White Paper

워크스테이션에 AI 기술을 개발하고 배포하는 것이 중요한 이유

Sponsored by: Dell Technologies

Peter Rutten Dave McCarthy
July 2023

IDC 의 견해

AI 는 모든 산업에서 차별화를 위한 중요 역량으로 자리 잡았으며, AI 실행에 필요한 하드웨어는 빠르게 발전하고 있다. 기술 산업은 최첨단 AI 모델의 기하급수적인 규모의 성장에 집중하는 경우가 많다. 현재 수백억 개의 매개변수, 정밀도 감소, 메모리 확장, AI 학습 및 추론에 필요한 HPC(High Performance Computing), 가속화된 서버 랙 등에 대한 논의가 이루어지고 있으나, 실제로 이러한 대규모의 AI 컴퓨팅은 예외적인 현상이며, 기업의 입장에서는 특히 그렇다.

오늘날 많은 기업들은 슈퍼컴퓨터가 필요하지 않은 생성형 AI 를 비롯한 AI 이니셔티브에 많은 노력을 기울이고 있다. 실제로 많은 AI 개발은 강력한 워크스테이션에서 이루어지고 있으며, 특히 엣지에서 AI 를 배포하는 사례가 늘어나고 있다. 워크스테이션은 AI 개발 및 배포에 많은 이점을 제공한다. AI 과학자나 개발자가 서버 시간을 조율할 필요가 없고, 데이터 센터에서 서버 기반 GPU 를 쉽게 이용할 수 없는 상황에서도 GPU 가속을 제공하고, 서버에 비해 매우 저렴하고, 빠르게 누적되는 클라우드 인스턴스 비용 대신 더 적은 일회성 비용으로 이용할 수 있으며, 기밀 데이터가 온프레미스에 안전하게 저장된다는 점에서 안심할 수 있는 등 다양한 장점이 있다. 따라서 과학자나 개발자는 AI 모델 실험에만 집중하면서 비용 부담에서 벗어날 수 있다.

IDC 는 AI 배포 시나리오에서 엣지가 온프레미스나 클라우드보다 빠르게 성장하는 것을 확인하고 있다. 여기서도 워크스테이션은 AI 추론 플랫폼으로 점점 더 중요한 역할을 하고 있으며, 심지어는 GPU 없이 소프트웨어에 최적화된 CPU 에서 추론을 하는 경우도 많다. 워크스테이션의 엣지에서 AI 추론을 하는 활용 사례가 빠르게 늘고 있으며, 예를 들어 여기에는 AI 옵스, 재해 대응, 방사선학, 석유 및 가스 탐사, 토지 관리, 원격 의료, 교통 관리, 제조 공장 모니터링, 드론 등이 포함된다.

이 백서에서는 AI 개발과 배포에서 워크스테이션의 역할이 확대되고 있는 양상을 살펴보고, 델(Dell)의 AI 용 워크스테이션 포트폴리오에 대해 간략하게 설명한다.

상황 개요

AI 의 폭발적인 성장과 인프라스트럭처에 미치는 영향

전 세계 기업에서 진행하고 있는 AI 프로젝트 수가 급격하게 늘어나고 있다. 이미 모든 산업에서 많은 작업이 부분적으로 또는 전체적으로 AI 모델 기반의 소프트웨어를 통해 수행되고 있다. IDC 는 여러 측면에서 AI 를 추적하고 있으며, 고려해야 할 유용한 지표 중 하나는 기업과 클라우드 서비스 공급업체가 AI를 개발하고 실행하기 위해 서버에 지출할 것으로 예상되는 금액이다. 2026 년에는 이 금액이 346 억 달러에 이를 것으로 전망되며, 이는 전 세계 서버 총지출의 약 22%를 차지할 것이다.

하지만, 서버만으로는 전체 상황을 파악할 수 없다. 많은 AI 준비(AI preparation), 개발, 프로토타입 제작 및 배포가 워크스테이션에서 이루어지고 있다. 규모와 상관없이 조직은 애플리케이션에 어느정도의 AI 기능을 주입하여 새로운 비즈니스 기회를 실현할 수 있다는 사실을 알게 되었다. 이에 따라 AI 모델에 대한 실험이 급증했는데, 데이터에 대한 즉각적인 가용성과 근접성을 갖춘 강력한 워크스테이션이 이러한 목적에 이상적이다.

AI 알고리즘이 수십 년 동안 배포되어 왔는데 왜 갑자기 AI 가 이렇게 널리 보급되게 되었을까? 그 이유는 지난 몇 년간 특히 성공적인 AI 알고리즘인 신경망을 구동하기 위한 두 가지 필수 조건이 실현되었기 때문이다. 즉, 비정형 및 반정형 데이터와 같은 방대하고 저렴하며 다양한 유형의 데이터를 쉽게 이용할 수 있게 되었고 병렬 모델로 선형 컴퓨팅을 강화하여 이러한 신경망을 적절한 시간 내에 처리할 수 있게 되었다. 이 두 가지 기본 조건이 충족되면서 데이터 과학자들은 점점 더 중요한 작업의 실행 방법을 자동으로 학습하는 신경망을 개발하여 엄청난 발전을 이루었다. 기존의 ML(Machine Learning)은 텍스트나 숫자 데이터에 적합하지만, DL(Deep Learning)은 비디오, 오디오, 언어 등에 더 효과적이다.

기존 머신 러닝 모델은 일반적으로 최대 수십 개의 코어를 갖춘 워크스테이션의 CPU 에서 개발할 수 있지만, 신경망 개발에는 수천 개의 코어에 걸쳐 처리를 병렬화하는 보조 프로세서가 필요하다. 그 주된 이유는 ML 에서는 특징 추출 및 분류가 수동 프로세스인 반면, DL 에서는 대규모 데이터 세트를 사용하여 지속적인 반복을 통해 모델을 학습시켜야 하는 자동 프로세스이기 때문이다. 현재 가장 일반적인 보조 프로세서는 GPU 이지만, 스타트업에서 개발한 새로운 AI 전용 프로세서도 출시되고 있다. 병렬 처리를 위해 개별 보조 프로세서를 사용하는 이러한유형의 가속은 서버 및 워크스테이션 시장에 대변혁을 일으켰으며, 이로써 IDC 가 대규모 병렬 컴퓨팅이라고 부르는 컴퓨팅이 탄생했다.

2022 년 전세계 가속 서버(accelerated servers)는 218 억 달러 규모의 시장을 형성했으며, 2026 년까지 434 억 달러 규모로 성장하고 그 중 57%가 AI 실행을 위한 가속 서버가 될 것으로 전망된다. 동시에 워크스테이션용으로 판매된 독립 GPU 수는 2022 년에 640 만 개로 증가했다. IDC 는 AI 개발로 점점 증가하는 과학 또는 소프트웨어 엔지니어링 목적으로 사용되는 워크스테이션 시장이 2026 년까지 20 억 달러에 가깝게 성장할 것으로 예상한다.

AI 개발 단계

앞서 언급했듯이, 데이터의 유형과 양이 늘어나고 컴퓨팅에 대한 새로운 접근 방식이 등장하면서 신경망이 실현 가능해졌다. 이러한 움직임의 첫 부분이라고 할 수 있는 데이터의 양과 유형은 절대 사소하지 않으며, 딥러닝 AI 이니셔티브에 투입되는 노력의 80%가 데이터 관리 및 준비라는 설명도 있다. 모델을 설계하고 학습하기 전에 데이터를 수집하고 관리하고 준비해야 한다. IDC 에 따르면 AI 개발 단계는 다음과 같다(그림 1 참조).

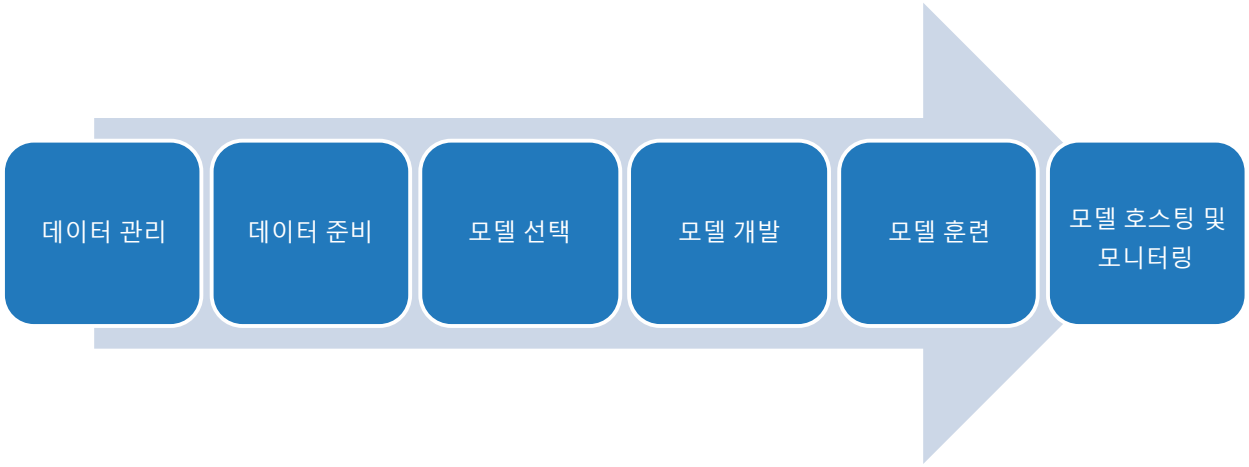
- **데이터 관리:** 데이터 센터와 엣지, 클라우드 전반에 걸쳐 조직이 수집하는 방대한 데이터에서 AI 모델에 맞는 데이터를 식별하고 관리한다. 이벤트 기반 또는 스트리밍 등 유형과 상관없이 모든 데이터가 여기에 해당되며, 그중 상당 부분에는 일종의 거버넌스가 필요하다.
- **데이터 준비:** 데이터 웨어하우스나 데이터 레이크에 데이터(파일, 블록 또는 오브젝트)를 저장하고, 데이터를 정리하고, 고품질의 완전한 데이터인지 확인한 다음, Pandas 같은 툴이나 Spark 를 사용하여 AI 모델에 사용할 수 있는 형태로 변환하는 작업을 말한다.
- **모델 선택:** 프로그래밍된 AI 작업을 오류율 및 성능 측면에서 최적으로 수행할 수 있는 모델을 결정한다.
- **모델 개발:** XGBoost, LightGBM, GLM, Keras, TensorFlow, PyTorch, Caffe, RuleFit, FTRL, Snap ML, scikit-learn, H2O 등의 프레임워크를 사용하여 AI 모델을 설계한다.

- **모델 훈련:** 병렬화에 충분한 프로세서 및 보조 프로세서 코어를 갖춘 컴퓨팅 인프라스트럭처에서 모델을 교육시킨다. 모델 학습에는 학습된 모델에서 추론을 실행하여 테스트하는 프로토타입 제작이 포함되며, 점차 공정성, 책임성, 투명성을 보장하기 위해 모델의 결정을 설명, 검증, 문서화하는 기능도 포함되고 있다.
- **모델 호스팅 및 모니터링:** 프로덕션 환경에 모델을 배포하여 해당 모델의 용도에 맞는 작업을 실행하게 하고(일반적으로 "AI 추론"이라고 함) 그 성능을 모니터링한다.

워크스테이션은 데이터 센터, 클라우드 또는 엣지 인프라스트럭처와 함께 이러한 여섯 단계 모두에서 중요한 역할을 할 수 있다.

그림 1

AI 개발 단계



출처: IDC, 2023 년

워크스테이션에서 AI 모델 개발

워크스테이션과 개인용 컴퓨터(PC) 비교

일반적으로 PC(Personal Computer)가 AI 를 개발할 만큼 강력하지 않다는 사실은 잘 알려져 있다. 데이터 과학자와 AI 개발자는 보통 소속 조직에서 전략적으로 중요한 프로젝트에 관여하고 있으며, 이들에게는 생산성이 저하되지 않는 것이 무엇보다 중요하다. 워크스테이션은 일반적으로 고성능 구성요소로 제작되고 실행 중인 소프트웨어에 최적화되어 있으므로 PC 보다 더 예측가능 한 성능을 발휘하는 경향이 있다.

여기에는 다음과 같은 구성요소가 포함된다.

- **고급 프로세서:** 인텔 제온 스케일러블(Intel Xeon Scalable) 프로세서를 예로 들 수 있다.
- **강력한 GPU:** NVIDIA RTX 6000 Ada 와 같은 엔비디아(NVIDIA)의 RTX 전문가용 GPU 를 예로 들 수 있다.
- **대용량 스토리지:** 워크스테이션에 따라서는 60TB 의 스토리지를 제공하기도 하며, I/O 속도는 PC 보다 훨씬 빠른 편이다.
- **대용량 메모리:** 이제 메모리를 6TB 까지 탑재한 워크스테이션을 사용할 수 있다.

- **냉각:** 고성능 구성 요소(components)에서는 많은 열이 발생하며, 데이터 과학자에게는 적절한 냉각 기능으로 과열을 방지하고 최적의 성능을 유지할 수 있는 워크스테이션이 필요하다.
- **NIC(Network Interface Card):** 원격 서버에 저장된 대규모 데이터 세트로 작업하는 데이터 과학자에게는 데이터를 빠르고 효율적으로 전송할 수 있는 고속 네트워크 인터페이스 카드가 필수다.
- **디스플레이**: 데이터 시각화 작업에는 고품질 디스플레이가 중요하며, 데이터 과학자에게는 색 정확도가 높고 화면이 큰 고해상도 모니터가 필요하다.
- **ECC(Error Correcting Code) 메모리:** ECC 는 가장 전형적인 내부 데이터 손상을 감지하고 수정하여 장시간의 AI 학습 실행 중 하드 오류(불량 비트)나 소프트 오류(비트 플립으로 잘못된 값 발생)로 인한 블루스크린을 방지하며, 의료분야 등 생명에 중요한 작업에 매우 중요한 결과 정확성도 보장한다.
- **특수 실리콘:** 유통, 보안, 산업 자동화 등의 환경에서 사용되는 컴퓨터 비전 및 엣지 AI 애플리케이션을 위한 병렬 프로세싱 보조 프로세서인 인텔® Movidius™ VPU(Vision Processing Unit)를 예로 들 수 있다. FPGA 도 워크스테이션에서 금융 애플리케이션 등에 사용된다.
- **최적화 소프트웨어:** 최적화 소프트웨어의 예로는 OneAPI 와 CUDA 가 있다. OneAPI 는 인텔의 표준 기반 프로그래밍 모델로 CPU, GPU, FPGA 및 기타 가속기 전반에 걸쳐 데이터 중심 워크로드의 개발과 배포를 간소화한다. CUDA 는 GPU 에서 일반 워크로드를 실행하기 위한 엔비디아의 병렬 컴퓨팅 플랫폼이자 애플리케이션 프로그래밍 인터페이스다.

AI 용 CPU 와 GPU 비교

워크스테이션은 AI 개발의 여러 단계에서 사용할 수 있으며, 일반적으로 다양한 기능을 갖추고 있다. 병렬 처리를 위한 GPU 가 강조되고 있지만, 워크스테이션에서 AI 모델을 개발할 때는 CPU 의 역할이 중요하다. GPU 와 마찬가지로 CPU 도 데이터 조작은 물론 기존 ML 모델을 개발하는 데 사용할 수 있다. 또한 CPU 는 데이터 세트의 시각적 표현을 사용하여 데이터 특성을 파악하는 프로세스인 데이터 탐색에도 사용된다.

DL 훈련에서는 실제 훈련 프로세스 과정에서 GPU 가 대신하기 때문에 호스트 CPU 의 역할이 다소 줄어들지만, 이 경우에도 CPU 는 계속 OS 나 CUDA 와 같은 중요한 소프트웨어의 처리 계층 역할을 하며, GPU 간에 또는 다른 실리콘과의 프로세스를 조율한다. 또한 프로덕션 환경에서 AI 모델을 실행하는 데 워크스테이션을 사용하는 경우, CPU 를 새로운 방식의 AI 추론 엔진으로 사용하는 사례가 점점 늘어나고 있다. IDC 는 2024 년까지 AI 추론을 위한 인프라스트럭처에 투입되는 지출이 AI 훈련을 위한 AI 인프라스트럭처에 투입되는 지출을 초과할 것이며, 이 추론의 상당 부분(39%)이 호스트 CPU 에서 이루어질 것으로 예상한다.

워크스테이션과 서버 비교: 공생 관계

대부분의 조직에서는 경험적으로 AI 개발을 위해 워크스테이션, 온프레미스 서버, 클라우드 인스턴스 또는 이 세 가지를 조합하여 배포할 때 실용주의를 우선한다. AI 프로젝트의 다양한 개발 단계에서 워크스테이션과 서버, 클라우드 인스턴스 간에는 공생 관계가 존재한다.

데이터 센터 서버와 비교하여 워크스테이션은 데이터 과학자가 원하는 곳 어디에서나 작업할 수 있다는 장점이 있으며, 이는 최근의 팬데믹 상황뿐 아니라 일반적인 상황에서도 중요한 요소다. 또한 강력한 GPU 를 탑재한 최신 워크스테이션의 성능 덕분에 필요한 만큼 반복하여 AI 모델을 자유롭게 실험할 수 있다. 또한 서버에 액세스를 요청하거나 기타 데이터 센터의 제약에 구애받지 않고 반복 프로세스를 진행하여 원활하게 상호작용하면서 즉각적인 피드백과 결과를 제공할 수 있다. 워크스테이션을 사용하면 컴퓨팅을 데이터와 더 가까운 방향으로 유연하게 이동할 수 있어 대역폭을 절감하고 네트워크 혼잡을 줄이며 처리량을 늘릴 수 있다. 또한, 워크스테이션은 기존의 ML 작업이나 DL 집약적인 작업 등 다양한 요구 사항에 맞게 구성할 수 있다.

또한 가속 서버 시장이 크게 성장했음에도, 아직 엔터프라이즈 데이터 센터에서 가속 서버를 널리 사용하고 있지는 않다. 이 백서가 작성될 당시에는 엔터프라이즈 데이터 센터 서버의 평균 4%가 가속화되었는데, 이는 즉시 사용할 수 있는 온프레미스 GPU 에서 AI 를 개발하거나 실행할 수단을 갖춘 조직이 많지 않다는 의미다. 이러한 이유로도 가속화된 워크스테이션은 AI 개발에 유용한 대안이 될 수 있다.

고도로 가속화된 워크스테이션은 이제 AI 모델이 과도하게 크지 않는 한 DL 을 훈련할 수 있을 정도로 강력해졌으므로 서버에서 교육할 필요가 없어졌다. 또한 GPU 를 탑재한 워크스테이션에서 교육한 모델은 GPU 가 탑재되지 않은 워크스테이션이나 서버에 배포하여 CPU 의 추론 기능을 활용하도록 할 수 있다. 인텔의 DL 부스트나 oneAPI 와 같은 소프트웨어 기술은 CPU 에서 AI 추론을 강화한다. 따라서 데이터 센터에 이미 배포되어 있는 비가속 서버(non-accelerated servers)로 AI 애플리케이션을 지원할 수 있다.

워크스테이션과 클라우드 비교

클라우드 컴퓨팅은 인프라스트럭처와 데이터, 애플리케이션에 대한 조직의 시각을 혁신적으로 바꾸어 놓았다. 무한에 가깝게 확장할 수 있는 클라우드를 사용하면 개발자는 필요에 따라 리소스를 프로비저닝할 수 있으므로 제약을 줄이면서 혁신을 가속화할 수 있다. 클라우드는 그 자체로 AI 개발을 위한 완벽한 패러다임으로 보인다.

하지만, 언제나 그런 것은 아니다. IDC 연구에 따르면 조직들은 특정 워크로드를 점차 퍼블릭 클라우드에서 온프레미스 인프라스트럭처로 되돌리고 있는 것으로 나타났다. 그 이유는 주로 다음과 같다.

- **클라우드 가용성:** 클라우드 서비스를 사용해 본 사람이라면 누구나 클라우드 공급업체 자체의 문제로 인해 하이퍼스케일 데이터 센터와 최종 사용자 간의 네트워크 연결이 끊어지는 등의 서비스 중단의 경험이 있을 것이다. 이러한 상황에서 사용자는 문제를 해결하는 데 서비스 공급업체에만 의존해야 하며, 문제가 해결될 때까지의 생산성은 저하된다.
- **보안 및 규정 준수:** 많은 산업에서 기업 거버넌스 정책에 따라 데이터가 전달되고 저장될 수 있는 위치가 결정되는데, 이에 따라 클라우드 서비스 사용이 제한된다. 유럽의 GDPR 과 캘리포니아 소비자 개인정보 보호법 같은 정부 규정에서도 데이터 주권에 관한 규정을 적용하고 있다.
- **비용:** 조직은 클라우드 서비스 요금이 얼마나 빨리 늘어날 수 있는지 과소평가하는 것이 일반적이며, 특히 고성능 컴퓨팅 기능과 대용량의 스토리지가 필요한 워크로드의 경우 더욱 그렇다. 클라우드 경제성은 모든 유형의 리소스 소비 측정을 기반으로 하며, 여기에는 기업 내부 인프라스트럭처로의 데이터 재송신도 포함된다.
- **시행착오에 대한 부담:** 대부분의 AI 이니셔티브는 많은 양의 실험으로부터 시작되며, 개발 과정에서 실패하는 모델이 나타나는 것은 당연한 일이다. 이 과정에서 아직 실행 가능한 결과를 보여줄 수 없는 상황인데도 클라우드 청구 비용이 누적되면 AI 과학자와 개발자는 심리적인 부담을 안게 된다.

워크스테이션은 마이크로서비스 기반 아키텍처 및 API 기반 자동화와 같은 클라우드 네이티브 기술을 활용하면서도 이러한 한계를 해결할 수 있다. 이를 통해 워크스테이션을 데이터 센터 서버와 비교했을 때와 동일한 이점 몇 가지를 얻을 수 있다:

- **어디에서나 작업 가능:** 퍼블릭 클라우드에 대한 의존도를 낮춤으로써 이제 연결하지 않고도 작업할 수 있다. 보안이 철저한 환경은 대체로 공용 네트워크와 물리적으로 격리되어 있으며, AI 워크스테이션은 이러한 요구 사항을 고유한 방식으로 해결할 수 있다. 또한 로컬 리소스로 인해 값비싼 네트워크 연결에 대한 수요도 줄어들게 된다.

- **데이터 지역성:** IoT 디바이스 및 기타 연결된 장비가 급증하면서 엣지 위치의 데이터가 기하급수적으로 증가하고 있다. 많은 상황에서 컴퓨팅 리소스를 전용 워크스테이션에 코로케이션하는 것이 합리적이다. 이렇게 하면 데이터 이동이 제한되어 다양한 규정 준수 요건 문제도 해결된다.
- **자유롭게 실험 가능:** AI 모델을 훈련하고 최적화하는 과정은 반복적인 프로세스로, 시행착오가 수반되는 경우가 많다. 개발자는 추가 서비스 수수료가 발생할 수 있다는 이유로 타협하지 않고 자유롭게 실험할 수 있어야 한다. 또한 워크스테이션을 사용하면 유연하게 맞춤형 툴을 구성할 수 있다.

맞춤형 툴의 경우, 대부분의 클라우드 서비스 공급업체에서 사용자가 배포하고자 하는 구성에 대해 즉시 비용 견적을 제공하므로 워크스테이션과 클라우드 배포 가격을 상대적으로 쉽게 비교할 수 있다. 예를 들어, 한 주요 클라우드 공급업체에서 하나의 NVIDIA T4 와 하나의 375GiB SSD 스토리지 인스턴스가 있는 단일 표준 VM(Virtual Machine)을 하루에 8 시간씩 주 5 일 사용하는 비용은 140 달러이다. VM, T4, SSD 를 두 배로 늘리면 비용은 월 365 달러로 증가한다. VM 은 두 대로 유지하면서 T4 4 개, 375GiB 스토리지 4 개로 두 배 늘리고 이 환경에서 온종일 AI 를 훈련하면 비용은 월 2,700 달러까지 증가한다. 따라서 AI 개발을 위한 클라우드 비용은 연간 수만 달러까지 쉽게 치솟을 수 있으며, 이는 고급(high-end) 워크스테이션의 연간 감가상각비보다 훨씬 많은 비용이다.

워크스테이션에서 AI 프로토타입 제작

온프레미스 서버 및 클라우드와 비교할 때 워크스테이션은 AI 모델 프로토타입 제작에서 뚜렷한 장점을 보인다. 데이터 센터의 서버는 AI 프로토타입 제작 및 테스트에 최대로 사용되거나 매우 중요한한 작업에 사용될 수 있다. 앞서 설명한 대로 클라우드 인스턴스는 테스트 환경으로 자유롭게 사용하는 경우 비용이 급격하게 치솟을 수 있는데, 워크스테이션을 사용하면 AI 과학자나 개발자는 서버 액세스를 조율하거나 프로토타입 제작 단계에서 클라우드 비용이 늘어나는 부담에서 벗어날 수 있다. 또한 추가 비용 없이 낮은 일회성 비용으로 언제 어디서나 자유롭게 프로토타입을 제작할 수 있다.

워크스테이션에 AI 모델 배포

워크스테이션에서 AI 모델을 개발하는 것은 수년간 일반적인 전략이었지만, IDC 는 최근 AI 모델을 워크스테이션(일반적으로 엣지)에 배포하는 사용 사례가 증가하고 있는 것을 확인했다. 즉, AI 모델에 대한 추론을 실행함으로써 AI 모델을 워크스테이션에서 생산에 투입하는 것이다. 2020 년부터 2024 년까지 연간 하드웨어 지출이 3 배 이상 증가할 정도로 서버용 AI 배포 위치로서 엣지는 빠르게 성장하고 있으며, 최종 사용자가 엣지의 장점을 인식함에 따라 워크스테이션도 크게 뒤처지지 않고 있다.

IDC 는 엣지를 분산 컴퓨팅 패러다임으로 정의한다. 이 패러다임에서는 인프라스트럭처와 애플리케이션을 중앙 집중식 클라우드 및 온프레미스 데이터 센터 외부에 데이터가 생성되고 소비되는 곳과 최대한 가까운 곳에 배포한다. 여기에는 원격 사무실과 지사는 물론 공장, 창고, 병원, 소매점 등 산업별 위치가 포함된다.

데이터 및 컴퓨팅 집약적인 워크로드는 점점 더 온프레미스나 엣지 위치에 배포되고 있다. 특히 상당한 양의 데이터 과학 실험이 필요한 상황에서 대규모 데이터 세트를 업로드하는 데 걸리는 시간과 AI 훈련에 소모되는 가변 비용 등, 퍼블릭 클라우드 특유의 제한을 완화하기 위해 엣지에 배포되는 경우가 더욱 증가하고 있다

IDC 연구에 따르면 엣지는 AI 배포 시나리오에서 빠르게 성장하고 있으며, 2023 년에 이루어진 조직들의 엣지 AI 컴퓨팅 투자 비용은 29 억 달러에서 2026 년에는 69 억 달러까지 증가할 것으로 예상된다(*Worldwide AI Hardware Forecast, 2022–2026: Strong Market Growth for AI Compute and Storage,* IDC #US49671722, 2022 년 9 월 참조). 또한 엣지는 엔지니어링 및 기술과 같은 HPC 워크로드 배포 옵션으로도 주목을 받고 있는데, 현재 기업들은 엣지에서 이러한 워크로드에 약 10 억 달러를 투자하고 있으며, 이 금액은 2027 년에 24 억 달러까지 증가할 것으로 예상된다(*Worldwide High-Performance Computing Server Forecast, 2023-2027: Enterprise Will Overtake HPC Labs,* IDC #US50525123, 2023 년 4 월 참조). 이처럼 엣지는 AI 워크스테이션을 배포하기에 합리적인 대상이다.

엣지의 워크스테이션에 AI 모델을 배포할 때 AI 를 개발할 때처럼 항상 고급 GPU 가 필요한 것은 아니다. 사양이 낮은 GPU 로도 AI 추론을 수행할 수 있으며, 상당한 경우 GPU 가 전혀 필요하지 않다. GPU 가 없는 경우, 특히 인텔 DL 부스트 같은 최적화와 함께 사용하면 CPU 가 추론 작업을 적절하게 수행할 수 있다. 인텔 DL 부스트는 AI 추론을 포함한 AI 워크로드를 가속하도록 설계된 인텔 마이크로프로세서의 명령어 집합 기능 세트이다. 이와 관련해서, 인텔에 따르면 인텔 DL 부스트를 지원하는 4 세대 인텔 제온 스케일러블 프로세서에서 이전 세대(BERT-Large SQuAD)보다 1.45 배 더 높은 INT8 실시간 추론 처리량을 기록했다. 또한 이러한 점 때문에 전력, 이동성, 발열 관리 등을 고려할 때 더 적은 전력을 요구하는 엣지에 배포하기에 적합한 워크스테이션—을 만드는데 도움이 된다. 인텔 M2(Movidius Myriad)는 12W 의 적은 에너지 사용량 덕분에 이러한 전력 엔벨로프에 적합하다.

워크스테이션에서의 AI 배포 활용 사례

로컬로 배포된 워크스테이션에 AI 를 배포하는 것이 자연스러운 몇 가지 상황이 있다. 공통된 특징은 머신에서 생성하는 대량의 시계열 데이터와 비디오 스트림 및 이미지와 같은 비정형 데이터라는 것이다. 또한 SME(Subject Matter Expert, 분야별 전문가)가 인간의 해석을 통해 AI 모델을 보강해야 하는 경우도 있다.

그 예는 다음과 같다.

- **AI 옵스:** IT 시스템의 규모와 복잡성이 증가함에 따라 사후 대응적 인시던트 관리에서 사전 예방적 모니터링으로 전환해야 할 필요성이 증가하고 있다. 특히 기술 인력이 거의 없거나 전혀 없는 엣지 위치에 인프라와 애플리케이션이 분산되어 있는 경우에는 더욱 그렇다. 정상 성능의 기준을 모델링함으로써 이상 징후를 식별하고 문제 해결 단계를 자동화할 수 있다.
- **재해 대응:** 긴급 상황 발생 시 긴급 대응팀은 신속하게 상황을 평가하고, 중요한 장비를 추적하며, 가장 도움이 필요한 곳을 지원할 수 있도록 리소스를 배치해야 한다. 이러한 작업은 네트워크에 연결되지 않은 환경에서 진행해야 하는 경우가 많으므로 데이터 피드를 집계하고, AI 모델로 추론하고, 핵심 인력과 커뮤니케이션을 자동화할 수 있는 로컬 워크스테이션이 필요하다.
- **방사선학:** 영상 처리 기술이 발전하면서 한 번의 검사에서 생성되는 데이터 크기가 증가했으며, 이로 인해 적시에 데이터를 분석하려면 현장에서 데이터가 유지되어야 한다. 수백만 개의 이전 예들로 교육된 AI 모델은 육안보다 정확하게 패턴을 식별하여 정확도를 높일 수 있다.
- **석유 및 가스 탐사:** 업스트림(upstream) 석유 및 가스 회사는 원격 측정, 지진 및 영상 데이터를 조합하여 천연자원의 매장량을 파악하고 시추 위치를 선정하며 생산 과정에서 장비의 성능을 최적화한다. 이를 위해서는 대개 고가의 위성 통신만 가능한 지역에서 정보를 분석해야 한다.
- **암 연구 및 신약 개발:** 연구 병원과 학계의 연구원들은 AI 와 자연어 처리를 사용하여 암 전문의가 각 환자에게 맞는 가장 효과적인 암 치료법을 결정할 수 있도록 지원한다. 또한 머신 러닝과 컴퓨터 비전을 결합하여 영상의학과 전문의가 환자의 종양 진행 상황을 효과적으로 이해하도록 지원한다. 그리고 알고리즘을 사용하여 암 진행 상황과 암을 퇴치하는 데 가장 효과적인 치료법이 무엇인지 한층 적절히 파악한다.

- **보험금 청구 심사:** 수작업으로 하는 보험금 청구 처리는 노동 집약적이며 인적 오류가 발생하기 쉽다. 보험금 청구가 유효한지를 평가할 수 있는 AI 를 사용하면 보험 손해사정사가 더 자세한 조사가 필요한 사건에 집중할 수 있어 비용이 절감된다. 그러면 정확성을 유지하면서 전반적인 운영 처리량이 늘어난다.
- **원격 의료:** AI 는 웨어러블 기기의 실시간 생체 신호를 기반으로 각 개인에 맞춰 치료 계획을 마련함으로써 환자의 회복 속도를 개선하고 있다. 이 정보는 환자의 병력 및 유사 사례에 대한 기술 자료와 결합된다. 이러한 정보는 원격 의료 서비스에 대한 의존도가 높은 외곽 지역에서 특히 중요하다.
- **리테일 보안(도난 방지):** 비디오 스트림에 적용된 실시간 분석은 범죄 행위로 이어질 수 있는 사람의 행동을 예측하는 데 사용되고 있다. 이를 위해서는 일반적으로 여러 개의 비디오 피드를 연결하여 매장 내에서 개인의 움직임을 추적해야 한다. 중요한 사건을 식별하는 작업은 최대한 빨리 이루어져야 하므로 이러한 프로세스는 로컬에서 실행되는 것이 가장 좋다.
- **교통 관리:** 교통 운영을 담당하는 정부 기관은 AI 를 사용해 신호등과 디지털 사이니지를 조정하여 차량 흐름을 개선하고 시민의 안전을 유지하는 데 한층 더 많은 노력을 기울이고 있다. 이를 위해서는 비디오 카메라, 도로 센서를 통한 원격 측정 등의 다양한 입력을 조합하여 교통 패턴을 최적화해야 한다.
- **제조 공장 모니터링:** 공장 관리자에게는 중요 공정의 가동 시간을 보장하고 생산 일정을 맞추는 것이 가장 중요하다. 이는 주요 장비의 예측 유지 관리와 자동화된 결함 감지, 현장 공급망 안팎의 최적화로 이어진다. 이 분야는 AI 가 안전 기준을 유지하면서 성능을 향상시킬 수 있도록 인간 작업자에게 도움을 줄 수 있는 영역이다.
- **드론:** 드론이 촬영한 이미지를 자동 분석하면 이전에는 불가능했던 규모의 광범위한 상황을 모니터링할 수 있다. 드론의 사용은 가스 및 전기 유틸리티 인프라 점검, 보험 조사, 수색 및 구조 활동, 정밀 농업, 어업 및 야생동물 보호구역 유지 관리 등에 큰 영향을 미치고 있다.
- **일상적인 사무실 환경:** 일상적인 사무실 환경은 Microsoft Copilot 과 같은 AI 기반 생산성 향상 툴을 통해 점차 개선되고 있다.
- **재생에너지:** 풍력 발전소, 수력 발전 댐, 태양광 발전소 같은 재생 에너지 사업장에서는 현지에서 생산하고 분석해야 하는 실시간 모니터링, 유지 보수 및 데이터 수집이 필요하다.

AI 를 위한 델(DELL) 워크스테이션

델(Dell)은 데이터 사이언스 워크스테이션(Data Science Workstation, DSW) 브랜드 산하에 다양한 수준의 AI 개발 및 구현을 위한 광범위한 워크스테이션을 제공한다. 이 섹션에서는 사양을 간략하게 소개한 다음, 데이터 과학자와 같은 다양한 AI 페르소나 및 애플리케이션, 그리고 델 DSW 기술의 이점에 대해 설명할 것이다. 이러한 AI 지원 데이터 사이언스 워크스테이션은 데이터 과학자를 위해 특별히 설계되었다. 최신 프리시전(Precision) 데이터 사이언스 워크스테이션은 AI 기능을 활용하여 데이터 과학자가 가장 많이 사용하는 애플리케이션이 최적의 성능을 구현하도록 디바이스를 미세 조정한다. 이를 통해 가장 중요한 작업을 빠르게 완료할 수 있다. 또한 델 프리시전 워크스테이션은 독립 ISV 의 테스트와 인증을 거쳐 델의 고객이 일상적인 작업을 완료하는 데 필요한 고성능 애플리케이션을 지원한다.

델 워크스테이션이 두각을 나타내는 영역

NVIDIA RTX GPU 기반의 델 프리시전 워크스테이션은 조직의 분석 및 AI 이니셔티브에 강력한 확장성과 성능을 제공하도록 설계되었다. 델 테크놀러지스(Dell Technologies)는 각 업계의 최신 AI 소프트웨어를 실행하는 데 최적화된 포괄적인 하드웨어 솔루션을 제공한다.

- **견고한 하드웨어 구성:** 델 프리시전 워크스테이션은 멀티 코어 프로세서, 대용량 RAM, 여러 개의 GPU 옵션 등 다양한 하드웨어 구성을 통해 강력한 성능을 지원한다. 이러한 구성 요소를 통해 효과적인 교육과 추론을 지원하는 AI 작업에 필요한 컴퓨팅 리소스를 제공한다.
- **확장성 및 사용자 정의:** 델 프리시전 워크스테이션은 확장 및 사용자 정의가 가능하므로 사용자가 구체적인 AI 요구 사항에 맞춰 하드웨어를 구성할 수 있다. 이러한 유연성 덕분에 워크스테이션을 특별한 AI 워크로드 요구 사항에 맞춰 최적화할 수 있다.
- **인증 및 최적화:** 델은 엔비디아와 협력하여 NVIDIA RTX 6000 Ada Generation 카드를 비롯한 NVIDIA RTX GPU 와의 호환성과 성능에 대해 프리시전 워크스테이션을 인증한다. 이 인증은 델 프리시전 워크스테이션과 NVIDIA RTX GPU 를 AI 작업에 사용하는 경우 원활하게 통합할 수 있고 최적화된 성능을 구현한다는 것을 보장한다.
- **강력한 처리 능력:** 인텔 프로세서가 탑재된 델 프리시전 워크스테이션은 AI 작업에 필요한 컴퓨팅 성능을 제공한다. 멀티 코어 프로세서와 높은 클록 속도를 갖춘 프리시전 워크스테이션은 AI 워크플로에서 교육과 추론에 필요한 성능을 제공한다.
- **소프트웨어 및 툴 지원:** 델 프리시전 워크스테이션에는 AI 개발 및 배포를 지원하는 소프트웨어와 툴이 미리 설치되어 있다. 여기에는 사용자가 AI 프로젝트를 더 쉽게 시작할 수 있도록 최적화된 소프트웨어 스택과 AI 프레임워크, NVIDIA RTX GPU 를 활용하는 라이브러리가 포함되어 있다.

또한 다음 섹션에서 설명하는 기술은 델 워크스테이션이 두각을 나타내는 또 다른 중요한 영역이다.

Reliable Memory Technology (RMT)

델은 가동 시간을 극대화하도록 설계된 RMT Pro(Reliable Memory Technology Pro)라는 ECC 기반 기술을 제공한다. 이 기술은 ECC 메모리와 연계하여 메모리 오류를 실시간으로 탐지하고 수정한다. 델에 따르면, RTM Pro 는 DIMM 전체를 사용 중일 때에도 불량 메모리로 다시 가지 않도록 방지하여 메모리 오류를 사실상 제거한다. 시스템 재부팅 후 RTM Pro 는 결함이 있는 메모리 영역을 격리하고 OS 에서 숨긴다. 결과적으로 불량 메모리로 인한 문제가 계속 해결되므로 AI 데이터 과학자와 개발자는 지속적으로 충돌 문제를 겪지 않게 되어 생산성이 크게 향상된다.

Dell Optimizer for Precision (DOP)

대부분의 델 워크스테이션에는 또한 델 옵티마이저의 프리시전 버전(Dell Optimizer for Precision)이 포함되어 있다. DOP(Dell Optimizer for Precision)는 워크스테이션에서 다양한 인기 상용 애플리케이션을 최대한 빠른 속도로 실행할 수 있도록 시스템 설정을 자동으로 조정한다. 따라서 데이터 과학자 또는 개발자의 생산성이 높아진다. 이 툴은 IT 부서를 위해 프로세서와 스토리지, 메모리, 그래픽 사용에 대한 실시간 성능 보고서도 생성한다. DOP 는 아직 리눅스에서 실행되지 않으며, AI 개발은 리눅스 기반 오픈 소스 소프트웨어로 이루어지는 경향이 있으므로 DOP 는 주로 AI 배포에 유용한다. 또한 DOP 는 워크스테이션을 정밀하게 조정하는 데 유용한 ExpressSign-in 와 Express Charge(모바일), Intelligent Audio, 보고 및 분석 툴도 제공한다.

당면 과제와 기회

기업을 위한 과제와 기회

IDC 는 AI 시장이 두 갈래로 나뉘는 분기점에 직면한 것을 목격하고 있다. 한 갈래의 기업들은 AI 를 대대적으로 도입하는 등 경쟁력을 유지하기 위해 본격적인 데이터 전략을 전개하고 있다. 예를 들면, 실제로 상위 100 대 슈퍼컴퓨터에 등록된 엔터프라이즈 AI 인프라스트럭처 제품을 사용하여 탁월한 작업을 수행한 기업이 있다. 반면, 부족한 예산과 성능이 떨어지는 하드웨어로 데이터 센터나 클라우드의 가용 서버에서 소규모 AI 이니셔티브를 시도하는 것이 일상인 기업도 있다.

많은 기업이 첫 번째 시나리오가 아닌 두 번째 시나리오가 현실인 상황이다. 이러한 기업들에게는 클라우드 인스턴스나 GPU 가속 데이터 센터 서버에 막대한 비용을 들이지 않고도 AI 데이터 과학자나 개발자가 적시에 AI 를 훈련할 수 있는 적절한 툴을 제공하는 것이 과제다. IDC 는 이러한 기업이 AI 데이터 과학자와 개발자에게 강력한 GPU 가속 워크스테이션을 제공하면 좋은 성과를 거둘 수 있다고 판단한다.

델(Dell)을 위한 과제와 기회

시장에서는 AI 개발과 배포를 위해서는 고가의 가속화된 서버 하드웨어가 필요하며, 심지어 클러스터에도 필요하다 오해가 있다. 이는 수십억 개의 매개변수가 있는 최대 규모의 AI 알고리즘에 대해서는 사실일 수 있지만, 대부분의 기업은 그렇게 거대한 알고리즘을 개발하고 있지 않다. 기업들이 AI 이니셔티브를 통해 유용하고 영향력 있으며 관리하기 쉬운 작업을 수행하는 상황에서, 많은 기업들은 워크스테이션에서 이러한 일반적인 규모의 AI 모델을 개발하고 배포할 수 있다는 사실을 인지하지 못하고 있다. 델의 과제는 이러한 선입견을 깨고 델 워크스테이션 포트폴리오의 성능과 기능을 시장에 알리는 것이다.

동시에 델은 워크스테이션이 원활하게 성능을 구현하고 시간이 지나면서 기술 병목 현상을 일으키지 않도록 해야 한다. 워크스테이션을 적절하게 사용하는 최종 사용자, 즉, 수십억 개의 매개변수 알고리즘을 실행하지 않는 사용자가 실망하는 일이 없도록 계속해서 혁신을 빠르게 추진해야 한다. 또한 갑자기 매우 빠르게 비즈니스를 확장하는 고객이나 알고리즘이 매우 커지는 고객은 워크스테이션에서 델의 AI 서버 제품군으로 원활하게 전환할 수 있다. 물론 여기에는 델이 AI 이니셔티브의 규모와 관계없이 모든 고객에게 적합한 솔루션을 제공할 기회 또한 있다.

결론

IDC 는 워크스테이션이 현재 과소평가되고 있으나, 많은 활용 사례에서 AI 개발 및 배포의 주역으로 활약하고 있다고 생각한다. IDC 는 워크스테이션이 현재 많은 사용 사례에서 AI 개발 및 배포의 핵심으로 과소평가되고 있다고 생각한다. 워크스테이션은 AI 과학자와 개발자에게 서버보다 적은 CAPEX 와 클라우드 인스턴스보다 획기적으로 낮은 OPEX 로 훨씬 더 자유롭게 AI 모델을 실험할 수 있는 강력한 GPU 가속 플랫폼을 제공하고 있다. 수십억 개의 매개변수 알고리즘이 필요하지 않은 AI 이니셔티브를 개발하는 기업(대부분의 기업)은 AI 팀에 워크스테이션을 제공하여 제약 없는 AI 개발과 간편한 엣지 기반 배포를 지원하는 것을 고려해야 한다.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com