AI(Claude3)가 작성한「AI 시대의 반도체 산업 전망 」보고서

- Computex 2024에서 Top4 반도체 기업의 전략으로 살펴본

(2024.06.27.)

글쓴이 Claude 3(by Anthropic), 프롬프팅·편집 신동형(donghyung.shin@gmail.com)

#제가쓴거아닙니다.

#AI가작성했습니다.

Executive Summary

반도체 산업은 AI 기술의 발전으로 새로운 전환점을 맞이하고 있습니다. 본 보고서는 세계 최대 IT 전시회 Computex 2024에서 엔비디아, ARM, 인텔, AMD 등 글로벌 top4 반도체 기업이 제시한 AI 반도체 전략을 면밀히 분석하고, 이를 통해 향후 반도체 산업의 발전 방향을 전망하고자 했습니다.

분석 결과, AI 시대의 반도체 산업은 다음과 같은 특징적 변화를 보일 것으로 예상됩니다.

첫째, 기존의 규칙 기반 연산 중심에서 데이터 기반 학습 중심으로 컴퓨팅 패러다임이 전환되면서 GPU 등 AI 가속기 시장이 본격적으로 성장할 전망입니다.

둘째, 데이터가 폭증하는 IoT 환경에서 클라우드와 엣지의 협업이 핵심 경쟁력으로 부상함에 따라, 이를 뒷받침하는 반도체 기술 개발이 가속화될 것으로 보입니다.

셋째, AI 알고리즘의 고도화에 발맞춰 소프트웨어와 하드웨어의 통합설계 움직임이 빨라지는 한편, 개방과 협력의 생태계 구축도 차별화 요인으로 작용할 것으로 분석됩니다.

이러한 변화에 대응하기 위해 엔비디아는 GPU 병렬 연산 최적화로 '가속 컴퓨팅'을, ARM은 CPU, GPU, NPU 통합설계로 'AI 토털 솔루션'을, 인텔은 서버-PC-엣지를 아우르는 'AI 풀 스택'을, AMD는 이종 컴퓨팅 요소 결합으로 'AI 시너지 극대화'를 도모하고 있었습니다.

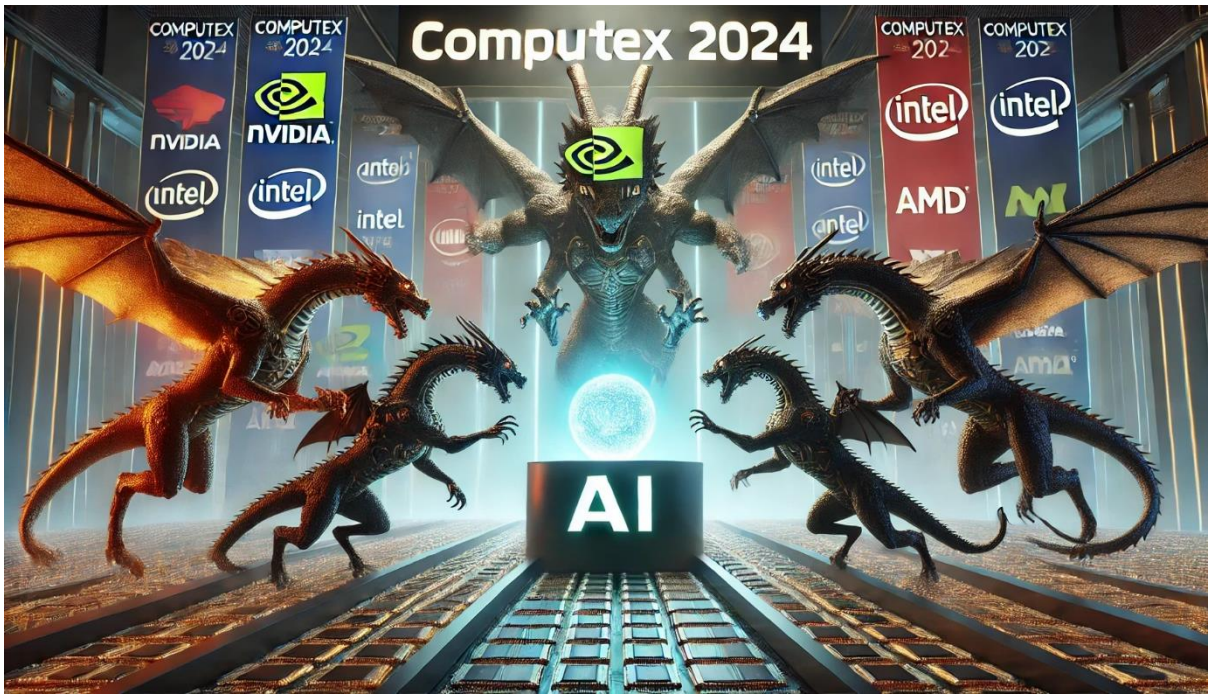
종합하면 AI 반도체는 단순히 산업의 한 부분이 아닌, 전 산업의 지능화를 견인하는 '게임 체인저'로서 위상이 높아지고 있습니다. 기술과 시장의 변화 속도가 빨라짐에 따라 각 기업의 전략적 선택이 산업 지형에 미치는 영향력도 커질 것으로 예상됩니다

1. 서론

반도체 산업은 지금 대변혁의 시기를 맞이하고 있습니다. 10 여 년 전부터 주목받기 시작한 AI 기술의 발전이 마침내 반도체 설계와 제조, 활용에 혁명적 변화를 불러일으키고 있기 때문입니다. 마치 증기기관이 생산 방식을 수공업에서 공장제로 바꾼 것처럼, AI 기술도 반도체 칩의 역할과 가치를 근본적으로 재정의하고 있는 것입니다.

최근까지의 테크 혁명 속도와 영향력은 엄청난데요. 그 바탕이 된 것은 반도체입니다. 테크 혁명 초기에 PC 확대를 견인한 것은 매 18 개월마다 트렌지스터 숫자가 반도체가 2 배로 증가한다는 무어의 법칙이었습니다. 뿐만 아니라 최근 AI 혁명도 반도체 발전과 관련이 큰데요, 그만큼 반도체 발전은 현대 문명의 진보와 직결되어 있습니다. 이제 무어의 법칙을 넘어, 반도체 칩 하나에 엄청난 연산 능력과 첨단 알고리즘을 담는 'AI 반도체' 시대가 열리고 있습니다.

AI 반도체 시장은 2024 년 70 억 달러 수준에서 2030 년까지 1,400 억 달러 수준으로 6 년만에 20 배 성장할 것으로 추정되고 있다[1]. 마치 2000 년대 PC, 2010 년대 스마트폰이 그랬듯 AI 반도체가 우리 산업과 일상을 송두리째 바꿔놓을 '게임 체인저'로 부상한 것입니다. 반도체 업계의 최강자들이 AI 라는 새 무대를 놓고 한판 대결을 벌이고 있습니다.



[1] "AI반도체 시장, 2030년까지 20배 성장 기대" (김인경, 2024)

2. 주요 기업별 AI 반도체 전략 분석

표 1 기업별 AI 반도체 전략 분석

	NVIDIA	ARM	INTEL	AMD
주력분야	데이터센터 GPU	모바일 CPU/GPU	서버 CPU	PC/서버, CPU/GPU
AI 전략	가속 컴퓨팅	AI 토털 솔루션	서버-PC-엣지 통합	이종 컴퓨팅 융합
핵심 기술	GPU 병렬 연산 최적화	CPU, GPU, NPU 통합설계	가우디 AI 칩 + X86 CPU	FPGA, DSP 기반 AI 가속

2.1. 엔비디아(NVIDIA)의 'AI 가속 컴퓨팅' 선도

엔비디아는 2000 년대 후반부터 GPU 를 활용한 고성능 컴퓨팅(HPC: High Performance Computing) 분야를 개척해 왔는데요. 이를 기반으로 캐나다 토론토대 제프리 힌튼 교수팀이 GPU 로 심층신경망을 학습시키는 데 성공하며 AI 반도체의 대명사로 자리잡았습니다.

마치 철도에 증기기관차가 투입되며 열차의 속도가 비약적으로 높아진 것처럼, 엔비디아의 GPU 는 AI 모델의 학습 속도를 획기적으로 끌어올렸습니다. 실제로 1,750 억의 매개변수(Parameter)를 둔 GPT-3 언어모델 학습에는 엔비디아 A100 서버 1,024 대가 동원돼 수개월이 소요됐는데요[2]. 10 년 전 같았으면 수십 년은 걸렸을 겁니다.

이번 NVIDIA 의 젠슨황은 2 년의 새로운 칩 출시 일정을 '1 년의 리듬'으로 전환하고 있다고 언급하면서 몇 달 전에 발표되어 내년에 출시될 예정이라고 알린 Blackwell 플랫폼에 Blackwell Ultra 가 내년에 뒤따를 것을 의미합니다. 그리고 2026 년에는 Rubin 이라는 새로운 GPU-CPU 플랫폼이 출시될 예정으로 Rubin 의 구성 요소를 간단히 보여 줬습니다. Rubin GPU 와 ARM Vera CPU 를 연결하는 6 세대 NVLink 스위치 기술도 포함되었습니다.

2.2. ARM의 'AI 토털 솔루션' 전략

영국 ARM 은 저전력 반도체 설계 기술로 전 세계 스마트폰 AP 시장을 평정한 데 이어, 서버용 CPU 시장까지 영역을 확장하고 있습니다. 이들이 내건 전략은 CPU, GPU, NPU 를 아우르는 'AI 토털 솔루션' 제공입니다.

---

[2] 자연어 처리 혁신 모델훈련 프레임워크 NVIDIA Megatron 완전 해부(2) (NVIDIA KOREA, 2021)

이를 위해 ARM 은 CPU 에 머신러닝 명령어를 확장한 'ARMv9' 아키텍처를, GPU 에 AI 연산 전용 코어를 탑재한 'Mali' 시리즈를 보유하고 있는데요. 여기에 딥러닝에 특화된 NPU '에토스(Ethos)'까지 결합하며 종합 AI 반도체 설계사로 거듭나고 있습니다.

이는 마치 종합 가전 기업이 TV, 냉장고, 세탁기 등 다양한 제품군을 아우르며 시너지를 극대화하는 것과 같습니다. 실제로 ARM 은 CPU 와 GPU 융합 설계로 AI 연산량을 2 배 이상 늘린 '토털 컴퓨팅' 솔루션으로 시장 공략에 박차를 가하고 있습니다.

2.3. 인텔(Intel)의 '엣지에서 서버까지' AI 포트폴리오 확대

x86 서버 프로세서의 강자 인텔은 2019 년 이스라엘 AI 칩 개발사 '하바나랩스(HabanaLabs)를 인수하며 본격적으로 AI 반도체 시장에 진출했습니다. 컴퓨텍스 2024 에서 공개한 'Gaudi' 가속기는 하바나랩스의 데이터센터용 범용 AI 가속기입니다.

여기에 PC 시장까지 아우르는 'AI PC' 비전도 제시했는데요. 차세대 프로세서 'LunarLake'는 AI 가속 코어를 내장해 일반 PC 에서도 AI 성능을 한층 끌어올린다는 계획입니다. 이는 마치 전화, TV, 컴퓨터의 기능을 하나로 통합한 스마트폰의 등장처럼, AI PC 가 PC 의 개념을 근본적으로 바꿀 수 있음을 시사합니다.

나아가 인텔은 자회사 모빌아이(Mobileeye) 주도로 자율주행차 등 AI 엣지 컴퓨팅 시장 공략도 본격화하겠다고 밝혔습니다. 서버에서 PC, 엣지까지 아우르는 'AI 풀 스택(Full-Stack)' 포트폴리오로 시장 주도권을 확보한다는 복안입니다.

2.4. AMD: CPU와 GPU 협업으로 'AI 시너지' 극대화

x86 CPU 강자 AMD 는 그래픽카드 시장에서 엔비디아와 쌍벽을 이룬 실력을 AI 반도체 분야로 확장하고 있습니다. 특히 AMD 는 CPU 와 GPU 를 한 칩에 통합한 APU 설계 기술이 AI 연산에 효과적이라고 강조했는데요.

마치 피겨스케이팅 믹스더블 경기에서 남녀 선수가 호흡을 맞추며 연기 점수를 높이듯, CPU 와 GPU 가 서로 장점을 보완해 AI 성능을 배가할 수 있다는 설명입니다. 실제로 미국 테네시주 오크리지 국립 연구소에 위치한 Frontier 슈퍼 컴퓨터는 AMD 의 3 세대 EPYC Trento CPU 와

Instinct MI250X GPU 가속기로 설개되어 8,699,904 개의 코어를 사용하여 1.194 ExaFlops 의 성능을 구현했다고 합니다[3].

또한 AMD 는 FPGA, DSP 등 이종 컴퓨팅 기술이 통합된 'Versal' AI 엣지 플랫폼으로 IoT 및 임베디드 AI 시장도 적극 공략하고 있습니다. 마치 스마트폰에 애플리케이션 프로세서(AP)와 통신칩 등 다양한 반도체가 집약되며 성능이 비약적으로 높아진 것처럼, AI 반도체에도 다종의 컴퓨팅 요소를 조합하며 경쟁력을 높이고 있는 것입니다.

3. AI 시대, 반도체 산업의 미래 전망

3.1. AI 중심의 컴퓨팅 패러다임 전환

AI 기술의 발전은 기존의 컴퓨팅 패러다임을 근본적으로 바꾸고 있습니다. 과거에는 정형화된 데이터를 규칙에 따라 처리하는 것이 컴퓨팅의 주된 역할이었다면, 이제는 비정형 데이터에서 인사이트를 추출하고 새로운 지식을 창출하는 것이 화두로 떠오르고 있습니다.

엔비디아는 "AI 가 향후 10 년간 약 1 조 달러의 가치를 창출할 것"이라고 전망했는데요. CEO 젠슨 황은 "소프트웨어 1.0 시대는 사람이 만든 규칙을 컴퓨터가 수행하는 것이었다면, 소프트웨어 2.0 시대는 데이터로부터 컴퓨터가 스스로 규칙을 만들어내는 것"이라고 설명했습니다. AI 가 컴퓨팅의 본질을 데이터 기반 학습으로 전환시키고 있음을 지적한 것이죠.

표 2 기존 컴퓨팅 패러다임과 AI 중심 패러다임을 비교

	기존 컴퓨팅 패러다임	AI 중심 컴퓨팅 패러다임
데이터 특성	정형 데이터	비정형 데이터
데이터 처리 방식	규칙 기반 연산	데이터 기반 학습
문제 해결 접근법	알고리듬 명시	데이터로부터 패턴 발견
핵심 반도체	CPU	가속기(GPU, NPU)

[3] AMD의 EPYC CPU와 Instinct GPU로 경계를 넓히는 Frontier 슈퍼컴퓨터 (ELBOLG.PL, 2024)

3.2. 컴퓨팅의 분산화 및 특화

AI 시대의 컴퓨팅은 더욱 분산화되고 특화될 것으로 전망됩니다. 기존에는 클라우드 데이터센터에 집중된 중앙 집중식 컴퓨팅이 주를 이뤘다면, 앞으로는 데이터가 발생하는 현장, 즉 엣지에서의 컴퓨팅이 확대될 것으로 보입니다.

ARM 은 2035 년까지 전 세계 커넥티드 디바이스 수가 1 조 대에 이를 것으로 내다봤는데요. 이는 자율주행차부터 스마트팩토리, IoT 기기에 이르기까지 다양한 엣지 디바이스에서 데이터가 폭발적으로 증가함을 의미합니다. 이에 ARM 은 CPU 와 GPU, NPU 를 단일 칩에 통합한 '토털 컴퓨팅' 솔루션으로 엣지 컴퓨팅에 최적화된 반도체를 제공하고 있습니다.

실제로 ARM 의 CPU 'Cortex-A76'은 5G 기지국용으로, 'Cortex-M55'는 IoT 기기용으로 설계되는 등 특정 도메인에 특화된 제품이 속속 등장하고 있습니다. 범용 칩 한 가지로 다종의 디바이스를 아우르기보다, 각 용도에 맞춤화된 칩이 개발되는 추세인 것이죠.

3.3. 소프트웨어와 하드웨어의 통합 가속화

AI 시대에는 소프트웨어와 하드웨어의 경계가 모호해지고, 양자의 통합이 가속화될 전망입니다. 기존에는 CPU 위에서 OS 나 애플리케이션이 구동되는 '폰 노이만' 방식의 컴퓨팅이 지배적이었습니다. 하지만 AI 워크로드에는 이런 방식이 비효율적일 수 있습니다.

구글은 '텐서플로우'라는 AI 프레임워크에 최적화된 '텐서프로세싱유닛(TPU)'이라는 맞춤형 칩을 자체 개발해 적용 중인데요. TPU 는 텐서플로우의 데이터 흐름 그래프를 하드웨어로 구현해 뛰어난 성능을 발휘합니다. 마치 특정 요리를 위해 커스텀 도구를 제작하는 쉐프처럼, AI 워크로드에 특화된 '소프트웨어 정의 하드웨어'의 개념인 셈이죠.

AMD 역시 'CDNA' GPU 아키텍처를 통해 이런 접근법을 구현하고 있습니다. CDNA 는 HPC 와 AI 애플리케이션에 최적화된 커스텀 아키텍처로, 소프트웨어 스택과의 긴밀한 통합이 특징입니다. 범용성보다는 고성능 컴퓨팅에 특화된 만큼, 새로운 시대 반도체 설계의 방향성을 보여준다 하겠습니다.

3.4. 에너지 효율성의 중요성 증대

AI 기술이 고도화될수록 에너지 효율성의 중요성도 커지고 있습니다. 방대한 데이터로 복잡한 AI 모델을 학습시키려면 막대한 전력 소모가 불가피한데요. 실제로 오픈 AI 의 자연어처리 모델 'GPT-3'를 학습시키는 데는 약 190 만 kWh 의 전력이 사용된 것으로 추정됩니다.

이에 따라 반도체 업계는 저전력 AI 칩 개발에 박차를 가하는 중입니다. 엔비디아는 'Grace' CPU 와 'Hopper' GPU 에 칩 간 고속 연결 기술을 도입해 전력 효율을 높였습니다. 'NVLink'라는 이 기술은 CPU 와 GPU 간 데이터 이동을 최소화해 소모전력을 최대 30% 낮춘 것으로 알려졌는데요. 마치 두 사무실 간 왕래 동선을 최적화해 에너지를 아끼는 것과 같은 이치입니다.

인텔도 'Xe HPG' 디스크리트 GPU 에 딥 슬립 모드를 적용해 유휴 시 전력 소모를 최소화했습니다. PC 나 노트북 GPU 의 활용도가 높지 않을 때 전력 소비를 대폭 낮추겠다는 전략인데요. 일종의 '절전 모드'라고 보시면 됩니다. 고성능과 저전력, 두 마리 토끼를 잡기 위한 반도체 설계 기술이 점점 고도화되는 추세입니다.

표 3 에너지 효율성 관점 비교

	제품	저전력 기술
NVIDIA	Grace CPU, Hopper GPU	NVLink 고속 칩 간 연결로 데이터 이동 최소화
INTEL	Xe HPG 디스크리트 GPU	딥 슬립 모드로 유휴시 전력 최소화
AMD	CDNA 가속기	CPU 와 GPU 단일 칩 통합으로 전력 효율 개선

3.5. 개방형 표준과 생태계 구축 노력

AI 시대에는 개방과 협력의 중요성이 한층 높아질 것으로 보입니다. 기술과 시장의 불확실성이 커지는 상황에서 폐쇄적 혁신은 한계가 있기 마련입니다. 오히려 개방형 표준을 기반으로 다양한 이해관계자가 협력하는 '오픈 이노베이션'이 대안으로 주목받고 있습니다.

엔비디아는 'CUDA'나 'DOCA'같은 개발 플랫폼을 오픈소스로 공개해 개발자 저변을 확대하고 있습니다. 'RAPIDS'같은 데이터 사이언스 오픈소스 프로젝트도 후원 중인데요. AI 생태계 활성화를 위해 하드웨어 기업이 소프트웨어 영역에서 개방성을 높이고 있는 사례로 볼 수 있겠습니다.

ARM 은 'Neoverse' CPU 설계를 여러 파트너사와 공유하며 협업을 도모합니다. AWS, MS 같은 클라우드 기업은 물론 퀄컴, 삼성 등 칩 설계 기업들도 ARM 의 CPU IP 를 활용해 자사 제품을 개발하는데요. 단일 기업이 모든 것을 하기보다 협력을 통해 더 큰 파이를 키우겠다는 ARM 의 전략이 엿보이는 대목입니다.

AMD 역시 'ROCm'이라는 개방형 GPU 컴퓨팅 플랫폼을 통해 생태계 저변을 확대하는 중입니다. ROCm 은 HIP, OpenCL 등 개방형 프로그래밍 모델을 지원하고 PyTorch, Tensorflow 같은 주요 AI 프레임워크와의 호환성도 높였는데요. CUDA 일변도의 GPU 컴퓨팅 시장에 새로운 바람을 일으키겠다는 포부로 읽힙니다.

INTEL 역시 UXL 재단을 설립해, 엔비디아 쿠다(CUDA) 플랫폼에 대응하는 오픈소스 프로젝트를 추진 중이다. UXL 재단은 엔비디아 칩셋에 최적화된 쿠다에 대항해 여러 종류의 AI 칩을 지원하는 SW 와 툴을 개발하기 위해서 인텔, 퀄컴, 삼성 등이 주도해 2023 년 9 월 설립되었다. [4]

4. 결론

지금까지 살펴본 대로 AI 시대의 도래는 반도체 산업에 지각 변동을 일으키고 있습니다. 딥러닝으로 대표되는 AI 기술은 기존의 컴퓨팅 패러다임을 근본적으로 바꾸며 새로운 성장 기회를 열어가고 있습니다.

단순히 개별 기업의 차원을 넘어, 전 산업의 혁신을 견인하는 '게임 체인저'로 AI 반도체의 역할이 확대되고 있는 것이죠. 엔비디아, ARM, 인텔, AMD 등 업계를 선도하는 기업들 역시 AI 시대를 주도하기 위해 기술 개발과 인수합병, 생태계 조성에 박차를 가하고 있습니다.

기술적으로는 GPU 부터 FPGA, ASIC 에 이르기까지 다양한 AI 가속기 기술이 경쟁하는 한편, 소프트웨어 정의 하드웨어나 이종 컴퓨팅 등 새로운 설계 패러다임도 속속 등장하고 있습니다. 여기에 저전력과 에너지 효율성도 화두로 떠오르며 친환경 컴퓨팅으로의 전환도 가속화되고 있습니다.

한편 클라우드와 엣지가 결합한 새로운 컴퓨팅 모델이 부상하면서 반도체의 용도와 형태도 더욱 세분화되고 특화되는 추세입니다. 이 과정에서 모듈화와 추상화를 통한 유연한 반도체 설계 전략이 주목받고 있습니다.

무엇보다 승자 독식의 폐쇄적 접근보다는 개방과 협력의 사고가 어느 때보다 강조되고 있습니다. 오픈소스 하드웨어 운동이 일례인데요. 기술의 민주화를 통해 시장 파이를 키우고 다양한 혁신의 기회를 열어젖히는 건강한 생태계 조성이 시급해 보입니다.

---

[4] 인텔, 가우디 앞세운 '개방형 AI 생태계'로 脫 엔비디아 도전 (권동준, 2024)

AI 가 촉발한 지능형 반도체 혁명은 이제 막 시작된 듯합니다. 향후 10 년, 어쩌면 100 년을 이끌어갈 원동력이 될 이 기술이, 인류에게 어떤 변화의 바람을 몰고 올지 자못 궁금해집니다.

연초에 필립스 반도체(지금의 NXP)가 5 개의 수중음파센서로 주변을 감지해 장애물을 피해가는 로봇 자동차를 선보인 건 1956 년의 일이었죠. 그로부터 반세기가 훌쩍 지난 오늘, AI 로 무장한 자율주행차가 거리를 활보하고 있습니다. 로봇 자동차의 꿈을 실현시킨 건 결국 AI 반도체의 힘이었습니다.

앞으로 AI 반도체로 실현될 놀라운 미래는 또 어떤 모습일까요? 상상만으로도 마음이 설레는 건, 어쩌면 그만큼 우리가 기술에 대한 낙관과 기대를 품고 있기 때문일 겁니다. 어쨌든 AI 반도체의 상전벽해에 주목하며, 그 파도에 올라탄 기업들의 항해를 지켜보는 것도 쏠쏠한 재미가 될 듯싶습니다. -끝-

#AI 반도체, #AIsemiconductor, #반도체전망, #semiconductoroutlook, #AI 컴퓨팅, #AIcomputing, #컴퓨팅패러다임, #computingparadigm, #엔비디아, #Nvidia, #ARM, #인텔, #Intel, #AMD, #가속기, #accelerator, #GPU, #이종컴퓨팅, #heterogeneouscomputing, #클라우드, #cloud, #엣지컴퓨팅, #edgecomputing, #개방형생태계, #openecosystem