채널기반형 네트웍에서의 IPoIB 프로토콜 성능평가

A Performance Evaluation for IPoIB Protocol in Channel based Network

전기만, 민수영, 김영환
Ki Man Jeon, Soo Young Min, Young Wan Kim

Abstract - As using of network increases rapidly, performance of system has been deteriorating because of the overhead and bottleneck. Nowadays, High speed I/O network standard, that is a sort of PCI Express, HyperTransport, InfiniBand, and so on, has come out to improve the limites of traditional I/O bus. The InfiniBand Architecture(IBA) provides some protocols to service the applications such as SDP, SRP and IPoIB. In our paper, We explain the architecture of IPoIB (IP over InfiniBand) and its features in channel based I/O network. And so we provide a performance evaluation result of IPoIB which is compared with current network protocol. Our experimental results also show that IPoIB is batter than TCP/IP protocol. For this test, We use the dual processor server systems and Linux Redhat 9.0 operating system.

Key Words :IPoIB, SDP, InfiniBand, 고속 I/O Network

1. 소개

인터넷 기술과 멀티미디어 처리 기술의 발달로 데이터 처리속도 및 처리량의 요구 또한 급상승하고 있다. 이러한 요구을 만족시키기 위해 기존 기업용 서버, 고속 연산용 서버와는 다른 인터넷 서버구축을 위한 제품과 환경구축이 필요하게 되었다. 또한 대용량 저장장치와 연동하여, 사용자의 요구에 따른 신속하고, 원활한 대용량의 데이터 제공을 가능하게 하는 고속 I/O 규격과 운영체계의 등장 및 대중화가 요구되고 있다.[2]

현재 서버 및 스토리지 시스템에서 사용되는 주된 입출력 방식은 PCI 버스를 통해 연결망에 접속하는 다중 계층 구조이다. 그러므로 각 노드의 전송 능력은 PCI 버스 성능에 의해 제한되고 있다. PCI 버스는 공유 버스 구조를 갖는 입출력의 구조적 단점으로 확장성 및 성능향상에 한계를 벗어나지 못한다. 이러한 한계를 극복하기 위해 모든 연결을 스위치 기반형 단일 연결망으로 구성하여, 가용성 및 확장성을 보장하고, 점대점 연결을 통한 고속 직렬 통신 방법을 사용 전송속도를 향상시키는 새로운 시스템 연결방식이 대두되기 시작했다. 현재 PCI Express, HyperTransport, RapidIO 규격 제품들 이 등장되고 있으며 전체 네트웍의 상태파악 및 효율적인 관리와 운영에 장점을 가진 인피니밴드 표준에 따른 제품들 또한 등장되고 있다.

인피니밴드에는 IPoIB(IP over InfiniBand), SDP, SRP등의 여러 응용프로토콜이 존재하며. IPoIB는 인피니밴드와 같은 새로운 네트웍 기술이 가진 특징을 기존 이더넷 망과 호환성 있게 사용되어 지도록 만든 프로토콜 이다. 인피니밴드는 프로세서와 I/O시스템, 그리고 장치에 대한 상호 연결을 위한 표준이므로, IP는 이 상호 연결망에서 사용할 수 있는 매우 중요한 트래픽의 한 종류이다. 또한, 인피니밴드 망 내부에서 IP 트래픽을 다루기 위한 표준화된 방식 가운데 기존 이더넷 망에서 보다 패킷 처리량에 있어 많은 이점을 가지고 있다. IPoIB는 데이터 전송에 있어 UD(Unreliable Datagram)전송 방식뿐만 아니라 멀티캐스팅 기능을 지원하기 때문에 망 내의 트래픽을 줄일 수 있는 장점이 있다.[3,4,7]

본 논문 구조는 다음과 같은 내용으로 구성되어 진다. 2장에서 인피니밴드의 구조와 망 관리 구성요소 그리고 IPoIB구조에 대해 소개하며, 3장에서는 Netperf 소프트웨어 툴을 이용하여 IPoIB의 성능평가 내용을 설명한다. 마지막으로 4장에서는 요약과 함께 성능 평가를 통한 결론을 정리한다.

2. 인피니밴드

2.1 인피니밴드 구조

차세대 서버 시스템구조 요구사항 변화에 따라 Compaq, Dell, IBM, HP, Microsoft, Sun, Microsystems 등이 주축이 되어 1999년 8월 IBTA(InfiniBand Trade Association)가 결성되었고, 200년 9월 발표된 InfiniBand Architecture 표준규

저자 소개
전 기 만 : 전자부품연구원 연구원
민 수 영 : 전자부품연구원 책임연구원
김 영 환 : 전자부품연구원 전임연구원

격1.0 이후 202년 11월 표준규격1.1까지 발표 되고있다.

인피니밴드는 상호 독립적인 프로세서 플랫폼, 입출력 처리 그리고 입출력 장치를 연결하는 시스템 연결망으로 통신 및 관리 메커니즘을 포함하며, 소규모 서버 시스템에서 대규모 인터넷 서버 및 슈퍼컴퓨터에 이르기까지 다양한 영역에서 사용가능하다. 더욱이 인터넷 통신 프로토콜에 친숙하게 설계되어 인터넷 접속, 인트라넷 접속원격 서버 접속이 쉽게 구현될 수 있다.[2,6]

인피니밴드 연결망은 스위치 기반의 비점형 연결망으로 종단에 프로세서 노드, 입출력 노드가 연결되는 여러 개의 서브넷(Subnet)으로 구성된다. 서브넷은 종단노드, 스위치, 라우터로 구성되며 서브넷간의 연결은 라우터를 통한다. 하나의 서브넷에는 최대 65,536개의 종단 노드가 연결 가능하며, 각 종단 노드는 인피니밴드 연결망 접속을 위한 채널 어뎁터를 가지며, 프로세서 노드 쪽은 호스트 채널 어뎁터(HCA: Host Channel Adapter), 입출력 처리 노드 및 장치 쪽에서는 타겟 채널 어뎁터(TCA: Target Channel Adapter)가 사용된다.

인피니밴드 링크는 양방향 점대점 통신 채널로 되어 있어, 1x, 4x, 12x의 독립적인 Tx, Rx 직렬연결 포트를 가지며, 2.5Gbps에서 최대 30Gbps의 데이터 전송 속도를 지원한다.[1]

2.2 인피니밴드 관리 구성요소

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

서브넷 매니저는 스위치, 라우터, 채널 어뎁터 등 인피니밴드 구성요소의 형상 관리를 수행하며, 인피니밴드 연결망의 모든 노드는 서브넷 매니저와의 통신을 위해서 서브넷 매니지먼트 에이전트(SMA: Subnet Management Agent)를 탑재하고 있어야 한다. 서브넷 매니저는 서브넷 형상 검출 관리 데이터베이스 구성 및 운영, LID 및 GID 주소 매핑 서비스 등의 기능을 수행한다. 모든 종단 노드는 서브넷 매니지먼트 에이전트를 제공해야 한다.

서브넷 매니지먼트 에이전트는 서브넷 매니저가 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 초기화되기 전에 MAD 패킷을 전송할 수 있게 한다.

종단 노드는 서브넷 매니지먼트 에이전트 이외에 부가적인 관리형상 지원을 위해 제너럴 서비스 에이전트(GSA: General Service Agent)를 가질 수 있으며, LID를 통한 경로 설정 방법을 사용하는 GSI(General Service Interface)를 통해서 통신한다. 제너럴 서비스의 종류는 노드에서 타 노드와 전송 서비스 정보를 찾고, 경로 탐색을 수행하는 서브넷 Administration, TCA 하단에 연결된 입출력 장치를 관리하는 디바이스 매니저, 입출력 리소스 관리에 사용되는 디바이스 Configuration등이 있다.[2]

2.3 IPoIB(IP over InfiniBand)

인피니밴드 네트웍에서 IP를 사용하기 위해서는 인피니밴드 프레임 내 페이로드 부분에 IP와 ARP에 대한 정보를 인캡슐레이션 해야 하며, 이 과정을 통해 IPoIB의 한 프레임이 만들어 진다. 다음 그림 1은 IP/ARP 패킷을 포함한 인피니밴드 프레임의 각 필드에 대한 정보를 나타내고 있다. IPoIB 프레임은 인피니밴드 헤더, 페이로드, 그리고 인피니밴드 트레일러의 3가지 항목으로 분류되며, 인피니밴드 프레임 헤더는 로컬 서브넷의 ID를 나타내는 LRH, 전체 서브넷에서의 ID인 GRH, 로컬 또는 목적지 QP번호, 서비스 클래스를 나타내는 SL등의 값을 포함하고 있는 BTH, 그리고 RDMA을 위한 ETH로 구성되어 있다. 페이로드 부분에는 실제 전송할 데이터를 포함하고 있는데, 4-octet 헤더와 실제 IP에 대한 정보를 가지고 있다. 마지막으로 인피니밴드 트레일러는 해당 프레임에 대한 CRC 정보를 갖는다.[4,8]

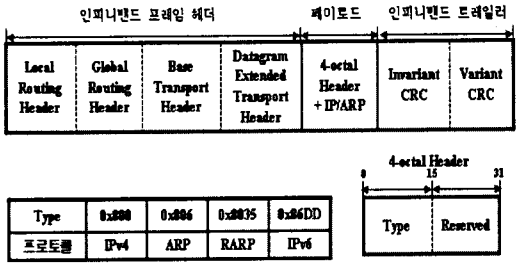

그림 1. IPoIB 프레임

3. IPoIB 성능평가 및 분석 결과

성능평가를 위해 사용된 구성은 듀얼 Xeon 서버 프로세서, 3GB 메모리의 H/W 시스템과 Red Hat Linux 9.0의 OS환경이며, Mellanox사의 MT23108 호스트 채널 어뎁터와 Intel사의 Pro/1000 XT NIC를 대상으로 테스트가 이루어졌다.

성능평가에 사용된 인피니밴드 소프트웨어로는 IPoIB 미들웨어 드라이버가 포함되어 있는 OpenIB 인피니밴드 스택을 사용하였다. OpenIB는 여러 업체들이 인피니밴드 관련 공동작업을 위해 만든 인피니밴드 프로젝트 단체이며, IBTA에서 표준화 작업이 진행되고 있는데 반해 실제 개발은 OpenIB에서 이뤄지고 있다. OpenIB에서 배포된 프로토콜 계층구조에는 커널을 통해 호스트 채널 어뎁터에 접근하는 방식과 RDMA를 이용한 커널을 by-pass하여 호스트 채널 어뎁터에 접근하는 방식을 이용하여 메시지를 전송하는 구조가 제안되어있다.

성능평가를 위한 테스트 툴은 HP에서 제작된 Netperf를 사용했다. 두 종류의 Netperf를 사용했으며, 그 중 하나는 기가빗 이더넷 측정을 위한 것이고 다른 하나는 Netperf를 수정하여 IPoIB를 측정하기 위한 Netperf_IB 버전이다. 전자는 원격지 노드와 연결을 시도하기 위해 소켓 구성시 어드레스 패밀리 가운데 AF_INET를 사용하도록 하고, 후자는 소켓을 구성할 때 새로운 어드레스 패밀리를 추가하였는데,

AF_INET_IB를 사용하도록 수정했다.[5]

그림 2-4는 IPoIB와 기가빗 이더넷에 의해 얻어지는 패킷 처리량 결과이다. 성능 평가 방법을 위해 메시지를 크기별로 구분하였으며, 임의적으로 패킷 처리량의 변화에 따라, 작은 사이즈의 메시지와 큰 사이즈의 메시지를 구분하여 측정하였다. 메시지 크기에 따른 처리량에 대해 두드러진 결과를 보이는 것이 IPoIB이다. 작은 사이즈의 패킷 처리량에서는 메시지 크기가 16바이트에서 128바이트 까지 기가빗 이더넷에 비해 패킷 처리량이 적지만 256바이트 부터는 기가빗에 비해 패킷처리량이 많다는 것을 알 수 있으며, 이유는 기가빗 이더넷에 비해 수신 가능 메시지나 송신 가능 메시지 등에 의해 처리해야 할 패킷이 많기 때문이다. 큰 사이즈의 메시지 패킷 처리량에서는 기가빗 이더넷에 비해 IPoIB가 약 1.5배 가량 많다는 것을 알 수 있다.

기가빗 이더넷과 IPoIB의 패킷 처리량에 대한 최적의 메시지 사이즈는 그림 2에서 기가빗 이더넷의 경우 512바이트에서 가장 많은 처리량을 보이고, 그림 3에서의 IPoIB는 128K 바이트일 경우 가장 효율적인 패킷 처리가 된다는 것을 확인할 수 있다.

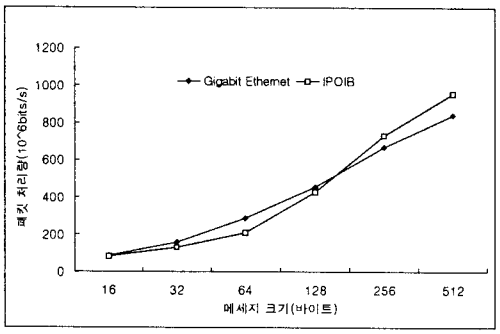

그림 2. 작은 사이즈 메시지에 대한 패킷 처리량

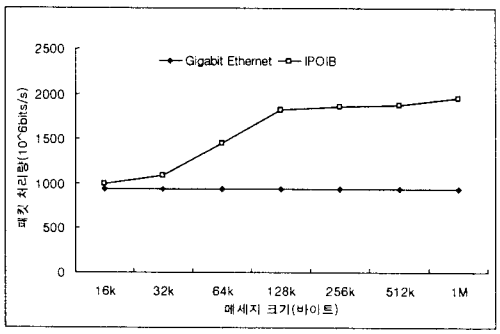

그림 3. 큰 사이즈 메시지에 대한 패킷 처리량

네트웍 트래픽의 처리는 호스트의 CPU 부하와 밀접한 관련이 있다. 기가빗 이더넷과 IPoIB의 프로토콜 프로세싱을 위한 CPU의 부하를 앞에서 측정한 패킷 처리량에 따라 각 메시지의 크기별로 CPU 부하를 측정하였다. 메시지 크기에 따른 CPU Utilization을 1%의 CPU Utilization 당 패킷 처리량으로 나타냈을 때 그림 4에서 보듯이 IPoIB에 비해 기가빗 이더넷이 CPU Utilization에 대한 패킷 처리량이 높은 것으로 확인되었다.

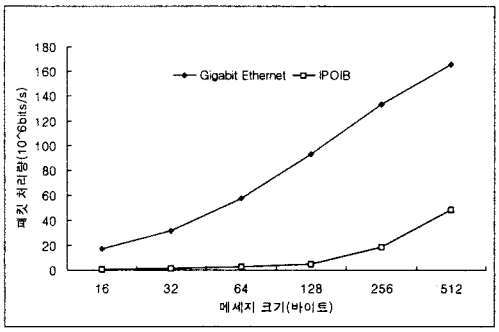

그림 4. CPU 1% Utilization 당 패킷 처리량

4. 결론

본 논문에서는 현재 개발진행 중인 인피니밴드 구성요소들이 어떠한 특성을 가지는지를 확인하고, 고성능 대용량 네트웍에서, 기존 네트웍 구성요소와의 차별점과 우월성에대해 인피니밴드의 IPoIB 프로토콜을 이용하여 알아보는 것 이었다. 평가결과에서 확인할 수 있듯이 인피니밴드 프로토콜 스텍에서 IPoIB 미들웨어 드라이버를 이용함으로, 기존 이더넷 프로토콜에 비해 상대적으로 높은 패킷 처리량을 보였다. 그러나 1%의 CPU Utilization 당 패킷 처리량에서는 다소 높은 수치를 보이고 있어 이에 대한 보다 심도 있는 연구가 요구된다.

참 고 문 헌

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

[2] 박경, 모상만, "InfiniBand: 차세대 시스템 연결망", 정보과학회지, vol. 19, no. 3, pp. 43-51, March 2001.

[3] J. Wu, P. Wyckoff, and D.K. Panda. PVFS over Infiniband: Design and Performance Evaluation. In ICPP, 2003.

[4] IPoIB Internet-draft. draft-ietf-ipoib-architecture-04.txt

[5] netperf. http://www.netperf.org/

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

[7] Cohen, A, "A performance analysis of the sockets direct protocol (SDP) with asynchronous I/O over 4X InfiniBand", 2004 IEEE International Conference, pp. 241-246, April 2004.

[8] IPoIB Working Group. http://www.ietf.org/html.charters/ipoib-charter.html