CCIX 연결망과 메모리 확장기술 동향

Trends of the CCIX Interconnect and Memory Expansion Technology

김선영 (S.Y. Kim, seonyoung8436@etri.re.kr) 슈퍼컴퓨팅기술연구센터 연구원
안후영 (H.Y. Ahn, ahnhy@etri.re.kr) 슈퍼컴퓨팅기술연구센터 선임연구원
전성익 (S.I. Jun, sijun@etri.re.kr) 데이터중심컴퓨팅시스템연구실 책임연구원
박유미 (Y.M. Park, parkym@etri.re.kr) 슈퍼컴퓨팅기술연구센터 책임연구원/센터장
한우종 (W.J. Han, woojong.han@etri.re.kr) 인공지능연구소 책임연구원/연구위원

ABSTRACT

With the advent of the big data era, the memory capacity required for computing systems is rapidly increasing, especially in High Performance Computing systems. However, the number of DRAMs that can be used in a computing node is limited by the structural limitations of the hardware (for example, CPU specifications). Memory expansion technology has attracted attention as a means of overcoming this limitation. This technology expands the memory capacity by leveraging the external memory connected to the host system through hardware interface such as PCIe and CCIX. In this paper, we present an overview and describe the development trends of the memory expansion technology. We also provide detailed descriptions and use cases of the CCIX that provides higher bandwidth and lower latency than cases of the PCIe.

KEYWORDS CCIX, 고성능컴퓨팅, 메모리 확장기술, 슈퍼컴퓨팅, 차세대 연결망

Ⅰ. 서론

컴퓨팅 시스템에서 처리되는 데이터의 크기는 인공지능을 비롯한 빅데이터를 활용하는 분야가 확대됨에 따라 급격히 증가하는 추세를 보인다. IDC(International Data Corporation)의 글로벌 데이터 동향 조사지에 따르면, 전 세계에서 처리되는 데이터의 크기는 2021년에 60ZB를 넘어섰으며, 2025년에는 175ZB에 이를 것으로 전망하고 있다[1]. 이처럼 컴퓨팅 시스템에서 처리되는 데이터의 크기가 늘어남에 따라, 시스템에 요구되는 메모리의 용량 또한 기하급수적으로 증가하고 있다.

특히, 이러한 대용량 메모리에 대한 수요는 HPC(High Performance Computing) 시스템에서 더욱 두드러진다. 2017년에 HPC 시스템의 CPU 코어 수와 메모리 용량에 따른 HPL(High Performance

* DOI: https://doi.org/10.22648/ETRI.2022.J.370105
* 본 연구 논문은 한국연구재단 슈퍼컴퓨터개발선도사업[2020M3H6A1084857, 초병렬 프로세서 기반 고집적 컴퓨팅 노드 및 시스템 개발] 및 한국전자통신연구원 내부연구과제(전략적선행투자사업)[21RS1100, 의료데이터 기반 인공지능 슈퍼컴퓨팅 기술 활성화 연구사업]의 일환으로 수행되었음.

Linpack) 벤치마크 점수를 분석한 연구의 결과에 따르면, HPL의 이론성능을 얻기 위한 CPU 코어당 메모리 용량은 시스템을 구성하는 전체 코어 수에 비례하여 증가하는 경향을 보인다[2]. 해당 연구 결과는 HPC 시스템을 구성하는 CPU 코어 수가 증가하는 추세 속에서 HPC 시스템에 요구되는 메모리 용량은 더욱 빠르게 증가할 것임을 시사한다.

이처럼 컴퓨팅 시스템에 요구되는 메모리 용량은 날이 갈수록 증가하고 있지만, 하드웨어적 특성(CPU의 specification 등)에 따라 컴퓨팅 노드에 장착 가능한 메모리의 용량에는 한계가 존재한다[3]. 이에 따라, 메모리 용량 한계를 극복하기 위해 컴퓨팅 노드가 자신의 로컬 메모리 외의 확장 메모리를 사용할 수 있도록 하는 메모리 확장기술이 제안되었다. 메모리 확장기술은 컴퓨팅 노드가 PCIe(Peripheral Component Interconnect express)와 같은 하드웨어 인터페이스를 통해 확장 메모리에 접근하여 이를 자신의 로컬 메모리처럼 사용할 수 있도록 함으로써 메모리 용량 한계를 극복하도록 한다.

하지만 기존에 널리 사용되는 PCIe 인터페이스를 통해 메모리 확장기술을 구현할 경우, PCIe의 낮은 대역폭이 병목으로 작용하게 된다. 그림 1에 보이는 바와 같이 CPU와 DDR4 메모리 사이의 대역폭은 100GB/s 이상인 반면, PCIe 4.0 인터페이스로 연결된 CPU와 메모리 확장장치 간 대역폭은 최대 64GB/s에 불과하다. 또한, 최근 HPC 시스템 분야에서 주목받고 있는 HBM(High Bandwidth Memory)의 경우 400GB/s 이상의 대역폭을 제공하기에, 이 경우 PCIe 인터페이스로 인한 대역폭 병목 현상은 더욱 심화된다.

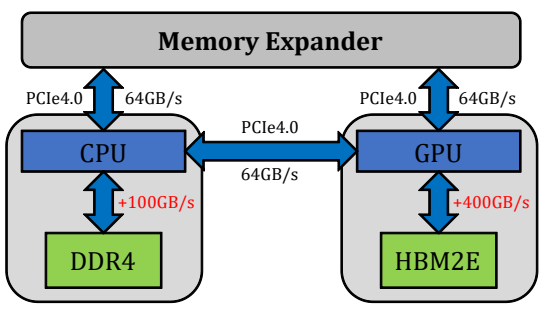


그림 1 PCIe 대역폭 병목 현상

위와 같은 PCIe의 한계점을 극복하기 위해 다수의 기업은 PCIe 대비 낮은 지연율, 높은 대역폭을 지향하는 차세대 연결망 개발에 착수하였다. 대표적으로 인텔 진영에서 개발한 CXL, ARM 진영에서 개발한 CCIX 등이 있으며, 최근 CXL 연결망을 기반으로 한 메모리 확장장치가 공개되는 등[3] 차세대 연결망을 활용한 메모리 확장기술 연구의 성과가 창출되고 있다.

본고에서는 메모리 확장기술과 대표적 차세대 연결망인 CCIX 연결망 기술에 대한 동향을 다루며, 본고의 구성은 다음과 같다. Ⅱ장에서는 메모리 확장기술의 개요, 차세대 연결망, 그리고 메모리 확장기술의 개발 동향에 대해 다룬다. Ⅲ장에서는 CCIX 연결망의 상세한 구조와 CCIX의 use case에 대해 다룬다. 마지막으로, Ⅳ장에서 전체적인 내용을 요약하며 본고를 마무리한다.

Ⅱ. 메모리 확장기술

1. 메모리 확장기술의 개요

메모리 확장기술은 컴퓨팅 노드가 하드웨어 인터페이스를 통해 노드 외부의 메모리에 접근하고 이를 사용할 수 있도록 하는 기술이다. 컴퓨팅 노드는 메모리 확장기술을 통해 로컬 메모리 외의 추가적인 메모리 용량을 확보할 수 있다. 그림 2는 메모리 확장기술을 적용한 시스템의 일반적인 구현 형태를 나타낸다. 그림 2에 보이는 바와 같이, 시

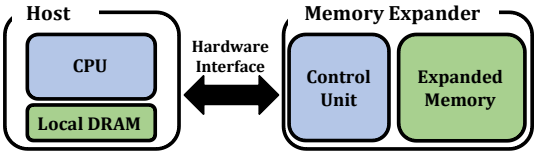


그림 2 메모리 확장기술의 일반적인 구현 형태

스템은 로컬 메모리 외의 추가적인 메모리 용량을 확보하고자 하는 호스트와 메모리 확장기술 구현을 위한 하드웨어인 메모리 확장장치로 구성된다. 메모리 확장장치는 호스트와의 통신, 장치 제어, 그리고 확장 메모리 제어를 위한 제어 유닛 및 확장 메모리 모듈로 구성된다. 메모리 확장장치와 호스트 간 연결은 CCIX를 비롯한 하드웨어 인터페이스를 통해 이루어진다.

메모리 확장기술은 그림 2의 구조를 기반으로 하여 다양한 형태로 구현될 수 있다. 예를 들어, 다수의 컴퓨팅 노드가 접근 가능한 공유 메모리 형태의 대용량 확장 메모리 풀을 중심으로 다수의 컴퓨팅 노드가 연결된 형태로 구현되거나[4,5], add-in 카드 형태의 메모리 확장장치가 노드에 직접 장착되는 형태로 구현될 수도 있다[3,6]. 만약 시스템에 속한 컴퓨팅 노드의 수가 많고, 노드 간 데이터 공유가 빈번하게 발생하는 경우, 공유 메모리 풀의 형태로 시스템의 메모리 용량을 확장하는 것이 적합할 것이다. 만약 특정 컴퓨팅 노드의 메모리 용량만을 확장하고자 하는 소규모 서버 운영자 또는 개인 사용자의 경우, add-in 카드 형태의 메모리 확장장치를 해당 컴퓨팅 노드에 장착하는 것이 적합할 것이다. 이처럼 메모리 확장기술은 사용자의 필요에 따라 다양한 형태로 구현되는 것이 가능하다.

2. 차세대 연결망

PCIe의 낮은 대역폭 및 높은 지연율 등의 문제를 해결하기 위해 인텔, ARM을 비롯한 세계 유수의 기업들은 PCIe 대비 고대역폭, 저지연을 지향하는 차세대 연결망 개발에 집중하고 있다. 메모리 확장장치가 차세대 연결망을 통해 호스트와 연결될 경우, PCIe 대비 로컬 메모리에 접근하는 속도와 확장 메모리에 접근하는 속도의 차이를 줄이는 것이 가능할 것이다.

표 1은 세 가지의 대표적인 차세대 연결망들을 비교한 표이다. 우선, CCIX[7]는 ARM이 주도하는 CCIX 컨소시엄에서 개발한 차세대 연결망으로 2018년 1.0 규격이 배포되었다. CCIX는 호스트와 가속기를 고속으로 연결하는 동시에 스눕 필터(Snoop Filter) 기반의 캐시 일관성 유지정책을 통해 프로토콜 레벨에서 장치 간 캐시 일관성을 보장한다. 또한, PCIe의 물리 계층을 기반으로 하므로 기존 시스템과의 호환성이 높으며, 메시, 링 등의 다양한 토폴로지를 지원한다는 장점이 있다. CCIX는 최대 128GB/s의 대역폭을 지원한다.

CXL[8]은 인텔이 주도하는 CXL 컨소시엄에서

표 1 차세대 연결망 비교표

	CCIX	CXL	Gen-Z
토폴로지 지원	p2p, mesh, ring, etc.	p2p	p2p, mesh, ring, etc.
대역폭	~128GB/s (x16)	~128GB/s (x16)	32GB/s ~ 400+GB/s
물리 계층	PCIe 5.0 PHY	PCIe 5.0 PHY	IEEE 802.3
캐시 일관성 지원 여부	O	O	Supported only in p2p
개발 주도 기업	ARM, Xilinx	Intel	Hewlett Packard (HP)

개발한 차세대 연결망으로, CCIX와 유사하게 호스트와 주변 장치들을 캐시 일관성을 유지하면서 고속으로 연결하는 것을 목적으로 한다. CXL 또한 PCIe의 물리 계층을 기반으로 하며, 최대 128GB/s의 대역폭을 지원한다. 이처럼 CCIX와 CXL은 기능 측면에서 유사한 점이 많지만, 구조적인 측면에서 많은 차이가 존재한다. 예를 들어, CCIX는 대칭형 CCI(Cache Coherent Interconnect) 구조를 채택하였지만, CXL은 비대칭형 CCI 구조를 채택하였다.

Gen-Z[9]는 확장성에 중심을 둔 연결망으로 최대 4,096개의 장치 연결이 가능한 서브넷 주소 공간을 제공한다. 또한, 유연한 장치 구성을 위해 단일 노드 레벨에서부터 클러스터 레벨에 이르기까지 메시, 링 등의 다양한 토폴로지를 지원한다. Gen-Z는 구성에 따라 32GB/s에서 400GB/s에 이르는 넓은 대역폭 지원이 가능하지만, p2p 구조에서만 캐시 일관성을 보장한다.

3. 메모리 확장기술의 개발 동향

차세대 연결망과 같은 기반 기술들의 발전에 따라, 메모리 확장기술을 구현한 다양한 메모리 확장장치들이 공개되고 있다. 표 2는 국내외 기업과 연구기관에서 개발한 메모리 확장장치들을 비교한 표이며, 이 절에서는 표 2의 메모리 확장장치들을 중심으로 메모리 확장기술 관련 최신 개발 동향을 살펴본다[10].

가. Samsung "CXL Memory Expander"

상용화를 앞둔 대표적인 메모리 확장장치로는 2021년 5월 삼성전자가 공개한 CXL 연결망 기반의 메모리 확장장치가 있다[3]. 해당 메모리 확장장치는 add-in 카드 형태로 컴퓨팅 노드에 연결되며, CXL 2.0 연결망을 통해 TB급의 DDR5 메모리를 컴퓨팅 노드가 사용할 수 있도록 한다. 공개된 구조에 따르면, 해당 장치의 컨트롤러는 Xilinx 사의 Kintex FPGA를 통해 구현되었으며, 이를 통해 메모리 어드레스 맵핑 기능 및 인터페이스 간 변환 기능이 수행된다.

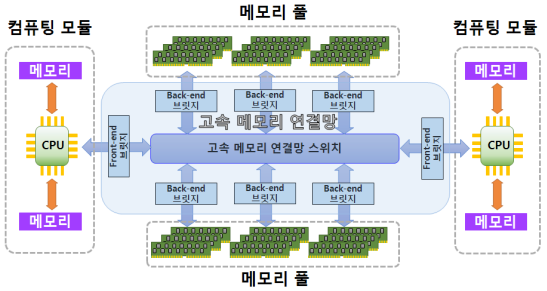
나. ETRI "Gen-Z Memory Pool System"

ETRI에서는 2020년 Gen-Z 연결망 기반의 메모리 풀 시스템을 개발하였다[4]. Gen-Z 메모리 풀 시스템은 대용량 메모리 모듈들이 장착된 Xilinx 사의 FPGA 보드를 기반으로 하는 확장 메모리 풀을 통해, 이에 연결된 컴퓨팅 모듈들에 총 4.5TB 크기의 확장 메모리 용량을 제공한다. 그림 3[11]에서 보이는 바와 같이 ETRI의 Gen-Z 메모리 풀 시스템은 크게 다음과 같은 기능 블록으로 구성된다: 1) 컴퓨팅 노드의 확장 메모리 접근 명령에 따라 Gen-Z 패킷을 생성하는 front-end 브릿지,

표 2 메모리 확장장치 비교표

	Samsung "CXL Memory Expander"	ETRI "Gen-Z Memory Pool System"	HPE "The Machine"
적용된 연결망 기술	CXL 2.0	Gen-Z 1.0	Gen-Z or HPE's own interface
메모리 인터페이스	DDR5	DDR4	DDR4
확장 메모리 용량	TB level	4.5TB ↑	160TB ↑
구현 형태	Add-in Card	Shared Memory Pool	Shared Memory Pool

출처 Reprinted with permission from [10].



출처 Reprinted with permission from [11].

그림 3 ETRI Gen-Z 메모리 풀 시스템의 구조

2) Gen-Z 패킷을 목적지로 라우팅하는 스위치, 3) Gen-Z 패킷을 수신하고 확장 메모리 풀을 제어하는 back-end 브릿지. 이와 같은 구조를 가진 ETRI의 Gen-Z 메모리 풀 시스템은 스위치에 연결된 컴퓨팅 노드들의 메모리 용량을 큰 폭으로 확장한다.

다. HPE "The Machine 프로토타입 보드"

HPE사는 2017년 메모리 드리븐 컴퓨팅(Memory-Driven Computing) 구현을 위한 The Machine 프로토타입 보드와 이를 적용한 시스템을 공개하였다[5]. The Machine 프로토타입 보드는 ARM 아키텍처 기반의 CPU와 256GB의 로컬 메모리, 그리고 4TB 용량의 확장 메모리 풀을 갖추고 있으며, 해당 보드로 구성된 시스템은 프로토타입 보드 40개가 모여 총 160TB의 공유 메모리 풀을 가진다. 프로토타입 보드에 사용된 하드웨어 인터페이스는 명확히 공개되지 않았으나, 다수의 보드 연결을 통해 메모리 풀을 구성했다는 점을 고려하면 Gen-Z 혹은 HPE의 독자적인 인터페이스가 적용된 것으로 추측된다.

Ⅲ. CCIX

ETRI 슈퍼컴퓨팅기술연구센터에서는 2021년부터 CCIX 연결망 기반의 메모리 확장기술을 개발 중이다. 이에, 이 장에서는 Ⅱ장에서 설명하였던 CCIX 연결망의 구조에 대한 상세한 설명과 더불어 CCIX의 use case에 관해 설명한다.

1. CCIX 연결망 개요

CCIX(Cache Coherent Interconnect for Accelerators)는 ARM, Xilinx가 주도하는 CCIX 컨소시엄에서 개발한 차세대 연결망으로, 2018년 1.0 규격이 배포되었으며 현재 1.1 버전까지 공개되었다[7]. CCIX는 호스트와 가속기 간 캐시 일관성을 프로토콜 레벨에서 보장하여 PCIe의 DMA(Direct Memory Access) 기반 캐시 일관성 유지정책 대비 캐시 일관성 유지에 소모되는 지연 시간을 줄인다. 또한, CCIX의 독자적인 물리 계층인 EDR(Extended Data Rate) PHY를 통해 최대 32GT/s의 대역폭을 제공하여, PCIe 대비 향상된 물리적 연결을 제공한다.

최근, CCIX 컨소시엄은 SC20에서 현재 개발 중인 CCIX 2.0 버전에 관한 내용을 발표하였다[12]. CCIX 2.0 버전에서는 기존 버전에서 지원하였던 보드 레벨에서의 상호 간 연결 외에도, 패키지 레벨에서의 칩 간 상호 연결이 지원될 예정이다. 또한, CCIX 2.0 버전에서는 FLIT(Flow Control

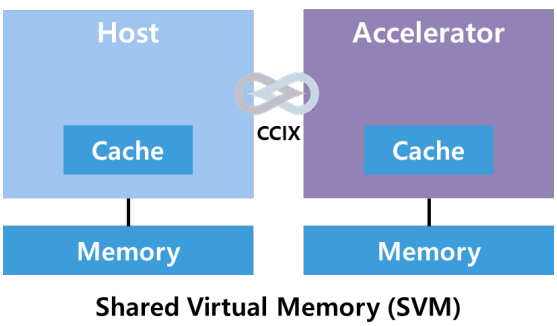


그림 4 호스트와 가속기가 CCIX를 통해 연결된 구조

Unit) 인코딩 적용 및 계층구조 간소화를 통해 이전 CCIX 버전에서 문제점으로 지적된 높은 지연 시간을 대폭 감소시켰으며, 최대 56GT/s의 높은 대역폭이 제공될 예정이다.

메모리 확장기술 측면에서 CCIX는 그림 4와 같이 호스트와 가속기가 CCIX를 통해 연결된 구조에서 호스트가 가속기 측에 연결된 메모리를 SVM(Shared Virtual Memory) 모델을 통해 자신의 로컬 메모리처럼 할당하고 관리하는 것을 가능하게 한다. 즉 호스트가 가속기 측 외부 메모리를 바라보는 시점은, 멀티노드 시스템에서 사용되는 NUMA(Non-Uniformed Memory Access) 메모리 모델에서의 원격 메모리를 보는 시점과 같다. 이처럼 CCIX 연결망을 통해 시스템의 메모리 용량을 호스트의 메모리 용량 한계 이상으로 확장할 수 있으며, 이기종(Heterogeneous)의 메모리들을 단일 컴퓨팅 노드가 사용할 수 있도록 한다. 이외에도 CCIX는 호스트와 가속기 간의 고속 연결을 통해 데이터베이스, 엣지 컴퓨팅, 스토리지 등 다양한 분야의 작업을 가속하는 데 활용될 수 있으며, CCIX의 다양한 use case에 대한 설명은 Ⅲ장의 3절에서 후술한다.

2. CCIX 연결망 구조

CCIX 연결망은 PCIe 인터페이스의 계층구조를 기반으로 하는 것을 특징으로 하며, 이 절에서는 CCIX 연결망의 상세한 구조에 관하여 설명한다. 우선, CCIX 1.1 및 2.0 연결망의 각 계층에 관해 설명한 후, 다양한 에이전트와 포트, 그리고 링크로 구성되는 CCIX 연결망의 구성요소에 관해 설명한다.

가. CCIX 연결망의 계층구조

그림 5는 CCIX 1.1 및 2.0의 계층구조를 나타내

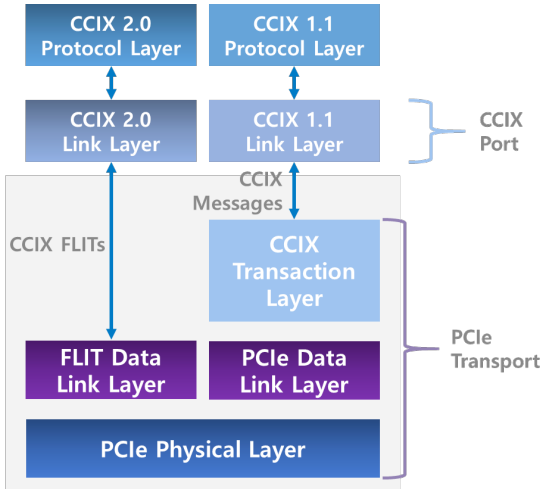


그림 5 CCIX 1.1 및 CCIX 2.0의 계층구조

며, 계층별 상세한 설명은 다음과 같다.

- 물리 계층(Physical Layer): PCIe의 물리 계층을 기반으로 하며, PLL(Phase Locked Loop) 등의 인터페이스 기능을 수행하기 위한 회로로 구성된다. CCIX 물리 계층은 데이터 링크 계층과 CCIX 링크 반대편에 있는 장치 간 인터페이스를 제공하며, 이를 위해 서로 다른 두 계층 간 패킷 포맷 변환 기능을 지원한다. CCIX는 PCIe 대비 넓은 대역폭 제공을 위해 PCIe와 호환 가능한 물리 계층 외에도 EDR(Extended Data Rate)이라는 독자적인 물리 계층을 제공한다.
- 데이터 링크 계층(Data Link Layer): 물리 계층과 트랜잭션 계층 간 데이터 전송을 담당한다. 데이터 링크 계층은 해당 계층을 통과하는 패킷의 무결성(Integrity)을 보장하며, 이를 위해 에러 검출 및 정정 기능을 수행한다. CCIX 1.1의 데이터 링크 계층은 PCIe의 데이터 링크 계층과 구성 및 동작이 완전히 같다는 특징이 있다. CCIX 2.0의 데이터 링크 계층은 CCIX 1.1의 데이터 링크 계층과 달리

패킷 레벨이 아닌 더 세분화된 단위인 FLIT 레벨에서의 데이터 무결성을 검증한다.

- 트랜잭션 계층(Transaction Layer): TLP(Transaction Layer Packet)를 처리하고 이를 다음 계층으로 전달하는 기능을 수행한다. CCIX 트랜잭션 계층은 PCIe와 호환 가능한 TLP(PCIe Compatible TLP)를 지원할 뿐만 아니라 PCIe의 표준 패킷 형식에서 불필요한 부분을 제거한 최적화된 TLP(Optimized TLP)를 지원한다. 해당 패킷 포맷을 사용할 경우 더욱 효율적인 데이터 전송이 가능하다. CCIX 2.0에서는 트랜잭션 계층을 계층구조에서 생략하여 데이터 전송 오버헤드를 줄이고자 하였다. 이를 통해 이전 버전의 CCIX 대비 지연 시간을 큰 폭으로 낮출 수 있다.
- 링크 계층(Link Layer): PCIe transport 계층의 포맷에 맞게 CCIX 패킷의 포맷을 변환하는 기능을 수행한다. CCIX 링크 계층은 현재 PCIe transport 계층을 기준으로 동작하지만, CCIX는 계층구조를 기반으로 하기에 향후 다른 프로토콜의 transport 계층을 기반으로 할 가능성이 존재한다[13]. CCIX 링크 계층은 다수의 CCIX 포트를 통합하여 더 넓은 대역폭을 사용할 수 있도록 하는 port aggregation 기능 또한 지원한다.
- 프로토콜 계층(Protocol Layer): CCIX 계층구조의 최상단에 위치하여, 메모리 읽기/쓰기 작업에 대한 일관성(Coherency) 유지 기능을 담당한다. CCIX 프로토콜 계층에서 정의된 캐시 상태에 따라 하드웨어는 메모리의 상태 정보(shared, dirty 등)를 알 수 있다.

나. CCIX 연결망의 구성요소

CCIX 연결망은 크게 CCIX 디바이스, 링크, 포트, 그리고 에이전트로 구성된다. 시스템 내부 CCIX 연결망의 엔드포인트에 위치하는 CCIX 디바이스는 모두 한 개 이상의 CCIX 포트를 가지며, 각 포트는 CCIX 링크를 통해 다른 포트와 연결된다. CCIX 디바이스의 기능은 해당 디바이스를 구성하는 CCIX 에이전트들에 의해 정의되며, CCIX 에이전트의 종류 및 각 에이전트의 기능은 다음과 같다.

- RA(Request Agent): CCIX 시스템 내부의 주소에 대한 읽기/쓰기 트랜잭션을 수행하며, 해당 주소를 캐시에 저장하는 기능을 수행한다. RA는 읽기/쓰기 트랜잭션의 시발점이 되는 가속기(Accelerator)와 연결되어 있으며, 이를 위해 가속기에 접근 가능한 인터페이스를 갖추고 있다.
- HA(Home Agent): 할당된 주소 영역에 대한 접근을 관리하고 캐시 일관성을 유지하는 기능을 수행한다. HA는 자신의 주소 영역 내 캐시 상태정보가 변경된 경우, 해당 캐시 정보를 가지고 있는 RA에게 스눕 트랜잭션(Snoop Transaction)을 통해 이를 알려, 캐시 일관성이 유지되도록 한다.
- SA(Slave Agent): 가속기를 비롯한 주변장치에 있는 확장 메모리를 의미한다. SA는 HA가 속한 칩 외부에 위치하면서 HA가 관리하는 주소 영역에 포함된 메모리이다.
- EA(Error Agent): CCIX 구성요소들로부터 전달된 프로토콜 레벨에서의 에러를 수신하고 이를 처리하는 기능을 수행한다.

3. CCIX use case

가. 호스트와 가속기 간 메모리 접근 성능 향상

CCIX 연결망을 활용할 경우 호스트와 가속기

간에 캐시 일관성을 보장하면서 메모리를 공유하여 장치 드라이버 없이 메모리 확장 효과를 얻는 것이 가능하다[13]. 예를 들어 가속기와 호스트가 호스트 측 메모리를 공유하는 상황에서, 호스트 측의 RA와 HA, 그리고 가속기 측 RA를 통해 호스트와 가속기가 캐시 일관성을 보장하면서 동일한 데이터 구조로 메모리를 공유하는 것이 가능하다.

또 다른 예로, 호스트 측에는 로컬 메모리가 장착되어 있고 가속기 측에는 영구 메모리인 PMEM(Persistent Memory)이 장착되어 상호 간 메모리를 공유하는 것 또한 가능하다. 가속기에 장착된 영구 메모리인 PMEM은 메모리 버스에 상주하고 바이트 단위 주소 지정이 가능하므로, DRAM과 같이 데이터에 대한 고속 접근이 가능하다. Redis와 MongoDB 같은 데이터베이스 시스템은 PMDK(Persistent Memory Development Kit)를 이용하여 가속기에 장착된 PMEM에 load/store 연산을 직접 수행하고, NUMA 모델로 메모리 확장 기능을 구현할 수 있다[14].

이처럼 메모리 확장기술은 호스트와 가속기의 메모리를 통합하여 Near-Memory Processing을 가능하게 한다. 이 기술을 이용하여 다수의 가속기로 구성된 가속기 풀을 구성하고, 시스템의 전체 데이터를 호스트와 가속기가 공유하도록 하면 데이터베이스, 엣지 컴퓨팅, 스토리지의 성능을 크게 향상하는 것이 가능하다.

나. 데이터베이스 장애 복구 성능 향상

데이터베이스는 트랜잭션 실행 시 오류가 발생하거나, 시스템 장애로 손상될 수 있으며, 특히 메인 메모리 데이터베이스의 경우 운영 중 메모리 부족으로 인해 결함이 발생할 수 있다. 예를 들어, PCIe 인터페이스와 SSD를 사용하는 컴퓨팅 구조에서 메인 메모리 데이터베이스인 Redis가 장애 복구를 위해 로깅과 체크포인팅을 수행하는 경우, Redis는 매초 마다 로그 파일을 백업하고 체크포인트 파일을 생성한다. 만약, 트랜잭션 처리가 실패하거나 시스템 운영 중 결함이 발생하면 데이터베이스가 재시작되고, SSD에 저장된 로그 파일을 읽어와 복구를 시작한다. 이러한 방식은 데이터의 유실에 대해서는 대비할 수 있지만, 상당한 수준의 입출력 오버헤드가 발생하고 이는 곧 CPU 사이클 낭비로 이어진다. CCIX를 활용하면 메인 메모리 데이터베이스의 로깅 및 체크포인팅 기능을 가속기로 오프로딩하여 트랜잭션 처리 및 장애 복구 성능을 향상할 수 있다.

CCIX 컨소시엄은 ARM 호스트와 Xilinx FPGA 기반의 가속기가 CCIX로 연결된 환경에서, 가속기에 장착된 PMEM에 직접 로그 파일과 체크포인트 파일을 저장하여 백업과 장애 복구 성능 향상을 동시에 달성할 수 있음을 보였다[15]. 가속기에 장착된 PMEM에 로그 파일과 체크포인트 파일을 저장하면 읽기/쓰기 연산을 모두 가속기에 오프로딩할 수 있으며, 이를 통해 호스트에서는 파일 시스템 처리 오버헤드를 줄이고, 시스템 장애 발생 시 재시작 시간을 절반 수준으로 단축할 수 있다.

또한, 앞선 예와 같이 CCIX를 통해 호스트와 가속기가 연결된 환경에서 호스트 메모리는 데이터베이스의 암호화와 압축 기능을 빠르게 처리하는 in-line acceleration을 수행하여 가속기에 장착된 PMEM에 데이터를 저장하고, 가속기는 직접 기계학습 알고리즘과 데이터 분석 질의 같은 고비용 연산을 처리하는 off-line acceleration을 수행하여 데이터베이스 전반의 성능을 향상하는 것 또한 가능하다.

다. 엣지 컴퓨팅 성능 향상

엣지 컴퓨팅은 데이터가 발생하는 소스에서 물

리적으로 가까운 곳에서 처리되도록 하므로, 전송해야 하는 데이터의 양이 줄어들어 네트워크 트래픽과 전송 비용이 절감되고, 보안 수준이 향상되며, 가공된 데이터만 중앙의 데이터센터로 전송되므로 스토리지 비용을 줄인다.

최근에는 인공지능 기술의 발전으로 일상 전반에서 다양한 형태의 지능형 서비스를 이용하게 되었다. 그러나 자율주행, 메타버스와 같은 엔드 포인트에서 실시간 대규모 데이터 처리가 필요한 서비스를 위해서는 더욱 짧은 지연 시간(Latency)과 높은 처리량(Throughput)이 요구되고 있다. 범용적인 컴퓨팅 기술로는 위와 같은 서비스들의 성능을 최적화하기에 한계가 있고, 네트워크, 스토리지, 기계학습, 영상 처리, 보안과 같은 이종 도메인에 최적화된 컴퓨팅을 수행해야만 만족할 수준의 성능을 얻을 수 있다.

Redis 진영은 엣지 데이터 처리 성능을 높이기 위해 CCIX 연결망으로 연결된 ARM 호스트와 Xilinx FPGA 기반 가속기를 활용하여 RedisEdge 플랫폼을 제시하였다[16]. RedisEdge 플랫폼을 엣지 단에서 발생하는 스트림 데이터에 대한 분석, 인공지능 추론, 데이터의 실시간 처리 등에 활용할 경우, 고화질 영상의 트랜스코딩, 네트워크 기능 가상화, 데이터 분석 질의, 기계학습 추론, 스토리지 압축과 같은 연산이 가속되어 엣지 워크로드의 성능이 향상될 수 있다.

라. 스토리지 성능 향상

CCIX 연결망을 활용한 메모리 확장기술을 통해 스토리지 데이터 분석 및 압축 성능이 향상될 수 있다. CCIX 컨소시엄은 메모리 확장기술을 통해 KVS(Key-Value Storage)의 key는 호스트의 메인 메모리에 할당하고, value는 CCIX를 통해 연결된 가속기의 PMEM에 할당하였다. 이를 통해 데이터 분석을 위한 고비용 연산 처리를 가속기로 오프로딩하여 KVS의 성능을 크게 향상할 수 있었다. 또한, 해당 시스템에서는 데이터베이스 연산량의 증가 대비 호스트 메모리 사용량이 거의 증가하지 않음을 보였다[17].

또 다른 사례로는 MongoDB의 입출력 및 압축 연산을 CCIX를 통해 연결된 가속기에 오프로딩하여 성능 향상을 도모한 사례가 있다[18]. 호스트와 가속기 간에 구축된 Split File System은 MongoDB의 메타데이터를 가속기 메모리에서 관리하고, 실제 데이터는 가속기에 장착된 SCM(Storage Class Memory)에 저장한다. 호스트에서는 fs_open, fs_exist, fs_rename 등의 연산을 처리하고, 가속기는 fs_read, fs_write와 같이 파일의 데이터를 실제로 읽고 쓰는 연산을 수행한다. 이 기술을 적용하면, 호스트의 CPU 사이클을 절약하면서도 가속기에 하드웨어적으로 구현된 압축 및 데이터 분석 기능을 활용하여 SCM에 저장된 데이터를 고속으로 처리하여 MongoDB의 성능이 크게 향상될 수 있다.

Ⅳ. 결론

본고에서는 메모리 확장기술의 개요 및 최신 개발 동향에 관해 기술하고, 대표적인 차세대 연결망인 CCIX의 상세한 구조와 메모리 확장기술 측면에서의 use case에 대해 살펴보았다. 4차 산업혁명을 기점으로 빅데이터를 활용하는 분야가 폭발적으로 증가함에 따라 컴퓨팅 시스템에 요구되는 메모리의 용량은 급격히 증가하고 있으며, 이러한 현상은 고성능 작업에 특화된 HPC 시스템에서 특히 도드라진다. 이러한 대용량 메모리에 대한 수요를 충족시키기 위해 시스템의 메모리 용량을 컴퓨팅 노드의 물리적 한계 이상으로 확장 가능케 하는 메모리 확장기술이 주목받고 있다.

대용량 메모리에 대한 수요가 폭증하고 있는 지금, 메모리 확장기술이 그 수요를 충족시킬 수 있는 적합한 솔루션임은 자명하다. 하지만, 하드웨어 인터페이스를 통해 확장 메모리에 접근하는 경우 로컬 메모리에 접근하는 경우 대비 상당한 수준의 지연 시간이 발생하는 등 기술의 성숙도 부족으로 인한 제약사항이 다수 존재한다. 따라서 메모리 확장기술의 한계점을 극복하기 위한 연구는 필수적으로 이루어져야 할 중요한 과제이며, ETRI에서도 슈퍼컴퓨터개발선도사업을 통해 CCIX 연결망 기반의 메모리 확장기술에 관한 연구를 수행하며 기술적 한계점을 극복하기 위한 노력을 다하고 있다.

용어해설

메모리 확장기술(Memory Expansion Technology) 컴퓨팅 노드가 자신의 로컬 메모리 외에 CCIX, PCIe와 같은 하드웨어 인터페이스를 통해 연결된 외부의 확장 메모리에 접근 및 이를 사용 가능하도록 하는 기술

PCIe(Peripheral Component Interconnect express) 인텔의 주도하에 2003년 개발된 CPU와 주변장치를 고속으로 연결하기 위한 시리얼 하드웨어 인터페이스

차세대 연결망 하드웨어 인터페이스로 널리 사용되고 있는 PCIe 인터페이스 대비 저지연, 고대역폭을 지향하는 차세대 하드웨어 인터페이스

고성능컴퓨팅(High-Performance Computing) 고성능의 하드웨어를 통해 대규모의 복잡한 연산을 빠르게 수행할 수 있도록 하는 기술

약어 정리

CCIX	Cache Coherent Interconnect for Accelerators
CPU	Central Processing Unit
CXL	Compute eXpress Link
DDR	Double Data Rate
FPGA	Field Programmable Gate Array
GPU	Graphic Processor Unit
HA	Home Agent
HBM	High Bandwidth Memory
HPC	High Performance Computing
HPL	High Performance Linpack
IDC	International Data Corporation
KVS	Key-Value Storage
NUMA	Non-Uniform Memory Access
PCIe	Peripheral Component Interconnect express
PMEM	Persistent Memory
PMDK	Persistent Memory Development Kit
RA	Request Agent
SCM	Storage Class Memory
SSD	Solid-State Drive

참고문헌

[1] IDC, "Data age 2025: The evolution of data to life-critical," white paper, 2017.

[2] D. Zivanovic et al., "Main memory in HPC: Do we need more or could we live with less?," ACM Trans. Archit. Optim., vol. 14, no. 1, 2017, pp. 1-26.

[3] Samsung Electronics, "Samsung unveils industry-first memory module incorporating new cxl interconnect standard," May 11, 2021, Available from: https://bit.ly/3uBo27J

[4] W. Kwon et al., "Gen-Z memory pool system architecture," in Proc. IEEE Int. Conf. Inf. Commun. Technol. Converg. (ICTC), (Jeju, South Korea), Oct. 2020.

[5] C. Hopkins, "The Machine seeks adventurous developers," June 8, 2017, Available from: https://www.hpe.com/us/en/insights/articles/new-computing-platform-seeks-adventurous-developers-1706.html [retrieved Dec. 13, 2021].

[6] Xilinx, "Versal: The first adaptive compute acceleration platform," white paper, 2020.

[7] CCIX Consortium, "CCIX base specification version 1.1," 2019.

[8] CXL Consortium, "CXL 2.0 specification," 2020.

[9] Gen-Z Consortium, "Gen-Z core specification 1.1," 2020.

[10] 김선영 외, "메모리 확장기술 동향," 대한전자공학회 추계학술대회, 2021. 11. 26.

[11] 김강호, "확장 메모리 풀 시스템 기술," Available from: https://itec.etri.re.kr/itec/sub02/sub02_01_1.do?t_id=1210-2021-00629#1

[12] CCIX Consortium, "SC20-CCIX 2.0: Transport agnostic interface in the intelligent accelerator era," Available from: https://www.youtube.com/watch?v=MUHgp3o35RA

[13] CCIX Consortium, “An introduction to CCIX,” white paper, 2019.

[14] Persistent Memory Programming, Available from: https://pmem.io/

[15] CCIX and Real-World Application, Dec. 2019, available from: https://www.ccixconsortium.com/ccix-and-real-world-application/

[16] J. Defilippi, “Accelerating RedisEdge with CCIX,” Arm TechCon, 2019.

[17] XilinxInc, Key-Value Store Acceleration with CCIX, Dec. 18, 2018, Available from: https://www.youtube.com/watch?v=drIu4vlubxE&list=PLRr5m7hDN9TLI3vuw1OqLbF7YcGi3UO9c&index=10

[18] M. Mittal, “Memory expansion and storag acceleration with CCIX Technology,” in Proc. Int. Conf. High Perform. Comput., Netw., Storage, Anal. (SC19), (Denver, CO, USA), Nov. 2019.