 중요한 진전임
- AI와 양자기술의 융합, 그리고 수직 통합형 플랫폼 전략은 향후 양자 응용 산업 전반에 걸친 파급 효과를 가질 것으로 전망됨
- 이번 개발은 인도가 글로벌 양자 기술 경쟁에서 전략적 우위를 확보하려는 노력의 일환으로, 중장기적으로 1000큐비트급 시스템으로의 기반 마련

GLOBAL HPC HORIZON



03 초고성능컴퓨팅 기술개발 동향

제1절 산업계 동향

01

Meta, 2030년까지 루이지애나 북동부에 100억 달러 규모의 AI 데이터 센터 건설

개요

Meta는 루이지애나 북동부에 새로운 인공지능 데이터 센터에 100억 달러를 투자할 계획이라고 발표

- 리치먼드 교구의 2,250에이커 부지에 위치하며 500개의 직접 일자리와 1,000개 이상의 간접 일자리를 창출할 것으로 예상
- 루이지애나주 역사상 가장 큰 민간 투자 중 하나로서 루이지애나 북동부 전역에 새로운 경제 활동과 투자를 촉진할 것으로 기대
- Meta는 전기 사용량을 100% 깨끗하고 재생가능한 에너지로 맞추겠다고 약속했으며 Geaux Zero 프로그램을 통해 최소 1,500MW의 새로운 재생 에너지를 공급하기 위해 Entergy와 협력할 예정
- 루이지애나 커뮤니티 및 기술 대학 시스템(LCTCS)는 델타 커뮤니티 대학과 협력하여 데이터 센터의 건설과 최종 운영을 위한 커리큘럼을 개발하여 제공할 예정
- Meta는 Entergy의 저소득층 납세자 지원 프로그램에 연간 최대 100만 달러를 기부하기로 약속했으며 도로와 수계 시스템을 포함한 지역 인프라 개선에도 2억 달러 이상을 투자하기로 약속
- Meta는 데이터 센터 장비 구매에 대한 주 및 지방 판매세 환급 혜택을 제공하는 루이지애나 인센티브 프로그램을 활용할 것으로 예상되며 주정부의 Quality Jobs 프로그램에도 참여할 것으로 예상

결론 및 시사점

- 인공지능의 급속한 성장으로 인한 거대 기업들의 초거대 데이터 센터 건설이 확산되고 있으며 초거대 데이터센터 건설을 위한 전력 확보 전략, 지역과의 협력을 통한 경제성장 및 인력양성 방안 등에 대해서 주시할 필요가 있음

02

AWS, HPC 서버용 대규모 인메모리 데이터베이스 EC2 U7inh 인스턴스 출시

개요

AWS는 HPE와 협력하여 EC2 고성능메모리 새 제품인 Amazon EC2 U7inh 인스턴스를 발표하였으며 해당 인스턴스는 16소켓 HPE Compute Scale-up Server 3200에서 실행되고 AWS Nitro System에 구축되어 제공함

- 4세대 Intel Xeon Scalable 프로세서 (Sapphire Rapids)
- 32TB 메모리, 1920개 vCPU

U7inh 인스턴스는 가장 큰 U7i 인스턴스의 2배 되는 vCPU와 1.6배의 EBS 대역폭 제공함

U7inh 인스턴스는 Amazon Linux, Red Hat Enterprise Linux, SUSE Enterprise Linux Server 지원함

Instance Name	vCPU	Memory (DDR5)	EBS Bandwidth	Network bandwidth
U7inh-32tb.480xlarge	1,920	32,768 GiB	160Gbps	200Gbps

U7inh 인스턴스의 프로덕션 환경에서 Business Suite on HANA (SoH), Business Suite S/4 HANA, Business Warehouse on HANA (BW), SAP BW/4HANA에 대한 SAP인증 받음

클러스터에 최대 4개의 U7inh 인스턴스 (128TB)를 배포가능함

결론 및 시사점

- U7inh 인스턴스의 출시는 클라우드 기반 대규모 워크로드 처리의 발전을 보여주고 있고 SAP 중심의 시장을 확대하고 대규모 데이터 분석 작업을 보다 효율적으로 처리할 수 있어 고성능 컴퓨팅 및 데이터 분석에서 실시간 비즈니스 인사이트를 도출하여 경쟁력을 확보함

03

OpenAI, 2024년 12월 20일 새로운 AI 모델 'o3'와 'o3-mini' 발표

개요 **OpenAI, 2024년 12월 20일 새로운 AI 모델「o3」[1]와「o3-mini」[2] 발표**

논리적 추론 능력 강화: ARC-AGI 벤치마크 성능: 다양한 활용 분야
- 단계적 문제 해결 및 맥락 이해 능력 향상
- 다양한 수학 및 논리적 도전과제를 해결하는 데 탁월한 성과
- 인공지능 추론 능력을 평가하는 ARC-AGI에서 이전 모델보다 3배 높은 정확도 달성
- 코딩 문제 해결, 복잡한 수학 공식 계산, 과학적 데이터 분석 등에서 좋은 성능을 보여 작업 환경에서 유용성을 검증

「o3」와「o3-mini」에 대한 내부 테스트 진행 중으로 조만간 출시 예정
- 모델의 안전성 및 성능 확인을 위해 제한된 외부 연구자 그룹에게 접근 권한을 제공하여 테스트 중
- o3-mini는 2025년 1월 말 공개 예정이며 전체 o3 모델은 이후 단계적으로 출시될 예정

「o3」와「o3-mini」은 추론 기술의 발전, 효율성 개선, 하드웨어 호환성 증대를 통해 기존 모델과 기술적 차별성 확보
- 단계적 프로세스를 통해 복잡한 질문을 나누고, 세부적으로 해결하는 능력 추가
- 기존 모델 대비 처리 속도 및 에너지 효율성을 개선
- 클라우드 및 엣지 디바이스에서도 원활히 작동 가능하도록 설계

OpenAI는 추론 모델에 대한 지속적인 혁신과 확장을 추구
- OpenAI는 사용자의 피드백과 테스트 데이터를 통해 모델을 지속적으로 개선해나갈 계획
- AI 모델의 활용 영역을 더욱 확장하여 산업 및 학술 연구 등 다양한 분야에서의 응용을 목표로 모델 개발 및 개선 추진 중

결론 및 시사점
- OpenAI의「o3」와「o3-mini」는 추론 능력과 효율성을 크게 개선하여 다양한 산업 및 학술 분야에서 활용 가능성을 확장, AI 발전을 선도

1) o3: 이전 'o1' 모델의 후속작으로, 복잡한 문제를 논리적으로 해결하는 능력이 대폭 개선된 모델

2) o3-mini: o3의 경량화된 버전으로 빠른 처리 속도와 낮은 하드웨어 요구 사양이 특징

04

Microsoft, Hugging Face에 Phi-4 언어 모델 출시

개요

Microsoft가 효율성과 정확성을 강조하는 소형 언어 모델인 AI의최신 언어 모델 Phi-4를 Hugging Face 플랫폼에 공개 (under the permissive MIT licence)

Phi-4의 주요 특징

- (소형 및 에너지 효율성) 경량 아키텍처로 소비자 하드웨어에서 작동 가능, 대규모 서버 인프라 불필요하고 에너지 소비 절감으로 지속 가능성과 그린 컴퓨팅에 부합
- (수학적 추론 능력) MATH 벤치마크에서 80.4점 획득하였고, 수학적 계산 및 추론에 뛰어난 성능 발휘 (금융, 엔지니어링, 데이터 분석 분야)
- (특화된 응용 분야) 큐레이션된 데이터셋으로 훈련되어 도메인별 활용도 높음 (헬스케어, 고객 서비스 등의 분야)
- (향상된 안전성) Azure AI의 콘텐츠 안전 도구를 활용해 유해한 프롬프트에 대응 가능하며 실시간 배포에 적합

공개와 활용

- (접근성) Hugging Face 플랫폼을 통해 공식적으로 공개 (MIT licence)
- (AI 진입 장벽 완화) 효율적 비용과 컴퓨팅 리소스가 제한된 환경에서도 활용 가능
- (혁신적인 훈련 기술) 합성 데이터와 유기적 데이터의 조합으로 훈련, 데이터 가용성 문제 해결하였고, 향후 확장 가능성과 정밀도 증가를 위한 기반 마련함

결론 및 시사점

- Phi-4는 고성능 AI 모델이 자원 제한이 있는 환경에서도 활용 가능하여 산업 및 학계의 혁신을 가속화
- 대규모 언어 모델에 대한 접근성을 높여, 다양한 기관과 사용자가 AI 기술을 보다 쉽게 채택하고 응용할 수 있는 환경 조성
- 오픈소스 및 라이선스를 통해 AI 커뮤니티 내 혁신과 발전 가능성 제고

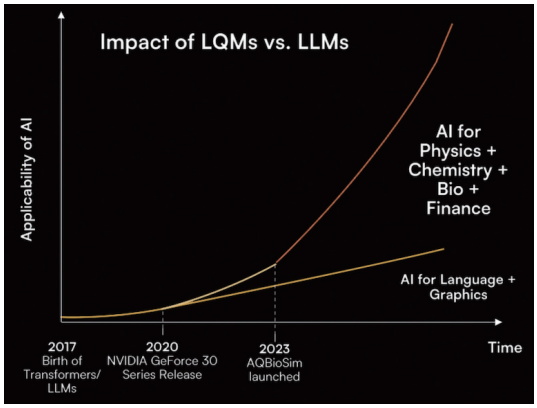
05

SandboxAQ가 과학과 의학에서 조용한 혁명을 주도하는 방법

개요

**SandboxAQ는 2022년 3월 Alphabet의 AI 양자 컴퓨팅 부서에서 독립한 스타트업으로,
현재 AI 시장에서 대규모 언어 모델이 주목받고 있지만
이들은 물리학 기반 방법으로 생성된 데이터로 훈련된 대규모 정량 모델(LQM)을 개발하고 있음**

- **정량적 AI는 기존에 금융 서비스 분야에서 주로 사용되었으나, SandboxAQ는 이를 과학 및 기술 분야로 확장하고 있는데 언어가 아닌 숫자, 방정식, 물리학, 화학의 언어를 통해 분자를 이해하고 모델링하는 방식을 채택하여 적용하고 있음**
- **신약 발견 분야에서 SandboxAQ의 LQM은 단백질에 효과적으로 결합하는 작은 분자를 식별하는 데 주력하고 있는데, 이를 위해 실험 데이터와 전산 화학 기술을 결합하여 사용하며 속도와 정밀도를 위해 자체 개발한 도구들을 활용하고 있음**
- **기술적 측면에서 SandboxAQ는 트랜스포머 아키텍처와 함께 텐서 네트워크를 활용하고 있고, 특히 텐서 네트워크는 물리적 시스템 모델링에 특화되어 있으며 GPU를 사용하여 양자 컴퓨터 없이도 복잡한 원자 및 분자 시스템을 모델링할 수 있음**
- **SandboxAQ는 고성능 컴퓨팅, AI, 빅데이터의 시너지를 활용하여 지속적인 발전을 이루고 있으며 특히 화학 및 재료 과학 분야에서는 각각 고유한 특성을 가진 무수한 시스템을 모델링할 수 있는 적응형 접근 방식을 채택하고 있음**
- **SandboxAQ는 최근 Google Cloud와의 파트너십 체결을 통해 LQM 플랫폼을 Google Cloud Marketplace를 통해 제공할 예정이며 향후에는 반자율적으로 작동하는 정량적 AI 에이전트 개발을 목표로 하고 있음**
- **결론 및 시사점**
  - 정량적 AI를 통한 SandboxAQ의 혁신은 기존 언어 모델 중심의 AI 산업과 달리, 과학, 의료, 신약 개발 등 실제 산업 분야에서 실질적인 변화를 이끌고 있으며, Google Cloud와의 파트너십을 통해 이러한 혁신 기술의 상용화가 가속화될 것으로 전망됨



LQM과 LLM의 영향

출처: SandboxAQ

06

Altair, 인공지능 기반 스케줄링이 가능한 클라우드 플랫폼 HPC웍스 업그레이드 출시

개요

Altair는 최근 자사의 고성능 컴퓨팅 및 클라우드 플랫폼인 「Altair HPC웍스」를 업그레이드하여 출시

Altair HPC웍스는 클라우드 확장, 고급 모니터링, 인공지능 기반 작업 스케줄링 기능 제공

- Altair 및 타사 워크로드 관리자를 위한 클라우드 확장을 확장하고, 고급 모니터링 및 보고 기능을 통합
- 향상된 인공지능 지원 작업 스케줄링 및 시각화와 GPU, Kubernetes, 머신러닝 워크플로우를 지원

Altair HPC웍스는 유닛 라이선스 시스템을 통해 유연한 동적 라이선싱을 제공하여 기업은 온프레미스 클러스터를 효율적으로 운영

- Altair HPC웍스는 이제 시장의 표준인 알타이어 유닛 라이선스 시스템 내에서 운영하여 유닛 시스템의 유연하고 확장 가능한 동적 라이선싱 가능
- 기업은 Altair의 라이선싱 시스템을 활용하여 Altair HPC웍스 플랫폼과 Altair 원 게이트웨이를 원활하게 연결하여 온프레미스 클러스터를 활성화 가능

Altair HPC웍스는 인공지능 기반 스케줄링, 광범위한 기술 지원, HPC 및 클라우드 모니터링을 통해 AI 워크로드 최적화와 IT 자원 운영을 효율화

- Altair HPC웍스는 인공지능 통합을 통해 작업 제출을 간소화하고 대기 시간을 줄이며, RapidMiner와의 연계를 통해 스마트한 스케줄링을 지원
- GPU와 Kubernetes를 포함한 광범위한 기술을 지원하여 AI 워크로드를 최적화하며, 상세한 HPC 및 클라우드 모니터링 기능을 제공해 IT 관리자의 효율적인 자원 운영을 지원

결론 및 시사점

- Altair는 자사의 HPC 및 클라우드 플랫폼인 HPC웍스의 업그레이드 발표를 통해 기업이 HPC 및 인공지능 워크로드를 보다 효율적으로 운영할 수 있도록 지원

07

세계 최초로 100% AI 생성 논문이 세계 탑 AI 학술대회 워크샵 동료평가 통과, Sakana ai「AI Scientist」가 달성

배경 **2024년 8월 Sakana ai는 AI Scientist 개발을 발표. AI Scientist는 과학논문 연구의 전과정을 자동화하는 시스템으로 연구 아이디어 생성부터, 실험, 논문작성, 피어 리뷰까지 수행 가능**

개요 **2025년 3월, 개선된'AI Scientist-v2'가 생성한 머신러닝 분야의 논문이 국제학회 ICLR 2025('25. 7. 13 ~ 19) 워크샵 세션의 더블 블라인드 리뷰 과정을 통과**

- **논문 작성을 위해 사람은 워크샵 테마에 맞는 대략적인 주제만 제공하였고 'AI Scientist-v2'가 가설 입안, 실험 설계, 코드 작성, 실험 실행, 데이터 분석 및 시각화, 논문 서식 설정과 집필의 모든 과정을 자동으로 실시함**
- **리뷰 과정은 ICLR[1] 주최자와 워크숍 관계자의 전반적인 협력하에 진행되었고, 브리티시 컬럼비아 대학에서 IRB(연구윤리심사위원회)의 승인도 받음**

리뷰 과정 및 결과

- AI Scientist-v2가 주어진 대략적인 연구 주제만을 바탕으로 논문(「Compositional Regularization: Unexpected Obstacles in Enhancing Neural Network Generalization」) 생성
- 해당 논문은 신경망의 구성적 일반화 능력을 높이기 위해 새로운 정규화 기법을 검토했을 때 발생한 장애에 대해 보고한 것으로, ICLR 워크숍에서 채택 수준을 웃도는 평균 검토자 점수 6.33을 획득
- ICLR 워크샵의 논문 리뷰 기준과는 별개로 회사에서 자체적으로 AI Scientist-v2가 생성한 3편의 논문에 대해 최고 수준 학술대회 본선의 논문 리뷰 기준으로 평가를 진행하며 저자(AI Scientist-v2)와 논문을 개선해 나감. 3편 모두 이 기준을 통과하지는 못했지만 향후 발전의 여지가 있는 것으로 판단됨
- AI 분야 최상위 학술대회 논문의 채택율 기준을 상회하는 품질의 논문을 생성하는 것이 과제이자 도전 테마

논문 쓰기는 과학 커뮤니티의 대표 활동으로서 윤리적 행동규범을 가지고 있음

- 전자동 AI 생성논문 작성이 가능해진 작금의 현실에서 언제 어떻게 이를 허용하고 활용할 것인지에 대한 논의가 필요(공개 여부와 바이어스된 판단, 리뷰 통과만을 위한 교정 등)

결론 및 시사점

- 과학 커뮤니티에 대규모 생성형 AI의 적절한 활용의 발전 가능성에 대한 주의 환기 및 인간과의 협력을 통한 최고 수준으로의 도약 가능성을 제시
- AI는 과학적 발견을 가속화하는 도구 또는 파트너로 진화할 가능성이 크며, 인간과 AI가 협력하여 인류의 번영과 지식의 확장을 위한 AI 과학 시대로의 발전을 위해 그 방향성에 대해 논의가 시작돼야 할 중요한 시점임을 시사

1) ICLR: NeuriPS, ICML와 더불어 기계 학습 및 AI 연구 분야 세계 3대 국제 학회

08

대형 언어 모델(LLM) 추론 최적화

개요

LLM 서비스의 확대로 인해 추론 기능의 최적화가 요구됨

- **LLM은 거대한 파라미터 수로 인해 연산 부담이 커 응답 지연과 운영비용 증가 문제가 발생함**
- **서비스 운영 시 사용자 요청이 폭증할 경우 추론 비용이 급증하여 수익성에 영향을 미침**
- **실시간 응답성이 중요한 애플리케이션에서는 고비용 구조가 비즈니스 모델 지속 가능성을 위협**

추론 최적화의 목적

- 운영 비용 절감: 계산 리소스 효율화로 비용 절감
- 성능 최적화: 모델 응답 속도 향상으로 사용자 만족도 증대
- 환경적 지속 가능성: 전력 소비 감소를 통한 탄소 발자국 저감
- 확장성 확보: 사용자 증가에도 안정적 서비스 제공

핵심 최적화 기술 방법론

- 양자화(Quantization): 부동소수점(32비트)을 정수형(8비트)로 변환하여 연산 속도와 메모리 효율 향상하여 모델 크기 감소, 메모리 사용량 절감, 전력 소모 감소함, GPT-3 등 초거대 모델에서 INT8로 양자화하여 처리 속도 2배 향상
- 지식 증류(Knowledge Distillation): 대형 모델의 성능을 경량화된 작은 모델로 전이하여 큰 모델 수준의 성능을 작고 빠른 모델로 구현, BERT를 경량화한 DistilBERT, OpenAI의 경량 모델들
- 배치 처리(Batch Processing): 여러 요청을 묶어 한 번에 처리하여 병렬성을 향상하여 단일 추론 비용 감소와 처리량 증가 달성, 챗봇 서비스에서 다수의 사용자 요청을 하나의 배치로 묶어 처리
- 하드웨어 최적화(Hardware Optimization): TPU와 같은 전용 칩을 활용하여 모델 연산을 가속화하여 GPU 대비 더 높은 연산 밀도로 속도 향상, Google의 TPU 및 NVIDIA의 TensorRT를 통한 추론 성능 극대화
- 캐싱(Caching): 반복 계산을 캐시에 저장하여 재사용하여 중복 연산 제거로 속도 향상

※ 적용 사례: 긴 대화 세션에서 이전 대화 히스토리를 캐싱하여 다음 응답에 활용

LLM 추론 최적화의 도전 과제

- 성능 저하 문제: 양자화나 경량화 과정에서 성능이 저하될 위험
- 모델 정확성 저하: 지나친 경량화로 인해 성능이 낮아질 우려
- 하드웨어 한계: TPU나 GPU 간 성능 격차로 최적화가 어려운 경우

결론 및 시사점

- 하드웨어와 소프트웨어 동시 최적화가 추세로 자리잡고 있음
- 자동화된 최적화 도구 개발이 활발하여 점점 더 쉽게 최적화를 수행할 수 있을 전망
- 향후 친환경 AI 요구가 커지며 에너지 효율성을 중시하는 연구가 지속될 것으로 예상됨

09

Fujitsu, 오픈소스 기반 양자 컴퓨터 운영 소프트웨어 출시

개요

Fujitsu는 오사카 대학, Systems Engineering Consultants 주식회사, TIS사와 협력하여 오픈소스 양자 컴퓨터 운영 소프트웨어인 'Open Quantum Toolchain for Operators and Users(OQTOPUS)'를 출시

- 이 소프트웨어는 GitHub(https://github.com/oqtopus-team)에서 제공되며, 양자 컴퓨터 설정부터 클라우드 운영까지 전 과정을 지원
- 기존에는 클라우드 기반 양자 컴퓨팅 환경을 구축하려면 복잡한 운영 시스템을 개발해야 했으나, OQTOPUS을 통해 이 과정을 간소화할 수 있음
- OQTOPUS는 자원 할당, 작업 스케줄링, 시스템 모니터링, 사용자 접근 관리 등의 기능을 제공하며 다양한 양자 하드웨어 플랫폼과 호환되도록 설계됨
- Fujitsu는 2025년 하반기부터 이 소프트웨어를 자사 클라우드 기반 양자 서비스에 통합할 계획이며 지속적으로 업데이트할 예정
- OQTOPUS 출시를 통해 Fujitsu는 양자 컴퓨팅 생태계에서 하드웨어 공급자이자 소프트웨어 제공자로서의 입지를 강화

결론 및 시사점

- Fujitsu의 오픈소스 양자 컴퓨터 운영 소프트웨어 출시는 양자 컴퓨팅의 대중화와 클라우드 기반 양자 서비스의 확산을 촉진하는 중요한 움직임
- 연구소와 기업이 복잡한 운영 소프트웨어를 직접 개발할 필요 없이 쉽게 양자 컴퓨팅 환경을 구축할 수 있어 기술적 진입 장벽이 낮아짐
- 클라우드와 양자 하드웨어의 통합이 가속화되면서 양자 컴퓨팅 활용 분야가 확대될 것으로 기대됨
- Fujitsu는 하드웨어뿐만 아니라 소프트웨어 부문에서도 영향력을 확대하며 글로벌 양자 컴퓨팅 시장에서 경쟁력을 강화하고 있음

10

UALink 컨소시엄, Ultra Accelerator LInk 200G 1.0 사양 발표

개요

UALink 컨소시엄은 AI 컴퓨팅 파드(pod) 내 가속기와 스위치 간 통신을 위한 저지연, 고대역폭 상호 연결을 정의하는 UALink 200G 1.0 사양의 비준을 발표

- **UALink 1.0 사양은 AI 컴퓨팅 파드(pod) 내 최대 1,024개의 가속기에 대해 레인당 200G의 스케일업 연결을 지원하여 차세대 AI 클러스터 성능을 위한 개방형 표준 상호 연결을 제공**

UALink의 주요 이점

- 고성능: PCIe 스위치의 지연 시간을 가지면서도 이더넷과 동일한 원시 속도를 제공하는 간단한 로드/저장 프로토콜을 제공
- 저전력: 전력과 복잡성을 줄이는 고효율 스위치 설계 가능
- 비용 효율적: 링크 스택에 사용되는 다이 면적이 작아 전력과 인수 비용 절감
- 개방적이고 표준화: 다양한 공급업체가 UALink 가속기와 스위치를 개발하며 상호 운용 가능한 제품을 시장에 출시

결론 및 시사점

- NVIDIA의 NVLink에 대항하는 UALink 출시로 시작된 반 NVIDIA 진영의 반격에 대해서 AI 가속기 시장의 변화 흐름에 주목할 필요가 있음

11

젠스파크, 마누스보다 뛰어난 '슈퍼 에이전트' 출시 - "진정한 첫 범용 에이전트"

개요

젠스파크(Genspark[1])는 4일, 사용자의 일상 업무를 자율적으로 수행할 수 있는 범용 AI 에이전트 '슈퍼 에이전트(Super Agent)' 출시를 발표

- 젠스파크의 슈퍼 에이전트는 기존 인공지능(AI) 도구를 넘어서는 혁신적인 플랫폼으로, 다양한 작업을 자동화하고 효율성을 극대화하며 사용자 경험을 향상시키는 것을 목표로 함

중국의 마누스를 규모면에서 압도하며 마누스 출시후 3주만에 성과 발표

- 9개의 대형언어모델(LLM)과 80개 이상의 AI 도구, 10개 이상의 데이터베이스를 결합하여 구축되었고 이 모든 요소가 협력하여 복잡한 워크플로우를 처리하고 결과를 제공함
- 젠스파크의 데모 영상에서는 슈퍼 에이전트가 5일간의 샌디에이고 여행을 계획하고, 명소 간의 도보 거리를 계산하고, 대중교통 옵션을 파악한 다음, 실감나는 음성 통화 에이전트를 사용하여 음식 알레르기 및 좌석 선호도를 고려하여 식당을 예약하는 모습을 보여줌
- 이러한 기능들이 소비자에게 초점을 맞춘 것처럼 보일 수 있지만, 실제로는 기술이 창의적인 생성과 실행 사이의 경계를 허물면서 다중 모드, 다단계 작업 자동화로 나아가고 있음을 보여주는 것

젠스파크의 성공비결은 대규모 도구 오케스트레이션을 해결한다는 점

- 중국 쑤저우 대학이 개발한 '도구사슬(CoTools)'와 MCP(Model Context Protocol) 표준을 결합
- MOA(Mixture-of-Agents) 시스템을 활용하여 여러 작업을 동시에 처리하며 복잡한 문제해결이 가능
- 자율성 강화로 단순한 명령 수행을 넘어 목표를 설정하고 이를 달성하기 위한 계획을 스스로 수립하고 실행하는 단계로 진입의

결론 및 시사점

- 마케터, 교사, 채용담당자, 디자이너 및 분석가가 최소한의 설정으로 현업에 활용가능해짐
- 범용에이전트가 현실화하며 빠르게 진화함을 보여주는 사례임

1) 젠스파크(Genspark): 캘리포니아 팔로 알토에 본사를 둔 인공지능(AI) 검색 엔진 개발 기업. 전 바이두 임원이자 마이크로소프트 김 팅의 개발 매니저였던 에릭 징(Eric Jing)과 구글 및 바이두 검색 개발 경력자 케이 주(Kay Zhu)가 설립

12

AMD, TSMC 2nm 공정 기반 첫 HPC 제품 'Venice' EPYC CPU 출시

개요

AMD는 자사의 차세대 AMD EPYC 프로세서 (코드명: Venice)가 TSMC의 첨단 2nm (N2) 공정 기술로 데이프아웃 및 생산될 업계 최초의 HPC 제품이라고 발표

- 최첨단 공정 기술을 활용, 새로운 설계 아키텍처를 공동 최적화하기 위한 AMD와 TSMC의 반도체 제조 파트너십의 강점 부각
  - 내년 출시 예정인 Venice를 통한 AMD 데이터센터 CPU 로드맵 실행의 중요한 진전을 의미

또한, AMD는 애리조나에 있는 TSMC의 새로운 제조 시설에서 5세대 AMD EPYC CPU 제품의 성공적인 출시 및 검증을 발표하면서 미국 제조업에 대한 회사의 의지를 강조

- AMD 회장 겸 CEO인 리사 수 박사는 AMD가 TSMC N2 공정과 TSMC 애리조나 Fab21의 주요 HPC 고객이라는 점은 혁신을 주도하고 컴퓨팅의 미래를 이끌어갈 첨단 기술을 제공하기 위해 양 사가 얼마나 긴밀하게 협력하고 있는지를 보여주는 좋은 사례임을 강조

결론 및 시사점

- AMD의 Venice 프로세서는 Intel의 Diamond Rapids 프로세서와 함께 CXL 3.0을 공식 지원하는 프로세서로 예상되는 바, 향후 CXL 확산과 관련한 HPC 시장의 변화 흐름에 주목할 필요가 있음

TSMC의 차세대 AMD EPYC CPU(베니스) 웨이퍼를 들고 있는 AMD Lisa SU 박사와 TSMC C.C. Wei 박사

출처: AMD

13

OpenAI, AI 에이전트 검사를 위한 벤치마크 BrowseComp 발표

개요

오픈AI(OpenAI)는 10일, 인공지능 에이전트의 검색 능력을 검증하는 벤치마크 BrowseComp를 오픈소스로 공개

- 2일의 PaperBench, 9일의 OpenAI Pioneers Program과 함께 일주일 사이에 3가지 벤치마크 관련 발표를 쏟아냄
- 올 하반기 본격적인 에이전트 출시 및 판매를 위해 자사 제품의 가치를 증명하기 위한 체계를 만들기 위한 포석으로 추정

BrowseComp는 정확성과 추론, 꾸준한 검색 능력을 확인하는 것을 목표로 1,266개의 검색 문제로 구성됨

- 벤치마크 데이터셋의 주제분포가 다양하고 난이도가 높아서 트레이너들도 검증을 시도한 1,255개의 문제 중 주어진 시간 내에 푼 문제는 29.2%에 불과했고 정답률은 86.4%였음
- 추론 기반의 Deep Research를 제외한 다른 모델들의 정확도는 0에 가까운 수치를 보임

※ GPT-4o 0.6%, GPT-4o w/브라우징 1.9%, GPT-4.5 0.9%, OpenAI o1 9.9%, Deep Research 51.5%

- 결론적으로 추론에 사용되는 컴퓨팅 시간이 늘어날수록 벤치마크에 대한 모델의 정확성도 높아짐

OpenAI, OpenAI Pioneers Program 발표

- (목표) AI를 실제 활용 사례에 적용하기 위한 평가 기준 개발 및 개발자 도구 제공
- 법률, 금융, 보험, 의료, 회계 등의 산업 분야별 벤치마크 평가를 설계하고 각 기업의 3대 핵심 사용 사례에 맞춘 맞춤형 모델 개발 예정
- 스타트업을 중심으로 소수의 고부가가치 응용 사례 개발 기업을 선정하여 도메인별 평가 지표 설계 및 맞춤 모델 개발을 진행할 예정

결론 및 시사점

- OpenAI는 Pioneers Program을 통해 산업 분야에 맞춤화된 벤치마크를 마련함으로써, AI 도입의 불확실성을 줄이고 적용 속도를 높일 수 있는 환경을 제공
- AI가 실무에서 유용하려면 단순 정확도뿐만 아니라 도메인별 현실 기반의 평가 프레임워크가 필수로 자리잡음
- 이렇듯 AI 평가 방식은 인위적인 테스트셋에서 벗어나 실제 사용 조건을 반영한 동적 평가로 진화하고 있음을 시사하며, 이렇듯 도메인별로 특화된 AI의 도입은 단순한 기술 수용이 아닌 경쟁력 향상의 전략적 수단이 될 것임을 시사

제2절 학계, 연구계 동향

01

ANL, AI에 대비하여 새로운 세대의 연구자를 양성하기 위한 교육 시리즈 운영

개요

미국 에너지부(DOE) 산하 ANL은 매년 “슈퍼컴퓨터를 활용한 AI 기반 과학 입문”이라는 교육 시리즈를 통해 AI와 고성능 컴퓨팅(HPC)의 힘을 과학 연구에 활용할 수 있도록 새로운 세대의 연구자들을 준비시키고 있으며, 이 가상 교육 프로그램은 2021년 시작된 이후 지금까지 700명 이상의 학부 및 대학원생들을 대상으로 진행됨

- **이번 시리즈는 미국 국가 인공지능 연구 자원(National Artificial Intelligence Research Resource, NAIRR) 파일럿 프로젝트를 통해 홍보되었는데, 이 파일럿 프로젝트는 미국 국립과학재단(NSF)이 DOE 및 여러 협력 기관과 손잡아 시작한 것으로, AI 교육, 훈련, 사용자 지원 및 홍보를 통해 새로운 커뮤니티를 참여시키는 데 초점을 맞추고 있음**
- **이 프로그램은 ALCF 직원들이 구성 및 교육을 진행하는데 올해는 대규모 언어 모델, 신경망, 모델 훈련 등 다양한 주제를 다룬 7부작 시리즈로 구성되었으며, 참가자들은 ALCF의 HPC 자원, 예를 들어 폴라리스(POLARIS) 슈퍼컴퓨터와 ALCF AI 테스트베드에서 실습을 진행할 수 있었음**
- **이번 교육 프로그램에 참여한 한 대학원생은 이전에 AI 도구를 작업한 경험은 많았지만, 슈퍼컴퓨터 자원에 대한 노출이 제한적인 상태였기 때문에 ALCF를 통해 큰 도움을 받았다 밝힘**
- **프로그램 참여 학생들은 AI 모델이 개인 컴퓨터보다 대규모 시스템에서 효과적으로 작동하는 것을 경험하였으며, 프롬프트 엔지니어링의 모범 사례에 대한 교육도 이루어짐**
- **결론 및 시사점**
  - ANL에서는 슈퍼컴퓨터를 활용한 AI 교육 프로그램을 운영하고 있으며, 이 프로그램을 통해 많은 학부생과 대학원생들이 평소 접하기 힘든 HPC에 접근하여 활용법을 익히고 운영 방식 등에 대한 교육을 받을 수 있게 됨

ANL에서 진행되는 AI 교육 세션

출처: ANL

02

Sandia 국립 연구소, 에너지 효율적 AI 및 컴퓨팅 기술 개발을 위해 연구소들과 협력

개요

샌디아 국립연구소가 인공지능(AI) 시대의 에너지 위기를 대비하기 위한 새로운 연구 프로젝트를 시작하였으며, 이는 컴퓨팅 기술이 향후 10년 내 전 세계 에너지 생산량의 상당 부분을 소비할 것으로 예상되기 때문

- **미국 에너지부는 이러한 도전과제를 해결하기 위해 세 개의 새로운 마이크로전자 과학연구센터를 설립한다고 발표하였으며, 그 중 하나인 MEERCAT(마이크로전자 에너지 효율 연구센터)는 센싱, 엣지 프로세싱, AI, 고성능 컴퓨팅을 아우르는 에너지 효율성 솔루션을 연구할 계획이며 이 프로젝트에는 총 1억 7,900만 달러가 투자되며, 최대 4년간 16개의 다학제 기초연구가 진행될 예정**
- **샌디아 연구소는 MEERCAT의 창립 멤버로서 8개의 에너지 효율 관련 연구 프로젝트 중 하나를 주도하게 되었는데, 특히 실리콘을 대체할 수 있는 새로운 물질(몰리브덴 이황화물, 갈륨 비소, 다이아몬드 등)을 기존 실리콘 제조 공정에 통합하는 '이종 통합' 연구를 진행할 예정**
- **이 프로젝트에는 미국 전역의 5개 나노스케일 과학연구센터와 페르미 국립가속기연구소, MIT, MIT 링컨 연구소 등이 참여하며, 연구진들은 산업계와 긴밀히 협력하여 더 효율적이고 강력한 컴퓨터 칩을 개발함으로써 경제 및 국가 안보에 기여하는 것을 목표로 하고 있음**
- **특히 이 연구는 2022년 CHIPS 법안의 일환으로 진행되는 것으로, AI와 양자 컴퓨팅 등 에너지 집약적 기술의 급속한 성장에 따른 시급한 과제를 해결하기 위한 것임. 기존 컴퓨터 알고리즘보다 더 많은 에너지를 사용하는 AI 기술이 가정과 직장에서 급속도로 보급되면서, 에너지 효율성 향상이 더욱 중요한 과제로 대두되고 있음**

통합나노센터

출처: Randy Montoya

- **결론 및 시사점**
  - AI와 컴퓨팅 기술의 급격한 확산으로 인한 에너지 위기에 대비해 미국이 1억 7,900만 달러 규모의 마이크로전자 연구센터들을 설립하였으며, 샌디아 국립 연구소를 중심으로 실리콘을 대체할 새로운 물질 개발과 에너지 효율성 향상 연구를 진행할 예정임

03

BSC, MareNostrum 5에 통합할 새로운 양자 시스템 출시

개요

바르셀로나 슈퍼컴퓨팅센터(BSC)는 100%로 유럽 기술로 개발된 최초의 양자 컴퓨터를 선보이며 기존 컴퓨팅과 양자 컴퓨팅을 결합한 하이브리드 컴퓨팅을 위한 기반을 마련

- 스페인 슈퍼컴퓨팅 네트워크(RES)의 14개 노드와 CSIC, ICFO 등의 기관, 바르셀로나 대학교, 마드리드 자치대학교, 발렌시아 폴리테크닉 대학교 등의 대학을 포함하여 스페인의 27개 주요 연구 및 슈퍼컴퓨팅 기관이 참여하는 협력 프로젝트인 퀀텀 스페인(Quantum Spain)의 일부로서 구축
- 양자컴퓨터는 MareNostrum 5 슈퍼컴퓨터에 통합되며 양자 기술과 고전적 기술의 결합은 연구와 혁신을 가속화하고 스페인의 산업 및 기술 발전을 촉진하고 고도로 자격을 갖은 일자리 창출에 기여할 것으로 예상
- 퀀텀 스페인의 양자 컴퓨터 건설은 스페인 회사 Qilimanjaro와 GMV가 설립한 합작 투자가 주도했으며 초전도 큐빗을 기반으로 하는 시스템을 개발
- 100% 유럽 기술로 구축된 이 시스템은 스페인의 양자 컴퓨팅 전략에서 결정적인 단계를 나타내며 유럽의 기술적 자율성을 강화하여 제3국의 핵심 인프라에 대한 의존도를 낮추려는 유럽 위원회의 전략과 일치

출처: BSC

결론 및 시사점

- 양자 기술과 고전 기술이 결합한 하이브리드 컴퓨팅의 향후 방향에 대해서 주목할 필요가 있음

04

중국, 양자 컴퓨터로 10억 개 매개변수 AI 모델 파인튜닝 완료

개요

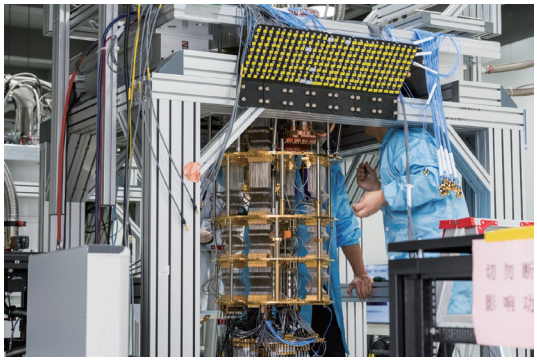
중국 과학자들이 독자적으로 개발한 3세대 초전도 양자 컴퓨터(Origin Wukong)를 사용하여 세계 최초로 10억 개의 매개변수를 가진 인공지능(AI) 대형 모델의 파인튜닝 작업을 성공적으로 수행

- 'Origin Wukong'은 72큐비트의 초전도 양자 칩 'Wukong'이 사용되었으며, 해당 칩은 안후이 퀀텀 컴퓨터 기술 연구센터(Anhui Quantum Computing Engineering Research Center)에서 개발됨
- 수행 결과, 최적화된 모델의 훈련 손실이 15% 감소하였고, 수학적 추론 작업 정확도가 68%에서 82%로 향상됨. 또한, 매개변수를 76% 줄였으나 훈련 효과는 8.4% 향상됨
- 이번 실험은 양자 컴퓨팅을 활용하여 대형 언어 모델(LLM)의 파인튜닝 작업을 수행할 수 있으며, 경량화까지 실현할 수 있음을 입증한 첫 사례

'Origin Wukong'은 2024년 1월 6일 가동 이후 유체 역학, 금융, 생명 의학 등 다양한 산업 분야에서 35만 건 이상의 양자 컴퓨팅 작업을 완료하였으며, 139개국의 원격 이용자가 사용중

결론 및 시사점

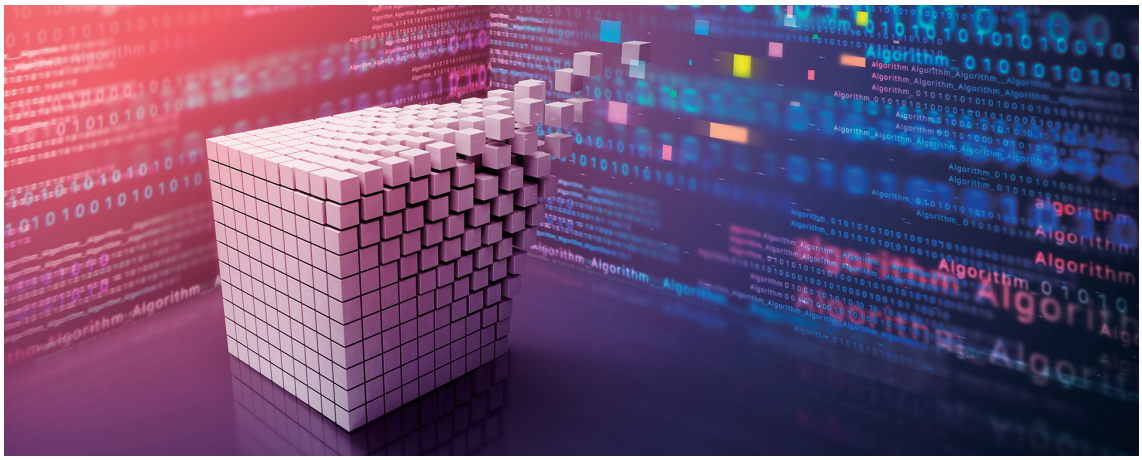
- 양자 컴퓨터가 실제 AI 모델의 파인튜닝 작업을 지원할 수 있음을 증명하였을 뿐 아니라, 양자-AI 융합 기술의 실용화 가능성을 입증
- 매개변수의 대폭적인 감소에도 불구하고 훈련 효과가 향상된 것은 양자 컴퓨팅이 AI 모델 경량화와 최적화에 기여할 수 있음을 시사함
- 이번 성과는 현재 직면한 계산 자원 부족 이슈를 해소하고, 다양한 산업 분야에서 AI 적용의 효율성을 높이는 데 중요한 역할을 할 것으로 예상됨



중국이 독자적으로 개발한 3세대 초전도 양자 컴퓨터

출처: 안후이 양자컴퓨팅공학연구센터

GLOBAL HPC HORIZON



04 초고성능컴퓨팅 응용 및 활용 동향

01 슈퍼컴퓨터를 활용한 연쇄 지진 활동 연구

개요 **다단계 지진 연쇄 반응 연구를 위한 슈퍼컴퓨터 시뮬레이션**

- Ludwigs-Maximillians Universität München, University of California at San Diego, Technical University of Munich의 연구진은 Leibniz Supercomputing Centre의 슈퍼컴퓨터를 활용하여 지진의 연쇄 반응 및 위험성을 연구
- 작은 지진과 큰 지진의 메커니즘 차이를 규명하고, 연쇄 지진(cascading earthquakes)의 불확실성을 탐구

(접근방안) 지진 시뮬레이션 및 데이터 통합

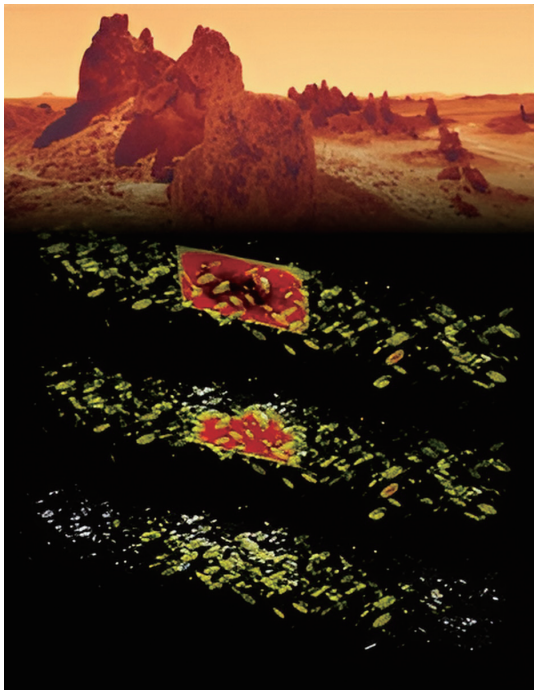
- SuperMUC-NG 슈퍼컴퓨터를 사용하여 지진의 물리적 동작과 규모 간 차이를 시뮬레이션
- 대규모 지진 데이터를 기반으로 선형적인 파괴 에너지와 단층 크기 간 관계를 발견
- 기계학습 모델을 결합하여 슈퍼컴퓨터-기계학습 하이브리드 시뮬레이션 방식을 도입하여 연구 효율성 향상

(결과) 지진 및 연쇄 반응 시뮬레이션 모델 개발

- 연쇄 지진의 특성과 메커니즘을 재정립하여 시뮬레이션의 정확도 및 위험 예측 능력 향상
- 공공 슈퍼컴퓨터 인프라를 통해 연구 활동으로부터 얻어진 데이터를 공유하고, 이를 돕는 "과학 게이트웨이" 구축에 기여

결론 및 시사점

- 공공 슈퍼컴퓨터 인프라를 기반으로 지진 및 연쇄 반응을 보다 정확히 모델링 가능
- 도시 계획 및 구조물 설계에서 지진 위험을 더 잘 평가하기 위한 과학적 데이터를 제공
- 공공 연구시설이 제공하는 데이터 공유 및 관리 지원이 상호 협력 연구에 기여



700개가 넘은 균열에 걸친 지진 시뮬레이션

출처: UCSD Jennifer Matthews

02

화학 소재분야를 위한 확장 가능한 기계학습 모델 개발

개요

Berkeley 연구진은 화학 시뮬레이션을 위한 확장 가능한 AI 방법론을 개발

- **UC Berkeley 및 Berkeley Lab 연구진이 원자 간 힘 계산을 가속화하는 새로운 인공 신경망 기반 기계 학습 모델 개발**
- **기존 모델 대비 메모리 사용량을 5배 이상 감소시키고, 계산 속도를 10배 이상 향상**
- **화학, 재료 과학 및 약물 개발 등 기초 과학 연구에서 분자 및 원자 상호작용을 더 효율적으로 분석 가능**

(접근방안) 기계 학습과 슈퍼컴퓨터의 융합

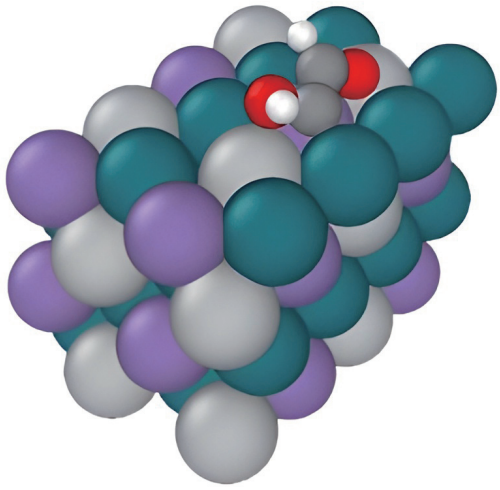
- 새로운 모델인 EScAIP(Efficiently Scaled Attention Interatomic Potential) 개발, 기존 NNIP 모델의 한계 극복
- 물리적 제약 없이 데이터를 통해 복잡한 패턴과 물리적 통찰을 직접 학습하는 기계 학습 모델 설계
- 美 에너지부의 NERSC 슈퍼컴퓨터 다중 GPU 자원을 활용하여 대규모 데이터셋에서 모델 훈련을 최적화

(결과) 과학 연구를 위한 확장 가능한 NNIP 모델 구현

- EScAIP 모델은 촉매 (Open catalyst 20, 22), 소재 (Materials project), 분자 (SPICE) 등 다양한 데이터셋에서 최고 성능 기록
- 기술 기업의 기여 없이 학계 및 공공 연구기관으로만 구성된 팀이 개발한 모델로는 최초로 Open catalyst 데이터셋에서 성능 1등을 달성
- 메모리 비용을 줄이고 계산 속도를 높여 화학 반응 경로 최적화 및 대규모 데이터 처리 가능
- 효율적인 확장 전략을 통해 대규모 훈련을 민주화하고 더 넓은 연구 커뮤니티에서 접근할 수 있도록 하는 방향성을 제시

결론 및 시사점

- EScAIP는 NNIP 확장성을 고려한 새로운 접근 방식을 제시하며 물리적 제약에 의존하지 않는 데이터 중심 모델의 가능성을 확인
- 포괄적인 데이터셋 생성과 이를 보완하는 평가 지표 개발이 필수
- Berkeley 연구진은 과학 커뮤니티에 확장 가능한 기계 학습 접근 방식을 제안하며, 해당 분야의 발전 방향을 모색
- 공공 GPU 자원과 협력 프로젝트를 통해 대규모 데이터 기반 연구의 효율성을 극대화



인공 신경망 기반 분자간 포텐셜 모델을 통해 촉매 표면에서 일어나는 반응을 빠르게 연산할 수 있다.

출처: Samuel Blau & Eric Yuan, Berkeley Lab

03

슈퍼컴퓨터를 활용한 연쇄 지진 활동 연구

개요

해양 풍력 터빈 설계를 위한 대와류 시뮬레이션 연구

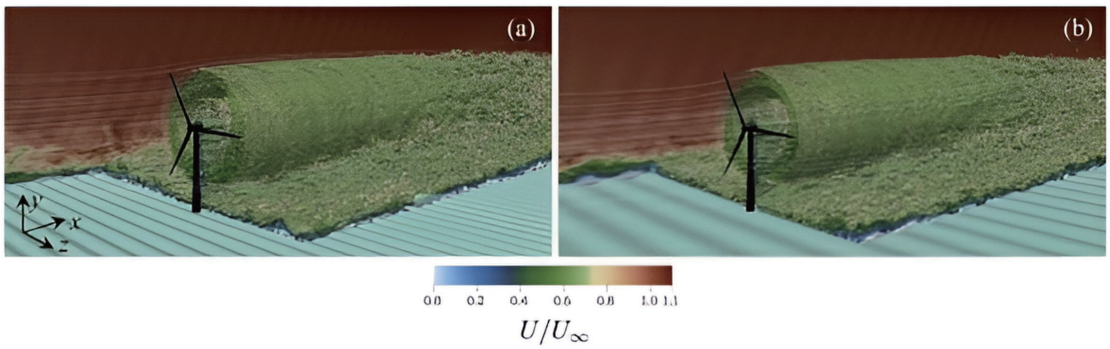
- University of Puerto Rico at Mayaguez의 기계공학 교수 Umberto Ciri가 SDSC의 Expanse 슈퍼컴퓨터를 활용하여 해양 풍력 터빈 주변의 바람과 파도의 상호작용을 연구
- 대와류 시뮬레이션(LES)을 통해 다양한 파도 유형(잔잔한 파도부터 긴 파랑까지)이 터빈의 효율성과 내구성에 미치는 영향을 분석

(접근방안) 공기-물 흐름의 통합 시뮬레이션

- 가상경계법(immersed boundary method)과 레벨셋 방법(level set method)을 혼합하여 바다와 바람의 경계를 시뮬레이션
- 회전하는 액츄에이터 디스크 모델(rotating actuator disk model)을 활용하여 풍력 터빈의 동작을 시뮬레이션
- 해양 표면 근처의 바람 속도 및 난류 운동 에너지 변화를 시뮬레이션으로 분석
- 바람 터빈 뒤쪽의 “wake recovery” 속도를 파도 유형별로 조사

(결과) 지진 및 연쇄 반응 시뮬레이션 모델 개발

- 파도가 대기 하층의 바람을 느리게 하고 해양 표면 근처의 난류를 증가시켜 터빈의 후류(wake)에 영향을 미침
- 성숙한 파도는 후류 회복을 늦추는 반면, 중간 나이의 파도는 회복 속도를 증가시킴
- 난류가 파도 주파수와 단순히 일치하지 않으며, 파도의 복잡하고 분산된 영향을 나타냄



파도의 파장에 따른 해양 풍력 터빈 주변 공기의 흐름

출처: University of Puerto Rico at Mayaguez

결론 및 시사점

- 슈퍼컴퓨터를 활용하여 해양 풍력 단지 설계에 파도 패턴을 고려하는 것이 중요하다는 점을 제시
- 슈퍼컴퓨팅이 신재생에너지 발전시설 효율 향상에 기여한 사례

04

혈관 내 단백질의 기계적 힘에 대한 반응 연구

개요

G 단백질 결합 수용체 (GPCR)[1]의 기계적 힘에 대한 반응을 슈퍼컴퓨터 시뮬레이션을 이용해 연구

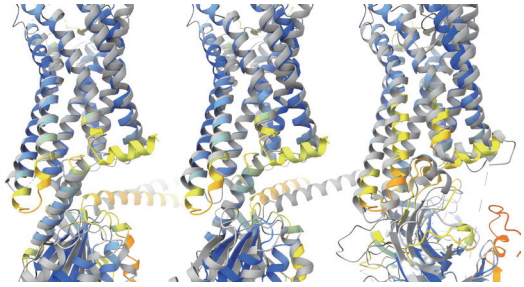
- Oregon State University(OSU)의 연구진이 Anton 슈퍼컴퓨터[2]를 활용하여 혈관 내 GPCR의 기계적 힘에 대한 반응 메커니즘 연구.
- Angiotensin II Type 1 Receptor (AT1R)[3]의 구조와 기계적 힘이 신호 전달 경로에 미치는 영향을 시뮬레이션으로 분석

(접근방안) 분자 동역학 시뮬레이션과 AI 도구의 통합 사용

- 수십 마이크로초 동안 AT1R의 구조적 변화를 시뮬레이션
- 막 두께 변화와 기계적 장력을 적용하여 AT1R의 활성화와 비활성화 상태 전환을 관찰
- AlphaFold 2[4]를 활용해 GPCR과 β-arrestin[5] AT1R 구조에 미치는 영향을 분석하여 시뮬레이션 결과의 정확성을 검증

(결과) GPCR 신호 전달 메커니즘에 대한 새로운 발견

- 막 두께가 두꺼워지면 AT1R의 활성 상태와 비활성 상태 간 전환이 용이해지며, 얇아지면 전환이 어려워짐을 확인
- 기계적 장력은 막을 얇게 만들어 AT1R을 활성 상태로 전환함
- AT1R의 구조적 변화가 GPCR 및 β-arrestin 신호 전달 경로의 활성화에 미치는 영향을 시뮬레이션으로 규명
- AT1R의 기계적 힘에 의한 활성화는 혈압 조절 및 심혈관 질환 연구에 중요한 통찰 제공



AT1 수용체와 β-arrestin과 생체 같은 분자들의 상호작용

출처: PSC, https://www.psc.edu/anton-blood-vessel/

결론 및 시사점

- 슈퍼컴퓨팅을 활용해 복잡한 생체 분자 기계의 화학적 및 기계적 특성을 규명
- 혈압 조절 및 심혈관 질환 연구에서 중요한 혈관 내 단백질의 특성에 대한 새로운 통찰 제공
- 인공지능 모델의 발전 속에서도 슈퍼컴퓨팅이 생체 분자 연구에서 여전히 핵심적인 역할을 함을 보여주는 사례

1) GPCR(G-Protein Coupled Receptor, G 단백질 결합 수용체): 세포막에 위치한 단백질로 외부 신호(예: 호르몬, 신경전달물질)를 세포 내부로 전달하여 다양한 생리적 반응을 조절하는 역할을 함

2) Anton 슈퍼컴퓨터: 미국의 D.E.Shaw Research에서 설계한 특수 목적 슈퍼컴퓨터로 분자 동역학 시뮬레이션에 특화되어 설계됨

3) AT1R(Angiotensin II Type 1 Receptor): 혈압과 체액 균현을 조절하는 주요 경로인 레닌-인조어센틴 시스템의 주요 구성요소로 GPCR의 일종

4) AlphaFold 2: DeepMind에서 개발한 단백질의 3차원 구조 예측 프로그램. 기계학습 기술을 활용, 높은 정확도를 보임

5) β-arrestin: GPCR 신호 전달에 핵심 조절자로 작용하는 단백질

05

ITER 토카막에서의 전자 행동 연구를 위한 슈퍼컴퓨터 시뮬레이션

개요

슈퍼컴퓨터를 활용하여 국제 핵융합 프로젝트 ITER에서 발생할 수 있는 폭주 전자[1] 문제를 연구

- ORNL의 Summit 슈퍼컴퓨터를 사용하여 토카막 내의 복잡한 전자-플라즈마 상호작용을 시뮬레이션
- 알펜파[2]을 이용하여 폭주 전자를 분산시켜 플라즈마 장치의 손상을 방지할 수 있음을 보임

(접근방안) 슈퍼컴퓨터를 활용한 전자 분산 시뮬레이션

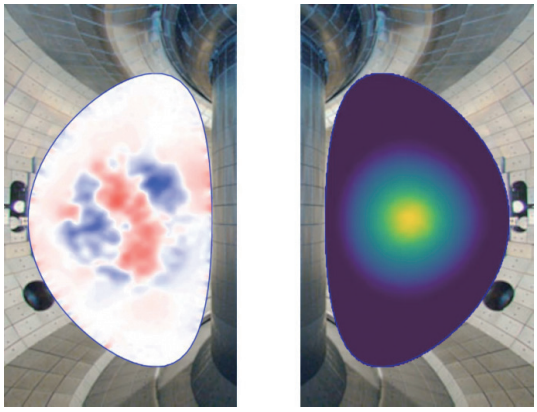
- Princeton Plasma Physics Laboratory, General Atomics, Columbia University의 연구진은 알펜파의 여기(excitation), 폭주 전자와의 상호작용, 플라즈마, 전자기장의 모델링을 결합한 시뮬레이션 모델 구축
- 플라즈마 내부의 자기장 변동(알프벤 파동)의 생성과 전자 분산 메커니즘을 모델링.
- Summit 컴퓨터의 기존 CPU 기반 시스템보다 30배 빠른 계산 속도를 통해 수많은 전자가 광속에 근접해 움직이는 현상을 정확하게 분석

(결과) 알펜파를 통한 폭주 전자 분산 가능성 확인

- 시뮬레이션 결과, 알펜파가 폭주 전자를 효과적으로 분산시켜 강력한 전자 빔 형성을 방지함을 확인
- 美 에너지부의 D-III National Fusion Facility에서 수행된 실험과 일치
- 더 많은 상호작용을 포함한 모델을 구축하여 Frontier 컴퓨터에서 시뮬레이션 할 예정

결론 및 시사점

- 슈퍼컴퓨터를 활용하여 핵융합 발전로의 안정성을 향상시키기 위한 전략 수립에 기여
- GPU 기반 슈퍼컴퓨팅 기술이 복잡한 실제 문제 해결에 효과적으로 적용된 사례



ORNL의 Summit 슈퍼컴퓨터에서 수행된 시뮬레이션은 핵융합로 토카막 내부의 폭주 전자 문제에 대한 해결책을 제시. 알펜파(좌)를 활용하여 전자(우)를 분산시킬 수 있음

출처: Chang Liu, 프린스턴 플라지마 물리 연구실

1) 폭주 전자(Runaway electron): 플라즈마가 운전을 마치는 순간, 플라즈마 전류가 가끔씩 떨어지는 순간에 발생하는 전자들의 운전속도가 제대로 제어되지 않는 현상. 핵융합로를 공격하여 구조 손상을 야기

2) 알펜파(Alfvén wave): 플라즈마 내부의 이온과 자기장의 저주파 진동

06

슈퍼컴퓨터를 활용한 유전 데이터 분석 가속화

개요

유전 데이터 분석의 100배 속도 향상 달성

- 미국 에너지부(DOE) 아르곤 국립연구소(Argonne National Laboratory) 연구진이 미국 재향군인 백만 프로그램 (MVP)[1] 유전 데이터 분석을 100배 가속화하는 방법을 개발
- 유전 변이와 질병 위험 간의 연관성을 빠르게 분석하여 맞춤형 의료 및 질병 예측 연구의 발전에 기여
- GPU 기반 슈퍼컴퓨팅을 활용하여 방대한 MVP 데이터의 분석 성능을 획기적으로 향상

(접근방안) GPU 기반 분석으로 데이터 병렬 처리 최적화

- 기존 CPU 기반 분석에서는 데이터 크기가 처리 능력을 초과하여 실행이 실패하는 문제가 발생
- 연구팀은 행렬 연산 최적화를 통해 GPU를 활용한 코드 재설계로 연산 성능을 극대화
- Summit 슈퍼컴퓨터 2,000 노드를 활용하여 14일 동안 분석 수행

(결과) 연산 속도 100배 증대

- MVP 데이터를 통해 다양한 인종 집단(아프리카계 87% 증가, 히스패닉계 45% 증가)에 대한 질병 연관성 연구 수행
- 4500만개의 바이오마커를 기반으로 MVP 데이터 전체에 대하여 3000억 개의 연관성 분석 수행
- 2024년 7월 Science 저널에 연구 결과 게재

결론 및 시사점

- 슈퍼컴퓨터를 활용한 거대 유전 데이터 분석 사례
- GPU 기반 고성능 병렬 컴퓨터 및 가속화 기술을 통해 거대 데이터의 효과적인 분석이 가능해짐

1) 미국 재향군인 백만 프로그램(Milion Veteran Program): 유전자, 생활 방식, 군 경험 및 노출이 재향군인의 건강에 미치는 영향을 연구하는 프로그램. 2011년 출범 이후 100만 명 재향군인의 유전 정보 데이터를 수입했으며, 유전 정보 중 29%가 비유럽계 조상을 지님

07

PSC: Bridges-2 시뮬레이션, 로켓 배출수가 달의 얼음 저장소에 미치는 영향 분석

개요

달 표면의 물 저장소에 미치는 로켓 배기가스의 영향 연구

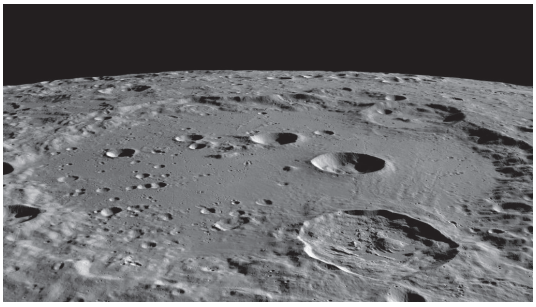
- 일반적으로 달 표면에는 물이 존재하지 않으나 극지방의 영구히 그늘진 분화구 속에는 얼음 형태로 물이 존재함
- Johns Hopkins Applied Physics Laboratory 연구진은 PSC의 Bridges-2 슈퍼컴퓨터를 활용하여 로켓 배기가스에서 생성된 물이 달의 극지방 얼음 저장소에 미치는 영향을 시뮬레이션
- 인간 탐사가 달의 수자원 연구를 오염시킬 가능성을 분석하고, 과거 및 미래 탐사선의 기여도를 정량화

(접근방안) 달의 물 순환 과정 시뮬레이션

- 달 표면의 영구히 그늘진 크레이터에 존재하는 물의 기원과 형성을 연구
- 로켓 연소 과정에서 생성된 물 분자의 이동 경로를 시뮬레이션, 태양광과의 상호작용 및 물 분자의 확산 과정을 모델링
- Bridges-2 슈퍼컴퓨터의 수만 개의 연산 코어를 활용하여 수천만 개의 물 분자의 이동과 화학 반응을 추적

(결과) 인간 탐사가 달의 물 저장소에 미치는 영향 분석

- 아폴로 탐사선은 달의 극지방에 최대 0.36톤의 물을 공급했을 가능성이 있음
- 스페이스X 社의 스타쉽 착륙선은 특정 착륙 시나리오에서 10톤 이상의 물을 극지방에 공급할 수 있어 연구를 오염시킬 위험이 큼



달의 Clavius 크레이터를 시각화한 이미지

출처: NASA Scientific Visualization Studio

결론 및 시사점

- 슈퍼컴퓨터를 활용하여 달 착륙 과정에서 생성되는 물의 거동을 규명
- 우주과학분야의 복잡한 시뮬레이션을 수행하기 위해선 슈퍼컴퓨터 자원이 필수적

08

슈퍼컴퓨터를 활용한 맞춤형 암 치료법 연구

개요

개인화된 암 치료법을 위한 분자 동역학 시뮬레이션

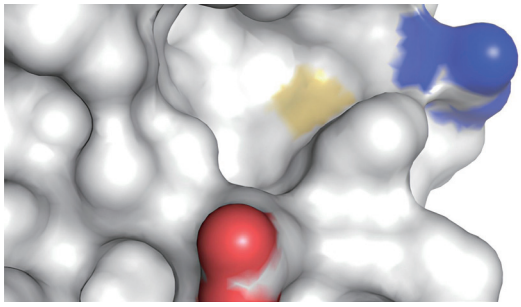
- High-Performance Computing Center Stuttgart (HLRS)의 슈퍼컴퓨터 자원을 이용하여 공격적인 암과 관련된 유전자 변이를 연구
- 분자 동역학 시뮬레이션을 통해 단백질 구조 변화와 돌연변이가 암세포에 미치는 영향을 분석
- 암 치료에 중요한 p53 단백질의 돌연변이 연구를 통해 맞춤형 치료법 개발 가능성 모색

(접근방안) 슈퍼컴퓨터 Hawk를 활용한 분자 동역학 시뮬레이션

- p53 단백질의 변이를 포함한 여러 돌연변이 모델에 대해 시뮬레이션 수행
- 36,000~39,000개의 원자가 포함된 모델로 0.8마이크로초 동안 단백질 변형 추적
- 돌연변이로 인한 구조적 변화가 약물 결합 가능성을 어떻게 변화시키는지 분석

(결과) p53 단백질의 돌연변이 Y163C의 약물 결합 가능성 발견

- Y163C 변이로 인해 생성된 구멍은 X선 결정학으로 관찰하였을 때는 너무 작았지만 분자 동역학 시뮬레이션에서는 이 구멍이 약물이 결합할 수 있을 만큼 커질 수 있음을 확인
- Y163C 변이에 대해 약물을 개발하는 방향으로 연구를 집중시킬 기회를 제공
- 가상 약물 스크리닝을 통해 약물 후보 물질을 빠르게 선정하고 실험을 통해 검증 가능



분자 동역학을 통해 시뮬레이션한 단백질 구조

출처: Balourdas et al, 2024

결론 및 시사점

- 슈퍼컴퓨팅을 활용한 분자 동역학 시뮬레이션을 통해 암 치료에 중요한 p53 단백질의 돌연변이 연구에 새로운 접근법 제시
- 맞춤형 암 치료를 위한 약물 개발을 가속할 가능성을 확인, 분자 수준에서의 정확한 표적화와 치료법 개발에 이바지할 수 있음을 시사

09

슈퍼컴퓨터를 활용한 배터리용 고체전해질 설계

개요

전고체배터리[1]를 위한 고체 전해질[2] 연구에 슈퍼컴퓨터 활용

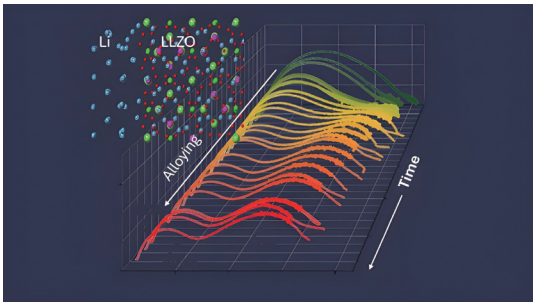
- 미국 에너지부(DOE) 아르곤 국립연구소(Argonne National Laboratory) 연구진이 고체 전해질의 원자 수준에서의 거동을 분석하기 위해 컴퓨터 모델링 기법을 적용
- 리튬 이온 배터리의 안전성 및 에너지 효율성을 개선하기 위한 연구로, 리튬 란타넘 지르코늄 가닛(LLZO) 기반 전해질의 성능을 최적화하는 방법을 탐색

(접근방안) 슈퍼컴퓨팅과 실험 기법을 결합한 연구

- 밀도 범함수 이론(DFT, Density Functional Theory)을 이용하여 원자 및 전자의 거동을 시뮬레이션하여 LLZO의 안정성 및 반응성을 예측
- X선 광전자 분광법(XPS)과 전기화학 임피던스 분광법(EIS)을 사용하여 LLZO의 표면 화학 및 리튬 이온 이동 특성을 분석
- 중성자 회절 실험을 통해 도펀트[3]의 구조적 안정성을 검증

(결과) LLZO 전해질 내 도펀트의 역할 및 영향 분석

- 갈륨 도핑 LLZO는 높은 이온 전도성을 보였지만, 구조적 안정성이 낮음
- 알루미늄 도핑 LLZO는 구조적으로 안정하며 리튬과의 반응성이 낮아 수명 연장에 유리
- 갈륨 도핑의 이점을 유지하면서도 반응성을 억제하기 위한 보호 계면층 필요성 확인
- 고체 전해질-전극 계면에서 발생하는 반응 메커니즘을 원자 수준에서 규명하여, 더 안정적인 배터리 설계 가능성 제시



LLZO-리튬 금속 계면의 구조와 합금 형성 과정

출처: Matt Klernk, Sanja Tepavcevic, Peter Zapol/ANL

결론 및 시사점

- 슈퍼컴퓨팅 기반 시뮬레이션과 실험 데이터를 결합하여 고체 전해질 내 원자 수준 반응을 분석하는 연구 접근법을 확립
- 고체 전해질의 반응성과 이온 전도성의 균형을 최적화하는 설계 방향 제시, 차세대 고체 전해질 배터리의 개발에 기여

1) 전고체배터리: 일반적인 배터리와 달리 액체 전해질을 사용하지 않는 배터리

2) 전해질: 배터리의 양극과 음극 사이 전류가 흐를 수 있게 해주는 물질

3) 도펀트(dopant): 물질의 성질을 변화하기 위해 반도체에서 의도적으로 첨가하는 불순물. 본 연구에서는 알루미늄이나 갈륨을 사용

10

슈퍼컴퓨터를 활용한 소규모 해양과정의 폭풍 발달 영향 확인

개요

소규모 해양과정[1]의 폭풍 발달 영향 연구에 슈퍼컴퓨터 활용

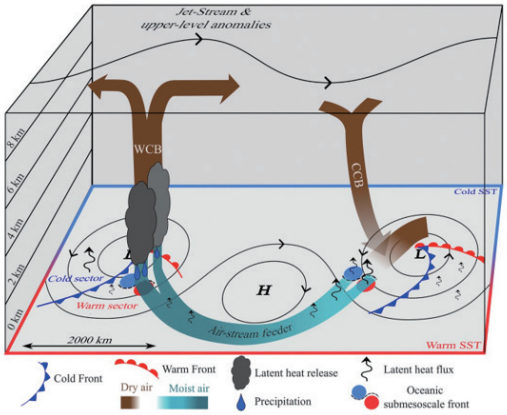
- 수십 년 동안 과학자들은 200km 이상을 포함하는 큰 해양 온도 패턴만이 폭풍에 지대한 영향을 미칠 수 있다고 판단해 왔음
- 고해상도 해양-대기 모델로의 슈퍼컴퓨팅을 통해 소규모 해양과정이 폭풍 발달에 큰 영향을 미칠 수 있음을 발견
- 미국의 UC 샌디에이고 스크립스해양연구소(Scripps Institution of Oceanography), NASA 제트추진연구소(Jet Propulsion Laboratory) 및 NASA 고다드우주비행센터(NASA Goddard Space Flight Center)의 과학자팀이 연구에 참여

(접근방안) 슈퍼컴퓨터를 활용한 해양-대기 시뮬레이션

- 슈퍼컴퓨터를 활용하여 km 규모의 해상도로 글로벌 결합 해양-대기 시뮬레이션을 수행
- 겨울철 3개월 동안 쿠로시오 확장역의 2,400km×1,100km 영역에서 시간당 데이터를 분석

(결과) 소규모 해양 전선의 영향력 확인

- 시뮬레이션을 통해 잠열 유속[2] 변동성의 절반은 겨울철 쿠로시오 확장역에서 중규모(~20km 크기, ~40%)와 아중규모(~10-20km 크기, <10%)의 해양 운동에 의해 주도됨을 확인
- 9km에 걸쳐 해수면 온도가 6℃ 변하는 소규모 전선이 바다의 열을 대기로 전달한 중위도 겨울 폭풍을 일으키는데 크게 기여함을 확인
- 소규모 전선이 바다에서 약 4km 떨어진 대기로 수분을 운반하여 중위도 폭풍의 일부 지역에서 강수량의 절반을 차지한다는 것도 밝힘



바다에서 공급되는 폭풍 트랙 수분

출처: Communications Earth & Environment (2025). DOI: 10.1038/s43247-025-02002-z

결론 및 시사점

- 고해상도 해양-대기 모델을 사용하고 슈퍼컴퓨팅 자원을 투입하여 소규모 해양과정이 폭풍 발달에 큰 영향을 미칠 수 있음을 발견하였으며 폭풍 강도와 강우량도 보다 정확하게 확인 가능함
- 소규모 해양과정 메커니즘으로 대기의 강[3]과 같은 다른 현상을 탐구할 수 있으며, 폭풍 강도에 미치는 영향도 정량화할 수 있음

1) 해양과정: 해양에 영향을 미치는 자연적 변화(물리적, 화학적, 생물학적 변화 포함)

2) 잠열 유속: 물이 표면에서 증발하거나 표면으로 응축될 때 발생하는 표면과 대기 사이의 에너지 교환

3) 대기의 강: 대기 중에 수증기가 포함된 길고 좁은 기류

11

ESA의 Space HPC를 통한 태양 폭풍 및 우주 기상 모델링 가속화

개요

유럽 우주 기상 등 모델링을 위한 ESA[1]의 Space HPC 슈퍼컴퓨터 도입

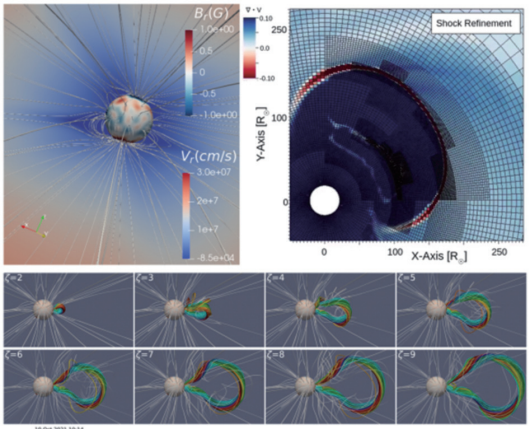
- 태양 폭풍 및 우주 기상 이벤트는 지구의 통신, 전력 전송 및 항법 시스템에 큰 영향을 줄 수 있으며, 이에 대한 유일한 방어책은 태양 활동을 지속적으로 모니터링하고 주요 인프라 운영자에게 정확하고 적절한 시점에 정보를 제공하는 것임
- 우주 기상 등 모델링과 관련 유럽 우주 산업의 증가하는 컴퓨팅 요구 사항이 충족되도록 설계된 ESA의 Space HPC(5PF 성능)가 이탈리아 ESRIN[2]에서 출범
- Space HPC를 우주 기상 시스템에 적용함으로써 우주 기상 등 모델링에 필요한 방대한 양의 다양한 데이터를 분석해서 필요시 적시 통보하는 데 도움이 되리라 예상

(접근방안) Space HPC의 우주 기상 전용 슈퍼컴퓨터 구축 및 HPE와의 시연

- ESA 우주기상청이 HPE와 협력하여 Space HPC에서 COCONUT, Icarus AMR 스킴, 플럭스 로프 CME 등 우주 기상 모델링의 시뮬레이션을 시연
- ESA가 여러 유럽 파트너와 공동으로 개발한 가상 공간 기상 모델링 센터(Virtual Space Weather Modelling Centre, VSWMC)의 우주에서 발생하는 광범위한 물리적 메커니즘을 설명하는 수많은 모델의 연결성을 확보 후 Space HPC에 마이그레이션 적용
- ESA Space HPC는 우주 기상 애플리케이션을 포함하여 우주 산업의 전체 에코시스템에 서비스를 제공하도록 설계됨

(결과) 우주 기상 등 모델링의 Space HPC 시연을 통한 시뮬레이션 시간 단축 확인 및 해당 시스템의 서비스 액세스 허용 시작

- 우주 기상 등 모델링(예로 태양계의 모든 방향으로 전파되는 태양 분출을 추적하는데 자주 사용되는 모델 중 하나인 EUHFORIA)의 시뮬레이션 시간이 Space HPC를 사용하여 몇 분 단위로까지 단축 가능해짐
- Space HPC 시스템에 대한 액세스는 EU 회원국의 산업계와 학계가 표명한 요구 사항을 사례별 접근 방식으로 평가한 후 부여되는데 현재 해당 서비스에 대한 액세스 요청이 가능한 상태



태양권 모델링 기술 활동의 일부인 성숙한 모델 시각화-(상단 왼쪽)COCONUT, (상단 오른쪽)Icarus AMR 스킴, (하단)플럭스 로프 CME

출처: ESA

1) ESA(European Space Agency): 유럽 각국이 공동으로 설립한 우주개발기구

2) ESRIN(ESA Centre for Earth Observation): ESA 소속 연구센터로 위성의 지구 관측 데이터와 관련된 연구 수행

결론 및 시사점

- 우주 기상 현상의 과학에 대한 이해를 높이고 태양 내부의 구조, 코로나 질량 방출, 우주 기상의 행성 영향성 등 미지의 영역에 대한 모델 생성에 Space HPC가 크게 기여할 것임
- 우주 기상 전문 분야에서의 유럽 내 전용 ESA의 Space HPC 슈퍼컴퓨터 도입으로 연구 효율 극대화를 추진하는 것임, 이는 국내의 초고성능컴퓨팅 국가센터와 연계된 전문센터 정책 방향과 맥을 같이 하는 것으로 우주 기상 분야의 적극적인 운영 정책이 필요해 보임

12

Ansys와 Baker Hughes, Frontier를 활용하여 획기적으로 CFD 시뮬레이션 시간 단축

개요

Frontier[1]에서 실행된 사상 최대 규모의 Ansys Fluent 전산유체역학(CFD)[2] 시뮬레이션에서 96%의 획기적인 시간 단축

- 에너지 회사인 Baker Hughes와 응용 소프트웨어 업체 Ansys는 DOE(미국 에너지부, Department of Energy)의 ORNL(오크리지 국립연구소, Oak Ridge National Laboratory)이 보유중인 Frontier 엑사스케일 슈퍼컴퓨터를 사용해 2억 셀 규모의 축방향 터빈 고정자 시뮬레이션을 수행
- AMD EPYC CPU 및 Instinct GPU로 구동되는 Frontier 엑사스케일 슈퍼컴퓨터의 성능을 활용하여 Ansys Fluent®를 1,024개의 AMD Instinct™ MI250X의 GPU로 확장, 짧은 시뮬레이션 시간 내에 큰 작동 압력에서의 열공력 물리학(aerothermal physics)에 대한 탁월한 통찰력을 확보

(접근방안) 2억 셀의 축방향 터빈 고정자 시뮬레이션에 대한 GPU와 CPU 코어와의 직접 비교

- Baker Hughes와 Ansys는 2억 셀의 축방향 터빈 고정자 시뮬레이션에 대해 Frontier 슈퍼컴퓨터에서 1,024개의 AMD Instinct MI250X GPU와 3,700개 이상의 CPU 코어를 사용한 시뮬레이션 간의 상호 계산 시간을 비교

(결과) Fluent의 GPU 솔버를 통해 시뮬레이션 실행 시간의 획기적 단축으로 터빈 제품 개발 속도 향상

- 2억 셀의 축방향 터빈 고정자 시뮬레이션을 수행하는데 Fluent의 CPU 솔버는 38.5시간 걸렸으나 GPU 솔버는 단 1.5시간 만에 수행
- 터빈 고정자 개발 단계에서 충실도가 높은 시뮬레이션 결과로 임계 유동(critical flow) 및 난류 구조를 식별할 수 있어 터빈 제품의 개발 속도를 높일 수 있음

ANSYS Fluent®를 사용한
2억 셀의 축방향 터빈 고정자 시뮬레이션

출처: Ansys

결론 및 시사점

- Fluent 등의 GPU 솔버 발전으로 중소기업에서도 작은 규모의 GPU시스템을 구축하면 충실도 높은 전산유체역학 시뮬레이션을 확보할 수 있게 됨
- 최첨단 GPU 컴퓨팅의 보급에 따라 고충실도의 해석 결과와 함께 시뮬레이션 속도를 높일 수 있는 고급 GPU 지원 솔버들의 개발 연구가 필요하며, 이를 통해 우수 제품 개발 검증을 신속하게 추진할 수 있음

1) Frontier: 오크리지 국립 연구소에 설치돼 있는 세계 최초의 엑사스케일 슈퍼컴퓨터

2) CFD(Ciomputational Fluid Dynamics): 유체의 움직임을 컴퓨터로 계산하는 과학

13

TACC Frontera를 활용한 미세소관 말단에서의 새로운 행동 확보

개요

TACC(텍사스첨단컴퓨팅센터)의 Frontera[1]를 활용, 미세소관(microtubules)[2] 말단에서의 단편 추가 또는 감소의 새로운 행동 확인

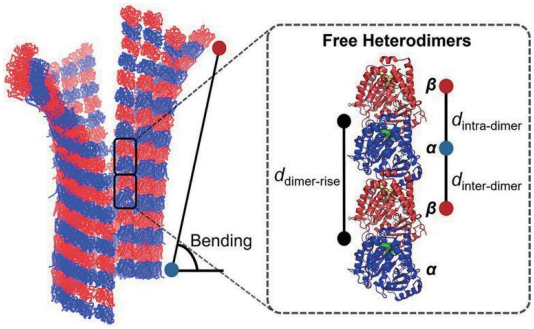
- 성인 한 명은 매일 평균 500억 개에서 700억 개의 세포를 잃으며, 세포 분열의 복잡한 과정을 통해 다시 잃어버린 세포를 대체함
- 미세소관이 세포 분열과 화학적 수송에 중요한 역할을 하며 분자 기계(molecular machines)[3]를 위한 길을 닦고 핵을 밀어 분리하여 분열시킴
- 시카고 대학(University of Chicago)과 유타 대학(University of Utah)의 과학자들은 Frontera를 통한 시뮬레이션에서 미세소관이 자라거나 줄어드는 위치인 미세소관 말단에서의 단편 추가 또는 감소의 행동을 새롭게 확보함

(접근방안) 리더십 연구 지원을 통한 Frontera 자원 확보 및 분자동역학 시뮬레이션을 수행 후 기계 학습으로 확장하여 미세소관 말단의 행동 확인

- 미국 국립과학재단(National Science Foundation)의 자금 지원을 받는 미국 최고의 학술 슈퍼컴퓨터인 Frontera에 대한 리더십 연구로 자원을 할당받음
- 5,600만 시간의 Frontera 중앙처리장치(CPU) 코어 시간을 사용하여 4μs의 AA MD(All-Atom Molecular Dynamics) 시뮬레이션을 수행
- 항상 펼쳐진 상태인 미세소관 말단에서의 뉴클레오시드 구아노신-5'-트리포스페이트(GTP)와 구아노신 이인산(GDP)에 따른 펼침 차이에 대한 분석
- AA MD를 통한 생성 데이터를 미세소관 말단의 이완 상태까지의 시뮬레이션을 수행할 수 있도록 기계 학습 방법에 공급함, 이는 2,100만 개에서 3,800만 개의 원자로 구성된 시스템에서 5.875μs가 걸리는 진행 과정임

(결과) 기계 학습을 포함한 다중 스케일 시뮬레이션 방법을 통해 계산시간 축소 및 미세소관 말단에서의 GTP 또는 GDP 연계에 따른 새로운 행동을 확보

- 기계 학습을 통합한 다중 스케일 시뮬레이션 방법으로 시뮬레이션 시간을 4μs에서 5.875μs로 확장하고 AA MD로만 수행 시의 추가 1,500만 CPU 시간을 절약
- 미세소관 말단에서 GTP 또는 GDP와의 연계 여부에 따라 해당 구조에서의 미묘한 행동 차이가 있음을 확인



튜불린(tubulin)의 빌딩 블록으로 구성된
동적 구조 폴리머인 미세소관(microtubules)

출처: DOI:10.1016/j.ihcs.2025.103460

1) Frontera: 38.75PF 성능(Rpeak)으로 Top500 52위(2024.11 기준)인 TACC의 슈퍼컴퓨터

2) 미세소관(microtubule): 길고 속이 빈 관 모양의 폴리머 구조로 세포골격의 일부이며 튜블린(tublin)의 단백질 구성 요소로 이뤄짐

3) 분자기계(molecular machines): 특정 자극에 반응하여 준기계적인 움직임을 생성하는 분자 성분

결론 및 시사점

- 슈퍼컴퓨터는 기계 학습을 활용하여 세포에서 매우 중요한 단백질 집합의 새로운 행동을 발견하는 데 필요한 적정 데이터를 제공해 줄 수 있음
- 알츠하이머병 및 파킨슨병과 같은 신경퇴행성 질환에 대한 이해를 높이고 암 약물 설계를 돕는 데에도 유용

14

LLNL Sierra 슈퍼컴퓨터를 활용한 둔감 고폭탄에서의 핫스팟 형성 시뮬레이션

개요

시에라(Sierra)[1]를 활용해 TATB[2] 둔감 고폭탄에서의 핫스팟(hot-spot)[3] 형성 관련 분자동력학(MD) 시뮬레이션 확보

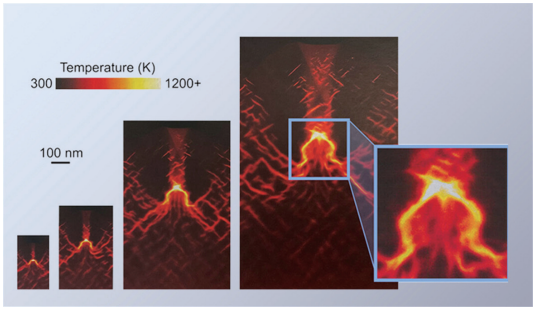
- 고폭탄이 갑작스러운 충격파를 받을 때 공극과 같은 미세 구조 결함으로 핫스팟이 형성
- LLNL(Lawrence Livermore National Laboratory) 연구팀은 Sierra 슈퍼컴퓨터를 사용하여 TATB 물질로 된 둔감 고폭탄에서의 미세한 핫스팟 형성 과정을 분자동력학 시뮬레이션으로 확인

(접근방안) 분자동력학을 기반으로 핫스팟 형성 과정 파악 및 더 큰 기공으로 확장 상태까지의 분석을 위해 연속체 기반 ALE3D 시뮬레이션을 추가 사용

- TATB 폭발성 재료를 최대 6억 개의 원자(현재까지 최대 규모)로 구성여 분자동역학 시뮬레이션을 수행
- 분자동력학 시뮬레이션에 적용된 원자 모델링과 연속체 기반의 다중 물리 유한요소 코드인 ALE3D 시뮬레이션에 사용된 미세 구조 규모 모델링 간의 격차를 확인

(결과) 분자동역학 시뮬레이션에서 관찰된 핫스팟 거동이 연속체 단위의 ALE3D 시뮬레이션에서도 동등 수준으로 나타남에 따라 공극 길이 스케일과 무관함을 확인

- 분자동역학 시뮬레이션을 통해 20nm보다 큰 기공에서도 핫스팟 형성은 공극 크기에 관계없이 불변함을 확인
- ALE3D 시뮬레이션을 통해 핫스팟이 공극 크기에 무관하고 TATB의 기계적 강도(응력-변형률 선도의 제어)에만 관련 있음을 확인함
- 공극 붕괴와 관련된 초고속 변형률에서 TATB 결정에 대한 응력-변형률 반응은 변형 속도에도 영향을 받지 않음



분자동역학 시뮬레이션을 통한 기공 붕괴 중의 핫스팟 형성 스냅샷

출처: LLNL

결론 및 시사점

- 분자동역학 시뮬레이션에서 핫스팟 모델링을 단순화시킬 수 있음을 확인하였으며, 더 큰 공극 길이의 스케일 적용 시 분자동역학 시뮬레이션이 가이드 역할을 할 수 있음
- 분자동역학 시뮬레이션으로부터 일반적인 다중 스케일 모델로 확장 가능하므로 고성능의 안전한, 새로운 폭발성 물질을 개발할 수 있을 것임

1) Sierra: LLNL이 보유한 125.71PF 성능의 Top500 14위(2024.11) 슈퍼컴퓨터

2) 핫스팟(hot-spot): 작은 공극 등 영역에서 발생한 강렬한 열

3) TATB(2,4,6-트리니트로벤젠-1,3,5-트리아미노): 방향족 폭발물로 매우 강력하지만 충격, 진동, 화재에 매우 둔감함

05 초고성능컴퓨팅 정책 동향

01
유럽의 AI 모델 개발을 발전시키기 위한 EuroHPC 자금 지원 MINERVA 프로젝트

개요

EuroHPC JU의 자금 지원을 받는 MINERVA 프로젝트, 유럽의 AI 생태계 강화를 목표로 HPC 기술을 활용해 AI 모델 개발 및 협업 촉진

- (목표) 최첨단 HPC 인프라를 ML/AI 연구자와 개발자에게 접근 가능하게 함으로써 AI 연구 개발 가속화
- (서비스 및 지원) EuroHPC 슈퍼컴퓨팅 인프라를 활용, 대규모 AI 모델을 다룰 수 있는 고급 HPC 기능 제공, AI 커뮤니티를 위한 서비스 및 교육 제공
- (주요 활동) AI 친화적인 과학, 사회, 산업 애플리케이션 개발 추진 및 AI 기술에 대한 인식 및 전문성 향상
- (협력 네트워크) 이탈리아, 프랑스, 스페인, 독일, 핀란드 등 유럽 국가 6개 연구 기관 및 기업들이 참여하며 CINECA Consorzio Interuniversitario가 프로젝트를 조정
- (예산 및 기간) Horizon Europe 자금 프로그램을 통해 5백만 유로 지원, 2027년까지 3년간 진행

결론 및 시사점

- 유럽의 AI 연구 및 개발을 촉진할 수 있는 중요한 기반을 마련하기 위한 장치로 특히 대규모 AI 모델을 구현하는 데 필요한 HPC 인프라를 제공
- AI와 HPC에 대한 전문성을 확장하고 교육을 통해 더 많은 인재를 양성할 수 있는 기회를 제공함으로써 유럽 AI 산업의 지속적인 발전에 기여할 것으로 예상
- 유럽이 HPC를 활용한 AI 분야에서 중요한 기술 중심지로 자리잡는데 기여할 수 있는 중요한 이니셔티브로, 향후 한국의 AI 및 HPC 기술 융합 환경 구축에 참고할만한 프로젝트

02

EuroHPC, 유럽의 독자적인 HPC 및 AI 개발을 위한 DARE 프로젝트 지원

개요

EuroHPC JU의 자금 지원을 받는 DARE(Digital Autonomy with RISC-V in Europe) 프로젝트 공식 출범

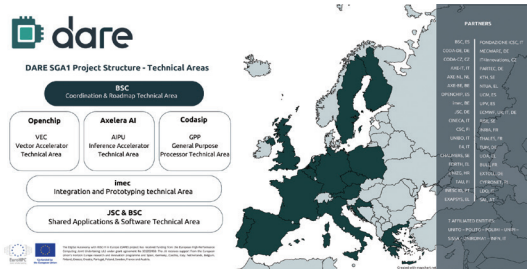
- (목표) RISC-V 기반의 최첨단 고성능 컴퓨팅(HPC) 하드웨어 및 소프트웨어 개발
- (조직) 13개국 38개 기관(중소기업 포함)이 참여(BSC, 바르셀로나 슈퍼컴퓨팅 센터가 프로젝트 총괄)
- (기간 및 예산) 6년간(~2030)진EuroHPC JU를 통해 최대 1억 2천만 유로(약 1,700억원)의 예산 지원 예정

RISC-V 아키텍처를 기반으로 1개 프로세서와 2개 가속기 설계 및 개발 예정

- 벡터 가속기(VEC): 정밀한 HPC 및 HPC-AI 융합 응용 프로그램 지원
- AI 프로세싱 유닛(AIPU): AI 추론 가속을 위한 전용 프로세서
- 범용 프로세서(GPP): 유럽 슈퍼컴퓨터에서 활용될 HPC 최적화 프로세서

유럽 프로세서 이니셔티브(EPI), MEEP, eProcessor, EUPEX 등 EU에서 추진해 온 기존 HPC 연구를 기반으로 진행

- 유럽은 프로세서, (AI) 가속기, 양자 칩 등 다양한 첨단 하드웨어 기술 개발을 지속적으로 확대해나가는 추세
- (하드웨어) EU에서 설계 및 개발한 산업표준 칩렛을 기반으로 프로토타입 HPC 및 AI 시스템 구축 예정
- (소프트웨어) 전체 RISC-V HPC 및 AI 소프트웨어 스택 개발



출처: EuroHPC JU

결론 및 시사점

- 기존의 미국, 아시아 반도체 기업(인텔, AMD, ARM 등)에 대한 의존도를 낮추고 개방형 아키텍처인 RISC-V를 이용한 유럽의 독자적인 프로세서 및 HPC 기술 생태계를 구축하려는 전략의 일환
- 유럽 주도의 하드웨어 및 소프트웨어 기술 발전을 통한 유럽의 디지털 주권 강화 시도이며, 유럽의 향후 HPC 및 AI 산업 경쟁력 강화 예상
- 장기적인 정책 지원을 바탕으로 유럽이 미래 반도체 및 슈퍼컴퓨팅 산업에서 독립적인 기술 역량을 갖추게 될 계기

03

DOE, 잠재적 AI 중심 데이터 센터 개발을 위한 16개 연방 부지 선정

개요

미 에너지부(DOE), AI 발전과 에너지 비용 절감을 위해 자사 부지에 데이터 센터와 새로운 에너지 인프라를 함께 구축하는 계획 발표

- 16개의 부지 선정: 핵과 같은 새로운 에너지 생산에 대한 허가를 빠르게 처리할 수 있는 기존 에너지 인프라 포함
- 공공-민간 파트너십 우선시: AI 및 에너지 인프라 개발을 위한 파트너십 확대 예정
- 개발 목표 시점: 2027년 말까지 AI 인프라 운영 개시
- 업계에 DOE의 연구 시설과의 협력 기회를 제공함으로써 차세대 데이터 센터 기술 공동 개발 기대

RFI(정보요청서) 공개

- (목표) AI 인프라 구축과 관련된 잠재적인 개발 접근 방식, 기술 솔루션, 운영 모델, 경제적 고려 사항에 대한 정보 수집
- (대상) 데이터 센터 개발자, 에너지 개발자 및 일반 대중의 의견을 구함
- (제공 정보) 위치, 사용 가능한 면적 및 기타 특성을 포함한 각 사이트에 대한 공개 정보를 포함

결론 및 시사점

- 미국은 AI 시대의 글로벌 기술 리더십을 놓지 않기 위해 전략적으로 대응하고 있으며, 인프라를 비롯하여 기술, 인재, 산업까지 전방위적 리드를 강화해나가고 있음
- AI 경쟁은 전력·서버·부지 등 하드웨어 인프라 확보가 핵심이며, 한국도 데이터 센터와 에너지 인프라를 함께 고려하는 전략적 접근이 필요한 상황임
- 한국도 국가 공공 부지를 연계할 수 있으며, 입지, 전력, 규제 혜택을 결합한 패키지형 AI 클러스터 조성 정책에 대한 고려가 필요함

04

EuroHPC, 국가 AI 생태계 지원을 위한 AI 팩토리 안테나 제안 개시

개요

EuroHPC JU는 유럽 전역에 선정된 AI 팩토리와 관련, AI 팩토리 안테나를 구축할 기업 또는 기업 컨소시엄을 선정하기 위한 제안 요청을 개시

• AI 팩토리 안테나는 해당 국가의 기존 AI 팩토리를 보조하는 분산 거점을 의미하며 본 AI 팩토리와의 긴밀한 협력을 위해 안테나와 팩토리간 양해각서(MoU) 체결 예정

유럽 스타트업과 중소기업(SME)의 AI 컴퓨팅 자원 접근성 강화

- 범용 대규모 AI 모델의 교육 및 개발뿐만 아니라 새로운 AI 애플리케이션의 개발, 테스트 및 검증 지원

(예산) 총 예산은 최대 7천만 유로로, 개별 안테나당 최대 500만 유로 규모로 지원

AI 팩토리 안테나의 역할

- EuroHPC AI 팩토리와의 긴밀한 협력을 통한 공동 미션 수행 및 목표 달성에 협력
- AI 관련 서비스의 확장 및 보완
- 원격으로 AI 최적화 슈퍼컴퓨팅 자원에 접근 제공
- 필요시에는 AI 애플리케이션 튜닝, 테스트 및 검증을 위한 소규모 AI 컴퓨팅 자원 제공

결론 및 시사점

- EuroHPC JU는 AI 팩토리 구축 뿐만 아니라 중소기업·스타트업의 AI기술 접근성 확대에 중점을 둔 AI 팩토리 안테나 구축을 추진하고 있으며, 이를 통해 상대적으로 자원이 부족한 국가나 기관도 AI 생태계에 참여할 수 있는 길을 열어줌으로써 전 유럽적 균형 발전을 꾀하고 있음
- AI 모델 학습, 튜닝, 검증을 위한 실험 공간 및 자원 제공 측면에서 중소기업 및 스타트업에 중요한 기회로 작용할 수 있는 측면이 있어서 추후 한국에서도 벤치마킹을 통한 유사한 전략을 고려할 필요가 있음

출 처

CHAPTER 01

초고성능컴퓨팅 시장 및 서비스 동향

1. https://www.dt.co.kr/contents.html?article_no=2024122302109931065001
2. https://www.digitimes.com/news/a20241226PD209/amd-qualcomm-intel-market-hpc.html
3. https://www.datacenterdynamics.com/en/opinions/four-key-trends-disrupting-data-centers-in-2025/
4. https://www.prnewswire.com/news-releases/sk-hynix-to-unveil-full-stack-ai-memory-provider-vision-at-ces-2025-302341613.html
5. https://www.hpcwire.com/off-the-wire/penguin-solutions-signs-ai-data-center-agreement-with-sk-telecom-and-sk-hynix/
6. https://www.prnewswire.com/news-releases/artificial-intelligence-ai-chips-market-to-grow-by-usd-902-65-billion-2025-2029-driven-by-ai-chip-innovation-for-smartphones-ai-driving-transformation---technavio-302354514.html
7. https://www.hpcwire.com/off-the-wire/multiverse-computing-partners-with-kinesis-to-optimize-ai-while-reducing-energy-demands/
8. https://www.hpcwire.com/off-the-wire/supermicro-announces-support-for-nvidia-rtx-pro-6000-blackwell-server-edition-gpus/
9. https://www.hpcwire.com/off-the-wire/softbank-group-to-acquire-ampere-computing/
10. https://www.hpcwire.com/2025/03/24/jensen-huang-charts-nvidias-ai-powered-future/
11. https://www.aiwire.net/2025/03/24/transforming-engineering-workflows-with-ai-driven-knowledge-management/
12. https://www.hpcwire.com/off-the-wire/ibm-unveils-z17-to-expand-ai-inference-system-integration/

CHAPTER 02

초고성능컴퓨팅 인프라 구축 동향

1절

1. https://www.hpcwire.com/2024/12/11/nvidias-blackwell-showcases-the-future-of-ai-is-water-cooled-for-now/
2. https://www.hpcwire.com/off-the-wire/aws-invests-10b-in-ohio-for-statewide-data-center-expansion/
3. https://www.hpcwire.com/2024/12/11/nvidias-blackwell-showcases-the-future-of-ai-is-water-cooled-for-now/
4. https://insidehpc.com/2025/01/eni-officially-unveils-e100m-hpc6-supercomputer-for-exploration-and-decarbonization/
5. https://insidehpc.com/2025/01/infortrend-unveils-storage-for-hpc-and-media-entertainment/
6. https://www.hpcwire.com/off-the-wire/seagate-readies-36tb-hamr-hard-drives-for-data-center-deployments/
7. https://www.hpcwire.com/off-the-wire/spectra-logic-introduces-24g-optical-sas-switch-for-expanded-tape-connectivity/
8. https://www.hpcwire.com/off-the-wire/vdura-showcases-next-gen-data-platform-at-rice-universitys-energy-hpc-conference/
9. https://www.hpcwire.com/2025/01/13/uss-doe-details-the-next-major-supercomputer-a-companion-to-el-capitan/
10. https://www.hpcwire.com/off-the-wire/fluidstack-to-build-1gw-ai-supercomputer-in-france/
11. https://insidehpc.com/2025/02/why-tier-0-is-a-game-changer-for-gpu-storage/
12. https://www.hpcwire.com/off-the-wire/coolit-debuts-high-capacity-row-based-cooling-solution-for-ai-and-hpc/
13. https://insidehpc.com/2025/03/report-oracle-to-deploy-ai-cluster-with-30000-amd-mi355x-accelerators/

2절

1. https://www.hpcwire.com/off-the-wire/hlrs-dynamic-power-capping-enables-better-energy-efficiency-in-hpc/
2. https://www.businesswire.com/news/home/20250317715500/en/Cineca-to-House-Italys-Most-Powerful-Quantum-Computer-IQM-Radiance-54
3. https://www.hpcwire.com/off-the-wire/pasqal-to-deliver-euroqcs-italy-quantum-system-under-eurohpc-procurement-deal/
4. https://thequantuminsider.com/2025/04/15/qpiai-launches-25-qubit-superconducting-system-under-indias-national-quantum-mission/

출 처

CHAPTER 03

초고성능컴퓨팅 기술개발 동향

1절

1. https://www.hpcwire.com/off-the-wire/meta-to-build-10b-ai-data-center-in-northeast-louisiana-by-2030/
2. https://hpcwire.com/off-the-wire/aws-launches-ec2-u7inh-instance-for-large-in-memory-databases-on-hpe-servers/
3. https://techcrunch.com/2024/12/20/openai-announces-new-o3-model/
4. https://www.artificialintelligence-news.com/news/microsoft-releases-phi-4-language-model-hugging-face/
5. https://www.hpcwire.com/2025/01/23/how-sandboxaq-is-leading-a-quiet-revolution-in-science-and-medicine/
6. https://www.hpcwire.com/off-the-wire/altair-expands-hpcworks-with-ai-driven-scheduling-and-cloud-scaling/
7. https://sakana.ai/ai-scientist-first-publication-jp/
   https://github.com/SakanaAI/AI-Scientist-ICLR2025-Workshop-Experiment
8. https://www.tredence.com/blog/llm-inference-optimization#:~:text=Reduced%20Operational%20Costs%3A
9. https://quantumcomputingreport.com/fujitsu-launches-open-source-quantum-computer-operations-software-on-github/
10. https://www.hpcwire.com/off-the-wire/ualink-consortium-releases-the-ultra-accelerator-link-200g-1-0-specification/
11. https://www.youtube.com/watch?v=mXJkGF37rAE
12. https://www.hpcwire.com/off-the-wire/amd-tapes-out-1st-hpc-product-on-tsmc-2nm-process-with-venice-epyc-cpu/
13. https://openai.com/index/browsecomp/
    https://openai.com/index/openai-pioneers-program/

2절

1. https://www.hpcwire.com/off-the-wire/argonne-training-series-helps-prepare-a-new-generation-of-ai-ready-researchers/
2. https://www.hpcwire.com/off-the-wire/sandia-partners-with-national-labs-to-develop-energy-efficient-ai-and-computing-tech/
3. https://www.hpcwire.com/off-the-wire/bsc-integrates-new-quantum-system-into-marenostrum-5/
4. https://www.globaltimes.cn/page/202504/1331580.shtml

CHAPTER 04

초고성능컴퓨팅 응용 및 활용 동향

1. https://www.hpcwire.com/off-the-wire/hpc-simulations-at-lrz-explore-cascading-earthquake-behavior-and-risks/
2. https://www.hpcwire.com/off-the-wire/berkeley-researchers-crack-open-ai-at-scale-method-for-chemical-science/
3. https://www.hpcwire.com/off-the-wire/access-supports-large-eddy-simulations-for-ocean-wind-turbine-design/
4. https://www.hpcwire.com/off-the-wire/psc-anton-studies-reveal-blood-vessel-response-to-mechanical-force/
5. https://www.hpcwire.com/off-the-wire/ornl-researchers-leverage-summit-to-study-electron-behavior-in-iter-tokamak/
6. https://www.hpcwire.com/off-the-wire/argonne-team-delivers-a-100x-speedup-of-genetic-data-analysis-from-the-million-veteran-program/
7. https://www.hpcwire.com/off-the-wire/psc-bridges-2-simulations-show-impact-of-rocket-exhaust-water-on-moons-ice-reservoirs/
8. https://www.hpcwire.com/off-the-wire/hlrs-simulation-supports-search-for-personalized-cancer-treatments/
9. https://www.hpcwire.com/off-the-wire/argonne-uses-computational-modeling-to-study-solid-electrolytes-for-batteries/
10. https://www.hpcwire.com/off-the-wire/supercomputers-reveal-how-small-ocean-processes-influence-storms/
    https://phys.org/news/2025-03-supercomputers-reveal-small-ocean-storms.html
11. https://www.hpcwire.com/off-the-wire/esa-space-hpc-to-accelerate-modeling-of-solar-storms-and-space-weather/
12. https://www.hpcwire.com/off-the-wire/ansys-and-baker-hughes-use-frontier-to-cut-cfd-runtime-from-38-5-to-1-5-hours/
    https://investors.ansys.com/news-releases/news-release-details/ansys-baker-hughes-and-oak-ridge-national-laboratory-set-new
13. https://www.hpcwire.com/off-the-wire/taccs-frontera-enables-new-insights-into-microtubule-tip-dynamics/
14. https://www.hpcwire.com/off-the-wire/llnl-simulating-hot-spot-formation-in-insensitive-high-explosives/

출 처

CHAPTER 05

초고성능컴퓨팅 정책 동향

1. https://www.hpcwire.com/off-the-wire/eurohpc-funded-minerva-project-to-advance-ai-model-development-in-europe/
2. https://www.hpcwire.com/off-the-wire/eurohpc-supports-dare-project-for-sovereign-hpc-and-ai-development/
3. https://www.hpcwire.com/off-the-wire/doe-identifies-16-federal-sites-for-potential-ai-focused-data-center-development/
4. https://www.hpcwire.com/off-the-wire/eurohpc-opens-proposals-for-ai-factory-antennas-to-support-national-ai-ecosystems/

한국과학기술정보연구원
Korea Institute of Science and Technology Information

발행일 2025. 05. 30

발행인 이 식

편집위원 서민호, 허영주, 고명주, 심형욱,
안설아, 엄정호, 온누리

발행처 한국과학기술정보연구원
초고성능컴퓨팅정책센터
34141 대전광역시 유성구 대학로 245
www.kisti.re.kr

ISSN 2733-7561