

물리적 AI의 등장과 로봇틱스 분야 동향



CONTENTS

Digitalservice Issue Report

디지털서비스 이슈리포트

01	물리적 AI의 등장과 로보틱스 분야 동향	3
	한국지능정보사회진흥원 임상혁 책임연구원	
02	마이크로소프트 빌드 2025: AI 에이전트의 시대	10
	Senior Program Manager 김영욱	
03	일본의 클라우드 정책 현황	19
	한상기 테크프론티어 대표	
04	2025 구글 I/O와 JAX 업데이트	28
	정채상 인이지 연구소 기술 책임자	
05	대만 엔비디아 컨스텔레이션 설립의 의미와 시사점	34
	윤대균 아주대학교	

본 저작물은 디지털서비스 이용지원시스템이 저작권을 보유하고 있습니다.

디지털서비스 이용지원시스템의 승인 없이 이슈리포트의 내용 일부 또는 전부를 다른 목적으로 이용할 수 없습니다.

01 물리적 AI의 등장과 로봇틱스 분야 동향

| 한국지능정보사회진흥원 임상혁 책임연구원

인식형 AI부터 물리적 AI로의 변화

AI 기술은 지난 수십 년간 단계적으로 진화를 거듭해 왔다. 엔비디아 젠슨 황 CEO는 CES 2025 기조연설에서 "물리적 AI의 시대"가 도래했다고 선언하며, AI가 단순한 인식과 생성을 넘어 물리적 세계와 직접 상호작용하는 새로운 패러다임으로 진입했음을 강조했다. 그는 AI가 '인식형 AI(Perception AI)', '생성형 AI(Generative AI)', '에이전틱 AI(Agentic AI)'를 거쳐, 물리적 세계에서 처리하고, 추론하고, 계획하며 행동하는 '물리적 AI(Physical AI)'로 진화하며, 이것이 AI 혁명의 다음 단계이자 궁극적인 목표라고 설명했다.

- **인식형 AI (Perception AI)** : 이미지 인식, 음성인식, 패턴 분석 등을 통해 데이터를 해석하고 분류하는 기능을 담당한다. 컴퓨터 비전, 자연어 처리, 머신러닝 알고리즘이 핵심 기술이며, 얼굴 인식 시스템이나 음성 어시스턴트 등에 활용된다.
- **생성형 AI(Generative AI)** : GPT, DALL-E 등과 같은 모델을 통해 텍스트, 이미지 등 새로운 콘텐츠를 생성한다. 트랜스포머 아키텍처와 대규모 언어 모델(LLM)이 핵심 기술이며, 창작 도구나 코딩 어시스턴트 등으로 구현된다. 최근 LLM 기술의 급격한 발전으로 대중적으로 친숙한 AI 서비스가 되었다.
- **에이전틱 AI(Agentic AI)** : 자율적 의사 결정과 계획수립이 가능한 AI로, 다중 도구 사용과 복잡한 워크플로우 실행을 특징으로 하며 목표를 설정하고 행동하는 AI를 의미한다. 강화학습, 계획 알고리즘, 멀티에이전트 시스템이 핵심 기술이며, 자동화된 업무 처리 등에 적용된다. 최근 Anthropic의 MCP, Google의 A2A 등이 정교한 에이전틱 AI 기술로 주목받는다.
- **물리적 AI(Physical AI)** : 디지털 세계를 넘어 실제 물리적 환경에서 작동하며, 로봇, 자율주행차 등을 통해 현실 세계와 직접 상호작용하는 특징을 가진다. 최근 발달하는 LLM 및 멀티모달 기술과 결합하여 더욱 빠르게 진화할 것으로 예상되는 분야이다. 센서 융합, 실시간 제어, 환경 적응 학습이 핵심 기술이며, 휴머노이드 로봇, 자율주행차, 스마트 팩토리 등의 형태로 구현된다.

디지털서비스 이슈리포트

| CES 2025에서 젠슨 황 CEO가 발표한 Physical AI로의 진화 과정 |

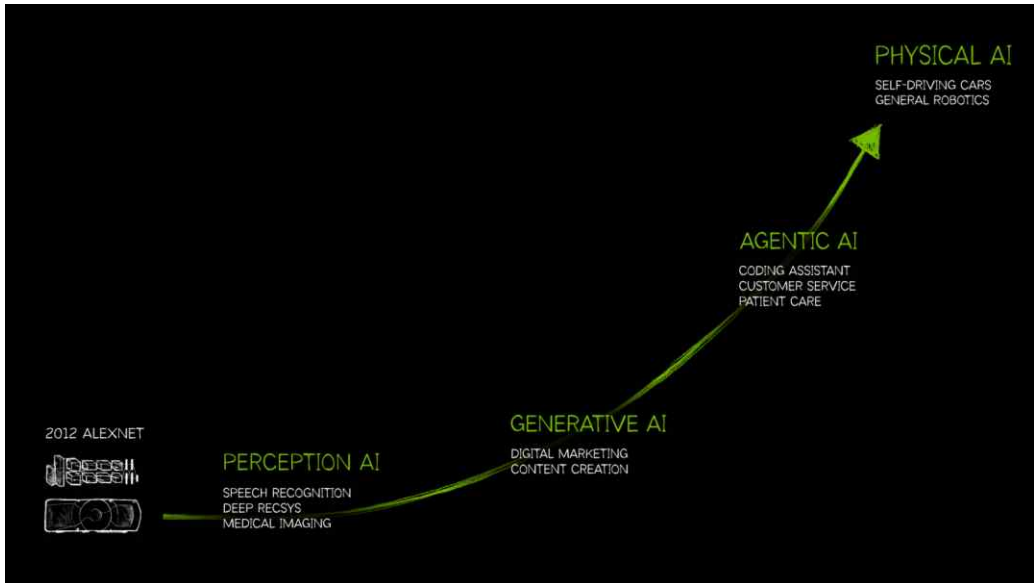


그림 1 출처 : nvidia.com, CES2025 Highlights Deck(2025)

물리적 AI(Physical AI)란?

물리적 AI는 인공지능(AI)의 한 분야로, 디지털 공간에 머무르던 기존 AI와 달리 물리적 세계에서 직접 행동하고 상호작용하도록 설계된 AI 시스템이다. 이는 센서로 실세계 환경을 인식하고, AI 모델로 상황을 판단하며, 액추에이터를 통해 물리적 행동을 수행하는 통합 시스템으로 구현된다. 특히 물리적 AI는 자율 주행 차량, 산업용 로봇, 휴머노이드 등 다양한 하드웨어와 결합하여 현실 세계의 복잡한 문제를 해결하고 광범위한 파급효과를 가져올 핵심 인공지능 기술로 주목받는다.

물리적 AI의 주요 특징은 다음과 같다. 첫째, 다양한 센서를 통해 주변 환경 변화를 감지하고 즉각 대응하는 실시간 환경 인식 및 적응 능력이다. 둘째, 로봇팔, 그리퍼 등을 통해 객체를 조작하고 이동하는 물리적 조작 능력이다. 셋째, 경험을 통해 지속적으로 성능을 개선하며 새로운 상황에 적응하는 학습과 진화 능력이다.

이러한 물리적 AI는 센서, 액추에이터, AI 모델, 그리고 안전한 학습과 검증을 위한 시뮬레이션 환경 등을 핵심 구성 요소(기술)로 한다.

디지털서비스 이슈리포트

- **센서 기술:** AI가 주위 환경을 정확하게 인식하기 위한 핵심으로, LiDAR, 카메라, 레이더, IMU(관성측정장치) 등이 사용된다. 이들 센서는 융합되어 360도 환경 인식, 정밀 거리 측정, 객체 인식을 가능하게 한다.
- **AI 모델:** 환경과의 상호 작용으로 최적 행동 패턴을 학습하는 강화학습이 핵심적이며, 전문가 행동을 모방하는 모방학습(Imitation Learning)도 빠른 학습을 위해 활용된다.
- **시뮬레이션 환경:** Unity, NVIDIA Isaac Sim 등의 플랫폼을 사용하여 실제계와 유사한 환경에서 안전한 학습과 검증을 수행한다. 이는 위험하거나 비용이 많이 드는 실제 테스트를 보완한다.
- **액추에이터 기술:** 물리적 동작을 실행하는 부분으로, 모터 등이 포함되며 정밀한 제어와 빠른 응답 속도가 요구된다. 최근에는 소프트 액추에이터 등 새로운 방식도 개발되고 있다.

이처럼 물리적 AI를 구성하는 AI 모델, 시뮬레이션 환경 등 핵심 기술들은 다양한 로봇틱스 분야에 응용되며 그 가능성을 확장하고 있다. 그중에서도 물리적 AI가 사회 전반에 큰 변화를 불러올 것으로 주목받는 대표적인 적용 영역은 자율 주행 기술과 휴머노이드 기술이다. AI와 결합함으로써, 이 두 분야는 큰 기능적 향상과 함께 가파른 시장 성장이 전망될 뿐 아니라, 선도 국가 및 글로벌 기업들의 전략적 관심 또한 집중되고 있는 핵심 영역이다.

이어지는 내용에서는 이 두 핵심 영역을 중심으로 물리적 AI가 어떻게 각 영역에 융합되어 기술적 발전을 이끌고 있는지, 그리고 관련 선도 기업들의 주요 동향은 무엇인지 구체적으로 살펴보고자 한다.

자율 주행과 AI의 융합

자율 주행 기술은 복잡하고 동적인 환경에서 실시간 의사 결정을 요구하므로, 기존 규칙 기반 시스템으로는 도로 위의 무수한 변수와 예외 상황 처리에 한계가 있어 AI와의 융합이 필수적이다. AI는 다중 센서 데이터 통합 분석을 통한 ▲환경 인식 정확성 향상, 다른 차량·보행자 행동 예측을 통한 ▲예측 능력 향상, 교차로·합류 구간 등에서의 ▲복잡한 상황 의사 결정 능력을 제공한다.

현재 발전 방향은 센서 입력부터 조향, 가속, 제동까지 전체 과정을 하나의 신경망으로 학습하는 End-to-End 학습 방식으로 모이고 있으며, 이는 기존 모듈화 시스템보다 더 자연스럽고 인간과 유사한 주행을 가능하게 한다.

초기 자율 주행 시스템은 미리 정의된 규칙에 따라 동작하는 Rule-Based 접근법을 사용했으나, 예측 불가능한 예외 상황에 취약했다. 대량의 주행 데이터를 통해 스스로 패턴을 학습하고 새로운 상황에

디지털서비스 이슈리포트

적응하는 딥러닝 기반 AI 학습으로의 전환은 자율 주행 기술의 패러다임을 바꾸었다. 특히 컨볼루션 신경망(CNN)을 통한 이미지 인식과 순환 신경망(RNN)을 통한 시계열 예측이 핵심 기술로 자리 잡았다.

LiDAR 센서는 정확한 3D 환경 인식을 제공하지만, 높은 비용과 악천후 성능 저하의 한계로, 카메라 기반 비전 시스템과 AI를 결합한 접근법(예: 테슬라)이 주목받고 있다. 테슬라 FSD는 v12부터 기존 C++ 코드를 제거하고 실제 주행 영상을 학습한 단일 신경망 모델로 전환, '퓨어 비전(Pure Vision)' 전략을 강화했다.

자율 주행 선도 기업의 현황은 다음과 같다.

- **Tesla FSD** : v12는 End-to-End 신경망 기반으로, 비전 트랜스포머(Vision Transformer) 아키텍처를 활용해 다중 카메라 영상을 동시 처리하고, 대규모 실제 주행 데이터 학습 및 OTA(Over-The-Air) 업데이트를 통해 성능을 지속 개선한다.
- **리오토(LiAuto)** : 33개 센서 융합 기술과 NVIDIA Orin 칩셋 기반 고성능 컴퓨팅, 배터리 교환 시스템 연계 자율 충전 기능이 특징이다.
- **화웨이(Huawei)** : 자체 개발 MDC 컴퓨팅 플랫폼과 고정밀 지도 기술을 결합, 5G 통신 기술 연계 V2X(Vehicle-to-Everything) 통신이 강점이다.
- **샤오펑(XPeng)** : 비전 기반 접근법과 자체 개발 X-Net 신경망, 대규모 중국 도로 데이터를 활용하여 LiDAR 없이도 도시 자율 주행을 구현했다.

또한 이 분야에서 특히 물리적 AI를 활용한 합성데이터(실제 데이터가 아닌 알고리즘이나 시뮬레이션을 통해 인위적으로 생성된 데이터)의 적용·활용이 활발하게 이루어지고 있다. NVIDIA Drive Sim은 포토리얼리스틱 렌더링과 정확한 물리 시뮬레이션으로 실제계와 유사한 테스트 환경을 제공하며, COSMOS는 생성형 AI를 활용해 실제 주행 영상 기반으로 새로운 시나리오의 주행 영상을 생성, 희귀하거나 위험한 상황의 안전한 시뮬레이션을 가능케 한다. 즉, 합성데이터를 통해 실제 데이터 수집의 한계(위험 상황 재현의 어려움, 희귀 시나리오 확보의 난점, 높은 비용 문제)를 극복하려 하고 있다.

합성데이터는 자율 주행 AI 모델 학습 및 검증에서 ▲희귀 상황(사고, 극한 날씨 등) 학습 ▲안전한 테스트 환경 제공 ▲비용 효율성 증대 등의 역할을 한다. 검증 과정에서는 심-투-리얼(Sim-to-Real) 전이 학습을 통해 시뮬레이션 학습 모델이 실세계에서도 잘 동작하도록 한다.

디지털서비스 이슈리포트

| AI가 실제 세계의 물리 법칙을 이해하고 예측할 수 있도록 돕는 월드 파운데이션 모델 NVIDIA Cosmos(우측) |

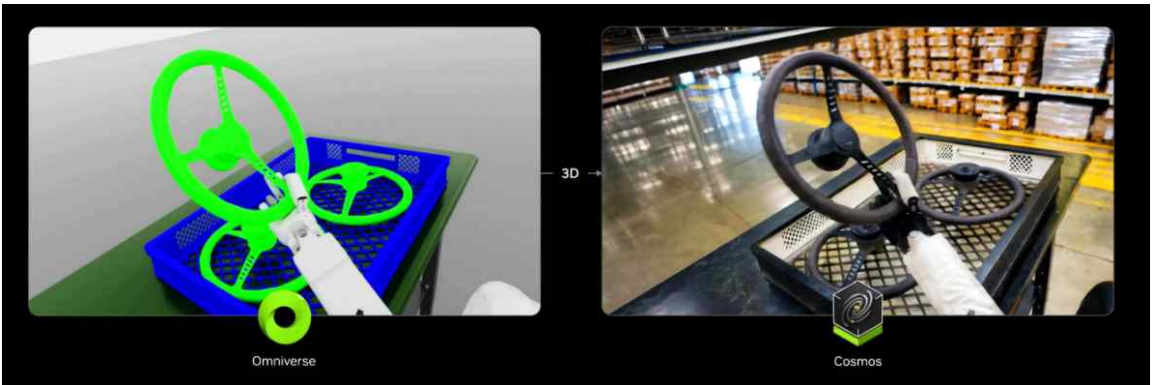


그림 2 출처 : nvidia.com

휴머노이드 로봇과 AI의 융합

휴머노이드 로봇은 인간과 유사한 형태와 기능을 가진 로봇으로, 이족 보행, 물체 조작, 언어 소통 등 인간의 능력을 모방한다. 핵심 기능으로는 환경 인식 및 내비게이션, 이족 보행 및 균형 제어, 양손을 이용한 정교한 물체 조작, 인간과 자연스러운 상호 작용이 있다. 휴머노이드 로봇이 지닌 핵심적 이점은 다른 산업용 로봇 등과 다르게 ‘인간을 닮았다’라는 직관적인 사실에 있다. 이는 인간의 손길이 필요했던 어떤 환경, 어떤 작업에서도 로봇이 투입되어 인간의 업무를 수행하거나 지원할 수 있음을 의미한다. 결과적으로 인간이 수행해 온 업무 영역에서 로봇이 그 역할을 대체하거나 인간과 협력할 수 있다는 유연성과 범용성을 부여한다. 가장 빠르게 적용될 수 있는 분야로는 다음을 예상할 수 있다.

- **산업 현장(제조, 물류)** : 제조업에서 기존 산업용 로봇이 수행하기 어려운 복합적이고 유연한 작업(조립, 검사, 포장 등)을 담당하며, BMW, 현대 등 자동차 제조사들이 검증 중이다. 물류 분야에서는 창고 내 피킹, 패킹, 분류 작업에 활용 기대된다.
- **서비스, 가사, 재난 대응 분야** : 서비스 분야(호텔, 병원 등 고객 응대), 가사 도우미(고령화 사회 핵심 솔루션), 재난 대응(위험 지역 수색, 구조) 등에서 활용될 전망이다.

물리적 AI의 태동은 휴머노이드의 지능을 혁신적으로 향상시키고 있다. ▲실시간 환경 적응 능력 향상으로 예상치 못한 상황 변화에 즉각 대응하고, ▲학습 능력 고도화로 새로운 작업을 빠르게 습득하며 기존 경험을 새로운 상황에 적용하고, ▲인간과의 협업 능력 강화로 복잡한 작업을 함께 수행할 수 있게 된다. 휴머노이드의 핵심 기술 요소는 다음이 거론된다.

디지털서비스 이슈리포트

- **이족 보행 및 균형 제어** : ZMP(Zero Moment Point) 제어, 관성 측정, 실시간 궤적 계획이 필요하며, 최근 머신러닝을 활용한 적응적 보행 제어로 다양한 지형에서의 안정성이 향상되었다. 모델 예측 제어(MPC), 강화학습 기반 제어 등도 활용된다.
- **물체 조작(Manipulation)** : 정밀한 그리п 제어, 힘 피드백 시스템, 객체 인식 및 추적 기술이 필요하다. 최근 딥러닝(이미테이션 러닝, 강화학습)과 촉각 센서 발전으로 다양한 물체 조작 능력이 향상되고 있다.
- **환경 인식 및 상호 작용(HRI) 기술**: SLAM(Simultaneous Localization and Mapping) 기술로 실시간 지도 생성 및 위치 파악을 하며, 다중 센서 융합으로 정확한 환경 인식을 수행한다. 자연어 처리, 표정·제스처 인식을 통한 인간-로봇 상호 작용(HRI)과 안전한 물리적 상호 작용 기술이 중요하다.
- **LLM 및 VLM (Vision-Language Model) 통합을 통한 고도화** : GPT-4o, Claude-3 등 멀티모달 AI 모델 통합은 휴머노이드 로봇의 지능을 획기적으로 향상시켜, 시각 정보와 언어적 지시를 동시에 이해하고 처리하게 한다. 이를 통해 자연어 명령 이해, 상황 이해, 학습 효율성이 증대된다.

휴머노이드 분야의 주요 선도 기업의 개발 현황 및 동향은 다음과 같다. (개발사/로봇명)

- **Tesla (Optimus)** : 범용 휴머노이드로 설계, 테슬라 FSD 칩 기반 AI 컴퓨팅과 360도 시각 인식이 특징이며, 자동차 제조 경험을 바탕으로 대량 생산 가능성을 목표로 한다.
- **Figure AI (Figure-03)** : 상업 배포 목표로 OpenAI와 협력하여 고도화된 언어 이해 능력과 정밀한 손가락 제어 능력을 갖추었으며, BMW 자동차 제조 설비에서 테스트 중이다.
- **Apptronik (Apollo)** : 물류와 제조업에 특화된 휴머노이드로, 모듈형 설계와 교체할 수 있는 배터리 시스템으로 연속 작업이 가능하며 안전 중심 설계가 특징이다.
- **Boston Dynamics (Atlas)** : 기술적 완성도가 높은 휴머노이드로, 뛰어난 이족 보행 능력과 동적 균형 제어, 복잡 지형 이동 능력을 보여주며 최근 전기 구동 방식으로 업그레이드되었다.
- **Unitree Robotics (H1/G1)** : 중국의 대표적인 로봇 기업으로, 뛰어난 동적 제어 기술과 높은 가성비로 바탕으로 휴머노이드 로봇 H1 및 G1을 발표했다. 특히 G1 모델은 비교적 저렴한 가격대로(약 16,000\$) 산업 및 서비스 분야에서의 빠른 상용화를 목표로 하고 있으며, 민첩한 움직임과 AI 기반 상호 작용 능력을 강조한다.

디지털서비스 이슈리포트

| 주요 휴머노이드 분야 기업이 발표한 최신 모델들 |

			
Tesla(Optimus)	Figure AI(Figure-01)	Boston Dynamics(Atlas)	Unitree Robotics(H1)

그림 3 출처 : 각 기업 자료

마무리

본 보고서에서 살펴본 바와 같이, 물리적 AI는 인식, 생성, 에이전트 단계를 넘어 현실 세계와 직접 상호작용하는 AI 기술의 새로운 지평을 열고 있다. 특히 자율 주행과 휴머노이드 로봇 분야에서 AI와의 융합을 통해 혁신적인 변화를 주도하고 있다.

자율 주행 분야에서는 End-to-End 학습 방식, 신경망을 활용한 카메라 중심의 비전 시스템, 그리고 합성데이터 활용이 기술 발전을 견인하며 주요 기업들의 경쟁을 치열하게 만들고 있다. 휴머노이드 로봇 또한 인간의 작업 환경에 투입될 수 있는 유연성과 범용성을 바탕으로, LLM 및 멀티모달 AI 기술과 결합하여 지능과 활용성이 빠르게 고도화되고 있으며, 테슬라, Figure AI 등 다양한 기업들이 상용화를 목표로 개발에 매진 중이다.

이러한 물리적 AI와 로봇틱스의 발전은 산업 현장의 자동화 수준을 높이고 새로운 서비스를 창출하는 등 긍정적 가능성을 제시하는 한편, 기술의 안전성, 사회적 수용성 및 윤리적 고려 사항 등 다각적인 검토가 필요한 과제도 함께 안고 있다. 앞으로 이 기술 분야의 지속적인 연구개발과 함께 사회적 논의가 병행될 것으로 전망된다.

02 마이크로소프트 빌드 2025: AI 에이전트의 시대

| Senior Program Manager 김영욱

빌드(Build)는 2025는 마이크로소프트가 매년 주최하는 개발자 중심의 행사로 올해는 5월 19일부터 5월 22일까지 미국 시애틀에서 개최된 연례 개발자 콘퍼런스이다. 이 행사에서 마이크로소프트는 AI 기술 발전이 가져올 미래 비전을 제시했으며, 특히 'AI 에이전트 시대'의 도래와 '개방형 에이전트 웹(open agentic web)' 구축이라는 핵심 메시지를 통해 새로운 플랫폼 전환의 시대를 선언했다.

사티아 나델라 CEO는 기조연설에서 현재 우리가 경험하고 있는 AI 기술 발전이 과거 윈도우즈, 웹 스택, 그리고 클라우드 및 모바일 시대에 비견되는 플랫폼 전환의 '중반전'에 진입했음을 강조하며, 추론 및 메모리 분야의 획기적인 발전 덕분에 AI 모델이 더욱 강력하고 효율적이 되었으며, 이 시기는 모든 것이 확장되고 개발자를 위한 기회가 모든 계층에서 폭넓게 확대되는 중요한 단계라고 언급했다.

나델라는 이러한 전환의 핵심 비전으로 '개방형 에이전트 웹'을 구축하는 것을 제시했다. 이는 소수의 수직 통합형 앱에서 벗어나 AI 에이전트들이 서로 협력하고 데이터를 공유하며 사용자를 대신하여 결정을 내리고 작업을 수행하는 개방적이고 확장 가능한 플랫폼을 의미한다.

이번 빌드 2025에서는 바로 이러한 'AI 에이전트 시대'를 가속화하고 '개방형 에이전트 웹'을 현실로 만들기 위한 마이크로소프트의 구체적인 전략과 이를 지원하는 강력한 플랫폼 도구들을 공개했다. 특히 다음 5가지 주요 주제는 AI 기술이 개발자의 역할과 소프트웨어 개발 라이프사이클, 그리고 엔터프라이즈 및 개인의 생산성에 어떤 근본적인 변화를 가져올 것인지를 설명한다.

이번 빌드 2025의 5가지 핵심 메시지를 상세히 살펴보고자 한다.

1. AI 에이전트 시대의 도래와 개방형 에이전트 웹 비전

AI 에이전트란 사용자나 조직을 대신하여 자율적으로 결정을 내리고 작업을 수행할 수 있는 AI 시스템을 의미한다. 이러한 에이전트의 등장은 추론 및 메모리 기술의 비약적인 발전 덕분에 가능해졌다고 할 수 있다. 마이크로소프트는 빌드 2025를 통해 우리가 AI 에이전트의 시대에 진입했음을 공식적으로 선언했다.

이러한 AI 에이전트의 등장은 인터넷의 미래 형태에 대한 비전으로 이어지는데 그것이 바로 '개방형 에이전트

디지털서비스 이슈리포트

웹'이다. 이는 현재와 같이 몇몇 거대 플랫폼의 수직 통합된 앱 중심 구조에서 벗어나, 다양한 에이전트들이 서로 유기적으로 연결되고 협력하며 데이터를 공유하는 개방적이고 분산된 웹 생태계를 목표로 한다.

의미와 중요성 : 이러한 변화는 역사적인 플랫폼 전환기의 '중반'에 비유된다. 이는 AI가 단순히 기존 도구를 개선하는 수준을 넘어, 컴퓨팅 및 소프트웨어 개발 방식 자체를 재정의하는 근본적인 변화이다. 개방형 에이전트 웹은 개발자들이 특정 플랫폼에 종속되지 않고 다양한 에이전트를 개발 및 배포하여 새로운 가치를 창출할 수 있는 거대한 기회를 열어준다. 에이전트가 사용자나 조직을 대신하여 복잡한 작업을 처리함으로써 생산성과 효율성을 극대화할 수 있고 이러한 개방적인 접근 방식은 에이전트 웹의 유용성을 높이는 데 필수적이며, 엔지니어 커뮤니티의 집단적인 노력으로 더욱 풍부하고 역동적인 생태계를 만들어 낼 수 있다.

변화 : 이러한 비전이 현실화하면, 웹사이트와 애플리케이션은 단순히 정보를 제공하거나 특정 기능을 수행하는 것을 넘어, 자연어 인터페이스(Natural Language Web과 같은 기술 활용)를 통해 사용자와 상호 작용하고 다른 에이전트와 협력하여 작업을 완료하는 '에이전트 애플리케이션'으로 진화하게 된다. 검색이나 정보 탐색 방식이 근본적으로 변화하고, 개인화된 AI 비서들이 복잡한 업무를 처리하는 것이 일상화될 것이다. 개발자는 특정 앱을 만드는 것과 더불어, 특정 도메인 지식이나 기능을 가진 에이전트를 설계하고, 이들이 상호 작용할 수 있도록 오케스트레이션 하는 능력이 중요해지고 있다. 또한, 에이전트 간의 안전하고 표준화된 통신을 위한 프로토콜(예: MCP)의 중요성이 커질 것이다.

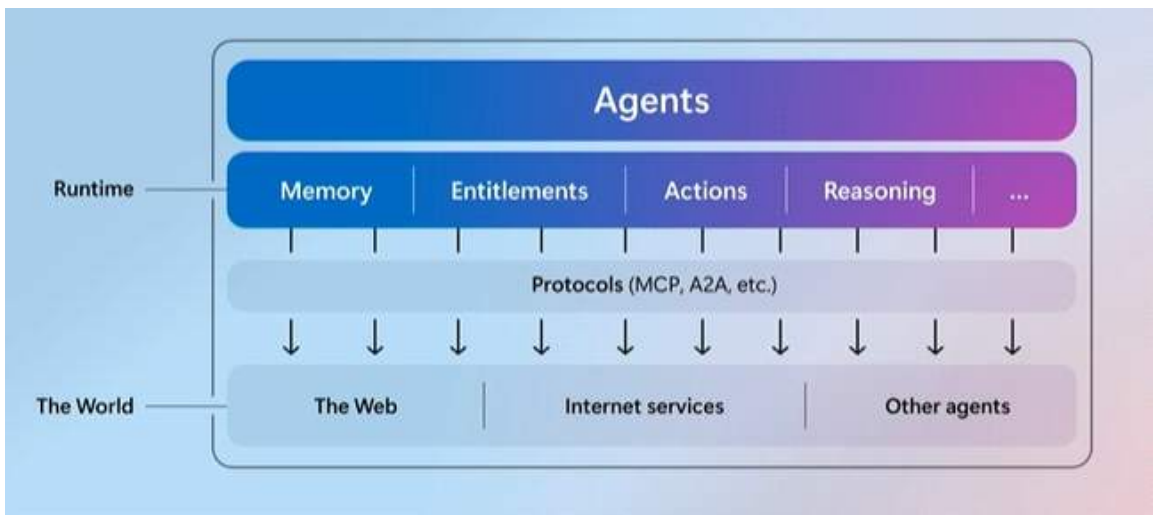


그림 4 개방형 에이전트 웹의 개념을 보여주는 아키텍처 다이어그램 (출처: 마이크로소프트)

디지털서비스 이슈리포트

2. AI 에이전트 역량 및 오케스트레이션의 발전

AI 에이전트 역량 발전에는 단일 에이전트의 성능 향상뿐만 아니라, 여러 에이전트가 협력하여 복잡한 작업을 처리하는 '멀티에이전트 오케스트레이션' 및 특정 회사 데이터, 워크플로우 및 스타일에 맞춰 모델을 조정하는 '코파일럿 튜닝(Tuning)' 기능의 도입이 포함된다. 또한 SRE(Site Reliability Engineering)를 위한 자율 에이전트와 같이 특정 전문 분야의 자율 에이전트도 발표되었다.

의미와 중요성 : 멀티에이전트 시스템은 에이전트들이 전문 기술을 결합하여 작업을 분산하고 보다 포괄적인 응답 및 경험을 제공함으로써 복잡한 시나리오를 처리할 수 있도록 한다. 예를 들어, 스탠포드 헬스케어는 의료 에이전트 오케스트레이터를 사용하여 종양 위원회 준비를 돕는 솔루션을 개발했으며, 이는 환자 기록, 방사선 이미지, PubMed 데이터 등 다양한 출처의 정보를 통합한다.¹⁾ 코파일럿 튜닝은 조직이 자체 데이터를 사용하여 AI 모델을 조정하여 회사 특유의 톤, 언어, 전문 지식 및 지식을 반영할 수 있게 한다. 이는 특정 도메인에 대해 매우 정확하고 관련성 높은 에이전트를 구축하는 데 필수적이며, 법률 회사나 컨설팅 회사와 같이 특정 전문성을 가진 조직에 특히 유용하다. 자율 에이전트는 개발자가 실시간 사이트 문제와 같은 반복적이고 시간이 많이 소요되는 작업에서 벗어나 더 전략적인 작업에 집중할 수 있도록 돕는다.

변화 : 코파일럿 스튜디오에서 멀티에이전트 오케스트레이션 기능이 도입되어 여러 에이전트가 팀으로 함께 작업할 수 있게 되었다. 코파일럿 튜닝은 조직이 로우 코드 방식으로 자체 데이터와 워크플로우를 사용하여 모델을 조정할 수 있게 되었다. 애저 AI 파운드리 에이전트 서비스는 시맨틱 커널 및 오토젠(AutoGen)을 단일 SDK로 통합하고 A2A(Agent-to-Agent) 및 MCP(Model Context Protocol)를 지원하여 전문 개발자가 복잡한 워크플로우를 여러 에이전트를 오케스트레이션할 수 있도록 한다. 또한 SRE를 위한 자율 에이전트와 같은 새로운 유형의 에이전트가 소개되었다. M365 코파일럿 앱은 채팅, 검색, 노트북, 생성 및 에이전트를 하나의 직관적인 스캐폴딩으로 통합하여 AI의 UI 역할을 한다. 이런 역할과 전문성을 나누어 프로세스화 하면, 연구자 에이전트와 분석가 에이전트와 같은 새로운 사전 구축된 에이전트를 통해 웹과 엔터프라이즈 소스를 합성하는 역할과, 원시 데이터를 분석하여 통찰력을 제공하는 역할로 나누어진 에이전트를 통해 비약적인 생산성 향상을 기대해 볼 수 있다.

1) 마이크로소프트, "Meet 4 developers leading the way with AI agents", May 19, 2025

디지털서비스 이슈리포트

Microsoft 365 Copilot Tuning



그림 5 M365 코파일럿 튜닝 워크프로세스 (출처: 마이크로소프트)

3. 애저 AI 파운드리를 에이전트 개발 플랫폼으로 활용

애저 AI 파운드리는 개발자가 AI 애플리케이션 및 에이전트를 설계, 사용자 정의 및 관리하기 위한 통합 플랫폼이다. 이는 모델 선택, 안전한 데이터 통합, 모델 사용자 정의, 엔터프라이즈급 거버넌스 기능을 제공한다. 파운드리 에이전트 서비스는 개발자가 코드를 적게 사용하여 선언적 에이전트를 구축할 수 있도록 하며 복잡한 워크플로우에 대한 멀티에이전트 오케스트레이션을 지원한다.

의미와 중요성 : 파운드리는 기업이 AI의 ROI를 진정으로 실현하기 위해 POC에서 전사적 배포로 이동하는 것을 지원하는 핵심 도구이다. 1,900개 이상의 모델을 지원하며, Grok 3 및 Grok 3 mini를 포함한 오픈AI 및 xAI의 최신 모델에 당일 액세스를 제공한다. 파운드리는 모델 선택을 자동화하는 모델 라우터와 같은 도구를 제공하여 개발자의 작업을 단순화한다. 또한 파운드리는 Entra ID (에이전트 식별 및 관리), Purview (데이터 보안 및 거버넌스), Defender (보안) 와 통합되어 엔터프라이즈급 보안, 규정 준수 및 거버넌스를 제공한다. AI 에이전트의 성능, 품질, 비용 및 안전성을 모니터링하기 위한 관찰 가능성(Observability) 기능도 눈에 띄는 기능이다.

변화 : 전문 개발자들이 에이전트를 오케스트레이션 할 수 있는 새로운 기능 애저 AI 파운드리 에이전트 서비스와 애저 AI 파운드리 모델 포트폴리오에 오픈AI 및 xAI의 최신 모델이 추가되었다. 모델 리더보드 및 모델 라우터와 같은 모델 평가 및 선택 도구가 도입되어 개발자의 작업이 간단해지고,

디지털서비스 이슈리포트

선택 실수가 줄어든다. Entra ID와의 통합을 통해 AI 파운드리 및 코파일럿 스튜디오에서 생성된 에이전트에게 고유한 ID가 자동으로 할당된다. 파운드리에서 모델을 미세 조정하거나 사전 학습시킨 후 코파일럿 스튜디오에서 가져와 워크플로우를 자동화하거나 에이전트를 구축할 수 있게 되었다. 또한 파운드리 로컬이란 기능을 통해 윈도우 및 맥 클라이언트에서도 파운드리 기능을 사용할 수 있게 되었다. 다양한 모델과 에이전트를 이용한 협업을 통해 생산성 향상을 자연스럽게 기대해 볼 수 있다.

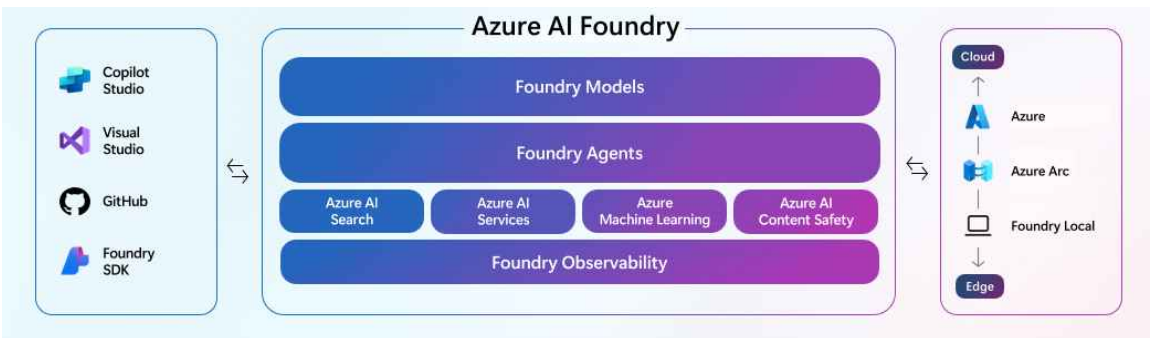


그림 6 애저 AI 파운드리 (출처: 마이크로소프트)

4. 윈도우에 AI 통합 (윈도우 AI 파운드리, MCP, NLWeb)

마이크로소프트는 윈도우 OS를 AI 개발과 실행을 위한 최고의 플랫폼으로 진화시키고 있다. '윈도우 AI 파운드리'는 개발자가 클라이언트 및 클라우드 전반의 AI 개발 수명 주기를 지원하는 통합 플랫폼이다. MCP는 AI 애플리케이션과 에이전트가 다른 앱, 웹 서비스 또는 윈도우의 일부와 통신할 수 있도록 하는 오픈 소스다. NLWeb은 웹사이트나 API를 에이전트가 자연어로 상호 작용할 수 있는 '에이전틱 애플리케이션'으로 쉽게 전환할 수 있도록 한다.

의미와 중요성 : 윈도우는 10억 명 이상의 사용자와 수많은 디바이스가 연결된 널리 사용되는 플랫폼 중 하나이다. 윈도우에 AI 기능을 통합함으로써 개발자는 이러한 방대한 사용자 기반에 도달할 수 있는 새로운 AI 기반 애플리케이션을 구축할 수 있게 된다는 것이 가장 큰 의미가 된다. 윈도우 AI 파운드리는 개발자가 CPU, GPU, NPU 등 다양한 실리콘에서 실행되는 AI 애플리케이션을 구축할 수 있도록 지원하며, 로컬에서 실행할 수 있는 사전 최적화된 오픈 소스 모델 카탈로그에 접근할 수 있도록 한다. MCP는 AI 앱 간의 상호 운용성을 가능하게 하며, 에이전트가 윈도우의 기본 기능에 안전하게 액세스하고 제어할 수 있도록 한다. NLWeb은 기존 웹사이트에 에이전트와의 자연어 인터페이스를 추가하여 웹을 에이전틱 경험의 일부로 통합하는 데 중요한 역할을 한다.

디지털서비스 이슈리포트

변화 : 개발자가 AI 개발 수명 주기를 지원하는 플랫폼을 사용할 수 있게 되었다. 윈도우에서 MCP에 대한 네이티브 지원이 발표되었으며, 파일 시스템, 설정, 앱 작업 같은 여러 내장 MCP 서버가 포함된다. 또한 안전하고 성능이 검증된 MCP 서버를 검색하기 위한 네이티브 MCP 레지스트리가 추가되었다. NLWeb으로 웹사이트나 API를 에이전틱 애플리케이션으로 쉽게 만들 수 있게 되었다. 파운드리 로컬은 윈도우 AI 파운드리에 내장되어 로컬 AI 앱 개발을 지원한다. 윈도우 11에서 Phone Link 패널 출시, 탐색기에 AI 액션 통합, Edge 브라우저에서 웹 앱을 위한 온-디바이스 AI 모델(Phi-4-mini) 공개 등 윈도우 사용자에게도 AI 통합이 확대되고 있다.

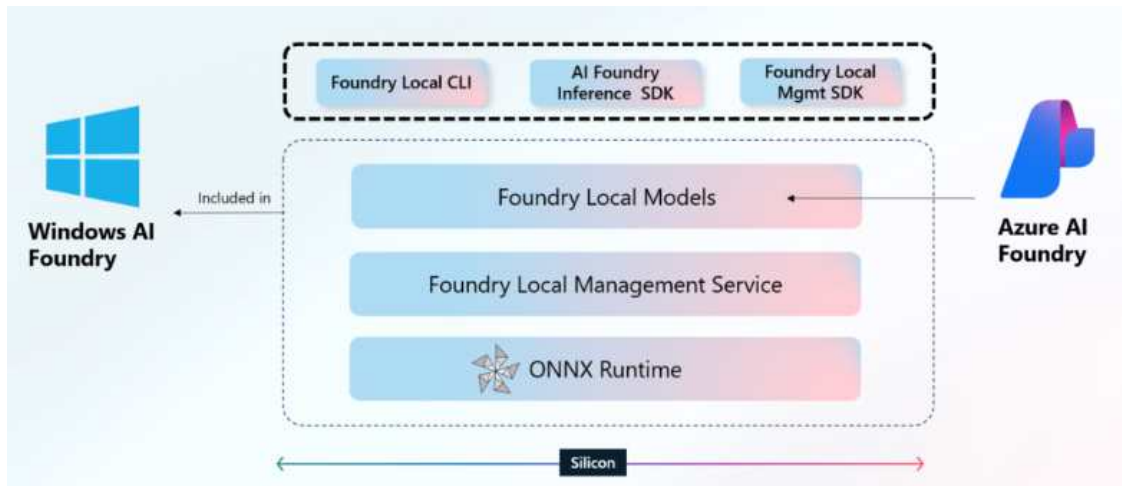


그림 7 파운드리 로컬 동작 구조 (출처: 마이크로소프트)

5. 개발 도구에 AI 통합 (GitHub Copilot, VS Code, WSL)

마이크로소프트는 개발자의 생산성을 높이기 위해 핵심 개발 도구에 AI를 깊숙이 통합하고 있다. 깃허브 코파일럿은 코드 완성 도구에서 코드 검토 및 자율적 작업 수행이 가능한 '코딩 에이전트'로 발전하고 있다. VS Code는 AI 기반 기능을 통합했으며, 특히 깃허브 코파일럿 챗 기능을 오픈 소스화 했다. 이제 WSL(Windows Subsystem for Linux)를 오픈 소스로 제공한다.

의미와 중요성 : 개발 도구에 AI를 통합함으로써 개발자는 더 오랫동안 개발 환경에 몰입하고 더 전략적인 작업에 집중할 수 있게 된다. 깃허브 코파일럿과 같은 코딩 에이전트는 버그 수정, 새 기능 추가, 코드 유지보수와 같은 작업을 자율적으로 완료하여 개발 워크플로우를 혁신한다. 또한 오래된 프레임워크 업그레이드나 온프레미스 앱 클라우드 마이그레이션과 같은 앱 현대화 작업을 돕는다. VS Code에 AI 기능을 직접 통합하고 코파일럿 챗을 오픈소스화하는 것은 개발자 커뮤니티와의 개방성과

디지털서비스 이슈리포트

협업에 대한 비전이다. WSL의 오픈 소스화는 개발자 커뮤니티의 오랜 요청에 부응하는 것으로, 커뮤니티 기여를 통해 WSL의 발전과 혁신을 가속할 수 있다. 이러한 도구들은 개발자들이 AI 에이전트 시대와 에이전틱 웹을 구축하는 데 필수적인 역할을 할 것으로 전망한다.

변화 : 깃허브 코파일럿은 페어 프로그래머에서 피어 프로그래머로 진화하고 있으며, 프레임워크 업그레이드, 앱 마이그레이션과 같은 자율적인 작업을 수행하고 SRE를 위한 자율 에이전트 역할을 할 수 있다. SRE 에이전트는 자동으로 문제를 분류, 근본 원인 분석 및 완화하고, 깃허브 이슈로 인시던트를 기록하며, 코파일럿에게 복구 항목을 할당할 수 있다. 이것은 전체 개발 프로세스의 처음부터 끝까지 AI와의 협업을 가능하게 하겠다는 변화를 시사한다.

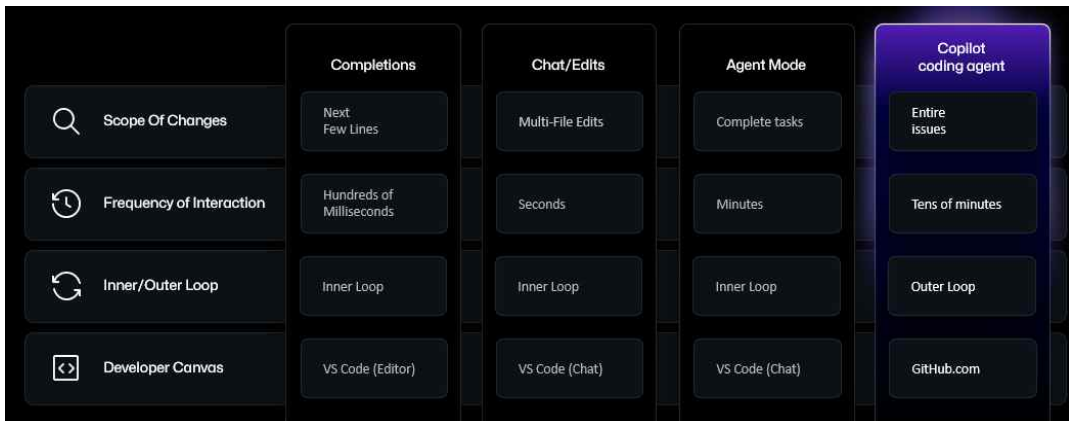


그림 8 코파일럿 코딩에이전트 (출처: 마이크로소프트)

마무리

이번 빌드에서 가장 중요한 메시지는 AI가 단순한 도우미 역할을 넘어 개발자와 사용자의 역량을 크게 확장하는 '에이전트'로 진화하고 있다는 점이다. 이는 소프트웨어 개발 프로세스 자체를 변화시키고 있으며, 개인이든 조직이든 문제를 해결하고 작업을 수행하는 방식을 근본적으로 바꾸고 있다.

마이크로소프트는 이러한 AI 에이전트 시대를 위한 포괄적인 플랫폼과 도구 스택을 제공함으로써 개발자들은 이전보다 훨씬 쉽고 효율적으로 AI 기능을 애플리케이션에 통합하거나 새로운 에이전트를 구축할 수 있게 된다.

개방성과 상호 운용성에 대한 강조는 주목할 만하다. 다양한 모델을 지원하고, 에이전트 간 통신을 위한 개방형 프로토콜(MCP, NLWeb)을 도입하며, 핵심 개발 도구(WSL, Copilot Chat in VS Code)를

디지털서비스 이슈리포트

오픈소스화하는 것은 AI 생태계의 성장을 가속화하고 혁신을 촉진할 잠재력이 매우 크다. '개방형 에이전트 웹'은 단순히 마이크로소프트만의 비전이 아니라, 수백만 명의 개발자가 참여하여 만들어갈 미래가 될 것으로 전망한다. 한편, 엔터프라이즈 환경에서의 AI 도입을 위해 보안, 거버넌스, 관리 기능이 기본적으로 제공된다는 점도 중요합니다. 이는 기업들이 AI 에이전트를 신뢰하고 실제 업무에 적용하는 데 필수적인 요소이다. 또한, 하드웨어부터 소프트웨어까지 풀 스택 최적화를 통해 성능과 비용 효율성을 극대화하려는 노력은 AI 워크로드의 확산에 기여할 것으로 기대한다.

이러한 급격한 변화 속에서 뒤처지지 않고 기회를 포착하기 위해 우리는 다음과 같은 다섯 가지 적극적인 태도가 필요하다고 생각한다.

첫째, 변화를 적극적으로 수용해야 한다. AI 에이전트와 새로운 플랫폼은 우리의 업무 방식과 개발 프로세스를 변화시킬 것이다. 새로운 도구와 워크플로우를 탐색하고 익히는 데 주저하지 않아야 한다. 초기에 약간의 회의론이 있을 수 있지만, 그 가능성을 직접 경험해보는 것이 무엇보다 중요하다.

둘째, 더 높은 목표를 설정해야 한다. 모델의 성능은 빠르게 향상되고 있으며, 현재는 어렵다고 생각되는 과제도 AI 에이전트를 통해 해결 가능해질 수 있다. 에이전트의 '추론 능력'의 발전은 이전에는 상상하기 어려웠던 복잡한 문제 해결을 가능하게 할 것이다.

셋째, 조합과 오케스트레이션에 집중해야 한다. AI 에이전트 시대는 단 하나의 강력한 도구나 에이전트가 모든 것을 해결하는 것이 아니라, 여러 전문 에이전트와 시스템이 협력하여(멀티에이전트 오케스트레이션) 더 크고 복잡한 가치를 창출하는 시대이다. 애플리케이션 아키텍처를 설계할 때 이러한 '조합 가능한 지능(distributed intelligence)'을 어떻게 활용할 것인지 고민해야 할 필요가 있다.

넷째, 개방형 플랫폼과 프로토콜을 적극 활용해야 한다. 마이크로소프트가 제시하는 NLWeb, MCP와 같은 개방형 표준과 오픈 소스 프로젝트는 개발자 생태계 전체에 기회를 제공한다. 이러한 개방성은 에이전트 웹의 성장을 가속화하고 더 많은 혁신을 가능하게 할 것이다.

다섯째, 보안과 거버넌스를 최우선으로 고려해야 한다. AI 에이전트는 강력한 기능을 제공하는 만큼 잠재적인 위험도 수반한다. 마이크로소프트가 제공하는 Entra Agent ID, Purview, Defender와 같은 도구를 활용하여 처음부터 에이전트의 보안과 규정 준수를 철저히 관리해야 한다.

마이크로소프트 빌드 2025는 AI 에이전트와 개방형 에이전트 웹이라는 명확한 방향성을 제시하며, 개발자들이 이 새로운 시대의 주역이 될 수 있도록 전방위적인 플랫폼과 도구를 공개했다. AI가

디지털서비스 이슈리포트

제공하는 무한한 가능성을 현실로 만들기 위한 기술적 기반과 생태계 조성을 향한 마이크로소프트의 노력은 매우 인상적이었다. 이제 이 강력한 도구를 활용하여 새로운 형태의 애플리케이션과 에이전트를 만들고, 우리 사회와 산업 전반에 걸쳐 의미 있는 변화를 일으킬 책임과 기회는 바로 우리에게 있다. 이러한 변화를 두려워하기보다는 흥미로운 도전으로 받아들이고, 적극적인 학습과 참여를 통해 AI 에이전트 시대의 가능성을 최대한 활용해야 할 때이다.

03 일본의 클라우드 정책 현황

| 한상기 테크프론티어 대표

들어가며

일본이 디지털 전환에서 뒤처졌다는 반성과 함께 AI 혁명의 시대를 맞아 새롭게 전열을 정비하고 정부 정책 개발과 민간 협업의 움직임이 빠르게 변화하고 있다. 그동안 일본의 디지털 정책은 우리에게 모범 사례가 아니라는 판단에 도외시해 왔으나 이제 전문가들 사이에서는 일본 정책 변화를 주의 깊게 살펴볼 필요가 있다는 진단이 나오기 시작한다.

일본이 디지털 영역에서 새로운 정책 방향을 제시한 것은 2016년의 ‘소사이어티 5.0’부터라고 볼 수 있다. 소사이어티 5.0 비전은 사이버 공간과 물리적 공간을 고도로 통합하여 경제 발전을 촉진하고 사회 문제를 해결할 수 있는 인간 중심 사회를 구현한다는 것이다.

그러나 챗지피티 이후 벌어진 혁명적 변화에서 일본이 더 이상 지금의 속도로 AI와 클라우드 정책을 추진해서는 안 된다는 생각을 바탕으로 더 적극적이고 민첩하게 움직이기 시작했다. 이런 움직임의 중심에는 기존 경제산업성(METI)와 총무성(MIC) 만으로는 안 된다는 생각으로 만든 디지털 장관 임명과 디지털청(Digital Agency)의 설립이다.

2025년 2월에 일본 정부의 내각부 산하 전문가 위원회인 AI 정책 연구회가 중간 보고서를 발표했으며, 이 보고서는 새로운 접근 방식을 제시했다.²⁾ 이전 정책 보고서들은 고급 AI 모델에 대한 더욱 엄격한 감독을 지향했지만, 이번 중간 보고서는 더욱 신중한 태도를 보이고 있다. 핵심 제안 중 하나는 정부 주도의 전략적 리더십 기구를 설립하는 것이었다. 이 기구는 AI 관련 위험 및 사고에 대한 정보 수집 및 조율 활동을 위한 것이 아니라, 집행을 위한 것이었다. 정부는 기업을 지도하고 책임 있는 AI 도입에 대한 장벽을 낮출 것이라고 한다.

따라서 본 글에서는 일본 클라우드와 나아가 AI 정책 변화의 주요 흐름이 어떻게 나타나고 있는지 여러 가지 자료를 기반으로 정리해 보고자 한다.

2) CSIS, “Unpacking Japan’s AI Policy with Horoki Habuka,” Mar 21, 2025

디지털서비스 이슈리포트

디지털 전환을 위한 부처 설립과 주요 서비스

일본은 디지털청 설치법을 2021년 9월 1일에 통과시켰으며 이에 따라 디지털 트랜스포메이션 부와 디지털청을 설립했다. 디지털 트랜스포메이션을 다루는 디지털부 장관은 일본의 디지털 경제로 전환하는 것을 총괄한다. 현재 2024년 10월 1일부터 마사키 타이아가 장관직을 맡고 있다.

같은 시점에 이의 실행을 담당할 디지털청을 2021년 9월 1일에 설립했다. 디지털청은 일본 사회 전반의 디지털 전환을 총괄하고 가속하기 위해 설립된 내각 직속 기관으로 전체 인력 1,180명 가운데 절반 이상인 600명을 민간 전문가로 채웠다. 주요 업무는 다음과 같은 것이 있다.

- 정부 시스템 표준화 및 공통 기반 구축: 부처별로 개별적으로 운영되던 정부 시스템을 표준화하고, 클라우드 기반의 공통 플랫폼을 구축하여 행정의 효율성을 높이고 국민 편의를 증진한다.
- 마이넘버(My Number) 카드 보급 및 활용 촉진: 개인 식별 번호 제도인 '마이넘버' 카드의 보급을 확대하고, 이를 활용한 다양한 행정 서비스 혁신을 추진하여 국민이 보다 쉽고 편리하게 행정 서비스를 이용할 수 있도록 한다.
- 디지털 인재 확보 및 육성: 정부 및 공공 부문의 디지털 전환을 이끌어갈 전문 인력을 확보하고 양성하여, 디지털 역량을 강화한다.
- 민간과의 협력 강화: 민간 기업 및 전문가들과의 협력을 통해 선진 디지털 기술의 도입과 혁신을 가속화하고, 공공 서비스에 민간의 노하우를 접목한다.
- 디지털 격차 해소: 고령층이나 디지털 소외 계층이 디지털 혜택에서 소외되지 않도록 접근성 향상, 정보 활용 교육 등을 통해 디지털 격차를 줄이고 모든 국민이 디지털 사회의 혜택을 누릴 수 있도록 지원한다.
- 개인 정보 보호 및 사이버 보안 강화: 디지털 사회에서 개인 정보의 안전을 확보하고, 사이버 공격으로부터 정부 시스템 및 국민의 정보를 안전하게 보호하기 위한 정책을 수립하고 실행한다.
- 디지털 사회의 이념 및 목표 설정: 장기적인 비전과 전략을 수립하여 일본이 나아가야 할 디지털 사회의 방향을 제시한다.
- 디지털화 관련 법률 및 제도 정비: 기존 법률을 검토하고 새로운 법률 및 제도를 제정하여 디지털 전환을 위한 제도적 기반을 마련한다.

디지털청은 이런 업무 내용을 바탕으로 다음과 같은 서비스를 제공하고 있다.

디지털서비스 이슈리포트

- e-Gov 관련 서비스

- 개인 정보 보호: 개인 정보 파일을 보유한 각 부처가 e-Gov에서 공표한 개인 정보 파일을 정부 횡단적으로 검색할 수 있다.
- 데이터 포털: 행정기관 등이 공개하고 있는 오픈 데이터를 검색할 수 있다.
- 전자 신청: 각 부처가 소관하는 다양한 행정 절차에 대해 검색, 전자 신청, 신고를 할 수 있다.
- 공개 코멘트: 행정기관이 정령(政令)이나 성령(省令) 등을 정하고자 할 때, 사전에, 널리 일반으로부터 의견을 모집해, 그 의견을 고려하는 것으로, 행정 운영의 공정함의 확보와 투명성의 향상을 도모해, 국민의 권리 이익의 보호에 도움을 주는 것을 목적으로 하고 있다.
- 문서 관리: 행정기관이 기재하고 공표한 행정 문서 파일을 검색할 수 있다.
- 법령 검색: 법령 데이터를 2차 사용하기 쉬운 오픈 데이터로 제공하는 웹 서비스
- e-Gov 포털: 각 행정기관이 제공하고 있는 행정정보에 대한 종합적인 검색과 안내에 더해 전자 신청, 신고 등을 일원화하여 제공하는 웹 서비스이다.

- 마이넘버 관련 서비스

- 마이너 포털: 육아 등 다양한 행정 수속을 위한 온라인 창구로 점차 제공 기능이 확대되고 있다. 건강 진단 같은 의료 정보, 운전면허 기능, 여권 신규 신청 등도 할 수 있다.
- 마이너 포털 앱: 마이넘버 카드를 이용해 스마트폰에서 마이너 포털을 사용한다.
- 디지털 인증 앱: 마이넘버 카드를 사용한 본인 확인을 안전하고 간단하게 하기 위한 앱이다.
- 마이넘버 카드 대면 확인 앱: 마이넘버 카드의 IC 칩을 읽어 저장된 이름 등 본인 정보를 확인하기 위한 앱이다.

- Visit Japan Web: 입국 절차 및 면세 구입에 필요한 정보를 등록할 수 있는 웹 서비스

- 법인의 절차 지원

- G 비즈 ID: 1개의 ID/패스워드로 복수의 사업자용 행정 서비스에 본인 확인 서류 필요 없이 로그인할 수 있는 서비스
- jGrants: 부성청, 지자체가 운영하는 보조금, 조성금에 온라인 신청하기 위한 웹 서비스
- 법인 설립 원스톱 서비스: 법인 설립에 관한 다양한 절차를 한꺼번에 정리해 온라인 신청할 수 있는 서비스

- 데이터 표준화, 데이터 연계

- 레지스트리 카탈로그: 오픈 데이터로 공개 가능한 정부 기관의 데이터를 게재하는 카탈로그 사이트

디지털서비스 이슈리포트

- 생활 지원
 - 혜택 지원 서비스: 주민에게 신속하고 효율적인 급여 실현을 지원하는 웹 서비스
- 정부 조달
 - 디지털 마켓플레이스(DMP): 행정기관의 클라우드 소프트웨어(SaaS)의 조달 신속화와 다양한 벤더의 진입에 의한 조달처의 다양화를 목적으로 하는 카탈로그 사이트
 - 조달 포털: 통일된 참가 자격 신청에서 공공 조달의 검색, 입찰, 계약을 할 수 있는 웹 서비스
- 디지털 추진 위원 포털: 디지털 기기, 서비스에 익숙하지 않은 분들을 지원 활동하는 디지털 추진 위원용 포털 사이트
- 지방 창생
 - 지역 행복도(웰빙) 지표 활용 서비스: 행복도 지표의 대시보드 및 가이드북을 제공하는 웹 서비스
 - 디지털 지방 창생 서비스 카탈로그: 우수한 디지털 구현을 지원하는 서비스를 카탈로그로 공개
- 아날로그 규제 검토
 - 기술 카탈로그: 기술 보유 사업자 등으로부터 응모를 통해 아날로그 규제에 대한 재검토에 활용 가능성이 있는 개별 제품 및 서비스 정보를 정리한 카탈로그 사이트
 - 기술 지도: 아날로그 규제 재검토에 활용할 수 있는 기술을 파악할 수 있도록, 규제와 기술 유형의 대응 관계를 정리한 맵을 제공

주요 서비스 내용을 살펴보면 많은 행정 서비스가 이제 포털이나 앱을 통해서 지원되고 있다는 것을 알 수 있다. 일본이 기본으로 추진하는 마이넘버 카드 보급도 이제는 78%를 넘어서 전체 9천7백70만 명 이상이 발급을 받았다고 한다.



1日あたり (表示最新月の平均)

11,257

表示最新月内の合計

348,972

前月の合計: 422,909 (-73,937)

그림 9 마이넘버 카드 보급 현황

디지털서비스 이슈리포트

디지털청의 슬로건은 "스타트업으로서의 정부(Government as a Startup)"로, 관료주의적 관행을 타파하고 빠르고 유연하게 디지털 전환을 추진하겠다는 의지를 담고 있다. 각 서비스를 안내하는 웹사이트 구성도 매우 일관성 있는 디자인으로 보이며 영국 정부 포털의 모습을 많이 닮았다는 느낌이다.

디지털청의 연간 보고서는 현재 2024년 9월에 나온 2023년 8월부터 2024년 8월까지의 업무 성과에 대한 것이 나와 있다.³⁾ 디지털청이 진행하는 다양한 프로젝트에 대한 정보는 이 연간 보고서를 참고하면 도움이 된다.

AI와 클라우드 역량을 강화하는 일본

디지털청은 2021년부터 정부 클라우드 구축을 추진해 왔다. 정부 클라우드는 12개 중앙 정부 부처, 1,788개 지방 정부 기관, 그리고 준공공 부문 기관을 위한 빠르고 유연하며 안전하고 비용 효율적인 클라우드 서비스 환경 구축을 목표로 했다. 2021년 디지털청은 AWS와 구글의 GCP를 정부 클라우드 이니셔티브의 파트너로 선정했으며, 2022년에 마이크로소프트 애저와 오라클을 추가 파트너로 선정했다.

2023년 '정부 정보시스템에서 클라우드 서비스의 적절한 사용을 위한 규범 지침'⁴⁾이 나오면서 디지털청은 부 클라우드 사용 시스템의 전면적인 개혁과 현대화를 추진해 왔다. 정부 클라우드 서비스 제공업체는 약 300개 기준에 명시된 엄격한 보안 및 데이터 저장 요건을 충족해야 했다. 이전에는 각 기업이 모든 기준을 개별적으로 충족해야 했는데, 그 결과 글로벌 네트워크를 보유한 주요 미국 기업만이 이러한 요건을 충족할 수 있었다.

일부 국내 기업들은 정부의 엄격한 요건을 재검토하고, 요건을 충족하기 위해 다른 기업의 서비스를 이용하도록 허용하거나, 여러 기업이 함께 서비스를 제공하도록 허용할 것을 요구했다. 새로운 기준에 따라 선정된 기업은 데이터 관리 및 인증 서비스와 같은 핵심 기능을 독립적으로 제공할 수 있는 한, 다른 기업이 제공하는 서비스를 통합할 수 있게 되었다. 따라서 데이터 분석 및 기타 기능을 다른 기관에 위탁하는 것도 허용된다.

3) Digital Agency, "Digital Agency Annual Report 2024," Sep 4, 2024

4) 일본어 버전은

https://www.digital.go.jp/assets/contents/node/basic_page/field_ref_resources/e2a06143-ed29-4f1d-9c31-0f06fca67afc/5167e265/20230929_resources_standard_guidelines_guideline_01.pdf 에서 찾아볼 수 있다.

디지털서비스 이슈리포트

Selection criteria for govt cloud providers



그림 10 정부 클라우드 선정 요건의 완화 [출처: 요미우리 신문]

2024년 디지털청은 변화를 꾀하는데, 2024년 9월에 미국 기업의 독점을 깨고 국내 생산 제품을 포함하도록 선정 절차를 수정했다. 여러 회사의 서비스를 사용하여 선정 요건을 충족할 수 있도록 허용하기로 한 것이다. 이로써 아마존 AWS, 구글 클라우드, 마이크로소프트 애저, 오라클이 제공하던 정부 클라우드 서비스 파트너사에 오사카 기반의 IT 회사인 사쿠라 인터넷이 참여하게 되었고, 이는 일본 국내 기업의 역량 강화와 함께 기술 독립을 위한 준비 단계로 해석할 수 있다.

정부 클라우드를 구축하려고 하는 디지털청의 노력은 인력과 공급업체 직원 부족, 클라우드 기술 경험 부족, 변화에 대한 문화적 저항과 같은 문제에 직면해 왔다. 이러한 어려움을 극복하기 위해 디지털청은 AWS를 포함한 클라우드 서비스 공급업체(CSP)에 추가 리소스와 전문 지식을 요청했다.

품질 보장, 생산성 향상, 그리고 현대적인 디자인의 모범 사례 제시를 위해 디지털청은 참조 아키텍처(RA)를 개발했다. 이 RA는 현재 행정 시스템의 특성에 맞춰 최적화된 클라우드 아키텍처 표준을 제시한다. RA는 일반 관리, 운영 블록, 그리고 비기능 블록으로 구성되며, 이 블록들을 결합하여 모델 시스템 패턴을 형성한다. RA가 처음 제공된 2023년 11월부터 2024년 6월까지 약 40건의 마이그레이션 프로젝트 시스템 계획에 RA가 사용되었고, 2024년 9월 기준으로 수백 건의 마이그레이션 프로젝트가 진행 중이다.⁵⁾

5) AWS, "Japan's Digital Agency accelerates government cloud migration with AWS generative AI-powered architecture reviews." Sep 3, 2024.

디지털서비스 이슈리포트

일본은 소버린 클라우드와 함께 소버린 AI에 대한 역량을 강화하기 위한 노력을 하고 있다. 한 예로 일본 산업기술종합연구소(AIST)는 수천 대의 엔비디아 H200 텐서 코어 GPU를 AI 브릿징 클라우드 인프라 3.0(ABCI 3.0) 슈퍼컴퓨터에 통합하기로 했다.⁶⁾ ABCI 3.0은 AI 연구 개발을 촉진하기 위해 설계된 일본의 대규모 개방형 AI 컴퓨팅 인프라의 최신 버전이다.

ABCI 3.0은 AIST, 그 자회사인 AIST Solutions, 그리고 시스템 통합자인 휴렛 팩커드 엔터프라이즈 (HPE)가 구축하고 운영한다. 이 프로젝트는 경제 안보 기금을 통해 컴퓨팅 리소스를 강화하기 위한 일본 경제산업성(METI)의 지원에 따른 것이며, ABCI 활동과 클라우드 AI 컴퓨팅에 대한 투자를 모두 포함하는 METI의 10억 달러 규모의 광범위한 이니셔티브의 일부이다.

2024년 11월에는 일본 소프트뱅크와 엔비디아 간 전략적인 협력을 발표했다. 먼저 세계 최초로 엔비디아 DGX B200 시스템을 도입해 AI 슈퍼컴퓨터를 만든다는 것이다. 이 슈퍼컴퓨터는 25억사플롭스의 AI 성능을 활용하여 생성 AI 프로젝트부터 일본 대학 및 기업의 복잡한 연구에 이르기까지 모든 것을 가속할 것이다. 두 번째는 세계 최초의 실시간 AI 및 5G 하이브리드 네트워크인 AI-RAN(인공지능 무선 접속 네트워크)을 선보일 예정이다. 엔비디아 AI 에어리얼 엔진을 소프트뱅크의 5G 네트워크에 통합해 통신 인프라를 AI 네트워크로 진화시킬 것이라고 했다. 세 번째는 AI 마켓플레이스로 소프트뱅크는 국내 및 외부 개발자들이 AI 애플리케이션을 공유할 수 있도록 AI 스토어를 구축하고 있다.

일본은 METI를 중심으로 하는 새로운 이니셔티브를 2024년 5월에 발표했다.⁷⁾ 이 이니셔티브는 AI의 경제적 잠재력을 실현하고 국가 인력을 강화하는 것을 목표로 하며 GMO 인터넷 그룹, 하이레소, KDDI 주식회사, 루틸레아, 사쿠라 인터넷 주식회사, 소프트뱅크 주식회사를 포함한 여러 디지털 인프라 제공업체를 첨단 AI 애플리케이션에 필요한 클라우드 인프라 개발의 핵심 파트너로 선정했다.

경제산업성(METI)은 AI 컴퓨팅 자원에 대한 보조금을 지원하고 지역 기업의 프로젝트 참여를 확대하기 위해 약 1,146억 엔(미화 7억 4천만 달러)을 배정했다. 또한 일본 정부는 2024년 11월 AI와 반도체 산업에 2030년까지 최소 10조 엔(약 97조 8,000억 원)을 투자하겠다고 밝혔다. AI 활용 소프트웨어를 개발하는 기업엔 2032년까지 최대 30%까지 법인세를 감면하며, 데이터센터 건설에도 최대 450억 엔 보조금을 지원한다. 민간 기업인 소프트뱅크는 자체적으로 AI에만 매년 90억 달러를 투자하겠다고 한다.

6) NVIDIA, "Japan Enhances AI Sovereignty With Advanced ABCI 3.0 Supercomputer," Jul 11, 2024

7) Tech Insight Zone, "Japan is investing \$740 Million to Achieve AI Supremacy," May 14, 2024

디지털서비스 이슈리포트

일본 AI 클라우드 시장에 투자하는 해외 기업

스태티스티카에 따르면 일본의 정부 클라우드 시장 규모는 2023 회계연도에 73억 엔으로 추산되었다. 점점 더 많은 지자체가 클라우드 서비스를 도입할 계획이기 때문에 2027 회계연도에는 시장이 1,230억 엔 이상으로 확대될 것으로 예상되었다.

오라클은 2024년 4월 일본 클라우드와 AI 분야에 향후 10년 동안 80억 달러를 투자하겠다고 했다.⁸⁾ 이번 투자는 일본 전역에서 오라클 클라우드 인프라스트럭처(OCI)의 입지를 확대하는데 기여할 것으로 보인다. 또한 같은 해 10월에는 NTT 데이터와 협력을 통해 일본의 소버린 클라우드 역량 강화를 도모하겠다고 했다.

마이크로소프트와 아마존도 각각 일본 데이터센터와 AI 관련 연구소 등에 투자를 약속했는데, 마이크로소프트는 2년간 약 29억 달러, 아마존은 3년간 약 150억 달러 규모의 투자를 진행할 예정이다. 구글 역시 해저 케이블 두 개를 포함한 10억 달러의 투자 계획을 발표했다.

알리바바는 자사의 오픈 소스 모델인 Qwen을 활용하는 일본 파트너를 늘리기 위해 노력을 하고 있다. 예를 들어, 물리적 공간 분석 시스템을 개발하는 아베자(Abeja)는 4월 17일, 새로운 QwQ-32B 추론 모델을 발표했다. 도쿄대학 출신들이 만든 엘라이자와 라이트블루, Axccept를 포함하여 다른 일본 스타트업도 Qwen 기반 모델을 출시했거나 개발 중이다. 알리바바 일본 사업부는 3년 안에 Qwen 사용 모델을 1,000개 수준으로 늘리겠다고 한다.

오픈AI와 소프트뱅크는 오사카에 있는 샤프의 LCD 공장을 철거하고 여기에 아시아 최대 규모의 데이터센터를 짓기로 했다. 여기에 엔비디아의 최첨단 GPU 10만 장을 투입하는데 총비용은 1조 엔이 들어갈 것으로 예상된다.

나가면서

일본이 디지털 정책을 2016년부터 소사이어티 5.0이라는 이름으로 추진했지만, 그 성과나 진행에 대해서는 우리에게 큰 인상을 주지 못하고 있었다. 그러나 2023년부터 변화의 조짐을 보이기 시작했고 이미 2021년에 설립한 디지털청을 통해 디지털 정책과 클라우드 정책의 급격한 변화를 꾀하기 시작했다는 것을 알아야 한다.

8) Oracle, "Oracle to Invest More Than \$8 Billion in Cloud Computing and AI in Japan," Apr 17, 2024

디지털서비스 이슈리포트

최근 각 국가의 AI 정책 변화와 맞물려 일본 역시 산업 진흥책과 적극적인 투자, 글로벌 기업과의 협력을 바탕으로 일본을 아시아의 새로운 AI 리더 국가로 만들고자 하는 도전을 보이고 있고, 소버린 클라우드와 소버린 AI 정책을 중심으로 적극적인 투자를 민간과 함께 추진하고 있음을 알 수 있다.

이제는 많은 전문가가 일본이 더 이상 클라우드와 AI에 있어 한국보다 뒤처진 나라가 아니라는 것을 인정해야 한다는 목소리가 나오고 있으며, 아날로그 기반에서 탈출하려고 하는 전략을 공공 섹터를 포함해 모든 분야에서 빠르게 구현하고 있음을 알 수 있다.

일찍부터 글로벌 기업과 협력을 바탕으로 디지털 전환을 꾀했으며 AI 시대에 있어서도 국내 기업을 양성함과 동시에 엔비디아나 오픈AI와 같은 기업과 규모 있는 투자와 적극적인 제휴를 추진하고 있음을 알아야 한다.

일본의 잠재성은 여전히 세계 3위의 경제 규모, 잘 발달된 제조 산업, 뛰어난 과학 연구 수준, 정부 정책에 적극 호응하는 기업과 시민들의 자세에 있다고 보인다. 몇 년 동안 움츠렸던 일본의 IT 산업이 새로운 도약을 위한 준비를 체계적으로 하고 있는 모습은 우리에게 시사하는 바가 크다고 본다.

04 2025 구글 I/O와 JAX 업데이트

| 정채상 인이지 연구소 기술 책임자

들어가며

매년 5월에 열리는 구글 I/O는 기술 혁신의 최전선을 조망하고, 구글의 미래 비전을 공유하는 자리이다. 2025년 행사 역시 예외는 아니었고, 전 세계 개발자들을 위한 다양한 업데이트가 발표되었다. AI 시대를 맞아 새롭게 선보인 제품과 서비스는 여러 매체에서 폭넓게 다뤄졌지만, 이 글에서는 JAX를 중심으로 AI 개발자들이 특히 주목할 만한 소식들을 정리해 본다.

관련 세션으로는 ‘JAX의 실제 적용(JAX in Action)⁹⁾’, ‘케라스 3에서 `model.fit(...)`으로 JAX 모델 학습하기(Train your JAX models using `model.fit(...)` in Keras 3)¹⁰⁾’, ‘JAX로 트랜스포머 만들기(Build a Transformer with JAX)¹¹⁾’ 등이 공개되었으며, JAX에 대한 더 자세한 정보는 <https://jax.dev/>에서 확인할 수 있다.

JAX 소개

JAX는 구글에서 개발한 고성능 수치 계산 라이브러리로, NumPy 스타일의 코드를 사용하면서도 자동 미분과 하드웨어 가속(GPU/TPU)을 손쉽게 활용할 수 있도록 설계되었다. 머신러닝과 과학 계산 등 고성능 계산이 필요한 분야에서 각광받아 왔으며, 최근 대규모 훈련 등 효율적인 자원 사용이 중요해지면서 사용이 늘고 있다.

JAX는 "Just After eXecution"의 줄임말로, 함수형 프로그래밍 패러다임을 바탕으로 설계되어 연구자와 개발자들이 복잡한 수치 계산을 효율적으로 수행할 수 있도록 돕는다. JAX의 주요 특징들은 다음과 같다.

- NumPy와 유사한 인터페이스: 데이터 분석에서 가장 널리 쓰이는 NumPy와 거의 동일한 문법과 기능을 제공한다. 따라서 기존 NumPy 코드에 익숙한 개발자들도 쉽게 JAX로 전환하여 하드웨어 가속의 이점을 누릴 수 있다.

9) <https://io.google/2025/explore/technical-session-1>

10) <https://io.google/2025/explore/technical-session-3>

11) <https://io.google/2025/explore/workshop-2>

디지털서비스 이슈리포트

- 자동 미분 (Automatic Differentiation): 복잡한 함수의 기울기를 자동으로 계산할 수 있다. 순방향과 역방향 자동 미분을 모두 지원하며, 고차 미분도 가능하다. 이는 딥러닝에서 역전파(back propagation) 구현에 필수적이며, 효율적인 최적화를 가능하게 한다.
- JIT (Just-In-Time) 컴파일: 파이썬 언어의 데코레이터를 사용하여 함수를 XLA(Accelerated Linear Algebra) 컴파일러로 컴파일할 수 있다. 이는 코드 실행 시점에서 최적화된 기계어 코드로 변환하여 연산 속도를 대폭 향상시키고, CPU와 GPU에서 상당한 성능 향상을 얻을 수 있다.
- 벡터화 및 병렬화: 간단한 함수를 이용해서 스칼라 함수를 벡터 함수로 자동 변환할 수 있다. 이는 배치 처리나 몬테카를로 시뮬레이션 등에서 코드를 간결하게 유지하면서도 높은 성능을 달성하게 해 주며, 병렬 계산을 통해 분산 학습이나 대규모 계산에서도 높은 성능을 가능하게 한다.
- 함수형 프로그래밍 패러다임: 순수 함수(pure functions)와 불변 배열(immutable arrays) 사용을 권장하는 함수형 프로그래밍 모델을 따른다. 이는 코드의 예측 가능성을 높여 테스트, 디버깅, 병렬화 및 최적화를 용이하게 한다.

아래는 구글 코랩¹²⁾에서 엔비디아 T4 칩에 연결해 성능을 비교한 결과로, 간단한 병렬 연산의 경우 1,000배 이상의 성능 차이를 볼 수 있다.

```
import numpy as np

x = np.ones((5000, 5000))
y = np.arange(5000)

%timeit z = np.sin(x) + np.cos(y)

352 ms ± 7.42 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)
```



```
import jax.numpy as jnp
x = jnp.ones((5000, 5000))
y = jnp.arange(5000)

%timeit z = jnp.sin(x) + jnp.cos(y)

223 µs ± 102 µs per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

그림 11 JAX 병렬 연산의 속도 비교

12) <https://colab.research.google.com/>

디지털서비스 이슈리포트

```
from jax import jit
import tensorflow as tf
```

```
def fn(x, y):
    z = np.sin(x)
    w = np.cos(y)
    return z + w
```

```
@jit
def fn_jit(x, y):
    z = jnp.sin(x)
    w = jnp.cos(y)
    return z + w
```

```
x = np.ones((5000, 5000))
y = np.ones((5000, 5000))
%timeit fn(x, y)
```

790 ms ± 186 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)

```
jx = jnp.ones((5000, 5000))
jy = jnp.ones((5000, 5000))
%timeit fn_jit(jx, jy)
```

1.22 ms ± 613 ns per loop (mean ± std. dev. of 7 runs, 1000 loops each)

그림 12 JIT를 사용한 JAX 병렬 연산의 속도 비교

구글이 지원하는 생태계: JAX와 케라스(Keras)

구글은 개발자들이 다양한 수준과 목적에 맞춰 AI 개발을 할 수 있도록 체계적인 생태계를 구축하고 지원해 왔다. 하드웨어 레이어에서는 자체 제작한 TPU를 포함해 CPU, GPU를 아우르는 광범위한 지원을 제공하며, 그 위에는 JAX, 텐서플로우와 같은 저수준 프레임워크를 지원한다. 특히 JAX는 구글의 TPU와 최적화된 연동을 통해 대규모 연구 프로젝트에서 진가를 발휘한다.

고수준 프레임워크로는 사용하기 쉬운 API를 제공하는 케라스가 핵심 역할을 한다. 또한 특정 도메인에 특화된 라이브러리들인 KerasTuner(하이퍼파라미터 튜닝), KerasHub(사전 훈련된 모델), Keras Recommenders(추천 시스템) 등이 개발자들의 구체적인 요구사항을 충족한다. 이러한 계층적 구조는 연구자들이 저 수준의 세밀한 제어가 필요할 때는 JAX를 활용하고, 빠른 프로토타이핑이나 프로덕션 배포가 목적인 때는 케라스를 선택할 수 있는 유연성을 제공한다.

디지털서비스 이슈리포트

Ecosystem

The Keras ecosystem also offers several domain-specific packages for training LLMs, recommender models, and more.

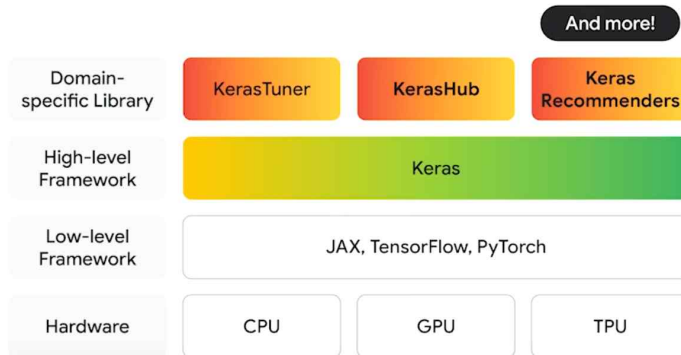


그림 13 케라스와 JAX를 포함하는 구글이 지원하는 생태계

JAX의 저 수준 성능 제어: 이스케이프 해치(Escape Hatch)와 팔라스(Pallas)

JAX는 강력한 자동 미분과 JIT 컴파일 기능을 제공하며 대부분의 머신러닝 워크로드에 충분한 성능을 제공한다. 하지만 때로는 극도로 최적화된 성능이 필요하거나, 특정 하드웨어의 저 수준 기능을 직접 제어해야 하는 경우가 생기는데, 이를 위해 JAX는 '이스케이프 해치(Escape Hatch)'를 위한 계층을 지원한다.

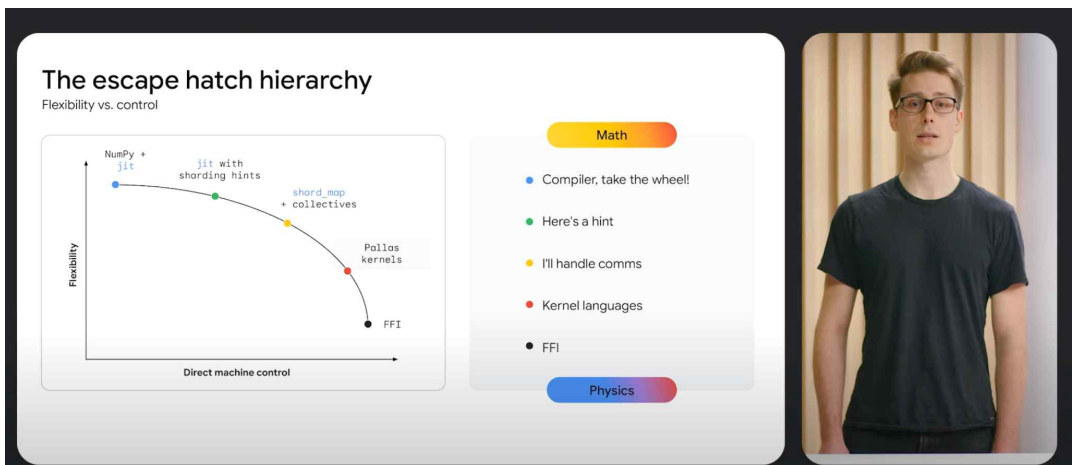


그림 14 JAX에서 성능과 유연성을 지원하는 이스케이프 해치 계층

디지털서비스 이슈리포트

간단한 함수를 이용하는 것, 그리고 JIT와 XLA를 이용하는 것보다 더 하드웨어에 가까운 계층이 팔라스(Pallas)라는 커널 언어다. 이는 JAX 생태계의 새로운 구성 요소로, GPU와 TPU에서 고성능 커스텀 커널을 작성할 수 있게 해 주는 도구다. 기존에는 쿠다(CUDA)나 트리톤(Triton) 같은 저 수준 언어를 배워야 했던 커널 프로그래밍을, JAX의 친숙한 문법으로 할 수 있게 만들어 준다.

이 팔라스의 핵심 아이디어는 “블록 수준 프로그래밍”이다. 전체 벡터 혹은 텐서를 한 번에 처리하는 대신, 메모리에 맞는 작은 블록들로 나누어 처리함으로써 메모리 계층 구조를 효율적으로 활용할 수 있게 한다. 이는 특히 대규모 행렬 연산이나 트랜스포머의 어텐션 메커니즘 같은 메모리 집약적인 연산에서 큰 성능 향상을 가져올 수 있다. 게다가 GPU의 공유 메모리나 TPU의 벡터 처리 유닛 등을 직접 제어할 수도 있어, 이론적 성능의 한계에 근접한 최적화가 가능하다. 이는 연구자와 개발자들이 하드웨어 성능의 한계를 탐험할 수 있는 새로운 가능성을 열어주었다.

```
import jax
from jax.experimental import pallas as pl

def matrix_multiply_kernel(a_ref, b_ref, o_ref):
    """Pallas로 작성한 행렬 곱셈 커널"""
    # 블록 단위로 데이터를 처리
    a_block = a_ref[:]
    b_block = b_ref[:]

    # 커스텀 연산 수행
    result = jnp.dot(a_block, b_block)
    o_ref[:] = result

@pl.pallas_call(
    out_shape=jax.ShapeDtypeStruct((1024, 1024), jnp.float32),
    grid=(8, 8), # 8x8 블록으로 분할
)
def custom_matmul(a, b):
    return matrix_multiply_kernel(a, b)
```

그림 15 팔라스로 작성한 커스텀 행렬 곱셈 함수 예제

디지털서비스 이슈리포트

JAX를 활용한 논문 동향

2025년 3월 PapersWithCode 집계에 따르면, 최근에 발행된 논문에서 JAX는 3%~4% 정도의 점유율을 보인다. 주로 대규모 언어 모델을 포함한 복잡한 연구 분야에서 효율적인 활용성을 인정받으며 사용되고 있다. 2021년의 1%대에 머물렀던 것과 비교하면 최근 들어 빠르게 점유율을 높이며 텐서플로우 영역을 대체하는 추세다.

Frameworks

Paper Implementations grouped by framework

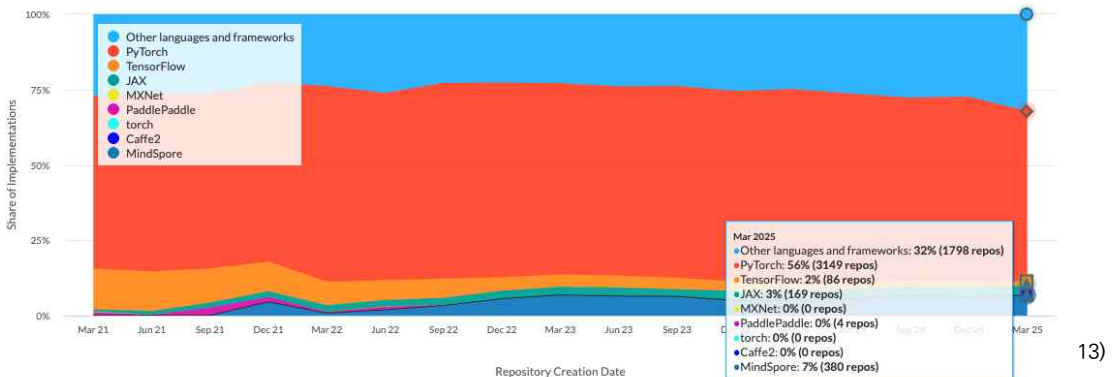


그림 16 AI 논문에 인용된 프레임워크 점유율

참고로 중국 연구의 약진과 함께 화웨이의 마인드스포어(MindSpore)가 7% 정도의 점유율로 파이토치 다음으로 널리 쓰이고 있다.

맺으며

AI 기술의 발전 속도가 더욱 빨라지는 이 시대에, JAX는 고성능 컴퓨팅의 한계를 뛰어넘어 혁신적인 AI 모델을 구현하는 데 필수적인 도구로 자리매김하고 있다. 파이토치의 강력한 경쟁 프레임워크로서 JAX가 제시하는 유연하고 강력한 생태계는 앞으로도 수많은 AI 연구와 실제 애플리케이션에서 핵심적인 역할을 할 것이다.

13) <https://paperswithcode.com/trends>

05 대만 엔비디아 컨스텔레이션 설립의 의미와 시사점

| 윤대균 아주대학교

1. 들어가며

5월 20일부터 23일까지 대만에서 컴퓨텍스 2025가 개최되었다. 컴퓨텍스(COMPUTEX)¹⁴⁾는 1981년부터 시작된 아시아 최대 규모의 IT 및 컴퓨터 박람회로, 매년 대만 타이베이에서 개최되는 글로벌 기술 전시회이다. PC 및 컴퓨터 부품 중심의 전시회로 출발하여 2000년대 초반까지 하드웨어 제조업체들의 신제품과 기술을 발표하는 주요 무대로 성장했다. 2010년대에는 모바일, IoT, 스마트 기기로 영역을 확장하여 2020년대에는 AI, 로봇틱스가 무대의 중심을 차지하고 있다.

2025년 컴퓨텍스는 'AI Next'를 핵심 주제로 삼아, AI와 이를 기반으로 한 다양한 기술 분야를 집중적으로 다루었다. 로봇 공학, 미래 모빌리티 등 여러 산업에서 AI의 응용 가능성을 선보였으며, 많은 기술 기업과 스타트업이 관련 솔루션을 발표했다. 그러나 뭐니 뭐니 해도 이번 컴퓨텍스의 하이라이트는 엔비디아 젠슨 황 대표의 키노트 연설이었다. 이 연설에서 자사의 새로운 제품 및 기술과 이를 중심으로 엮이는 생태계를 상세히 소개했다. 이 키노트의 클라이맥스는 '엔비디아 컨스텔레이션(Nvidia Constellation)¹⁵⁾' 설립 발표이다.

이는 단순히 대만에 대규모의 엔비디아 오피스를 설립한다는 의미를 넘어 키노트 내내 강조했던 정부 및 기업과의 긴밀한 파트너십 그리고 이를 기반으로 한 AI 생태계 구축을 완성한다는 관점에서 매우 의미 있다. 특히 젠슨 황이 이끄는 엔비디아가 GPU를 공급하는 업체에서 AI 컴퓨팅 생태계의 중심으로 자리잡아 이끌겠다는 미래 비전을 제시하고 있다. 그의 키노트에서 언급한 부분에서도 잘 나타나 있다.

"AI는 이제 인프라입니다. 그리고 이 인프라는 인터넷처럼, 전기처럼, 공장이 필요합니다. 이러한 공장들이 본질적으로 우리가 오늘날 구축하는 것입니다. 이것들은 과거의 데이터센터가 아니며, 에너지를 투입하면 믿을 수 없을 정도로 가치 있는 무언가를 생산해 내며, 이런 것들을 토큰이라고 부릅니다."

14) <https://www.computextaipei.com.tw/en/index.html>

15) 2018년 GTC에서 처음 소개된 자율주행차 개발을 위한 시뮬레이션 시스템인 엔비디아 드라이브 컨스텔레이션과 혼동하기 쉽다.

디지털서비스 이슈리포트



그림 17 컴퓨텍스 2025 키노트에서 엔비디아 컨스텔레이션을 소개하는 젠슨 황

젠슨 황이 말하는 ‘AI 팩토리’ 비전 및 전략에 그리고 엔비디아 컨스텔레이션 설립 발표에 담겨 있는 대만에서의 AI 생태계에 대해서 살펴보고자 한다.

2. AI 팩토리 - 데이터센터에서 AI 인프라로

젠슨 황이 제시한 ‘AI 팩토리’는 기존 데이터센터의 개념을 뛰어넘는 혁신적인 비전을 담고 있다. 이는 단순히 데이터를 저장하고 처리하는 공간을 넘어, 방대한 양의 데이터를 투입하여 새로운 형태의 가치, 즉 ‘지능’을 생산하는 산업 규모의 컴퓨팅 인프라를 의미한다. 젠슨황은 AI 팩토리가 전기를 투입하여 ‘토큰(tokens)’이라는 단위로 측정되는 지능을 산출한다고 설명하며, 이는 마치 과거 전기나 인터넷이 그랬던 것처럼 AI가 사회 전반의 필수적인 인프라로 자리매김할 것임을 시사한다. 이러한 AI 팩토리의 궁극적인 목표는 모두가 접근할 수 있는 AI 기술을 통해 전 세계 모든 산업 분야에서 디지털 전환을 가속화하고 혁신을 촉진하는 것이다.

엔비디아는 이러한 비전을 통해 자사를 단순한 칩 제조사가 아닌 "AI 인프라 기업"으로 재정의하고 있다. 특히 젠슨 황이 AI 팩토리에서 생산되는 것을 '토큰'이라고 강조한 것은 주목할 만하다. 이는 AI의 결과물을 마치 전력량이나 데이터 대역폭처럼 정량화하고 표준화할 수 있음을 의미하고 AI 팩토리의 최종 제품이 거래 가능한 상품이 될 수 있음을 암시한다. 즉, AI가 생성하는 ‘지능’에 대한 새로운 시장이 형성되고, AI 투자에 대한 ROI 계산이 더욱 정밀해지는 등 새로운 비즈니스 모델의 등장을 촉진할 수 있다. 이를 바탕으로 엔비디아는 AI 팩토리 비전을 통해 AI 시대의 핵심 유틸리티

디지털서비스 이슈리포트

공급자로서의 입지를 구축하려 하고 있다. 이는 전력 회사나 인터넷 서비스 제공자와 유사한 수준의 필수 불가결인 핵심 인프라 역할을 목표로 하는 전략으로 해석될 수 있다. 젠슨 황이 "우리는 이제 전 세계적으로 필수적인 AI 인프라 회사임을 깨달았다"고 언급하면서 장기적인 인프라 구축 계획을 시사한 것은 엔비디아의 아키텍처를 AI 생태계 깊숙이 통합하여 사실상의 표준으로 만들고 장기적인 시장 지배력을 확보하려는 의도로 풀이된다.¹⁶⁾

AI 팩토리 비전의 기술적 기반은 엔비디아의 차세대 GPU 아키텍처인 블랙웰(Blackwell)이다. 컴퓨텍스 2025에서 공개된 GB300 및 블랙웰 울트라(Blackwell Ultra)는 이전 세대 대비 추론 성능 1.5배, HBM(고대역폭 메모리) 1.5배, 네트워킹 대역폭 2배 향상 등 획기적인 성능 개선이 이루어졌다고 한다. 이러한 고성능 칩들을 집약하여 랙 단위의 액침 기술을 구현한 GB300 NVL72 랙 스케일 솔루션은 AI 팩토리의 주요 구성 요소가 될 것이다.¹⁷⁾ 또한, 기업들이 GPU를 유연하게 결합하여 맞춤형 AI 인프라를 구축할 수 있도록 지원하는 NVLink 퓨전 기술도 중요한 역할을 한다.¹⁸⁾

NVLink 퓨전은 엔비디아의 고속 인터커넥트 기술인 NVLink를 타사 칩 제조업체들도 사용할 수 있도록 개방한 것이다. 이를 통해 고객들은 엔비디아 GPU와 다른 회사의 CPU, ASIC(주문형 반도체) 등을 하나의 시스템에서 통합하여 맞춤형 AI 인프라를 구축할 수 있게 된다. 미디어텍, Marvell, 컬컴, 후지쯔, Alchip 등 주요 반도체 기업들이 참여하였으며, 기존의 폐쇄적 생태계에서 개방형 협력 모델로 전환했다는 것에 큰 의미를 부여할 수 있다. 이 기술은 엔비디아가 마이크로소프트, 아마존 같은 대형 클라우드 기업들이 자체 AI 칩 개발을 늘리고 있는 상황에 대응하는 전략으로 판단되며, 고객들이 엔비디아 GPU를 완전히 대체하는 대신, NVLink 퓨전을 통해 자체 칩과 엔비디아 제품을 함께 사용하도록 유도하는 것이다.

엔비디아는 단순한 GPU 칩 제조사를 넘어 '지능형 인프라의 설계자' 또는 'AI 인프라 기업'으로 진화하고 있다. 이는 글로벌 데이터센터 및 소프트웨어 생태계 전반에 걸친 깊은 통합을 의미한다. 젠슨 황은 AI의 발전 단계를 인식 AI, 생성형 AI, 그리고 물리적 AI(Physical AI, 로봇틱스 및 자율 시스템)의 세 단계로 구분하며, 엔비디아의 전략은 이 모든 단계를 지원하는 것을 목표로 한다고 선언했다. 특히, 물리적 AI 훈련을 위한 시뮬레이션 플랫폼 '코스모스(Cosmos)'와 디지털 작업자로서의 '에이전틱 AI(Agentic AI)' 개념은 엔비디아가 지향하는 미래를 보여준다.

16) Nvidia, "NVIDIA CEO Envisions AI Infrastructure Industry Worth 'Trillions of Dollars'" May 18, 2025 <https://blogs.nvidia.com/blog/computex-2025-jensen-huang/>

17) <https://www.nvidia.com/en-us/data-center/gb300-nvl72/>

18) <https://www.nvidia.com/en-us/data-center/nvlink-fusion/>

3. AI 팩토리의 시험대로서의 대만 AI 생태계

대만 AI 생태계의 가장 큰 강점은 세계 최고 수준의 반도체 제조 능력과 강력한 제조 기반이다. TSMC는 2024년 4분기 전 세계 반도체 위탁 생산(pure-play wafer foundry) 시장에서 무려 67%의 점유율¹⁹⁾을 차지하고 있으며 GPU와 같은 첨단 칩의 경우 90%에 달할 것이라고도 한다. 대만 ODM(Original Design Manufacturer) 업체들은 글로벌 시장에서 매우 높은 점유율을 확보하고 있으며 이 중 폭스콘 한 회사가 전세계 ODM 시장의 40%를 차지하고 있다.²⁰⁾ 글로벌 서버 시장에서는 폭스콘이 델을 누르고 가장 많이 서버를 공급한 회사로 기록될 것이라는 전망도 나왔다.²¹⁾ 이러한 하드웨어 제조 역량은 AI 인프라 구축의 견고한 토대가 된다. 즉, AI 하드웨어 부품에 대한 강력한 공급망 역시 대만의 핵심 경쟁력이다.

컴퓨텍스 2025에서 발표한 엔비디아, 폭스콘, 대만 정부 간의 협력은 대만을 AI 혁신의 최전선에 세우려는 야심 찬 계획을 보여준다. 이들은 10,000개의 엔비디아 블랙웰 GPU를 탑재한 최첨단 AI 팩토리 슈퍼컴퓨터를 공동 구축하여 대만의 연구자, 스타트업, 산업계에 막대한 연산 능력을 제공할 예정이다.²²⁾ 세계 최대 파운드리 기업인 TSMC는 이 시스템의 주요 초기 수혜자로서, 반도체 제조 분야의 R&D 역량을 한층 강화할 것으로 기대한다.

이 AI 슈퍼컴퓨터 구축 프로젝트는 단순한 인프라 확충을 넘어, 대만 내에서의 긴밀한 파트너십을 바탕으로 하는 AI 혁신 생태계를 조성하려는 전략의 실현으로 해석할 수 있다. 최첨단 하드웨어 개발(엔비디아, TSMC), 시스템 통합(폭스콘), 그리고 AI 애플리케이션 개발이 지리적으로 근접한 환경에서 이루어짐으로써 혁신 주기를 단축시킬 수 있기 때문이다. 칩 설계(엔비디아), 칩 제조(TSMC), 서버/시스템 통합(폭스콘), 그리고 정부의 지원 및 방향 설정이라는 가치 사슬의 핵심 주체들이 모두 참여하는 이 프로젝트는 특히 TSMC가 반도체 R&D의 초기 수혜자가 됨으로써 더 나은 칩이 더 나은 AI 시스템을 만들고, 이는 다시 더 나은 칩 설계에 기여하는 직접적인 선순환 구조를 창출할 수 있다. 이러한 국지적으로 집중된 협력은 글로벌 분산된 모델에 비해 혁신 및 제품 개발 지연 시간을 크게 줄일 수 있다.

19) Focus Taiwan, "TSMC market share tops 67% in Q4: Advisory firm", Mar. 15, 2025
<https://focustaiwan.tw/business/202503150014>

20) <https://www.mk.co.kr/en/it/11322206>

21) Hosting Journalist, "Foxconn Projected to Become World's Largest Server Vendor by Omdia", Dec, 27, 2024

22) <https://nvidianews.nvidia.com/news/foxconn-ai-factory-tsmc-taiwan-nvidia>

디지털서비스 이슈리포트

엔비디아는 대만과의 파트너십을 더욱 공고히 하기 위해 타이베이 베이터우 시린 또는 시린-베이터우 과학단지에 새로운 대만 본사인 '엔비디아 컨스텔레이션(Nvidia Constellation)'을 설립한다. 이는 단순한 사무 공간을 넘어 AI 연구 개발, 인재 양성, 교육 프로그램 운영, 그리고 델타(Delta), 기가바이트(GIGABYTE)와 같은 현지 파트너와의 공급망 협력을 위한 중심 허브로 기능할 것이다. 이러한 투자는 엔비디아의 글로벌 운영에서 대만이 차지하는 핵심적인 역할을 다시 한번 강조하는 것이다.

엔비디아의 대규모 투자와 협력은 대만 산업 및 AI 지형에 광범위한 영향을 미칠 것으로 예상된다. 우선, 반도체, 헬스케어, 제조, 교육 등 주요 산업 분야의 연구 개발(R&D) 속도가 빨라질 것이다. 특히, 현지 AI 스타트업(예: Appier, Viscovery)은 최첨단 인프라 접근성 향상을 통해 성장 동력을 얻게 될 것이다. 이는 아시아 지역의 AI 인프라 부족 문제를 해결하는 데도 기여하며, 궁극적으로 대만의 글로벌 AI 경쟁력을 한 단계 끌어올릴 것이다. 이러한 움직임은 산업 변화를 주도하고 노동력 부족, 탄소 배출과 같은 사회적 문제를 해결하기 위해 AI를 활용하려는 대만의 국가 AI 전략, 'AI 대만 액션 플랜 2.0'과도 정확히 부합한다.²³⁾ AI 기반 기술이 대만의 GDP 성장에도 기여할 충분한 잠재력을 제공할 것이다.

엔비디아의 투자는 대만이 강점을 가지고 있는 기존 산업(예: 첨단 제조, 반도체 설계, 스마트 헬스케어)에 특화된 AI 애플리케이션 개발을 촉진하여, 대만 산업계에 글로벌 경쟁 우위를 제공할 수 있다. 최첨단 AI 인프라에 대한 접근은 이들 산업이 AI 도구를 단순히 채택하는 것을 넘어, 기존 역량을 강화하는 맞춤형 AI 솔루션을 개발할 수 있게 한다. 다만, 엔비디아와의 이러한 깊은 파트너십은 전략적 종속에 기인한 잠재적 취약성을 야기할 수 있으며, 따라서 대만 정부 담당자는 이를 신중하게 관리해야 할 것이다. 엔비디아의 하드웨어와 쿠다 소프트웨어 생태계는 매우 지배적이며, 국가 AI 인프라를 한 회사의 기술에 크게 의존하여 구축하는 것은 일종의 공급업체 종속을 초래할 수 있기 때문이다. 또한, 미중 기술 패권 경쟁과 같은 지정학적 이슈가 대만의 미래 기술에 대한 접근성에 영향을 미칠 수 있으므로, 균형 잡힌 국가 AI 전략이 필요하다.

4. 시사점

엔비디아는 최근 사우디, UAE에 대규모 칩 공급을 하기로 한 것에 이어 프랑스가 지난 2월 발표한 유럽 최대의 데이터센터 구축 계획에도 참여하기로 했다.²⁴⁾ 엔비디아와 아랍에미리트(UAE)의 국부펀드 MGX, 프랑스 국영 투자사 Bpi프랑스, 그리고 미스트랄 AI가 협력해 유럽 최대 규모의 AI 데이터센터

23) <https://ai.taiwan.gov.tw/news/executive-yuan-announces-artificial-intelligence-plan/>

24) <https://www.aitimes.com/news/articleView.html?idxno=170592>

디지털서비스 이슈리포트

캠퍼스를 프랑스 파리 인근에 조성한다는 것인데 프랑스는 이를 기반으로 대표 스타트업 미스트랄 AI를 유럽의 간판으로 키운다는 계획이다. 전세계 곳곳에서 '소버린 AI'를 통한 엔비디아의 사업 확대가 전개되고 있는 양상이다.

우리나라 정부에서는 'AI G3 도약 실현을 위한 국가 AI 역량 강화 방안'을 발표했다.²⁵⁾ 여기에 가장 핵심이 되는 것은 역시 AI 컴퓨팅 인프라의 확충이다. 첨단 GPU 1.8 만장을 확보하여 국가 AI 컴퓨팅센터에 1만 장, 그리고 8천 장 규모 슈퍼컴퓨터 6호기를 구축한다는 것이다. 이 밖에도 국산 AI 반도체 성장 지원, 인재 양성, 월드베스트 LLM 개발 같은 것이 포함되어 있다. 곧 있을 대선의 모든 후보도 AI에 대한 대규모 투자를 최우선 공약으로 내걸고 있다. 대선 이후 우리나라의 AI 전략 수정도 불가피할 것이다. 누가 당선이 되더라도 AI 컴퓨팅 인프라 확충에 대한 투자는 최우선 순위가 될 것으로 예상된다.

우리나라에도 대만의 TSMC나 폭스콘까지는 아니더라도 엔비디아와 파트너십을 맺고 있는 혹은 향후 맺기를 원하는 잠재력 있는 기업이 많이 있다. 이미 하이닉스와 같은 HBM 분야에서 엔비디아 파트너가 있고 삼성전자도 충분히 경쟁력을 갖추고 있는 반도체 분야의 강자다. 또한 AI 반도체에서 승부를 보고자 하는 여러 스타트업도 왕성하게 활동 중이며 네이버의 클로바x, LG 엑사원, 업스테이지 솔라프로와 같은 잘 알려진 파운데이션 모델을 보유하고 있다. 하지만 이들을 모두 아우르는 AI 생태계를 구축하기 위해서는 아직은 절대 강자로 군림하고 있는 엔비디아와의 협력을 피하기는 어렵다. 현재 대만에서 벌이고 있는 엔비디아의 다양한 생태계 구축 노력은 향후 우리나라 AI 생태계 구축을 위한 좋은 참고서가 될 것으로 판단된다. 대만 엔비디아 컨스텔레이션 설립 발표로 마무리한 젠슨 황 2025년 컴퓨텍스 키노트의 핵심 주제인 AI 팩토리를 주목해야 하는 이유다.

25) <https://www.korea.kr/multi/visualNewsView.do?newsId=148939865>

