인피니밴드 네트워크에 대한 응용 레벨 성능 분석

차광호° 김성호
한국과학기술정보연구원 슈퍼컴퓨팅센터
{khocha°, sungho}@kisti.re.kr

Application level performance evaluation of Infiniband Network

Kwangho Cha°, Sungho Kim
Supercomputing Center
Korea Institute of Science and Technology Information

요 약

클러스터 시스템의 성능과 관련되어 중요성이 강조되는 분야로 SAN(System Area Network)을 이야기할 수 있다. 특히 Infiniband 제품의 출시는 이 분야에 대한 관심을 더욱 집중시키는 계기가 되었다. 이에 본 논문에서는 Infiniband가 보여주는 단순한 네트워크적인 특징 이외에 응용 프로그램의 성능에 어떠한 영향을 미치는가에 대하여 실험하고 분석하였다.

1. 서론

1999 년, 차세대 I/O 아키텍쳐라 불려지던 Future I/O 와 Next Generation I/O 의 통합은 Infiniband 연구의 시발점이 되었다[1]. 그 후 초고속 통신을 지원하는 Infiniband 제품이 출시되면서 클러스터시스템을 위한 고성능 네트워크로서의 관심이 커지고 있다.

본 논문에서는 Infiniband 를 이용한 클러스터 시스템에서의 응용 프로그램들의 성능을 측정하여 기존 SAN (System Area Network)을 이용하는 경우와 비교하였다. 즉 Infiniband 가 갖는 네트워크적인 특성이 응용 프로그램의 성능 개선에 어느 정도 영향을 미치는가에 대하여 살펴보았다. 본 논문의 구성은 다음과 같다. 2 장에서는 성능 측정에 사용된 응용 프로그램들에 대하여 설명하고, 3 장에서는 실험에 사용된 하드웨어 및 소프트웨어적인 특성을, 4 장에서는 실험 결과, 5 장에서는 결론 및 향후 계획에 대하여 기술한다.

2. 관련 성능 분석 도구

2.1 MPI 성능 테스트

메시지 전달 방식을 지원하는 가장 대표적인 병렬 프로그래밍 환경으로 MPI(Message Passing Interface)를 들 수 있으며 네트워크 성능과 직접적으로 연관된다고 할 수 있다. 이러한 MPI의 주요 함수에 대한 성능을 측정하는 도구로서 Pallas MPI Benchmarks(PMB)[1] 가 있다[2]. 주요 MPI함수를 대상으로 전송되는 메시지 크기를 증가시키면서 성능을 측정하는 방식이다.

2.2 병렬 응용 프로그램 벤치 마크

병렬 프로그램의 유형별 성능을 측정하는 벤치마크 프로그램으로 NPB(NAS Parallel Benchmarks)가 있다[3]. 주요 병렬 프로그램을 대표적인 8 종류로 분류하여 이에 대한 성능을 측정하는 벤치마크 프로그램의 집합이다.

클러스터 시스템뿐만 아니라 슈퍼 컴퓨터의 성능을 측정하는 대표적인 응용 프로그램으로 HPL(High Performance Linpack Benchmark)이 있다[4]. 선형 시스템을 풀기 위한 소프트웨어 패키지로서 매년 2 번씩 발표되는 Top500 슈퍼컴퓨터를 선정하는 기초 자료로 쓰이고 있다[5].

2.3 병렬 응용 프로그램 벤치 마크

지금까지 클러스터 시스템에서 고성능 네트워크는 주로 계산용 네트워크의 역할을 담당하여 왔다. 그러나 각 계산 노드에 대한 파일 서비스의 역할도 중요시되는 상황이어서 파일 서비스 네트워크로의 활용도 중요하게 고려되고 있다. 이에 본 논문에서는 연구용 병렬파일 시스템인 PVFS (Parallel Virtual File System)2[6]를 대상으로 네트워크의 종류에 따른 파일 시스템의 성능 변화를 측정하였다.

3. 실험 환경

본 장에서는 성능 측정에 사용된 하드웨어 및 소프트웨어의 구성과 종류에 대하여 설명한다.

3.1 하드웨어 구성

테스트를 위한 시스템은 크게 단위 노드와 각 노드를 연결하는 네트워크로 구분할 수 있다. 이때 단위 노드는 표 1 과

* 본 연구는 '㈜지티엠코리아'와 '㈜디지털헨지'의 장비 협조로 진행되었다.

[1] 현재는 Intel의 IMB(Intel MPI Benchmarks)라는 명칭으로 배포되고 있다.

같이 x86 기반의 아키텍쳐를 사용하였다. HPL 테스트의 메모리 사용량을 고려하여 각 노드에 3GB의 메모리를 장착하였고, Infiniband는 PCI-Express를 Myrinet은 PCI-X를 이용하여 장착하였다. 각 노드의 연결은 표 2, 3 과 같이 Mellanox사의 Infiniband장비와 Myricom사의 Myrinet장비를 이용하여 구성하였다[7,8].

[표 1] 단위 노드 사양

CPU	Intel Xeon 2.8 GHz (Nocona)
# CPU / node	2
Memory	3 GB
I/O Interface	PCI Express, PCI-X

[표 2] Infiniband 구성

Host Channel Adapter	MHEL-CF128-T - Dual 10Gb/s - 128MB Local SDRAM - PCI Express x8
Switch	MTS-2400 - 24 Ports (10Gb/s)

[표 3] Myrinet 구성

Interface Card	M3F-PCI64B-2 - 2MB Local Memory - PCI 64/32
Switch	Myrinet 2000 - 16 Ports (Fiber)

3.2 소프트웨어 구성

시스템 소프트웨어의 구성은 표 4 와 같다. 주요 컴파일러는 Intel 컴파일러를 이용하였으며, Infiniband를 사용시에는 mvapich를 Myrinet을 사용시에는 mpichgm을 각각 MPI 라이브러리로 사용하였다. HPL 테스트시 사용되는 BLAS 라이브러리는 기본적으로 goto 라이브러리를 사용하였다[9].

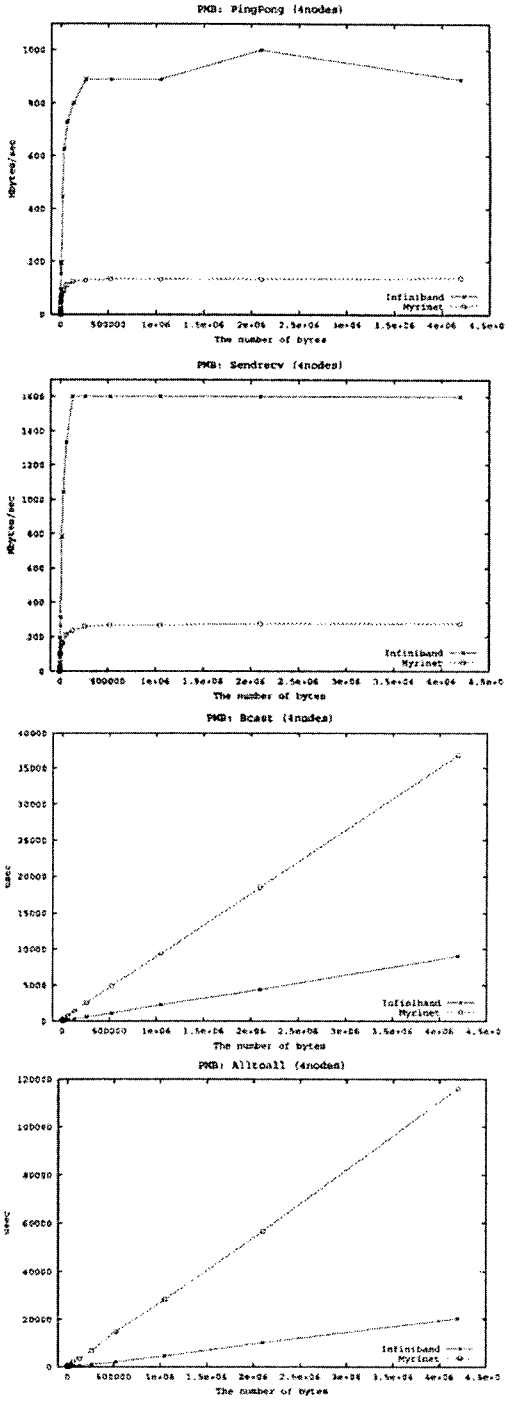
[표 4] 소프트웨어 구성

Linux	2.4.21(smp)
Compiler	Intel Compiler 9.0 (C++, Fortran)
Infiniband Driver	IBGD-1.7.0
Myrinet Driver	GM-2.1.9
MPI Library	mvapich-0.9.4, mpichgm-1.2.6
BLAS Library	libgoto_prescott-64-r0.99-3
Parallel File System	pvfs2-1.2.0

4. 실험 결과

4.1 MPI 성능 테스트

그림 1 은 PMB 테스트 결과의 주요 내용을 보여주고 있다. 결과에서 보듯이 MPI의 각 함수는 통신 네트워크의 성능에 직접적인 영향을 많이 받아 Infiniband를 사용한 경우, 상대적으로 좋은 결과를 보여주고 있다.

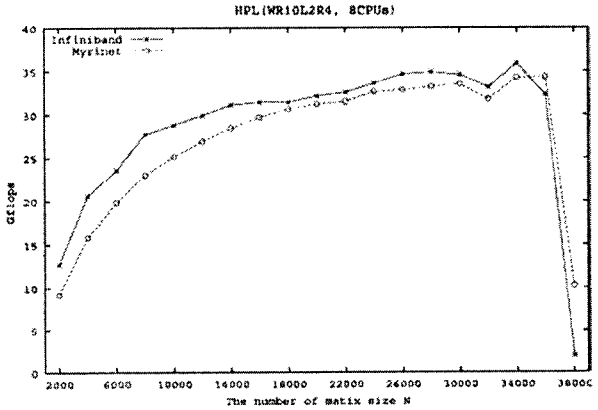


[그림 1] PMB 테스트 결과

4.2 병렬 응용 프로그램 벤치 마크

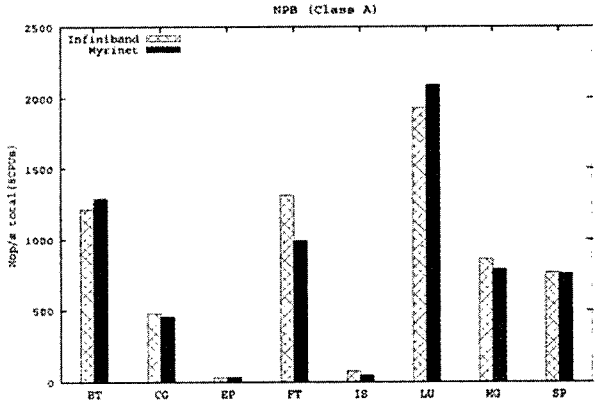
클러스터 시스템의 성능을 평가할 때는 실행되는 응용 프로그램의 특성도 주요 고려 사항이 된다. 즉 실행되는 응용 프로그램의 통신대 계산비율(Communication to Computation Ratio)가 상당히 낮은 경우에는 네트워크 시스템의 영향력이 약해질 수 있기 때문이다.

그림 2는 클러스터 시스템을 포함한 슈퍼컴퓨터의 성능 측정에 기준이 되는 HPL테스트의 결과이다. 표 1에 기술된 4대의 단위 노드를 이용하여 테스트를 진행하였다. 결과와 같이 Infiniband를 사용한 경우 최대 실측 성능이 Myrinet을 이용하였을 때 보다 4% 정도 우수하게 나타났다.



[그림 2] HPL 테스트 결과

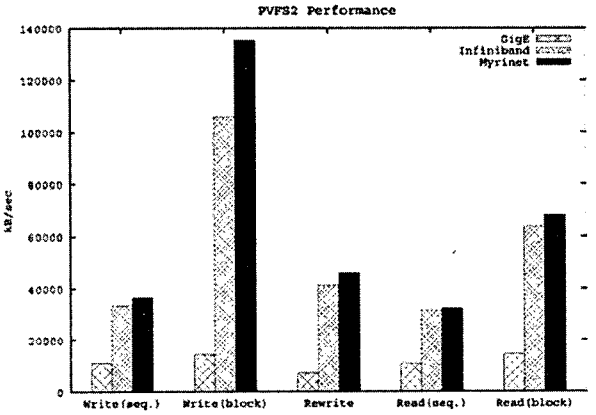
그림 3은 NPB 테스트 결과의 보여주고 있다. 통신이 적은 응용 프로그램에서는 성능 차이가 미비하였으며, 일부 응용 분야에서는 Myrinet을 사용한 결과가 오히려 더 좋은 결과를 보이는 경우도 확인할 수 있었다. 이는 MPI 함수의 구현 차이에서 발생한 것으로 예측되나 정확한 원인 파악은 향후 연구 내용으로 계획하고 있다.



[그림 3] NPB 테스트 결과

4.3 병렬 파일 시스템 테스트

그림 4는 병렬 파일시스템의 일종인 PVFS2를 서로 다른 3종류의 네트워크를 이용하여 구성하고 파일 시스템의 기본 명령에 대한 성능을 측정한 결과이다. 일반적인 Gigabit Ethernet에 비하여 Infiniband와 Myrinet이 좋은 성능을 보여주고 있으면 앞서 언급한 일부 테스트와 같이 Myrinet이 보다 우수한 성능을 보여주었다. 이 역시 PVFS2 의 특징과 연계하여 확실한 원인 파악이 필요한 부분이라고 예측된다.



[그림 4] PVFS2 성능

5. 결론 및 향후 계획

본 논문에서는 Infiniband를 이용한 클러스터 시스템에서의 응용 프로그램 성능을 조사하였다. 고성능 네트워크 성능이 직접적으로 반영되는 MPI 함수의 경우 가장 확실한 성능 향상을 확인할 수 있었으며 주요 응용 프로그램의 성능 향상도 확인할 수 있었다. 그러나 일부 응용 프로그램의 경우 기존의 Myrinet을 이용한 성능이 보다 우수하게 나오는 경우가 있어서 Infiniband를 사용하는 부분에 있어서 개선되어야 하는 부분이 있다고 예측되며 차후 연구 내용으로 계획 중이다.

참고문헌

[1] InfiniBand Trade Association, http://www.infinibandta.org
[2] Pallas MPI Benchmarks, http://www.pallas.de/e/products/index.htm
[3] Rob Van Der Wijngaart, "NAS Parallel Benchmarks, Version 2.4," NAS Technical Report NAS-02-007, 2002.
[4] HPL - A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers, http://www.netlib.org/benchmark/hpl
[5] TOP500 Supercomputer sites, http://www.top500.org
[6] Parallel Virtual File System Version 2, http://www.pvfs.org/pvfs2
[7] Mellanox Technologies, http://www.mellanox.com
[8] Myricom, http://www.myri.com/
[9] High-Performance BLAS by Kazushige Goto, http://www.cs.utexas.edu/users/flame/goto