(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2016년11월09일
(11) 등록번호 10-1669700
(24) 등록일자 2016년10월20일

(51) 국제특허분류(Int. Cl.)
H04L 29/06 (2006.01) *H04L 12/46* (2006.01)
H04L 12/715 (2013.01) *H04L 12/803* (2013.01)
(21) 출원번호 10-2011-7028254
(22) 출원일자(국제) 2010년05월28일
심사청구일자 2015년04월30일
(85) 번역문제출일자 2011년11월25일
(65) 공개번호 10-2012-0026516
(43) 공개일자 2012년03월19일
(86) 국제출원번호 PCT/US2010/036758
(87) 국제공개번호 WO 2010/138937
국제공개일자 2010년12월02일
(30) 우선권주장
61/182,063 2009년05월28일 미국(US)
12/578,608 2009년10월14일 미국(US)
(56) 선행기술조사문헌
US20070280243 A1
US6766371 B1
US20090106529 A1
KR1020070023697 A

(73) 특허권자
마이크로소프트 테크놀로지 라이센싱, 엘엘씨
미국 워싱턴주 (우편번호 : 98052) 레드몬드 원 마이크로소프트 웨이
(72) 발명자
그린버그 알버트
미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션
라히리 파란탑
미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션
(뒷면에 계속)
(74) 대리인
제일특허법인

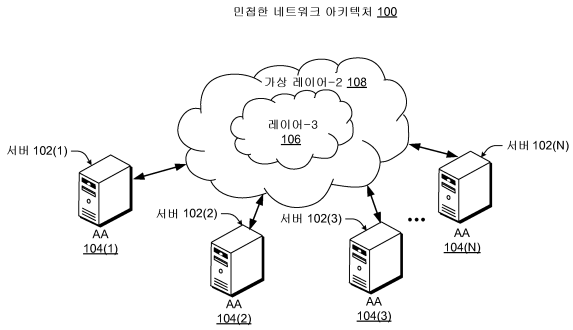
전체 청구항 수 : 총 22 항 심사관 : 문형섭

(54) 발명의 명칭 **민첩한 데이터 센터 네트워크 아키텍처**

(57) 요 약

본 특허 출원은 무엇보다도 데이터 센터 내에 이용될 수 있는 민첩한 네트워크 아키텍처에 관한 것이다. 일 구현예는 레이어-3 인프라구조의 머신에 접속하는 가상 레이어-2 네트워크를 제공한다.

대 표 도



(72) 발명자

말츠 데이비드 에이

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션

파텔 파빈 케이

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션

센굽타 수디프타

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션

제인 나벤두

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션

김 창훈

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션

명 세 서

청구범위

청구항 1

제1 머신 및 제2 머신을 포함하는 복수의 머신 중의 개별 머신들에 애플리케이션 어드레스를 할당하고 레이어-3 인프라구조의 구성 요소에 위치 어드레스를 할당함으로써, 상기 복수의 머신을 연결하는 가상 레이어-2 네트워크를 상기 레이어-3 인프라구조를 통해 제공하기 위한 방법으로서,

상기 제1 머신과 연관된 애자일 에이전트(agile agent)에 의해, 상기 제2 머신과 연관된 목적지 서버(destination server)의 할당된 애플리케이션 어드레스를 갖는 가상 레이어-2 패킷을 인터셉트(intercept)하는 단계 - 상기 복수의 머신 중의 상기 개별 머신들은 상이한 애자일 에이전트와 연관됨 - 와,

상기 제1 머신과 연관된 상기 애자일 에이전트에 의해, 서비스를 위한 정의된 서버 그룹 내에 상기 목적지 서버가 존재한다는 것을 결정하는 단계 - 상기 정의된 서버 그룹은 애자일 디렉토리 서비스(agile directory service)에 저장되어 있음 - 와,

상기 정의된 서버 그룹 내에 상기 목적지 서버가 존재한다는 것에 기초하여, 상기 애자일 에이전트에 의해, 상기 할당된 애플리케이션 어드레스와 연관된 개별 위치 어드레스를 검색하기 위해 상기 애자일 디렉토리 서비스를 활용하는 단계와,

상기 애자일 에이전트에 의해, 상기 개별 위치 어드레스에 상기 패킷을 전송할 상기 레이어-3 인프라구조의 스위치를 선택하는 단계 - 상기 스위치는 로드 밸런싱(load balancing)을 제공하기 위해 상기 레이어-3 인프라구조의 복수의 스위치들로부터 선택됨 - 와,

상기 애자일 에이전트에 의해, 상기 가상 레이어-2 패킷을 레이어-3 패킷 내에 캡슐화(encapsulate)하는 단계 - 상기 레이어-3 패킷에 상기 스위치의 상기 개별 위치 어드레스가 할당됨 - 와,

상기 레이어-3 패킷을 상기 스위치에 전송하는 단계 - 상기 스위치는 상기 캡슐화된 가상 레이어-2 패킷을 역캡슐화(decapsulate)하고 상기 역캡슐화된 가상 레이어-2 패킷을 물리 네트워크 연결을 통해 상기 제2 머신에 전송하도록 구성됨 -

를 포함하고,

상기 목적지 서버가 상기 정의된 서버 그룹 내에 존재하지 않는 경우, 상기 애자일 디렉토리 서비스는 상기 개별 위치 어드레스를 제공하는 것을 거부하는

방법.

청구항 2

제 1 항에 있어서,

상기 복수의 머신 중의 개별 머신 사이에 조기 선회 경로(early turnaround paths)를 사용하는 단계를 더 포함하는

방법.

청구항 3

제 1 항에 있어서,

상기 애자일 에이전트는 상기 제1 머신과 연관된 레이어-2 인프라구조 내에 위치한 하드웨어와 상기 제1 머신과 연관된 상기 레이어-2 인프라구조 내에서 실행되는 소프트웨어의 조합을 포함하는

방법.

청구항 4

제 1 항에 있어서,

상기 가상 레이어-2 네트워크를 제공하는 단계는 복수의 가상 레이어-2 네트워크를 제공하는 단계를 포함하는

방법.

청구항 5

제 1 항에 있어서,

상기 애자일 에이전트는 상기 제1 머신과 연관된 소스 서버에 위치하거나 상기 제1 머신과 연관된 ToR(Top of Rack) 스위치에 위치하며, 상기 ToR 스위치는 상기 스위치가 아닌

방법.

청구항 6

제 1 항에 있어서,

상기 스위치를 선택하는 단계는 상기 제1 머신과 상기 제2 머신 사이의 상기 레이어-3 인프라구조의 개별 경로를 랜덤하게 선택하는 단계를 더 포함하는

방법.

청구항 7

제 6 항에 있어서,

상기 개별 경로를 선택하기 위해 발리언트 로드 밸런싱(valiant load balancing)을 이용하는 단계를 더 포함하는

방법.

청구항 8

제 6 항에 있어서,

상기 애자일 에이전트에 의해, 적어도 하나의 네트워크 성능 파라미터를 모니터링하는 단계와,

상기 적어도 하나의 네트워크 성능 파라미터에의 값에 기초하여 상기 개별 경로를 재선택하는 단계를 더 포함하는

방법.

청구항 9

시스템으로서,

애플리케이션 어드레스와 연관된 개별 컴퓨팅 장치 및 복수의 스위치를 포함하는 물리 네트워크 연결을 통해 연결된 복수의 컴퓨팅 장치, 및

상기 복수의 컴퓨팅 장치 중의 소스 컴퓨팅 장치와 연관된 애자일 에이전트

를 포함하고,

상기 애자일 에이전트는,

상기 복수의 컴퓨팅 장치 중의 목적지 컴퓨팅 장치의 할당된 애플리케이션 어드레스를 갖는 패킷을 인터셉트하고,

상기 소스 컴퓨팅 장치 중의 정의된 통신 그룹 내에 상기 목적지 컴퓨팅 장치가 존재하는지 여부를 결정 - 상기 정의된 통신 그룹은 애자일 디렉토리 서비스에 저장되어 있음 - 하고,

상기 목적지 컴퓨팅 장치가 상기 정의된 통신 그룹 내에 존재하는 경우, 상기 애자일 디렉토리 서비스를 이용하여 상기 할당된 애플리케이션 어드레스와 연관된 개별 스위치의 위치 어드레스를 검색하고,

상기 위치 어드레스로 상기 패킷을 전송할 상기 복수의 스위치 중의 상이한 개별 스위치를 선택 - 상기 상이한 개별 스위치는 로드 밸런싱을 제공하기 위해 선택됨 - 하고,

상기 패킷을 캡슐화하고 상기 캡슐화된 패킷에 상기 개별 스위치의 상기 위치 어드레스를 할당하고,

상기 상이한 개별 스위치를 통해 상기 개별 스위치로 상기 캡슐화된 패킷을 전송하되, 상기 개별 스위치는 상기 캡슐화된 패킷의 수신 시에 상기 패킷을 역캡슐화하고 상기 역캡슐화된 패킷을 상기 목적지 컴퓨팅 장치의 상기 애플리케이션 어드레스로 물리 네트워크 연결을 통해 전송

하도록 구성되고,

상기 목적지 컴퓨팅 장치가 상기 정의된 통신 그룹 내에 존재하지 않는 경우, 상기 애자일 디렉토리 서비스는 상기 개별 스위치의 위치 어드레스를 제공하는 것을 거부하는

시스템.

청구항 10

제 9 항에 있어서,

상기 할당된 애플리케이션 어드레스와 연관된 상기 개별 스위치는 ToR 스위치인

시스템.

청구항 11

제 9 항에 있어서,

상기 애자일 에이전트는 복수의 애자일 에이전트를 포함하고, 상기 애자일 에이전트의 부분 집합(sub-set)들 및 상기 복수의 컴퓨팅 장치는 서버 랙(server rack) 내에 조직화되어 있고, 상기 애자일 에이전트는 복수의 서버 랙에 통신 가능하게 연결되도록 구성되는 네트워크 케이지(network cage)를 포함하는

시스템.

청구항 12

제 9 항에 있어서,

상기 복수의 스위치는 상기 복수의 컴퓨팅 장치 중 적어도 일부와 함께 플러깅 가능한 컨테이너(pluggable container)에 배치되는 중간 스위치 및 집선 스위치(aggregate switch)를 포함하는

시스템.

청구항 13

제 9 항에 있어서,

상기 복수의 스위치는 레이어-3 인프라구조를 포함하고, 상기 패킷을 캡슐화하는 것은 상기 레이어-3 인프라구조를 통한 전송을 위해 레이어-3 패킷을 이용해 가상 레이어-2 패킷을 캡슐화하는 것을 포함하는

시스템.

청구항 14

제 9 항에 있어서,

상기 시스템은 애자일 디렉토리 서비스를 더 포함하고, 상기 애자일 디렉토리 서비스는 상기 복수의 스위치의 부분 집합을 컴퓨팅 장치의 개별 그룹에 할당하도록 구성되는

시스템.

청구항 15

제 9 항에 있어서,

상기 시스템은 애자일 디렉토리 서비스를 더 포함하고, 상기 애자일 디렉토리 서비스는 특정 애플리케이션 어드레스 및 특정 위치 어드레스를 개별 고객들에게 할당하고 어느 애플리케이션 어드레스들이 서로 간에 통신하도록 허용되는지 지정하도록 구성되는

시스템.

청구항 16

제 9 항에 있어서,

상기 시스템은 애자일 디렉토리 서비스를 더 포함하고, 상기 애자일 디렉토리 서비스는 상기 소스 컴퓨팅 장치 및 상기 목적지 컴퓨팅 장치 사이의 경로를 애니캐스트 어드레스(anycast address) 또는 발리언트 로드 밸런싱을 이용해 개별 스위치를 통해 맵핑(map)하도록 구성되는

시스템.

청구항 17

제 9 항에 있어서,

상기 애자일 에이전트는 상기 패킷의 포워딩 경로 상에 위치한 하드웨어를 포함하고, 상기 하드웨어는 상기 패킷을 인터셉트하도록 구성되며, 상기 포워딩 경로는 상기 소스 컴퓨팅 장치와 상기 목적지 컴퓨팅 장치 사이에 위치하고, 상기 애자일 에이전트는 상기 포워딩 경로에서 실행되어 상기 패킷의 인터셉트가 수행되도록 하는 소프트웨어 인스트럭션을 더 포함하는

시스템.

청구항 18

제 17 항에 있어서,

상기 복수의 스위치는 상기 포워딩 경로를 따라 상기 애자일 에이전트보다 상기 목적지 컴퓨팅 장치로 향해 있

으며, 상기 복수의 스위치는 상기 애플리케이션 어드레스는 인식하지 못하고 상기 위치 어드레스만 인식하는

시스템.

청구항 19

서버로서,

컴퓨터 판독가능 명령어들을 실행하기 위한 적어도 하나의 프로세서, 및

상기 적어도 하나의 프로세서에 의해 실행 가능한 애자일 에이전트

를 포함하고,

상기 애자일 에이전트는

애플리케이션 어드레스를 가지는 다른 서버로의 전송을 위해 패킷을 수신하고,

애자일 디렉토리 서비스에 액세스하여 상기 서버의 정의된 서버 그룹이 상기 다른 서버를 포함하는지 여부를 결정 - 상기 정의된 서버 그룹은 상기 애자일 디렉토리 서비스에 저장됨 - 하고,

상기 정의된 서버 그룹이 상기 다른 서버를 포함하는 경우, 상기 애자일 디렉토리 서비스를 이용하여 상기 다른 서버와 연관된 스위치의 위치 어드레스를 검색하고, 로드 밸런싱을 제공하기 위해 복수의 중간 스위치 중에서 선택된 중간 스위치를 통한 물리 네트워크 연결을 거치는 전송을 위해 상기 패킷을 캡슐화하되, 상기 캡슐화된 패킷에는 상기 다른 서버와 연관된 상기 스위치의 상기 위치 어드레스가 할당되고, 상기 캡슐화된 패킷의 수신 시에 상기 다른 서버와 연관된 상기 스위치는 상기 캡슐화된 패킷을 역캡슐화하고 상기 역캡슐화된 패킷을 물리 네트워크 연결을 거쳐 상기 애플리케이션 어드레스를 갖는 상기 다른 서버로 전송하도록 구성되고,

상기 정의된 서버 그룹이 상기 다른 서버를 포함하지 않는 경우, 상기 애자일 디렉토리 서비스는 상기 위치 어드레스를 거부하는

서버.

청구항 20

제 19 항에 있어서,

상기 애자일 에이전트는 상기 로드 밸런싱과 연관된 경로 제어를 제공하도록 구성되는

서버.

청구항 21

제 19 항에 있어서,

상기 정의된 서버 그룹은 상기 애자일 디렉토리 서비스에 의해 저장되는

서버.

청구항 22

제 19 항에 있어서,

상기 애자일 디렉토리 서비스는 상기 애플리케이션 어드레스의 상기 위치 어드레스로의 맵핑을 저장하는

서버.

발명의 설명

배 경 기 술

[0001] 종래의 데이터 센터 네트워크 아키텍처는 이들의 민첩성(agility)(데이터 센터 네트워크의 임의의 서버를 임의의 서비스에 할당하는 이들의 능력)을 손상시킬 수 있는 다수의 디자인 결점을 겪는다. 첫째로, 종래의 네트워크의 구성은 통상적으로 사실상 트리형(tree-like)이고, 비교적 고가의 장비로 이루어진다. 이는 예비 용량이 네트워크 내의 다른 장소에서 이용 가능할 때에도 연산 핫스팟의 혼잡 및 전개를 초래할 수 있다. 둘째로, 종래의 데이터 센터 네트워크는 하나의 서비스 내의 트래픽 플러드(traffic flood)가 그 주위의 다른 서비스에 영향을 미치는 것을 방지하는데 쓸모가 없다. 하나의 서비스가 트래픽 플러드를 경험할 때, 동일한 네트워크 서브 트리를 공유하는 모든 이들 서비스가 부수적인 손상을 겪게 되는 것이 통상적이다. 셋째로, 종래의 데이터 센터 네트워크 내의 라우팅 디자인은 통상적으로 인터넷 프로토콜(IP) 어드레스에 위상적으로 중요한 서버를 할당하고 가상 근거리 통신망(VLAN) 사이에 서버를 분할함으로써 스케일을 성취한다. 그러나, 이는 서버가 서비스들 사이에 재할당될 때 막대한 구성 부담을 생성할 수 있고, 따라서 데이터 센터의 리소스를 추가로 분해한다. 더욱이, 인간 개입이 통상적으로 이들 재구성에 요구될 수 있어, 따라서 이 프로세스의 속도를 제한한다. 마지막으로, 종래의 데이터 센터 네트워크를 구성하는데 있어서의 어려움 및 이러한 네트워크 내에 사용되는 장비의 비용과 같은 다른 고려 사항이 또한 이들 네트워크의 민첩성에 부정적인 영향을 미칠 수 있다.

발명의 내용

과제의 해결 수단

[0002] 이 특허 출원은 무엇보다도, 데이터 센터 내에 이용될 수 있는 민첩한 네트워크 아키텍처에 관한 것이다. 일 구현예는 레이어-3 인프라구조의 서버와 같은 가상 레이어-2 네트워크 접속 머신을 제공한다.

[0003] 다른 구현예는 복수의 스위치를 경유하여 통신적으로 결합된 복수의 컴퓨팅 디바이스를 포함한다. 개별 컴퓨팅 디바이스는 애플리케이션 어드레스와 관련될 수 있다. 개별 컴퓨팅 디바이스가 소스로서 작용하도록 구성 가능할 수 있고, 다른 개별 컴퓨팅 디바이스가 목적지로서 작용하도록 구성 가능할 수 있다. 상기 소스 컴퓨팅 디바이스는 상기 목적지 컴퓨팅 디바이스의 애플리케이션 주소에 패킷을 송신하도록 구성될 수 있다. 이 구현예는 또한 패킷을 인터셉트하고 목적지 컴퓨팅 디바이스와 관련된 위치 어드레스를 식별하고 패킷을 위치 어드레스에 송신하기 위해 이를 통해 개별 스위치를 선택하도록 구성된 애자일 에이전트를 포함할 수 있다.

[0004] 상기 열거된 구현예는 소개를 목적으로 제공된 것이고, 모든 청구된 요지를 포함하고 그리고/또는 한정하는 것은 아니다.

도면의 간단한 설명

[0005] 첨부 도면은 본 출원에 시사된 개념의 구현예를 도시한다. 도시된 구현예의 특징은 첨부 도면과 관련하여 취한 이하의 설명을 참조하여 보다 쉽게 이해될 수 있다. 다양한 도면에서 유사한 도면 부호가 유사한 요소를 지시하기 위해 실행 가능한 경우마다 사용된다. 또한, 각각의 도면 부호의 최좌측 숫자는 도면 부호가 처음으로 소개된 도면 및 관련 설명을 시사한다.

1 내지 도 6은 본 발명의 개념의 몇몇 구현예에 따른 민첩한 네트워크 아키텍처의 예를 도시하는 도면.

도 7 내지 도 9는 본 발명의 개념의 몇몇 구현예에 따른 민첩한 네트워크 데이터 센터 레이아웃의 예를 도시하는 도면.

도 10은 본 발명의 개념의 몇몇 구현예에 따라 성취될 수 있는 민첩한 네트워크 방법의 흐름도.

발명을 실시하기 위한 구체적인 내용

[0006] 개요

[0007] 본 특허 출원은 무엇보다도, 데이터 센터 내에 이용될 수 있는 민첩한 네트워크 아키텍처에 관한 것이다. 클라우드 서비스(cloud service)가 수만 또는 수십만개의 서버를 잠재적으로 보유하는 거대한 데이터 센터의 생성을 추진한다. 이들 데이터 센터는 큰 동적 수의 별개의 서비스(웹 앱들(apps), 이메일, 맵 리듀스(map-reduce) 클

러스터 등)를 동시에 지원할 수 있다. 클라우드 서비스 데이터 센터의 구현예는 스케일 아웃 디자인, 즉 요구되는 바에 따라 서비스들 사이에 신속하게 재할당될 수 있는 리소스(예를 들어, 서버)의 큰 풀(pool)을 통해 성취된 신뢰성 및 성능에 의존할 수 있다. 임의의 서비스에 데이터 센터의 임의의 서버를 할당하는 능력은 데이터 센터 네트워크의 민첩성을 고려할 수 있다. 막대한 비용과 관련될 수 있는 데이터 센터의 이득을 효과적으로 지레 작용하기 위해, 네트워크 민첩성이 가치가 있을 수 있다. 네트워크 민첩성이 없이, 데이터 센터 서버 리소스는 궁지에 몰리게 될 수 있고, 따라서 돈이 낭비된다.

[0008] 제 1 예시적인 민첩한 네트워크 아키텍처

[0009] 소개를 목적으로, 민첩한 네트워크 아키텍처(100)의 예를 도시하는 도 1 내지 도 2를 고려한다. 민첩한 데이터 네트워크 아키텍처(100)는 서버(102(1), 102(2), 102(3), 102(N))와 같은 복수의 서버측 컴퓨팅 디바이스를 포함할 수 있다.

[0010] 용어 서버 및 머신은 데이터를 송신하거나 수신할 수 있는 임의의 디바이스를 칭하는 것으로 이해되어야 한다. 예를 들어, 이들 용어는 물리적 서버, 서버 상에서 실행되는 가상 머신(예를 들어, 가상화 기술을 사용하는), 단일 운영 체제를 실행하는 컴퓨팅 디바이스, 하나 초과의 운형 체제를 실행하는 컴퓨팅 디바이스, 상이한 운영 체제(예를 들어, 마이크로소프트 윈도우, 리눅스, FreeBSD)를 실행하는 컴퓨팅 디바이스, 서버 이외의 컴퓨팅 디바이스(예를 들어, 랩탑, 어드레스 가능한 전원) 또는 컴퓨팅 디바이스의 부분(예를 들어, 네트워크 연결 디스크, 네트워크 연결 메모리, 저장 서브시스템, 저장 영역 네트워크(SAN), 그래픽 프로세싱 유닛, 수치 가속기, 양자화 컴퓨팅 디바이스)를 칭하는 것이라는 것이 이해되어야 한다.

[0011] 민첩한 네트워크 아키텍처(100)는 서버의 수에 대한 확장성(scalibility)을 촉진할 수 있다. 확장성이 성취될 수 있는 일 방식은 애플리케이션 어드레스를 이용하는 서버(102(1) 내지 102(N))에 대한 이더넷형 플랫 어드레싱을 생성하는 것이다. 이더넷 레이어-2 시맨틱(semantic)은 임의의 인터넷 프로토콜(IP) 어드레스가 서버가 근거리 통신망(LAN) 상에 있는 것처럼 임의의 네트워크 포트에 접속된 임의의 서버에 할당될 수 있는 플랫 어드레싱을 지원하는 네트워크 상태를 성취하는 것과 관련될 수 있다.

[0012] 이 경우에, 애플리케이션 어드레스(AA)(104(1), 104(2), 104(3), 104(N))는 각각의 서버(102(1), 102(2), 102(3), 102(N))에 각각 할당될 수 있다. 서버 관점으로부터, 임의의 서버는 관련된 애플리케이션 어드레스(104(1), 104(2), 104(3), 104(N))를 경유하여 임의의 다른 서버에 토크(talk)할 수 있다. 이는 서버(102(1), 102(2), 102(3), 102(N))를 포함하는 근거리 통신망(LAN)에 대해 유효할 수 있는 모든 것들을 포함하는 애플리케이션 어드레스가 임의의 방식으로 배열될 수 있기 때문에 레이어-2 기능성인 것으로 고려될 수 있다. 그러나, 이하에 설명되는 바와 같이, 몇몇 구현예에서, 민첩한 네트워크 아키텍처의 기초 인프라구조는 도면 부호 106에 지시된 바와 같이 레이어-3일 수 있다. 따라서, 이들 구현예는 레이어-3 인프라구조(106) 상에(또는 이용하여) 가상 레이어-2 네트워크(108)를 생성할 수 있다. 동일한 레이어-3 인프라구조(106) 상에 생성된 하나 초과의 가상 레이어-2 네트워크(108)가 존재할 수 있고, 각각의 서버는 이들 가상 레이어-2 네트워크(108)의 하나 이상에 속할 수 있다.

[0013] 도 2는 인터넷(204)을 경유하여 민첩한 네트워크 아키텍처(100)에 접속되는 외부 클라이언트(202)를 소개한다. 민첩한 네트워크 아키텍처(100)는, 외부 클라이언트가 애플리케이션 어드레스(104(1) 내지 104(N))의 지식을 갖는 외부 클라이언트 없이 서버(102(1) 내지 102(N)) 중 하나 이상에 할당된 글로벌 또는 위치 어드레스(206)와 통신할 수 있게 할 수 있다. 이들 개념은 도 3 내지 도 5의 설명과 관련하여 이하에 더 상세히 설명된다.

[0014] 제 2 예시적인 민첩한 네트워크 아키텍처

[0015] 도 3은 전술된 개념이 구현될 수 있는 예시적인 민첩한 네트워크 아키텍처(300)를 도시한다. 이 경우에, 외부 클라이언트(302)는 인터넷(306) 및/또는 다른 네트워크를 경유하여 민첩한 시스템(304)과 통신할 수 있다. 이 구현예에서, 민첩한 시스템(304)은 일반적으로 도면 부호 308로 지시되고 도면 부호 308(1) 내지 308(N)으로 구체적으로 지시된 라우터의 세트와, 도면 부호 310으로 일반적으로 지시되고 도면 부호 310(1), 310(2) 및 310(N)으로 구체적으로 지시된 복수의 중간 스위치, 도면 부호 312로 일반적으로 지시되고 도면 부호 312(1), 312(2) 및 312(N)으로 구체적으로 지시된 복수의 집선 스위치, 일반적으로 도면 부호 314로 지시되고 도면 부호 314(1), 314(2) 및 314(N)으로 구체적으로 지시된 복수의 탑 오브 랙(top of rack)(TOR 또는 ToR) 스위치, 및 일반적으로 도면 부호 316으로 지시되고 도면 부호 316(1), 316(2), 316(3), 316(4), 316(5) 및 316(N)으로 구

체적으로 지시된 복수의 서버를 포함한다. 도면 페이지의 공간 제약에 기인하여, 단지 6개의 서버(316(1) 내지 316(N))만이 여기에 도시되어 있지만, 민첩한 시스템(304)은 수천, 수만, 수십만개 또는 그 이상의 서버를 즉시 수용할 수 있다. 간략화를 위해 그리고 도면 페이지의 공간 제약에 기인하여, 구성 요소들 사이의 모든 접속부(즉, 통신 경로)가 도 3 내지 도 8에 도시되어 있지는 않다.

[0016] 서버(316(1), 316(2))는 서버 랙(318(1))으로서 TOR 스위치(314(1))와 관련된다. 유사하게, 서버(316(3), 316(4))는 서버 랙(318(2))으로서 TOR 스위치(314(2))와 관련되고, 서버(316(5), 316(N))는 서버 랙(318(N))으로서 TOR 스위치(314(N))와 관련된다. 재차, 이는 도면 페이지의 공간 제약에 기인하고, 종종 서버 랙은 10개 이상의 서버를 포함한다. 또한, 개별 서버는 애자일 에이전트와 관련될 수 있다. 예를 들어, 서버(316(1))는 애자일 에이전트(320(1))와 관련된다. 유사한 관계가 서버(316(2) 내지 316(N))와 애자일 에이전트(320(2) 내지 320(N)) 각각 사이에 표시되어 있다.

[0017] 애자일 에이전트(321(1) 내지 320(N))의 기능이 이하에 더 상세히 설명된다. 간략하게, 애자일 에이전트는 개별 서버들 사이의 통신을 용이하게 할 수 있다. 이 특정 예에서, 애자일 에이전트는 컴퓨터 판독 가능 명령으로서 서버 상에 저장된 논리적 모듈로서 고려될 수 있다. 다른 구현예는 서버의 세트를 서빙하는 애자일 에이전트(320)가 예를 들어 TOR 스위치(314) 또는 중간 스위치(310)와 같은 스위치 상에 위치되는 구성을 포함할 수 있다. 스위치 상에 위치될 때, 애자일 에이전트는 서버(316)로부터 중간 스위치(310)를 향해 네트워크 위로 흐르기 때문에 패킷을 프로세싱할 수 있다. 이러한 구성에서, 애자일 에이전트(320)는 포워딩 경로에서 또는 스위치의 제어 프로세서 내에서 실행되는 패킷 포워딩 경로 및 소프트웨어 명령 상에서 주문형 하드웨어의 조합을 사용하여 구현될 수 있다.

[0018] 민첩한 시스템(304)은 3개의 디렉토리 서비스 모듈(322(1) 내지 322(N))을 추가로 포함한다. 디렉토리 서비스 모듈의 예시된 수는 민첩한 시스템에 임계적인 것은 아니고, 다른 구현예가 더 적거나 더 많은 디렉토리 서비스 모듈(및/또는 다른 예시된 구성 요소)을 이용할 수 있다. 디렉토리 서버의 기능이 이하에 더 상세히 설명된다. 간략하게, 디렉토리 서비스 모듈은 다른 정보 중에서도, 민첩한 시스템(304)을 통한 통신을 용이하게 하기 위해 애자일 에이전트(320(1) 내지 320(N))(및/또는 다른 구성 요소)에 의해 이용될 수 있는 애플리케이션 어드레스 대 위치 어드레스 맵핑(정방향 및 역방향 맵핑 중 하나 또는 모두)을 포함할 수 있다. 이 경우, 디렉토리 서비스 모듈(322(1) 내지 322(N))은 특정 서버(316(1), 316(3), 316(5))와 관련된다. 다른 구성에서, 디렉토리 서비스 모듈은 데이터 센터 제어 서버, 스위치 및/또는 전용 컴퓨팅 디바이스와 같은 다른 구성 요소에서 발생할 수 있다.

[0019] 민첩한 시스템(304)은 2개의 논리적 그룹을 포함하는 것으로서 고려될 수 있다. 제 1 논리적 그룹은 도면 부호 326으로 지시된 바와 같이 위치 또는 글로벌 어드레스를 전송하는 링크 상태 네트워크이다. 제 2 논리적 그룹은 도면 부호 328로 지시된 바와 같이 애플리케이션 어드레스를 소유하는 대체 가능한 서버의 풀이다. 간략하게, 링크 상태 네트워크(326)의 구성 요소는 서버(328)의 풀 내의 어느 서버가 어느 애플리케이션 어드레스를 현재 사용하는지를 추적하기 위해 정보를 교환할 필요가 없다. 또한, 서버의 관점으로부터, 서버는 다른 서버의 애플리케이션 어드레스를 경유하여 서버 풀(328) 내의 임의의 다른 서버와 통신할 수 있다. 이 프로세스는 서버에 투명하게 되는 방식으로 애자일 에이전트, 디렉토리 서비스 및/또는 다른 구성 요소에 의해 용이해진다. 다른 방식으로 말하면, 프로세스는 서버 상에서 실행하는 애플리케이션에 투명할 수 있지만, 서버 상의 다른 구성 요소가 프로세스를 인식할 수 있다.

[0020] 라우터(308), 중간 스위치(310), 집선 스위치(312), TOR 스위치(314) 및 서버(316(1) 내지 316(N))는 예를 들어 레이어-3 기술을 사용하여 통신적으로 결합될 수 있다. 개별 서버의 관점으로부터, 다른 서버와의 통신은 레이어-2 통신(즉, 가상 레이어-2)으로서 나타난다. 그러나, 서버 랙(318(1))의 소스 서버(316(1))로부터 서버 랙(318(2))의 목적지 서버(316(3))로와 같은 랙간 통신(inter-rack communication)은 실제로 레이어-3 인프라 구조를 통해 발생한다. 예를 들어, 애자일 에이전트(320(1))가 통신{즉, 서버(316(3))의 애플리케이션 어드레스에 어드레스되는 패킷}을 인터셉트할 수 있고, 그 전송을 용이하게 할 수 있다.

[0021] 애자일 에이전트(320(1))는 서버(316(3))와 관련된 위치 어드레스로의 애플리케이션 어드레스의 맵핑을 얻기 위해 디렉토리 서비스 모듈(322(1) 내지 322(N)) 중 하나 이상에 액세스할 수 있다. 예를 들어, 맵핑된 위치 어드레스는 TOR 스위치(314(2))일 수 있다. 애자일 에이전트는 위치 어드레스로 패킷을 캡슐화(encapsulation)할 수 있다. 애자일 에이전트는 이어서 캡슐화된 패킷을 송신하거나 바운스하기 위해 개별(또는 세트) 집선 및/또는 중간 스위치(들)를 선택할 수 있다. 이 선택 프로세스의 특징은 이하에 더 상세히 설명된다. TOR 스위치(314(2))에서 캡슐화된 패킷의 수신시에, TOR 스위치는 패킷을 역캡슐화(decapsulation)하고 서버(316(3)) 상에

패킷을 송신할 수 있다. 대안 실시예에서, 위치 어드레스는 서버(316(3)) 또는 서버(316(3)) 상에서 실행되는 가상 머신과 관련될 수 있고, 패킷은 목적지 서버 자체 상에 역캡슐화될 수 있다. 이 실시예에서, 서버 또는 가상 머신에 할당된 위치 어드레스는 애플리케이션 어드레스가 다른 호스트가 이들과 통신하는데 사용하는 어드레스인 LAN에 의해 접속되는 애플리케이션에 대한 환영을 유지하기 위해 서버 상에서 동작하는 다른 애플리케이션으로부터 은폐될 수 있다.

[0022] 대안 실시예에서, 패킷은 레이어-3/레이어-2 경계를 교차할 때 다른 구성 요소에 의해 역캡슐화될 수 있다. 예를 들어, 역캡슐화를 수행할 수 있는 구성 요소의 예는 가상 머신 모니터의 하이퍼바이저 및/또는 루트 파티션을 포함할 수 있다.

[0023] 이 구성은 서버가 많은 수로 서버 풀(328)에 추가될 수 있게 하고, 또한 서버의 관점으로부터 다른 서버가 이들이 동일한 서브 네트워크 상에 있는 것처럼 나타날 수 있다. 대안적으로 또는 추가적으로, 링크 상태 네트워크(326)의 구성 요소는 서버 애플리케이션 어드레스를 인식할 필요가 없다. 또한, 서버가 추가되거나 제거될 때와 같이 어드레스 정보가 변경될 때마다, 디렉토리 서버(들)는 다수의 상이한 유형의 구성 요소를 업데이트하는 것보다는 간단히 업데이트될 수 있다.

[0024] 요약하면, 레이어-2 시맨틱은 임의의 IP 어드레스가 서버가 LAN 상에 있는 것처럼 임의의 네트워크 포트에 접속된 임의의 서버에 할당될 수 있는 플랫 어드레싱을 지원하는 네트워크 상태를 성취하는 것과 관련될 수 있다. 또한, 링크 상태 네트워크(326) 내의 구성 요소(즉, 스위치)는 링크 상태 네트워크 내의 다른 구성 요소를 인식할 수 있지만, 서버 풀(328)의 구성 요소를 인식할 필요는 없다. 또한, TOR 스위치는 이들의 각각의 랙 내의 서버에 대해 인지할 수 있지만, 다른 랙의 서버에 대해 인지할 필요는 없다. 또한, 애자일 에이전트는 서버 애플리케이션 어드레스(AA) 패킷을 인터셉트하고 AA의 목적지 컴퓨팅 디바이스와 관련된 위치 어드레스(LA)를 식별할 수 있다. 애자일 에이전트는 이어서 LA에 패킷을 전송하기 위해 이를 통해 개별 스위치(또는 스위치의 세트)를 선택할 수 있다. 이 경우에, 개별 스위치는 이용 가능한 스위치 중 임의의 하나 이상을 포함할 수 있다.

[0025] 이 구성은 또한 서비스와 관련하는 다른 서버 특징을 용이하게 한다. 예를 들어, 디렉토리 서비스 모듈(322(1) 내지 322(N)) 내에 포함될 수 있는 것과 같은 데이터 센터 관리 소프트웨어는 임의의 서비스에 임의의 서버(316(1) 내지 316(N))를 할당할 수 있고 서비스가 예측되는 어떠한 IP 어드레스를 갖는 그 서버를 구성할 수 있다. 각각의 서버의 네트워크 구성은 접속되는 경우 LAN을 경유하는 것일 수 있는 것과 동일할 수 있고, 링크-로컬 브로드캐스트와 같은 특징이 지원될 수 있다. 서비스들 사이의 통신 격리의 목적은 서비스 및 통신 그룹을 규정하기 위해 용이하고 일관적인 애플리케이션 프로그램 인터페이스(API)를 제공하는 것과 관련될 수 있다. 이와 관련하여, 디렉토리 서비스는 서비스와 관련된 서버의 그룹(예를 들어, 고객)을 규정할 수 있다. 완전 접속성이 그룹 내의 서버들 사이에 허용될 수 있고, 액세스 제어 리스트(ACL)와 같은 정책이 상이한 그룹 내의 어느 서버가 통신하도록 허용되어야 하는지를 지배하기 위해 지정될 수 있다.

[0026] 상기 구성은 또한 트래픽 관리에 적합하다. 설명의 목적으로, 제 1 고객이 민첩한 시스템(304)의 서버에 의해 수행될 서비스에 대해 비교적 높은 요금을 지불하고 이에 따라 비교적 높은 품질의 서비스 동의를 얻는 것을 가정한다. 또한, 제 2 고객은 비교적 낮은 요금을 지불하고 이에 따라 비교적 낮은 품질의 서비스 동의를 수신하는 것을 가정한다. 이러한 경우에, 중간 스위치(310(1) 내지 310(N))의 비교적 높은 비율, 또는 모두가 제 1 고객을 위한 트래픽을 취급하기 위해 할당될 수 있고, 반면에 더 적은 수의 스위치가 제 2 고객에 할당될 수 있다. 다른 방식으로 말하면, 제 1 서브 세트의 스위치는 제 1 고객에 할당될 수 있고, 제 2 서브 세트의 스위치는 제 2 고객에 할당될 수 있다. 제 1 및 제 2 세트는 서로 배제적이거나 중첩할 수 있다. 예를 들어, 몇몇 구현예에서, 개별 스위치는 특정 고객에 전용되거나 또는 다수의 고객에 할당될 수 있다. 예를 들어, 중간 스위치(310(1))는 양 고객에 할당될 수 있고, 반면 중간 스위치(310(2), 310(N))가 제 1 고객에 독점적으로 할당될 수 있다.

[0027] 요약하면, 이하에 더 상세하게 설명되는 바와 같이, 민첩한 네트워크 아키텍처(300)는 이하의 목적, 즉 서버들 사이의 균일한 높은 용량, 서비스들 사이의 성능 격리, 이더넷 레이어-2 시맨틱 및/또는 서비스들 사이의 통신 격리 중 하나 이상과 관련될 수 있다. 서버들 사이의 균일한 높은 용량의 목적은 네트워크 내의 트래픽 흐름의 전송율이 송신 및 수신 서버의 네트워크 인터페이스 카드 상의 이용 가능한 용량에 의해서를 제외하고는 주로 제한되지 않는 네트워크 상태를 성취하는 것과 관련될 수 있다. 이와 같이, 개발자의 관점으로부터, 이 목적을 성취함으로써, 네트워크 토폴로지는 서비스에 서버를 추가할 때 더 이상 주요 관심 사항이 아닐 수 있다. 서비스들 사이의 성능 격리의 목적은 하나의 서비스의 트래픽이 각각의 서비스가 개별 물리적 스위치에 의해 접속되어 있는 것처럼 임의의 다른 서비스에 의해 취급된 트래픽에 의해 영향을 받지 않는 네트워크 상태를 성취하는

것과 관련될 수 있다. 이더넷 레이어-2 시맨틱의 목적은 거의 임의의 IP 어드레스가 서버가 LAN 상에 있는 것처럼 임의의 네트워크 포트에 접속된 임의의 서버에 할당될 수 있는 플랫 어드레싱을 지원하는 네트워크 상태를 성취하는 것과 관련될 수 있다. 이와 같이, 데이터 센터 관리 소프트웨어는 임의의 서비스에 임의의 서버를 할당할 수 있고, 서비스가 예측하는 어떠한 IP 어드레스를 갖는 이 서버를 구성한다.

[0028] 각각의 서버의 네트워크 구성은 접속되는 경우 LAN을 경유하는 것일 수 있는 것과 동일할 수 있고, 링크-로컬 브로드캐스트와 같은 특징이 지원될 수 있다. 서비스들 사이의 통신 격리의 목적은 서비스 및 통신 그룹을 규정하기 위해 용이하고 일관적인 API를 제공하는 것과 관련될 수 있다. 이와 관련하여, 서버의 그룹을 규정하는 디렉토리 시스템{즉, 예를 들어 디렉토리 서비스 모듈(322(1) 내지 322(N))을 경유하여}이 제공될 수 있다. 완전 접속성이 그룹 내의 서버들 사이에 허용될 수 있고, 정책이 상이한 그룹 내의 어느 서버가 통신하도록 허용되어야 하는지를 지배하기 위해 지정될 수 있다.

[0029] 설명된 민첩한 네트워크 아키텍처를 이용함으로써, 이하의 네트워크 특징, 즉 (1) 서비스 인스턴스(instance)가 네트워크 내의 임의의 위치에 배치될 수 있게 하는 플랫 어드레싱, (2) 네트워크 경로를 가로질러 균일하게 트래픽을 확산시키기 위해 랜덤화를 사용하는 로드 밸런싱(예를 들어, 발리언트 로드 밸런싱(VLB)) 및 (3) 큰 서버 풀에 스케일링하면서 레이어-2 이더넷 시맨틱을 성취하기 위한 새로운 종료 시스템 기반 어드레스 결정 서비스 중 하나 이상과 관련되는 데이터 센터 네트워크가 제공될 수 있다.

[0030] 전술된 목적을 성취하기 위해, 적어도 몇몇 실시예에서, 이하의 민첩한 네트워크 아키텍처 디자인 원리 중 하나 이상이 다양한 구현예에서 이용될 수 있다.

[0031] 광대한 경로 다이버시티를 갖는 토폴로지의 이용

[0032] "메시(meshy)" 토폴로지를 이용함으로써, 서버의 개별 세트들 사이의 다수의 경로가 제공될 수 있다. 예를 들어, 서버 랙(318(1))의 서버와 서버 랙(318(N))의 서버 사이의 통신은 TOR 스위치(314(1))로부터 임의의 집선 스위치(312(1) 내지 312(2))를 통해 임의의 중간 스위치(310(1) 내지 310(N))로 진행할 수 있다. 중간 스위치로부터, 통신은 집선 스위치(312(2) 내지 312(N)) 중 어느 하나를 통해 TOR 스위치(314(N))로 진행할 수 있다.

[0033] 이 구성은 다수의 이득을 생성할 수 있다. 예를 들어, 다수의 경로의 존재는 파라미터의 조정 또는 명시적인 트래픽 엔지니어링을 위한 요구 없이 네트워크로부터 혼잡의 감소 및/또는 제거를 가능하게 할 수 있다. 또한, 다수의 경로는 "스케일 아웃(scale-out)" 네트워크 디자인을 허용한다. 달리 말하면, 더 많은 용량이 더 많은 저비용 스위치를 추가함으로써 추가될 수 있다. 대조적으로, 통상의 이력 네트워크 디자인은 이력의 높은 레벨에 하나 또는 매우 소수의 링크에 트래픽을 집중시킨다. 그 결과, 통상의 네트워크는 고밀도의 트래픽에 대처하기 위해 고가의 "빅 아이언(big iron)" 스위치의 구매를 필요로 할 수 있다.

[0034] 또한, "메시" 토폴로지를 이용함으로써, 다수의 경로가 링크 또는 스위치가 고장남에 따라 적절한 열화를 허용할 수 있다. 예를 들어, 소정의 레이어에서 "n개의" 스위치를 갖는 설명된 민첩한 데이터 센터 네트워크 아키텍처에 따라 구현된 민첩한 네트워크는 그 용량의 50%를 손실할 수 있는 통상의 네트워크에 비교할 때 스위치가 고장날 때 그 용량의 단지 1/n만을 손실할 수 있다. 설명된 민첩한 데이터 네트워크 아키텍처에 따라 구현된 민첩한 네트워크는 잠재적으로 완전한 이분형(bipartite) 토폴로지를 이용할 수 있다.

[0035] 어드레스 휘발성의 랜덤화

[0036] 데이터 센터는 이들의 작업부하, 이들의 트래픽 및 이들의 고장 패턴에 거대한 휘발성을 가질 수 있다. 따라서, 리소스의 큰 풀이 생성될 수 있다. 작업은 이어서 이들 상에 랜덤하게 확산될 수 있는데, 최선의 경우의 몇몇 성능이 절충되어 최악의 경우를 평균 경우로 향상시킬 수 있다. 적어도 몇몇 실시예에서, 광대한 경로 다이버시티와 관련된 토폴로지(예를 들어, 도 3에서 명백한 바와 같이)가 이용될 수 있다. 작업 흐름은 발리언트 로드 밸런싱(VLB) 기술과 같은 로드 밸런싱 기술을 사용하여 토폴로지를 가로질러 라우팅될 수 있다. 간략하게, VLB 기술은 데이터 전송을 수행하는데 사용되는 경로 또는 경로들을 랜덤하게 선택하는 것을 수반할 수 있고, 여기서 경로는 일련의 링크 및/또는 스위치로 구성된다. 그 후에, 경로가 재선택될 수 있고, 여기서 재선택은 원래 경로를 포함하는 스위치 또는 링크 중 하나 이상을 변경하는 것을 수반한다. 재선택은 예를 들어 지정된 수의 바이트/패킷을 송신/수신한 후 또는 선택된 경로, 반응적으로 스위치 또는 링크와 관련된 전송 문제의 지시시와 같이 주기적으로 발생할 수 있다. 예를 들어, 패킷 지연 또는 다른 통신 장애가 검출되면, 선택

프로세스가 반복될 수 있다. 이 원리의 적용을 통해, 균일한 용량 및 성능 격리 목적이 부합될 수 있다.

[0037] 더 구체적으로, 데이터 센터 트래픽 매트릭스 내의 휘발성 및 불확실성을 처리하기 위해, 로드 밸런싱 기술(예를 들어, VLB)이 네트워크 경로를 가로질러 랜덤하게 흐름을 해시하는데 이용될 수 있다. 이 접근법의 목적은 호스트 트래픽 모델에서와 같이 네트워크 진입-진출 제약을 받는 중재 트래픽 편차에 대한 대역폭 보장을 제공하는 것일 수 있다. 간략하게, 소정의 경로 상의 데이터 전송율이 경로의 최저 또는 최대 제약된 부분을 초과할 수 없는 호스 모델이 지정된다.

[0038] 유리할 수 있는 흐름 입도(대부분의 흐름의 패킷이 경로를 재선택할 때를 제외하고는 네트워크를 통해 동일한 경로를 따르는 것을 의미함)에서 VLB와 같은 로드 밸런싱 기술을 사용하여, 흐름의 패킷이 재순서화되고 또는 목적지에서 인식되는 급속하게 변경하는 지연 시간을 경험할 수 있는 기회가 감소되고, 그리고/또는 흐름 내에서 MTU 차이에 기인하여 경로 최대 전송 유닛(MTU) 발견 프로토콜의 동작을 붕괴시킬 수 있다. 몇몇 유형의 트래픽(예를 들어, 패킷 재순서화에 의해 손상되지 않는 것들) 및 몇몇 환경(예를 들어, 모든 경로를 따라 매우 균일한 지연을 갖는 것들)은 패킷 입도에서 VLB와 같은 로드 밸런싱을 사용하도록 선호할 수 있다(잠재적으로 상이한 경로가 패킷의 시퀀스에서 각각의 패킷에 대해 사용되는 것을 의미함). 예를 들어 IP 5-튜플 흐름, IP 2-튜플 흐름 또는 2개의 서브넷 또는 어드레스 범위 사이의 패킷의 세트와 같은 흐름의 통상적으로 수용된 임의의 정의가 사용될 수 있다.

[0039] 민첩한 데이터 센터 네트워크를 제공하는 것과 관련하여, 진입-진출 제약은 서버라인 카드 속도에 대응할 수 있다. 높은 양분 대역폭 토폴로지(예를 들어, 폴디드 클로스(folded-clos) 토폴로지)와 조합하여, 로드 밸런싱 기술은 비간섭 패킷 교환된 네트워크(비차단 회로 교환된 네트워크의 대응부)를 생성하도록 이용될 수 있고, 서버 진입-진출 포트 속도를 초과하는 로드를 지속하지 않는 트래픽 패턴을 위한 핫스팟이 없는 성능을 제공할 수 있다. 이와 관련하여, 몇몇 구현예에서, 전송 제어 프로토콜(TCP 종단간 혼잡 제어 메커니즘은 호스 모델을 실행하고 과실행 서버 포트 속도를 회피하는데 이용될 수 있다. 이 원리는 스위치의 3개의 상이한 레이어, 즉 TOR(314), 집선(312) 및 중간(310)으로 이루어질 수 있는 도 3에 도시된 논리 토폴로지로 유도될 수 있다. 하나의 서버로부터 다른 서버로의 흐름은 TOR 및 집선 스위치를 가로질러 랜덤 중간 스위치를 경유하여 랜덤 경로를 취할 수 있다. 따라서, VLB와 같은 로드 밸런싱 기술은 지속적인 트래픽 혼잡을 배제하면서 이용을 원활하게 하기 위해 데이터 센터의 스위치간 패브릭과 관련하여 이용될 수 있다.

[0040] 위치로부터 명칭 분리

[0041] 위치로부터 명칭 분리는 새로운 특징을 구현하는데 사용될 수 있는 자유도를 생성할 수 있다. 이 원리는 데이터 센터 네트워크 내의 민첩성을 가능하게 하고 어드레스와 위치 사이의 바인딩이 발생할 수 있는 분열을 감소시킴으로써 이용을 향상시키기 위해 지레 작용될 수 있다. 이 원리 및 이하에 설명되는 종단 시스템 포함의 원리를 통해, 레이어-2 시맨틱 목적이 부합될 수 있다. 이와 같이, 개발자들은 네트워크 토폴로지에 관련하지 않고 이들의 애플리케이션 또는 네트워크 스위치를 재구성할 필요 없이 IP 어드레스를 할당하도록 허용될 수 있다.

[0042] 네트워크 민첩성을 향상시키기 위해(임의의 서버 상의 임의의 서비스의 지원, 서버 풀의 동적 성장 및 수축 및 작업부하 마이그레이션을 지원함), AA라 명명된 명칭 및 LA라 명명된 로케이터를 분리하는 IP 어드레싱 체계가 사용될 수 있다. 디렉토리 서비스 모듈(322(1) 내지 322(N))로서 명시될 수 있는 것과 같은 민첩한 디렉토리 서비스는 스케일 가능하고 신뢰적인 방식으로 AA와 LA 사이의 맵핑을 관리하도록 규정될 수 있다. 민첩한 디렉토리 서비스는 개별 서버 상의 네트워킹 스택에서 실행되는 심 레이어(shim layer)에 의해 호출될 수 있다. 도 3에 표현된 구현예에서, 이 심 레이어는 애자일 에이전트(320(1) 내지 320(N))로서 명시될 수 있다.

[0043] 종단 시스템의 포함

[0044] 데이터 센터 서버 상의 운영 체제를 포함하는 소프트웨어는 통상적으로 데이터 센터 내부에서 사용을 위해 광대하게 수정된다. 예를 들어, 새로운 또는 수정된 소프트웨어는 서버를 가로질러 데이터를 저장하기 위해 가상화 또는 블롭 파일 시스템을 위한 하이퍼바이저를 생성할 수 있다. 스위치 상에 소프트웨어를 변경하기보다는, 이 소프트웨어의 프로그램 가능성이 지레 작용될 수 있다. 더욱이, 스위치 또는 서버의 하드웨어로의 변경은 회피되거나 제한되고, 레가시 애플리케이션이 수정되지 않고 유지될 수 있다. 현재 이용 가능한 저가의 스위치 응용 특정 집적 회로(ASIC)의 제한 내에서 동작하기 위해 서버 상에 소프트웨어를 사용함으로써, 현재 구성되고

전개될 수 있는 디자인이 생성될 수 있다. 예를 들어, 브로드캐스트 어드레스 결정 프로토콜(ARP) 패킷에 의해 생성된 확장성 문제점은, 스위치 상의 소프트웨어 또는 하드웨어 변경을 경유하여 ARP를 제어하려고 시도하기보다는, 서버 상의 ARP 요구를 인터셉트하고 이들을 디렉토리 시스템으로의 룩업 요구로 변환함으로써 감소되고 그리고/또는 배제될 수 있다.

[0045] 도 4는 예시적인 애자일 에이전트(320(1))를 더 상세히 도시한다. 이 경우에, 애자일 에이전트(320(1))는 사용자 모드(406) 및 커널 모드(408)를 포함하는 서버 머신(402) 상에서 동작한다. 서버 머신은 사용자 모드에서 사용자 모드 에이전트(410)를 포함한다. 커널 모드는 TCP 구성 요소(412), IP 구성 요소(414), 캡슐화기(416), NIC(418) 및 라우팅 정보 캐시(420)를 포함한다. 서버 머신은 디렉토리 서비스(322(1))를 포함하고 그리고/또는 통신할 수 있다. 디렉토리 서비스는 서버 역할 구성 요소(422), 서버 건강 구성 요소(424) 및 네트워크 건강 구성 요소(426)를 포함할 수 있다. 애자일 에이전트(320(1))는 사용자 모드 에이전트(410), 캡슐화기(416) 및 라우팅 정보 캐시(420)를 포함할 수 있다. 캡슐화기(416)는 ARP를 인터셉트하고 이를 사용자 모드 에이전트(410)에 송신할 수 있다. 사용자 모드 에이전트는 디렉토리 서비스(322(1))에 질의할 수 있다. 커널 모드 구성 요소 내로의 사용자 모드 에이전트를 포함하고 예를 들어 IP 테이블 또는 IP 체인과 같은 메커니즘을 경유하여 또는 라우팅 테이블 룩업 중에 ARP 이외의 메커니즘을 경유하여 디렉토리 룩업을 호출하는 것과 같은 이들 블록의 다른 배열이 가능하다는 것이 이해되어야 한다.

[0046] 도 3의 민첩한 네트워크 아키텍처에서, 종단 시스템 제어가 새로운 기능성을 급속하게 주입하기 위한 메커니즘을 제공할 수 있다. 이와 같이, 민첩성 에이전트는 로드 밸런싱에 사용되는 랜덤화를 제어함으로써 최소 단위 경로 제어를 제공할 수 있다. 게다가, 명칭 및 로케이터의 분리를 실현하기 위해, 애자일 에이전트는 민첩한 디렉토리 서비스로의 질의로 이더넷의 ARP 기능성을 대체할 수 있다. 민첩한 디렉토리 서비스 자체는 스위치보다는 서버 상에서 실현될 수 있다. 이 민첩한 디렉토리 서비스는 서버 도달 가능성, 그룹화, 액세스 제어, 리소스 할당(예를 들어, 중간 스위치의 용량), 격리(예를 들어, 비중첩 중간 스위치) 및 동적 성장 및 수축의 미세 단위 제어를 허용한다.

[0047] 네트워크 기술의 지레 작용

[0048] 네트워크 스위치 내의 강인한 구현을 갖는 하나 이상의 네트워크 기술을 이용하는 것은 민첩한 네트워크의 디자인을 간단화하고 이러한 네트워크를 전개하기 위한 조작자 자발성을 증가시킬 수 있다. 예를 들어, 적어도 몇몇 실시예에서, 링크 상태 라우팅 프로토콜은 서버로부터 특정 고장을 은폐하기 위해 네트워크 스위치 상에 구현될 수 있고 또한 민첩한 디렉토리 서비스 상의 로드를 감소시키는 것을 돕도록 지레 작용될 수 있다. 이들 프로토콜은 토폴로지를 유지하도록 이용될 수 있고, 민첩한 네트워크를 위해 라우팅하는데, 이는 민첩한 디렉토리 서비스와 네트워크 제어 평면 사이의 결합을 감소시킬 수 있다. 스위치 상에 애니캐스트 어드레스를 규정하는 라우팅 디자인을 통해, 설명된 민첩한 아키텍처는 서버로부터 스위치의 고장을 은폐하기 위해 등가 다중 경로(ECMP)를 지레 작용할 수 있다. 다중 경로의 사용을 지원하는 다른 라우팅 프로토콜이 또한 적합하다.

[0049] 가상 레이어 2개의 네트워킹 예에 관한 구현 상세

[0050] 스케일 아웃 토폴로지

[0051] 통상의 네트워크는 통상적으로 네트워크의 최고 레벨에서 소수의 스위치 내로 트래픽을 집중시킨다. 이는 모두 이들 디바이스의 용량에 양분 대역폭을 제한하고 이들이 고장날 때 네트워크에 상당히 충돌할 수 있다. 그러나, 이들 문제점을 회피하기 위해, 트래픽 휘발성을 대처하기 위한 랜덤화를 사용하는 원리에 의해 구동된 민첩한 네트워크 토폴로지가 이용될 수 있다. 이와 관련하여, 네트워크 디바이스를 스케일 아웃하는 접근법이 취해질 수 있다. 이는 도 3에 도시된 바와 같이 고속 포워딩에 전용될 수 있는 낮은 복잡성 스위치의 비교적 넓은 네트워크를 생성할 수 있다. 이는 중간 스위치(310(1) 내지 310(N))와 집선 스위치(312(1) 내지 312(N)) 사이의 링크가 완전한 이분형 그래프를 형성할 수 있는 폴디드 클로스 네트워크의 예이다. 통상의 토폴로지에서와 같이, TOR은 2개의 집선 스위치에 접속할 수 있다. 그러나, 임의의 2개의 집선 스위치 사이의 다수의 경로는 n개의 중간 스위치가 존재하는 경우 이들 중 임의의 것의 고장이 대역폭의 적절한 열화라 칭할 수 있는 단지 1/n 바람직한 특성에 의해 양분 대역폭을 감소시키는 것을 의미한다. 또한, 클로스 네트워크와 같은 네트워크는 과잉가입이 없도록 설계될 수 있다. 예를 들어, 도 3에서, D 인터페이스 포트의 카운트를 갖는 집선 및 중간 스위치가 사용될 수 있다. 이들 스위치는 스위치의 각각의 레이어 사이의 용량이 링크 용량의 D*D/2배가

되도록 접속될 수 있다.

[0052] 클로스 네트워크(Clos network)와 같은 네트워크는 상부 티어 또는 네트워크의 "중추"에서 중간 스위치를 통해 바운스함으로써, 네트워크가 서버 라인 카드에서 진입-진출 바운드를 받게 되는 잠재적으로 모든 가능한 트래픽 매트릭스를 위한 대역폭 보장을 제공할 수 있는 점에서 로드 밸런싱(예를 들어, VLB)에 예외적으로 양호하게 적합될 수 있다. 라우팅은 간단하고 탄력적일 수 있다(예를 들어, 랜덤 경로는 랜덤 중간 노드를 올릴 수 있고 랜덤 경로는 낮출 수 있음).

[0053] 설명된 민첩한 아키텍처는 통상의 네트워크 아키텍처로 성취될 수 있는 더 큰 경로 제어를 제공할 수 있다. 더 구체적으로, 중간 노드는 분할될 수 있고, 트래픽 클래스는 몇몇 트래픽 클래스로 더 높은 전체 대역폭을 할당하기 위해 상이한 분할에 전용된다. 혼잡 지시는 미국 전기 전자 학회(IEEE) 802,1 Qau 혼잡 제어에서와 같이 명시적 혼잡 통지(ECN) 또는 유사한 메커니즘을 통해 송신기로 재차 신호화될 수 있다. 이와 같이, ECN 신호를 누적하는 송신기는 네트워크를 통한 대안 경로를 선택하는데 사용되는 소스 패킷 내의 필드를 변경함으로써 응답할 수 있다(경로 위 재선택이라 칭함).

[0054] 민첩한 라우팅

[0055] 로케이터로부터 명칭을 분리하는 원리를 구현하기 위해, 민첩한 네트워크는 2개의 IP 어드레스군을 사용할 수 있다. 도 3은 이러한 분리를 도시한다. 네트워크 인프라구조는 LA의 견지에서 동작할 수 있다. 스위치 및 인터페이스(310(1) 내지 310(N), 312(1) 내지 312(N), 314(1) 내지 314(N))가 LA에 할당될 수 있다. 스위치는 이들 LA를 전송하는 링크 상태 IP 라우팅 프로토콜을 실행할 수 있다.

[0056] 서버(316(1) 내지 316(N)) 상에서 실행되는 것과 같은 애플리케이션은 LA를 인식하지 못하지만 AA를 인식할 수 있다. 이 분리는 다수의 이득과 관련될 수 있다. 먼저, 패킷은 AA에 직접 송신되기보다는 적절한 LA에 터널링될 수 있다(스위치는 이들을 전달하기 위해 호스트마다 라우팅 엔트리를 유지할 필요가 없음). 이는 AA를 LA로 변환하는 민첩한 디렉토리 서비스가 어느 서비스가 통신되도록 허용되어야 하는지에 관한 정책을 구현할 수 있다는 것을 의미한다. 둘째로, 저가의 스위치는 종종 모든 LA 라우트를 유지할 수 있는 작은 라우팅 테이블(예를 들어, 12K 엔트리)을 갖지만, AA의 수에 의해 과장될 수 있다. 이 개념은 이것이 스위치가 유지할 수 있는 라우팅 엔트리의 수보다 크게 네트워크가 구성될 수 있게 하는 점에서 특히 가치가 있을 수 있다. 셋째로, 분리는 임의의 AA가 토폴로지에 관련되지 않고 임의의 서버에 할당될 수 있기 때문에 민첩성을 가능하게 한다. 넷째로, AA로부터 개별적으로 LA를 할당하기 위한 자유도는 LA가 이들이 위상적으로 중요한 방식으로 요약될 수 있는 이러한 방식으로 할당될 수 있어, 데이터 센터 또는 데이터 센터의 조작자 내부에서 실행되는 서비스에 의해 어떠한 방식이 요구되던간에 애플리케이션 어드레스를 할당하는 능력을 방해하지 않으면서 스위치가 전송해야 하는 라우팅 상태의 양을 추가로 제한한다.

[0057] 본 발명의 대안적인 실시예는 LA 및 AA 어드레스를 위한 다른 유형의 데이터를 사용할 수 있다. 예를 들어, LA 어드레스는 IPv4일 수 있고, AA 어드레스는 IPv6일 수 있고, 또는 그 반대도 마찬가지이고, 또는 IPv6 어드레스는 AA 및 LA 어드레스의 모두를 위해 사용될 수 있고, 또는 IEEE 802.1 MAC 어드레스가 AA 어드레스로서 사용될 수 있고, IP 어드레스(v4 또는 v6)가 LA 어드레스에 대해 사용될 수 있고, 또는 그 반대도 마찬가지이다. 어드레스는 또한 IP 어드레스를 갖는 VLAN 태그 또는 VRF 식별자와 같은 상이한 유형의 어드레스를 함께 조합함으로써 생성될 수 있다.

[0058] 이하의 설명은 어떻게 토폴로지, 라우팅 디자인, 애자일 에이전트 및 민첩한 디렉토리 서비스가 조합되어 기초 네트워크 패브릭을 가상화하고 이들이 레이어-2 LAN 내의 이들의 그룹의 다른 서버(316(1) 내지 316(N))에 접속되는 민첩한 네트워크의 서버(316(1) 내지 316(N))로의 환영을 생성하고, 호스트는 비교적 큰 데이터 센터폭 레이어-2 LAN의 부분이다.

[0059] 어드레스 결정 및 패킷 포워딩

[0060] 적어도 몇몇 구현예에서, 큰 이더넷을 성가시게 할 수 있는 브로드캐스트 ARP 스케일링 보틀넥을 배제하면서 동일한 서비스 내의 다른 서버와 단일의 큰 VLAN을 공유하는 것을 서버(316(1) 내지 316(N))가 고려하는 것을 가능하게 하기 위해, 이하에 언급되는 해결책이 제공된다. 예비적으로, 이하의 해결책은 하위 호환성이 있을 수 있고 현존하는 데이터 센터 애플리케이션에 투명할 수 있다.

[0061] 패킷 포워딩

[0062] AA는 통상적으로 네트워크의 라우팅 프로토콜 내로 고지될 수 없다. 따라서, 서버가 패킷을 수신하게 하기 위해, 패킷의 소스는 먼저 패킷을 캡슐화하여, 호스트를 위한 LA에 외부 헤더의 목적지를 설정할 수 있다. LA 어드레스를 유지하는 디바이스에 도달할 때, 패킷은 역캡슐화되어 목적지 서버에 전달된다. 일 실시예에서, 목적지 서버를 위한 LA는 목적지 서버가 위치되는 TOR에 할당된다. 일단 패킷이 그 목적지 TOR에 도달하면, TOR 스위치는 패킷을 역캡슐화할 수 있고, 일반적인 레이어-2 전달 규칙에 따라 내부 헤더 내의 목적지 AA에 기초하여 이를 전달할 수 있다. 대안적으로, LA는 서버 상에서 실행하는 물리적 목적지 서버 또는 가상 머신과 관련될 수 있다.

[0063] 어드레스 결정

[0064] 서버는 AA 어드레스가 이들과 동일한 LAN 내에 있는 것을 고려하도록 구성될 수 있어, 따라서 애플리케이션이 처음으로 AA에 패킷을 전송할 때, 호스트 상의 커널 네트워크 스택이 목적지 AA를 위한 브로드캐스트 ARP 요구를 생성할 수 있다. 소스 서버의 네트워킹 스택 내에서 실행되는 애자일 에이전트는 ARP 요구를 인터셉트하고 이를 민첩한 디렉토리 서비스로의 유니캐스트 질의로 변환할 수 있다. 민첩한 디렉토리 서비스가 질의에 대답할 때, 이는 패킷이 터널링되어야 하는 LA를 제공할 수 있다. 이는 또한 패킷을 바운스하는데 사용될 수 있는 중간 스위치 또는 중간 스위치의 세트를 제공할 수 있다.

[0065] 디렉토리 서비스에 의한 서비스간 액세스 제어

[0066] 서버는 이들이 AA에 대해 패킷을 터널링해야 하는 TOR의 LA를 얻을 수 없으면 AA에 패킷을 송신하는 것이 불가능할 수 있다. 따라서, 민첩한 디렉토리 서비스(322(1) 내지 322(N))는 통신 정책을 실행할 수 있다. 룩업 요구를 취급할 때, 민첩한 디렉토리 서비스는 어느 서버가 요구를 행하는지, 소스 및 목적지의 모두가 속하는 서비스 및 이들 서비스 사이의 격리 정책을 인지한다. 정책이 "거절"되면, 민첩한 디렉토리 서비스는 LA를 제공하는 것을 간단히 거부할 수 있다. 설명된 민첩한 네트워크 아키텍처의 장점은 서비스간 통신이 허용될 때, 패킷은 IP 게이트웨이에 우회되지 않고 송신 서버로부터 수신 서버로 직접 흐를 수 있다. 이는 통상의 아키텍처의 2개의 VLAN의 접속과는 다르다.

[0067] 인터넷과의 상호 작용

[0068] 종종, 데이터 센터에 의해 취급되는 트래픽의 대략 20%는 인터넷으로 또는 인터넷으로부터일 수 있다. 따라서, 데이터 센터 네트워크가 이들 큰 용적을 취급하는 것이 가능한 것이 유리하다. 설명된 민첩한 네트워크 아키텍처는 가상 레이어-2 네트워크를 구현하기 위해 레이어-3 패브릭을 이용하는 것이 처음에는 이상해보일 수 있지만, 이러한 것의 하나의 장점은 몇몇 통상의 제안된 네트워크 환경에서 요구되는 바와 같이 재기록되는 이들의 헤더를 갖도록 게이트웨이 서버를 통해 강요되지 않고 이 아키텍처를 갖는 민첩한 데이터 센터 네트워크를 구성할 수 있는 스위치의 고속 실리콘을 직접 가로질러 외부 트래픽이 흐를 수 있다는 것이다.

[0069] 인터넷으로부터 직접적으로 도달 가능해야 할 필요가 있는 서버(예를 들어, 프론트엔드 웹 서버)는 2개의 어드레스, 즉 LA 및 AA를 할당될 수 있다. LA는 네트워크간 통신을 위해 사용될 수 있다. AA는 백엔드 서버와의 인트라 데이터 센터 통신을 위해 사용될 수 있다. LA는 경계 게이트웨이 프로토콜(BGP)을 경유하여 고지되고 외부에서 도달 가능한 풀로부터 유도될 수 있다. 인터넷으로부터의 트래픽은 이어서 서버에 직접 도달할 수 있다. 서버로부터 외부 목적지로의 패킷은 ECMP에 의해 이용 가능한 링크 및 코어 라우터를 가로질러 확산되면서 코어 라우터를 향해 라우팅될 수 있다.

[0070] 브로드캐스트 취급

[0071] 설명된 민첩한 네트워크 아키텍처는 하위 호환성을 위한 애플리케이션에 레이어-2 시맨틱을 제공할 수 있다. 이는 브로드캐스트 및 멀티캐스트를 지원하는 것을 포함할 수 있다. 민첩한 네트워크 아키텍처의 접근법은 ARP

및 동적 호스트 구성 프로토콜(DHCP)과 같은 브로드캐스트의 가장 공통적인 소스를 완전히 배제하는 것이다. ARP는 애자일 에이전트(320) 내의 ARP 패킷을 인터셉트하고 전술된 바와 같이 민첩한 디렉토리 서비스로부터 정보를 참고한 후에 응답을 제공함으로써 취급될 수 있고, DHCP 패킷은 DHCP 서버에 포워딩된 유니캐스트 및 통상의 DHCP 릴레이 에이전트를 사용하여 TOR에서 인터셉트될 수 있다. 다른 브로드캐스트 패킷을 취급하기 위해, 이 세트 내의 다른 호스트에 의해 송신된 브로드캐스트 패킷을 수신하는 것이 가능해야 하는 호스트의 각각의 세트가 IP 멀티캐스트 어드레스를 할당될 있다. 이 어드레스는 디렉토리 시스템에 의해 할당될 수 있고, 민첩성 에이전트는 디렉토리 시스템에 질의함으로써 이를 학습할 수 있다.

[0072] 브로드캐스트 어드레스에 송신된 패킷은 대신에 서비스의 멀티캐스트 어드레스로 진행하도록 수정될 수 있다. 민첩한 네트워크 아키텍처의 민접한 에이전트는 스톰(storm)을 방지하기 위해 브로드캐스트 트래픽을 전송율 제한할 수 있다. 애자일 에이전트는 서버가 최근의 시간 간격(예를 들어, 과거 1초 및 과거 60초)에 걸쳐 송신되는 브로드캐스트 패킷의 전송율의 추정을 유지하고, 각각의 간격 중에 서버가 구성된 수의 브로드캐스트 패킷보다 많이 송신하는 것을 방지할 수 있다. 허용되는 것을 초과하여 송신된 패킷은 다음의 간격까지 드롭되거나 지연될 수 있다. 네이티브 IP 멀티캐스트가 또한 지원될 수 있다.

[0073] 스위치가 레이어-3 라우터로서 동작하는 실시예의 잠재적인 장점은 멀티캐스트 그룹에 속하는 모든 호스트 또는 머신으로의 멀티캐스트 그룹에 어드레스된 패킷의 전달을 구현하는 것이 특히 용이하다는 것이다. PIM-BIDIR과 같은 임의의 현존하는 IP 멀티캐스트 라우팅 프로토콜이 스위치 상에 구성될 수 있다. 이는 이들이 멀티캐스트 그룹에 속하는 각각의 호스트 또는 머신에서 종단점을 갖는 멀티캐스트 분배 트리를 계산할 수 있게 할 것이다. 호스트, 머신 또는 서버 상의 애자일 에이전트는 통상적으로 그 디폴트 게이트웨이에 IGMP 연결 메시지를 송신함으로써 적절한 멀티캐스트 그룹의 부분인 것으로서 호스트, 머신 또는 서버를 등록한다. 멀티캐스트 라우팅 프로토콜은 이어서 이 멀티캐스트 그룹을 위한 분배 트리에 호스트, 머신 또는 서버를 추가하는 것을 주의할 것이다. 레이어-2에서 동작하는 스위치는, 에이전트의 호스트, 머신 도는 서버가 수신되지 않아야 하는 패킷을 필터링하는 각각의 호스트, 머신 또는 서버 상의 민첩성 에이전트를 갖는, 멀티캐스트 그룹마다의 VLAN 또는 네트워크를 통한 플러드 필링(flood filling) 패킷과 같은 다양한 메커니즘을 사용할 수 있다.

[0074] 다중 경로 라우팅에 의한 랜덤화

[0075] 설명된 민첩한 네트워크 아키텍처는 적어도 몇몇 실시예에서, 2개의 관련 메커니즘, 즉 VLB 및 등가 다중 경로(ECMP)를 사용하여 휘발성에 대처하기 위해 랜덤화를 사용하는 원리를 지레 작용/이용할 수 있다. 이들 양자의 목표는 유사한데, VLB는 중간 노드를 가로질러 랜덤하게 트래픽을 분배하고, ECMP는 지속적인 혼잡을 감소시키거나 방지하기 위해 등가 경로를 가로질러 트래픽을 송신한다. 이하에 더 상세히 설명되는 바와 같이, VLB 및 ECMP는 각각이 다른 것의 제한을 극복하는데 사용될 수 있는 점에서 상보적일 수 있다. 양 메커니즘은 패킷의 송신기가 네트워크를 가로질러 경로의 선택에 영향을 미치는데 사용될 수 있는 제어를 제공할 수 있다. 애자일 에이전트는 이들 제어가 혼잡을 회피하도록 지레 작용할 수 있게 한다.

[0076] 도 5는 도 3에 소개되어 있는 민첩한 네트워크 아키텍처(300)의 서브세트를 도시한다. 도 5는 서버간 통신의 추가의 상세를 제공한다. 이 예는 서버(316(5))와 통신하는 서버(316(1))를 포함한다. 송신 서버(316(1)) 및 목적지 서버(316(5))는 VLAN으로서 기능하는 서버 풀(328) 내에서 기능할 수 있고, 10.128/9의 애플리케이션 어드레스를 가질 수 있다. 중간 스위치(310(1) 내지 310(N))는 링크 상태 네트워크(326) 내에 위치한다.

[0077] 민첩한 네트워크 아키텍처(300)는 VLB의 이득이 랜덤하게 선택된 중간 노드로부터 바운스하기 위해 패킷을 강요함으로써 성취되게 할 수 있다. 이 경우에, 송신기의 애자일 에이전트(320(1))는 중간 스위치(310(1) 내지 310(N))에 각각의 패킷을 캡슐화함으로써 이를 구현할 수 있다. 중간 스위치는 목적지의 TOR(이 경우에, 314(N))에 패킷을 터널링한다. 따라서, 패킷은 먼저 310(2)와 같은 중간 스위치 중 하나에 전달되고, 스위치에 의해 역캡슐화되고, TOR(314(N))의 LA에 전달되고, 재차 역캡슐화되고, 마지막으로 목적지 서버(316(5))로 송신될 수 있다.

[0078] 애자일 에이전트(320(1))가 능동 중간 스위치(310(1) 내지 310(N))의 어드레스를 인지하면, 이는 패킷을 송신할 때 이들 사이에서 랜덤하게 선택할 수 있다. 그러나, 이는 중간 스위치가 고장날 때 잠재적으로 수십만개의 애자일 에이전트를 업데이트하는 것을 필요로 할 수 있다. 대신에, 동일한 LA 어드레스가 다수의 중간 스위치(이 경우, LA 어드레스 10.0.0.5)에 할당될 수 있다. 민첩한 디렉토리 서비스(도 3에 도시됨)는 하나 이상의 룩업 결과의 부분으로서 애자일 에이전트(320(1))에 이 애니캐스트 어드레스를 복구시킬 수 있다. ECMP는 능동 중간

스위치(310(1) 내지 310(N)) 중 하나에 애니캐스트 어드레스에 캡슐화된 패킷을 전달하는 것을 주의할 수 있다. 스위치가 고장나면, ECMP는 반응할 수 있어 애자일 에이전트를 통지할 필요성을 배제한다.

[0079] 그러나, ECMP는 스케일링 제한을 가질 수 있다. 현재 통상의 스위치는 16-웨이 ECMP를 지원할 수 있고, 256-웨이 ECMP 스위치 ECMP 스위치가 또한 이용 가능할 수 있고 또는 곧 이용 가능할 수 있다. ECMP가 사용할 수 있는 것보다 많은 경로가 이용 가능하게 되면, VLB 캡슐화는 보상될 수 있다. 일 해결책은 다수의 애니캐스트 어드레스를 규정하는 것이고, 개별 애니캐스트 어드레스는 ECMP가 수용할 수 있는만큼과 동수의 중간 스위치(310(1) 내지 310(N))와 관련된다. 송신기는 로드를 분배하기 위해 애니캐스트 어드레스를 가로질러 해시할 수 있고, 스위치가 고장날 때 애니캐스트 어드레스는 디렉토리 시스템에 의해 다른 스위치에 재할당될 수 있어 개별 서버가 통지될 필요가 없게 된다. 설명의 목적으로, 이 양태는 디렉토리 시스템에 의해 제공된 네트워크 제어 기능성으로서 고려될 수 있다.

[0080] 설명된 VLB 기반 불확정 라우팅은 폴디드-클로스 네트워크 토폴로지 상에 순수 OSPF/ECMP 메커니즘을 사용하여 구현될 수 있다. 이러한 구성은 중간 스위치에서 역캡슐화 지원을 필요로 하지 않는다. 예를 들어, N이 각각의 TOR 상의 상향링크의 수이면, 집선 스위치는 세트로 그룹화될 수 있다. 몇몇 경우에, 이들 세트의 각각은 정확히 N개의 스위치를 포함할 수 있다. 각각의 TOR은 세트 내의 모든 N개의 스위치로의 상향링크를 가질 수 있거나 또는 세트 내의 스위치 중 어느 것으로도의 상향링크를 갖지 않을 수 있다. TOR의 이 배선에 의해, 서버 진입/진출 제약을 받게 되는 중재 트래픽을 위한 대역폭 보장은 OSPF 및/또는 ECMP와 같은 프로토콜이 TOR 사이의 라우팅을 위해 사용될 때에도 계속 유지되는 것을 나타낼 수 있다.

[0081] TOR 사이의 라우팅을 위한 OSPF 또는 ECMP의 사용은 집선 스위치의 동일한 세트 내의 2개의 TOR 사이의 패킷과 같은 몇몇 패킷(들)이 중간 스위치를 통해 진행하지 않는 경로를 취할 수 있게 한다. 따라서, 이들 경로는 이들이 소스와 목적지 사이의 최단 경로를 따르고 동일한 집선 스위치 또는 스위치들에 접속된 동일한 TOR 하에서 또는 TOR들 하에서 서버들 사이의 트래픽의 조기 선회를 허용하기 때문에 "조기 선회 경로"라 명명될 수 있다. 이들 트래픽 흐름은 코어 집선/중간 네트워크에 진입할 필요가 없다.

[0082] 조기-선회 경로의 잠재적인 이득은 다른 클래스의 트래픽(예를 들어, 외부)을 위한 코어 내의 용량을 확보하는 것을 포함할 수 있다. 확보된 용량은 예를 들어 현존하는 애플리케이션이 교차-TOR 트래픽을 최소화하기 위해 기록되어 있을 때 "평균" 경우에 대해 상당할 수 있다. 다른 방식으로 볼 때, 이는 코어가 몇몇 팩터에 의해 프로비저닝 하에 있게 할 수 있고, 여전히 서버간 트래픽에 대해 마찬가지로 동작할 수 있다. 조기-선회 경로의 사용은 또한 더 넓은 범위의 디바이스가 중간 스위치로서 사용될 수 있게 하여, 이들 스위치에 대한 낮은 비용을 초래한다.

[0083] 혼잡의 대처

[0084] ECMP 및 VLB의 모두에 의해, 큰 흐름이 동일한 링크 및 중간 스위치에 각각 해시될 수 있고, 이는 혼잡을 발생시킬 수 있다. 이러한 것이 발생하면, 송신 애자일 에이전트는 ECMP가 다음-홉, 즉 패킷이 통과되어야 하는 다음 스위치를 선택하기 위해 사용하는 필드의 값을 변경함으로써 민첩한 네트워크를 통해 그 흐름이 취하는 경로를 변경할 수 있다. 이와 관련하여, 애자일 에이전트는 큰 흐름을 주기적으로 재해싱하는 것 또는 심각한 혼잡 이벤트(예를 들어, 완전한 윈도우 손실) 또는 명시적 혼잡 통지가 TCP에 의해 검출될 때 또는 임계수의 바이트/패킷을 송신/수신한 후에와 같은 간단한 메커니즘을 갖는 이러한 상황을 검출하고 취급할 수 있다.

[0085] 호스트 정보 유지

[0086] 설명된 민첩성 네트워크 아키텍처에 따라 구현된 네트워크 시스템은 데이터 센터 작업부하를 위해 설계된 스케일 가능한, 신뢰적인 및/또는 높은 성능의 저장 또는 디렉토리 시스템을 사용할 수 있다. 민첩성 네트워크 아키텍처에 따라 구현된 네트워크는 이들 4개의 특성, 특 균일한 높은 용량, 성능 격리, L-2 시맨틱 및 서비스 사이의 통신 격리 중 하나 이상을 가질 수 있다. 네트워크는 또한 네트워크가 용량이 고장 후에 유지되는 모든 것을 계속 사용하는 경우에 적절한 열화를 나타낼 수 있다. 이와 같이, 네트워크는 고장에도 불구하고 신뢰성/탄성일 수 있다. 이와 관련하여, 이러한 네트워크의 디렉토리 시스템은 2개의 잠재적으로 주요 기능, (1) AA-대-LA 맵핑을 위한 룩업 및 업데이트 및 (2) 예를 들어 라이브 가상 머신 마이그레이션과 같은 지연 시간 민감 동작을 지원할 수 있는 반응성 캐시 업데이트 메커니즘을 제공할 수 있다.

[0087] 특징화 요구

[0088] 디렉토리 시스템을 위한 룩업 작업부하는 빈번하고 버스트성(bursty)일 수 있다. 서버는 각각의 흐름이 AA-대-LA 맵핑을 위한 룩업을 생성하는 상태로 짧은 시간 기간 이내에 최대 수천 또는 수만의 다른 서버와 통신할 수 있다. 업데이트를 위해, 작업부하는 고장 및 서버 시동 이벤트에 의해 구동될 수 있다. 다수의 고장이 통상적으로 크기가 작고, 큰 상관된 고장은 가능하게는 드물다.

[0089] 성능 요구

[0090] 작업부하의 버스트성 성질은 룩업이 다수의 접속을 신속하게 설정하기 위해 높은 처리량 및 낮은 응답 시간을 필요로 할 수 있다. 룩업은 제 1 시간에 서버와 통신하도록 요구되는 시간을 증가시키고, 응답 시간은 가능한 한 작게 유지되어야 하는데, 수십 밀리초가 적당한 값이다. 그러나, 업데이트를 위해, 잠재적으로 주요 요건은 신뢰성이고, 응답 시간은 덜 중요할 수 있다. 또한, 업데이트는 통상적으로 시간에 앞서 스케쥴링되기 때문에, 높은 처리량이 업데이트를 묶음화(batch)함으로써 성취될 수 있다.

[0091] 일치성 고려

[0092] 통상의 레이어-2 네트워크에서, ARP는 ARP 타임아웃에 기인하여 최종 일치성을 제공할 수 있다. 게다가, 호스트는 근거 없는(gratuitous) ARP를 발행함으로써 그 도착을 고지할 수 있다. 극단적인 예로서, 설명된 민첩성 네트워크 아키텍처에 따라 구현된 네트워크 내의 라이브 가상 머신(VM) 마이그레이션을 고려한다. VM 마이그레이션은 스테일 맵핑(AA-대-LA)의 고속 업데이트를 이용할 수 있다. VM 마이그레이션의 잠재적인 목표는 위치 변화를 가로질러 진행중인 통신을 보존하기 위한 것일 수 있다. 이들 고려는 AA-대-LA 맵핑의 약한 또는 최종 일치성이 신뢰적인 업데이트 메커니즘이 제공될 수 있는 한 허용될 수 있다는 것을 암시한다.

[0093] 민첩한 디렉토리 시스템 또는 서비스 디자인

[0094] 룩업의 성능 파라미터 및 작업부하 패턴은 업데이트의 것들과 상당히 상이할 수 있다. 이와 같이, 도 6에 도시된 2단 민첩한 디렉토리 서비스 아키텍처(600)를 고려한다. 이 경우에, 민첩한 디렉토리 서비스 아키텍처(600)는 애자일 에이전트(602(1) 내지 602(N)), 디렉토리 서비스 모듈(604(1) 내지 604(N)) 및 복제된 상태 머신(RSM) 서버(606(1) 내지 606(N))를 포함한다. 이 특정 경우에, 개별 디렉토리 서비스 모듈은 전용 컴퓨터(608(1) 내지 608(N)) 상에서 구현된다. 다른 구현예에서, 디렉토리 서비스 모듈은 다른 시스템 기능을 수행하는 컴퓨터 상에서 명시될 수 있다. 이 구현예에서, 디렉토리 서비스 모듈의 수는 전체 시스템 크기에 대해 일반적으로 적당하다. 예를 들어, 일 구현예는 100K 서버{즉, 도 3의 서버(316(1) 내지 316(N))}를 위한 대략 50 내지 100개의 디렉토리 서비스를 이용할 수 있다. 이 범위는 설명을 위해 제공된 것으로서 임계적인 것이 아니다.

[0095] 디렉토리 서비스 모듈(604(1) 내지 604(N))은 AA-대-LA 맵핑을 캐시할 수 있는 판독-최적화된 복제된 디렉토리 서버인 것으로 고려될 수 있다. 디렉토리 서비스 모듈(604(1) 내지 604(N))은 애자일 에이전트(602(1) 내지 602(N)) 및 Aa-대-LA 맵핑의 매우 일치성인 신뢰적인 저장을 제공할 수 있는 소수의 기록 최적화된 복제된 상태 머신(RSM) 서버(606(1) 내지 606(N))와 통신할 수 있다.

[0096] 디렉토리 서비스 모듈(604(1) 내지 604(N))은 낮은 지연 시간, 높은 처리량 및 높은 룩업 레이트를 위한 높은 이용 가능성을 보장할 수 있다. 한편, RSM 서버(606(1) 내지 606(N))는 적어도 몇몇 실시예에서 적당한 업데이트의 레이트를 위해 팍소스(Paxos) 일치 알고리즘 등을 사용하여 강한 일치성 및 내구성을 보장할 수 있다.

[0097] 개별 디렉토리 서비스 모듈(604(1) 내지 604(N))은 RSM 서버(606(1) 내지 606(N))에 저장된 AA-대-LA 맵핑을 캐시할 수 있고, 캐시된 상태를 사용하여 애자일 에이전트(602(1) 내지 602(N))로부터 룩업에 독립적으로 응답할 수 있다. 강한 일치성은 요건이 아닐 수 있기 때문에, 디렉토리 서비스 모듈은 규칙적인 기초로(예를 들어, 매 30초마다) RSM 서버와 그 로컬 맵핑을 느리게 동기화할 수 있다. 높은 이용 가능성 및 낮은 지연 시간을 동시에 성취하기 위해, 애자일 에이전트(602(1) 내지 602(N))는 랜덤하게 선택된 디렉토리 서비스 모듈(604(1) 내지 604(N))의 수(k)(예를 들어, 2개)에 룩업을 송신할 수 있다. 다수의 응답이 수신되면, 애자일 에이전트는

가장 빠른 응답을 간단히 선택하고 이를 그 캐시에 저장할 수 있다.

[0098] 디렉토리 서비스 모듈(604(1) 내지 604(N))은 또한 네트워크 프로비저닝 시스템으로부터 업데이트를 취급할 수 있다. 일치성 및 내구성을 위해, 업데이트가 단일의 랜덤하게 선택된 디렉토리 서비스 모듈에 송신될 수 있고, RSM 서버(606(1) 내지 606(N))에 기록될 수 있다. 구체적으로, 업데이트시에, 디렉토리 서비스 모듈은 RSM에 업데이트를 먼저 포워딩할 수 있다. RSM는 개별 RSM 서버에 업데이트를 신뢰적으로 복제할 수 있고, 이어서 디렉토리 서비스 모듈에 확인 응답을 갖고 응답하고, 이는 이어서 발신 클라이언트에 확인 응답을 재차 포워딩할 수 있다.

[0099] 일치성을 향상시키기 위한 잠재적인 최적화로서, 디렉토리 서비스 모듈(604(1) 내지 604(N))은 선택적으로 소수의 다른 디렉토리 서비스 모듈에 확인 응답된 업데이트를 유포할 수 있다. 발신 클라이언트가 타임아웃(예를 들어, 2초) 이내에 확인 응답을 수신하지 않으면, 클라이언트는 다른 디렉토리 서비스 모듈에 동일한 업데이트를 송신할 수 있어, 따라서 신뢰성 및/또는 이용 가능성을 위해 응답 시간을 희생한다.

[0100] 디렉토리 시스템의 다른 실시예가 또한 가능하다. 예를 들어, 분산형 해시 테이블(DHT)이 디렉토리 서버 및 DHT 내의 엔트리로서 저장된 AA/LA 맵핑을 사용하여 구성될 수 있다. 능동 디렉토리 또는 경량 디렉토리 시스템과 같은 다른 현존하는 디렉토리 시스템이 또한 사용될 수 있지만, 성능은 양호하지 않고 또는 일치는 전술된 실시예에서와 같이 강하지 않을 수 있다.

[0101] 최종 일치성의 보장

[0102] AA-대-LA 맵핑은 디렉토리 서비스 모듈에서 그리고 애자일 에이전트의 캐시에서 캐시될 수 있기 때문에, 업데이트가 불일치성을 유도할 수 있다. 서버 및 네트워크 리소스를 폐기하지 않고 불일치성을 해결하기 위해, 반응성 캐시-업데이트 메커니즘이 동시에 확장성 및 성능의 모두를 보장하기 위해 이용될 수 있다. 캐시-업데이트 프로토콜은 주요 관찰을 지레 작용할 수 있는데, 스테일 호스트 맵핑은 단지 이 맵핑이 트래픽을 전달하는데 사용될 때에만 보정될 필요가 있다. 구체적으로, 스테일 맵핑이 사용될 때, 몇몇 패킷은 스테일 LA - 더 이상 목적지 서버에 호스팅되지 않는 TOR 또는 서버에 도달할 수 있다. TOR 또는 서버는 이러한 전달 불가능한 패킷을 디렉토리 서비스 모듈에 포워딩할 수 있어, 예를 들어 유니캐스트를 경유하여 소스 서버의 캐시 내의 스테일 맵핑을 선택적으로 보정하기 위해 디렉토리 서비스 모듈을 트리거링할 수 있다. 업데이트의 다른 실시예에서, 디렉토리 서비스는 영향을 받은 서버와 통신하도록 허용되는 모든 서버 그룹에 업데이트를 멀티캐스트할 수 있다.

[0103] 다른 구현예

[0104] 로드 밸런싱의 최적성

[0105] 전술된 바와 같이, VLB와 같은 로드 밸런싱 기술은 휘발성에 대처하기 위해 랜덤화를 사용할 수 있다 - 잠재적으로, 트래픽 패턴(최선의 경우 및 최악의 경우의 모두를 포함함)을 평균 경우로 변환함으로써 최선의 경우 트래픽 패턴에 대한 몇몇 성능을 희생시킴. 이 성능 손실은 더 최적의 트래픽 엔지니어링 시스템 하에 있을 수 있는 것보다 높은 몇몇 링크의 이용으로서 자신을 명시할 수 있다. 그러나, 실제 데이터 센터 작업부하의 평가는 VLB와 같은 로드 밸런싱 기술의 간단성 및 범용성이 더 복잡한 트래픽 엔지니어링 체계에 비교할 때 비교적 적은 용량 손실과 관련될 수 있다는 것을 나타낸다.

[0106] 레이아웃 구성

[0107] 도 7 내지 도 9는 설명된 민첩한 네트워크 아키텍처에 따라 구현되는 데이터 센터 네트워크를 위한 3개의 가능한 레이아웃 구성을 도시한다. 도 7 내지 도 9에서, 도면 페이지 상의 공간 제약에 기인하여, TOR은 관련 서버 없이 도시되어 있다.

[0108] 도 7은 개방 플로어 플랜 데이터 센터 레이아웃(700)을 도시한다. 데이터 센터 레이아웃(700)은 TOR(702(1) 내지 702(N)), 집선 스위치(704(1) 내지 704(N)) 및 중간 스위치(706(1) 내지 706(N))를 포함한다. 도 7에서, TOR(702(1) 내지 702(N))은 중앙 "네트워크 케이지"(708)를 둘러싸는 것으로서 도시되어 있고, 접속될 수 있다(예를 들어, 구리 및/또는 파이버 케이블 등을 사용하여). 집선 및 중간 스위치(704(1) 내지 704(N), 706(1) 내지 706(N))는 각각 네트워크 케이지(708) 내부에 매우 근접하여 레이아웃될 수 있어, 이들의 사용 작용을 위

해 구리 케이블의 사용을 허용한다(구리 케이블은 파이버에 대해 저가이고, 더 두껍고, 낮은 도달 거리를 가질 수 있음). 네트워크 케이지 내부의 케이블의 수 뿐만 아니라 이들의 비용(예를 들어, 약 2의 팩터만큼)은, 예를 들어 쿼드 스몰 폼 플러깅 가능(QSFP) 표준과 같은 적절한 표준을 사용하여 단일 케이블로 다수의(예를 들어, 4개) 10 G 링크를 함께 번들화함으로써 감소될 수 있다(예를 들어, 4의 팩터만큼).

[0109] 개방 플로어 플랜 데이터 센터 레이아웃(700)에서, 중간 스위치(706(1) 내지 706(N))는 네트워크 케이지(708) 내에 중앙에 배열되고, 집선 스위치(704(1) 내지 704(N))는 중간 스위치(706(1) 내지 706(N))와 TOR 스위치(702(1) 내지 702(N))(및 관련된 서버) 사이에 개지된다.

[0110] 개방 플로어 플랜 데이터 센터 레이아웃(700)이 원하는 바에 따라 스케일링 가능할 수 있다. 예를 들어, 추가의 서버 랙이 서버 랙을 생성하기 위해 서버의 형태의 컴퓨팅 디바이스를 TOR(702(1) 내지 702(N))과 관련시킴으로써 추가될 수 있다. 서버 랙은 이어서 네트워크 케이지(708)의 집선 스위치(704(1) 내지 704(N))에 접속될 수 있다. 다른 서버 랙 및/또는 개별 서버가 개방 플로어 플랜 데이터 센터 레이아웃에 의해 제공된 서비스를 중단하지 않고 제거될 수 있다.

[0111] 도 8은 모듈화된 컨테이너 기반 레이아웃(800)을 도시한다. 레이아웃(800)은 TOR(802(1) 내지 802(N)), 집선 스위치(804(1) 내지 804(N)) 및 중간 스위치(806(1) 내지 806(N))를 포함한다. 이 경우에, 중간 스위치(806(1) 내지 806(N))는 레이아웃의 데이터 센터 인프라구조(808) 내에 포함된다. 집선 스위치 및 TOR 스위치는 데이터 센터 인프라구조에 접속된 플러깅 가능한 컨테이너로서 관련될 수 있다. 예를 들어, 집선 스위치(804(1), 804(2))는 데이터 센터 인프라구조(808)에 접속될 수 있는 플러깅 가능한 컨테이너(810(1)) 내의 TOR 스위치(802(1), 802(2))와 관련된다. 유사하게, 집선 스위치(804(3), 804(4))는 플러깅 가능한 컨테이너(810(2)) 내의 TOR 스위치(802(3), 802(4))와 관련되고, 집선 스위치(804(5), 804(N))는 플러깅 가능한 컨테이너(810(N)) 내의 TOR 스위치(802(5), 802(N))와 관련된다.

[0112] 도 7에서와 같이, 도 8에서, 서버 랙을 구상하기 위해 TOR과 관련될 수 있는 서버가 도면 페이지의 공간 제약에 기인하여 도시되어 있지 않다. 또한, 공간 제약에 기인하여, 단지 2개의 집선 스위치 및 2개의 TOR 스위치만이 플러깅 가능한 컨테이너마다 도시되어 있다. 물론, 다른 구현예가 더 많거나 적은 이들 구성 요소 중 하나 또는 양자를 이용할 수 있다. 또한, 다른 구현예는 여기에 도시된 3개보다 많거나 적은 플러깅 가능한 컨테이너를 이용할 수 있다. 일 관심 특징은 레이아웃(800)이 각각의 플러깅 가능한 컨테이너(810(1) 내지 810(N))로부터 데이터 센터 중추(즉, 데이터 센터 인프라구조(808))로 하나의 케이블 번들(812)을 유도하기에 적합할 수 있다. 요약하면, 데이터 센터 인프라구조(808)는 레이아웃(800)이 개별 플러깅 가능한 컨테이너(810(1) 내지 810(N))를 추가하거나 제거함으로써 크기를 팽창하거나 수축하게 할 수 있다.

[0113] 도 9는 "인프라구조가 없는" 및 "컨테이너 수납형(containerized)" 데이터 센터 레이아웃(900)을 도시한다. 레이아웃은 다수의 컨테이너(908(1) 내지 908(N)) 내에 배열되어 있는 TOR(902(1) 내지 902(N)), 집선 스위치(904(1) 내지 904(N)) 및 중간 스위치(906(1) 내지 906(N))를 포함한다. 예를 들어, TOR(902(1) 내지 902(2)), 집선 스위치(904(1) 내지 904(2)) 및 중간 스위치(906(1))는 컨테이너(908(1)) 내에 배열되어 있다.

[0114] 컨테이너(908(1) 내지 908(N))는 "인프라구조가 없는" 및 "컨테이너 수납형" 데이터 센터 레이아웃(900)의 실현을 허용할 수 있다. 이 레이아웃(900)은 개별 쌍의 컨테이너(908(1), 908(3)) 사이에 케이블 번들(910(1))을 연장하는 것과 관련될 수 있다. 다른 케이블 번들(910(2))이 개별 쌍의 컨테이너(908(2), 908(N)) 사이로 연장될 수 있다. 개별 케이블 번들(910(1), 910(2))이 컨테이너(908(1)) 내의 집선 스위치(904(1), 904(2))를 컨테이너(908(3)) 내의 중간 스위치(906(3))에 접속하고 그 반대도 마찬가지인 링크를 가질 수 있다.

[0115] 요약하면, 개별 컨테이너(908(1) 내지 908(N))는 복수의 스위치를 포함할 수 있다. 이들 스위치는 상보적인 플러깅 가능한 컨테이너 내에 배열되어 있는 TOR 스위치(902(1) 내지 902(N)), 집선 스위치(904(1) 내지 904(N)) 및 중간 스위치(906(1) 내지 906(N))을 포함할 수 있다. 상보적인 플러깅 가능한 컨테이너의 쌍은 제 2 플러깅 가능한 컨테이너의 중간 스위치에 제 1 플러깅 가능한 컨테이너의 집선 스위치를 접속함으로써 결합될 수 있고, 케이블 번들을 경유하여 그 반대도 마찬가지이다. 예를 들어, 컨테이너(908(1))는 케이블 번들(910(1))을 경유하여 컨테이너(908(3))에 접속될 수 있다. 구체적으로, 번들은 컨테이너(908(1))의 집선 스위치(904(1), 904(2))를 컨테이너(908(3))의 중간 스위치(906(3))에 접속할 수 있다. 유사하게, 번들(910(1))은 컨테이너(908(3))의 집선 스위치(904(5), 904(6))를 컨테이너(908(1))의 중간 스위치(906(1))에 접속할 수 있다.

[0116] 적어도 몇몇 실시예에서, 민첩한 네트워크 아키텍처는 이하의 구성 요소, 즉 (1) 토폴로지로 함께 접속된 스위치의 세트, (2) 스위치 중 하나 이상에 각각 접속된 서버의 세트, (3) 서버가 패킷(들)을 다른 서버에 송신하기

를 원할 때 요구가 이루어지고 이들이 스위치의 토폴로지를 횡단하는 것이 가능할 수 있도록 송신하기를 원하는 패킷을 어드레스하거나 캡슐화하는데 서버(또는 서버의 대표적인 애자일 에이전트)가 사용하는 정보에 응답하는 디렉토리 시스템, (4) 임의의 링크 상의 이용이 패킷이 이 링크 내에 송신되는 스위치(들)에 의해 드롭되기에 너무 높게 성장하는 것을 감소시키고/방지하는 네트워크 내의 혼잡을 제어하기 위한 메커니즘, 및 (5) 디렉토리 서비스와 통신하고, 즉 필요에 따라 패킷을 캡슐화하고, 어드레스하거나 역캡슐화하고 필요에 따라 혼잡 제어에 참여하는 서버 상의 모듈로 이루어질 수 있다.

[0117] 적어도 일 실시예에서, (1) 패킷을 목적지에 포워딩하는데 이용된 캡슐화 정보를 검색하기 위해 민첩한 디렉토리와 통신하고, 이 서버를 시스템에 정합시키는 것 등, (2) 요구되는 바와 같이 대안의 세트 중에서(예를 들어, 중간 스위치 중에서) 랜덤 선택을 행하고 이들 선택을 캐시하는 것, (3) 패킷을 캡슐화하고/역캡슐화하는 것 및 (4) 네트워크로부터 혼잡 지시를 검출하고 응답하는 것과 같은 기능을 제공하는 애자일 에이전트가 각각의 서버 상에 존재할 수 있다. 대안적으로, 적어도 몇몇 실시예에서, 이들 기능은 네트워크 내의 서버와 스위치 사이에 분배될 수 있다. 예를 들어, 디폴트 라우팅이 스위치의 세트(중간 스위치와 같은)에 패킷을 지향시키는데 사용될 수 있고, 상기 열거된 기능이 패킷이 횡단하는 중간 스위치 상의 각각의 패킷에 대해 구현된다.

[0118] 적어도 몇몇 실시예에서, 본 명세서에 설명된 민첩한 네트워크 아키텍처는 데이터 센터 내의 스위치의 세트 사이에 네트워크를 생성하여 네트워크 내의 각각의 스위치가 네트워크 내의 임의의 다른 스위치에 패킷을 송신하는 것이 가능하게 하는 것을 포함할 수 있다. 이들 스위치 또는 이 네트워크가 다른 서버와 통신하기 위해 서버에 의해 사용된 어드레스와 동일한 유형의 이들 자신들 사이에 패킷을 지향시키기 위한 어드레스를 사용할 필요는 없다. 예를 들어, MAC 어드레스, IPv4 어드레스 및/또는 IPv6 어드레스가 모두 적합할 수 있다.

[0119] 민첩한 네트워크의 적어도 하나의 실시예에서, 데이터 센터 내의 스위치의 세트 사이의 일 고려 사항은 IPv4 또는 IPv6와 같은 IP 어드레스를 갖고 이들의 각각을 구성하고, 하나 이상의 표준 레이어-3 라우팅 프로토콜을 실행하도록 이들을 구성하는 것이고, 통상의 예는 개방형 최단 경로 우선(OSPF), 중간 시스템-중간 시스템(IS-IS) 또는 경계 게이트웨이 프로토콜(BGP)이다. 이러한 실시예의 이득은 네트워크와 디렉토리 시스템 사이의 결합이 스위치들 사이에 패킷을 포워딩하는 네트워크의 능력을 유지하는 그 라우팅 프로토콜에 의해 생성된 네트워크의 제어 평면을 갖고 감소되어, 디렉토리 시스템이 토폴리지의 대부분의 변화의 서버에 반응하고 통지할 필요가 없게 된다는 것이다.

[0120] 대안적으로 또는 추가적으로, 디렉토리 시스템은 네트워크의 토폴로지를 모니터링할 수 있고(예를 들어, 스위치 및 링크의 건강 상태를 모니터링함), 토폴로지가 변경함에 따라 서버에 제공하는 캡슐화 정보를 변경한다. 디렉토리 시스템은 또한 이들 응답이 더 이상 유효하지 않다는 것을, 이전에 송신된 응답을 갖는 서버에 통지할 수 있다. 대안예에 비한 제 1 실시예의 잠재적인 이득은 네트워크와 디렉토리 시스템 사이의 결합이 스위치들 사이에 패킷을 포워딩하는 네트워크의 능력을 유지하는 그 라우팅 프로토콜에 의해 생성된 네트워크의 제어 평면을 갖고 감소되어, 디렉토리 시스템이 토폴리지의 대부분의 변화의 서버에 반응하고 통지할 필요가 없게 된다는 것이다. 요약하면, 패킷 전달 지연은 네트워크 성능에 관련된 하나 이상의 파라미터를 모니터링함으로써 감소되거나 회피될 수 있다. 파라미터는 특정 경로를 통한 통신 장애와 같은 네트워크 이벤트를 지시할 수 있다.

[0121] 일 실시예에서, 네트워크의 스위치는 LA 어드레스의 서브넷으로부터 유도된 IPv4 어드레스를 갖고 구성된다. 스위치는 OSPF 라우팅 프로토콜을 실행하도록 구성된다. 스위치의 어드레스는 OSPF 프로토콜에 의해 스위치들 사이에 분배된다. OSPF로의 넘버링되지 않은 인터페이스 확장부가 OSPF 프로토콜에 의해 분배된 정보의 양을 감소시키는데 사용될 수 있다. 각각의 탑 오브 랙(TOR) 스위치의 서버 지향 포트가 가상 근거리 통신망(VLAN)의 부분이 되도록 스위치 상에 구성된다. AA 공간을 포함하는 서브넷(들)은 서버 지향 VLAN에 할당되는 것으로서 스위치 상에 구성된다. 이 VLAN의 어드레스는 OSPF 내에 분배되지 않고, VLAN은 통상적으로 트렁크되지 않는다. 서버에 예정된 패킷은 서버가 접속되는 TOR에 캡슐화된다. 이 TOR은 이것이 패킷들을 수신함에 따라 패킷들을 역캡슐화할 수 있고, 이어서 이들을 서버의 목적지 어드레스에 기초하여 서버 지향 VLAN 상에 포워딩할 수 있다. 서버는 이어서 정상 LAN에서와 같이 패킷을 수신할 수 있다.

[0122] 다른 실시예에서, TOR 스위치의 서버 지향 VLAN 상에 AA 서브넷(들)을 구성하는 대신에, 각각의 TOR에 고유한 LA 서브넷이 서버 지향 VLAN에 할당된다. 이 LA 서브넷은 OSPF에 의해 분배된다. TOR에 접속된 서버는 적어도 2개의 어드레스를 갖고 구성된다. LA 서브넷으로부터 유도된 LA 어드레스는 그 부분인 서버 지향 VLAN 및 AA 어드레스에 할당된다. 서버에 예정된 패킷은 서버 상에 구성되어 있는 LA에 캡슐화된다. 서버 상의 모듈은 이것이 패킷들을 수신함에 따라 패킷들을 역캡슐화할 수 있고, 이들을 국부적으로 가상 머신에 전달하거나 이들이 패킷 내에 포함된 AA 어드레스에 기초하여 예정되는 서버 상에서 프로세싱할 수 있다.

[0123] 다른 실시예에서, TOR 스위치는 레이어-2 스위치로서 동작할 수 있고, 반면 집선 레이어 스위치는 레이어-3로서 동작할 수 있다. 이 디자인은 잠재적으로는 더 저가의 레이어-2 스위치가 TOR 스위치로서 사용될 수 있게 하고 (그리고, 다수의 TOR 스위치가 존재함), 레이어-3 기능성은 비교적 적은 수의 집선 레이어 스위치에 구현될 수 있다. 이 디자인에서, 역캡슐화 기능성이 레이어-2 스위치, 레이어-3 스위치, 목적지 서버 또는 목적지 가상 머신에서 수행될 수 있다.

[0124] 임의의 실시예에서, 추가의 어드레스는 스위치 상에 구성될 수 있거나 또는 OSPF와 같은 라우팅 프로토콜을 경유하여 분배될 수 있다. 이들 어드레스는 통상적으로 위상적으로 중요할 수 있다(즉, LA). 어드레스는 통상적으로 인프라구조 서비스-즉, 추가의 서비스로서 알려져 있는 것을 제공하는 서버, 스위치 또는 네트워크 서비스에 패킷을 지향시키는데 사용될 수 있다. 이러한 서비스의 예는 로드 밸런서(이들은 F5로부터의 BigIP와 같은 하드웨어 기반 또는 소프트웨어 기반 로드 밸런서일 수 있음), 소스 네트워크 어드레스 트랜스레이터(S-NAT), 디렉토리 시스템의 부분인 서버, DHCP 서비스를 제공하는 서버, 또는 다른 네트워크로의 게이트웨이(인터넷 또는 다른 데이터 센터와 같은)를 포함한다.

[0125] 일 실시예에서, 각각의 스위치는 BGP 프로토콜을 사용하여 라우트 리플렉터로서 구성될 수 있다. 추가의 어드레스는 라우트 리플렉터(들) 상에 이들을 구성하고 BGP가 이들을 스위치에 분배하게 함으로써 스위치에 분배된다. 이 실시예는 추가의 어드레스를 추가하거나 제거하는 것이 스위치의 라우팅 프로세서를 오버로드할 수 있는 OSPF 재연산을 발생하지 않는 이득을 갖는다.

[0126] 다른 실시예에서, 네트워크 내의 혼잡을 제어하기 위한 메커니즘이 서버 자체 상에 구현된다. 적합한 메커니즘은 서버에 의해 목적지로 송신된 트래픽이 네트워크가 전달하는 것이 가능한 것으로 나타나는 레이트로 서버에 의해 제한되는 전송 제어 프로토콜(TCP)과 같은 것이다. TCP와 같은 프로토콜의 사용의 개량이 다음에 설명될 것이다. 대안 실시예에서, 스위치 상의 서비스 품질 메커니즘이 혼잡 제어를 위해 사용될 수 있다. 이러한 메커니즘의 예는 가중된 공평 큐잉(WFQ) 및 그 파생물, 랜덤 조기 검출(RED), RSVP, 명시적 제어 프로토콜(XCP) 및 레이트 제어 프로토콜(RCP)을 포함한다.

[0127] 적어도 하나의 실시예에서, 서버 상의 모듈은 민첩한 네트워크로부터 수신되는 패킷을 관찰하고, 수신된 패킷으로부터 얻거나 암시되는 정보에 기초하여 패킷의 송신 또는 패킷의 캡슐화를 변경한다. 애자일 에이전트는 (1) 이들이 송신되는 레이트를 감소시키기 위해 패킷의 송신을 변경하고, 또는 (2) 패킷(들)의 캡슐화 및 어드레싱을 먼저 선택할 때 가능한 대안들 중에서 임의의 또는 모든 랜덤 선택을 주목함으로써 성취될 수 있는, 이들이 네트워크를 통한 상이한 경로를 취하도록 패킷의 캡슐화를 변경함으로써 혼잡을 감소시킬 수 있다.

[0128] 애자일 에이전트가 수행할 수 있는 관찰 및 그 반응의 예는 이하를 포함한다. (1) 애자일 에이전트가 TCP 패킷의 전체 윈도우의 손실을 검출하면, 애자일 에이전트가 네트워크를 통해 패킷이 취할 수 있는 경로를 재랜덤화한다. 이는 흐름 상에서 이전에 송신된 모든 패킷이 네트워크로부터 나오게 되는 것으로 고려되어 패킷에 의해 취해진 경로의 변경이 재순서화된 패킷이 목적지에 의해 수신되게 하지 않을 수 있게 하는 것과 동시에 상이한 (바람직하게는, 비혼잡된) 경로 상에 흐름을 배치하기 때문에 특히 유리하다. (2) 애자일 에이전트는 패킷에 의해 취해진 경로를 주기적으로 재랜덤화할 수 있다. (3) 애자일 에이전트는 흐름에 의해 성취되는 유효 레이트를 컴퓨팅하고, 레이트가 예측된 임계값 미만이면 재랜덤화할 수 있다. (4) 애자일 에이전트는 명시적 혼잡 통지(Explicit Congestion Notification: ECN) 마크에 대한 수신된 패킷을 감시하고 레이트를 감소시키거나 이 목적지로 임의의 패킷의 경로를 재랜덤화할 수 있다. (5) 스위치는 혼잡 상태로 진입되거나 막 진입되려고 하는 링크를 검출하기 위해 논리를 실행할 수 있고(예를 들어, IEEE QCN 및 802.1au), 상류측 스위치 및/또는 서버에 통지를 송신할 수 있다. 이들 지시를 수신하는 애자일 에이전트는 이들의 패킷의 레이트를 감소시키거나 패킷의 경로를 재랜덤화할 수 있다.

[0129] 설명된 실시예의 일 장점은, VM이 동일한 IP 어드레스의 사용을 유지하면서 하나의 서버로부터 다른 서버로 재위치 지정될 수 있기 때문에 가상 머신(VM)의 라이브 마이그레이션을 허용한다는 것이다. 디렉토리 시스템은 VM이 이동 중에 재위치 지정되는 서버에 VM의 IP 어드레스로 예정된 패킷을 지향시키도록 간단하게 업데이트될 수 있다. 위치의 물리적 변화는 진행중인 통신을 방해할 필요는 없다.

[0130] 적어도 하나의 실시예에서, 네트워크의 용량의 부분이 분할비의 불균일한 연산에 의해 네트워크 상에서 동작하는 서비스의 세트에 보류되거나 우선적으로 할당될 수 있어, 바람직한 서비스가 더 큰 또는 더 적은 수의 경로 또는 서비스의 다른 세트에 의해 사용된 경로로부터 분리된 경로의 세트 상에서 확산된 이들의 패킷을 갖게 된다. 다수의 클래스의 선호도 또는 QoS가 이 동일한 기술을 사용하여 생성될 수 있다.

[0131] 방법 예

[0132] 도 10은 본 발명의 개념의 적어도 몇몇 구현예에 따른 민첩한 네트워킹 기술 또는 방법(1000)의 흐름도를 도시한다. 방법(1000)이 설명되는 순서는 한정으로서 해석되도록 의도된 것은 아니고, 임의의 수의 설명된 블록은 방법 또는 대안적인 방법을 구현하기 위해 임의의 순서로 조합될 수 있다. 더욱이, 방법은 컴퓨팅 디바이스가 방법을 구현할 수 있도록 임의의 적합한 하드웨어, 소프트웨어, 펌웨어 또는 이들의 임의의 조합으로 구현될 수 있다. 일 경우에, 방법은 컴퓨팅 디바이스의 프로세서에 의한 실행이 컴퓨팅 디바이스가 방법을 수행할 수 있게 하도록 하는 명령의 세트로서 컴퓨터 판독 가능 저장 매체 상에 저장된다. 다른 경우에, 방법은 ASIC에 의한 실행을 위해 ASIC의 컴퓨터 판독 가능 저장 매체 상에 저장된다.

[0133] 블록 1002에서, 방법은 패킷을 목적지에 포워딩하는데 이용된 캡슐화 정보를 얻는다.

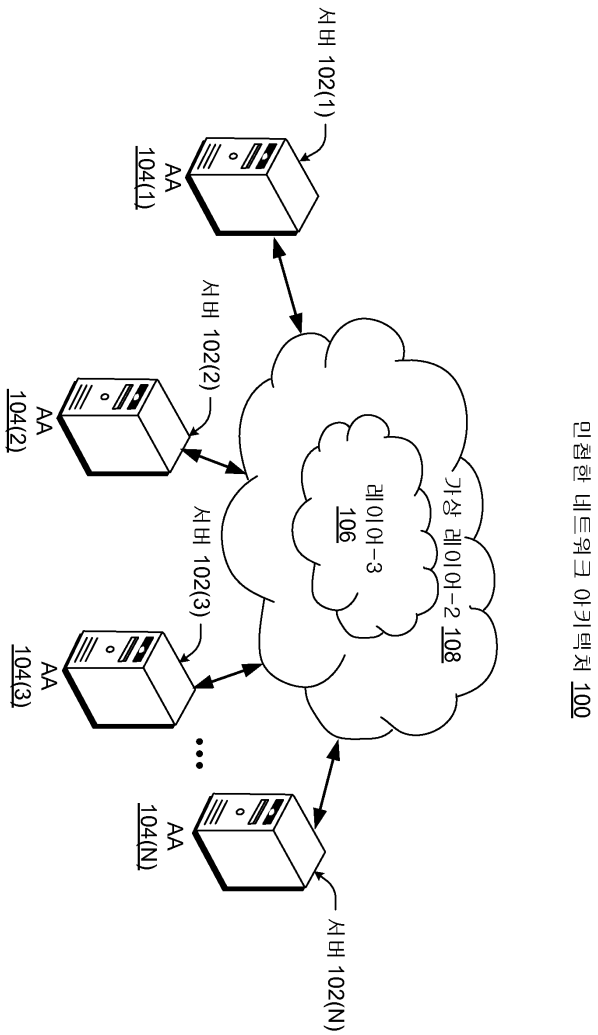
[0134] 블록 1004에서, 방법은 스위치와 같은 이용 가능한 하드웨어를 통해 경로를 선택한다.

[0135] 블록 1006에서, 방법은 경로 상에서 전달을 위해 패킷을 캡슐화한다.

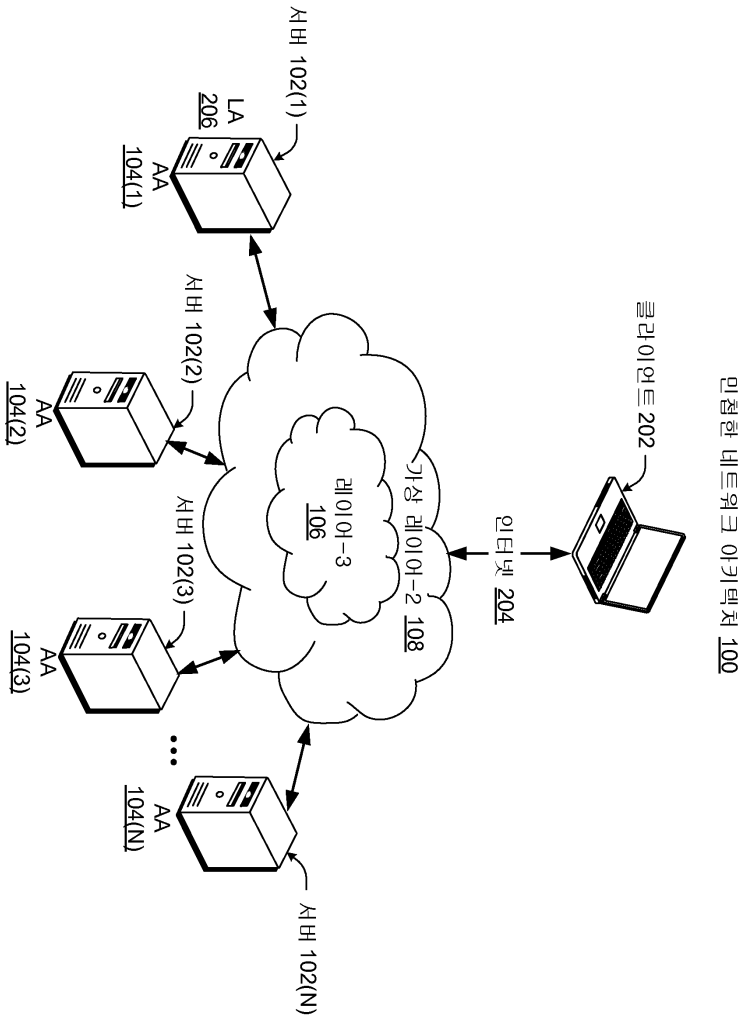
[0136] 블록 1008에서, 방법은 혼잡의 지시를 모니터링한다. 예를 들어, 방법은 네트워크 성능과 관련된 파라미터를 모니터링할 수 있다. 예를 들어, TCP는 혼잡과 관련하는 네트워크 파라미터로서 작용할 수 있는 네트워크 구성 요소 상의 패킷 전송율 및/또는 부하와 관련된 업데이트를 제공할 수 있다. 방법은 혼잡이 검출될 때 경로를 재선택하고 그리고/또는 다른 작용을 취할 수 있다.

도면

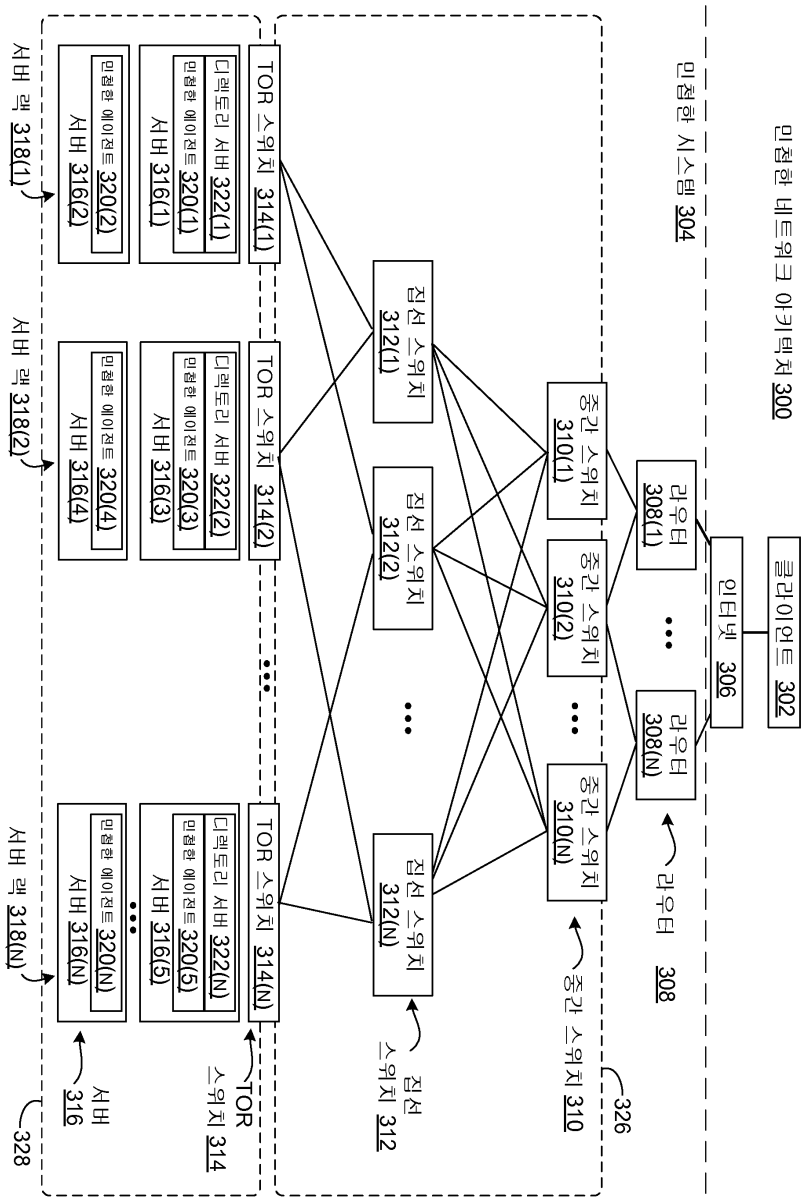
도면1



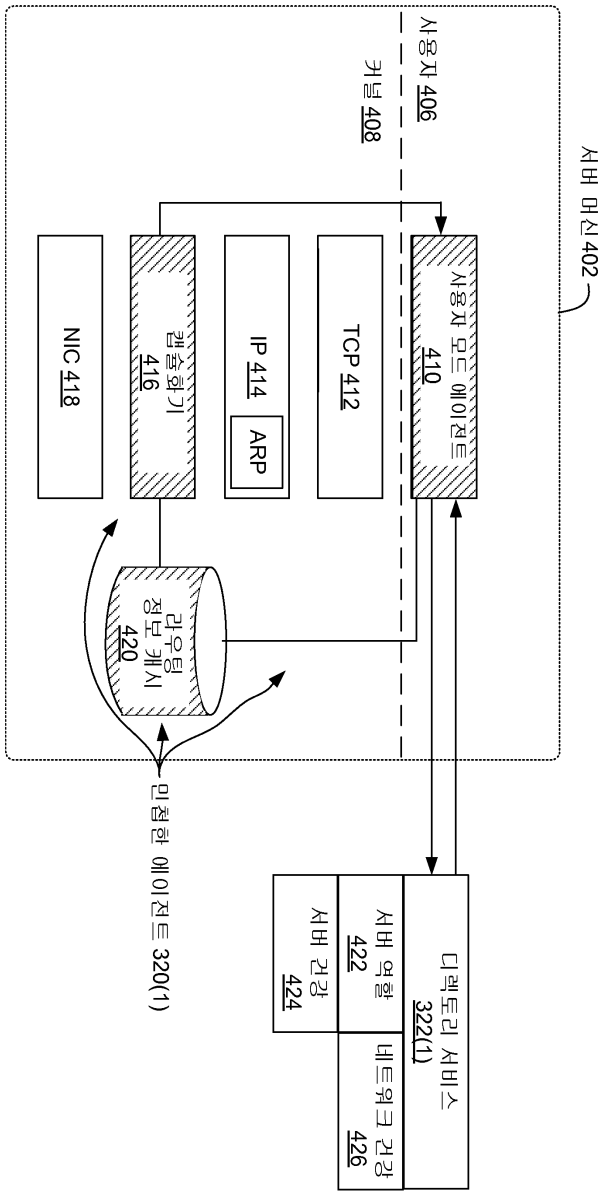
도면2


민첩한 네트워크 아키텍처 100
클라이언트 202
인터넷 204
가상 레이어-2 108
레이어-3 106
서버 102(1)
LA 206
AA 104(1)
서버 102(2)
AA 104(2)
서버 102(3)
AA 104(3)
AA 104(N)
서버 102(N)

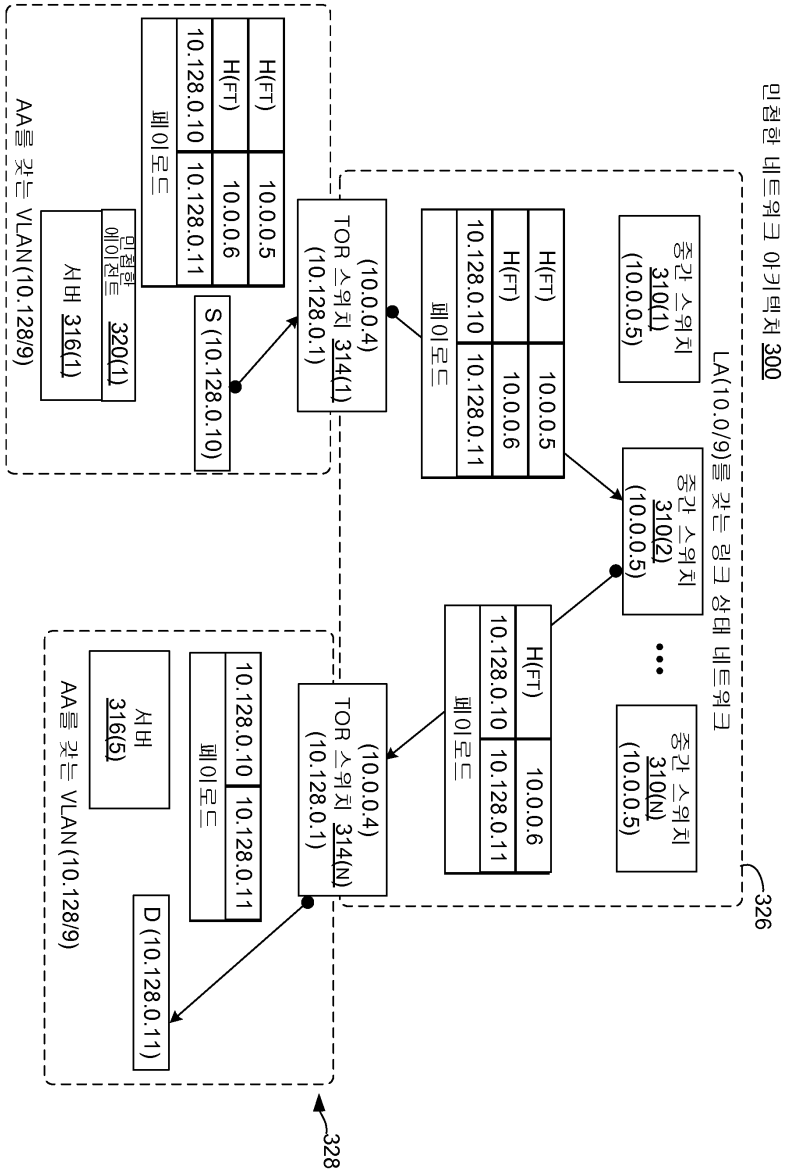
도면3


민첩한 네트워크 아키텍처 300
민첩한 시스템 304
클라이언트 302
인터넷 306
라우터 308(1)
라우터 308(N)
라우터 308
중간 스위치 310(1)
중간 스위치 310(2)
중간 스위치 310(N)
중간 스위치 310
326
집선 스위치 312(1)
집선 스위치 312(2)
집선 스위치 312(N)
집선 스위치 312
TOR 스위치 314(1)
TOR 스위치 314(2)
TOR 스위치 314(N)
TOR 스위치 314
서버 랙 318(1)
서버 랙 318(2)
서버 랙 318(N)
디렉토리 서버 322(1)
민첩한 에이전트 320(1)
서버 316(1)
민첩한 에이전트 320(2)
서버 316(2)
디렉토리 서버 322(2)
민첩한 에이전트 320(3)
서버 316(3)
민첩한 에이전트 320(4)
서버 316(4)
디렉토리 서버 322(N)
민첩한 에이전트 320(5)
서버 316(5)
민첩한 에이전트 320(N)
서버 316(N)
서버 316
328

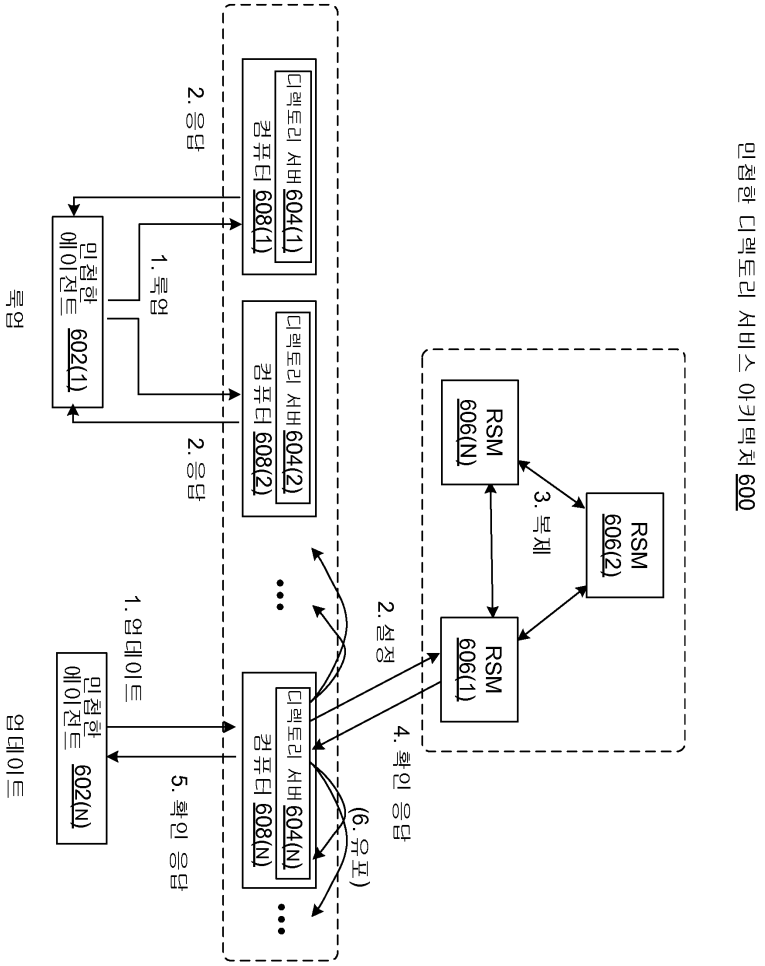
도면4


서버 머신 402
사용자 406
커널 408
사용자 모드 에이전트 410
TCP 412
IP 414
ARP
캡슐화기 416
NIC 418
라우팅 정보 캐시 420
민첩한 에이전트 320(1)
디렉토리 서비스 322(1)
서버 역할 422
서버 건강 424
네트워크 건강 426

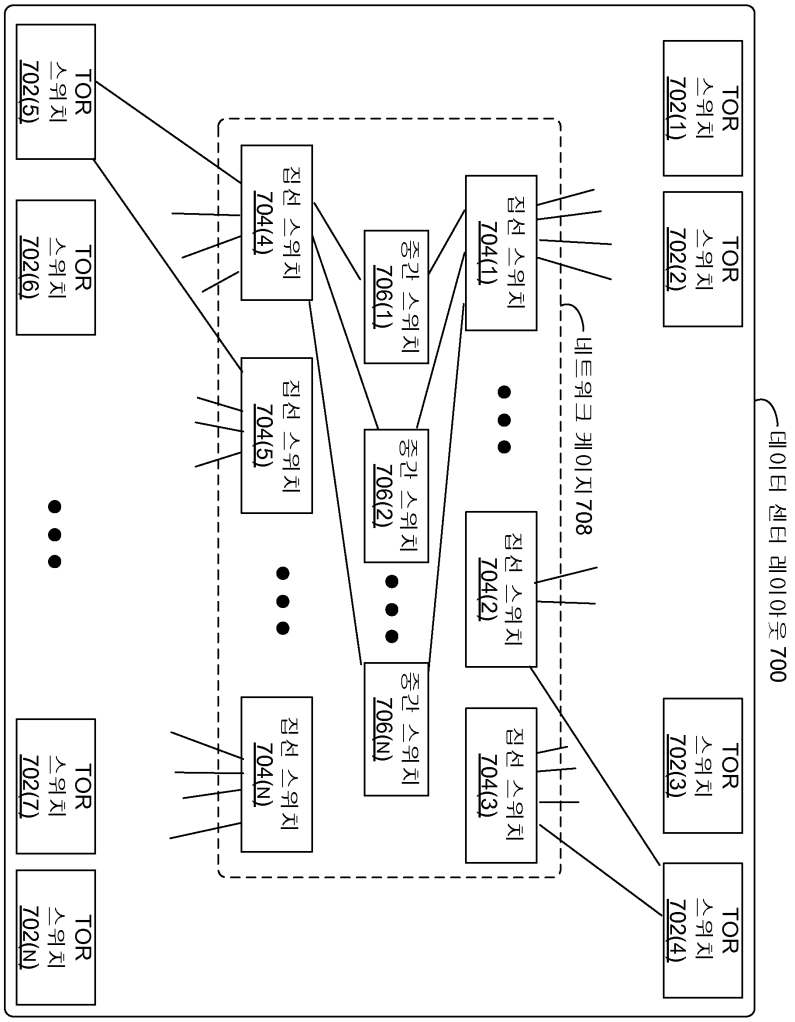
도면5


민첩한 네트워크 아키텍처 300
LA(10.0/9)를 갖는 링크 상태 네트워크
중간 스위치 310(1) (10.0.0.5)
중간 스위치 310(2) (10.0.0.5)
중간 스위치 310(N) (10.0.0.5)
H(FT) 10.0.0.5
H(FT) 10.0.0.6
10.128.0.10 10.128.0.11
페이로드
(10.0.0.4) TOR 스위치 314(1) (10.128.0.1)
(10.0.0.4) TOR 스위치 314(N) (10.128.0.1)
S (10.128.0.10)
민첩한 에이전트 320(1)
서버 316(1)
서버 316(5)
D (10.128.0.11)
AA를 갖는 VLAN (10.128/9)
326
328

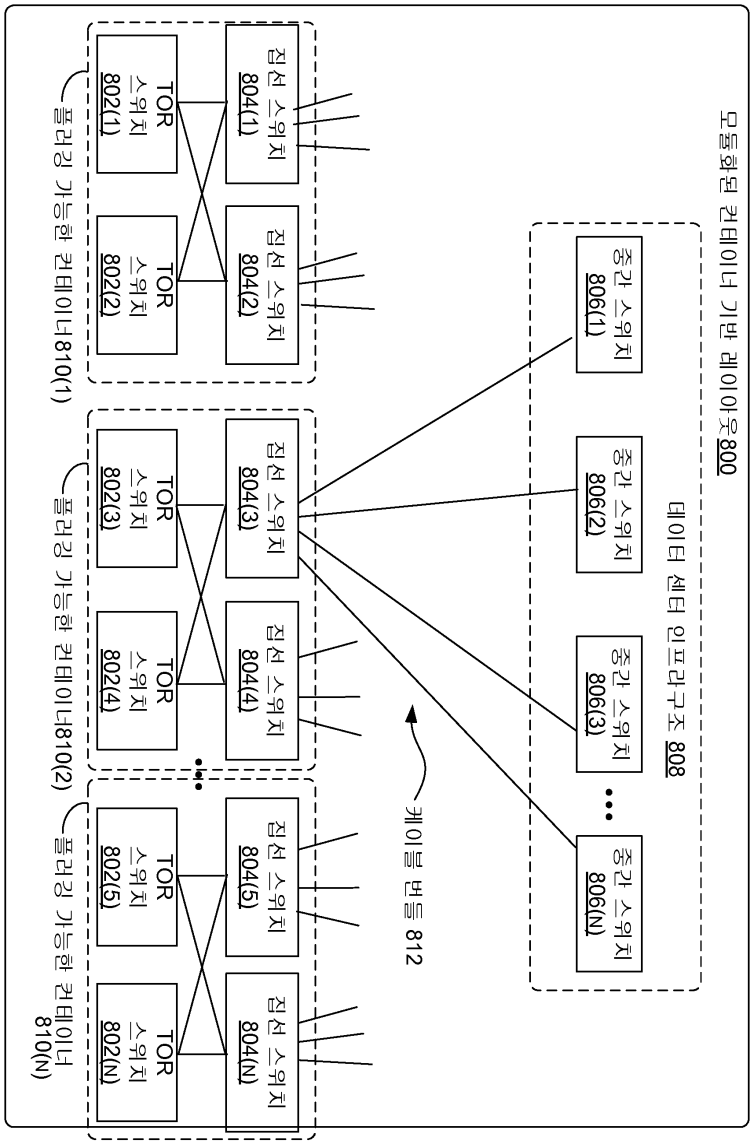
도면6


민첩한 디렉토리 서비스 아키텍처 600
RSM 606(1)
RSM 606(2)
RSM 606(N)
3. 복제
디렉토리 서버 604(1)
컴퓨터 608(1)
디렉토리 서버 604(2)
컴퓨터 608(2)
디렉토리 서버 604(N)
컴퓨터 608(N)
2. 응답
2. 응답
1. 룩업
룩업
민첩한 에이전트 602(1)
1. 업데이트
2. 설정
4. 확인 응답
(6. 유포)
5. 확인 응답
업데이트
민첩한 에이전트 602(N)

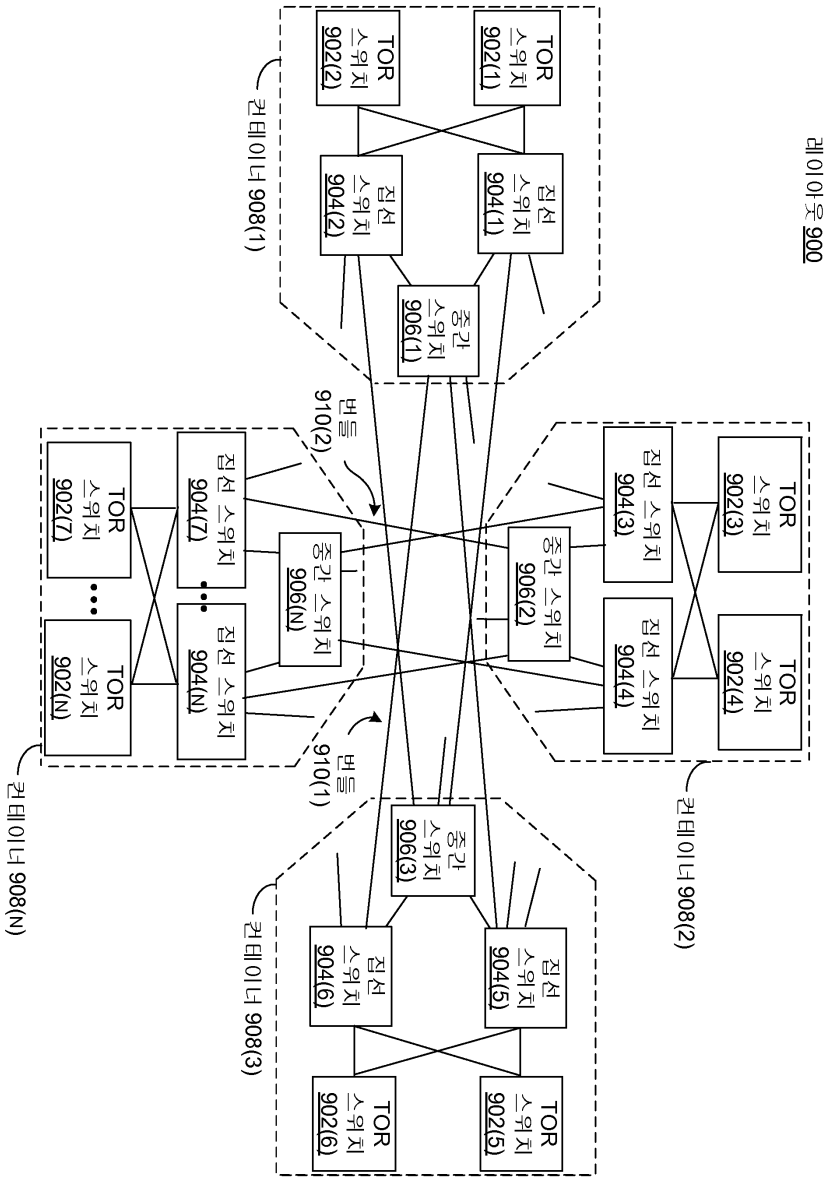
도면7


데이터 센터 레이아웃 700
네트워크 케이지 708
TOR 스위치 702(1)
TOR 스위치 702(2)
TOR 스위치 702(3)
TOR 스위치 702(4)
TOR 스위치 702(5)
TOR 스위치 702(6)
TOR 스위치 702(7)
TOR 스위치 702(N)
집선 스위치 704(1)
집선 스위치 704(2)
집선 스위치 704(3)
집선 스위치 704(4)
집선 스위치 704(5)
집선 스위치 704(N)
중간 스위치 706(1)
중간 스위치 706(2)
중간 스위치 706(N)

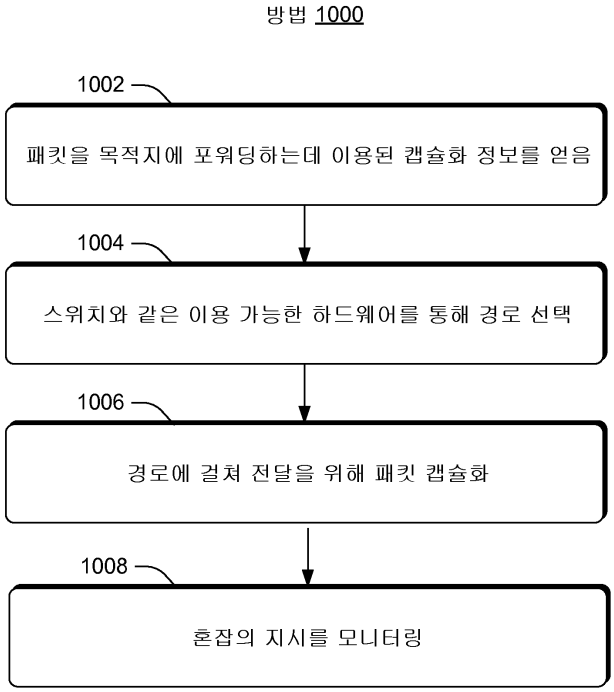
도면8


모듈화된 컨테이너 기반 레이아웃 800
데이터 센터 인프라구조 808
중간 스위치 806(1)
중간 스위치 806(2)
중간 스위치 806(3)
중간 스위치 806(N)
케이블 번들 812
플러깅 가능한 컨테이너 810(1)
TOR 스위치 802(1)
TOR 스위치 802(2)
집선 스위치 804(1)
집선 스위치 804(2)
플러깅 가능한 컨테이너 810(2)
TOR 스위치 802(3)
TOR 스위치 802(4)
집선 스위치 804(3)
집선 스위치 804(4)
플러깅 가능한 컨테이너 810(N)
TOR 스위치 802(5)
TOR 스위치 802(N)
집선 스위치 804(5)
집선 스위치 804(N)

도면9


레이아웃 900
TOR 스위치 902(1)
TOR 스위치 902(2)
TOR 스위치 902(3)
TOR 스위치 902(4)
TOR 스위치 902(5)
TOR 스위치 902(6)
TOR 스위치 902(Z)
TOR 스위치 902(N)
집선 스위치 904(1)
집선 스위치 904(2)
집선 스위치 904(3)
집선 스위치 904(4)
집선 스위치 904(5)
집선 스위치 904(6)
집선 스위치 904(Z)
집선 스위치 904(N)
중간 스위치 906(1)
중간 스위치 906(2)
중간 스위치 906(3)
중간 스위치 906(N)
컨테이너 908(1)
컨테이너 908(2)
컨테이너 908(3)
컨테이너 908(N)
번들 910(1)
번들 910(2)

도면10



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 19

【변경전】

상기 정의된 통신 그룹이

【변경후】

상기 정의된 서버 그룹이

【직권보정 2】

【보정항목】 청구범위

【보정세부항목】 청구항 9

【변경전】

상기 개별 위치 어드레스

【변경후】

상기 개별 스위치의 위치 어드레스