

AI를 통한 기업 역량 강화: 현실로 다가오는 선택의 순간



목차

- AI로 산업을 혁신할 기회 1
- 산업 분야의 AI 4
- IT 의사 결정권자가 고려해야 할 사항 5
- 시작하기: AI에 대한 이해 5
- 중요한 선택..... 6
 - 성능 6
 - 데이터 보안 6
- 솔루션 규모 조절 7
 - 비용과 혁신의 균형 유지 7
 - 단순성과 유연성 7
 - 설명 가능성 보장 7
- 실제 시나리오..... 8
- 소매 8
- 의료..... 9
- Dell Technologies 솔루션 10
- 모두를 위한 AI: Dell과 AMD의 AI 민주화 10
- Hugging Face 협업..... 11
- AMD EPYC™ 프로세서..... 11
- AMD Instinct™ MI300X 가속기 11
- AMD ROCm™ 6 오픈 소스 소프트웨어 플랫폼 12
- Dell PowerEdge™ 서버 포트폴리오 12
- 요약..... 13

AI로 산업을 혁신할 기회

오늘날 산업계는 AI 덕분에 미래에 대비해 비즈니스를 혁신할 절호의 기회를 맞이했습니다. Accenture Vision Technology 2023에서 수집한 데이터에 따르면 전 세계 경영진의 98%가 향후 3~5년 이내에 AI 파운데이션 모델이 조직의 전략에서 중요한 역할을 할 것으로 봅니다.¹

AI는 업무의 효율성을 높이고 혁신을 지원하며 의사 결정 프로세스를 개선해 준다는 점에서 소매, 의료, 금융 서비스 등 비즈니스 부문에서 매우 유용해졌습니다. 이러한 장점이 있지만, 흔히 있는 몇 가지 오해 때문에 AI를 활용하는 데는 여전히 진입 장벽이 있습니다.



시작하려면 AI 개발자 팀이 필요합니다.

고급 AI 솔루션을 개발하고 기본 원칙을 이해하는 데 여전히 데이터 사이언스 분야에 대한 전문 지식이 필요하지만, 더 이상 필수 요구 사항은 아닙니다. 사용자 친화적인 AI 툴, Hugging Face와 같은 플랫폼, 그리고 AI 솔루션 개발과 관련된 복잡성을 상당 부분 해소하는 작업별 모델이 널리 사용되고 있습니다.

성과를 얻으려면 하드웨어에 수천만 달러를 투자해야 합니다.

이러한 오해는 오늘날 이용 가능한 AI 리소스의 다양성을 심각하게 훼손합니다. 널리 알려진 리소스가 강력하고 지원이 잘 되는 경우도 많지만, 언제나 모든 기업에 가장 적합하거나 비용 효율적인 선택이 되는 것은 아닙니다.

가속기를 얻기 위해 다음과 같이 끊임없이 노력해야 합니다.

가속기는 과중한 AI 워크로드를 처리하는 데 탁월하지만, AI 애플리케이션에 그만큼의 컴퓨팅 능력을 필요로 하지 않는 기업들도 있습니다. 시장을 선도하는 가속기를 이용하기 위해 지나치게 오랜 시간을 기다리는 것도 현실적이지 않습니다. 많은 경우에 AI 최적화 CPU는 AI 지원 분석 및 의사 결정을 실시간으로 생성하는 데 필요한 성능과 효율성을 실제로 제공할 뿐만 아니라 훨씬 더 비용 효율적이고 적응력이 뛰어난 솔루션입니다.

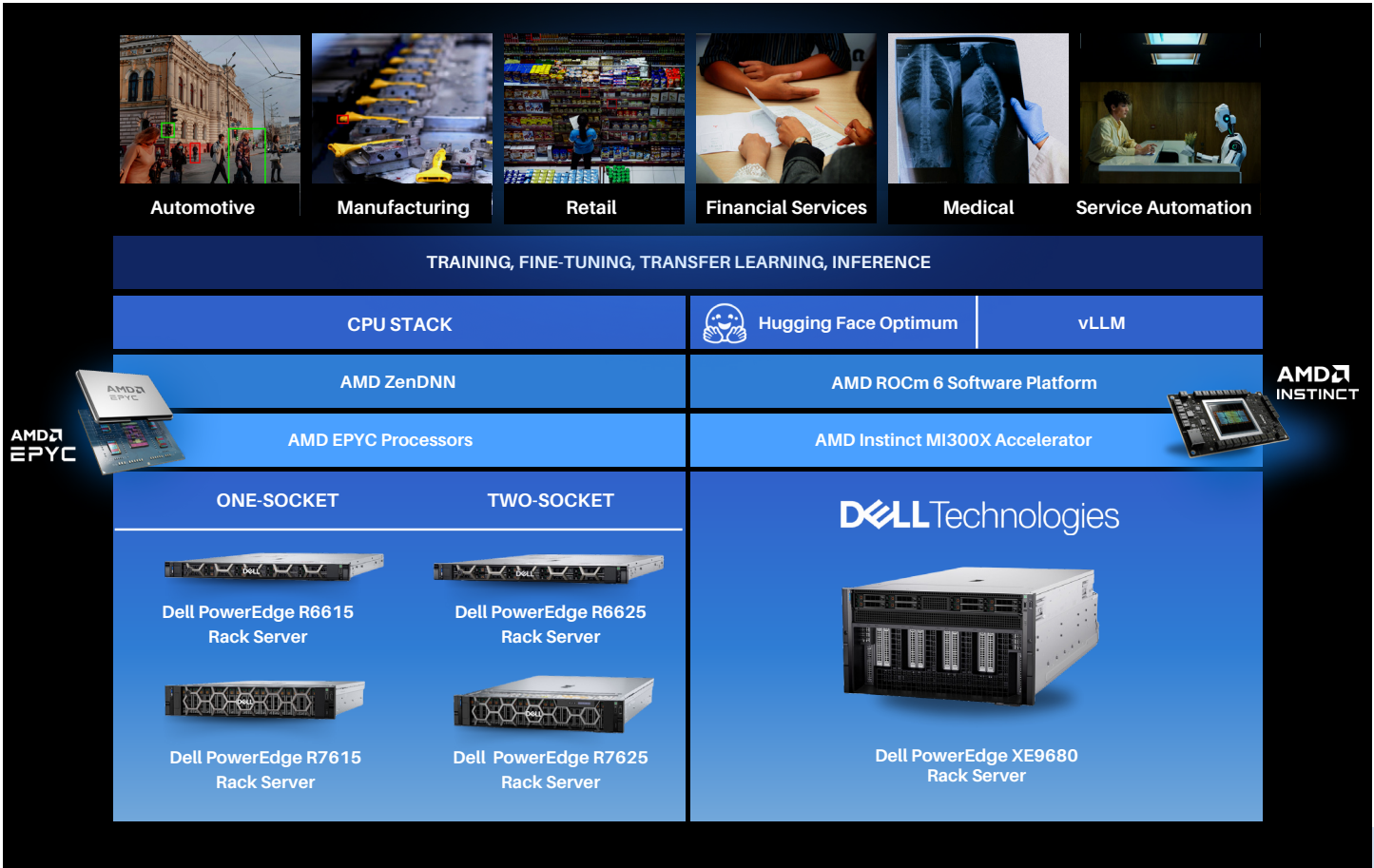
¹ Accenture, 2023년 3월 30일, "Accenture Technology Vision 2023: Generative AI to Usher in a Bold New Future for Business, Merging Physical and Digital Worlds", <https://newsroom.accenture.com/news/2023/accenture-technology-vision-2023-generative-ai-to-usher-in-a-bold-new-future-for-business-merging-physical-and-digital-worlds>



다행히 AI 환경은 진화하고 있습니다. **Dell**과 **AMD**는 오늘날의 AI 수요를 지원하도록 설계된 포괄적인 인프라스트럭처를 통해 다양한 사용자들에게 AI 기술과 툴을 제공하기 위해 함께 협력하면서 이러한 오해를 불식시키고 있습니다.

이미 최적화되어 있는 모델, 신뢰할 수 있는 소프트웨어 스택, 다용도 하드웨어 시스템을 활용하여 시작할 수 있으며, 이 모든 것은 Dell과 AMD의 파트너십을 통해 쉽게 이용할 수 있습니다. 갈수록 부족해지는 가속기, 상당한 규모의 숙련된 AI 엔지니어 그룹, 또는 대규모 클라우드 클러스터를 구축하기 위한 리소스에 대한 접근성은 더 이상 AI 활용의 필수 요구 사항이 아닙니다.

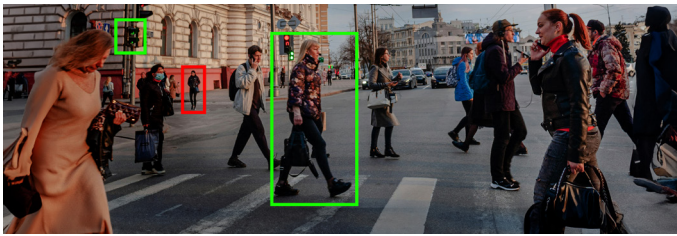
Dell과 AMD의 협업은 개발자가 전이 학습, 미세 조정, 추론을 쉽고 효율적으로 통합하는 포괄적인 AI 솔루션을 만들 수 있도록 설계된 하드웨어와 소프트웨어의 통합 생태계를 제공합니다. **Hugging Face**의 지원으로 이제 AMD EPYC™ 프로세서 또는 AMD Instinct™ MI300X 가속기가 탑재된 Dell PowerEdge 서버에서 실행되는 모델 포트폴리오를 확장하게 되었습니다. 덕분에 개발자가 미세 조정을 수행하고 전이 학습을 적용하며 추론을 위해 배포할 수 있습니다. AMD ROCm™ 및 AMD ZenDNN™에 대한 투자와 PyTorch, Tensorflow, ONNX Runtime 프레임워크와의 파트너십은 응용 AI 개발자가 AI 민주화를 경험하기 위한 필수적인 기반입니다. 아래의 스택 다이어그램은 Dell과 AMD의 통합 AI 생태계를 구성하는 구성 요소를 자세히 보여줍니다.



산업 분야의 AI

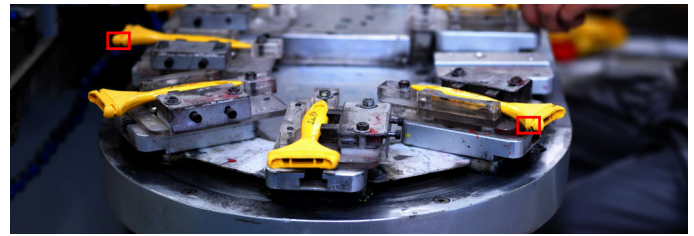
AI는 리소스의 다양화와 오픈 소스 혁신에 중점을 두고 고객 서비스, 금융 및 बैं킹, 의료, 소매 등 다양한 산업으로 마이그레이션되고 있습니다. 이러한 산업 전반에서 AI는 조직이 데이터 분석, 자동화, 개인화, 예측 분석 등의 주요 기능을 활용하여 독점 데이터의 잠재력을 발굴하고 AI 워크플로를 혁신할 수 있도록 종합적으로 지원합니다. AMD ROCm 및 ZenDNN 라이브러리는 이러한 AI 워크플로를 가속하여 거의 실시간으로 결과를 제공합니다.

AI가 다양한 산업에 정확히 어떤 영향을 미치는지 아래에서 자세히 살펴보십시오.



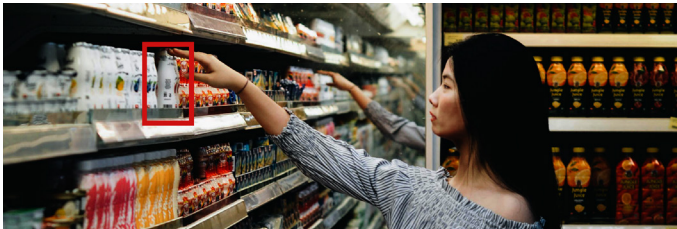
자동차

AI는 자율 주행 차량의 물체 감지, 차선 추적, 의사 결정에 활용됩니다. 또한 AI는 차량 부품에서 고장이 발생할 가능성을 예측하여 사전 예방적 유지 보수를 수행하고 다운타임을 줄일 수 있게 해줍니다.



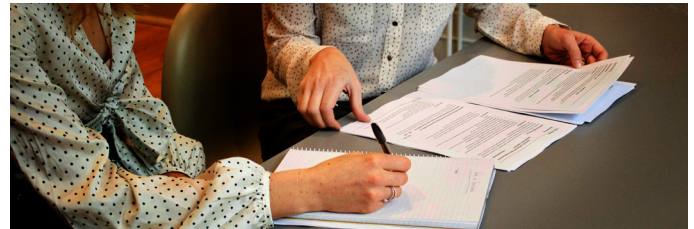
제조 및 산업

제조 및 산업 부문에서 AI는 예측 유지 보수, 품질 관리, 공정 최적화 및 공급망 관리에 활용되어 효율성을 높이고 다운타임을 줄일 수 있습니다.



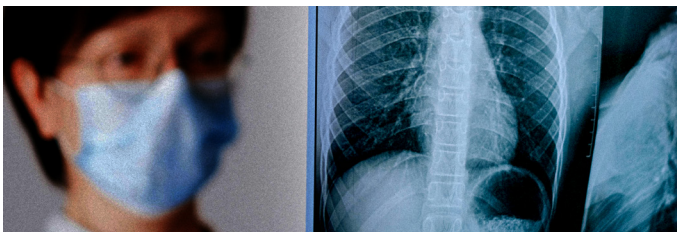
소매

AI는 고객 행동을 분석하여 개인화된 제품을 추천함으로써 고객 참여와 영업을 개선할 수 있습니다. 또한 수요를 예측하고 재고 과잉 또는 재고 소진을 최소화하여 재고 수준을 최적화할 수 있습니다.



금융 서비스

금융 및 बैं킹 부문에서 AI는 사기 탐지, 위험 진단, 고객 서비스 및 투자 분석에 사용되어 보안을 강화하고 더욱 합리적인 의사 결정을 내릴 수 있게 해줍니다.



의료 직원

의료 분야에서 AI는 의료 영상 분석, 질병 진단, 맞춤형 치료 계획, 약물 발견 등 다양한 용도로 사용되어 환자의 치료 예후를 개선하고 비용을 절감할 수 있습니다.



서비스 자동화

AI 기반 챗봇은 고객 문의를 처리하고 지원을 제공함으로써 인간이 개입해야 할 필요성을 줄일 수 있습니다. 또한 AI는 데이터 입력 또는 문서 처리와 같은 반복적인 작업을 자동화하여 효율성을 높이고 오류를 줄일 수 있습니다.

IT 의사 결정권자가 고려해야 할 사항

시작하기: AI에 대한 이해

활용 사례를 확인하기 전에 먼저 AI 수명주기를 자세히 살펴보겠습니다. AI(Artificial Intelligence) 수명주기는 AI 시스템의 개발, 배포 및 유지 보수와 관련된 단계를 말합니다. 구체적인 방법론과 용어는 다를 수 있지만, 일반적인 AI 수명주기에는 항상 모델 훈련과 추론이 포함됩니다.

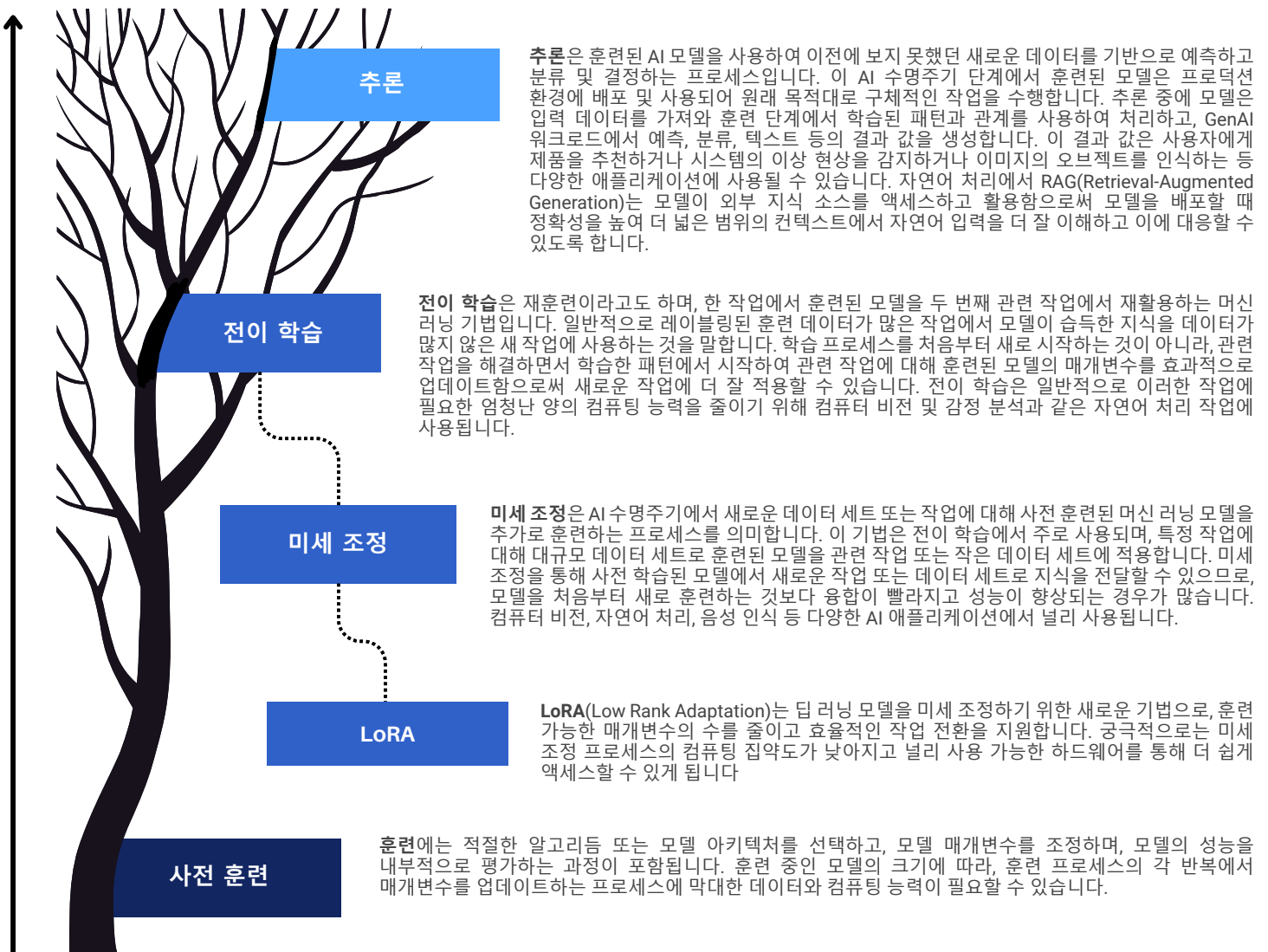


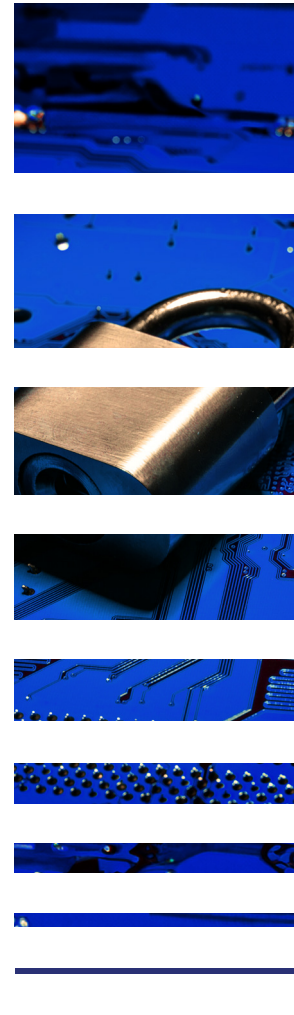
그림 1: AI 수명주기



중요한 선택

| 성능

많은 실제 애플리케이션에서 실시간 또는 거의 실시간에 가까운 의사결정은 성공의 중요한 조건이 됩니다. 예를 들어 금융 거래 또는 보험 청구에서 금전적 손실을 방지하고 기업 자산을 보호하기 위해서는 사기 행위를 신속하게 식별해야 합니다. 제조 시나리오에서는 품질 보증을 위해 조립 라인 또는 공장 조건의 결함을 동적으로 모니터링해야 합니다. 추론 워크로드를 처리하는 프로세서는 들어오는 데이터 스트림을 빠르고 효율적으로 처리할 수 있도록 최적화되어야 합니다. AMD EPYC 프로세서가 탑재된 Dell PowerEdge 서버는 고성능 컴퓨팅, 클라우드 컴퓨팅 및 빅데이터 분석과 관련한 작업뿐만 아니라 엣지 추론 워크로드를 처리하는 데에도 적합한 다기능 시스템입니다.



| 데이터 보안

데이터 보안은 AI 시스템, 특히 생성형 AI를 활용하는 시스템의 성공에 중요하며, AI를 운영에 통합하려는 기술 리더의 주요 관심사입니다. AI 시스템은 일반적으로 대량의 데이터에 의존하며, 여기에는 개인 정보, 재무 데이터 또는 독점 정보와 같은 민감한 기밀 정보가 포함될 수 있습니다. 이러한 데이터를 보호하는 것은 무단 액세스나 데이터 도용을 방지하고 AI 모델 및 예측의 정확성, 신뢰성, 일관성을 보장하는 데 매우 중요합니다.

기밀 컴퓨팅은 안전한 환경에서 데이터 처리를 지원하여 클라우드 공급업체 및 기타 사용자를 비롯한 권한이 없는 당사자의 무단 액세스 또는 조작으로부터 데이터를 보호하는 기술입니다.² 처리 중에 데이터를 격리하는 데 암호화 및 기타 보안 조치가 사용됩니다. AMD EPYC 프로세서에 통합된 정교한 보안 기능 모음인 AMD Infinity Guard는 프로세서에만 알려진 키를 사용하여 VM(Virtual Machine)을 암호화하는 SEV(Secure Encrypted Virtualization)를 통해 기밀 컴퓨팅을 지원합니다. 이러한 서비스는 AMD SEV-SNP(SEV-Secure Nested Paging)를 사용하여 하드웨어 기반의 신뢰할 수 있는 실행 환경을 제공하는 것을 목표로 하며, 게스트 보호를 강화하여 외부 위협으로부터 보호합니다.

페더레이션 학습은 데이터 보안을 유지하는 또 하나의 방법입니다. 분산된 디바이스 또는 서버 전반에 걸쳐 중앙의 모델을 훈련합니다.³ 모든 데이터를 중앙 위치로 전송하는 것이 아니라 각 디바이스가 모델을 로컬로 훈련하고 모델 업데이트만 공유합니다. 이 접근 방식은 프라이버시를 보호하며 원시 데이터를 공유하지 않고도 협업 학습을 가능하게 합니다. Dell Technologies의 페더레이션 AI 플랫폼은 데이터 세트가 수집되는 동안 네트워크 엣지의 데이터 세트에 대해 컴퓨팅 프로세스, AI 및 ML 알고리즘이 실행되도록 합니다. 따라서 수학적 모델, 메타데이터 및 쿼리 결과만 네트워크를 통해 다른 엣지 디바이스, 데이터 센터 또는 클라우드에 공유됩니다. 이러한 교환을 통해 데이터와 지적 재산을 공개하지 않고도 대규모 분산 데이터 세트에서 유용한 통찰력을 실시간에 가깝게 도출할 수 있으므로 결과가 향상됩니다.

² Advanced Micro Devices, Inc. 2023년 8월 30일, "AMD shares the technical details of technology Powering Innovative Confidential Computing Leadership Cloud Offerings", <https://www.AMD.com/en/newsroom/press-releases/2023-8-30-AMD-shares-the-technical-details-of-technology-pow.html>
Advanced Micro Devices, Inc., 2021년, "Data Center Solutions, Confidential Computing" 솔루션 브리프, <https://www.AMD.com/content/dam/AMD/en/documents/EPYC-business-docs/solution-briefs/confidential-computing-solution-brief.pdf>

³ Analytics Vidhya, 2023년 12월, "Federated Learning: A Beginner's Guide", <https://www.analyticsvidhya.com/blog/2021/05/federated-learning-a-beginners-guide/#:~:text=Federated%20learning%20works%20by%20training,learning%20without%20sharing%20raw%20data>
Dell Technologies, 2021년, "A federated learning platform for real-time artificial intelligence" 솔루션 브리프, <https://www.Delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/dt-sb-analytics-anywhere.pdf>

솔루션 규모 조절

| 비용과 혁신의 균형 유지

비용과 혁신 사이의 적절한 균형을 유지하면, 재정적으로 실현 가능할 뿐만 아니라 영향력도 있으며 기업과 사용자 모두에게 진정한 가치를 창출하는 AI 솔루션을 보장할 수 있습니다. 이러한 균형을 찾기 위한 핵심 구성 요소는 활용 사례를 해결하고 기존 인프라스트럭처에 쉽게 통합할 수 있는 하드웨어를 찾는 것입니다. 최신 AI 하드웨어 시장에서는 생산 능력의 제약, 물류 문제, 반도체 부족 문제에 더해 다양한 산업에서 가속기에 대한 수요가 증가하는 문제로 인해 가속기 부족 현상이 더욱 가중되고 있습니다.

반면, CPU는 이미 대부분의 데이터 센터에서 표준 구성 요소로 사용되고 있으므로 완전히 새로운 가속기 하드웨어를 추가하는 것과 비교했을 때, 통합이 더 간편하고 비용 효율적입니다. AI에 최적화된 CPU는 기존 소프트웨어와 툴을 활용할 수 있으므로 광범위한 재정부 또는 재훈련의 필요성이 줄어듭니다. 또한 CPU는 AI 외에도 다양한 작업에서 뛰어난 유연성과 효율성을 제공하므로, 데이터 센터 내에서 리소스를 더욱 다양한 목적에 사용할 수 있습니다. AMD EPYC 프로세서를 실행하는 Dell PowerEdge 서버로 데이터 센터를 교체하면 기존 워크로드를 충족하는 동시에 AI 기반으로 혁신을 실현하고 효율성을 개선할 수 있습니다.

| 단순성과 유연성

AI 시스템의 단순성과 유연성은 장기적으로 효과적이고 적응력이 뛰어나며 확장 가능한 AI 솔루션을 구축하는 데 필수적입니다. 하드웨어를 보완하는 소프트웨어 프레임워크 및 최적화 제품군을 이용할 수 있다면 교차 플랫폼 통합에 추가로 시간과 노력을 들이지 않고도 성능을 높일 수 있습니다. 이러한 특성은 훈련, 추론 및 데이터 처리와 같은 다양한 유형의 AI 작업이 결합된 혼합 AI 워크로드를 지원하는 데 특히 중요합니다.

AMD와 Dell Technologies는 하드웨어와 소프트웨어 솔루션의 조합을 통해 혼합 AI 워크로드를 처리합니다. AMD EPYC 프로세서는 SMT(Simultaneous Multithreading) 및 많은 코어 수와 같은 특성을 바탕으로 뛰어난 컴퓨팅 성능을 제공하므로, AI 워크로드를 위한 효율적인 병렬 프로세싱이 가능합니다. 이러한 프로세서는 AI 작업에 최적화되어 훈련 및 추론 워크로드 모두에서 강력한 성능을 제공합니다. AMD EPYC 프로세서가 장착된 Dell PowerEdge 서버는 AI 워크로드를 배포하기 위한 확장 가능하고 유연한 플랫폼을 제공합니다. 또한 Dell OpenManage 소프트웨어 제품군은 혼합 AI 워크로드에 대한 리소스 할당 및 성능 모니터링을 최적화하는 관리 툴을 제공합니다.

아울러 AMD는 성능이 향상된 최신 소프트웨어 스택 버전을 활용하고 AMD EPYC 프로세서용 AMD ZenDNN 라이브러리, AMD Instinct Accelerator용 오픈 소스 AMD ROCm 스택, AMD 적응형 SoC용 소프트웨어 스택을 이용하는 UIF(Unified Inference Frontend)를 제공합니다. 또한 AMD ROCm은 전문가용 및 소비자용 제품을 비롯하여 다양한 AMD CPU 및 가속기와 함께 작동하도록 설계되었습니다.

| 설명 가능성 보장

설명 가능한 AI(Explainable AI)는 인공지능 애플리케이션에서 투명성, 신뢰성 및 효율성을 보장하는 데 핵심적인 역할을 합니다. 설명 가능한 AI는 AI 모델이 어떻게 의사 결정을 내리는지에 대한 통찰력을 제공하여 기반 요인과 추론 프로세스를 보여줍니다. 이러한 투명성은 이해 관계자의 신뢰를 얻는 데 매우 중요하며, 의사 결정이 개인의 삶에 직접적으로 영향을 미치는 의료, 금융 및 형사 사법과 같은 민감한 영역에서 특히 중요합니다.

휴먼 인 더 루프(Human-in-the-loop) AI 시스템은 인간 지능을 활용하여 AI 성능을 높이고 알고리즘 편향을 완화합니다. 이러한 시스템에는 인간의 감시가 통합되어 있기 때문에 복잡하고 모호한 상황을 더욱 효과적으로 처리할 수 있으며, AI 솔루션이 윤리적/사회적 규범에 부합하도록 할 수 있습니다. 또한 인간의 참여 덕분에 실제 피드백을 기반으로 AI 모델의 지속적인 개선과 적응이 이루어지므로, 반복적인 개선과 장기적인 신뢰성이 촉진됩니다. 이러한 접근법은 최선의 공익에 봉사하는 책임 있고 포괄적인 AI 시스템을 구축하는 데 필수적입니다.

실제 시나리오

Scalers AI는 Dell, AMD와 협력하여 AMD 프로세서가 탑재된 Dell PowerEdge 서버의 기능을 선보였습니다. 이러한 기술이 소매와 의료 시나리오에서 훈련, 전이 학습, 추론에 어떻게 활용되는지 알아보십시오.

소매

Scalers AI는 오브젝트 감지 AI 모델을 구현하여 소매 진열대의 재고 수준을 모니터링하고 관리하도록 설계된 Retail Inventory Management Reference Solution이라는 시스템을 구축했습니다. 이 레퍼런스 솔루션은 SSD_MobileNet_V2 모델을 활용하여 매장 진열대의 제품을 식별하고 인식함으로써, 궁극적으로 재고 수량을 자동으로 계산하고 재고 수준을 정밀하게 모니터링할 수 있습니다. 이 모델은 23,000개 이미지로 구성된 Roboflow의 SKU110K 이미지 데이터 세트를 사용하여 전이 학습을 거쳤습니다. 이 시스템은 컴퓨터 비전 및 머신 러닝 알고리즘을 활용하여 물품의 재고가 부족하거나 소진된 상황을 감지하고 매장 직원에게 알림으로써 적시에 재입고하거나 보충할 수 있도록 합니다.

이 솔루션은 AMD EPYC 9354P 32코어 프로세서가 탑재된 Dell PowerEdge R7615 서버를 사용합니다.

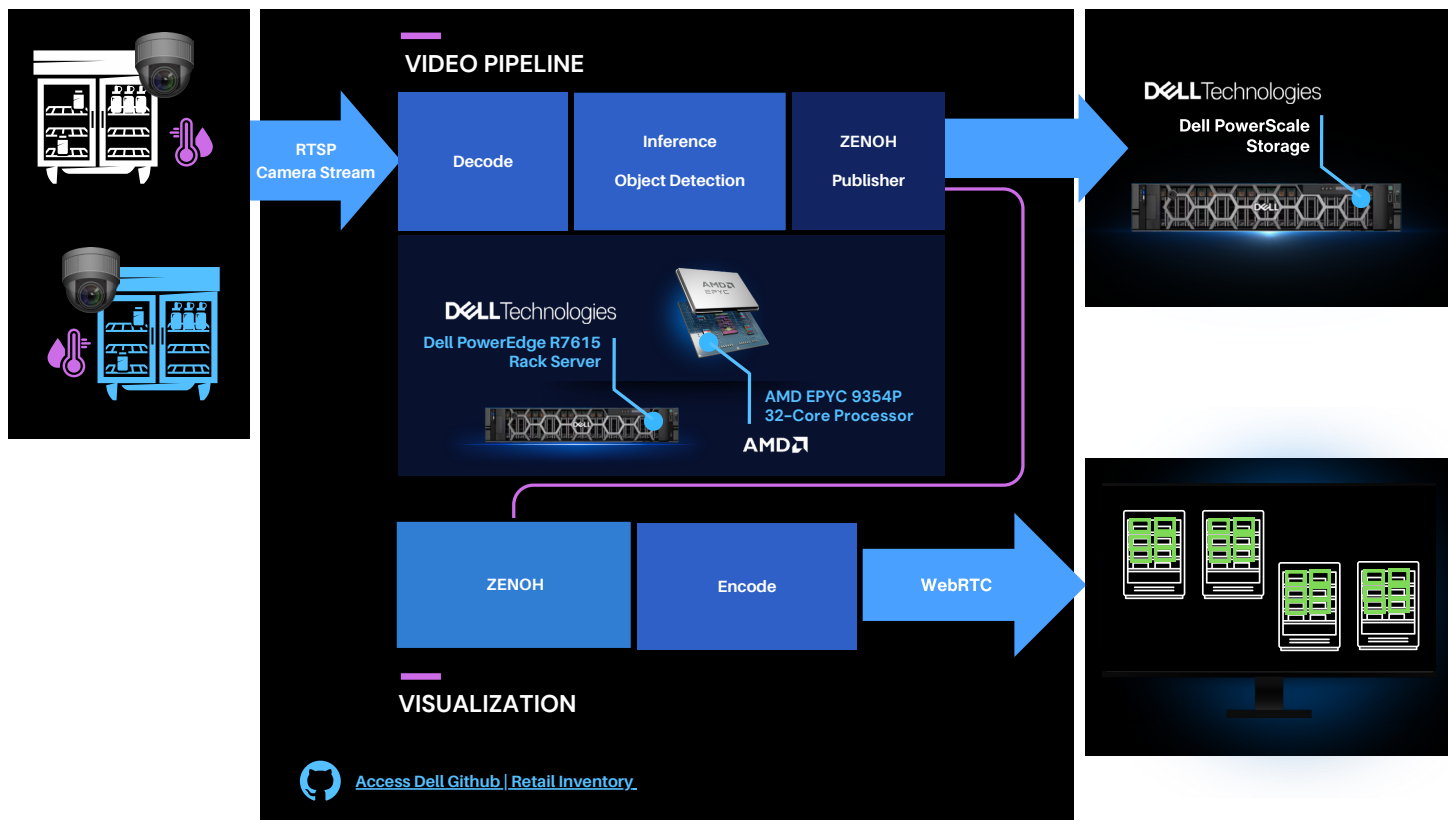


그림 2: Retail Inventory Management Reference Solution의 아키텍처 다이어그램

의료

AI 기반 의료 이미징은 진단의 정확성과 효율성을 개선하고 의료 전문가가 육안으로 포착하기 어려울 수 있는 증상에 대한 정확한 통찰력을 제공함으로써 의료 서비스를 개선하는 매우 가치 있는 기능을 제공합니다. AI는 의료 이미지 분석을 자동화하여 진단에 걸리는 시간을 단축합니다. 덕분에 더욱 신속한 치료 결정을 내릴 수 있고, 그 결과 환자 예후가 개선됩니다.

Scalers AI는 AMD EPYC 9554 64 코어 프로세서가 탑재된 Dell PowerEdge R7625 서버의 기능을 활용하여 폐렴 감지를 위한 AI 기반 의료 이미징 솔루션을 만들었습니다. 이 솔루션은 고급 알고리즘과 머신 러닝 기술을 사용하여 X선 또는 CT 스캔과 같은 의료 영상을 분석함으로써 환자의 폐렴 진단 속도와 정확도를 높이는 데 도움을 줍니다. 궁극적으로 컴퓨터 지원 검토라는 추가적인 계층을 도입함으로써 의료 전문가가 대량의 영상 데이터를 더욱 효율적으로 처리할 수 있도록 합니다.

이 레퍼런스 솔루션은 ResNet50 모델을 사용하여 NIH Clinical Center 데이터 세트에서 얻은 흉부 X선 영상을 분석합니다. 폐렴의 유무를 진단하는 것이 주요 목적이며, 기본적으로 이원성 분류를 수행합니다. 이 모델은 NIH Clinical Center 데이터 세트의 X선 DICOM 데이터 세트를 사용하여 훈련되었으며 ResNet50 아키텍처를 통한 전이 학습을 거쳤습니다.

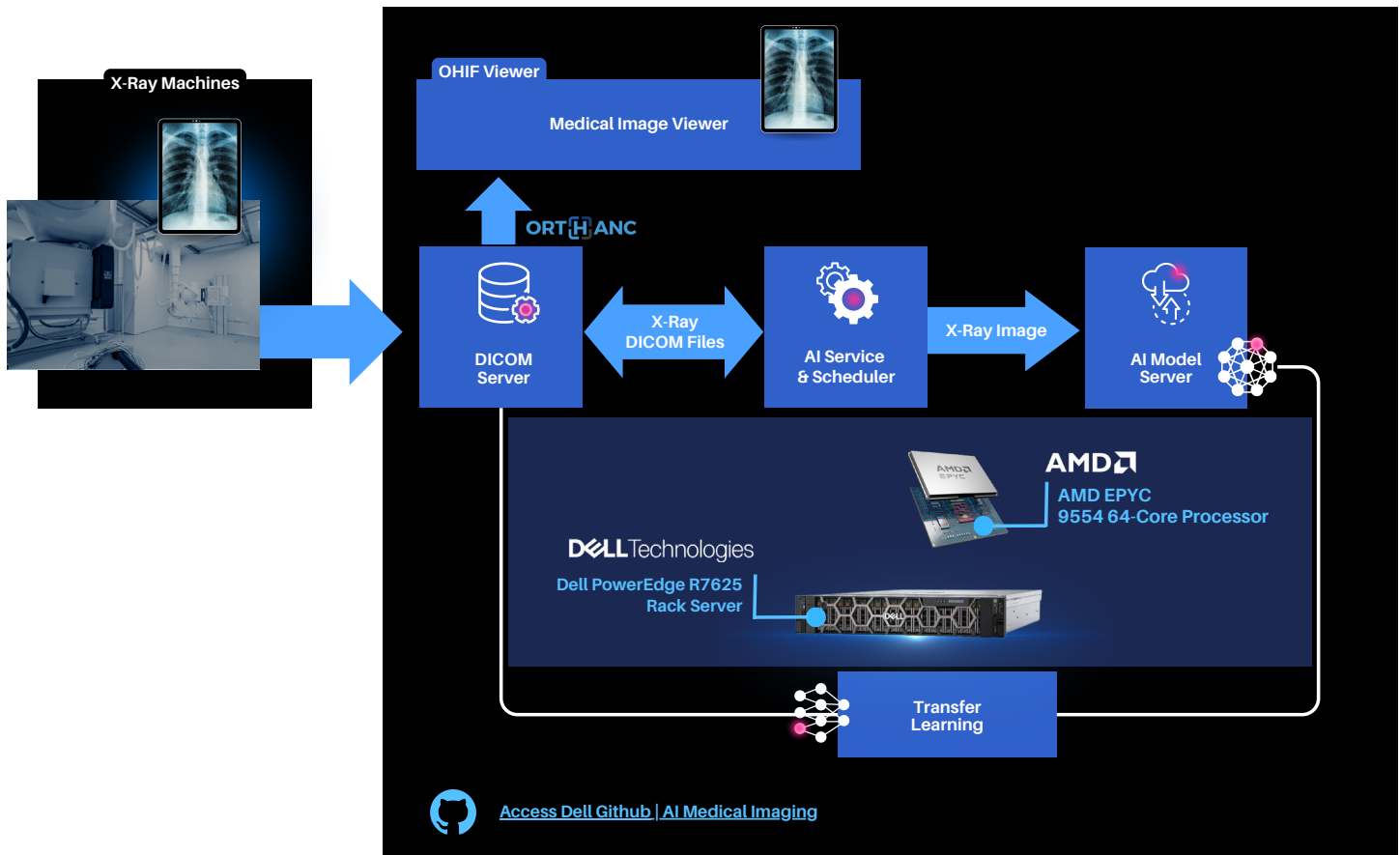
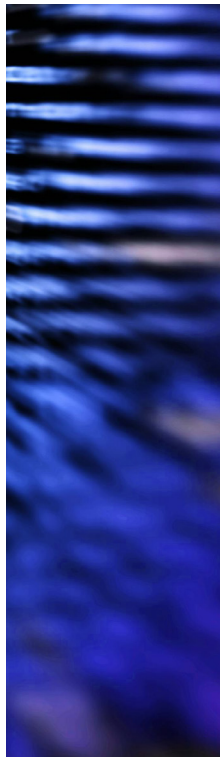
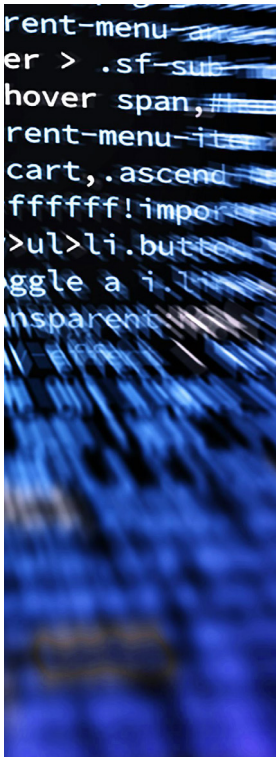


그림 3: 의료 AI 이미징 솔루션의 아키텍처 다이어그램

Dell Technologies 솔루션

모두를 위한 AI: DELL과 AMD의 AI 민주화

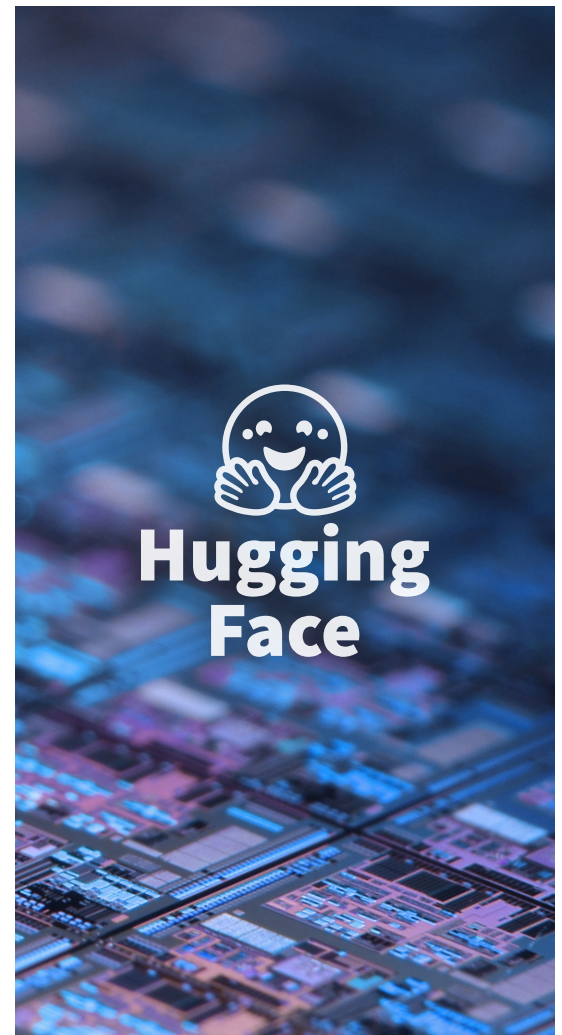
이러한 협업은 AI 생태계에서 혁신을 촉진하고 포용성을 증진하는 데 필수적인 AI 민주화를 위한 토대를 구축합니다. Dell과 AMD는 개인과 조직이 첨단 AMD CPU 및 가속기 기술을 갖춘 강력한 서버 제품군을 통해 AI를 활용하고 각 분야의 고유한 과제를 해결할 수 있도록 지원함으로써 이러한 성과를 실현하고 있습니다. AMD Instinct MI300X 가속기가 탑재된 Dell PowerEdge 서버는 LLM(Large Language Models) 훈련 및 미세 조정과 같은 대규모 AI 워크로드를 처리할 수 있으며, AMD EPYC 프로세서가 탑재된 Dell PowerEdge 서버는 엣지 추론 워크로드를 처리하는 데 탁월합니다. 기본 하드웨어 플랫폼 외에도 AMD는 AMD CPU에서 딥 러닝 추론을 최적화하기 위한 ZenDNN 소프트웨어 라이브러리와 AMD ROCm 소프트웨어 라이브러리를 제공하여 AMD Instinct Accelerator의 훈련, 미세 조정, 추론 기능을 개선합니다. 이 모든 옵션은 AMD의 UIF(Unified Inferencing Model)에서 원활하게 연동되므로 사용자는 소프트웨어 프레임워크, 소프트웨어 최적화 및 하드웨어 플랫폼을 원하는 대로 선택하면서 포괄적인 AI 솔루션을 구성할 수 있습니다.



HUGGING FACE 협업

AI를 도입하려는 기업은 데이터 사이언스 및 머신 러닝 전용 오픈 소스 플랫폼인 Hugging Face에서 기존 모델 또는 구체적인 요구 사항에 맞게 조정된 AI 워크플로를 직접 활용하는 것으로 시작할 수 있습니다. AMD는 이미 AMD 플랫폼과 원활하게 통합되어 있는 소프트웨어 라이브러리와 프레임워크에 AMD 전용 소프트웨어 최적화 기술을 추가하여 최고 수준의 혁신적인 성능을 제공한다는 공통의 목표를 가지고 Hugging Face와 협력을 시작했습니다. Hugging Face는 AMD의 엔지니어링 팀과 적극적으로 협력하여 최고의 성능을 내도록 주요 모델을 최적화하고, AMD ROCm을 Transformers 라이브러리에 통합하고, AMD 플랫폼을 위해 특별히 설계된 라이브러리인 Optimum-AMD를 개선하여 코드 변경을 최소화하면서 Hugging Face 사용자가 이 플랫폼을 활용할 수 있도록 지원하고 있습니다.

또한 Dell Technologies는 기업이 Hugging Face 커뮤니티를 활용하여 최고의 Dell 인프라스트럭처 제품 및 서비스를 기반으로 자체 오픈 소스 생성형 AI(Gen AI) 모델을 개발하고 미세 조정하며 적용하는 프로세스를 간소화할 수 있도록 하기 위해 최근 Hugging Face와 힘을 합쳤습니다. 새로운 Dell 포털은 Hugging Face 플랫폼에서 개발되고 있습니다. 이 플랫폼에는 Dell의 서버와 데이터 스토리지 시스템을 사용하여 Hugging Face에서 제공되는 오픈 소스 모델을 안전하고 손쉽게 배포할 수 있도록 도와주는 맞춤형 전용 컨테이너와 스크립트가 포함되어 있습니다. 이제 기업은 Hugging Face 리소스를 최대한 활용하여 AMD 프로세서가 탑재된 Dell PowerEdge 서버에 모델을 직접 배포하고 자체 독점 데이터를 사용하여 포괄적인 AI 솔루션을 구축할 수 있습니다.



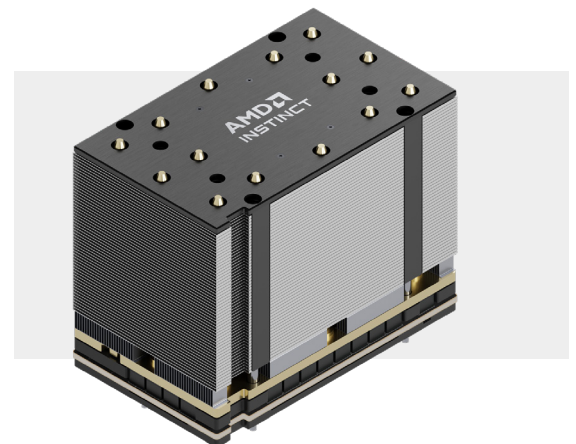
AMD EPYC 프로세서

AMD는 AMD EPYC 프로세서를 통해 최신 클라우드 기반 데이터 센터에 필요한 기술을 발전시키고 있습니다. 이러한 프로세서는 현재와 미래의 데이터 센터 수요를 효율적으로 해결하기 위해 처음부터 새로 설계된 SoC(System On Chip)입니다. AMD EPYC 9000 Series 프로세서는 데이터 센터에 최대 128개의 코어, 256개의 스레드, 소켓당 최대 6TB의 메모리를 지원하는 12개의 메모리 채널과 128개의 PCIe Gen5 레인을 갖추고 있습니다. 이 솔루션은 업계를 선도하는 하드웨어 내장 x86 서버 보안 솔루션과 결합되어 있습니다. AMD EPYC 프로세서는 필수 컴퓨팅, 메모리, I/O 및 보안 리소스를 SoC에 통합함으로써 최고 수준의 성능을 제공하고 TCO(Total Cost of Ownership)를 낮춥니다.



AMD INSTINCT MI300X 가속기

첨단 AMD cDNA 3 아키텍처를 기반으로 구축된 AMD Instinct MI300X 가속기는 가장 컴퓨팅 집약적인 AI 및 HPC 애플리케이션에 필요한 업계 최고 수준의 효율성과 성능을 제공합니다. 304개의 고성능 컴퓨팅 유닛을 갖추고 있으며, 새로운 데이터 유형, 사진 및 비디오 디코딩 지원, 단일 가속기에서 독보적인 192GB의 HBM3 메모리와 같은 AI 관련 기능을 갖추고 있습니다.

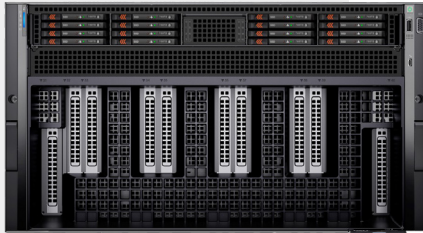


AMD ROCm 6 오픈 소스 소프트웨어 플랫폼

AMD ROCm 6 오픈 소스 소프트웨어 플랫폼은 AMD Instinct MI300X 가속기의 HPC(High Performance Computing) 및 AI 워크로드 성능을 극대화하도록 최적화되어 있습니다. 또한 AMD Instinct MI300X 가속기에 대한 지원을 확장하여 업계 소프트웨어 프레임워크와의 호환성을 보장합니다. AMD ROCm 플랫폼은 커널 수준부터 최종 사용자 애플리케이션에 이르기까지, 가속기 프로그래밍을 지원하는 다양한 드라이버, 개발 툴 및 API를 캡슐화하며 특정 요구 사항에 따라 맞춤화할 수 있습니다. AMD ROCm은 HPC(High Performance Computing), AI(Artificial Intelligence) 및 과학 컴퓨팅의 애플리케이션에 특히 적합합니다. 또한 AMD ROCm 플랫폼은 서버 노드 통신을 위한 RDMA(Remote Direct Memory Access)를 비롯하여 다중 가속기 컴퓨팅을 지원합니다.

AMD
ROCm

DELL POWEREDGE 서버 포트폴리오



Dell은 AMD 투자는 AI를 민주화하기 위한 노력의 일환으로서, 중요한 선택 옵션을 제공합니다. 이러한 사실은 EPYC를 탑재한 4개의 서버 플랫폼과 AMD Instinct MI300X 가속기를 탑재한 Dell PowerEdge XE9680 랙 서버를 보면 알 수 있습니다. 데이터베이스 및 분석, 가상화, 소프트웨어 정의 스토리지, VDI(Virtual Desktop Infrastructure), 컨테이너화, 고성능 컴퓨팅(HPC), AI, ML(Machine Learning) 등의 혁신적인 워크로드를 지원하는 AMD EPYC 프로세서 기반의 최신 세대 Dell PowerEdge 서버를 사용하면 비즈니스 민첩성을 높이고 출시 시간을 단축할 수 있습니다. 1 소켓(단일 CPU) 랙 서버는 성능과 스토리지 용량의 비용 효율적인 균형을 제공하여 비즈니스의 성장에 따라 원활하게 확장되도록 설계되었으며, 2 소켓(듀얼 CPU) 랙 서버는 다양한 기능을 통해 더욱 까다로운 워크로드를 지원합니다.

Dell PowerEdge XE9680 랙 서버는 AI 작업에 맞춤화된 강력한 데이터 처리 시스템입니다. ML(Machine Learning)/DL(Deep Learning) 훈련 및 추론 워크로드, 특히 LLM(Large Language Model) 훈련에 적합한 8개의 가속기를 지원합니다. 각각 192GB의 5.3TB/s 고대역폭 메모리(HBM3)를 장착한 MI300X 가속기 8개를 탑재하여 서버당 1.5TB의 총 HBM3 용량과 21페타플롭 이상의 FP16 성능을 제공하며, AMD Instinct MI300X 가속기가 탑재된 Dell PowerEdge XE9680 랙 서버는 생성형 AI 접근성을 엔터프라이즈까지 확장할 준비가 되어 있습니다. 따라서 대규모 모델을 훈련하고, 데이터 센터 설치 공간을 최소화하고, TCO를 절감하고, 경쟁 우위를 확보할 수 있습니다.

요약

AI가 주도하는 급격한 혁신은 다른 어떤 기술 혁신보다 빠르게 데이터 센터 워크로드에 변화를 불러오고 있습니다. 이러한 기술 발전을 지원하기 위해 Dell과 AMD는 모든 산업 분야의 개발자들이 오픈 소스 리소스를 협업하고 오늘날의 생성형 AI 혁신에 나서도록 장려하는 더욱 포괄적이고 혁신적이며 윤리적으로 발전된 AI 생태계를 만들기 위해 노력하고 있습니다. AI 솔루션의 성능 요구 사항을 충족하기 위해 AMD EPYC 프로세서와 AMD Instinct 가속기 중 무엇이 탑재된 서버를 사용하면, Dell은 하드웨어 플랫폼에서 AI 워크로드를 실행할 수 있는 유연성을 제공합니다. 따라서 Dell과 AMD가 제공하는 최고의 기능을 활용하는 것이 가능합니다.

참조

AMD 이미지: AMD.com, AMD Partner Resource Library,
<https://www.amd.com/en/partner/resources/resource-library.html>

Dell 이미지: [Dell.com](https://www.dell.com)