



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2020년06월05일

(11) 등록번호 10-2116244

(24) 등록일자 2020년05월22일

(51) 국제특허분류(Int. Cl.)

H04L 12/24 (2006.01)

(52) CPC특허분류

H04L 41/147 (2013.01)

H04L 41/0896 (2013.01)

(21) 출원번호 10-2018-0166297

(22) 출원일자 2018년12월20일

심사청구일자 2018년12월20일

(56) 선행기술조사문헌

KR1020160073892 A

KR1020170099710 A

(73) 특허권자

경희대학교 산학협력단

경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)

(72) 발명자

허의남

경기도 용인시 기흥구 이현로29번길 86-21 201동401호(보정동, e편한세상대림아파트)

손동영

서울특별시 양천구 목동동로 130, 1412동 104호

(74) 대리인

김홍석

전체 청구항 수 : 총 21 항

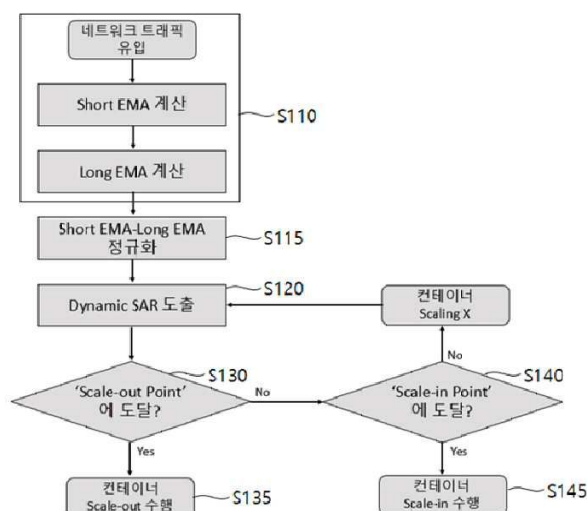
심사관 : 윤태섭

(54) 발명의 명칭 네트워크 트래픽 예측 기반의 효율적인 컨테이너 오토 스케일링 방법 및 장치

(57) 요약

본 발명에 따른 네트워크 트래픽 예측 기반의 효율적인 컨테이너 오토 스케일링 방법이 제공된다. 상기 방법은 오토 스케일링 장치에 의해 수행되고, 상기 방법은, 유입된 네트워크 트래픽에 대하여 단기지수이동평균(Short EMA)과 장기지수이동평균(Long EMA)을 계산하는 단기/장기지수이동평균 계산 과정; 상기 Short EMA와 상기 Long EMA의 차이 값을 정규화하는 정규화 과정; 상기 정규화된 차이 값을 동적 변수로 하여 동적 SAR (Stop and Reversal) 값을 도출하는 동적 SAR 도출 과정; 및 상기 동적 SAR 값에 기반하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하였는지를 판단하는 스케일 아웃 지점 도달 판단 과정을 포함한다. 이에 따라, 효율적으로 네트워크 트래픽을 예측할 수 있고, 서비스 제공 도중 품질 저하를 일으키지 않고 사용자들에게 서비스를 제공할 수 있다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 1711065146

부처명 과학기술정보통신부

연구관리전문기관 정보통신기술진흥센터

연구사업명 SW컴퓨팅산업원천기술개발사업

연구과제명 (대학ICT기초연구실) 서비스 이동 지원을 위한 분산형 클라우드 핵심원천기술 연구

기 여 율 1/1

주관기관 경희대학교 산학협력단

연구기간 2018.01.01 ~ 2018.12.31

명세서

청구범위

청구항 1

네트워크 트래픽 예측 기반의 효율적인 컨테이너 오토 스케일링 방법에서,

유입된 네트워크 트래픽에 대하여 단기지수이동평균(Short EMA)과 장기지수이동평균(Long EMA)을 계산하는 단기/장기지수이동평균 계산 과정;

상기 Short EMA와 상기 Long EMA의 차이 값을 정규화하는 정규화 과정;

상기 정규화된 차이 값을 동적 변수로 하여 동적 SAR (Stop and Reversal) 값을 도출하는 동적 SAR 도출 과정; 및

상기 동적 SAR 값에 기반하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하였는지를 판단하는 스케일 아웃 지점 도달 판단 과정을 포함하는, 오토 스케일링 방법.

청구항 2

제1 항에 있어서,

상기 유입된 네트워크 트래픽의 양은 상기 에지 클라우드 내의 특정 콘텐츠를 요청하는 사용자 단말의 수와 상기 사용자 단말 별로 요청된 특정 콘텐츠의 용량에 기반하여 결정되고,

상기 트래픽이 상기 스케일 아웃 지점에 도달하면 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시키는 스케일 아웃 수행 과정을 더 포함하는, 오토 스케일링 방법.

청구항 3

제1 항에 있어서,

t+1 번째 시각에서의 상기 동적 SAR (Dynamic SAR) 값은,

$\text{Dynamic SAR} = \text{sar}[t+1] = \text{sar}[t] + (\text{af} + \text{Dynamic af}) * (\text{ep} - \text{sar}[t])$ 으로 도출되고,

af는 비례상수이고, Dynamic af는 상기 정규화된 차이 값과 연관된 상기 동적 변수로서 네트워크 트래픽이 유입될 때마다 새로운 값으로 갱신되고,

상기 동적 변수인 Dynamic af에 의해 상기 갱신된 동적 SAR 값에 기반하여, 상기 에지 클라우드에서 상기 스케일 아웃을 수행하고 상기 컨테이너의 수를 증가시키는 것을 특징으로 하는, 오토 스케일링 방법.

청구항 4

제2 항에 있어서,

상기 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하는 스케일 인 지점 도달 판단 과정을 더 포함하는, 오토 스케일링 방법.

청구항 5

제4 항에 있어서,

상기 스케일 인 지점에 도달하지 않은 경우, 상기 컨테이너의 수를 유지하면서 상기 동적 SAR 도출 과정을 반복하는 것을 특징으로 하는, 오토 스케일링 방법.

청구항 6

제4 항에 있어서,

상기 스케일 아웃 지점과 상기 스케일 인 지점은 상기 컨테이너의 수, 사용자 단말로부터 요청된 트래픽의 지연

민감 트래픽(delay sensitive traffic) 여부 및 최대 지연 허용 시간에 따라 동적으로 변경되고,

상기 최대 지연 허용 시간이 감소함에 따라 상기 스케일 아웃 지점의 임계값을 상기 최대 지연 허용 시간과 제1 비례 상수에 기반하여 낮게 변경하고, 상기 스케일 인 지점의 제2 임계값을 상기 최대 지연 허용 시간과 제2 비례 상수에 기반하여 낮게 변경하고,

상기 제2 비례 상수는 상기 제1 비례 상수보다 더 크게 설정되어 상기 스케일 인 지점이 더 낮게 변경되는, 오토 스케일링 방법.

청구항 7

네트워크 트래픽 예측 기반의 효율적인 컨테이너 오토 스케일링 방법에서,

에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하는 스케일 인 지점 도달 판단 과정;

상기 스케일 인 지점에 도달하지 않은 경우, 단기지수이동평균(short EMA)과 장기지수이동평균(Long EMA)의 차이값을 정규화하고, 정규화된 차이값을 동적변수로 하여 복수의 컨테이너를 기반으로 동적 SAR (Stop and Reversal) 값을 도출하는 동적 SAR 도출 과정; 및

상기 동적 SAR 값에 기반하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하였는지를 판단하는 스케일 아웃 지점 도달 판단 과정을 포함하는, 오토 스케일링 방법.

청구항 8

제7 항에 있어서,

상기 트래픽이 상기 스케일 아웃 지점에 도달하면 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시키는 스케일 아웃 수행 과정을 더 포함하는, 오토 스케일링 방법.

청구항 9

제7 항에 있어서,

t+1 번째 시각에서의 상기 동적 SAR (Dynamic SAR) 값은,

Dynamic SAR = sar[t+1] = sar[t] + (af+Dynamic af) * (ep-sar[t])으로 도출되고,

af는 비례상수이고, Dynamic af는 상기 정규화된 차이 값과 연관된 상기 동적 변수로서 네트워크 트래픽이 유입될 때마다 새로운 값으로 갱신되고,

상기 동적 변수인 Dynamic af에 의해 상기 갱신된 동적 SAR 값에 기반하여, 상기 에지 클라우드에서 상기 스케일 아웃을 수행하고 상기 컨테이너의 수를 증가시키는 것을 특징으로 하는, 오토 스케일링 방법.

청구항 10

제8 항에 있어서,

상기 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하는 스케일 인 지점 도달 판단 과정을 반복하고,

상기 스케일 인 지점에 도달하지 않은 경우, 상기 컨테이너의 수를 유지하면서 상기 동적 SAR 도출 과정을 반복하는 것을 특징으로 하는, 오토 스케일링 방법.

청구항 11

제7 항에 있어서,

상기 스케일 아웃 지점과 상기 스케일 인 지점은 상기 컨테이너의 수, 사용자 단말로부터 요청된 트래픽의 지연 민감 트래픽(delay sensitive traffic) 여부 및 최대 지연 허용 시간에 따라 동적으로 변경되고,

상기 최대 지연 허용 시간이 감소함에 따라 상기 스케일 아웃 지점의 임계값을 상기 최대 지연 허용 시간과 제1 비례 상수에 기반하여 낮게 변경하고, 상기 스케일 인 지점의 제2 임계값을 상기 최대 지연 허용 시간과 제2 비

레 상수에 기반하여 낮게 변경하고,

상기 제2 비례 상수는 상기 제1 비례 상수보다 더 크게 설정되어 상기 스케일 인 지점이 더 낮게 변경되는, 오토 스케일링 방법.

청구항 12

네트워크 트래픽 예측 기반의 컨테이너 오토 스케일링 장치에서,

복수의 사용자 단말로부터 서비스 요청을 수신하고, 상기 서비스 요청에 기반하여 에지 클라우드에 접속하는 상기 사용자 단말의 서비스 품질을 측정하는 사용자 단말 제어부; 및

유입된 네트워크 트래픽에 대하여 단기지수이동평균(Short EMA)과 장기지수이동평균(Long EMA)을 계산하고,

상기 Short EMA와 상기 Long EMA의 차이 값을 정규화하고,

상기 정규화된 차이 값을 동적 변수로 하여 동적 SAR (Stop and Reversal) 값을 도출하고,

상기 동적 SAR 값에 기반하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하였는지를 판단하는 스케일링 제어부를 포함하는, 오토 스케일링 장치.

청구항 13

제12 항에 있어서,

상기 유입된 네트워크 트래픽의 양은 상기 에지 클라우드 내의 특정 콘텐츠를 요청하는 사용자 단말의 수와 상기 사용자 단말 별로 요청된 특정 콘텐츠의 용량에 기반하여 결정되고,

상기 스케일링 제어부는,

상기 트래픽이 상기 스케일 아웃 지점에 도달하면 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시키는, 오토 스케일링 장치.

청구항 14

제12 항에 있어서,

t+1 번째 시각에서의 상기 동적 SAR (Dynamic SAR) 값은,

Dynamic SAR = sar[t+1] = sar[t] + (af+Dynamic af) * (ep-sar[t])으로 도출되고,

af는 비례상수이고, Dynamic af는 상기 정규화된 차이 값과 연관된 상기 동적 변수로서 네트워크 트래픽이 유입될 때마다 새로운 값으로 갱신되고,

상기 동적 변수인 Dynamic af에 의해 상기 갱신된 동적 SAR 값에 기반하여, 상기 에지 클라우드에서 상기 스케일 아웃을 수행하고 상기 컨테이너의 수를 증가시키는 것을 특징으로 하는, 오토 스케일링 장치.

청구항 15

제13 항에 있어서,

상기 스케일링 제어부는,

상기 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하고,

상기 스케일 인 지점에 도달하지 않은 경우, 상기 컨테이너의 수를 유지하면서 상기 동적 SAR 값을 도출하는 동작을 반복하는 것을 특징으로 하는, 오토 스케일링 장치.

청구항 16

제13 항에 있어서,

상기 스케일링 제어부는,

상기 컨테이너의 수, 사용자 단말로부터 요청된 트래픽의 지연 민감 트래픽(delay sensitive traffic) 여부 및

최대 지연 허용 시간에 따라 상기 스케일 아웃 지점과 상기 스케일 인 지점을 동적으로 변경하고,

상기 최대 지연 허용 시간이 감소함에 따라 상기 스케일 아웃 지점의 임계값을 상기 최대 지연 허용 시간과 제1 비례 상수에 기반하여 낮게 변경하고, 상기 스케일 인 지점의 제2 임계값을 상기 최대 지연 허용 시간과 제2 비례 상수에 기반하여 낮게 변경하고,

상기 제2 비례 상수는 상기 제1 비례 상수보다 더 크게 설정되어 상기 스케일 인 지점이 더 낮게 변경되는, 오토 스케일링 장치.

청구항 17

네트워크 트래픽 예측 기반의 컨테이너 오토 스케일링 장치를 포함하는 네트워크 시스템에서,

복수의 사용자 단말로부터 서비스 요청을 수신하고, 상기 서비스 요청에 기반하여 에지 클라우드에 접속하는 상기 사용자 단말의 서비스 품질을 측정하는 사용자 단말 제어부;

복수의 컨테이너를 포함하고, 상기 복수의 컨테이너 별로 콘텐츠를 동적으로 저장하도록 구성된 에지 클라우드; 및

상기 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하고,

상기 스케일 인 지점에 도달하지 않은 경우, 단기지수이동평균(short EMA)과 장기지수이동평균(Long EMA)의 차이값을 정규화하고, 정규화된 차이값을 동적변수로 하여 복수의 컨테이너를 기반으로 동적 SAR (Stop and Reversal) 값을 도출하고,

상기 동적 SAR 값에 기반하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하였는지를 판단하는 오토 스케일링 장치를 포함하는, 네트워크 시스템.

청구항 18

제17 항에 있어서,

상기 오토 스케일링 장치는,

상기 트래픽이 상기 스케일 아웃 지점에 도달하면 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시키는, 네트워크 시스템.

청구항 19

제17 항에 있어서,

t+1 번째 시각에서의 상기 동적 SAR (Dynamic SAR) 값은,

Dynamic SAR = sar[t+1] = sar[t] + (af+Dynamic af) * (ep-sar[t])으로 도출되고,

af는 비례상수이고, Dynamic af는 상기 정규화된 차이 값과 연관된 상기 동적 변수로서 네트워크 트래픽이 유입될 때마다 새로운 값으로 갱신되고,

상기 동적 변수인 Dynamic af에 의해 상기 갱신된 동적 SAR 값에 기반하여, 상기 에지 클라우드에서 상기 스케일 아웃을 수행하고 상기 컨테이너의 수를 증가시키는 것을 특징으로 하는, 네트워크 시스템.

청구항 20

제18 항에 있어서,

상기 오토 스케일링 장치는,

상기 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하는 스케일 인 지점 도달 판단 과정을 반복하고,

상기 스케일 인 지점에 도달하지 않은 경우, 상기 컨테이너의 수를 유지하면서 상기 동적 SAR 도출 과정을 반복하는 것을 특징으로 하는, 네트워크 시스템.

청구항 21

제17 항에 있어서,

상기 오토 스케일링 장치는,

상기 컨테이너의 수, 사용자 단말로부터 요청된 트래픽의 지연 민감 트래픽(delay sensitive traffic) 여부 및 최대 지연 허용 시간에 따라 상기 스케일 아웃 지점과 상기 스케일 인 지점을 동적으로 변경하고,

상기 최대 지연 허용 시간이 감소함에 따라 상기 스케일 아웃 지점의 임계값을 상기 최대 지연 허용 시간과 제1 비례 상수에 기반하여 낮게 변경하고, 상기 스케일 인 지점의 제2 임계값을 상기 최대 지연 허용 시간과 제2 비례 상수에 기반하여 낮게 변경하고,

상기 제2 비례 상수는 상기 제1 비례 상수보다 더 크게 설정되어 상기 스케일 인 지점이 더 낮게 변경되는, 네트워크 시스템.

발명의 설명

기술 분야

[0001] 본 발명은 네트워크 트래픽 예측 기반의 컨테이너 오토 스케일링 방법 및 장치에 관한 것이다. 보다 상세하게는, 네트워크 트래픽 예측 기반의 동적 SAR 산출과 이에 따른 효율적인 컨테이너 오토 스케일링 방법 및 장치에 관한 것이다.

배경 기술

[0002] 기존의 컨테이너 오토 스케일링 발명은 급격하게 값이 요동치는 네트워크 트래픽을 감당하지 못하여서 본래의 감당 가능한 네트워크 트래픽 양보다 더 많은 네트워크 트래픽을 맞이하곤 했다. 그 결과, 서비스 품질이 떨어지는 문제가 발생하였다. 뿐만 아니라 네트워크 트래픽이 급격하게 감소하는 경우, 이에 빠르게 대응하지 못하여 유입되는 네트워크 트래픽의 양이 적음에도 불구하고 컨테이너의 수를 유지하여 불필요하게 리소스 자원을 낭비하는 문제점이 있었다.

[0003] 이러한 문제를 해결하기 위해 EMA(지수이동평균)과 Parabolic SAR 등 기존에 존재하던 예측 지표를 사용할 수 있다. 하지만, 기존 방법들 역시 정확하게 네트워크 트래픽을 예측을 하지 못하는 문제점이 있었다. EMA(지수이동평균)의 경우, 짧은 기간 동안의 이동평균을 적용할 경우, 조금의 값의 변동에도 예측에 많은 영향을 주어 거짓 예측을 알려주는 문제가 발생하였고, 긴 기간 동안의 이동평균을 적용할 경우, 값의 추세를 실제보다 느리게 반영하는 문제가 발생하였다.

[0004] Parabolic SAR (Stop and Reversal)의 경우, 가속도 변수에 따라 실제 값의 추세를 전혀 예측하지 못하거나, 느리게 추세를 예측하여서 실제 값의 추세보다 추세전환이 늦어지는 문제가 발생하였다. 이처럼 기존의 방법으로는 값 또는 값의 흐름을 예측하는데 느리다는 한계점이 분명하게 존재하였다.

발명의 내용

해결하려는 과제

[0005] 따라서, 본 발명의 목적은 추세 예측 지표인 EMA(지수이동평균) 지표와 Parabolic SAR 지표를 결합하여 만든 새로운 추세 예측 지표인 Dynamic SAR를 기반으로 하여 효율적인 컨테이너 오토 스케일링을 위한 발명을 제시한다.

[0006] 또한, 본 발명의 목적은 Dynamic SAR를 통해 보다 정확하게, 유입되는 네트워크 트래픽의 양과 네트워크 트래픽 흐름의 추세를 파악하는 것을 목적으로 한다.

[0007] 또한, 본 발명의 목적은 변동하는 네트워크 트래픽이 컨테이너로 유입되는 상황에서 신속하고 정확하게 컨테이너 스케일링을 수행함으로써 서비스 품질을 보전하는 방법 및 장치를 제공하는 것을 목적으로 한다.

[0008] 또한, 본 발명의 목적은 불필요하게 컨테이너 스케일링을 수행하는 과정을 감소시킴으로써 리소스 자원이 낭비되는 것을 방지하는 것을 목적으로 한다.

과제의 해결 수단

- [0009] 상기와 같은 과제를 해결하기 위한 본 발명에 따른 네트워크 트래픽 예측 기반의 효율적인 컨테이너 오토 스케일링 방법이 제공된다. 상기 방법은 오토 스케일링 장치에 의해 수행되고, 상기 방법은, 유입된 네트워크 트래픽에 대하여 단기지수이동평균(Short EMA)과 장기지수이동평균(Long EMA)을 계산하는 단기/장기지수이동평균 계산 과정; 상기 Short EMA와 상기 Long EMA의 차이 값을 정규화하는 정규화 과정; 상기 정규화된 차이 값을 동적 변수로 하여 동적 SAR (Stop and Reversal) 값을 도출하는 동적 SAR 도출 과정; 및 상기 동적 SAR 값에 기반하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하였는지를 판단하는 스케일 아웃 지점 도달 판단 과정을 포함한다. 이에 따라, 효율적으로 네트워크 트래픽을 예측할 수 있고, 서비스 제공 도중 품질 저하를 일으키지 않고 사용자들에게 서비스를 제공할 수 있다.
- [0010] 일 실시 예에서, 상기 유입된 네트워크 트래픽의 양은 상기 에지 클라우드 내의 특정 콘텐츠를 요청하는 사용자 단말의 수와 상기 사용자 단말 별로 요청된 특정 콘텐츠의 용량에 기반하여 결정될 수 있다. 이에 따라, 상기 트래픽이 상기 스케일 아웃 지점에 도달하면 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시키는 스케일 아웃 수행 과정을 더 포함할 수 있다.
- [0011] 일 실시 예에서, $t+1$ 번째 시각에서의 상기 동적 SAR (Dynamic SAR) 값은, $\text{Dynamic SAR} = \text{sar}[t+1] = \text{sar}[t] + (\text{af} + \text{Dynamic af}) * (\text{ep} - \text{sar}[t])$ 으로 도출될 수 있다. 한편, af는 비례상수이고, Dynamic af는 상기 정규화된 차이 값과 연관된 상기 동적 변수로서 네트워크 트래픽이 유입될 때마다 새로운 값으로 갱신될 수 있다. 이에 따라, 상기 동적 변수인 Dynamic af에 의해 상기 갱신된 동적 SAR 값에 기반하여, 상기 에지 클라우드에서 상기 스케일 아웃을 수행하고 상기 컨테이너의 수를 증가시킬 수 있다.
- [0012] 일 실시 예에서, 상기 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하는 스케일 인 지점 도달 판단 과정을 더 포함할 수 있다.
- [0013] 일 실시 예에서, 상기 스케일 인 지점에 도달하지 않은 경우, 상기 컨테이너의 수를 유지하면서 상기 동적 SAR 도출 과정을 반복할 수 있다.
- [0014] 일 실시 예에서, 상기 스케일링 아웃 지점과 상기 스케일링 인 지점은 상기 컨테이너의 수, 사용자 단말로부터 요청된 트래픽의 지연 민감 트래픽(delay sensitive traffic) 여부 및 최대 지연 허용 시간에 따라 동적으로 변경될 수 있다. 한편, 상기 최대 지연 허용 시간이 감소함에 따라 상기 스케일링 아웃 지점의 임계값을 상기 최대 지연 허용 시간과 제1 비례 상수에 기반하여 낮게 변경하고, 상기 스케일링 인 지점의 제2 임계값을 상기 최대 지연 허용 시간과 제2 비례 상수에 기반하여 낮게 변경될 수 있다. 이때, 상기 제2 비례 상수는 상기 제1 비례 상수보다 더 크게 설정되어 상기 스케일링 인 지점이 더 낮게 변경될 수 있다.
- [0015] 본 발명의 다른 측면에 따른 네트워크 트래픽 예측 기반의 효율적인 컨테이너 오토 스케일링 방법이 제공된다. 상기 방법은 오토 스케일링 장치에 의해 수행되고, 상기 방법은, 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하는 스케일 인 지점 도달 판단 과정; 상기 스케일 인 지점에 도달하지 않은 경우, 복수의 컨테이너를 기반으로 동적 SAR (Stop and Reversal) 값을 도출하는 동적 SAR 도출 과정; 및 상기 동적 SAR 값에 기반하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하였는지를 판단하는 스케일 아웃 지점 도달 판단 과정을 포함한다.
- [0016] 일 실시 예에서, 상기 트래픽이 상기 스케일 아웃 지점에 도달하면 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시키는 스케일 아웃 수행 과정을 더 포함할 수 있다.
- [0017] 일 실시 예에서, $t+1$ 번째 시각에서의 상기 동적 SAR (Dynamic SAR) 값은, $\text{Dynamic SAR} = \text{sar}[t+1] = \text{sar}[t] + (\text{af} + \text{Dynamic af}) * (\text{ep} - \text{sar}[t])$ 으로 도출될 수 있다. 한편, af는 비례상수이고, Dynamic af는 상기 정규화된 차이 값과 연관된 상기 동적 변수로서 네트워크 트래픽이 유입될 때마다 새로운 값으로 갱신될 수 있다. 이에 따라, 상기 동적 변수인 Dynamic af에 의해 상기 갱신된 동적 SAR 값에 기반하여, 상기 에지 클라우드에서 상기 스케일 아웃을 수행하고 상기 컨테이너의 수를 증가시킬 수 있다.
- [0018] 일 실시 예에서, 상기 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하는 스케일 인 지점 도달 판단 과정을 반복할 수 있다. 한편, 상기 스케일 인 지점에 도달하지 않은 경우, 상기 컨테이너의 수를 유지하면서 상기 동적 SAR 도출 과정을 반복할 수 있다.

- [0019] 일 실시 예에서, 상기 스케일링 아웃 지점과 상기 스케일링 인 지점은 상기 컨테이너의 수, 사용자 단말로부터 요청된 트래픽의 지연 민감 트래픽(delay sensitive traffic) 여부 및 최대 지연 허용 시간에 따라 동적으로 변경될 수 있다. 한편, 상기 최대 지연 허용 시간이 감소함에 따라 상기 스케일링 아웃 지점의 임계값을 상기 최대 지연 허용 시간과 제1 비례 상수에 기반하여 낮게 변경하고, 상기 스케일링 인 지점의 제2 임계값을 상기 최대 지연 허용 시간과 제2 비례 상수에 기반하여 낮게 변경될 수 있다. 이때, 상기 제2 비례 상수는 상기 제1 비례 상수보다 더 크게 설정되어 상기 스케일링 인 지점이 더 낮게 변경될 수 있다.
- [0020] 본 발명의 또 다른 측면에 따른 네트워크 트래픽 예측 기반의 컨테이너 오토 스케일링 장치가 제공된다. 상기 오토 스케일링 장치는, 복수의 사용자 단말로부터 서비스 요청을 수신하고, 상기 서비스 요청에 기반하여 에지 클라우드에 접속하는 상기 사용자 단말의 서비스 품질을 측정하는 사용자 단말 제어부를 포함한다. 또한, 유입된 네트워크 트래픽에 대하여 단기지수이동평균(Short EMA)과 장기지수이동평균(Long EMA)을 계산하고, 상기 Short EMA와 상기 Long EMA의 차이 값을 정규화하고, 상기 정규화된 차이 값을 동적 변수로 하여 동적 SAR (Stop and Reversal) 값을 도출하고, 상기 동적 SAR 값에 기반하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하였는지를 판단하는 스케일링 제어부를 더 포함한다.
- [0021] 일 실시 예에서, 상기 유입된 네트워크 트래픽의 양은 상기 에지 클라우드 내의 특정 콘텐츠를 요청하는 사용자 단말의 수와 상기 사용자 단말 별로 요청된 특정 콘텐츠의 용량에 기반하여 결정될 수 있다. 이에 따라, 상기 스케일링 제어부는, 상기 트래픽이 상기 스케일 아웃 지점에 도달하면 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시킬 수 있다.
- [0022] 일 실시 예에서, $t+1$ 번째 시각에서의 상기 동적 SAR (Dynamic SAR) 값은, $\text{Dynamic SAR} = \text{sar}[t+1] = \text{sar}[t] + (\text{af} + \text{Dynamic af}) * (\text{ep} - \text{sar}[t])$ 으로 도출될 수 있다. 한편, af는 비례상수이고, Dynamic af는 상기 정규화된 차이 값과 연관된 상기 동적 변수로서 네트워크 트래픽이 유입될 때마다 새로운 값으로 갱신될 수 있다. 이에 따라, 상기 동적 변수인 Dynamic af에 의해 상기 갱신된 동적 SAR 값에 기반하여, 상기 에지 클라우드에서 상기 스케일 아웃을 수행하고 상기 컨테이너의 수를 증가시킬 수 있다.
- [0023] 일 실시 예에서, 상기 스케일링 제어부는, 상기 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단할 수 있다. 한편, 상기 스케일 인 지점에 도달하지 않은 경우, 상기 컨테이너의 수를 유지하면서 상기 동적 SAR 값을 도출하는 동작을 반복할 수 있다.
- [0024] 일 실시 예에서, 상기 스케일링 제어부는, 상기 컨테이너의 수, 사용자 단말로부터 요청된 트래픽의 지연 민감 트래픽(delay sensitive traffic) 여부 및 최대 지연 허용 시간에 따라 상기 스케일링 아웃 지점과 상기 스케일링 인 지점을 동적으로 변경할 수 있다. 한편, 상기 최대 지연 허용 시간이 감소함에 따라 상기 스케일링 아웃 지점의 임계값을 상기 최대 지연 허용 시간과 제1 비례 상수에 기반하여 낮게 변경하고, 상기 스케일링 인 지점의 제2 임계값을 상기 최대 지연 허용 시간과 제2 비례 상수에 기반하여 낮게 변경될 수 있다. 이때, 상기 제2 비례 상수는 상기 제1 비례 상수보다 더 크게 설정되어 상기 스케일링 인 지점이 더 낮게 변경될 수 있다.
- [0025] 한편, 본 발명의 또 다른 측면에 따른 네트워크 트래픽 예측 기반의 컨테이너 오토 스케일링 장치를 포함하는 네트워크 시스템이 제공된다. 상기 네트워크 시스템은, 복수의 사용자 단말로부터 서비스 요청을 수신하고, 상기 서비스 요청에 기반하여 에지 클라우드에 접속하는 상기 사용자 단말의 서비스 품질을 측정하는 사용자 단말 제어부; 및 복수의 컨테이너를 포함하고, 상기 복수의 컨테이너 별로 콘텐츠를 동적으로 저장하도록 구성된 에지 클라우드를 포함한다. 또한, 상기 네트워크 시스템은, 상기 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하고, 상기 스케일 인 지점에 도달하지 않은 경우, 복수의 컨테이너를 기반으로 동적 SAR (Stop and Reversal) 값을 도출하고, 상기 동적 SAR 값에 기반하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하였는지를 판단하는 오토 스케일링 장치를 더 포함한다.
- [0026] 일 실시 예에서, 상기 오토 스케일링 장치는, 상기 트래픽이 상기 스케일 아웃 지점에 도달하면 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시킬 수 있다.
- [0027] 일 실시 예에서, $t+1$ 번째 시각에서의 상기 동적 SAR (Dynamic SAR) 값은, $\text{Dynamic SAR} = \text{sar}[t+1] = \text{sar}[t] + (\text{af} + \text{Dynamic af}) * (\text{ep} - \text{sar}[t])$ 으로 도출될 수 있다. 한편, af는 비례상수이고, Dynamic af는 상기 정규화된 차이 값과 연관된 상기 동적 변수로서 네트워크 트래픽이 유입될 때마다 새로운 값으로 갱신될 수 있다. 이에 따라, 상기 동적 변수인 Dynamic af에 의해 상기 갱신된 동적 SAR 값에 기반하여, 상기 에지 클라우드에서 상기 스케일 아웃을 수행하고 상기 컨테이너의 수를 증가시킬 수 있다.

[0028] 일 실시 예에서, 상기 오토 스케일링 장치는, 상기 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하는 스케일 인 지점 도달 판단 과정을 반복할 수 있다. 이에 따라, 상기 스케일 인 지점에 도달하지 않은 경우, 상기 컨테이너의 수를 유지하면서 상기 동적 SAR 도출 과정을 반복할 수 있다.

[0029] 일 실시 예에서, 상기 오토 스케일링 장치는, 상기 컨테이너의 수, 사용자 단말로부터 요청된 트래픽의 지연 민감 트래픽(delay sensitive traffic) 여부 및 최대 지연 허용 시간에 따라 상기 스케일링 아웃 지점과 상기 스케일링 인 지점을 동적으로 변경할 수 있다. 한편, 상기 최대 지연 허용 시간이 감소함에 따라 상기 스케일링 아웃 지점의 임계값을 상기 최대 지연 허용 시간과 제1 비례 상수에 기반하여 낮게 변경하고, 상기 스케일링 인 지점의 제2 임계값을 상기 최대 지연 허용 시간과 제2 비례 상수에 기반하여 낮게 변경될 수 있다. 이때, 상기 제2 비례 상수는 상기 제1 비례 상수보다 더 크게 설정되어 상기 스케일링 인 지점이 더 낮게 변경될 수 있다.

발명의 효과

[0030] 본 발명의 적어도 일 실시예에 따르면, 사용자들의 접속에 의해서 네트워크 트래픽이 짧은 시간 안에 급작스럽게 증가하거나, 네트워크 트래픽이 요동치는 흐름을 보이는 분야에서 효율적으로 네트워크 트래픽을 예측할 수 있다는 장점이 있다.

[0031] 또한, 본 발명의 적어도 일 실시예에 따르면, 미리 정의한 ‘Scale-out Point’를 통해 컨테이너의 스케일 아웃을 수행함으로써 서비스 제공 도중 품질 저하를 일으키지 않고 사용자들에게 서비스 제공할 것으로 기대된다.

[0032] 또한, 본 발명의 적어도 일 실시예에 따르면, 정확한 예측을 기반으로 신속하게 컨테이너 스케일링을 하는 것 이외에도, 네트워크 트래픽이 급격하게 줄어든 상황에서도 불필요하거나 서비스 제공을 하지 않는 컨테이너를 정리함으로써 자원이 낭비되는 것을 방지할 수 있다는 장점이 있다.

[0033] 또한, 본 발명의 적어도 일 실시예에 따르면, 컨테이너의 리소스 관리 측면에서도 효율적인 방법을 제공할 수 있다는 장점이 있다.

도면의 간단한 설명

[0034] 도 1은 본 발명에 따른 네트워크 트래픽 예측을 기반으로 한 효율적인 컨테이너 오토 스케일링 방법의 흐름도를 나타낸다.

도 2는 본 발명에 따른 컨테이너 당 허용 가능 트래픽 양과 ‘Scale-out Point’를 나타낸다.

도 3은 본 발명에 따른 컨테이너 당 허용 가능 트래픽 양과 ‘Scale-out Point’ 및 ‘Scale-in Point’를 나타낸다.

도 4는 본 발명에 따른 네트워크 트래픽 예측을 기반으로 한 효율적인 컨테이너 오토 스케일링 장치를 포함하는 네트워크 시스템의 개념도를 나타낸다.

도 5는 본 발명에 따른 네트워크 트래픽 예측을 기반으로 한 효율적인 컨테이너 오토 스케일링 장치의 상세 구성과 네트워크 시스템을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0035] 상술한 본 발명의 특징 및 효과는 첨부된 도면과 관련한 다음의 상세한 설명을 통하여 보다 분명해 질 것이며, 그에 따라 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 본 발명의 기술적 사상을 용이하게 실시할 수 있을 것이다.

[0036] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시 예를 가질 수 있는바, 특정 실시 예들을 도면에 예시하고 상세한 설명에 구체적으로 설명하고자 한다. 그러나 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0037] 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용한다.

[0038] 제1, 제2등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된

다.

- [0039] 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.
- [0040] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미가 있다.
- [0041] 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않아야 한다.
- [0042] 이하의 설명에서 사용되는 구성요소에 대한 접미사 모듈, 블록 및 부는 명세서 작성의 용이함만이 고려되어 부여되거나 혼용되는 것으로서, 그 자체로 서로 구별되는 의미 또는 역할을 갖는 것은 아니다.
- [0043] 이하, 본 발명의 바람직한 실시 예를 첨부한 도면을 참조하여 당해 분야에 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 설명한다. 하기에서 본 발명의 실시 예를 설명함에 있어, 관련된 공지의 기능 또는 공지의 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략한다.
- [0044] 이하에서는, 본 발명에 따른 네트워크 트래픽 예측 기반의 컨테이너 오토 스케일링 방법 및 장치, 보다 상세하게는, 네트워크 트래픽 예측 기반의 동적 SAR 산출과 이에 따른 효율적인 컨테이너 오토 스케일링 방법 및 장치에 대해 살펴보기로 한다.
- [0045] 한편, 본 발명의 목적은 추세 예측 지표인 EMA(지수이동평균) 지표와 Parabolic SAR 지표를 결합하여 만든 새로운 추세 예측 지표인 Dynamic SAR를 기반으로 하여 효율적인 컨테이너 오토 스케일링을 위한 발명을 제시하는 것을 목적으로 한다.
- [0046] 또한, 본 발명의 목적은 Dynamic SAR를 통해 보다 정확하게, 유입되는 네트워크 트래픽의 양과 네트워크 트래픽 흐름의 추세를 파악하는 것을 목적으로 한다.
- [0047] 또한, 본 발명의 목적은 변동하는 네트워크 트래픽이 컨테이너로 유입되는 상황에서 신속하고 정확하게 컨테이너 스케일링을 수행함으로써 서비스 품질을 보전하는 방법 및 장치를 제공하는 것을 목적으로 한다.
- [0048] 또한, 본 발명의 목적은 불필요하게 컨테이너 스케일링을 수행하는 과정을 감소시킴으로써 리소스 자원이 낭비되는 것을 방지하는 것을 목적으로 한다.
- [0049] 이러한 목적을 달성하기 위한 본 발명에 따른 구체적인 내용은 다음과 같다. 네트워크 트래픽 예측 기반의 효율적인 컨테이너 오토 스케일링 방법을 위한 Dynamic SAR(이하 본 제안 발명)은 여러 가격 예측 및 가격 추세 예측 발명들 중에서 대표적인 발명 중 하나인 EMA(지수이동평균) 지표와 Parabolic SAR 지표를 활용하여 새로운 예측 지표(Dynamic SAR)를 생성한다.
- [0050] 한편, 이러한 Dynamic SAR를 기반으로 컨테이너를 스케일링하기 위한 새로운 방법을 제시한다. 본 제안 발명은 컨테이너가 스케일링을 수행하기 위한 임계값인 ‘Scale-out Point’를 미리 정의하였을 뿐만 아니라, 요동치는 네트워크 트래픽의 흐름 내부에서 컨테이너가 빈번하게 스케일링을 수행하여 불필요하게 리소스 자원을 소비하는 것을 방지하기 위해 ‘Scale-in Point’을 별도로 지정함으로써 보다 효율적인 컨테이너 오토 스케일링 발명을 제시한다.
- [0051] 한편, 기존의 컨테이너 오토 스케일링 발명은 급격하게 값이 요동치는 네트워크 트래픽을 감당하지 못하여서 본래의 감당 가능한 네트워크 트래픽 양보다 더 많은 네트워크 트래픽을 맞이하곤 했다. 그 결과, 서비스 품질이 떨어지는 문제가 발생하였다. 뿐만 아니라 네트워크 트래픽이 급격하게 감소하는 경우, 이에 빠르게 대응하지 못하여 유입되는 네트워크 트래픽의 양이 적음에도 불구하고 컨테이너의 수를 유지하여 불필요하게 리소스 자원을 낭비하는 문제점이 있었다.
- [0052] 이러한 문제를 해결하기 위해 EMA(지수이동평균)과 Parabolic SAR 등 기존에 존재하던 예측 지표를 사용할 수 있다. 하지만, 기존 방법들 역시 정확하게 네트워크 트래픽을 예측을 하지 못하는 문제점이 있었다. EMA(지수이동평균)의 경우, 짧은 기간 동안의 이동평균을 적용할 경우, 조금의 값의 변동에도 예측에 많은 영향을 주어 거짓 예측을 알려주는 문제가 발생하였고, 긴 기간 동안의 이동평균을 적용할 경우, 값의 추세를 실제보다 느리게

반영하는 문제가 발생하였다.

- [0053] Parabolic SAR (Stop and Reversal)의 경우, 가속도 변수에 따라 실제 값의 추세를 전혀 예측하지 못하거나, 느리게 추세를 예측하여서 실제 값의 추세보다 추세전환이 늦어지는 문제가 발생하였다. 이처럼 기존의 방법으로는 값 또는 값의 흐름을 예측하는데 느리다는 한계점이 분명하게 존재하였다.
- [0054] 이러한 문제점을 해결하기 위해, 본 발명에서 제시하는 Dynamic SAR 값을 도출하기 위해 추세 예측지표인 Parabolic SAR와 지수이동평균을 결합하였다. Parabolic SAR의 경우 아래 수식과 같이 가속도 변수 'af'에 따라서 SAR값의 기울기가 달라진다.
- [0055] [수학식 1]
- [0056]
$$\text{Parabolic SAR} = \text{sar}[t+1] = \text{sar}[t] + \text{af} * (\text{ep} - \text{sar}[t])$$
- [0057] 여기서, $\text{sar}[t]$ 및 $\text{sar}[t+1]$ 은 각각 시각 t 및 시각 $t+1$ 에서 SAR 값이고, af (acceleration factor)는 가속도 변수로서 가중치 요소(weighting factor)에 해당한다. 이때, 가속도 변수인 af 는 비례 상수로서 최적화된 상수일 수 있다. 예를 들어, af 는 0.02로 고정(fix)할 수 있지만, 이에 한정되는 것은 아니고 응용에 따라 변경 가능하다. 한편, ep (extreme price)는 일정 시간 동안의 고점과 저점을 나타내는 변수이다. 예를 들어, ep 의 값은 최대값 또는 최소값을 나타낼 수 있다. 반면에, 일정 시간 구간 내에 전체 시간 구간의 최대값 또는 최소값과 같은 의미 있는 값이 존재하지 않으면, ep 의 값은 직전의 ep 값을 재사용할 수 있다.
- [0058] 하지만 가속도 값이 작을 경우, 적절한 타이밍에 추세전환이 이루어지지 않는 문제 뿐만 아니라 신속하게 추세를 따라가지 못할 수 있다. 가속도 값이 지나치게 클 경우, SAR 값의 기울기가 너무 커져버려서, 잘못된 추세전환이 발생할 수 있다. 따라서, 가속도 변수 'af'를 적절하게 조정하여 네트워크 트래픽 예측을 최적화할 수 있다.
- [0059] Parabolic SAR의 가속도 변수를 조정하기 위해서 본 발명에서는 짧은 기간동안의 평균 값을 구하는 단기 지수이동평균(이하 Short EMA)와 긴 기간 동안의 평균 값을 구하는 장기 지수이동평균(이하 Long EMA)를 이용할 수 있다. Short EMA와 Long EMA는 유입되는 네트워크 트래픽의 지수이동평균을 통해서 아래의 수식과 같이 구할 수 있다.
- [0060] [수학식 2]
- [0061]
$$\text{Difference} = \text{Short EMA} - \text{Long EMA}$$
- [0062] 여기서, 'Difference'는 Short EMA와 Long EMA의 차이를 구한 값이다. 'Difference'의 값 그 자체는 너무 크기 때문에, Parabolic SAR의 가속도의 값과 비슷한 규모로 조정해주기 위하여 정규화 과정을 거칠 필요가 있다. 이때 'Difference'의 범위는 EMA의 값에 따라 0 내지 500의 범위의 값을 갖는다. 따라서, 정규화 식은 'Difference'에 기존 Parabolic 수식의 최대 가속도 값인 0.2를 곱한 후 500으로 나누어 주는 것으로 계산할 수 있다.
- [0063] [수학식 3]
- [0064]
$$\text{정규화 값} = (\text{Difference} * 0.2) / 500$$
- [0065] 한편, 이하에서는 이러한 'Difference'의 정규화 값을 이용하는 본 발명에 따른 네트워크 트래픽 예측 기반의 컨테이너 오토 스케일링 방법에 대해 상세하게 살펴보기로 한다. 이와 관련하여, 도 1은 본 발명에 따른 네트워크 트래픽 예측을 기반으로 한 효율적인 컨테이너 오토 스케일링 방법의 흐름도를 나타낸다.
- [0066] 도 1을 참조하면, 본 발명에 따른 오토 스케일링 방법은 단기/장기지수이동평균 계산 과정(S110), 정규화 과정(S115), 동적 SAR 도출 과정(S120), 스케일 아웃 지점 도달 판단 과정(S130) 및 스케일 아웃 수행 과정(S135)을 포함한다. 또한, 본 발명에 따른 오토 스케일링 방법은 스케일 인 지점 도달 판단 과정(S140)을 더 포함할 수 있다.
- [0067] 단기/장기지수이동평균 계산 과정(S110)에서, 유입된 네트워크 트래픽에 대하여 단기지수이동평균(Short EMA)과 장기지수이동평균(Long EMA)을 계산한다. 한편, 정규화 과정(S115)에서, 상기 Short EMA와 상기 Long EMA의 차이 값을 정규화할 수 있다.
- [0068] 이에 따라, 동적 SAR 도출 과정(S120)에서, 상기 정규화된 차이 값을 동적 변수로 하여 동적 SAR (Stop and Reversal) 값을 도출할 수 있다. 한편, 스케일 아웃 지점 도달 판단 과정(S130)에서, 상기 동적 SAR 값에 기반

하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하였는지를 판단할 수 있다.

- [0069] 한편, 상기 유입된 네트워크 트래픽의 양은 상기 에지 클라우드 내의 특정 콘텐츠를 요청하는 사용자 단말의 수와 상기 사용자 단말 별로 요청된 특정 콘텐츠의 용량에 기반하여 결정될 수 있다. 이에 따라, 스케일 아웃 수행 과정(S135)에서, 상기 트래픽이 상기 스케일 아웃 지점에 도달하면 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시킬 수 있다.
- [0070] 한편, 전술한 동적 SAR (Dynamic SAR)는 아래의 수식에 의해 연산될 수 있다. 즉, 정규화 과정을 거친 ‘Difference’는 Parabolic SAR의 가속도를 조정할 ‘동적 가속도’ 변수로서 기존의 Parabolic SAR 값에 추가된다. 아래 수식에서 ‘Dynamic af’는 정규화 과정을 거친 후 기존의 Parabolic SAR에 추가된 ‘Difference’이다.
- [0071] [수학식 4]
- [0072]
$$\text{Dynamic SAR} = \text{sar}[t+1] = \text{sar}[t] + (\text{af} + \text{Dynamic af}) * (\text{ep} - \text{sar}[t])$$
- [0073] 여기서, $\text{sar}[t]$ 및 $\text{sar}[t+1]$ 은 각각 시각 t 및 시각 $t+1$ 에서 SAR 값이고, af (acceleration factor)는 가속도 변수로서 가중치 요소(weighting factor)에 해당한다. 이때, 가속도 변수인 af 는 비례상수이고, Dynamic af 는 정규화된 차이 값과 연관된 동적 변수로서 네트워크 트래픽이 유입될 때마다 새로운 값으로 갱신될 수 있다. 한편, ep (extreme price)는 전술한 바와 같이 일정 시간 동안의 고점과 저점을 나타내는 변수이다. 예를 들어, ep 의 값은 최대값 또는 최소값을 나타낼 수 있다. 반면에, 일정 시간 구간 내에 전체 시간 구간의 최대값 또는 최소값과 같은 의미 있는 값이 존재하지 않으면, ep 의 값은 직전의 ep 값을 재사용할 수 있다.
- [0074] 따라서, ‘Dynamic af’는 매 번 새롭게 계산되는 ‘Difference’를 통해 도출되기 때문에 네트워크 트래픽이 유입됨에 따라 항상 새로운 값으로 수식에 추가되어 계산된다. 이를 통해 기존 보다 신속하고 정확하게 예측을 수행할 수 있는 ‘Dynamic SAR’ 값을 구할 수 있다.
- [0075] 이에 따라, 동적 변수인 Dynamic af 에 의해 상기 갱신된 동적 SAR 값에 기반하여, 에지 클라우드에서 스케일 아웃을 수행하고 컨테이너의 수를 증가시킬 수 있다.
- [0076] 한편, 유입되는 네트워크 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점 도달 판단 과정(S140)에서, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단할 수 있다. 이때, 상기 스케일 인 지점에 도달하지 않은 경우, 상기 컨테이너의 수를 유지하면서 상기 동적 SAR 도출 과정(S120)을 반복할 수 있다. 한편, 상기 스케일 인 지점에 도달한 경우, 상기 컨테이너의 수를 감소시키는 스케일 인 수행 과정(S145)이 수행될 수 있다. 이와 같이, 컨테이너 수가 감소된 경우, 단기/장기 지수이동평균 계산 과정(S110)부터 해당 과정들이 다시 수행될 수 있지만, 응용에 따라, 특정 과정부터 수행될 수 있다.
- [0077] 이와 관련하여, 도 2는 본 발명에 따른 컨테이너 당 허용 가능 트래픽 양과 ‘Scale-out Point’를 나타낸다.
- [0078] 도 2를 참조하면, 간단한 예시로서 컨테이너 스케일링을 통해 늘어날 수 있는 컨테이너의 수는 최대 3개로 가정하였고, 컨테이너 1개당 최대로 허용 가능한 트래픽의 양을 400이라고 가정하였다. 즉, 컨테이너가 1대인 경우, 최대 허용 가능 네트워크 트래픽의 양은 400이 되고, 컨테이너가 2대인 경우는 800, 컨테이너가 3대인 경우는 최대 1200만큼의 네트워크 트래픽이 허용 가능하다고 가정한다. 따라서, 도 1은 컨테이너의 수에 따른 네트워크 트래픽의 최대 허용가능 양을 그래프를 통해 보여준다.
- [0079] 본 발명에서는 컨테이너의 스케일 아웃을 수행하기 위해서 ‘Scale-out Point’를 미리 정의하였다. 컨테이너에 유입되는 네트워크 트래픽을 예측한 Dynamic SAR의 값이 미리 지정해 둔 ‘Scale-out Point’를 넘어서는 순간 컨테이너의 스케일 아웃이 수행된다. 컨테이너의 스케일 아웃을 수행하게 하는 ‘Scale-out Point’는 컨테이너 수에 따라 각기 따로 설정되어 있으며, 도 1과 같이 컨테이너의 수가 1대인 경우 300, 컨테이너의 수가 2대인 경우 700, 컨테이너의 수가 3대인 경우 1100으로 설정할 수 있다. 해당 지점에 Dynamic SAR값이 도달하거나 또는 해당 지점을 넘어서 Dynamic SAR 값이 산출되면 예외없이 컨테이너의 스케일 아웃이 수행될 수 있다.
- [0080] 한편, 도 3은 본 발명에 따른 컨테이너 당 허용 가능 트래픽 양과 ‘Scale-out Point’ 및 ‘Scale-in Point’를 나타낸다.
- [0081] 구체적으로, 도 3은 요동치는 네트워크 트래픽에 불필요한 컨테이너 스케일링이 발생하는 것을 방지하기 위해서 미리 지정해 놓은 ‘Scale-in Point’를 보여준다. ‘Scale-out Point’에 의해 한번 스케일 아웃을 한 컨테이너가 약간의 변동에 의해 컨테이너 스케일 인을 수행하는 것을 방지하기 위해 그림 2와 같이 ‘Scale-in Point

’를 지정해두어, 해당 지점을 내려가지 않는 경우에는 컨테이너의 수를 유지할 수 있다.

- [0082] ‘Scale-out Point’와 같이 ‘Scale-in Point’도 컨테이너의 수에 따라 각각 지정이 되어있으며, 본 지점보다 Dynamic SAR 값이 위에 있다면 컨테이너의 수는 유지되고, 컨테이너 스케일 인은 수행되지 않는다. 이를 통해서 무분별하게 컨테이너 스케일링이 발생하는 상황을 방지하고, 불필요하게 리소스가 낭비되는 것을 피할 수 있다.
- [0083] 한편, 도 1 내지 도 3을 참조하면, 스케일링 아웃 지점과 스케일링 인 지점은 컨테이너의 수, 사용자 단말로부터 요청된 트래픽의 지연 민감 트래픽(delay sensitive traffic) 여부 및 최대 지연 허용 시간에 따라 동적으로 변경될 수 있다.
- [0084] 본 발명의 일 실시 예에 따르면, 최대 지연 허용 시간이 감소함에 따라 상기 스케일링 아웃 지점의 임계값을 상기 최대 지연 허용 시간과 제1 비례 상수에 기반하여 낮게 변경될 수 있다. 또한, 상기 스케일링 인 지점의 제2 임계값을 상기 최대 지연 허용 시간과 제2 비례 상수에 기반하여 낮게 변경할 수 있다.
- [0085] 이때, 상기 제2 비례 상수는 상기 제1 비례 상수보다 더 크게 설정되어 상기 스케일링 인 지점이 더 낮게 변경될 수 있다. 따라서, 지연 민감 트래픽에 대해 일단 스케일링 아웃이 수행되어 특정 콘텐츠를 수용하는 컨테이너 수가 증가하면 다시 트래픽 요청이 감소하더라도 컨테이너 수를 일시적으로 유지할 수 있다는 장점이 있다. 이에 따라, 지연 민감 트래픽에 대한 요청이 일시적으로 감소하더라도 이에 대한 요청이 다시 증가하여 컨테이너 수를 다시 증가시켜야 하는 네트워크 혼잡을 미리 방지할 수 있다.
- [0086] 한편, 최대 지연 허용 시간은 도 4에 표시되는 바와 같이 최대 허용 네트워크 트래픽 볼륨(Maximum Acceptable Network Traffic Volumes)로 대체될 수 있다. 또한, 최대 지연 허용 시간과 최대 허용 네트워크 트래픽 볼륨을 모두 고려하여 임계값 및 제2 임계값을 변경할 수 있다.
- [0087] 또한, 본 발명의 다른 측면에 따른 오토 스케일링 방법은 스케일 인 지점 도달 판단 과정(S140), 동적 SAR 도출 과정(S120), 스케일 아웃 지점 도달 판단 과정(S130)을 포함하여 수행하도록 구성될 수 있다.
- [0088] 이와 관련하여, 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점 도달 판단 과정(S140)에서, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단할 수 있다.
- [0089] 한편, 동적 SAR 도출 과정(S120)에서, 상기 스케일 인 지점에 도달하지 않은 경우, 복수의 컨테이너를 기반으로 동적 SAR (Stop and Reversal) 값을 도출할 수 있다. 이때, 동적 SAR (Dynamic SAR) 값의 산출은 전술한 수학적 식 2를 이용할 수 있다. 따라서, 동적 변수인 Dynamic af에 의해 갱신된 동적 SAR 값에 기반하여, 에지 클라우드에서 스케일 아웃을 수행하고 컨테이너의 수를 증가시킬 수 있다. 또는, 동적 변수인 Dynamic af에 의해 갱신된 동적 SAR 값에 기반하여, 에지 클라우드에서 스케일 인을 수행하고 컨테이너의 수를 감소시킬 수 있다.
- [0090] 이에 따라, 스케일 아웃 지점 도달 판단 과정(S130)에서, 상기 동적 SAR 값에 기반하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하였는지를 판단할 수 있다.
- [0091] 또한, 스케일 아웃 지점 도달 판단 과정(S130)에서, 트래픽이 상기 스케일 아웃 지점에 도달하면, 스케일 아웃 수행 과정(S135)에서, 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시킬 수 있다.
- [0092] 한편, 네트워크 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하는 스케일 인 지점 도달 판단 과정(S140)을 반복할 수 있다. 이때, 상기 스케일 인 지점에 도달하지 않은 경우, 상기 컨테이너의 수를 유지하면서 동적 SAR 도출 과정(S120)을 반복할 수 있다.
- [0093] 한편, 전술한 바와 같이, 최대 지연 허용 시간이 감소함에 따라 상기 스케일링 아웃 지점의 임계값을 상기 최대 지연 허용 시간과 제1 비례 상수에 기반하여 낮게 변경될 수 있다. 또한, 상기 스케일링 인 지점의 제2 임계값을 상기 최대 지연 허용 시간과 제2 비례 상수에 기반하여 낮게 변경할 수 있다.
- [0094] 이때, 상기 제2 비례 상수는 상기 제1 비례 상수보다 더 크게 설정되어 상기 스케일링 인 지점이 더 낮게 변경될 수 있다. 따라서, 지연 민감 트래픽에 대해 일단 스케일링 아웃이 수행되어 특정 콘텐츠를 수용하는 컨테이너 수가 증가하면 다시 트래픽 요청이 감소하더라도 컨테이너 수를 일시적으로 유지할 수 있다는 장점이 있다. 이에 따라, 지연 민감 트래픽에 대한 요청이 일시적으로 감소하더라도 이에 대한 요청이 다시 증가하여 컨테이너 수를 다시 증가시켜야 하는 네트워크 혼잡을 미리 방지할 수 있다.
- [0095] 한편, 최대 지연 허용 시간은 도 4에 표시되는 바와 같이 최대 허용 네트워크 트래픽 볼륨(Maximum Acceptable Network Traffic Volumes)로 대체될 수 있다. 또한, 최대 지연 허용 시간과 최대 허용 네트워크 트래픽 볼륨을

모두 고려하여 임계값 및 제2 임계값을 변경할 수 있다.

- [0096] 이상에서는 본 발명에 따른 네트워크 트래픽 예측 기반의 컨테이너 오토스케일링 방법에 대해 살펴보았다. 이하에서는 본 발명의 다른 측면에 따른 네트워크 트래픽 예측 기반의 컨테이너 오토스케일링 장치 및 이를 포함하는 네트워크 시스템에 대해 살펴보기로 한다. 한편, 전술된 오토스케일링 방법에 대한 설명이 이하의 컨테이너 오토스케일링 장치 및 이를 포함하는 네트워크 시스템에 적용될 수 있다.
- [0097] 이와 관련하여, 도 4는 본 발명에 따른 네트워크 트래픽 예측을 기반으로 한 효율적인 컨테이너 오토 스케일링 장치를 포함하는 네트워크 시스템의 개념도를 나타낸다. 도 4를 참조하면, 네트워크 시스템은 오토 스케일링 장치(100), 복수의 사용자 단말(200) 및 에지 클라우드(300)를 포함한다.
- [0098] 한편, 도 5는 본 발명에 따른 네트워크 트래픽 예측을 기반으로 한 효율적인 컨테이너 오토 스케일링 장치의 상세 구성과 네트워크 시스템을 나타낸다. 도 5를 참조하면, 오토 스케일링 장치는 사용자 단말 제어부(110) 및 스케일링 제어부(120)를 포함한다. 이때, 사용자 단말 제어부(110)는 복수의 단말(200)과 정보를 교환하는 인터페이스를 포함할 수 있다. 한편, 에지 클라우드(300)는 복수의 컨테이너들을 포함할 수 있다.
- [0099] 한편, 응용에 따라 오토 스케일링 장치(100)는 스케일링 제어부(120)만을 포함하거나 또는 사용자 단말 제어부(110) 및 스케일링 제어부(120)를 모두 포함할 수 있다.
- [0100] 도 4 및 도 5를 참조하면, 사용자 단말 제어부(110)는 복수의 사용자 단말로부터 서비스 요청을 수신하고, 상기 서비스 요청에 기반하여 에지 클라우드에 접속하는 상기 사용자 단말의 서비스 품질을 측정할 수 있다. 한편, 에지 클라우드(300)는 복수의 컨테이너를 포함하고, 상기 복수의 컨테이너 별로 콘텐츠를 동적으로 저장하도록 구성된다.
- [0101] 또한, 스케일링 제어부(120)에 해당하는 오토스케일링 장치(100)는 상기 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단한다. 이때, 상기 스케일 인 지점에 도달하지 않은 경우, 복수의 컨테이너를 기반으로 동적 SAR (Stop and Reversal) 값을 도출할 수 있다. 한편, 상기 동적 SAR 값에 기반하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하였는지를 판단할 수 있다.
- [0102] 이때, 오토스케일링 장치(100)는 상기 트래픽이 상기 스케일 아웃 지점에 도달하면 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시킬 수 있다. 한편, t+1 번째 시각에서의 상기 동적 SAR (Dynamic SAR) 값은 전술한 수학적 식 2와 관련된 설명이 적용될 수 있다. 이에 따라, 동적 변수인 Dynamic af에 의해 갱신된 동적 SAR 값에 기반하여, 에지 클라우드에서 스케일 아웃을 수행하고 컨테이너의 수를 증가시킬 수 있다.
- [0103] 한편, 오토스케일링 장치(100)는 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단하는 스케일 인 지점 도달 판단 과정을 반복할 수 있다. 이때, 스케일 인 지점에 도달하지 않은 경우, 상기 컨테이너의 수를 유지하면서 상기 동적 SAR 도출 과정을 반복할 수 있다.
- [0104] 또한, 오토스케일링 장치(100)는 컨테이너의 수, 사용자 단말로부터 요청된 트래픽의 지연 민감 트래픽(delay sensitive traffic) 여부 및 최대 지연 허용 시간에 따라 상기 스케일링 아웃 지점과 상기 스케일링 인 지점을 동적으로 변경할 수 있다. 이때, 최대 지연 허용 시간이 감소함에 따라 상기 스케일링 아웃 지점의 임계값을 상기 최대 지연 허용 시간과 제1 비례 상수에 기반하여 낮게 변경할 수 있다. 또한, 상기 스케일링 인 지점의 제2 임계값을 상기 최대 지연 허용 시간과 제2 비례 상수에 기반하여 낮게 변경할 수 있다. 한편, 상기 제2 비례 상수는 상기 제1 비례 상수보다 더 크게 설정되어 상기 스케일링 인 지점이 더 낮게 변경될 수 있다.
- [0105] 한편, 본 발명의 또 다른 측면에 따른 사용자 단말 제어부와 스케일링 제어부를 포함하는 오토 스케일링 장치에 대해 살펴보면 다음과 같다. 이와 관련하여, 도 4 및 도 5를 참조하면, 사용자 단말 제어부(110)는 복수의 사용자 단말로부터 서비스 요청을 수신하고, 상기 서비스 요청에 기반하여 에지 클라우드에 접속하는 상기 사용자 단말의 서비스 품질을 측정한다.
- [0106] 한편, 스케일링 제어부(120)는 유입된 네트워크 트래픽에 대하여 단기지수이동평균(Short EMA)과 장기지수이동평균(Long EMA)을 계산하고, 상기 Short EMA와 상기 Long EMA의 차이 값을 정규화한다. 또한, 스케일링 제어부(120)는 상기 정규화된 차이 값을 동적 변수로 하여 동적 SAR (Stop and Reversal) 값을 도출할 수 있다. 또한, 스케일링 제어부(120)는 상기 동적 SAR 값에 기반하여 에지 클라우드의 트래픽이 스케일 아웃 지점(Scale Out

Point)에 도달하였는지를 판단할 수 있다.

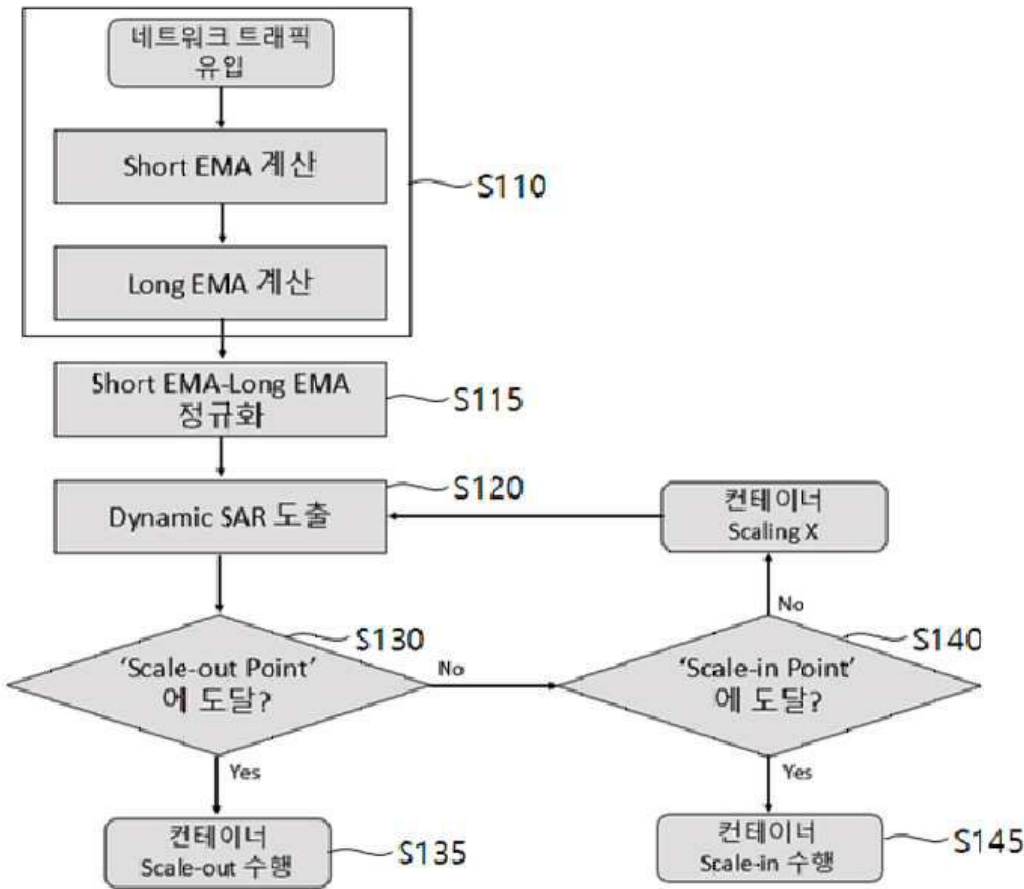
- [0107] 한편, 상기 유입된 네트워크 트래픽의 양은 상기 에지 클라우드 내의 특정 콘텐츠를 요청하는 사용자 단말의 수와 상기 사용자 단말 별로 요청된 특정 콘텐츠의 용량에 기반하여 결정할 수 있다.
- [0108] 이에 따라, 스케일링 제어부(120)는 상기 트래픽이 상기 스케일 아웃 지점에 도달하면 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시킬 수 있다. 이때, 스케일링 제어부(120)는 상기 트래픽이 상기 스케일 아웃 지점에 도달하면 상기 에지 클라우드의 복수의 컨테이너에 대한 스케일-아웃을 수행하여 상기 컨테이너의 수를 증가시킬 수 있다. 한편, t+1 번째 시각에서의 상기 동적 SAR (Dynamic SAR) 값은 전술한 수학적 2와 관련된 설명이 적용될 수 있다. 이에 따라, 동적 변수인 Dynamic af에 의해 갱신된 동적 SAR 값에 기반하여, 에지 클라우드에서 스케일 아웃을 수행하고 컨테이너의 수를 증가시킬 수 있다.
- [0109] 또한, 스케일링 제어부(120)는 네트워크 트래픽이 스케일 아웃 지점(Scale Out Point)에 도달하지 않은 경우, 스케일 인 지점(Scale In Point)에 도달하였는지를 판단할 수 있다. 이에 따라, 상기 스케일 인 지점에 도달하지 않은 경우, 상기 컨테이너의 수를 유지하면서 상기 동적 SAR 값을 도출하는 동작을 반복할 수 있다.
- [0110] 또한, 스케일링 제어부(120)는 상기 컨테이너의 수, 사용자 단말로부터 요청된 트래픽의 지연 민감 트래픽(delay sensitive traffic) 여부 및 최대 지연 허용 시간에 따라 상기 스케일링 아웃 지점과 상기 스케일링 인 지점을 동적으로 변경할 수 있다. 이때, 상기 최대 지연 허용 시간이 감소함에 따라 상기 스케일링 아웃 지점의 임계값을 상기 최대 지연 허용 시간과 제1 비례 상수에 기반하여 낮게 변경할 수 있다. 또한, 상기 스케일링 인 지점의 제2 임계값을 상기 최대 지연 허용 시간과 제2 비례 상수에 기반하여 낮게 변경할 수 있다. 이때, 상기 제2 비례 상수는 상기 제1 비례 상수보다 더 크게 설정되어 상기 스케일링 인 지점이 더 낮게 변경될 수 있다.
- [0111] 이상에서는 본 발명에 따른 네트워크 트래픽 예측 기반의 컨테이너 오토 스케일링 방법 및 장치, 보다 상세하게는, 네트워크 트래픽 예측 기반의 동적 SAR 산출과 이에 따른 효율적인 컨테이너 오토 스케일링 방법 및 장치와 네트워크 시스템에 대해 살펴보았다. 한편, 본 발명에 따른 기술적 구성 및 특징에 대해 요약하면 다음과 같다.
- [0112] 이와 관련하여, 본 발명에서는 기존의 EMA(지수이동평균)과 Parabolic SAR가 가지는 단점을 보완하기 위해 EMA(지수이동평균) 지표와 Parabolic SAR 지표를 결합하여 보다 신속하고 정확하게 값과 값의 추세를 예측할 수 있는 지표인 Dynamic SAR를 도출할 수 있다.
- [0113] 한편, 본 발명에 따른 Dynamic SAR는 유입되는 네트워크 트래픽의 값과 네트워크 트래픽 흐름의 추세를 정확하게 예측하는 것을 가능하게 한다. 정확한 예측을 수행하는 Dynamic SAR 지표와 하나의 컨테이너가 수용 가능한 최대 네트워크 트래픽의 양을 기반으로 'Scale-out Point'를 미리 정의하여 컨테이너가 과부하상태에 들어서기 전에 네트워크 트래픽의 양이 'Scale-out Point'에 근접하게 되면 컨테이너 스케일링을 수행할 수 있다.
- [0114] 또한 'Scale-out Point' 아래에 'Scale-in Point'를 설정하여 요동치는 네트워크 트래픽에 컨테이너가 하나 하나 반응하지 않고 상태를 유지할 수 있도록 하여 불필요한 컨테이너 스케일링 수행을 방지할 수 있다.
- [0115] 또한, 본 발명은 새로운 추세 예측 지표인 Dynamic SAR를 사용하여 더 빠르고 정확한 예측을 수행한다. 또한 'Scale-out Point'를 통해 컨테이너 스케일링을 빠르게 수행함으로써 서비스 품질이 떨어지는 상황을 피하고, 'Scale-in Point'를 통해서 불필요한 컨테이너 스케일링을 감소시킴으로써 리소스 자원의 낭비를 방지한다는 점에서 기존과 비교하여 더 효율적인 컨테이너 오토 스케일링 발명이다.
- [0116] 이러한 본 발명의 기술적 구성과 특징에 따른 본 발명의 기술적 효과에 대해 살펴보면 다음과 같다.
- [0117] 본 발명의 적어도 일 실시예에 따르면, 사용자들의 접속에 의해서 네트워크 트래픽이 짧은 시간 안에 급작스럽게 증가하거나, 네트워크 트래픽이 요동치는 흐름을 보이는 분야에서 효율적으로 네트워크 트래픽을 예측할 수 있다는 장점이 있다.
- [0118] 또한, 본 발명의 적어도 일 실시예에 따르면, 미리 정의한 'Scale-out Point'를 통해 컨테이너의 스케일 아웃을 수행함으로써 서비스 제공 도중 품질 저하를 일으키지 않고 사용자들에게 서비스 제공할 것으로 기대된다.
- [0119] 또한, 본 발명의 적어도 일 실시예에 따르면, 정확한 예측을 기반으로 신속하게 컨테이너 스케일링을 하는 것 이외에도, 네트워크 트래픽이 급격하게 줄어든 상황에서도 불필요하거나 서비스 제공을 하지 않는 컨테이너를 정리함으로써 자원이 낭비되는 것을 방지할 수 있다는 장점이 있다.
- [0120] 또한, 본 발명의 적어도 일 실시예에 따르면, 컨테이너의 리소스 관리 측면에서도 효율적인 방법을 제공할 수

있다는 장점이 있다.

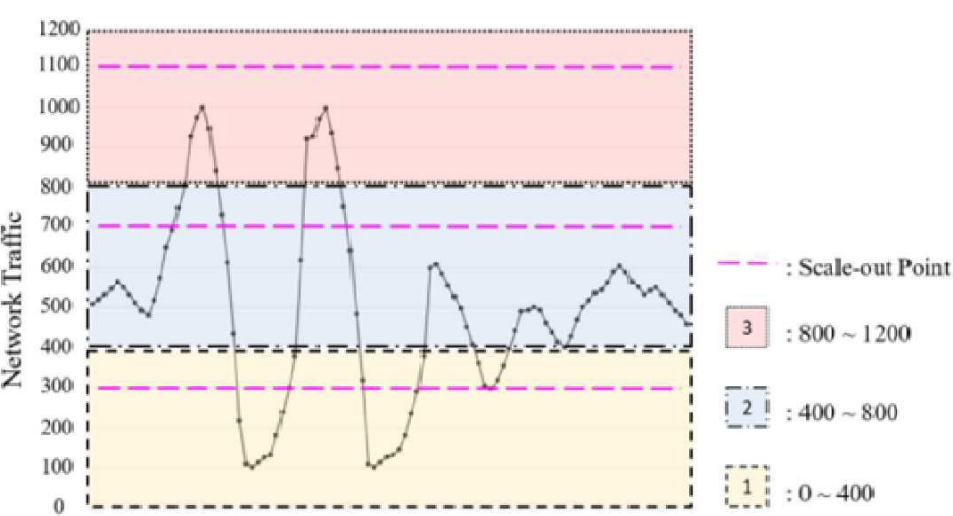
- [0121] 다음으로, 본 발명의 기술적 구성 및 특징과 기술적 효과에 따른 본 발명의 활용분야에 대해 살펴보면 아래와 같다.
- [0122] 이와 관련하여, 사용자와 멀리 떨어진 중앙 클라우드와 비교하여 에지 클라우드는 사용자와 가까운 위치에 배치되어 있다. 하지만 에지 클라우드는 중앙 클라우드와 비교하여 상대적으로 부족한 컴퓨팅 자원을 가지고 있기 때문에, 허용 가능한 네트워크 트래픽 양 보다 더 많은 네트워크 트래픽이 들어올 경우, 서비스 품질이 떨어지는 문제가 발생 할 수 있다. 특히 조금이라도 급격하게 많은 네트워크 트래픽이 에지 클라우드로 집중되는 경우 혹은 빈번하게 요동치는 네트워크 트래픽이 에지 클라우드로 유입되는 경우, 과부하로 인해 올바른 서비스 품질을 기대하는 것은 어렵다.
- [0123] 따라서, 에지 클라우드는 중앙 클라우드와 달리 많은 양의 네트워크 트래픽을 수용하기 어렵기 때문에, 어느 정도 양의 네트워크 트래픽이 유입되는 상황에서 빠르게 컨테이너 스케일링을 수행하여 동적으로 유입되는 네트워크 트래픽을 수용할 준비를 해야 한다. 이러한 상황을 만족해야, 네트워크 트래픽이 빈번하게 요동치거나 급격하게 네트워크 트래픽이 유입되는 경우, 서비스 품질을 떨어뜨리지 않고 신속한 서비스 제공이 가능하다. 또한 상대적으로 부족한 컴퓨팅 자원을 가진 에지 클라우드에서 불필요하게 많은 컨테이너를 구동하는 것은 자원낭비이기 때문에 네트워크 트래픽이 적게 유입되는 상황에서는 빠르게 컨테이너 수를 줄여 쓸데없이 자원이 낭비되는 것을 방지해야 한다. 이러한 측면에서 볼 때 본 발명은 에지 클라우드에 적용되어 사용자들에게 품질 좋은 서비스를 제공하게끔 해주는 방법으로 활용될 수 있다.
- [0124] 한편, 본 발명의 기술 사업화 전망과 관련하여, 네트워크 트래픽 예측을 기반으로 한 효율적인 컨테이너 오토 스케일링 방법은 네트워크 트래픽이 갑자기 급격하게 증가하는 웹 사이트(쇼핑몰 사이트, 수강신청 사이트) 또는 실시간 영상 스트리밍 분야에 적용될 수 있다.
- [0125] 또한, 사용자의 접속이 집중될 때, 유입될 네트워크 트래픽을 신속하게 파악하여 컨테이너가 감당할 수 있는 네트워크 트래픽을 초과하기에 앞서서 미리 컨테이너 스케일링을 수행함으로써 서비스가 중단되는 일 없이 원활하게 사용자들에게 서비스를 제공할 수 있다.
- [0126] 소프트웨어적인 구현에 의하면, 본 명세서에서 설명되는 절차 및 기능뿐만 아니라 각각의 구성 요소들에 대한 설계 및 파라미터 최적화는 별도의 소프트웨어 모듈로도 구현될 수 있다. 적절한 프로그램 언어로 쓰여진 소프트웨어 어플리케이션으로 소프트웨어 코드가 구현될 수 있다. 상기 소프트웨어 코드는 메모리에 저장되고, 제어부(controller) 또는 프로세서(processor)에 의해 실행될 수 있다.

도면

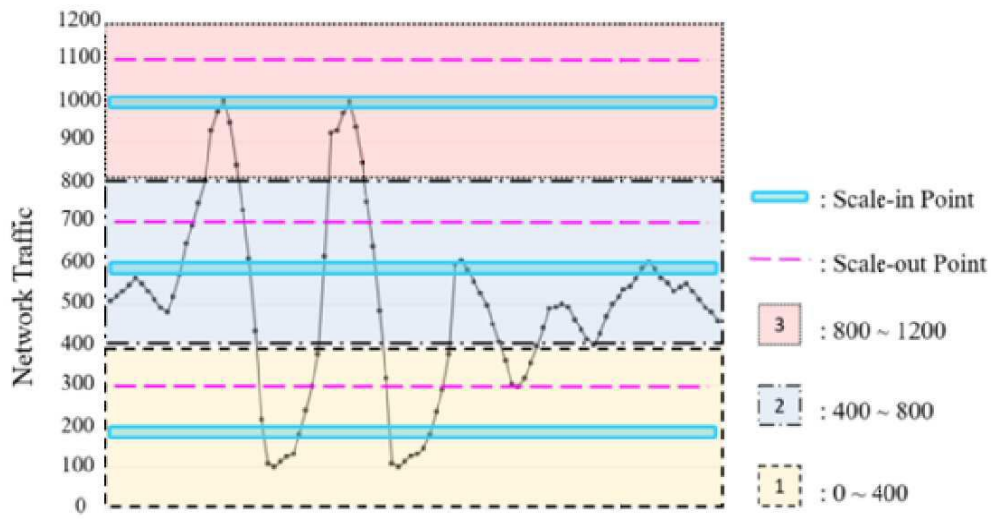
도면1



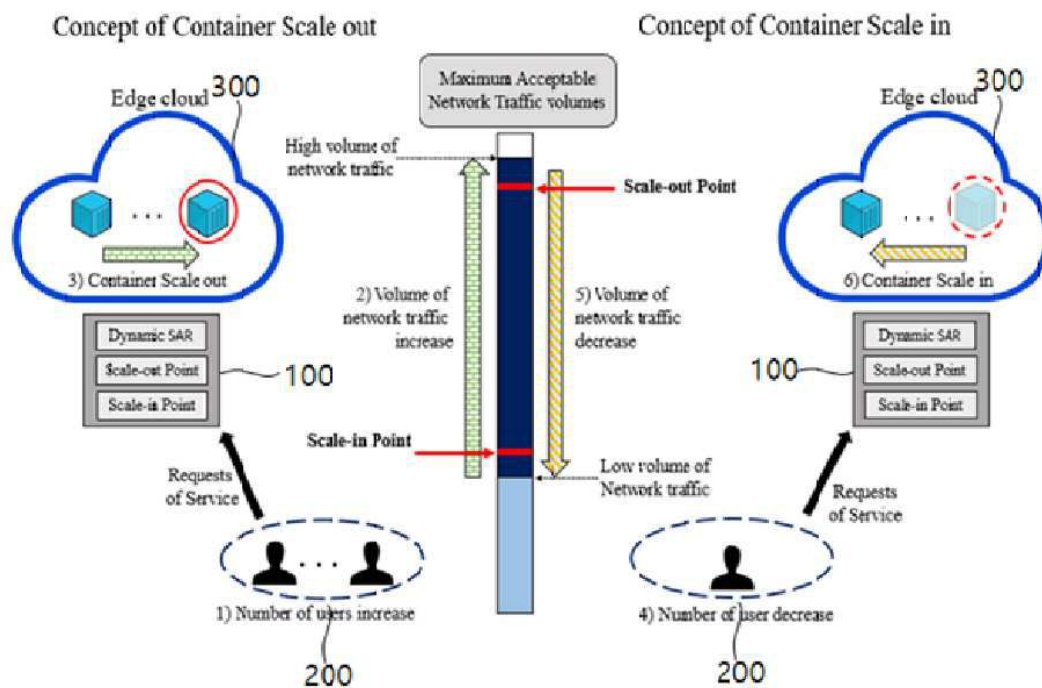
도면2



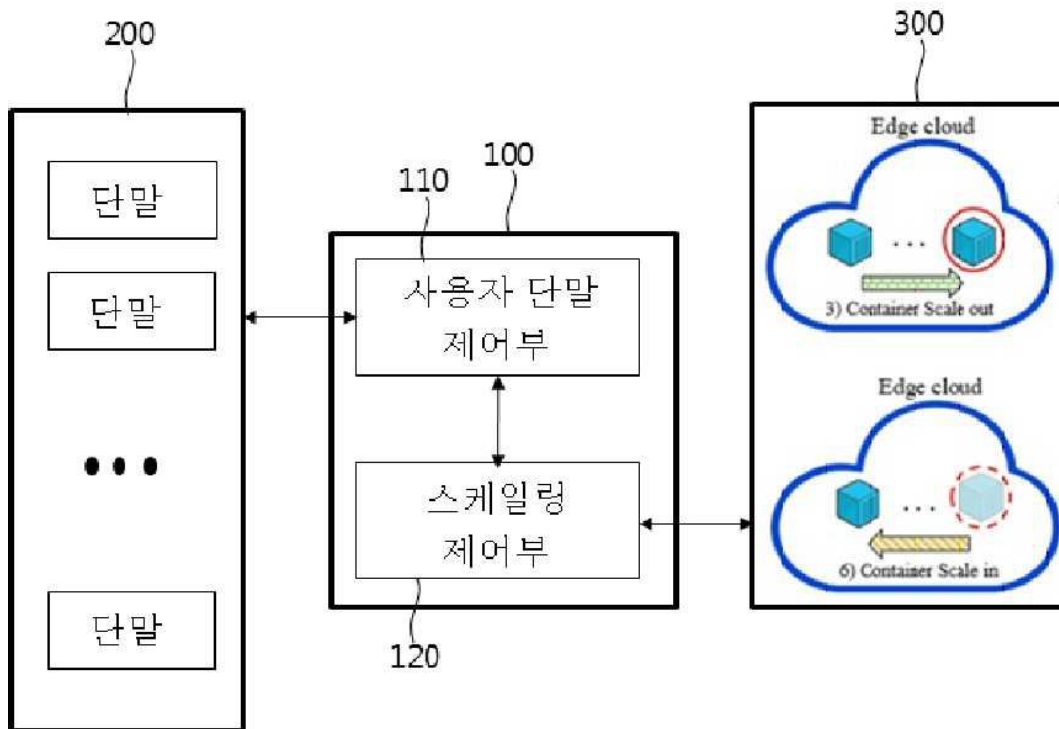
도면3



도면4



도면5



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 제 6, 11, 16 및 21 항

【변경전】

스케일링 인

【변경후】

스케일 인

【직권보정 2】

【보정항목】 청구범위

【보정세부항목】 제 6, 11, 16 및 21 항

【변경전】

스케일링 아웃

【변경후】

스케일 아웃