(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2014-0002013
(43) 공개일자 2014년01월07일

(51) 국제특허분류(Int. Cl.)
H04L 12/851 (2013.01) *H04L 12/937* (2013.01)
(21) 출원번호 10-2013-7028257
(22) 출원일자(국제) 2012년04월12일
심사청구일자 2013년11월21일
(85) 번역문제출일자 2013년10월25일
(86) 국제출원번호 PCT/IB2012/051803
(87) 국제공개번호 WO 2012/156832
국제공개일자 2012년11월22일
(30) 우선권주장
13/107,893 2011년05월14일 미국(US)

(71) 출원인
인터내셔널 비지네스 머신즈 코포레이션
미국 10504 뉴욕주 아몬크 뉴오차드 로드
(72) 발명자
캠블, 케샤브, 고빈드
미국 캘리포니아 95054, 산타 클라라, 미션 칼리지 불러바드 2051, 아이비엠 코포레이션
판데이, 비조이
미국 캘리포니아 95054, 산타 클라라, 미션 칼리지 불러바드 2051, 아이비엠 코포레이션
(뒷면에 계속)
(74) 대리인
허정훈

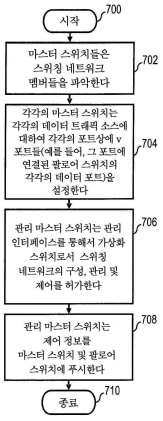
전체 청구항 수 : 총 29 항

(54) 발명의 명칭 **분산형 패브릭 프로토콜(ＤＦＰ) 스위칭 네트워크 아키텍처에서 우선순위 기반 플로우 제어**

(57) 요 약

스위칭 네트워크는 상위 층과 복수의 하위 층 엔티티들을 포함하는 하위 층을 포함한다. 각각의 포트가 각각의 하위 층 엔티티에 결합된 복수의 포트들을 갖는, 상위 층의 마스터 스위치는 상기 포트들 각각에 복수의 가상 포트들을 구현하며 상기 복수의 가상 포트들 각각은 그 포트에 결합된 하위 층 엔티티에서 복수의 원격 물리적 인터페이스들(RPI들) 각각에 대응한다. 상기 마스터 스위치와 RPI들 사이에서 전달되는 데이터 트래픽은 그 데이터 트래픽이 저달되는 하위 층 엔티티들상의 RPI들에 대응하는 가상 포트들 내에서 대기한다(queued). 상기 마스터 스위치는 주어진 가상 포트의 데이터 트래픽 상에서 우선순위-기반 플로우 제어(PFC)를 실행하는데, 대응 RPI가 상주하는 하위 층 엔티티에, 특정 RPI에 의해서 전달되는 데이터 트래픽의 적어도 두 개의 다른 클래스에 대해 우선순위를 명시하는 PFC 데이터 프레임을 전송함으로써 실행한다.

대 표 도



(72) 발명자

카맛, 데이야반티, 고팔

미국 캘리포니아 95054, 산타 클라라, 미션 칼리지 불러바드 2051, 아이비엠 코포레이션

리우, 다렌

미국 캘리포니아 95054, 산타 클라라, 미션 칼리지 불러바드 2051, 아이비엠 코포레이션

키담비, 자야크리쉬나

미국 캘리포니아 95054, 산타 클라라, 미션 칼리지 불러바드 2051, 아이비엠 코포레이션

멘든, 첸다라니

미국 캘리포니아 95054, 산타 클라라, 미션 칼리지 불러바드 2051, 아이비엠 코포레이션

특허청구의 범위

청구항 1

상위 층(an upper tier)과 복수의 하위 층 엔티티들을 포함하는 하위 층(a lower tier)을 포함하는 스위칭 네트워크에서 플로우 제어(flow control)를 구현하는 방법에 있어서, 상기 방법은:

복수의 포트들을 갖고 각각의 포트는 복수의 하위 층 엔티티들 각각에 결합 가능한 상위 층 내의 마스터 스위치에서, 복수의 가상 포트들을 상기 복수의 포트들의 각각 상에 구현하는 단계(implementing) - 상기 복수의 가상 포트들 각각은 그 포트에 결합 가능한 하위 층 엔티티에서 복수의 원격 물리적 인터페이스들(RPI들) 각각에 대응함 -;

상기 마스터 스위치와 상기 복수의 하위 층 엔티티들상의 RPI들 사이에서 전달되는(communicated) 데이터 트래픽을 상기 복수의 가상 포트들 중에서 그 데이터 트래픽이 전달되는 하위 층 엔티티들상의 RPI들에 대응하는 가상 포트들 내에 대기시키는 단계(queuing); 및

상기 마스터 스위치가 우선순위-기반 플로우 제어(PFC)를 주어진 가상 포트의 데이터 트래픽 상에 실행하는 단계에 있어서, 상기 주어진 가상 포트에 대응하는 특정 RPI가 상주하는 하위 층 엔티티에, 상기 특정 RPI에 의해서 전달되는 데이터 트래픽의 적어도 두 개의 다른 클래스들에 대한 우선순위를 명시하는 PFC 데이터 프레임을, 전송함으로써 실행하는 단계(enforcing)를 포함하는,

방법.

청구항 2

제1항에 있어서,

상기 마스터 스위치와 상기 복수의 하위 층 엔티티들 사이에서 전달되는 데이터 트래픽은 그 트래픽이 전달되는 상기 하위 층 엔티티들 상의 RPI들을 식별하는 RPI 식별자들(RPI identifiers)을 포함하고; 그리고

상기 대기시키는 단계는 상기 데이터 트래픽을 상기 RPI 식별자들에 기초하여 상기 마스터 스위치상의 가상 포트들에 대기시키는 단계를 포함하는,

방법.

청구항 3

제1항 또는 제2항에 있어서,

상기 하위 층 엔티티는 플로우 스루 모드(flow through mode)로 구성된 팔로어 스위치(a follower switch)를 포함하고;

상기 특정 RPI는 데이터 포트를 포함하며; 그리고

상기 방법은 상기 팔로어 스위치가 상기 PFC 데이터 프레임을 수신하는 단계, 상기 PFC 데이터 프레임으로부터 상기 특정 RPI를 식별하는 RPI 식별자를 제거하여 표준 PFC 데이터 프레임을 획득하는 단계, 그리고 상기 표준 PFC 데이터 프레임을 상기 특정 RPI에 대응하는 데이터 포트를 통해서 전송하는 단계를 더 포함하는

방법.

청구항 4

제1항 내지 제3항 중 어느 한 항에서,

상기 하위 층 엔티티는 네트워크 인터페이스와 가상 머신 모니터를 갖는 호스트 플랫폼을 포함하고;

상기 특정 RPI는 상기 호스트 플랫폼상에서 실행하는 가상 머신을 포함하고; 그리고

상기 방법은 상기 네트워크 인터페이스가 상기 PFC 데이터 프레임을 상기 가상 머신 모니터로 전달하는 단계를 더 포함하는

방법.

청구항 5

제4항에 있어서, 상기 방법은:

상기 PFC 데이터 프레임을 수신하는 것에 응답하여, 상기 호스트 플랫폼상의 상기 가상 머신 모니터가 상기 PFC 데이터 프레임을 상기 PFC 데이터 프레임에서 명시한 상기 특정 RPI에 기초하여 상기 가상 머신에 전송하는 단계를 더 포함하는

방법.

청구항 6

제1항 내지 제5항 중 어느 한 항에서, 상기 실행하는 단계는 상기 마스터 스위치가 PFC를 우선순위별로(per-priority), 애플리케이션별로(per-application) 실행하는 단계를 포함하는

방법.

청구항 7

제1항 내지 제6항 중 어느 한 항에 있어서, 상기 특정 RPI는 물리적 포트, 링크 집합 그룹(LAG: a link aggregation group) 인터페이스 및 가상 포트를 포함하는 세트 중 하나를 포함하는,

방법.

청구항 8

제7항에 있어서, 상기 가상 포트는 가상 네트워크 인터페이스 카드(NIC), 단일 루트(Single Root) I/O 가상화(SR-IOV) NIC 파티션, 및 파이버채널 오버 이더넷(FCoE) 포트(FibreChannel over Ethernet (FCoE) port)를 포함하는 세트 중 하나를 포함하는,

방법.

청구항 9

제1항 내지 제8항 중 어느 한 항에 있어서, 상기 PFC 프레임은:

상기 특정 RPI를 식별하는 RPI 필드; 및

상기 데이터 트래픽의 적어도 두 개의 다른 클래스에 대한 상대적 우선순위를 명시하는 복수의 필드들을 포함하는,

방법.

청구항 10

제1항 내지 제9항 중 어느 한 항에서, 상기 실행하는 단계는 상기 마스터 스위치가 상기 하위 층 엔티티 상의 두 개의 다른 RPI들에 대하여 상이한 우선순위-기반 플로우 제어(differing priority-based flow control)를 실행하는 단계를 포함하는,

방법.

청구항 11

스위칭 네트워크를 위한 마스터 스위치에 있어서, 상기 스위칭 네트워크는 상기 마스터 스위치를 포함하는 상위 층과 복수의 하위 층 엔티티들을 포함하는 하위 층을 포함하고, 상기 마스터 스위치는:

각각이 상기 복수의 하위 층 엔티티들 각각에 결합할 수 있고, 각각이 복수의 가상 포트들을 포함하는 복수의 포트들 - 상기 복수의 가상 포트들 각각은 그 포트에 결합 가능한 상기 하위 층 엔티티에서 복수의 원격 물리적 인터페이스들(RPI들) 각각에 대응하고, 상기 마스터 스위치와 상기 복수의 하위 층 엔티티들상의 RPI들 사이에서 전달되는 데이터 트래픽은 상기 복수의 가상 포트들 중에서 그 데이터 트래픽이 전달되는 하위 층 엔티티

들 상의 RPI들에 대응하는 가상 포트들에 대기함 -; 및

상기 복수의 포트들 사이에서 데이터 트래픽을 스위치하는 스위치 컨트롤러를 포함하되,

상기 마스터 스위치는 우선순위-기반 플로우 제어(PFC)를 주어진 가상 포트의 데이터 트래픽상에 실행하는데 있어서, 상기 주어진 가상 포트에 대응하는 특정 RPI가 상주하는 하위 층 엔티티에, 상기 특정 RPI에 의해서 전달되는 데이터 트래픽의 적어도 두 개의 다른 클래스에 대한 우선순위를 명시하는 PFC 데이터 프레임을, 전송함으로써 실행하는,

마스터 스위치.

청구항 12

제11항에 있어서,

상기 마스터 스위치와 상기 복수의 하위 층 엔티티들 사이에서 전달되는 상기 데이터 트래픽은 그 트래픽이 전달되는 상기 하위 층 엔티티들 상의 RPI들을 식별하는 RPI 식별자들을 포함하고; 그리고

상기 마스터 스위치는 데이터 트래픽을 상기 RPI 식별자에 기초하여 자신의 가상 포트들에 대기시키는,

마스터 스위치.

청구항 13

제11항에 있어서, 상기 마스터 스위치는 PFC를 우선순위별로(per-priority), 애플리케이션별로(per-application) 실행하는,

마스터 스위치.

청구항 14

제11항 또는 제12항에 있어서, 상기 PFC 프레임은:

상기 특정 RPI를 식별하는 RPI 필드; 및

상기 데이터 트래픽의 적어도 두 개의 다른 클래스에 대한 상대적 우선순위를 명시하는 복수의 필드들을 포함하는,

마스터 스위치.

청구항 15

제11항 내지 제14항 중 어느 한 항에 있어서, 상기 마스터 스위치는 상기 하위 층 엔티티 상의 두 개의 다른 RPI들에 대하여 상이한 우선순위-기반 플로우 제어(differing priority-based flow control)를 실행하는,

마스터 스위치.

청구항 16

스위칭 네트워크에서, 상기 스위칭 네트워크는:

제11항 내지 제15항의 어느 한 항의 마스터 스위치; 및

상기 복수의 하위 층 엔티티들을 포함하는

스위칭 네트워크.

청구항 17

제16항에 있어서,

상기 하위 층 엔티티는 플로우 스루 모드로(in flow through mode) 구성된 팔로어 스위치(a follower switch)를 포함하고; 그리고

상기 특정 RPI는 상기 팔로어 스위치상에 데이터 포트를 포함하며; 그리고

상기 팔로어 스위치는 상기 PFC 데이터 프레임을 수신하고, 상기 PFC 데이터 프레임으로부터 상기 특정 RPI를 식별하는 RPI 식별자를 제거하여 표준 PFC 데이터 프레임을 획득하고, 그리고 상기 표준 PFC 데이터 프레임을 상기 특정 RPI에 대응하는 상기 데이터 포트를 통해서 전송하는,

스위칭 네트워크.

청구항 18

제16항 또는 제17항에 있어서,

상기 하위 층 엔티티는 호스트 플랫폼을 포함하고; 그리고

상기 특정 RPI는 상기 호스트 플랫폼 상에서 실행하는 가상 머신을 포함하는,

스위칭 네트워크.

청구항 19

제18항에 있어서,

상기 호스트 플랫폼은 가상 머신 모니터를 실행하고;

상기 호스트 플랫폼은 상기 PFC 데이터 프레임을 수신하여 이 PFC 데이터 프레임을 상기 가상 머신 모니터로 전달하는 네트워크 인터페이스를 포함하고; 그리고

상기 가상 머신 모니터는 상기 PFC 데이터 프레임에서 명시한 상기 특정 RPI에 기초하여 상기 가상 머신에 상기 PFC 데이터 프레임을 전달하는,

스위칭 네트워크.

청구항 20

제16항 내지 제19항 중 어느 한 항에 있어서, 상기 특정 RPI는 물리적 포트, 링크 집합 그룹(LAG) 인터페이스 및 가상 포트를 포함하는 세트 중 하나를 포함하는,

스위칭 네트워크.

청구항 21

제20항에 있어서, 상기 가상 포트는 가상 네트워크 인터페이스 카드(NIC), 단일 루트(Single Root) I/O 가상화(SR-IOV) NIC 파티션, 및 파이버채널 오버 이더넷(FCoE) 포트를 포함하는 세트 중 하나를 포함하는,

스위칭 네트워크.

청구항 22

프로그램 제품에 있어서, 상기 프로그램 제품은:

머신-판독가능 스토리지 디바이스; 및

상기 머신-판독가능 스토리지 매체 내에 저장된 프로그램 코드를 포함하되, 상기 프로그램 코드는 머신에 의해 처리될 때 상기 머신이:

마스터 스위치를 갖는 상위 층(an upper tier)과 복수의 하위 층 엔티티들을 갖는 하위 층(a lower tier)을 포함하는 스위칭 네트워크에서, 상기 마스터 스위치는 복수의 포트들을 갖고 각각은 상기 복수의 하위 층 엔티티들 각각에 결합 가능하며, 상기 마스터 스위치가 복수의 가상 포트들을 상기 복수의 포트들 각각 상에 구현하는 단계(implementing) – 상기 복수의 가상 포트들 각각은 그 포트에 결합 가능한 하위 층 엔티티에서 복수의 원격 물리적 인터페이스들(RPI들) 각각에 대응함 –;

상기 마스터 스위치가 상기 마스터 스위치와 상기 복수의 하위 층 엔티티들상의 RPI들 사이에서 전달되는 데이터 트래픽을 상기 복수의 가상 포트들 중에서 그 데이터 트래픽이 전달되는 하위 층 엔티티들상의 RPI들에 대응하는 가상 포트들 내에 대기시키는 단계(queuing); 및

상기 마스터 스위치가 우선순위-기반 플로우 제어(PFC)를 주어진 가상 포트의 데이터 트래픽 상에 실행하는 단

계에 있어서, 상기 주어진 가상 포트에 대응하는 특정 RPI가 상주하는 하위 층 엔티티에, 상기 특정 RPI에 의해서 전달되는 데이터 트래픽의 적어도 두 개의 다른 클래스들에 대한 우선순위를 명시하는 PFC 데이터 프레임을, 전송함으로써 실행하는 단계(enforcing)를 수행하게 하는,

프로그램 제품.

청구항 23

제22항에 있어서,

상기 마스터 스위치와 상기 복수의 하위 층 엔티티들 사이에서 저달되는 상기 데이터 트래픽은 그 트래픽이 전달되는 상기 하위 층 엔티티들상의 RPI들을 식별하는 RPI 식별자들을 포함하고; 그리고

상기 대기시키는 단계는 상기 데이터 트래픽을 상기 RPI 식별자들에 기초하여 상기 마스터 스위치상의 가상 포트들에 대기시키는 단계를 포함하는,

프로그램 제품.

청구항 24

제22항 또는 제23항에 있어서,

상기 하위 층 엔티티는 플로우 스루 모드로 구성된 팔로어 스위치를 포함하고; 그리고

상기 특정 RPI는 상기 팔로어 스위치상에 데이터 포트를 포함하고; 그리고

상기 프로그램 코드는 상기 팔로어 스위치가 더(further) 상기 PFC 데이터 프레임을 수신하고, 상기 PFC 데이터 프레임으로부터 상기 특정 RPI를 식별하는 RPI 식별자를 제거하여 표준 PFC 데이터 프레임을 획득하고, 그리고 상기 표준 PFC 데이터 프레임을 상기 특정 RPI에 대응하는 데이터 포트를 통해서 전송하도록 하게하는,

프로그램 제품.

청구항 25

제22항 내지 제24항 중 어느 한 항에 있어서,

상기 하위 층 엔티티는 네트워크 인터페이스와 가상 머신 모니터를 갖는 호스트 플랫폼을 포함하고;

상기 특정 RPI는 가상 머신을 포함하며;

상기 프로그램 코드는 상기 호스트 플랫폼이 상기 PFC 데이터 프레임을 상기 가상 머신 모니터를 통하여 상기 특정 RPI에 의해 명시되는 상기 가상 머신에 전달하게 하는,

프로그램 제품.

청구항 26

제22항 내지 제25항 중 어느 한 항에 있어서, 상기 프로그램 코드는 상기 마스터 스위치가 PFC를 우선순위별로(per-priority), 애플리케이션별로(per-application) 실행하게 하는,

프로그램 제품.

청구항 27

제22항 내지 제26항 중 어느 한 항에 있어서, 상기 PFC 프레임은:

상기 특정 RPI를 식별하는 RPI 필드; 및

상기 적어도 두 개의 다른 데이터 트래픽 클래스에 대한 상대적 우선순위를 명시하는 복수의 필드들을 포함하는,

프로그램 제품.

청구항 28

제22항 내지 제27항 중 어느 한 항에 있어서, 상기 실행하는 단계는 상기 마스터 스위치가 상기 하위 층 엔티티상의 두 개의 다른 RPI들에 대하여 상이한 우선순위-기반 플로우 제어를 실행하는 단계를 포함하는,

프로그램 제품.

청구항 29

상위 층과 복수의 하위 층 엔티티들을 포함하는 하위 층을 포함하는 스위칭 네트워크에서 플로우 제어를 구현하기 위한 장치(apparatus)에 있어서, 상기 장치는:

각각이 상기 복수의 하위 층 엔티티들 각각에 결합 가능한 복수의 포트들을 갖는 상위 층 내의 마스터 스위치에서, 상기 복수의 포트들 각각 상에 복수의 가상 포트들을 구현하기 위한 수단 - 상기 복수의 가상 포트들 각각은 그 포트에 결합 가능한 하위 층 엔티티에서 복수의 원격 물리적 인터페이스들(RPI들) 각각에 대응함 -;

상기 마스터 스위치와 상기 복수의 하위 층 엔티티들 상의 RPI들 사이에서 전달되는 데이터 트래픽을 상기 복수의 가상 포트들 중에서 그 데이터 트래픽이 전달되는 하위 층 엔티티들상의 RPI들에 대응하는 가상 포트들 내에 대기시키기 위한 수단; 및

상기 마스터 스위치에서, 주어진 가상 포트의 데이터 트래픽상에 우선순위-기반 플로우 제어(PFC)를 실행시키기 위한 수단 - 상기 실행시키기 위한 수단은, 상기 주어진 가상 포트에 대응하는 특정 RPI가 상주하는 하위 층 엔티티에, 상기 특정 RPI에 의해서 전달되는 데이터 트래픽의 적어도 두 개의 다른 클래스에 대한 우선순위를 명시하는 PFC 데이터 프레임을, 전송하기 위한 수단을 더 포함함 - 을 포함하는

장치.

명 세 서

기 술 분 야

[0001] 본 발명은 일반적으로 네트워크 통신에 관한 것이고, 더 자세하게는 컴퓨터 네트워크를 위한 개선된 스위칭 네트워크 아키텍처(switching network architecture)에 관한 것이다.

배 경 기 술

[0002] 이 기술분야에서 알려진 바와 같이, 네트워크 통신은 통상적으로 잘 알려진 7개의 계층 개방형 시스템 간 상호접속(OSI: Open Systems Interconnection) 모델에 전제를 두는데, 이 모델은 여러 프로토콜 계층들의 기능들을 정의하지만, 계층 프로토콜들 자체는 명시하지 않는다. 때때로 여기에서 계층 7 ~ 계층 1로 불리는, 상기 7개의 계층들은, 각각 애플리케이션 계층(application layer), 프리젠테이션 계층(presentation layer), 세션 계층(session layer), 전송 계층(transport layer), 네트워크 계층(network layer), 데이터 링크 계층(data link layer), 및 물리적 계층(physical layer)이다.

[0003] 소스 스테이션(source station)에서, 데이터 통신은 데이터가 소스 프로세스로부터 평션들 스택의 최상위 (애플리케이션) 계층에 수신될 때 시작된다. 데이터는 상기 스택의 연속적으로 더 낮은 각각의 계층에서(at each successively lower layer of the stack) 순차적으로 포맷되어 비트들의 데이터 프레임(a data frame of bits)이 데이터 링크 계층에서 획득된다. 최종적으로, 물리적 계층에서, 데이터는 네트워크 링크를 통해서 전자기 신호들의 형태로 목적지 스테이션 쪽으로 전송된다. 목적지 스테이션에서 수신될 때, 전송된 데이터는 데이터가 소스 스테이션에서 처리된 역순으로 평션들의 대응 스택에 보내지고(passed up), 그렇게 하여 그 정보가 목적지 스테이션의 수신 프로세스(receiving process)에 공급된다.

[0004] OSI 모델에 의해 지원되는 것들과 같은, 계층형 프로토콜들(layered protocols)의 원리는, 데이터가 모델 계층들을 수직으로 가로지르는 동안 소스 및 목적지 스테이션들(the source and destination stations)에서 계층들은 피어-투-피어(peer-to-peer)(즉, 계층 N 대 계층 N) 방식으로 상호작용하고, 각 개별 계층의 평션들은 개별 계층의 평션과 그것의 바로 위 아래의 프로토콜 계층들 사이의 인터페이스에 영향을 주지 않고 수행된다는 것이다. 이러한 효과를 달성하기 위해, 소스 스테이션 내 프로토콜 스택의 각 계층은 전송 프로세스(sending process)에 의해 생성된 데이터가 스택을 내려감에 따라 그 데이터에 통상적으로 정보를(캡슐화된 헤더의 형태로) 추가한다. 목적지 스테이션에서, 이들 캡슐화된 헤더들은 상기 프레임이 스택의 계층들로 상향 전달됨(propagates up)에 따라 하나씩 벗겨져서(stripped off) 캡슐화가 해제된 데이터가 수신 프로세스에 전달된다.

[0005] 소스 및 목적지 스테이션들을 연결하고 있는 물리적 네트워크는 하나 또는 그 이상의 무선 또는 유선 네트워크 링크들에 의해 상호 연결되는 많은 네트워크 노드들을 포함할 수 있다. 네트워크 노드들에는 보통 네트워크 트래픽을 생산하고 소비하는 호스트들(예를 들어, 서버 컴퓨터들, 클라이언트 컴퓨터들, 모바일 디바이스들 등), 스위치들 및 라우터들이 포함된다. 종래의 네트워크 스위치들은 여러 네트워크 세그먼트들을 상호 연결하고 OSI 모델의 데이터 링크 계층(계층 2)에서 데이터를 처리 및 포워드한다. 스위치들은 통상적으로 적어도 기본 브릿지 평션들을 제공하며, 이 평션들은 계층 2 MAC(Media Access Control) 주소에 의해서 데이터 트래픽을 필터링하는 것(filtering), 프레임들의 소스 MAC 주소들을 습득하는 것(learning), 및 목적지 MAC 주소들에 기초하여 프레임들을 포워드하는 것(forwarding)을 포함한다. OSI 모델의 네트워크(계층 3)에서 여러 네트워크들을 상호 연결하는, 라우터들은 통상적으로 라우트 처리(route processing), 경로 결정(path determination) 및 경로 스위칭(path switching)과 같은 네트워크 서비스들을 구현한다.

[0006] 큰 네트워크는 통상적으로 많은 수의 스위치들을 포함하고, 이 스위치들은 관리 평면(management plane), 제어 평면(control plane) 및 데이터 평면(data plane)에서 독립적으로 동작한다. 따라서, 각 스위치는 개별적으로 구성되어야 하고(configured), 데이터 트래픽에 대한 독립적인 제어(예를 들어, 접근 제어 목록들 (ACLs: access control lists))을 구현해야 하며, 그리고 다른 스위치들에 의해 처리되는 데이터 트래픽과는 상관 없이 데이터 트래픽을 포워드해야 한다.

발명의 내용

과제의 해결 수단

[0007] 적어도 하나의 실시 예에 따라서, 컴퓨터 네트워크에서 복수의 스위치들의 관리(management), 제어(control), 및 데이터 처리(data handling)가 개선된다.

[0008] 적어도 하나의 실시 예에서, 스위칭 네트워크(switching network)는 마스터 스위치(master switch)를 포함하는 상위 층(upper tier)과 복수의 하위 층 엔티티들(lower tier entities)을 포함하는 하위 층(lower tier)을 포함한다. 상기 마스터 스위치는 복수의 포트들을 포함하고 이 포트들 각각은 상기 복수의 하위 층 엔티티들 각각에 연결된다. 상기 복수의 포트들 각각은 복수의 가상 포트들을 포함하고 이 가상 포트들 각각은 그 포트에 결합된 하위 층 엔티티에서 복수의 원격 물리적 인터페이스들(RPIs: remote physical interfaces) 각각에 대응한다. 상기 복수의 포트들 각각은 또한 수신 인터페이스(a receive interface)를 포함하고, 이 수신 인터페이스는, 상기 복수의 하위 층 엔티티들 중에서 특정(particular) 하위 층 엔티티로부터 데이터 트래픽을 수신하는 것에 응답하여, 상기 데이터 트래픽을 상기 복수의 가상 포트들 중에서, 상기 데이터 트래픽의 소스였던 상기 특정 하위 층 엔티티상의 RPI에 대응하는 가상 포트에 대기시킨다(queues). 상기 마스터 스위치는 데이터 트래픽을 상기 가상 포트로부터, 상기 복수의 포트들 중에서 이그레스 포트(egress port)로 스위치하는 스위치 컨트롤러(switch controller)를 더 포함하는데, 상기 이그레스 포트로부터 상기 데이터 트래픽이 포워드된다.

[0009] 적어도 하나의 실시 예에서, 스위칭 네트워크는 상위 층과, 복수의 하위 층 엔티티들을 포함하는 하위 층을 포함한다. 상기 상위 층 내의 마스터 스위치는, 각각의 포트가 각각의 하위 층 엔티티에 결합된 복수의 포트들을 가지며, 상기 포트들 각각에 복수의 가상 포트들을 구현하고, 상기 복수의 가상 포트들 각각은 그 포트에 결합된 하위 층 엔티티에서 복수의 원격 물리적 인터페이스들(RPI들) 각각에 대응한다. 상기 마스터 스위치와 RPI들 사이에서 전달되는 데이터 트래픽은 그 데이터 트래픽이 전달되는 하위 층 엔티티들상의 상기 RPI들에 대응하는 가상 포트들 내에서 대기한다(queued). 상기 마스터 스위치는 주어진 가상 포트의 데이터 트래픽 상에서, 우선순위-기반 플로우 제어(PFC)를 실행하는데, 대응 RPI가 상주하는 하위 층 엔티티에, 특정 RPI에 의해서 전달되는 데이터 트래픽의 적어도 두 개의 다른 클래스에 대한 우선순위를 명시하는 PFC 데이터 프레임을 전송함으로써, 실행한다.

[0010] 적어도 하나의 실시 예에서, 스위칭 네트워크는 마스터 스위치를 갖는 상위 층과 복수의 하위 층 엔티티들을 포함하는 하위 층을 포함한다. 상기 마스터 스위치는, 각각의 포트가 각각의 하위 층 엔티티에 결합된 복수의 포트들을 갖고, 상기 포트들 각각에 복수의 가상 포트들을 구현하는데, 상기 복수의 가상 포트들 각각은 그 포트에 결합된 상기 하위 층 엔티티에서 복수의 원격 물리적 인터페이스들(RPIs) 각각에 대응한다. 상기 마스터 스위치와 RPI들 사이에서 전달되는 데이터 트래픽은 그 데이터 트래픽이 전달되는 RPI들에 대응하는 가상 포트들 내에서 대기한다(queued). 상기 마스터 스위치는 적어도, 상기 데이터 트래픽이 대기하는, 상기 가상 포트상에 기초한 제어 폴리시(control policy)에 따라 상기 데이터 트래픽에 대한 데이터 처리(data handling)를 적용하

여서 상기 마스터 스위치가 상기 마스터 스위치의 동일한 포트상의 두 개의 가상 포트들에 대기하는 데이터 트래픽에 다른 폴리시들을 적용하도록 한다.

도면의 간단한 설명

[0011] 다음과 같은 내용으로 첨부되는 도면들을 참조하여 이어지는 예시적인 실시 예의 구체적인 내용을 읽는다면 본 발명뿐 아니라 본 발명의 바람직한 이용 방식 및 장점들까지도 가장 잘 이해할 수 있을 것이다.

도 1은 일 실시 예에 따른 데이터 처리 환경의 높은 수준 블록 다이어그램이다.

도 2는 도 1의 데이터 처리 환경 내에서 구현될 수 있는 분산형 패브릭 프로토콜(DFP) 스위칭 네트워크 아키텍처의 일 실시 예의 고 수준 블록 다이어그램이다.

도 3은 도 1의 데이터 처리 환경 내에서 구현될 수 있는 DFP 스위칭 네트워크 아키텍처의 또 다른 실시 예의 고 수준 블록 다이어그램이다.

도 4는 일 실시 예에 따른 도 3에 있는 호스트의 더 상세한 블록 다이어그램이다.

도 5a는 일 실시 예에 따른 DFP 스위칭 네트워크의 마스터 스위치의 예시적 실시 예의 고 수준 블록 다이어그램이다.

도 5b는 일 실시 예에 따른 DFP 스위칭 네트워크의 팔로어 스위치(follower switch)의 예시적 실시 예의 고 수준 블록 다이어그램이다.

도 6은 일 실시 예에 따른 관리 인터페이스를 경유하는 가상화 스위치(virtualized switch)로서 제시되는, 도2 또는 도 3의 DFP 스위칭 네트워크 아키텍처의 도면이다.

도 7은 일 실시 예에 따른 DFP 스위칭 네트워크를 관리하기 위한 예시적 프로세스의 고 수준 논리 순서도이다.

도 8은 일 실시 예에 따라서, 네트워크 트래픽이 가상화 스위치로서 동작하도록 구성된 DFP 스위칭 네트워크의 하위 층으로부터 상위 층으로 포워드되는 예시적 프로세스의 고 수준 논리 순서도를 도시한다.

도 9는 일 실시 예에 따라서, 상기 상위 층에 있는 마스터 스위치가 DFP 스위칭 네트워크의 상기 하위 층으로부터 수신한 데이터 프레임을 처리하는 예시적 프로세스의 고 수준 논리 순서도이다.

도 10은 일 실시 예에 따라서, 상기 하위 층에 있는 팔로어 스위치 또는 호스트가 DFP 스위칭 네트워크의 상기 상위 층에 있는 마스터 스위치로부터 수신한 데이터 프레임을 처리하는 예시적 프로세스의 고 수준 논리 순서도이다.

도 11은 일 실시 예에 따른 DFP 스위칭 네트워크에서 링크 집합 그룹(LAG: a link aggregation group)을 운영하는 예시적 방법의 고 수준 논리 순서도이다.

도 12는 일 실시 예에 따른, LAG의 멤버십(membership)을 기록하는 데 이용되는 LAG 데이터 구조의 예시적 실시 예를 도시한다.

도 13은 일 실시 예에 따른 DFP 스위칭 네트워크에서 멀티캐스팅(multicasting)하는 예시적 방법의 고 수준 논리 순서도이다.

도 14는 계층 2 멀티캐스트 색인 데이터 구조 및 계층 3 멀티캐스트 색인 데이터 구조의 예시적 실시 예들을 도시한다.

도 15는 일 실시 예에 따른 DFP 스위칭 네트워크에서 향상된 전송 선택(ETS)의 예시적 방법의 고 수준 논리 순서도이다.

도 16은 일 실시 예에 따른 DFP 스위칭 네트워크의 마스터 스위치를 위한 향상된 전송 선택(ETS)을 구성하는 데 이용될 수 있는 예시적 ETS 데이터 구조를 도시한다.

도 17은 DFP 스위칭 네트워크가 하위 층에서 우선순위-기반 플로우 제어(PFC) 및/또는 기타 서비스들을 구현하는 예시적 방법의 고 수준 논리 순서도이다.

도 18은 일 실시 예에 따른 DFP 스위칭 네트워크의 하위 층에서 우선순위-기반 플로우 제어(PFC) 및/또는 기타 서비스들을 구현하는 데 이용될 수 있는 예시적 PFC 데이터 프레임(1800)을 도시한다.

도 19a는 일 실시 예에 따른 DFP 스위칭 네트워크의 하위 수준 팔로어 스위치가 마스터 스위치로부터 수신된 PFC 데이터 프레임을 처리하는 예시적 프로세스의 고 수준 논리 순서도이다.

도 19b는 일 실시 예에 따른 DFP 스위칭 네트워크에서 하위 수준 호스트가 마스터 스위치로부터 수신된 PFC 데이터 프레임을 처리하는 예시적 프로세스의 고 수준 논리 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0012] 본 출원의 공개 내용은 컴퓨터 네트워크에서 복수의 상호 연결된 스위치들에 통합된 관리 평면, 제어 평면 및 데이터 평면을 도입하는 스위칭 네트워크 아키텍처이다.

[0013] 이제 도면들을 참조하며 구체적으로 도 1을 참조하면, 일 실시 예에 따른 예시적 데이터 처리 환경(100)의 고 수준 블록 다이어그램이 도시된다. 도시된 바와 같이, 데이터 처리 환경(100)은 일단의 리소스들(102)를 포함한다. 여러 호스트들, 클라이언트들, 스위치들, 라우터들, 스토리지 등이 포함될 수 있는 리소스들(102)는 통신을 위해 상호 연결되며 하나 또는 그 이상의 공공, 사설, 커뮤니티, 또는 클라우드 네트워크들 또는 이들의 조합으로 물리적 또는 가상적으로 그룹화될(도시되지는 않음) 수 있다. 이와 같은 방식으로, 데이터 처리 환경(100)은 퍼스널(예를 들어, 데스크탑, 랩탑, 넷북, 태블릿 또는 핸드헬드) 컴퓨터(110a), 스마트폰(110b), 서버 컴퓨터 시스템(110c) 및 미디어 플레이어(예를 들어, 셋톱박스, 디지털 다목적 디스크(DVD) 플레이어, 또는 디지털 비디오 녹화기(DVR))(110d) 등의 가전제품 등의 여러 클라이언트 디바이스들이 액세스 가능한 인라스트럭처, 플랫폼, 소프트웨어 및/또는 서비스를 제공할 수 있다. 도 1에 도시된 클라이언트 디바이스들(110)의 타입들은 단지 예시적인 것이며 클라이언트 디바이스들(110)은 패킷 네트워크를 통해서 리소스들(102)과 통신 및 리소스들(102)에 액세스할 수 있는 모든 타입의 전자 디바이스일 수 있다는 것을 이해해야 한다.

[0014] 이제 도 2를 참조하면, 일 실시 예에 따른 리소스들(102) 내에서 구현될 수 있는 예시적 분산형 패브릭 프로토콜(DFP) 스위칭 네트워크 아키텍처의 고 수준 블록 다이어그램이 도시된다. 상기 도시된 예시적 실시 예에서, 리소스들(102)는 복수의 물리적 및/또는 가상 네트워크 스위치들을 포함하며 이 스위치들은 DFP 스위칭 네트워크(200)을 형성한다. 각각의 스위치가 독립적인 관리 평면, 제어 평면 및 데이터 평면(independent management, control and data planes)을 구현하는 종래의 네트워크 환경들과는 대조적으로, DFP 스위칭 네트워크(200)은 통합된 관리 평면, 제어 평면 및 데이터 평면(unified management, control and data planes)을 구현하여, 모든 구성 스위치들이 통합된 가상화 스위치로 보이도록 해주며, 그렇게 함으로써 상기 네트워크 패브릭의 배치(deployment), 구성(configuration), 및 관리(management)를 간소화한다.

[0015] DFP 스위칭 네트워크(200)은 둘 또는 그 이상의 스위치들의 층들(tiers of switches)을 포함하며, 상기 실시 예에서 이들은 팔로어 스위치들(follower switches)(202a-202d)를 포함하는 복수의 팔로어 스위치들을 갖는 하위 층과, 마스터 스위치들(master switches)(204a-204b)를 포함하는 복수의 마스터 스위치들을 갖는 상위 층을 포함한다. 도시된 바와 같은 두 개의 층이 있는 실시 예에서, 각각의 마스터 스위치(204)의 포트는 층-간(inter-tier) 링크들(206) 중 하나에 의해 각각의 팔로어 스위치(202)의 포트들 중 하나에 직접적으로 연결되고, 각각의 마스터 스위치(204)의 포트는 마스터 링크(208)에 의해 적어도 하나의 다른 마스터 스위치(204)의 포트에 직접 또는 간접적으로 결합된다. 이러한 차이들을 구별하기 위해, 층-간 링크들(206)을 통하여 스위치-대-스위치(switch-to-switch) 통신을 지원하는 포트들은 본 출원에서 "스위치-간 포트들(inter-switch ports)"로 불리고, 다른 포트들(예를 들어, 팔로어 스위치(202a-202d)의 포트들)은 "데이터 포트들(data ports)"로 불린다.

[0016] 바람직한 일 실시 예에서, 팔로어 스위치들(202)는 패스-스루 모드(pass-through mode)에서 데이터 평면상에서 동작하도록 구성되며, 이것은 (예를 들어, 호스트들로부터) 팔로어 스위치들(202)의 데이터 포트들(210)에 수신되는 모든 인그레스(ingress) 데이터 트래픽이 스위치-간 포트들과 층-간 링크들(206)을 통해서 마스터 스위치들(204) 중 하나로 포워드된다는 것을 의미한다. 결국 마스터 스위치들(204)는 데이터 트래픽을 위한 패브릭으로서 기능하고(이런 이유로 분산형 패브릭의 개념이 나옴) 데이터 트래픽을 위한 모든 패킷 스위칭 및 라우팅을 구현한다. 이러한 구성 때문에 데이터 트래픽은, 예를 들어, 화살표들(212a-212d)로 표시되는 제1의 예시적 플로우와 화살표들(214a-214e)로 표시되는 제2의 예시적 플로우로 포워드될 수 있다.

[0017] 따라서, 마스터 스위치들(204)에서 팔로어 스위치들(202)를 위한 스위칭 및 라우팅이 집중화하는 것은 마스터 스위치들(204)가 데이터 트래픽이 수신된 팔로어 스위치들(202)의 인그레스 데이터 포트들에 관한 지식을 갖고 있음을 의미한다는 것을 알 수 있을 것이다. 바람직한 일 실시 예에서, 링크들(206, 208)을 통한 스위치-대-스위치 통신은 계층 2 프로토콜을 채용하는데, 그 예에는 Cisco Corporation이 개발한 ISL(Inter-Switch Link)

프로토콜 또는 IEEE 802.1QnQ가 있으며, 이 계층 2 프로토콜은 명시적인 태그달기(explicit tagging)를 이용하여 다수의 계층 2 가상 근거리 통신망(VLAN)들을 DFP 스위칭 네트워크(200)에 걸쳐서 구축한다. 각각의 팔로어 스위치(202)는 데이터 프레임이 수신된 팔로어 스위치(202)상의 인그레스 데이터 포트(210)을 수신 마스터 스위치(204)에 통신하기 위해 데이터 프레임들에 VLAN 태그들(서비스 태그들(S-태그들)로 알려짐)을 적용하는 것이 바람직하다. 다른 실시 예들에서, 인그레스 데이터 포트는 예를 들면, MAC-in-MAC 헤더, 고유 MAC 주소, IP-in-IP 헤더 등의 또 다른 식별자(another identifier)에 의해 통신할 수 있다. 아래에서 더 논의되는 바와 같이, 각 팔로어 스위치(202)상의 각 데이터 포트(210)은 각 마스터 스위치(204)상에 대응 가상 포트(다른 말로 하면, v포트)를 가지며, 팔로어 스위치(202)의 데이터 포트(210)상에 진입하는(ingressing) 데이터 프레임들은 마치 수신 마스터 스위치(204)의 대응 v포트상에 진입하는 것처럼 처리된다.

[0018] 이제 도 3을 참조하면, 일 실시 예에 따른 리소스들(102) 내에서 구현될 수 있는 또 다른 예시적 분산형 패브릭 프로토콜(DFP) 스위칭 네트워크 아키텍처의 고 수준 블록 다이어그램이 도시된다. 도 3에서 도시하는 DFP 아키텍처는, DFP 스위칭 네트워크(300)에 걸쳐서 통합된 관리 평면, 제어 평면 및 데이터 평면을 구현하고, 도 2에 도시된 DFP 스위칭 네트워크 아키텍처의 대안으로서 또는 그에 추가하여 리소스들(102) 내에서 구현될 수 있다.

[0019] 상기 예시적인 실시 예에서, DFP 스위칭 네트워크(300) 내의 리소스들(102)는 상위 층 내 마스터 스위치들(204a-204b) 중 적어도 하나를 구현하는 하나 또는 그 이상의 물리적 및/또는 가상 네트워크 스위치들을 포함한다. 스위칭 네트워크(300)은 또한 하위 층에 복수의 물리적 호스트들(302a-302d)를 포함한다. 도 4에 도시된 바와 같이, 예시적 실시 예에서, 각 호스트(302)는 하나 또는 그 이상의 네트워크 인터페이스들(404)(예를 들어, 네트워크 인터페이스 카드(NIC: network interface card), 융합 네트워크 어댑터(CNA: converged network adapter) 등)을 포함하고, 이 인터페이스들(404)는 호스트(302)가 마스터 스위치(들)(204)와 통신할 때 인터페이스를 제공한다. 호스트(302)는 또한 (통상적으로 하나 또는 그 이상의 직접 회로들을 포함하는) 하나 또는 그 이상의 프로세서들(402)를 포함하며 이 프로세서들(402)는 예를 들어 데이터 처리 환경(100)에서 데이터 또는 소프트웨어를 관리, 액세스 및 조작하기 위해 데이터 및 프로그램 코드를 처리한다. 호스트(302)는 또한 포트들, 디스플레이들, 유저 입력 디바이스들 및 부속 디바이스들 등과 같은 입력/출력(I/O) 디바이스들(406)을 포함하며 이 I/O 디바이스들은 데이터 처리 환경(100)에서 호스트 (302) 및/또는 다른 리소스(들)에 의해 수행되는 처리의 입력들을 수신하고 출력들을 제공한다. 끝으로, 호스트(302)는 데이터 스토리지(410)을 포함하며 이 데이터 스토리지(410)에는 메모리, 고체 상태 드라이브, 광 또는 자기 디스크 드라이브, 테이프 드라이브 등을 포함한 하나 또는 그 이상의 휘발성 또는 비휘발성 스토리지 디바이스들이 포함된다. 데이터 스토리지(410)은 예를 들어 (소프트웨어, 펌웨어 또는 이들의 조합 등을 포함한) 프로그램 코드 및 데이터를 저장할 수 있다.

[0020] 도 3으로 돌아가서, 각 호스트(302)에 의해 실행되는 프로그램 코드는 ("하이퍼바이저"라고도 불리는) 가상 머신 모니터(VMM)(304)를 포함하며 이 모니터는 자신의 각각의 물리적 호스트(302)의 리소스들을 가상화하고 관리한다. 각 VMM(304)는 하나 또는 그 이상의 아마도 이종 운영 체제(heterogeneous operating system) 파티션들에서 하나 또는 그 이상의 가상 머신(VM)들(306)에 리소스들을 할당하고 이 가상 머신들의 실행을 지원한다. VM들(304)의 각각은 하나의 (어떤 경우들에는 다수의) 가상 네트워크 인터페이스(virtual NIC: VNIC)를 가지며 이 인터페이스는 적어도 OSI 모델의 계층 2와 계층 3에서 네트워크 연결을 제공한다.

[0021] 도시된 바와 같이, 하나 또는 그 이상의 VMM들(304a-304d)는 선택에 따라(optionally) VM들(306)이 접속될 수 있는 하나 또는 그 이상의 가상 스위치(VS)들(310)(예를 들어, 파이버 채널 스위치(들), 이더넷 스위치(들), 파이버 채널 오버 이더넷(FCoE) 스위치들 등)을 제공할 수 있다. 이와 비슷하게, 호스트들(302)의 하나 또는 그 이상의 네트워크 인터페이스들(404)는 선택에 따라 VM들(306)이 연결될 수 있는 하나 또는 그 이상의 가상 스위치(VS)들(312)(예를 들어, 파이버 채널 스위치(들), 이더넷 스위치(들), FCoE 스위치들 등)을 제공할 수 있다. 그러므로, VM들(306)은 층-간 링크들(206), 네트워크 인터페이스들(404), VMM들(304)에 의해 제공되는 가상화 계층, 및 선택에 따라 프로그램 코드 및/또는 하드웨어에서 구현되는 하나 또는 그 이상의 가상 스위치들(310, 312)을 통해서 마스터 스위치들(204)와 네트워크 통신을 한다.

[0022] 도 2에서와 같이, 가상 스위치들(310, 312)는, 존재한다면, 패스-스루 모드에서 데이터 평면상에서 동작하도록 구성되는 것이 바람직하며, 이것은 가상 스위치들(310, 312)의 가상 데이터 포트들에서 VM들(306)으로부터 수신된 모든 인그레스 데이터 트래픽이 네트워크 인터페이스들(404)와 층-간 링크들(206)을 통해 가상 스위치들(310, 312)에 의해서 마스터 스위치들(204) 중 하나로 포워드된다는 것을 의미한다. 그 다음에 마스터 스위치들(204)는 데이터 트래픽을 위한 패브릭으로서 기능하며 데이터 트래픽을 위한 모든 스위칭 및 라우팅을 구현한다.

[0023] 위에서 논의한 바와 같이, 마스터 스위치(들)(204)에서 호스트들(302)를 위한 스위칭 및 라우팅을 집중화하는 것은 호스트(302)로부터 데이터 트래픽을 수신하는 마스터 스위치(204)가 데이터 트래픽의 소스(예를 들어, 링크 집합 그룹(LAG) 인터페이스, 물리적 포트, 가상 포트 등)에 관한 지식을 갖고 있음을 의미한다. 다시 말하면, 이러한 트래픽 소스 정보의 통신을 허용하기 위해서는, 층-간 링크들(206)을 통한 통신은, Cisco Corporation이 개발한 ISL(Inter-Switch Link) 프로토콜 또는 IEEE 802.1QnQ 등의, 계층 2 프로토콜을 이용하는 것이 바람직하며 이 계층 2 프로토콜은 DFP 스위칭 네트워크(300)에 걸쳐서 다수의 계층 2 가상 근거리 통신망(VLAN)들을 설정하기 위해 명시적인 태그달기(tagging)를 포함한다. 각 호스트(302)는 데이터 프레임을 수신받았던 데이터 트래픽 소스(예를 들어, 물리적 포트, LAG 인터페이스, 가상 포트(예를 들어, VM 가상 네트워크 인터페이스 카드(VNIC), 단일 루트 I/O 가상화(SR-IOV) NIC 파티션 또는 FCoE 포트 등)를 수신 마스터 스위치(204)에 통신하기 위해 데이터 프레임들에 VLAN 태그들을 적용하는 것이 바람직하다. 이러한 데이터 트래픽 소스 각각은 대응 v포트를 각각의 마스터 스위치(204)상에 가지고 있어서, 호스트(302)상의 데이터 트래픽 소스에서 발신하는(originating) 데이터 프레임들은 마치 수신 마스터 스위치(204)의 대응 v포트상에 진입하는(ingressing) 것처럼 처리된다. 일반화를 위해서, 호스트들(302)상의 데이터 트래픽 소스들과 팔로어 스위치들(202)상의 데이터 포트들(210)은, RPI들의 여러 타입들 사이에 구별해야 하는 경우가 아니면, 이후로 원격 물리적 인터페이스(RPI)들로 불릴 것이다.

[0024] DFP 스위칭 네트워크들(200과 300)에서, 부하균형(load balancing)은 팔로어 스위치들(202) 및/또는 호스트들(302)의 구성(configuration)을 통해서 이루어질 수 있다. 예를 들면, 하나의 가능한 정적 구성(static configuration)의 일 실시 예에서, 데이터 트래픽은 소스 RPI에 기초하여 마스터 스위치들(204) 사이에서 분할될 수 있다. 이 예시적 실시 예에서, 만일 두 개의 마스터 스위치들(204)가 배치된다면, 각각의 팔로어 스위치(202) 또는 호스트(302)는 각각 총 RPI 수의 절반을 보유하는 두 개의 정적 RPI 그룹들을 구현하여서 RPI 그룹들 각각의 트래픽을 두 개의 마스터 스위치들(204) 중 다른 하나로 전송하도록 구성될 수 있다. 이와 비슷하게, 만일 네 개의 마스터 스위치들(204)가 배치된다면, 각각의 팔로어 스위치(202) 또는 호스트(302)는 각각 총 RPI 수의 사분의 일(one-fourth)을 보유하는 네 개의 정적 RPI 그룹들을 구현하여서 RPI 그룹들 각각의 트래픽을 네 개의 마스터 스위치들(204) 중 다른 하나로 전송하도록 구성될 수 있다.

[0025] 이제 도 5a를 참조하면, 스위치(500a)의 예시적 실시 예의 고 수준 블록 다이어그램이 도시되며, 스위치(500a)는 도 2-3의 마스터 스위치들(204) 중 어느 하나를 구현하는 데 이용될 수 있다.

[0026] 도시된 바와 같이, 스위치(500a)는 복수의 물리적 포트들(502a-502m)을 포함한다. 각 포트(502)는 복수의 수신(Rx) 인터페이스들(504a-504m) 각각을 포함하고 연관된 Rx 인터페이스(504)에 의해 수신되는 데이터 프레임들을 버퍼링하는(buffer) 복수의 인그레스 큐들(ingress queues)(506a-506m) 각각을 포함한다. 포트들(502a-502m) 각각은 복수의 이그레스 큐들(egress ques)(514a-514m) 각각을 더 포함하고, 연관된 이그레스 큐(514)로부터 데이터 프레임들을 전송하는 복수의 전송(Tx) 인터페이스들(520a-520m) 각각을 더 포함한다.

[0027] 한 실시 예에서, 각각의 포트(502)의 인그레스 큐들(506) 및 이그레스 큐들(514)의 각각은 DFP 스위칭 네트워크(200, 300)의 하위 층 내의 RPI(이로부터 인그레스 데이터 트래픽이 그 포트(502)상에 수신될 수 있음) 마다 다수의(예를 들면, 8개의) 큐 엔트리들을 제공하도록 구성된다. 하위 층 RPI를 위해 정의된 마스터 스위치(204) 내의 다수의 큐 엔트리들의 그룹은 본 출원서에서 가상 포트(v포트)로 정의되며, v포트 내 각 큐 엔트리는 VOQ에 대응된다. 예를 들면, 도 2에 도시한 바와 같은 DFP 스위칭 네트워크(200)에 있어서, 스위치(500a)의 포트(502a)는, 포트(502a)에 연결된 팔로어 스위치(202)의 k+1 데이터 포트들(210)의 각각에 대하여, 인그레스 v포트들(522a0-522ak) 각각 및 이그레스 v포트들(524a0-524ak) 각각을 구현하도록 구성된다. 만일 스위치(500a)가 도 3에 도시한 바와 같이 DFP 스위칭 네트워크(300)에서 구현된다면, 포트(502a)는 층-간 링크(206)에 의해 포트(502a)에 연결된 호스트(302) 내의 k+1 데이터 트래픽 소스들 각각에 대하여 각각의 v포트(522)를 구현하도록 구성된다. 이와 비슷하게, 도 2에 도시된 바와 같은 DFP 스위칭 네트워크(200)에 있어서, 스위치(500a)의 포트(502m)은, 포트(502m)에 연결된 팔로어 스위치(202)의 p+1 데이터 포트들(210)의 각각에 대하여, 인그레스 v포트들(522m0-522mp) 각각 및 이그레스 v포트들(524m0-524mp) 각각을 구현하도록 구성된다. 만일 스위치(500a)가 도3에 도시된 바와 같은 DFP 스위칭 네트워크(300)에서 구현된다면, 포트(502a)는 층-간 링크(206)에 의해 포트(502a)에 연결된 호스트(302) 내의 k 데이터 트래픽 소스들의 각각에 대하여 각각의 v포트(522)를 구현한다. 따라서, 포트들(502)의 각각에 구현되는 인그레스 v포트들의 수는 포트들(502)의 각각에 연결된 특정 하위 층 엔티티(예를 들어, 팔로어 스위치(202) 또는 호스트(302))상의 RPI 수에 따라 다를 수 있음을 이해해야 할 것이다. 그러므로, DFP 스위칭 네트워크(200 또는 300)의 하위 층에 있는 각 RPI는 각 마스터 스위치(204)의 물리적 포트(502)상의 인그레스 및 이그레스 v포트들(522, 524)의 세트로 매핑되고, 그 RPI로부터 데이터 프레임

들이 물리적 포트(502)상에 수신될 때, 포트(502)의 수신 인터페이스(504)는 데이터 트래픽 내 RPI 식별자에 기초하여 데이터 프레임들을 적절한 인그레스 v포트(522)로 보낼(direct)수 있다.

[0028] 마스터 스위치(204)는 자신의 물리적 포트들(502)에 걸쳐있는 v포트들(522, 524)를 필요에 따라서, 예를 들어 하위 층 엔티티들(202, 302)와의 연결 상태에 따라서 생성(create), 파괴(destroy), 디스에이블(disable) 또는 마이그레이션(migrate)시킬 수 있다. 예를 들면, 만일 팔로어 스위치(202)가 더 큰 수의 포트들을 갖는 대체 팔로어 스위치(a replacement follower switch) (202)에 의해 대체된다면, 마스터 스위치들(204)는 대체 팔로어 스위치(202)상의 추가 RPI들을 수용하기 위해 관련 물리적 포트(502)상에 추가 v포트들(522, 524)를 자동적으로 생성할 것이다. 이와 비슷하게, 만일 마스터 스위치(204)의 제1 물리적 포트에 연결된 호스트(302)상에서 실행중인 VM(306)이 상기 마스터 스위치(204)의 다른 제2 물리적 포트에 연결된 다른 호스트(302)로 마이그레션하면(즉, 마이그레이션이 그 스위치 영역 내에서 유지되면), 상기 마스터 스위치(204)는 상기 VM(306)에 대응하는 v포트들(522, 524)를 상기 마스터 스위치(204)의 제1 물리적 포트(502)로부터 상기 마스터 스위치(204)의 제2 물리적 포트(502)로 자동적으로 마이그레이션시킬 것이다. 만일 상기 VM(306)이 자신의 마이그레이션을 미리 정해진 플러시 구간(flush interval) 내에서 완료하면, 상기 VM(306)에 대한 데이터 트래픽은 스위치 컨트롤러(530a)에 의해 인지되어(remarked) 상기 제2 물리적 포트(502)상의 이그레스 v포트(524)로 포워드될 수 있다. 이러한 방식으로, VM(306)의 마이그레이션은 트래픽 인터럽션 또는 데이터 트래픽의 손실 없이 이루어질 수 있으며, 이것은 손실에 민감한(loss-sensitive) 프로토콜들에는 특히 장점이다.

[0029] 각 마스터 스위치(204)는 하위 층 엔티티에 대한 스위치-간 링크(206)의 손상(예를 들어, 링크 상태가 업(up)에서 다운(down)으로 변경됨, 스위치-간 링크(206)이 연결 해제됨, 또는 하위 층 엔티티가 고장남(fail))을 추가로 감지한다. 만일 스위치-간 링크(206)의 손상이 감지되면, 그 마스터 스위치(204)는 스위치-간 링크(206)의 복원이 감지될 때까지 연관된 v포트들(522, 524)를 자동적으로 디스에이블시킬 것이다. 만일 상기 스위치-간 링크(206)이 미리 정해진 플러시 구간 내에 복원되지 않으면, 마스터 스위치(204)는 큐 용량을 복구시키기 위해 통신이 손상된 하위 층 엔티티와 연관된 v포트들(522, 524)를 파괴할 것이다. 플러시 구간 동안에, 스위치 컨트롤러(530a)는 디스에이블된 이그레스 v포트(524)로 가기로 되어 있는 데이터 트래픽이 인그레스 쪽에 버퍼링되도록 허용한다. 만일 상기 스위치-간 링크(206)이 복원되어 디스에이블된 이그레스 v포트(524)가 재인에이블되면, 상기 버퍼링된 데이터 트래픽은 손실 없이 이그레스 v포트(524)로 포워드될 수 있다.

[0030] 스위치(500a)는 스위치 컨트롤러(530a)의 지휘 아래 인그레스 큐들(506a-506m) 중 어느 하나에서 이그레스 큐들(514a-514m) 중 어느 하나로 (그러므로 임의 인그레스 v포트(522)와 임의 이그레스 v포트(524) 사이에) 데이터 프레임들을 지능적으로(intelligently) 스위치하도록 동작 가능한 크로스바(510)을 추가로 포함한다. 따라서, 스위치 컨트롤러(530a)는 하나 또는 그 이상의 집중형 또는 분산형, 특별목적용 또는 범용 처리 엘리먼트들 또는 논리 디바이스들로 구현될 수 있으며, 이 엘리먼트들 또는 디바이스들은 제어를 전적으로 하드웨어에서, 또는 더 일반적으로 처리 엘리먼트에 의한 펌웨어 및/또는 소프트웨어의 실행을 통해 구현할 수 있음을, 이해할 수 있을 것이다.

[0031] 데이터 프레임들을 지능적으로 스위치하기 위해, 스위치 컨트롤러(530a)는 하나 또는 그 이상의 데이터 평면 데이터 구조들을 구축 및 유지하는데, 예를 들면 통상적으로 내용 주소화 메모리(CAM: content-addressable memory) 내의 포워딩 테이블로 구현되는 포워딩 정보 베이스(FIB: forwarding information base)(532a)를 구축 및 유지한다. 도시된 예에서, FIB(532a)는 복수의 엔트리들(534)를 포함하며, 이 엔트리들은 예를 들어 MAC 필드(536), 포트 식별자(PID) 필드(538) 및 가상 포트(v포트) 식별자(VPID) 필드(540)을 포함한다. 그러므로 각각의 엔트리(534)는 데이터 프레임의 목적지 MAC 주소를 그 데이터 프레임에 대한 특정 이그레스 포트(502)상의 특정 v포트(520)과 연관시킨다. 스위치 컨트롤러(530a)는 관찰된 데이터 프레임들로부터 포트들(502) 및 v포트들(520)과 데이터 프레임들에 의해 명시된 목적지 MAC 주소들 사이의 연관(association)을 습득함으로써 그리고 습득한 연관들을 FIB(532a)에 기록함으로써 자동화된 방식으로 FIB(532a)를 구축한다. 스위치 컨트롤러(530a)는 그 후 FIB(532a)에 기록된 연관들에 따라서 데이터 프레임들을 스위치하도록 크로스바(510)을 제어한다. 그러므로, 각 마스터 스위치(204)는 하위 층에서 RPI들에 대응하는 v포트마다(per vport) 자신의 계층 2 및 계층 3의 QoS, ACL 및 기타 관리 데이터 구조들을 관리 및 액세스한다.

[0032] 스위치 컨트롤러(530a)는 통합된 가상화 스위치를 위한 관리 및 제어 센터로 기능하는 관리 모듈(550)을 추가로 구현한다. 한 실시 예에서, 각 마스터 스위치(204)는 관리 모듈(550)을 포함하지만, 주어진 DFP 스위칭 네트워크(200 또는 300)의 유일한 마스터 스위치(204)(본 출원에서는 관리 마스터 스위치(204)로 불림)만의 관리 모듈(550)은 임의의 시간에 동작 가능하다. 관리 마스터 스위치(204)로 기능하던 마스터 스위치(204)에 고장이 발생한 경우에(예를 들어, 마스터 링크(208)을 통해서 관리 마스터 스위치(204)에 의한 하트비트 메시징(heartbeat

messaging)의 상실에 의해서 감지된 경우), 나머지 동작 가능한 마스터 스위치들(204) 중에서 미리 정해지거나 선정되는 또 다른 마스터 스위치(204)가 관리 마스터 스위치(204)의 역할을 자동적으로 맡아서 자신의 관리 모듈(550)을 이용하여 DFP 스위칭 네트워크(200 또는 300)의 집중형 관리 및 제어를 제공하는 것이 바람직하다.

[0033] 관리 모듈(550)은 관리 인터페이스(552)를 포함하는 것이 바람직한데, 예를 들면, 로그인 및 관리자 인증 값의 입력(entry of administrative credentials)에 응답하여, 네트워크에 연결된 관리자 콘솔(예를 들어, 클라이언트들(110a-110c) 중 하나)에 배치된 관리자가 액세스할 수 있는 XML 또는 HTML 인터페이스를 포함하는 것이 바람직하다. 관리 모듈(550)은 DFP 스위칭 네트워크(200 또는 300) 내의 모든 스위치들(예를 들어, 스위치들(204 및/또는 202))상에 상주하는 모든 포트들의 전역 뷰(global view)를 관리 인터페이스(552)를 통해서 제공하는 것이 바람직하다. 예를 들면, 도 6은 일 실시 예에 따라 관리 인터페이스(552)를 통해서 가상화 스위치(600)으로서 제공되는 도 2의 DFP 스위칭 네트워크(200)의 한 뷰(view)이다. 이 실시 예에서, 마스터 스위치(204)는 가상 회선 카드들(virtual line cards)로 기능하는 팔로어 스위치들(202)를 갖는 가상 스위칭 섀시(virtual switching chassis)로 간주될 수 있다. 이 예에서, 예를 들어 관리자 콘솔의 디스플레이에 그래픽적으로 및/또는 표형식으로 표현될 수 있는, 가상화 스위치(600)은 팔로어 스위치(202a)의 데이터 포트들과 스위치-간 포트들에 대응하는 가상화 포트들(Pa-Pf)(602a)와, 팔로어 스위치(202b)의 데이터 포트들과 스위치-간 포트들에 대응하는 P1-Pp(602b)와, 팔로어 스위치(202c)의 데이터 포트들과 스위치-간 포트들에 대응하는 Pq-Ps(602c), 및 팔로어 스위치(202d)의 데이터 포트들과 스위치-간 포트들에 대응하는 Pw-Pz(602d)를 제공한다. 추가로, 가상화 스위치(600)은 Pg-Pk(602e)로 마스터 스위치(204a)의 스위치-간 포트들을 표현하고, Pt-Pv(602f)로 마스터 스위치(204b)의 스위치-간 포트들을 표현한다. 또한, 가상화 스위치(600)은 마스터 스위치(204)상에 구현된 각각의 v포트(522, 524)를 각각 일 세트의 가상 출력 큐들(VOQ들)(604)로 표현한다. 예를 들면, 마스터 스위치들(204a, 204b)상에 구현된 v포트들(522, 524)의 각각은 VOQ 세트들(604a-604k) 각각으로 표현된다. 가상화 스위치(600)과 상호작용함으로써, 관리자는 통합된 인터페이스를 통해서 DFP 스위칭 네트워크(200)에서 하나 또는 그 이상의 (또는 모든) 팔로어 스위치들(202)와 마스터 스위치들(204)의 하나 또는 그 이상의 (또는 모든) 포트들 또는 v포트들에 대하여 원하는 제어를 (예를 들어, 그래픽, 텍스트, 숫자 및/또는 기타의 입력들을 통해서) 관리하고 설정할 수 있다. 가상화 포트들 Pa-Pf(602a), P1-Pp(602b), Pq-Ps(602c) 및 Pw-Pz(602d)에 더하여 가상화 스위치(600) 내에서 VOQ들(604a-604k)의 세트들을 구현하면 DFP 스위칭 네트워크(200 또는 300)의 둘 중 하나의 층 (또는 두 층 모두)에서 각 RPI의 데이터 트래픽(및 RPI의 데이터 트래픽의 각 트래픽 분류)에 대한 개별화된 제어를 구현할 수 있음에 유의해야 한다. 그러므로, 아래에서 더 논의되는 바와 같이, 관리자는 가상화 스위치(600)의 가상화 포트(Pa)와 상호작용을 통해서 팔로어 스위치(202a)의 특정 데이터 포트(210)의 구체적인 트래픽 분류에 대한 원하는 제어를 구현할 수 있다. 이와는 달리 또는 추가적으로, 관리자는 데이터 포트(210)에 대응하는 인그레스 v포트(522) 또는 이그레스 v포트(524)를 표현하는 VOQ 세트(604)상에서 그 트래픽 분류에 대응하는 특정 VOQ와 상호작용함으로써 그 데이터 포트(210)의 트래픽 분류에 대하여 원하는 제어를 설정할 수 있다.

[0034] 도 5a로 돌아가면, 스위치 컨트롤러(530a)는 DFP 스위칭 네트워크(200 또는 300)을 순회하는 데이터 프레임들에 대한 원하는 제어를 구현하는 데 이용될 수 있는 제어 모듈(560a)를 더 포함한다. 제어 모듈(560a)는 v포트당 기준으로(on a per-vport basis) 인그레스 및/또는 이그레스에서 스위치(500a)에 대한 원하는 세트의 제어 폴리시들을 구현하는 로컬 폴리시 모듈(562)를 포함한다. 제어 모듈(560a)는 v포트당 기준으로 스위치(500a)에 대한 인그레스 액세스를 제한하는 로컬 액세스 제어 목록(ACL)(564)를 더 포함한다. 관리 마스터 스위치(204)는 선택에 따라 원격 폴리시 모듈(566)과 원격 ACL(568)을 더 포함할 수 있으며, 이들은 데이터 포트당 기준으로 인그레스 및/또는 이그레스상에서 하나 또는 그 이상의 팔로어 스위치들(202) 또는 가상 스위치들(310, 312)에 대한 원하는 일 세트의 제어 폴리시들 및 액세스 제어를 구현한다. 관리 마스터 스위치(204)는 또 다른 마스터 스위치(204), 팔로어 스위치(202) 또는 가상 스위치(310, 312)에 대하여 새롭게 추가되거나 갱신된 제어 정보(예를 들어, 제어 폴리시 또는 ACL)를 예약된 관리 VLAN을 통해서 타겟 스위치로 푸시해줄 수 있는 것이 장점이다. 그러므로, 가상화 스위치를 통과하는 트래픽에 대한 ACL들, 제어 폴리시들 및 기타 제어 정보는 마스터 스위치들(204)의 v포트들(522, 524)에서 마스터 스위치들(204)에 의해서, 데이터 포트들(210)에서 팔로어 스위치들(202)에 의해서, 그리고/또는 가상 스위치들(310, 312)의 가상 포트들에서 실행될 수 있다.

[0035] DFP 스위칭 네트워크(200 또는 300) 내의 하나 또는 그 이상의 원하는 위치들에서 폴리시 및 액세스 제어를 전역적으로(globally) 구현하는 능력은 다수의 관리 특징들에 도움이 된다(facilitate). 예를 들면, 마스터 스위치들(204) 사이에서 원하는 부하균형을 달성하기 위해서, 동종의(homogeneous) 또는 이종의(heterogeneous) 제어 폴리시들이 팔로어 스위치들(202) 및/또는 가상 스위치들(310, 312)에 의해 구현되어서, 스위칭 및 라우팅을 위해 마스터 스위치(들)(204)로 가는 데이터 트래픽의 원하는 배분(distribution)을 달성할 수 있다. 한 특정

구현에서, 부하 배분(load distribution)은 다른 통신 프로토콜들이 다른 마스터 스위치들(204)상에서 실행되는 상황에서 여러 트래픽 타입들에 따라서 이루어질 수 있다. 그러므로 마스터 스위치들(204)에 연결된 팔로어 스위치들(202) 및 호스트들(302)는 복수의 다양한 트래픽 타입들 각각의 프로토콜 데이터 유닛들(PDU들)을 그 프로토콜을 담당하는 마스터 스위치(204)에 보냄으로써 원하는 부하 배분을 구현할 수 있다.

[0036] 도 5에 분명하게 도시되어 있지는 않지만, 적어도 일부 실시 예들에서 스위치 컨트롤러(530a)는 계층 2 프레임 스위칭에 더하여 이 기술분야에서 알려진 바와 같이 계층 3에서 라우팅 및 기타 패킷 처리 (및 위의 것)를 추가로 구현한다는 것을 인식해야 한다. 이러한 경우에, 스위치 컨트롤러(530a)는 경로(route)들을 계층 3의 주소들과 연관시키는 라우팅 정보 베이스(RIB)를 포함할 수 있다.

[0037] 이제 도 5b를 참조하면, 도 2의 팔로어 스위치들(202) 중 어느 하나를 구현하는 데 이용될 수 있는 스위치(500b)의 예시적 실시 예의 고 수준 블록 다이어그램이 도시된다. 같은 참조 번호들로 표현된 바와 같이, 스위치(500b)는 복수의 포트들(502a-502m), 스위치 컨트롤러(530b), 및 스위치 컨트롤러(530b)에 의해 제어되는 크로스바 스위치(510)를 갖는 스위치(500a)와 유사한 구조일 수 있다. 그러나, 스위치(500b)는 프레임들을 포워드하는 최종 책임을 마스터 스위치들(204)에 맡기는 패스-스루 모드에서 동작하도록 의도되어 있기 때문에, 스위치 컨트롤러(530b)는 단순화되어 있다. 예를 들면, 도시된 실시 예에서, FIB(532b)의 각 엔트리(534)는 프레임들을 분류하는 데(프레임 분류들은 관리 모듈(550)에 의해 스위치 컨트롤러(530b)로 푸시됨) 이용되는 하나 또는 그 이상의 프레임 필드들(예를 들어, 목적지 MAC 주소, RPI 등)에 대한 값들을 식별하기 위한 제어 필드(570) 및 데이터 트래픽의 해당 분류를 포워드하기 위해 마스터 스위치(204)에 연결된 스위치(530b)의 인그레스 데이터 포트(502)를 식별하는 연관 PID 필드(538)을 포함한다. 제어 모듈(560b)도 마찬가지로 원격 폴리시(566)이나 원격 ACL들(568)이 지원되지 않음으로써 단순화되어 있다. 끝으로, 스위치(500b)는 마스터 스위치(204)로서 기능하도록 장치될 필요가 없으므로, 관리 모듈(550)은 완전히 생략될 수 있다.

[0038] 이제 도 7을 참조하면, 일 실시 예에 따른 DFP 스위칭 네트워크를 관리하기 위한 예시적 프로세스의 고 수준 논리 순서도가 도시된다. 편의상 도 7의 프로세스는 도 2와 도 3의 DFP 스위칭 네트워크들(200 및 300)을 참조하여 기술된다. 본 출원에서 도시되는 다른 논리 순서도와 마찬가지로, 단계들은 엄격한 발생 순서보다는 논리적 순서로 도시되며, 적어도 일부 단계들은 도시된 것과 다른 순서로 또는 동시에 수행될 수 있다.

[0039] 프로세스는 블록(700)에서 시작하고 그 다음에 블록(702)로 진행하며, 이 블록은 마스터 스위치들(204a, 204b)의 각각이 자신이 위치한 DFP 스위칭 네트워크(200 또는 300)의 멤버십(membership: 귀속관계)과 토폴로지(topology)를 파악하는 것(learning)을 도시한다. 여러 실시 예에서, 마스터 스위치들(204a, 204b)는 DFP 스위칭 네트워크(200 또는 300)의 토폴로지와 멤버십을 파악할 수 있는데, 그 방법은 예를 들면 클라이언트 디바이스들(110a-110c) 중 하나에 배치된 네트워크 관리자로부터 구성(configuration)을 수신함으로써, 또는 이와는 달리, 마스터 스위치들(204a, 204b)의 각각의 스위치 컨트롤러(530a)에 의한 자동화된 스위치 발견 프로토콜(an automated switch discovery protocol)의 구현을 통해서 파악할 수 있다. DFP 스위칭 네트워크(200 또는 300)에서 발견된 멤버십에 기초하여, 마스터 스위치들(204)의 각각의 스위치 컨트롤러(530a)는 각각의 포트(502)상에서, DFP 스위칭 네트워크(200, 300)의 하위 층 내의 각 RPI에 대한 각각의 인그레스 v포트(522)와 각각의 이그레스 v포트(524)를 구현하며, 상기 포트로부터 인그레스 데이터 트래픽이 그 포트(502)상에 수신될 수 있다(블록 704). 관리 마스터 스위치(204), 예를 들면, 마스터 스위치(204a)는 그 후에 관리 인터페이스(552)를 통해서 가상화 스위치(600)으로서 DFP 스위칭 네트워크(200 또는 300)의 구성(configuration), 관리(management) 및 제어(control)를 허가한다(블록 706). 가상화 스위치(600)으로서 DFP 스위칭 네트워크(200 또는 300)은 마치 가상화 스위치(600)의 모든 가상화 포트들(602)가 단일한 물리적 스위치 내에 있는 것처럼 동작하도록 구성, 관리 및 제어될 수 있다는 것을 인식해야 한다. 그러므로, 예를 들어 포트 미러링(port mirroring), 포트 트렁킹(port trunking), 멀티캐스팅(multicasting), ETS(enhanced transmission selection)(예를 들어, 규격 초안 표준 IEEE 802.1Qaz에 따른 레이트 리미팅(rate limiting) 및 셰이핑(shaping)), 및 우선순위 기반 플로우 제어가 대응 RPI들이 속하는 스위치들(202, 310, 312) 또는 호스트들(302)와 상관 없이 가상화 포트들(602)에 대하여 구현될 수 있다. 그 후에, 관리 마스터 스위치(예를 들어, 마스터 스위치 204a)의 스위치 컨트롤러(530a)의 관리 모듈(550)이 제어 정보를 다른 마스터 스위치들(204), 팔로어 스위치들(202) 및/또는 가상 스위치들(310, 312)로 푸시해주는데, 이는 다른 스위치들의 제어 모듈(560) 및 FIB(532)를 적절히 구성하기(properly configure) 위해서이다(블록 708). 도 7의 프로세스는 그 후 블록(710)에서 종료된다. 이제 도 8을 참조하면, 일 실시 예에 따라서, 네트워크 트래픽이 가상화 스위치로서 동작하도록 구성된 DFP 스위칭 네트워크의 하위 층으로부터 상위 층으로 포워드되는 예시적 프로세스의 고 수준 논리 순서도가 도시된다. 편의상 도 8의 프로세스 또한 도 2의 DFP 스위칭 네트워크(200)과 도 3의 DFP 스위칭 네트워크(300)을 참조하여 기술한다.

[0040] 상기 도시된 프로세스는 블록(800)에서 시작하고 그 후 블록(802)로 진행하며, 블록(802)는 DFP 스위칭 네트워크의 하위 층에서 RPI가 마스터 스위치(204)로 전송될 데이터 프레임을 수신하는 것을 도시한다. 블록(804)에서 점선 예시로 표시된 바와 같이, RPI가 위치하고 있는 팔로어 스위치(202) 또는 호스트(302)는, 이전에 관리 마스터 스위치(204)에 의해 그렇게 하도록 명령을 받았다면, 선택적으로 상기 데이터 프레임에 폴리시 제어 또는 액세스 제어(ACL을 참조하여)를 실행할 수 있다.

[0041] 블록(806)에서, 하위 층에서 팔로어 스위치(202) 또는 호스트(302)가 RPI 식별자(예를 들어, S-태그)를 상기 데이터 프레임에 적용하여 상기 데이터 프레임이 수신된 인그레스 RPI를 식별한다. 하위 층에서 팔로어 스위치(202) 또는 호스트(302)는 그 다음으로 상기 데이터 프레임을 DFP 스위칭 네트워크(200 또는 300)의 상위 층에 있는 마스터 스위치(204)로 포워드한다(블록 808). 팔로워 스위치(202)의 경우에, 상기 데이터 프레임은 블록(808)에서 FIB(532b)에 의해 표시되는 스위치-간 이그레스 포트를 통해서 포워드된다. 그 후 도 8에 도시된 프로세스는 블록(810)에서 종료된다.

[0042] 도 9를 참조하면, 일 실시 예에 따라서, 상기 상위 층에 있는 마스터 스위치가 DFP 스위칭 네트워크의 상기 하위 층으로부터 수신한 데이터 프레임을 처리하는 예시적 프로세스의 고 수준 논리 순서도가 도시된다. 도시된 프로세스는 블록(900)에서 시작하고 그 다음으로 블록(902)로 진행되며, 블록(902)는 DFP 스위칭 네트워크(200 또는 300)의 마스터 스위치(204)가 팔로어 스위치(202) 또는 호스트(302)로부터 자신의 포트들(502) 중 하나에 데이터 프레임을 수신하는 것을 도시한다. 상기 데이터 프레임의 수신에 응답하여, 상기 데이터가 수신되었던 포트(502)의 수신 인터페이스(504)는 상기 데이터 프레임에 의해 명시되는 RPI 식별자(예를 들어, S-태그)에 따라서 상기 데이터 프레임을 사전-분류하고 상기 데이터 프레임을 그 RPI와 연관된 인그레스 v포트(522)에 대기시킨다(블록 904). 블록(904)로부터, 도 9에 도시된 프로세스는 블록(910)과 블록(920) 양쪽으로 진행한다.

[0043] 블록(910)에서, 스위치 컨트롤러(530a)는 상기 데이터 프레임에 의해 명시되는 목적지 MAC 주소를 이용하여 FIB(532a)에 액세스한다. 만일 일치하는 MAC 필드(536)을 갖는 FIB 엔트리(534)를 찾으면, 프로세스는 블록들(922-928)에서 계속되며, 이 블록들은 아래에서 기술된다. 그러나 만일 스위치 컨트롤러(530a)가 블록(910)에서 목적지 MAC 주소가 알려져 있지 않다고(unknown) 결정을 내리면, 스위치 컨트롤러(530a)는 종래의 발견 기술을 이용하여 목적지 MAC 주소와, 이그레스 포트(502)와 목적지 RPI 사이의 연관을 파악하고 그에 따라서 FIB(532a)를 갱신한다. 그 다음으로 프로세스는 블록들(922-928)로 진행한다.

[0044] 블록(920)에서, 스위치 컨트롤러(530a)는 제어 모듈(560a)에 의해 상기 인그레스 v포트(522)에 대하여 명시되는 임의의 로컬 폴리시(562) 또는 로컬 ACL(564)를 상기 데이터 프레임에 적용한다. 또한, 스위치 컨트롤러(530a)는 인그레스상에서 상기 데이터 프레임에 대하여 기타 다른 특별 처리(special handling)를 수행한다. 아래에서 상세하게 논의되는 바와 같이, 이 특별 처리는 예를 들어 포트 트렁킹, 우선순위 기반 플로우 제어, 멀티캐스팅, 포트 미러링 또는 ETS의 구현을 포함할 수 있다. 각 타입의 특별 처리가 인그레스에서 및/또는 이그레스에서 데이터 트래픽에 적용될 수 있으며, 아래에서 더 기술되는 바와 같다. 프로세스는 그 다음으로 블록들(922-928)로 진행한다.

[0045] 이제 블록들(922-924)를 참조하면, 스위치 컨트롤러(530a)는 상기 데이터 프레임의 RPI 식별자를 일치하는 FIB 엔트리(534)의 VPID 필드(540)에 명시된(또는 상기 발견 프로세스에 의해 파악된) 것과 같도록(equal) 갱신하고 그리고 상기 데이터 프레임을 일치하는 FIB 엔트리(534)의 PID 필드(538)에 의해 식별된(또는 상기 발견 프로세스에 의해 파악된) 대응 이그레스 v포트(524)에 대기시킨다. 블록(926)에서, 스위치 컨트롤러(530a)는 제어 모듈(560a)에 의해 이그레스 v포트(524)에 대하여 명시된 임의의 로컬 폴리시(562) 또는 로컬 ACL(564)를 상기 데이터 프레임에 적용한다. 또한, 스위치 컨트롤러(530a)는 이그레스상에서 상기 데이터 프레임에 대하여, 예를 들어, 포트 트렁킹, 우선순위 기반 플로우 제어, 멀티캐스팅, 포트 미러링 또는 ETS의 구현을 포함한 기타 다른 특별 처리를 수행한다. 그 후에 마스터 스위치(204)가 상기 데이터 프레임을 스위치-간 링크(206)을 통하여 DFP 스위칭 네트워크(200 또는 300)의 하위 층(예를 들어, 팔로어 스위치(202) 또는 호스트(302))으로 포워드한다(블록 928). 도 9에 도시된 프로세스는 그 후 블록(930)에서 종료된다.

[0046] 이제 도 10을 참조하면, 일 실시 예에 따라서, 상기 하위 층에 있는 팔로어 스위치(202) 또는 호스트(302)가 DFP 스위칭 네트워크(200 또는 300)의 상기 상위 층에 있는 마스터 스위치로부터 수신한 데이터 프레임을 처리하는 예시적 프로세스의 고 수준 논리 순서도가 도시된다. 도 10에 도시된 프로세스는 블록(1000)에서 시작하고 그 다음으로 블록(1002)로 진행하며, 블록(1002)는 팔로어 스위치(202) 또는 호스트(302) 등의 하위 층 엔티티가 마스터 스위치(204)로부터, 예를 들어 상기 팔로어 스위치(202)의 스위치-간 포트(502)에 또는 상기 호스트(302)의 네트워크 인터페이스(404)나 VMM(304)에 수신하는 것을 도시한다.

[0047] 상기 데이터 프레임을 수신한 것에 응답하여, 상기 하위 수준 엔티티는 상기 데이터 프레임으로부터 상기 마스터 스위치(204)에 의해 갱신된 상기 RPI 식별자를 제거한다(블록 1004). 상기 하위 수준 엔티티는 그 다음으로 상기 데이터 프레임을 상기 추출된 RPI 식별자에 의해 식별된 RPI로 흘려보낸다(블록 1006). 그러므로, 예를 들면, 스위치 컨트롤러(530b)는 상기 RPI 및/또는 상기 데이터 프레임의 목적지 MAC 주소로 자신의 FIB(532b)에 액세스하여 일치하는 FIB 엔트리(534)를 식별하고 그 다음으로 크로스바(510)을 제어하여 상기 데이터 프레임을 상기 일치하는 FIB 엔트리(534)의 PID 필드(538)에서 명시된 포트로 포워드 한다. 호스트의 네트워크 인터페이스(404) 또는 VMM(304)는 이와 유사하게 상기 데이터 프레임을 RPI 식별자에 의해 표시되는 RPI로 보낸다. 그 후에, 프로세스는 블록(1008)에서 종료된다.

[0048] 이제 도 11을 참조하면, 일 실시 예에 따른 DFP 스위칭 네트워크에서 링크 집합 그룹(LAG: link aggregation group)을 운영하는 예시적 방법의 고 수준 논리 순서도가 도시된다. 링크 집합(link aggregation)은 이 기술분야에서 트렁킹(trunking), 링크 묶음(link bundling), 결합(bonding), 팀화(teaming), 포트 채널, 이더채널(EtherChannel), 및 멀티-링크 트렁킹 등으로 다양하게 불릴 수도 있다.

[0049] 도 11에 도시된 프로세스는 블록(1100)에서 시작하고 그 다음으로 블록(1102)로 진행하며, 블록(1102)는 DFP 스위칭 네트워크(200 또는 300)의 마스터 스위치(204)에서 복수의 RPI들을 포함하는 LAG의 설정(establishement)을 도시한다. 종래의 LAG들과는 달리, DFP 스위칭 네트워크(200 또는 300)에서 설정된 LAG는 다수의 다른 (그리고 아마도 이종의) 팔로어 스위치들(202) 및/또는 호스트들(302)로 된 RPI들을 포함할 수 있다. 예를 들면, 도 2-도 3의 DFP 스위칭 네트워크들(200 및 300)에서, 단일 LAG는 하나 또는 그 이상의 팔로어 스위치들(202a-202d) 및/또는 호스트들(302a-302d)로 된 RPI들을 포함할 수 있다.

[0050] 적어도 일부 실시 예들에서, LAG는, 예를 들면 관리 마스터 스위치(204)의 관리 인터페이스(552)와 상호작용하는 클라이언트 디바이스들(110a-110c) 중 하나에 배치된 시스템 관리자에 의해서, 마스터 스위치(204)의 정적 구성(static configuration)으로 마스터 스위치(204)에 설정될 수 있다. 이와는 달리 또는 추가적으로, LAG는 본 출원에서 참조로 포함된 IEEE 802.1AX-2008에서 정의된 LACP(Link Aggregation Control Protocol)를 통해서 마스터 스위치(204)와 하나 또는 그 이상의 하위 층 엔티티들(예를 들어, 팔로어 스위치들(202) 또는 호스트들(302)) 사이의 메시지들의 교환으로 마스터 스위치(204)에서 설정될 수 있다. LAG는 마스터 스위치(204)에서 설정되기 때문에, LAG에 속한 스위치-간 링크(206)에 연결된 모든 하위 수준 엔티티들이 LAG에 대한 지원을 제공할 (또는 LAG의 존재도 알) 필요는 없다는 것을 인식해야 한다.

[0051] 블록(1102)에 도시된 바와 같이 마스터 스위치(204)에서 LAG의 설정은 도 12에 도시된 바와 같이 스위치 컨트롤러(530a) 내 LAG 데이터 구조(1200)에 LAG의 멤버십을 기록하는 것을 포함하는 것이 바람직하다. 상기 도시된 예시적 실시 예에서, LAG 데이터 구조(1200)은 하나 또는 그 이상의 LAG 멤버십 엔트리들(1202)을 포함하며 이들 각각은 각각의 LAG에서 멤버십을 명시한다. 한 바람직한 실시 예에서, LAG 멤버십 엔트리들(1202)는 RPI들 또는 상기 LAG를 형성하는 상기 RPI들과 연관된 v포트들(520)의 관점에서 LAG 멤버십을 표현한다. 다른 실시 예들에서, 상기 LAG는 이와는 다르게 또는 추가로 상기 마스터 스위치(204)와 RPI들을 연결하는 스위치-간 링크들(206)의 관점에서 표현될 수도 있다. 따라서, LAG 데이터 구조(1200)은 독자적인(stand alone) 데이터 구조로 구현되거나 FIB(532a) 등과 같은 또 다른 데이터 구조의 하나 또는 그 이상의 필드들에 구현될 수도 있음을 이해할 수 있을 것이다.

[0052] 상기 LAG의 설정에 이어서, 마스터 스위치(204)는, 도 9의 블록들(920-926)을 참조하여 위에서 이미 언급한 바와 같이, LAG 내의 RPI들로 보내진 데이터 프레임들에 대한 특별 처리를 수행한다. 구체적으로, 블록(1104)에 도시된 바와 같이, 스위치 컨트롤러(530a)는 포워딩을 위해 수신된 데이터 프레임들을 모니터하고, 예를 들어 FIB(532a) 및/또는 LAG 데이터 구조(1200)을 참조하여 상기 데이터 프레임에 포함된 목적지 MAC 주소가 LAG에 속한 RPI와 연관된 것으로 알려지는지 아닌지를 결정한다. 블록(1104)에서 부정적인 결정에 응답하여, 프로세스는 블록(1112)로 넘어가며, 블록(1112)는 아래에 기술된다. 그러나 만일 스위치 컨트롤러(532a)가 블록(1104)에서, 데이터 프레임은 LAG에 속한 RPI와 연관된 목적지 MAC으로 주소지정되어 있다고 결정하면, 스위치 컨트롤러(532a)는 상기 LAG의 멤버십 중에서 상기 데이터 프레임을 위한 이그레스 RPI를 선택한다.

[0053] 블록(1110)에서, 스위치 컨트롤러(532a)는 라운드-로빈(round-robin) LAG 폴리시, 브로드캐스트(broadcast) LAG 폴리시, 부하균형(load balancing) LAG 폴리시, 또는 해시(hashed) LAG 폴리시를 포함한 복수의 LAG 폴리시들 중 어느 하나에 기초하여 LAG 멤버십 중에서 이그레스 RPI를 선택할 수 있다. 해시 LAG 폴리시의 한 구현에서, 스위치 컨트롤러(532a)는 소스 및 목적지 MAC 주소들을 XOR(배타적 논리합)연산하고 주어진 목적지 MAC 주소에 대하여 항상 동일한 RPI를 선택하기 위해 LAG의 사이즈로 연산 결과에 대하여 모듈로 연산(modulo

operation)을 수행한다. 다른 실시 예들에서, 해시 LAG 폴리시는 소스 IP 주소, 목적지 IP 주소, 소스 MAC 주소, 목적지 주소, 및/또는 소스 RPI 등을 포함한 서로 다른 또는 추가적인 요소들(factors)에 기초하여 이그레스 RPI를 선택할 수 있다.

[0054] 블록(1112)에 표시된 바와 같이, LAG에 걸친 데이터 프레임들의 "분사(spraying)" 또는 배분(distribution)은, 예를 들어 마스터 스위치(204)의 정적 구성을 제거함으로써 또는 LCAP를 통해서 LAG가 구성해제(deconfigure)될 때까지 계속된다. 그 후, 도 11에 도시된 프로세스는 블록(1120)에서 종료된다.

[0055] 마스터 스위치(204)에서 상이한 하위 수준 엔티티들을 포괄하는(span) 분산형(distributed) LAG를 구현하는 능력은 추가적인 네트워크 능력들을 가능하게 한다. 예를 들면, 동일한 서비스를 제공하는 다수 VM들(306)을 포함하는 DFP 스위칭 네트워크(300)에서, 그러한 모든 VM들을 멤버들(members)로 갖는 LAG를 형성하는 것은, VMM들(304)에 의한 어떠한 관리가 없어도, 서비스 태그 필드 및 다른 투플(other tuple) 필드에 기초하여 상기 VM들(306)에 걸쳐 자동으로 부하균형이 될 수 있게 해준다. 또한, 이러한 부하균형은 다른(different) VMM들(304)과 다른(different) 호스트들(302)에서 실행중인 VM들(306)에 걸쳐 달성될 수 있다.

[0056] 위에서 언급한 바와 같이, 도 9의 블록들(920-926)에서 선택적으로 수행되는 특별 처리는 프레임들을 LAG에 배분하는 것과 데이터 트래픽의 멀티캐스팅을 포함한다. 이제 도 13을 참조하면, 일 실시 예에 따른 DFP 스위칭 네트워크에서 멀티캐스팅하는 예시적 방법의 고 수준 논리 순서도가 도시된다. 프로세스는 블록(1300)에서 시작하고 그 다음으로 블록들(1302-1322)로 진행하며, 이 블록들은 도 9의 블록들(920-926)을 참조하여 앞에서 기술된 바와 같이, 멀티캐스트 데이터 트래픽에 대하여 마스터 스위치에 의해 수행되는 특별 처리를 도시한다.

[0057] 구체적으로, 블록(1310)에서, 마스터 스위치(204)의 스위치 컨트롤러(530a)는 데이터 트래픽 내에서 명시된 목적지 MAC 주소 또는 IP 주소를 참조하여 상기 데이터 트래픽이 멀티캐스트 전달(multicast delivery)을 요청할지를 결정한다. 예를 들면, IP는 멀티캐스트 주소들을 위해 224.0.0.0부터 239.255.255.255까지 유보하고(reserve), 이더넷은 적어도 표 1에 요약된 멀티캐스트 주소들을 이용한다.

표 1

[0058]

멀티캐스트 주소	프로토콜
01:00:0C:CC:CC:CC	CDP(Cisco Discovery Protocol) 또는 VTP(VLAN Trunking Protocol)
01:00:0C:CC:CC:CD	시스코 공유 스패닝 트리 프로토콜 주소(Cisco Shared Spanning Tree Protocol Addresses)
01:80:C2:00:00:00	IEEE 802. 1D 스패닝 트리 프로토콜(Spanning Tree Protocol)

[0059] 블록(1310)에서 데이터 트래픽이 멀티캐스트 처리를 필요로하지 않는다는 결정에 응답하여, 상기 데이터 트래픽에 대하여 멀티캐스트 처리는 수행되지 않고(그럼에도 불구하고 다른 특별 처리는 수행될 수 있음), 프로세스는 블록(1310)에서 반복된다. 그러나, 만일 스위치 컨트롤러(530a)가 블록(1310)에서 진입하는 데이터 트래픽이 멀티캐스트 트래픽이라고 결정하면, 프로세스는 블록(1312)로 넘어간다.

[0060] 블록(1312)에서, 스위치 컨트롤러(530a)는 멀티캐스트 인덱스 데이터 구조 내 상기 멀티캐스트 데이터 트래픽에 대한 검색(lookup)을 수행한다. 예를 들면, 도 14에 도시된 예시적 일 실시 예에서, 스위치 컨트롤러(530a)는 계층 2 멀티캐스트 프레임들을 위한 계층 2 멀티캐스트 인덱스 데이터 구조(1400) 및 계층 3 멀티캐스트 패킷들을 위한 계층 3 멀티캐스트 인덱스 데이터 구조(1410)을 구현한다. 상기 도시된 예시적 실시 예에서, 계층 2 멀티캐스트 인덱스 데이터 구조(1400)은, 예를 들어 표(table)로 구현될 수 있는데, 복수의 엔트리들(1402)를 포함하며, 이 복수의 엔트리들(1402) 각각은 인그레스 RPI, 소스 MAC 주소, 목적지 MAC 주소 및 VLAN으로 형성된 4-투플 필드(1404)를, 멀티캐스트 목적지 데이터 구조(1420)에 인덱스를 명시하는, 인덱스 필드(1406)과 연관시킨다. 계층 3 멀티캐스트 인덱스 데이터 구조(1410)도, 비슷하게 표(table)로 구현될 수 있고, 복수의 엔트리들(1412)를 포함하며, 이 복수의 엔트리들(1412) 각각은 소스 계층 3(예를 들어, IP) 주소와 멀티캐스트 그룹 ID로 형성된, 2-투플 필드(1414)를, 멀티캐스트 목적지 데이터 구조(1420)에 인덱스를 명시하는, 인덱스 필드(1416)과 연관시킨다. 멀티캐스트 목적지 데이터 구조(1420)은, 표(table) 또는 링크된 목록(linked list)으로 구현될 수도 있으며, 복수의 멀티캐스트 목적지 엔트리들(1422)를 포함하며, 이 복수의 멀티캐스트 목적지 엔트리들(1422) 각각은 데이터 트래픽이 전송될 하위 층에 있는 하나 또는 그 이상의 RPI들을 식별한다. 계층 2 멀티캐스트 데이터 구조(1400), 계층 3 멀티캐스트 인덱스 데이터 구조(1410) 및 멀티캐스트 목적지 데이터 구조

(1420)은 모두 종래의 MC 학습 프로세스에서 제어 평면이 이식되는(populated) 것이 바람직하다.

[0061] 그러므로, 블록(1312)에서, 스위치 컨트롤러(530a)는 만일 상기 데이터 트래픽이 계층 2 멀티캐스트 프레임이면 계층 2 멀티캐스트 인덱스 데이터 구조(1400)에서 멀티캐스트 목적지 데이터 구조(1420)에 대한 인덱스를 획득하기 위해 검색을 수행하고, 만일 상기 데이터 트래픽이 L3 멀티캐스트 패킷이면 계층 3 멀티캐스트 인덱스 데이터 구조(1410)에서 상기 검색을 수행한다. 블록(1314)에서 표시되는 바와 같이, 마스터 스위치(204)는 인그레스 복제 또는 이그레스 복제 중 하나를 통해서 원하는 구현으로 상기 데이터 트래픽의 멀티캐스트를 처리할 수 있으며, 원하는 구현은 스위치 컨트롤러(530a)에서 구성되는 것이 바람직하다. 만일 이그레스 복제가 마스터 스위치(204)상에서 구성되면, 프로세스는 블록(1316)으로 진행하며, 블록(1316)은 스위치 컨트롤러(530a)가 상기 데이터 트래픽의 단일 복제본(copy)이 크로스바(510)을 가로질러서 각각의 이그레스 큐(514)에서 복제되게 하는 것을 도시하며, 이 각각의 이그레스 큐(514)는 블록(1312)에서 획득된 인덱스에 의해 식별되는 멀티캐스트 목적지 엔트리(1422)에서 식별된 RPI에 대응한다. 따라서, 멀티캐스트 트래픽의 이그레스 복제는 HOL(head-of-line) 블록킹을 희생하여 크로스바(510)의 대역폭(bandwidth)의 이용을 감소시킨다는 것을 이해할 것이다. 블록(1316)에 이어서, 마스터 스위치(204)에 의해 복제된 데이터 트래픽의 처리는, 도 9에서 이미 기술된 바와 같이 계속된다(블록 1330).

[0062] 반면에, 만일 마스터 스위치(204)가 인그레스 복제를 위해 구성되면, 프로세스는 블록(1314)에서 블록(1320)으로 진행하고, 블록(1320)은 스위치 컨트롤러(530a)가 상기 멀티캐스트 데이터 트래픽이 포트들(502)의 인그레스 큐들(506) 각각 내에서 복제되도록 하는 것을 도시하며, 이 포트들(502)의 인그레스 큐들(506) 각각은 인덱스된 멀티캐스트 목적지 엔트리(1422)에서 식별된 RPI들과 연관된 출력 큐들(514)를 갖는다. 따라서, 이러한 방식의 인그레스 복제는 HOL 블록킹을 제거한다는 것을 이해할 것이다. 블록(1320)에 이어서, 상기 데이터 트래픽은 도 9를 참조하여 위에서 논의된 바와 같이 추가적인 처리를 거친다. 이러한 처리에서, 스위치 컨트롤러 (530a)는 크로스바(510)을 제어하여 인그레스상에서 복제된 상기 멀티캐스트 데이터 트래픽을 인그레스 큐들(506)으로부터 동일 포트들(502)의 이그레스 큐들(514)로 직접 전송하게 한다.

[0063] 따라서, 팔로어 스위치들(202)에서보다는 도시된 바와 같이 DFP 스위칭 네트워크(200)의 마스터 스위치(204)에서 MC 처리의 구현이, 데이터 트래픽의 멀티캐스트 배분을 가능하게 할 필요가 없는, 간소화된 팔로어 스위치들(202)의 사용을 가능하게 함을 이해할 것이다.

[0064] 도 9의 블록들(920-926)을 참조하여 위에서 기술된 바와 같이, DFP 스위칭 네트워크에서 데이터 트래픽의 특별 처리는 선택적으로 데이터 트래픽에 ETS를 적용하는 것을 포함할 수 있다. 도 15는 일 실시 예에 따른 DFP 스위칭 네트워크(200 또는 300)에서 향상된 전송 선택(ETS)의 예시적 방법의 고 수준 논리 순서도이다.

[0065] 도 15에 도시된 프로세스는 블록(1500)에서 시작하고 그 다음으로 블록(1502)로 진행하며, 블록(1502)는 예를 들면 DFP 스위칭 네트워크(200 또는 300)의 관리 마스터 스위치(204)상의 관리 인터페이스(552)를 통해서, ETS를 구현하기 위한 마스터 스위치(204)의 구성(configuration)을 도시한다. 여러 실시 예들에서, ETS는 마스터 스위치(204)의 인그레스 및/또는 이그레스에서 구현되도록 구성된다.

[0066] 규격 초안 표준 IEEE 802.1Qaz에서 정의되는 ETS는 다수 TCG(traffic class group)들을 설정하고 그 TCG들 중에서 링크 이용의 원하는 균형(balance)을 달성하기 위해 트래픽 큐들(예를 들어, 인그레스 v포트들(522) 또는 이그레스 v포트들(524))로부터 여러 TCG들 내 데이터 트래픽의 전송 우선순위(즉, 스케줄링)를 명시한다. ETS는 각각의 TCG에 대하여 최소의 보장된 대역폭을 설정할 뿐 아니라, 하위-순위 트래픽으로 하여금 상위-순위 TCG들이 공칭적으로(nominally) 이용할 수 있는 이용되는 대역폭을 소비할 수 있도록 허용하며, 그렇게 함으로써 하위 순위 트래픽의 기아현상(starvation)을 방지하면서 링크 이용성 및 유연성을 개선시킨다. 마스터 스위치(204)에서 ETS의 구성은, 예를 들면, 상기 마스터 스위치(204)의 스위치 컨트롤러(530a) 내에 도 16에 도시된 바와 같이 ETS 데이터 구조(1600)을 설정(establishing) 및/또는 이식(populating)하는 것을 포함할 수 있다. 도 16에 도시된 예시적 실시 예에서, ETS 데이터 구조(1600)은, 예를 들면 표(table)로 구현될 수 있는데, 복수의 ETS 엔트리들(1602)를 포함한다. 상기 도시된 실시 예에서, 각각의 ETS 엔트리(1602)는 주어진 TCG에 속하는 트래픽 타입(들)(예를 들어, 파이버 채널(FC), 이더넷, FC 오버 이더넷(FCoE), iSCSI 등)을 정의하는 TCG 필드(1604)와, TCG 필드(1604)에서 정의된 TCG에 대한 보장된 최소 대역폭을 (예를 들어, 절대치 또는 백분율로) 정의하는 최소 필드(1606), 및 TCG 필드(1604)에서 정의된 TCG에 대한 최대 대역폭을 (절대치 또는 백분율로) 정의하는 최대 필드(1608)을 포함한다.

[0067] 도 15로 돌아가서, 마스터 스위치상에서 ETS의 구성(1502) 이후에, 프로세스는 블록들(1504-1510)으로 진행하고, 블록들(1504-1510)은 도 9의 블록들(920-926)에서 ETS에 대하여 선택적으로 수행되는 특별 처리를 도

시한다. 구체적으로, 블록(1504)는 마스터 스위치(204)가 인그레스 v포트(520) 또는 이그레스 v포트(522)에 수신된 데이터 프레임이 예를 들어 ETS 데이터 구조(1600)에 의해 정의된 바와 같은, 현재 구성된 ETS TCG에 속한 트래픽 클래스에 속하는지 아닌지를 결정하는 것을 도시한다. 따라서, 상기 데이터 프레임은 종래의 이더넷 프레임 또는 그와 유사한 것의 이더타입(Ethertype) 필드에 기초하여 분류될 수 있음을 이해할 것이다. 블록(1504)에서 상기 수신된 데이터 프레임이 현재 구성된 ETS TCG에 속하지 않는다는 결정에 응답하여, 상기 데이터 프레임은 최상의 최선 노력 스케줄링(best efforts scheduling)을 수신하고, 프로세스는 블록(1512)로 진행하며, 이 블록은 아래에 기술된다.

[0068] 블록(1504)로 돌아가서, 상기 수신된 데이터 프레임이 현재 구성된 ETS TCG에 속한다는 결정에 응답하여, 마스터 스위치(204)는 상기 데이터 프레임에 레이트 리미팅(rate limiting) 및 트래픽 셰이핑(traffic shaping)을 적용하여 ETS 데이터 구조(1600)의 관련 ETS 엔트리(1602)의 필드들(1606, 1608) 내에서 ETS TCG에 대하여 명시된 최소 대역폭 및 최대 대역폭을 준수하게 한다(블록 1510). 위에서 언급된 바와 같이, 구성에 따라, 마스터 스위치(204)는 ETS를 인그레스 v포트들(522) 및/또는 이그레스 v포트들(524)에서 VOQ들에 적용할 수 있다. 그 다음으로 프로세스는 블록(1512)로 진행하며, 블록(1512)는 마스터 스위치(204)가 블록들(1504와 1510)에서 도시된 바와 같은 트래픽 클래스에 대하여 ETS를 구현하되, ETS가 그 트래픽 클래스에 대하여 구성해제될 때까지 하는 것을 도시한다. 그 후, 도 15에 도시된 프로세스는 블록(1520)에서 종료된다.

[0069] DFP 스위칭 네트워크(200 또는 300)에서, 플로우 제어는 도 15와 도 16을 참조하여 기술된 바와 같이 마스터 스위치들(204)에 구현될 뿐 아니라, 팔로어 스위치들(202)와 호스트들(302)와 같은 하위 층 엔티티들의 RPI들에서도 구현될 수 있다는 것이 장점이다. 이제 도 17을 참조하면, DFP 스위칭 네트워크(200 또는 300)이 하위 층에서 우선순위-기반 플로우 제어(PFC) 및/또는 기타 서비스들을 구현하는 예시적 방법의 고 수준 논리 순서도가 도시된다.

[0070] 도 17에 도시된 프로세스는 블록(1700)에서 시작하고 그 다음으로 블록(1702)로 진행하며, 블록(1702)는 마스터 스위치(204)가 DFP 스위칭 네트워크(200 또는 300)의 하위 층에서 한 엔티티에 대한 우선순위-기반 플로우 제어(PFC)를 구현하는 것을 표시하는데, 이것은, 예를 들면 (1) 관리 모듈(550)을 실행하는 관리 마스터 스위치(204)에, 하위 층 엔티티의 적어도 하나의 RPI에 대응하는 가상화 포트(602a-602d)에 대한 PFC 구성을 수신한 것에 응답하여 또는 (2) 마스터 스위치(204)에, 상기 네트워크 내 다운스트림 엔티티에 의해 발신되고 패스-스루 팔로어 스위치(202)를 통하여 상기 마스터 스위치(204)에 수신된 표준-기반(standards-based) PFC 데이터 프레임을 수신하는 것에 응답하여, 표시한다. 이 기술분야에서 통상의 지식을 가진 자들은 인식할 수 있는 바와 같이, 표준-기반 PFC 데이터 프레임은 업스트림 엔티티로부터 데이터 트래픽 플로우(flow)를 수신하여 상기 업스트림 엔티티에 그 트래픽 플로우에 대한 혼잡(congestion)을 알리는 다운스트림 엔티티에 의해 생성될 수 있다. 블록(1702)에서 마스터 스위치(204)가 하위 층 엔티티에 대한 PFC 구성을 수신했다는 긍정(affirmative) 결정에 응답하여, 프로세스는 블록(1704)로 진행하며, 이 블록은 상기 마스터 스위치(204)가 PFC에 대하여 하위 층 엔티티를 구성하기 위해 적어도 하나의 하위 층 엔티티(예를 들어, 팔로어 스위치(202) 또는 호스트(302))에 PFC 구성 필드들을 갖는 개선된 소유권 있는(proprietary) 데이터 프레임(이하 소유권 있는 PFC 데이터 프레임이라 불림)을 구축 및 전송하는 것을 도시한다. 그 후, 도 17에 도시된 프로세스는 블록(1706)에서 종료된다.

[0071] 이제 도 18을 참조하면, 일 실시 예에 따른 예시적 소유권 있는 PFC 데이터 프레임(1800)의 구조가 도시된다. 도 17의 블록(1704)를 참조하여 앞에서 기술된 바와 같이, 소유권 있는 PFC 데이터 프레임(1800)은 하위 층 엔티티에서 PFC를 구현하기 위해 마스터 스위치(204)에 의해 구축되고 팔로어 스위치(202) 또는 호스트(302) 등과 같은 DFP 스위칭 네트워크의 하위 층 엔티티에 전송될 수 있다.

[0072] 상기 도시된 예시적 실시 예에서, 소유권 있는 PFC 데이터 프레임(1800)은 확장(expanded) 이더넷 MAC 제어 프레임으로 구현된다. 소유권 있는 PFC 데이터 프레임(1800)은 따라서 상기 하위 층 엔티티에서 RPI – 이곳으로부터 상기 마스터 스위치(204)가 데이터 프레임들을 수신할 수 있음 – 의 상기 MAC 주소를 명시하는 목적지 MAC 주소 필드(1802)와 마스터 스위치(204) 상의 이그레스 v포트 – 이곳으로부터 상기 소유권 있는 PFC 데이터 프레임(1800)이 전송됨 – 를 식별하는 소스 MAC 주소 필드(1804)를 포함한다. 주소 필드들(1802, 1804) 다음으로 PFC 데이터 프레임(1800)을 MAC 제어 프레임으로 (예를 들어, 0x8808의 값에 의해) 식별하는 이더타입(Ethertype) 필드(1806)가 온다.

[0073] 소유권 있는 PFC 데이터 프레임(1800)의 데이터 필드는 MAC 제어 오피코드(opcode) 필드(1808)부터 시작하며 이 필드는 소유권 있는 PFC 데이터 프레임(1800)은 (예를 들어, 0x0101의 PAUSE 커맨드 값에 의해) 플로우 제어를 구현하기 위한 것임을 표시한다. MAC 제어 오피코드 필드(1808) 다음으로 인에이블 필드(1812)와 클래스 벡터

필드(1814)를 포함하는 우선순위 인에이블 벡터(1810)가 온다. 한 실시 예에서, 인에이블 필드(1812)는 최하위 비트(least significant bit)의 상태에 의해 소유권 있는 PFC 데이터 프레임(1800)이 소유권 있는 PFC 데이터 프레임(1800)의 목적지인 하위 층 엔티티에 있는 RPI에서 플로우 제어를 구현하기 위한 것인지 아닌지를 표시한다. 클래스 벡터(1814)는, 예를 들어 멀티-핫 인코딩(multi-hot encoding)을 이용하여, N개의 트래픽 클래스들 중 어떤 것을 위하여, 플로우 제어가 소유권 있는 PFC 데이터 프레임(1800)에 의해 구현되는지를 더 표시한다. 우선순위 인에이블 벡터(1810) 다음에, 소유권 있는 PFC 데이터 프레임(1800)은 N개의 시간 양자(time quanta) 필드들(1820a-1820n)을 포함하고 이 필드들 각각은 N개의 트래픽 클래스들 중 각각의 트래픽 클래스에 대응하며 이 N개의 트래픽 클래스들을 위해 플로우 제어가 구현될 수 있다. 인에이블 필드(1812)는 RPI들에 대한 플로우 제어를 가능하게 하도록 설정되고 클래스 벡터(1814) 내 그 대응 비트는 특정 트래픽 클래스에 대한 플로우 제어를 표시하도록 설정된다고 가정하면, 주어진 시간 양자 필드(1820)은 연관된 트래픽 클래스 내의 데이터의 RPI에 의한 최대 전송 대역폭(maximum bandwidth of transmission)을 (예를 들어, 백분율로서 또는 절대값으로서) 명시한다. 소유권 있는 PFC 데이터 프레임(1800)에 의해 플로우 제어가 구성되는 RPI는 RPI 필드(1824)에 의해 더 명시된다.

[0074] 상기 데이터 필드 다음에, 소유권 있는 PFC 데이터 프레임(1800)은 PFC 데이터 프레임(1800)의 미리 정해진 사이즈를 획득하기 위한, 선택가능한 패딩(1826)을 포함한다. 끝으로, PFC 데이터 프레임(1800)은 PFC 데이터 프레임(1800)에서 에러들을 감지하는 데 이용되는 종래의 체크섬 필드(1830)을 포함한다.

[0075] 따라서, 소유권 있는 PFC 데이터 프레임(1800)은 RPI들에 대한 플로우 제어 이외의 평션들을 트리거하는 데 이용될 수 있음을 이해할 것이다. 예를 들면, 소유권 있는 PFC 데이터 프레임(1800)은 또한 명시된 RPI에 대한 서비스들을 (예를 들어, 시간 양자 필드들(1820)의 특별 예약된 값들을 이용하여) 트리거하는 데 이용될 수 있다. 이러한 추가적인 서비스들에는 예를 들어 서버 부하균형 폴리시 재해싱(rehashing), 파이어월 제한사항 업데이트, 서비스 거부(DOS) 공격 검사의 실행 등이 포함될 수 있다.

[0076] 도 19a를 참조하면, 일 실시 예에 따른 팔로어 스위치(202)와 같은 DFP 스위칭 네트워크(200 또는 300)의 하위 수준 엔티티가 마스터 스위치로부터 수신된 소유권 있는 PFC 데이터 프레임(1800)을 처리하는 예시적 프로세스의 고 수준 논리 순서도가 도시된다.

[0077] 프로세스는 블록(1900)에서 시작하고 그 다음으로 블록(1902)로 진행하며, 이 블록은 소유권 있는 PFC 데이터 프레임(1800)의 수신을 모니터하는, 팔로어 스위치(202)와 같은, 패스 스루 하위 수준 엔티티(a pass through lower level entity)를 도시한다. 예를 들어 MAC 제어 오피코드 필드(1808)에 기초한 분류에 의해 감지되는 소유권 있는 PFC 데이터 프레임(1800)의 수신에 응답하여, 프로세스는 블록(1902)에서 블록(1904)로 진행한다. 블록(1904)는, 팔로어 스위치(202)(예를 들어, 스위치 컨트롤러(530b))가 예를 들어 비-표준 필드들(non-standard fields) (1810, 1820 및 1824)를 추출(extracting)함으로써, 상기 소유권 있는 PFC 데이터 프레임(1800)을 표준-기반 PFC 데이터 프레임으로 변환하는 것을 도시한다. 팔로어 스위치(202)는 그 다음으로, 예를 들어 상기 RPI 필드(1824)로부터 추출된 RPI를 FIB(532b)를 참조하는 포트 ID로 변환함으로써, 상기 표준-기반 PFC 데이터 프레임에 대한 이그레스 데이터 포트(210)을 결정하고, 그 결과로 나온 표준-기반 PFC 데이터 프레임을 상기 결정된 이그레스 데이터 포트(210)을 통하여 혼잡을 야기하는 데이터 트래픽의 소스 쪽으로 포워드한다(블록 1906). 그 후, 도 19a에 도시된 프로세스는 블록(1910)에서 종료된다. PFC는 RPI별로 개별적으로 구현될 수 있기 때문에, 상기 기술된 프로세스는 동일한 하위 층 엔티티(예를 들어, 팔로어 스위치(202) 또는 호스트(302)) 상에서 다른 RPI들에 대하여 다른 PFC를 구현하는 데 이용될 수 있다는 것에 유의해야 한다. 또한, 하위 층 엔티티들에서 RPI들은 VOQ들(604)로 표시되기 때문에, 하나 또는 그 이상의 RPI들에 대한 개별화된 PFC는, 동일한 포트(502)가 다른 v포트들(522, 524)의 데이터 트래픽에 대하여 다른 PFC를 구현하도록, 마스터 스위치들(204)에서 대안적으로 그리고 선택적으로 구현될 수 있다.

[0078] 이제 도 19b를 참조하면, 일 실시 예에 따른 호스트 플랫폼(302)와 같은 DFP 스위칭 네트워크(200 또는 300)의 하위 수준 엔티티가 마스터 스위치(204)로부터 수신된 소유권 있는 PFC 데이터 프레임(1800)을 처리하는 예시적 프로세스의 고 수준 논리 순서도가 도시된다.

[0079] 프로세스는 블록(1920)에서 시작하고 그 다음으로 블록(1922)로 진행하며, 이 블록은 예를 들어 MAC 제어 오피코드 필드(1808)에 기초하여 진입하는 데이터 프레임들을 분류함으로써 소유권 있는 PFC 데이터 프레임(1800)의 수신을 모니터하는 호스트 플랫폼(302)의 네트워크 인터페이스(404)(예를 들어, CNA 또는 NIC)를 도시한다. 소유권 있는 PFC 데이터 프레임(1800)의 수신을 감지한 것에 응답하여, 프로세스는 블록(1922)에서 블록(1930)으로 진행한다. 블록(1930)은 네트워크 인터페이스(404)가 소유권 있는 PFC 데이터 프레임(1800)을 예를 들어 인

터럽트 또는 다른 메시지를 통해서 처리를 위해 VMM(304)로 전송하는 것을 도시한다. 소유권 있는 PFC 데이터 프레임(1800)의 수신에 응답하여, 하이퍼바이저(304)는 소유권 있는 PFC 데이터 프레임(1800)을 소유권 있는 PFC 데이터 프레임(1800)의 RPI 필드(1824)에서 표시된 RPI와 연관된 VM(306)으로 전송한다(블록 1932). 이에 대한 응답으로, VM(306)은 소유권 있는 PFC 데이터 프레임(1800)에 의해 표시되는 특정 애플리케이션과 트래픽 우선순위에 대한 PFC(또는 소유권 있는 PFC 데이터 프레임(1800)에 의해 표시되는 다른 서비스)를 적용한다(블록 1934). 그러므로, PFC는 우선순위-별로, 애플리케이션-별로 구현되어, 예를 들면 데이터 센터 서버 플랫폼과의 통신에서 비디오 스트리밍 클라이언트로부터의 백 프레셔에 응답하여, 예를 들면, 데이터 센터 서버 플랫폼이 제2 VM(306)(예를 들어, FTP 서버)에보다 제1 VM(306)(예를 들어, 비디오 스트리밍 서버)에 다른 PFC를 적용하는 것이 가능하게 해준다. 블록(1934)에 이어서, 도 19b에 도시된 프로세스는 블록(1940)에서 종료된다.

[0080] 기술된 바와 같이, 일부 실시 예들에서, 스위칭 네트워크는 마스터 스위치를 포함하는 상위 층과 복수의 하위 층 엔티티들을 포함하는 하위 층을 포함한다. 마스터 스위치는 각각 복수의 하위 층 엔티티들 중 각각의 하위 층 엔티티에 결합된 복수의 포트들을 포함한다. 상기 복수의 포트들의 각각의 포트는 복수의 가상 포트들을 포함하며 이 가상 포트들 각각은 그 포트에 결합된 하위 층 엔티티에서 복수의 원격 물리적 인터페이스들(RPI들) 중 각각의 원격 물리적 인터페이스에 대응한다. 상기 복수의 포트들의 각각의 포트는 또한 상기 복수의 하위 층 엔티티들 중에서 특정 하위 층 엔티티로부터 데이터 트래픽의 수신에 응답하여 상기 데이터 트래픽을 상기 복수의 가상 포트들 중에서 상기 데이터 트래픽의 소스였던 상기 특정 하위 층 엔티티상의 RPI에 대응하는 가상 포트에 대기시키는, 수신 인터페이스를 포함한다. 상기 마스터 스위치는 데이터 트래픽을 상기 가상 포트로부터 이 데이터 트래픽이 포워드되어 나오는 복수의 포트들 중에서 이그레스 포트로 스위치하는 스위치 컨트롤러를 더 포함한다.

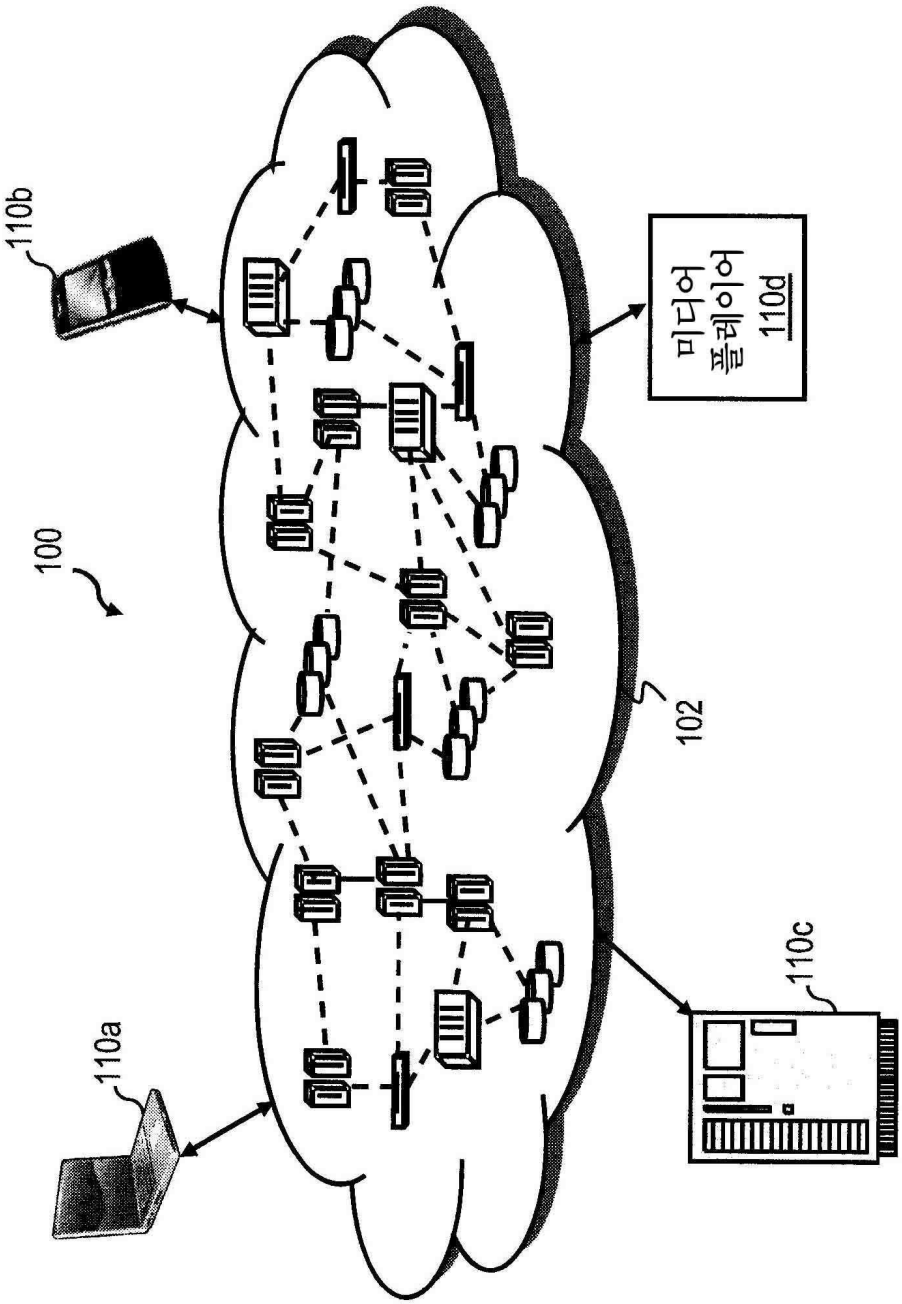
[0081] 상위 층과 하위 층을 포함하는 스위칭 네트워크의 일부 실시 예들에서, 각각의 포트가 각각의 하위 층 엔티티에 결합된 복수의 포트들을 갖는 상위 층에서 마스터 스위치는 상기 포트들 중 각각의 포트상에서 그 포트에 결합된 하위 층 엔티티에 있는 복수의 원격 물리적 인터페이스들(RPI들) 중 각각의 원격 물리적 인터페이스에 각각의 가상 포트가 대응하는 복수의 가상 포트들을 구현한다. 상기 마스터 스위치와 RPI들 사이에서 전달되는 데이터 트래픽은 그 데이터 트래픽이 전달되는 하위 층 엔티티들상의 RPI들에 대응하는 가상 포트들 내에서 대기한다. 상기 마스터 스위치는 주어진 가상 포트의 데이터 트래픽상에서, 대응 RPI가 상주하는 하위 층 엔티티에, 특정 RPI에 의해 전달되는 적어도 두 개의 다른 데이터 트래픽의 클래스에 대한 우선순위를 명시하는 PFC 데이터 프레임을 전송함으로써 우선순위-기반 플로우 제어(PFC)를 실행한다.

[0082] 상위 층과 하위 층을 포함하는 스위칭 네트워크의 일부 실시 예에서, 각각의 포트가 각각의 하위 층 엔티티에 결합된 복수의 포트들을 갖는 상위 층에서 마스터 스위치는 상기 포트들 중 각각의 포트상에서 그 포트에 결합된 하위 층 엔티티에 있는 복수의 원격 물리적 인터페이스들(RPI들) 중 각각의 원격 물리적 인터페이스에 각각의 가상 포트가 대응하는 복수의 가상 포트들을 구현한다. 상기 마스터 스위치와 RPI들 사이에서 전달되는 데이터 트래픽은 그 데이터 트래픽이 전달되는 RPI들에 대응하는 가상 포트들 내에서 대기한다. 상기 마스터 스위치는 적어도 상기 데이터 트래픽이 대기하는 가상 포트에 기초한 제어 폴리시에 따라서 상기 데이터 트래픽에 데이터 처리를 적용하여, 상기 마스터 스위치가 이 마스터 스위치의 동일한 포트상의 두 개의 가상 포트들에 대기하는 데이터 트래픽에 다른 폴리시들을 적용하게 한다.

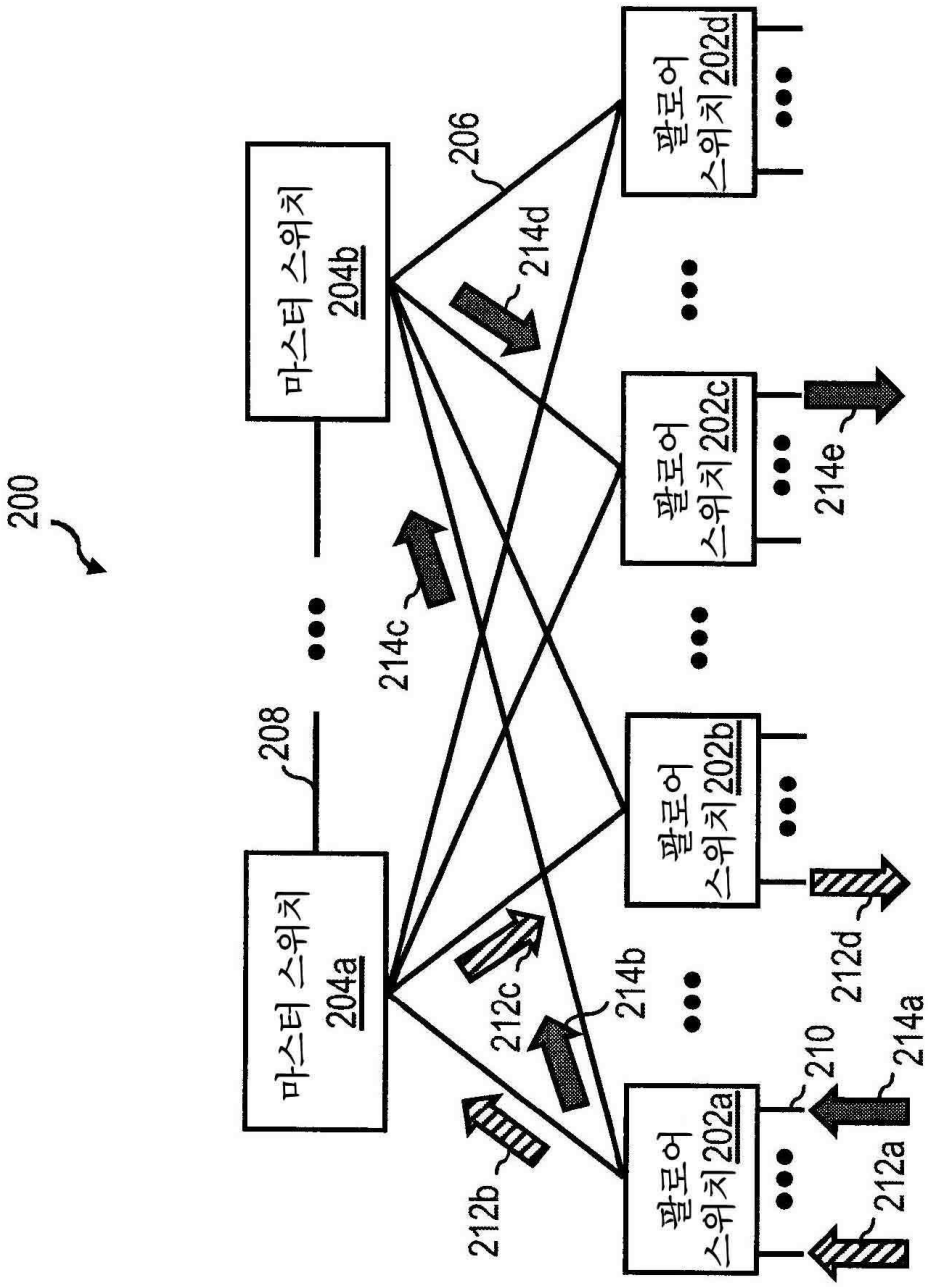
[0083] 본 발명은 하나 또는 그 이상의 실시 예들을 참조하여 기술하면서 구체적으로 도시하고 있지만, 이 기술분야에서 통상의 지식을 가진 자들은 본 발명의 정신과 범위에서 벗어나지 않으면서 형태와 세부사항에서 여러 변경들이 이루어질 수 있다는 것을 이해할 수 있을 것이다. 예를 들면, 본 발명의 특징들은 본 출원에 기술된 평션들에 지시를 내리는 프로그램 코드(예를 들어, 소프트웨어, 펌웨어 또는 이들의 조합)를 실행하는 하나 또는 그 이상의 머신들(예를 들어, 호스트 및/또는 네트워크 스위치)과 관련하여 기술되고 있지만, 실시 예들은 이와는 다르게, 머신에 의해 처리되어 그 머신이 하나 또는 그 이상의 기술된 평션들을 수행하게 할 수 있는 프로그램 코드를 저장하는 유형의(tangible) 머신-판독가능 스토리지 매체 또는 스토리지 디바이스(예를 들어, 광학 스토리지 매체, 메모리 스토리지 매체, 디스크 스토리지 매체 등)를 포함하는 프로그램 제품으로서 구현될 수도 있다.

도면

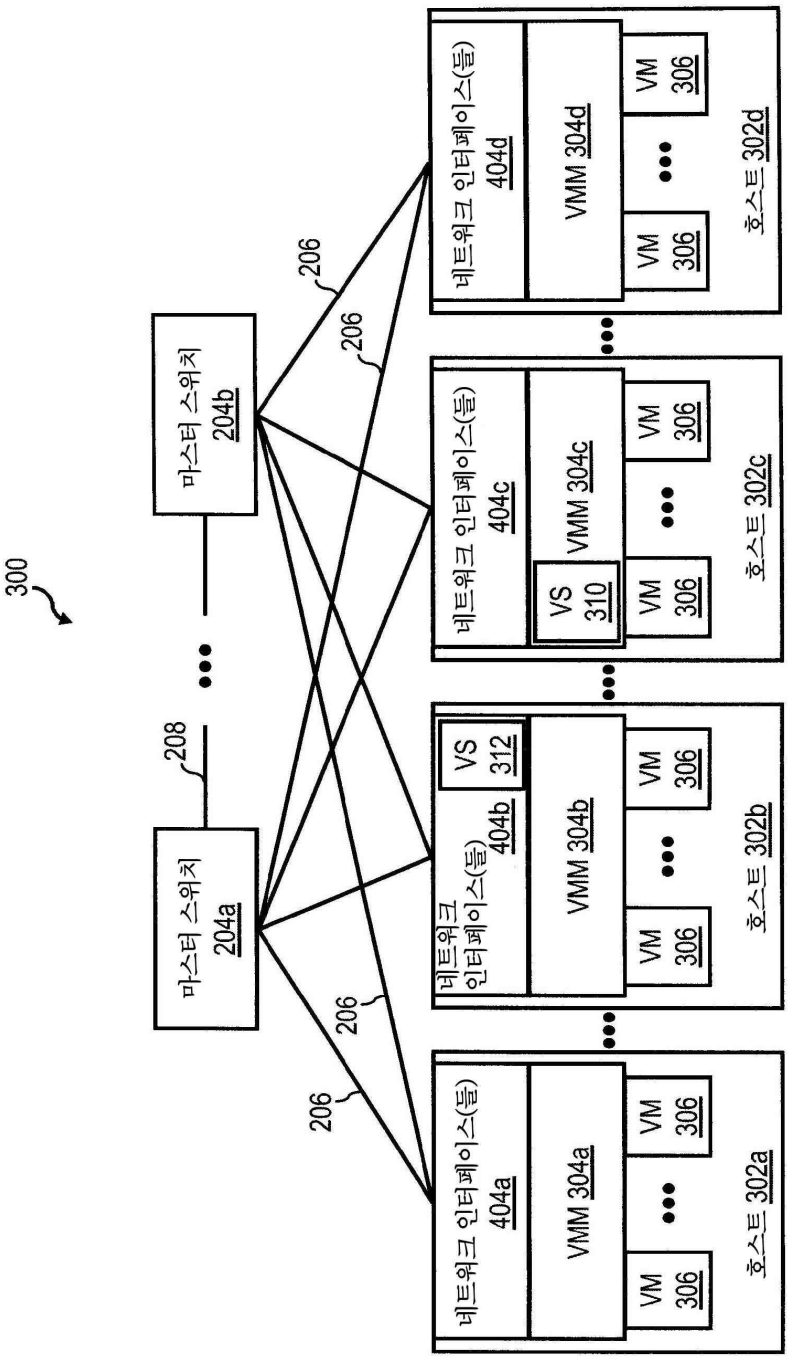
도면1



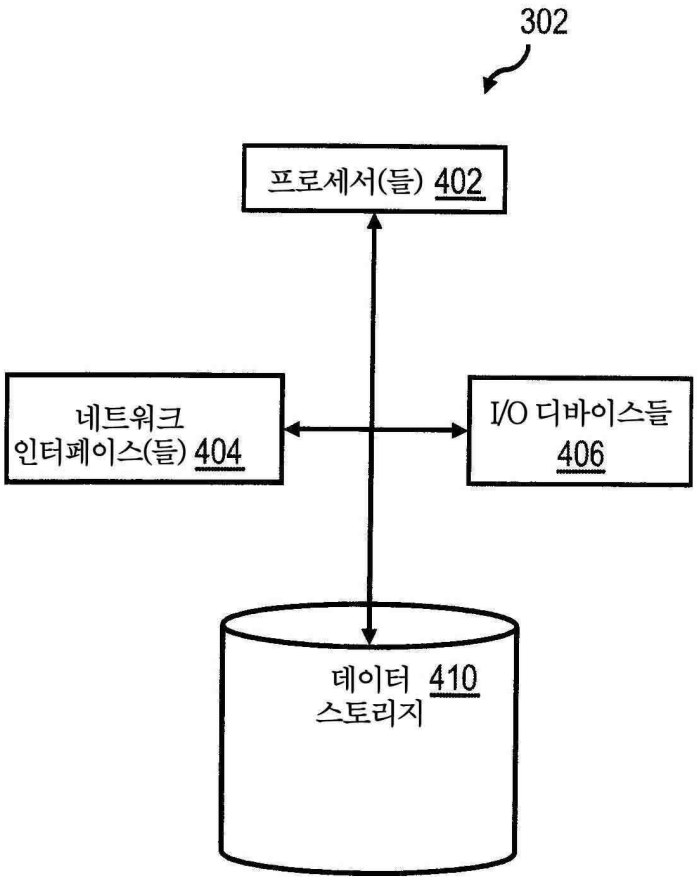
도면2


200
마스터 스위치 204a
마스터 스위치 204b
208
206
팔로어 스위치 202a
팔로어 스위치 202b
팔로어 스위치 202c
팔로어 스위치 202d
210
212a
212b
212c
212d
214a
214b
214c
214d
214e

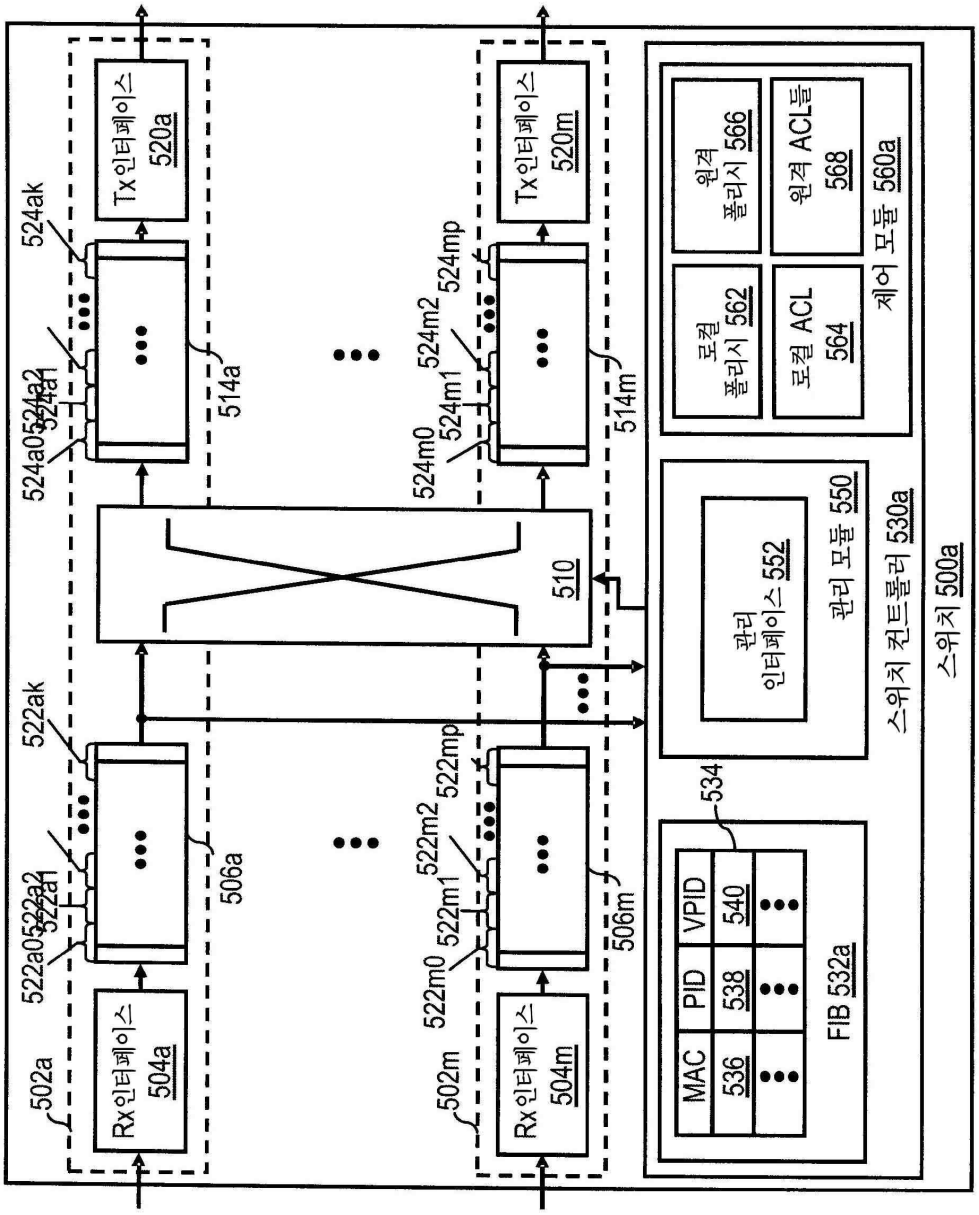
도면3


300
마스터 스위치 204a
마스터 스위치 204b
208
206
206
206
206
네트워크 인터페이스(들) 404a
VMM 304a
VM 306
VM 306
호스트 302a
네트워크 인터페이스(들) 404b
VS 312
VMM 304b
VM 306
VM 306
호스트 302b
네트워크 인터페이스(들) 404c
VS 310
VMM 304c
VM 306
VM 306
호스트 302c
네트워크 인터페이스(들) 404d
VMM 304d
VM 306
VM 306
호스트 302d

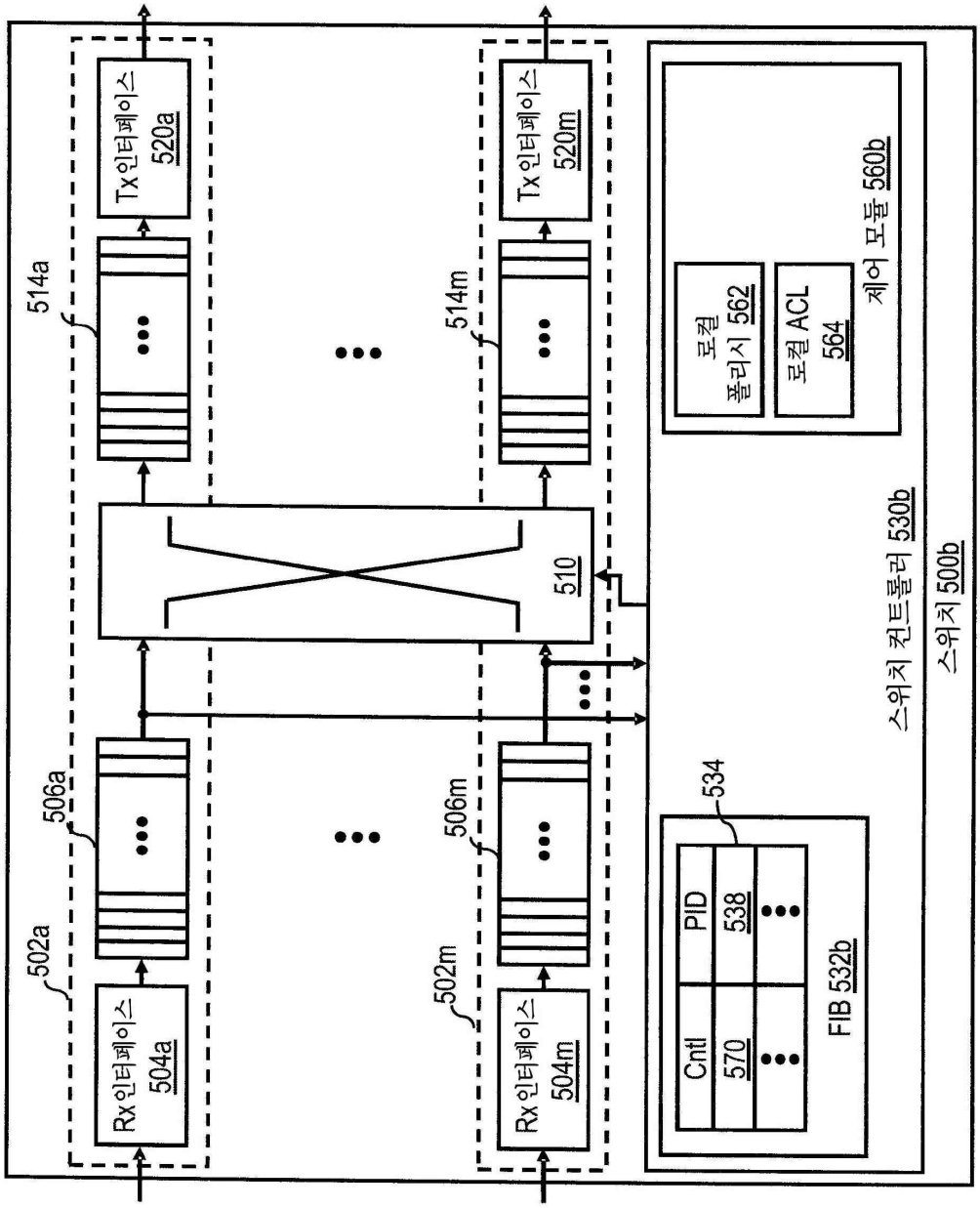
도면4



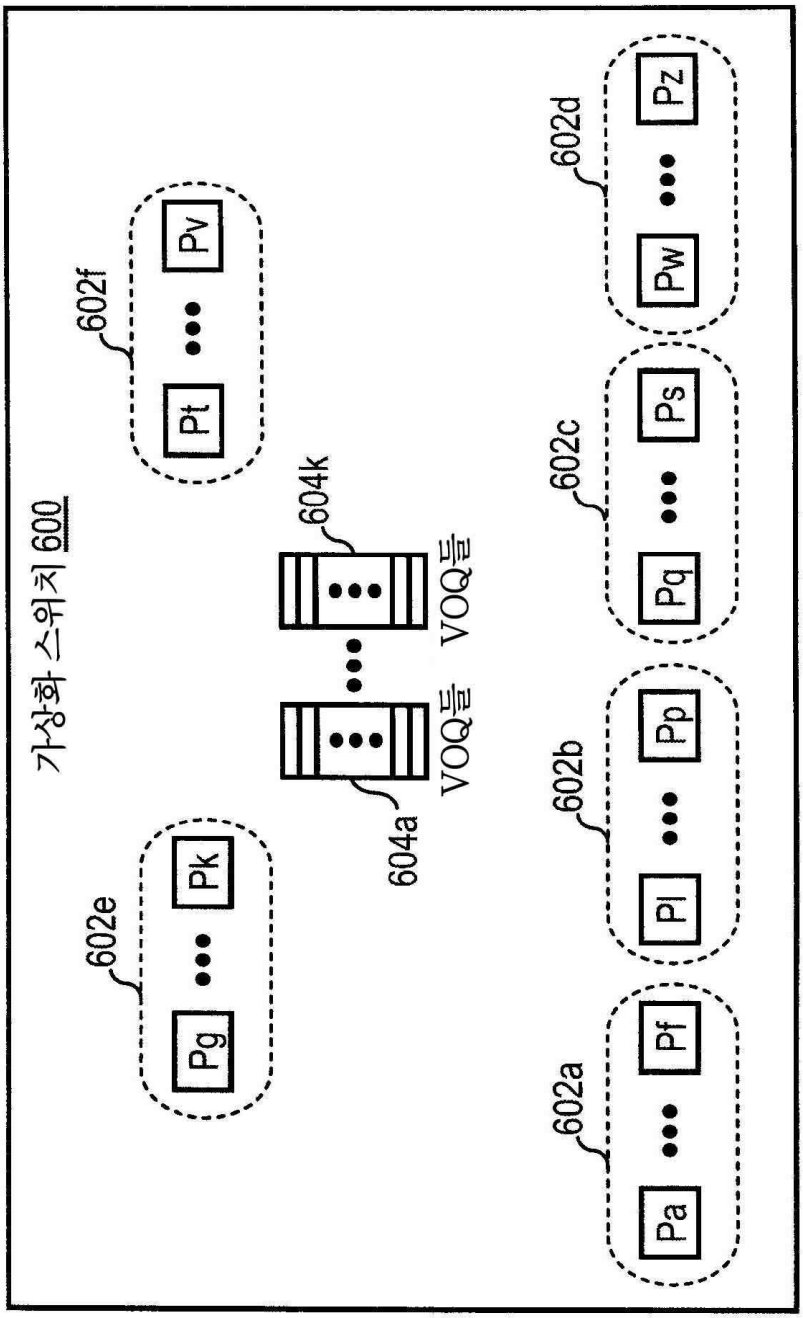
도면5a


스위치 500a
스위치 컨트롤러 530a
제어 모듈 560a
원격 폴리시 566
원격 ACL들 568
로컬 폴리시 562
로컬 ACL 564
관리 모듈 550
관리 인터페이스 552
FIB 532a
MAC 536
PID 538
VPID 540
534
Rx인터페이스 504a
Rx인터페이스 504m
Tx인터페이스 520a
Tx인터페이스 520m
502a
502m
506a
506m
510
514a
514m
522ak
522a0
522a1
522a2
522m0
522m1
522m2
522mp
524ak
524a0
524a1
524a2
524m0
524m1
524m2
524mp

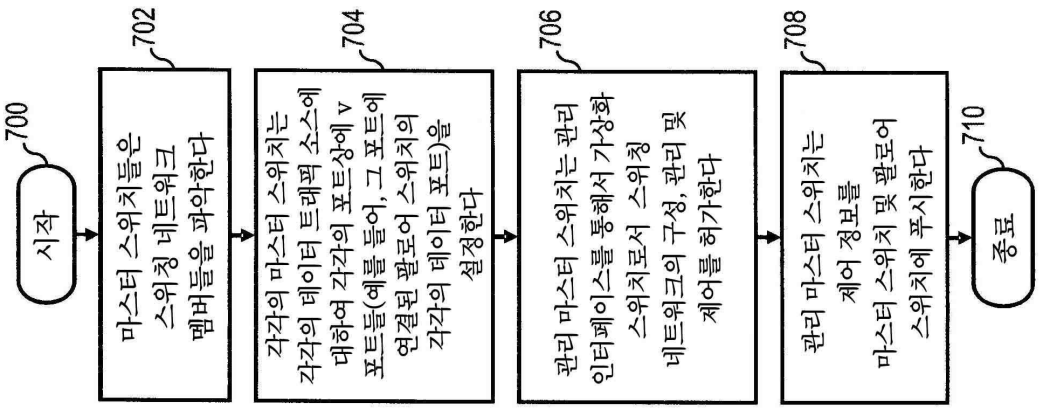
도면5b


Rx 인터페이스 504a
Tx 인터페이스 520a
502a
506a
514a
Rx 인터페이스 504m
Tx 인터페이스 520m
502m
506m
514m
510
스위치 500b
스위치 컨트롤러 530b
제어 모듈 560b
로컬 폴리시 562
로컬 ACL 564
FIB 532b
534
Cntl
570
PID
538

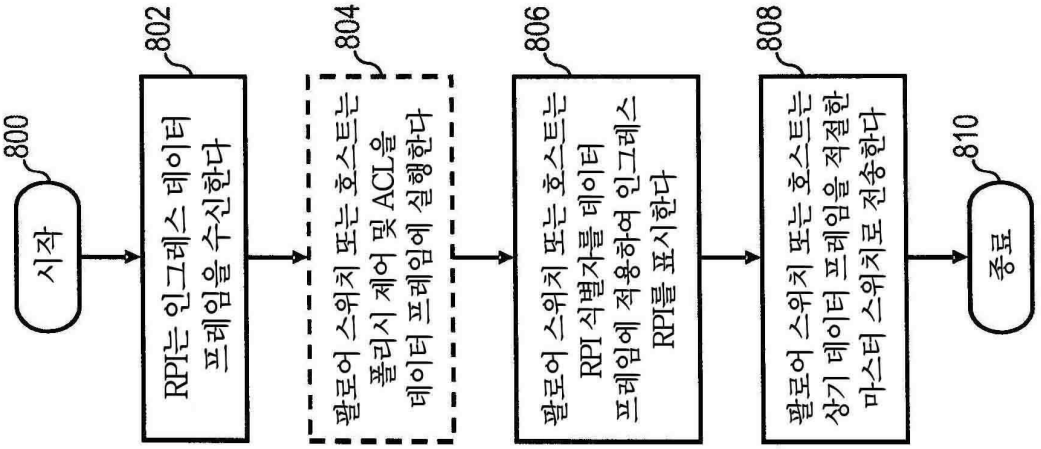
도면6


가상화 스위치 600
602a
Pa
Pf
602b
Pl
Pp
602c
Pq
Ps
602d
Pw
Pz
602e
Pg
Pk
602f
Pt
Pv
604a
VOQ들
604k
VOQ들

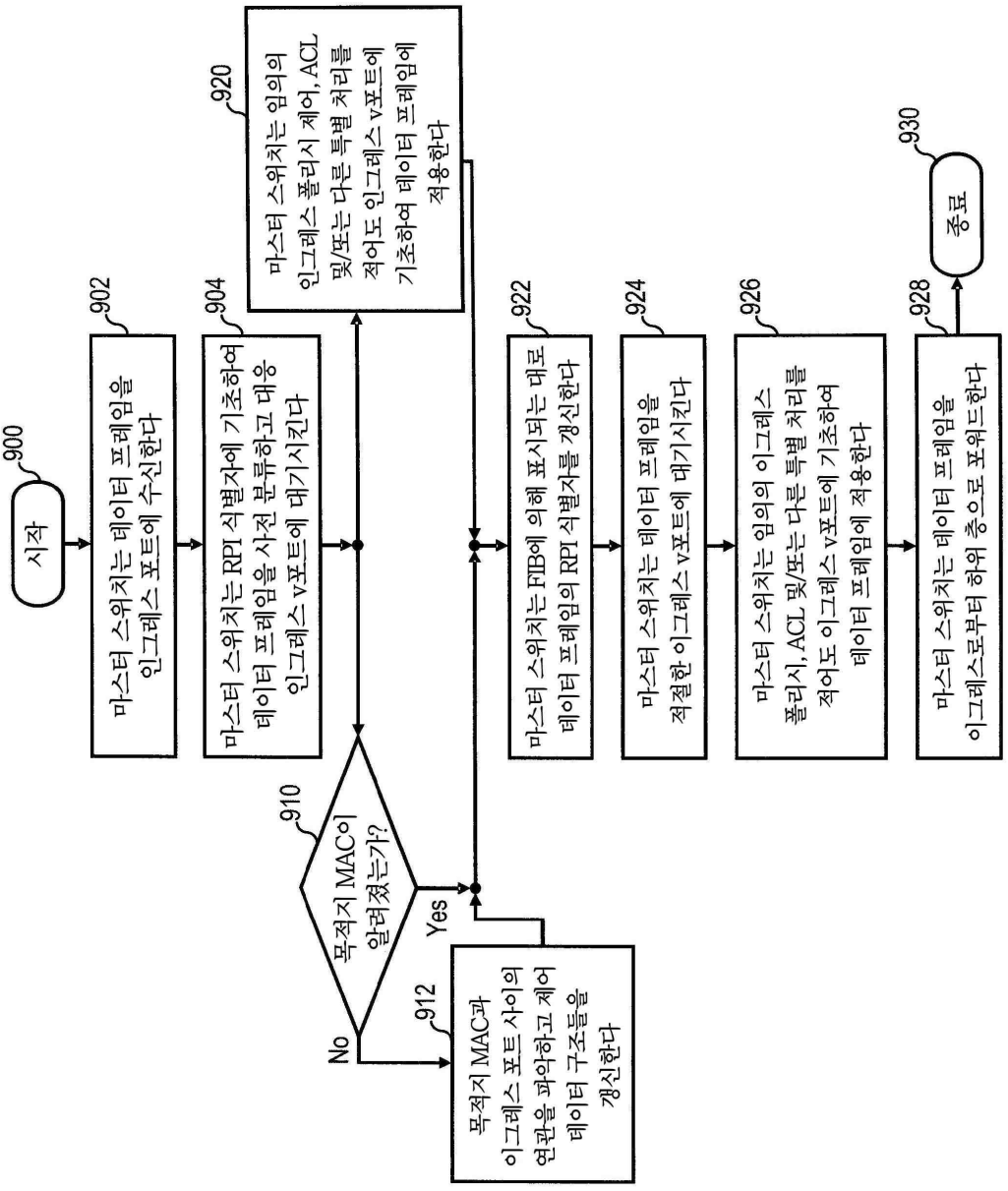
도면7


700
시작
702
마스터 스위치들은 스위칭 네트워크 멤버들을 파악한다
704
각각의 마스터 스위치는 각각의 데이터 트래픽 소스에 대하여 각각의 포트상에 v 포트들(예를 들어, 그 포트에 연결된 팔로어 스위치의 각각의 데이터 포트)을 설정한다
706
관리 마스터 스위치는 관리 인터페이스를 통해서 가상화 스위치로서 스위칭 네트워크의 구성, 관리 및 제어를 허가한다
708
관리 마스터 스위치는 제어 정보를 마스터 스위치 및 팔로어 스위치에 푸시한다
710
종료

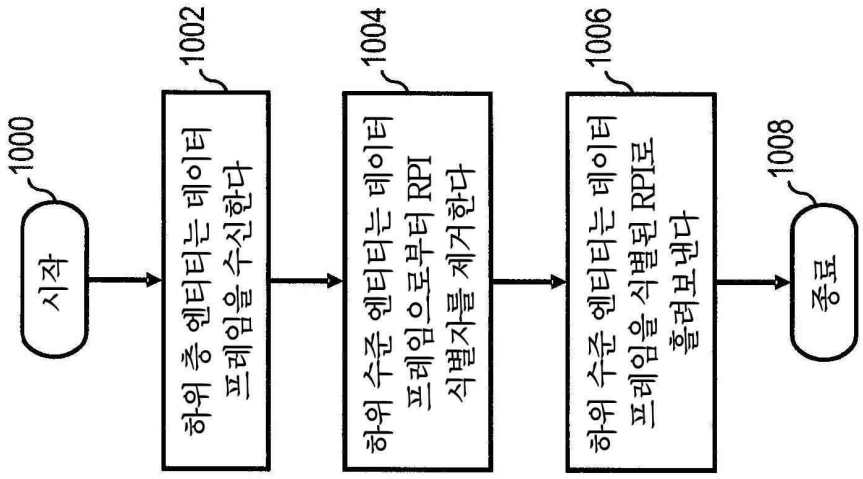
도면8


800
시작
802
RP는 인그레스 데이터 프레임을 수신한다
804
팔로어 스위치 또는 호스트는 폴리시 제어 및 ACL을 데이터 프레임에 실행한다
806
팔로어 스위치 또는 호스트는 RPI 식별자를 데이터 프레임에 적용하여 인그레스 RPI를 표시한다
808
팔로어 스위치 또는 호스트는 상기 데이터 프레임을 적절한 마스터 스위치로 전송한다
810
종료

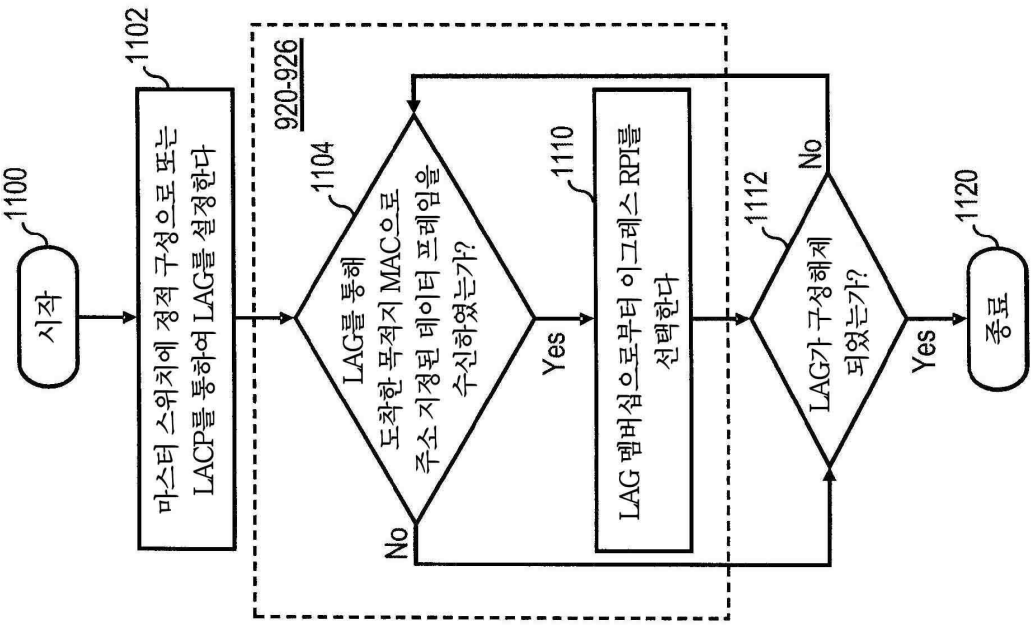
도면9


900
시작
902
마스터 스위치는 데이터 프레임을 인그레스 포트에 수신한다
904
마스터 스위치는 RPI 식별자에 기초하여 데이터 프레임을 사전 분류하고 대응 인그레스 v포트에 대기시킨다
910
목적지 MAC이 알려졌는가?
Yes
No
912
목적지 MAC과 이그레스 포트 사이의 연관을 파악하고 제어 데이터 구조들을 갱신한다
920
마스터 스위치는 임의의 인그레스 폴리시 제어, ACL 및/또는 다른 특별 처리를 적어도 인그레스 v포트에 기초하여 데이터 프레임에 적용한다
922
마스터 스위치는 FIB에 의해 표시되는 대로 데이터 프레임의 RPI 식별자를 갱신한다
924
마스터 스위치는 데이터 프레임을 적절한 이그레스 v포트에 대기시킨다
926
마스터 스위치는 임의의 이그레스 폴리시, ACL 및/또는 다른 특별 처리를 적어도 이그레스 v포트에 기초하여 데이터 프레임에 적용한다
928
마스터 스위치는 데이터 프레임을 이그레스로부터 하위 층으로 포워드한다
930
종료

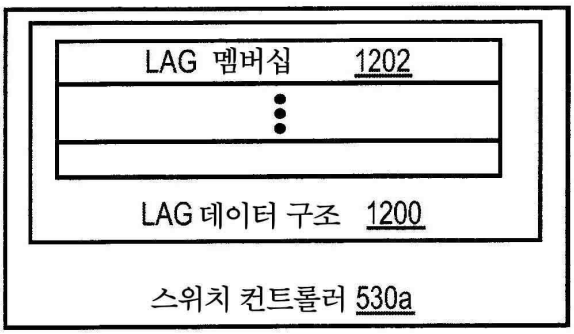
도면10



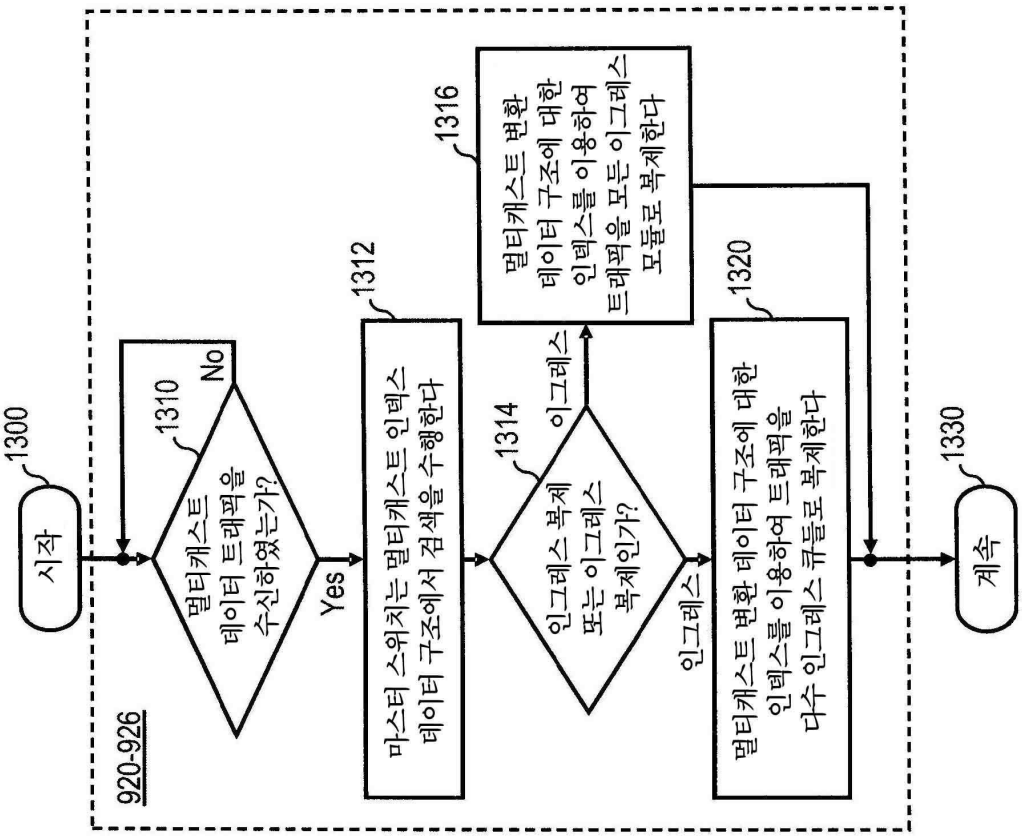
도면11



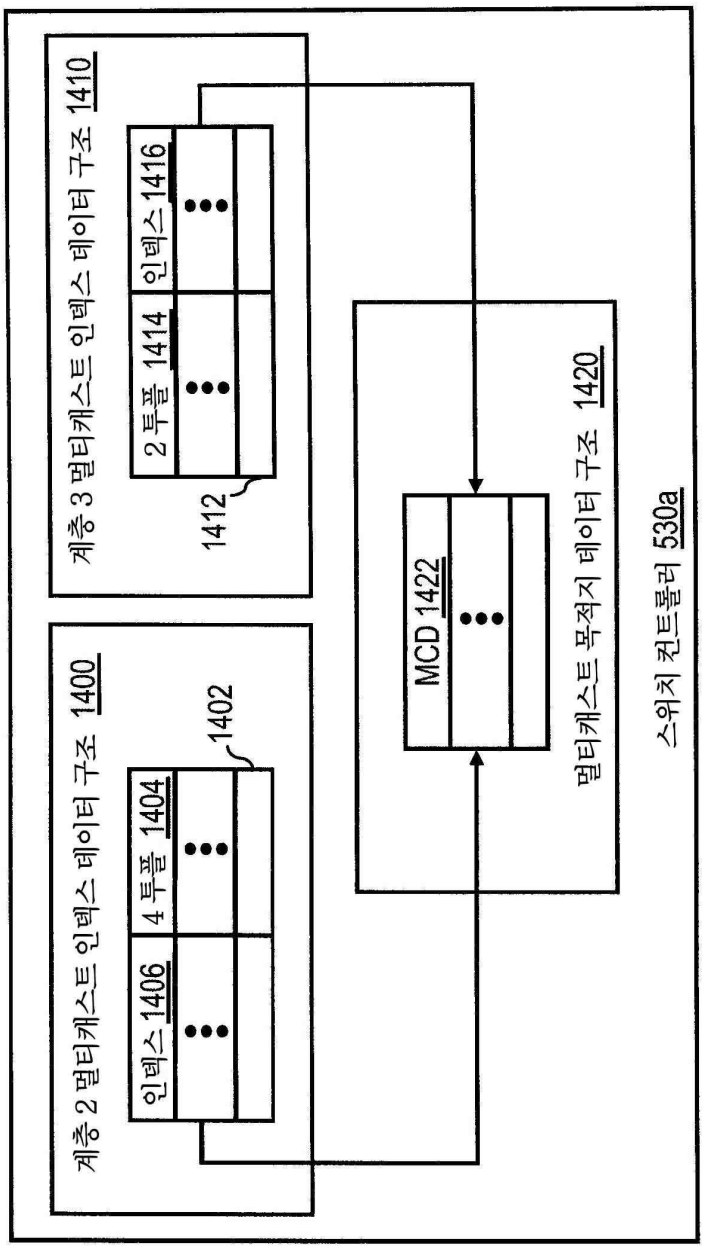
도면12



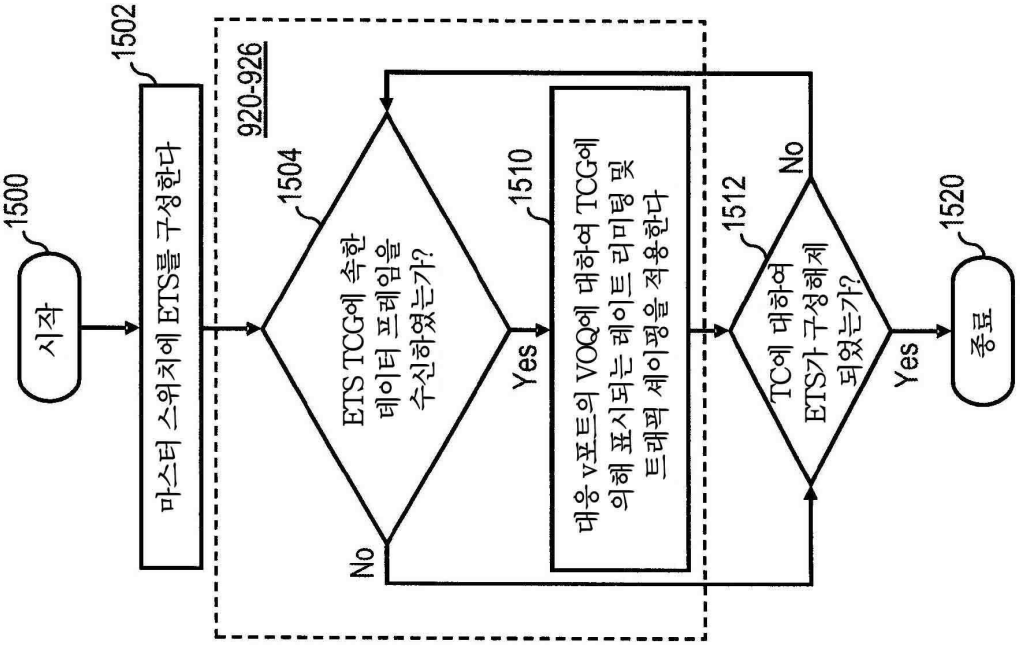
도면13


1300
시작
1310
멀티캐스트 데이터 트래픽을 수신하였는가?
No
Yes
1312
마스터 스위치는 멀티캐스트 인덱스 데이터 구조에서 검색을 수행한다
1314
인그레스 복제 또는 이그레스 복제인가?
이그레스
1316
멀티캐스트 변환 데이터 구조에 대한 인덱스를 이용하여 트래픽을 모든 이그레스 모듈로 복제한다
인그레스
1320
멀티캐스트 변환 데이터 구조에 대한 인덱스를 이용하여 트래픽을 다수 인그레스 큐들로 복제한다
1330
계속
920-926

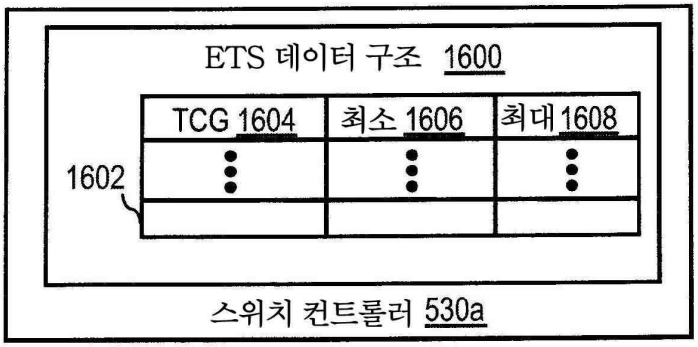
도면14


스위치 컨트롤러 530a
계층 2 멀티캐스트 인덱스 데이터 구조 1400
인덱스 1406
4 투플 1404
1402
계층 3 멀티캐스트 인덱스 데이터 구조 1410
2 투플 1414
인덱스 1416
1412
멀티캐스트 목적지 데이터 구조 1420
MCD 1422

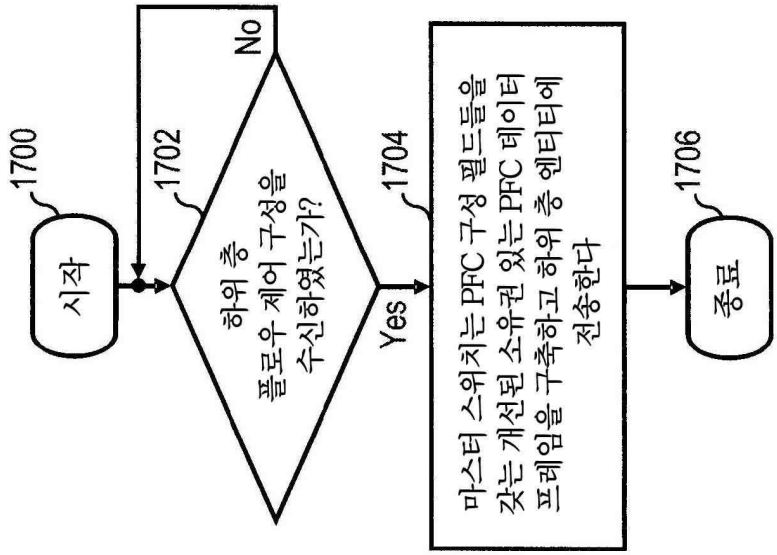
도면15


1500
시작
1502
마스터 스위치에 ETS를 구성한다
920-926
1504
ETS TCG에 속한 데이터 프레임을 수신하였는가?
No
Yes
1510
대응 v포트의 VOQ에 대하여 TCG에 의해 표시되는 레이트 리미팅 및 트래픽 셰이핑을 적용한다
1512
TC에 대하여 ETS가 구성해제 되었는가?
No
Yes
1520
종료

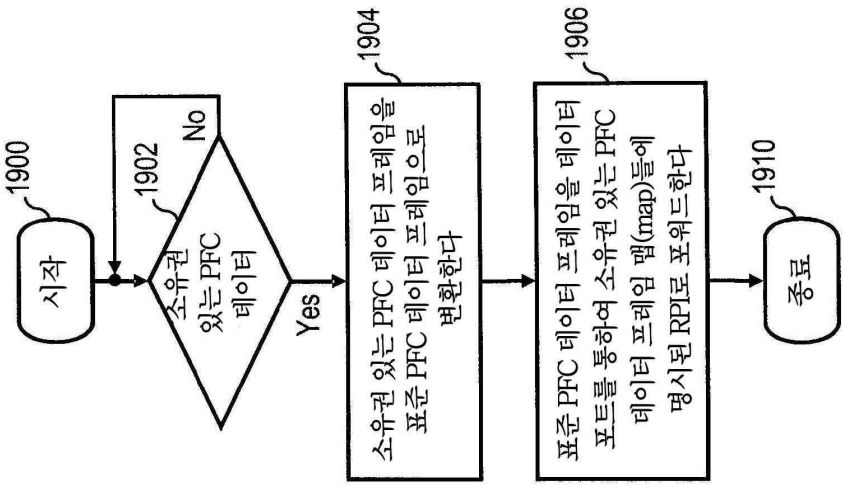
도면16


ETS 데이터 구조 1600
TCG 1604
최소 1606
최대 1608
1602
스위치 컨트롤러 530a

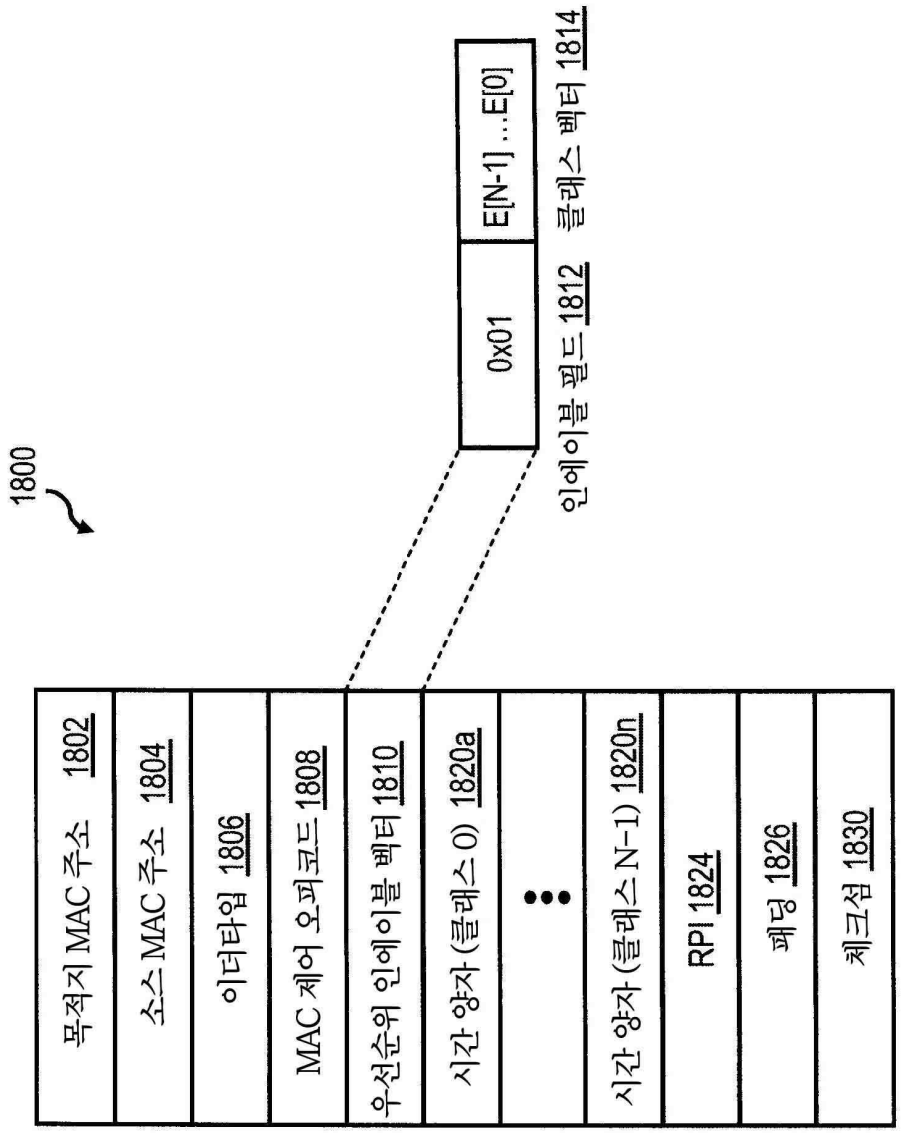
도면17


1700
시작
1702
하위 층
플로우 제어 구성을
수신하였는가?
No
Yes
1704
마스터 스위치는 PFC 구성 필드들을
갖는 개선된 소유권 있는 PFC 데이터
프레임을 구축하고 하위 층 엔티티에
전송한다
1706
종료

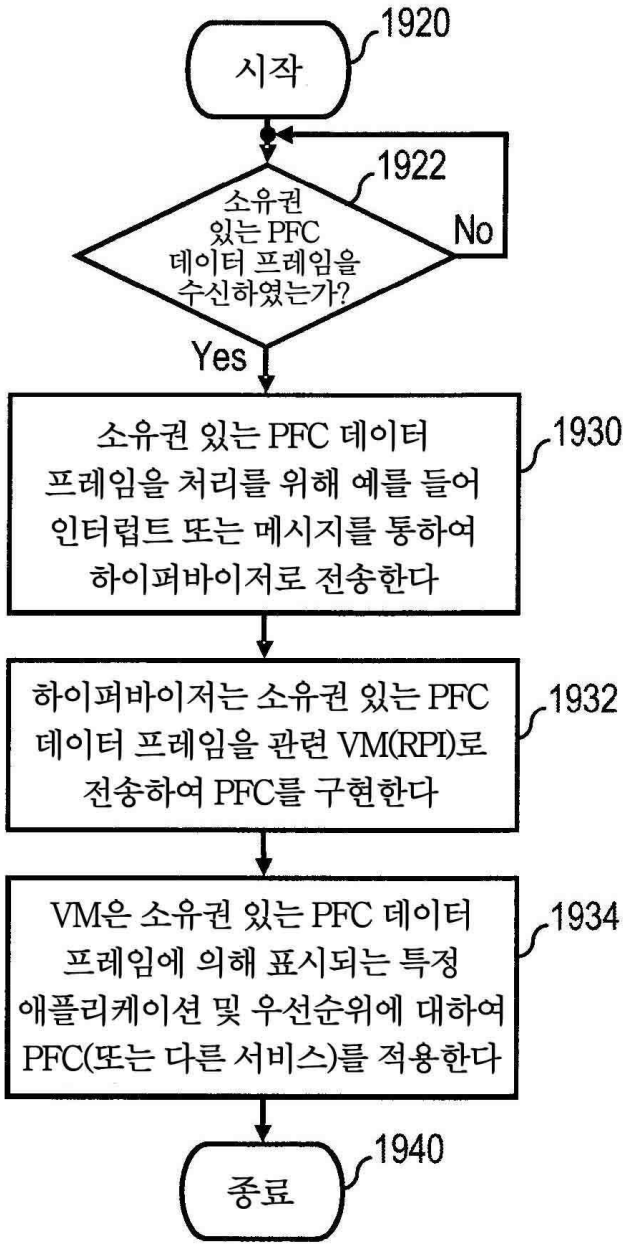
도면19a


1900
시작
1902
소유권
있는 PFC
데이터
No
Yes
1904
소유권 있는 PFC 데이터 프레임을
표준 PFC 데이터 프레임으로
변환한다
1906
표준 PFC 데이터 프레임을 데이터
포트를 통하여 소유권 있는 PFC
데이터 프레임 맵(map)들에
명시된 RP로 포워드한다
1910
종료

도면18


1800
목적지 MAC 주소 1802
소스 MAC 주소 1804
이더타입 1806
MAC 제어 오피코드 1808
우선순위 인에이블 벡터 1810
시간 양자 (클래스 0) 1820a
...
시간 양자 (클래스 N-1) 1820n
RPI 1824
패딩 1826
체크섬 1830
인에이블 필드 1812
0x01
E[N-1] ...E[0]
클래스 벡터 1814

도면19b


1920
시작
1922
소유권 있는 PFC 데이터 프레임을 수신하였는가?
No
Yes
소유권 있는 PFC 데이터 프레임을 처리를 위해 예를 들어 인터럽트 또는 메시지를 통하여 하이퍼바이저로 전송한다
1930
하이퍼바이저는 소유권 있는 PFC 데이터 프레임을 관련 VM(RPI)로 전송하여 PFC를 구현한다
1932
VM은 소유권 있는 PFC 데이터 프레임에 의해 표시되는 특정 애플리케이션 및 우선순위에 대하여 PFC(또는 다른 서비스)를 적용한다
1934
종료
1940