능동적 즉시 대응 추론 서비스를 위한 경량 배포 프레임워크 기술의 설계 및 개발

박종빈, 박효찬

한국전자기술연구원

jpark@keti.re.kr

Design and development of a lightweight deployment framework technology for active response services

Jongbin Park, Bak Hyo-Chan*

Korea Electronics Technology Institute

요 약

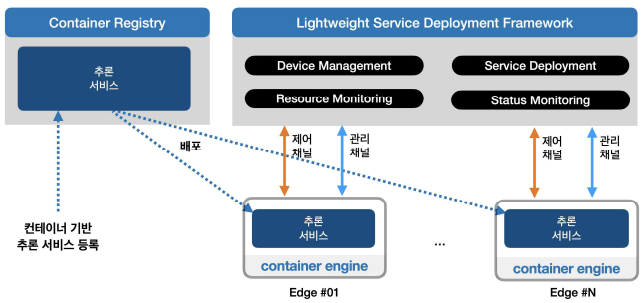
본 논문에서는 대상 에지 기기에 컨테이너 이미지로 패키징한 추론 서비스를 수행이 필요한 시점에 신속하고 간편하게 배포할 수 있게 하는 경량 프레임워크 기술을 설계하고 구현한다. 개발 기술을 사용하면 제어 노드와 연결된 이기종 에지 노드로 추론서비스를 전달하고 실행할 수 있다. 앤서블(Ansible)이 설치된 제어 노드는 에지 노드를 Secure Shell(SSH)을 통해 사전에 부여된 권한 만큼 제어할 수 있으며, 에지 노드는 제어 노드로부터 전달받은 정보를 기초로 컨테이너 이미지를 내려받고 이를 실행하는 과정을 거친다. 제안 기술은 기존에 제안되었던 클라우드-엣지 연동분석 프레임워크의 경량화된 버전으로써 프로메테우스 기반 모니터링 기능, 컨테이너 이미지 빌드 기능 들을 제거하고, 추론 서비스 배포 기능에 집중하여 가벼우면서 단일화된 웹 애플리케이션 형태로 구동할 수 있게 한 것이 주요 특징이다.

Ⅰ. 서 론

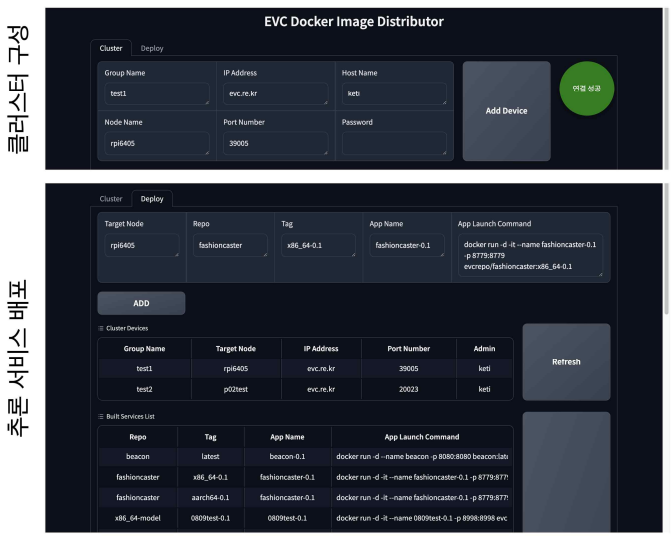
AI(Artificial intelligence) 기술의 발전과 함께 다양한 추론 서비스들이 AI 모델을 탑재함으로써 종래의 서비스 한계를 크게 개선하고 사용자 만족도를 높이고 있다. 또한 추론 서비스를 운용하는 대상이 종래에는 주로 클라우드 환경에서 이뤄졌으나, 컴퓨팅 자원의 성능이 향상되고 비용이 저렴해짐에 따라 현장에서는 에지 컴퓨팅(Edge computing) 환경을 사용하는 사례가 증가하고 있다 [1]. 다만, 에지 기기들은 물리적으로 분산 배치된 경우가 많고 종류도 다양함에 따라 개발한 서비스를 능동적으로 즉시 배포하는 작업은 여전히 도전적이다. 따라서 본 논문에서는 서비스 운용이 필요한 대상 에지 기기에 컨테이너 이미지로 패키징한 추론 서비스를 신속하고 간편하게 배포할 수 있게 하는 경량 프레임워크 기술을 설계하고 구현한다. 제안 기술은 [2][3]에서 제안한 클라우드-엣지 연동분석 프레임워크의 경량화된 버전으로써 프로메테우스 기반 모니터링 기능, 컨테이너 이미지 빌드 기능, 쿠버네티스 연동 기능들을 제거하고, 컨테이너 이미지를 통한 추론 서비스 배포 기능에 집중하여 단일한 웹 애플리케이션 형태로 구동할 수 있게 한 것이 가장 큰 특징이다.

Ⅱ. 본론

그림 1은 이러한 경량 배포 프레임워크 기반 배포 프로세스를 보여준다. 사용자가 컨테이너 레지스트리(Container Registry)에 Docker [4]와 같은 컨테이너 이미지로 추론 서비스를 패키징하여 등록하면, 제안한 경량 배포 프레임워크는 이를 대상 에지 기기에 전달하고, 필요시 서비스를 실행한다. 배포 프레임워크는 일종의 제어 노드로써 앤서블(Ansible) [5]이 사전에 설치되어 있다. 개발한 배포 프레임워크 기술은 내부적으로 Ansible을 호출하여 에지 기기를 Secure Shell (SSH) 방식으로 제어하며, 원격지의 컨테이너 레지스트리(Container registry)에 등록된 추론서비스를 내려받고 실행할 수 있게 하는 사용자 인터페이스를 제공한다.



<그림 1> 경량 프레임워크 기반 배포 프로세스



<그림 2> 개발한 경량 프레임워크 사용자 인터페이스

그림 2는 이러한 경량 프레임워크 사용자 인터페이스 예시를 나타낸다. 그라디오(gradio) [6] 프레임워크 기술과 python 언어로 웹 기반 그래픽 애플리케이션을 구현했으며, 핵심 기능으로 클러스터 구성과 추론 서비스 배포 기능을 포함한다. 클러스터 구성 단계에서는 서비스를 배포하고자 하는 대상 에지 기기 정보를 등록하고, 앤서블(Ansible)을 통한 제어권 설정 처리를 진행하며, 연결이 정상적으로 확인되면 서비스를 배포할 수 있는 대상 에지 기기 목록에 등록한다. 에지 기기 목록은 sqlite 데이터베이스로 관리하며, 서비스 배포 영역에서 DB에 접속하여 기기 목록과 상태를 확인한다. 기기 정보 이외에 배포할 수 있는 서비스 정보는 다양한 컨테이너가 등록된 컨테이너 레지스트리에서 전달받아 목록으로 표시한다. 사용자는 배포하려는 서비스와 운용 대상 에지 기기를 선택하고, 배포 버튼을 클릭하여 대상 기기에 서비스를 로딩하고 실행한다.

Ⅲ. 결론

본 논문에서는 사용자가 설정한 대상 에지 기기에 컨테이너 이미지로 패키징한 추론 서비스를 신속하고 간편하게 배포할 수 있게 하는 경량 프레임워크 기술을 설계하고 구현했다. 종래의 클라우드-엣지 연동분석 프레임워크에서 서비스 배포와 실행 기능에 집중하여 경량화를 추구했다.

ACKNOWLEDGMENT

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-00907, 능동적 즉시 대응 및 빠른 학습이 가능한 적응형 경량 엣지 연동분석 기술개발).

참 고 문 헌

[1] K. Cao, Y. Liu, G. Meng, Q. Sun, “An overview on edge computing research,” IEEE access, vol. 8, pp. 85714-85728. 2020.

[2] 박종빈, 박효찬, “에지 클러스터에 추천서비스를 배포하는 파이프라인 기술의 설계 및 개발,” 대한전자공학회 ICS 2024 정보 및 제어 심포지엄, pp. 64-65, 2024.

[3] 송무현, 김규민, 문지훈, 김유림, 남채원, 박종빈, 이경용, “KubEVC-Agent : 머신러닝 추론 엣지 컴퓨팅 클러스터 관리 자동화 시스템,” 대한임베디드공학회논문지, vol. 18, no. 6, pp. 293-301, 2023.

[4] Docker URL, “www.docker.com”

[5] Ansible URL, “https://www.ansible.com/”

[6] Gradio URL, “https://www.gradio.app/”