쿠버네티스 클러스터 환경에서 분산 AI 애플리케이션의 SLO를 효율적으로 지원하는 모니터링 시스템

*김재환 김경훈 노재춘 박성순

세종대학교, 글루시스

*bibe0829@gmail.com ystar001@gmail.com jano@sejong.ac.kr sspark@gluesys.com

A monitoring system that efficiently supports SLO of distributed AI applications in Kubernetes cluster environment

*Jaehwan Kim, Gyeonghoon Kim, Jaechun No,

Seongsoon Park Sejong University, Gluesys

요약

쿠버네티스는 컨테이너를 사용하는 분산 클라우드에서 컨테이너화를 쉽고 빠르게 배포/확장할 수 있어 유용한 플랫폼이다.쿠버네티스에서 다양한 애플리케이션들이 동작하며 서비스를 제공하고 있다. 서비스의 원활한 제공을 위하여 고객과 서비스수준에 대한 약속인 SLA와 SLA의 기준이 되는 SLO에 필요한 지표를 확인하는 것은 중요하다.

본 논문은 쿠버네티스 클러스터로 구성된 분산 클라우드 DECENTER를 소개하고 DECENTER에서 분산 AI 애플리케이션의 효율적인 SLO를 지원하는 모니터링 시스템을 제안한다.

1. 서론

분산 클라우드는 클라우드 네이티브 컴퓨팅 기술을 이용하여 서로 다른 위치에 있는 데이터들과 애플리케이션들을 연결한다. 클라우드 네이티브 컴퓨팅은 클라우드 컴퓨팅 기술에 컨테이너 기술을 접목한 새로운 컴퓨팅 기술이다. 쿠버네티스[1]는 컨테이너화를 쉽고 빠르게 배포/확장하고 관리를 자동화해주는 플랫폼이로 분산 클라우드에서 유용하다.

분산 클라우드에서 애플리케이션들은 주로 마이크로서비스로 구성된다. 마이크로서비스의 예로 웹 스트리밍 서비스인 넷플릭스가 있다. 넷플릭스를 사용하는 사용자들은 실시간으로 영상을 보기 때문에스트리밍 서비스가 조금이라도 지연된다면 서비스의 신뢰도가 낮아질것이다. 신뢰도 유지를 위해 넷플릭스 관리자는 서비스 수준에 대한 약속인 SLA의 관련 지표를 모니터링 하는 것이 중요하다.

본 논문은 쿠버네티스 클러스터 환경인 DECENTER를 소개하고DECENTER에서 분산 AI 애플리케이션의 효율적인 SLO를 지원하기위한 시스템을 제안한다.

2. 관련연구

2-1 마이크로서비스

마이크로서비스는 소프트웨어를 만들기 위한 아키텍처이자 하나의 접근 방식이다. 일반적인 애플리케이션들은 모든 요소를 넣는 전통적인 모놀리식 접근 방식이었다. 반면 마이크로서비스 아키텍처는 애플리케이션을 상호 독립적인 최소 구성 요소로 분할하여 모든 요소가독립적이며 연동되어 동일한 작업을 수행한다. 애플리케이션의 상호독립적인 최소 구성 요소 하나하나를 마이크로서비스라 한다. 일반적인 애플리케이션은 하나의 기능에 문제가 생기면 애플리케이션 전체에 문제가 생겨 애플리케이션이 동작하지 못하며 문제가 있는 부분을찾기도, 고치기도 힘들다. 반면 마이크로서비스로 이루어진 분산형 애플리케이션은 마이크로서비스 하나하나가 상호 독립적으로 동작하여마이크로서비스 하나가 문제가 생기더라도 문제가 있는 마이크로서비스의 기능이 동작하지 않을 뿐 분산형 애플리케이션은 계속하여 동작한다. 또한 문제가 생기더라도 어느 부분이 문제인지 파악하기가 용이하다.

2-2 SLI, SLO, SLA

SLI는 서비스에 대한 수준을 측정하여 정량적으로 정의한 지표이다. Monitoring을 통해 얻을 수 있는 모든 지표를 취급할 필요는 없으며 너무 많은 지표를 선택하게 되면 중요한 것을 놓칠 수 있고 집중하기 어려울 수 있다. 적절한 SLI의 선정을 위해 다양한 시스템 환경 중에서 크게 3가지로 나누어보았다. 첫 번째로 사용자가 직접 대면하는시스템인 경우 가용성, 응답 시간, 처리량이 중요 지표이고, 두 번째로저장소 시스템은 응답 시간, 가용성, 내구성이 중요 지표이고, 마지막으로 빅데이터 시스템은 처리량, 종단간 응답 시간이 중요 지표이다.

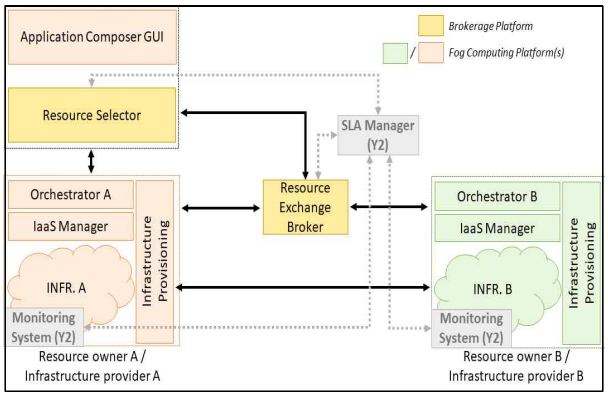
SLO는 고객과 서비스 수준에 대한 목표이고 SLI에 의해 측정된 서비스 수준의 목표 값 또는 일정 범위의 값을 의미한다. SLO는 $LI \leq$ 표치 혹은 최솟값$\leq LI \leq$최댓값으로 표현할 수 있다. 명확성을 극대화하기 위해 SLO는 측정 방식과 유효한 기준이 반드시 명시되어야 한다. SLA는 고객과 서비스 수준에 대한 약속이며 SLO를기준으로 작성한다.

3. 모니터링 시스템

3-1 DECNENTER

DECENTER는 분산 AI 애플리케이션을 지원하는 분산 클라우드시스템이며 쿠버네티스 클러스터 환경으로 [그림 1]과 같이 구성되어져 있다.

DECENTER의 Use Case로 Smart City에서 횡단보도를 건너는보행자의 안전을 알려주는 시스템이 있다. 이 시스템은 보행자가 횡단보도를 건널 때 자동으로 주변 이미지와 소리, 날씨 정보를 IoT 장비로수집하여 분산 AI 애플리케이션으로 보행자의 위험도를 측정하고 보행자에게 알린다.



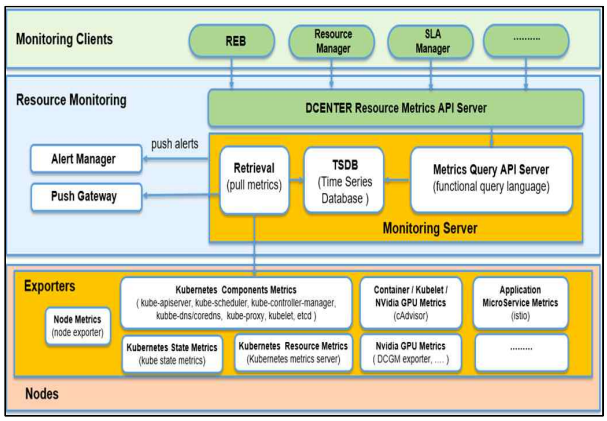
[그림 1] DECENTER 구조

3-2 모니터링 시스템

DECENTER에 있는 서비스들은 실시간 정보를 통하여 동작한다.서비스들이 지속적으로 동작하기 위해서는 분산 AI 애플리케이션의SLA에 필요한 SLO 지표를 확인할 수 있어야한다.

DECENTER의 모니터링 시스템은 SLO 지표 중 쿠버네티스 클러스터의 노드와 컨테이너 정보, 노드와 컨테이너 상태 지표를 얻기 위해 Prometheus[1]를 사용하였고 분산 AI 애플리케이션을 구성하는 마이크로서비스들의 응답시간 등 지표를 얻기 위하여 Istio[1]를 사용하였다.

[그림 2]를 보면, 쿠버네티스 클러스터 노드에 여러 지표를 수집하는 Exporter들을 설치한다. Monitoring Server는 Exporter가 제공하는 HTTP End Point인 /metrics을 통해 지표들을 수집하고TSDB(Time Series Database)에 저장한다.
Prometheus 사용자들은 TSDB에 저장된 지표를 Query로 가져올 수 있다. AlertManager는Query를 통하여 가져온 지표들이 특정 상황이 됐을 경우 알림을 보낸다. Monitoring Client들은 SLO에 필요한 지표를 DECENTER Resource Metrics API Server를 통하여 가져올 수 있고 가져온 지표로 서비스를 확인할 수 있다.



[그림 2] 모니터링 시스템 구조

4. 결론

본 논문은 쿠버네티스 클러스터 환경인 DECENTER를 소개하고 DECENTER에서 분산 AI 애플리케이션의 효율적인 SLO를 지원하는모니터링 시스템을 제안했다. 향후 다양한 상황에서의 시스템 검증이진행될 예정이다.

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 1711075689, AI 애플리케이션을 지원하는 IoT 연동 분산 Edge 클라우드 기술 개발)

참고문헌

[1] https://kubernetes.io/ko/docs/concepts/overview/[2] https://prometheus.io/docs/introduction/overview/[3] https://istio.io/docs/concepts/what-is-istio/