(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년10월11일
(11) 등록번호 10-2715528
(24) 등록일자 2024년10월04일

(51) 국제특허분류(Int. Cl.)
H04L 49/00 (2022.01) *G06F 12/06* (2006.01)
G06F 12/0815 (2016.01) *H04J 14/02* (2006.01)
H04Q 11/00 (2006.01)
(52) CPC특허분류
H04L 49/9078 (2013.01)
G06F 12/0607 (2013.01)
(21) 출원번호 10-2021-0153138
(22) 출원일자 2021년11월09일
심사청구일자 2022년06월09일
(65) 공개번호 10-2023-0067254
(43) 공개일자 2023년05월16일
(56) 선행기술조사문헌
KR1020060062576 A*
KR1020200066893 A*
KR1020210132348 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
송종태
대전광역시 유성구 가정로 218 (가정동)
김대업
대전광역시 유성구 가정로 218 (가정동)
(뒷면에 계속)
(74) 대리인
특허법인 무한

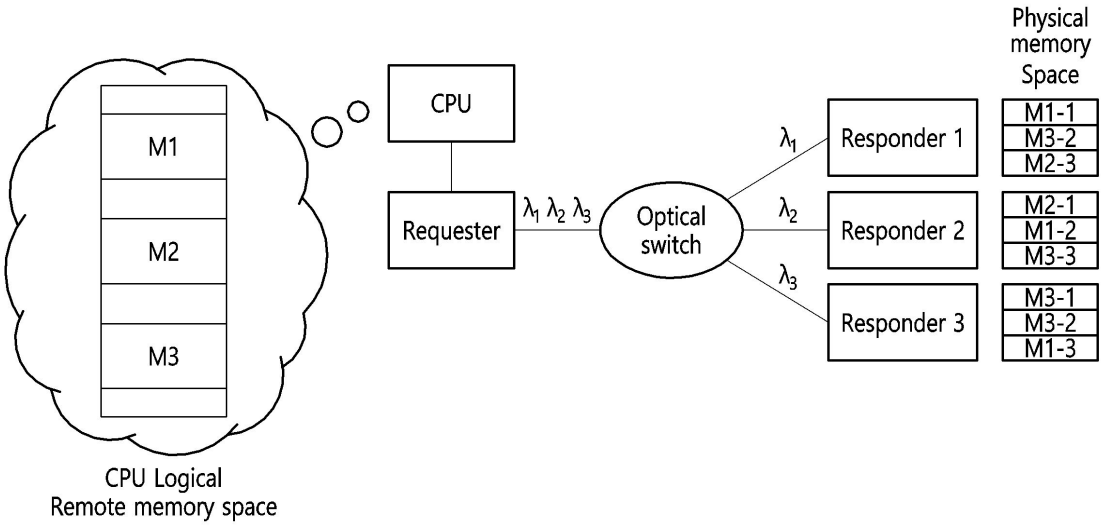
전체 청구항 수 : 총 16 항 심사관 : 윤태섭

(54) 발명의 명칭 **메모리 액세스 방법 및 이를 수행하는 서버**

(57) 요 약

메모리 액세스 방법 및 이를 수행하는 서버가 개시된다. 서버를 구성하는 광인터리버가 수행하는 메모리 액세스 방법은 상기 서버를 구성하는 요청자 처리 엔진(Requester Processing Engine)으로부터 요청 메시지를 수신하는 단계; 상기 서버와 연결된 외부 저장장치들의 개수에 대응하여 서로 다른 파장에 대응하는 수신 버퍼들을 설정하는 단계; 상기 서로 다른 파장에 동일한 요청 메시지를 파장 분할 다중화(Wavelength Division Multiplexing, WDM) 방식에 따라 다중화 하는 단계; 및 상기 다중화된 요청 메시지를 상기 외부 저장장치들 각각으로 전송하는 단계를 포함하고, 상기 서버에서 관리하는 가상 메모리의 주소는 상기 외부 저장장치들 각각에 포함된 응답자(Responder)에 의해 인터리빙(Interleaving) 방식에 따라 분리되어 저장될 수 있다.

대 표 도



(52) CPC특허분류
G06F 12/0815 (2013.01)
H04J 14/0254 (2023.08)
H04L 49/9084 (2013.01)
H04Q 11/0001 (2013.01)
(72) 발명자
윤지욱
대전광역시 유성구 가정로 218 (가정동)
한경은
대전광역시 유성구 가정로 218 (가정동)
이준기
대전광역시 유성구 가정로 218 (가정동)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711125992
과제번호	2019-0-00002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	ETRI연구개발지원사업
연구과제명	[전문연구실/통합과제]광 클라우드 네트워킹 핵심기술 개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

서버를 구성하는 광인터리버가 수행하는 메모리 액세스 방법에 있어서,

상기 서버를 구성하는 요청자 처리 엔진(Requester Processing Engine)으로부터 타겟 가상 메모리에 대한 읽기 작업 또는 쓰기 작업에 대한 요청 메시지를 수신하는 단계;

상기 서버와 연결된 외부 저장장치들의 개수와 동일한 개수의 수신 버퍼들을 서로 다른 파장에 대응하여 설정하는 단계;

상기 서로 다른 파장에 대응하여 설정된 수신 버퍼들 각각에 상기 수신된 요청 메시지를 저장하는 단계;

상기 수신 버퍼들 각각에 저장된 요청 메시지를 파장 분할 다중화(Wavelength Division Multiplexing, WDM) 방식에 따라 다중화 하는 단계; 및

상기 다중화된 요청 메시지를 상기 외부 저장장치들 각각으로 전송하는 단계

를 포함하고,

상기 서버에서 관리하는 가상 메모리의 주소는,

상기 외부 저장장치들 각각에 포함된 응답자(Responder)에 의해 인터리빙(Interleaving) 방식에 따라 분리되어 저장되는 메모리 액세스 방법.

청구항 2

제1항에 있어서,

상기 서버와 외부 저장장치들은,

광스위치를 통해 서로 연결되는 메모리 액세스 방법.

청구항 3

제1항에 있어서,

상기 요청자 처리 엔진은,

상기 외부 저장장치들 각각에 포함된 응답자와 요청 메시지를 교환하여 상기 외부 저장장치들 각각에 포함된 리모트 메모리에 읽기 작업 또는 쓰기 작업이 수행되도록 하는 메모리 액세스 방법.

청구항 4

제1항에 있어서,

상기 서버는,

상기 서버에 포함된 로컬 메모리와 상기 외부 저장장치들 각각에 포함된 리모트 메모리 간의 캐시 일관성을 제공하기 위한 코히어런스 패브릭(Coherence Fabric)을 더 포함하는 메모리 액세스 방법.

청구항 5

제1항에 있어서,

상기 서버는,

상기 외부 저장장치들 각각에 포함된 리모트 메모리의 주소체계를 구성하여 데이터의 캐시공유를 위한 기능을 지원하는 홈에이전트(Home Agent)를 더 포함하는 메모리 액세스 방법.

청구항 6

서버를 구성하는 광인터리버가 수행하는 메모리 액세스 방법에 있어서,

상기 서버와 연결된 외부 저장장치들로부터 타겟 가상 메모리에 대한 읽기 작업 또는 쓰기 작업에 대한 요청 메시지에 대응하는 응답 메시지를 수신하는 단계;

상기 외부 저장장치들의 개수와 동일한 개수의 서로 다른 파장에 대응하는 수신 버퍼들을 식별하는 단계;

상기 응답 메시지를 전송하는데 사용된 파장에 기초하여 상기 응답 메시지를 상기 서로 다른 파장에 대응하여 식별된 수신 버퍼들에 저장하는 단계; 및

상기 수신 버퍼들에 상기 응답 메시지가 모두 채워진 경우, 상기 수신 버퍼들에 저장된 응답 메시지들을 상기 서버를 구성하는 요청자 처리 엔진으로 전달하는 단계

를 포함하고,

상기 서버에서 관리하는 가상 메모리의 주소는,

상기 외부 저장장치들 각각에 포함된 응답자(Responder)에 의해 인터리빙(Interleaving) 방식에 따라 분리되어 저장되는 메모리 액세스 방법.

청구항 7

제6항에 있어서,

상기 서버와 외부 저장장치들은,

광스위치를 통해 서로 연결되는 메모리 액세스 방법.

청구항 8

제6항에 있어서,

상기 요청자 처리 엔진은,

상기 외부 저장장치들 각각에 포함된 응답자와 요청 메시지를 교환하여 상기 외부 저장장치들 각각에 포함된 리모트 메모리에 읽기 작업 또는 쓰기 작업이 수행되도록 하는 메모리 액세스 방법.

청구항 9

제6항에 있어서,

상기 서버는,

상기 서버에 포함된 로컬 메모리와 상기 외부 저장장치들 각각에 포함된 리모트 메모리 간의 캐시 일관성을 제공하기 위한 코히어런스 패브릭(Coherence Fabric)을 더 포함하는 메모리 액세스 방법.

청구항 10

제6항에 있어서,

상기 서버는,

상기 외부 저장장치들 각각에 포함된 리모트 메모리의 주소체계를 구성하여 데이터의 캐시공유를 위한 기능을 지원하는 홈에이전트(Home Agent)를 더 포함하는 메모리 액세스 방법.

청구항 11

메모리 액세스 방법을 수행하는 서버에 있어서,

프로세서를 포함하고,

상기 프로세서는,

상기 서버를 구성하는 요청자 처리 엔진(Requester Processing Engine)으로부터 타겟 가상 메모리에 대한 읽기

작업 또는 쓰기 작업에 대한 요청 메시지를 수신하고, 상기 서버와 연결된 외부 저장장치들의 개수와 동일한 개수의 수신 버퍼들을 서로 다른 파장에 대응하여 설정하며, 상기 서로 다른 파장에 대응하여 설정된 수신 버퍼들 각각에 상기 수신된 요청 메시지를 저장하고, 상기 수신 버퍼들 각각에 저장된 요청 메시지를 파장 분할 다중화(Wavelength Division Multiplexing, WDM) 방식에 따라 다중화 하고, 상기 다중화된 요청 메시지를 상기 외부 저장장치들 각각으로 전송하며, 상기 서버에서 관리하는 가상 메모리의 주소는 상기 외부 저장장치들 각각에 포함된 응답자(Responder)에 의해 인터리빙(Interleaving) 방식에 따라 분리되어 저장되는 서버.

청구항 12

제11항에 있어서,

상기 프로세서는,

상기 요청 메시지에 대응하여 외부 저장장치부터 응답 메시지가 수신된 경우, 상기 수신된 응답 데이터를 파장에 따라 구분된 수신 버퍼에 저장하고, 상기 수신 버퍼가 응답 메시지가 모두 채워진 경우, 상기 수신 버퍼에 저장된 응답 메시지들을 상기 요청자 처리 엔진으로 전달하는 서버.

청구항 13

제11항에 있어서,

상기 서버와 외부 저장장치들은,

광스위치를 통해 서로 연결되는 서버.

청구항 14

제11항에 있어서,

상기 요청자 처리 엔진은,

상기 외부 저장장치들 각각에 포함된 응답자와 요청 메시지를 교환하여 상기 외부 저장장치들 각각에 포함된 리모트 메모리에 읽기 작업 또는 쓰기 작업이 수행되도록 하는 서버.

청구항 15

제11항에 있어서,

상기 서버는,

상기 서버에 포함된 로컬 메모리와 상기 외부 저장장치들 각각에 포함된 리모트 메모리 간의 캐시 일관성을 제공하기 위한 코히어런스 패브릭(Coherence Fabric)을 더 포함하는 서버.

청구항 16

제11항에 있어서,

상기 서버는,

상기 외부 저장장치들 각각에 포함된 리모트 메모리의 주소체계를 구성하여 데이터의 캐시공유를 위한 기능을 지원하는 홈에이전트(Home Agent)를 더 포함하는 서버.

발명의 설명

기 술 분 야

[0001] 본 발명은 메모리 액세스 방법 및 이를 수행하는 서버에 관한 것으로, 보다 구체적으로는 광스위치를 사용한 인터리브(Interleave) 방식의 광대역 메모리 액세스 기술에 관한 것이다.

배 경 기 술

[0002] 5G를 기점으로 클라우드 서비스 대중화가 예상되며 기존 클라우드 기능도 단순한 데이터 제공 서비스에서 점차 다양한 초저지연 AI 서비스로 발전될 것으로 예상된다. 클라우드 컴퓨팅 구조는 중앙처리장치(Central

Processing Unit, 이하 CPU) 성능이 중요시되는 호모지니어스 컴퓨팅(Homogeneous computing) 구조에서 특화된 엔진간 빠른 데이터교환이 중요한 헤테로지니어스 컴퓨팅(Heterogeneous computing)으로 진화하고 있다. 이와 같은 헤테로지니어스 컴퓨팅으로의 진화에 따라 데이터센터 네트워크는 서버를 연결하는 구조에서 CPU, 메모리, 액셀러레이터 (Accelerator), 스토리지 등 클라우드 자원을 네트워크 상에서 연결하는 디스어그리게이션(Disaggregation) 기술로 발전하고 있다.

[0003] 이러한 컴퓨팅자원의 디스어그리게이션 기술의 이슈는 자원을 네트워크 상에 분산배치 하면서도 자원 간 연결을 빠르게 하여 응용의 성능이 저하되지 않게 하는 것이다. 궁극적으로 자원간 연결 지연과 대역폭을 물리적으로 동일 서버 수준으로 제공하여야 하는게 목표인데 액셀러레이터와 스토리지의 경우 전기스위치를 사용하여 자원 풀을 구성하여도 요구되는 지연 및 대역폭을 만족 수 있어 성능저하가 발생하지 않는다.

[0004] 그러나 메모리의 경우 지연/대역폭은 성능저하를 최소화하기 위해 1 μs 지연과 100 Gbps 대역폭 보장이 필요한데 이는 전기스위치로는 해결이 불가능하여 광스위치 적용이 필수적이다. 특히 향후 수 Tbps 의 대역폭이 요구되는 HBM(High Bandwidth Memory)을 지원하기 위해서는 광기반으로 연결이 필요할 것으로 예상된다.

[0005] 또한 현재 데이터센터의 메모리와 CPU의 사용율이 작다. 실제 데이터센터의 메모리와 CPU 이용율을 측정한 결과 서버별로 메모리/CPU 이용 비율이 크게는 1000배까지 차이가 나며 2%의 응용이 98%의 자원을 사용하는 자원 이용 불균형이 심각한 실정이며 이에 따라 현재 데이터센터의 자원이용율은 40% 수준에 머물고 있다. 따라서, 이러한 자원 이용 불균형을 해소하기 위한 방법이 요구되고 있다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 데이터센터 내 서버의 메모리 자원을 인터리브 방식을 이용하여 네트워크 상에 분산 배치함으로써 광대역의 메모리 액세스 기술을 제공하는 시스템 및 방법을 제공한다.

[0007] 또한, 본 발명은 서버의 CPU가 네트워크 상에 분산 배치된 메모리 자원들과 광스위치를 사용하여 액세스함으로써 네트워크 스위칭 지연을 최소화할 수 있다.

과제의 해결 수단

[0008] 본 발명의 일실시예에 따른 서버를 구성하는 광인터리버가 수행하는 메모리 액세스 방법은 상기 서버를 구성하는 요청자 처리 엔진(Requester Processing Engine)으로부터 요청 메시지를 수신하는 단계; 상기 서버와 연결된 외부 저장장치들의 개수에 대응하여 서로 다른 파장에 대응하는 수신 버퍼들을 설정하는 단계; 상기 서로 다른 파장에 동일한 요청 메시지를 파장 분할 다중화(Wavelength Division Multiplexing, WDM) 방식에 따라 다중화하는 단계; 및 상기 다중화된 요청 메시지를 상기 외부 저장장치들 각각으로 전송하는 단계를 포함하고, 상기 서버에서 관리하는 가상 메모리의 주소는 상기 외부 저장장치들 각각에 포함된 응답자(Responder)에 의해 인터리빙(Interleaving) 방식에 따라 분리되어 저장될 수 있다.

[0009] 상기 서버와 외부 저장장치들은 광스위치를 통해 서로 연결될 수 있다.

[0010] 상기 요청자 처리 엔진은 상기 외부 저장장치들 각각에 포함된 응답자와 요청 메시지를 교환하여 상기 외부 저장장치들 각각에 포함된 리모트 메모리에 읽기 작업 또는 쓰기 작업이 수행되도록 할 수 있다.

[0011] 상기 서버는 상기 서버에 포함된 로컬 메모리와 상기 외부 저장장치들 각각에 포함된 리모트 메모리 간의 캐시 일관성을 제공하기 위한 코히어런스 패브릭(Coherence Fabric)을 더 포함할 수 있다.

[0012] 상기 서버는 상기 외부 저장장치들 각각에 포함된 리모트 메모리에 데이터의 캐시공유를 위한 기능을 제공하는 홈에이전트(Home Agent)를 더 포함할 수 있다.

[0013] 본 발명의 일실시예에 따른 서버를 구성하는 광인터리버가 수행하는 메모리 액세스 방법은 상기 서버와 연결된 외부 저장장치들로부터 응답 메시지를 수신하는 단계; 상기 응답 메시지를 전송하는데 사용된 파장에 기초하여 상기 응답 메시지를 파장에 따라 구분된 수신 버퍼에 저장하는 단계; 및 상기 수신 버퍼에 응답 메시지가 모두 채워진 경우, 상기 수신 버퍼에 저장된 응답 메시지들을 상기 서버를 구성하는 요청자 처리 엔진으로 전달하는 단계를 포함하고, 상기 서버에서 관리하는 가상 메모리의 주소는 상기 외부 저장장치들 각각에 포함된 응답자(Responder)에 의해 인터리빙(Interleaving) 방식에 따라 분리되어 저장될 수 있다.

[0014] 상기 서버와 외부 저장장치들은 광스위치를 통해 서로 연결될 수 있다.

[0015] 상기 요청자 처리 엔진은 상기 외부 저장장치들 각각에 포함된 응답자와 요청 메시지를 교환하여 상기 외부 저장장치들 각각에 포함된 리모트 메모리에 읽기 작업 또는 쓰기 작업이 수행되도록 할 수 있다.

[0016] 상기 서버는 상기 서버에 포함된 로컬 메모리와 상기 외부 저장장치들 각각에 포함된 리모트 메모리 간의 캐시 일관성을 제공하기 위한 코히어런스 패브릭(Coherence Fabric)을 더 포함할 수 있다.

[0017] 상기 서버는 상기 외부 저장장치들 각각에 포함된 리모트 메모리에 데이터의 캐시공유를 위한 기능을 제공하는 홈에이전트(Home Agent)를 더 포함할 수 있다.

[0018] 본 발명의 일실시예에 따른 메모리 액세스 방법을 수행하는 서버는 프로세서를 포함하고, 상기 프로세서는 상기 서버를 구성하는 요청자 처리 엔진(Requester Processing Engine)으로부터 요청 메시지를 수신하고, 상기 서버와 연결된 외부 저장장치들의 개수에 대응하여 서로 다른 파장에 대응하는 수신 버퍼들을 설정하며, 상기 서로 다른 파장에 동일한 요청 메시지를 파장 분할 다중화(Wavelength Division Multiplexing, WDM) 방식에 따라 다중화 하고, 상기 다중화된 요청 메시지를 상기 외부 저장장치들 각각으로 전송하며, 상기 서버에서 관리하는 가상 메모리의 주소는 상기 외부 저장장치들 각각에 포함된 응답자(Responder)에 의해 인터리빙(Interleaving) 방식에 따라 분리되어 저장될 수 있다.

[0019] 상기 프로세서는 상기 리퀘스트 메시지에 대응하여 외부 저장장치부터 응답 메시지가 수신된 경우, 상기 수신된 응답 데이터를 파장에 따라 구분된 수신 버퍼에 저장하고, 상기 수신 버퍼가 응답 메시지가 모두 채워진 경우, 상기 수신 버퍼에 저장된 응답 메시지들을 상기 요청자 처리 엔진으로 전달할 수 있다.

[0020] 상기 서버와 외부 저장장치들은 광스위치를 통해 서로 연결될 수 있다.

[0021] 상기 요청자 처리 엔진은 상기 외부 저장장치들 각각에 포함된 응답자와 요청 메시지를 교환하여 상기 외부 저장장치들 각각에 포함된 리모트 메모리에 읽기 작업 또는 쓰기 작업이 수행되도록 할 수 있다.

[0022] 상기 서버는 상기 서버에 포함된 로컬 메모리와 상기 외부 저장장치들 각각에 포함된 리모트 메모리 간의 캐시 일관성을 제공하기 위한 코히어런스 패브릭(Coherence Fabric)을 더 포함할 수 있다.

[0023] 상기 서버는 상기 외부 저장장치들 각각에 포함된 리모트 메모리에 데이터의 캐시공유를 위한 기능을 제공하는 홈에이전트(Home Agent)를 더 포함할 수 있다.

발명의 효과

[0024] 본 발명의 일실시예에 의하면, 데이터센터 내 서버의 메모리 자원을 인터리브 방식을 이용하여 네트워크 상에 분산 배치함으로써 광대역의 메모리 액세스 기술을 제공할 수 있다.

[0025] 또한, 본 발명의 일실시예에 의하면, 서버의 CPU가 네트워크 상에 분산 배치된 메모리 자원들과 광스위치를 사용하여 액세스함으로써 네트워크 스위칭 지연을 최소화할 수 있다.

도면의 간단한 설명

[0026] 도 1은 본 발명의 일실시예에 따른 메모리 액세스 시스템의 개요를 나타낸 도면이다.

도 2는 본 발명의 일실시예에 따른 메모리 액세스 시스템의 구성을 나타낸 도면이다.

도 3a은 종래 기술에 따른 메모리 액세스 과정을 나타낸 도면이다.

도 3b는 본 발명의 일실시예에 따른 메모리 액세스 과정을 나타낸 도면이다.

도 4는 본 발명의 일실시예에 따른 메모리 액세스 시스템의 구성도를 나타낸 도면이다.

도 5는 본 발명의 일실시예에 따른 광인터리버의 읽기 요청 전송 동작을 나타낸 도면이다.

도 6는 본 발명의 일실시예에 따른 광인터리버의 응답 메시지 수신 동작을 나타낸 도면이다.

도 7는 본 발명의 일실시예에 따른 광인터리버의 쓰기 요청 전송 동작을 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0027] 이하, 본 발명의 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0028] 도 1은 본 발명의 일실시예에 따른 메모리 액세스 시스템의 개요를 나타낸 도면이다.

[0029] 본 발명은 데이터센터 내의 서버가 사용하는 메모리를 동일 서버내의 메모리버스에 배치하여야 하는 물리적 한계를 해결하는 방법을 제공한다. 이를 위해 본 발명의 메모리 액세스 시스템(100)은 도 1과 같이 리모트 메모리를 포함하는 외부 저장장치들(110~130)을 네트워크 상에 분산 배치하고, 이를 공유하기 위한 메모리 디스어그리게이션(Memory Disaggregation) 기술을 제공할 수 있다.

[0030] 그러나, 외부 저장장치들(110~130)에 포함된 리모트 메모리의 경우 대용량 트래픽을 병렬 고속으로 교환하기 때문에 네트워크의 링크를 통해 상대적으로 작은 대역폭을 통해 액세스 할 경우 서버(140~150) 내에 포함된 로컬 메모리와 같은 성능을 보장할 수 없다. 예를 들어, 고대역폭 메모리(High Bandwidth Memory 2, HBM2) 경우 2Gbps의 대역폭을 1024개 병렬 연결함으로써 총 2Tbps의 메모리 액세스 대역폭을 제공한다. 하지만 이러한 광대역 메모리 액세스는 고집적화 된 칩렛(Chiplet) 형태로 가능하며 네트워크를 통해 스위칭이 될 경우 병렬연결을 통해 광대역 메모리 액세스를 제공하기 어려운 문제가 있다.

[0031] 본 발명에서는 이러한 한계를 극복하기 위해 광스위치를 사용함으로써 외부 저장장치들(110~130)에 포함된 로컬 메모리와 서버(140~150)에 포함된 CPU 간을 고속 광대역으로 연결할 수 있다.

[0033] 도 2는 본 발명의 일실시예에 따른 메모리 액세스 시스템의 구성을 나타낸 도면이다.

[0034] 서버 내의 CPU는 특정 응용프로그램을 실행할 수 있으며, 해당 특정 응용프로그램을 실행하기 위하여 물리적으로 분리되어 분산 배치된 외부 저장장치들의 리모트 메모리를 사용할 수 있다. 이때, 분산 배치된 외부 저장장치들에 포함된 리모트 메모리는 CPU의 관리하에서 가상 주소체계상의 연속된 주소로 인식될 수 있다.

[0035] 이와 같은 리모트 메모리의 물리적 위치는 복수의 외부 저장장치들에 분산되어 저장될 수 있다. 일례로, 도 2를 참고하면, 가상 메모리 M1 은 가상 주소 체계상에서는 연속된 주소를 갖지만 물리적으로는 M1-1, M1-2, M1-3의 형태로 인터리브(Interleave) 되어 외부 저장장치들 각각에 포함된 응답자들(Responder 1,2,3)에 분리되어 저장될 수 있다.

[0036] 특정 응용프로그램이 리모트 메모리에 대해 읽기와 쓰기를 수행할 경우 서버 내의 CPU는 리모트 메모리를 연속된 메모리로 인식하여 읽기 쓰기를 요청할 수 있다. 이때, 읽기 쓰기 요청은 CPU로부터 읽기 쓰기 요청을 받은 요청자(Requester)에 의해 물리적으로 분리된 복수의 외부 저장장치들 내의 응답자(Responder)로 전달될 수 있다.

[0037] 서버 내의 요청자와 외부 저장장치 내의 응답자는 네트워크를 통해 물리적으로는 단일 광링크로 연결되나 파장분할 다중화(Wavelength Division Multiplexing, WDM) 방식을 통해 여러 개의 파장이 다중화되어 단일 광링크를 통해 전달될 수 있다.

[0038] 그러나 도 2와 같이 응답자의 응답속도는 미디어의 특성에 따라 제한될 수 있다. 예를 들어, 현재 DRAM 메모리 컨트롤러(memory controller)가 10 Gbps의 성능을 가질 경우 네트워크에서 이보다 큰 대역폭이 제공되더라도 활용할 수 없다.

[0039] 하지만 도 2와 같이 병렬로 연결된 3개의 응답자에 메모리를 인터리브 형태로 배치하게 될 경우 요청자는 30 Gbps의 메모리 액세스 성능을 얻을 수 있다.

[0041] 도 3a은 종래 기술에 따른 메모리 액세스 과정을 나타낸 도면이다.

[0042] 도 3a는 가상 메모리 M1의 가상 주소가 외부 저장장치에 인터리브 되지 않은 상태에서 메모리를 읽어오는 경우를 나타낸다. 이 경우, 요청자가 가상 메모리 M1에 대한 읽기 요청을 외부 저장 장치 내의 응답자에 전달하고 응답자는 메모리 컨트롤러(Media controller)를 통해 해당 가상 메모리 M1에 대응하는 리모트 메모리로부터 응답 데이터를 읽어올 수 있다.

[0043] 응답자는 이와 같이 가상 메모리 M1에 대응하는 리모트 메모리로부터 읽어온 응답 데이터를 패킷의 형태로 요청자에게 전달할 수 있다. 이때, 응답자는 요청자에게 전달되는 패킷의 크기가 최대 패킷사이즈 이상일 경우 도 3a와 같이 하나의 요청에 대해 여러 개의 응답으로 나누어 전달하게 된다.

[0044] 이로 인해 응답자는 요청자로부터 읽기 요청을 수신한 후 응답 데이터를 요청자에게 전송 완료할 때까지 메모리 컨트롤러 지연(Media Controller Delay)과 데이터 전송 지연(Data Transmission Delay)을 겪게 된다.

[0045] 또한 이와 같은 읽기 요청 및 응답 데이터를 스위칭 할 때 전기 기반의 스위치 패브릭(fabric) 구조를 이용하는

경우, 2번의 스위칭 지연(Switching Delay)이 추가적으로 발생될 수 있다.

[0047] 도 3b는 본 발명의 일실시예에 따른 메모리 액세스 과정을 나타낸 도면이다.

[0048] 본 발명의 메모리 액세스 시스템은 서버 내의 요청자가 동일한 요청 메시지를 서로 다른 외부 저장장치의 응답자에게 동시 전달하는 구조를 제공할 수 있다. 도 3b의 예에서는 응답자가 3개로 설정되어 있으나 이는 하나의 예시일 뿐 응답자의 개수에는 제한이 없을 수 있다.

[0049] 일례로, 요청자는 서로 다른 파장에 3개의 동일한 읽기 요청 메시지를 파장 분할 다중화 방식에 따라 다중화하여 하나의 동일한 광링크를 통해 서로 다른 3개의 응답자에게 각각 전송할 수 있다. 이때, 다중화된 3개의 읽기 요청 메시지는 광스위치를 통해 파장 별로 각기 다른 응답자에게 도착할 수 있다.

[0050] 예를 들어, 광스위치가 AWGR(Arrayed Waveguide Grating Routers)인 경우 파장에 따라 서로 다른 경로로 스위칭 하는 수동 광소자를 이용함으로써 스위칭 지연이 발생하지 않을 수 있다. 도 3에서와 같이 가상 메모리 M1에 대한 읽기 요청은 서로 다른 응답자들(Responder 1, 2, 3)에 전달되고, 응답자들은 가상 메모리 M1-1, M1-2, M1-3의 정보를 동시에 응답 메시지를 통해 요청자에게 전달할 수 있다. 이때 응답자는 읽기 요청이 수신된 파장과 동일한 파장을 이용하여 응답 메시지를 해당 요청자에게 전달할 수 있다.

[0052] 도 4는 본 발명의 일실시예에 따른 메모리 액세스 시스템의 구성도를 나타낸 도면이다.

[0053] 도 4를 참고하면, 하나의 서버는 파장 분할 다중화를 통해 여러 개의 외부 저장장치에 접속하며 서버와 외부 저장장치는 각각 하나의 파장을 통해 연결될 수 잇다.

[0054] 보다 구체적으로 서버는 CPU 코어, 로컬 메모리, 코히어런스 패브릭(Coherence Fabric), 홈에이전트(Home Agent), 요청자 처리 엔진(Requester Processing Engine) 및 광인터리버(Optical Interleaver)로 구성될 수 있다. 이때, 코히어런스 패브릭은 서버에 포함된 로컬 메모리와 외부 저장장치들 각각에 포함된 리모트 메모리 간의 캐시 일관성을 제공할 수 있고, 홈에이전트는 외부 저장장치들 각각에 포함된 리모트 메모리를 하나의 주소 체계로 구성하여 데이터의 캐시공유를 위한 기능을 제공할 수 있다.

[0055] 또한, 요청자 처리 엔진은 외부 저장장치들 각각에 포함된 응답자와 요청 메시지를 교환하여 해당 외부 저장장치들 각각에 포함된 리모트 메모리에 읽기 작업 또는 쓰기 작업이 수행되도록 제어할 수 있고, 광인터리버는 인터리브된 복수의 외부 저장장치들과의 통신을 위해 읽기/쓰기 요청을 파장 분할 다중화 방식을 통해 동시 전송하고, 응답 메시지를 동시 수신할 수 있다. 이때, 홈에이전트 아래에 광인터리버를 배치함으로써 여러 개의 홈에이전트 간 데이터 일관성 유지(Cache Coherence) 부하를 단순화 시킬 수 있다.

[0056] 외부 저장장치는 서버로부터 읽기/쓰기 요청을 수신하여 요청된 메모리 주소에 대응하는 리모트 메모리에 읽기 작업 또는 쓰기 작업을 수행할 수 있다. 보다 구체적으로 외부 저장장치는 응답자, 메모리 컨트롤러 및 리모트 메모리로 구성될 수 있다.

[0057] 먼저, 응답자는 서버로부터 수신된 요청 메시지를 분석하여 읽기 작업 또는 쓰기 작업을 수행할 리모트 메모리를 식별할 수 있다. 이후 응답자는 식별된 리모트 메모리에 대해 메모리 컨트롤러를 통해 읽기 작업 또는 쓰기 작업을 수행할 수 있다. 만약 리모트 메모리에 대해 읽기 작업이 수행되는 경우, 응답자는 해당 읽기 작업을 통해 획득된 데이터를 응답 메시지를 이용하여 요청자에게 전달할 수 있다.

[0059] 도 5는 본 발명의 일실시예에 따른 광인터리버의 읽기 요청 전송 동작을 나타낸 도면이다.

[0060] 본 발명은 N 개의 파장(λ_1, λ_2, λ_3, … λ_N)을 활용한 광인터리버의 동작을 나타낸다. 단계(510)에서, 광인터리버는 서버를 구성하는 요청자 처리 엔진(Requester Processing Engine)으로부터 읽기 요청을 수신할 수 있다.

[0061] 단계(520)에서, 광인터리버는 서버와 연결된 외부 저장장치들의 개수에 대응하여 서로 다른 파장에 대응하는 각각의 수신 버퍼들을 설정할 수 있다. 일례로, 광인터리버는 송신 파장의 수만큼 N 개의 수신 버퍼 (B_1, B_2, B_3, … B_N)를 준비할 수 있다.

[0062] 단계(530)에서, 광인터리버는 서로 다른 파장에 동일한 읽기 요청을 파장 분할 다중화(Wavelength Division Multiplexing, WDM) 방식에 따라 다중화하여 광스위치를 통해 외부 저장장치들 각각으로 전송할 수 있다.

[0064] 도 6는 본 발명의 일실시예에 따른 광인터리버의 응답 메시지 수신 동작을 나타낸 도면이다.

[0065] 단계(610)에서, 광인터리버는 서버와 연결된 외부 저장장치들로부터 응답 메시지를 수신할 수 있다.

[0066] 단계(620)에서, 광인터리버는 응답 메시지를 전송하는데 사용된 파장에 기초하여 해당 응답 메시지를 파장에 따라 구분된 수신 버퍼에 저장할 수 있다. 일례로, 광인터리버는 파장 λ_1을 이용하여 수신된 응답 메시지를 파장 λ_1에 대응하는 설정된 수신 버퍼 B_1에 저장할 수 있다.

[0067] 단계(630)에서, 광인터리버는 수신 버퍼에 응답 메시지가 모두 채워진 경우, 즉, N 개의 수신 버퍼에 응답 메시지가 채워진 것으로 판단되면, 단계(640)에서, 수신 버퍼에 저장된 응답 메시지들을 서버를 구성하는 요청자 처리 엔진으로 전달할 수 있다.

[0069] 도 7는 본 발명의 일실시예에 따른 광인터리버의 쓰기 요청 전송 동작을 나타낸 도면이다.

[0070] 단계(710)에서, 광인터리버는 서버를 구성하는 요청자 처리 엔진(Requester Processing Engine)으로부터 쓰기 요청을 수신할 수 있다.

[0071] 단계(720)에서, 광인터리버는 서버와 연결된 외부 저장장치들의 개수에 대응하여 서로 다른 파장에 대응하는 각각의 송신 버퍼들을 설정할 수 있다. 일례로, 광인터리버는 송신 파장의 수만큼 N 개의 수신 버퍼 (B_1, B_2, B_3, … B_N)를 준비할 수 있다.

[0072] 단계(730)에서, 광인터리버는 설정된 송신 버퍼들에 데이터를 인터리브할 수 있다. 읽기 요청과는 달리 쓰기 요청의 경우 해당 주소에 쓸 데이터를 포함하고 있기 때문에 데이터를 버퍼링하는 단계가 추가로 필요하며 버퍼링된 데이터가 여러 개의 파장으로 전송이 완료된 후 버퍼를 비우게 된다.

[0073] 마지막으로 단계(740)에서, 광인터리버는 서로 다른 파장에 동일한 쓰기 요청을 파장 분할 다중화(Wavelength Division Multiplexing, WDM) 방식에 따라 다중화 하여 광스위치를 통해 외부 저장장치들 각각으로 전송할 수 있다.

[0075] 한편, 본 발명에 따른 방법은 컴퓨터에서 실행될 수 있는 프로그램으로 작성되어 마그네틱 저장매체, 광학적 판독매체, 디지털 저장매체 등 다양한 기록 매체로도 구현될 수 있다.

[0076] 본 명세서에 설명된 각종 기술들의 구현들은 디지털 전자 회로조직으로, 또는 컴퓨터 하드웨어, 펌웨어, 소프트웨어로, 또는 그들의 조합들로 구현될 수 있다. 구현들은 데이터 처리 장치, 예를 들어 프로그램가능 프로세서, 컴퓨터, 또는 다수의 컴퓨터들의 동작에 의한 처리를 위해, 또는 이 동작을 제어하기 위해, 컴퓨터 프로그램 제품, 즉 정보 캐리어, 예를 들어 기계 판독가능 저장 장치(컴퓨터 판독가능 매체) 또는 전파 신호에서 유형적으로 구체화된 컴퓨터 프로그램으로서 구현될 수 있다. 상술한 컴퓨터 프로그램(들)과 같은 컴퓨터 프로그램은 컴파일된 또는 인터프리트된 언어들을 포함하는 임의의 형태의 프로그래밍 언어로 기록될 수 있고, 독립형 프로그램으로서 또는 모듈, 구성요소, 서브루틴, 또는 컴퓨팅 환경에서의 사용에 적절한 다른 유닛으로서 포함하는 임의의 형태로 전개될 수 있다. 컴퓨터 프로그램은 하나의 사이트에서 하나의 컴퓨터 또는 다수의 컴퓨터들 상에서 처리되도록 또는 다수의 사이트들에 걸쳐 분배되고 통신 네트워크에 의해 상호 연결되도록 전개될 수 있다.

[0077] 컴퓨터 프로그램의 처리에 적절한 프로세서들은 예로서, 범용 및 특수 목적 마이크로프로세서들 둘 다, 및 임의의 종류의 디지털 컴퓨터의 임의의 하나 이상의 프로세서들을 포함한다. 일반적으로, 프로세서는 판독 전용 메모리 또는 랜덤 액세스 메모리 또는 둘 다로부터 명령어들 및 데이터를 수신할 것이다. 컴퓨터의 요소들은 명령어들을 실행하는 적어도 하나의 프로세서 및 명령어들 및 데이터를 저장하는 하나 이상의 메모리 장치들을 포함할 수 있다. 일반적으로, 컴퓨터는 데이터를 저장하는 하나 이상의 대량 저장 장치들, 예를 들어 자기, 자기-광 디스크들, 또는 광 디스크들을 포함할 수 있거나, 이것들로부터 데이터를 수신하거나 이것들에 데이터를 송신하거나 또는 양쪽으로 되도록 결합될 수도 있다. 컴퓨터 프로그램 명령어들 및 데이터를 구체화하는데 적절한 정보 캐리어들은 예로서 반도체 메모리 장치들, 예를 들어, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM(Compact Disk Read Only Memory), DVD(Digital Video Disk)와 같은 광 기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-Optical Media), 롬(ROM, Read Only Memory), 램(RAM, Random Access Memory), 플래시 메모리, EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM) 등을 포함한다. 프로세서 및 메모리는 특수 목적 논리 회로조직에 의해 보충되거나, 이에 포함될 수 있다.

[0078] 또한, 컴퓨터 판독가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용매체일 수 있고, 컴퓨터 저장매체 및 전송매체를 모두 포함할 수 있다.

[0079] 본 명세서는 다수의 특정한 구현물의 세부사항들을 포함하지만, 이들은 어떠한 발명이나 청구 가능한 것의 범위

에 대해서도 제한적인 것으로서 이해되어서는 안되며, 오히려 특정한 발명의 특정한 실시형태에 특유할 수 있는 특징들에 대한 설명으로서 이해되어야 한다. 개별적인 실시형태의 문맥에서 본 명세서에 기술된 특정한 특징들은 단일 실시형태에서 조합하여 구현될 수도 있다. 반대로, 단일 실시형태의 문맥에서 기술한 다양한 특징들 역시 개별적으로 혹은 어떠한 적절한 하위 조합으로도 복수의 실시형태에서 구현 가능하다. 나아가, 특징들이 특정한 조합으로 동작하고 초기에 그와 같이 청구된 바와 같이 묘사될 수 있지만, 청구된 조합으로부터의 하나 이상의 특징들은 일부 경우에 그 조합으로부터 배제될 수 있으며, 그 청구된 조합은 하위 조합이나 하위 조합의 변형물로 변경될 수 있다.

[0080] 마찬가지로, 특정한 순서로 도면에서 동작들을 묘사하고 있지만, 이는 바람직한 결과를 얻기 위하여 도시된 그 특정한 순서나 순차적인 순서대로 그러한 동작들을 수행하여야 한다거나 모든 도시된 동작들이 수행되어야 하는 것으로 이해되어서는 안 된다. 특정한 경우, 멀티태스킹과 병렬 프로세싱이 유리할 수 있다. 또한, 상술한 실시형태의 다양한 장치 컴포넌트의 분리는 그러한 분리를 모든 실시형태에서 요구하는 것으로 이해되어서는 안되며, 설명한 프로그램 컴포넌트와 장치들은 일반적으로 단일의 소프트웨어 제품으로 함께 통합되거나 다중 소프트웨어 제품에 패키징 될 수 있다는 점을 이해하여야 한다.

[0081] 한편, 본 명세서와 도면에 개시된 본 발명의 실시 예들은 이해를 돕기 위해 특정 예를 제시한 것에 지나지 않으며, 본 발명의 범위를 한정하고자 하는 것은 아니다. 여기에 개시된 실시 예들 이외에도 본 발명의 기술적 사상에 바탕을 둔 다른 변형 예들이 실시 가능하다는 것은, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 자명한 것이다.

부호의 설명

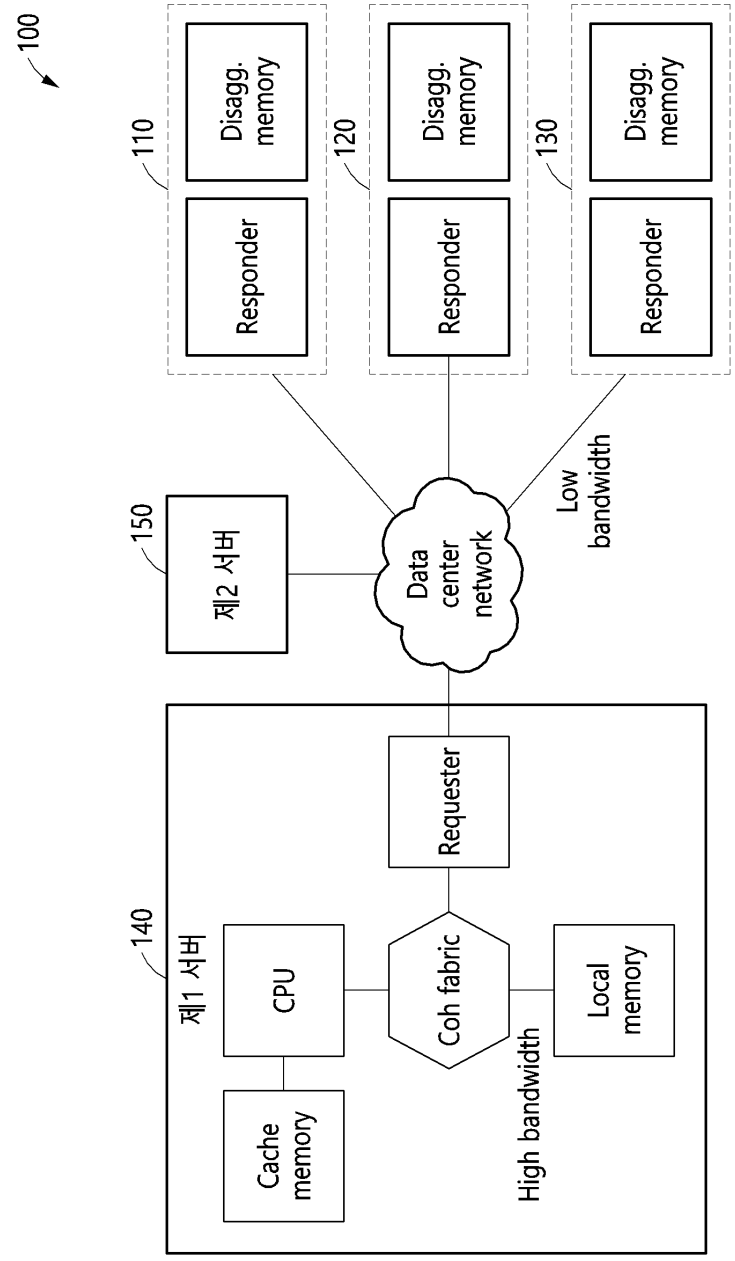
[0082] 100 : 메모리 액세스 시스템

110~130 : 외부 저장장치들

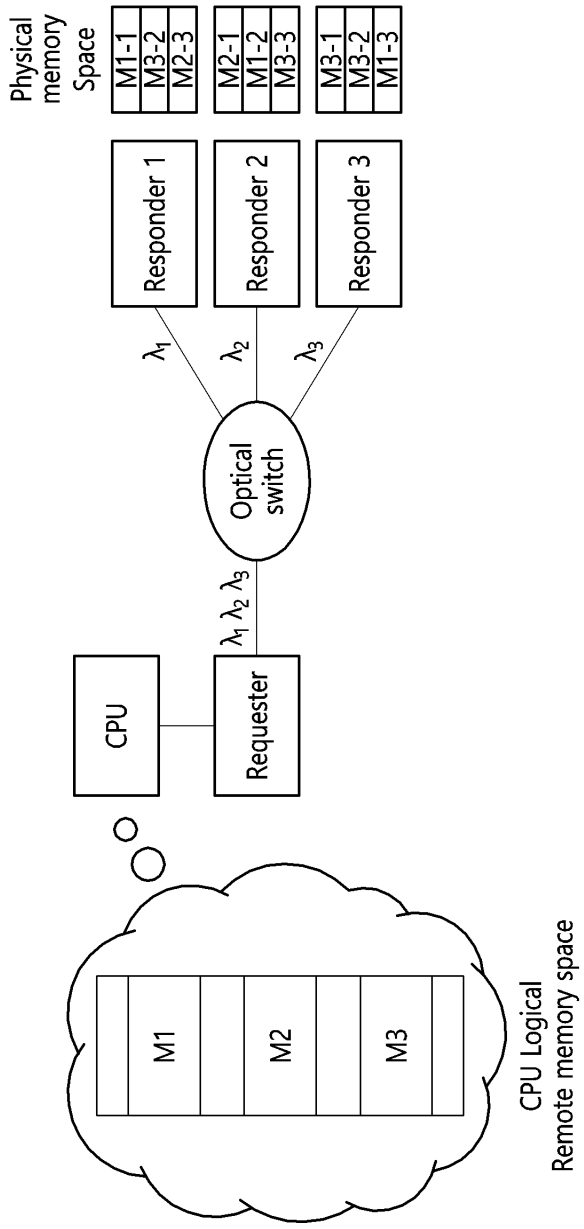
140~150 : 서버

도면

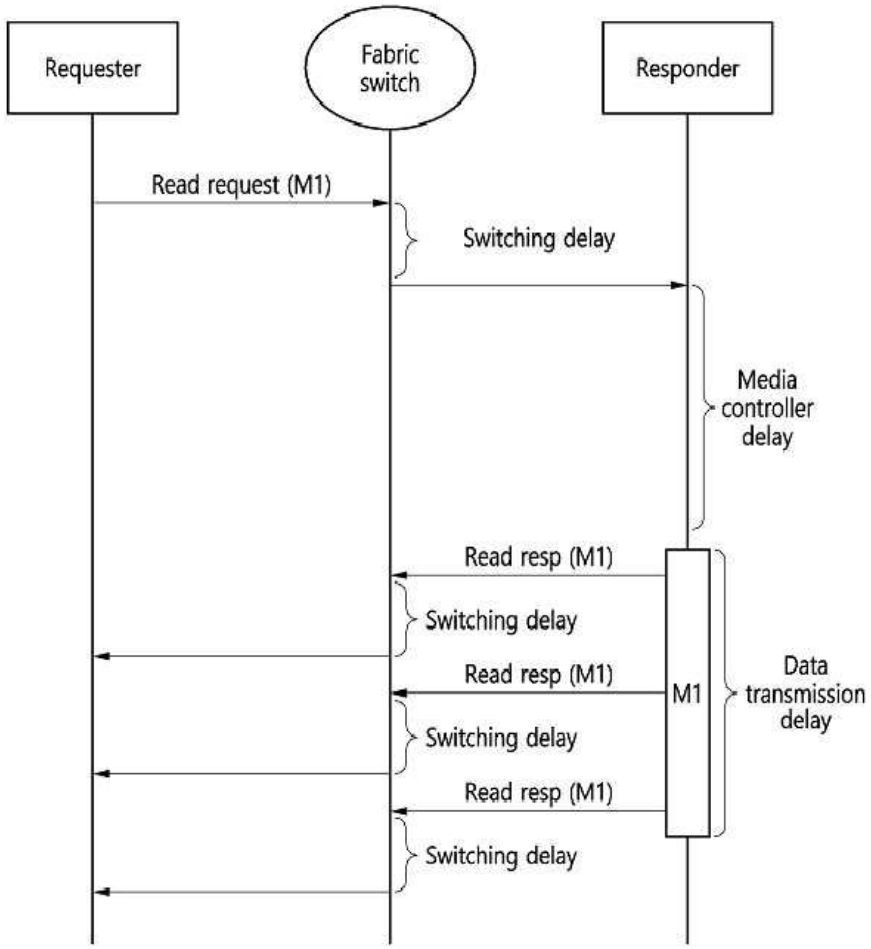
도면1


100
110
Disagg. memory
Responder
120
Disagg. memory
Responder
130
Disagg. memory
Responder
Low bandwidth
150
제2 서버
Data center network
140
제1 서버
Requester
CPU
Coh fabric
Local memory
Cache memory
High bandwidth

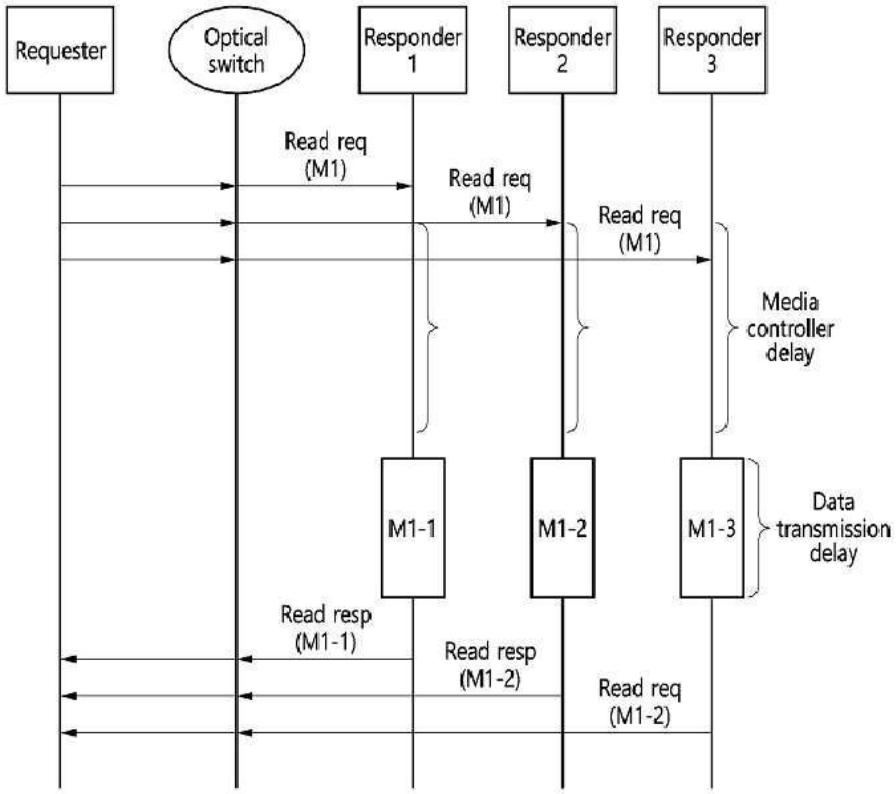
도면2


Physical memory Space
M1-1
M3-2
M2-3
M2-1
M1-2
M3-3
M3-1
M3-2
M1-3
Responder 1
Responder 2
Responder 3
λ1
λ2
λ3
Optical switch
λ1 λ2 λ3
CPU
Requester
M1
M2
M3
CPU Logical
Remote memory space

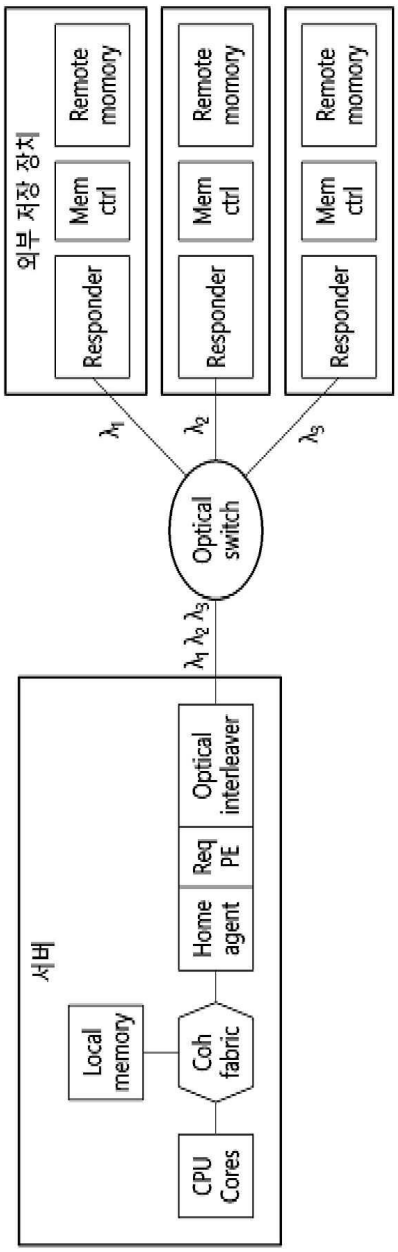
도면3a


Requester
Fabric switch
Responder
Read request (M1)
Switching delay
Media controller delay
Read resp (M1)
Switching delay
Read resp (M1)
M1
Data transmission delay
Switching delay
Read resp (M1)
Switching delay

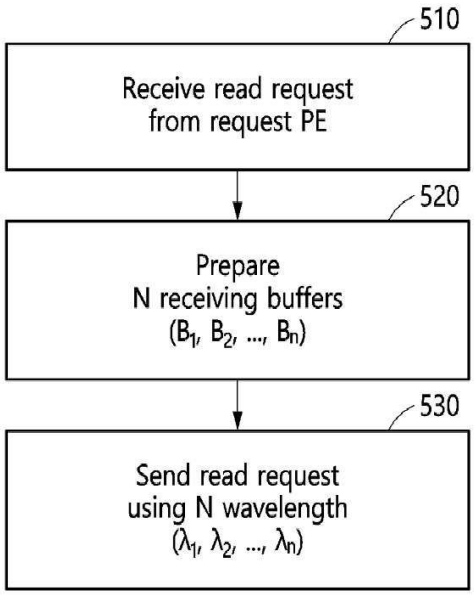
도면3b


Requester
Optical switch
Responder 1
Responder 2
Responder 3
Read req (M1)
Read req (M1)
Read req (M1)
Media controller delay
M1-1
M1-2
M1-3
Data transmission delay
Read resp (M1-1)
Read resp (M1-2)
Read req (M1-2)

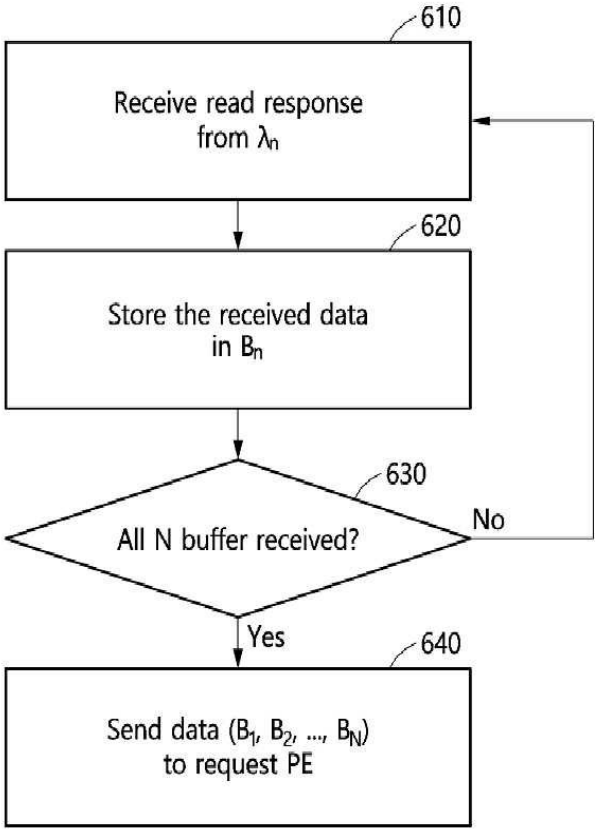
도면4


외부 저장 장치
Remote momory
Mem ctrl
Responder
Remote momory
Mem ctrl
Responder
Remote momory
Mem ctrl
Responder
λ1
λ2
λ3
Optical switch
λ1 λ2 λ3
서버
Optical interleaver
Req PE
Home agent
Local memory
Coh fabric
CPU Cores

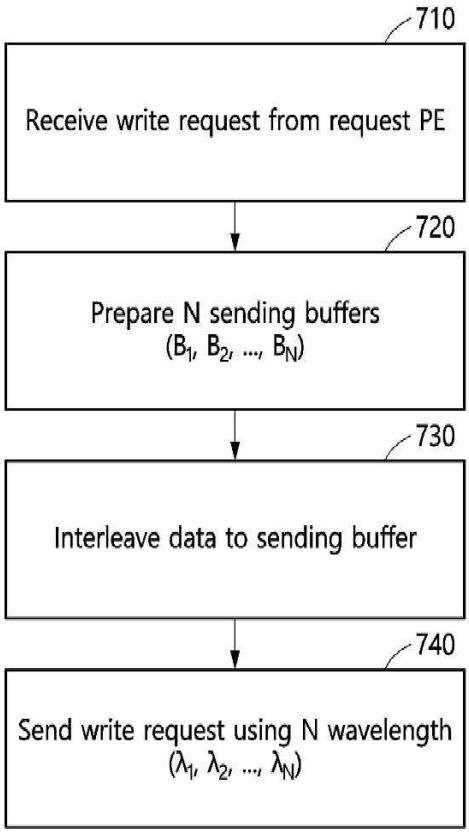
도면5


510
Receive read request from request PE
520
Prepare N receiving buffers (B_1, B_2, ..., B_n)
530
Send read request using N wavelength (λ_1, λ_2, ..., λ_n)

도면6


610
Receive read response from λ_n
620
Store the received data in B_n
630
All N buffer received?
No
Yes
640
Send data (B_1, B_2, ..., B_N) to request PE

도면7


710
Receive write request from request PE
720
Prepare N sending buffers
(B_1, B_2, ..., B_N)
730
Interleave data to sending buffer
740
Send write request using N wavelength
(λ_1, λ_2, ..., λ_N)