InfiniBand: 차세대 시스템 연결망

한국전자통신연구원 박 경* · 모상만*

1. 서 론

인터넷 기술과 멀티미디어 처리 기술의 발전은 인터넷 기반형 서비스의 대중화를 가능하게 했으며, 이로 인하여 인터넷 접속 서비스, 전자상거래, 인터넷 게임, 인터넷 방송국, 인터넷 데이터 센터, 사이버 대학 등과 같은 인터넷 기반 상용 서비스가 하나의 산업 분야로 부상하였다. 이와 같은 인터넷 서비스의 급성장과 대중화에 따라 기업용 서버(Enterprise Server)와 고속 연산용 서버(High Performance Computing Server)로 양분되어 오던 컴퓨터 서버 시장에 인터넷 서버 라는 새로운 시장 영역을 창출하게 하였으며, 인터넷 서버 구축을 위한 제품군과 사용 환경이 컴퓨터 서버 시장의 큰 영역으로 급부상 하였다[1,2].

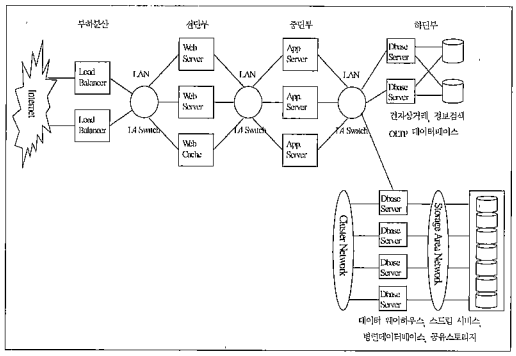


그림 1 3계층 연결 구조에 의한 인터넷 서버

현재 인터넷 서버는 기능별로 그림 1과 같이 복수 개의 컴퓨터를 다계층으로 연결하는 구조를 일반적으로 사용하고 있으며, 고속 LAN을 통한 이기종 통합 환경에 전용 클러스터 시스템과 스토리지 연결망 기반 대용량 자료 저장 시스템이 연동되는 구조로 요약된다. 3계층 연결 구조를 갖는 인터넷 서버는 사용자의 서비스 요구를 받아서 해석하는 전단부(Front-end)와 서비스별로 응용 프로그램을 수행하는 중단부(Mid-Tier), 응용 서비스 처리를 위한 데이터베이스와 대용량 데이터를 관리하는 후단부(Back-end)로 구성된다. 각 계층은 고속 LAN을 통해서 연결되고, 데이터베이스와 대용량 데이터를 관리하는 후단부는 분산 처리를 위한 전용 클러스터 연결망(Cluster Network)과 RAS (Reliability, Availability, Scalability) 특성이 뛰어난 스토리지 연결망(Storage Area Network)으로 구성된다[1,2,3].

인터넷 서버는 기능상 분산 처리 시스템과 원격 저장 장치를 고속 LAN, 클러스터 연결망, 스토리지 연결망 등을 사용하여 연결한 이기종 클러스터 시스템으로 정의할 수 있다. 클러스터 기술은 여러 개의 프로세싱 노드를 클러스터 연결망으로 연결하여 가용성과 확장성 높은 시스템을 우수한 가격대 성능비로 구축하게 한다. 특히 근래에 들어 SHV(Standard High Volume) 노드의 출현, 클러스터 연결망 기술, 파일 시스템 및 운영체제 기술, 클러스터 관리 소프트웨어 기술들이 개발되면서 클러스터 구조는 고성능 연산용 서버에서부터 기업용 서버, 인터넷 서버에 이르기까지 그 영역을 확대해 가면서 서버 구조의 주류로 자리잡고 있다.

클러스터 기술의 발전[5,6,7]은 그림 2에 나타낸 바와 같이 서버 팜(Server Farm)으로 불리는 대규모 클러스터 구조로의 변화를 이끌고 있으며, 클러스터 연결망[5,6,7], 스토리지 연결망[8], 이기종간 클러스터 연결망 등과 같은 기존의 연결망들이

* 정회원

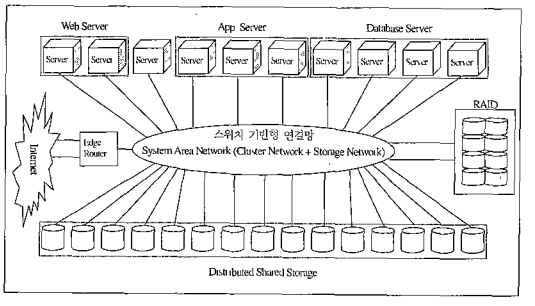


그림 2 서버 팜 구조를 이용한 인터넷 서버

하나의 시스템 연결망(System Area Network)으로 통합되어 가는 추세이다[9]. 이에 따라 시장 지배력이 큰 회사들을 주축으로 새로운 시스템 연결망 개발이 진행되어 왔으며 2000년 10월에 이르러 InfiniBand[4]라는 업계 표준 연결망 규격을 제정하고 상품화 개발을 추진 중이다.

본 고에서는 업계 표준으로 제정된 InfiniBand 연결망 기술을 소개하고 업체들의 개발 동향에 대해서 기술한다. 서론에 이어 제2장에서는 InfiniBand 연결망 기술의 태동 배경을 설명하고, 제3장에서는 InfiniBand 연결망 구조에 대해서 개괄적으로 기술하며, 제4장에서는 관련 업체들의 InfiniBnd 기술 개발 동향에 대해서 기술한다. 마지막 제5장에서는 결론 및 향후 기술 개발 방향을 기술한다.

2. InfiniBand 기술의 태동

2.1 통합형 연결망의 대두

서론에서 언급한 바와 같이 서버 구조가 분산 공유 저장 장치를 포함하는 대규모 클러스터 시스템으로 진화함에 따라 Servernet , Myrinet 등으로 대표되는 클러스터 연결망[5,6,7]과 FC-AL로 대표되는 스토리지 연결망[8]을 하나의 시스템 영역 연결망으로 통합할 필요성이 제기되었다. 클러스터 연결망과 스토리지 연결망은 공통적으로 고가용성과 고확장성을 보장하는 구조를 가지며, 고속 데이터 전송을 지원하고 있으나, 전송 서비스 특성 및 구현 방식이 다르게 발전되어 왔다. 이에 따라 컴퓨터 서버 시장을 주도하는 업체들은 클러스터 연결망과 스토리지 연결망을 동시에 충족하는 통합형 시스템 연결망 개발을 시도하게 되었고, 이를 통하여 대규모 클러스터 시스템 구조를 차세대 서버 구조로 지향하고 있다[3,9].

2.2 단일 계층 연결망의 필요성

현재 클러스터 연결망이나 스토리지 연결망에 접속된 노드는 내부적으로 PCI 버스[10]를 통하여 연결망에 접속하는 다중 계층 구조이다. 따라서 각 노드의 전송 능력은 PCI 버스 성능에 의해서 제한되고 있다. PCI 버스는 공유 버스 구조를 갖는 입출력 버스로 공유 버스라는 구조적 단점으로 확장성 및 성능 향상에 한계를 갖는다. 또한 PCI 버스는 동작중 장/탈착(Live Insertion/Withdrawl) 및 고장 분리(Fault Isolation) 기능이 제공되지 않아서 시스템 상에서 단일점 고장(Single Point Failure) 지점으로 지적되고 있다. 따라서 시스템 상에서 PCI 버스를 제거하고 프로세서 인터페이스 이하의 모든 연결을 스위치 기반형 단일 연결망으로 구성하여 가용성 및 확장성을 보장하고, 점대점 연결을 통한 고속 직렬 통신 방법을 사용하여 전송 속도를 향상시킨 새로운 시스템 영역 연결망이 필요하게 되었다[3,9,11,12,13].

PCI 버스를 통한 입출력 처리는 지정된 공유 메모리 공간에 프로세서가 읽기와 쓰기를 반복해야 하는 로드/스토어(Load/Store) 방식으로 프로세서의 데이터 복제 오버헤드가 과도한 단점을 갖는다. 따라서 프로세서와 입출력 장치 사이에 구성된 독립적인 메시지 패싱 인터페이스로 주기억 장치와 입출력 장치간의 직접적인 데이터 전송을 보장하여 공유 메모리 매핑 방식에서 발생하는 데이터 복제 오버헤드를 제거할 수 있는 채널 기반형 통신 방식으로의 전환이 성능 향상을 위해서 필수적이다. 채널 기반형 메시지 패싱은 중형 클러스터 컴퓨터에서 사용되던 VIA(Virtual Interface Architecture)를 적용한 것으로 프로세서 상호간 또는 프로세서와 입출력 장치간에 독립적인 사용자 수준 메시지 패싱 인터페이스를 제공하여 운영체제의 간섭 없이 통신을 가능하게 한다. 채널은 점대점 연결망 사용을 위한 사용자 프로그래머블 레지스터 레벨 인터페이스로 정의할 수 있으며, 채널을 제어하는 하드웨어는 DMA(Direct Memory Access) 엔진을 사용하여 주기억 장치와 통신 상대방 사이의 직접 데이터 송수신을 가능하게 한다. 따라서 프로세서는 입출력 장치의 제어와 접근 및 데이터 전송에 관련된 작업 부하를 줄임으로써 시스템 성능을 향상시킬

수 있다[3,9,11,12,13]. 표 1은 기존의 PCI 버스, 성능이 개선된 PCI-X 버스, 그리고 InfiniBand 연결망 기술의 주요 특징을 비교 요약한 것이다.

표 1 PCI 버스와 InfiniBand의 비교

	PCI 2.2	PCI-X 1.0	InfiniBand
동작속도	66MHz	133MHz/1 slot	2.5Gbps
데이터 신호 폭	32/64비트	32/64 비트	1x, 4x, 12x
최대 대역폭	533 Mbyte/sec	1064Mbyte/sec	2.5Gbps/포트
연결 방식	공유 버스	공유 버스	스위치 기반 점대점
최대 연결 노드	10개 이하	1개/133MHz[1]	2^{16}개/서브넷
통신 방식	공유 메모리 매핑	공유 메모리 매핑	채널 기반 메시지 패싱

2.3 업계의 표준화 노력

서버의 기능과 요구사항이 변화함에 따라 요구 기능에 적합한 서버 구조가 제안되었고, 새롭게 제안된 서버 구조는 새로운 연결망 구조의 필요성을 부각시키고 급기야 InfiniBand 기술의 탄생을 촉발시켰다. 인텔(Intel)과 썬(Sun Microsystems)을 주축으로 하는 NGIO(Next Generation IO) 그룹과 HP(Hewlet Packard)와 IBM, Compaq을 주축으로 하는 FIO(Future IO) 그룹이 차세대 시스템 연결망을 목표로 각각 구성되었다. 양 진영은 모두 스위치 기반형 점대점 연결망이라는 유사한 연결망 구조와 이를 이용한 서버 연결 구조 및 스토리지 연결 구조를 제안하고 있었으며, 지난 1999년 8월 두 진영이 통합하여 IBTA(InfiniBand Trade Association)라는 국제적인 표준화 단체를 결성하였다[11]. IBTA는 컴팩(Compaq), 델(Dell), IBM, HP, 인텔(Intel), 마이크로소프트(Microsoft), 썬(Sun Microsystems) 등 7개 회사로 구성된 운영위원회를 주축으로 11개 회사로 구성된 스폰서 회원사와 220여 개의 일반 회원사로 운영되고 있다.

IBTA는 지난 2000년 10월 InfiniBand 표준 규격을 제정하여 Las Vegas에서 동월 24일부터 3일간 개최된 IBTA 개발자 회의에서 InfiniBand 표준 규격을 공개하였다. IBTA는 세부 기술 분과 위원회의 지속적인 활동과 기술 공개를 통해서 회원사들의 상품화를 지원하고 있으며, 2001년 말 또는 2002년 초경에는 선도 기업부터 InfiniBand 관련 제품을 출시할 것으로 예측되고 있다[12,13, 14,15].

1) 1 슬롯 / 133MHz, 2 슬롯 / 100 MHz, 4 슬롯이상 / 100 MHz 이하

3. InfiniBand 구조(Architecture)

InfiniBand 구조(IBA: InfiniBand Architecture)는 상호 독립적인 프로세서 플랫폼, 입출력 처리 플랫폼 그리고 입출력 장치를 연결하는 시스템 연결망(SAN: System Area Network)으로 입출력 장치 구성 및 입출력 처리에 최적화된 스토리지 연결망과 프로세서 상호간 통신에 최적화된 클러스터 연결망을 통합한 새로운 연결망으로 통신 및 관리 메커니즘을 포함하고 있으며, 소규모 서버 시스템에서 대규모 인터넷 서버 및 수퍼 컴퓨터에 이르기까지 다양한 영역에서 사용 가능하다. 더욱이 인터넷 통신 프로토콜에 친숙하게 설계되어 있어 인터넷 접속, 인트라넷 접속, 원격 서버 접속을 손쉽게 구현할 수 있게 한다. 그림 3은 IBA 구성도를 나타낸 것이다.

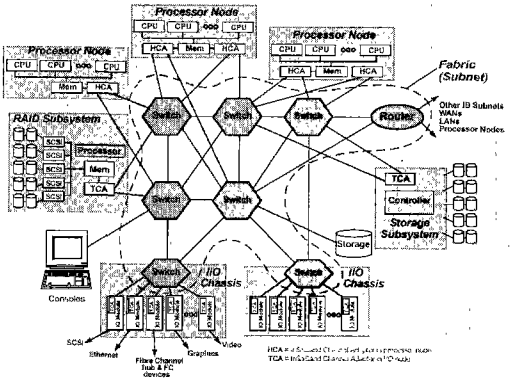


그림 3 InfiniBand Architecture 구성도

IBA는 복수개의 디바이스간에 동시 전송을 가능하게 하며, 높은 전송 대역폭과 낮은 전송 지연시간을 제공하고 높은 확장성 보장을 위하여 스위치 기반형 연결망으로 정의되었다. IBA는 통신 처리에 있어서 프로세서의 부하를 최소화할 수 있도록 프로토콜 및 프로토콜 처리 하드웨어를 정의하고 있으며, 이를 통하여 무복사(zero processor-copy) 데이터 전송, 커널 오버헤드 최소화, DMA (Direct Memory Access), 고확장성 및 고가용성 기능을 제공하고 있다[3,9,12,13,14,15].

3.1 토폴로지와 연결망 구성 요소

IBA 연결망은 스위치 기반 비정형 연결망으로 종단 노드(프로세서 노드, 입출력 노드)가 연결되는 여러 개의 서브넷(subnet)으로 구성된다. 서브넷은 종단 노드와 스위치, 라우터로 구성되며 서브넷 간 연결은 라우터를 통해서 이루어 진다. 하나의 서브넷에는 최대 65,536개의 종단 노드를 연결할 수 있는 고확장성이 제공된다. 각 종단 노드는 IBA 연결망 접속을 위한 채널 어댑터를 가지며, 프로세서 노드쪽에서는 호스트 채널 어댑터(HCA: Host Channel Adapter)를 사용하고 입출력 처리 노드 및 입출력 장치쪽에서는 타겟 채널 어댑터(TCA: Target Channel Adapter)를 사용한다.

3.1.1 채널 어댑터

채널 어댑터는 지역 또는 원격 DMA 기능을 수행하는 프로그래머블 DMA 엔진으로 정의 될 수 있으며, 가상 메모리 보호 메커니즘을 포함하고 있다. 전송 요구를 해독하여 해당 전송 요구를 처리하기 위한 패킷을 발생시키고 수신하는 기능을 수행하며 기능에 따라서 HCA와 TCA로 구분된다.

HCA는 다른 호스트 또는 입출력 장치와 통신을 위하여 호스트에게 채널 인터페이스를 제공하는 장치로 사용자 수준 프로그램에서 Verb 라는 소프트웨어 인터페이스를 사용하여 메시지를 송출하고 수신하는 기능을 수행한다. 호스트 프로세서에 의해서 발생된 메시지 전송 요구를 해독하여 호스트 메모리에서 데이터를 읽어서 IBA 연결망으로 송출하고 또한 수신된 메시지를 해독하여 호스트 메모리에 직접 쓰는 작업을 수행한다. HCA는 스위치와 TCA 모두에 직접 연결할 수 있으므로 서버 구성 시 서버의 규모에 따라 유연하게 시스템을 구성할 수 있다. TCA는 입출력 장치와 연결망을 이어주는 장치로 디스크 콘트롤러, 네트워크 콘트롤러, RAID 콘트롤러 등과 같은 다양한 입출력 장치를 IBA 연결망에 정합할 수 있게 한다.

채널 어댑터는 그림 4와 같은 개념적 구조를 갖는다. 채널 어댑터(HCA)의 구성, 관리 및 동작 명령 전달을 위해서 Verb를 정의하고 있다. Verb는 사용자(응용 프로그램)의 메시지 전송 또는 데이터 서비스 요구를 채널 어댑터에 전송하기 위해서 다양한 기능을 정의하고 각 기능에서 사용되는 파라미터를 정의하고 있다. 채널 어댑터는 복수개의

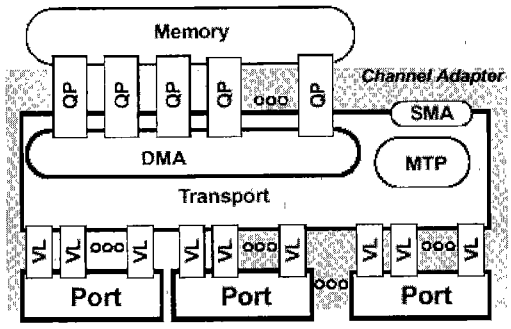


그림 4 채널 어댑터 개념도

포트를 가질 수 있으며, 각 포트는 하나의 LID (Local ID) 또는 특정 영역의 LID를 할당 받는다. 각 포트는 송신 및 수신을 동시에 수행할 수 있도록 독립적인 송신 버퍼와 수신 버퍼를 가지며, 독립적인 흐름제어에 의해서 동작하는 가상 레인(VL: Virtual Lane)을 통해서 데이터 송수신을 위한 채널을 구성한다.

채널 어댑터는 가상 주소와 물리 주소간의 변환 및 접근 권한을 관리하기 위한 MTP(Memory Translation & Protection) 메커니즘을 지원 하며, 이 기능을 통하여 채널 어댑터는 운영체제의 간섭 없이 메시지 버퍼로 사용되는 사용자 영역의 호스트 메모리를 직접 접근한다.

서브넷 매니저는 IBA 연결망의 형상 관리 및 채널 어댑터 형상을 제어하며, 각 포트에 LID를 할당한다. 이를 위해서 채널 어댑터는 서브넷 메니지먼트 에이전트(SMA: Subnet Management Agent)를 가지고 있으며, 이를 통해서 서브넷 매니저의 요청에 응답한다.

3.1.2 스위치와 라우터

IBA 연결망은 스위치와 라우터로 구성된다. IBA 연결망은 글로벌넷과 서브넷 두단계로 구성되며 각 단계에 따라서 경로 설정에 사용되는 식별자를 구분하고 있다. 스위치는 서브넷 연결망에 연결된 노드간의 경로 설정을 담당하고, 라우터는 글로벌넷 연결망을 구성하는 서브넷 연결망간의 경로 설정을 담당한다.

스위치는 서브넷을 구성하여 경로를 제공하는 장치로서 경로 설정 정보에 따라 패킷을 전달하는 기능을 수행하며, 패킷을 발생시키거나 직접 수신하지 않는다. 스위치는 내부에 패킷 포워딩을 위한

테이블을 구성하고 이를 이용하여 패킷의 목적지 LID에 따라서 패킷 경로 설정 및 전송을 수행한다. 패킷 포워딩 테이블은 서브넷 매니저에 의해서 구성되며 다중 경로가 존재할 경우 서브넷 매니저는 부하 분산 또는 오류에 의한 자동 경로 변경이 가능하도록 패킷 포워딩 테이블의 재설정을 수행한다.

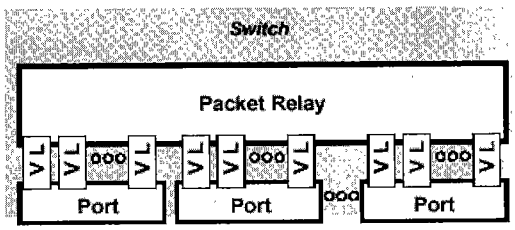


그림 5 스위치의 개념적 구조

스위치는 그림 5와 같이 복수개의 포트를 가지며, 하나의 패킷을 하나의 목적지로 전송하는 유니캐스트 전송을 지원한다. 사용자 구현 선택으로 하나의 패킷을 복수의 목적지로 전송하는 멀티캐스트를 구현할 수 있다.

라우터는 서브넷과 서브넷을 연결하는 장치로 기능 및 구조는 스위치와 유사하다. 패킷 경로 설정에 있어서 스위치와 달리 GID(Global Identification)를 사용한다.

3.1.3 관리 구성요소

IBA는 서브넷 매니저(SM: Subnet Manager)와 제너럴 서비스(General Service)로 구성되는 관리 형상을 제공하며, 각 매니저는 해당 에이전트와 MAD(Management Datagram) 패킷을 사용하여 연결망 관리 동작을 수행한다.

가. 서브넷 매니저

서브넷 매니저(SM)는 스위치, 라우터, 채널 어댑터등 IBA 구성 요소의 형상 관리를 수행하며, IBA 연결망상의 모든 노드는 SM과의 통신을 위해서 SMA(Subnet Management Agent)를 탑재하고 있어야 한다. SM의 기능은 다음과 같다.

- 서브넷 형상(Topology) 검출
- 채널 어댑터 포트에 LID, GID, Subnet-prefix, 망 분할 정보 할당
- 스위치에 LID, Subnet-prefix 할당, 패킷 포워드 테이블 구성
- 서브넷 관리 데이터베이스 구성 및 운영, LID 및 GID 주소 매핑 서비스

나. 서브넷 매니저 에이전트

모든 종단 노드는 SMA를 제공해야 한다. SMA는 SM이 SMI(Subnet Management Interface)를 통해서 접근하는 기능 모듈이다. SMI는 LID를 통한 경로 설정 방법과 직접 경로 설정 방법을 지원하며, 직접 경로 설정 방식은 스위치나 종단 노드가 초기화 되기 전에 MAD 패킷을 전송할 수 있게 한다.

다. 제너럴 서비스 에이전트

종단 노드는 SMA외에 부가적인 관리 형상 지원을 위하여 GSA(General Service Agent)를 가질 수 있으며, LID를 통한 경로 설정 방법을 사용하는 GSI(General Service Interface)를 통해서 통신한다. 제너럴 서비스의 종류는 다음과 같이 분류된다.

- Subnet Administration : SM에 의해서 제공되는 서비스로 노드에서 타 노드와 전송 서비스 정보를 찾고, 경로 경로 탐색 및 서비스 등록을 수행한다.
- Performance Manager : Performance Counter의 정보를 관리한다.
- Device Management : TCA 하단에 연결된 입출력 장치의 관리에 사용된다.
- Baseboard Management : 새시의 전기 기계적 환경을 관리한다.
- SNMP Tunneling : SNMP 프로토콜을 사용한 관리 환경에 사용된다.
- Vendor Defined : 개발자가 관리 형상을 확장하는데 사용한다.
- Communication Management : 종단 노드간에 연결 설정 및 해제를 수행한다.
- Device Configuration : 입출력 리소스 관리에 사용된다.

3.2 통신 방법 및 통신 스택

IBA 연결망에 연결된 노드는 그림 6과 같은 방식에 의해서 통신하게 된다. 사용자 프로그램은 메

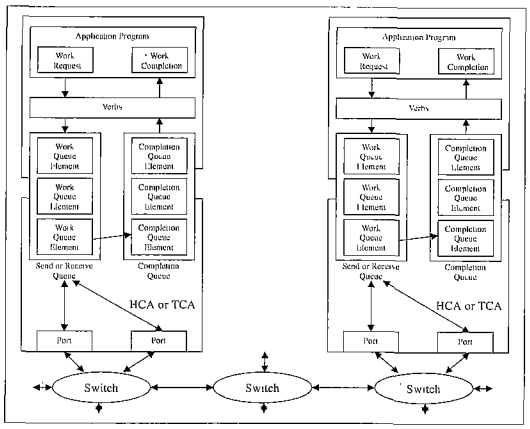


그림 6 InfiniBand 통신 방법

시지 패싱을 위한 메시지 전송 요구와 전송 종료 정보 취득을 Verb를 통해서 수행한다.

Verb를 통해서 생성된 메시지 전송 요구는 호스트 메모리에 위치한 작업 큐(송신큐와 수신큐로 구성되는 Queue Pair)에 FIFO(First-in First-out) 방식으로 저장된다. HCA 또는 TCA는 작업 큐의 내용을 검색하여 해당 메시지를 호스트 메모리에서 읽어내어 패킷으로 변환 후에 IBA연결망으로 전송하게 된다. 전송된 패킷은 목적지 노드에서 다시 메시지로 조립되어 작업 큐에 저장된다. 노드의 전송 요구가 작업 큐에 저장될 때는 해당 전송 요구의 전송 상태 정보를 관리하는 완료 큐를 하나 할당 받게 되며, 전송이 종료되면 해당 완료 큐의 내용에 전송 완료 정보를 기록한다. 노드는 전송 요구 후 전송 완료 여부를 완료 큐의 상태를 점검함으로써 전송의 완료 여부 및 상태를 알 수 있다.

각 노드는 메시지 전송을 위하여 사전에 호스트 메모리에 메시지 전송을 위해서 사용할 영역을 가상주소를 사용하여 지정해야 한다. 사용자 프로그램은 메시지 전송 요구 및 메시지 데이터를 지정된 영역에 위치시키게 되고, HCA 또는 TCA는 운영체제의 간섭없이 지정된 영역에서 메시지 전송 요구와 해당 데이터를 가상 주소를 사용하여 읽어내어 IBA연결망으로 전송한다. 따라서 메시지 전송에 있어서 통신 계층을 통과하며 발생하는 데이터 복제 현상을 제거하여 메시지 전송 처리 속도를 향상 시키고 있으며, 데이터 복제시에 발생하는 프로세서의 오버헤드도 제거한다.

IBA 연결망을 통한 통신은 송신/수신 프리미티브(Send/Receive primitive)를 사용하는 메시지 패싱 기법 이외에 RDMA (Remote Direct Memory Access) 전송을 제공하여대용량 데이터의 전송을 지원하고 있으며, 응용 프로그램의 특성에 따라서 메시지의 신뢰도와 대역폭을 조절할 수 있도록 다양한 형태의 전송 서비스를 제공한다. 표 2는 IBA 연결망에서 정의하고 있는 전송 서비스를 요약한 표이다.

표 2 IBA 전송 서비스 분류

	Reliable Connection	Reliable Datagram	Unreliable Datagram	Unreliable Connection	Raw Datagram
데이터 오류 검출	있음	있음	있음	없음	있음
패킷 전송 순서 확인	있음	있음	없음	없음	없음
패킷 전송 누락 확인	있음	있음	없음	있음	없음
오류 복구	있음	있음	전송 누락	오류 로깅	전송 누락
메시지 크기	제한 없음	제한 없음	단일 패킷	단일 패킷	제한 없음
연결 확인	확인	미확인	미확인	확인	미확인

3.3 InfiniBand 프로토콜 계층 구조

IBA 연결망을 구성하는 HCA, TCA 및 스위치와 라우터는 그림 7과 같은 통신 프로토콜 계층을 처리한다. HCA와 TCA는 Verb를 통하여 전달되는 메시지를 IBA 연결망에 패킷 단위로 전송해야 하므로, 메시지 전송 요구를 입력으로 받아서 IBA 패킷으로 변환하여 IBA 연결망으로 전기적 신호를 전송하는 기능을 수행해야 한다. 따라서 HCA와 TCA는 Transport, Network, Link, Physical 계층을 모두 가져야 한다. 스위치와 라우터는 Infini

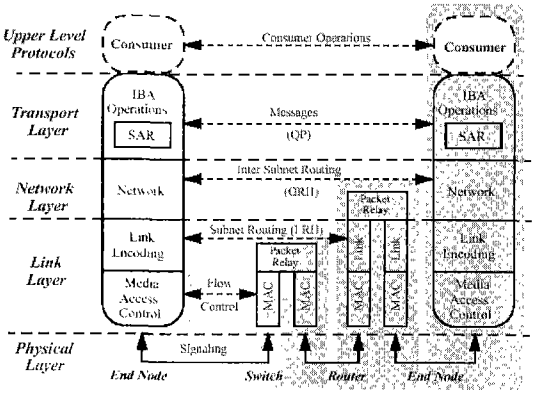


그림 7 InfiniBand 프로토콜 계층 구조

Band 연결망에서 Inifniband 패킷의 경로 설정 및 전송을 담당하며, Network, Link, Physical 계층으로 구성된다.

3.3.1 Physical 계층

Physical 계층은 연결 매체상에 비트로 표현되는 데이터 정보가 심볼로 구성되어 전달되는 계층으로 비트 전송 속도, 전송 매체, 신호 전달 방식 등을 정의한다. 8b/10b 코딩 방식에 의한 직렬 차동 신호 전송 방식을 사용하여 동축 케이블이나 광케이블 매체를 통해서 2.5Gbps 전송 속도를 제공하고 있다. 사용자의 성능 요구에 따라 포트 당 연결 매체의 개수를 1x, 4x, 12x 중에서 선택할 수 있게 한다.

3.3.2 Link 계층

Link 계층은 패킷 포맷과 패킷 단위 흐름제어, 서브넷 내부에서의 패킷 경로 설정을 포함하는 프로토콜을 정의한다. 패킷은 링크 관리 패킷(Link Management Packet)과 데이터 패킷으로 구분되며 각 패킷의 용도 및 기능은 다음과 같다.

• 링크 관리 패킷 : 포트와 포트간의 점대점 연결로 구성된 링크상에서 전송 속도, 전송 데이터 폭 등을 상호 협상하는 링크 트레이닝 동작에 사용된다. 또한 링크상에서 흐름제어 정보를 전달하고 링크의 무결성을 유지 관리한다.

• 데이터 패킷 : IBA 연결망을 통해서 전송 서비스 동작을 수행하기 위해서 사용하는 패킷으로 전송 서비스 종류에 따라서 다양한 헤더를 가질 수 있다. 서브넷 내부에서의 경로 설정을 위해서 기본적으로 LRH (Local Routing Header)를 가져야 한다. 오류 방지를 위해서 ICRC (Invariant CRC)와 VCRC (Variant CRC)를 사용하여 데이터 페이로드(payload)를 포함한 패킷 전체를 보호한다.

링크 흐름제어는 Credit-based 방식을 사용하며, 수신단은 송신단에게 가상 레인 별로 수신 가능한 데이터 패킷 수를 credit으로 알려주고 송신단은 credit을 확인한 후 전송 여부를 결정하게 된다.

3.3.3 Network 계층

Network 계층은 서브넷간 패킷 경로 설정 및 패킷 전달을 정의하는 계층으로 GRH(Global Routing Header)를 사용한다. GRH는 IPv6 어드레스 포맷으로 표현되는 GID를 사용하며 GRH의 내용을 사용하여 서브넷간 경로 설정 및 패킷 전달을 수행한다.

3.3.4 Transport 계층

Transport 계층은 데이터 전송 서비스의 종류별로 전송 동작을 정의하고, 전송 메시지를 패킷으로 분할 또는 재결합 기능을 수행한다. BTH (Base Transport Header)를 사용하여 목적지 QP (Queue Pair), PSN(Packet Sequence Number), 망분할 등과 같은 전송 서비스를 위한 기본 정보를 제공하고 전송 서비스 종류에 따라서 ETH (Extened Transport Header)를 사용하여 전송 서비스별 상세 정보를 제공한다.

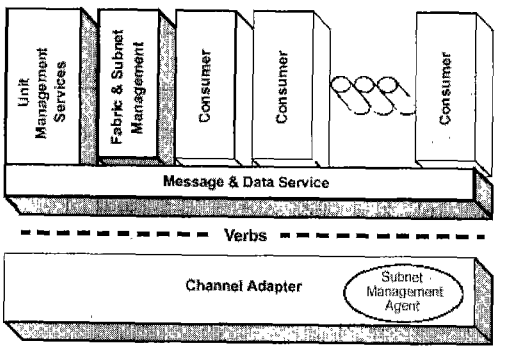


그림 8 상위 계층

3.3.5 상위 계층

상위 계층은 IBA 전송 서비스를 사용하는 사용자 프로그램(프로세스)과 관리 형상 프로그램(프로세스)으로 IBA 연결망 사용자에 해당하며, Verb를 통해서 IBA 연결망을 사용하거나 관리하는 작업을 수행한다. 그림 8은 상위 계층을 구성하는 프로그램이 Verb 인터페이스를 통하여 채널 어댑터와 연동되는 관계를 보여준다.

4. InfiniBand 기술 개발 동향

인텔, IBM 등 강력한 시장 지배력을 가지고 있는 IBTA 운영위원회 회원사를 주축으로 현재 InfiniBand 시제품 개발이 진행 중에 있는 것으로 알려지고 있다. 표 3은 현재 InfiniBand 관련 제품을 개발하고 있는 것으로 알려진 업체와 제품들을 요약한 것이다.

표 3 InfiniBand 관련 제품 개발 동향

업체명	제품	출시 예정 시기
인텔	HCA, TCA, Switch InfiniBand 서버	2001년 말 또는 2002년
IBM	HCA, TCA, Swtich	2001년 중
RedSwitch	Switch	2001년 말 또는 2002년
Agilent	SerDes(Serial/Deserialize, InifiniBand 계측장비	2001년 중
Rucent	SerDes, Switch	2001년 말 또는 2002년
Crossroad	Switch, InfiniBand 자료 저장 시스템	2001년 말 또는 2002년
Banderacom	HCA, TCA, Switch	2001년 말 또는 2002년
LANE15	InfiniBand 연결망 관리 소프트웨어	2001년 말 또는 2002년
LSI Logic	InfiniBand 설계용 ASIC IP	-
Adaptec	TCA(Storage 시스템용)	2001년 말 또는 2002년

표 3에서 언급된 회사들 외에도 Dell, Compaq, Sun 등 시스템 업체들은 InfiniBand 연결망을 사용한 서버, 자료 저장 시스템 등을 현재 개발하고 있으며, Adaptec, 3Com과 같은 입출력 장치용 제어기 생산 업체들은 자사용 입출력 제어기를 TCA로 재설계하고 있는 것으로 알려지고 있다. 그외에 현재 220여개의 회원사들도 InfiniBand 기술 개발에 관심을 기울이고 있는 것으로 알려지고 있다.

InifiniBand 관련 제품은 2001년 말 또는 2002년에 이르러 시장에 출시될 것으로 예측되고 있으며, 초기 시장 진입은 대규모 클러스터 구조를 갖는 인터넷 데이터 센터 시스템과 자료 저장장치 분야일 것으로 예상된다. 2000년 가을 인텔 개발자 포럼에서 발표된 IDC의 시장 예측에 의하면, 인텔 프로세서를 채택하는 서버를 기준으로 예측할 때 2001년부터 형성되는 InfiniBand 서버 시장은 전체 서버 시장의 10% 미만으로 예측되지만, 2003년에 이르러 40% 이상의 시장 점유율을 갖는 성장기를 거쳐 2004년에는 서버 시장의 80%를 차지하는 주력 제품이 될 것으로 예상하고 있다.

IBTA는 InfiniBand 표준 규격을 계속해서 발전시켜 갈 예정이며, 이와 더블어서 반기별로 계획되고 있는IBTA 개발자 회의를 통해서 선행 기술을 확보하고 있는 선도 기업들은 기술 공개 및 지원을 계속할 예정이다. 아울러서 InfiniBand 연결망 기술 확산을 위하여 각종 개발 환경을 제공할 예정이며, 특히 각 개발 업체들의 제품간에 호환성 및 상호 운영성 보장을 위한 CIWG(Compliance & Interoperability Working Group) 활동은 InfiniBand 기술 확산에 기폭제가 될 전망이다.

현재 국내에서는 ETRI를 중심으로 몇몇 벤처기업이 InfiniBand 기술 개발을 준비 중에 있지만, 아직 국내에는 널리 알려져 있지 않은 상태이다. 하지만 강력한 시장 지배력을 가진 기업군이 추진하는 기술인 만큼 그 파급 효과는 매우 클 것으로 예상되며, 국내에서도 InfiniBand 기술 개발을 더 이상 늦추어서는 안될 것으로 예상된다.

5. 결 론

인터넷의 발전 및 인터넷 서비스의 대중화는 서버의 기능을 연산능력 중심에서 입출력 처리 중심으로 변화시켰으며 고가용성과 높은 확장성을 요구하고 있다. 서버 기능의 변화 및 요구 사항의 변화를 수용하기 위하여 차세대 서버는 스위치 기반형 점대점 연결망을 사용하는 클러스터 구조로 변화하고 있으며, 이 큰 흐름을 주도하고 있는 것이 InfiniBand 기술이다.

InfiniBand 구조는 스위치 기반 점대점 연결망과 기존의 클러스터 컴퓨팅에서 사용되어진 채널 기반형 메시지 패싱을 접목한 형태로 클러스터 연결망과 스토리지 연결망을 통합한 시스템 연결망을 표방하고 있다. InfiniBand 표준화 기구인 IBTA에는 인텔을 비롯한 컴퓨터 분야 선진 대기업들이 대거 참여하여 선도 기술과 강력한 시장 지배력으로 InfiniBand 기술의 확산을 시도하고 있으며, 2000년 10월 표준 규격 발표에 이어 2001년 말부터는 이를 지원하는 칩셋이 상용화되기 시작할 것으로 예상되고 있다.

InfiniBand의 상용화 초기 제품은 현재 설치된 서버에서 InfiniBand 연결망을 사용할 수 있도록 지원하는 PCI-to-InfiniBand 브릿지 칩셋과 연결망을 구성하는 스위치로 예상되고 있으며, 이러한 칩셋을 이용한 인터넷 데이터 센터 시스템과 자료 저장장치 시스템이 초기 시장 진입 제품으로 예상된다.

InfiniBand의 상용화는 서버 구조의 진화를 의미하는 것이며, 아울러 현재 PCI 인터페이스로 제품화된 입출력 장치들이 InfiniBand 입출력 장치

로 서서히 대체될 것으로 예상된다. 서버 구조 역시 모듈 단위로 InfiniBand 연결망에 접속되는 형태로 구성될 것이며, 이를 통하여 입출력 기능이 향상되는 클러스터 컴퓨팅 기술이 서버 기술의 주류로 자리 잡을 것으로 예견된다.

참고문헌

[1] Borland's Golden Gate Architecture: Bridging Client/Server and Internet Technologies for Corporate Developers, http://www.inprise.com/about/papers/bgg/, Jan. 2001.

[2] Distributed Internet Server Array Architecture, Technical White Paper from Compaq, 1998, ftp://ftp.compaq.com/pub/supportinformation/papers/ecg0550498.pdf, Jan. 2001.

[1] InfiniBand Architectural Technology, Technical White Paper from Compaq, ftp://ftp.compaq.com/pub/supportinformation/papers/tc000702tb.pdf, June 2000.

[3] InfiniBand Trade Association Official Homepage, http://www.infinibandta.org, Jan. 2001.

[4] 김진미, 온기원, 김학영, 지동해, 클러스터 컴퓨팅 기술동향 , 전자통신동향분석, 제14권, 제1호, pp.1~12, 1999. 2.

[5] Virtual Interface Architecture for Cluster system, Technical White Paper from Dell, http://www.dell.com/downloads/global/vectors/via.pdf, 1998.

[6] ServerNet - A High Bandwith, Low Latency Cluster Interconnection, Technical White Paper from Compaq, ftp://ftp.compaq.com/pub/supportinformation/papers/tc000602wp.pdf, Aug. 1998.

[7] 김정환, 강회일, 이동일, SAN 기술 및 시장동향 , 전자통신동향분석, 제15권, 제1호, pp.24~37, 2000. 2.

[8] 박경, InfiniBand의 개요 , 주간기술동향, 통권967호, pp.13~22, 2000. 10.

[9] PCI-X : An Evalution of the PCI bus, TC990903TB, White Paper of Compaq Computer System Co., ftp://ftp.compaq.com/pub/supportinformation/papers/tc990903tb.pdf, Sep. 1999.

[10] Future I/O and Next Generation I/O Merge, http://www.infinibandta.org/press/merger.html, Aug. 1999.

[11] InfiniBand Technology Prototypes White Paper Spring 2000, Technical white paper from Intel, ftp://download.intel.com/design/servers/future_server_io/documents/Final_Whitepaperxx.pdf, Feb. 2000.

[12] InfiniBand Architecture: Next-Generation Server I/O, Technical White paper from Dell, http://www.dell.com/downloads/ global/vectors/infiniband.pdf, 2000.

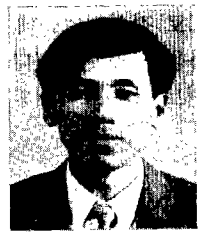
[13] Get on the Fabric: InfiniBand Fabric Prototype Demonstration White paper Fall 2000 IDF, Technical white paper from Intel, ftp://download.intel.com/design/servers/future_server_io/documents/get_on_fabric.pdf, Aug. 2000.

[14] To InfiniBand and Beyond, The Presentation From IBM, http://www. chips.ibm.com/products/infiniband/presentations/To_ InfiniBand_and_Beyond.pdf, 2000.

박 경

1991 전북대학교 컴퓨터공학 학사
1993 전북대학교 컴퓨터공학 석사
1993~현재 한국전자통신연구원 선임연구원
관심분야:컴퓨터구조, 마이크로프로세서구조, 병렬처리, 상호연결망
E-mail:kyoung@etri.re.kr

모 상 만



1991 연세대학교 대학원 컴퓨터과학과 졸업(석사)
1991~현재 한국전자통신연구원 병렬시스템연구팀장
1993 정보처리기술사
1998~현재 한국정보통신대학원(ICU) 박사과정(수료)
관심분야:컴퓨터구조, 병렬컴퓨팅, 클러스터컴퓨팅, ASIC설계
E-mail:smmoh@etri.re.kr