

AWS re:Invent 2024, '클라우드와 생성형 AI' 결합 주목

이승민 정보통신기획평가원 동향분석팀 팀장

1. 머리말

아마존웹서비스(AWS) re:Invent 2024는 클라우드와 생성형 AI의 결합을 통한 디지털 전환 가속화를 제시하며 많은 주목을 받은 주요 행사로 평가된다. 2024년 13회를 맞이한 이 행사는 AWS가 클라우드 운영과 서비스 혁신을 위해 매년 개최하는 글로벌 컨퍼런스로, 작년 12월 2일부터 6일까지 라스베이거스에서 진행됐다. AWS는 행사에서 생성형 AI, AI 반도체, 클라우드의 최신 기술 트렌드를 발표하며 산업별 활용 사례와 미래 비전을 제시했다. 더불어, 주요 인사들의 기조연설(Keynote)과 다채로운 세션, 전시가 이어지며 클라우드와 AI의 융합 전략을 강조했다.

기조연설에선 AWS와 아마존의 주요 기술리더들이 등장해 생성형 AI, 자체 AI칩, 클라우드 관련 신규 서비스를 발표했다. 피터 디샌티스(Peter DeSantis) AWS 유틸리티컴퓨팅 수석 부사장, 매트 가먼(Matt Garman) AWS CEO, 앤디 제시(Andy Jassy) 아마존(Amazon) CEO 등이 참석해 AI 반도체, 생성형 AI 솔루션, 데이터 혁신 사례 등을 발표하며 이번 행사의 기술적 비전을 제시했다. 매트 가먼 수석부사장은 “기업의 모든 애플리케이션에 AI가 통합될 것”이라 전망하며 이를 위한 AWS의 사업 전략을 설명했고, 워너 보겔스(Werner Vogels) 아마존 CTO는 “단순함을 통한 혁신”을 핵심 키워드로 제시하며 2025년 기술발전 전망을 공유했다.

	기간	2024.12.2~6(미국, 현지 시간)
	장소	미국 라스베이거스 베네시안 호텔
	기조 연설	• 피터 디샌티스(AWS 서비스를 구동하는 기술혁신) • 매트 가먼(AI와 함께 하는 AI 혁신의 여정) • 앤디 제시(자체 AI 모델 아마존 노바 공개) • 스와미 시바수브라마니안(데이터와 AI의 융합) • 워너 보겔스(2025년 이후 기술 산업의 미래상 제시)
	주요 내용	• 생성형 AI 중심의 혁신(아마존 베드록, 아마존 세이지 메이커, 아마존 노바 등) • AI 인프라-데이터센터(트레이니엄3, Trn2 울트라서버 등)

출처 : AWS, 2024.12.2.

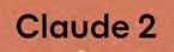
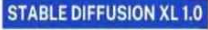
[그림 1] AWS re:Invent 2024 개요

2. AWS re:Invent 2024 주요 내용

2.1 주요 발표 기술 및 서비스

AWS는 생성형 AI 애플리케이션을 빠르고 쉽게 구축할 수 있도록 지원하는 완전 관리형 서비스 플랫폼 '아마존 베드록(Amazon Bedrock)'에 세 가지 신규 기능을 추가해 주목을 받았다. 특히, 아마존 자체 AI 모델 '아마존 타이탄(Amazon Titan)'을 비롯해 앤트로픽(Anthropic)의 '클로드(Claude)', 메타(Meta)의 '라마(Llama)', 코히어(Cohere)의 '커맨드(Command)', 미스트랄AI(Mistral AI)의 '미스트랄(Mistral)', 스테빌리티 AI(Stability AI)의 '스테이블 디퓨전(Stable Diffusion)' 등 다양한 파운데이션 모델을 동시에 제공하는데, 이는 의료·금융 서비스처럼 정확도가 중요한 분야에서 적합한 논리적 추론 기능을 바탕으로, 잘못된 정보 생성을 방지하는 특징을 가졌다. 또한, 여러 AI 에이전트를 활용한 다중 협업 환경을 지원해 복잡한 문제를 효율적으로 해결하고, 대규모 모델 지식을 소규모 모델로 이전해 비용과 응답 속도를 개선하도록 설계한 것이 돋보인다.

AWS는 베드록 플랫폼을 확장하기 위해 100개 이상 AI 모델 중 필요한 서비스를 선택할 수 있는 '베드록 마켓플레이스(Bedrock Marketplace)'를 공개해 개발자들이 특정 언어와 목적에 맞는 모델을 쉽게 활용할 수 있도록 했다. 이번에 루마AI(Luma AI)의 '레이2(Ray2)', 풀사이드(Poolside)의 '포인트(Point)'와 '말리부(Malibu)', 스테빌리티AI의 '스테이블 디퓨전 3.5 라지(Stable Diffusion 3.5 Large)' 등이 새롭게 추가됐고, 국내 기업인 업스테이지의 '솔라프로', '솔라미니', LG AI연구원의 '엑사원', 엔씨소프트의 '바르코' 등도 포함돼 다양한 선택지를 제공한다. 특히, 업스테이지의 '솔라프로'는 매개변수 220억 개를 갖춰 금융·법률·헬스케어 등 다양한 산업 분야에서 우수한 추론 능력을 제공한다. 이를 통해, 최신 AI 기술을 더욱 쉽게 도입할 수 있도록 지원하며, 베드록 플랫폼 확장을 바탕으로 AWS 생태계 내 생성형 AI 활용 폭이 크게 넓어질 것으로 기대된다.

				
				
텍스트 요약, 생성, 분류, 개방형 Q&A, 정보 추출, 임베딩 검색	스페인어, 프랑스어, 독일어, 포르투갈어, 이탈리아어 및 네덜란드어로 된 텍스트 생성을 위한 다국어 LLM	헌법적 AI 및 무해성 훈련을 기반으로 사려 깊은 대화, 콘텐츠 생성, 복잡한 추론, 창의성 및 코딩을 위한 LLM	비즈니스 애플리케이션을 위한 텍스트 생성 모델 및 100개 이상 언어로 검색, 클러스터링 또는 분류를 위한 임베딩 모델	독특하고 사실적인 고품질 이미지, 예술, 로고 및 디자인 생성

출처 : AWS, 2024.12.2.

[그림 2] AWS 베드록에서 지원하는 AI모델

AWS는 생성형 AI 기반 머신러닝(ML)을 지원하는 서비스인 '세이지메이커(SageMaker)'를 대대적으로 업그레이드했다. 이는 복잡한 인프라 없이도 ML 알고리즘 개발·학습·배포를 지원하는 도구로, 새롭게 추가된 '세이지메이커 유니파이드 스튜디오(SageMaker Unified Studio)'를 통해 조직 전체 데이터를 쉽게 탐색·접근 가능하도록 통합 솔루션을 제공한다. 이에 더해, '세이지메이커 레

이크하우스(SageMaker Lakehouse)'는 다양한 데이터 소스를 통합해 빠른 데이터 분석과 작업 수행을 지원함으로써 기업의 비즈니스 혁신을 가속화하는 동력이 되고 있다.

AWS는 업무 정보 검색, 인사이트 도출, 조치 실행을 위한 업계 최고 수준 생성형 AI 기반 어시스턴트인 '아마존 Q(Amazon Q)' 비즈니스의 새로운 기능과 개선 사항을 발표했다. 아마존 Q는 출시 이후 아마존 직원들의 업무 효율성을 높이는 데 기여하고 있으며, AWS 영업팀을 위해 10만 건 이상의 계정 요약을 생성하고, 기술 조사에 소요되는 시간을 45만 시간 이상 절감하는 성과를 보였다. 반복적인 작업을 자동화하기 위해 새롭게 공개된 '아마존 Q 앱스(Amazon Q Apps)'는 직원들의 업무 혁신을 지속적으로 지원하며, 이를 통해 생산성을 높이고 작업의 효율성을 강화했다. 또한, 아마존 Q 앱스는 구글 드라이브(Google Drive), 셰어포인트(SharePoint) 등 다양한 기업용 데이터 소스를 통합해 분석하고, 아마존 퀵사이트(Amazon QuickSight)와 연동함으로써 시각화된 데이터를 제공해 더욱 심층적인 분석과 사용자 맞춤형 서비스를 가능케 했다.

2.2 자체 AI 모델 및 하드웨어 개발

AWS는 자체적으로 개발한 새로운 생성형 AI 모델인 '아마존 노바(Amazon Nova)'를 공개하며 생성형 AI 시장에서의 입지를 더욱 강화했다. 아마존 노바는 AWS가 베드록 플랫폼에서 독점적으로 제공하는 멀티모달 모델로, 텍스트, 이미지, 비디오 입력을 처리하며 다양한 생성형 AI 작업에서 비용과 지연 시간을 획기적으로 줄이는 것을 목표로 한다. 노바는 텍스트 전용 '노바 마이크로(Nova Micro)'부터 '노바 라이트(Nova Lite)', '노바 프로(Nova Pro)', 2025년 1분기에 출시 예정인 '노바 프리미어(Nova Premier)', 그리고 이미지와 영상 생성에 특화된 '노바 캔버스(Nova Canvas)'와 '노바 릴(Nova Reel)'까지 총 6가지로 구성돼 있다. 이러한 모델들은 작업 내용과 특성에 따라 최적화된 옵션을 제공한다.

<표 1> AWS 새로운 생성형 AI 노바

노바 마이크로	노바 라이트	노바 프로	프리미어	캔버스	릴
텍스트. 빠른 응답과 비용 효율 강조	저렴한 멀티모달. 텍스트, 이미지, 동영상 이해·생성	가격·성능 균형 맞춘 멀티모달. 라이트보다 높은 정확도와 성능	최고 사양 멀티모달. 복잡한 추론, 사용자 맞춤 모델 교육에 최적화	전문가 수준의 이미지 생성. 유해한 콘텐츠 생성 제한가능	텍스트·이미지 기반으로 고품질 동영상 쉽게 생성

출처 : AWS, 2024.12.2.

AWS 발표에 따르면, 노바 라이트는 오픈AI(Open AI)의 GPT-4o 미니와 비교해 19개 벤치마크 중 17개에서, 구글(Google) 제미니(Gemini)와 비교해선 21개 벤치마크 중 17개에서 동등하거나 우수한 성능을 보여주며, 노바 프로는 20개 언어 이해와 추론 능력 면에서 GPT-4o와 동등하거나 뛰어나다는 평가를 받고 있다. 또한, 노바는 영어, 한국어, 일본어를 포함한 200개 이상의 언어를 지원한다. 특히, 비즈니스 환경에서 시간을 단축하고 비용을 줄일 수 있는 인텔리전스 기능과 함께 콘텐츠 생성 및 제공을 함으로써 AWS의 경쟁력이 한층 강화될 것으로 예상된다.

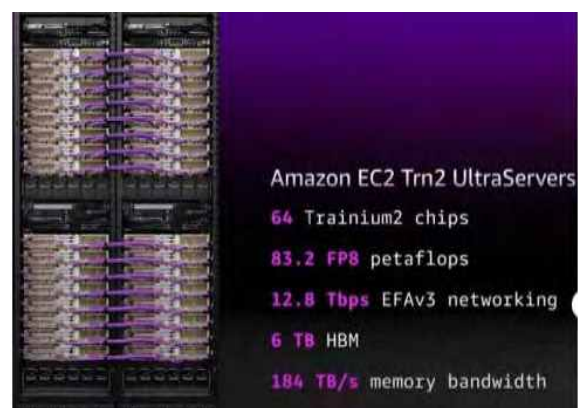
AWS는 AI 모델을 지원하는 자체 칩인 '트레이니엄(Trainium) 시리즈'를 통해 엔비디아(NVIDIA)에 도전장을 내밀고 있다. 특히 3나노 공정으로 제작된 트레이니엄3는 엔비디아 GPU에 필적하는 성

능으로 평가받고 있으며, 전작인 트레이니엄2 대비 성능이 2배 향상되고 에너지 효율성이 40% 개선될 것으로 기대된다. 이 칩은 2025년 말까지 개발 완료를 목표로 하고 있다. 트레이니엄2 칩은 AWS의 EC2 인스턴스와 새로운 데이터센터 서버인 Trn2 울트라서버(Trn2 UltraServer)에 적용됐으며, 엔트로픽 및 애플(Apple)과 협업해 초대형 AI 클러스터 구축과 AI 모델 훈련에 활용되고 있다. 단일 트레이니엄2 칩은 최대 1.3PF(PetaFlops, 1초당 1,000조 번 연산) 성능을 제공하며, 이를 탑재한 EC2 인스턴스는 20.8PF 성능을 발휘해 대규모 LLM(거대언어모델, Large Language Model)학습 및 배포에 적합하다. 또한, 기존 GPU 기반 서버 대비 30~40% 더 나은 가격 대비 성능을 제공하고 메모리 대역폭이 크게 개선돼 비용 효율성을 높였다.

Trn2 울트라서버는 64개 트레이니엄2 칩을 연결해 초고속 뉴런링크(NeuronLink) 기술로 최대 83.2PF 수준의 연산 성능을 확보하며, 기존 AI 서버 대비 5배 높은 컴퓨팅 용량과 10배 더 큰 메모리를 제공한다. 이러한 성능을 바탕으로, 엔비디아 블랙웰(Blackwell)칩 72개를 탑재한 서버와 경쟁이 가능할 것으로 예상된다. AWS는 자사의 칩 연결 기술을 통해 더 많은 칩을 효율적으로 연결할 수 있다는 점을 강조하고 있으며, 트레이니엄2를 활용한 학습 비용을 엔비디아 칩 대비 최대 40%까지 절감할 수 있다고 주장한다. 이러한 비용 절감은 데이터 센터 운영 기업에 큰 이점을 제공할 수 있다.



(가) 트레이니엄2



(나) Trn2 울트라서버

출처 : AWS, 2024.12.2.

[그림 3] Trn2 울트라서버

엔트로픽은 AWS와 함께 EC2 울트라클러스터를 구축 중이며, 이 클러스터는 수십만 개의 트레이니엄2 칩을 탑재해 엔트로픽의 AI모델 '클로드'의 학습 프로젝트에 활용될 예정이다. 애플 또한 트레이니엄2 칩을 사용해 AI모델을 훈련함으로써 최대 50% 효율성 향상을 기대하고 있다. 이와 같은 기술혁신을 통해 Trn2 울트라서버와 울트라클러스터는 2025년 중 가동될 예정이다. AWS는 안정적인 공급망을 기반으로 시장수요에 대응하고, 엔비디아 의존도를 줄이면서 독자적인 AI 인프라 역량을 강화하고자 하는 모습이다.

2.3 클라우드 인프라와 지속 가능성 강화

AWS는 오브젝트 스토리지 서비스인 '아마존 S3'와 관계형 데이터베이스 서비스(RDS)의 혁신을

통해 클라우드 인프라 성능과 효율성을 대폭 개선했다. '아마존 S3 테이블(Amazon S3 Table)'은 고성능 오픈소스 테이블 형식인 아파치 아이스버그(Apache Iceberg)를 도입해 초당 트랜잭션 효율을 10배 높이고 확장성을 강화함으로써 대규모 데이터 분석의 효율성을 극대화했다. 또한, '아마존 S3 메타데이터(Amazon S3 Metadata)' 기능은 데이터를 더욱 쉽고 빠르게 분류할 수 있도록 해 데이터 관리의 편의성을 증대시켰다.

관계형 데이터베이스 서비스인 '아마존 RDS(Relational Database Service)'는 클라우드 환경에서 관계형 데이터베이스 설정·운영·확장을 손쉽게 할 수 있도록 지원하며, 높은 가용성과 안정성을 제공한다. 특히, '아마존 오로라 DSQL(Amazon Aurora DSQL)'은 분산형 기술을 기반으로 99.99% 가용성을 유지하며, 구글의 클라우드 데이터베이스 서비스 '스패너(Spanner)'보다 4배 빠른 속도를 자랑해 데이터 처리 성능 면에서 차별화를 이루고 있다.

AWS는 에너지 효율을 극대화한 새로운 데이터 센터를 설계해 지속가능성과 성능을 모두 개선하고 있다. 이를 위해 전력, 냉각, 하드웨어 설계의 혁신을 결합해 새로운 데이터센터에 적용할 예정이다. 구체적으로, 고가용성을 위한 간소화된 전기 및 기계 설계, 냉각과 랙 설계의 개선, 제어 시스템의 혁신 등을 통해 기계 에너지 소비를 46% 감소시키고, 콘크리트 사용 시 체화 탄소를 35% 줄이는 등 에너지 효율성과 지속가능성 면에서 한 단계 더 발전했다.

이 데이터 센터 설계는 AWS가 운영 중인 34개 리전, 108개 가용 영역, AWS 로컬 존과 같은 글로벌 인프라 전반에 걸쳐 확장이 가능하도록 설계됐다. 이러한 혁신적 구성 요소를 갖춘 새로운 데이터 센터 건설은 2025년 초 미국에서 시작될 예정이다. AWS는 이와 같은 인프라 혁신을 통해 성능과 복원력이 뛰어나고, 안전하며, 지속가능한 클라우드 환경을 구축하고 있다. 이는 에너지 효율성을 높이고 새로운 워크로드를 유연하게 지원한다는 점에서 중요한 진전으로 평가받고 있다.

3. 맺음말

AWS re:Invent 2024는 클라우드를 넘어 생성형 AI로 전선을 확장하는 AWS의 여정을 보여준다. 이 행사를 통해 AWS는 클라우드 기업의 성장 모멘텀이 생성형 AI에 초점이 맞춰져 있음을 분명히 했다. 클라우드 서비스, AI칩, AI모델 등 다양한 비즈니스 영역에서의 확장을 시도하며, 엔비디아와 앤트로픽에 대한 의존도를 줄이고 AI 풀스택 최적화에 집중하고 있다. 이는 경쟁사인 마이크로소프트(Microsoft), 구글의 전략과 차별화된 모델을 구축하기 위한 노력으로 볼 수 있다.

2024년 마이크로소프트 빌드, 애플 WWDC, 구글 I/O 등 주요 빅테크 행사들은 생성형 AI 기술 트렌드와 비즈니스 전략을 중심으로 진행됐다. AWS re:Invent 2024 역시 생성형 AI가 클라우드 업계에서도 차세대 성장 동력으로 자리 잡고 있음을 확인시켜 줬다. 이러한 변화 속에서, 한국 정부와 기업은 생성형 AI를 기반으로 한 지속가능한 수익 모델을 개발하고, 이를 실제 비즈니스에 적용하며 효과를 입증하는 노력이 필요하다.

전 세계적으로 생성형 AI 주도권 경쟁이 치열해지면서 클라우드 내에서 생성형 AI의 확산도 가속화되고 있다. 이에 따라, AI모델의 현장 적용 성과와 비즈니스 모델을 결합해 통합 생태계를 구축하는 기업이 경쟁 우위를 차지할 것으로 보이며, 국내 기업들도 생성형 AI 특화 모델 개발 및 클라우드와의 연계 전략을 고도화해야 할 시점이다. AI 하드웨어, 데이터 센터, 통합 플랫폼을 포괄

하는 AI 종합 솔루션으로 발전해 나가는 글로벌 흐름에 대응하기 위해, 정부와 기업은 기술개발과 운영 전략을 강화해 이러한 변화에 적극 대비해야 한다.

[참고문헌]

- [1] AWS reinvent, 'See you at re:Invent 2025' (2024.12.1.) : <https://reinvent.awsevents.com/>
- [2] Analyticsvidhya, '7 Major AI Updates from AWS re:Invent 2024 that You Can't MISS' (2024.12.4.) : <https://www.analyticsvidhya.com/blog/2024/12/reinvent/>
- [3] NDS, 'Amazon BedRock이란?' (2024.10.18.) : <https://tech.cloud.nongshim.co.kr/blog/aws/2311/>
- [4] 삼성증권, '아마존, AWS re:Invent 2024, 아마존이 달라졌다' (2024.12.4.) : https://www.samsungpop.com/common.do?cmd=down&contentType=application/pdf&inlineYn=Y&saveKey=research.pdf&fileName=3010/2024120415502043K_02_03.pdf
- [5] 이코노믹리뷰, '구름 위 현자의 2025년 전망 "SMR부터 디지털의 재발견" [AWS 리인벤트 2024]' (2024.12.6.) : <https://www.econovill.com/news/articleView.html?idxno=675923>
- [6] 디지털데일리, '[re:Invent2024] 맷 가먼의 첫번째 메시지는 'AI칩'...애플 깜짝등장 이유는?' (2024.12.4.) : <https://www.ddaily.co.kr/page/view/2024120408283316850>
- [7] 디지털데일리, '[re:Invent2024] AI 백화점으로 오세요, AWS의 '타임플레이'는 계속된다(종합)' (2024.12.4.) : <https://www.ddaily.co.kr/page/view/2024120715221037260>
- [8] 브런치스토리, '아마존 노바(Amazon Nova) 출시 의의와 시사' (2024.12.4.) : <https://brunch.co.kr/@mentats1/161>
- [9] 헤럴드경제, '돈보다 의미 따지는 20대 온다' 아마존 CTO가 본 미래 예측' (2024.12.6.) : <https://biz.heraldcorp.com/article/10009692?ref=naver>

※ 출처: TTA 저널 제217호