OpenShift Container Platform 4.17

하드웨어 가속기

하드웨어 가속기

Last Updated: 2024-12-04

OpenShift Container Platform 4.17 하드웨어 가속기

하드웨어 가속기

법적 공지

초록

이 문서에서는 인공 지능 및 머신 러닝(AI/ML) 애플리케이션을 생성하기 위해 제공된 하드웨어 가속 기능을 위해 Red Hat OpenShift AI에서 지원하는 GPU Operator를 설치 및 구성하는 방법을 설명합니다.

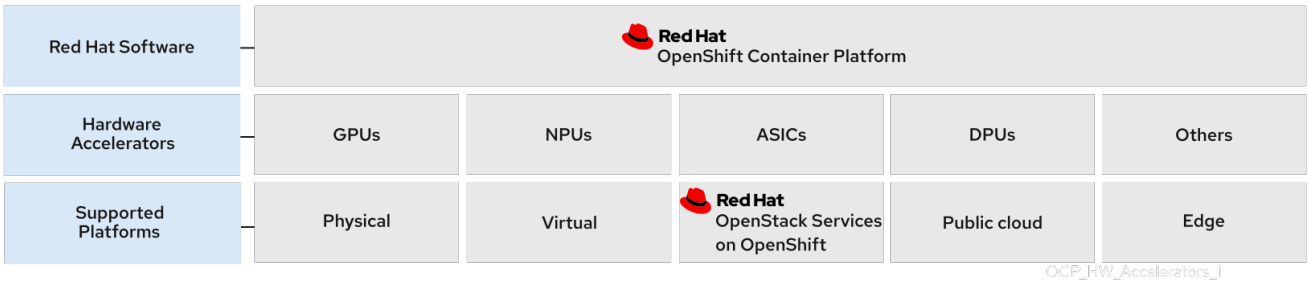
차례

1장. 하드웨어 가속기 정보

특수 하드웨어 가속기는 새롭게 등장하는 인공 지능 및 머신 러닝(AI/ML) 업계에서 중요한 역할을 합니다. 특히 하드웨어 가속기는 이 새로운 기술을 지원하는 대규모 언어 및 기타 기본 모델을 교육하고 제공하는 데 필수적입니다. 데이터 과학자, 데이터 엔지니어, ML 엔지니어 및 개발자는 데이터 집약적인 변환 및 모델 개발 및 서비스를 위한 특수 하드웨어 가속을 활용할 수 있습니다. 이러한 에코시스템의 대부분은 오픈 소스이며, 여러 파트너 및 오픈 소스 기반이 포함됩니다.

Red Hat OpenShift Container Platform은 하드웨어 가속기를 구성하는 처리 장치를 추가하는 카드 및 주변 하드웨어를 지원합니다.

- 그래픽 처리 단위(GPU)
- 통신 처리 단위(NPU)
- Application-specific integrated circuit (ASIC)
- 데이터 처리 단위(DPU)



특수 하드웨어 가속기는 AI/ML 개발을 위한 다양한 이점을 제공합니다.

모두를 위한 하나의 플랫폼

개발자, 데이터 엔지니어, 데이터 과학자 및 DevOps를 위한 협업 환경

Operator를 통한 확장 기능

Operator를 사용하면 OpenShift Container Platform에 AI/ML 기능을 가져올 수 있습니다.

하이브리드 클라우드 지원

모델 개발, 제공 및 배포를 위한 온프레미스 지원

AI/ML 워크로드 지원

모델 테스트, 반복, 통합, 승격 및 프로덕션을 서비스로 제공

Red Hat은 Linux(커널 및 사용자 공간) 및 Kubernetes 계층에서 RHEL(Red Hat Enterprise Linux) 및 OpenShift Container Platform 플랫폼에서 이러한 특수 하드웨어 가속기를 활성화하는 최적화된 플랫폼을 제공합니다. 이를 위해 Red Hat은 Red Hat OpenShift AI와 Red Hat OpenShift Container Platform의 검증된 기능을 엔터프라이즈급 AI 애플리케이션 플랫폼에 결합합니다.

하드웨어 Operator는 Kubernetes 클러스터의 운영 프레임워크를 사용하여 필요한 가속기 리소스를 활성화합니다. 제공된 장치 플러그인을 수동으로 또는 데몬 세트로 배포할 수도 있습니다. 이 플러그인은 클러스터에 GPU를 등록합니다.

특정 특수 하드웨어 가속기는 개발 및 테스트를 위해 보안 환경을 유지해야 하는 연결이 끊긴 환경에서 작동하도록 설계되었습니다.

1.1. 하드웨어 가속기

Red Hat OpenShift Container Platform은 다음과 같은 하드웨어 가속기를 활성화합니다.

- NVIDIA GPU
- AMD Instinct® GPU
- Intel® Gaudi®

추가 리소스

- Red Hat OpenShift AI 소개
- Red Hat OpenShift Container Platform의 NVIDIA GPU Operator
- AMD Instinct Accelerators
- Intel Gaudi AI Accelerators