HPC 클러스터 구축을 위한 다양한 네트워크 성능 분석

홍정우*, 이보성**, 박형우*, 이상산*
*한국과학기술정보연구원 슈퍼컴퓨팅센터
**리눅스원(주)
e-mail:jwhong@hpcnet.ne.kr

Performance Analysis of Network Devices for High Performance Computing Cluster

Jeong-Woo Hong*, Bo-Sung Lee**, Hyung-Woo Park*,
Sang-San Lee*
*Supercomputing Center, Korea Institute of Science and Technology Informations
**Linux One, Inc.

요약

최근 주목받고 있는 그리드 컴퓨팅 연구등에 주요한 요소로서 기대되어지는 고성능 클러스터 시스템들은 주로 과학 기술 응용연구를 위해 사용되어진다. 이러한 종류의 병렬 시스템은 특정 부품들을 사용하는데 그중 네트워크를 구성하는 부품들이 통상의 분산/병렬컴퓨팅에 주요한 역할요소로서 주목을 받아오고 있다. 이 논문에서는 myrinet, Gbit ethernet, Fast ethernet 장비에 대하여 각각 Netpipe, Linpack, NPB 등의 벤치마크를, 성능 실험을 통해 선정한 Pentium IV 1.7Mhz/1Gb Mem 16노드로 구성한 클러스터에 대하여 2종의 컴파일러를 사용하여 테스트하고 그 결과를 분석하였다. 상이한 성능 차를 보이는 장비간의 성능 비교를 통해 2002년 2월 현재 가능한 응용문제가 사용하고 있는 알고리즘에 따른 최적의 클러스터 시스템의 최적 구성을 도출 할 수 있다.

1. 서론

최근 고성능 컴퓨팅을 필요로 하고 있는 과학 기술 응용분야에서는 PC급 또는 PC서버급 컴퓨터들을 클러스터링 하여 컴퓨터를 구성하여 대용량 계산을 위해 사용하는 경향이 많이 나타나고 있다.

이러한 클러스터 컴퓨터들은 COTS(commodity off the shelf)부품들을 사용하여 제작되어지는 탓에, 시장상황에 따라 등장과 쇠퇴 주기가 몹시 빠른 기술들을 사용하게 된다. 따라서, 한정된 예산을 사용하여 자신의 응용에 적합한 클러스터를 구성하기 위해서는 해결하고자 하는 문제 특성에 적합한 장비들을 선택하여야하고 이를 위해 적절한 성능 분석 결과를 바탕으로 할 필요성이 높다. 본 실험논문에서는 분산 병렬형 계산 모델인 클러스터 시스템에서 가장 영향력을 지니는 주요 구성요소인 3종의 네트워크 장비들과 컴파일러, 2종의 CPU 및 시스템 보드에 대해 Linpack 및 NPB(Nas Paralle Benchmark)를 테스트하고 그 결과들에 대한 분석을 통하여 비교 대상 시스템들의 문제별 효용성을 검토한다.

2. 실험 환경

본 실험을 위하여 사용된 장비들과 그 구성은 다음과 같다. 표 1에는 테스트를 위해 구성한 클러스터 시스템에 사용된 단위 시스템의 사양을 나타내었다.

표 1. 테스트 시스템(단위 시스템 성능 측정)

시스템 사양
Intel1.7 MHz256KB /133MHz /Intel D850MV Board/1GB (256MB * 4 800MHz ECC RDRAM)/RedHat 7.1 Kernel 2.4.2/SSE2

표 2에서는 네트워크 장비에 대해 성능 측정을 위해 사용한 컴파일러와 벤치마크를 정리하였다. 클러스

표 2. Single Processor LINPACK Result

테스트 기계 / 문제 크기	Intel PIII 1GHz/256KB/133MHz Coppermine Dual/VIA Apollo Pro Familly/1GB MEM/RedHat 2.4.2SMP Kernel/SSE1	Intel PIII 1.26GHz/512KB /133Mhz Tualatin Dual/VIA Apollo Pro Familly/1GB MEM/RedHat 2.4.2SMP Kernel/SSE1	Intel PIII 1.26GHz/512KB /133MHz Tualatin Dual/ServerWorks Enterprise ServerSet III HE-SL/2GB MEM/RedHat 2.4.2SMP Kernel/SSE1	Intel PIII 1.13GHz/512KB /133MHz Tualatin Dual/ServerWorks Enterprise ServerSet III HE-SL/2GB MEM/RedHat 2.4.2SMP Kernel/SSE1	Intel PIII 1.26GHz/512KB /133MHz Tualatin Dual/ServerWorks Enterprise ServerSet III HE-SL/2GB MEM/RedHat 2.4.2SMP Kernel/SSE1	IIntel P4 1.4GHz/256KB /133MHz/Intel 850/1GB (256MB * 4 800MHz ECC RDRAM)/Red Hat 7.1 Kernel 2.4.2 /SSE2	AMD Athlon MP1500 (1.33GHz/256KB) /1GB (512MB * 2 PC2100 DDR SDRAM)/RedHat 7.1, Kernel 2.4.2smp/SSE1
1000	333.92	548.41	570.17	559.38	623.73	946.26	710.96
2000	368.81	598.97	638.24	597.56	666.28	1187.01	795.75
3000	404.28	630.08	665.14	624.11	695.88	1328.00	810.35
4000	422.04	645.39	681.74	637.34	710.63	1431.34	902.49
5000	436.35	670.86	707.08	649.71	724.43	1503.00	923.48
6000	442.25	681.62	714.43	658.29	733.99	1555.58	928.82
7000	440.17	694.82	727.66	666.32	742.95	1578.38	953.07

표 3. 클러스터 시스템의 구성 및 벤치마크별 적용 컴파일러

네트워크장비	단위 시스템 구성	테스트 규모	벤치마크와 사용 컴파일러 및 통신 라이브러리	
			Linpack : Scalarpack 2.3.2	NPB 3.2
Fast Ethernet	Intel 1.7 MHz, Intel D850MV Board, RDRAM PC800 1GB/노드	1,2,4,8,16 노드 테스트	gcc mpich 1.2.3	gcc, icc mpich1.2.3
Gigabit Ethernet			gcc, lam 6.5.6	gcc, icc mpich 1.2.3 lam 6.5.6
Mrinet2000			gcc mpich-gm 1.5.1	gcc mpich-gm 1.5.1
Cray T3E 내부 네트워크	Alpha 21164, 128MB/노드		Cray사 제공 컴파일러	Cray사 제공 컴파일러

터 시스템 선정은 표 3에서 정리한 여러성능의 프로세서 및 시스템 보드를 사용하여 Single Processor Linpack등의 성능을 테스트 한 결과를 근거로 한 결정이다. 표3에서는 Intel Pentium IV프로세서 및 동급의 AMD 프로세서 및 기존의 Intel Pentium III 프로세서들에 대해서 벤치마킹한 결과인데, SSE2를 기반으로 한 Pentium IV가 최상의 성능치를 보이고 있다. 이러한 실험치들을 근거로 선택한 Pentium IV 시스템은 주변장치를 위하여 32bit 33Mhz의 PCI Slot들을 제공한다.

본실험에 사용된 PC 클러스터 용 장비의 주된 구성요소들이 PC급 장비임을 고려하여 실험 시점의 시장 상황에서 가격대비 성능을 고려하여 선택한 것이다.

3. 실험

3.1 Netpipe

Netpipe[1]는 A Network Protocol Independent Performance Evaluator로서 point-to-point통신의 경우에 대하여 통신 대역폭 및 latency측정을 위하여 사용되어지는 벤치마크이다. 이를 사용하여 FastEthernet, Gigabit, Myrinet 3종의 네트워크에 대해 point-to-point 대역폭 및 latency를 측정하였다.

그림 1, 2, 3에서 볼 수 있듯이 latency와 대역폭(12~13usec/640Mb) 면에서 myrinet이 가장 우수한 성능을 보이며, Gbit Ethernet(45~61usec/410Mb), Fast Ethernet(75~80usec/89Mb)를 보인다.

그림 1. Netpipe Results (Myrinet2000)

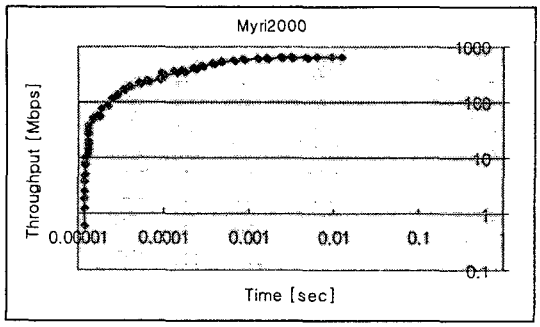


그림 2. Netpipe Results (FastEthernet: Intel 410T)

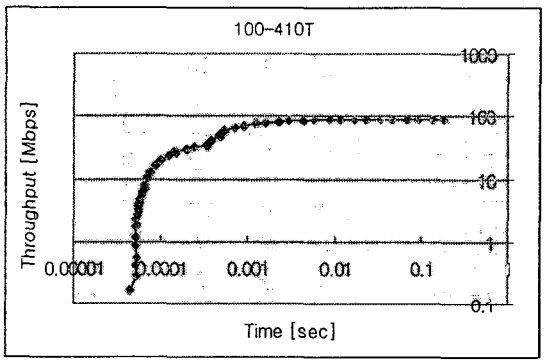
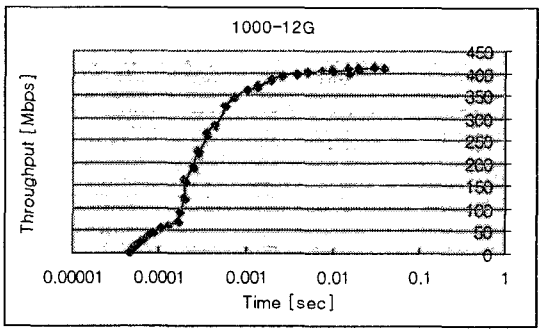


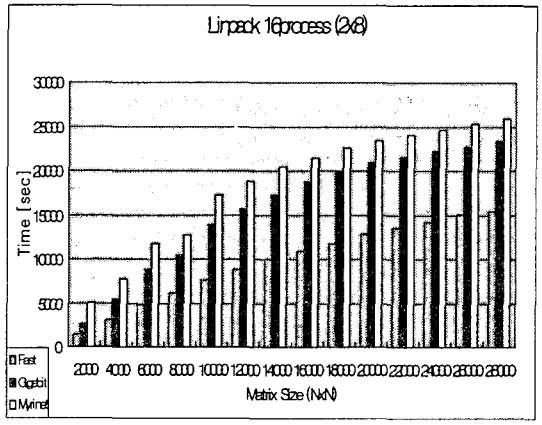
그림 3. Netpipe Results (Gbit Ethernet : Asante)



3.2 Linpack

그림 ?에서 볼 수 있듯이 계산에 사용되어지는 노드 수가 늘어감에 따라 대역폭의 한계에 따른 포화상황을 기대할 수 있다. 벤치마킹의 기준으로 사용되어지는 LU 계산의 통신 특성에 따라 누그러진 포화곡선을 그린다. myrinet 장비와 gigabit의 특성을 비교하면 우선 그래프 좌측의 모습에서 latency가 좋은 myrinet이 작은 문제 크기에 따른 잦은 통신을 상대적으로 잘 지원해 주고 있다는 것을 볼 수 있다. 그러나, 문제의 크기가 커짐에 따라 gigabit이 myrinet의 성능에 80~90%선에 육박하는 계산특성을 보여 주고 있다. 이는 일반 LU형태의 계산을 활용하는 응용프로그램에서 통신패킷의 크기를 크게 하는 방향으로 프로그램 함으로써 최상의 성능을 도출할 수 있을 가능성을 보여주는 근거가 될 수 있을 것으로 판단된다.

그림 4. Linpack 성능 그래프



3.3 NPB

NPB[3]는 NAS parallel benchmark 프로그램으로서 일반적인 프로그래머들이 사용하는 알고리즘 및 숙련도로 제작한 실제 응용 문제를 벤치마킹 문제로 제작한 것이다. 실험에 사용된 코드는 버전 3.2이다.

그림 5. NPB CG class B

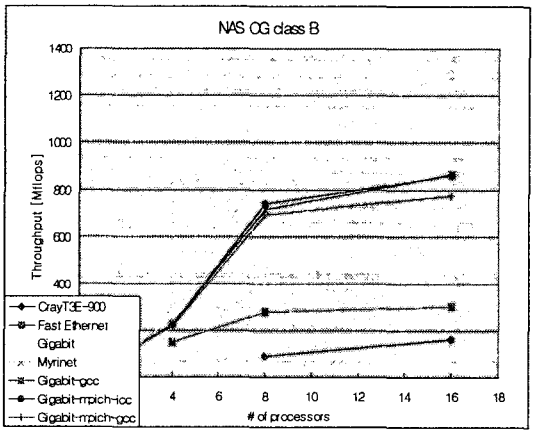


그림 5, 6, 7, 8는 각각 NAS 벤치마크중 CG (conjuagated gradiant problem), SP(scalar pentadiagonal solver), BT(block tridiagonal solver), LU(LU solver)의 벤치마킹 결과 그래프이다. 여기서 icc 라고 표현하지 않은 곡선의 경우 Pentium IV용 intel 컴파일러가 아닌 gcc를 사용한 경우이다.

그림 5의 CG 문제의 경우 작은 패킷이 다수 발생하는 문제로서 myrinet이 가장 우수하다. 그림 6의 SP

문제는 그림 7의 BT문제에 비하여 그 통신 패턴은 비슷하나 통신량이 25배 이상의 많은 코드이다[4]. 이에 따라 비슷한 형태의 곡선을 그리나 좀 더 좋은 성능을 보이고 있다.

그림 6. NAS SP class B

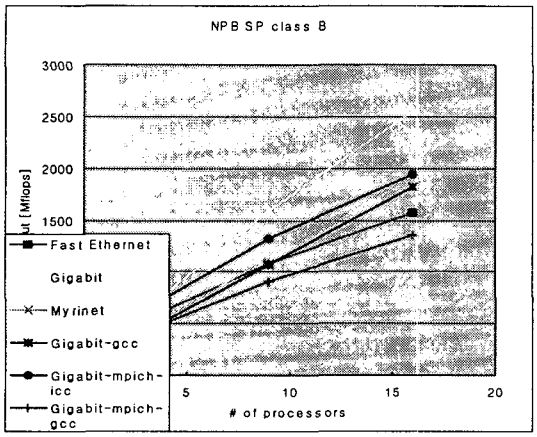


그림 8의 LU 문제의 경우 앞서 Linpack 벤치마크 결과와는 다르게 Gbit ethernet이 myrinet에 비하여 더 우수한 결과를 보여주고 있다. 이는 사용된 LU 코드가 최적화 과정을 거친 코드가 아니라 통상의 프로그래머들이 코딩하였음을 가정한 코드를 사용하였음에 근거 하는 것으로 판단된다.

5. 맺음말

벤치 마크 결과들을 통해 볼때, 통상의 자체 제작 응용코드의 수준으로 평가되는 NPB등에 사용된 코드들이 메쉬 구조에 적합한 형태로 작성이 되어있으며 SP와 BT같은 경우는 Hypercube구조에 적합하도록 되어있으나, 단순한 스위치 구조를 가진 클러스터에서도 Myrinet, Gbit ethernet등의 장비를 채택할 경우 적정규모의 시스템에서 비교적 좋은 결과를 얻을 수 있음을 볼 수 있다. 또한, 지속적인 가격 인하가 기대되는 Gbit의 경우 myrinet에 비하여 절반 이하의 작은 비용임에도 통상의 응용 코드에 대해서 훌륭한 가격대비 성능을 기대할 수 있음을 살펴 볼 수 있다.

또한, 클러스터 상에서는 MPI 프로그램의 수행 시 Processor들 사이에 통신 횟수를 줄이고 통신량을 늘이도록 알고리즘을 고안하면 현재 클러스터용 네트워크 장비들로도 보다 효율적인 시스템 구성이 가능할 것으로 기대됨을 볼 수 있다.

이상의 결론은 2002년 2월 이 실험 논문이 작성된 시점에서의 테스트 결과에 대한 요약이며 향후 새로운 시스템과 프로세서 등에 대해서는 지속적인 실험을 수행할 예정이다.

그림 7. NAS BT class B

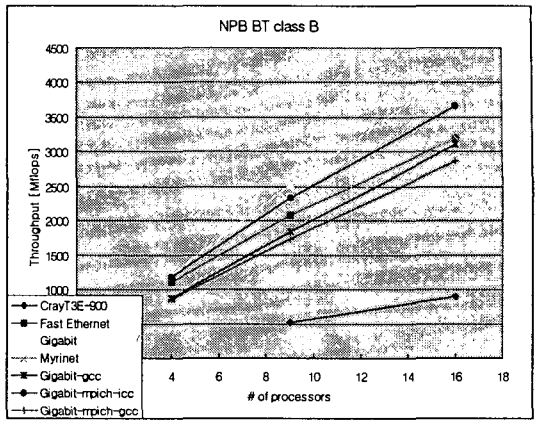
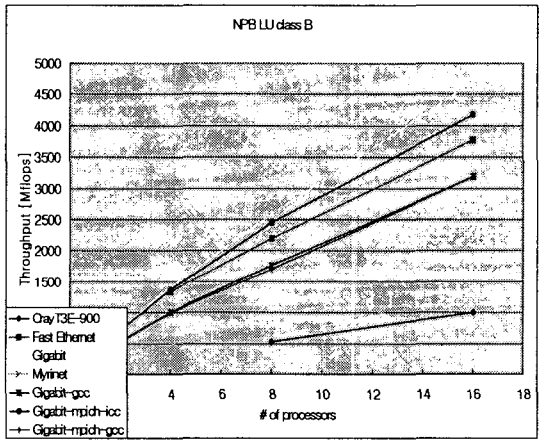


그림 8. NPB LU class B



참고문헌

[1] Quinn O. Snell, Armin R. Mikler and John L. Gustafson, "NetPIPE: A Network Protocol Independent Performance Evaluator", http://www.scl.ameslab.gov/netpipe/paper/full.html,

[2] Jack Dongarra, Jim Bunch, Cleve Moler and Pete Stewart, "LINPACK", http://www.netlib.org/linpack/

[3] Bailey, D. H, et al., " The NAS Parallel Benchmarks", NASA Technical Memorandum 103863, NASA Research Center, Moffett Field, CA, 94035-1000, July 1993, (http://www.nas.nasa.gov/NAS/NPB)

[4] 권오영, "고성능 컴퓨터 성능 측정 도구 NPB의 소스코드 구성 및 병렬화 방법 분석", 슈퍼컴퓨팅소식 Vol. 4, 2001.3. ISSN 1339-7838, 한국과학기술정보연구원, 슈퍼컴퓨팅센터