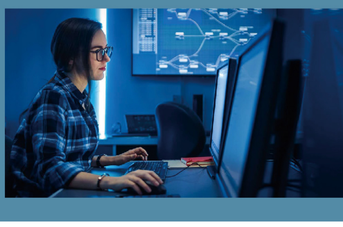
intel

강력한 조합을 이루는 HPC와 AI

머신 러닝(Machine learning)과 딥 러닝(Deep learning) 은
HPC를 보강하여 새로운 통찰력을 더 빠르게 얻을 수 있도록 합니다



HPC에 AI 도입을 통한 대용량 데이터 처리 및 심층 분석

- AI 구현은 HPC의 컴퓨팅 아키텍처와 친화력이 있으며 AI와 HPC는 모두 고성능 인텔® 하드웨어를 기반으로 한 유사한 구성의 이점을 누릴 수 있습니다.
- CERN(유럽 원자핵 공동 연구소)의 연구원들은 물리 법칙을 AI 모델에 통합하는 인텔® 지원 컨볼루션 신경망(Intel-enabled convolutional neural networks)을 사용하여 실제 사용 사례에 대해 보다 정확한 결과를 이끌어 내고 있습니다.
- 인텔®은 하드웨어 가속 AI와 oneAPI을 사용한 통합 플랫폼 프로그래밍, 그리고 시스템 설계 및 배포의 마찰을 줄이는 데 도움이 되는 간소화된 인텔®의 셀렉트 솔루션(Intel® Select Solution)을 통해 전용 AI와 HPC 플랫폼 간의 트레이드 오프(Trade off) 상황을 줄이는 데 도움을 줍니다.

하드웨어 및 소프트웨어 가속 AI가 뒷받침하는 인텔® 지원 HPC는 연구자와 데이터 과학자의 기능을 확장하고 세계 최대의 데이터 세트 및 복잡한 모델에 대한 결과와 통찰력에 대한 더 빠른 시간을 단축하고 있습니다.

AI 증강 HPC(AI-Augmented HPC)

HPC 구현에 필요한 아키텍처는 AI 구현과 많은 유사점을 가지고 있습니다. 둘 다 높은 수준의 컴퓨팅 및 스토리지, 대용량 메모리 용량 및 대역폭, 고대역폭 패브릭을 사용하여 일반적으로 크기가 증가하는 대용량 데이터 세트를 처리하여 결과를 얻습니다. 딥 러닝은 매우 큰 다차원 데이터 세트를 포함하는 HPC에서 해결한 문제와 매우 잘 어울립니다. 예를 들어, Quantifi는 인텔® 지원 AI를 사용하여 금융 시장에서 기존 방법에 비해 파생 상품 평가 속도를 700배 이상 가속화했으며, 일반적인 평가 워크로드에 대해 거의 실시간으로 결과를 제공했습니다.

HPC에서 AI의 역할은 AI 모델이 데이터 세트의 전문가 분석을 보강하여 동일한 수준의 정확도로 더 빠른 결과를 생성할 수 있다는 것입니다. 주요 HPC 사용 사례는 다음과 같은 고급 AI 기능의 이점을 누리고 있습니다.

- 위험 및 사기 탐지, 물류 및 제조와 같은 금융 서비스(FSI)에 대한 분석
- 산업 제품 설계, 전산 유체 역학(CFD), 컴퓨터 지원 엔지니어링(CAE) 및 컴퓨터 지원 설계(CAD)
- 고에너지 물리학과 같은 분야에서의 과학적 시각화 및 시뮬레이션
- 패턴 클러스터링, 생명 과학, 게놈 시퀀싱(genomic sequencing) 및 의학 연구
- 지구 과학 및 에너지 분야 탐사
- 날씨, 기상학 및 기후 과학
- 천문학 과 천체 물리학

워크로드가 변화된 방식

AI에 대한 현재 사용 사례의 대부분은 AI 객체 인식을 위해 스마트 카메라에 크게 의존하는 지능형 교통 시스템과 같은 엣지 또는 데이터 센터 배포로 제한됩니다. AI 모델을 뒷받침하는 알고리즘은 훨씬 더 복잡해졌으며 과학적 발견, 혁신, 산업 및 비즈니스 응용 분야에 대한 더 큰 잠재력뿐만 아니라 더 큰 계산 요구 사항을 제공합니다. 문제는 AI 추론을 HPC 수준으로 확장하는 방법이나 교차로에서 트래픽 패턴을 인식하는 것에서 몇 주가 아닌 몇 시간 만에 게놈을 시퀀싱하는 방법입니다.

다행스럽게도 HPC 업계에서는 더 많은 병렬 처리의 필요성, 대규모 데이터 세트를 위한 빠른 I/O, 분산 컴퓨팅 환경의 효율적인 탐색과 같은 대규모 AI의 과제를 해결하는 방법에 대한 수십 년의 경험을 제공합니다. 이와 같은 HPC 기능은 딥 러닝 추론을 통해 전문가 수준의 휴리스틱을 초당 수천 개의 트랜잭션, 워크로드 또는 시뮬레이션에 적용하는 등 유용한 결과를 얻기 위해 AI를 가속화하는 데 도움이 될 수 있습니다.

물리 정보 신경망(Physics-Informed Neural Networks, PINNs)

AI-증강 HPC 사용 사례 중 한 가지 예는 물리 법칙을 추론 모델에 통합하여 보다 사실적인 결과를 생성하는 것입니다. 이러한 응용 분야에서 신경망은 질량, 에너지 및 속도 보존과 같은 알려진 법칙을 준수해야하며 물리 정보 신경망(Physics-Informed Neural Networks PINN)이라고 합니다. PINN은 유체 흐름 분석, 분자 역학, 에어포일 및 제트 엔진 설계, 고에너지 물리학과 같은 사용 사례에 대한 HPC 모델링 및 시뮬레이션을 보강하거나 대체하는 데 사용할 수 있습니다.

예를 들어, CERN 연구원들은 입자 충돌에 대한 몬테카를로 시뮬레이션을 대체하기 위해 인텔® 제온® 스케일러블 프로세서를 탑재한 시스템에서 인텔® 딥 러닝 부스트(Intel DL Boost)를 사용했습니다. 저정밀도 int8 양자화(Low-precision int8 quantization)는 소프트웨어 시뮬레이션 대비 약간의 정확성 향상과 함께 최대 68,000배 빠른 처리가 될 수 있게 지원하였습니다.

데이터 성장에 의해 주도되는 HPC의 AI

HPC 및 AI 워크로드의 주요 동인은 데이터의 지속적인 성장과 HPC 규모의 분석 속도에 부합해야 하는 필요성입니다. AI 알고리즘은 정교함이 증가하고 있으며, 특히 딥 러닝 방법론이 도입된 이후 이전보다 훨씬 큰 데이터 세트를 처리할 수 있습니다. 게놈 시퀀싱과 같은 분야는 엄청난 양의 데이터를 생성하고 있으며, MIT와 하버드의 브로드 인스티튜트(Broad Institute of MIT)와 하버드(Harvard)와 같은 기관은 매일 약 24 테라 바이트의 새로운 데이터를 만들고 있습니다.

AI는 중요한 워크로드를 가속화하는 데 도움이 되므로, 결과를 발견하는 시간이 뒤처지지 않습니다. 예를 들어, 인텔®은 Broad Institute와 협력하여 GATK(Genomics Analytics Toolkit)를 위한 인텔® 셀렉트 솔루션(Intel Select Solution)을 개발했으며, 이 솔루션은 하드웨어 지원 AI 가속을 통합하여 주요 유전체학 툴킷에 대한 HPC 워크로드를 구동합니다. Broad Institute는 GATK Select Solution을 사용하여 BWA(Burrow-Wheeler Aligner) 애플리케이션의 경우 1.75배의 속도 향상을, HaplotypeCaller 애플리케이션의 경우 2배의 속도 향상을 달성할 수 있었습니다.

샌디에이고 슈퍼컴퓨터 센터(SDSC)는 세계에서 가장 큰 학술 데이터 센터 중 하나를 호스팅 하며 데이터 사용, 관리, 저장 및 보존 분야에서 국제적인 리더로 인정받고 있습니다. AI-focused 시스템을 통해 과학자들은 가속화된 훈련 및 추론을 위한 새로운 접근법을 개발할 수 있습니다.

AI를 HPC에 도입하는 데 있어 여러가지 도전을 극복하는 방법

AI를 위한 HPC 구성의 경우 전통적으로 CPU 아키텍처 내에서 AI와 HPC 요구 사항 간에 절충점(Trade-off)이 있습니다. AI가 많은 워크로드는 일반적으로 코어 수를 속도로 교환하는 반면, HPC의 워크로드는 종종 높은 코어 수와 더 많은 코어-투-코어 대역폭으로 더 높은 컴퓨팅 성능을 선호합니다. 세대별 개선이 계속됨에 따라 인텔®은 인텔® 제온®스케일러블 프로세서에 내장된 가속을 포함한 솔루션을 제공하고 있습니다.

하드웨어 및 소프트웨어 계층에서 다음과 같은 주요 혁신으로 AI 솔루션을 쉽게 설계하고 구축할 수 있습니다.

- 인텔® 제온® 스케일러블 프로세서는 AI 가속이 내장된 높은 수준의 AI 성능을 제공합니다. 인텔® 프로세서에 특화되었으며, 인텔® DL 부스트 벡터 신경망 명령어(VNNI)가 탑재된 인텔® AVX-512는 최적화된 AI 성능을 제공하여 짧은 시간에 빠른 인사이트를 제공합니다.
- 인텔® oneAPI AI analytics 툴킷 내의 저정밀 최적화 라이브러리를 사용하면 HPC 및 AI 플랫폼에 대한 코딩이 더 쉬워지며, 성능을 높이고 정확도 임계값을 유지할 수 있습니다.
- 머신 러닝을 위한 인텔® FPGA는 높은 병렬화를 지원하며 HPC 및 AI 워크로드에 대한 결과 및 인사이트 확보 시간을 단축하는 데 도움이 됩니다.
- 딥 러닝 프로세서 기술에 중점을 둔 인텔®의 데이터 센터 팀인 Habana Labs의 Gaudi 플랫폼을 사용하면 데이터 과학자와 기계 학습 엔지니어가 교육을 가속화하고 몇 줄의 코드로 새로운 모델을 구축하거나 기존 모델을 마이그레이션하여 생산성을 높일 수 있을 뿐만 아니라 낮은 운영 비용. Habana 가속기는 대규모 AI 모델 교육 및 추론을 위해 특별히 제작되었습니다.
- HPC AI 클러스터용 인텔® Select Solution은 GPU를 배포하지 않고 컨버지드 HPC 플랫폼에 AI 워크로드를 배포할 수 있는 경로를 제공합니다.
- AI 개발자는 HPC 클러스터에서 보다 효과적으로 실행하기 위해 기술과 코드를 개선하고 있습니다. 새로운 최적화는 데이터 로드에서 사전 처리, 교육 및 추론에 이르기까지 워크로드를 종단 간 가속화하고 있습니다.

복잡성은 HPC 및 AI 채택에 있어 마찰의 주요 원인이기도 합니다. 필요한 기술 세트는 도메인별로 매우 다르며, 기업은 성공하려면 HPC 및 AI에서 훈련된 인재를 확보해야 합니다. 인텔의 업계 리더십은 인텔이 HPC 및 AI 커뮤니티와 긴밀하게 협력하여 전문 지식과 아이디어를 공유함으로써 길을 닦는 데 도움이 될 수 있습니다.

결론: HPC에 AI 인텔리전스를 도입하는 것은 큰 흐름

AI는 빠른 발견과 통찰력을 위해 AI 분석의 속도와 규모를 증가시키는 새로운 기술과 방법론으로 HPC 애플리케이션에 점점 더 많이 사용되고 있습니다. 이러한 혁신을 통해 데이터 과학자와 연구원은 AI에 의존하여 더 많은 데이터를 처리하고, 보다 사실적인 시뮬레이션을 생성하고, 종종 더 짧은 시간에 더 정확한 예측을 수행할 수 있습니다.

(출처: 인텔 홈페이지 - https://www.intel.com/content/www/us/en/high-performance-computing/hpc-artificial-intelligence.html)