(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2012-0093208
(43) 공개일자 2012년08월22일

(51) 국제특허분류(Int. Cl.)
H04L 12/28 (2006.01)
(21) 출원번호 10-2012-7009062
(22) 출원일자(국제) 2010년10월06일
심사청구일자 없음
(85) 번역문제출일자 2012년04월06일
(86) 국제출원번호 PCT/US2010/051698
(87) 국제공개번호 WO 2011/044288
국제공개일자 2011년04월14일
(30) 우선권주장
12/723,697 2010년03월15일 미국(US)
61/250,013 2009년10월09일 미국(US)

(71) 출원인
마이크로소프트 코포레이션
미국 워싱턴주 (우편번호 : 98052) 레드몬드 원 마이크로소프트 웨이
(72) 발명자
파드예 지텐드라 디
미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션
칸둘라 스리칸스
미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션
바흘 파람비르
미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션
(74) 대리인
제일특허법인

전체 청구항 수 : 총 15 항

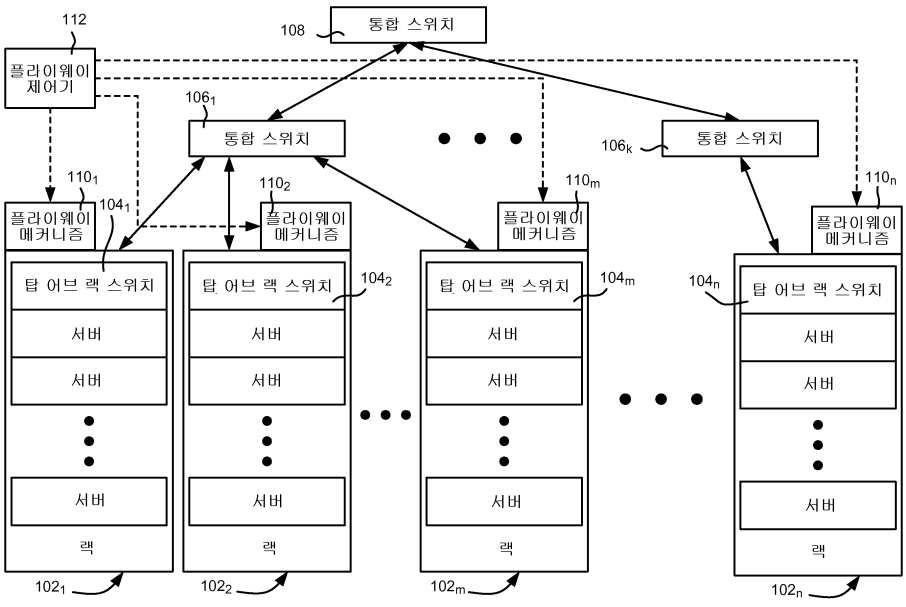
(54) 발명의 명칭 **데이터 센터에서의 플라이웨이**

(57) 요 약

추가적인 네트워크 통신 용량이 플라이웨이로 지칭되는 동적으로 제공된 통신 링크의 사용을 통해, 필요한 곳에서 초과 수용된 베이스 네트워크에 제공되는 기술이 기재되어 있다. 제어기는 2개의 네트워크 머신 사이, 예를 들어 서버의 2개의 랙 사이의

추가적인 네트워크 통신 용량에 대한 요구를 탑 어브 랙 스위치로 검출한다. 제어기는 랙의 머신들 사이에서 네트워크 트래픽의 적어도 일부를 반송하기 위해 플라이웨이 메커니즘(예를 들어, 랙 당 1개)을 구성함으로써 추가적인 네트워크 통신 용량을 제공한다. 플라이웨이 메커니즘은 60GHz 기술, 광 링크, 802.11n 또는 유선 커머더티 스위치를 포함하는 어떤 무선 또는 유선 기술에 기초할 수 있다.

대 표 도



특허청구의 범위

청구항 1

컴퓨터 네트워킹 환경에서, 플라이웨이 메커니즘들의 세트를 포함하는 시스템으로서,

상기 플라이웨이 메커니즘 세트는 서버의 랙에 연결되는 하나의 플라이웨이 메커니즘을 포함하고, 상기 하나의 플라이웨이 메커니즘은 추가적인 네트워크 통신 용량을 초과 수용된(oversubscribed) 베이스 네트워크에 제공하기 위해 다른 플라이웨이 메커니즘과 통신하는

시스템.

청구항 2

제 1 항에 있어서,

추가적인 네트워크 통신 용량에 대한 요구를 검출하는 제어기를 더 포함하며, 상기 플라이웨이 메커니즘들은 상기 제어기에 의해 서로 통신하도록 동적으로 제공되는

시스템.

청구항 3

제 1 항에 있어서,

상기 초과 수용된 베이스 네트워크는 트리형(tree-like) 토폴로지 또는 메시 토폴로지로 구성되는

시스템.

청구항 4

제 1 항에 있어서,

상기 플라이웨이 메커니즘들은 상기 네트워크의 물리적 레이아웃에 적어도 부분적으로 기초하는

시스템.

청구항 5

제 1 항에 있어서,

상기 플라이웨이 메커니즘들은 물리적 와이어를 통해 상기 다른 플라이웨이 메커니즘과 통신하거나, 상기 플라이웨이 메커니즘은 무선 기술을 통해 상기 다른 플라이웨이 메커니즘과 통신하는

시스템.

청구항 6

제 1 항에 있어서,

상기 플라이웨이 메커니즘은 무선 기술을 통해 상기 다른 플라이웨이 메커니즘과 통신하며, 상기 무선 기술은 60 GHz 주파수 대역 기술, 광 링크 또는 802.11 기반 Wi-Fi 기술, 또는 60 GHz 주파수 대역 기술, 광 링크 또는 802.11 기반 Wi-Fi 기술의 임의의 조합을 포함하는

시스템.

청구항 7

실행시에, 네트워크 트래픽(402)을 결정하는 단계와, 상기 네트워크 트래픽에 기초하여, 네트워크 내의 머신들 사이에서 상기 네트워크 트래픽의 적어도 일부를 반송하기 위해 상기 네트워크 내의 머신들 사이에 통신 링크를 제공하는 단계를 수행하는 컴퓨터 실행 가능 명령어를 포함하는

한 개 이상의 컴퓨터 판독 가능 매체.

청구항 8

제 7 항에 있어서,

상기 통신 링크를 해체하는 단계를 수행하는 컴퓨터 실행 가능 명령어를 더 포함하는

한 개 이상의 컴퓨터 판독 가능 매체.

청구항 9

제 7 항에 있어서,

머신들 사이에 다른 통신 링크를 제공하는 단계와, 및 멀티 홉 통신에 대한 적어도 1개의 중개 머신을 사용하는 단계를 수행하는 컴퓨터 실행 가능 명령어를 더 포함하는

한 개 이상의 컴퓨터 판독 가능 매체

청구항 10

제 7 항에 있어서,

상기 네트워크 트래픽을 결정하는 단계는 서버 머신 또는 폴링 스위치와 통신하는 단계, 또는 서버 머신 및 폴링 스위치 둘 다와 통신하는 단계를 포함하는

한 개 이상의 컴퓨터 판독 가능 매체

청구항 11

제 7 항에 있어서,

상기 통신 링크를 제공하는 단계는 상기 통신 링크가 제공되는 상기 머신을 선택하기 위해 배치 알고리즘을 구동하는 것을 포함하는

한 개 이상의 컴퓨터 판독 가능 매체

청구항 12

컴퓨팅 환경에서, 적어도 1개의 프로세서 상에서 수행되는 방법으로서,

네트워크 머신 쌍 사이에서 네트워크 트래픽을 결정하는 단계,

상기 네트워크 트래픽에 기초하여, 머신 쌍 사이에 제공되는 플라이웨이를 갖고 있지 않은 소정의 머신 쌍을 선택하는 단계, 및

상기 선택된 머신 쌍 사이에 플라이웨이를 제공하는 단계를 포함하는

방법.

청구항 13

제 12 항에 있어서,

다른 머신 쌍 사이의 플라이웨이를 해체하는 단계를 더 포함하는

방법.

청구항 14

제 12 항에 있어서,

상기 소정의 머신 쌍을 선택하는 단계는 최적 라우팅을 컴퓨팅하기 위해 배치 알고리즘을 실행하는 단계를 포함하는

방법.

청구항 15

제 12 항에 있어서,

단일 플라이웨이를 통해 통신이 불가능한 2개의 머신 사이에서 멀티 홉 네트워크 통신을 가능하게 하기 위해 적어도 2개의 플라이웨이를 제공하는 단계를 더 포함하는

방법.

명 세 서

배 경 기 술

[0001] 대규모 네트워크 데이터 센터는 규모의 경제, 대규모 자원 풀, 간략화된 IT 관리 및 대규모 데이터 마이닝 작업을 운영하는 능력을 제공한다. 네트워크 비용을 포함시키는 것은 대규모 데이터 센터를 구축할 때 중요한 고려사항이다. 네트워크 비용은 주요 지출 중 하나이며; 알려진 바와 같이, 서버 클러스터 내의 임의의 서버 쌍 사이에서 회선 속도 통신 대역폭을 제공하는 것과 관련된 비용은 통상 서버 클러스터의 사이즈에 따라 초선형적으로 증가한다.

[0002] 제조 데이터 센터 네트워크는 필요한 용량을 제공하기 위해 고대역폭 링크 및 하이 엔드 네트워크 스위치를 사용하지만, 여전히 초과 수용되어 있으므로(때때로 용량이 결핍되므로) 산발적인 성능 문제를 겪는다. 과다 할당은 통상 기술 제한의 조합, 고가의 "빅 아이언(big-iron)" 스위치를 필요로 하는 네트워크의 토폴로지(예를 들어, 트리형) 및 비용을 낮게 유지하기 위한 네트워크 관리자의 압박의 결과이다. 다른 네트워크 토폴로지는 유사한 문제를 갖는다.

발명의 내용

과제의 해결 수단

[0003] 본 요약은 간략화된 형태로 이하 발명을 실시하기 위한 구체적인 내용에 더 설명되는 대표적인 개념의 선택을 소개하기 위해 제공된다. 본 요약은 청구된 발명의 대상의 중요한 특징 또는 본질적인 특징을 식별하도록 의도되지 않으며, 어떠한 청구된 발명의 대상의 범위를 제한하는 방식으로 사용되도록 의도되지 않는다.

[0004] 간단히, 여기에 기재된 주제의 각종 양상은 추가적인 네트워크 통신 용량이 플라이웨이(flyway)로 지칭되는 통

신 링크의 사용을 통해 과도하게 할당된 베이스 네트워크에 제공되는 기술에 관한 것이다. 통상, 각 플라이웨이는 필요에 따라 동적으로 설정되며, 요구대로 해체될 수 있는 네트워크 머신들 사이에 추가적인 네트워크 통신 경로를 포함한다.

[0005] 하나의 구현에서, 네트워크 내의 서버의 각 랙은 다른 플라이웨이 메커니즘과 통신할 수 있는 관련 플라이웨이 메커니즘을 갖는다. 플라이웨이 메커니즘은 무선 또는 유선 기술에 기초할 수 있다. 제어기는 서로 통신하는 플라이웨이 메커니즘을 가짐으로써 2개의 랙들 사이에서 추가적인 네트워크 통신 용량에 대한 요구를 검출할 시에 플라이웨이를 동적으로 제공한다. 일단 제공되면, 머신들 사이의 추가적인 통신 링크가 그 머신들 사이에서 네트워크 트래픽의 적어도 일부를 반송하는데 사용된다.

[0006] 다른 장점은 도면과 함께 검토될 때 이하의 상세한 설명으로부터 분명해질 수 있다.

도면의 간단한 설명

[0007] 본 발명은 예로서 예시되며, 동일 참조 번호가 동일 요소를 지시하는 첨부 도면에 한정되지 않는다.

도 1은 플라이웨이가 설정될 수 있는 플라이웨이 메커니즘을 통합한 예시적 데이터 센터를 도시하는 블록도이다.

도 2는 네트워크 머신들 사이의 플라이웨이 설정의 도면이다.

도 3은 서버 랙 및 랙 세트의 센터 근방의 랙 위에 실장된 예시적 60GHz 장치의 범위를 도시하는 부분 데이터 센터의 도면이다.

도 4는 네트워크 머신 쌍 사이에 플라이웨이를 제공하기 위한 예시적 단계를 도시하는 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0008] 여기에 설명된 기술의 각종 양태는 통상 기존 네트워크의 이익은 유지하면서, 그 성능은 실질적으로 개선하는 네트워크 설계에 관한 것이다. 이 목적을 위해, 평균적인 경우에는 (트리형(tree-like) 또는 메시형(mesh-like) 등의) 베이스 네트워크가 제공되지만(따라서 초과 수용(oversubscribed)되지만), 네트워크에 필요한 경우 추가적인 용량을 제공하도록 주문에 기초하여 추가되는 여분의 링크를 포함하는 것을 통해 핫 스폿을 다루는 하이브리드 아키텍처가 기재되어 있다. 여기에 사용되는 바와 같이, 그러한 링크는 플라이웨이로 지칭된다. 또한, 장래의 네트워크 설계(예를 들어, Clos 네트워크)가 여기에 기재된 기술에 의해 강화될 수 있는 것에 주목하라. 이해되는 바와 같이, 강화된 설계는 네트워크 장비의 수량 및 비용의 상당한 증가 없이 네트워크가 서버 카운트에서 증가됨에 따라 그 성능을 유지하게 한다.

[0009] 여기에 기재된 예 중 어느 것도 비한정적인 예라는 것이 이해되어야 한다. 이와 같이, 본 발명은 여기에 기재된 어떤 특정 실시예, 양태, 개념, 구조, 기능성 또는 예에 한정되지 않는다. 오히려, 여기에 기재된 실시예, 양태, 개념, 구조, 기능성 또는 예 중 어느 하나는 비한정적이고, 본 발명은 통상 컴퓨팅 및 컴퓨터 네트워크에 이익 및 장점을 제공하는 각종 방식으로 사용될 수 있다.

[0010] 도 1은 트리형 토폴로지에 기초한 제조 네트워크를 도시한다. 복수의 랙(102_1-102_n) 각각은 탑 어브 랙(top of rack) 스위치(104_1-104_n)를 통해 통신하는 서버를 갖는다. 전형적인 네트워크는 랙 당 20 내지 40개의 서버를 갖는데, 이들은 트리 위쪽으로 갈수록 더 강력해지는(increasingly powerful) 링크 및 스위치를 갖는다. 플라이웨이가 트리형 토폴로지에 한정되지 않지만, Clos 네트워크와 다른 형태의 메시 토폴로지, 및 패트트리(FatTree) 토폴로지를 포함하는 어떤 토폴리지에도 사용될 수 있는 것에 주목하라.

[0011] 도 1에 나타낸 바와 같이, 각 탑 어브 랙 스위치(104_1-104_n)는 1개 이상의 통합(aggregation) 스위치(106_1-106_k)를 통해 서로 연결된다. 이와 같이, 각 서버는 다른 랙 내의 서버를 포함하는 임의의 다른 서버와 통신할 수 있다. 이 예에서, 하이 레벨 통합 스위치(108)는 랙 레벨 통합 스위치(106_1-106_k)를 연결하며 통합 스위치 연결의 1개 이상의 추가적인 레벨이 존재할 수 있는 것에 주목하라.

[0012] 애플리케이션 요구(application demand)는 통상 초과 수용된 네트워크에 의해 충족될 수 있지만, 때때로 네트워크는 "핫 스폿"을 처리하기 위해 충분한 용량을 갖지 못한다. 여기에 기재된 기술은 필요에 따라 여분의 데이

터 트래픽을 처리하기 위해 플라이웨이의 사용을 통하여 추가적인 용량을 제공한다.

[0013] 도 1 및 2에 나타낸 바와 같이, 플라이웨이(도 2의 만곡된 화살표)는 플라이웨이 제어기(112)에 의해 제어되는 플라이웨이 메커니즘(110_1-110_n)을 통해 구현된다. 플라이웨이는 필요성에 기초하여 플라이웨이 제어기(112)에 의해 동적으로 설정될 수 있고, 필요하지 않을 때(또는 다른 경우에 필요할 때) 제거될 수 있다. 도 2는 한쪽 플라이웨이(220)가 링크 랙(102_1 및 102_n) 및 그 각각의 탑 어브 랙 스위치(104_1 및 104_n)를 링크하는데 사용될 수 있는 한편, 다른 쪽 플라이웨이(222)가 랙(102_2 및 102_m) 및 그 각각의 탑 어브 랙 스위치(104_2 및 104_n)를 링크하는 방법을 도시한다. 랙/탑 어브 랙 스위치(rack/top-of-rack switch)는 랙(102_1 및 102_2) 사이의 플라이웨이(221)에 의해 나타낸 바와 같이, 언제든지 1개 이상의 플라이웨이를 가질 수 있는 것에 주목하라. 단일 플라이웨이 메커니즘이 랙마다 도시되어 있지만, 랙 당 2개 이상의 플라이웨이 메커니즘(또는 단일 플라이웨이 메커니즘 내에 다수의 장치)이 존재할 수도 있으며 다른 통신 기술(예를 들어, 무선 및 광)을 사용하는 것도 가능하다.

[0014] 데이터 센터 네트워크로부터의 트레이스의 분석은 언제든지 수 개의 탑 어브 랙 스위치만이 "핫"이며, 즉 대량의 트래픽을 송신 및/또는 수신하고 있는 것을 보여준다. 더욱이, 핫일 때, 탑 어브 랙 스위치는 통상 단지 수 개의 다른 탑 어브 랙 스위치와만 많은 데이터를 교환한다. 이것은 편향된 병목으로 변형되며, 여기서 수 개의 탑 어브 랙 스위치는 나머지를 지연시켜 전체 네트워크를 정지시킨다. 여기에 기재된 플라이웨이는 여분의 용량을 이 수 개의 탑 어브 랙 스위치에 제공하며 따라서 전체 성능을 상당히 개선시킨다. 실제로, 비교적 저 대역폭을 갖는 수 개의 플라이웨이만이 초과 수용된 데이터 센터 네트워크의 성능을 상당히 개선시킨다.

[0015] 플라이웨이는 비교적 적은 추가 비용으로 네트워크에 추가될 수 있다. 이것은 용량을 랜덤화된 방식으로 추가하도록 무선 링크(예를 들어, 60GHz, 광 링크 및/또는 802.11n) 및/또는 커머더티(commodity) 스위치를 사용하는 것에 의해 달성될 수 있다. 통상, 어떤 플라이웨이 메커니즘은 연결 요건(예를 들어, 무선 범위 내에서 옵티컬 등에 대한 가시선을 갖는)을 충족하기만 하면 어떤 다른 플라이웨이 메커니즘에 링크될 수 있다.

[0016] 따라서, 플라이웨이는 플라이웨이 메커니즘(예를 들어, 적당한 무선 장치) 사이의 요구에 따라 설정된 무선 링크를 통하며/통하거나 탑 어브 랙 스위치의 서브세트를 상호 접속하는 커머더티 스위치(commodity switch)를 통하는 것을 포함하는 각종 방식으로 구현될 수 있다. 이하에 기재된 바와 같이, 60 GHz 무선 기술은 근거리(1-10 미터), 고대역폭(1Gbps) 무선 링크를 지원하기 때문에 플라이웨이를 생성하기 위한 하나의 구현예이다. 게다가, 60GHz의 고용량 및 제한된 간섭 범위는 이익을 제공한다.

[0017] 플라이웨이 강화 초과 수용된 네트워크의 성능은 초과 수용되지 않은 네트워크의 성능에 근접하거나 심지어 같을 수 있다. 최대 이익을 달성하는 하나의 방법은 플라이웨이를 적절한 장치에 배치하는 것이다. 네트워크 트래픽 요구는 통상 단시간 조정으로 예측가능하고/결정가능해서, 변화하는 요구에 맞추어 플라이웨이를 제공할 수 있음에 주목하라. 여기에 기재된 바와 같이, 중앙 플라이웨이 제어기(112)는 요구 데이터를 수집하고, 플라이웨이를 동적 방식으로 적응시키며, 트래픽을 라우팅하는 경로를 스위치한다.

[0018] 무선 플라이웨이는 요구에 따라 링크를 형성할 수 있으므로, 이용가능한 용량을 중앙 플라이웨이 제어기(112)에 의해 결정된 대로 그것을 필요로 하는 어느 탑 어브 랙 스위치 쌍에도 분배하는데 사용될 수 있다. 유선 플라이웨이는 상응하는 이익을 제공한다; 염가의 스위치가 탑 어브 랙 스위치의 서브세트에 접속될 때, 이 스위치에서의 제한된 백플레인 대역폭은 접속된 다수의 탑 어브 랙 스위치 중 그것을 필요로 하는 어느 것들 사이에서 분할될 수 있다. 유선 플라이웨이는 탑 어브 랙 스위치 쌍이 플라이웨이 스위치 중 하나를 통해 접속되는 경우에만, 플라이웨이로 직접 이익을 얻을 수 있다는 점에서(멀티 홉 플라이웨이가 이하에 기재된 바와 같이 간접적인 이익을 허용하는 하나의 대안일지라도) 다소 더 제한적이다. 그러나, 유선 플라이웨이는 유선 속도를 더 유지하기 쉽다(예를 들어, NIC가 10Gbps까지 상승하고 링크가 40Gbps까지 상승하므로). 여하튼 간에, 플라이웨이는 그것을 필요로 하지 않는 링크에 사용되지 않는 만큼, 모든 랙에 걸쳐 동일 대역폭을 확산시키는 대안에 대하여 이익을 제공한다. 게다가, 플라이웨이는 현재 제조 데이터 센터의 기존 토폴로지상에서 전개될 수 있다.

[0019] 트래픽을 교환하는 핫 탑 어브 랙 스위치와 어떤 다른 탑 어브 랙 스위치 사이에 여분의 용량을 제공하는 플라이웨이를 추가하기 위해, 최대 스피드업을 제공하는 플라이웨이 메커니즘 쌍이 선택될 수 있다. 또한, 각 플라이웨이는 충분한 용량을 필요로 한다. 가장 혼잡한 탑 어브 랙 스위치를 갖는 랙과 스위치가 최대 데이터를 교환하는 다른 랙 사이에 플라이웨이를 설정하는 것은 간단하다. 그러나, 후속 선택은 예를 들어 동일 랙에 다른 플라이웨이를 설정하거나 또는 다른 곳이 고려될 필요가 있는지는 덜 명확하다.

[0020] 통상, 플라이웨이를 매우 얇게 확산시키거나, 그것을 수 개의 탑 어브 랙 스위치에 집중시키는 것은 특히 잘 작동하지 않는다. 예를 들어, 상부 50개의 탑 어브 랙 스위치와 그 가장 큰 대응 사이에 각각 1개의 플라이웨이를 배치하는 것은 핫 탑 어브 랙 스위치의 완료 시간을 충분히 감소시키지 못한다. 반대로, 상부 5개의 탑 어브 랙 스위치와 그 10개의 가장 큰 대응 각각 사이에 플라이웨이를 배치하는 것은 상부 5개에서 혼잡을 제거하지만, 6번째 탑 어브 랙 스위치가 병목으로 종료되는 것으로 귀결된다. 통상, 더 많은 탑 어브 랙 스위치를 지원하는 것과 핫 탑 어브 랙 스위치의 모두에서 충분한 혼잡을 감소시키는 것 사이에서 적절한 밸런스를 달성하는 것이 최대의 능률 향상을 얻는다. 플라이웨이 배치에 대한 하나의 적당한 알고리즘이 이하에 설명된다.

[0021] 각 플라이웨이가 얼마나 많은 용량을 필요로 하는지에 대해서는, 최대 데이터를 교환하는 상부 10개의 탑 어브 랙 스위치와 5개의 다른 탑 어브 랙 스위치 각각 사이에 플라이웨이를 추가하는 예를 고려한다(즉, 50개의 플라이웨이 전체). 대부분의 플라이웨이는 유용하게 될 탑 어브 랙 스위치의 업링크 대역폭을 10퍼센트보다 작게 요구한다; 하나의 이유는 탑 어브 랙 스위치의 업링크가 트래픽을 다른 탑 어브 랙 스위치의 모두에 반송하는 동안, 플라이웨이가 트래픽을 하나의 다른 탑 어브 랙 스위치에만 반송해야 한다는 것이다.

[0022] 플라이웨이의 유용성은 통상 스파스(sparse) 수요 매트릭스를 야기시키는 응용 특징으로부터 기인된다. 플라이웨이가 모든 수요를 위해 이익을 제공할 수는 없을지라도, 많은 실제 애플리케이션 세트는 플라이웨이의 이익을 얻는다. 예를 들어, 웹 서비스를 지원하는 데이터 센터에서, 요청 트래픽은 서버에 걸쳐 로드밸런스되고, 그 각각은 응답의 부분(예를 들어, 광고)을 생성하기 위해 수 개의 다른 서버를 요구할 수 있음으로써 응답 페이지를 차례로 어셈블리한다. 데이터 마이닝 작업에서의 맵 리듀스의 리듀스 부분이 - 각 리듀서가 모든 맵퍼로부터 데이터를 풀링함 - 아마 최악의 시나리오인데 그 이유는 모든 리듀서가 종료될 때까지 업무가 병목되기 때문이다. 그렇다 해도, 모든 탑 어브 랙 스위치가 동시에 혼잡되는 그렇게 많은 맵퍼 및 리듀서를 갖는 것은 드물다.

[0023] 60 GHz 무선 통신 등으로 돌아가면, 밀리미터 파장 무선 통신은 빠르게 향상되어왔고, 데이터 센터에 플라이웨이를 구성하기 위한 적당한 기술이다. 60 GHz 주파수 대역은 비허가 장치의 사용을 위하여 2001년에 FCC에 의해 제외된 7 GHz의 광대역(57-64 GHz)이다. ISO 및 IEEE에 의해 지정되는 채널 대역폭은 2.160 GHz이다. 60 GHz 대역에서 실질적으로 큰 대역폭은 고용량 링크를 쉽게 한다. 예를 들어, 1 bps/Hz로 달성되는 간단한 인코딩 방식에 의해 7 Gbps의 공칭 대역폭으로 링크를 구성하는 것이 가능해진다. 보다 나은 변조 및 코딩 기술이 구현되므로, 60 GHz 장치는 스펙트럼적으로 효율적이기 쉽고 7개의 직교 1 Gbps 채널은 공통일 것이며, 1 Gbps 링크 속도는 플라이웨이를 구성하는데 적당하다. 4-5 Gbps 링크 속도는 60GHz 대역에서 다수의 채널을 사용하여 이미 이용가능하다는 것에 주목하라.

[0024] 용량에 더하여, 60 GHz 통신의 다른 특징은 플라이웨이를 예시하는데 특히 매력적이다. 한 예로서는, 60 GHz 장치는 비교적 근거리(5 내지 10 미터)를 갖는다. 무선 통신 경로에서, 손실은 주파수의 제곱에 직접 비례하므로 60 GHz 신호는 2.4 및 5 GHz Wi-Fi 신호와 비교할 때 거리에 따라 더 빨리 감쇠된다. 다른 예로서는, 고 지향성 60 GHz 통신이 달성되기 쉽다. 신호 파장이 주파수에 반비례하기 때문에, 60 GHz 신호는 매우 짧은 파장(5 ㎜)을 가져서 이것은 RF 칩 설계자가 안테나 어레이를 송수신기 칩으로 통합하게 한다. 적절한 위상을 갖는 다수의 안테나 및 송수신기를 단일 칩으로 집적하는 능력은 빔 형성 및 그에 따른 지향성 통신을 허용한다. 게다가, 위상 어레이 안테나는 단거리에서 로우 비트 에러율을 갖는 클린 신호를 야기하는 신호 품질을 부스트한다. 예를 들어, 1제곱 인치 안테나는 60 GHz에서 25dBi의 신호 부스트를 제공한다. 칩 상의 안테나의 집적의 부가적인 이익은 이것이 신호를 칩에 그리고 칩으로부터 반송하기 위해 와이어의 필요성을 회피해서, 100배까지 패키징의 비용을 감소시키는 것이다.

[0025] 저범위 및 고용량 지향성 링크는 물리적 공간이 프리미엄 자원이며 이동성 및 서버 랙 당 대략 10 와트의 전력 인출은 이슈가 되지 않으므로 서버 랙이 통상 꽉 차있는 데이터 센터에 유익하다. 지향성은 네트워크 설계자가 공간 재사용을 통해 전체 스펙트럼 효율을 증가시키게 한다. 예를 들어, 4개의 탑 어브 랙 스위치 사이의 2개의 통신 세트가 지향성 및 범위 때문에 동시에 발생될 수 있다.

[0026] 전형적인 데이터 센터 서버에서, 랙은 도 3에 도시된 바와 같이 폭이 약 24 인치이며 조밀하고 규칙적인 패턴으로 배치되어 있다. 도 3에서, 각 박스는 예를 들어 대개 행으로 배치된 24x48 인치 랙을 나타낸다. 원은 약 70개의 다른 랙을 포함하는 센터에서 랙 위에 실장된 60 GHz 장치의 10m 범위를 나타낸다. 지향성 안테나를 사용해서, 랙의 상부에 실장된 작은(2-3 입방 인치) 60 GHz 네트워크 카드는 필요성에 기초하여 수 개의 랙들 사이에 다수의 비간섭 기가비트 플라이웨이를 예시할 수 있다. 약 60도의 빔 폭을 갖는 전자적으로 조종 가능한 위상 어레이 안테나는 밀리초의 레이턴시로 조종될 수 있다.

[0027] 지향성이 결과에 수반되는 것에 주목하라. 예를 들어, 발견은 더 이상 수월하지 않으며, 무 지향성 통신에 잘 작동되는 대중적인 매체 액세스 제어 프로토콜이 수정되어야 한다. 그럼에도 불구하고, 이것은 연구계에서 잘 이해하는 문제이며 다수의 해결방안이 존재한다. IEEE 및 ISO 60 GHz 워킹 그룹은 지향성 통신 및 공간 재사용을 위한 지원을 통합하기 위해 적극적으로 노력하고 있다. 그 주요 시나리오가 가정 내 멀티미디어 통신일 지라도, 그들이 논의하고 있는 해결방안은 데이터 센터 네트워크에 적용되도록 수정될 수 있다. 특히, 네트워크가 구성되는 방식 및 네트워크의 규모로 인해, 그 문제는 제어 채널로 변화될 수 있는 모든 발견가능한 탑 어브 랙 스위치 사이에서 종종 유선인 백채널 접속이 존재하므로 해결될 수 있다. 제어 채널은 이 때 탑 어브 랙 스위치를 발견하고, 60 GHz 빔을 정렬시키고, 간섭 완화를 위한 채널을 할당하고, 충돌 회피 매체 액세스를 조정하는데 사용될 수 있다.

[0028] 60 GHz에 있어서의 다른 문제는 이 고주파수에서, 신호가 물체에 의해 흡수되고 비가시선(NLOS) 통신이 가능하지 않다는 것이다. 그러나, 데이터 센터에서, 장치는 랙의 상부에 실장될 수 있고 인간 운영자의 방해를 받지 않으며, 이로 인해 가시선이 차단되지 않는다. 더욱이, 기술에 있어서 최근의 향상은 이 제한을 극복하기 위해 시작되었다.

[0029] 상술한 바와 같이, 스펙트럼 효율의 개선과 함께 대량의 스펙트럼은 다수의 플라이웨이가 독립적으로 구성되고 운영되게 한다(10 bits/Hz 스펙트럼 효율로, 1Gbps에서 각각 동작하는 70개까지의 플라이웨이가 구성될 수 있다). 지향성으로 인한 저간섭, 및 경로 손실로 인한 저범위의 성질과 결합되는 것은 데이터 센터에 존재하는 수천 개의 서버 랙을 통해 상당한 주파수 재사용을 허용한다.

[0030] 더욱이, 다른 고대역폭 무선 기술과 다르게, 60 GHz 대역은 전세계에서 이용가능하다. 따라서, 칩셋 및 장치는 전세계에 사용가능하고 규모의 경제를 제공하고 비용을 내린다.

[0031] 네트워크를 플라이웨이로 구현하는 것으로 돌아가면, 탑 어브 랙 스위치 쌍 사이에 플라이웨이를 형성하기 위해, 대응하는 랙 위에 배치된 1개 이상의 장치는 무선 링크를 생성한다. 기술의 선택은 이용가능한 대역폭, 공간 재사용을 위해 이용가능한 채널의 수, 간섭 패턴 및 플라이웨이의 범위에 영향을 미친다. 안테나 기술은 설정에 요구되는 시간을 지정하고 플라이웨이를 해체한다. 탑 어브 랙 스위치 당 수 개의 무선 장치를 추가하는 것은 비용을 아주 조금만 증가시키는 것에 주목하라.

[0032] 유선 플라이웨이는 예컨대 탑 어브 랙 스위치의 임의 서브세트를 상호 접속하는 현대 탑 어브 랙 스위치와 동일 구성의 추가적인 스위치를 사용함으로써 구성될 수 있으며, 예를 들어 커머더티 스위치(commodity switch)는 20개의 탑 어브 랙 스위치를 1Gbps 링크로 각각 상호 접속할 수 있다. 링크를 짧게 유지하기 위해, 플라이웨이 스위치는 데이터 센터에서 상호 근접한 랙을 접속할 수 있다. 유선 플라이웨이를 전개할 때, 스펙트럼 할당 또는 간섭은 문제가 되지 않는다. 그러나, 그 임의 구성은 유선 플라이웨이를 제한한다; 예를 들어, 많은 트래픽을 교환하고 과잉 용량의 이익을 얻을 수 있는 탑 어브 랙 스위치 쌍은 유선 플라이웨이 없이 종료될 수 있다.

[0033] 여하튼 간에, 유선 또는 무선 플라이웨이 구성은 동일량의 대역폭을 모든 탑 어브 랙 스위치 쌍에 걸쳐 균일하게 분할하는 것에 비해서 개선된 것이다. 대역폭을 균일하게 확산시키고 그것을 많이 소모하기보다는 오히려, 수요 매트릭스가 스파스될 때 일어나므로, 플라이웨이는 과잉 용량으로부터 최대로 이익을 얻을 수 있는 수요 매트릭스의 부분을 목표로 정하기 위해 여분의 대역폭을 사용하는 방법을 제공한다.

[0034] 중앙 플라이웨이 제어기(112)(도 1)는 랙 스위치 쌍 사이에서 수요의 추정을 수집한다. 예를 들어, 정보는 엔드 서버 그 자체에서 경량의 기구로부터 수집되거나, 스위치에서 SNMP 카운터를 폴링함으로써 수집될 수 있다. 이 추정을 사용하면, 제어기(112)는 이용가능한 플라이웨이를 배치하기 위해 배치 알고리즘(이하 기재되는)을 주기적으로 또는 다르게 실행할 수 있다.

[0035] 따라서, 플라이웨이 기반 네트워크의 토폴로지는 동적이며, 다중 경로 라우팅을 필요로 한다. 유사한 문제에 대한 해결방안이 알려져 있다. 제어기(112)는 탑 어브 랙 스위치 쌍 사이의 트래픽이 베이스 네트워크를 따라 어느 정도 진행되는지, 그 대신에 1개가 존재하면 플라이웨이를 송신 탑 어브 랙 스위치로부터 수신 탑 어브 랙 스위치로 가져가는지를 판단한다. 탑 어브 랙 스위치는 다른 흐름을 다른 MPLS(Multiprotocol Label Switching) 라벨 교환 경로 상으로 할당함으로써 트래픽을 이 비율에 따라 분할한다. 수 개의 플라이웨이만이 있다면 이들은 각 탑 어브 랙 스위치에서 이용가능하다는 것에 주목하라. 그러므로, 각 스위치에서 요구되는 LSP(label switched path)의 수는 적고 하나의 홉이 긴 베이스 및 플라이웨이에 걸쳐 트래픽을 분할하는 문제는 다른 트래픽 분할 기술보다 상당히 더 간단하다.

[0036] 최적 플라이웨이를 생성하는 문제는 최적화 문제로 간주될 수 있다. 탑 어브 랙 스위치(i, j) 사이의 D_{ij} 수요

가 주어지면 그리고 C_l이 링크(l)의 용량이면, 최적 라우팅은 최대 완료 시간을 최소화하는 것이다:

$$\text{최소최대}\ \frac{D_{ij}}{r_{ij}} \qquad (1)$$

$$\sum_{l\in\text{인커밍}} r_{ij}^{l} - \sum_{l\in\text{아웃고잉}} r_{ij}^{l} = \begin{cases} D_{ij} & \text{ToR } j \text{ 에서} \\ -D_{ij} & \text{ToR } i \text{ 에서} \\ 0 & \text{모든다른 ToRs 에서} \end{cases}$$

$$\sum_{ij} r_{ij}^{l} \leq C^{l} \quad \forall \text{ 링크} l$$

[0037]

[0038] 여기서 rij는 탑 어브 랙 스위치 쌍(i, j)에 대해 달성된 비율이고 r_{ij}^{l}는 링크(l) 상의 그 쌍의 트래픽의 부분이다.

[0039] 최적/근사 최적 플라이웨이 배치를 컴퓨팅하는 것은 토폴로지를 적절히 변경하는 것 및 상술한 최적화 문제를 해결하는 것을 수반한다. 예를 들어, 일정 수만이 동시에 활성될 수 있거나, 플라이웨이들 중 어느 것도 일정량보다 더 큰 용량을 가질 수 없다는 제약에 의해 모든 가능한 플라이웨이가 추가될 수 있다. 기술 혁신 제약으로 인해 유도되는 상술한 최적화 문제의 모든 변형이 취급되기 쉬울 수 있는 것은 아니다. 그 대신에, 완료 시간을 최고로 감소시키는 플라이웨이를 발견하기 위해 상술한 최적화 문제를 해결함으로써 하나의 플라이웨이를 동시에 추가하는 그리디(greedy) 절차가 사용될 수 있다. 다른 절차가 더 좋은 근사일 수 있다.

[0040] 일례는 고려되는 쌍에 대한 네트워크 트래픽 데이터를 수집하는 것을 나타내는 단계 402에서 시작되는 도 4의 예시적 흐름도에 전체적으로 도시되어 있다. 통상, 트래픽 데이터에 기초하여, 2개의 선택된 머신 쌍(탑 어브 랙 스위치) 사이에서 필요한(예를 들어, 스위치가 핫이지만 플라이웨이가 이미 제공되지 않은) 가장 최적인 플라이웨이가 단계 404에서 선택된다.

[0041] 플라이웨이는 랙의 나머지로의 1차 접속이 실패되는 랙에 접속성을 제공하는데 사용될 수도 있는 것에 주목하라. 그러한 실패 보상 플라이웨이(또는 플라이웨이들)는 예를 들어 단계 404에서 필요한 가장 최적인 플라이웨이로 간주될 수 있거나, 용량에 기초하여 다른 것을 선택하기 전과 같이 개별적으로 간주될 수 있다.

[0042] 단계 406은 가장 최적인 플라이웨이가 제공될 수 있는지를 평가하는 것을 나타낸다; 예를 들어, 플라이웨이의 전체 수가 제약이면, 이 시간에 이용가능한 다른 플라이웨이가 존재할 수 없다. 그러나, 이용가능한 플라이웨이가 존재하지 않으면, 아직 가장 최적인 제공되지 않은 플라이웨이는 현재 제공되는 다른 플라이웨이보다 더 필요하고, 단계 408은 플라이웨이가 이용가능하도록 다른 플라이웨이(또는 플라이웨이들, 멀티 홉 개요가 이하에 기재되는 바와 같이 사용중이면)를 해체할 수 있다. 예를 들어, 최소로 필요한 플라이웨이가 해체될 수 있다.

[0043] 그러나 그것은 플라이웨이를 해체하는데 적절하지 않을 수도 있다는 것에 주목하라. 예를 들어, 그것은 이미 제공된 플라이웨이 모두가 심지어 더 많은 트래픽을 취급하고 있으므로 단계 404에서 선택된 쌍에 대한 플라이웨이보다 더 필요할 수 있으며, 그 경우에 어떤 액션도 취해지지 않는다(예를 들어, 단계 408로부터의 파선).

[0044] 단계 410은 가장 최적인 플라이웨이(적절하다면)를 제공한다. 제공되면, 플라이웨이는 단계 412에 의해 나타낸 바와 같이, 선택된 쌍 사이에서 네트워크 트래픽의 적어도 일부를 취급하는데 사용될 수 있다. 그 다음에 처리가 반복된다. 도 4는 일례만인 것에 주목하라; 대안으로서, 상부 N개까지 이미 제공되지 않은 가장 최적인 플라이웨이는 단계 402에서 쌍의 각 폴링 사이에 추가될 수 있다.

[0045] 플라이웨이에 있어서, 기술로 인한 제약뿐만 아니라 네트워크의 물리적 토폴로지도 고려될 필요가 있다. 무선 플라이웨이는 범위에 의해 제약되는 한편, 탑 어브 랙 스위치의 임의 서브세트를 상호 접속함으로써 구성되는 유선 플라이웨이는 이 임의 서브세트들 사이에 과잉 용량만을 제공할 수 있다. 60 GHz 플라이웨이는 10 미터의 거리를 스팬하고 제조 데이터 센터로부터의 데이터 센터 레이아웃(예를 들어, 도 3) 사용한다. 유선 플라이웨이에 대해서는, 24 포트, 1Gbps 스위치는 포트의 일부가 플라이웨이 채널에 사용된 상태로 사용될 수 있다. 그러한 제약이 플라이웨이의 이익에 영향을 줄지라도, 그 이익은 여전히 중요하다.

[0046] 더욱 많은 유선 플라이웨이가 무선 플라이웨이의 동일 이익을 얻는데 추가될 필요가 있는 것에 주목하라. 예를 들어, 50개의 24 포트 스위치가 추가될 때, 1,200 듀플렉스 링크가 네트워크에 추가된다. 무선 플라이웨이는 등가 성능을 50개의 하프 듀플렉스 링크에만 제공한다. 이것은 무선 플라이웨이가 목표된 방식으로 추가되기 때문이다: 그것은 추가적인 용량을 필요로 하는 탑 어브 랙 스위치 쌍을 스피드업하는데 원조한다. 유선 플라

이웨이는 임의로 추가되고, 선택된 것 중에서 발생되면 단지 쌍의 이득이 된다.

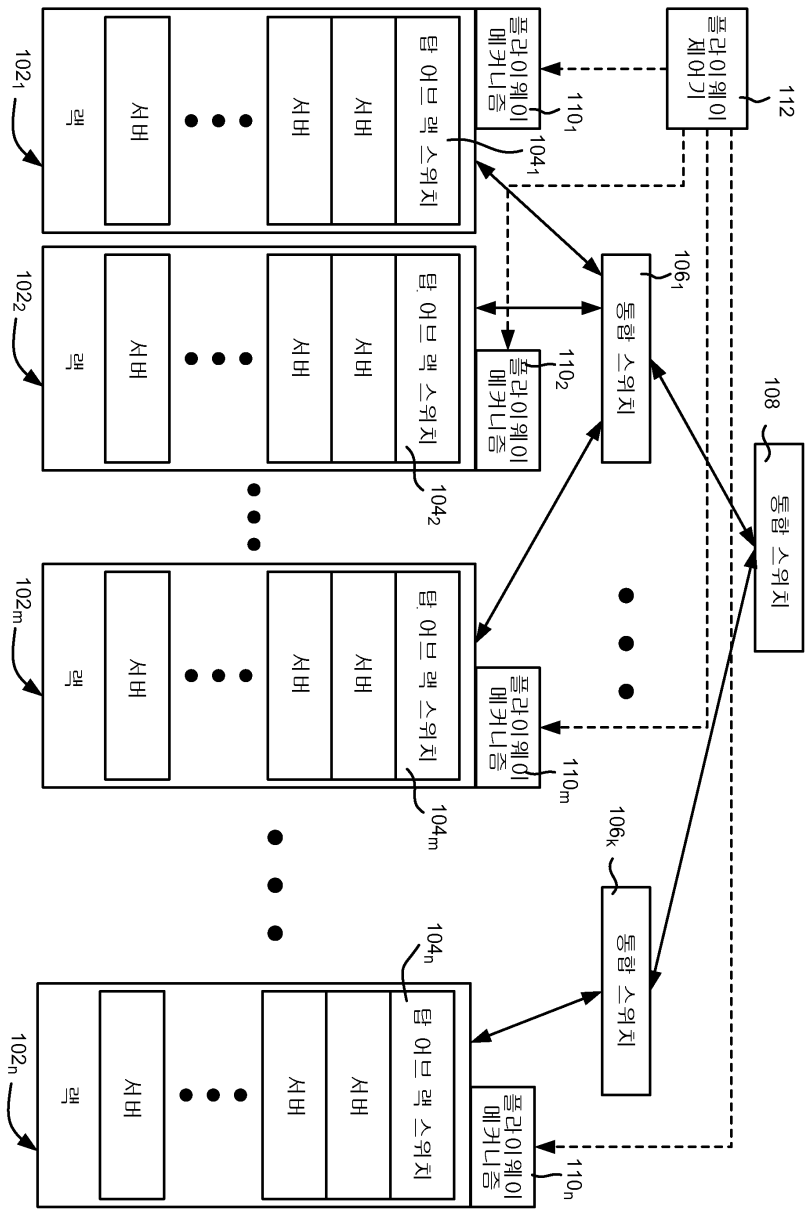
[0047] 멀티 홉 플라이웨이가 가능하다. 도 3에서 원 내의 머신이 원 외의 머신과 통신하기를 원하는 것을 고려하자. 그러나, 다른 머신과 충분히 가까운 원 내의 머신으로부터의 제 2 홉은 통신을 달성하는데 사용될 수 있다.

[0048] 게다가, 제어기(112)(도 1)는 2개의 머신(예를 들어, A 및 Z)이 단일 홉 범위 외에 있는지를 고려할 수 있다. 그렇다면, 플라이웨이는 통신이 플라이웨이에 교환된 상태로 근접하지만 탑 어브 랙 스위치를 통해 통신하는 2개의 다른 머신(예를 들어, C 및 M) 사이에 구성될 수 있다. C 및 M에 대한 새롭게 구성된 플라이웨이에 의해 제공되는 대역폭 절감은 이 때 머신(A 및 Z)이 탑 어브 랙 스위치를 사용하여 통신하게 하기 위해 탑 어브 랙 스위치를 제거하는데 사용될 수 있다.

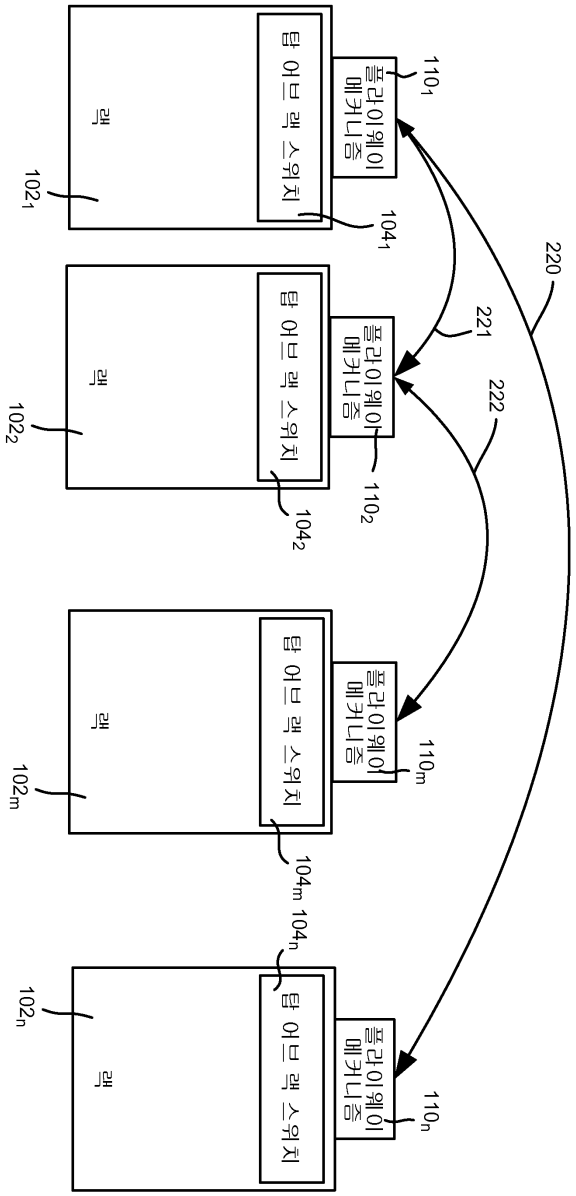
[0049] 본 발명은 각종 수정 및 대체 구성이 가능할지라도, 어떤 예시된 실시예는 도면에 도시되어 있고 자세히 상술되어 있다. 그러나, 개시된 특정 형태에 본 발명을 제한하는 의도가 존재하는 것이 아니라, 반대로 그 의도는 본 발명의 정신 및 범위 내에 포함되는 모든 수정, 대체 구성, 및 등가물을 커버하는 것이 이해되어야 한다.

도면

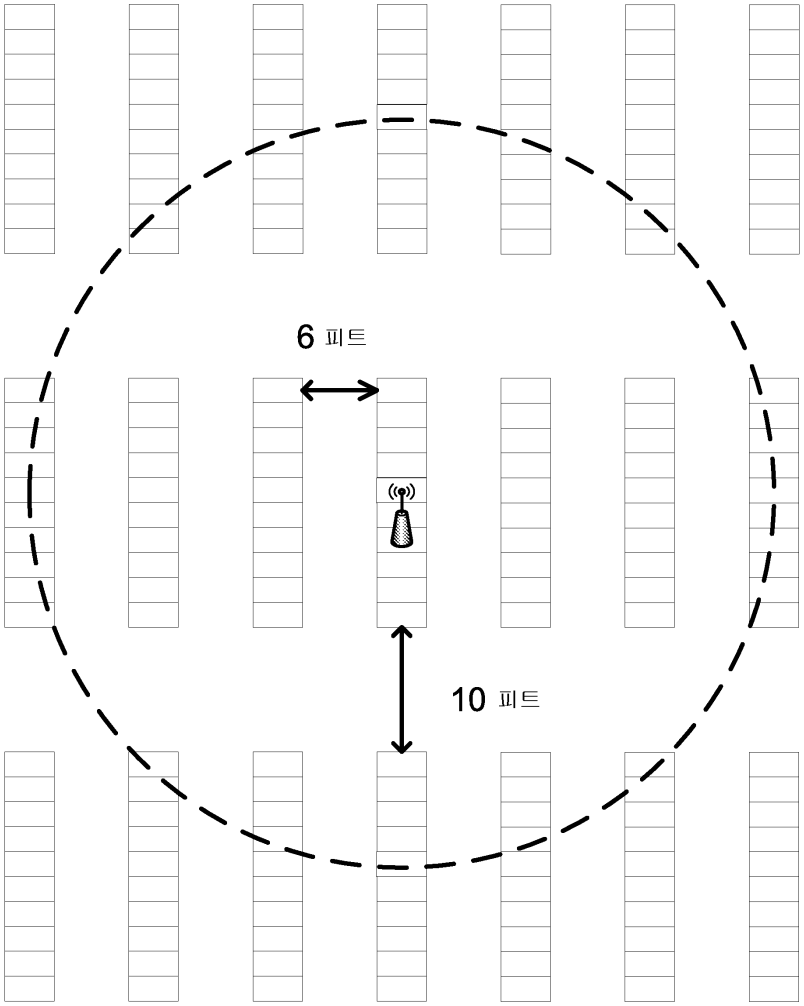
도면1



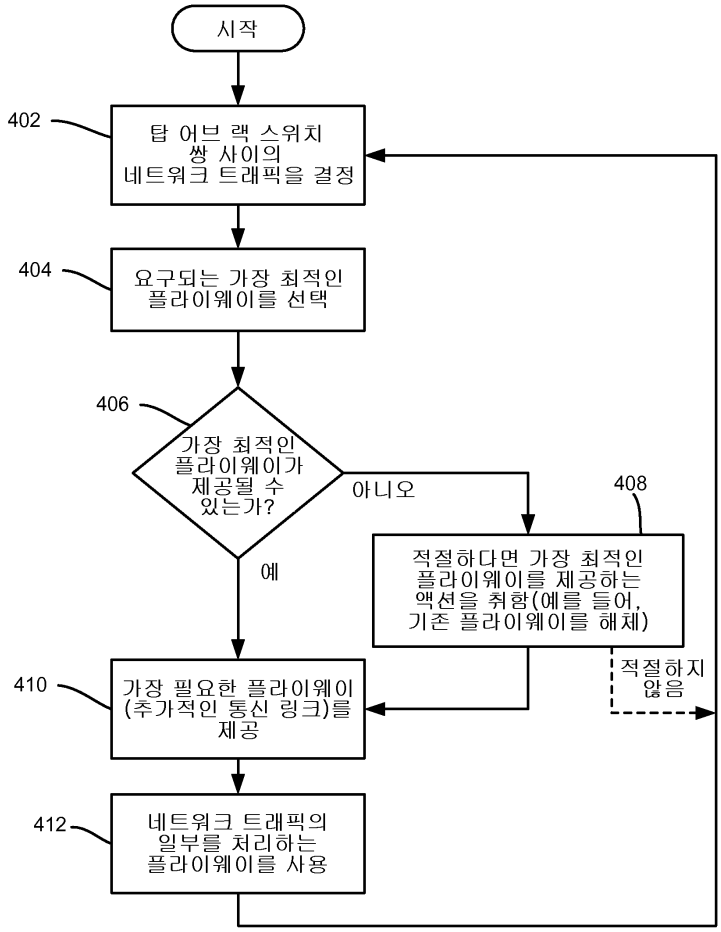
도면2


랙
탑 어브 랙 스위치
플라이웨이 메커니즘
102_1
104_1
110_1
102_2
104_2
110_2
102_m
104_m
110_m
102_n
104_n
110_n
220
221
222

도면3



도면4


시작
402
탑 어브 랙 스위치 쌍 사이의 네트워크 트래픽을 결정
404
요구되는 가장 최적인 플라이웨이를 선택
406
가장 최적인 플라이웨이가 제공될 수 있는가?
아니오
예
408
적절하다면 가장 최적인 플라이웨이를 제공하는 액션을 취함(예를 들어, 기존 플라이웨이를 해체)
적절하지 않음
410
가장 필요한 플라이웨이(추가적인 통신 링크)를 제공
412
네트워크 트래픽의 일부를 처리하는 플라이웨이를 사용