Tachyon 클러스터 시스템에서의 네트워크 성능 평가

차광호, 김성호, 이식
한국과학기술정보연구원 슈퍼컴퓨팅본부
e-mail : khocha@kisti.re.kr

Network performance evaluation of Tachyon cluster systems

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

요 약

멀티 코어 또는 매니 코어 기반 시스템을 클러스터 시스템의 단위 노드로 활용하면서 클러스터 시스템은 다양한 형태의 노드내(Intra-node)및 노드간(Inter-node)네트워크를 가지게 되었다. 최적화된 어플리케이션의 개발을 위해서는 해당 시스템의 이러한 네트워크적 특징을 미리 파악하는 것이 중요하다고 할 수 있다. 본 논문에서는 서로 다른 계산 노드를 사용하는 클러스터 시스템에서 네트워크 성능을 비교 분석하였다.

1. 서론

하드웨어 기술의 발달로 인하여 일반 PC 뿐만 아니라 클러스터 시스템의 단위 노드로 사용되는 서버에서도 멀티 코어 또는 매니코어 기술을 쉽게 찾아볼 수 있게 되었다[1].

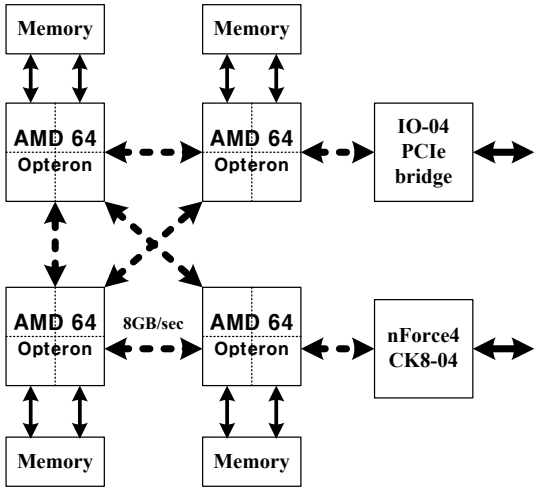
시스템 별로 다양한 CPU 소켓간 연결 기술이 개발되어 다양한 형태의 노드내(Intra-node)및 노드간(Inter-node)네트워크를 구성하고 있다. 본 논문에서는 KISTI 에서 보유하고 있는 Tachyon 클러스터 시스템에서의 네트워크 성능을 측정하고 분석하였다. 1,2 차 시스템으로 구성된 Tachyon 시스템의 상이한 네트워크 특징을 MPI(Message Passing Interface)환경[2]에서 측정하고 비교 분석하였다.

2. Tachyon 클러스터 시스템

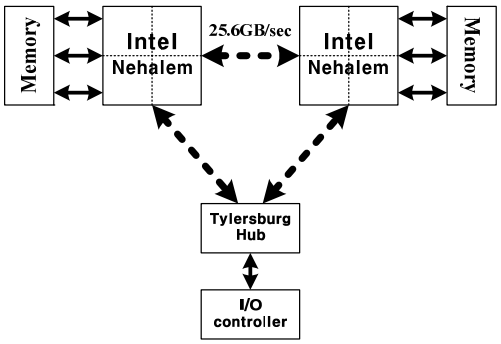
Tachyon 클러스터 시스템은 x86 기반 리눅스 클러스터 시스템으로 2 단계로 구축되었다. 각 단계별로 구축된 두 시스템은 서로 다른 시스템 아키텍쳐를 가지고 있으며, 단위 노드를 연결하는 인피니밴드 네트워크에도 성능 차이가 존재한다.

<표 1> 하드웨어 및 소프트웨어 구성

System	Tachyon 1 차	Tachyon 2 차
CPU	AMD Barcelona	Intel Nehalem
CPU Clock Speed	2.3GHz	2.93GHz
노드당 코어수	16(4core/4socket)	8(4core/2soket)
노드수	192	3200
총메모리	6TB(32GB/node)	76.8TB(24GB/node)
Interconnection N/W	Infiniband 4X DDR	Infiniband 4X QDR
OS	CentOS 4.6	RedHat 5.3
File System	Lutre 1.6.5.1	Lustre 1.8.1.1
MPI	MVAPICH2 1.4	MVAPICH2 1.4
Scheduler	SGE 6.1	SGE 6.2 u4



(그림 1) Tachyon 1 차 시스템 단위 노드



(그림 2) Tachyon 2 차 시스템 단위 노드

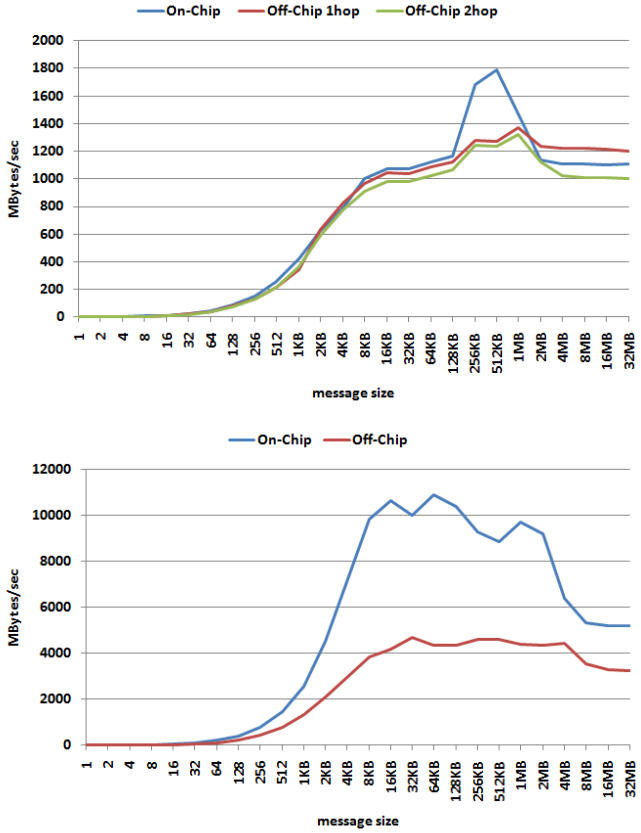
표 1 과 그림 1,2 에서 보는 바와 같이 두 시스템 간에는 단위 구조와 연결 네트워크에 변화가 생겨서 노드내 통신과 노드간 통신에 성능 차이가 있을 수 있음을 예상할 수 있다.

3. 성능 분석

본 실험에서는 대표적인 병렬처리 라이브러리인 MPI 기반의 테스트 코드를 사용하였다. 관련된 벤치마크 프로그램으로 IMB 를 고려할 수 있으나[3], 코드 수정의 용이성과 사용된 MPICH 라이브러리가 MVAPICH 임을 감안하여 OSU 에서 제공하는 OMB[4]를 수정하여 테스트 하였다.

3.1. 노드내 대역폭

단위 노드 내부에서의 프로세스간 대역폭을 측정하였다. CPU 의 Affinity 를 설정하여 소켓 내부에서의 대역폭, 소켓간의 대역폭에 대해서 측정하였다. 특히 1 차 시스템의 경우, 소켓간 통신도 1hop 거리와 2hop 거리로 나뉘기 때문에 이를 고려하여 실험하였다.

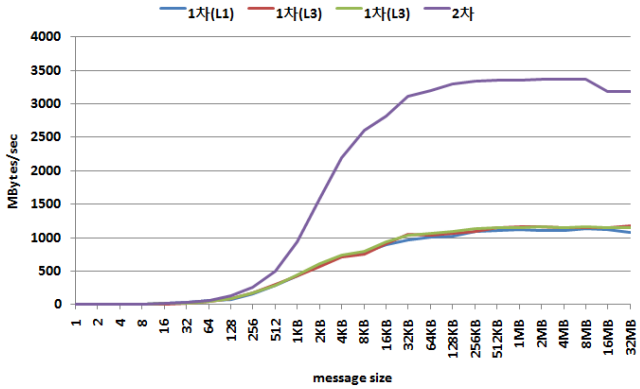


(그림 3) 노드내 대역폭: 위=1 차시스템, 아래=2 차시스템

그림 3 에서 보는 바와 같이 2 차 시스템에서의 증가된 대역폭을 확인할 수 있다. 특히 1 차 시스템의 경우, 소켓 내부와 소켓간 통신의 차이가 상대적으로 적은 반면, 2 차 시스템은 둘간에 확실한 차이를 보여주었다. 최고 성능을 기준으로 2 차 시스템이 1 차 시스템 보다 소켓 내부 통신은 약 5 배, 소켓간 통신은 약 2.4 배 향상된 대역폭을 보여주고 있다.

3.2. 노드간 대역폭

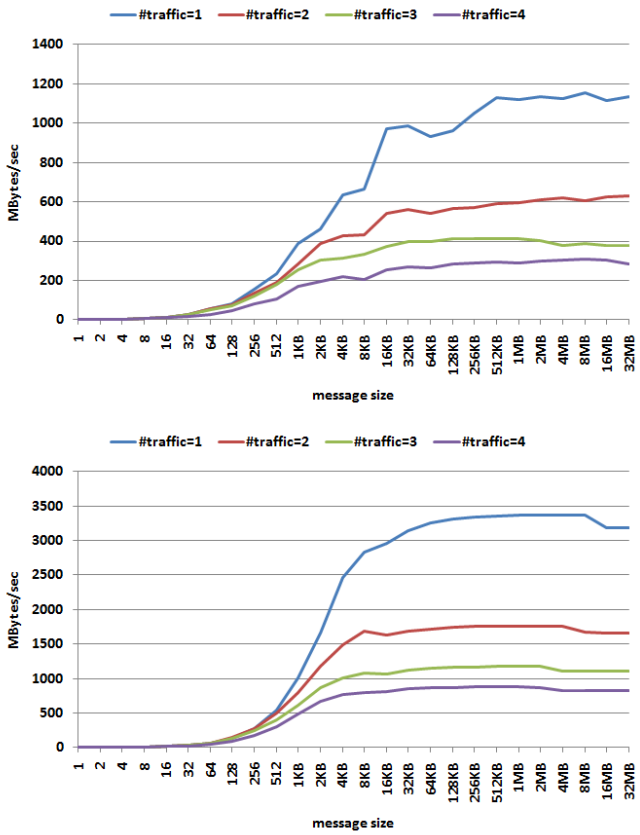
노드간 연결 네트워크인 인피니밴드의 성능을 확인하기 위하여 노드간 대역폭을 측정하였다. 1 차 시스템의 경우에는 인피니밴드 스위치 내에서 칩, 라인카드, 스위치로의 레벨을 나누어 측정하였고, 2 차 시스템은 실험 당시 인피니밴드 네트워크의 구성에 레벨에 따른 차이를 두지 않는 상황이어서 이를 구분하지 않았다.



(그림 4) 노드간 대역폭 비교

3.3.노드간 대역폭 성능 저하

기본적인 대역폭 측정에 추가하여 벤치마크 코드를 수정하여 노드간 통신량이 증가하는 경우의 성능 저하 정도를 확인하였다. 멀티코어 환경에서 두 노드 사이에 속한 프로세스들을 위한 복수의 통신이 존재 할 수 있다는 가정[5]에서 노드간 통신 링크의 수를 증가하면서 성능 저하 정도를 확인하였다.



(그림 5) 노드간 대역폭의 성능 저하:
위=1 차 시스템, 아래=2 차 시스템

두 시스템 모두에서 두 노드간 최대 대역폭을 트래픽 수로 나눈 만큼의 대역폭을 각각의 트랙픽별로 제공하는 것을 확인 할 수 있었다.

4. 결론

본 연구에서는 서로 다른 계산 노드를 사용하는 클러스터 시스템의 네트워크 성능을 MPI 환경에서 비교하였다. 노드내 통신과 노드간 통신으로 나누어 실행된 실험에서 전체적으로 Tachyon 2 시스템이 우수한 성능을 보여주었다. 2 차 시스템의 경우, 주의할 부분도 존재하였는데 1 차 시스템과 달리 노드내 통신에 있어서, 소켓 내부와 소켓간 통신의 성능 차이가 크게 나타나고 있었다. 즉, 동일 노드 내에 할당된 병렬처리 작업이라 하더라도, CPU Affinity 를 고려한 프로세스 할당이 필요함을 보여주는 결과였다.

참고문헌

[1] Top 500 Supercomputer Sites, Retrieved Aug. 17, 2010, from http://www.top500.org/

[2] Marc Snir, Steve Otto, Steven Huss-Lederman, David Walker, and Jack Dongarra, "MPI-The Complete Reference," 1998, The MIT Press.

[3] Intel® MPI Benchmarks 3.2 http://software.intel.com/en-us/articles/intel-mpi-benchmarks/

[4] OSU Micro-Benchmarks 3.2 http://mvapich.cse.ohio-state.edu/benchmarks/

[5] Vienne Jérôme, Martinasso Maxime, Vincent Jean-Marc, Méhaut Jean-François, ``Predictive models for bandwidth sharing in high performance clusters," in Proc. of the IEEE International Conference on Cluster Computing, pp. 286~291, Oct. 2008.