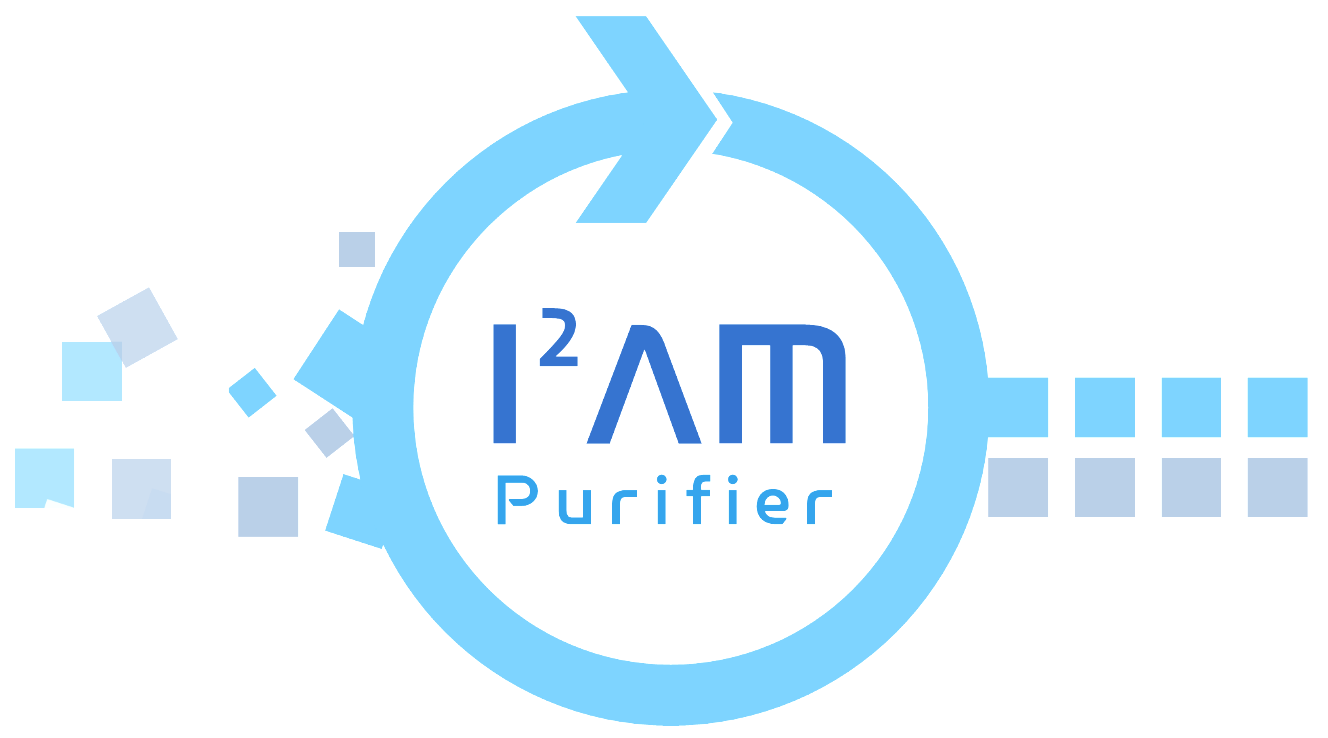
****

2016. 11.

강원대학교

인피니밴드 기술 동향 보고서

문서 정보

|  |  |
| --- | --- |
| **버 전** | 0.1 |
| **작성일** | 2016-11-15 |
| **상 태** | 🞎 완료 🞎 진행 중 🗹 초안 |
| **작성자** | 최성윤 (3팀) |
| **검토자** |  |
| **승인자** |  |

**목 차**

[1. 인피니밴드 개요 1](#_Toc467530759)

[1.1. 인피니밴드란? 1](#_Toc467530760)

[1.1.1. 인피니밴드 정의 1](#_Toc467530761)

[1.1.2. 인피니밴드 역사 2](#_Toc467530762)

[1.2. OFED란? 3](#_Toc467530763)

[1.2.1. OFED의 정의 3](#_Toc467530764)

[1.2.2. OFA란? 4](#_Toc467530765)

[1.3. 관련 기술 및 경쟁 기술 6](#_Toc467530766)

[1.3.1. iWARP 6](#_Toc467530767)

[1.3.2. RDMA over Converged Ethernet (RoCE) 7](#_Toc467530768)

[2. 인피니밴드 기술 7](#_Toc467530769)

[2.1. 인피니밴드 기술 설명 7](#_Toc467530770)

[2.1.1. 인피니밴드 7](#_Toc467530771)

[2.1.2. OFED 14](#_Toc467530772)

[2.2. 버전별 성능 15](#_Toc467530773)

[3. 인피니밴드 적용 사례 17](#_Toc467530774)

[3.1. 데이터 센터 17](#_Toc467530775)

[3.1.1. 오라클 빅데이터 솔루션 17](#_Toc467530776)

[3.2. 슈퍼컴퓨터 18](#_Toc467530777)

[3.2.1. www.top500.org 18](#_Toc467530778)

[3.3. 논문 사례 18](#_Toc467530779)

[3.3.1. The Ohio State University 18](#_Toc467530780)

[3.3.2. 인하대학교 19](#_Toc467530781)

[4. 구축 및 동작 확인 19](#_Toc467530782)

[4.1. 서버 구성 환경 19](#_Toc467530783)

[4.2. 성능 21](#_Toc467530784)

[4.2.1. 인피니밴드 동작 확인 21](#_Toc467530785)

[4.2.2. 인피니밴드 성능 측정 21](#_Toc467530786)

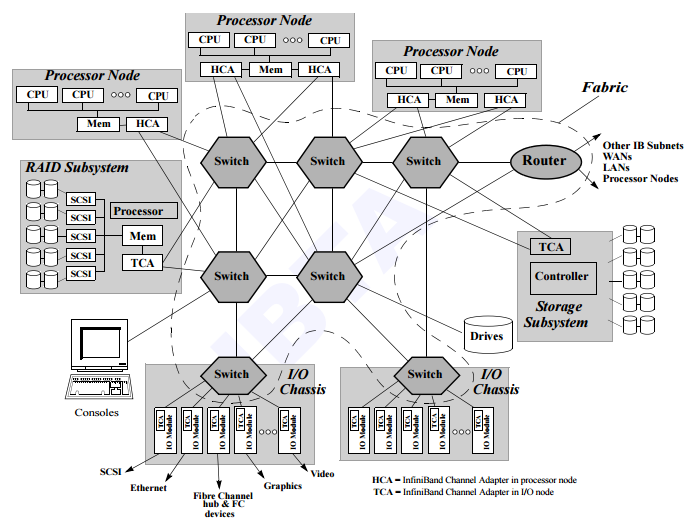
[5. 참고 문헌 23](#_Toc467530787)

# 인피니밴드 개요

## 인피니밴드란?

### 인피니밴드 정의

인피니밴드는 서버, 통신 장비, 스토리지, 임베디드 시스템들의 상호 연결에 사용하는 입출력 아키텍쳐를 정의하는 산업 표준 사양서이다. 높은 대역폭과 낮은 지연시간, 높은 안정성, 확장성을 특징으로 가지고 있다. 고성능 컴퓨팅과 기업용 데이터 센터에서 많이 사용하는 스위치 패브릭 방식의 통신 연결방식이다 (IBTA, 2016). 스위치 패브릭 방식을 사용하여 패킷을 수평적으로 전달할 수 있다. 노드와 스위치는 Channel Adapter (CA)라는 네트워크인터페이스 카드와 연결된다. Host Channel Adapter (HCA)는 계산노드에서 사용하며 Target Channel Adapter (TCA)는 입출력노드에서 사용한다. 그림 1은 인피니밴드 컨셉을 보여준다 (IBTA, 2007).



<그림 > IBA System Area Network

### 인피니밴드 역사

1999년에 Future I/O와 Next Generation I/O의 두 가지 기술이 병합되어 인피니밴드가 탄생했다. Compaq, Dell, Hewlett-Packard, IBM, Intel, Microsoft, Sun 등이 멤버로 가입한 인피니밴드 Trade Association (IBTA)에서 표준화를 이끈다. PCI bus를 이용한 상호 연결에는 병목현상이 발생하여 PCI-X를 이용한 인피니밴드 사양서 Version 1.0이 2000년에 처음으로 공개되었다. 2001년에 멜라녹스가 10Gbit/s 장비를 출시하였고, 2002년에 Microsoft는 인피니밴드 개발을 멈추었다. 2003년에 인피니밴드 기반의 클러스터가 그 해에 TOP500 프로젝트에서 3위를 기록하였다. 이 클러스터는 Virginia Tech에서 구축하였다. TOP500은 세계에서 비분산형 컴퓨터시스템을 1위부터 500위까지로 성능의 순위를 매기는 프로젝트이다. 2004년에 OpenFabrics Alliance에서 표준화된 리눅스 기반의 인피니밴드 스택을 개발함으로 리눅스가 인피니밴드를 지원하였다. 2009년에 TOP500에서 기가비트 이더넷은 259개에 설치되었고 인피니밴드는 181개에서 사용되었다. 2010년에 인피니밴드 시장을 주도하는 멜라녹스와 볼테어가 합병했고, 오라클은 멜라녹스에 대규모로 투자하였다. 2011년 International Supercomputing Conference에서 FDR 스위치가 발표되었다 (infiniband, 2016). 현재 IBTA에 가입한 멤버에는 "Advanced Photonics, Inc.", "Amphenol Interconnect Products", "Anritsu", "Applied Optoelectronics, Inc.", "Bay Microsystems", "Broadcom", "Bull SAS", "ConnPro Industries Inc.", "Cray, Inc.", "Finisar Corporation", "Foxconn Interconnect Technology, Ltd.", "Fujitsu Limited", "Genesis Technology USA, Inc.", "Hewlett-Packard", "IBM", "Intel Corporation", "Jess-Link Products Co., Ltd.", "Keysight Technologies, Inc.", "Lawrence Livermore National Laboratory (LLNL)", "LUXSHARE-ICT Co., Ltd.", "Mellanox Technologies, Inc.", "Microsoft", "Molex Inc.", "NetApp", "Oracle", "QLogic", "Samtec", "Semtech", "Shanghai Net Miles Fiber Optics Technology Co. Ltd.", "Software Forge, Inc.", "TE Connectivity", "Tektronix", "UNH InterOperability Lab", "Unisys Corporation", “Volex inc.", "Woodward McCoach, Inc.", "Yamaichi Electronics USA" 등이 있다. 그림 2는 IBTA를 주도하는 스티어링 위원회이다.

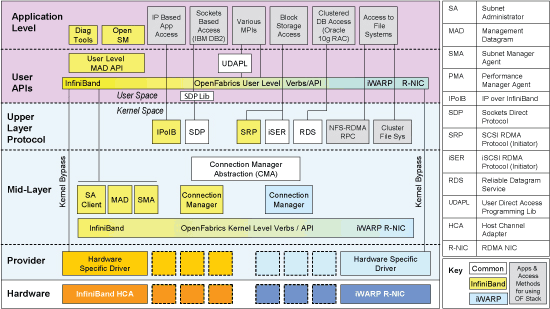


<그림 2> IBTA 스티어링 위원회

## OFED란?

### OFED의 정의

OFED(OpenFabrics Enterprise Distribution) 혹은 OFS(OpenFabrics Software)는 RDMA(Remote Direct Memory Access)와 커널 바이패스 애플리케이션을 위한 오픈소스 소프트웨어이다. OFED는 커널 레벨 드라이버, 채널 지향 RDMA send/receice 오퍼레이션, 커널 바이패스를 위한 유저 레벌 API, 병렬 처리 시스템인 MPI, 소켓 데이터 교환(RDS, SDP), file system/database system(iSER, NFS-RDMA, SRP)를 포함한다. 높은 효율의 네트워크, 스토리지 연결, 병렬 컴퓨팅 등을 필요로 하는 데이터 센터, HPC에서 사용된다. 이러한 OFED는 10 기가비트 이더넷, iWARP 기술, RDMA over Converged Ethernet (ROCE), 인피니밴드 등의 네트워크 기술에서 사용할 수 있다. OFED는 Red Hat Enterprise Linux (RHEL), Novell SUSE Linux Enterprise Distribution (SLES), Oracle Enterprise Linux (OEL), 마이크로소프트 Windows Server 등의 다양한 운영체제에서 사용이 가능하다 (OFA, 2016). 그림 3은 OFED 스택 구조를 보여준다. 2.1.2절에서 OFED에 대해 자세히 설명한다.



<그림 3> OFED 스택

### OFA란?

OpenFabrics Alliance (OFA)는 오픈소스 기반의 기관으로 OFED의 개발, 테스트, 라이선스 관리, 배포 등을 수행한다. 이 기관의 미션은 최소한의 CPU 부하로 최대의 대역폭, 낮은 지연시간 빠른 유선 메시징으로 고효율의 애플리케이션 개발을 가능하게 하는 것이다. 2004년 6월에 OpenIB Alliance로 벤더와 독립적으로 리눅스 기반의 인피니밴드 스택 개발을 목표로 시작되었다. 2005년에 마이크로소프트 Windows를 지원하였다. 2006년에 iWARP를 지원하여 기관을 확장하였고 2010년에 RoCE를 지원하기 시작했다. 멤버쉽의 종류는 Promoter, Adopter, Supporter, Academic, Individual, Consulting Participant로 나누어진다. Promoter는 가장 강한 영향력을 가진 것으로 가입비가 3000달러이고 매년 회비는 10000달러이다. Adopter는 기관에 개발과정에 공헌하고 참여하지만 강한 영향력을 가지고 있지 않다. 가입비는 3000달러이고 연회비는 5000달러이다. Supporter는 OFED를 사용하는 멤버들로 실제적으로 공헌하지는 않는다. 가입비는 3000달러이고 연회비는 1500달러이다. Academic은 교육기관들로 OFED에 공헌하고 개발 프로세스, 테스팅, 문제 해결의 역할을 수행하는 멤버다. 연회비는 2000달러이다. Individual은 OFED 사용자, 개발자들로 이루어진 멤버로 연회비는 200달러이다. Consulting Participant는 OFED 기반의 기술에 공헌한 기관, 회사들로 명예멤버로 연회비는 없다. 다음 그림 4, 그림 5, 그림 6, 그림 7은 OFA 멤버쉽 명단을 보여준다 (OFA, 2016).



<그림 > OFA Promoters 멤버



<그림 > OFA Adopters 멤버



<그림 > OFA Supporters 멤버



<그림 7> OFA Consulting Participants 멤버

## 관련 기술 및 경쟁 기술

### iWARP

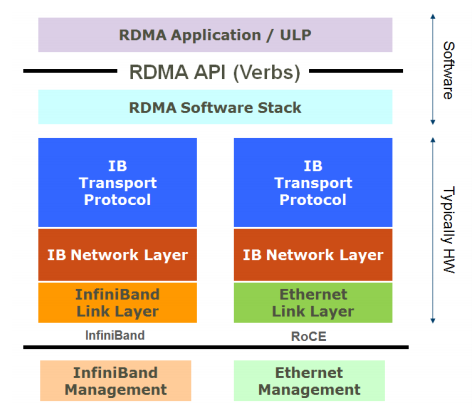
iWARP는 Internet 프로토콜 네트워크에서 효율적인 데이터 전송을 위한 RDMA 네트워크 프로토콜이다. 인피니밴드와 경쟁기술로 자리잡고 있다. IETF에 RFC 문서가 등록되어있다. iWARP를 정의하는 RFC 문서는 표 1에서 보여준다.

<표 > iWARP에 관한 RFC 문서

|  |  |
| --- | --- |
| RFC 번호 | RFC 내용 |
| RFC 5040 | A Remote Direct Memory Access Protocol Specification |
| RFC 5041 | Direct Data Placement over Reliable Transports |
| RFC 0543 | Stream Control Transmission Protocol (SCTP) Direct Data Placement (DDP) Adaptation |
| RFC 5044 | Marker PDU Aligned Framing for TCP Specification |
| RFC 5042 | Direct Data Placement Protocol (DDP) / Remote Direct Memory Access Protocol (RDMAP) Security |
| RFC 6580 | IANA Registries for the Remote Direct Data Placement (RDDP) Protocols |
| RFC 6581 | Enhanced Remote Direct Memory Access (RDMA) Connection Establishment |
| RFC 7306 | Remote Direct Memory Access (RDMA) Protocol Extensions |

### RDMA over Converged Ethernet (RoCE)

RDMA는 CPU 개입 없이 응용프로그램 메모리 간에 서버 간 데이터 이동을 실현하는 기능이다.이더넷 네트워크에서 RDMA를 사용가능하게 해주는 메커니즘으로 인피니밴드 기반의 애플리케이션을 이더넷에서 사용가능하게 해준다. 이더넷에서 낮은 지연시간과 효율적인 데이터 전송을 가능하게 한다. IBTA에서 표준화 작업 및 관리하고 있다. 그림 8에서 RoCE 프로토콜 스택을 보여준다 (IBTA, 2014).



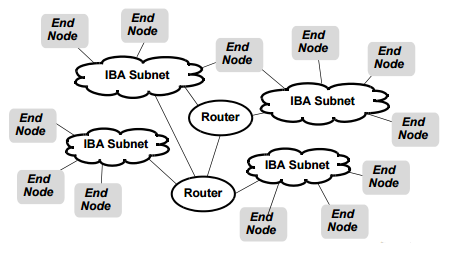
<그림 8> 인피니밴드와 RoCE 프로토콜 스택

# 인피니밴드 기술

## 인피니밴드 기술 설명

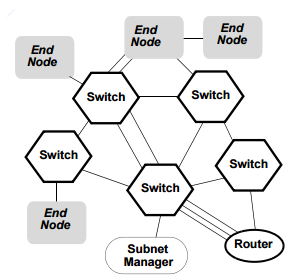
### 인피니밴드

인피니밴드는 크게 스위치 패브릭 구조의 특징과 RDMA 오퍼레이션을 통해 커널을 거치지 않아 CPU 부하를 낮추는 특징이 있다. 이를 통해 인피니밴드는 낮은 지연시간과 높은 대역폭을 지원한다. 또한 물리적 한 개의 링크를 Virtual Lane (VL) 0~ VL 15로 논리적으로 나누어 우선순위가 높은 것을 VL15, 가장 낮은 것을 VL0으로 패킷을 보내어 QoS를 보장한다 (IBTA, 2007). 인피니밴드 서브넷은 라우터를 통해서 연결을 한다. 그림 9는 인피니밴드의 네트워크 토폴로지를 보여준다.



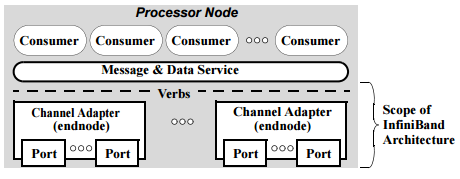
<그림 >IBA 네트워크 컴포넌트

한 개의 서브넷 안에는 여러 개의 스위치가 존재할 수 있으며 반드시 한 개의 서브넷매니저가 필요하다. 스위치, 라우터, 노드 간의 링크가 여러 개가 존재할 수 있다. 또한 토폴로지 구조가 이더넷의 계층적 구조가 아닌 패브릭 방식으로 이루어진 것을 알수있다. 기존의 계층적 구조의 이더넷은 서버간의 통신에서 상위 스위치 장비를 거쳐 수직적으로 패킷이 전달되었다. 패브릭 방식은 패킷이 수평적으로 전달되어 안정적인 메시지 전달을 할 수 있다. 그림 10은 인피니밴드 서브넷의 토폴로지를 보여준다.



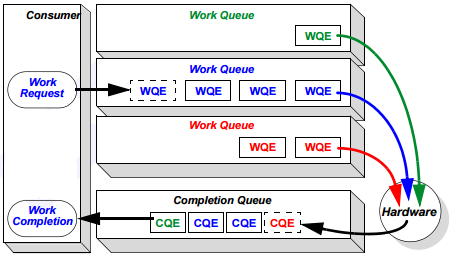
<그림 >IBA 네트워크 서브넷 컴포넌트

각 노드에는 Channel Adapter를 통해 인피니밴드 네트워크와 연결한다. 노드의 애플리케이션에서 IBA Verbs를 통해서 설정, 관리, 오퍼레이션을 할 수 있는 API 이다. 그림 11은 노드의 구조도이다.



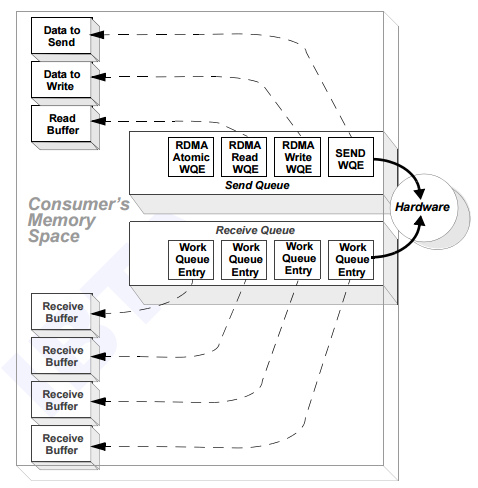
<그림 > 각 노드의 구조도

통신에서 오퍼레이션들은 큐잉 모델을 통해서 이루어진다. 오퍼레이션을 담는 큐를 work queue라고 하며 만들어 질 때 기본적으로 쌍으로 만들어진다. 이것을 Queue Pair (QP)라고 한다. 한 개는 다른 소비자에게 보내는 오퍼레이션을 위한 것이고 한 개는 다른 소비자에게서 받는 오퍼레이션을 위한 것이다. 소비자는 Work Request를 통해서 지시하면 Work queue에 Work Queue Element (WQE) 를 넣는다. 그리고 Channel Adapter에서 처리하면 Completion Queue에 Completion Queue Element (CQE) 를 넣는다. 그림 12에서 큐잉 모델을 보여준다.

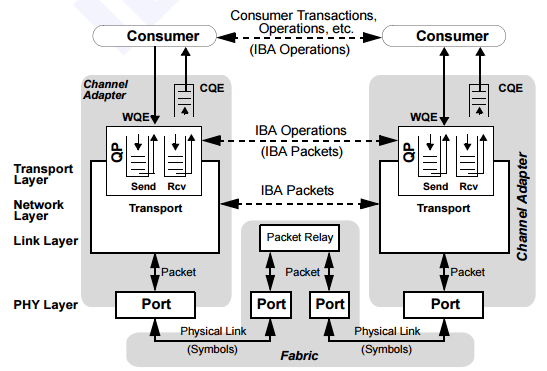


<그림 > 소비자 큐잉 모델

인피니밴드에서 오퍼레이션은 SEND, RDMA, MEMORY BINDING 등이 있다. SEND 오퍼레이션은 소비자의 메모리에서 데이터 블록을 네트워크를 통해 목적지 소비자에게 데이터를 전송하는 WQE를 만든다. RDMA 오퍼레이션은 RDMA-WRITE, RDMA-READ, ATOMIC의 3가지 타입으로 WQE를 만든다. RDMA-WRITE는 소비자의 메모리에서 원격 소비자의 메모리로 데이터를 전송한다. RDMA-READ는 원격 메모리에서 소비자의 메모리로 데이터를 전송한다. ATOMIC는 원격 메모리의 특정 위치의 값을 읽어오고 다시 그곳에 값을 수정 또는 대체하는 오퍼레이션이다. MEMORY BINDING 오퍼레이션은 소비자가 전송할 때 사용할 수 있는 Memory Region을 설정할 수 있다. 이 메모리 영역에는 R\_KEY가 생성되는데 이것은 RDMA 오퍼레이션을 수행할 때 권한이 있는지 판단하기위해 사용한다. 그림 13은 Work Queue 오퍼레이션을 보여주고 그림 14는 인피니밴드 통신 구조를 보여준다.



<그림 > Work Queue 오퍼레이션



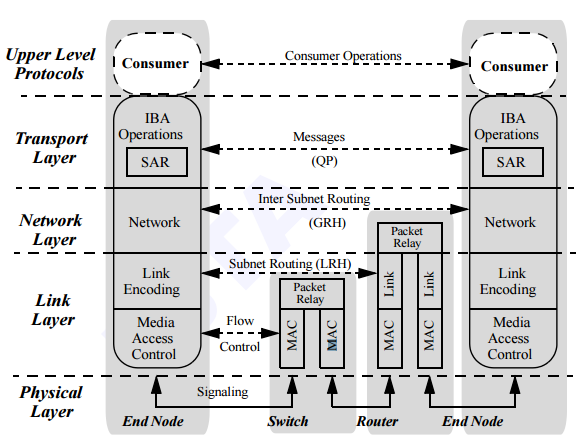
<그림 14> 인피니밴드 통신 구조

종단 간의 노드에서 각각의 QP는 같은 서비스 타입으로 만들어진다. Connection Oriented와 Datagram의 차이는 한 노드의 QP가 다른 노드의 QP와 연결이 맺어지는 것이 Connection Oriented 이고 Datagram은 한 노드의 QP에서 어떤 노드의 QP던지 송신 혹은 수신이 가능하다. Acknowledged는 신뢰성이 있는 데이터 통신을 위해 ACK 메시지를 응답으로 보내는 것을 말한다. 표 2는 인피니밴드 서비스 타입을 보여준다.

<표 > 인피니밴드 서비스 타입

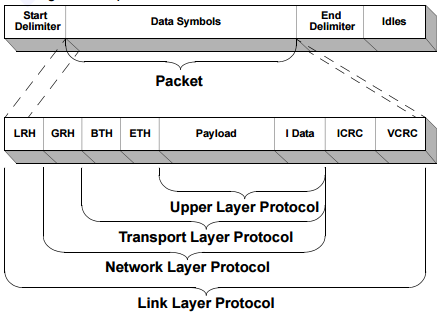
|  |  |  |  |
| --- | --- | --- | --- |
| Service Type | Connection Oriented | Acknowledged | Transport |
| Reliable Connection | Yes | Yes | IBA |
| Unreliable Connection | Yes | No | IBA |
| Reliable Datagram | No | Yes | IBA |
| Unreliable Datagram | No | No | IBA |
| RAW Datagram | No | No | Raw |

인피니밴드 프로토콜 스택은 물리 계층, 링크 계층, 네트워크 계층, 전송 계층, 상위 계층으로 나누어 진다. 그림 15는 인피니밴드 프로토콜 스택을 보여준다.



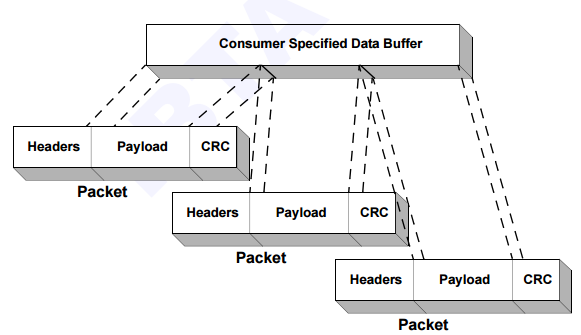
<그림 15> 인피니밴드 프로토콜 스택

물리 계층은 데이터의 전송을 비트 단위로 전송한다. 링크 계층은 링크 관리 패킷과 데이터 패킷으로 나누어 진다. 링크 관리 패킷은 링크를 만들고 유지하는 오퍼레이션을 위해 사용된다. 데이터 패킷은 인피니밴드 오퍼레이션, 데이터 등을 전달할 때 사용한다. 그림 16은 데이터 패킷의 포맷을 보여준다.



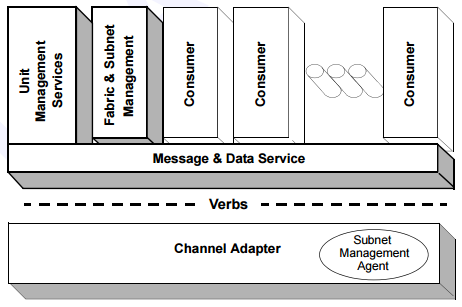
<그림 > 인피니밴드 데이터 패킷 포맷

Local Route Header (LRH)는 로컬 출발지 포트에서 로컬 목적지 포트의 전송에 관한 부분을 담당하며 스위치가 전달하여준다. LRH 안에는 LID 정보가 들어있는데 LID는 서브넷 매니저가 포트별로 할당하여 준다. 또한 ICRC와 VCRC를 통해 데이터의 무결성을 검사한다. 네트워크 레이어는 서브넷 간의 데이터 전송을 위해 존재한다. Global Route Header (GRH)가 패킷에 존재한다. GID라는 IPv6 형식의 주소가 헤더 안에 존재하여 출발지에서 목적지로 전송한다. 라우터는 GRH를 기반으로 패킷은 전달한다. 전송 계층에서는 Maximum Transfer Unit (MTU)에 따라 데이터를 나누어서 전송하고 다시 합치는 역할을 수행한다. Base Transport Header에는 오퍼레이션 코드, 패킷 넘버 등의 정보가 들어 있다. 그림 17은 전송 계층에서 데이터를 나누고 합치는 것을 보여준다.



<그림 > 전송계층에서 데이터 나누기/합치기

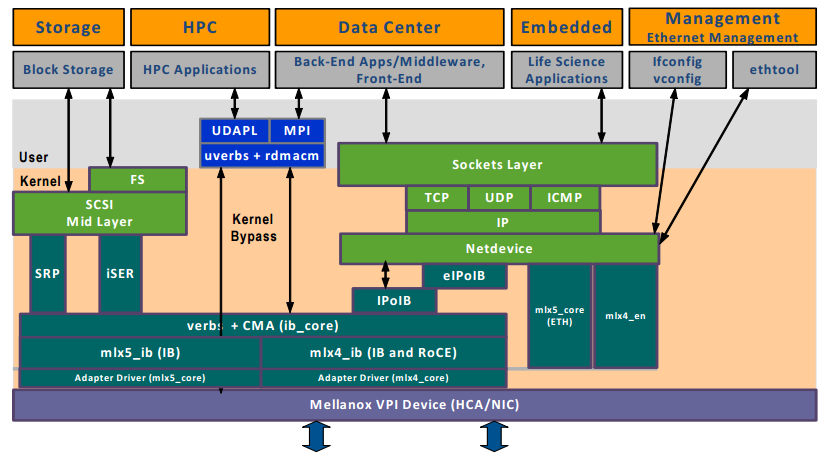
상위 계층에서는 다양한 소비자들이 존재한다. 소비자들은 Verbs를 통해서 데이터를 송수신할 수 있다. 그림 18은 상위 계층의 구조를 보여준다 (IBTA, 2007).



<그림 > 상위 계층

### OFED

인피니밴드를 사용하기 위해서는 OFED를 설치해야한다. OFED에는 드라이버까지 포함되어서 배포된다. 그림 19는 멜라녹스 OFED 스택으로 상위 계층 프로토콜이 어떻게 자원을 사용하는지 보여준다.



<그림 > 멜라녹스 OFED 스택

mlx4\_ib, mlx4\_ib는 장비의 드라이버 역할을 수행한다. 상위 계층으로는 IPoIB가 있다. IPoIB는 인피니밴드 위에서 IP 데이터그램을 송수신 가능하게 해주는 것으로 쉽게 인피니밴드를 이용할 수 있다는 장점이 있다. IPoIB는 두종류의 MTU가 존재한다. Datagram mode의 경우는 MTU 사이즈가 4KByte이고 Connected Mode는 MTU가 64KByte이다. iSCSI Extensions for RDMA (iSER) 은 iSCSI를 RDMA로 확장한 것으로 데이터를 SCSI 버퍼에서 데이터 카피가 없이 바로 전송이 가능하게 해준다. User Direct Access Programming Library (uDAPL)은 표준 API로 인피니밴드, RoCE에서 높은 안정성, 확장성, 고성능 메시지 전송을 위해 사용한다. Message Passing Interface (MPI)는 병렬 소프트웨어 개발을 가능하게 해주는 라이브러리로 병렬 컴퓨터, 클러스터, 헤테로지니어스 네트워크 등에서 사용한다. 멜라녹스 OFED는 MPI를 포함하고 있다. Open MPI로 오픈소스 MPI-2 버전을 구현한 것이다 (Mellanox, 2016). OFED 소프트웨어 배포는 크게 Channel Adaapter 드라이버, 중간 계층의 Verbs, CM, uVerbs 등, ULPs의 IPoIB, SRP 등, MPI, OpenSM(인피니밴드 서브넷 매니저), 진단 프로그램, 성능 테스트 프로그램, 펌웨어 툴 등이 포함되어 있다.

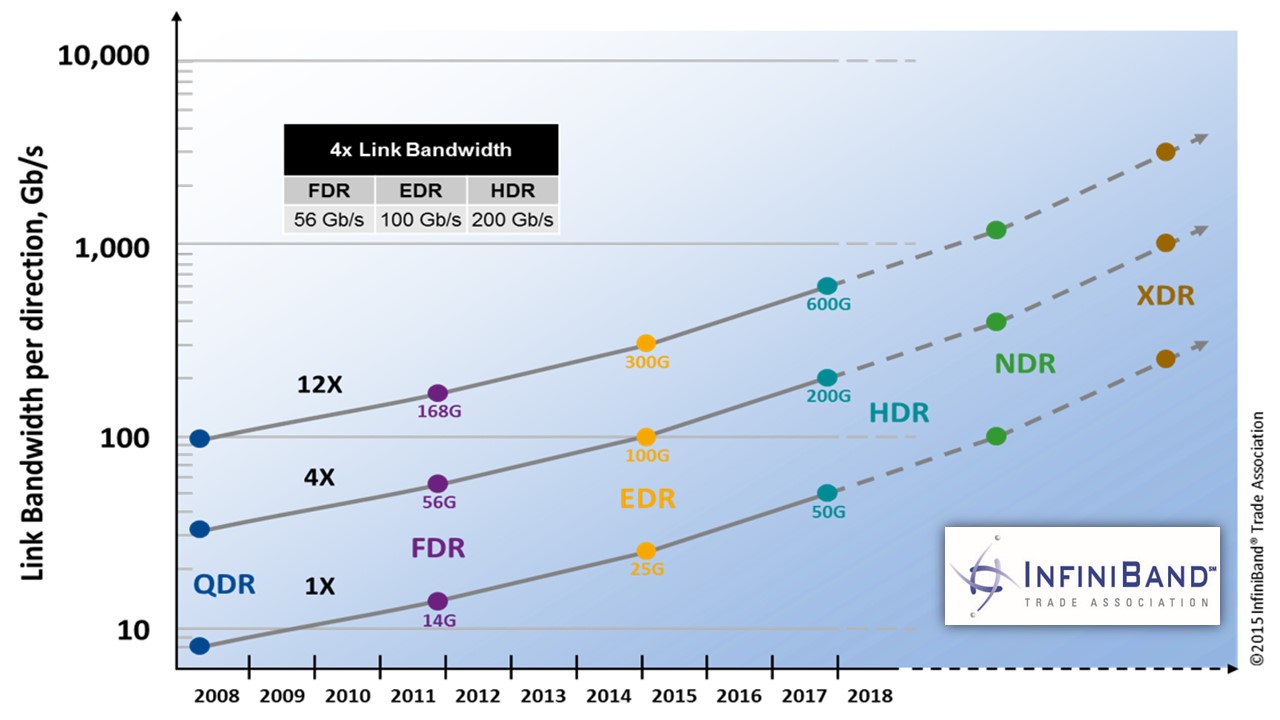
## 버전별 성능

인피니밴드는 시그널 속도에 따라 SDR, DDR, QDR, FDR-10, FDR, EDR이라고 버전을 부른다. 표 3에서 자세한 속도를 확인할 수 있다.

<표 > 인피니밴드 속도 버전

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | SDR | DDR | QDR | FDR-10 | FDR | EDR |
| 1X | 2 Gbit/s | 4 Gbit/s | 8 Gbit/s | 10 Gbit/s | 13.64 Gbit/s | 25 Gbit/s |
| 4X | 8 Gbit/s | 16 Gbit/s | 32 Gbit/s | 40 Gbit/s | 54.54 Gbit/s | 100 Gbit/s |
| 12X | 24 Gbit/s | 48 Gbit/s | 96 Gbit/s | 120 Gbit/s | 163.64 Gbit/s | 300 Gbit/s |

인피니밴드는 시그널링 속도와, 링크 수로 전송 속도가 결정된다. 링크 수는 케이블 자체에 와이어 수를 의미한다. 와이어 수는 기본 1x는 4개이고 4x는 16개, 12x는 48개이다. 지금 현재 최신 인피니밴드 시그널링 속도는 EDR(Enhanced Date Rate)로 4x 기준으로 100Gbps이고 12x라면 300Gbps이다. 일반적으로 4x가 많이 쓰인다 (PfisterG., 2001). 속도 버전에 대한 개발은 IBTA 워킹그룹이 주관하고 있다. 그림 20은 IBTA에서 제공하는 인피니밴드 개발 로드맵이다 (IBTA, 2016).



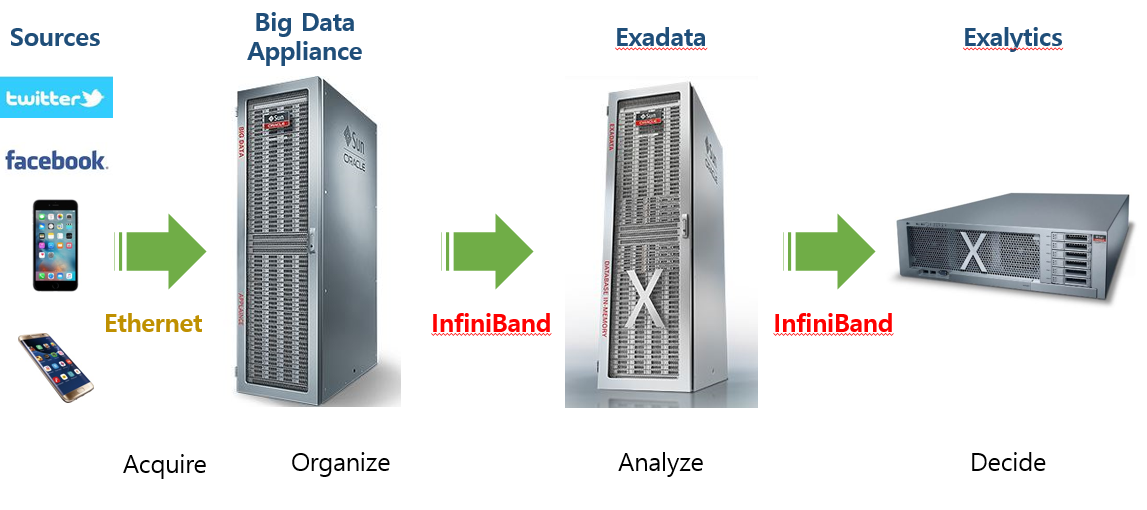
<그림 > 인피니밴드 개발 로드맵

# 인피니밴드 적용 사례

## 데이터 센터

### 오라클 빅데이터 솔루션

오라클은 인피니밴드 관련 네트워크 제품들을 판매하고 있다. 또한, 인피니밴드 기반의 솔루션, 서비스들을 판매한다. 오라클 빅데이터 솔루션은 데이터를 수집하여 정제하고 분석하여 저장하는 것이다. 다음 그림 21은 오라클 빅데이터 솔루션을 나타낸다.



<그림 > 오라클 빅데이터 솔루션

빅데이터 어플라이언스는 Hadoop을 탑재하여 빅데이터 수집, 저장 처리를 담당하는 제품이다.데이터 연결폭 한계를 극복하기 위해 인피니밴드 네트워크를 사용한다. 엑사데이터 베이스 머신은 오라클 데이터베이스 소프트웨어를 위해 컴퓨트시스템과 스토리지 시스템을 결합한 것이다. 스케일 아웃 컴퓨트 서버와 스케일 아웃 스토리지 서버, 인피니밴드, 특화된 소프트웨어가 결합된 것을 엑사데이터라고 한다. 이 제품을 사용하는 기업들은 다음과 같다 (Oracle, 2016).

* 신세계 푸드 – 수불데이터 집계, 분석을 위해 사용한다.
* SK Telecom – 실시간 과금 신뢰도 및 업무효율성 향상에 사용한다.
* 친애저축은행 – 기업의 정보계 시스템과 개인 계정계 시스템에 사용한다.
* IBK – 온라인 금융 검래 지원 시스템 속도 향상과 전산시스템 통합에 사용한다.

## 슈퍼컴퓨터

### www.top500.org

1993년에 시작된 프로젝트로 세계에서 가장 강력한 비분산 컴퓨터 시스템을 1위부터 500위까지 순위를 매긴다. 일년에 두번 업데이트를 한다. 첫번째는 International Supercomputing Conference에서 6월에 업데이트하고 두번째는 ACM/IEEE Supercomputing Conference에서 11월에 업데이트를 한다. 2016년 11월 기준, 전세계 슈퍼컴퓨터 500위에서 187개 슈퍼컴퓨터가 인피니밴드를 사용하여 인피니밴드가 가장 큰 비중을 차지한다. 표 4는 슈퍼컴퓨터 500위에서 사용한 네트워크를 종류별로 나타낸다 (TOP500.org, 2016).

<표 > TOP500 네트워크 종류별 사용 비율

|  |  |
| --- | --- |
| 이름 | 비율 |
| Infiniband | 37.4 % |
| 10G Ethernet | 35.6 % |
| Custom Interconnect | 14.2 % |
| Gigabit Ethernet | 5.6 % |
| omnipath | 5.6 % |
| 기타 | 1.6 % |

## 논문 사례

### The Ohio State University

#### 논문

* Islam, N. S., Rahman, M. W., Jose, J., Rajachandrasekar, R., Wang, H., Subramoni, H., ... & Panda, D. K. (2012, November). High performance RDMA-based design of HDFS over InfiniBand. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* (p. 35). IEEE Computer Society Press.
* Wasi-ur-Rahman, Md, et al. "High-performance rdma-based design of hadoop mapreduce over infiniband." *Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2013 IEEE 27th International*. IEEE, 2013.
* Lu, Xiaoyi, et al. "High-performance design of Hadoop RPC with RDMA over InfiniBand." *Parallel Processing (ICPP), 2013 42nd International Conference on*. IEEE, 2013.

#### 연구 내용

인피니밴드 네트워크에서 높은 성능의 MPI 구현하여 이를 바탕으로 Hadoop HDFS, Hadoop MapReduce 등에 적용하여 성능을 향상시키는 연구를 주로 한다. 또한, 클러스터, High Performance Computing에 관한 연구를 많이 한다.

### 인하대학교

#### 논문

* 허희성, 이광수, 메히디, & 김덕환. (2015). 인피니밴드 스토리지 네트워크를 적용한 오픈스택 클라우드 스토리지 시스템의 설계 및 스토리지 가상화 성능평가. *정보과학회 컴퓨팅의 실제 논문지*, *21*(7), 470-475.

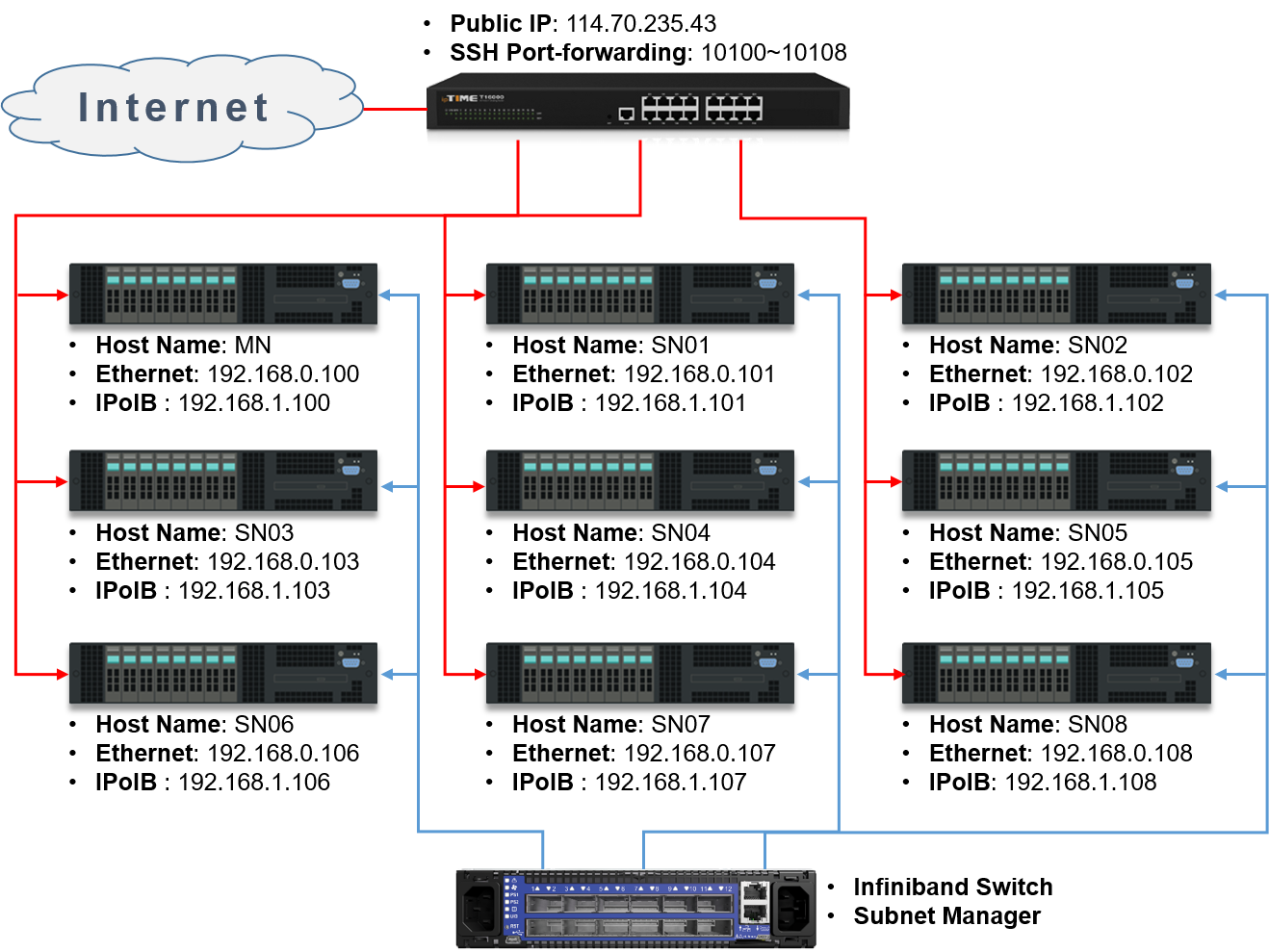
#### 연구내용

입출력 연산 시 서버와 스토리지 노드 간에 발생할 수 있는 병목현상을 해결하기 위해 고속 통신을 지원하는 인피니배늗 스토리지 네트워크를 적용한 오픈스택 클라우드 스토리지 시스템의 구조를 제안하고, 오픈스택에 적용한 인피니밴드 스토리지 네트워크가 고성능 스토리지 가상화에 적합함을 연구한다.

# 구축 및 동작 확인

## 서버 구성 환경

본 과제에서는 마스터노드 한 개와 데이터 노드 8개로 클러스터를 구성하였다. 네트워크는 1Gb/s 이더넷 공유기와 56Gb/s 인피니밴드 스위치를 사용하여 구축하였다. 인피니밴드의 IP는 IPoIB의 사용을 위한 IP주소이다. 인피니밴드 스위치는 서브넷 매니저 기능을 같이 수행한다. 이더넷 공유기를 통해 Internet과 연결되고 외부와 통신이 가능하다. 그림 22는 네트워크 구성도를 보여준다.



<그림 22> 네트워크 구성도

CPU가 Master Node는 8 Core이고 Slave Node는 6 Core인 점을 제외하고 사양은 모두 같다. 각 노드의 사양은 표 5에서 보여준다.

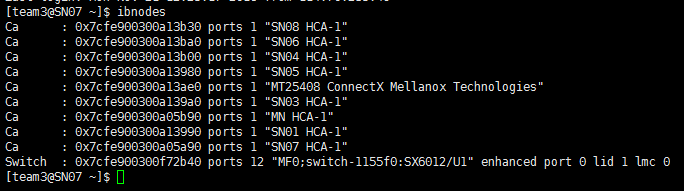
<표 > 노드 사양

|  |  |
| --- | --- |
| CPU | Xeon E5-2630V3 2.5GHz, 8 Core(Master Node) / 6 Core(Slave Node) |
| Memory | PC4-1700 REG ECC 32GB |
| SSD | 256G SSD 850 PRO |
| HDD | 기업용 1TB ST1000NM0033 |
| NIC | Gbit LAN on Board \* 2 / ConnectX-3 adapter card, FDR IB (56Gb/s) |

## 성능

### 인피니밴드 동작 확인

인피니밴드 네트워크가 올바르게 동작하고 있는지 확인이 가능한 명령어는 OFED에 포함되어 있는 ibnodes이다. 현재 인피니밴드 네트워크에 연결되어 있는 노드를 알 수 있다. 스위치와 노드 9개가 확인 가능함으로 인피니밴드 네트워크에 모두 접속했음을 알 수 있다. 이 명령어의 결과화면은 그림 23과 같다.



<그림 23> ibnodes 명령어 결과

### 인피니밴드 성능 측정

#### qperf 명령어

한쪽 노드에서는 qperf 서버를 동작하고 다른 노드에서는 qperf 클라이언트 역할을 수행하여 대역폭, 지연시간을 측정하는 명령어이다. tcp\_bw, tcp\_lat는 이더넷 성능과 IPoIB의 성능을 측정하는 것으로 IP주소를 다르게 사용하면 된다. 이 명령어는 OFED에 포함되어 있다. 표 6에서 측정할 테스트 목록을 보여준다.

<표 > qperf 테스트 목록

|  |  |  |
| --- | --- | --- |
| 통신 종류 | 테스트 이름 | 설명 |
| 소켓 기반 | tcp\_bw | TCP 대역폭 측정 |
| tcp\_lat | TCP 지연시간 측정 |
| RDMA Send/Receive | rc\_bw | RC 전송 대역폭 측정 |
| rc\_lat | RC 전송 지연시간 측정 |
| RDMA | rc\_rdma\_read\_bw | RC RDMA read 오퍼레이션 대역폭 측정 |
| rc\_rdma\_read\_lat | RC RDMA read 오퍼레이션 지연시간 측정 |
| rc\_rdma\_write\_bw | RC RDMA write 오퍼레이션 대역폭 측정 |
| rc\_rdma\_write\_lat | RC RDMA write 오퍼레이션 지연시간 측정 |

이더넷 tcp\_bw는 1 GbE 이더넷을 사용한 것이고 IPoIB tcp\_bw는 IPoIB tcp\_bw는 OFED에서 제공하는 IPoIB를 사용한 것이다. RDMA를 사용했을 때 대역폭이 약 6.28 GB/s로 가장 크고, RDMA READ에서 지연시간이 4.94 μs로 가장 낮다. 표 7은 qperf를 이용한 성능 측정 결과이다.

<표 > qperf 성능 측정 결과

|  |  |  |  |
| --- | --- | --- | --- |
| 이더넷 tcp\_bw | 118 MB/s | 이더넷 tcp\_lat | 30.6 μs |
| IPoIB tcp\_bw | 2.52 GB/s | IPoIB tcp\_lat | 10.4 μs |
| rc\_bw | 6.28 GB/s | rc\_lat | 7.44 μs |
| rc\_rdma\_read\_bw | 6.29 GB/s | rc\_rdma\_read\_lat | 4.94 μs |
| rc\_rdma\_write\_bw | 6.29 GB/s | rc\_rdma\_write\_lat | 6.98 μs |

#### 파일 전송 프로그램

파일 전송 프로그램을 작성하여서 파일 전송 속도를 측정할 때 사용하였다. 클라이언트에서 서버에게 파일을 전송하고 서버에서는 파일을 받으면서 속도를 측정하는 프로그램이다. 프로그램의 간략한 코드는 표 8에서 보여준다.

<표 > 파일 전송 프로그램

|  |  |
| --- | --- |
| 서버 | 클라이언트 |
| int main()  {  clnt\_sock=accept();  fd = open();  clock\_gettime();  while( (str\_len=read(clnt\_sock)) != 0)  {  write(fd);  total += str\_len;  }  close(fd);  clock\_gettime();  } | main()  {  sock=socket();  connect();  fd = open(argv[3], O\_RDONLY);  while( (str\_len=fileRead(fd)) > 0)  {  networkWrite(sock);  }  close();  } |

파일 전송은 1GB 더미 파일을 전송하였다. 네트워크가 이더넷과 IPoIB에서 파일시스템을 각각HDD, SSD, RamDisk에서 전송 속도를 측정하였다. IPoIB, RamDisk를 사용했을 때 1376 MB/s 속도로 가능 높다. 성능 결과는 표 9에서 보여준다.

<표 > 파일 전송 속도 결과

|  |  |  |  |
| --- | --- | --- | --- |
|  | HDD | SSD | RamDisk |
| Ethernet | 112 MB/s | 112 MB/s | 112 MB/s |
| IPoIB | 153 MB/s | 357 MB/s | 1376 MB/s |

# 참고 문헌

IBTA. (2007년 11월). Infiniband Architecture Specificaiton Volume 1. Release 1.2.1.

IBTA. (2014년 10월). Supplement to InfiniBand Architecture Specification Volume 1 Relaese 1.2.1 Annex A17: RoCEv2.

IBTA. (2016년 11월 18일). “About Infiniband”. infinibandta.org.에서 검색됨

“infiniband”. (2016). wikipedia: https://en.wikipedia.org/wiki/InfiniBand에서 검색됨

Mellanox. (2016). Mellanox OFED for Linux User Manual Rev 3.40.

OFA. (2016). “OFA Overview”. openfabrics: https://www.openfabrics.org/에서 검색됨

OFA. (2016). “OFED Overview”. Openfabrics: https://www.openfabrics.org/에서 검색됨

Oracle. (2016). Oracle: http://www.oracle.com/에서 검색됨

PfisterF.G. (2001). An introduction to the infiniband architecture. , High Performance Mass Storage and Parallel I/O (페이지: 617-632).

TOP500.org. (2016년 11월). “top500 list”. TOP500.org: https://www.top500.org/에서 검색됨