(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2014년06월20일
(11) 등록번호 10-1410292
(24) 등록일자 2014년06월16일

(51) 국제특허분류(Int. Cl.)
H04L 12/28 (2006.01) *H04L 12/801* (2013.01)
H04L 12/12 (2006.01) *G06F 1/32* (2006.01)
(21) 출원번호 10-2011-7029353
(22) 출원일자(국제) 2010년04월22일
심사청구일자 2011년12월08일
(85) 번역문제출일자 2011년12월08일
(65) 공개번호 10-2012-0024734
(43) 공개일자 2012년03월14일
(86) 국제출원번호 PCT/EP2010/055336
(87) 국제공개번호 WO 2010/130545
국제공개일자 2010년11월18일
(30) 우선권주장
09160076.7 2009년05월12일
유럽특허청(EPO)(EP)
(56) 선행기술조사문헌
US06748435 B1*
US20040136712 A1*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
알까뗄 루슨트
프랑스 92100 불론뉴-비영꾸르 루뜨 들 라 렌느 148/152
(72) 발명자
슐렝크, 랄프
독일 91054 에를랑겐 마르쿠아르트센스트라쎄 9
헤름스마이어, 크리스티안
독일 90542 에켄탈 클라인게샤이테르스트라쎄 20
(74) 대리인
장훈

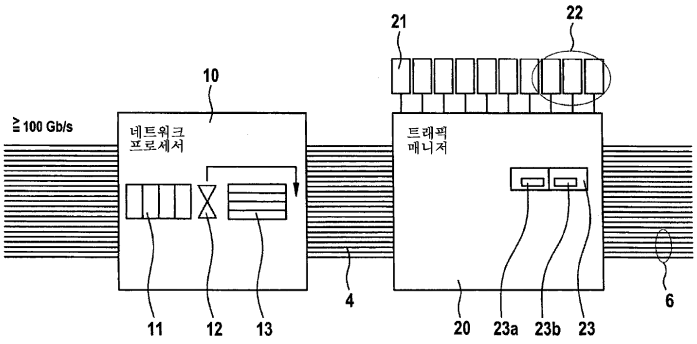
전체 청구항 수 : 총 14 항 심사관 : 이철수

(54) 발명의 명칭 **고속 패킷 교환 시스템들에서의 트래픽-부하 의존 전력 감소**

(57) 요 약

본 발명은 패킷 교환 시스템들에서의 트래픽-부하 의존 전력 감소를 위한 패킷 교환 시스템에 대한 방법 및 디바이스들에 관한 것이다. 패킷 교환 시스템의 전력 소비를 감소시키기 위해, 방법은 업스트림 패킷 프로세싱 디바이스에서 인입하는 데이터 패킷들에 대한 트래픽 레이트를 결정하는 단계; 트래픽 레이트가 결정된 프로세싱된 데이터 패킷들이 다운스트림 패킷 프로세싱 디바이스로 송신되는 기간 이전의 시간에 결정된 트래픽 레이트의 표시를 업스트림 패킷 프로세싱 디바이스로부터 다운스트림 패킷 프로세싱 디바이스로 송신하는 단계; 및 수신된 트래픽 레이트 표시에 기초하여 다운스트림 패킷 프로세싱 디바이스에서 이용가능한 패킷 프로세싱 리소스들을 조정하는 단계를 포함하는 방법이 제안된다.

대 표 도



특허청구의 범위

청구항 1

네트워크 백본(backbone)의 업스트림 패킷 프로세싱 디바이스(upstream packet processing device) 및 병렬 패킷 프로세싱 케이퍼빌리티(capability)들을 포함하는 네트워크 백본의 다운스트림 패킷 프로세싱 디바이스를 가지는 패킷 교환 시스템의 전력 소비를 감소시키는 방법에 있어서:

- 상기 업스트림 패킷 프로세싱 디바이스에서 인입하는 데이터 패킷들에 대한 트래픽 레이트(traffic rate)를 결정하는 단계;

- 상기 인입하는 데이터 패킷들을 프로세싱하는 단계;

- 트래픽 레이트가 결정된 프로세싱된 데이터 패킷들이 상기 다운스트림 패킷 프로세싱 디바이스로 송신되는 시간 이전의 시간에 상기 결정된 트래픽 레이트의 표시를 상기 업스트림 패킷 프로세싱 디바이스로부터 상기 다운스트림 패킷 프로세싱 디바이스로 송신하는 단계;

- 수신된 트래픽 레이트 표시에 기초하여, 상기 다운스트림 패킷 프로세싱 디바이스에서 이용가능한 병렬 패킷 프로세싱 리소스들을 적응식으로 스위치 온(switch on) 및 스위치 오프(switch off)하는 단계; 및

- 상기 프로세싱된 데이터 패킷들을 상기 업스트림 패킷 프로세싱 디바이스로부터 상기 다운스트림 패킷 프로세싱 디바이스로 송신하는 단계를 포함하는, 패킷 교환 시스템의 전력 소비를 감소시키는 방법.

청구항 2

제 1 항에 있어서,

상기 트래픽 레이트 표시를 송신할 때, 상기 업스트림 패킷 프로세싱 디바이스는, 상기 업스트림 패킷 프로세싱 디바이스에서 계속해서 프로세싱되고 있는 인입하는 데이터 패킷들에 대하여 결정되는 트래픽 레이트를 포함하는 헤더 필드(header field)를, 상기 업스트림 패킷 프로세싱 디바이스에 의해 이미 프로세싱된 적어도 하나의 데이터 패킷에 추가하는, 패킷 교환 시스템의 전력 소비를 감소시키는 방법.

청구항 3

제 1 항에 있어서,

상기 트래픽 레이트 표시의 송신과 상기 트래픽 레이트에 대응하는 프로세싱된 데이터 패킷들의 송신 사이의 시간차는 최소한 상기 다운스트림 패킷 프로세싱 디바이스의 패킷 프로세싱 리소스를 활성화 또는 비활성화하는데 필요한 시간인, 패킷 교환 시스템의 전력 소비를 감소시키는 방법.

청구항 4

제 1 항에 있어서,

상기 트래픽 레이트는 상기 인입하는 데이터 패킷들에 대한 정보 레이트 값 및/또는 버스트 레이트(burst rate) 값을 결정함으로서 결정되는, 패킷 교환 시스템의 전력 소비를 감소시키는 방법.

청구항 5

제 4 항에 있어서,

상기 정보 레이트 값 및/또는 버스트 레이트 값은 각각의 인입하는 패킷 플로우(flow)에 대해 별개로 결정되는, 패킷 교환 시스템의 전력 소비를 감소시키는 방법.

청구항 6

제 4 항에 있어서,

상기 정보 레이트 값 및/또는 버스트 레이트 값은 취합 패킷 대역폭(aggregate packet bandwidth)에 대하여 결정되는, 패킷 교환 시스템의 전력 소비를 감소시키는 방법.

청구항 7

제 1 항에 있어서,

병렬 프로세싱 유닛(unit)들을 가지는 패킷 프로세싱 리소스에 걸쳐 상기 데이터 패킷들을 부하-분배(load-distributing)하는 단계를 추가로 포함하고, 이용되는 병렬 패킷 프로세싱 유닛들의 수는 최소화되는, 패킷 교환 시스템의 전력 소비를 감소시키는 방법.

청구항 8

네트워크 백본의 업스트림 패킷 프로세싱 디바이스에 있어서:

- 인입하는 데이터 패킷들에 대한 트래픽 레이트를 결정하도록 구성되는 트래픽 미터링 유닛(traffic metering unit);

- 데이터 패킷 프로세서;

- 트래픽 레이트가 결정된 프로세싱된 데이터 패킷들이 다운스트림 패킷 프로세싱 디바이스로 송신되는 시간 이전의 시간에 상기 결정된 트래픽 레이트의 표시를 상기 업스트림 패킷 프로세싱 디바이스로부터 상기 다운스트림 패킷 프로세싱 디바이스로 송신하기 위한 수단; 및

- 상기 프로세싱된 데이터 패킷들을 상기 업스트림 패킷 프로세싱 디바이스로부터 상기 다운스트림 패킷 프로세싱 디바이스로 송신하기 위한 데이터 패킷 송신기를 포함하는, 업스트림 패킷 프로세싱 디바이스.

청구항 9

제 8 항에 있어서,

상기 업스트림 패킷 프로세싱 디바이스는:

상기 인입하는 데이터 패킷들의 사전-분류 및 초과 신청 관리(oversubscription management)를 버퍼링하기 위한 버퍼링 유닛(buffering unit)(11); 및

패킷 분류, 포워딩, 필터링 및 태깅(tagging)을 실행하기 위한 병렬 프로세싱 케이퍼빌리티(capability)들(13)을 포함하는 네트워크 백본의 네트워크 프로세서(10)인, 업스트림 패킷 프로세싱 디바이스.

청구항 10

네트워크 백본의 다운스트림 패킷 프로세싱 디바이스에 있어서:

- 트래픽 레이트가 결정된 프로세싱된 데이터 패킷들이 상기 다운스트림 패킷 프로세싱 디바이스에서 수신되는 시간 이전의 시간에 업스트림 패킷 프로세싱 디바이스에서 결정되는 트래픽 레이트에 대하여 송신되는 표시를 수신하기 위한 수단; 및

- 수신된 트래픽 레이트 표시에 기초하여 상기 다운스트림 패킷 프로세싱 디바이스에서 이용가능한 병렬 패킷 프로세싱 리소스들을 적응식으로 스위치 온 및 스위치 오프하도록 구성되는 리소스 매니저(23)를 포함하는, 다운스트림 패킷 프로세싱 디바이스.

청구항 11

제 10 항에 있어서,

상기 다운스트림 패킷 프로세싱 디바이스는,

- 트래픽을 구현하도록 구성되는 병렬 프로세싱 케이퍼빌리티들,

- 데이터 패킷들의 관리, 큐잉(queuing), 분할화 및 재조립, 및

- 스위칭 패브릭 및 네트워크 프로세서로의 인터페이스 관리를 포함하는 네트워크 백본의 트래픽 매니저(20)이고,

상기 트래픽 매니저(20)는 다중-링크 송신 인터페이스를 통해 상기 업스트림 패킷 프로세싱 디바이스에 접속되고;

상기 리소스 매니저(23)는 상기 수신된 트래픽 레이트 표시에 기초하여 패킷 프로세싱 리소스들의 상태를 변경하는 상태 기계를 포함하는, 다운스트림 패킷 프로세싱 디바이스.

청구항 12

제 11 항에 있어서,

상기 수신된 트래픽 레이트 표시에 기초하여 관리되는 상기 트래픽 매니저의 병렬 프로세싱 요소들은 병렬 메모리 뱅크들(21) 및/또는 파이프라인 루프(pipeline loop)들 및/또는 네트워크 프로세서로의 인터페이스의 다중 송신 레인들인, 다운스트림 패킷 프로세싱 디바이스.

청구항 13

제 10 항에 있어서,

상기 다운스트림 패킷 프로세싱 디바이스는 복수의 병렬 트래픽 매니저들(20)에 교차 접속되는 다수의 교차 디바이스들(32)을 포함하는 네트워크 백본의 스위치 패브릭(30)이고, 상기 스위치 패브릭(30)의 리소스 매니저(31)는 복수의 네트워크 프로세서들(10)로부터 상기 스위치 패브릭(30)으로 송신되는 트래픽 레이트들을 수신하도록 구성되고 상기 수신된 트래픽 레이트 표시들로부터 결정되는 취합된 트래픽 레이트 정보에 기초하여 상기 스위치 패브릭(30)의 패킷 프로세싱 리소스들을 관리하는, 다운스트림 패킷 프로세싱 디바이스.

청구항 14

제 13 항에 있어서,

상기 스위치 패브릭의 교차 접속 교환 디바이스는 상기 교환 디바이스(32)를 복수의 트래픽 매니저들(20)과 연결시키는 모든 송신 링크들(52)이 상기 트래픽 매니저들(20)에서 상기 수신된 트래픽 레이트 표시자들에 기초하여 상기 트래픽 매니저들(20)에 의해 비활성화되었을 경우에, 비활성화되도록 구성되는, 다운스트림 패킷 프로세싱 디바이스.

청구항 15

삭제

명 세 서

기 술 분 야

[0001] 본 발명은 패킷 교환 시스템들에서의 트래픽-부하 의존 전력 감소를 위한 트래픽 교환 시스템에 대한 방법 및 디바이스들에 관한 것이다.

배 경 기 술

[0002] 인터넷 트래픽은 무어의 법칙에 따라 더욱더 지수적으로 증가할 것이라고 예상되어 왔고 예상되고 있다. 결과적으로, 네트워크 회선 속도는 과거에는 약 매 2년에 배가되어 왔다. 그러나, 집적 회로 및 메모리 클럭 레이트(clock rate)들은 그 정도가 동일하게 개선되지 않았는데, 하나의 이유는 디바이스 로직(device logic) 사이의 칩 배선 지연들이 기하학적 크기들의 비율에 따라 비례하지 않고 오히려 일정하게 유지되는 점이다. 이 고속 설계에서의 상호접속 지연들의 문제를 처리할 공통 해법은 트래픽 매니저들의 병렬 메모리 뱅크(memory bank)들 또는 아주 많은 수의 상대적 저속 칩-대-칩 인터페이스들과 같은 리소스들의 병렬화이다.

[0003] 고속 데이터 패킷 프로세싱에 대한 그와 같은 리소스들의 병렬화는 필수 공간이 줄어들고 전력이 소비되며, 궁극적으로는 비용이 더 높아진다. 더욱이, 전력 소비의 증가는, 오늘날의 컴퓨터들 및 패킷-프로세싱 네트워크 디바이스들의 점차 더 소형화하는 하드웨어 설계와 결합됨으로써, 고 전력 밀도들이 발생하는 결과를 초래한다. 이 고 전력 밀도들은 칩 신뢰도를 해치고 예상 수명을 저하시키고, 냉각 비용들을 증가시키며, 큰 데이터 센터들의 경우, 심지어 환경 문제들을 발생시킨다.

[0004] 최신의 설계들에서, 전력 소비의 두 형태들, 즉 동적 전력 소비 및 정적 전력 소비는, 회로 및 로직 레벨 기술들(예를 들면, 트랜지스터 설계, 저전력 상호 접속들), 캐싱 아키텍처(caching architecture)들(예를 들면, 적응형 캐시들) 및 동적 전압 스케일링(dynamic voltage scaling: DVS)에 의해 감소될 수 있다.

[0005] 그러나, 이 기술들 중 다수는 고속, 즉, 100Gb/s 이상의 패킷 프로세싱 디바이스들, 예를 들면, 네트워크 프로세서(network processor: NP)들, 트래픽 매니저(traffic manager: TM)들 및 스위치 패브릭(switch fabric: SF)들에 대해 너무 복잡해진다. 예를 들면, 칩의 클럭 주파수 및 공급 전압을 변조하는 DVS 방법은 대역폭, 프로세싱 레이턴시(latency) 및 지터(jitter)에 대하여 특수한 요건을 가지는 고속 패킷 프로세싱 디바이스들에 통합되기 매우 어렵다.

[0006] 그러므로, 상술한 문제들을 해결하여 고속 패킷 교환 시스템들에 대한 더욱 효율적이고 비용 효율적인 전력 감소를 제공할 필요가 있다.

발명의 내용

해결하려는 과제

[0007] 종래 기술의 상기 문제점들을 고려하면, 본 발명의 목적은 패킷 교환 시스템들, 특히 100Gb/s 이상의 고속 패킷 교환 시스템들의 전력 소비를 감소시킬 수 있는 더욱 효율적인 방법 및 패킷 교환 시스템을 제공하는 것이다.

과제의 해결 수단

[0008] 이 목적은 독립 청구항들에 따른 발명 대상에 의해 달성된다. 종속 청구항들은 본 발명의 바람직한 실시예들을 칭한다.

[0009] 본 발명의 양태에 따르면, 업스트림 패킷 프로세싱 디바이스 및 다운스트림 패킷 프로세싱 디바이스를 가지는 패킷 교환 시스템이 제안된다. 업스트림 패킷 프로세싱 디바이스는 특히 네트워크 프로세서일 수 있다. 다운스트림 패킷 프로세싱 디바이스는 트래픽 매니저 및/또는 스위치 패브릭일 수 있다.

[0010] 상기 방법은 업스트림 패킷 프로세싱 디바이스에서 인입하는 데이터 패킷들에 대한 트래픽 레이트(traffic rate)를 결정하는 단계를 포함할 수 있다. 예를 들면, 인입하는 데이터 패킷들은 인터페이스 초과 신청(oversubscription)을 관리하고, 버스트(burst)들을 필터링하고, 초기 분류를 실행하고, 패킷 길이를 결정하기 위해 업스트림 패킷 프로세싱 디바이스의 사전-분류 버퍼에 저장된다. 그 다음 결정된 패킷 길이는 인입하는 데이터 패킷들에 대한 트래픽 레이트의 후속 측정에 이용될 수 있고 이 후속 측정은 인그레스(ingress) 데이터 패킷들에서의 수를 측정하는 것을 포함한다. 트래픽 레이트는 트래픽 미터링 유닛(traffic metering unit)에 의해 결정될 수 있다. 본 발명의 추가적인 양태에 따르면, 인입하는 데이터 패킷들은 업스트림 데이터 프로세싱 디바이스에서 프로세싱되고나서 업스트림 패킷 프로셋이 디바이스로부터 다운스트림 패킷 프로세싱 디바이스로 송신될 수 있다. 예를 들면, 업스트림 데이터 프로세싱 디바이스에서의 패킷 프로세싱은 다음: 패킷 분류, 큐잉(queuing), 어드레스 학습(address learning) 및 포워딩 테이블의 관리, 어드레스 및 서비스 등급(class of service: CoS) 매핑을 포함하는 브릿지 기능(bridging functionality), MPLS 레이블(label) 생성 또는 교환(swapping) 중 적어도 하나를 포함할 수 있다.

[0011] 상기 방법은 업스트림 패킷 프로세싱 디바이스로부터 다운스트림 패킷 프로세싱 디바이스로 시간 이전에, 즉 트래픽 레이트가 결정된 프로세싱된 데이터 패킷들이 다운스트림 패킷 프로세싱 디바이스로 송신되는 시간 이전의 시간에 결정된 트래픽 레이트의 표시를 송신하는 단계를 포함할 수 있다. 트래픽 레이트 표시의 송신과 트래픽 레이트에 대응하는 프로세싱된 데이터 패킷들의 송신 사이의 시간차는 적어도 다운스트림 패킷 프로세싱 디바이스의 패킷 프로세싱 리소스를 활성화하거나 비활성화하는데 필요한 시간일 수 있다. 이 트래픽 레이트 표시는 인그레스 데이터 패킷들의 결정된 트래픽 레이트로부터 도출되는 값일 수 있고, 트래픽 레이트를 기술하는 정보를 포함할 수 있다. 본 발명의 추가적인 양태에 따르면, 다운스트림 패킷 프로세싱 디바이스에서의 이용가능한 패킷 프로세싱 리소스들은 수신된 트래픽 레이트에 기초하여 조정될 수 있다. 이용가능한 리소스들은 다운스트림 패킷 프로세싱 디바이스의 병렬 패킷 프로세싱 리소스들의 일부를 스위치 온(switch-on)하거나 스위치 오프(switch-off)하여 활성 리소스들을 트래픽 부하에 적응시킴으로써 조정하여, 여분의 프로세싱 케이퍼빌리티(capability)들에 전력 공급을 중단하여 에너지를 절약함으로써 열 발생이 감소하고 프로세싱 유닛들의 수명이 증가한다.

[0012] 본 발명의 상황에서, 패킷 프로세싱 리소스는 데이터 메모리, 특히 병렬 메모리 뱅트(memory bank)들, 데이터 패킷 송신 회선들, 또는 데이터 패킷들을 처리, 예를 들면, 데이터 패킷들을 저장, 큐잉, 포워딩 또는 변경하기 위한 임의의 다른 리소스를 포함할 수 있다.

[0013] 즉, 부하 미터링 정보가 측정되었던 데이터 패킷들이 여전히 업스트림 패킷 프로세싱 디바이스에서 프로세싱되

고 있는 동안에 상기 데이터 패킷들의 세트에 대하여 결정된 부하 미터링 정보를 업스트림으로부터 다운스트림 데이터 패킷 프로세싱 디바이스로 송신함으로써, 본 발명에 따른 방법 및 시스템은 트래픽 예측이 결정된 데이터 패킷들이 다운스트림 패킷 프로세싱 디바이스에 도달하기 전에 다운스트림 패킷 프로세싱 디바이스가 병렬 패킷 프로세싱 리소스들을 활성화시킬 수 있는지 또는 비활성화시킬 수 있는지의 여부에 기초하여 트래픽 기대값을 다운스트림 패킷 프로세싱 디바이스에 송신할 수 있다. 결정된 트래픽 레이트를 송신하는 단계들은 트래픽 레이트의 송신과 대응하는 데이터 패킷들의 송신 사이의 시간차가 다운스트림 패킷 프로세싱 디바이스에 충분한 시간을 제공하여 송신되는 트래픽 레이트 정보에 기초하여 상기 디바이스의 리소스들을 조정하는 한, 업스트림 패킷 프로세싱 디바이스에서 인입하는 데이터 패킷들을 프로세싱하는 단계 이전에 또는 상기 단계와 동시에 실행될 수 있다. 업스트림 패킷 프로세싱 디바이스에서 측정되는 부하 미터링 정보를 이용함으로써, 다운스트림 디바이스는 통상적으로 버퍼를 요구하며 디바이스의 네트워크 레이턴시 및 복잡도를 증가시키는 자체의 내장 미터링 유닛을 구비할 필요성을 방지한다. 프로세싱 부하를 미리 다운스트림 패킷 프로세싱 디바이스에 송신함으로써, 다운스트림 패킷 프로세싱 디바이스는 트래픽 레이트가 측정되었던 프로세싱된 데이터 패킷들이 다운스크림 디바이스에 도달하기 전에 충분한 시간을 갖고 상기 디바이스의 리소스들을 조정한다. 이 방식으로, 다운스트림 패킷 프로세싱 디바이스는 패킷 프로세싱으로 인하여 업스트림 패킷 프로세싱 디바이스에서 발생하는 지연 시간을 이용하여, 데이터 패킷들 이전에 수신된 트래픽 레이트 정보에 기초하여 자신의 리소스 구성을 최적화할 수 있다. 결과적으로, 다운스트림 패킷 프로세싱 디바이스는 자신의 리소스들을 더욱 효율적으로 관리함으로써 에너지 소비를 감소시킬 수 있다.

[0014] 본 발명의 다른 양태에 따르면, 상기 방법은 이용되는 병렬 패킷 프로세싱 유닛의 수를 최소화함으로써 데이터 패킷들을 병렬 프로세싱 유닛들을 가지는 패킷 프로세싱 리소스들에 걸쳐 부하-분배하는 단계를 포함할 수 있다. 고속 패킷 교환 시스템들에서의 종래의 부하-밸런싱(load-balancing) 방법들은 프로세싱 부하가 모든 프로세싱 리소스들에 의해 공유되도록 데이터 패킷들을 이용가능한 병렬 패킷 프로세싱 리소스들에 걸쳐 균등하게 분배함으로써, 데이터 패킷들을 병렬 프로세싱 리소스들에 걸쳐 부하-밸런싱하는 것이다. 대조적으로, 본 발명은 본 부하를 가능한 적은 수의 프로세싱 유닛들에 분배하여 유휴 프로세싱 유닛들의 수를 최대화하고나서 결정된 트래픽 레이트가 더 낮은 트래픽 예측을 표시하는 경우 유휴 프로세싱 유닛들에 전력 공급을 중단하여 전력 소비를 절약하는 것을 제안한다. 예를 들면, 이용도가 낮은 패킷 프로세싱 디바이스의 모든 이용가능한 패킷 송신 회선들을 이용하는 대신, 본 발명은 데이터 패킷을 가능한 적은 수의 병렬 송신 회선들에 걸쳐 분배하여, 트래픽 레이트에 의해 표시되는 바에 따라 다수의 이용되지 않은 송신 회선들을 스위치 오프하는 것을 용이하게 행하는 것을 제안한다.

[0015] 본 발명의 다른 양태에 따르면, 트래픽 레이트 정보를 송신하는 단계에서, 업스트림 패킷 프로세싱 디바이스는, 업스트림 패킷 프로세싱 디바이스에서 업스트림 패킷 프로세싱 디바이스에 의해 이미 프로세싱된 적어도 하나의 패킷에, 계속해서 프로세싱되고 있는 인입하는 데이터 패킷들에 대하여 결정되는 트래픽 레이트를 포함하는 헤더 필드(header field)를 추가할 수 있다. 결과적으로, 업스트림과 다운스트림 패킷 프로세싱 디바이스 사이의 시그널링 부하는 트래픽 레이트를 결정하는 정보가 다운스트림 데이터 패킷들과 함께 반송될 수 있을 때 증가하지 않을 것이다. 대안으로, 업스트림 패킷 프로세싱 디바이스는 별개의 제어 메시지 또는 데이터 패킷을 이용하여 트래픽 레이트를 송신할 수 있다. 부하-미터링 정보를 전송하고 평가하는 수단은 대역 내 송신으로만 결부되는 것은 아니고, 또한 대역 외에서 직접, 또는 다른 시설들에 대한 간접 관찰에 의해 발생할 수 있다.

[0016] 본 발명의 추가적인 양태에 따르면, 트래픽 레이트는 인입하는 데이터 패킷들에 대한 정보 레이트 값 및/또는 버스트 레이트 값을 결정함으로써 결정될 수 있다. 상기 정보 레이트 및 버스트 레이트에 대하여 결정된 값들은 어떤 이용가능한 리소스가 전력 효율을 증대시키는데 동적으로 이용될 수 있는지에 기초하여 패킷 프로세싱 디바이스의 리소스 필요성의 정확한 결정에 이용될 수 있다. 이것은 각각의 인입하는 패킷 플로우에 대해 또는 다운스트림 패킷 프로세싱 디바이스의 복잡도에 따라 취합된 패킷 대역폭에 대해 개별적으로 정보 레이트 값 및/또는 버스트 레이티를 결정하는데 유용할 수 있다.

[0017] 본 발명의 추가적인 양태는 패킷 프로세싱 디바이스의 업스트림 패킷 프로세싱 디바이스 및 다운스트림 패킷 프로세싱 디바이스에 관한 것으로서, 업스트림 패킷 프로세싱 디바이스는 인입하는 데이터 패킷들에 대한 트래픽 레이트를 결정하도록 구성되는 트래픽 미터링 유닛을 포함할 수 있다. 업스트림 패킷 프로세싱 디바이스는 트래픽 레이트가 결정된 프로세싱된 데이터 패킷들이 다운스트림 패킷 프로세싱 디바이스로 송신되는 시간 이전의 시간에 결정된 트래픽 레이트를 업스트림 패킷 프로세싱 디바이스로부터 다운스트림 패킷 프로세싱 디바이스로 송신하기 위한 수단을 포함할 수 있다. 업스트림 패킷 프로세싱 디바이스는 인그레스 데이터 패킷들을 프로세싱하기 위한 데이터 패킷 프로세서를 추가로 포함할 수 있고, 프로세싱된 데이터 패킷들을 업스트림 패킷 프로세

싱 디바이스로부터 다운스트림 패킷 프로세싱 디바이스로 송신하기 위한 데이터 패킷 송신기를 포함할 수 있다. 다운스크림 패킷 프로세싱 디바이스는 수신된 트래픽 레이트에 기초하여 다운스트림 패킷 프로세싱 디바이스에서 이용가능한 패킷 프로세싱 리소스들을 조정하도록 구성되는 리소스 매니저를 포함할 수 있다.

[0018] 본 발명의 상황에서, 업스트림 패킷 프로세싱 디바이스는 다수의 다운스트림 패킷 프로세싱 디바이스들, 예를 들면, 업스트림 패킷 프로세싱 디바이스로부터 직접적으로 데이터 패킷들을 수신하도록 구성되어 있는 제 1 다운스트림 패킷 프로세싱 디바이스, 및 데이터 패킷들을 업스트림 패킷 프로세싱 디바이스로부터 제 1 다운스트림 패킷 프로세싱 디바이스를 통해 수신하도록 구성되어 있는 제 2 다운스트림 패킷 프로세싱 디바이스에 서빙할 수 있다. 업스트림 패킷 프로세싱 디바이스는 또한 일련의 다운스트림 패킷 프로세싱 디바이스에 서빙할 수 있다. 본 발명의 상황에서의 다운스트림 패킷 프로세싱 디바이스는 적어도 하나의 업스트림 패킷 프로세싱 디바이스에 의해 결정되고 송신되었던 트래픽 레이트 정보를 수신하고 이용하여 다운스트림 패킷 프로세싱 디바이스에서 패킷 프로세싱 리소스들을 조정한다.

[0019] 바람직하게도, 업스트림 패킷 프로세싱 디바이스는 인입하는 데이터 패킷들의 사전-분류 및 초과 신청 관리를 버퍼링(buffering)하기 위한 버퍼링 유닛을 포함할 수 있다. 이것은 이 버퍼링이 버스트들을 필터링하고 후속 트래픽 미터링에 대한 패킷 길이를 결정하기 위해 수백개의 플레임들을 보유할 수 있다는 사실로 인해 유용할 수 있다.

[0020] 본 발명의 다른 양태에 다르면, 업스트림 패킷 프로세싱 디바이스는 패킷 분류, 전송 및 태깅(tagging)을 실행하기 위하여 프로세싱 케이퍼빌리티들, 바람직하게는 병렬 패킷 프로세싱 케이퍼빌리티를 포함하는 고속 네트워크 백본(backbone)의 네트워크 프로세서일 수 있고, 다운스트림 패킷 프로세싱 디바이스는 데이터 패킷들의 트래픽 관리, 큐잉, 분할화 및 재조립을 구현하도록 구성되는 병렬 프로세싱 케이퍼빌리티들 및 스위칭 패브릭으로의 인터페이스 관리를 포함하는 네트워크 백본의 트래픽 매니저일 수 있고, 트래픽 매니저는 다중-레인(lane) 송신 인터페이스를 통해 네트워크 프로세서 디바이스로 연결될 수 있다. 트래픽 매니저가 스위칭 서브시스템의 총 전력 소모에 대한 주요 원인이라는 사실로 인해 본 발명을 트래픽 매니저에 이용하는 것이 특히 유용하다. 예를 들면, 트래픽 매니저 디바이스의 리소스 매니저는 다운스트림 패킷 프로세싱 디바이스의 병렬 메모리 뱅크들의 세트를 제어하기 위한 메모리 제어기 또는 다운스크림 패킷 프로세싱 디바이스의 다수의 송신 레인들의 세트를 제어하기 위한 인터페이스 제어기일 수 있다.

[0021] 예를 들면, 네트워크 프로세서 및 트래픽 매니저는 고속 상호 접속을 통해 스위칭 카드(switching card)에 접속되는 라인 카드 상에서 구현될 수 있다. 본 발명의 방법 및 시스템은 바람직하게도 고속 패킷 스위칭 네트워크, 예를 들면, 100 Gb/s 이상의 네트워크들에서 구현된다.

[0022] 본 발명의 또 다른 양태에 따르면, 패킷 교환 시스템의 리소스 매니저는 병렬 프로세싱 요소들 중 하나 이상의 병렬 리소스들을 활성화하거나 비활성화할 수 있다. 본 발명의 추가적인 양태에 따르면, 고속 트래픽 매니저 디바이스는 수십 개의 병렬 메모리 뱅크들, 파이프라인 루프(pipeline loop)들, 및 다중-레인 인터페이스들을 포함하여, 에너지를 절약하고, 열 발생 및 냉각 요건들을 감소시키고, 패킷 프로세싱 유닛들의 수명을 증가시키도록 결정되는 트래픽 레이트에 기초하여 부분적으로 전력 공급이 중단될 수 있는 원하는 처리량을 달성할 수 있다.

[0023] 본 발명의 다른 양태에 따르면, 다운스트림 패킷 프로세싱 디바이스의 리소스 매니저가 수신된 트래픽 레이트에 기초하여 패킷 프로세싱 리소스들의 상태를 변경하는 상태 기계(state machine)를 포함할 수 있다. 예를 들면, 이 상태 기계는 상태를 선택하고 수신된 정보 레이트 및 버스트 레이트 값들에 기초하여 디바이스의 병렬 프로세싱 구성요소들의 각각에 대한 상태들 사이의 전이(transition)를 관리하도록 구성될 수 있고, 상기 상태는 병렬 프로세싱 구성요소들의 이용가능한 병렬 프로세싱 리소스들을 결정한다.

[0024] 업스트림 패킷 프로세싱 디바이스에서 패킷 교환 시스템을 통하는 송신 플로우를 따라서 결정되는 부하-미터링 정보를 이용하여 송신 성능을 실제 요구에 적응시키기 위해서, 본 발명의 추가적인 양태는 복수의 네트워크 프로세서들 및 복수의 트래픽 매니저들을 포함할 수 있는 패킷 교환 시스템에 관한 것이고, 추가 다운스트림 패킷 프로세싱 디바이스로서, 스위치 패브릭은 복수의 패킷 트래픽 매니저들에 교차 섭속되는 다수의 스위치 디바이스들을 포함한다. 스위치 패브릭의 리소스 매니저는 복수의 업스트림 네트워크 프로세서들로부터 트래픽 매니저를 통해 스위치 패브릭으로 송신되는 모든 트래픽 레이트들을 수신하고 모든 수신된 트래픽 레이트들로부터 결정되는 취합된 트래픽 제이트 정보에 기초하여 스위치 패브릭의 패킷 프로세싱 리소스들을 관리하도록 구성될 수 있다. 취합된 트래픽 레이트 정보는 스위치 패브릭에 대한 전체 인그레스 트래픽 부하 예측을 결정하고, 이 예측에 기초하여 스위치 패브릭 리소스 매니저는 자신의 패킷 프로세싱 리소스들의 일부를 스위치 온(switch

on) 또는 스위치 오프(switch off)하여 자체의 전력 소비 효율을 증가시킬 수 있다.

[0025] 본 발명의 다른 양태에 따르면, 스위치 패브릭의 교차 접속되는 교환 디바이스는, 교환 디바이스를 복수의 트래픽 매니저들과 연결시키는 모든 송신 링크들이 트래픽 매니저들에서의 수신된 트래픽 레이트들에 기초하여 트래픽 매니저들에 의해 비활성화되었던 경우에, 비활성화도록 구성될 수 있다.

[0026] 본 발명의 다른 양태에 따르면, 리소스 매니저는 병렬 패킷 프로세싱 요소들을 부분적으로 전력 공급을 중단하여 정적 전력 소모를 감소시키거나 패킷 프로세싱 리소스들을 클럭 게이팅(clock gating)하여 동적 전력 소모를 감소시킴으로써 이용가능한 패킷 프로세싱 리소스들을 조정할 수 있다.

[0027] 본 발명은 첨부 도면들을 참조하는 예시적인 방식으로 아래에 설명된다.

도면의 간단한 설명

[0028] 도 1은 네트워크 프로세서 및 트래픽 매니저를 포함하는 고속 패킷 스위칭 서브시스템의 블록도.

도 2는 본 발명의 실시예에 따른 업스트림 패킷 프로세싱 디바이스 및 다운스트림 패킷 프로세싱 디바이스의 블록도.

도 3은 본 발명의 실시예에 따른 패킷 교환 시스템에서 트래픽-부하 의존 전력 감소에 수반되는 단계들의 흐름도.

도 4는 본 발명의 실시예에 따라 결정된 트래픽 레이트 정보를 포함하는 추가 헤더 필드가 추가되는 데이터 패킷의 예를 도시하는 도면.

도 5는 메모리 제어기 및 복수의 병렬 메모리 유닛들을 포함하는 트래픽 매니저의 패킷 버퍼를 도시하는 도면.

도 6은 본 발명의 다른 실시예에 따라 복수의 네트워크 프로세서들, 트래픽 매니저들 및 스위칭 엔티티(entity)들을 포함하는 고속 패킷 교환 시스템의 블록도.

도 7은 네트워크 프로세서의 예시적인 아키텍처를 도시하는 도면.

도 8은 트래픽 매니저의 예시적인 아키텍처를 도시하는 도면.

발명을 실시하기 위한 구체적인 내용

[0029] 도 1은 네트워크 프로세서 및 트래픽 매니저를 포함하는 고속 패킷 교환 서브 시스템의 블록도를 도시한다. PHY 디바이스(2) 및 MAC/프레이머 디바이스(framer device)(3)는 이더넷 전송을 위한 것으로 포워딩, 분류, 우선화 및 플로우 제어를 위해서 인입하는 네트워크 데이터를 네트워크 프로세서(network processor: NP)(10) 및 트래픽 매니저(traffic manager: TM)(20)로 계속 통과시킨다. PHY 디바이스(2), MAC/프레이머 디바이스(3), NP(10), 및 TM(20) 및 백플레인(backplane)과 인터페이싱하는 패브릭 매니저(5)는 전형적으로 라인 카드(1)를 구현한다. 라인 카드(1)는 세이터를 스위치 패브릭 디바이스(SF)(30)에 송신하고, 스위치 패브릭 디바이스(30)는 데이터를 다른 라인 카드들로 통과시킨다. NP(10), TM(20), 및 SF(30)는 이더넷 스위칭 기능들과 관련된다. 디바이스들은 100Gb/s 이상의 고속 패킷 송신들을 위해 병렬 다중-레인 송신 회선들을 통해 접속된다. 본 발명의 방법은 도 2에 더 기술되고 네트워크 프로세서 및 트래픽 매니저를 포함하는, 그러한 고속 패킷 교환 서브시스템에서 구현될 수 있다.

[0030] 도 2는 본 발명의 실시예에 따른 업스트림 패킷 프로세싱 디바이스 및 다운스트림 패킷 프로세싱 디바이스의 블록도를 도시한다. 업스트림 패킷 프로세싱 디바이스는 인그레스 네트워크 프로세서(10)이고 다운스트림 패킷 프로세싱 디바이스는 트래픽 매니저(20)이다.

[0031] 네트워크 프로세싱은 일부 규칙들의 세트에 따라 인입하는 네트워크 데이터 패킷들을 프로세싱하고, 프로세싱된 패킷들을 발신하는 송신 회선 상으로 전송하는 것을 칭한다. 본 발명의 실시예에 따른 NP(10)는 인입하는 데이터 패킷들을 버퍼링하기 위한 사전-분류 버퍼(11)를 포함하여 인터페이스 초과 신청을 관리하고, 버스트들을 필터링하고, 인그레스 데이터 패킷들에 대한 패킷 길이를 결정한다. 패킷 길이는 인입하는 데이터 패킷들의 트래픽 레이트를 결정하는 트래픽 미터(12)에 의해 후속 미터링하는데 이용된다. NP(10)는 패킷들이 미터링 유닛(12)를 통과한 후에 인입하는 데이터 패킷들을 프로세싱하기 위해서 다수의 프로세싱 요소(Processing Element: PE)들(13)을 추가로 포함한다. PE들(13)은 명령 세트가 네트워크 데이터 패킷들을 처리하고 프로세싱하도록 맞춤화되었던 축소된 명령어 축약형 세트 컴퓨터(Reduced Instruction Set Computer: RISC) 코어들로 구현될 수

있다. 특히, 자체의 PE들(13)을 가지는 NP(10)는 패킷 분류, 필터링, 포워딩, 분류화, 미터링, 표시, 정책 및 계수와 같은 다양한 기능들을 실행한다. 예를 들면, NP는 어드레스 학습 및 포워딩 테이블의 관리, 어드레스 및 서비스 등급(CoS) 매핑, VLAN 태그 프로세싱, MPLS 레이블 생성 또는 교환을 포함하는 브리징 기능을 위해 구성될 수 있다. 전형적으로, NP들은 다수의 PE들(13)을 포함하는 병렬 프로세싱 케이퍼빌리티들을 위해 구성된다. 고속 패킷 교환 시스템들의 NP들에 대한 상이한 구현예들이 예를 들면, 아키텍처, 복잡도, 또는 프로그램 가능성에 관하여 존재한다. NP(10)의 예시적인 아키텍처가 도 7에 도시된다.

[0032] 도 2에 도시되는 바와 같이, 인그레스 NP(10)는 병렬의 다수-레인 송신 회선들을 통해 다운스트림 TM(20)에 접속된다. TM들(20)은 전형적으로 패브릭 매니저 옆에 있는 라인 카드 상에 상주하는데 왜냐하면 이것들은 스위치에 필요한 출력 큐잉을 구현하기 때문이다. TM(20)은 전형적으로 백플레인 옆에 있는 가상 출력 큐(Virtual Output Queue: VOQ)들을 가지는 인그레스 카드(1) 상에서 실현되는 패킷 큐잉을 실행한다. 대안으로, TM 및 FM은 하나의 디바이스로 통합될 수 있다. TM(20)은 흔히 교환 서브시스템의 총 전력 소모의 주요 원인이다. 데이터 패킷들을 저장하기 위해 트래픽 매니저들이 상당한 양의 메모리(21)를 요구하는 것이 하나의 이유이다. 메모리(21)는 내장형 메모리 뿐만 아니라, 고속 외부 메모리들에 대한 지원을 포함할 수 있다. 네트워크 프로세서 및 스위치 패브릭으로의 다중-레인 인터페이스들을 위한 다수의 송수신기들이 다른 이유다.

[0033] TM(20)은 트래픽 미터(12)에 의해 결정되는 트래픽 페이트에 기초하여 TM(20)의 패킷 프로세싱 리소스들을 조정하도록 구성되는 리소스 매니저(23)를 추가로 포함한다.

[0034] 예를 들면, 트래픽 매니저 디바이스의 리소스 매니저(23)는 트래픽 부하에 따라 병렬 메모리 모듈들(22)의 일부를 임시로 비활성화하여 TM(20)의 전력 소비를 감소시키는 메모리 제어기(23a)를 포함하고 NP 및/또는 FM으로의 송신 인터페이스들의 특정한 레인들을 비활성화하기 위한 인터페이스 제어기(23b)를 포함한다.

[0035] TM은 외부 메모리에 저장되는 패킷 데이터에 대한 포인터들인, 트래픽 큐들을 유지하는 큐 매니저와 같은 구성요소들(도시되지 않음)를 추가로 포함한다. 별개의 큐는 각각의 트래픽 등급 또는 트래픽 플로우에 대하여 유지될 수 있다. 그러나, NP(10)와는 대조적으로, TM(20)에는 전형적으로 미터링 케이퍼빌리티들이 없는데, 왜냐하면 이것은 전형적으로 패킷 교환 서브시스템의 레이턴시 및 복잡도를 증가시킬 것이기 때문이다. 도 8은 트래픽 매니저의 예시적인 아키텍처를 도시한다.

[0036] NP(10) 및 TM(20)의 설명은 예를 통해 제공된다. NP들 및 TM은 설계 및 구현되는 기능들에 따라 변할 수 있다. 게다가, NP의 기능들의 일부는 TM 내에서 구현될 수 있고 이 역도 마찬가지이다.

[0037] 도 3은 본 발명의 실시예에 따른 패킷 교환 시스템에서 트래픽-부하 의존 전력 감소에 수반되는 단계들의 흐름도를 도시한다.

[0038] 단계 100에서, NP(10)의 사전-분류 버퍼(11)는 인터페이스 초과 신청을 관리하고, 버스트들을 필터링하고, 단계 S110에 설명되는 바와 같이, 인그레스 데이터 패킷들에 대한 패킷 길이를 결정하기 위해 인입하는 데이터 패킷들을 버퍼링한다. 패킷 길이는 후속 미터링에 이용되고, 패킷들이 - 잠재적으로 혼잡되는 - 프로세싱 파이프라인, 즉, 프로세싱 요소들(13)에 진입하기 전에 기본 서비스 품질(Quality of Service: QoS)를 실행하는데 이용된다. 초과 신청/사전-분류 버퍼(11)를 통과시키고, 트래픽 미터(12)는 단계 S120에서 인입하는 데이터 패킷들의 트래픽 레이트를 결정한다. 예를 들면, 트래픽 레이트는 정보 레이트 및 그 결과에 따른 (형성된) 트래픽 플로우의 버스트 레이트를 측정함으로써 결정될 수 있다. 이 측정된 레이트들은 트래픽 레이트를 프로세싱 요소들의 파이프라인으로 제한하는데 이용되고 또한 후속 전력 감소 메커니즘에 이용될 것이다. 다운스트림 TM(20)의 복잡도에 따라, 트래픽 레이트는 각각의 패킷 플로우에 대해 또는 취합된 패킷 대역 폭에 대해 측정될 수 있다. 트래픽 레이트가 측정되었던 패킷들이 NP의 프로세싱 요소들(13)에서 여전히 프로세싱되고 있는 동안, NP로부터 TM까지의 작업부하 정보의 송신은 단계 S130에서 추가 "기대 레이트" 확장 헤더 필드를 이미 프로세싱된 패킷들에 선첨부(prepend)함으로서 달성된다. 그 다음, 단계 140에서, 결정된 트래픽 레이트는 트래픽 레이트가 결정된 프로세싱된 데이터 패킷들이 TM(20)에 송신되는 시간 이전의 시간에 NP(10)로부터 TM(20)으로 송신된다.

[0039] 이것은 트래픽 부하 정보가 송신 경로 내의 실제 데이터에 앞서 이동하는 것을 보장한다. 부하 미터링 정보를 송신하고 평가하는 수단은 반드시 대역 내 송신에만 결부되는 것은 아니고, 또한 대역 외에서 직접, 또는 다른 시설들의 간접 측척에 의해 발생할 수 있다. 대안으로, 업스트림 패킷 프로세싱 디바이스는 개별 제어 메시지 또는 데이터 패킷을 이용하여 트래픽 레이트를 송신할 수 있다.

[0040] 도 4는 본 발명의 실시예에 따라 결정된 트래픽 레이트 정보를 포함하는 추가 "기대 레이트(Expected Rate)" 헤더 필드(41)가 추가되었던 데이터 패킷(40)의 예를 도시한다. '기대 레이트' 헤더 외에도, 데이터 패킷(40)은

추가 내부 메타-정보를 포함하는데 왜냐하면 NP(10)는 인그레스 패킷들에 대하여 전송 결정을 행하고 있고 TM(20)에 대해 정확한 큐를 선택하며, 이것은 전형적으로 모든 패킷 앞에는 여분의 헤더가 추가될 것을 요구하기 때문이다. 인입하는 데이터 트래픽의 수정된 헤더로 인해 트래픽 매니저(20)는 어떤 패킷들이 탈락되고 재송신되어야만 하는지, 패킷들이 스위치 패브릭(30)으로 언제 전송되어야 하는지, 그리고 네트워크 상으로 송신될 때 트래픽이 어떠한 형상이어야만 하는지를 우선순위화하고 결정하는 것을 인에이블(enable)한다.

[0041] 특히, 트래픽 레이트가 결정된 데이터 패킷들 이전에 그리고 이 데이터 패킷들이 여전히 도 3의 단계 S145에서 NP(10)의 프로세싱 요소들(13)에 의해 프로세싱되고 있는 동안에 트래픽 레이트 정보가 송신되므로 송신되는 트래픽 레이트는 TM(20)에 대한 트래픽 예측을 나타낸다. TM(20)의 리소스 매니저(23)는 단계 150에서 송신되는 트래픽 레이트를 분석하여 요구되는 프로세싱 리소스들을 결정한다. 단계 S160에서, 리소스 매니저(23)는 그 후에 자신의 이용가능한 패킷 프로세싱 리소스를 수신된 트래픽 레이트에 기초하여 조정한다. 예를 들면, 병렬 메모리 뱅크들(21), 인터페이스 레인들(4), 파이프라인 루프들 또는 다른 프로세싱 요소들은 트래픽 부하에 따라 (부분적으로) 차단될 수 있다.

[0042] 단계 145에서 데이터 패킷들이 NP(10)에서 프로세싱된 후에, 데이터 패킷들은 데이터 패킷들이 단계 180에서 프로세싱되는 단계 170에서 TM(20)으로 송신되고, 여기서 TM은 인입하는 데이터 패킷들에 대하여 자체의 프로세싱 리소스들을 이미 조정했으므로, 불필요한 예비 회선들을 방지하고 자체의 전력 효율을 증가시킨다. 업스트림 패킷 프로세싱 디바이스에서 패킷 프로세싱에 대한 전형적인 레이턴시 시간들은 100㎲의 범위 내에 있고, 때로는 수 ms에 이르고, 반면에 전력-웨이크업(wake-up) 회로들에 대한 전형적인 활성 시간들은 하위 레벨 회로들에 대해서는 서브-㎲ 내에 있고 메모리 모듈과 같은 상위 논리 블록들에 대해서는 10㎲ 이상의 범위 내에 있다. 패킷 프로세싱에 대한 업스트림 패킷 프로세싱 디바이스의 레이턴시가 다운스트림 패킷 프로세싱 디바이스에서의 전력-웨이크-업 회로들에 대한 활성/비-활성 시간보다 더 크다면, 트래픽 레이트가 결정된 데이터 패킷들이 NP에서 프로세싱되기 전에 트래픽 레이트가 송신되는 경우에 레이트 정보를 송신함으로써 다운스트림 디바이스에서 패킷 프로세싱 리소스들을 조정하는데에 충분한 시간이 제공될 것이다.

[0043] 단계들 S130, 140, 150 또는 S160 중 임의의 단계는 트래픽 레이트의 송신과 대응하는 데이터 패킷들의 송신 사이의 시간차가 다운스트림 패킷 프로세싱 디바이스에 충분한 시간차를 제공하여 자체의 리소스들을 송신되는 트래픽 레이트 정보에 기초하여 조정하는 한, 단계 S145 이전에 또는 동시에 실행될 수 있다.

[0044] 추가 장점은 NP(10)에서 측정되는 트래픽 레이트 정보를 이용함으로서 TM(20)가 임의의 복잡한 패킷 특성화 케이퍼빌리티들이 자체의 프로세싱 리소스들을 조정할 필요가 없다는 것이다.

[0045] 도 5는 메모리 제어기 및 다수의 병렬 메모리 유닛들을 포함하는 트래픽 매니저의 패킷 버퍼를 도시한다. 메모리 성능의 한계들로 인해, 패킷 버퍼 메모리(21)는 다수의 병렬 '뱅크들' 및 메모리 제어기를 이용하여 실현된다. 최신의 800 MHz DDR3-SDRAM들이 100 Gb/s 패킷 버퍼의 메모리 모듈들에 대해 이용되는 경우조차도, 10 내지 20 뱅크들에 걸쳐 패킷 데이터를 분배하는 것이 원하는 처리량을 달성하는데 필요하다. 이 수 및 이 결과에 따른 전력 소모는 심지어 차세대 장비의 경우에서도 증가할 것이다. 병렬 메모리 모듈들은 메모리 뱅크들을 활성화 또는 비활성화하는 메모리 제어기에 의해 송신되는 트래픽 레이트 값들에 기초하여 제어된다. 그러므로 메모리 제어기는 리소스 매니저의 기능을 실행한다.

[0046] 리소스 매니저(23)는 수신된 트래픽 레이트에 기초하여 패킷 프로세싱 리소스들의 상태를 변경하는 상태 기계를 포함할 수 있다. 예를 들면, 상태 기계는 수신된 트래픽 레이트의 하한 및 상한 값에 의해 규정되는 다수의 트래픽 레이트 간격들을, 예를 들면, 정보 레이트 및 버스트 레이트 값들에 기초하여 규정하여, 요구되는 패킷 프로세싱 리소스들을 결정할 수 있고, 여기서 각각의 간격은 TM(20)의 사전-결정된 리소스 상태에 대응한다. 이 리소스 상태가 필요한 시간 지점은 TM(20)에서 송신되는 트래픽 레이트의 도착 및 트래픽 레이트가 결정되어 송신된 데이터 패킷들의 도착 사이의 평균 시간차에 의해 규정된다.

[0047] 그 다음 상태 기계는 프로세싱 리소스들의 상태, 예를 들면, 병렬 메모리 뱅크들 중 능동 메모리 뱅크들의 수를 각각의 트래픽 레이트 간격으로 할당할 수 있다. 트래픽 레이트 값이 사전 결정된 트래픽 레이트 간격 내에 있다면, 상태 기계는 이 간격에 대응하는 상태로의 전이를 개시한다. 본 발명에 따른 패킷 교환 시스템은 트래픽 레이트가 측정되었던 실제 데이터 이전에 TM(20)에서 트래픽 레이트 정보를 제공하도록 구성되고 여기서 트래픽 레이트의 송신과 트래픽 레이트에 대응하는 프로세싱된 데이터 패킷들의 송신 사이의 시간차는 적어도 TM 디바이스의 적응형 데이터 프로세싱 리소스들을 활성화 또는 비활성화하는데 필요한 시간이다.

[0048] 도 6은 본 발명의 다른 실시예에 따라, 복수의 네트워크 프로세서들(10)을 업스트림 패킷 프로세싱 디바이스들

로서, 복수의 트래픽 매니저들(20)을 제 1 다운스트림 패킷 프로세싱 디바이스들로서, 스위칭 패브릭(30)을 스위칭 엔티티들(32)을 가지는 제 2 패킷 프로세싱 디바이스로서 포함하는 고속 패킷 교환 시스템의 블록도를 도시한다. 교환 디바이스(32)는 복수의 병렬 트래픽 매니저들(20)에 교차 접속된다. 이 실시예에 따르면, 트래픽 레이트 정보 또는 부하-미터링 정보는 송신 성능을 실제 요구에 적응시키기 위해 패킷 교환 시스템을 통과하는 완성된 송신 플로우를 따라 송신되고 이용되어, 전 시스템의 이용되지 않는 송신 리소스들의 전력 감소가 인에이블된다.

[0049] 실시예에 따르면, 트래픽 매니저와 네트워크 프로세서 사이의 다수의 송신 레인들의 서브세트는 패킷 대역폭의 트래픽 매니저로의 취합에 따라 인터페이스 제어기에 의해 활성화되거나 비활성화될 수 있다.

[0050] 추가적인 실시예에 따르면, 트래픽 매니저의 다운스트림을 시작하고 이것을 (중앙집중식) 스위치 패브릭에 접속시키는 다수의 송신 레인들의 서브세트에는 패킷 대역폭의 중앙 스위치 패브릭으로의 실제 취합에 따라 전력 공급이 중단될 수 있다. 전형적으로, 중앙 패킷 교환 매트릭스(matrix)는 크기가 더 작은 상호-접속 디바이스들의 어레이로부터 구성되고, 회선 종료 디바이스(패킷 프로세싱 라인 카드)로부터 진입하는 트래픽은 상기 어레이에 걸쳐 부하-밸런싱된다. 모든 라인 카드들에 걸쳐 조직화되는 방식으로 다수의 레인들의 서브세트를 비활성화함으로써, 스위칭 구성요소들의 중앙집중화된 어레이의 부분들에는 전력이 최종적으로 하나의 디바이스마다 동적으로 완전히 공급 중단될 수 있다.

[0051] 트래픽 매니저들(20)은 제 1 다운스트림 디바이스들로서, NP들(10)으로부터의 패킷들의 헤더와 함께 전송되는 부하-미터링 정보를 추정한다. 그 다음 TM들(20)의 상태 기계들을 조정하여 메모리 뱅크들(23) 및/또는 송신 링크들(52)을 비활성화시킨다. 중앙 스위칭 매트릭스의 리소스 매니저(31)는 스위칭 엔티티들에 접속되는 링트들의 상태들을 관찰하고, 일단 예를 들면, 하나의 디바이스의 모든 링크들이 비활성되되면, 디바이스들에 전력 공급을 자동으로 중단한다. 대안으로, 복수의 병렬 트래픽 매니저들(20)에 교차 접속되는 스위치 디바이스들(32)은 스위치 디바이스의 업스트림 트래픽 매니저로의 모든 링크들이 비활성되는 경우, 전력 을 공급받는 것이 자동으로 중단되도록 구성된다.

[0052] 추가적인 실시예에 따르면, 스위치 패브릭 디바이스(30)의 리소스 매니저(31)는 복수의 네트워크 프로세서들(10)로부터 트래픽 매니저들(20)을 통해 스위치 패브릭으로 송신되는 모든 트래픽 레이트들을 수신하도록 구성되고, 모든 수신된 트래픽 레이트들로부터 결정되는 취합된 트래픽 레이트 정보에 기초하여 스위치 패브릭의 패킷 프로세싱 리소스들을 관리한다. 모든 인그레스 트래픽 레이트 값들을 취합함으로써, 리소스 매니저(31)는 전체 스위치 패브릭에 대한 트래픽 부하를 결정할 수 있고 능동 프로세싱 리소스들의 수를 조정하여 자체의 에너지 효율을 증가시킬 수 있다. 예를 들면, 리소스 매니저(31)는 병렬 패킷 프로세싱 요소들에 전력 공급을 부분적으로 중단시킴으로써 정적 전력 소모를 감소시키고/시키거나 패킷 프로세싱 리소스들을 클럭-게이팅함으로써 동적 전력 소모를 감소시킴으로서 이용가능한 패킷 프로세싱 리소스들을 조정할 수 있다.

[0053] 본 발명에 따른 방법 및 패킷 교환 시스템은 패킷 프로세싱 디바이스들이 수신된 트래픽 레이트 정보를 이용하여, 이용되는 프로세싱 리소스들의 수를 최소화하거나, 즉, 유휴 프로세싱 리소스들, 예를 들면, 메모리 뱅크들 또는 송신 링크들의 수를 최대화하는 방식으로, 데이터 프로세싱을 이용가능한 프로세싱 유닛들에 걸쳐 부하-밸런싱하도록 구성되는 경우, 에너치 효율이 최대가 된다. 예를 들면, 프로세싱 구성요소들을 50% 부하로 동작시키는 대신, 절반의 프로세싱 유닛들을 100%로 동작시키고 프로세싱 유닛들의 나머지 절반이 송신된 트래픽 예측에 의해 표시되는 바에 따라 필요하지 않은 경우 상기 나머지 절반의 프로세싱 유닛들에 전력 공급을 중단시키는 것이 유익하다. 바람직하게도, 다운스트림 패킷 프로세싱 디바이스의 리소스 매니저는 이용되는 병렬 패킷 프로세싱 유닛들의 수를 최소화함으로써 트래픽 부하를 병렬 프로세싱 유닛들을 가지는 패킷 프로세싱 리소스에 걸쳐 분배하도록 구성된다. 데이터 패킷들을 이용가능한 병렬 패킷 프로세싱 리소스들에 걸쳐 균등하게 분배하는 대신 부하를 가능한 적은 수의 병렬 프로세싱 유닛들에 걸쳐 분배함으로써, 수신된 트래픽 레이트 값에 기초하여 이용되지 않는 패킷 프로세싱 리소스들에 전력 공급을 중단하는 것이 더 빠를 것이고 전력 공급이 중단될 수 있는 리소스들의 수가 증가할 것이다. 도 6에 도시되는 바와 같이, TM들의 리소스 매니저(23)는 데이터 패킷들을 가능한 적은 수의 메모리 모듈들(21) 및 송신 회선들(51)에 분배하도록 구성된다. 메모리 모듈들(22) 및 송신 회선들(52)은 트래픽 부하가 메모리 모듈들(21) 및 송신 회선들(51)의 프로세싱 용량보다 더 작은 한 리소스 매니저(23)에 의해 서빙되지 않는다. 수신된 트래픽 레이트 정보를 분석함으로써, 리소스 매니저(23)는 유휴 프로세싱 유닛들(22 및 52)에 전력 공급이 중단되어 에너지 소비를 절약할 수 있는지를 결정할 수 있다. 향후의 트래픽 부하 상에 이용가능한 추정치가 없는 경우, 프로세싱 리소스를 비활성화하고 재활성화하는데 필요한 시간 및 트래픽 부하의 예측 불가능한 특성으로 인해 유휴 프로세싱 유닛들에는 용이하게 전력이 공급될 수 없다. 트래픽 부하를 완전히 이용되는 서너 개의 프로세싱 리소스들에 걸쳐 할당함으로써 후속 다운스트림 패킷 프로

세싱 디바이스의 리소스 관리가 또한 용이해진다. 예를 들면, 데이터 패킷들을 스위치 패브릭(30)으로 송신하는 TM들(20)이 자신들의 패킷들을 우선 상위 송신 회선들(51)에 분배하고 상위 송신 회선들이 완전히 이용되는 경우에만 하위 송신 회선들(52)을 이용하도록 구성되면, 하위의 교차-접속 교환 디바이스(32)는 더 용이하고 더 신속하게 스위치 오프될 수 있다.

[0054] 도 7은 본 발명의 실시예의 구현에 적합한 네트워크 프로세서(NP)의 예시적인 아키텍처를 도시한다. 이 실시예에 따른 NP는 데이터 패킷 프로세서(즉, 패킷 엔진 유닛들), 및 결정된 트래픽 레이트 표시 및 프로세싱된 데이터 패킷을 트래픽 매니저로 송신하기 위한 수단을 포함하는데, 상기 트래픽 레이트 표시 및 프로세싱된 데이터 패킷은 제어 프로세서 및 트래픽 매니저로의 인터페이스 내에서 구현된다. NP는 트래픽 미터링 유닛(도시되지 않음)을 추가로 포함한다.

[0055] 도 8은 본 발명의 실시예의 구현에 적합한 트래픽 매니저 TM(20)의 예시적인아키텍처를 도시한다. 예를 들면, 본 발명에 따른 리소스 매니저는 메모리 제어기 및/또는 NP 인터페이스 제어기에 의해 구현될 수 있다. 리소스 매니저는 전형적으로 프로그램 가능 칩들 또는 하드웨어 구성요소들을 이용하여 구성된다.

[0056] 상술한 실시예들의 구성들의 특징들, 구성요소들, 및 특정 세부사항들은 각각의 응용예에 최적화되는 추가 실시예들을 형성하도록 교환 또는 결합될 수 있다. 그러한 수정들이 당업자에게 명백한 한에서, 상기 수정들은 모든 가능한 결합을 명시적으로 지정하지 않고 상기 설명에 의해 암시적으로 드러날 것이다.

부호의 설명

[0057]

1: 라인 카드

2: PHY 디바이스

3: MAC/프레이머 디바이스

5: 패브릭 매니저

10: 인그레스 네트워크 프로세서

20: 트래픽 매니저

22: 메모리 모듈들

23: 메모리 뱅크들

23b: 인터페이스 제어기

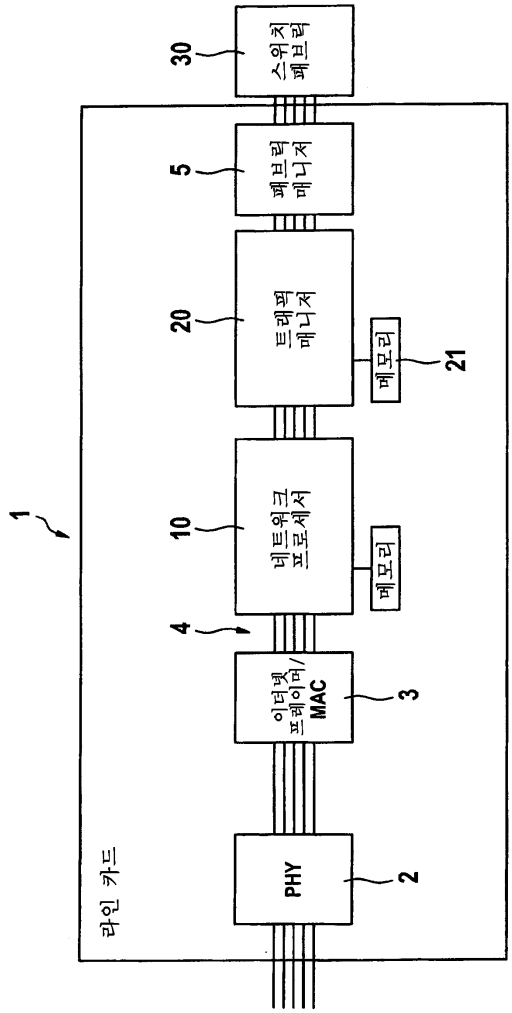
30: 스위치 패브릭 디바이스

31: 리소스 매니저

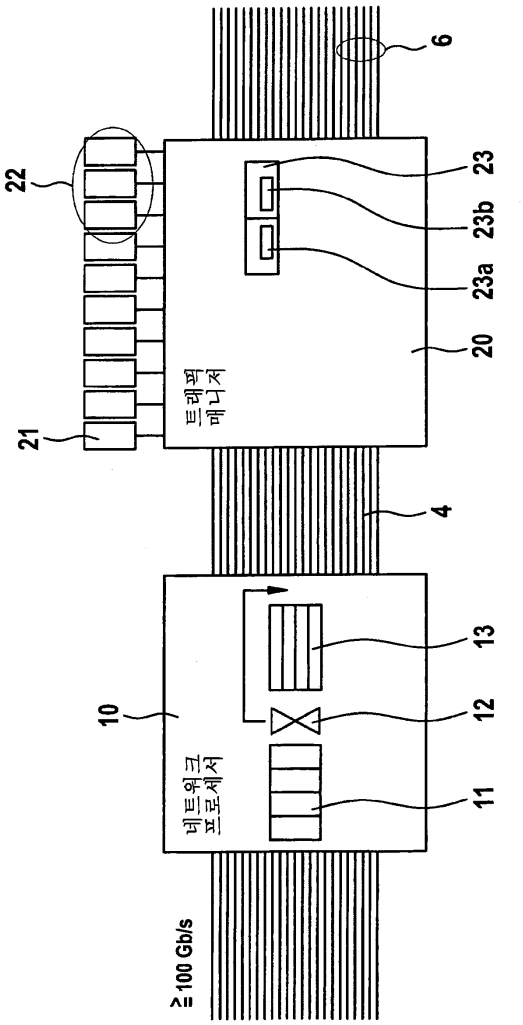
32: 교환 디바이스

도면

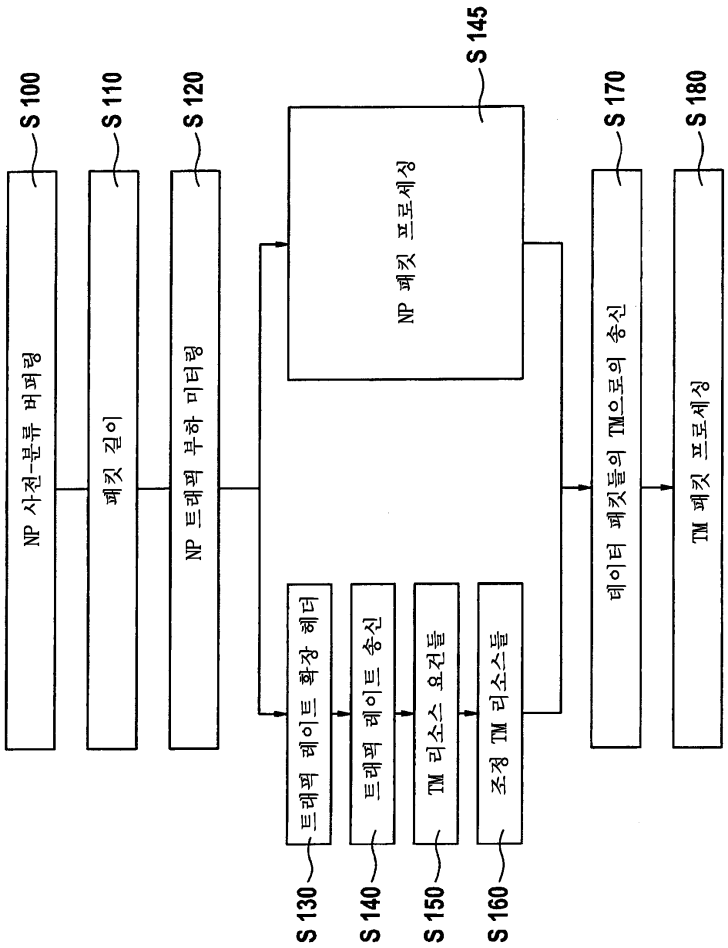
도면1


1
라인 카드
PHY
2
이더넷 프레이머/ MAC
3
4
네트워크 프로세서
10
메모리
트래픽 매니저
20
메모리
21
패브릭 매니저
5
스위치 패브릭
30

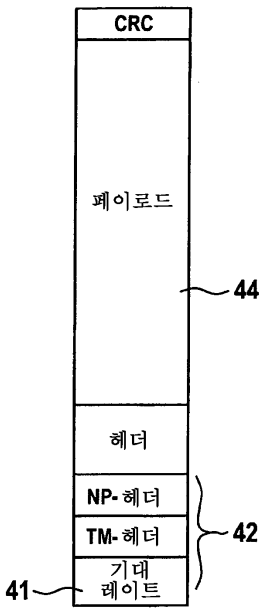
도면2



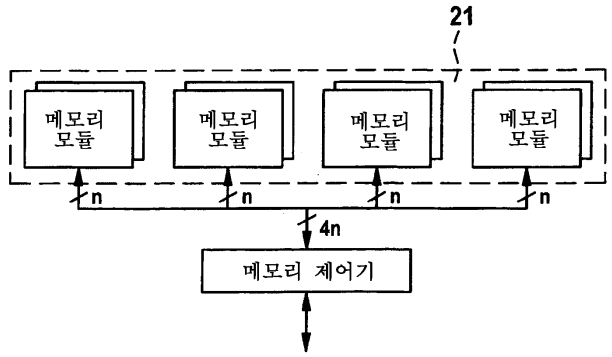
도면3


S 100
NP 사전-분류 버퍼링
S 110
패킷 길이
S 120
NP 트래픽 부하 미터링
S 145
NP 패킷 프로세싱
S 130
트래픽 레이트 확장 헤더
S 140
트래픽 레이트 송신
S 150
TM 리소스 요건들
S 160
조정 TM 리소스들
S 170
데이터 패킷들의 TM으로의 송신
S 180
TM 패킷 프로세싱

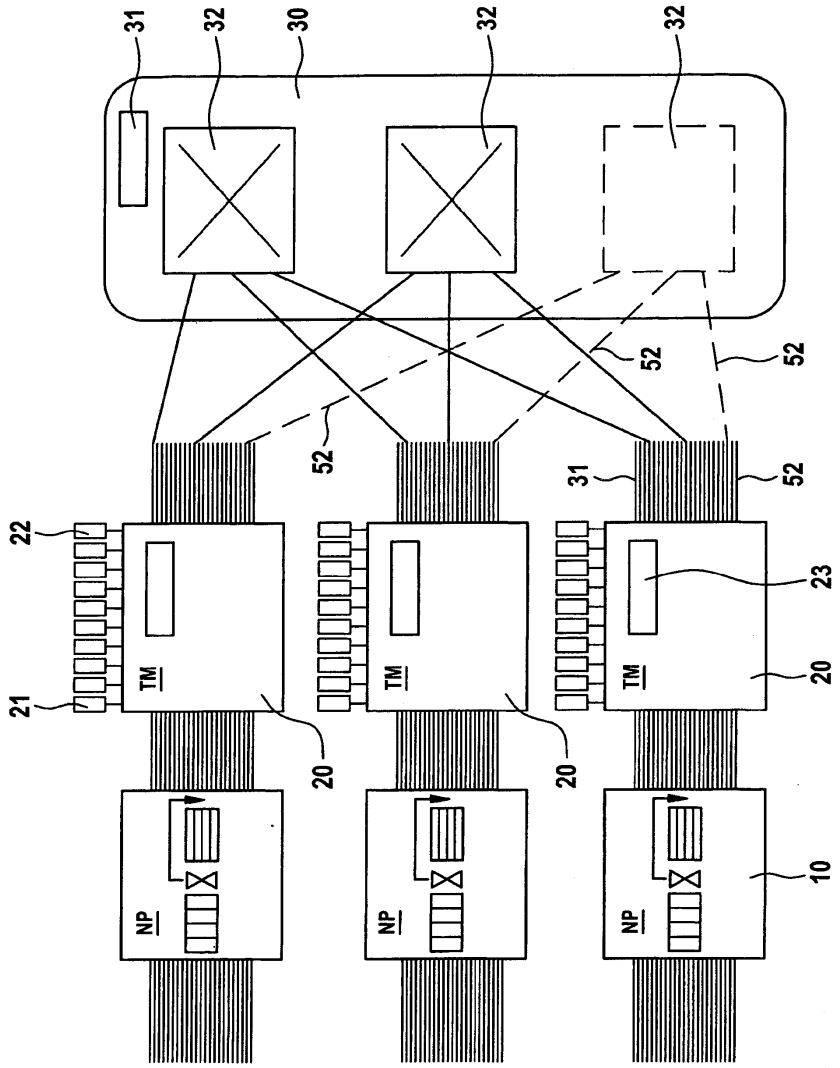
도면4


CRC
페이로드
44
헤더
NP-헤더
TM-헤더
42
기대 레이트
41

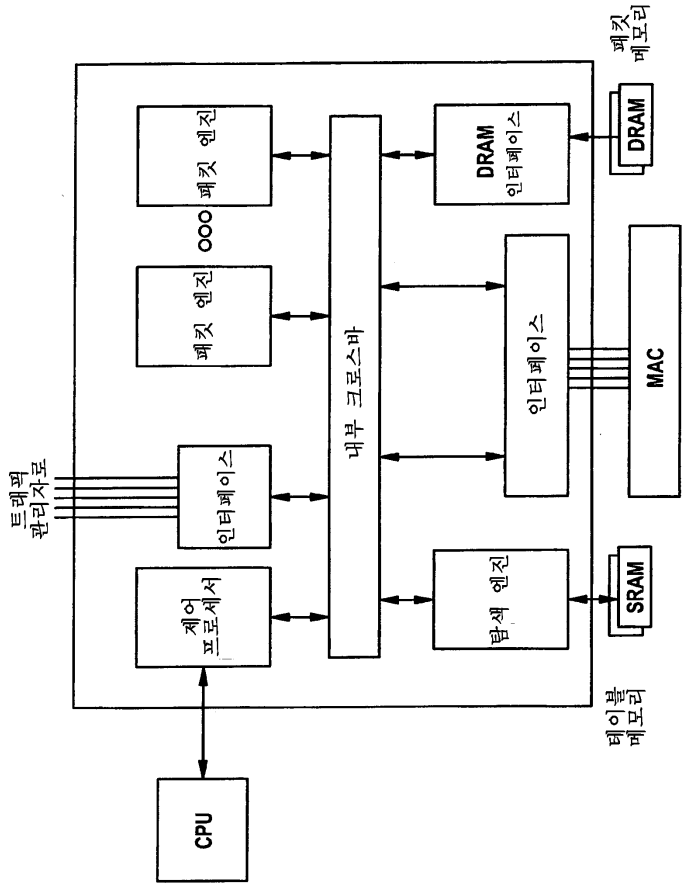
도면5


21
메모리 모듈
메모리 모듈
메모리 모듈
메모리 모듈
n
n
n
n
4n
메모리 제어기

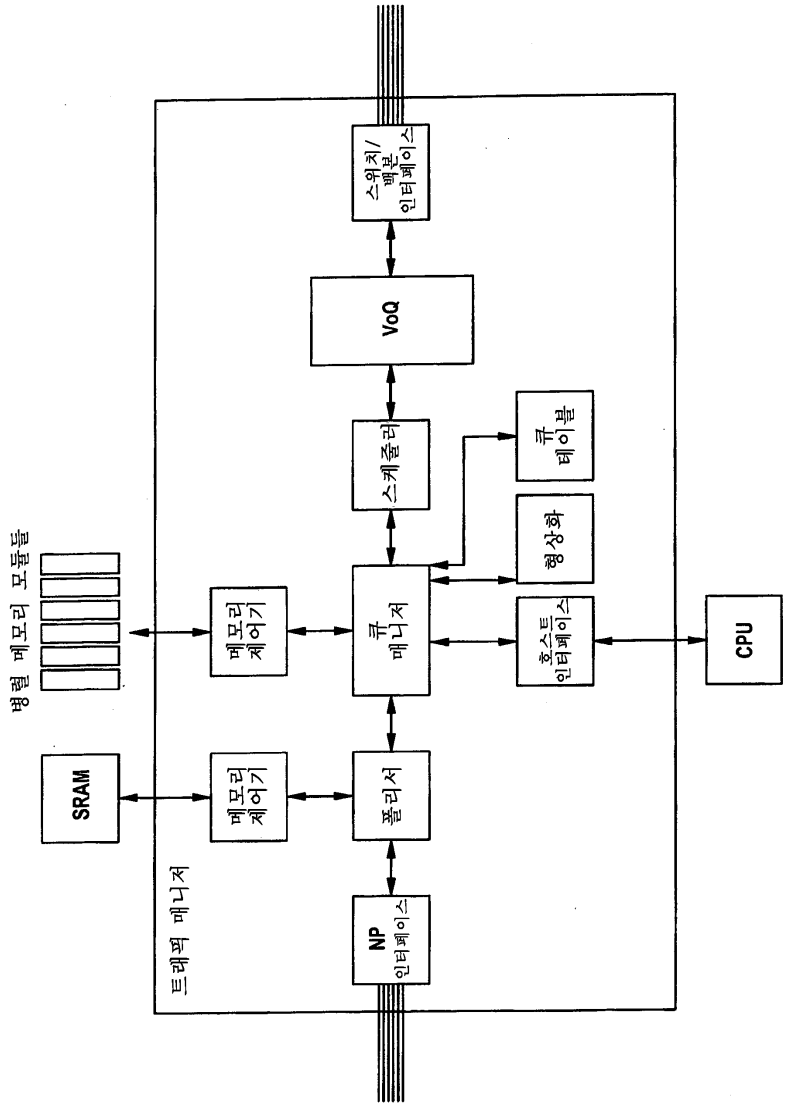
도면6


31
32
30
32
32
52
52
52
31
52
22
21
TM
TM
TM
23
20
20
20
NP
NP
NP
10

도면7


패킷 메모리
DRAM
DRAM 인터페이스
패킷 엔진
OOO
패킷 엔진
내부 크로스바
인터페이스
MAC
트래픽 관리자로
인터페이스
제어 프로세서
탐색 엔진
SRAM
테이블 메모리
CPU

도면8


병렬 메모리 모듈들
SRAM
메모리 제어기
메모리 제어기
큐 매니저
폴리셔
스케줄러
VoQ
스위치/백본 인터페이스
NP 인터페이스
큐 테이블
형상화
호스트 인터페이스
CPU
트래픽 매니저