특집

인피니밴드 망에서의 저장장치 관련 기술

박창원 (성균관대학교, 전자부품연구원), 김영환 · 손재기 · 전기만 (전자부품연구원)

요 약

IT 산업의 발전과 더불어 네트워크 사용자의 급속한 증가로 정보의 양은 폭발적으로 증가하고 있으며, 네트워크 내의 부하를 감당하기에는 많은 어려움을 가져왔다. 이와 같은 이유로 대규모의 정보를 저장할 수 있는 대용량의 저장 시스템 연구와 기존의 TCP/IP에서 세션을 통하여 노드들 간의 통신을 연결하는 방식에서 현재는 하나의 채널을 통해 고속의 I/O가 가능하도록 하는 기술이 많이 연구되고 있다. 그 대표적인 것으로 인피니밴드가 있다. 인피니밴드는 프로세싱 노드와 입출력 장치 사이의 통신, 프로세스간 통신에 대한 사업 표준이 되고 있고 프로세싱 노드와 입출력 장치를 연결하기 위해 스위치 기반의 상호 연결은 전통적인 버스 입출력을 대체하는 새로운 입출력 방식을 사용한다. 본 고에서는 앞서 말한 바와 같이 현재 TCP/IP기반의 네트워크 부하를 새로운 채널 기반의 인피니밴드를 이용하여 해소하고 이 기술을 네트워크 저장장치에 적용하기 위한 관련 기술에 대한 논하고자 한다.

I. 서 론

인터넷의 발전으로 인해 정보화 사회로의 변화가 급속히 이루어지고 있다. 다시 말하면, 모든 정보가 컴퓨터화하여 인터넷이라는 매개체를 통해 누구든지, 언제, 어디서라도 접근이 가능하게 된 것이다. 이를 위해서, 정보를 체계적으로 저장하고 이를 제공하는 저장 장치의 중요성은 새삼 언급할 필요조차 없다. 컴퓨터 기술의 발전과 더불어 저장 장치 또한 빠른 속도로 발전하여 왔으나 과거의 텍스트 위주의 데이터에서 현재에는 비디오, 오디오 등의 멀티미디어 데이터로 변화함에 따라 그 크기가 기하 급수적으로 증가하고 있는 실정이다.

최근에는 저장 매체 기술의 발달로 저장장치의 가격이 급속하게 떨어지고 있다. 또한 네트워크 및 인터넷 기술의 발달에 따라 데이터의 크기가 증가함으로 해서, 대용량 저장장치에 대한 요구가 높아지고 있다. IDC의 분석에 의하면 2003~2005 년 사이에 저장장치의 용량은 해마다 75% 이상 증가하고 있고, 이러한 증가추세에는 두 가지 중요한

이슈를 가지고 있는데, 데이터 중요도의 증가와 저장장치 자원 관리의 어려움이다. 또한 IDC 는 2005 년도에는 전체 저장장치의 70% 이상이 네트워크화 될 것으로 전망하고 있다. 그러나 현재의 TCP/IP 기반의 네트워크 저장장치 시스템에서 사용되고 있는 입출력 버스 방식은 디스크 접근, 특히 고성능의 서버에 있어서 병목현상의 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. 뿐만 아니라 대용량의 데이터를 다수의 사용자에게 서비스하기에는 많은 문제점을 가지고 있다.

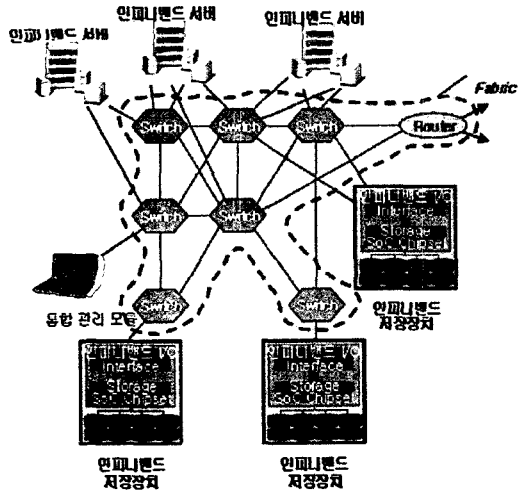
결과적으로 TCP/IP 기반의 네트워크 저장장치의 문제점을 해결하기 위해 인피니밴드 기반의 네트워크 저장장치 기술이 등장하게 되었다. 인피니밴드 기술은 대용량 저장장치와 서버사이 입출력 분야에서 한 개의 프로세서와 여러 개의 입출력 장치를 가진 소규모의 서버에서부터 수백개의 프로세서와 수천개의 입출력 장치를 가진 대규모의 슈퍼컴퓨터까지 모두 사용이 가능하다. 또한 현재 대부분의 TCP/IP 기반의 네트워크 제품들은 최고의 패킷 처리량과 최소의 전송 지연, 그리고 전송 대역폭에 대한 보장을 요구해 왔다. 이를 인피니밴드에서는 전송 계층을 하드웨어를 통해 가능하게 하고, 소프트웨어에서는 커널 바이패싱, Zero-Copy 메커니즘을 적용하였다. 그리고 네트워킹에서는 신뢰성 있는 전송 프로토콜을 사용하여 앞서 언급한 TCP/IP에서의 문제점을 해결했다. 결국 인피니밴드는 전통적인 TCP/IP 기반의 네트워크 저장장치에서 불가능했던 초고속의 네트워크 데이터 서비스를 가능하게 함으로서 다수의 사용자에게 대용량의 데이터 서비스 할 수 있게 되었다.

본 고에서는 인핀밴드 기반의 네트워크 저장장치 기술에 관하여 고찰하였다. 인피니밴드 기반의 저장장치 기술은 기존의 공유 버스 시스템에서의 문제점과 TCP/IP 프로토콜 프로세싱에서의 문제점을 해결하여 기존의 NAS(Network Attached Storage)와 SAN(Storage Area Network) 사이에 고성능의 데이터 전송을 가능하게 하는 네트워크 저장장치 기술이다. 또한 대규모의 컴퓨팅 파워를 필요로 하는 서버 클러스터 분야에서도 적용되고 있다. 2장에서는 인피니밴드 시스템 구조와 관련 기술에 관해 분석하고, 3장에서는 TCP/IP 기반에서의 저장장치 기술과 인피니밴드 기반의 기술에 대해 알아볼 것이다. 그리고 마지막 장에서는 인피니밴드 기반의 네트워크 저장장치에 대한 향후 전망에 대해 언급한다.

II. 인피니밴드

1. 시스템 구조

그림 1과 같이 인피니밴드 연결망은 스위치 기반 비정형 연결망으로 종단 노드(프로세서 노드, 입출력 노드)가 연결되는 여러 개의 서브넷으로 구성된다. 서브넷은 종단 노드와 스위치, 라우터로 구성되며 서브넷 간 연결은 라우터를 통해 이루어진다. 하나의



〈그림 1〉 인피니밴드 시스템 구조

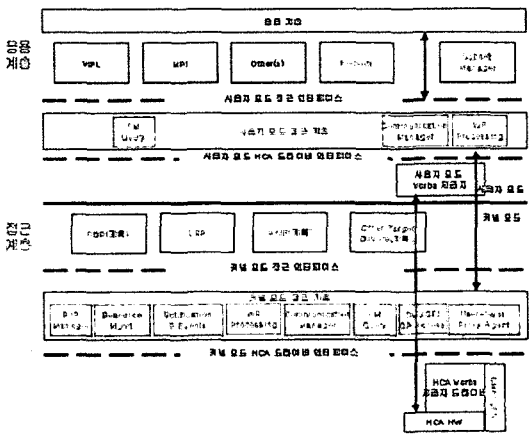
서브넷에는 최대 6만5536(216)개의 종단 노드를 연결할 수 있는 고 확장성이 제공된다. 각 종단 노드는 인피니밴드 연결망 접속을 위한 채널 어댑터를 가지며, 프로세서 노드 쪽에서는 호스트 채널 어댑터(HCA: Host Channel Adapter)를 사용하고 입출력 처리 노드 및 입출력 장치 쪽에서는 타겟 채널 어댑터(TCA: Target Channel Adapter)를 사용한다.

인피니밴드는 고속의 점대점 링크에 스위치를 기반으로 노드가 서로 연결된 망으로 설계되었다. 인피니밴드망은 하나 또는 그 이상의 스위치와 프로세싱 노드, 입출력 디바이스 장치로 구성된 서브넷이 라우터에 의해서 상호 연결되어 있다. 인피니밴드에서 라우팅은 각 스위치에 저장된 포워딩 테이블를 기반으로 하며, 유연성과 확장성을 제공하기 위하여, 사용자에 의해 정의된 어떠한 토폴로지도 지원한다.

인피니밴드 링크는 양방향 점대점 통신 채널로 되어 있고, 링크의 신호 발생율은 현재 2.5GHz이며 물리적 링크 레벨에서는 보다 높은 대역폭을 얻기 위해 병렬적으로 사용되는데, 가장 낮은 대역폭은 1X로서 2.5Gbps이고 최대12X인 30Gbps까지 지원한다.

인피니밴드 스위치는 서브넷 매니져에 의해 토폴로지의 초기화와 네트워크의 변화에 따른 수정에 의해 얻어진 포워딩 정보를 기반으로 만들어진 포워딩 테이블을 통해서 로컬에서 목적지로 메시지를 전송하게 된다. 그 메시지는 링크 상에서 스위치를 통해 전송 되도록 패킷의 형태로 세그먼트화 된다. 그 패킷의 크기는 헤더를 제외하고 256bytes, 1KB, 2KB, 4KB까지 될 수 있다

2. 인피니밴드 S/W 구조



〈그림 2〉 인피밴드 S/W 구조

인피니밴드는 그림 2와 같은 S/W 구조를 갖는다. 각 노드는 메세지 전송을 위해 사전에 호스트 메모리에 메세지 전송을 위해서 사용할 영역을 가상 주소로 사용 · 지정해야 한다. 사용자 프로그램은 메세지 전송 요구 및 메세지 데이터를 지정된 영역에 위치시키

게 되고 HCA 또는 TCA는 운용 체계의 간섭 없이 지정된 영역에서 메세지 전송 요구와 해당 데이터를 가상 주소를 사용해 읽어 내어 인피니밴드 연결망으로 전송한다. 인피니밴드 전송 메커니즘은 종단노드 간에 여러 가지 타입의 통신 서비스를 제공하며 이와 같은 타입은 연결 지향성과 데이터 그램 서비스 그리고 신뢰성과 비 신뢰성으로 분류될 수 있다. 또한 대역폭 보장과 최대 지연 마감과 같은 QoS요구 조건을 만족하도록 하기 위해 어플리케이션은 자원 할당을 할 수 있도록 신뢰성 있는 연결을 사용해야 한다.

통신 방식으로는 Send/Receive와 RDMA 그리고 Atomic의 세가지 방식을 사용한다. Send는 어플리케이션이 자신의 버퍼에 저장된 데이터를 원격지의 어플리케이션에게 보내는 것이고 Receive는 일단 원격지 어플리케이션이 Send를 동작을 하기 전에 로컬의 큐에 저장된다. 큐에는 저장될 메모리 위치에 대한 정보를 갖고 있다.

RDMA방식은 어플리케이션이 원격지 어플리케이션의 메모리에 데이터를 직접 쓰거나 읽을 수 있는데 이를 위해 RDMA 메세지 내에 원격지의 가상 메모리 주소를 포함하고 있다. 따라서 원격지의 어플리케이션은 RDMA를 위해 어떤한 동작도 필요하지 않으며 RDMA는 원격지 노드의 CPU 간섭 없이 데이터 전송이 수행된다. 결국, 그 데이터는 커널 버퍼로의 복사 없이 사용자 레벨의 버퍼 간에 복사를 하게 된다.[4]

III. 네트워크 기반의 저장장치 기술

본 고에서는 네트워크 기반의 저장장치 기술로 크게 두 가지로 구분한다. 현재 가장 많이 쓰고 있는 TCP/IP 기반의 저장장치 기술과 새로운 시스템 연결망인 인피니밴드 기반의 저장장치 기술이다. TCP/IP 기반의 네트워크 저장장치도 전통적인 NIC, TOE(TCP Offload Engine) NIC, RDMA NIC 등 하드웨어의 진화에 따라 고속의 저장장치 서비스를 제공해 왔고, 소프트웨어에서는 저장장치 기술을 구현하는 프로토콜 기술도 하드웨어와 함께 개발되었다. 그러나 클러스터링 분야나 대용량의 I/O를 요구하는 데이터 센터 분야에서는 TCP/IP 기반의 저장장치에 대한 기술적인 한계에 도달하였다. 그 결과로 인피니밴드 기술이 탄생하게 되었다.

1. TCP/IP 기반의 저장장치 기술

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. IP 스위치 포트는 파이버 채널 스위치 포트보다 2~3 배 저렴하며 파이버 채널 회선 대신 저렴한 네트워크 회선을 이용하면 된다. 이론상으로는 TCP/IP 기반 저장장치 기술을 구현해 SCSI 블록 레벨의 프로토콜을 보다 멀리, 그리고 장비를 무제한 연결할 수 있다. 또한 SAN 을 관리하기 위한 전문 기술

을 요하지 않는다. 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 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[3,6,7].

● iSCSI (internet Over SCSI)

iSCSI 는 SCSI 프로토콜을 사용하여 IP 기반 네트워크에 블록 데이터를 전송하기 위한 IETF(Internet Engineering Task Force)의 표준이다. SCSI 는 높은 데이터 전송률, 신뢰성, 낮은 지연(latency)을 강점으로 저장장치와 서버를 연결하는 이상적인 프로토콜로 평가 받았다. 그러나 서버와 저장장치 환경이 커지고 복잡해짐에 따라 SCSI는 물리적으로 속도, 저장장치 간 공유, 짧은 연결거리 등이 문제로 부각되었다. 물리적인 한계가 있긴 해도 SCSI 프로토콜은 새로운 기술과 쉽게 접목되는 유연성과 높은 성능을 바탕으로 앞으로도 계속 살아남을 것으로 전문가들은 예측하고 있다. iSCSI 는 TCP/IP 네트워크를 이용해 저장장치 데이터를 전송하는 기술이다. 이 기술은 TCP/IP 네트워크상에서 SCSI 프로토콜이 바로 전송될 수 있도록 한다. 즉, iSCSI를 도입한 기업 네트워크는 SCSI 의 명령어와 데이터를 원거리통신망(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 iSCSI는 TCP/IP 와 SCSI를 결합함으로써 SAN과 NAS의 이점을 갖춘 기술로 각광받고 있다.

SAN은 전형적으로 SCSI 와 Fibre Channel 프로토콜(SCSI-FCP)을 사용한다. Fibre Channel은 SCSI 와 마찬가지로 블록 단위로 데이터를 전송하지만 SCSI 와 달리 거리의 제한을 덜 받는다. 파이버채널이 가지는 가장 큰 장점은 빠른 전송 속도(초당 최대 100MB 전송률)를 가지지만 실제 속도는 60MB/s~80MB/s 로 기가비트의 40MB/s 평균 전송률보다 빠르다. iSCSI 는 SAN 과 같이 블록 단위로 데이터I/O(Input/Output)가 가능해 빠른 속도를 내며 랜 상에 TCP/IP 를 사용해 데이터를 저장 및 관리공유, 파일 액세스 등을 할 수 있다.

SCSI 인터페이스가 있는 시스템은 SCSI 명령어를 발동시키며, 명령어는 레이어 4 패킷으로 캡슐화되어 내보내진다. 수신 시스템은 패킷에서 SCSI 명령어를 분석하여 실행시킨다. 수신 유닛은 돌아오는 SCSI 명령어와 데이터를 IP 패킷 안으로 캡슐화한 다음 이들을 첫 번째 시스템으로 다시 돌려보낸다. 이 시스템은 데이터나 명령어를 분석하여 이들을 다시 SCSI 서브시스템으로 전달한다. 이러한

모든 작동은 사용자의 개입이 없이 이루어지며, 엔드유저에게 완전히 투명하다.

iSCSI 의 사양 중 상당 부분은, 호환성 유지를 위해 표준 SCSI 실행들을 따라야 하고, SCSI 를 깨뜨리지 않도록 해야 하기 때문에, 있는 그대로 구성되었다. 이것은 또한 처음부터 IPv4나 IPv6용으로 만들어졌다.

보안을 유지하기 위해, iSCSI에는 자체의 로그인 절차가 있다. 첫 번째 작동 시에 초기자(Initiator) 노드가 타겟(Target) 노드로 로그인을 한다. 로그인 프로세스를 수행하지 않은 초기자로부터 iSCSI PDU를 받은 타겟 노드는 어떤 것이건 프로토콜 에러를 발생시키고 접속을 종료시킬 것이다. 단 이 때 타겟 노드는 세션을 종료하기 전에 아마도 거부 iSCSI PDU를 다시 보낼 것이다. 이것은 기본적인 보안 형태인데, 왜냐하면 통신의 처음만을 보호하며, 모든 패킷 기반에서의 보안을 제공하지 않기 때문이다. 하지만 IPsec 의 이용 등, iSCSI 를 위한 보안을 제공하는 다른 방법들이 있다. 제어와 데이터 패킷 모두의 측면에서, IPsec 은 무결성, 재생 보호 및 인증을 준비해줄 것이다. 또한 개별적 패킷들을 위한 암호화도 제공할 것이다.

이에 비해, 파이버 채널은 그만큼 안전하지 못하지만, 파이버 채널 패브릭으로 접속을 하려면 물리적 액세스와 파이버 채널에 대한 철저한 지식이 필요하다. 물론, 파이버 채널 보안의 핵심은 다른 모든 네트워크 접속을 하지 않는 것이다. 하지만 파이버 채널에는 암호화가 없으며, 프로토콜 레벨의 보안이 거의 없다. 거의 모든 사람들이 알고 있는 것과 마찬가지로, TCP/IP 네트워크는 공격당하기가 쉽다. 여기에는 외부 및 잘 알려진 프로토콜로의 접속성이 있다. 시스템 관리자들에게는 IP 네트워크를 안전하게 지킬 수 있는 많은 툴이 있고 다양한 경험을 갖고 있다. 이와 함께, 대다수 iSCSI SAN이 어떠한 공중 액세스도 없이 네트워크를 분리시킬 것이라는 가능성은 보안 유지의 확률을 더욱 높여준다.

● FCIP (Fibre Channel Over IP)

TCP/IP 기반 저장장치 네트워크의 또 하나의 축인 FCIP(Fibre Channel over IP)는 IP를 통해 파이버채널 터널링을 제공하며 IP 패킷 안에 파이버 채널 프로토콜을 캡슐화(Encapsulation)한다. 파이버 채널을 사용하는 SAN은 비용이 많이 든다. 또한 숙련된 관리자를 찾는 일도 만만치 않다. 이에 반해 TCP/IP 네트워크는 전 세계적인 네트워크 프로토콜로 관리자 훈련비용이 상대적으로 적게 든다. 또한 IP를 이용하면 거리 제한 없이 SAN 과 SAN 을 연결할 수 있고 가상사설망(VPN), IPSec과 같은 기술을 이용해 쉽게 보안 문제도 해결한다.

FCIP는 특성상 원격 백업 · 복구 등 재해복구 시스템에 활용 여지가 많아 저장장치 업체들이 특히 눈독을 들이고 있다. 2 개 사이트의 완전 미러링을 통해 뜻하지 않은 사고 발생 시 미러 사이트에 있는 오프라인 테이프를 통해 신속하게 대체 사이트를 온라인화 할 수 있기 때문이다.

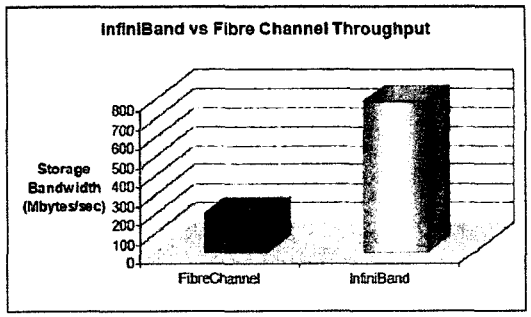
● iFCP (internet Fibre Channel Protocol)

iFCP는 IP 와 SCSI 혹은 Fibre Channel 사이에서 게이트웨이 역할을 한다. SCSI 나 Fibre Channel 서버와 저장장치가 iFCP 스위

치를 통해 LAN 이나 WAN 으로 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. 정확한 목적 주소로 라우트하기 위해 Fibre Channel 프레임을 파이버 채널과 IP 주소 변환을 조합한다. FCIP 의 주소변환 방식과는 다르게 현재의 iFCP 주소변환 방식은 각자의 독립적인 네임 스페이스를 유지하며 상호 연결된 SAN 을 연결할 수 있다.

2. 인피니밴드 기반의 저장장치 기술

Storage Networking World 2004에서 인피니밴드 칩셋 제조 기업인 멜라녹스가 처음으로 초기 인피니밴드 저장장치 플랫폼을 선보였다. 저장장치 관련 산업 표준을 기반으로 한 초기 인피니밴드 저장장치 플랫폼은 디스크까지 데이터 처리량이 거의 800MB/sec에 이르렀다. 그와 같은 저장장치 플랫폼은 시장에서 저장장치와 관련된 OEM의 제품 개발 주기와 시간을 가속화시킬 수 있었다. 다음 그림 3은 초기 인피니밴드 저장장치 플랫폼과 기존 Fibre Channel 기반의 저장장치에 대한 데이터 처리량의 대한 성능 분석 결과이다. 거의 Fibre Channel 기반의 저장장치에 비해 4배 이상의 데이터 처리량을 보이고 있음을 알 수 있다.[5]



〈그림 3〉 인피니밴드 vs. Fibre Channel 데이터 처리량

인피니밴드 저장장치는 서버 클러스터 시장에서 우수한 성능을 인정받고 인피니밴드 저장장치에 대한 많은 요구를 이끌어 내게 되었다. 이와 함께 저장장치 성능에 매우 민감한 응용 서비스에 대해서도 많은 해결책을 내놓았다. 다음은 인피니밴드 저장장치와 관련해서 가격대 성능비가 우수한 응용에 대한 예이다.

- 백업/ 무디스크 백업
- 미러링/스냅샷/체크포인팅
- 비디오 스트리밍/그래픽
- 재난 복구를 위한 클러스터 스토리지
- 데이터 저장

현재 인피니밴드는 데이터 센터와 같은 고성능의 컴퓨팅 능력과 I/O를 요구하는 분야에서 시장을 주도하고 있다. 또한 10Gb/s의 성능과 전송계층 offload 기술을 갖춘 산업 표준의 기술들 가운데 가장 우수한 가격대 성능비를 나타내고 있다. 결과적으로 고성능, 저비용을 요구하는 관련 업체는 인피니밴드 저장장치 시장을 선점하기 위해 서로

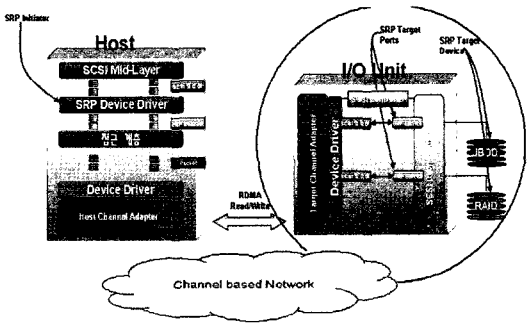
앞 다투어 인피니밴드 저장장치를 내 놓고 있다.

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·보완해왔고, 현재는 인터넷 Draft로 제안 중에 있다. 또한 인피니밴드 기반의 저장장치에도 iSER 프로토콜을 적용하기 위해 IBM Storage의 John Huffer가 IETF에 인터넷 Draft로 제안하고 있는 상황이다.

● SRP(SCSI RDMA Protocol)

인피니밴드망에서 호스트 시스템이 원격지의 저장 장치에 접근을 원할 때 그에 맞는 I/O 프로토콜이 정의 되어야 한다. SRP(SCSI RDMA Protocol)는 원래 ANSI NCITS T10 워킹 그룹에 의해 개발되었다. SRP는 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 급속도록 적용되고 있는 프로토콜이다.

그림 4는 SRP 프로토콜에 대한 전체 블럭도이다. SRP는 초기자가 SCSI 작업을 생성하고 이를 타겟에서 수행하도록 하는 기본적인 서버-클라이언트 모델이 가능한 전송 서비스를 제공하는 프로토콜이다. 또한 SRP와 관련한 모든 통신은 신뢰성을 기반으로 한 연결 서비스를 제공해야 한다. SRP는 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 4〉 SRP 전체 블락도

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 포함시킴으로서 타겟으로부터 RDMA 접근이 가능하다. 다음은 SRP 프로토콜의 I/O 과정을 단계별로 설명한다.

1. 초기자는 SCSI 미들웨어로부터 SCSI 명

령어와 LUN(Logical Unit Number) 그리고 데이터 버퍼 디스크립터를 포함한 SRP 요구 메시지를 생성하고, 타겟으로 해당 메시지를 전송한다.

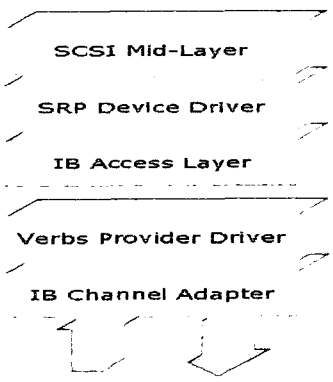
2. 타겟은 SRP 요구 메시지를 받고 메시지에 포함되어있는 초기자의 버퍼 공간 정보를 기반으로 RDMA 전송을 수행한다.

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.

-초기자(Initiator)

초기자는 그림 5와 같이 구성되어 있다. 인피니밴드 망과 직접 연결되어 있는 인피니밴드 어댑터가 최하단에 위치하며, 그 위로 어댑터 접근을 위한 Verbs 프로바이더 드라이버가 있다. 이 둘은 상호 긴밀한 관계를 가지며, Verbs 프로바이더는 보통 어댑터 벤더에서 제공한다.



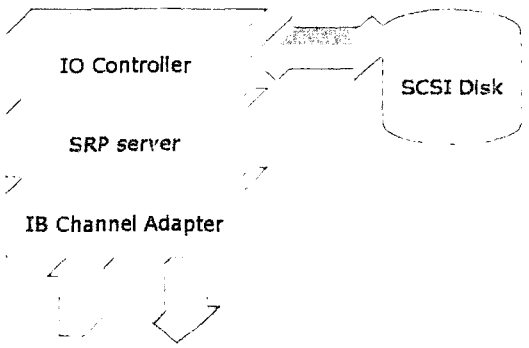
〈그림 5〉 초기자 구성도

인피니밴드 접근 계층은 다시 두개의 계층으로 나뉜다. 커널 모드와 사용자 모드가 그것이다. 커널 모드는 사용자 모드 하단에 위치하여 Verbs 프로바이더 드라이버와 연결이 된다. 사용자 모드는 사용자의 접근을 위해서 커널 모드로 진입하기 전 사용자 수준에서 가용한 인터페이스들을 말한다. SRP 디바이스 드라이버는 SRP 서비스를 타겟에서 받기 위하여 존재하며, SRP 서비스를 받기 위해서 이루어져야 할 작업들을 처리한다. SRP 디바이스 드라이버와 연결되어 SCSI 기기들에게 명령어를 내리는 SCSI 미들웨어는 SRP 디바이스 드라이버 상단에 위치한다. 실제 처리하는 명령어 집합의 수준은 저수준들이다.

SRP 디바이스 드라이버는 실제적으로 원격의 저장 장치와 연결이 된다. 하지만 기존의 저수준 SCSI 드라이버와는 RDMA를 사용한다는데 그 차이가 있고, 원격의 저장 장치에 입출력 제어를 가능하게 한다. 이러한 제어들은 초기자에서 수행되며, 제어를 받는 것은 타겟이 된다.

-타겟(Target)

타겟은 SRP 서비스를 제공하는 서버이다. 하지만 서버임에도 불구하고 구성은 그림 3과 같이 간단하다. 초기자와 마찬가지로 인피니밴드 망과 직접 닿아 있는 인피니밴드 어댑터가 최하단에 위치한다. SRP 타겟은 인피니밴드 어뎁터로 들어오는 프레임들을 읽어 들인다. 그리고 읽어 들인 프레임을 SRP 프로토콜에 맞게 파싱하여 각각의 명령어를 입출력 제어기로 넘겨준다.



〈그림 6〉 타겟 구성도

입출력 제어기는 저수준의 SCSI 명령어들을 처리하게 된다. SRP 헤더로 인캡슐레이션 되어 도착한 프레임을 SRP 타겟에서 디캡슐레이션하여 입출력 제어기로 넘기면 그때 SCSI 저장 장치에 저수준 명령어가 내려진다. SRP 서비스는 기본적으로 서버/클라이언트 모델을 하고 있다. 하지만 인피니밴드 구조를 가졌기 때문에 신뢰성을 보장 받을 수 있다.

SRP 서비스는 기본적으로 명령어를 일정 큐에 저장을 하고 큐에서 순차적으로 프레임을 읽어 들이는 방법을 사용한다. 즉, 비동기적으로 이벤트가 일어나게 된다. 많은 수의 요청이 오더라도 큐에 저장이 되고, 저장된 큐는 순서대로 SRP 서버에서 읽어 들여 처리를 하게 된다[10].

● iSER(iSCSI Extensions for RDMA)

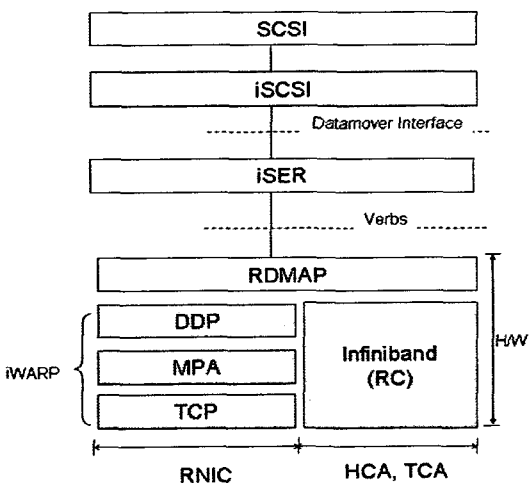
앞서 설명한 SRP는 순수하게 인피니밴드 기반의 저장장치를 접근하기 위한 프로토콜로 저장장치를 Discovery하거나 Management를 하기 위한 프로토콜이 별도로 정의되어 있지 않다. 또한, 인피니밴드 망에서만 사용이 가능하기 때문에 TCP/IP 기반의 저장장치와의 연동에는 어려움이 있다. iSER는 iSCSI의 Scalability, Manageability, Completeness를 보장하고, RDMA 전송 기능을 포함하고 있으며 브릿지를 통해서 TCP/IP 기반의 저장장치와도 연동이 가능해 SRP의 단점을 보완하기에 문제가 없다. 다음 표 1은 iSER와 SRP에 대한 비교표이다. 기능적인 면에서 SRP에 비해 우수함을 알 수 있다. iSER의 대표적 장점으로 프로토콜를 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.

〈표 1〉 iSER vs. SRP

	iSER	SRP
성능	v	v
확장성	v	x
표준/포괄적인 관리	v	x
FC 연동	v	x
GbE iSCSI 연동	v	x
에러핸들링,복구, 다중 경로	v	x
보안/인증	v	x
기존 소프트웨어 사용/자원관리	v	x

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. 다음 그림 7은 TCP/IP와 인피니밴드 기반의 iSER 프로토콜 스택에 대한 구조를 나타낸 그림이다. 우선 TCP/IP 기반의 iSER 스택을 보면 TOE를 기반으로 하고 있기 때문에 TCP/IP에서의 프로토콜 처리에 대한 오버헤드나 사용자 · 커널

메모리간의 복사 오버헤드를 줄일 수 있다. TCP 계층은 바이트 스트림과 양방향 전송을 제공하며, 원격지 노드와 신뢰성 있는 연결을 맺는데 이용된다. 그리고 그림 7에서 보듯이 RDMAP, DDP, MPA 모듈이 하나의 iWARP 프로토콜을 구성하게 되는데 iWRAP의 주된 역할은 원격지 노드의 메모리에 직접 접근하여 읽고 쓰기를 한다. MAP (Markers with PDU Alignment)계층은 TCP를 기반으로 해서 메시지 프레임화와 데이터 무결성 기능을 제공한다. DDP(Data Direct Placement) 계층은 RDMAP/DDP메세지에 대해 SAR(Segmentation And Reassembly), 직접 데이터 배치 및 메시지의 정렬 전송 기능을 제공한다. 그리고 RDMAP(Remote Direct Memory Protocol) 계층은 TCP/IP와 인피니밴드 기반의 저장장치에서도 같은 기능을 제공하는데 RDMA에 관한 시멘틱을 제공한다. 예를 들어 RDMA 읽기 · 쓰기 및 Send 기능을 지원한다. 인피니밴드 기반의 저장장치에서도 H/W 자체에 RDMA기능을



〈그림 7〉 iSER 프로토콜 스택

지원하고 있지만 아직 인피니밴드 기반의 iSER 명세서에는 해당 프로토콜 스택에 대한 내용이 포함되어 있지 않다. 다음으로 SCSI 계층이 있는데 이는 SCSI 관련 명령어를 생성하는 OS 관련 미들웨어이다. iSCSI는 3.1절을 참조하기 바란다. 마지막으로 iSER 계층은 iSCSI 프로토콜을 iWARP 프로토콜 또는 인피니밴드 프로토콜로 변환하기 위한 계층이다. 즉, iSCSI PDU를 적절한 RDMAP 메시지로 변환및 인캡슐하는 기능을 제공한다.[2, 8, 9]

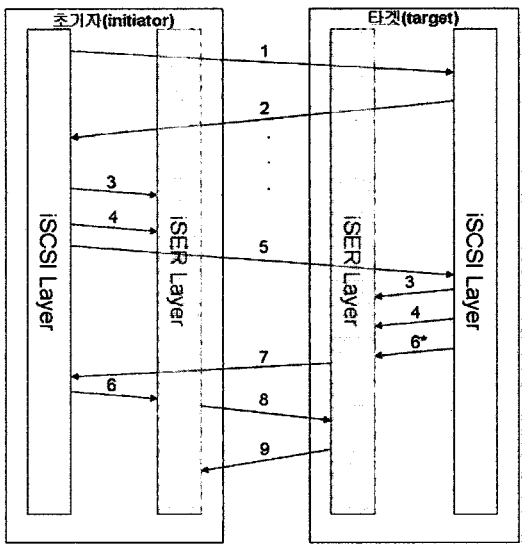
– 연결(Connection Setup) & 로그인(Login)

TCP/IP 기반의 저장장치에서 iSCSI 와 iSER over TCP/IP 프로토콜은 연결이나 로그인을 위해서 TCP/IP 바이트 스트림 프로토콜을 사용하는데 iSCSI는 초기자와 타겟 사이의 모든 동작에 바이트 스트림 프로토콜을 사용하는데 반해 iWARP는 단지 로그인이나 연결을 위한 파라미터를 주고 받을 때만 사용하고, 로그인 및 연결 이후에는 RDMA를 이용하여 통신하게 된다.

인피니밴드 기반의 iSER에서는 인피니밴드 RC(Reliable Connection)을 통해 연결을 하게 되고 그 이후에는 인피니밴드 RC기반의 RDMA 을 통해 통신을 하게 된다. 인피니밴드에서 두 노드 사이에 연결을 위해서 CM (Communication management)이라는 별도의 프로토콜을 이용하는데 초기자측의 CM이 Connection Request (REQ) 메세지를 타겟으로 보내고, 타겟의 CM은 Connection Response(REP) 또는 Connection Reject (REJ) 메세지를 초기자에게 보낸다. 그러나 인피니밴드는 TCP와 같은 역할을 할 수 있는 프로토콜에 대한 세부 기술은 iSER 명세서에

는 없다. 현재 인핀밴드는 TCP/IP 망과 연동을 위해 IPoIB(IP Over Infiniband)라는 프로토콜은 있지만 IPoIB는 비신뢰성 데이터그램으로 동작하기 때문에 신뢰성있는 연결을 필요로 하는 인피니밴드 기반의 iSER 에는 적용할 수 없다.

다음 그림 8은 인피니밴드 망에서 저장장치에 접근하기 위한 초기자와 타겟 사이의 과정을 나타낸 것이다. 여기서 초기자는 인피니밴드망 내의 저장장치 서비스를 이용할 임의의 HCA 노드이고 타겟은 저장장치 서비스를 제공하는 TCA 노드이다. 다음은 연결 및 로그인 과정을 설명한 것이다.

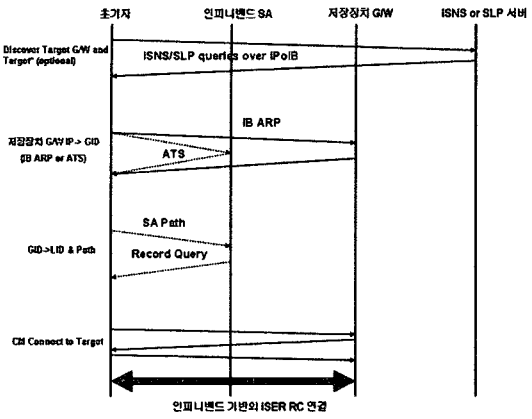


〈그림 8〉 iSER/IB 로그인 과정

1. 초기자가 타겟으로 SCSI Login Request PDU를 전송
2. 타겟은 초기자에게 SCSI Login Response PDU를 전송
3. 선택적으로 1, 2 단계에서 주고 받은 keys를 iSER 계층에 전달
4. RDMA 리소스 할당
5. T=1, NSG=FullFeaturePhase를 포함한 SCSI Login Request PDU 전송
6. iSER 모드를 전환하기 위해 Datamover를 Enable(*=바이트 스트림 모드로 마지막 iSCSI PDU 전송)
7. 바이트 스트림 모드에서 SCSI Login Response 전송
8. iSER Hello 메시지를 담은 RDMA Send 메시지 전송
9. iSER HelloReply 메시지를 담은 RDMA Send 메시지 전송

– iSER/IB 기반의 저장장치 Dicovery 과정

인피니밴드 기반의 iSER에서 저장장치를 Discovery하기 위한 과정을 다음 그림 9를 통해서 설명한다. 초기자는 iSNS(internet Storage Name Server) 나 SLP(Service Location Protocol) DB(DataBase)로 저장장치 이름, 영역(Zone), 사용자 인증서에 대한 질의를 하고 응답으로 타겟 IP을 얻게 된다. IPoIB을 이용해 인피니밴드 ARP(Address Resolution Protocol)에 대한 응답을 저장장치 게이트웨이(타겟:Target)로부터 받는다. 또는 SA(Subnet Administrator)로부터 ATS(Address Translation Service)에 대한 응답을 받는다. 초기자는 다시 SA에게 Path record에 질의 및 응답을 받게 되는데 이 과정에서는 인피니밴드에서 사용하는 LID(Local Identifier)를 얻게 된다. 마지막으로 CM은 해당 LID 정보를 가지고 타겟 저장장치로의 연결을 시도하게 된다.[1]



〈그림 9〉 저장장치 Discovery 과정

Ⅳ. 결론 및 향후 과제

인피니밴드 기반의 저장장치 기술은 기존의 TCP/IP 기반의 저장장치 기술에 비해 많은 장점을 가지고 있다. 실제로 데이터 센터나 대규모 컴퓨팅 자원을 필요로 하는 클러스터링 분야에서는 기존의 기가빗 이더넷 장비를 인피니밴드 시스템으로 전향하고 있는 추세이다. 먼저 네트워크 기반의 저장장치를 비교하기 위해 TCP/IP와 인피니밴드 망 사이의 차이를 비교해본다. 서비스 대역폭을 보면 TCP/IP는 Best Effort 서비스를 제공하고 인피니밴드는 Loss Less 서비스를 제공한다. 그리고 TCP/IP는 프로토토콜 스택에서 거의 80%가 소프트웨어로 구성되어 CPU 사용도가 높은 반면 인피니밴드는 60~70% 까지 하드웨어로 구성되어 CPU 사용도가 낮다. 그리고 QoS(Quality of Service) 측면에서 보면 TCP/IP는 대역폭이 낮고 높은 지연시간을 가지고 있지만 인피니밴드는 VL(Virtula Lane)을 통하여 하드웨어 영역에서 QoS를 보장하고 있다. 마지막으로 망 이용 형태를 보면 TCP/IP는 인터넷, LAN(Local Area Network), WAN(Wide Area Network)을 구성하는데 이용되고 인피니밴드는 컴퓨팅과 저장장치 클러스터링에 이용하기 위해 만들어졌다.

네트워크기반의 저장장치를 구현하는 프로토콜에 대해서는 이미 언급을 했다. 다음 표2는 TCP/IP 기반의 iSER와 인피니밴드 기반의 iSER를 비교한 표이다. 표준화 정도와 하드웨어 차이, 그리고 적용 분야, OS 지원여부로 구분하였다.[5]

〈표 2〉 인피니밴드 vs. TCP/IP 기반의 iSER

	인피니밴드 기반 iSER	TCP/IP 기반 iSER
표준화	IBTA로부터 인피니밴드 표준 프로토콜, 인터넷 Draft 단계	IEEE로부터 표준 이더넷 사양이 아님, 인터넷 Draft 단계
10Gbps 대역폭 10u sec 지연	HCA,TCA는 Full Offload 지원 10Gbps 지원 2.7usec 지연	보도와 TOE엔진 10Gbps 지원 10usec 지연
표준 툴셋즘 개발 공동체	IBTA (하드웨어와 소프트웨어) OpenIB	RDMA Consortium (하드웨어와 소프트웨어) OpenRDMA
저장장치, 네드워킹 및 컴퓨팅 관련 어플리케션 적용	MPI(컴퓨팅), SRP(저장장치)에 적용가능	TCP 관련 법률적 문제
사용분야 관련 Market OS지원	컴퓨팅, 저장장치 파러스터링 분야 Enterprise Financial 관성 증대 리눅스 커널에 포함	기존 표준 이더넷망 초기 단계 지원 OS 없음

인피니밴드는 고성능 컴퓨팅 시장과 데이터 센터 분야에서 급속한 성장을 보이고 있다. 현재 대부분의 주요 서버 제조업체는 시장에서 인피니밴드 솔루션을 제공하고 있고, 광범위한 어플리케이션 부분에서도 인피니밴드 기술을 적용 · 배치하고 있다. 이와 같은 성공은 곧 인피니밴드 기반의 저장장치에 대한 요구를 불러일으키게 될 것이다. 시장은 이미 10Gb/s 성능과 전송 계층의 offload는 산업 표준이 되고 있고, 이에 인피니밴드는 이미 10Gb/s 성능과 RDMA를 지원하는 어뎁터를 시장에 내놓았고, 30Gb/s의 성능을 갖는 어뎁터를 개발 중이다. 또한 클러스터 환경에서 가장 중요한 제품 선택 요인 가운데

하나인 가격대 성능비는 타 기술에 비해 가장 우수한 것으로 나타났다. 저장장치 측면에서 인피니밴드 소프트웨어, 플랫폼, 관리, 어플리케이션 지원을 통한 많은 이득은 저장장치 시장에서의 인피니밴드의 요구를 더욱더 가속화 시킬 것으로 전망하고 있다.

참고 문헌

[1] J. Hefferd., "iSER Over Infiniband", Internet Draft draft-hufferd-iser-ib-00.txt, July 2005

[2] M. Ko., "iSER Over TCP/IP", Internet Draft draft-ietf-ips-04.txt, June 2005

[3] M. Chadalapaka, H. Shah, U. Elzur, P. Thaler, and M. Ko. A study of iSCSI extensions for RDMA (iSER). In ACM SIGCOMM workshop on Network-I/O convergence: experience, lessons, implications, August 2003.

[4] InfiniBand Trade Association. InfiniBand Architecture Spec., Release 1.1, October 24 2004.

[5] Mellanox Technologies. Mellanox InfiniBand Storage, July 2004.

[6] K. Z. Meth and J. Satran. Design of the iSCSI Protocol. In 20th IEEE Symposium on Mass Storage Systems, 2003.

[7] J. C. Mogul. TCP Offload Is a Dumb Idea whose Time Has Come. In 9th Workshop on Hot Topics in Operating Systems (HotOS IX), May 2003.

[8] RDMA Consortium. iSCSI Extensions for RDMA (iSER) and Datamover Architecture for iSCSI (DA) Specifications, 2004.

[9] RDMA Consortium. iWARP Protocol Suite Specifications, 2004.

[10] Technical Committee T10. SCSI RDMA Protocol, 2002.

저자소개

김 영 환

2001년 부경대학교 정보통신 공학과 학사
2003년 성균관대학교 대학원 컴퓨터 공학과 석사
2003년 – 현재 전자부품연구원 지능형 정보시스템 연구센터 전임연구원
주관심 분야 클러스터링, 저장장치, 인피니밴드, 센서 네트워크

전 기 만

2000년 한양대학교 전기공학과 학사
2000년 – 2001년 삼보컴퓨터 연구소 연구원
2001년 – 현재 전자부품연구원 지능형 정보시스템 전임연구원
주관심 분야 인피니밴드, 센서네트워크

박 창 원

1986년 – 1988년 동양정밀(주) 중앙연구소 주임연구원
1988년 – 1993년 효성컴퓨터(주) 중앙연구소 선임연구원
1993년 – 현재 전자부품연구원 지능형 정보시스템 연구센터장
주관심 분야 저장장치, 센서네트워크

손 재 기

1998년 경기대학교 전자계산학과 학사
2001년 경기대학교 대학원 전자계산학과 석사
2001년 – 현재 전자부품연구원 지능형 정보시스템 연구센터 선임연구원
주관심 분야 저장장치, 임베디드, 센서네트워크