인피니밴드기반 저장장치에서의 iSER(iSCSI Extension for RDMA) 성능평가

김영환*, 손재기*, 정혜동*
*전자부품연구원 지능형 IDC 사업단
e-mail: yhkim93@keti.re.kr

Performance Evaluation of iSER on Storage system using Infiniband fabric

Young Hwan Kim*, Jae-Gi Son*, Hye-Dong Jung*
*Intelligent IDC R&D Division, Korea Electronics Technology Institute

요 약

최근 TCP/IP에서 세션을 통하여 노드들 간의 통신을 연결하는 방식에서 현재는 하나의 채널을 통해 고속의 I/O가 가능하도록 하는 인피니밴드 같은 기술이 많이 연구되고 있다. 인피니밴드는 프로세싱 노드와 입출력 장치 사이의 통신, 프로세스간 통신에 대한 산업 표준이 되고 있고 프로세싱 노드와 입출력 장치를 연결하기 위해 스위치 기반의 상호 연결은 전통적인 버스 입출력을 대체하는 새로운 입출력 방식이다. 또한 인피니밴드에서는 현재 이슈가 되고 있는 RDMA 방식을 이용해 원격지 서버들 간에 직접 메모리 접근 방식을 통해 CPU와 OS의 로드를 최소화하고 있다. 본 논문에서는 인피니밴드 네트워크를 이용하는 저장장치 접근 프로토콜인 iSER(iSCSI Extension RDMA Protocol)와 기존 이더넷 망에서 사용되는 iSCSI(Internet SCSI) 프로토콜을 이용하여 서버와 저장장치 간의 IOPS 와 초당 데이터 전송량에 대한 성능을 평가한다. 우리는 성능평가를 위해 Intel에서 제공하는 저장장치 I/O 성능평가 도구인 IO meter를 이용했다.

1. 서론

네트워크 기술의 급속한 발달과 정보의 효율적 공유에 대한 요구의 증가로 거의 모든 컴퓨팅 장비들이 네트워크에 연결되어 있다. 따라서 복잡하게 구성되어 있는 전체 네트워크의 상태 및 장비들을 효율적으로 파악하고, 관리하기 위한 필요성이 대두되고 있다. 특히 대용량 스토리지와 서버사이 입출력 분야에서는 한 개의 프로세서와 여러 개의 입출력 장치를 가진 소규모의 서버에서 수백개의 프로세서와 수천개의 입출력 장치를 가진 대규모의 슈퍼컴퓨터까지 사용 가능한 인피니밴드는 더욱 더 중요성이 대두 되고 있다.

현재 대부분의 네트워크 제품들은 최고의 패킷 처리량과 최소의 전송 지연, 그리고 전송 대역폭에 대한 보장을 요구해 왔고, 이는 하드웨어와 커널 바이패싱, Zero-Copy 네트워킹에서 신뢰성 있는 전송 프로토콜 사용을 통해 가능하게 되었다. 이들 메커니즘은 전통적인 TCP/IP방식에서는 불가능했던 초고속의 네트웍 데이터 프로세싱을 가능하게 한다. 그러나 이와 같은 메커니즘이 소프트웨어에 의해서 처리되는 것으로 성능향상에 한계를 갖고 있었다. 이를 해결하기 위한 방안으로 채널 기반의 네트워크 기술이 대두되게 되었고, 관련 업계에서는 IBTA라는 기술 표준화 단체를 통해 표준화를 진행해 왔다. 그 대표적인 기술이 인피니밴드이다. 인피니밴드에는 iSER, SRP(SCSI RDMA Protocol), SDP(Socket Direct Protocol), IPOIB(IP Over Infiniband) 등과 같은 여러 응용 프로토콜을 사용하고 있다. 여기서 iSER와 SRP는 인피니밴드 패킷 내에 SCSI 명령어와 데이터를 인캡슐레이션하여 저장장치에 블록단위로 송수신하는 프로토콜이다. 또한 둘 다 RDMA(Remote Direct Memory Access)를 지원하기 때문에 원격지 DMA버퍼에 직접 읽기와 쓰기가 가능하다. 더 자세한 내용은 다음의 [1] 인피니밴드 명세서 및 ANSI T10에서 iSER, SRP Draft 문서를 참고하기 바란다.

기존 이더넷 기반의 저장장치에서는 TCP/IP 계층에서 프로토콜 프로세싱를 위해 하나의 패킷을 처리하는데 커널 영역과 사용자 영역의 메모리에 읽기/쓰기를 반복한다. 이 결과로 전송 지연이 커지게 되는 단점이 있다. 그러나 인피니밴드에서는 커널 바이패싱이나 RDMA 기법을 사용하여 이 CPU에 대한 부하와 전송 지연을 최소화 하고 있다. 본 논문에서는 인피니밴드 기반의 저장장치를 사용하기 위한 접근 프로토콜로 iSER과 이더넷기반의 iSCSI를 사용하는 서버와 저장장치 간 IOPS와 초당 패킷 처리량에 대한 성능을 비교 분석할 것이다. 이는 대규모 OLTP(On-Line Transaction Processing)와 같은 데이터베이스 환경에서는 매우 중요한 성능평가 요소이기 때문이다[3].

2. 관련 연구

2.1. 이더넷기반의 블록전송 프로토콜

TCP/IP 기반 저장장치 기술을 구현하는 프로토콜은 IETF (Internet Engineering Task Force) 산하 IP스토리지 워킹그룹에서 주도하고 있는데 현재iSCSI(Internet SCSI), iFCP (Internet Fibre Channel Protocol), FCIP(Fibre Channel over IP), mFCP(Metro Fibre Channel Protocol), iSNS (Internet Storage Name Service) 등이 있다.

iSCSI는 이더넷망을 이용해 저장장치상의 블록데이터를 전송하는 기술이다. iSCSI는 이더넷 망에서 SCSI 프로토콜이 바로 전송될 수 있도록 한다. 즉, iSCSI를 사용하는 저장장치는 SCSI 명령어와 데이터를 원거리통신망(WAN)에 접속되어있는 장치(인터넷 경유 방식인 경우는 인터넷에 접속돼 있는 장치)에 전송, 데이터를 저장할 수도 있다. 또한 공통의 이더넷망을 사용해 소규모의 SAN을 복수 구축하는 것도 가능하다. 이에 따라 iSCSI 환경에서는 프로토콜 변환에 따르는 부하가 감소해 저장장치 성능 효과를 얻을 수 있다.

iFCP는 IP 와 SCSI 혹은 Fibre Channel 사이에서 게이트웨이 역할을 한다. SCSI 나 Fibre Channel 서버와 저장장치가 iFCP 스위치를 통해 LAN 이나 WAN으로 접근 가능하다. FCIP 처럼 iFCP 는 파이버 채널 프레임을 캡슐화하여 TCP/IP 네트워크를 통해 전송한다. IETF에서는 일반적인 Fibre Channel 포맷을 정의하고 있다. FCIP 와 iFCP 의 중요한 차이점은 두 프로토콜 간에 강조하고 있는 면에서 차이를 보이고 있다. FCIP 프로토콜의 경우 두 Fibre Channel SAN 을 연결하기 위해 점대점(Point-to-Point) 연결을 설정한다. 반면 iFCP 는 게이트웨이 대 게이트웨이 프로토콜이다.

2.2. 인피니밴드기반의 블록전송 프로토콜

인피니밴드 기반 저장장치 기술을 구현하는 프로토콜은 SCSI 저장장치에 대한 접근 인터페이스 기술을 정의하는 Technical Commitee T10의 SRP(SCSI RDMA Protocol)와 마이크로소프트, IBM, HP, Intel 등 주요 시스템 업체들로 구성된 RDMA (Remote Direct Memory Access) Consortium에서 공동 개발을 하고 있는 iSER(iSCSI Extension for RDMA)가 있다. 현재 iSER는 TCP/IP 기반의 네트워크 저장장치에서 IETF (Internet Engineering Task Force) 산하 IPS 워킹그룹에서 핵심 기술에 대해 수정·보완 해왔다. IBM Storage의 John Huffer가 인피니밴드 기반의 저장장치에 iSER 프로토콜을 적용하기 위해 IETF에 제안하였다[4].

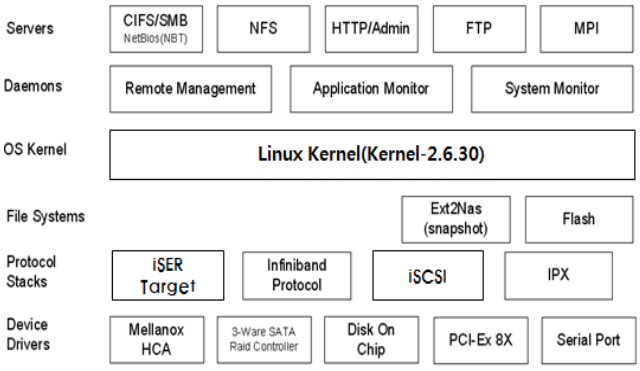
SRP는 순수하게 인피니밴드 기반의 저장장치를 접근하기 위한 프로토콜로 저장장치를 Discovery하거나 Management를 하기 위한 프로토콜이 별도로 정의되어 있지 않다. 또한, 인피니밴드 망에서만 사용이 가능하기 때문에 TCP/IP 기반의 저장장치와의 연동에는 어려움이 있다. iSER는 iSCSI의 Scalability, Manageability, Completeness를 보장하고, RDMA 전송 기능을 포함하고 있으며 브릿지를 통해서 인피니밴드에서 TCP/IP 기반의 저장장치와도 연동이 가능해 SRP의 단점을 보완하기에 문제가 없다. iSER의 대표적 장점으로 프로토콜을 처리하는데 있어 zero-copy 메커니즘을 사용하기 때문에 지연(latency)이 짧고, H/W에 의해 CRC 값이 계산된다. 또한, H/W에서 전송 프로토콜이 동작함으로 I/O에 대한 CPU 사이클 사용을 최소화 할 수 있다. iSER는 현재 TCP/IP 기반의 저장장치와 인피니밴드 기반의 저장장치에서 모두 사용된다. 그러나 TCP/IP 기반의 저장장치에서는 TOE(TCP Offload Engine) 포함한 iWARP 프로토콜이 지원되는 별도의 NIC(Network Interface Card)가 필요하다.

3. 인피니밴드 저장장치 시스템

3.1. 인피니밴드 저장장치 소프트웨어 구조

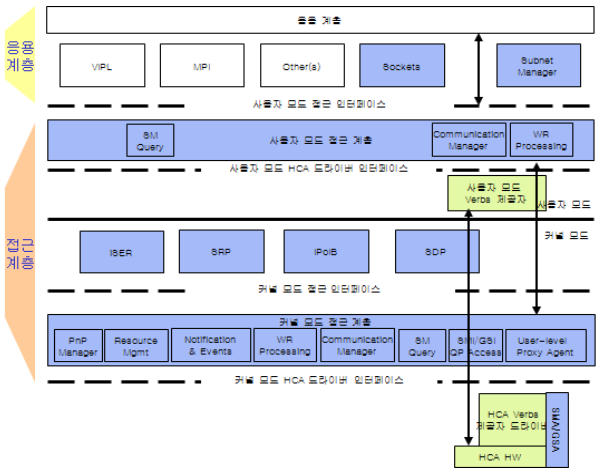
리눅스 커널은 버전 2.6.30을 택하여 개발하였다. 본 논문에서 사용된 저장장치의 경우 400KB 내외의 커널 크기를 Disk On Chip에 설치하였다. Disk On Chip의 크기를 고려하여 램을 이용하여 가상의 디스크를 생성하여 시스템의 오동작을 줄일 수 있도록 하였다.

(그림 1)은 인피니밴드 기반의 저장장치에 내장된 OS 내부 구조이다. 네트웍 저장장치를 위한 파일 시스템은 높은 신뢰성 및 가용성을 제공하여야 한다. 기존의 파일 시스템도 일관성 복구 기능을 제공하고는 있으나, 복구에 걸리는 시간은 파일 시스템의 크기와 데이터의 양에 따라 증가한다. 따라서 인피니밴드 저장장치에서 기존의 파일 시스템을 사용하는 경우 복구 시간은 더욱 증가하게 된다. 느린 복구 속도는 시스템의 가용성을 현저히 저하시키며, 가용성을 요구하는 저장장치에 적당하지 않다.



(그림 1) 인피니밴드 저장장치 소프트웨어 구조

(그림 2)는 OpenFabric[2]에서 배포한 인피니밴드 프로토콜 계층 구조에 대한 그림이다. 응용 계층에는 VIPL(Virtual Interface Provider Library), MPI(Message Passing Interface) 그리고 소켓이 있고, 미들웨어 드라이버로는 SDP, SRP, IPoIB 등이 있다. 커널을 통해 HCA에 접근하는 방식과 RDMA를 이용한 커널을 by-passing하여 HCA에 접근하는 방식을 이용해 메세지를 전송한다.



(그림 2) OFED 인피니밴드 스택 구조

3.2. 인피니밴드 저장장치 하드웨어 구조

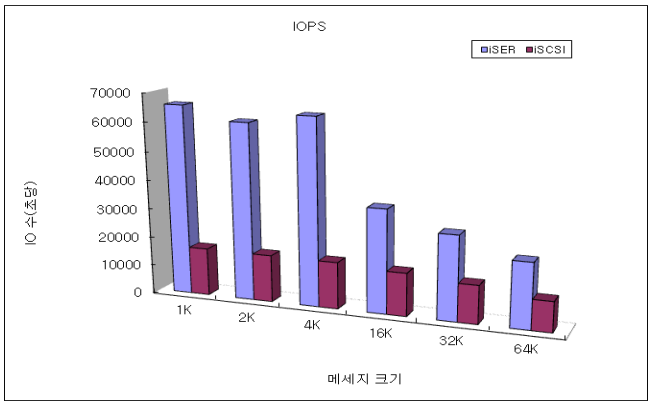
인피니밴드 저장장치 하드웨어 구조는 인텔 계열의 X86 시스템에 Cent OS 6.0 리눅스 시스템을 설치하였으며 Target 인피니밴드 저장장치 개발 시스템은 인텔 E7520 Lindenhurst 칩셋을 사용하는 메인보드를 택하였으며 저장장치 인터페이스 카드는 Mellanox사의 8X HCA, 로컬인터페이스로는 PCI-Express 8X를 지원한다. 저장장치 디스크로는 WD 10,000RPM 74.3GB 16개를 3-Ware Raid controller를 이용해서 1.2TB의 용량을 갖는다. 그리고 호스트와 타겟 저장장치 간에는 인피니밴드 프로토콜 스택을 이용해서 iSER로 연결되어 있다. 추가로 중간 버퍼를 위한 NVRAM 카드를 설치하였다. 또한 운영체제를 설치하기 위해 Disk On Chip을 사용하여 리눅스를 설치하였다.

인피니밴드기반의 저장장치 시스템 연결구조는 Dual 3GHz 프로세서에 두개의 4GB DDR3 메모리를 탑재하고 있으며, North Bridge는 8X PCI Express와 PCI-X Bridge가 연결되어 있다. 8X PCI Express슬롯에는 인피니밴드 8X HCA 카드가 연결되어 있고, 16개의 SATA를 연결하기 위해 두 개의 3-ware SATA RAID Controller PCI Express 슬롯에 연결되어 있다.

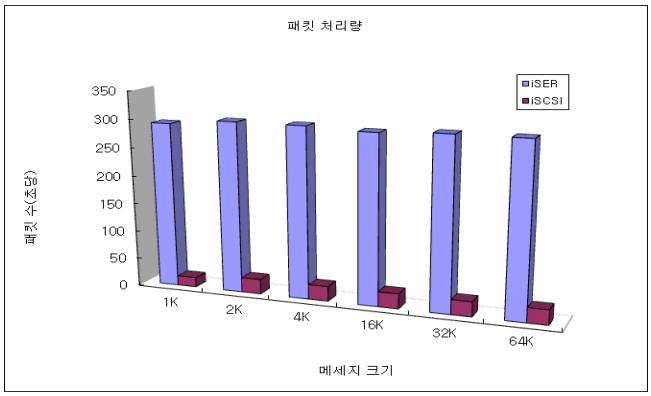
마지막으로, 성능평가를 위한 테스트 도구은 Intel에서 제공하는 IOMeter를 사용했는데 리눅스에 대해서는 별도의 GUI 환경을 제공하고 있지 않아서 IOMeter의 타겟 모듈을 리눅스에 설치하고 Windows 클라이언트에서 성능평가에 대한 내용을 제공하고 있다. 우리는 본 논문에서 인피니밴드 기반의 저장장치와 성능을 비교하기 위해 기존에 많이 사용되고 있는 이더넷 기반의 저장장치인 iSCSI 프로토콜과 함께 성능평가를 수행했다.

4. 성능평가

본 논문에서 성능 평가를 위해 메시지를 크기별로 구분하였는데 절대적인 크기로 구분한 것이 아니고 임의적으로 패킷 처리량의 변화에 따라 작은 사이즈의 메세지, 중간 사이즈의 메세지 그리고 큰 사이즈의 메시지로 구분하여 측정하였다. 그리고 Raw Disk IO방식을 이용해 Disk Cache에 읽기와 쓰기 작업을 각각50% 할당하였다. 또한 읽기와 쓰기 작업은 3분 동안 순차적인 방식으로 수행하도록 했다. 다음 (그림 3), (그림 4)는 성능평가 결과를 나타낸 그래프이다.



(그림 3) iSER & iSCSI IOPS



(그림 4) iSER & iSCSI 패킷처리량

본 논문에서는 OLTP와 같이 대규모 Transaction 많이 발생하는 망에서 인피니밴드를 이용한 고성능 저장장치에 어떠한 장점과 단점을 가지고 있는지를 알아보는 것이었다. 성능평가 결과를 보면 인피니밴드기반의 iSER 프로토콜을 이용함으로서 다른 프로토콜에 비해 상대적으로 높은 패킷 처리량을 보였다. 현재 성능평가는 실제 디스크를 읽고 쓰는 것이 아니라 Disk cache에 읽고 쓰기를 수행한다. 따라서 실제 디스크에도 읽기와 쓰기를 수행했을 때 성능을 평가해야 할 것이다.

본 논문은 지식경제부 산업원천기술개발사업(10043896)의 지원을 받아 수행된 연구임

참고문헌

[1] Infiniband Architecture Specification, Release 1.1 Infiniband Trade Association

[2] https://www.openfabrics.org/index.php

[3] http://www.rdmaconsortium.org/home

[4] iSER Storage Target for Object-based Storage Devices from ohio supercomputing center, http://www.osc.edu/research/network_file/projects/object/papers/dalessandro_snapi07.pdf