인피니밴드 네트웍에서 RDMA 기반의 저장장치 서비스 프로토콜 개 발

Implementation of Storage Service Protocol on Infiniband based Network

김 영 환*, 전 기 만, 박창원
(Young-Hwan Kim, Ki-Man Joen and Chang-Won Park)

Abstract : Because of the rapid increasing of network user, there are some problems to tolerate the network overhead. Recently, the research and technology of the user-level for high performance and low latency than TCP/IP which relied upon the kernel for processing the messages. For example, there is an Infiniband technology. The Infiniband Trade Association (IBTA) has been proposed as an industry standard for both communication between processing node and I/O devices and for interprocessor communication. It replaces the traditional bus-based interconnect with a switch-based network for connecting processing node and I/O devices. Also Infiniband uses RDMA (Remote DMA) for low latency of CPU and OS to communicate between Remote nodes. In this paper, we develop the SRP (SCSI RDMA Protocol) which is Storage Access Protocol on Infiniband network. And will compare to FC (Fibre Channle) based I-SCSI (Internet SCSI) that it is used to access storage on Etherent Fabric.

Keywords: Infiniband, I-SCSI, RDMA, SRP

I. 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 슈퍼 컴퓨터까지 사용 가능한 인피니밴드는 더욱 더 중요성이 대두 되고 있다.

현재 컴퓨터 시스템에서 사용되고 있는 입출력 버스 방식은 디스크 접근, 특히 고 성능의 서버에 있어서 병목 현상의 주요원인으로 나타나고 있다. 이러한 버스 방식은 구조가 단순하다는 큰 이점을 가지고 있어 지금까지 산업 전반적으로 사용되어 왔지만 버스 기반의 입출력 시스템은 현재의 디바이스 장치들이 요구하는 데이터 전송 대역폭을 처리할 수 있을 만큼의 시스템 입출력 성능을 가지고 있지 않다. 또한 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/IP 방식에서는 불가능했던 초고속 네트웍 데이터 프로세싱을 가능하게 한다.[2-4]

* 책임저자(Corresponding Author)
김영환: 전자부품연구원 지능형정보 시스템 연구센터
(yhkim93@keti.re.kr)
※ 본 연구는 산자부 중기거점개발사업비의 지원을 받아 연구되었음.

앞서 설명한 시스템 상의 문제는 DAS(Dsik Attached Storage), NAS(Network Attached Storage), SAN(Storage Area Network) 등 현존하는 모든 시스템 기반의 저장장치가 공통적으로 가지고 있다. 이를 해결하기 위한 방안으로 채널 기반의 네트웍 저장장치가 대두되게 되었고, 관련 업계에서는 IBTA 라는 기술 표준화 단체를 통해 표준화를 진행해 왔다. 그 대표적인 기술로 인피니밴드가 있다. 인피니밴드에는 SRP(SCSI RDMA Protocol), SDP(Socket Direct Protocol), IPOIB(IP Over Infiniband), 등과 같은 여러 응용 프로토콜이 있다. 여기서 SRP 는 인피니밴드 패킷 내에 SCSI 명령어와 데이터를 인캡슐레이션하여 저장장치에 데이터를 전송한다. 또한 RDMA(Remote Direct Memory Access)를 지원하기 때문에 원격지 DMA 버퍼에 직접 읽기와 쓰기가 가능하다. 그리고 SRP 를 사용하기 위해서는 신뢰성 연결(RC: Reliable Connection) 메세지 기반의 전송 프로토콜 계층에서 ULP(User Level Protocol) 인터페이스를 통해 접근한다. 더 자세한 내용은 다음의 [1] 인피니밴드 명세서 및 ANSI T10 에서 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 CPU 에 대한 부하와 전송 지연을 최소화 하고 있다. 본 논문에서는 인피니밴드 기반의 저장장치를 사용하기 위한 접근 프로토콜로 SRP 를 구현하고 이를 기존의 FC(Fibre Channel)기반의 i-SCSI 와 패킷 처리량 및 CPU utilization 에 대한 성능을 비교 분석할 것이다[5-7].

본 논문의 구조는 다음과 같다. 2 장에서는 이더넷 기반의 iSCSI 저장장치와 인피니밴드 저장장치기반의 SRP 프로토

콜에 대해 설명하며 3 장에서는 IOMeter 를 이용해서 SRP 와 I-SCSI 에대한 성능평가를 한다. 마지막으로 4 장에서는 요약과 함께 성능 평가를 통해 얻은 결과로 결론을 내린다.

II. 관련 연구

1. 기존 저장장치 시스템

최근 인터넷과 같은 개방된 네트워크 상황에서 기하급수적으로 증가하고 있는 데이터의 저장과 관리를 위해서 스토리지 네트워킹이 등장하였다. 스토리지 네트워크화를 통해서 전통적인 스토리지 접속 방법인 DAS 구조가 가지고 있는 병목 현상 의 단점을 해결해 줄 수 있게 되었다.

DAS 는 저장장치 가운데 가장 심플한 형태로, 서버에 직접 연결된 디스크 드라이브로 구성된다. 데이터는 컴퓨터와 하드 드라이브 사이의 가장 보편적인 입출력 통신 수단인 SCSI 명령을 사용함으로써 전송된다. SCSI 명령은 LAN 상의 가장 보편적인 전송 수단인 파일과 상반되는 블록으로 데이터를 전송한다. 그러나 DAS 방식에는 고나리의 고비용, 거리의 제한, 그리고 한정된 확장성의 문제 등의 단점이 있다. 특히 증가하는 스토리지의 확장에 따라, 기업은 더 많은 서버를 구입해야 한다. 더구나 스토리지 장비는 SCSI 장비가 최대 12 미터의 병렬 케이블에서 작동하도록 디자인되어 있기 때문에 서버에 가까이 위치해야 한다. 이러한 한계는 네트워크 스토리지의 요구를 급격하게 증가시켜왔다.

NAS 는 NIC (Network Interface Card)를 통해 기존 IP 네트워크에 직접 연결되는 새로운 개념의 스토리지를 의미한다. 기존의 서버 중심형 방식과 달리 데이터 저장 장치가 서버에 종속되지 않고 독립적으로 네트워크에 직접 연결되므로 다수의 이기종 클라이언트가 동일한 파일에 접근하여 파일을 공유하도록 지원한다. 그러나 파일 I/O 기반의 서비스가 이루어지기 때문에 데이터베이스나 응용 프로그램과 같은 블록 I/O 기반의 서비스에 부적합한 특징을 갖는다.

SAN 은 스토리지 전용 네트워크의 구축을 통해 스토리지에 대한 빠른 접근과 일원화된 관리, 그리고 확장성을 높이고 있다. 또한 SAN 구조는 높은 비용이 걸림돌이기는 하지만 서버 중심의 시스템 환경이 가지는 문제점을 근본적으로 해결해 줄 수 있는 최선의 방책으로 인식되고 있다. 그러나 SAN 이 가지는 가장 큰 단점은 클러스터 노드들이 스토리지에 존재하는 파일을 전역적으로 공유할 수 없다는 것에 있다. 이를 극복하기 위해 SAN 이 가지는 장점을 충분히 살리면서 노드들 간에 파일 공유가 가능한 파일 시스템이 필요하게 되었고 그 결과 전역적인 데이터의 공유를 효과적으로 할 수 있는SAN 기반의 클러스터 파일시스템이 등장하게 되었다.

2. TCP/IP 기반의 저장장치 접근 프로토콜

TCP/IP 기반 저장장치 기술의 가장 큰 이점은 지금 당장 사용 가능할 뿐만 아니라 비용 절감효과가 높다는 것이다. 이와 같은 장점을 발휘하는데 가장 큰 공을 세운 것은 이더넷의 급성장이다. 파이버 채널에 근접하는 수준으로 고속을 구현했기 때문이다. TCP/IP 기반 저장장치의 이점으로는 우선 상호 연동성이 있다. 비용 절감 측면에서는 파이버 채널 HBA(Host Bus Adapter)와 파이버 채널 스위치가 필요 없기 때문에 비용 지출이 줄어든다. 기존 IP 네트워크의 장점을 물려받는 TCP/IP 기반 저장장치 기술의 혜택으로는 QoS(Quality of Service)와 보안을 빼놓을 수 없다. IPSec, 3DES, 방화벽, ACL(Access Contol List), VPN(Virtual Private Network) 등 표준 IP 보안기능을 그대로 사용할 수 있다. 또 기존의 NMS(Network Management Software)도 사용 가능하다.

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

- iSCSI(Internet SCSI)

iSCSI 는 TCP/IP 네트워크를 이용해 저장장치 데이터를 전송하는 기술이다. 이 기술은 TCP/IP 네트워크상에서 SCSI 프로토콜이 바로 전송될 수 있도록 한다. 즉, iSCSI 를 도입한 기업 네트워크는 SCSI 의 명령어와 데이터를 원거리 통신망(WAN)에 접속되어있는 장치(인터넷 경유 방식인 경우는 인터넷에 접속돼 있는 장치)에 전송, 보관할 수 있다. 또한 공통의 이더넷 기반을 사용해 소규모의 SAN 을 복수 구축하는 것도 가능하다. 이에 따라 iSCSI 환경에서는 프로토콜 변환에 따르는 부하가 감소해 저장장치 성능 효과를 얻을 수 있다. 이처럼 iSCSI 는 TCP/IP 와 SCSI 를 결합함으로써 SAN 과 NAS 의 이점을 갖춘 기술로 각광받고 있다.

3. 인피니밴드 기반의 저장장치 접근 프로토콜

인피니밴드 저장장치는 서버 클러스터 시장에서 우수한 성능을 인정받고 인피니밴드 저장장치에 대한 많은 요구를 이끌어내게 되었다. 이와 함께 저장장치 성능에 매우 민감한 응용 서비스에 대해서도 많은 해결책을 내놓았다.

인피니밴드 기반 저장장치 기술을 구현하는 프로토콜은 SCSI 저장장치에 대한 접근 인터페이스 기술을 정의하는 Technical Commitee T10 의 SRP(SCSI RDMA Protocol)와 마이크로소프트, IBM, HP, Intel 등 주요 시스템 업체들로 구성된 RDMA (Remote Direct Memory Access) Consortium 에서 공동 개발을 하고 있는 iSER(iSCSI Extension for RDMA)가 있다. 현재 iSER 는 TCP/IP 기반의 네트워크 저장장치에서 IETF (Internet Engineering Task Force) 산하 IPS 워킹그룹에서 핵심 기술에 대해 수정. 보완해왔고, 현재는 인터넷 Draft 로 제안 중에 있다. 또한 인피니밴드 기반의 저장장치에도 iSER 프로토콜을 적용하

기 위해 IBM Storage 의 John Huffer 가 IETF 에 인터넷 Draft 로 제안하고 있는 상황이다.

- SRP(SCSI RDMA Protocol)

인피니밴드망에서 호스트 시스템이 원격지의 저장장치에 접근을 원할 때 그에 맞는 I/O 프로토콜이 정의 되어야 한다. SRP(SCSI RDMA Protocol)는 원래 ANSI NCITS T10 워킹 그룹에 의해 개발되었다. SRP 는 원격의 SCSI 장비를 제어하기 위한 프로토콜로 제안되었고, 인피니밴드 기술의 특성에 맞게 사용되도록 구현되었다. 일반적으로 SCSI 명령어는 저장장치 관련 산업에서는 광범위하게 사용되고, 다양한 타입의 장비에 적용할 수 있다. 현재 블럭 단위 전송 저장장치에 급속도로 적용되고 있는 프로토콜이다.

그림 1 은 SRP 프로토콜에 대한 전체 블럭도이다. SRP 는 초기자가 SCSI 작업을 생성하고 이를 타겟에서 수행하도록 하는 기본적인 서버-클라이언트 모델이 가능한 전송 서비스를 제공하는 프로토콜이다. 또한 SRP 와 관련한 모든 통신은 신뢰성을 기반으로 한 연결 서비스를 제공해야 한다. SRP 는 메시지 흐름 제어 메커니즘을 제공하는데 초기자에 의해 생성된 작업 요구에 대한 디스크립터를 큐에 넣을 수 있는 수를 타겟이 제한할 수 있도록 하고 있는데, 이 메커니즘은 다중 초기자에 의해 필요한 메시지 버퍼를 동적으로 할당할 수 있어 내부 자원을 관리하는데 사용된다. 따라서 제한된 자원에 대한 적절한 이용을 통해 전체 시스템 성능을 향상시킬 수 있다.

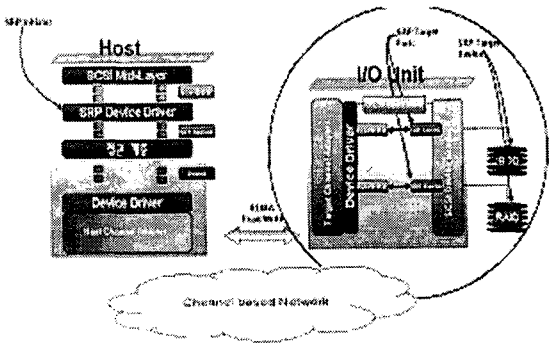

그림 1 . SRP 전체 블럭도

SRP 타겟은 모든 데이터 전송을 초기자 메모리에 직접 읽고 쓰기가 가능하도록 RDMA 기능을 포함하고 있다. 초기자는 자신의 데이터 버퍼를 등록하고, 그 내용을 전송할 SRP 명령어 내에 포함 시킴으로서 타겟으로부터 RDMA 접근이 가능하다. 다음은 SRP 프로토콜의 I/O 과정을 단계별로 설명한다[8,9].

1. 초기자는 SCSI 미들웨어로부터 SCSI 명령어와 LUN(Logical Unit Number) 그리고 데이터 버퍼 디스크립터를 포함한 SRP 요구 메시지를 생성하고, 타겟으로 해당 메시지를 전송한다.
2. 타겟은 SRP 요구 메시지를 받고 메시지에 포함되어있는 초기자의 버퍼 공간 정보를 기반으로 RDMA 전송을 수행한다.
3. 타겟은 해당 요구 작업에 대한 완료 내용을 담은 SRP 응답 메시지를 생성하고 초기자에게 전송한다.

또한, 초기자는 타겟 상에 존재하는 작업(task)을 무시할 수 있는 SRP 작업 관리에 동작을 수행할 수 있다. 게다가, 타겟은 새로운 미디어 추가와 같은 비동기적으로 발생하는 이벤트에 대한 메세지를 초기자에게 전송할 수 있다.

III. 인피니밴드 저장장치 플랫폼

1. 운영체제

리눅스 커널은 버전 2.6.14.2 를 택하여 개발하였다. 최근 정식 발표된 2.6.14.2 는 기존의 버전에 비해 임베디드 시스템을 위해 많은 부분을 최적화 시켰다. 기존 버전에서 모듈화 되지 않은 부분을 모듈화 시켰다. 연구 개발된 네트웍 저장장치의 경우 400KB 내외의 커널 크기를 Disk On Chip 에 설치하였다. Disk On Chip 의 크기를 고려하여 램을 이용하여 가상의 디스크를 생성하여 시스템의 오동작을 줄일 수 있도록 하였다.

그림 2 는 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 네트웍 저장장치에서 기존의 파일 시스템을 사용하는 경우 복구 시간은 더욱 증가하게 된다. 느린 복구 속도는 시스템의 가용성을 현저히 저하시키며, 가용성을 요구하는 네트웍 저장장치에 적당하지 않다.

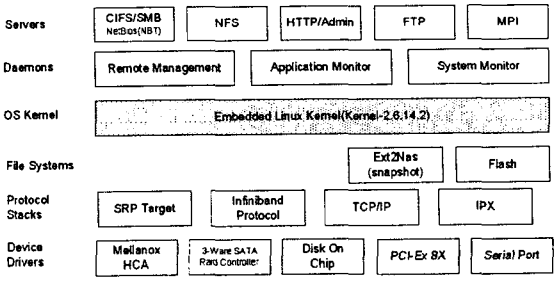

그림 2 . 인피니밴드 저장장치에 내장된 OS

2. 인피니밴드 저장장치 플랫폼

네트웍 저장장치 개발 환경은 호스트 컴퓨터로 인텔 계열의 시스템에 Fedora Core 3 리눅스 시스템을 설치하였으며 Target 인피니밴드 저장장치 개발 시스템은 인텔 E7520 Lindenhurst 칩셋을 사용하는 메인보드를 택하였으며 저장장치 인터페이스 카드는 Mellanox 사의 4X HCA 로 로컬인터페이스로는 PCI-Express 8X 를 지원한다. 저장장치 디스크로는 WD 1000RPM 74.3GB 16 개를 3-Ware Raid controller 를 이용해서 1.2TB 의 용량을 갖는다. 그리고 호스트와 타겟 저장장치 간에는 인피니밴드 프로토콜 스택을 이용해서 SRP 로 연결되어 있다. 그리고, 네트웍 버퍼를 위한 NVRAM 카드를 설치하였다. 또한 운영체제를 설치하기 위해 Disk On Chip 을 사용하여 임베디드 리눅스를 설치하였

다. 그리고, 인피니밴드 저장장치 개발에 맞도록 수정하여 사용하였으며 시리얼 포트를 이용하여 시스템을 디버깅하였다.

다음 그림 3 은 인피니밴드 저장장치 플랫폼의 블럭도이다. 우선 Dual 3GHz 프로세서에 두개의 512MB DDR2 메모리를 탑재하고 있으며, North Bridge 는 8X PCI Express 와 PCI-X Bridge 가 연결되어 있다. 8X PCI Express 슬롯에는 인피니밴드 4X HCA 카드가 연결되어 있고, 16 개의 SATA 를 연결하기 위해 두 개의 3-ware SATA RAID Controller PCI-X 슬롯에 연결되어 있다.

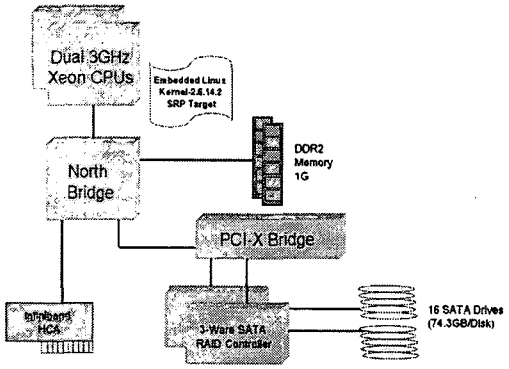

그림 3 . 인피니밴드 저장장치 플랫폼 블럭도

마지막으로, 성능평가를 위한 테스트 툴은 Intel 에서 제공하는 IOMeter 를 사용했는데 리눅스에 대해서는 별도의 GUI 환경을 제공하고 있지 않아서 IOMeter 의 타겟 모듈을 리눅스에 설치하고 Windows 클라이언트에서 성능평가에 대한 내용을 제공하고 있다.

3. 성능 평가

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 IO 방식을 이용해 Disk Cache 에 읽기와 쓰기 작업을 각각 50% 할당하였다. 또한 읽기와 쓰기 작업은 3 분 동안 순차적인 방식으로 수행하도록 했다. 다음 그림 4-6 는 성능평가 결과를 나타낸 그래프이다.

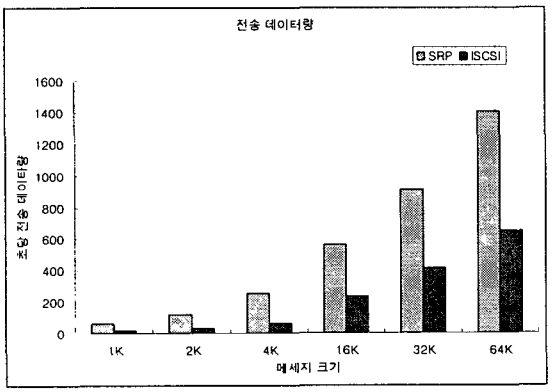

그림 4 . 전송 데이터량

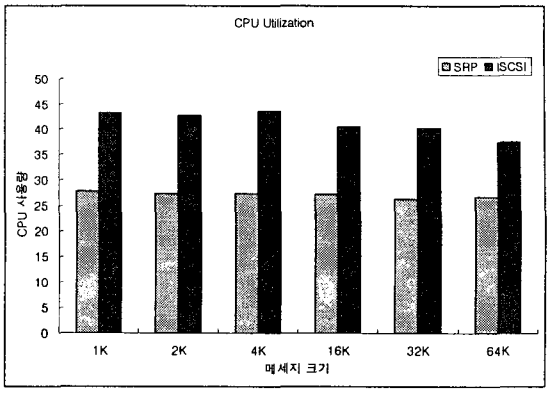

그림 5 . CPU Utilization

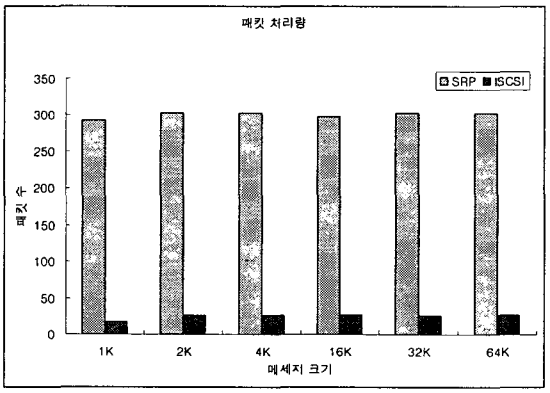

그림 6 . 패킷 처리량

VI. 결론

본 논문에서는 현재 개발단계에 있는 인피니밴드가 고성능의 네트웍에서 어떠한 장점과 단점을 가지고 있는지를 알아보는 것이었다. 특히, 인피니밴드의 목적은 데이터 입출력이 많은 IDC(Internet Data Center)에 사용될 계획으로 있어 사용자 요구에 따른 데이터 처리에 중점을 두고 있다[10]. 여기서는 인피니밴드 프로토콜 스택에서 SRP 미들웨어 드라이버를 이용함으로서 다른 프로토콜에 비해 상대적으로 높은 패킷 처리량을 보였다. 현재 성능평가는 실제 디스크를 읽고 쓰는 것이 아니라 Disk cache 에 읽고 쓰기를 수행한다. 따라서 실제 디스크에도 읽기와 쓰기를 수행했을 때 성능을 평가해야 할 것이다. 그러나 실제 디스크에 읽기와 쓰기를 수행하는 것은 디스크의 성능에 따라서 다양한 결과를 가져 올 수 있기 때문에 다양한 방법으로 테스트를 수행할 예정이다.

참고문헌

[1]Infiniband Architecture Specification, Release 1.1 Infiniband Trade Association, 2002.

[2]S. Bhattacharya, S. Pratt, B. Pulavarty, and J. Morgan. Asynchromous I/O Support inLinux 2.5. In Proceedings of the Linux Symposium, pp.371-386, July 2003

[3]*Cohen, A,* " A performance analysis of the sockets direct protocol (SDP) with asynchronous I/O over 4x infiniband" , 2004 IEEE Inter-

national Conference , pp. 241-246, April 2004.
[4]S.N. Damianakis, C. Dubnicji, and E.W. Felten. "Stream Sockets on SHRIMP". In Lecture Notes in Computer Science 1199, pp.16-30, 1997.
[5]T. von Eiken, A Basu, V.Buch, and W. Vogels. "U-Net: A User-Level Network Interface for parallel and Distributed Computing". In Proceedings of the ACM Symposium on Operating Systems Principles, PP. 40-53, December 1995.
[6]P.Balaji, S. Narravula, K. Vaidyanathan, S. Krishnamoorthy, J. Wu, and D. K. Panda. "Socket Direct Protocol over infiniband: Is it Beneficial?" Technical Report OSU-CISRC-10/03-TR54, The Ohio State University, 2003.
[7] P. Balaji, P. Shivam, P. Wyckoff, and D.K. Panda, and J. Saltz. Impact of High Performance Sockets on Gigabit Ethernet. In Cluster Computing, September 2002.
[8] The Public Netperf, http://www.netperf.org
[9]J.Wu, P. Wyckoff, and D. K. Panda, PVFS over Infiniband: Design and Performace Evaluation. In ICPP, 2003.

김 영 환

2003년 성균관대학교 컴퓨터 공학과 석사 졸업. 2003년~현재 전자부품연구원 지능형 정보시스템 연구센터 전임연구원. 관심분야는 인피니밴드, 임베디드 OS.

박 창 원

1986년 중앙대학교 전자공학과 학사 졸업. 2002년 광운대학교 전자통신공학과 석사 1986년~1988년 동양정밀 중앙연구소 주임연구원 1988년~1993년 효성컴퓨터 중앙연구소 선임연구원 1993년~현재 전자부품연구원 지능형 정보시스템 연구센터 센터장. 관심분야는 저장장치, 센서네트워크.

전 기 만

2000년 한양대학교 전기공학과(공학사). 2000년~2001년 삼보컴퓨터 연구소 연구원 2001년~현재 전자부품연구원 지능형정보시스템 연구센터 전임연구원 관심분야는 인피니밴드, 센서네트워크