(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년01월22일
(11) 등록번호 10-2205882
(24) 등록일자 2021년01월15일

(51) 국제특허분류(Int. Cl.)
H04L 12/741 (2013.01) *H04L 12/721* (2013.01)
(21) 출원번호 10-2014-7034088
(22) 출원일자(국제) 2013년05월08일
심사청구일자 2018년01월19일
(85) 번역문제출일자 2014년12월03일
(65) 공개번호 10-2015-0016309
(43) 공개일자 2015년02월11일
(86) 국제출원번호 PCT/US2013/040212
(87) 국제공개번호 WO 2013/169949
국제공개일자 2013년11월14일
(30) 우선권주장
13/889,088 2013년05월07일 미국(US)
(뒷면에 계속)
(56) 선행기술조사문헌
US20090141728 A1*
US20070189310 A1*
US20050213585 A1*
JP2002521860 A
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
오라클 인터내셔날 코포레이션
미국, 캘리포니아 94065, 레드우드 쇼어스 엠에스 5오피7, 오라클 파크웨이 500
(72) 발명자
복단스키 바르토즈
노르웨이 오슬로 엔-0275 에이치0203 호프 테라쎄 15
존센 본 닥
노르웨이 오슬로 엔-0687 빌베르그렌다 9
(74) 대리인
박장원

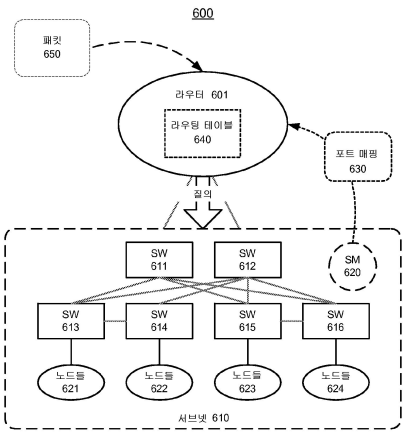
전체 청구항 수 : 총 21 항 심사관 : 이성영

(54) 발명의 명칭 **팻-트리 라우팅에 기반하여 별개의 인피니밴드 서브넷들 사이에서 트래픽을 라우팅하기 위한 시스템 및 방법**

(57) 요 약

시스템 및 방법은 네트워크 환경에서 별개의 서브넷들 사이에서 트래픽을 라우팅할 수 있다. 인피니밴드(IB) 서브넷들과 같은 별개의 서브넷들을 연결하는 라우터는 목적지들의 리스트를 수신할 수 있으며, 라우터는 그 리스트 내의 목적지들에 하나 이상의 패킷들을 라우팅하는 것을 담당한다. 더욱이, 라우터는 하나 이상의 패킷들을 라우팅하기 위하여 라우터의 어느 하향 출력 포트들이 사용될 수 있는지에 대한 정보를 적어도 하나의 서브넷의 하나 이상의 스위치들로부터 획득하며 획득된 정보에 기반하여 라우팅 테이블을 구축할 수 있다.

대 표 도



(30) 우선권주장
13/889,123 2013년05월07일 미국(US)
61/646,107 2012년05월11일 미국(US)

명 세 서

청구범위

청구항 1

제1 팻-트리 토폴로지를 갖는 제1 서브넷의 복수의 소스 노드들과 제2 팻-트리 토폴로지를 갖는 제2 서브넷의 복수의 목적지 노드들 사이에서 트래픽을 라우팅하기 위한 방법으로서, 상기 제1 서브넷과 제2 서브넷은 상기 제1 서브넷에 연결된 복수의 입구(ingress) 포트들과 상기 제2 서브넷에 연결된 복수의 출구(egress) 포트들을 갖는 라우터에 의해 직접 연결되며, 상기 방법은:

상기 라우터에서, 상기 제1 서브넷으로부터, 상기 목적지 노드들의 리스트를 수신하는 단계;

상기 라우터에서, 상기 제2 서브넷의 서브넷 관리자로부터, 상기 라우터에서 상기 제2 서브넷의 상기 제2 팻-트리 토폴로지의 포트 매핑 리스트를 수신하는 단계, 상기 포트 매핑 리스트는 상기 복수의 목적지 노드들을 포함하며;

상기 포트 매핑 리스트로부터, 상기 제2 서브넷에서 복수의 주 스위치들(primary switches)을 식별하는 단계, 상기 복수의 주 스위치들 각각은 상기 제2 서브넷에서 상기 복수의 목적지 노드들 중 하나 이상에 대한 주 경로를 포함하며;

상기 제1 서브넷에서 상기 소스 노드들 각각으로부터 상기 라우터의 상기 입구 포트들 중 하나로의 경로를 설정하기 위해 입구 포트 매핑 파일을 사용하는 단계, 상기 입구 포트 매핑 파일은 상기 복수의 목적지 노드들의 유사한 수를 라운드 로빈 분배를 사용하여 상기 복수의 입구 포트들 각각과 연관시키며;

상기 라우터에서, 상기 복수의 주 스위치들 각각을 통해 상기 라우터의 상기 복수의 출구 포트들 각각 마다 하나의 경로를 설정하여, 그에 의해 상기 제2 서브넷에서 상기 복수의 목적지 노드들 각각에 대한 경로를 설정하는 라우팅 테이블을 구축하는 단계를 포함하는 것을 특징으로 하는 방법.

청구항 2

제1항에 있어서,

상기 제1 서브넷과 제2 서브넷은 상기 제1 서브넷에 연결된 복수의 입구 포트들과 상기 제2 서브넷에 연결된 복수의 출구 포트들을 각각 가지는 복수의 라우터들에 의해 직접 연결되며, 상기 방법은:

상기 제1 서브넷과 상기 제2 서브넷에 연결되는 상기 복수의 라우터들 사이에서 라운드-로빈 경로 분배(round-robin path distribution)를 수행하는 단계를 더 포함하는 것을 특징으로 하는 방법.

청구항 3

제1항에 있어서,

상기 라우터에서, 상기 제2 서브넷의 상기 서브넷 관리자가 상기 제1 서브넷의 서브넷 관리자와 통신하는 인터페이스를 제공하는 단계를 더 포함하는 것을 특징으로 하는 방법.

청구항 4

제1항에 있어서,

상기 라우팅 테이블에 기반하여 지연을 최소화하는 것에 있어서 최대 성능을 위하여 상기 라우터의 어느 출구 포트들이 사용될 수 있는지를 결정하는 단계를 더 포함하는 것을 특징으로 하는 방법.

청구항 5

제1항에 있어서,

상기 라우터의 펌웨어에서, 상기 주 스위치들을 식별하고, 상기 포트 매핑 리스트를 사용하여 상기 복수의 출구 포트들로부터 출구 포트를 선택하기 위한 시스템을 구현하는 단계를 더 포함하는 것을 특징으로 하는 방법.

청구항 6

제1항에 있어서,

하나 이상의 추가 라우터들을 사용하여 하나 이상의 추가 인접 서브넷들에 상기 제1 서브넷을 연결하는 단계를 더 포함하는 것을 특징으로 하는 방법.

청구항 7

제1항에 있어서,

상기 입구 포트 매핑 파일은 상기 목적지 노드들의 목적지 주소들에 기초하여 상기 복수의 입구 포트들에 걸쳐 경로들을 분배하는 것을 특징으로 하는 방법.

청구항 8

제1항에 있어서,

상기 입구 포트 매핑 파일은 상기 라우터의 상기 복수의 입구 포트들을 선택하는 것에 있어서 최대 입상도(full granularity)를 제공하는 것을 특징으로 하는 방법.

청구항 9

제1항에 있어서,

소스 노드와 목적지 노드의 각 조합에 대해, 모든 메시지들은 동일한 입구 포트와 출구 포트를 통해 라우팅되어, 메시지들의 비순차 전달을 방지하는 것을 특징으로 하는 방법.

청구항 10

제1항에 있어서,

상기 제1 서브넷을 라우팅하기 위하여 팻-트리 라우팅 알고리즘을 사용하는 단계를 더 포함하는 것을 특징으로 하는 방법.

청구항 11

컴퓨터 프로그램이 저장된 컴퓨터 판독가능 저장 매체로서, 상기 컴퓨터 프로그램은 제1항 내지 제10항 중 어느 한 항의 방법의 모든 단계들을 수행하도록 하나 이상의 마이크로프로세서들에서 실행하기 위한 프로그램 명령어들을 포함하는 것을 특징으로 하는 컴퓨터 판독가능 저장 매체.

청구항 12

삭제

청구항 13

제1 팻-트리 토폴로지를 갖는 제1 서브넷의 복수의 소스 노드들과 제2 팻-트리 토폴로지를 갖는 제2 서브넷의 복수의 목적지 노드들 사이에서 트래픽을 라우팅하기 위한 시스템으로서, 상기 제1 서브넷과 제2 서브넷은 상기 제1 서브넷에 연결된 복수의 입구(ingress) 포트들과 상기 제2 서브넷에 연결된 복수의 출구(egress) 포트들을 갖는 라우터에 의해 직접 연결되며, 상기 시스템은:

하나 이상의 마이크로프로세서들 및

상기 하나 이상의 마이크로프로세서들에서 실행되는 라우터를 포함하며, 상기 라우터는 단계들을 수행하도록 구성되며, 상기 단계들은:

상기 라우터에서, 상기 제1 서브넷으로부터, 상기 목적지 노드들의 리스트를 수신하는 단계;

상기 라우터에서, 상기 제2 서브넷의 서브넷 관리자로부터, 상기 라우터에서 상기 제2 서브넷의 상기 제2 팻-트리 토폴로지의 포트 매핑 리스트를 수신하는 단계, 상기 포트 매핑 리스트는 상기 복수의 목적지 노드들을 포함하며;

상기 포트 매핑 리스트로부터, 상기 제2 서브넷에서 복수의 주 스위치들(primary switches)을 식별하는 단계, 상기 복수의 주 스위치들 각각은 상기 제2 서브넷에서 상기 복수의 목적지 노드들 중 하나 이상에 대한 주 경로를 포함하며;

상기 제1 서브넷에서 상기 소스 노드들 각각으로부터 상기 라우터의 상기 입구 포트들 중 하나로의 경로를 설정하기 위해 입구 포트 매핑 파일을 사용하는 단계, 상기 입구 포트 매핑 파일은 상기 복수의 목적지 노드들의 유사한 수를 라운드 로빈 분배를 사용하여 상기 복수의 입구 포트들 각각과 연관시키며;

상기 라우터에서, 상기 복수의 주 스위치들 각각을 통해 상기 라우터의 상기 복수의 출구 포트들 각각 마다 하나의 경로를 설정하여, 그에 의해 상기 제2 서브넷에서 상기 복수의 목적지 노드들 각각에 대한 경로를 설정하는 라우팅 테이블을 구축하는 단계를 포함하는 것을 특징으로 하는 시스템.

청구항 14

제13항에 있어서,

라운드-로빈 경로 분배가 상기 제1 서브넷과 상기 제2 서브넷에 연결되는 상기 복수의 라우터들 사이에서 수행되는 것을 특징으로 하는 시스템.

청구항 15

제13항에 있어서,

상기 라우터는 어떤 스위치가 특정 목적지로 향하는 주 경로로서 역할을 할 수 있는지를 알기 위하여 상기 라우터에 연결되는 스위치들에 질의하도록 동작하는 것을 특징으로 하는 시스템.

청구항 16

제13항에 있어서,

상기 라우터는 상기 라우팅 테이블에 기반하여 지연을 최소화하는 것에 있어서 최대 성능을 위하여 상기 라우터의 어느 출구 포트들이 사용될 수 있는지를 결정하도록 동작하는 것을 특징으로 하는 시스템.

청구항 17

제13항에 있어서,

상기 라우터는 상기 주 스위치들을 식별하고, 상기 포트 매핑 리스트를 사용하여 상기 복수의 출구 포트들로부터 출구 포트를 선택하기 위한 시스템을 구현하는 펌웨어를 포함하는 것을 특징으로 하는 시스템.

청구항 18

제13항에 있어서,

상기 제1 서브넷은 하나 이상의 추가 라우터들을 사용하여 하나 이상의 추가 인접 서브넷들에 연결되는 것을 특징으로 하는 시스템.

청구항 19

제13항에 있어서,

상기 입구 포트 매핑 파일은 상기 목적지 노드들의 목적지 주소들에 기초하여 상기 복수의 입구 포트들에 걸쳐 경로들을 분배하는 것을 특징으로 하는 시스템.

청구항 20

제13항에 있어서,

상기 입구 포트 매핑 파일은 상기 시스템으로 하여금 상기 라우터의 상기 복수의 입구 포트들을 선택하는 것에서 최대 입상도(full granularity)를 제공하게 하는 것을 특징으로 하는 시스템.

청구항 21

제13항에 있어서,

소스 노드와 목적지 노드의 각 조합에 대해, 모든 메시지들은 동일한 입구 포트와 출구 포트를 통해 라우팅되어, 메시지들의 비순차 전달을 방지하는 것을 특징으로 하는 시스템.

청구항 22

제1 팻-트리 토폴로지를 갖는 제1 서브넷의 복수의 소스 노드들과 제2 팻-트리 토폴로지를 갖는 제2 서브넷의 복수의 목적지 노드들 사이에서 트래픽을 라우팅하기 위한 명령어들이 저장된 비-일시적 머신 판독가능 저장 매체로서, 상기 제1 서브넷과 제2 서브넷은 상기 제1 서브넷에 연결된 복수의 입구(ingress) 포트들과 상기 제2 서브넷에 연결된 복수의 출구(egress) 포트들을 갖는 라우터에 의해 직접 연결되며, 상기 명령어들은 실행될 때 시스템으로 단계들을 수행하게 하며, 상기 단계들은:

상기 라우터에서, 상기 제1 서브넷으로부터, 상기 목적지 노드들의 리스트를 수신하는 단계;

상기 라우터에서, 상기 제2 서브넷의 서브넷 관리자로부터, 상기 라우터에서 상기 제2 서브넷의 상기 제2 팻-트리 토폴로지의 포트 매핑 리스트를 수신하는 단계, 상기 포트 매핑 리스트는 상기 복수의 목적지 노드들을 포함하며;

상기 포트 매핑 리스트로부터, 상기 제2 서브넷에서 복수의 주 스위치들(primary switches)을 식별하는 단계, 상기 복수의 주 스위치들 각각은 상기 제2 서브넷에서 상기 복수의 목적지 노드들 중 하나 이상에 대한 주 경로를 포함하며;

상기 제1 서브넷에서 상기 소스 노드들 각각으로부터 상기 라우터의 상기 입구 포트들 중 하나로의 경로를 설정하기 위해 입구 포트 매핑 파일을 사용하는 단계, 상기 입구 포트 매핑 파일은 상기 복수의 목적지 노드들의 유사한 수를 라운드 로빈 분배를 사용하여 상기 복수의 입구 포트들 각각과 연관시키며;

상기 라우터에서, 상기 복수의 주 스위치들 각각을 통해 상기 라우터의 상기 복수의 출구 포트들 각각 마다 하나의 경로를 설정하여, 그에 의해 상기 제2 서브넷에서 상기 복수의 목적지 노드들 각각에 대한 경로를 설정하는 라우팅 테이블을 구축하는 단계를 포함하는 것을 특징으로 하는 비-일시적 머신 판독가능 저장 매체.

청구항 23

삭제

발명의 설명

기 술 분 야

[0001] 본 특허문헌의 개시내용의 일부분은 저작권 보호를 받는 자료를 포함한다. 저작권자는 특허상표청의 특허 파일 또는 기록들에 나타난 대로 본 특허 문헌 또는 특허 개시내용을 어느 누군가가 팩시밀리 재생하는 것은 반대하지는 않지만, 그 밖의 모든 것은 저작권으로 보호된다.

[0002] 본 발명은 일반적으로 컴퓨터 시스템들에 관한 것이며, 특히 미들웨어 머신 환경에 관한 것이다.

배 경 기 술

[0003] 대규모 클라우드 컴퓨팅 아키텍처들이 도입됨에 따라, 기존의 네트워크 및 스토리지와 연관된 성능 및 관리 병목(bottleneck)들은 중요한 문제가 되어 왔다. 인피니밴드(IB: InfiniBand) 기술은 클라우드 컴퓨팅 패브릭의 토대로서 증가된 디플로이먼트를 보여 왔다. 이는 본 발명의 실시예들이 다루도록 의도된 일반적인 분야이다.

발명의 내용

[0004] 본원에서 설명된 것은 별개의 인피니밴드(IB) 서브넷들 사이에서 트래픽을 라우팅하는 시스템 및 방법이다. 인피니밴드(IB) 서브넷들과 같은 별개의 서브넷들을 연결하는 라우터는 목적지들의 리스트를 수신할 수 있으며, 라우터는 그 리스트 내의 목적지들에 하나 이상의 패킷들을 라우팅하는 것을 담당한다. 더욱이, 라우터는 하나 이상의 패킷들을 라우팅하기 위하여 라우터의 어느 하향 출력 포트들이 사용될 수 있는지에 대한 정보를 적어도 하나의 서브넷의 하나 이상의 스위치들로부터 획득하고 그리고 획득된 정보에 기반하여 라우팅 테이블을 구축할 수 있다.

[0005] 본원에서 또한 설명된 것은 네트워크 환경에서 별개의 서브넷들 사이에서 트래픽을 라우팅하기 위한 시스템이다. 시스템은 라우터에서 목적지들의 리스트를 수신하기 위한 수단을 포함하며, 라우터는 그 리스트 내의 목적지들에 하나 이상의 패킷들을 라우팅하는 것을 담당하며, 라우터는 네트워크 환경의 적어도 하나의 서브넷에 연결된다. 시스템은 하나 이상의 패킷들을 라우팅하기 위하여 라우터의 어느 하향 출력 포트들이 사용될 수 있는지에 대한 정보를 적어도 하나의 서브넷의 하나 이상의 스위치들로부터 획득하기 위한 수단을 더 포함한다. 시스템은 획득된 정보에 기반하여 라우팅 테이블을 구축하기 위한 수단을 더 포함한다.

[0006] 본원에서 추가로 설명된 것은 네트워크 환경에서 별개의 서브넷들 사이에서 트래픽을 라우팅하기 위한 시스템이다. 시스템은 목적지들의 리스트를 수신하도록 구성된 라우터를 포함하며, 라우터는 그 리스트 내의 목적지들에 하나 이상의 패킷들을 라우팅하는 것을 담당하며, 라우터는 네트워크 환경의 적어도 하나의 서브넷에 연결된다. 시스템은 또한 하나 이상의 패킷들을 라우팅하기 위하여 라우터의 어느 하향 출력 포트들이 사용될 수 있는지에 대한 정보를 적어도 하나의 서브넷의 하나 이상의 스위치들로부터 획득하며 그리고 획득된 정보에 기반하여 라우팅 테이블을 구축하도록 구성된다.

도면의 간단한 설명

[0007] 도 1은 본 발명의 실시예에 따라, 네트워크 환경에서 별개의 인피니밴드(IB) 서브넷들 사이에서 트래픽을 라우팅하는 예시를 도시한다.

도 2는 본 발명의 실시예에 따라, 네트워크 환경에서 라우터를 통한 패킷 포워딩을 지원하는 예시를 도시한다.

도 3은 본 발명의 실시예에 따라, 네트워크 환경에서 상이한 목적지들에 대한 입구 라우터 포트들을 선정하는 예시를 도시한다.

도 4는 본 발명의 실시예에 따라, 네트워크 환경의 라우터에서, 수신된 패킷의 출구 라우터 포트를 선택하는 예시를 도시한다.

도 5는 본 발명의 실시예에 따라, 네트워크 환경의 라우터에서 서브넷 간 소스 라우팅(ISSR) 알고리즘을 지원하기 위한 예시적인 흐름도를 예시한다.

도 6은 본 발명의 실시예에 따라, 네트워크 환경의 라우터에서 서브넷 간 팻-트리 라우팅(ISFR) 알고리즘을 지원하는 예시를 도시한다.

도 7-9는 본 발명의 실시예에 따라, 상이한 토폴로지들을 가진 네트워크 환경에서의 라우팅을 예시한다.

도 10은 본 발명의 실시예에 따라, 네트워크 환경의 라우터에서 서브넷 간 팻-트리 라우팅(ISFR) 알고리즘을 지원하기 위한 예시적인 흐름도를 예시한다.

도 11은 본 발명의 실시예에 따른 특징들을 보여주는 기능 블록도를 예시한다.

발명을 실시하기 위한 구체적인 내용

[0008] 본 발명은 유사한 참조 부호들이 유사한 엘리먼트들을 표시하는 첨부 도면들의 도면들에서 제한이 아닌 예로서 예시된다. 이 개시내용에서 "하나" 또는 "일부" 실시예(들)에 대한 참조들이 반드시 동일한 실시예에 대한 것이 아니며 이러한 참조들이 실시예(들) 중 적어도 하나를 의미한다는 것에 유의해야 한다.

[0009] 이하의 본 발명의 설명은 고성능 네트워크에 대한 일례로서 인피니밴드(IB) 네트워크를 사용한다. 다른 타입들의 고성능 네트워크들이 제한없이 사용될 수 있다는 것이 이 기술분야의 숙련자에게 명백할 것이다. 또한, 이하의 본 발명의 설명은 네트워크 토폴로지 모델에 대한 일례로서 팻-트리(fat-tree)를 사용한다. 다른 타입들의 네트워크 토폴로지 모델들이 제한없이 사용될 수 있다는 것이 이 기술분야의 숙련자에게 명백할 것이다.

[0010] 본원에서 설명된 것은 별개의 네트워크들 사이 내에서 트래픽을 라우팅하는 것을 지원할 수 있는 시스템들 및 방법들이다.

[0011] **인피니밴드(IB) 아키텍처**

[0012] IB 아키텍처는 직렬 포인트-투-포인트 풀-듀플렉스 기술(serial point-to-point full-duplex technology)이다. IB 클러스터들의 크기와 복잡성이 증가함에 따라, 네트워크는 서브넷들로 지칭될 수 있는 관리가능 섹션들로 세그먼트화(segment)될 수 있다. IB 서브넷은 스위치들 및 포인트 투 포인트 링크들을 사용하여 상호 연결되는 호스트들의 세트를 포함할 수 있다. IB 서브넷은 또한 서브넷의 모든 스위치들, 라우터들 및 호스트 채널 어댑

터(HCA: host channel adapter)들의 구성을 포함하여, 네트워크를 초기화하고 브링-업(bring up)하는 것을 담당하는 적어도 하나의 서브넷 관리자(SM)를 포함할 수 있다.

[0013] IB는 원격 직접 메모리 액세스(RDMA: remote direct memory access) 및 통상적인 송신/수신 시맨틱(semantic)들 모두를 제공하기 위해 전송 서비스들의 다양한 세트를 지원한다. 사용되는 전송 서비스와 관계 없이, IB HCA들은 큐 페어(QP: queue pair)들을 사용하여 통신한다. QP는 통신 셋업동안 생성되고, QP 번호, HCA 포트, 목적지 LID, 큐 크기들, 및 제공되는 전송 서비스와 같은 초기 속성들의 세트를 가질 수 있다. HCA는 많은 QP들을 핸들링할 수 있으며, 각각의 QP는 송신 큐(SQ) 및 수신 큐(RQ)와 같은 큐들의 쌍으로 구성되며, 통신에 참여하는 각각의 엔드-노드에 이러한 하나의 쌍이 존재한다. 송신 큐는 원격 노드로 전송될 작업 요청을 보유하는 반면에, 수신 큐는 원격 노드로부터 수신된 데이터를 어떻게 처리할지에 대한 정보를 보유한다. QP들에 부가하여, 각각의 HCA는 송신 및 수신 큐들의 세트와 연관되는 하나 이상의 완료 큐(CQ)들을 가진다. CQ는 송신 및 수신 큐에 포스팅되는 작업 요청들에 대한 완료 통지들을 보유한다. 통신의 복잡성들을 사용자가 알지 못할지라도, QP 상태 정보는 HCA내에서 유지된다.

[0014] **별개의 인피니밴드(IB) 서브넷들 사이에서 트래픽 라우팅**

[0015] 도 1은 본 발명의 실시예에 따라, 네트워크 환경에서 별개의 인피니밴드(IB) 서브넷들 사이에서 트래픽을 라우팅하는 예시를 도시한다. 도 1에 도시 된 바와 같이, 네트워크 환경(100)은 인피니밴드 아키텍처(IBA)에 기반할 수 있으며, 2-계층 토폴로지 분할을 지원한다.

[0016] IB 네트워크 또는 IB 패브릭(100)의 하위 계층은 서브넷들, 예를 들어 서브넷들(101-104)로 지칭될 수 있으며, 이들 서브넷들 각각은 스위치들 및 포인트-투-포인트 링크들을 사용하여 상호연결된 호스트의 세트를 포함한다. 상위 레벨에서, IB 패브릭(100)의 하나 이상의 서브넷들(101 -104)은 라우터들, 예를 들어 라우터들(105-106)을 사용하여 상호연결될 수 있다. 더욱이, 각각의 서브넷(101-104)은 로컬 서브넷상의 포트들만을 구성하는 자기 자신의 서브넷 관리자(SM)를 실행할 수 있으며, 라우터들(105-106)은 서브넷 관리자(SM)들에게 투과적이지 않다(non-transparent).

[0017] 서브넷들(101-104) 각각 내의 호스트들 및 스위치들은 지정된 로컬 식별자(LID)들을 사용하여 어드레싱될 수 있다. 대규모 인스톨레이션들의 크기는 이용가능한 로컬 식별자(LID)들의 수에 의해 제한될 수 있다. 예를 들어, 단일 서브넷은 49151개의 유니캐스트 LID들로 제한될 수 있다. IB 어드레스 공간을 확장하는 하나의 접근법은 32 비트까지 LID 어드레싱 공간을 확장하는 것이며, 이의 유용성은 이러한 접근법이 구(older) 하드웨어와 역호환가능하지 않을 수 있기 때문에 제한될 수 있다.

[0018] 본 발명의 실시예에 따르면, IB 라우터들(105-106)은 IB 패브릭(100)에서 어드레스 공간 확장성(address space scalability)을 제공할 수 있다. 도 1에 도시된 바와같이, 더 많은 엔드-포트들이 요구될 때, 다수의 서브넷들(101-104)은 하나 이상의 IB 라우터들(105-106)을 사용하여 함께 결합될 수 있다. 서브넷들(101-104) 각각은 로컬 식별자(LID) 어드레스 공간(111-114)을 사용할 수 있다. LID 어드레스들이 각각의 LID 어드레스 공간(111-114)내에서 로컬 가시성(local visibility)을 가지고 있기 때문에, LID 어드레스들은 라우터들(105-106)에 의해 연결되는 상이한 서브넷들(101-104)에서 재사용될 수 있다. 따라서, IB 패브릭(100)에 주소 공간 확장성이 제공될 수 있으며, 이러한 접근법은 이론적으로 무제한 어드레싱 공간을 생성할 수 있다.

[0019] 더욱이, 크고 복잡한 네트워크(100)를 다수의 서브넷들(101-104)로 세그먼트화함으로써, 시스템은 패브릭 관리 컨테인먼트(containment)를 제공할 수 있다. 패브릭 관리 컨테인먼트는 1) 결함 격리, 2) 증가된 보안 및 3) 서브넷 내 라우팅 플렉서빌리티을 제공함에 있어 유용할 수 있다.

[0020] 첫째, 큰 네트워크(100)를 여러개의 보다 작은 서브넷들(101-104)로 분할함으로써, 결함들 또는 토폴로지 변화들이 단일 서브넷에 포함될 수 있으며, 서브넷 재구성은 라우터들(105-106)을 통해 다른 서브넷들로 전해지지 않을 수 있다. 이는 재구성 시간을 단축하고 결함의 영향을 제한한다.

[0021] 둘째, 보안 관점에서, 라우터들(105-106)을 사용하여 큰 패브릭(100)을 서브넷들(101-104)로 세그먼트화하는 것은 대부분의 공격들의 범위를 공격당한 서브넷으로 제한하다.

[0022] 셋째, 라우팅 관점에서, 패브릭 관리 컨테인먼트는 보다 플렉서블한 라우팅 방식들을 지원함에 있어 유용할 수 있는데, 이는 2개 이상의 정규 토폴로지들을 포함하는 하이브리드 패브릭의 경우에 특히 유리할 수 있다.

[0023] 예를 들어, 하이브리드 패브릭(100)은 메시(mesh) 또는 토러스(torus) 부분(또는 임의의 다른 정규 토폴로지)과 상호연결된 팻-트리 부분을 포함할 수 있다. 서브넷 내의 라우팅 알고리즘들이 단지 서브넷 범위를 가질 수 있

기 때문에 하이브리드 패브릭(100)의 상이한 부분들을 개별적으로 라우팅하기 위한 간단한 방식은 존재하지 않을 수 있다. 더욱이, 하이브리드 토폴로지에 대하여 최적 성능을 제공할 수 있는 IB에 대한 범용적인 관용적 라우팅 알고리즘(general purpose agnostic routing algorithm)들이 존재하지 않는다.

[0024] 본 발명의 실시예에 따르면, 하이브리드 토폴로지는 보다 작은 정규 서브넷들(101-104)로 분할될 수 있으며, 이들 정규 서브넷들 각각은 특정 서브넷에 대하여 최적화된 상이한 라우팅 알고리즘을 사용하여 라우팅 될 수 있다. 예를 들어, 팻-트리 라우팅 알고리즘은 팻-트리 부분을 라우팅할 수 있으며, 치수-순서 라우팅(dimension-order routing)은 토폴로지의 메시 부분을 라우팅할 수 있다.

[0025] 더욱이, 슈퍼 서브넷 관리자는 로컬 서브넷 관리자들 간을 조직화하는데 사용될 수 있으며, 최종 목적지가 다른 서브넷 배후에 배치되는 (예를 들어, 적어도 2개의 라우터 홉들이 필요한) 더 비정규적인 네트워크들의 경우에 트랜지트 서브넷(transit subnet)을 통과하는 경로를 설정할 수 있다.

[0026] 따라서, IB 서브넷들 사이에서 라우팅을 사용함으로써, 시스템은 IB 네트워크에서 어드레스 공간 스케일러빌리티 및 패브릭 관리 컨테인먼트를 제공할 수 있다.

[0027] **네이티브 인피니밴드(IB) 라우터들**

[0028] 도 2는 본 발명의 실시예에 따라, 네트워크 환경에서 라우터를 통한 패킷 포워딩을 지원하는 예시를 도시한다. 도 2에 도시된 바와 같이, 네트워크 환경 (200)은 다수의 서브넷들, 예를 들어 라우터(210)를 사용하여 상호연결되는 IB 서브넷들 A-B(201-202)를 포함할 수 있다. 더욱이, IB 서브넷 A(201)는 호스트 채널 어댑터(HCA)(213)와 연관되는 엔드 노드 X(203)에 연결되며, IB 서브넷 B(202)는 호스트 채널 어댑터(HCA)(214)와 연관되는 엔드 노드 Y(204)에 연결된다.

[0029] 본 발명의 실시예에 따르면, IB 라우터(210)는 IB 어드레싱 계층구조(hierarchy)의 계층(layer)-3에서 동작할 수 있다. 더욱이, LID들에 부가하여, 각각의 IB 디바이스는 또한 GID, 즉 IB 계층-3 어드레스를 형성하기 위하여 사용될 수 있는 64-비트 글로벌 고유 식별자(GUID: global unique identifier)를 가질 수 있다. 예를 들어, GID는 IPv6-형 128-비트 어드레스를 형성하기 위해 64-비트 GUID와 64-비트 서브넷 ID를 연접함으로써 생성될 수 있다. 부가적으로, 용어 GUID는 또한 포트 GUID들, 즉 IB 패브릭의 모든 각각의 포트에 할당된 GUID들을 지칭하기 위하여 사용될 수 있으며, GUID는 비-휘발성 메모리로 구워질 수 있다(burn-into).

[0030] 도 2에 도시된 바와같이, 엔드-노드, 예를 들어 엔드 노드 X(203)는 라우터(210)를 통해 다른 엔드-노드, 예를 들어 엔드 노드 Y(204)에 패킷을 송신할 수 있다. 어드레스 분석 메커니즘은 로컬 라우터(210)를 엔드 노드들 X-Y(203-204)가 알아볼 수 있게 만들 수 있다.

[0031] 예를 들어, 라우터(210)에서 수신되는 패킷은 로컬 라우팅 헤더(LRH)(223) 및 글로벌 라우팅 헤더(GRH)(225)를 포함하는 인입 라우팅 헤더(221)를 포함할 수 있다. 패킷을 포워딩하기 전에, 라우터(210)는 인입 라우팅 헤더(221)를 인출 라우팅 헤더(222)로 수정할 수 있으며, 인출 라우팅 헤더(222)는 LRH(224) 및 GRH(226)를 포함한다.

[0032] 도 2에 도시된 바와같이, 엔드-노드 X(203)는, LRH(223)에서, 로컬 HCA의 LID 어드레스, 예를 들어 X를 소스 LID, 즉 srcLID로 둘 수 있고, 로컬 라우터의 LID 어드레스, 예를 들어 A를 목적지 LID, 즉 dstLID로 둘 수 있다. 더욱이, 엔드-노드 X(203)는, GRH(225)에서, 자신의 계층-3 어드레스(GID), 예를 들어 1234를 소스 GID, 즉 srcGID로 둘 수 있으며, 최종 목적지 어드레스(GID), 예를 들어 5678를 목적지 GID, 즉 dstGID로 둘 수 있다.

[0033] 패킷이 라우터(210)에 도달할 때, 패킷 포워딩 메커니즘(220)은 패킷 필드들을 업데이트하여 교체할 수 있다. 예를 들어, 시스템은, 로컬 라우팅 헤더(LRH)(224)에서, 소스 LID, 즉 srcLID를 라우터의 출구(egress) 포트의 LID, 예를 들어 B로 교체할 수 있다. 더욱이, 시스템은 목적지 LID, dstLID를 목적지의 LID 어드레스, 예를 들어 Y로 교체할 수 있다.

[0034] 대안적으로, 시스템은 패킷이 다른 라우터로부터 포워딩되는 경우에 이전-홉 라우터의 출구 포트의 LID를 소스 LID, 즉 srcLID로서 사용할 수 있다. 또한, 시스템은 추가 패킷 포워딩이 필요한 경우에 목적지 LID, 즉 dstLID를 다음-홉 포트의 LID와 교체할 수 있다. 각각의 경우에, 시스템은 다음 홉으로 패킷을 포워딩하기 전에 순환 리던던시 검사(CRC)들을 재컴퓨팅할 수 있다.

[0035] 본 발명의 실시예에 따르면, 별개의 인피니밴드(IB) 서브넷들 사이에서 트래픽을 라우팅하기 위하여 상이한 방법들이 사용될 수 있다. 이들 방법들은 서브넷 간 라우팅에 관한 다양한 문제들, 예를 들어 특정 목적지에 대

하여 어떤 라우터가 선정되어야 하는지(제 1 라우팅 페이즈(phase)) 그리고 목적지에 도달하기 위하여 라우터에 의해 어떤 경로가 선정되어야 하는지(제 2 라우팅 페이즈)를 해결하기 위하여 사용될 수 있다.

[0036] 예를 들어, 이들 방법들은 다양한 토폴로지들에 대하여 양호한 성능을 제공하는 단순한 전형적인 라우팅 방법, 즉 소스 라우팅 방법, 및 팻-트리 기반 토폴로지들에 대하여 전문화되어 최적의 성능을 제공하는 최적화식 라우팅 방법을 포함할 수 있다. 소스 라우팅 방법 및 최적화식 라우팅 방법 둘다는 잠재적으로, 전형적인 라우팅 방법보다 양호한 성능을 초래할 수 있다. 더욱이, 최적화된 라우팅 방법은 근본적인 팻-트리 토폴로지에 대해 최적 성능을 획득하는 것을 가능하게 한다. 부가적으로, 양 방법들은 임의적 멀티-포트 라우터들의 사용을 지원하는 반면에, 전형적인 라우팅은 그것이 이용가능한 포트들의 수를 제한하지 않을 때조차 모든 포트들을 효율적으로 활용하지 못할 수 있다.

[0037] 따라서, 이들 방법들을 사용하면, 네이티브 IB 라우터들은 성능을 크게 저하시키지 않고 서브넷들을 함께 연결시킴으로써 더 복잡한 IB 패브릭들을 구축하는 것을 가능하게 한다.

[0038] **서브넷 간 소스 라우팅(ISSR)**

[0039] 본 발명의 실시예에 따르면, 범용 라우팅 알고리즘, 예를 들어 서브넷 간 소스 라우팅(ISSR) 라우팅 알고리즘이 하이브리드 서브넷들을 가진 복잡한 네트워크를 라우팅하기 위하여 사용될 수 있다. ISSR 라우팅 알고리즘은 2개의 페이즈들, 즉 특정 목적지에 대한 입구(ingress) 라우터 포트를 선정하기 위한 제 1 페이즈 및 출구 라우터 포트를 선택하기 위한 제 2 페이즈를 포함할 수 있다.

[0040] 도 3은 본 발명의 실시예에 따라, 네트워크 환경에서 상이한 목적지들에 대한 입구 라우터 포트들을 선정하는 예시를 도시한다. 도 3에 도시된 바와같이, 네트워크 환경(300)은 하나 이상의 라우터들, 예를 들어 라우터 A(301) 및 라우터 B(302)와 연결되는 서브넷(310)을 포함할 수 있다.

[0041] 서브넷 관리자(SM), 예를 들어 서브넷(310)의 SM(303)은 상이한 목적지들에 대한 입구 라우터 포트들을 선정하기 위하여 사용될 수 있는 매핑 파일(304)을 포함할 수 있다. 더욱이, 입구 라우터 포트를 선정하는 것은 로컬 서브넷에 연결되는 모든 이용가능한 라우터들 사이의 라운드-로빈 경로 분배에 기반할 수 있다. 예를 들어, ISSR 알고리즘에 대한 로컬 라우터 포트를 선정하는 find_router() 함수는 OpenSM 라우팅 알고리즘과 유사한 방식으로 구현될 수 있다.

[0042] 하기는 매핑 파일(304)의 상위 레벨 예이다.

[0043] 테이블 1: ISSR 및 ISFR 알고리즘들에 대한 매핑 파일의 상위 레벨 예

[0044]

1: dst_gid1		router A port 1 guid
2: dst_gid2		router B port 1 guid
3: dst_gid3		router A port 2 guid
4: dst_gid4		router B port 2 guid
5: #default route		
6:	*	router A port 1 guid
7:	*	router B port 1 guid

[0045] 단지 단일 라우터 포트와 전체 서브넷을 매칭시킬 수 있는 OpenSM 라우팅과 다르게, 시스템은 상이한 라우터 포트들에 목적지 엔드-포트들을 매핑시킬 수 있다. 더욱이, 매핑 파일의 셋업은 OpenSM에 제공되는 매핑 파일과 상이할 수 있다. 앞의 테이블 1에 도시된 바와같이, 매핑 파일(304)은 OpenSM 서브넷 간 라우팅에 대해서는 단지 서브넷 프리픽스(subnet prefix) 대신에 완전하게 자격이 부여된 포트 GID를 포함하는데, 이는 입구 라우터 포트들을 선정할 때 시스템이 최대 입상도(full granularity)를 제공하도록 한다.

[0046] 더욱이, 도 3에 도시된 바와같이, 동일한(또는 유사한) 수의 목적지들이 예를 들어 라운드 로빈 방식으로 다수의 포트들에 매핑될 수 있다. 예를 들어, 목적지 A(311)(dst_gid1) 및 목적지 C(313)(dst_gid3)는 라우터 A(301)상의 포트 1 및 포트 2를 통해 라우팅될 수 있으며, 목적지 B(312)(dst_gid2) 및 목적지 D(314)(dst_gid4)는 라우터 B(302)상의 포트 1 및 포트 2를 통해 라우팅될 수 있다. 부가적으로, 매핑 파일(304)은 백업 및 디폴트 루트들을 특정할 수 있다.

[0047] 본 발명의 실시예에 따르면, 출구 라우터 포트를 선택하는 것은 라우터 펌웨어의 모듈로 기반 해시(modulo based hash)를 사용하여 구현될 수 있다. 해시는 소스 및 목적지 LID들(또는 라우터상의 입구 포트 번호와 같은 임의의 다른 유용한 파라미터)을 사용하여 생성되는 난수를 취할 수 있다. 해시를 사용하면, 출력 라우터 포트들 중 하나의 포트가 선택될 수 있다. 더욱이, ISSR 라우팅 알고리즘은 노드들의 동일한 쌍에 대하여 항상

동일한 경로를 사용하는 결정론적인 망각 라우팅 알고리즘(deterministic oblivious routing algorithm)이다.

[0048] 더욱이, 라우터가 3개 이상의 서브넷들에 부착될 수 있기 때문에, 출구 라우터 포트를 선택하기 위하여 2-단계 포트 검증 방법이 사용될 수 있다. 첫째, 시스템은 목적지가 배치되는 서브넷에 (또는 서브넷의 방향으로) 부착되는 가능한 포트들의 세트를 선정할 수 있다. 이후, 시스템은 출구 포트를 선정하기 위하여 모듈로 함수에 기반한 단순한 해시를 사용할 수 있다.

[0049] 도 4는 본 발명의 실시예에 따라, 네트워크 환경의 라우터에서, 수신된 패킷에 대한 출구 라우터 포트를 선택하는 예시를 도시한다. 도 4에 도시된 바와같이, 네트워크 환경(400)은 상이한 서브넷들의 다양한 목적지들에 하나 이상의 패킷들을 라우팅할 수 있는 라우터, 예를 들어 라우터 A(401)를 포함할 수 있다.

[0050] 예를 들어, 라우터 A(401)는, 포트 1에서, 소스 A(411)로부터의 패킷을 수신할 수 있으며, 이 패킷은 서브넷(403)의 목적지 X(421)로 향한다. 부가적으로, 라우터 A(401)는, 포트 2에서, 소스 C(413)로부터의 다른 패킷을 수신할 수 있으며, 이 다른 패킷은 서브넷(402)의 목적지 Y(422)로 향한다.

[0051] 하기의 알고리즘 1은 출구 라우터 포트를 선택하기 위해 라우터 펌웨어 및 SM 둘다에서 구현될 수 있다.

알고리즘 1 : choose_egress_port() function in ISSR

```
1: if received_intersubnet_packet() then
2:       dstLID = get_next_LID(dGID)
3:       srand(srcLID + dstLID)
4:       port_set = choose_possible_out_ports()
5:       e_port = port_set[(rand()%por_set:size)]
6: end if
```

[0052]

[0053] 라우팅 결정은 소스 LID 및 목적지 LID 둘다에 기반할 수 있다. 소스 LID는 트랜지트 서브넷 시나리오에서 이전-홉 라우터의 출구 포트 또는 원래 소스일 수 있고, 목적지 LID는 다음-홉 입구 라우터 포트의 LID 또는 최종 목적지 LID일 수 있다.

[0054] 라우터 A(401)는 배속(attach)된 서브넷들을 인식할 수 있기 때문에 소스 LID 및 목적지 LID 둘다에 대한 값들을 안다. 목적지 LID를 획득하기 위하여, 알고리즘 1(라인 2)에서 get_next_LID()와 같은 매핑 함수는 목적지 GID를 목적지 LID에 매핑시키기 위하여 사용될 수 있거나 또는 GID에 배치된 서브넷 프리픽스에 기반하여 다음-홉 LID를 리턴시키는 것이 요구된다. 매핑 함수는 하기의 사항, 컨텐트-어드레스가능 메모리 룩업, 글로벌 라우팅 헤더(GRH) 필드로부터 최종 목적지 로컬 식별자(LID)를 디코딩하는 것, 글로벌 라우팅 헤더(GRH) 필드로부터 다음-홉 로컬 식별자를 디코딩하는 것 중 하나 이상을 수행할 수 있다.

[0055] 본 발명의 실시예에 따르면, 출구 라우터 포트를 선택하는 알고리즘은 소스 및 목적지 LID들에 기반하여 난수를 계산할 수 있다. 이러한 난수는 가능한 포트들의 세트로부터 단일 출구 포트를 선택하기 위해 사용될 수 있다. 난수의 생성은 주어진 소스-목적지 쌍이 항상 동일한 수를 생성할 수 있도록 결정론적 방식으로 행해질 수 있다. 따라서, 시스템은 서브넷들 사이에서 라우팅할 때 비순차 전달을 방지할 수 있으며, 출구 포트를 선택하기 위한 라운드-로빈 방법과 다르게 반드시 각각의 소스-목적지 쌍이 항상 네트워크를 통해 동일한 경로를 사용하도록 한다.

[0056] 도 5는 본 발명의 실시예에 따라, 네트워크 환경의 라우터에서 서브넷 간 소스 라우팅(ISSR) 알고리즘을 지원하기 위한 예시적인 흐름도를 예시한다. 도 5에 도시된 바와같이, 단계(501)에서, 라우터는 목적지들의 리스트를 수신할 수 있으며, 그 리스트의 목적지들에 하나 이상의 패킷들을 라우팅하는 것을 담당하며, 라우터는 네트워크 환경의 적어도 하나의 서브넷에 연결된다. 이후, 단계(502)에서, 라우터는 하나 이상의 패킷들과 연관된 목적지 LID 및 소스 로컬 식별자(LID)에 기반하여 난수를 생성할 수 있다. 더욱이, 단계(503)에서, 라우터는 라우터의 출력 라우터 포트들로부터 하나의 포트를 선택하기 위하여 모듈로 기반 해시를 사용할 수 있다.

[0057] **서브넷 간 팻-트리 라우팅(ISFR)**

[0058] 본 발명의 실시예에 따르면, 서브넷 간 팻-트리 라우팅(ISFR) 라우팅 알고리즘과 같은 최적화된 라우팅 방법은 팻-트리 토폴로지들을 가진 서브넷들 사이에서 라우팅하는데 사용될 수 있다.

[0059] 최적화된 라우팅 방법은 3개의 페이즈들, 예를 들어 로컬 서브넷에 연결되는 모든 이용가능한 라우터 사이에서의 라운드-로빈 경로 분배의 제 1 페이즈, 어떤 스위치가 특정 목적지로 향하는 주 경로(primary path)로서 역할하는지를 알기 위하여 각각의 라우터에 연결된 하향 스위치들에 질의하는 제 2 페이즈 및 제 1 페이즈 및 제 2 페이즈로부터 획득되는 정보를 사용하여 라우팅 테이블을 구축하는 제 3 페이즈를 포함할 수 있다.

[0060] 도 6은 본 발명의 실시예에 따라, 네트워크 환경(600)의 라우터에서 서브넷 간 팻-트리 라우팅(ISFR) 라우팅 알고리즘을 지원하는 예시를 도시한다. 도 6에 도시된 바와같이, 라우터(601)는 하나 이상의 엔드 노드들(621-624)에 연결되는 서브넷(610)에 패킷(650)을 라우팅하는 것을 담당한다. 서브넷(610)은 팻-트리 토폴로지의 SW들(611-616)과 같은 하나 이상의 스위치들을 포함할 수 있다.

[0061] 라우터(601)는 연관된 GID에 따라 특정 목적지에 패킷들을 라우팅하기 위하여 자신의 포트들 중 어떤 포트가 사용될 수 있는지를 알 수 있다. 예를 들어, 라우터(601)는 서브넷(610)내의 서브넷 관리자(SM)(620)로부터 포트 매핑들(630)을 수신할 수 있다. 포트 매핑들(630)은 라우터(601)가 서브넷(610)으로부터 발신하는 패킷들을 라우팅하는 목적지의 리스트를 포함할 수 있다.

[0062] 더욱이, 라우터(601)는 패킷(650)을 라우팅할 때 최대 성능을 실현하기 위하여 어느 하향 출력 포트들을 사용할지를 스위치들(611-612)로부터 알 수 있다. 예를 들어, 라우터(601)는 서브넷(610)의 목적지 노드로의 주 경로에 대하여 서브넷(610)의 스위치들(611-612)에 질의할 수 있다. 질의는 라우터가 패킷(650)에 대한 주 경로(스위치(611) 또는 스위치(612))를 선정하도록 하는데, 왜냐하면 이들 스위치들 중 단지 하나만이 원하는 목적지로 향하는 전용 주 경로를 가질 수 있기 때문이다.

[0063] 이후, 라우터(601)는 앞서 획득된 정보를 사용하여 라우팅 테이블(640)을 구축할 수 있다. 예를 들어, 로컬 OpenSM 라우팅은 스위치(611) 또는 스위치(612)의 라우팅 테이블에서 목적지 노드로의 주 경로를 표시할 수 있다.

[0064] 이하의 알고리즘 2는 라우터상의 출구 포트를 선택하기 위한 라우터 펌웨어로 구현될 수 있다.

알고리즘 2 : query_down_for_egree_port() function in ISSR

```
1: if received_mapping_files then
2:       for all_port_in_down ports do
3:             down_switch = get_node(port)
4:             lid = get_LID_by_GID(GID)
5:             if down_switch:routing_table[lid] == primary_path then
6:                   e_port = port
7:             end if
8:       end for
9: end
```

[0065]

[0066] 앞의 ISFR 알고리즘은 테이블 1에서 GID-대-라우터 포트 매핑들을 포함하는 이전에 정의된 파일 포맷을 사용할 수 있다. ISSR 알고리즘과 유사하게, ISFR 알고리즘은 라우터 디바이스에서 구현될 수 있다. 더욱이, ISFR 알고리즘은 단지 팻-트리로 그리고 모든 각각의 서브넷에서 국부적으로 실행되는 팻-트리 라우팅으로 작동할 수 있다. 또한, ISFR 알고리즘은 필요한 경우에 ISSR로 폴백(fall-back)할 수 있다.

[0067] 도 6에 도시된 바와같이, 라우터(601)는 적절한 팻-트리 토폴로지의 최상부에 배치된 것으로 표현될 수 있다. 따라서, 질의가 수행된 이후에, 라우터(601)는 특정 서브넷에 배치된 각각의 목적지에 대하여 하향 방향에서 포트당 하나의 경로를 가질 수 있다. 과도 가입된(oversubscribed) 팻-트리들의 경우들에서, 경로들의 수는 과도 가입비(oversubscription ratio)와 동일할 수 있다.

[0068] 부가적으로, 서브넷(610)에 둘보다 많은 라우터들이 연결될 수 있다. ISFR 라우팅의 특성은 모든 스파인(spine)(최상부) 라우터들이 스위치들로 교체된 경우에 (라우터들을 사용한) ISFR 라우팅 및 (스위치들을 사용한) 로컬 라우팅에 대한 라우팅 테이블들이 동일하도록 한다.

[0069] 본 발명의 실시예에 따르면, ISFR 라우팅 알고리즘은 비-투과 라우터들을 통해 연결되는 상이한 서브넷들 내의 서브넷 관리자(SM)들 사이에서 설정된 통신에 기반하여 구현될 수 있다. 예를 들어, 인터페이스는 라우터들에 제공될 수 있으며 이 라우터들을 통해 SM들이 통신할 수 있으며, 인접 서브넷들 내에 배치된 2개의 SM들 사이에

서 핸드셰이킹(handshaking)이 구현될 수 있다.

[0070] 도 7-9는 본 발명의 실시예에 따라 상이한 토폴로지들을 가진 네트워크 환경에서의 라우팅을 예시한다. 토폴로지들 각각은 멀티-스테이지 팻-트리, 예를 들어 최상부에 라우터들이 배치된 3-스테이지 팻-트리로서 표현될 수 있다. 예를 들어, 3-스테이지 팻-트리는 3개의 라우팅/스위칭 스테이지들 및 하나의 노드 스테이지를 가질 수 있으며, 각각의 서브넷(2-스테이지 팻-트리)은 브랜치로서 나타난다. 더욱이, ISFR 라우팅 알고리즘에 기반하여 보다 큰 토폴로지들이 또한 지원될 수 있다.

[0071] 본 발명의 실시예에 따르면, 네트워크 환경은 각각의 서브넷이 서브넷들 사이내에서 어느 트랜지트 서브넷들도 없이 다른 서브넷들에 직접 배속될 수 있는 팻-트리 토폴로지이도록 구성될 수 있다.

[0072] 예를 들어, 도 7에서, 시스템(700)은 2개의 팻-트리 서브넷들, 예를 들어 서브넷들 A-B(701-702)를 연결하기 위하여 6개의 라우터들(710)을 사용할 수 있다. 도 8에서, 시스템(800)은 3개의 팻-트리 서브넷들, 예를 들어 서브넷들 A-C(801-803)를 연결하기 위하여 6개의 라우터들(810)을 사용할 수 있다. 도 9에서, 시스템(900)은 6개의 팻-트리 서브넷들(901-906)을 연결하기 위하여 18개의 라우터들(910)을 사용하여 648-포트 3-스테이지 팻-트리를 생성할 수 있다.

[0073] 도 10은 본 발명의 실시예에 따라 네트워크 환경의 라우터에서 서브넷 간 팻-트리 라우팅(ISFR) 알고리즘을 지원하기 위한 예시적인 흐름도를 예시한다. 도 10에 도시된 바와같이, 단계(1001)에서, 라우터는 목적지들의 리스트를 수신할 수 있고 그 리스트 내의 목적지들에 하나 이상의 패킷들을 라우팅하는 것을 담당하며, 라우터는 네트워크 환경의 적어도 하나의 서브넷에 연결된다. 이후, 단계(1002)에서, 라우터는 하나 이상의 패킷들을 라우팅하기 위하여 라우터의 어느 하향 출력 포트들이 사용될 수 있는지에 관한 정보를 적어도 하나의 서브넷의 하나 이상의 스위치들로부터 획득할 수 있다. 더욱이, 단계(1003)에서, 라우터는 획득된 정보에 기반하여 라우팅 테이블을 구축할 수 있다.

[0074] 도 11은 본 발명의 실시예의 특징들을 보여주는 기능 블록도를 예시한다. 본 특징들은 네트워크 환경에서 별개의 서브넷들 사이에서 트래픽을 라우팅하기 위한 시스템(1100)으로서 구현될 수 있다. 시스템(1100)은 하나 이상의 마이크로프로세서들 및 하나 이상의 마이크로프로세서들상에서 실행되는 라우터(1110)를 포함한다. 라우터(1110)는 수신 유닛(1120)으로부터 목적지들의 리스트를 수신하며 – 상기 라우터는 그 리스트 내의 목적지들에 하나 이상의 패킷들을 라우팅하는 것을 담당하며, 라우터(1110)는 네트워크 환경의 적어도 하나의 서브넷에 연결됨 –; 하나 이상의 패킷들을 라우팅하기 위하여 라우터의 어느 하향 출력 포트들이 사용될 수 있는지에 관한, 적어도 하나의 서브넷의 하나 이상의 스위치들로부터의 정보를 획득 유닛(1130)으로부터 획득하며; 그리고 획득된 정보에 기반하여 라우팅 테이블을 구축 유닛(1140)상에 구축하도록 동작한다.

[0075] 일 실시예에 따르면, 하나 이상의 마이크로프로세서들에서 네트워크 환경에서 별개의 서브넷들 사이에서 트래픽을 라우팅하기 위한 시스템이 개시된다. 시스템은 목적지들의 리스트를 라우터에서 수신하기 위한 수단을 포함하며, 라우터는 그 리스트 내의 목적지들에 하나 이상의 패킷들을 라우팅하는 것을 담당하며 라우터는 네트워크 환경의 적어도 하나의 서브넷에 연결된다. 시스템은 하나 이상의 패킷들을 라우팅하기 위하여 라우터의 어느 하향 출력 포트들이 사용될 수 있는지에 관한 정보를 적어도 하나의 서브넷의 하나 이상의 스위치들로부터 획득하기 위한 수단을 포함한다. 시스템은 획득된 정보에 기반하여 라우팅 테이블을 구축하기 위한 수단을 포함한다.

[0076] 바람직하게, 시스템은 적어도 하나의 서브넷에 연결되는 모든 이용가능한 라우터들 사이에서 라운드-로빈 경로 분배를 수행하기 위한 수단을 포함한다.

[0077] 바람직하게, 시스템은 어떤 스위치가 특정 목적지로 향하는 주 경로로서 역할을 할 수 있는지를 알기 위하여 각각의 라우터에 연결되는 스위치들에 질의하기 위한 수단을 포함한다.

[0078] 바람직하게, 시스템은 획득된 정보에 기반하여 최대 성능을 위하여 라우터의 어느 하향 출력 포트들이 사용될 수 있는지를 결정하기 위한 수단을 포함한다.

[0079] 바람직하게, 시스템은 하나 이상의 목적지 글로벌 식별자(GID)들과 하나 이상의 라우터 포트들 사이의 포트 매핑을 서브넷 관리자로부터 수신하기 위한 수단을 포함한다.

[0080] 바람직하게, 시스템은 하나 이상의 추가 라우터들을 사용하여 네트워크 환경에서 하나 이상의 인접 서브넷들에 적어도 하나의 서브넷을 연결하기 위한 수단을 포함한다.

[0081] 바람직하게, 시스템은 연결된 서브넷들이 상이한 토폴로지들을 이루도록 하기 위한 수단을 포함한다.

[0082] 바람직하게, 시스템은 각각의 연결된 서브넷이 멀티-스테이지 팻-트리의 브랜치를 이루도록 하기 위한 수단을 포함한다.

[0083] 바람직하게, 시스템은 적어도 하나의 서브넷이 팻-트리 토폴로지로 구성되도록 하기 위한 수단을 포함한다.

[0084] 바람직하게, 시스템은 적어도 하나의 서브넷을 라우팅하기 위하여 팻-트리 라우팅 알고리즘을 사용하기 위한 수단을 포함한다.

[0085] 일 실시예에 따르면, 네트워크 환경에서 별개의 서브넷들 사이에서 트래픽을 라우팅하기 위한 시스템이 개시된다. 시스템은 목적지들의 리스트를 수신하도록 구성된 라우터를 포함하며, 라우터는 그 리스트 내의 목적지들에 하나 이상의 패킷들을 라우팅하는 것을 담당하며, 라우터는 네트워크 환경의 적어도 하나의 서브넷에 연결된다. 라우터는 하나 이상의 패킷들을 라우팅하기 위하여 라우터의 어느 하향 출력 포트들이 사용될 수 있는지에 대한 정보를 적어도 하나의 서브넷의 하나 이상의 스위치들로부터 획득하도록 구성되며 그리고 획득된 정보에 기반하여 라우팅 테이블을 구축하도록 또한 구성된다.

[0086] 바람직하게, 시스템은 라운드-로빈 경로 분배가 적어도 하나의 서브넷에 연결되는 모든 이용가능한 라우터들 사이에서 수행되게 할 수 있다.

[0087] 바람직하게, 시스템은 어떤 스위치가 특정 목적지로 향하는 주 경로로서 역할을 할 수 있는지를 알기 위하여 상기 라우터에 연결되는 스위치들에 상기 라우터가 질의하게 동작하도록 할 수 있다.

[0088] 바람직하게, 시스템은 획득된 정보에 기반하여 최대 성능을 위하여 라우터의 어느 하향 출력 포트들이 사용될 수 있는지를 상기 라우터가 결정하게 동작하도록 할 수 있다.

[0089] 바람직하게, 시스템은 하나 이상의 목적지 글로벌 식별자(GID)들과 하나 이상의 라우터 포트들 사이의 포트 매핑을 상기 라우터가 서브넷 관리자로부터 수신하게 동작하도록 할 수 있다.

[0090] 바람직하게, 시스템은 적어도 하나의 서브넷이 하나 이상의 추가 라우터들을 사용하여 네트워크 환경의 하나 이상의 인접 서브넷들에 연결되게 할 수 있다.

[0091] 바람직하게, 시스템은 연결된 서브넷들이 상이한 토폴로지들을 이루도록 할 수 있다.

[0092] 바람직하게, 시스템은 각각의 연결된 서브넷이 멀티-스테이지 팻-트리의 브랜치이도록 할 수 있다.

[0093] 바람직하게, 시스템은 적어도 하나의 서브넷이 팻-트리 토폴로지로 구성되도록 하고 적어도 하나의 서브넷을 라우팅하기 위하여 팻-트리 라우팅 알고리즘이 사용되도록 할 수 있다.

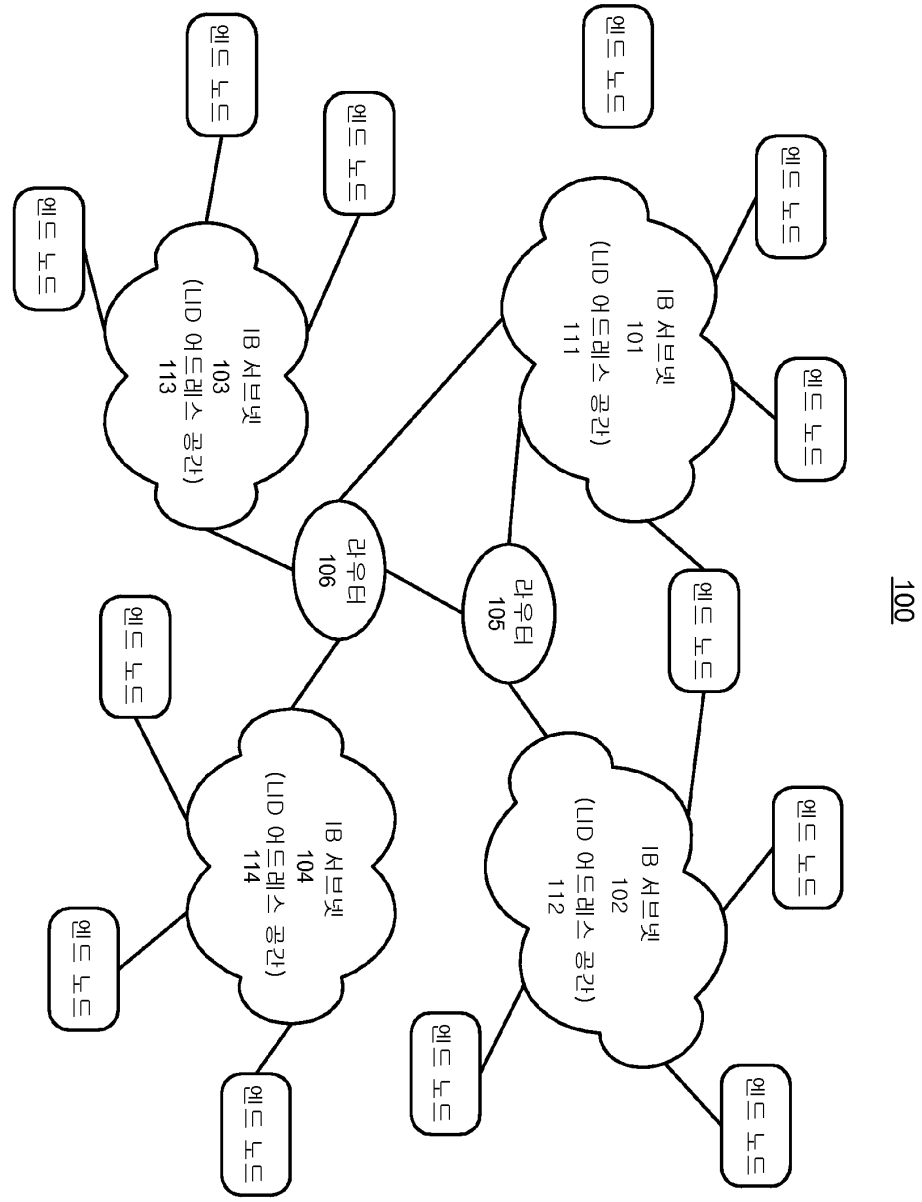
[0094] 본 발명은 본 개시내용의 교시들에 따라 프로그램된 하나 이상의 프로세서들, 메모리 및/또는 컴퓨터 판독가능 저장 매체를 포함하는 하나 이상의 종래의 범용 또는 특수화된 디지털 컴퓨터, 컴퓨팅 디바이스, 머신 또는 마이크로프로세서를 이용하여 통상적으로 구현될 수 있다. 적절한 소프트웨어 코딩은 소프트웨어 기술 분야의 숙련자들에게 분명할 바와 같이, 본 개시내용의 교시들에 기반하여 숙련된 프로그래머들에 의해 쉽게 준비될 수 있다.

[0095] 일부 실시예들에서, 본 발명은 본 발명의 프로세스들 중 임의의 것을 수행하도록 컴퓨터를 프로그램하는데 사용될 수 있는 명령어들이 저장된 저장 매체 또는 컴퓨터 판독가능 매체(매체들)인 컴퓨터 프로그램 제품을 포함한다. 저장 매체는 이들로만 한정되는 것은 아니지만, 플로피 디스크(disk)들, 광학 디스크(disc)들, DVD, CD-ROM들, 마이크로드라이브 및 자기-광학 디스크(disk)들을 포함하는 어떤 타입의 디스크, ROM들, RAM들, EPROM들, EEPROM들, DRAM들, VRAM들, 플래시 메모리 디바이스들, 자기 또는 광학 카드들, (분자 메모리 IC들을 포함하는) 나노시스템들, 또는 명령어들 및/또는 데이터를 저장하기에 적절한 어떤 타입의 매체 또는 디바이스를 포함할 수 있다.

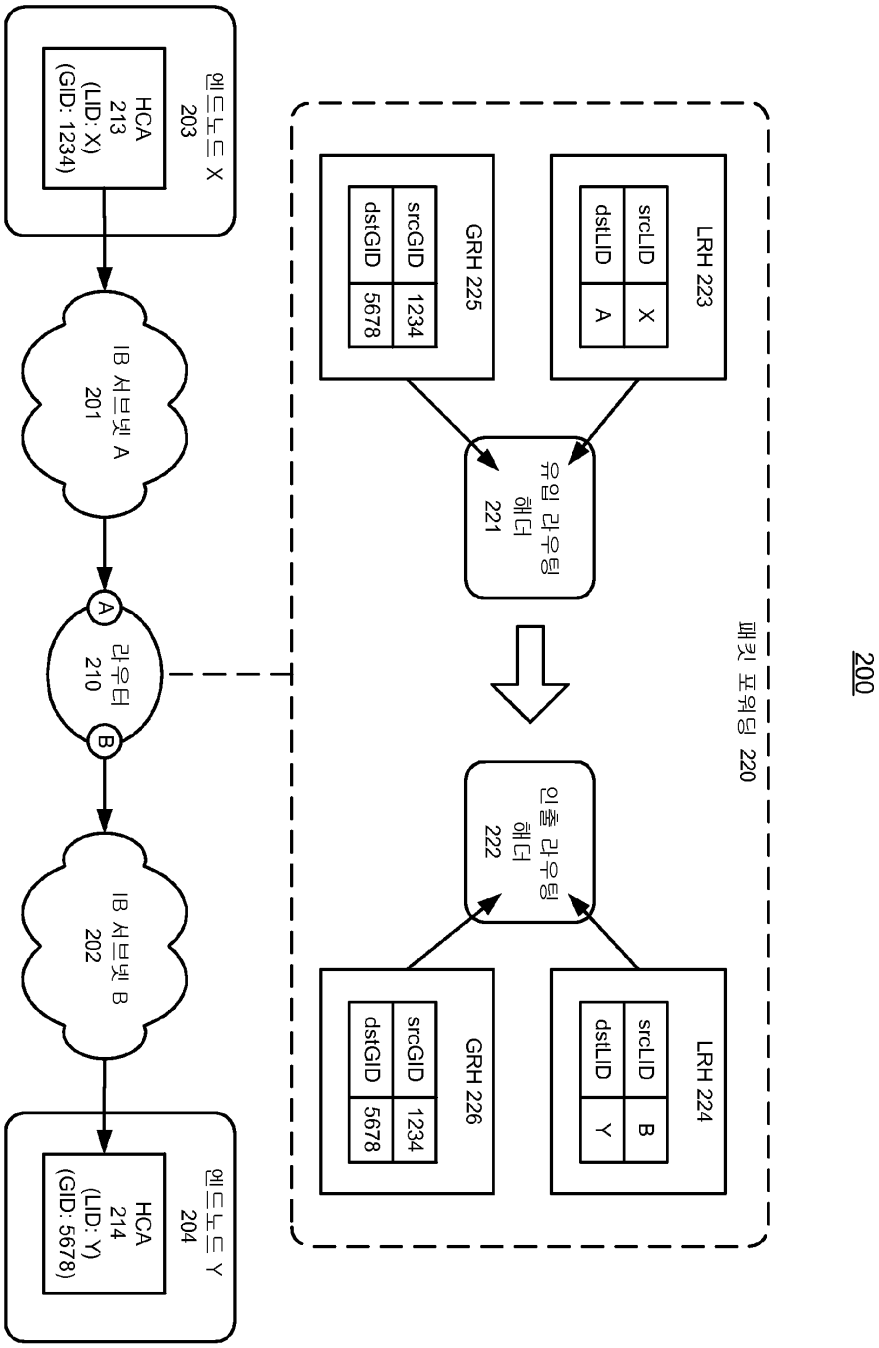
[0096] 본 발명의 전술한 설명은 예시 및 설명을 위하여 제공되었다. 본 설명은 완전한 것으로 의도되거나 또는 개시된 정확한 형태들로만 본 발명을 제한하고자 의도된 것이 아니다. 많은 수정들 및 변형들이 이 기술분야의 숙련자에게 명백할 것이다. 실시예들은 본 발명의 원리들 및 이의 실용적 응용을 가장 잘 설명하기 위하여 선택 및 설명되었으며, 그럼으로써 이 기술분야의 숙련자는 다양한 실시예들에 대한 발명 및 고려되는 특별한 용도에 적합한 다양한 수정들을 이해할 수 있다. 본 발명의 범위는 다음의 특허 청구범위 및 이의 균등물에 의해 한정되어야 함이 의도된다.

도면

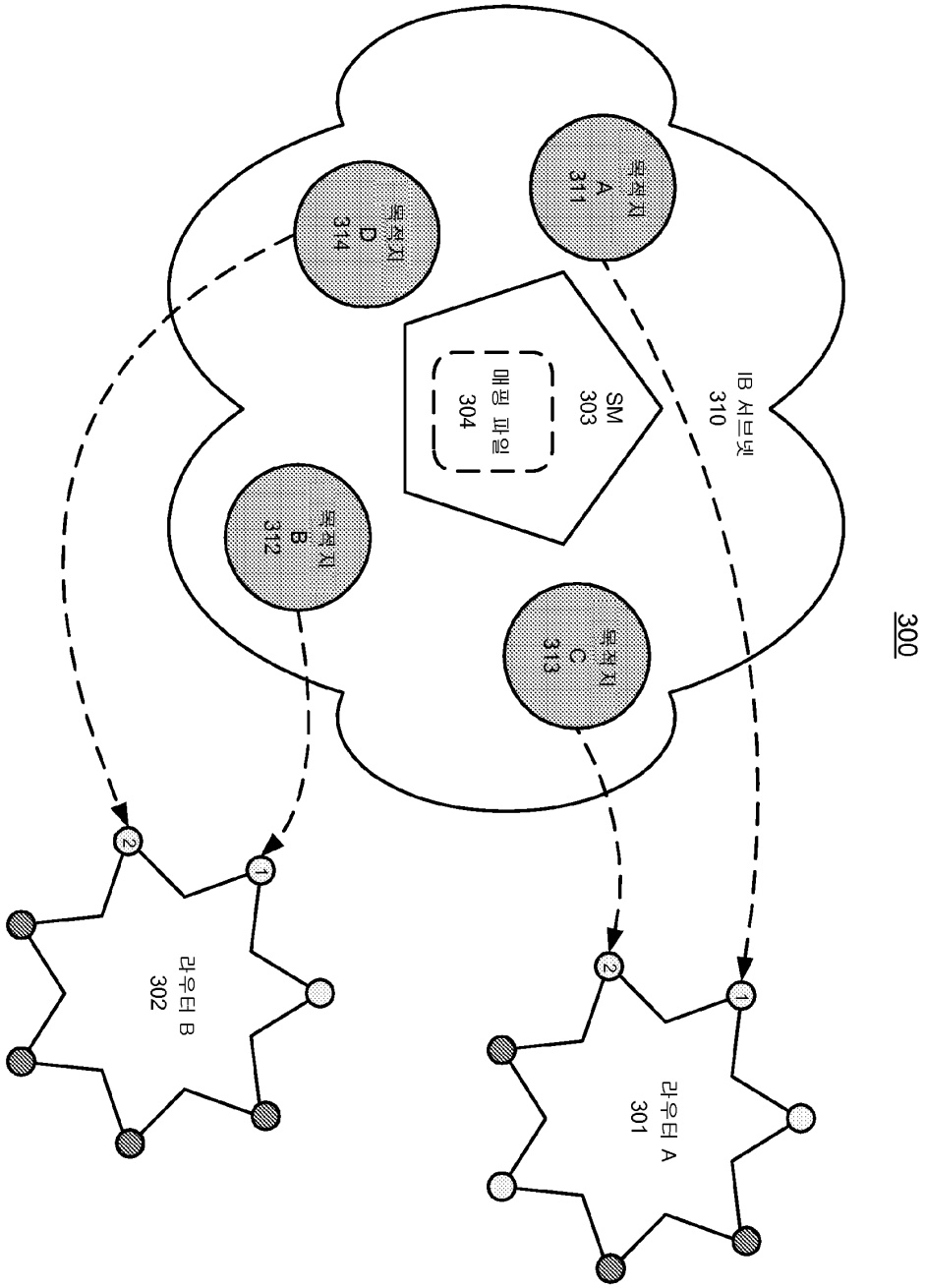
도면1


100
IB 서브넷
101
(LID 어드레스 공간)
111
IB 서브넷
102
(LID 어드레스 공간)
112
IB 서브넷
103
(LID 어드레스 공간)
113
IB 서브넷
104
(LID 어드레스 공간)
114
라우터
105
라우터
106
엣지 노드

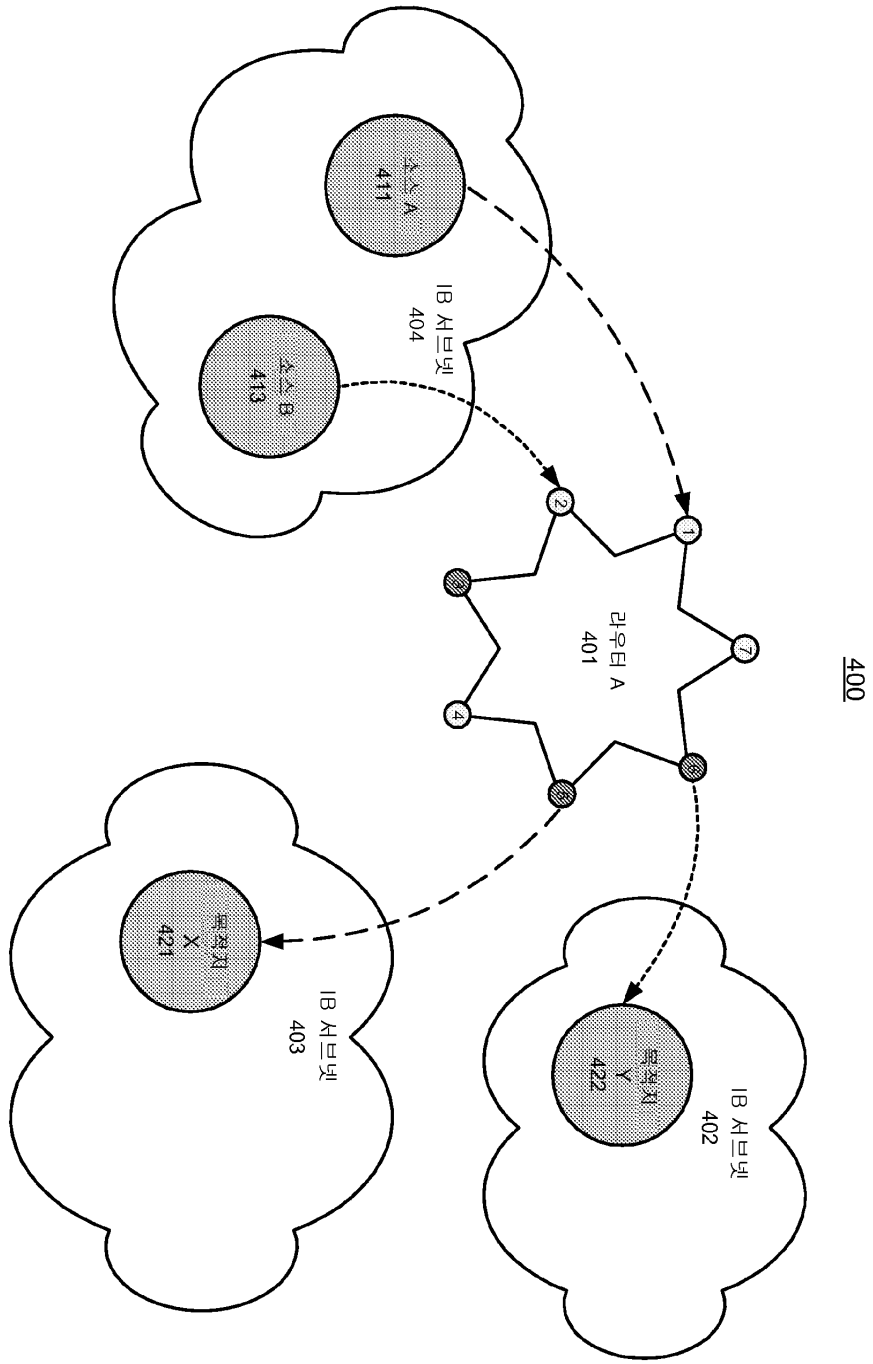
도면2


200
엔드노드 X
203
HCA
213
(LID: X)
(GID: 1234)
IB 서브넷 A
201
A
라우터
210
B
IB 서브넷 B
202
엔드노드 Y
204
HCA
214
(LID: Y)
(GID: 5678)
패킷 포워딩 220
LRH 223
srcLID
X
dstLID
A
GRH 225
srcGID
1234
dstGID
5678
유입 라우팅
헤더
221
인출 라우팅
헤더
222
LRH 224
srcLID
B
dstLID
Y
GRH 226
srcGID
1234
dstGID
5678

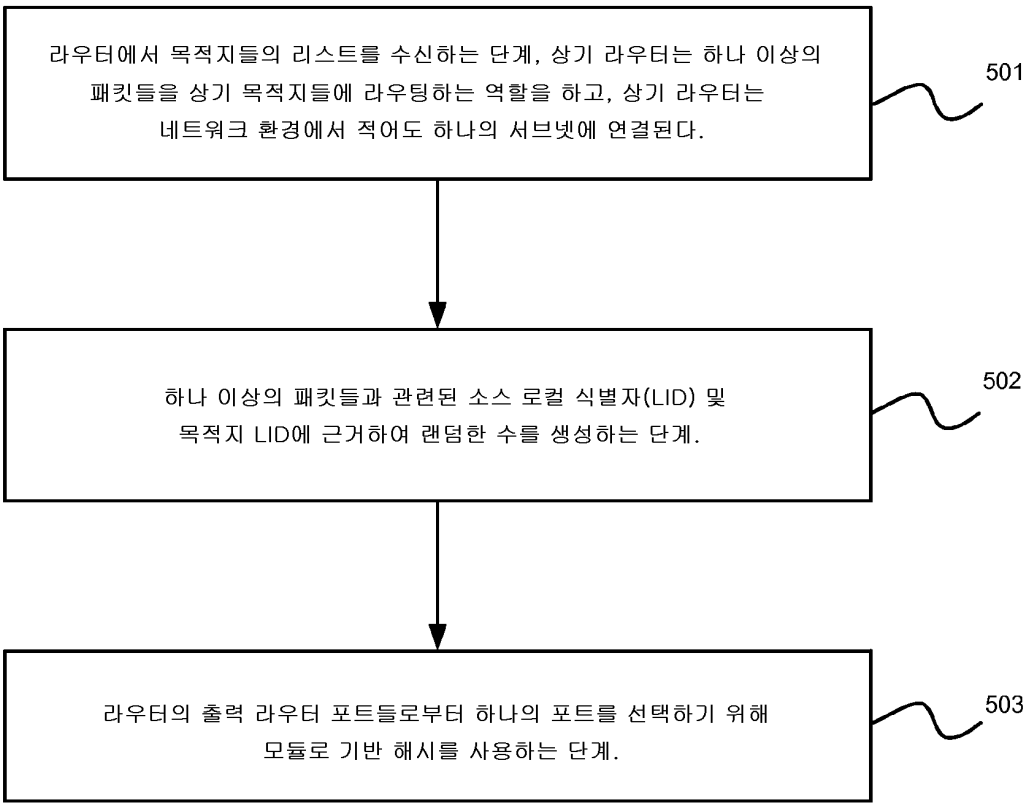
도면3


300
IB 서브넷
310
목적지
A
311
목적지
B
312
목적지
C
313
목적지
D
314
SM
303
매핑 파일
304
라우터 A
301
라우터 B
302
1
2
1
2

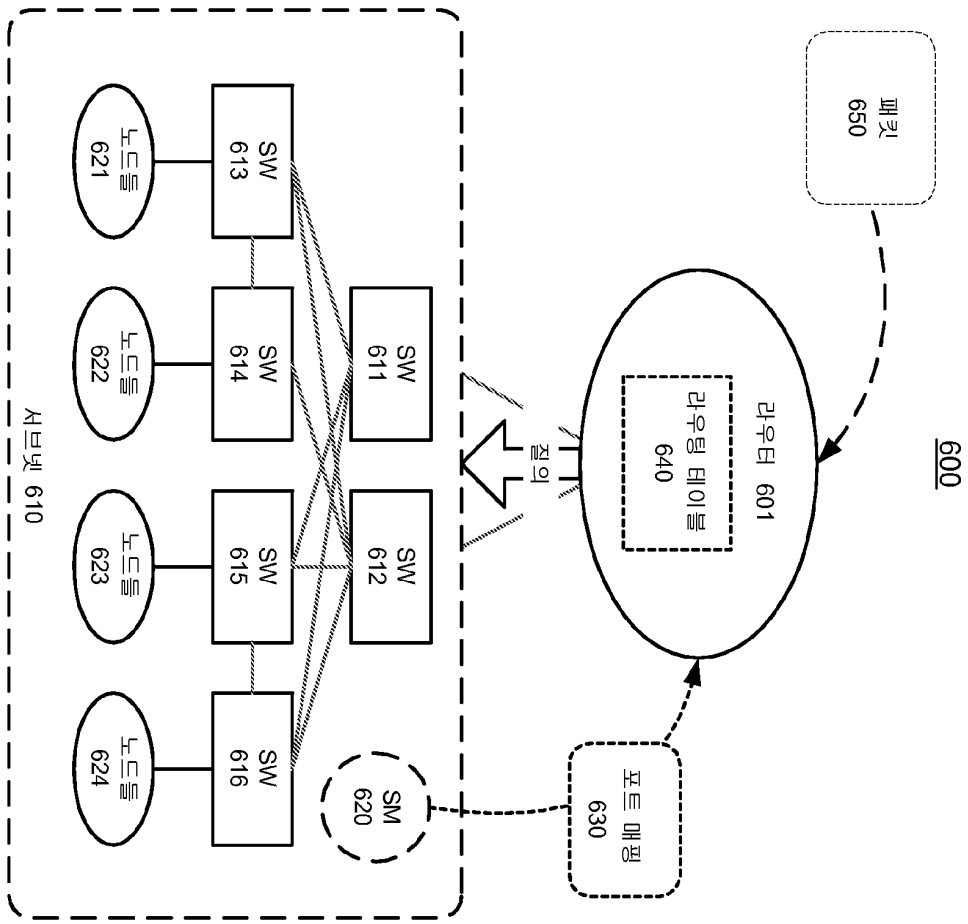
도면4


400
소스 A
411
소스 B
413
IB 서브넷
404
라우터 A
401
목적지
X
421
IB 서브넷
403
목적지
Y
422
IB 서브넷
402

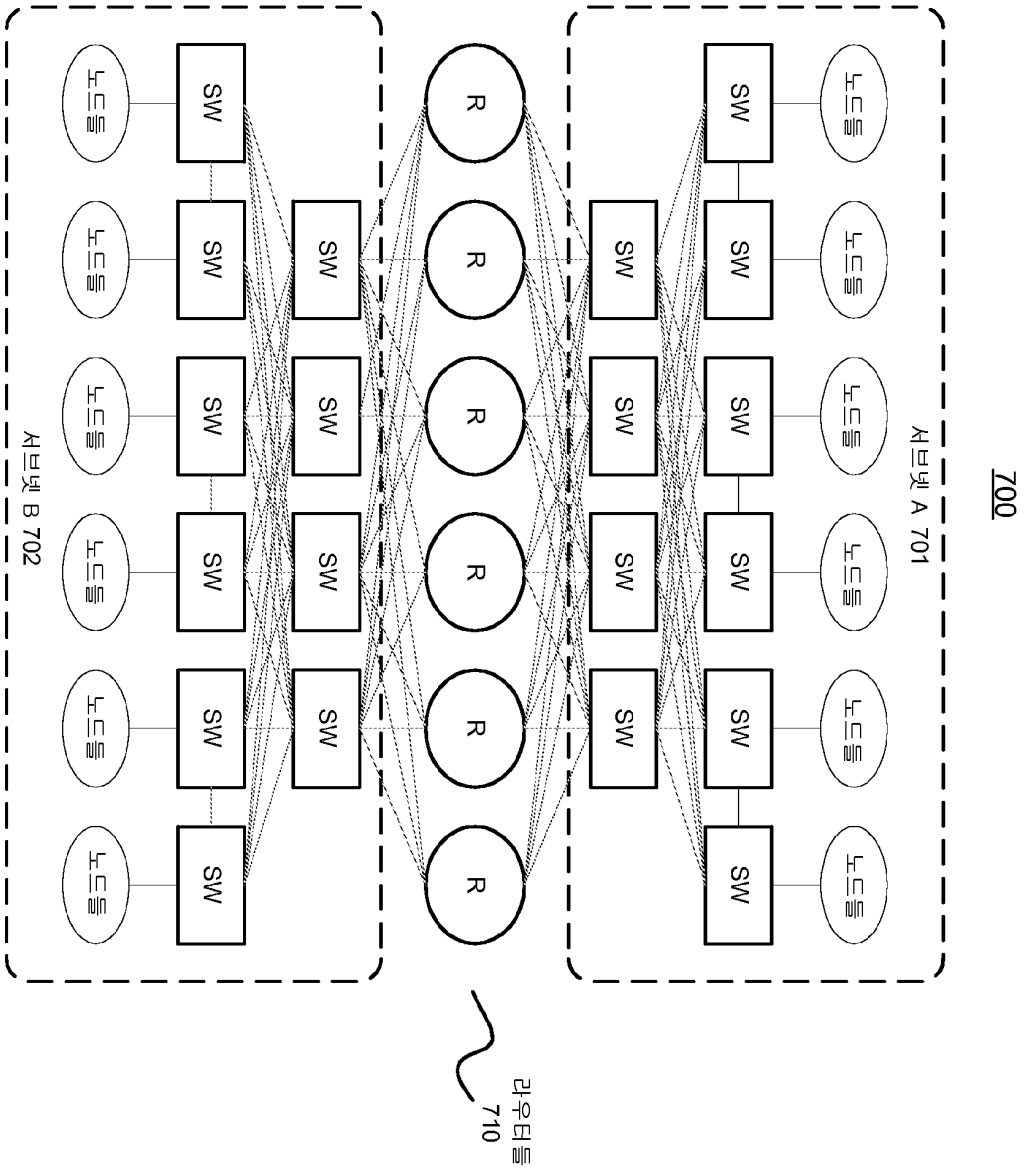
도면5


라우터에서 목적지들의 리스트를 수신하는 단계, 상기 라우터는 하나 이상의 패킷들을 상기 목적지들에 라우팅하는 역할을 하고, 상기 라우터는 네트워크 환경에서 적어도 하나의 서브넷에 연결된다.
501
하나 이상의 패킷들과 관련된 소스 로컬 식별자(LID) 및 목적지 LID에 근거하여 랜덤한 수를 생성하는 단계.
502
라우터의 출력 라우터 포트들로부터 하나의 포트를 선택하기 위해 모듈로 기반 해시를 사용하는 단계.
503

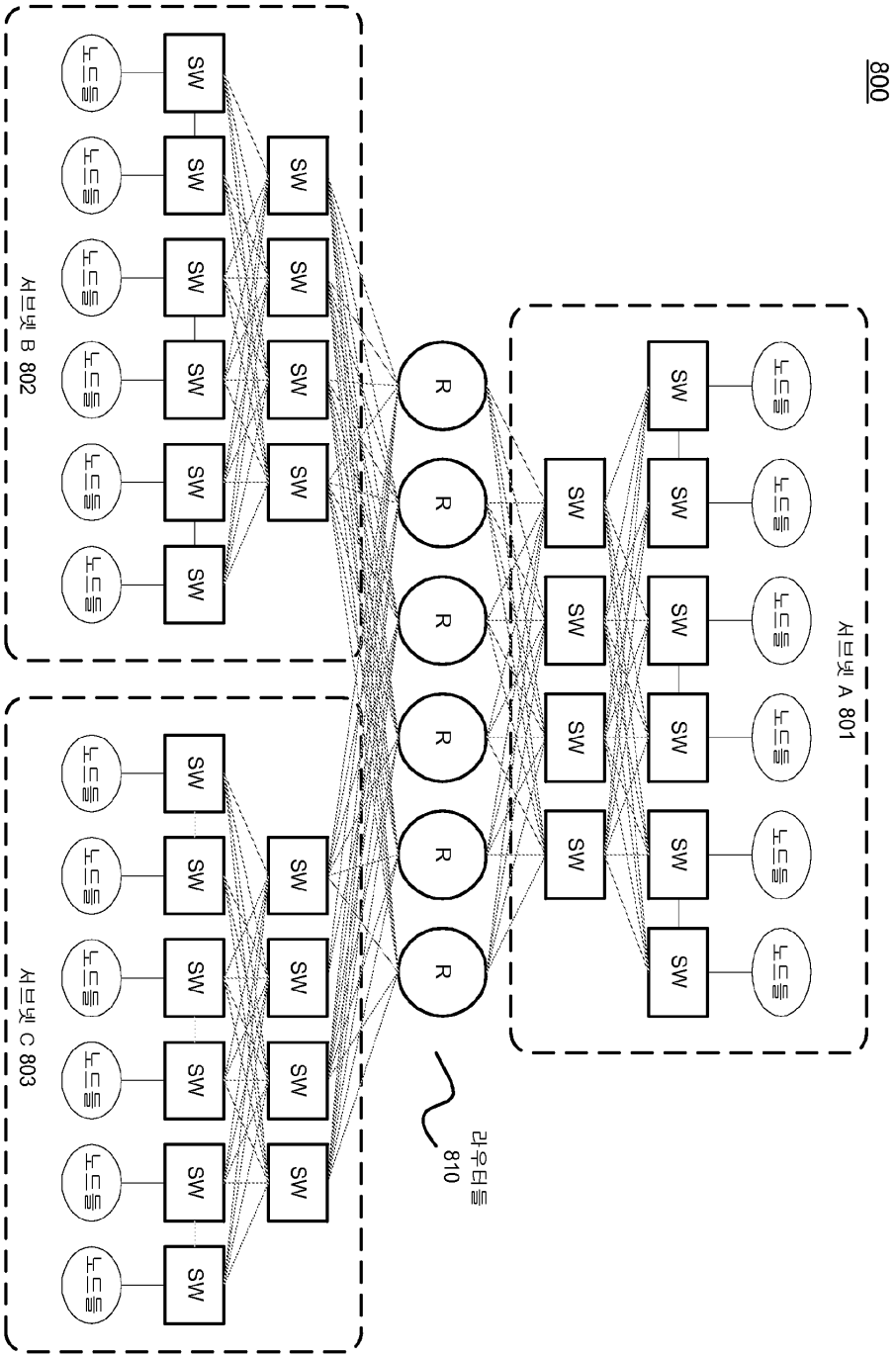
도면6


600
패킷
650
라우터 601
라우팅 테이블
640
질의
서브넷 610
SW
611
SW
612
SW
613
SW
614
SW
615
SW
616
노드들
621
노드들
622
노드들
623
노드들
624
SM
620
포트 매핑
630

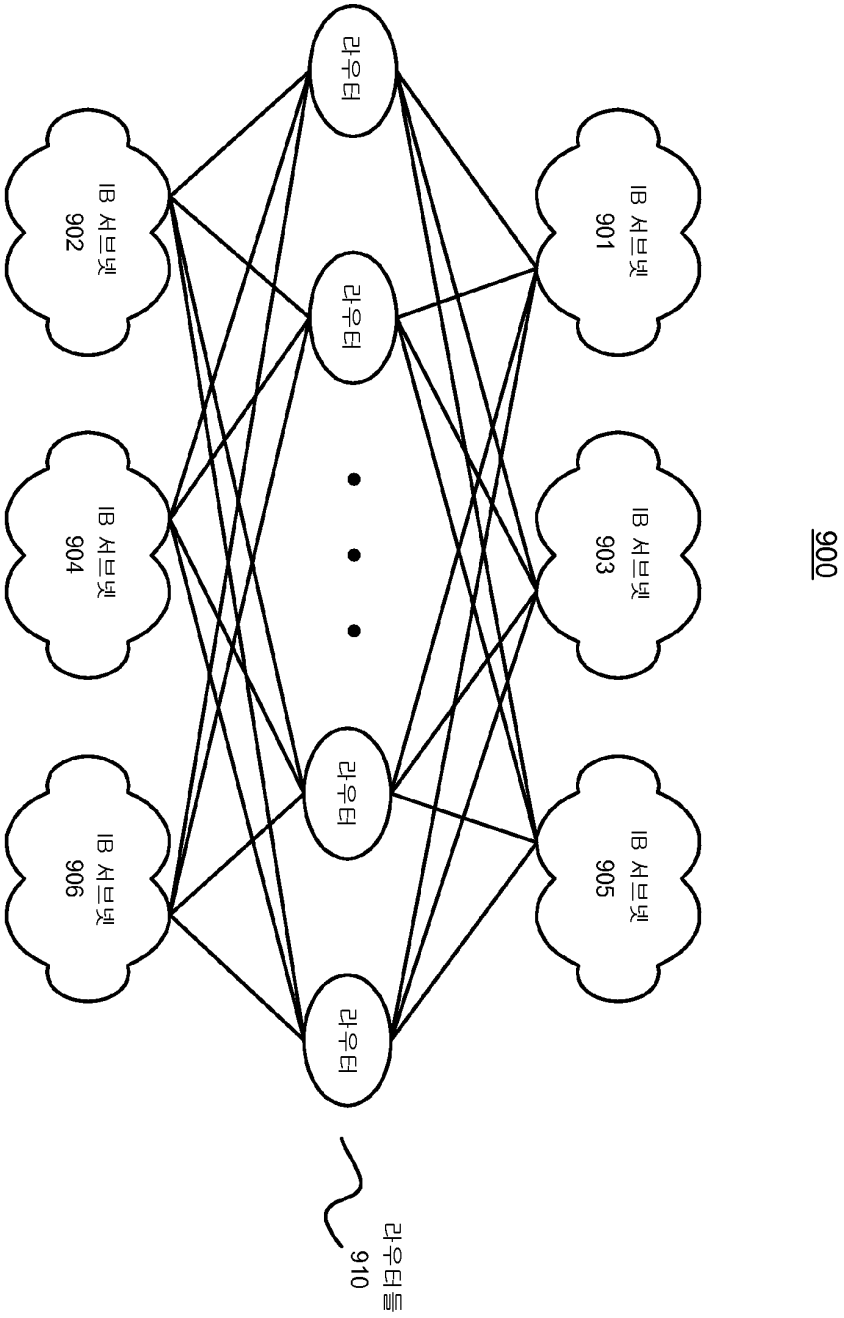
도면7



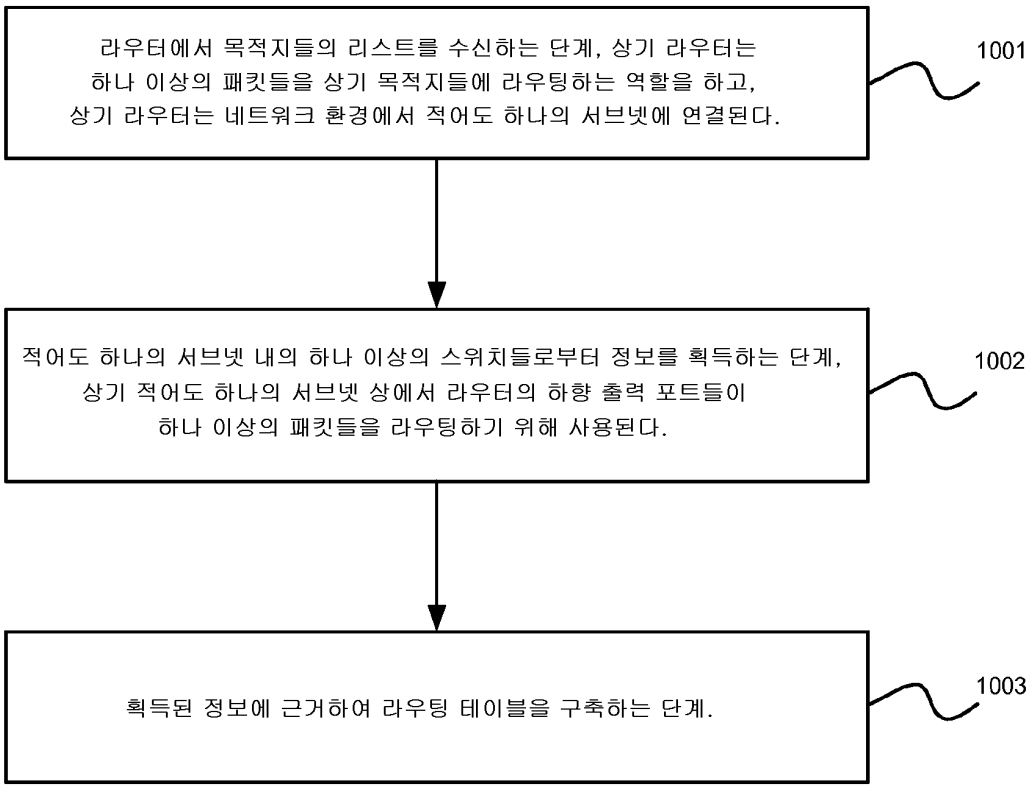
도면8



도면9


900
IB 서브넷 901
IB 서브넷 902
IB 서브넷 903
IB 서브넷 904
IB 서브넷 905
IB 서브넷 906
라우터
라우터
라우터
라우터
라우터들 910

도면10


라우터에서 목적지들의 리스트를 수신하는 단계, 상기 라우터는
하나 이상의 패킷들을 상기 목적지들에 라우팅하는 역할을 하고,
상기 라우터는 네트워크 환경에서 적어도 하나의 서브넷에 연결된다.
1001
적어도 하나의 서브넷 내의 하나 이상의 스위치들로부터 정보를 획득하는 단계,
상기 적어도 하나의 서브넷 상에서 라우터의 하향 출력 포트들이
하나 이상의 패킷들을 라우팅하기 위해 사용된다.
1002
획득된 정보에 근거하여 라우팅 테이블을 구축하는 단계.
1003

도면11

