



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2020년05월22일

(11) 등록번호 10-2113749

(24) 등록일자 2020년05월15일

(51) 국제특허분류(Int. Cl.)  
H04L 12/947 (2013.01) H04L 12/46 (2006.01)  
H04L 12/937 (2013.01)

(21) 출원번호 10-2014-7032476

(22) 출원일자(국제) 2013년05월08일

심사청구일자 2018년03월08일

(85) 번역문제출일자 2014년11월19일

(65) 공개번호 10-2015-0009550

(43) 공개일자 2015년01월26일

(86) 국제출원번호 PCT/US2013/040210

(87) 국제공개번호 WO 2013/169948

국제공개일자 2013년11월14일

(30) 우선권주장

13/889,088 2013년05월07일 미국(US)

(뒷면에 계속)

(56) 선행기술조사문헌

US20040024903 A1\*

US20080285562 A1\*

\*는 심사관에 의하여 인용된 문헌

(73) 특허권자

오라클 인터내셔널 코퍼레이션

미국, 캘리포니아 94065, 레드우드 쇼어스 엠에스 5오피7, 오라클 파크웨이 500

(72) 발명자

복단스키 바르토즈

노르웨이 오슬로 엔-0275 에이치0203 호프 테라쎬 15

존센 본 닥

노르웨이 오슬로 엔-0687 빌베르그렌다 9

(74) 대리인

박장원

전체 청구항 수 : 총 20 항

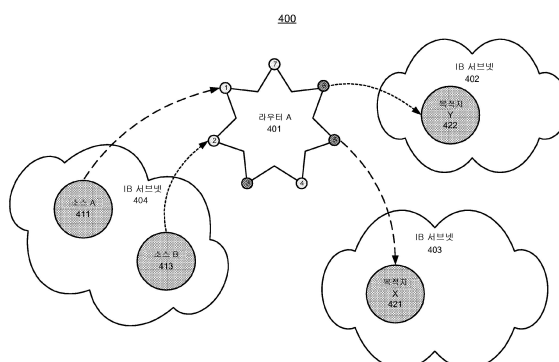
심사관 : 오수정

(54) 발명의 명칭 **소스 라우팅에 근거한 별개의 인피니넷 서브넷들 간에 트래픽을 라우팅하기 위한 시스템 및 방법**

### (57) 요약

시스템 및 방법은 네트워크 환경에서 별개의(distinct) 서브넷들 간의 트래픽을 라우팅할 수 있다. 인피니넷(IB) 서브넷들과 같은 별개의 서브넷들을 연결하는 라우터는 목적지들의 리스트를 수신할 수 있는 바, 상기 라우터는 하나 이상의 패킷들을 상기 목적지들에 라우팅하는 역할을 한다. 그 다음, 상기 라우터는 하나 이상의 패킷들과 관련된 소스 로컬 식별자(LID) 및 목적지 LID에 근거하여 랜덤한 수를 생성하고, 상기 라우터의 복수의 출력 라우터 포트들로부터 하나의 라우터 포트를 선택하기 위해 모듈로 기반 해시(modulo based hash)를 사용할 수 있다.

### 대표도



(30) 우선권주장

13/889,123 2013년05월07일 미국(US)

61/646,107 2012년05월11일 미국(US)

---

## 명세서

### 청구범위

#### 청구항 1

제 1 서브넷의 제 1 복수의 노드들과 제 2 서브넷의 제 2 복수의 노드들 사이에서 복수의 메시지들을 라우팅하기 위한 방법으로서, 상기 제 1 서브넷에 연결된 복수의 진입 포트들(ingress ports) 및 상기 제 2 서브넷에 연결된 복수의 출구 포트들(egress ports)을 갖는 하나 이상의 라우터들에 의해서 상기 제 1 서브넷과 제 2 서브넷이 직접 연결되며, 상기 방법은,

상기 복수의 메시지들의 각각의 메시지를 라우팅하되,

상기 메시지에서 소스 노드 식별자를 판독함으로써 상기 제 1 서브넷의 제 1 복수의 노드들 중에서 소스 노드를 결정하고,

상기 메시지에서 목적지 노드 식별자를 판독함으로써 상기 제 2 서브넷의 제 2 복수의 노드들 중에서 목적지 노드를 결정하고,

상기 목적지 노드 식별자에 응답하여 상기 메시지를 수신하도록 상기 라우터 상의 상기 복수의 진입 포트들 중에서 하나의 진입 포트를 선택하고,

랜덤 넘버 생성기 함수를 시딩(seed)하도록 상기 소스 노드 식별자와 조합하여 상기 목적지 노드 식별자를 사용함에 의해서 그리고 상기 출구 포트를 선택하도록 상기 랜덤 넘버 생성기 함수의 출력을 이용함에 의해서, 상기 라우터 상의 상기 복수의 출구 포트들 중에서 하나의 출구 포트를 선택하고,

상기 소스 노드로부터 상기 진입 포트로 그리고 상기 출구 포트로부터 상기 목적지 노드로 상기 메시지를 전송함에 의해서 상기 복수의 메시지들의 각각의 메시지를 라우팅하는 단계를 포함하고,

상기 소스 노드와 목적지 노드의 각각의 조합에 대하여 모든 메시지들이 상기 하나 이상의 라우터들의 동일한 진입 포트 및 출구 포트를 통해 라우팅되어, 메시지들의 순서를 벗어난 배송(out of order delivery)을 방지할 수 있는 것을 특징으로 하는 메시지들을 라우팅하기 위한 방법.

#### 청구항 2

제1항에 있어서,

상기 랜덤 넘버 생성기 함수를 시딩(seed)하도록 상기 소스 노드 식별자와 조합하여 상기 목적지 노드 식별자를 사용함에 의해서 그리고 상기 출구 포트를 선택하도록 상기 랜덤 넘버 생성기 함수의 출력을 이용함에 의해서, 상기 라우터 상의 상기 복수의 출구 포트들 중에서 하나의 출구 포트를 선택하는 것은,

상기 출구 포트를 선택하도록 모듈로 기반 해시 함수(modulo based hash function)에서 상기 랜덤 넘버 생성기 함수의 출력을 이용하는 것을 더 포함하는, 메시지들을 라우팅하기 위한 방법.

#### 청구항 3

제2항에 있어서,

라우터 펌웨어에서 상기 모듈로 기반 해시 함수를 구현하는 것을 더 포함하는, 메시지들을 라우팅하기 위한 방법.

#### 청구항 4

제1항에 있어서,

상기 목적지 노드 식별자에 응답하여 상기 메시지를 수신하도록 상기 라우터 상의 상기 복수의 진입 포트들 중에서 하나의 진입 포트를 선택하는 것은, 상기 제 2 서브넷의 제 2 복수의 노드들을 상기 복수의 진입 포트들과 연관시키는 포트 매핑 파일을 이용하여 수행되는 것을 특징으로 하는 메시지들을 라우팅하기 위한 방법.

#### 청구항 5

제1항에 있어서,

상기 제 1 서브넷은 상기 제 2 서브넷과는 상이한 네트워크 토폴로지를 갖는 것을 특징으로 하는 메시지들을 라우팅하기 위한 방법.

#### 청구항 6

제1항에 있어서,

상기 목적지 노드 식별자에 응답하여 상기 메시지를 수신하도록 상기 라우터 상의 상기 복수의 진입 포트들 중에서 하나의 진입 포트를 선택하는 것은, 상기 제 1 서브넷의 서브넷 관리자를 이용하여 수행되는 것을 특징으로 하는 메시지들을 라우팅하기 위한 방법.

#### 청구항 7

제1항에 있어서,

상기 목적지 노드 식별자에 응답하여 상기 메시지를 수신하도록 상기 라우터 상의 상기 복수의 진입 포트들 중에서 하나의 진입 포트를 선택하는 것은, 라우터 포트들을 선택함에 있어서 충분한 입도(full granularity)를 제공하는 포트 매핑 파일을 이용하여 수행되는 것을 특징으로 하는 메시지들을 라우팅하기 위한 방법.

#### 청구항 8

제4항에 있어서,

상기 포트 매핑 파일은, 유사한 개수의 상기 제 2 복수의 노드들을 상기 복수의 진입 포트들 각각과 연관시키는 것을 특징으로 하는 메시지들을 라우팅하기 위한 방법.

#### 청구항 9

제4항에 있어서,

상기 포트 매핑 파일은, 라운드 로빈 분배를 이용하여 유사한 개수의 상기 제 2 복수의 노드들을 상기 복수의 진입 포트들 각각과 연관시키는 것을 특징으로 하는 메시지들을 라우팅하기 위한 방법.

#### 청구항 10

제1항에 있어서,

상기 제 1 서브넷 및 제 2 서브넷 중 하나는 팻트리 토폴로지(fat-tree topology)를 갖는 것을 특징으로 하는 메시지들을 라우팅하기 위한 방법.

#### 청구항 11

메시지들을 라우팅하기 위한 시스템으로서,

제 1 복수의 노드들을 갖는 제 1 서브넷;

제 2 복수의 노드들을 갖는 제 2 서브넷; 및

상기 제 1 서브넷과 제 2 서브넷을 직접 연결하는 하나 이상의 라우터들을 포함하고, 상기 하나 이상의 라우터들은 상기 제 1 서브넷에 연결된 복수의 진입 포트들 및 상기 제 2 서브넷에 연결된 복수의 출구 포트들을 가지며,

상기 시스템은 각각의 메시지를 라우팅하도록 구성되되,

상기 메시지에서부터 소스 노드 식별자를 판독함으로써 상기 제 1 서브넷의 제 1 복수의 노드들 중에서 소스 노드를 결정하고,

상기 메시지에서부터 목적지 노드 식별자를 판독함으로써 상기 제 2 서브넷의 제 2 복수의 노드들 중에서 목적지 노드를 결정하고,

상기 목적지 노드 식별자에 응답하여 상기 메시지를 수신하도록 상기 라우터 상의 상기 복수의 진입 포트들 중에서 하나의 진입 포트를 선택하고,

랜덤 넘버 생성기 함수를 시딩(seed)하도록 상기 소스 노드 식별자와 조합하여 상기 목적지 노드 식별자를 사용함에 의해서 그리고 상기 출구 포트를 선택하도록 상기 랜덤 넘버 생성기 함수의 출력을 이용함에 의해서, 상기 라우터 상의 상기 복수의 출구 포트들 중에서 하나의 출구 포트를 선택하고,

상기 소스 노드로부터 상기 진입 포트에 그리고 상기 출구 포트로부터 상기 목적지 노드로 상기 메시지를 전송함에 의해서 각각의 메시지를 라우팅하도록 구성되고,

상기 소스 노드와 목적지 노드의 각각의 조합에 대하여 모든 메시지들이 상기 하나 이상의 라우터들의 동일한 진입 포트 및 출구 포트를 통해 라우팅되어, 메시지들의 순서를 벗어난 배송(out of order delivery)을 방지할 수 있는 것을 특징으로 하는 메시지들을 라우팅하기 위한 시스템.

## 청구항 12

제11항에 있어서,

상기 랜덤 넘버 생성기 함수를 시딩(seed)하도록 상기 소스 노드 식별자와 조합하여 상기 목적지 노드 식별자를 사용함에 의해서 그리고 상기 출구 포트를 선택하도록 상기 랜덤 넘버 생성기 함수의 출력을 이용함에 의해서, 상기 라우터 상의 상기 복수의 출구 포트들 중에서 하나의 출구 포트를 선택하는 것은,

상기 출구 포트를 선택하도록 모듈로 기반 해시 함수(modulo based hash function)에서 상기 랜덤 넘버 생성기 함수의 출력을 이용하는 것을 더 포함하는, 메시지들을 라우팅하기 위한 시스템.

## 청구항 13

제12항에 있어서,

상기 모듈로 기반 해시 함수는 상기 하나 이상의 라우터들의 라우터 펌웨어를 이용하여 구현되는 것을 특징으로 하는 메시지들을 라우팅하기 위한 시스템.

## 청구항 14

제11항에 있어서,

상기 제 2 서브넷의 제 2 복수의 노드들을 상기 복수의 진입 포트들과 연관시키는 포트 매핑 파일을 더 포함하며,

상기 목적지 노드 식별자에 응답하여 상기 메시지를 수신하도록 상기 라우터 상의 상기 복수의 진입 포트들 중에서 하나의 진입 포트를 선택하는 것은, 상기 포트 매핑 파일을 이용하여 수행되는 것을 특징으로 하는 메시지들을 라우팅하기 위한 시스템.

## 청구항 15

제11항에 있어서,

상기 제 1 서브넷은 상기 제 2 서브넷과는 상이한 네트워크 토폴로지를 갖는 것을 특징으로 하는 메시지들을 라우팅하기 위한 시스템.

## 청구항 16

제11항에 있어서,

상기 제 1 서브넷의 서브넷 관리자를 더 포함하며,

상기 서브넷 관리자가, 상기 목적지 노드 식별자에 응답하여 상기 메시지를 수신하도록 상기 라우터 상의 상기 복수의 진입 포트들 중에서 하나의 진입 포트를 선택하는 것을 특징으로 하는 메시지들을 라우팅하기 위한 시스템.

## 청구항 17

제11항에 있어서,

충분한 선택 입도(full selection granularity)로 상기 제 2 서브넷의 제 2 복수의 노드들을 상기 복수의 진입 포트들과 연관시키는 포트 매핑 파일을 더 포함하며,

상기 목적지 노드 식별자에 응답하여 상기 메시지를 수신하도록 상기 라우터 상의 상기 복수의 진입 포트들 중에서 하나의 진입 포트를 선택하는 것은, 상기 포트 매핑 파일을 이용하여 수행되는 것을 특징으로 하는 메시지들을 라우팅하기 위한 시스템.

#### 청구항 18

제11항에 있어서,

상기 제 2 서브넷의 제 2 복수의 노드들을 상기 복수의 진입 포트들과 연관시키는 포트 매핑 파일을 더 포함하고, 상기 포트 매핑 파일은 유사한 개수의 상기 제 2 복수의 노드들을 상기 복수의 진입 포트들 각각과 연관시키며,

상기 목적지 노드 식별자에 응답하여 상기 메시지를 수신하도록 상기 라우터 상의 상기 복수의 진입 포트들 중에서 하나의 진입 포트를 선택하는 것은, 상기 포트 매핑 파일을 이용하여 수행되는 것을 특징으로 하는 메시지들을 라우팅하기 위한 시스템.

#### 청구항 19

제14항에 있어서,

상기 포트 매핑 파일은, 라운드 로빈 분배를 이용하여 유사한 개수의 상기 제 2 복수의 노드들을 상기 복수의 진입 포트들 각각과 연관시키는 것을 특징으로 하는 메시지들을 라우팅하기 위한 시스템.

#### 청구항 20

제 1 서브넷의 제 1 복수의 노드들과 제 2 서브넷의 제 2 복수의 노드들 사이에서 복수의 메시지들을 라우팅하기 위한, 명령들이 저장되어 있는 비일시적 컴퓨터 판독가능 매체로서, 상기 제 1 서브넷에 연결된 복수의 진입 포트들 및 상기 제 2 서브넷에 연결된 복수의 출구 포트들을 갖는 하나 이상의 라우터들에 의해서 상기 제 1 서브넷과 제 2 서브넷이 직접 연결되며, 상기 명령들은 실행되는 때에 컴퓨터로 하여금,

상기 복수의 메시지들의 각각의 메시지를 라우팅하게 하되,

상기 메시지에서부터 소스 노드 식별자를 판독함으로써 상기 제 1 서브넷의 제 1 복수의 노드들 중에서 소스 노드를 결정하고,

상기 메시지에서부터 목적지 노드 식별자를 판독함으로써 상기 제 2 서브넷의 제 2 복수의 노드들 중에서 목적지 노드를 결정하고,

상기 목적지 노드 식별자에 응답하여 상기 메시지를 수신하도록 상기 라우터 상의 상기 복수의 진입 포트들 중에서 하나의 진입 포트를 선택하고,

랜덤 넘버 생성기 함수를 시딩(seed)하도록 상기 소스 노드 식별자와 조합하여 상기 목적지 노드 식별자를 사용함에 의해서 그리고 상기 출구 포트를 선택하도록 상기 랜덤 넘버 생성기 함수의 출력을 이용함에 의해서, 상기 라우터 상의 상기 복수의 출구 포트들 중에서 하나의 출구 포트를 선택하고,

상기 소스 노드로부터 상기 진입 포트로부터 그리고 상기 출구 포트로부터 상기 목적지 노드로 상기 메시지를 전송함에 의해서 상기 복수의 메시지들의 각각의 메시지를 라우팅하게 하고,

상기 소스 노드와 목적지 노드의 각각의 조합에 대하여 모든 메시지들이 상기 하나 이상의 라우터들의 동일한 진입 포트 및 출구 포트를 통해 라우팅되어, 메시지들의 순서를 벗어난 배송(out of order delivery)을 방지할 수 있는 것을 특징으로 하는 메시지들을 라우팅하기 위한 비일시적 컴퓨터 판독가능 매체.

#### 청구항 21

삭제

## 청구항 22

삭제

## 청구항 23

삭제

## 발명의 설명

### 기술 분야

[0001] 저작권 공지

[0002] 본 명세서에서 개시된 부분은 저작권 보호를 받는 내용을 포함한다. 저작권자는 미국특허상표청의 특허 파일 또는 기록에 나타난 대로 본 특허 문서 또는 특허 개시내용을 어느 누군가가 복사하여 재생하는 것은 반대하지 않지만, 그 밖의 모든 것은 저작권으로 보호된다.

[0003] 기술분야

[0004] 본 발명은 일반적으로, 컴퓨터 시스템들에 관한 것이며, 특히 미들웨어 머신 환경에 관한 것이다.

### 배경 기술

[0005] 거대한 클라우드 컴퓨팅 아키텍처들이 도입됨에 따라, 종래의 네트워크 및 스토리지와 관련된 성능 및 관리적인 병목(administrative bottleneck)들이 상당한 문제가 되어 왔다. 인피니밴드(IB) 기술은 클라우드 컴퓨팅 패브릭(fabric)을 위한 토대(foundation)로서 증가된 디플로이먼트(deployment)를 보여 왔다. 이는 본 발명의 실시예들이 해결하고자 의도한 일반적인 영역이다.

### 발명의 내용

[0006] 별개의(distinct) 인피니밴드(IB) 서브넷들 간의 트래픽을 라우팅할 수 있는 시스템 및 방법이 본 명세서에 기술된다. 인피니밴드 서브넷들과 같은 별개의 서브넷들을 연결하는 라우터는 목적지들의 리스트를 수신할 수 있는 바, 상기 라우터는 하나 이상의 패킷들을 상기 목적지들에 라우팅하는 역할을 한다. 그 다음, 상기 라우터는 하나 이상의 패킷들과 관련된 소스 로컬 식별자(LID) 및 목적지 LID에 근거하여 랜덤한 수를 생성하고, 상기 라우터의 복수의 출력 라우터 포트들로부터 하나의 라우터 포트를 선택하기 위해 모듈로 기반 해시(modulo based hash)를 사용할 수 있다.

[0007] 또한, 네트워크 환경에서 별개의 서브넷들 간의 트래픽을 라우팅하기 위한 시스템이 본 명세서에 기술된다. 상기 시스템은 라우터에서 목적지들의 리스트를 수신하기 위한 수단을 포함하는 바, 상기 라우터는 하나 이상의 패킷들을 상기 목적지들에 라우팅하는 역할을 하며, 상기 라우터는 네트워크 환경에서 적어도 하나의 서브넷에 연결된다. 상기 시스템은 하나 이상의 패킷들과 관련된 소스 LID 및 목적지 LID에 근거하여 랜덤한 숫자를 생성하기 위한 수단을 더 포함한다. 상기 시스템은 상기 라우터의 복수의 출력 라우터 포트들로부터 하나의 라우터 포트를 선택하기 위해 모듈로 기반 해시를 이용하기 위한 수단을 더 포함한다.

[0008] 또한, 네트워크 환경에서 별개의 서브넷들 간의 트래픽을 라우팅하기 위한 시스템이 또한 본 명세서에 기술된다. 상기 시스템은 목적지들의 리스트를 수신하도록 구성된 라우터를 포함하는 바, 상기 라우터는 하나 이상의 패킷들을 상기 목적지들에 라우팅하는 역할을 하며, 상기 라우터는 네트워크 환경에서 적어도 하나의 서브넷에 연결된다. 상기 시스템은 또한, 하나 이상의 패킷들과 관련된 소스 LID 및 목적지 LID에 근거하여 랜덤한 숫자를 생성하고, 상기 라우터의 출력 라우터 포트들로부터 하나의 라우터 포트를 선택하기 위해 모듈로 기반 해시를 이용하도록 구성된다.

### 도면의 간단한 설명

[0009] 도 1은 본 발명의 실시예에 따른 네트워크 환경에서 별개의 인피니밴드 서브넷들 간에 트래픽을 라우팅하는 예를 도시한다.

도 2는 본 발명의 실시예에 따른 네트워크 환경에서 라우터 상에서의 패킷 포워딩을 지원하는 예를 도시한다.

도 3은 본 발명의 실시예에 따른 네트워크 환경에서 서로 다른 목적지들에 대한 진입 라우터 포트들을 선택하는 예를 도시한다.

도 4는 본 발명의 실시예에 따른 네트워크 환경에서, 라우터에서 수신된 패킷에 대한 출구 라우터 포트를 선택하는 예를 도시한다.

도 5는 본 발명의 실시예에 따른 네트워크 환경에서, 라우터에서 서브넷 간 소스 라우팅(ISSR)을 지원하기 위한 예시적인 순서도를 도시한다.

도 6은 본 발명의 실시예에 따른 네트워크 환경에서, 라우터에서 서브넷 간 팻트리(fat-tree) 라우팅(ISFR) 알고리즘을 지원하는 예를 도시한다.

도 7 내지 9는 본 발명의 실시예에 따른 서로 다른 토폴로지들을 갖는 네트워크 환경에서의 라우팅을 도시한다.

도 10은 본 발명의 실시예에 따른 네트워크 환경에서, 라우터에서 서브넷 간 팻트리 라우팅(ISFR) 알고리즘을 지원하는 예시적인 순서도를 도시한다.

도 11은 본 발명의 실시예에 따른 특징들을 도시하기 위한 기능 블록도를 도시한다.

### 발명을 실시하기 위한 구체적인 내용

[0010] 본 발명은 예로서 그리고 제한이 아닌 것으로서 첨부 도면들의 도해들에서 예시되며, 상기 도면들에서 유사한 도면 부호들은 유사한 요소들을 나타낸다. 주목할 점으로서, 본 발명에서 "일" 또는 "한" 또는 "일부" 실시예(들)에 대한 참조들은 반드시 동일한 실시예에 대한 것이 아니며, 이러한 참조들은 적어도 일 실시예를 의미한다.

[0011] 다음의 본 발명의 설명은 인피니밴드 네트워크를 고성능 네트워크에 대한 예로서 이용한다. 다른 타입의 고성능 네트워크들이 제한 없이 이용될 수 있음이 이 기술분야의 숙련자들에게 분명할 것이다. 또한, 다음의 본 발명의 설명은 팻트리를 네트워크 토폴로지 모델에 대한 예로서 이용한다. 다른 타입의 네트워크 토폴로지 모델들이 제한 없이 이용될 수 있음이 이 기술분야의 숙련자들에게 분명할 것이다.

[0012] 별개의 네트워크들 간에 트래픽을 라우팅하는 것을 지원할 수 있는 시스템들 및 방법들이 본 명세서에 기술된다.

[0013] 인피니밴드 아키텍처

[0014] IB 아키텍처는 직렬 단대단 풀-듀플렉스 기술(serial point-to-point full-duplex technology)이다. IB 클러스터들의 사이즈 및 복잡도가 커짐에 따라, 네트워크는 관리가능한 섹션들로 세그멘테이션(segmentation)될 수 있는 바, 상기 섹션들은 서브넷들로 지칭될 수 있다. IB 서브넷은 스위치들 및 단대단 링크들을 이용하여 상호연결되는 호스트들의 세트를 포함할 수 있다. IB 서브넷은 또한, 적어도 하나의 서브넷 관리자(SM)를 포함할 수 있고, 상기 SM은 서브넷 내의 모든 스위치들, 라우터들 및 호스트 채널 어댑터들(HCA)의 구성을 포함하여, 상기 네트워크를 개시하고 및 실행하는(bringing up) 역할을 한다.

[0015] IB는 원격 다이렉트 메모리 액세스(RDMA) 및 종래의 전송/수신 시맨틱(semantic)들 모두를 제공하기 위해 전송 서비스들의 리치 세트(rich set)를 지원한다. 사용되는 전송 서비스에 관계없이, IB HCA들은 큐 쌍들(QP)을 이용하여 통신한다. QP는 통신 설정 동안 생성되며, 공급되는 QP 수, HCA 포트, 목적지 LID 큐 사이즈들 및 전송 서비스와 같은 초기 속성들의 세트를 가질 수 있다. HCA는 많은 QP들을 다룰 수 있으며, 각각의 QP는 전송 큐(SQ) 및 수신 큐(RQ)와 같은 큐들의 쌍으로 구성되고, 통신에 참여하는 각각의 엔드-노드에 이러한 하나의 쌍이 존재한다. 전송 큐는 원격 노드에 전달될 작업 요청들을 유지(hold)하고 수신 큐는 원격 노드로부터 수신되는 데이터로 무엇을 할지에 관한 정보를 유지한다. QP들에 추가적으로, 각각의 HCA는 하나 이상의 완료 큐들(CQ)을 가지며, 상기 완료 큐들은 전송 및 수신 큐들의 세트와 관련된다. CQ는 전송 및 수신 큐에 포스팅되는 작업 요청들에 대한 완료 통지들을 유지한다. 비록, 통신의 복잡성들이 사용자로부터 감추어 지지만, QP 상태 정보는 HCA에 유지된다.

[0016] 별개의 인피니밴드 서브넷들 간의 트래픽 라우팅

[0017] 도 1은 본 발명의 실시예에 따른 네트워크 환경에서 별개의 인피니밴드 서브넷들 간에 트래픽을 라우팅하는 예를 도시한다. 도 1에 도시된 바와 같이, 네트워크 환경(100)은 인피니밴드 아키텍처(IBA)에 근거할 수 있고, 2-층 토폴로지 분할(two-layer topological division)을 지원한다.



- [0018] IB 네트워크 또는 IB 패브릭(100)의 하위 계층은 서브넷들, 예컨대 서브넷들(101 내지 104)로서 지칭될 수 있고, 이들 각각은 스위치들 및 단대단 링크들을 이용하여 상호연결된 호스트들의 세트를 포함한다. 상위 계층에서, IB 패브릭(100) 내의 하나 이상의 서브넷들(101 내지 104)은 라우터들, 예컨대 라우터들(105 내지 106)을 이용하여 상호연결될 수 있다. 더욱이, 각각의 서브넷(101 내지 104)은 로컬 서브넷 상의 포트들만을 구성하는 자신 소유의 서브넷 관리자(SM)를 실행시킬 수 있고, 라우터들(105 내지 106)은 서브넷 관리자들(SM)에 투명하지 않다(non-transparent).
- [0019] 서브넷들(101 내지 104)의 각각 내의 호스트들 및 스위치들은 지정된 로컬 식별자들(LID)을 이용하여 어드레스될 수 있다. 큰 인스톨레이션(installation)의 사이즈는 이용가능한 로컬 식별자(LID)들의 수에 의해 제한될 수 있다. 예를 들어, 단일 서브넷은 49151개의 유니캐스트 LID들에 제한될 수 있다. IB 어드레스 공간을 확장하기 위한 일 접근법은 LID 어드레싱 공간을 32 비트로 확장하는 것이며, 이의 유용성(usability)은 이 접근법이 오래된 하드웨어와 호환가능하지(backward compatible) 않을 수 있기 때문에 제한될 수 있다.
- [0020] 본 발명의 실시예에 따르면, IB 라우터들(105 내지 106)은 IB 패브릭(100) 내의 어드레스 공간 스케일러빌리티(scalability)를 제공할 수 있다. 도 1에 도시된 바와 같이, 더 많은 엔드 포트들이 요구될 때, 복수의 서브넷들(101 내지 104)은 하나 이상의 IB 라우터들(105 내지 106)을 이용하여 함께 결합될 수 있다. 서브넷들(101 내지 104) 각각은 로컬 식별자 어드레스 공간(111 내지 114)을 이용할 수 있다. LID 어드레스들이 각각의 LID 어드레스 공간(111 내지 114) 내에 로컬한 가시도(visibility)를 가지기 때문에, LID 어드레스들은 라우터들(105 내지 106)에 의해 연결되는 서로 다른 서브넷들(101 내지 104)에서 재사용될 수 있다. 따라서, 어드레스 공간 스케일러빌리티가 IB 패브릭(100)에 제공될 수 있고, 이러한 접근법은 이론적으로는 비제한적 어드레싱 공간을 생산(yield)할 수 있다.
- [0021] 더욱이, 크고 복잡한 네트워크(100)를 복수의 서브넷들(101 내지 104)로 세그멘테이션함으로써, 시스템은 패브릭 관리 봉쇄(fabric management containment)를 제공할 수 있다. 패브릭 관리 봉쇄는 1) 결합 격리, 2) 증가된 보안성, 3) 서브넷 간 라우팅 플렉서빌리티(flexibility)를 제공한다는 점에서 이익일 수 있다.
- [0022] 첫째로, 큰 네트워크(100)를 여러 작은 서브넷들(101 내지 104)로 분할함으로써, 결합들 및 토폴로지 변경들이 단일 서브넷에 가두어질(contained) 수 있으며, 서브넷 구성은 라우터들(105 내지 106)을 통해 다른 서브넷들에 패스될 수 없다. 이는 재구성 시간을 단축하고 결합의 영향(impact)을 제한한다.
- [0023] 둘째로, 보안의 관점에서, 라우터들(105 내지 106)을 이용하여 큰 패브릭(100)을 서브넷들(101 내지 104)로 세그멘테이션하는 것은 대부분의 어택들의 범위를 어택당한 서브넷에 제한한다.
- [0024] 셋째로, 라우팅의 관점에서, 패브릭 관리 봉쇄는 더욱 플렉서블한 라우팅 기법들을 지원함에 있어 이익일 수 있는 바, 이는 둘 이상의 정규 토폴로지들을 포함하는 하이브리드 패브릭의 경우 특히 장점적일 수 있다.
- [0025] 예를 들어, 하이브리드 패브릭(100)은 메쉬(mesh) 또는 토러스(torus) 부분(또는 어떤 다른 정규 토폴로지)과 상호연결된 팻트리 부분을 포함할 수 있다. 서브넷 간 라우팅 알고리즘들이 서브넷 범위만을 가질 수 있기 때문에 하이브리드 패브릭(100)의 서로 다른 부분들을 각각 라우팅하는 간단한 방식은 존재하지 않을 수 있다. 더욱이, 하이브리드 토폴로지에 최적 성능을 제공할 수 있는 IB에 대한 범용 불가지론(agnostic) 라우팅 알고리즘들은 존재하지 않는다.
- [0026] 본 발명의 실시예에 따르면, 하이브리드 토폴로지는 작은 정규 서브넷들(101 내지 104)로 분할될 수 있으며, 이들 각각은 특별한 서브넷에 대해 최적화되는 서로 다른 라우팅 알고리즘을 이용하여 라우팅될 수 있다. 예를 들어, 팻트리 라우팅 알고리즘은 팻트리 부분을 라우팅할 수 있고, 차원 순서(dimension-order) 라우팅은 토폴로지의 메쉬 부분을 라우팅할 수 있다.
- [0027] 더욱이, 최종 목적지가 다른 서브넷 뒤에 위치되는(예컨대, 적어도 두 개의 라우터 홉들이 요구되는) 더 많은 비정규 네트워크들의 경우, 슈퍼 서브넷 관리자가 로컬 서브넷 관리자들 간을 조직화(coordinate)하기 위해 사용될 수 있고, 전송 서브넷을 통해 경로를 확립할 수 있다.
- [0028] 따라서, IB 서브넷들 간을 라우팅하는 것을 이용함으로써, 시스템은 IB 네트워크에서 어드레스 공간 스케일러빌리티 및 패브릭 관리 봉쇄를 제공할 수 있다.
- [0029] 네이티브 인피니밴드 라우터들
- [0030] 도 2는 본 발명의 실시예에 따른 네트워크 환경에서, 라우터 상에서의 패킷 포워딩을 지원하는 예를 도시한다. 도 2에 도시된 바와 같이, 네트워크 환경(200)은 라우터(210)를 이용하여 상호연결되는 복수의 서브넷들, 예컨

대 IB 서브넷들 A 내지 B(201 내지 202)을 포함할 수 있다. 더욱이 IB 서브넷 A(201)은 호스트 채널 어댑터(HCA)(213)와 관련된 엔드 노드 X(203)에 연결되며, IB 서브넷 B(202)는 호스트 채널 어댑터(HCA)(214)와 관련된 엔드 노드 Y(204)에 연결된다.

[0031] 본 발명의 실시예에 따르면, IB 라우터(210)는 IB 어드레싱 계층구조(hierarchy)의 계층-3에서 동작할 수 있다. 더욱이, LID들에 추가적으로, 각각의 IB 디바이스는 또한, GID, IB 계층-3 어드레스를 형성하기 위해 사용될 수 있는 64-비트 글로벌 고유 식별자(GUID)를 가질 수 있다. 예를 들어, GID는 IPv6-유사 128 비트 어드레스를 형성하기 위해 64 비트 GUID와 64 비트 서브넷 ID를 연결(concatenating)함으로써 생성될 수 있다. 추가적으로, 용어 GUID는 또한, 포트 GUID들, 즉 IB 패브릭 내의 모든 포트에 할당된 GUID를 나타내기 위해 사용될 수 있고, GUID는 비 휘발성 메모리 내에 버닝(burn)될 수 있다.

[0032] 도 2에 도시된 바와 같이, 엔드 노드(예컨대 엔드 노드 X(203))는 라우터(210)를 통해 다른 엔드 노드(예컨대 엔드 노드 Y(204))에 패킷을 전송할 수 있다. 어드레스 분해(resolution) 메커니즘은 로컬 라우터(210)가 엔드 노드들 X 내지 Y(203 내지 204)에 가시적(visible)이게 할 수 있다.

[0033] 예를 들어, 라우터(210)에 수신된 패킷은 유입 라우팅 헤더(221)를 포함할 수 있고, 상기 유입 라우팅 헤더는 로컬 라우팅 헤더(LRH)(223) 및 글로벌 라우팅 헤더(GRH)(225)를 포함한다. 패킷을 포워딩하기 전에, 라우터(210)는 유입 라우팅 헤더(221)를 인출(outgoing) 라우팅 헤더(222)로 수정할 수 있으며, 상기 인출 라우팅 헤더는 LRH(224) 및 GRH(226)를 포함한다.

[0034] 도 2에 도시된 바와 같이, 엔드-노드 X(203)는 LRH(223)에 로컬 HCA의 LID 어드레스(예컨대, X)를 소스 LID, srcLID로서 입력(put)하고, 로컬 라우터의 LID 어드레스(예컨대, A)를 목적지 LID, dstLID로서 입력한다. 더욱이, 엔드 노드 X(203)는 GRH(225)에 자신의 계층-3 어드레스(GID)(예컨대, 1234)를 소스 LID, srcLID로서 입력하고, 최종 목적지 어드레스(GID)(예컨대, 5678)를 목적지 LID, dstLID로서 입력한다.

[0035] 패킷이 라우터(210)에 도달할 때, 패킷 포워딩 메커니즘(220)은 패킷 필드들을 갱신 및 교체할 수 있다. 예를 들어, 시스템은 LRH(224) 내의 소스 LID, srcLID를 라우터의 출구 포트의 LID(예컨대, B)와 교체할 수 있다. 더욱이, 시스템은 목적지 LID, dstLID를 목적지의 LID 어드레스(예컨대, Y)와 교체할 수 있다.

[0036] 대안적으로는, 시스템은 패킷이 다른 라우터로부터 포워딩되면 이전의 홉 라우터의 출구 포트의 LID를 소스 LID, srcLID로서 사용할 수 있다. 또한, 시스템은 추가의 패킷 포워딩이 필요하면 목적지 LID, dstLID를 다음 홉 포트의 LID와 교체할 수 있다. 각각의 경우, 시스템은 다음 홉에 패킷을 포워딩하기 전에 순환 리던던시 체크(CRC)를 재컴퓨팅할 수 있다.

[0037] 본 발명의 실시예에 따르면, 서로 다른 방법들이 별개의 인피니밴드 서브넷들 간에 트래픽을 라우팅하기 위해 사용될 수 있다. 이 방법들은 서브넷 간 라우팅에 관한 다양한 문제들 가령, 어느 라우터가 특별한 목적지에 대해 선택되어야 하는지(제1 라우팅 단계) 그리고 어느 경로가 목적지에 도달하기 위해 라우터에 의해 선택되어야 하는지(제2 라우팅 단계)를 답해 주기(answer) 위해 사용될 수 있다.

[0038] 예를 들어, 이 방법들은 단순한 클래식 라우팅 방법, 다양한 토폴로지들에 양호한 성능을 제공하는 소스 라우팅 방법 및 팻트리 기반 토폴로지들에 특화되고 최적의 성능을 제공하는 최적화식 라우팅 방법을 포함할 수 있다. 소스 라우팅 방법 및 최적화식 라우팅 방법 모두는 잠재적으로는, 클래식 라우팅 방법보다 양호한 성능을 전할 수 있다. 더욱이, 최적화식 라우팅 방법은 기저(underlying) 팻트리 토폴로지에 대한 최적의 성능을 획득할 수 있게 한다. 추가적으로, 두 방법들 모두는 임의의(arbitrary) 다중-포트 라우터들의 사용을 지원하지만, 클래식 라우팅은 심지어 이용가능한 포트들의 수에 제약되지 않을 때에도 모든 포트들을 효과적으로 이용하지 못할 수 있다.

[0039] 따라서, 이 방법들을 이용하여, 네이티브 IB 라우터들은 성능의 상당한 감소 없이 서브넷들을 함께 연결함으로써 보다 복잡한 IB 패브릭들을 구축할 수 있게 한다.

#### [0040] 서브넷 간 소스 라우팅(ISSR)

[0041] 본 발명의 실시예에 따르면, 서브넷 간 소스 라우팅(ISSR) 라우팅 알고리즘과 같은 범용 라우팅 알고리즘은 하이브리드 서브넷들을 갖는 복잡한 네트워크를 라우팅하기 위해 사용될 수 있다. ISSR 알고리즘은 두 단계들을 포함할 수 있는 바, 제1 단계는 특별한 목적지에 대한 진입 라우터 포트를 선택하기 위한 것이고, 제2 단계는 출구 라우터 포트를 선택하기 위한 것이다.

[0042] 도 3은 본 발명의 실시예에 따른 네트워크 환경에서 서로 다른 목적지들에 대한 진입 라우터 포트들을 선택하는

예를 도시한다. 도 3에 도시된 바와 같이, 네트워크 환경(300)은 하나 이상의 라우터들, 예컨대 라우터 A(301) 및 라우터 B(302)와 연결되는 서브넷(310)을 포함할 수 있다.

[0043] 서브넷 관리자(SM)(예컨대, 서브넷(310) 내의 SM(303))는 서로 다른 목적지들을 위한 진입 라우터 포트들을 선택하기 위해 사용될 수 있는 매핑 파일(304)을 포함할 수 있다. 더욱이, 진입 라우터 포트의 선택은 로컬 서브넷에 연결되는 모든 이용가능한 라우터들 간의 라운드 로빈 경로 분배에 근거할 수 있다. 예를 들어, ISSR 알고리즘에 대한 로컬 라우터 포트를 선택하는 find\_router() 함수는 OpenSM 라우팅 알고리즘과 유사한 방식으로 구현될 수 있다.

[0044] 다음은 매핑 파일(304)의 상위-레벨 예이다.

[0045] 표 1: ISSR 및 ISFR 알고리즘들에 대한 매핑 파일의 상위-레벨 예

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

[0046]

[0047] 오직 단일 라우터 포트에 전체 서브넷을 매치시킬 수 있는 OpenSM 라우팅과 달리, 시스템은 서로 다른 라우터 포트들에 목적지 엔드 포트들을 매핑시킬 수 있다. 더욱이, 매핑 파일의 설정은 OpenSM에 제공되는 매핑 파일과 다를 수 있다. 상기 표 1에 도시된 바와 같이, 매핑 파일(304)은, 오직 OpenSM 서브넷 간 라우팅에 대한 서브넷 프리픽스(prefix) 대신, 충분히 적격화된 포트 GUID를 포함하는 바, 이는 진입 라우터 포트들을 선택함에 있어서 시스템으로 하여금 충분한 입도(granularity)를 제공할 수 있게 한다.

[0048] 더욱이, 도 3에 도시된 바와 같이, 동일한(또는 유사한) 수의 목적지들이 예컨대 라운드 로빈 방식으로 다수의 포트들에 매핑될 수 있다. 예를 들어, 목적지 A(311)(dst\_gid1) 및 목적지 C(313)(dst\_gid3)은 라우터 A(301) 상의 포트 1 및 포트 2를 통해 라우팅될 수 있고, 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. 추가적으로, 매핑 파일(304)은 백업 및 디폴트 라우트들을 명시할 수 있다.

[0049] 본 발명의 실시예에 따르면, 출구 라우터 포트를 선택하는 것은 라우터 펌웨어 내의 모듈로 기반 해시를 이용하여 구현될 수 있다. 상기 해시는 소스 및 목적지 LID들(또는 라우터 상의 진입 포트 수와 같은 어떤 다른 유용한 파라미터)을 이용하여 생성되는 랜덤한 수를 취할 수 있다. 상기 해시를 이용하여, 출력 라우터 포트들 중 하나가 선택될 수 있다. 더욱이, ISSR 라우팅 알고리즘은 동일한 쌍의 노드들에 대해 동일한 경로를 항상 이용하는 결정론적 오블리비우스 라우팅(deterministic oblivious routing) 알고리즘이다.

[0050] 더욱이, 라우터가 둘보다 많은 서브넷들에 배속(attach)될 수 있기 때문에 2-단계 포트 검증 방법이 출구 라우터 포트를 선택하기 위해 사용될 수 있다. 먼저, 시스템은 목적지가 위치한 서브넷에 배속되는 (또는 서브넷의 방향으로의) 가능한 포트들의 세트를 선택할 수 있다. 그 다음, 시스템은 출구 포트를 선택하기 위해 모듈로 함수에 근거한 단순한 해시를 이용할 수 있다.

[0051] 도 4는 본 발명의 실시예에 따른 네트워크 환경에서, 라우터에서 수신된 패킷에 대한 출구 라우터 포트를 선택하는 예를 도시한다. 도 4에 도시된 바와 같이, 네트워크 환경(400)은 서로 다른 서브넷들 내의 다양한 목적지들에 하나 이상의 패킷들을 라우팅할 수 있는 라우터(예컨대, 라우터 A(401))를 포함할 수 있다.

[0052] 예를 들어, 라우터 A(401)는 포트 1에서, 서브넷(403) 내의 목적지 X(421) 행의(destined) 소스 A(411)로부터 패킷을 수신할 수 있다. 추가적으로, 라우터 A(401)는 포트 2에서, 서브넷(402) 내의 목적지 Y(422) 행의 소스 C(413)로부터 다른 패킷을 수신할 수 있다.

[0053] 다음의 알고리즘 1은 출구 라우터 포트를 선택하기 위해 SM 및 라우터 펌웨어 둘 모두 내에서 구현될 수 있다.

[0054] 알고리즘 1: ISSR 내의 choose\_egress\_port() 함수

---

```

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())%port_set.size]
6: end if

```

[0055]

[0056] 라우팅 선택은 소스 LID 및 목적지 LID 둘 모두에 근거할 수 있다. 소스 LID는 전송 서브넷 시나리오에서 본래 소스 또는 이전의 홉 라우터의 출구 포트일 수 있고, 목적지 LID는 최종 목적지 LID 또는 다음 홉 진입 라우터 포트의 LID일 수 있다.

[0057]

라우터 A(401)는 배속된 서브넷들을 볼 수 있기 때문에 소스 LID 및 목적지 LID 둘 모두에 대한 값들을 알 수 있다. 목적지 LID를 획득하기 위해 알고리즘 1 내의 (2 제줄) get\_next\_LID()과 같은 매핑 함수는 목적지 GID를 목적지 LID에 매핑하기 위해 사용될 수 있거나 또는 GID에 위치한 서브넷 프리픽스에 근거한 다음 홉 LID로 리턴하는 것이 요구된다. 매핑 함수는 다음 중 하나 이상을 수행할 수 있다: 콘텐츠-어드레싱가능 메모리 록업, 글로벌 라우팅 헤더(GRH) 필드로부터 최종 목적지 로컬 식별자(LID)를 디코딩하는 것, 글로벌 라우팅 헤더(GRH) 필드로부터 다음 홉 로컬 식별자를 디코딩하는 것.

[0058]

본 발명의 실시예에 따르면, 출구 라우터 포트를 선택하는 알고리즘은 소스 및 목적지 LID들에 근거하여 랜덤한 수를 계산할 수 있다. 이 랜덤한 수는 가능한 포트들의 세트로부터 단일의 출구 포트를 선택하기 위해 사용될 수 있다. 랜덤한 수의 생성은 결정론적 방식으로 행해질 수 있어서, 주어진 소스-목적지 쌍은 항상 동일한 수를 생성할 수 있다. 따라서, 시스템은 서브넷들 간에 라우팅을 할 때 순서를 벗어난 배송(out of order delivery)을 방지할 수 있으며, 출구 포트를 선택하기 위한 라운드 로빈 방법과는 달리, 각각의 소스-목적지 쌍이 네트워크를 통한 동일한 경로를 항상 사용하도록 한다.

[0059]

도 5는 본 발명의 실시예에 따른 네트워크 환경에서, 라우터에서 서브넷 간 소스 라우팅(ISSR) 알고리즘을 지원하기 위한 예시적인 순서도를 도시한다. 도 5에 도시된 바와 같이, 단계(501)에서, 라우터는 목적지들의 리스트를 수신할 수 있는 바, 상기 라우터는 상기 목적지들에 하나 이상의 패킷들을 라우팅하는 역할을 하며, 상기 라우터는 상기 네트워크 환경에서 적어도 하나의 서브넷에 연결된다. 그 다음, 단계(502)에서, 라우터는 하나 이상의 패킷들과 관련된 소스 로컬 식별자(LID) 및 목적지 LID에 근거하여 랜덤한 수를 생성할 수 있다. 더욱이, 단계(503)에서, 라우터는 상기 라우터의 출력 라우터 포트들로부터 하나의 포트를 선택하기 위해 모듈로 기반 해시를 사용할 수 있다.

[0060]

서브넷 간 팻트리 라우팅(ISFR)

[0061]

본 발명의 실시예에 따르면, 서브넷 간 팻트리 라우팅(ISFR) 라우팅 알고리즘과 같은 최적화식 라우팅 방법은 팻트리 토폴로지들을 갖는 서브넷들 간의 라우팅을 위해 이용될 수 있다.

[0062]

최적화식 라우팅 방법은 세 개의 단계들을 포함할 수 있는 바, 이는 가령, 로컬 서브넷에 연결된 모든 이용가능한 라우터 간에 라운드 로빈 경로 분배의 제1 단계, 어느 스위치가 특별한 목적지를 향하는 주요 경로로서 역할을 하는지를 알기 위해 각각의 라우터에 연결된 하향 스위치들에 질의하는 제2 단계, 및 상기 제1 단계 및 제2 단계로부터 획득되는 정보를 이용하여 라우팅 테이블을 구축하는 제3 단계이다.

[0063]

도 6은 본 발명의 실시예에 따른 네트워크 환경(600)에서, 라우터에서 서브넷 간 팻트리 라우팅(ISFR) 알고리즘을 지원하는 예를 도시한다. 도 6에 도시된 바와 같이, 라우터(601)는 하나 이상의 엔드 노드들(621 내지 624)에 연결되는 서브넷(610)에 패킷(650)을 라우팅하는 역할을 한다. 서브넷(610)은 팻트리 토폴로지에 SW들(611 내지 616)과 같은 하나 이상의 스위치들을 포함할 수 있다.

[0064]

라우터(601)는 자신의 어느 포트들이 패킷들을 특별한 목적지로, 따라서 관련 GID로 라우팅하기 위해 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 목적지의 리스트를 포함할 수 있고, 상기 라우터(601)는 서브넷(610)으로부터 비롯되는 패킷들을 상기 목적지들에 라우팅하는 역할을 한다.

[0065]

더욱이, 라우터(601)는 패킷(650)을 라우팅할 때 최대 성능을 달성하기 위해 어느 하향 출력 포트들을 사용할지를 스위치들(611 내지 612)로부터 알 수 있다. 예를 들어, 라우터(601)는 서브넷(610) 내의 스위치들(611 내지



612)에게 상기 서브넷(610) 내의 목적지 노드로의 주요 경로를 질의할 수 있다. 질의는 라우터가 패킷(650)에 대한 주요 경로(스위치(611) 또는 (스위치(612)))를 선택하게 하는 바, 그 이유는 이 스위치들 중 단 하나만이 요구되는 목적지로 향하는 전용 주요 경로를 가질 수 있기 때문이다.

[0066] 그 다음, 라우터(601)는 상기에 획득된 정보를 이용하여 라우팅 테이블(640)을 구축할 수 있다. 예를 들어, 로컬 OpenSM 라우팅은 스위치(611) 또는 스위치(612)의 라우팅 표에 목적지 노드로의 주요 경로를 마킹할 수 있다.

[0067] 다음 알고리즘 2는 라우터 상에서 출구 포트를 선택하기 위한 라우터 펌웨어에서 구현될 수 있다.

[0068] 알고리즘 2: ISSR 내의 query\_down\_for\_egree\_port() 함수

---

```

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

```

[0069]

[0070] 상기 ISFR 알고리즘은 표 1 내의 GID-투-라우터(GID-to-router) 포트 매핑들을 포함하는 이전에 정의된 파일 포맷을 사용할 수 있다. ISSR 알고리즘과 같이, ISFR 알고리즘은 라우터 디바이스에서 구현될 수 있다. 더욱이, ISFR 알고리즘은 오직 팻트리들 상에서, 그리고 모든 서브넷에서 로컬하게 동작하는 팻트리 라우팅과 함께 작동할 수 있다. 또한, ISFR 알고리즘은 필요하면 ISSR로 폴백(fall back)할 수 있다.

[0071] 도 6에 도시된 바와 같이, 라우터(601)는 적절한 팻트리 토폴로지의 상단(top) 상에 위치하는 것으로서 표시될 수 있다. 따라서, 질의가 수행된 후, 라우터(601)는 특별한 서브넷에 위치된 각각의 목적지에 대해 하향으로 포트 당 하나의 경로를 가질 수 있다. 과도가입된(oversubscribed) 팻트리들의 경우, 경로들의 수는 과도가입 비(oversubscription ratio)와 동일할 수 있다.

[0072] 추가적으로, 서브넷(610)에 연결된 하나보다 많은 라우터들이 존재할 수 있다. ISFR 라우팅의 속성(property)은 모든 스핀 (탭) 라우터들이 스위치들과 교체된 경우 (라우터들을 이용한) ISFR 라우팅 및 (스위치들을 이용한) 로컬 라우팅을 위한 라우팅 테이블들이 동일할 수 있도록 한다.

[0073] 본 발명의 실시예에 따르면, ISFR 라우팅 알고리즘은 비-투명한 라우터들을 통해 연결되는 서로 다른 서브넷들 내의 서브넷 관리자들(SM) 간에 확립되는 통신에 근거하여 구현될 수 있다. 예를 들어, 인터페이스가 라우터들에 제공될 수 있고, 상기 인터페이스들을 통해 SM들이 통신할 수 있으며, 핸드셰이킹(handshaking)이 이웃한 서브넷들에 위치된 두 개의 SM들 사이에서 구현될 수 있다.

[0074] 도 7 내지 9는 본 발명의 실시예에 따른 서로 다른 토폴로지들을 갖는 네트워크 환경에서의 라우팅을 도시한다. 토폴로지들 각각은 다중-단계 팻트리 예컨대, 상단 상에 위치된 라우터들을 갖는 3-단계 팻트리로서 표시될 수 있다. 예를 들어, 3-단계 팻트리는 세 개의 라우팅/스위칭 단계들 및 하나의 노드 단계를 가질 수 있고, 각각의 서브넷(2-단계 팻트리)은 브랜치(branch)로서 나타난다. 더욱이, 큰 토폴로지들은 또한, ISFR 라우팅 알고리즘에 근거하여 지원될 수 있다.

[0075] 본 발명의 실시예에 따르면, 네트워크 환경은 각각의 서브넷이 중간에 어떤 전송 서브넷들 없이 다른 서브넷들에 직접적으로 배속될 수 있는 팻트리 토폴로지가 되도록 구성될 수 있다.

[0076] 예를 들어, 도 7에서, 시스템(700)은 두 개의 팻트리 서브넷들(예컨대, 서브넷들 A 내지 B(701 내지 702))을 연결하기 위해 6 개의 라우터들(710)을 사용할 수 있다. 도 8에서, 시스템(800)은 세 개의 팻트리 서브넷들(예컨대, 서브넷들 A 내지 C(801 내지 803))을 연결하기 위해 6개의 라우터들(810)을 사용할 수 있다. 도 9에서, 시스템(900)은 6개의 팻트리 서브넷들(901 내지 906)을 연결하기 위해 18개의 라우터들(910)을 이용하여 648-포트 3-단계 팻트리를 생성할 수 있다.

[0077] 도 10은 본 발명의 실시예에 따른 네트워크 환경에서, 라우터에서 서브넷 간 팻트리 라우팅(ISFR) 알고리즘을 지원하기 위한 예시적인 순서도를 도시한다. 도 10에 도시된 바와 같이, 단계(1001)에서, 라우터는 목적지들의

리스트를 수신할 수 있는 바, 상기 라우터는 하나 이상의 패킷들을 상기 목적지들에 라우팅하는 역할을 하며, 상기 라우터는 네트워크 환경에 적어도 하나의 서브넷에 연결된다. 그 다음, 단계(1002)에서, 라우터는 적어도 하나의 서브넷에서 하나 이상의 스위치들로부터 정보를 획득할 수 있는 바, 상기 적어도 하나의 서브넷 상에서 라우터의 하향 출력 포트들은 하나 이상의 패킷들을 라우팅하기 위해 사용될 수 있다. 더욱이, 단계(1003)에서, 라우터는 획득된 정보에 근거하여 라우팅 테이블을 구축할 수 있다.

[0078] 도 11은 본 구성을 도시하기 위한 기능 블록도를 예시한다. 본 구성은 네트워크 환경에서 별개의 서브넷들 간에 트래픽을 라우팅하기 위한 시스템(1100)으로서 구현될 수 있다. 시스템(1100)은 하나 이상의 마이크로프로세서들 및 상기 하나 이상의 마이크로프로세서들에서 동작하는 라우터(1110)를 포함한다. 상기 하나 이상의 마이크로프로세서들은 목적지들 - 라우터(1110)는 하나 이상의 패킷들을 상기 목적지들에 라우팅하는 역할을 하고, 상기 라우터(1110)는 네트워크 환경에서 적어도 하나의 서브넷에 연결되며 - 의 리스트를 수신하기 위한 수신 유닛(1120)과, 상기 하나 이상의 패킷들과 관련된 소스 로컬 식별자(LID) 및 목적지 LID에 근거하여 랜덤한 수를 생성하기 위한 생성 유닛(1130)과, 그리고 라우터의 출력 라우터 포트들로부터 일 포트를 선택하기 위해 모듈로 기반 해시를 사용하기 위한 사용 유닛(1140)을 포함한다.

[0079] 일 실시예에 따르면, 하나 이상의 마이크로프로세서들에서 동작하는 네트워크 환경에서 별개의 서브넷들 간의 트래픽을 라우팅하기 위한 시스템이 개시된다. 상기 시스템은 라우터에서 목적지들의 리스트를 수신하기 위한 수단을 포함하며, 상기 라우터는 하나 이상의 패킷들을 상기 목적지들에 라우팅하는 역할을 하며, 상기 라우터는 네트워크 환경에서 적어도 하나의 서브넷에 연결된다. 상기 시스템은 하나 이상의 패킷들과 관련된 소스 로컬 식별자(LID) 및 목적지 LID에 근거하여 랜덤한 수를 생성하기 위한 수단을 포함한다. 시스템은 라우터의 복수의 출력 라우터 포트들로부터 하나의 라우터 포트를 선택하기 위해 모듈로 기반 해시를 사용하기 위한 수단을 포함한다.

[0080] 바람직하게는, 시스템은 적어도 하나의 서브넷에 연결되는 모든 이용가능한 라우터 간에 라운드-로빈 경로 분배를 수행하기 위한 수단을 포함한다.

[0081] 바람직하게는, 시스템은 라우터 펌웨어에서 모듈로 기반 해시를 구현하기 위한 수단을 포함하며, 상기 모듈로 기반 해시는 소스 LID 및 목적지 LID를 이용하여 생성되는 랜덤한 수를 취할 수 있다.

[0082] 바람직하게는, 시스템은 서브넷 관리자로부터 하나 이상의 목적지 글로벌 식별자들(GID)과 하나 이상의 라우터 포트들과의 사이의 포트 매핑을 수신하기 위한 수단을 포함한다.

[0083] 바람직하게는, 시스템은 라우터에 의해 연결되는 서브넷들이 서로 다른 네트워크 토폴로지들에서 구성되도록 하기 위한 수단을 포함한다.

[0084] 바람직하게는, 시스템은 소스 LID가 이전의 홉 라우터의 출구 포트와 관련되도록 하기 위한 수단을 포함한다.

[0085] 바람직하게는, 시스템은 주어진 소스 및 목적지 쌍에 대해 동일한 랜덤 수를 생성하기 위한 수단을 포함한다.

[0086] 바람직하게는 시스템은 하나 이상의 패킷들과 관련된 목적지 글로벌 식별자(GID)에 근거하여 목적지 로컬 식별자(LID)를 획득하기 위한 수단을 포함한다.

[0087] 바람직하게는 시스템은 글로벌 식별자(GID)에 위치된 서브넷 프리픽스에 근거하여 다음 홉 LID를 매핑 함수를 통해 리턴하기 위한 수단을 포함한다.

[0088] 바람직하게는, 시스템은 적어도 하나의 서브넷이 팻트리 토폴로지에 구성되도록 하기 위한 수단을 포함한다.

[0089] 일 실시예에 따르면, 네트워크 환경에서 별개의 서브넷들 간의 트래픽을 라우팅하기 위한 시스템이 개시된다. 상기 시스템은 라우터를 포함하며, 상기 라우터는 목적지들의 리스트를 수신하도록 구성되고, 상기 라우터는 하나 이상의 패킷들을 상기 목적지들에 라우팅하는 역할을 하며, 상기 라우터는 네트워크 환경에서 적어도 하나의 서브넷에 연결된다. 상기 라우터는 하나 이상의 패킷들과 관련된 소스 로컬 식별자(LID) 및 목적지 LID에 근거하여 랜덤한 수를 생성하도록 구성되고, 또한 상기 라우터의 출력 라우터 포트들로부터 하나의 포트를 선택하기 위해 모듈로 기반 해시를 사용하도록 구성된다.

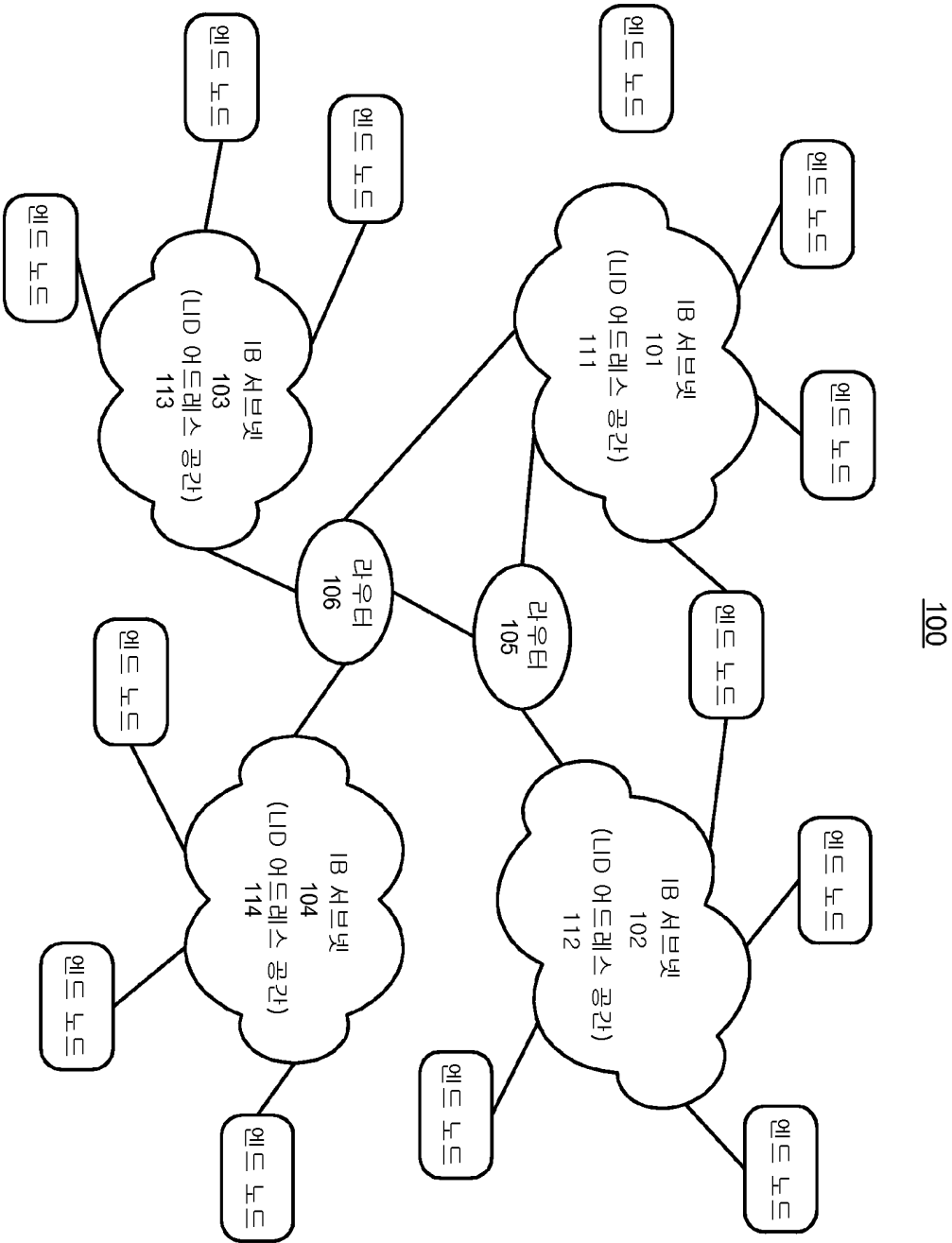
[0090] 바람직하게는, 시스템은 라운드 로빈 경로 분배가 적어도 하나의 서브넷에 연결된 모든 이용가능한 라우터 간에 수행되게 할 수 있다.

[0091] 바람직하게는, 시스템은 모듈로 기반 해시가 라우터 펌웨어에서 구현되도록 할 수 있으며, 상기 모듈로 기반 해시는 소스 LID 및 목적지 LID를 이용하여 생성된 랜덤한 수를 취할 수 있다.

- [0092] 바람직하게는, 시스템은 상기 라우터가 하나 이상의 목적지 글로벌 식별자들(GID)과 하나 이상의 라우터 포트들과의 사이의 포트 매핑을 서브넷 관리자로부터 수신하는 동작을 하게 할 수 있다.
- [0093] 바람직하게는, 시스템은 상기 라우터가, 상기 라우터에 의해 연결되는 서브넷들이 서로 다른 네트워크 토폴로지들로 구성되도록 하는 동작을 하게 할 수 있다.
- [0094] 바람직하게는, 시스템은 소스 LID가 이전의 홉 라우터의 출구 포트와 관련되게 할 수 있다.
- [0095] 바람직하게는, 시스템은 상기 라우터가 주어진 소스 및 목적지 쌍에 대해 동일한 랜덤 수를 생성하는 동작을 하게 할 수 있다.
- [0096] 바람직하게는, 시스템은 상기 라우터가 하나 이상의 패킷들과 관련된 목적지 글로벌 식별자(GID)에 근거하여 목적지 로컬 식별자(LID)를 획득하는 동작을 하게 할 수 있다.
- [0097] 바람직하게는, 시스템은 상기 라우터가 글로벌 식별자(GID)에 위치한 서브넷 프리픽스에 근거하여 다음 홉 LID를 획득하는 동작을 하게 할 수 있다.
- [0098] 본 발명은 통상적으로, 본 발명의 교시들에 따라 프로그램된 하나 이상의 프로세서들, 메모리 및/또는 컴퓨터 판독가능 스토리지 매체를 포함하는 하나 이상의 종래의 범용 또는 특수화된 디지털 컴퓨터, 컴퓨팅 디바이스, 머신 또는 마이크로프로세서를 이용하여 구현될 수 있다. 적절한 소프트웨어 코딩은 소프트웨어 기술 분야의 숙련자들에게 분명할 바와 같이, 본 발명의 교시들에 근거하여 숙련된 프로그래머들에 의해 쉽게 준비될 수 있다.
- [0099] 일부 실시예들에서, 본 발명은 본 발명의 프로세스들 중 어느 것을 수행하도록 컴퓨터를 프로그램하는 데 사용될 수 있는 명령어들이 저장된 스토리지 매체 또는 컴퓨터 판독가능 매체(들)인 컴퓨터 프로그램 물체를 포함한다. 스토리지 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.
- [0100] 본 발명의 상기 상세한 설명은 예시 및 설명을 위해 제공되었다. 본 설명은 완전한 것이거나 또는 정확히 개시된 형태들로만 본 발명을 제한하고자 의도된 것이 아니다. 많은 수정들 및 변형들이 이 기술분야의 숙련자에게 분명할 것이다. 위 실시예들은 본 발명의 원리 및 이의 실용적 응용을 가장 잘 설명하기 위해 선택 및 기술되었으며, 그럼으로써 이 기술분야의 숙련자들은 본 발명에 대한 다양한 실시예들 및 고려되는 특별한 사용에 적합한 다양한 수정들을 이해할 수 있다. 본 발명의 범위는 다음의 특허 청구 범위 및 이의 균등물에 의해 한정되어야 함이 의도된다.

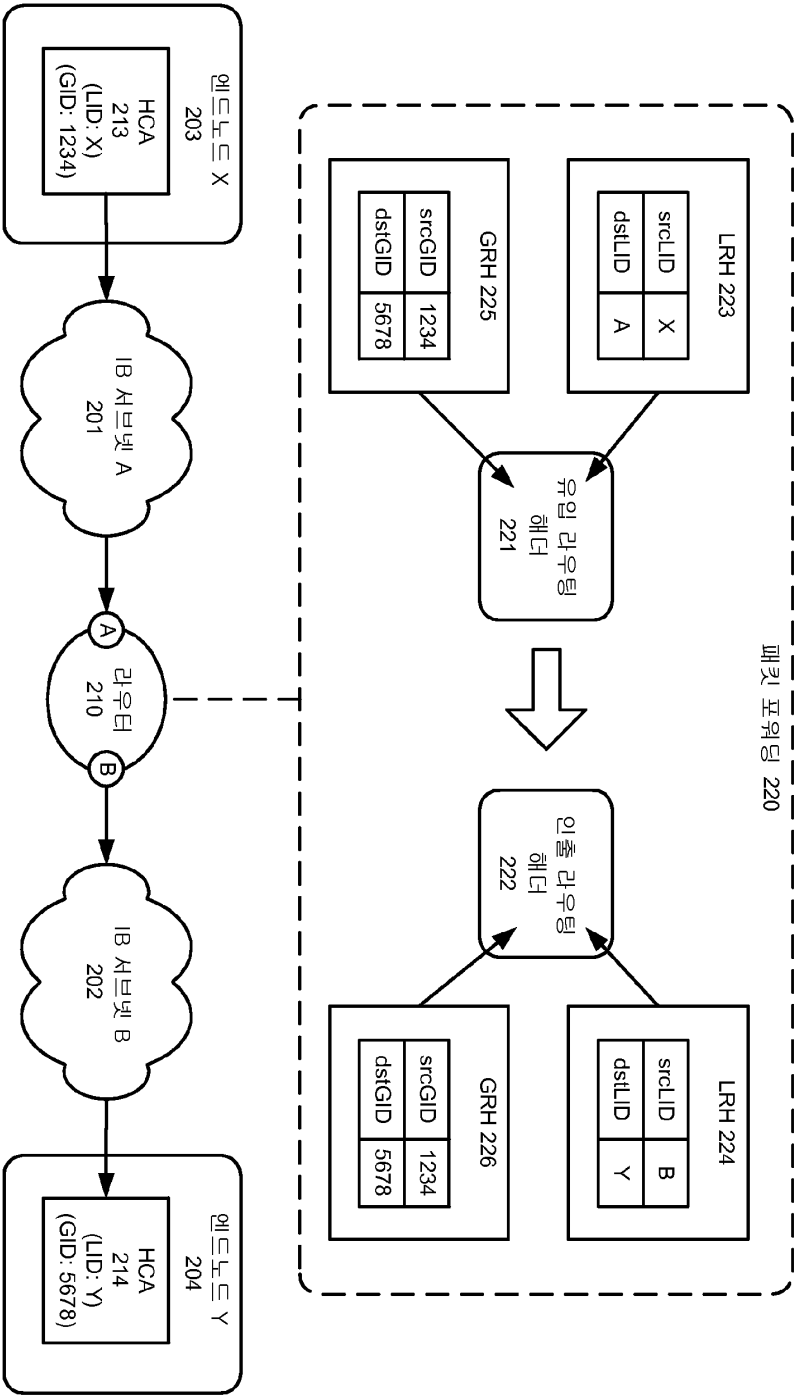
도면

도면1



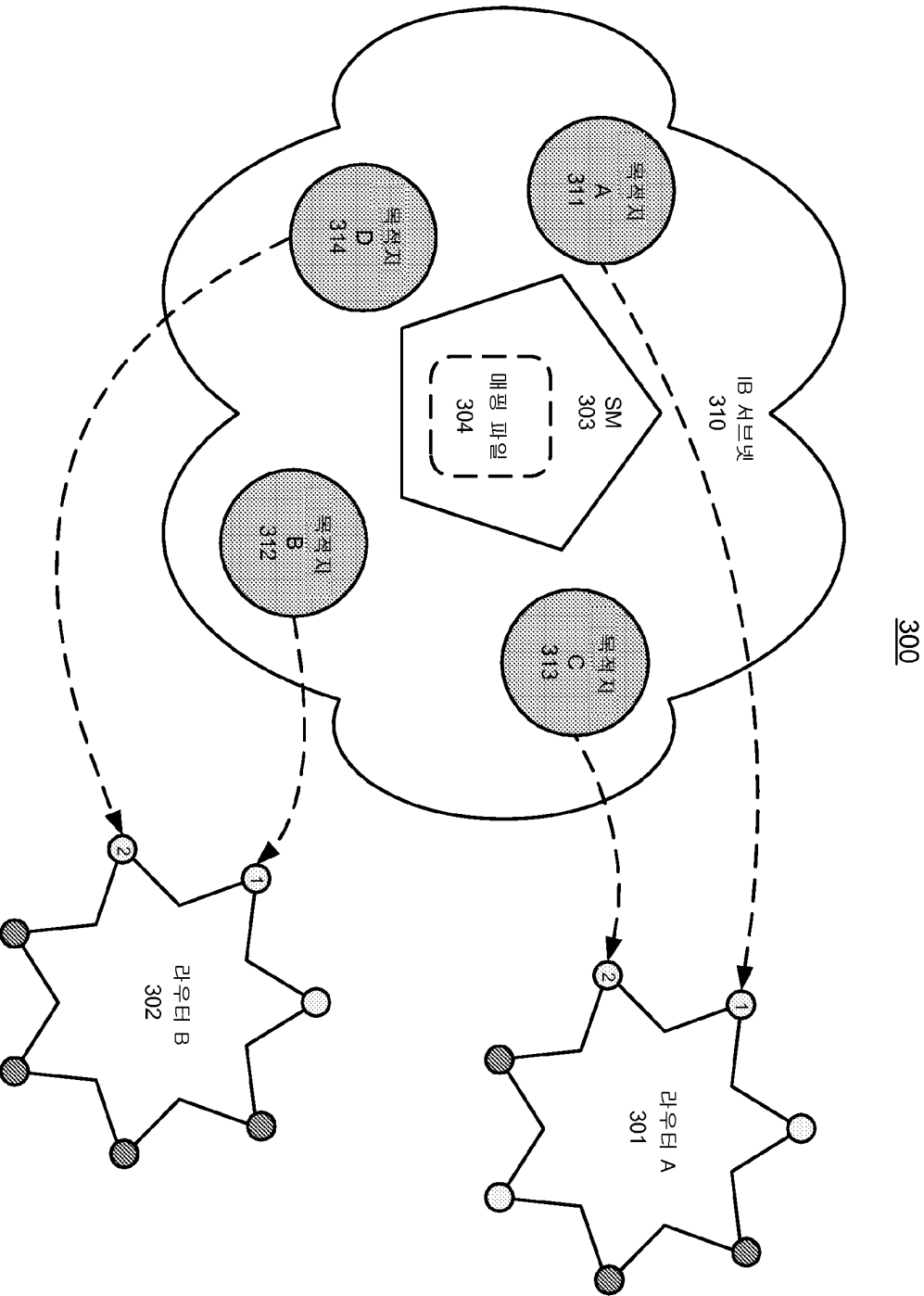


도면2

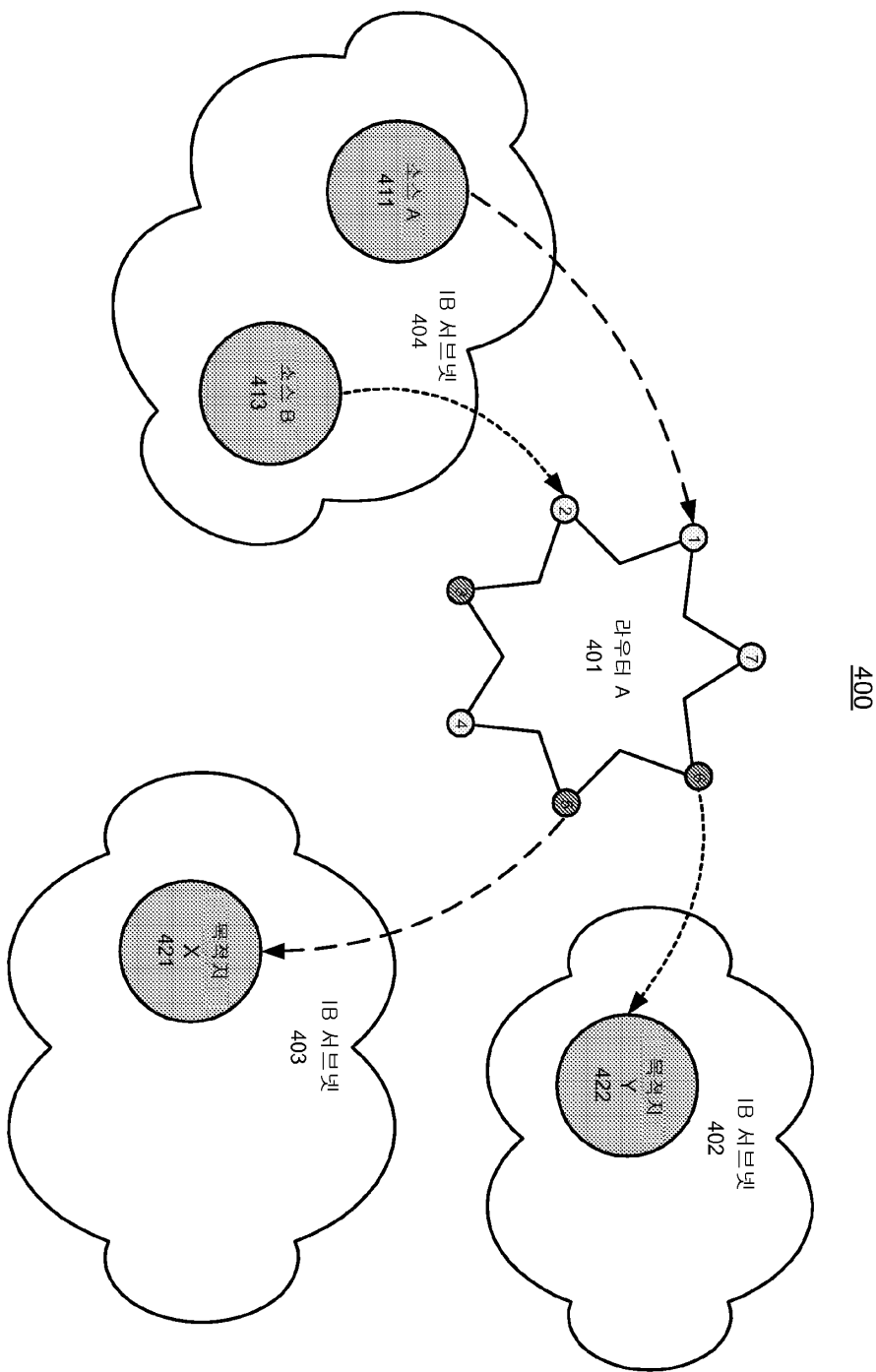


200

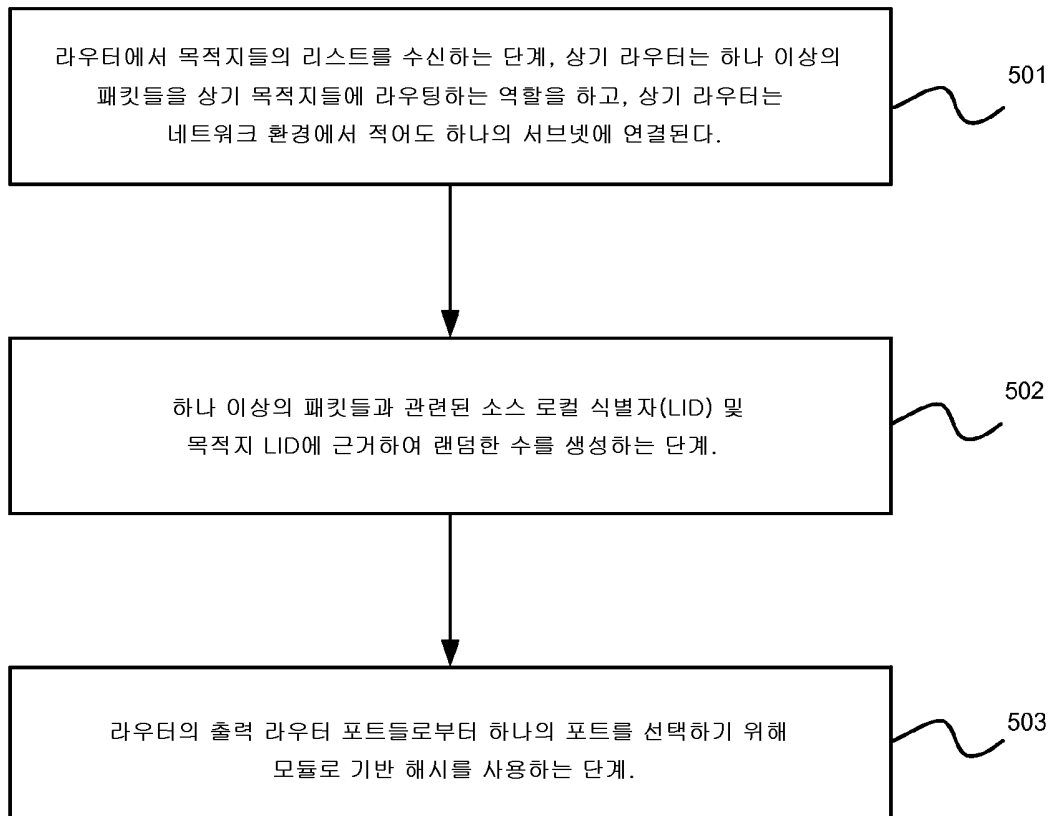
도면3



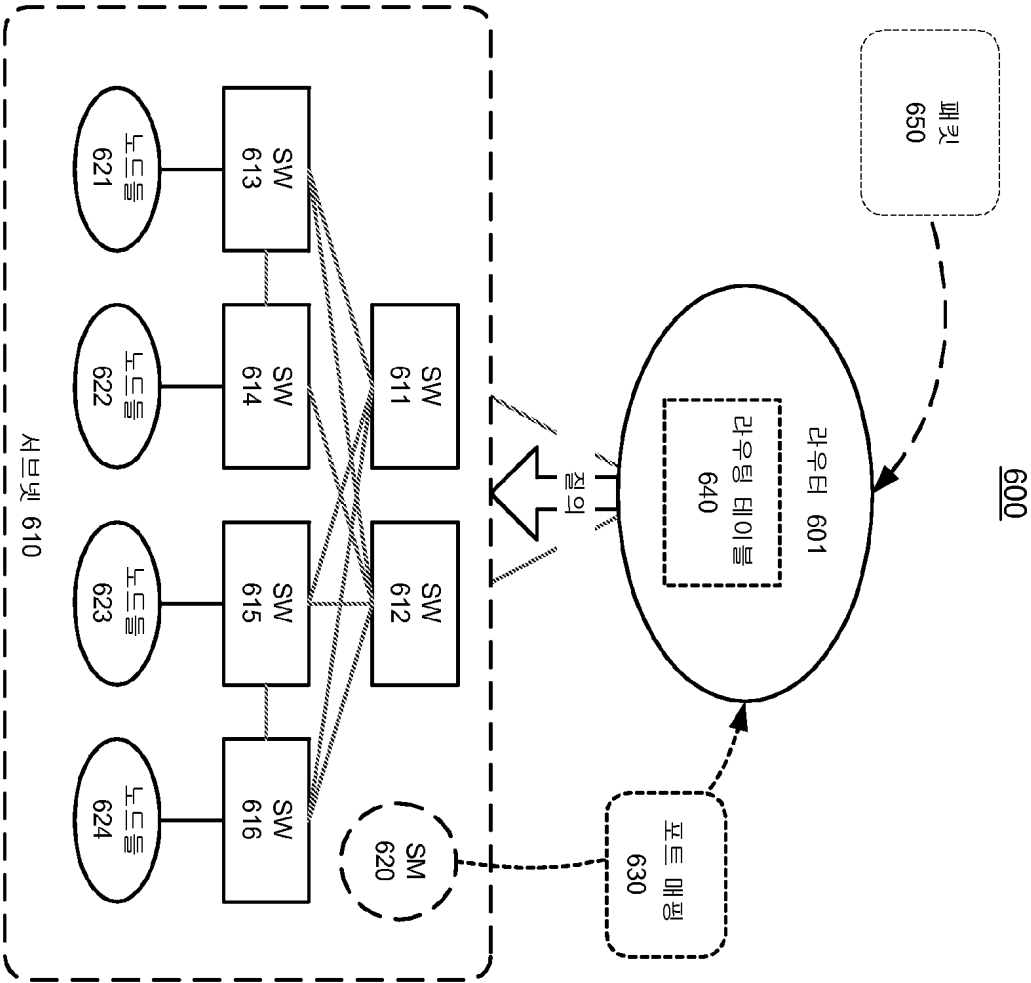
도면4



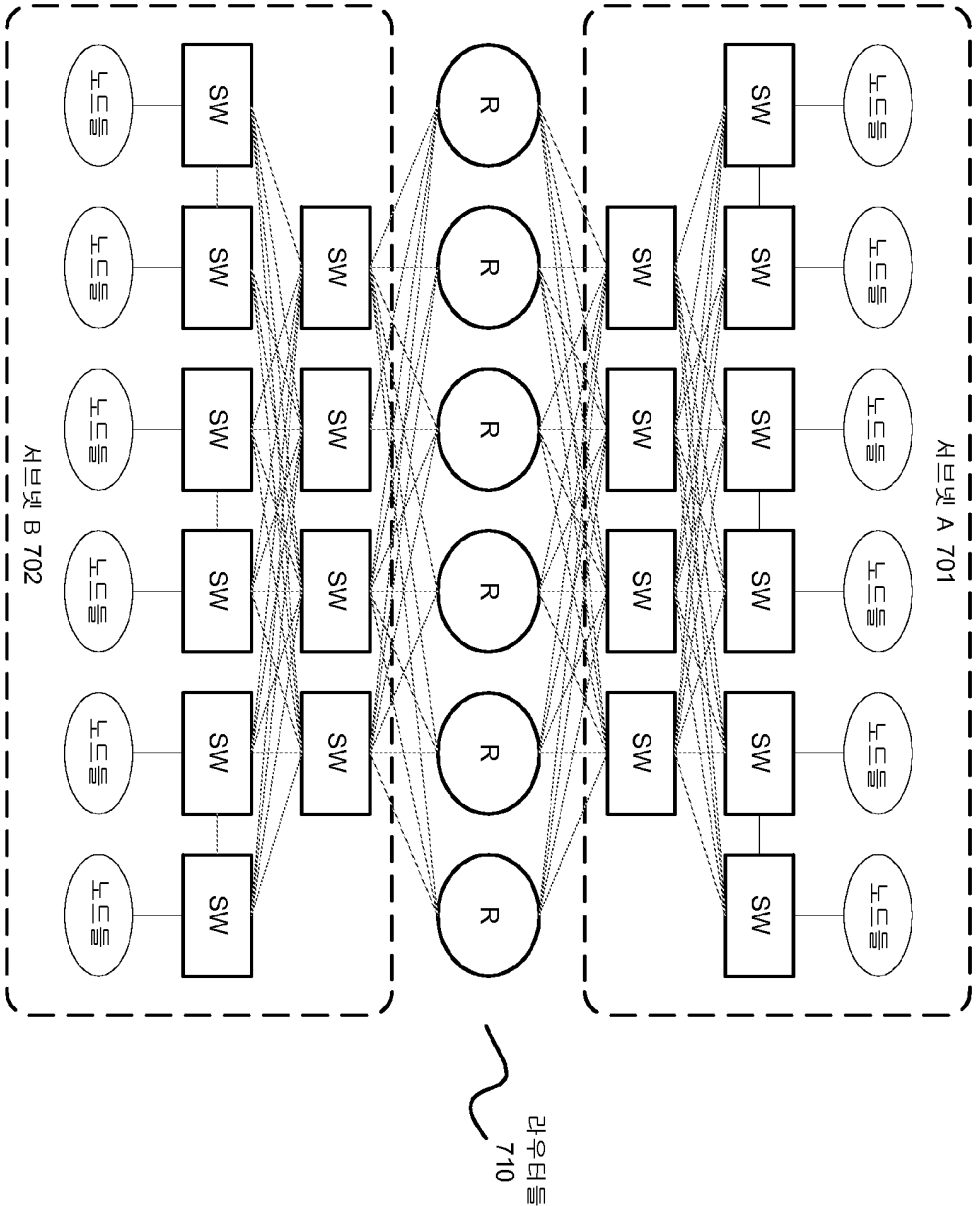
도면5



도면6

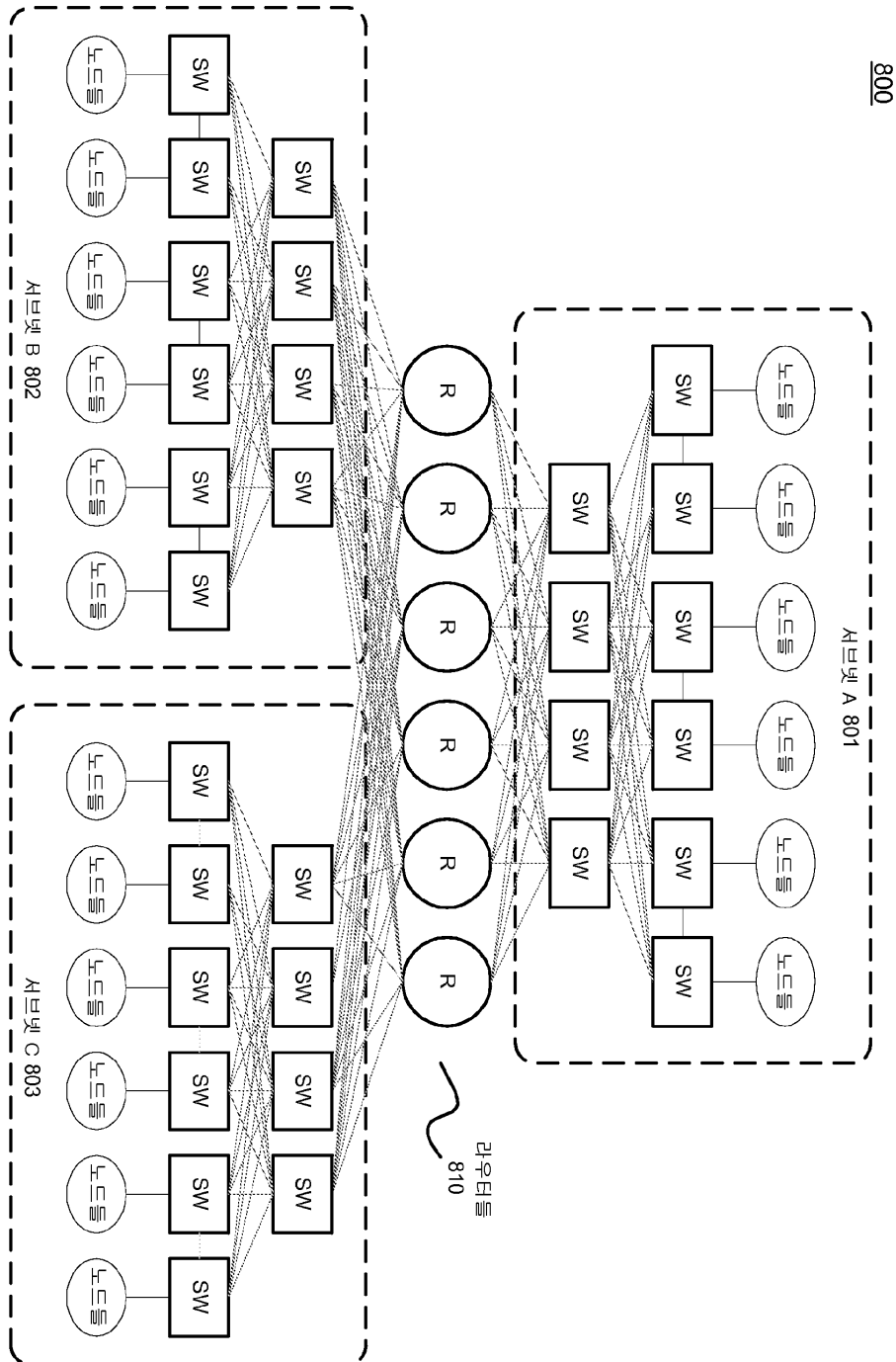


700

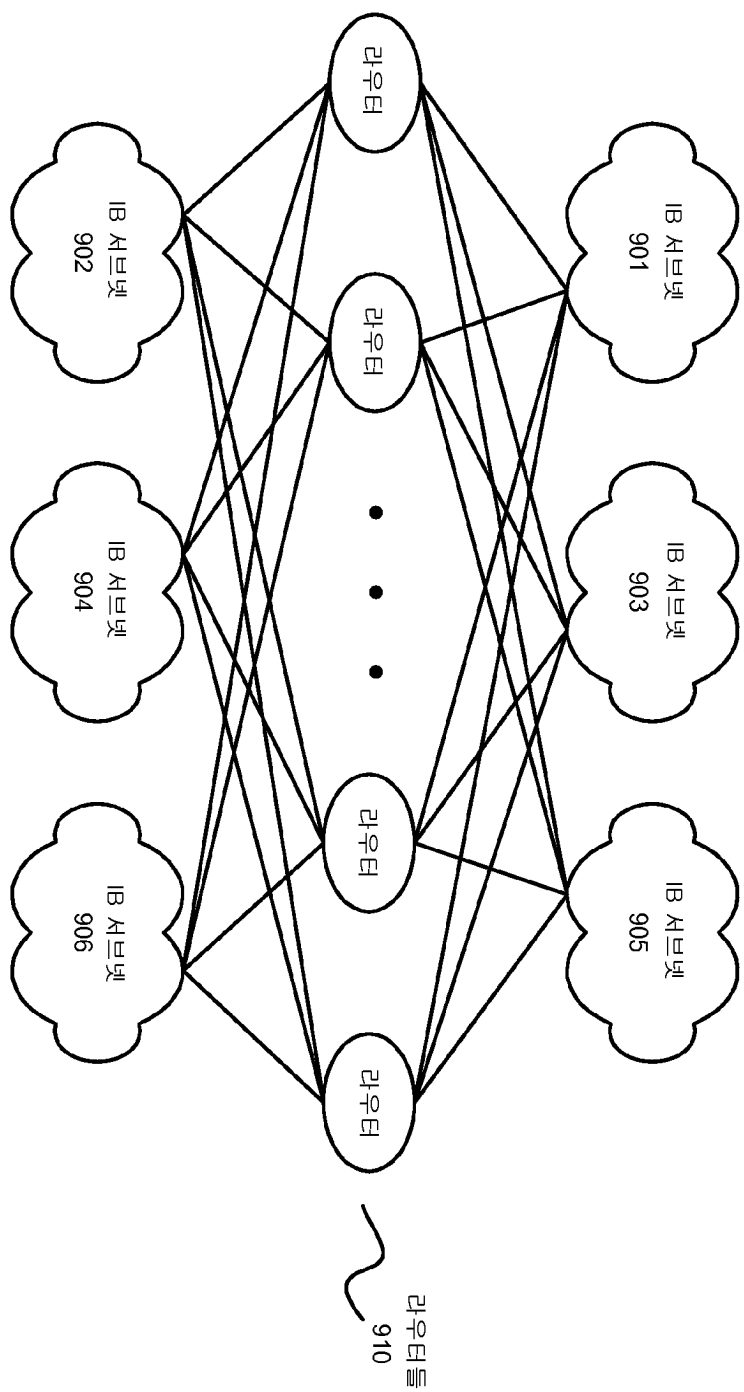


도면7

도면8



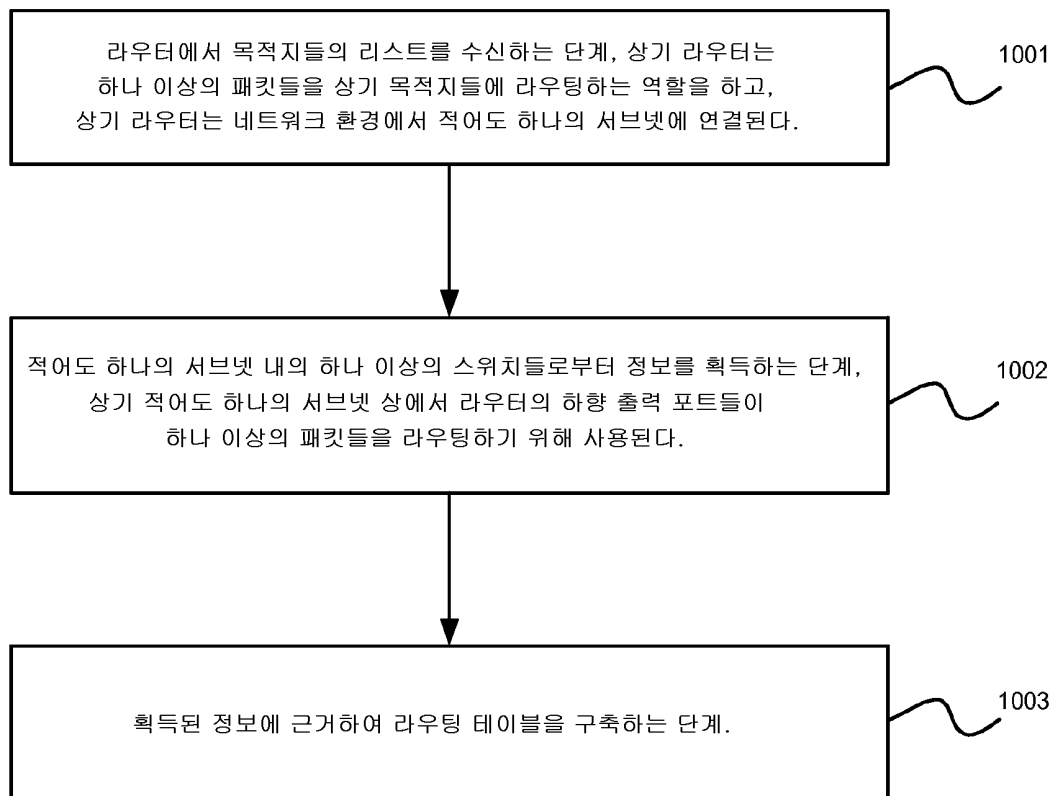
도면9



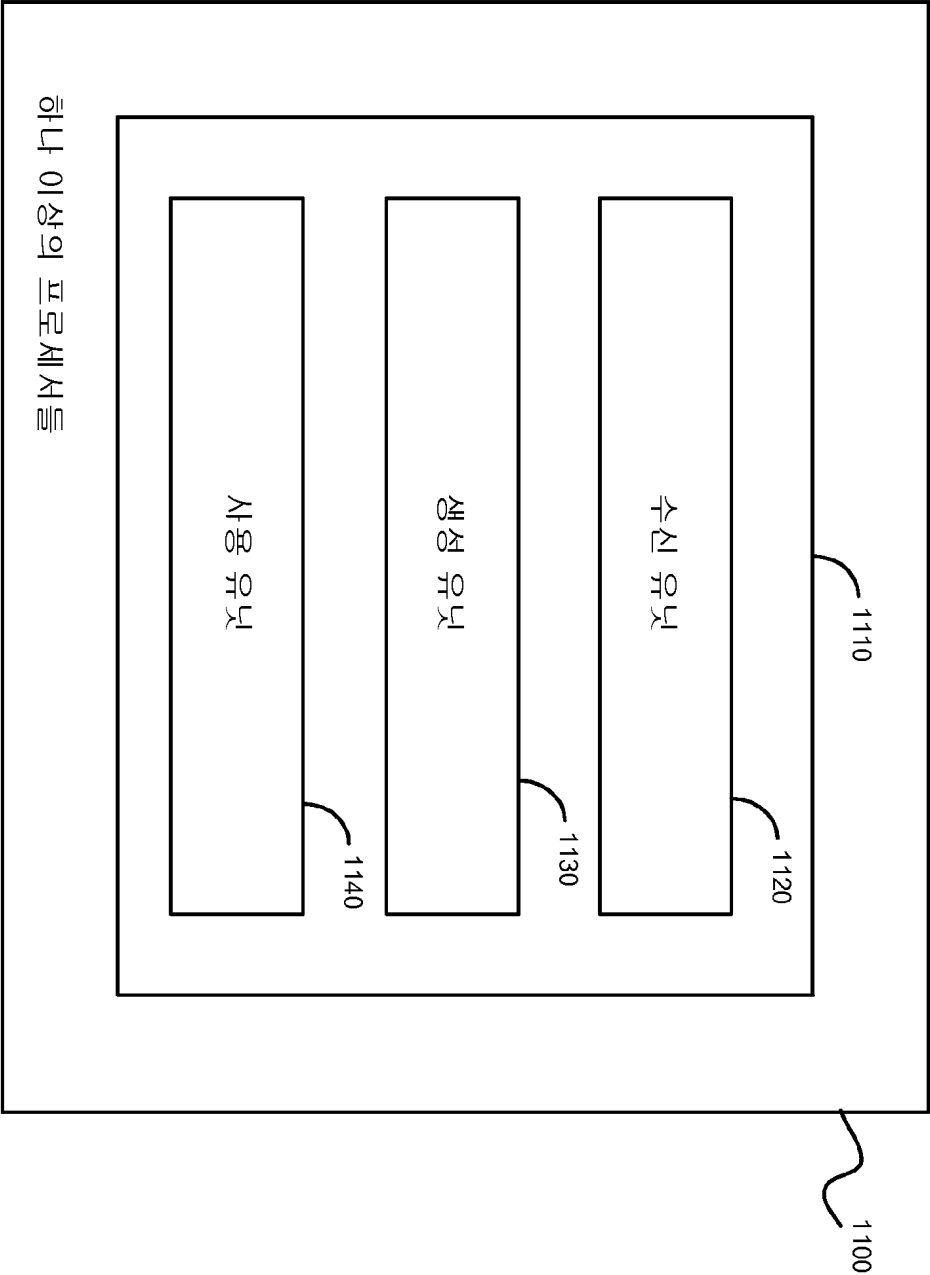
900



도면10



도면11



하나 이상의 프로세서들