초고속 네트워크를 이용한 PC 클러스터의 구현과 성능 평가

(Implementation and Performance Analysis of PC Clusters using Fast PCs & High Speed Network)

김 영 태 † 이 용 희 †† 최 준 태 †† 오 재 호 †††

(Youngtae Kim) (Yonh Hee Lee) (Jun-Tae Choi) (Jai-Ho Oh)

요 약 본 연구에서는 고속의 개인용 컴퓨터와 초고속 통신장비를 이용한 분산/병렬 계산을 위한 클러스터를 구현하였다. 먼저 16 PC의 1세대의 클러스터를 구현하여 현재 현업(제주 지방기상청)에서 실시간으로 운영중이며, 1세대 클러스터의 성능 분석을 통하여 효율적으로 성능이 개선된 2 CPU의 16 PC로 구성된 2세대 클러스터를 구현하였다. 본 연구에서는 다른 속도의 CPU 및 통신 장비로 구성된 두 가지 형태의 클러스터를 실제로 기상 예보를 위하여 사용되는 병렬모델인 MM5를 이용하여 계산시간 및 통신에 대한 다양한 분석을 하였다.

키워드 : 분산 병렬 처리, 네트워크, 성능분석

Abstract We implemented two fast PC clusters using fast PCs and high speed network. First, we built the first generation of 16 PC cluster and have used it for real-time operation at Cheju Regional Meteorological Office. Next, we built the second generation of 16 PC with dual CPUs cluster which was efficiently improved based on performance analysis of the first generation of cluster. In this research, we also analyzed performance of two different clusters, which have different CPUs and communication devices, using the parallel model MM5 which has been used for the real-time weather forecasting.

Key words : Distributed and Paralled Processing, Network, Performance Analysis

1. 서 론

오늘날 산업 기술의 획기적인 발달과 더불어 전반적인 분야에서 다양한 정보의 필요성과 함께 초고속으로 원하는 시간 내에 정보를 제공할 수 있는 수요가 요구되고 있다. 이러한 정보의 처리를 위하여 초고속 계산을 하는 컴퓨터가 어느 때보다 필요하게 되었으며, 지금까지는 상업용 슈퍼컴퓨터가 그 역할을 해 주었다. 또한 다양한 형태의 슈퍼컴퓨터들 중 한 대의 컴퓨터의 성능의 발달에는 많은 비용과 노력이 필요하며 또한 속도의 제한이 있기 때문에, 컴퓨터내의 노드의 수만 증가하면 성능이 향상되는 분산형 병렬 컴퓨터가 많이 사용되고 있는 추세이다. 그러나 뛰어난 성능에도 불구하고 비싼 가격 때문에 병렬 프로그램의 연구와 개발 및 실시간 운영에 어려움이 있었다. 국내에서는 분산형 병렬 컴퓨터로서는 지금까지 국내에서는 일반사용자들은 주로 한국과학기술정보연구원(KISTI)에서 제공되는 Cray T3E 병렬 컴퓨터를 사용해 왔으며, 제한적인 자원으로 인하여 일반 사용자들은 병렬 프로그램의 연구와 개발에 많은 제한이 있어 왔다.

1940년대에 세계 최초로 범용 컴퓨터가 개발된 이후, 오늘날에 이르러서는 컴퓨터 계산 성능의 급진적인 발달로 인하여 개인용 컴퓨터의 성능이 10여년 전의 슈퍼컴퓨터의 성능을 초과하는 계산 처리 능력을 갖추게 되었으

· 이 논문은 기상청에서 시행한 기상지진기술개발사업의 지원으로 수행되었습니다.

† 정 회 원 : 강릉대학교 컴퓨터과학과 교수
ykim@kangnung.ac.kr
†† 비 회 원 : 기상연구소 수치예보실 기상연구사
yhlee@metri.re.kr
cjt@metri.re.kr
††† 비 회 원 : 부경대학교 환경대기과학과 교수
jhoh@pknu.ac.kr

논문접수 : 2000년 8월 24일
심사완료 : 2001년 10월 10일

며, 아울러 컴퓨터 통신 기술의 발달로 네트워크를 통하여 초고속으로 자료 교환이 가능하게 되었다. 따라서 고속 네트워크를 이용한 '개인용 컴퓨터의 클러스터'를 통한 분산처리 시스템이 발달하게 되었으며, 이러한 클러스터를 이용한 분산처리 시스템의 운영은 저가의 하드웨어를 이용한 초고속 환경을 제공할 수 있으므로 병렬 프로그램의 연구 및 운영에 있어서 최적의 환경을 제공하게 되었다. 초고속 병렬 클러스터의 구축은 일반 컴퓨터를 연결하는 고속 네트워크와 범용의 프로세서들을 활용하는 기술로서 기존의 네트워크로 연결된 개인용 컴퓨터 및 워크스테이션들을 이용하여 가격대비의 뛰어난 성능과 더불어 경제성을 가진 병렬처리 환경을 구축하고, 나아가 병렬 프로그램의 개발 및 운영 환경을 제공할 수 있다.

본 연구에서는 중규모 기상 재해 예측시스템을 위하여 2000년에 1세대 클러스터를 개발하여 제주 지방기상청에서 실시간으로 운영중이며, 2001년에는 1세대 클러스터의 성능 분석을 이용한 다양한 개선을 통하여 2 CPU를 이용한 2세대 클러스터를 구축하였다. 본 논문에서는 두 가지 다른 형태의 클러스터에서 다른 장비를 사용하는 통신 등을 실제의 운영중인 병렬 모델을 이용하여 분석하였으며, 이에 따른 최적의 성능을 가진 클러스터의 형태를 제시한다. 본 논문은 2장에서는 국내외의 클러스터 연구동향, 3장에서는 두 가지 다른 형태의 클러스터 구현, 4장에서는 성능 분석, 그리고 5장의 결론으로 구성된다.

2. 클러스터의 국내외 연구 동향

최근까지 국내에서는 분산 처리를 이용한 병렬화에 대한 연구는 초보 단계였으며 이제 새로이 도약하는 단계에 와 있다. 1998년부터 삼성 종합연구소 및 한국과학기술정보연구원(KISTI) 등의 연구기관에서 알파칩을 내장한 워크스테이션의 클러스터와 시스템 소프트웨어에 관한 연구를 진행중이다. 또한 각 대학교에서도 클러스터를 이용한 병렬처리에 관한 연구가 활발하게 진행중이며, 이미 컴퓨터 업체에서도 다양한 형태의 클러스터를 출시하고 있다. 하지만 국내에서는 아직까지 세계적인 수준과 비교하여 괄목할만한 연구 성과가 발표되지 않고 있고, 안정성 내지는 가장 핵심적인 부분인 통신 시간 등에 관한 분석이 이루어지지 않고 있기 때문에 국내의 기술 동향을 판단하기에는 아직까지는 어렵다고 여겨진다.

국외에서는 미국을 중심으로 연구소 및 대학교에서 초고속 네트워크와 고속 개인용 컴퓨터(워크스테이션 포함)를 이용한 클러스터의 병렬 컴퓨터 구현을 통하여 분산/병렬 컴퓨팅 환경을 제공하기 위한 연구가 활발하게 진행중이다. 미국 Berkeley소재의 캘리포니아 주립대학의 NOW(Network of Workstations) 프로젝트는 분산 병렬 시스템의 대표적인 연구로서 SUN 워크스테이션들과 인텔 펜티엄 PC들을 연결하여 병렬처리 환경을 구현하였다. 이 프로젝트는 보다 많은 수의 워크스테이션들을 이용하여 보다 안정된 분산 시스템을 구현하기 위한 연구를 진행중이다(http://now.cs.berkeley.edu). NASA 연구소와 Maryland 주립대학교에서는 공동으로 지구과학 프로젝트를 대상으로 하는 BEOWULF 프로젝트를 수행중인데, 여기서는 순수한 분산 병렬 환경을 제공하기보다는 프로세서들간에 작업을 공유함으로써 빠른 작업을 수행하고 있다. PC용 유닉스 운영체제인 리눅스를 이용한 네트워크 클러스터의 한 형태인 BEOWULF 시스템은 1994년 NASA의 CESDIS에서 16노드 클러스터를 리눅스 표준 소프트웨어 패키지를 이용하여 개발하였다. 당시 NASA에서 사용중인 Cray의 임대 기간 종료와 더불어 새로운 병렬처리용 슈퍼컴퓨터를 직접 개발하게 되었으며, 리눅스 네트워크 클러스터 드라이버의 개발이 CESDIS에 의해 주도적으로 이루어져 현재 널리 사용되고 있다(http://www.beowulf.org). Colorado주에 있는 FSL(Forecast Systems Laboratory)에서는 2000년 260대의 Alpha DS-10 CPU와 Myrinet을 이용한 병렬처리 환경의 구현에 성공하였으며, 사용자를 위한 시스템 소프트웨어 및 일반 프로그램을 개발 중에 있다(http://www.fsl.noaa.gov). 상업용으로는 미국의 Compaq에서 AlphaServer SC Interconnect를 개발하여 미국, 일본 및 독일 등에 제공하고 있다(http://www.compaq.com/hpc/systems/sys_sc.html).

전 세계적으로 보았을 때 클러스터의 연구는 대부분 표준화된 라이브러리인 MPI(Message Passing Interface)[1]를 사용하기 때문에 클러스터의 기술적인 접근은 거의 일치하지만 아직까지 분산 병렬처리의 프로그램 사용에 대한 뚜렷한 연구 결과가 나오지 않은 상태이며, 이는 분산 시스템의 구현에 앞서 클러스터를 이용한 병렬 프로그램의 개발에 어려움이 있기 때문이라고 분석된다.

3. 클러스터의 구성

본 연구에서는 먼저 범용 네트워크 장비를 이용한 1세대 클러스터를 구축하였으며, 여기에서 나타난 문제점을 개선 발전시켜 2세대 클러스터를 구축하고 각종 수치 모델의 적용 실험을 실시하였다. 이 2가지 다른 형태의 클러스터는 각각 다른 형태의 CPU의 구조 및 통신의 형태를 가지고 있으며, 안정성을 보유할 수 있도록 설계되었다.

3.1 1세대 클러스터의 구성

1세대 클러스터는 일반적으로 사용되는 범용 네트워크 장비를 사용하여 클러스터를 구성하고자 하였다. 1세대 클러스터는 Beowulf 계열의 네트워크 클러스터로서 16개의 노드와 노드간의 통신을 위해서 100Mbps switch를 사용하였으며, 호스트 노드에는 2개의 NIC (Network Interface Card)를 설치하여 eth0는 외부 네트워크에서 접근이 가능하도록 하고, eth1은 클러스터 노드들간의 네트워크를 구성하였으며, 호스트 노드에서만 접근이 가능하도록 하였다(그림 1 참조). 호스트 노드는 하드디스크가 없는 일반 노드를 위해 NFS-root 서버가 되게 하였고, 각 노드들을 위한 /tftpboot 디렉토리를 만들어 설정파일을 제외한 파일은 공유하도록 하였다. 각 노드는 RARP를 이용하여 자신의 IP를 서버로부터 제공받아 diskless-booting이 되도록 하였다.

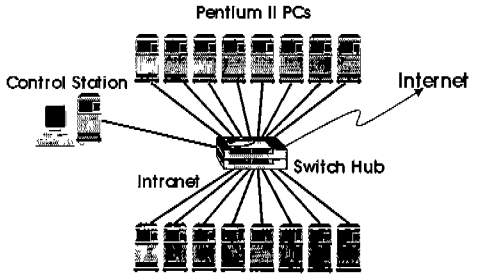


그림 1 Structure of the 1st generation cluster

운영체제는 PC용 유닉스 운영체제인 Linux Redhat 6.1을 사용하였으며, 메시지 전송 라이브러리로 MPICH, LAM MPI, PVM[2]을 설치하여 각 노드간 메시지 전송을 통해 병렬처리를 가능하게 하였으며 또한 기타 병렬처리용 라이브러리도 설치하였다(표 1 참조).

표 1 Hardware and software for the 1st generation cluster

Hardware		Sofeware	
CPU	Pentium II 400 Mhz×16	Kernel	Linux 2.2.13-tcpfix (Redhat 6.1)
internal cache	512 Kb	Compiler	Portland Group pgf77, pgf90, pgcc
RAM	128 Mb		
Network	10/100 Mbps	Message Passing	MPICH, LAM MPI, PVM
Switch	Intel 460T 10/100 Mbps Switch	Math Library	BLAS, LAPACK, FFT

현재 1세대 클러스터는 제주 지방기상청에서 운영중인 '한라단시간 예측시스템'의 운영에 직접 활용되고 있으며, 그림 2는 제주 지방기상청에 설치된 모습을 보여주고 있다.

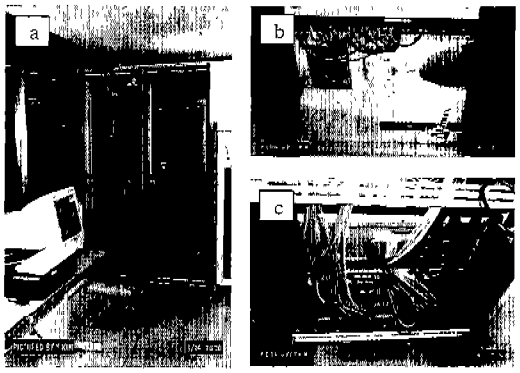


그림 2 The 1st generation cluster in Cheju Regional Meteorological Office : a) Front view of the cluster, b) switch and c) rear side view

3.2 2세대 클러스터의 구성

1세대 클러스터의 사용 결과, 범용장비 사용에 따른 문제점을 해결하고 클러스터의 노드의 확장성을 고려하여 2세대 클러스터의 구축과 보다 진보된 클러스터 관리 기법의 개발이 필요하였다. 먼저 1세대 클러스터의 성능의 분석을 통하여 노드간의 통신이 전체 성능에 큰 영향을 주게 되는 것을 알게 되었다 (4장 클러스터의 성능 분석 참조). 이는 슈퍼 컴퓨터급의 고성능 컴퓨터에서의 통신의 최적화(계산 대 통신 비)에 비교하여 버스 시스템을 사용하는 현재의 네트워크 장비들은 이런 것을 고려하지 않고 있기 때문이다(http://www. beowulf.org). 따라서 2세대 클러스터에서 사용한 네트워크 장비는 Myrinet SAN(System Area Network)으로, Myrinet은 일반적인 네트워크 보다 초고속 병렬 슈퍼컴퓨터에서 볼 수 있는 내부 접속 네트웍에 가까운 성질들을 많이 가지고 있어 계산 대 통신비의 획득에 유리하다. 또한 다양한 통신형태를 지원(any topology acceptable)하며 진보된 통신방법(wormhole routing, switch concept 등)을 가지고 있으며 통신 형태와 방법을 사용자가 선택할 수 있도록 유연성을 부여하고 있어 병렬 모델의 특성에 적합한 통신형태와 통신방법을 적용할 수 있다.

2세대 클러스터의 특성을 Table 2에 요약하였다. Linux 커널은 1세대 클러스터와 동일하나 각종 Linux

배포본을 사용하여 클러스터를 구성해본 결과 중규모 기상모형에는 Mandrake가 가장 우수한 것으로 나타났다. 일반적으로 많이 사용하는 Redhat 6.1과 Mandrake 6.1을 비교하여 약 15%의 속도 향상되는 것으로 나타났다. 메시지 전송 라이브러리의 경우에는 MPICH가 7층 구조를 가지고 있어 다소 복잡한 반면, Myrinet에 최적화된 'MPICH over GM'은 단일층 구조로 단순화되어 있어 보다 우수한 성능을 보이고 있었다. 또한 통신부하를 줄이기 위해서 NFS(Network File System)을 범용 네트워크인 Ethernet으로 분리하고 Myrinet을 계산 전용으로 활용할 수 있도록 구성하였다. 이러한 구성은 Myrinet에 이상이 발생할 경우 Ethernet으로 전환할 수 있게 되어 클러스터 네트워크의 안정성을 확보하는데 크게 기여하게 되었다.

표 2 Characteristics of the 2nd generation cluster

Hardware		Sofeware	
CPU	Pentium III 700 Mhz×32	Kernel	Linux 2.2.13-tcpfix (Mandrake 6.1)
Internal Cache	256 Kb	Compiler	Portland Group pgf77, pgf90, pgcc
RAM	256 Mb		
Network	SAN interface, Ehternet (100 Mpbs)	Message Passing	MPICH, MPICH over GM
Switch	Myrinet SAN Switch, Ehternet Switch	Math Library	BLAS, LAPACK, FFT

3.3 웹 기반의 클러스터 모니터링 기법 개발

웹 기반 클러스터 모니터는 클러스트 내의 각 프로세서들의 운영 형태를 한 곳에 모아서 쉽게 모니터 하기 위하여 웹 상에서 그래픽을 이용하여 다음의 정보를 사용자에게 실시간으로 보여 주는 프로그램이다. 현재 이용 가능한 정보는 swap space 및 free swap space, main memory 및 free memory space, number of users, CPU load, 누적된 CPU load 등이다(그림 3 참조).

Linux 운영체제에서는 시스템에 관한 정보가 /proc 디렉토리에 실시간으로 기록된다. 우선 클러스터의 각 노드(프로세서)는 데몬(deamon)의 형태로 시스템 데이타를 읽어서 전달해 주는 서버 프로그램을 수행한다. 웹 서버는 cgi-bin 프로그램을 이용하여 일정한 주기로(현재 5초 간격) 각 노드에 시스템 정보를 요청하는 클라이언트로서 정보를 요청한다. 이 때 각 노드의 데몬 프로그램은 실시간으로 요청한 정보를 클라이언트(웹서버)에게 전달하며, 웹서버는 필요한 시스템 정보를 전달받은 후 그 정보를 가공하여 화면에 나타낸다. 따라서 접근할 수 있는 네트워크만 존재한다면 언제 어디서나 클러스터의 운영 상황을 모니터할 수 있게 된다.

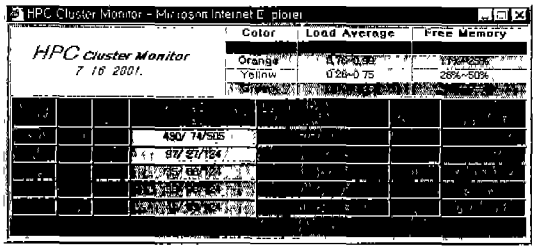


그림 3 Cluster usage statistics on the web

4. 클러스터의 성능 분석

이 장에서는 기상 모델인 병렬 MM5의 실행을 통하여 클러스터의 성능을 분석한다.

4.1 MM5(Mesoscale Model 5)

병렬 MM5는 약 40,000줄 정도의 메시지 전송을 위하여 MPI를 사용하는 Fortran 코드로서 FDM(Finite Difference Method)를 계산의 기반으로 하는 전세계적으로 가장 많은 사용자를 보유한 대표적인 기상 모델이다[3,4,5,6]. 병렬 MM5는 FDM외에도 Matrix Inverse, Gaussian Elimination 등의 다양한 수치 알고리즘을 포함하고 있으며, 또한 통신 시간을 분리할 수 있도록 설계되어 있으며 MPI를 호출하기 위하여 C 언어로 쓰여진 인터페이스를 사용하기 때문에 장시간의 실행시간을 이용한 성능 분석을 위한 최적의 코드로 사료된다. 실행시간의 비교를 위해서는 1 time-step당 135초의 24시간의 시뮬레이션을 하였으며 기타의 분석을 위해서는 3시간의 시뮬레이션을 한 wall-clock시간으로 분석 및 비교를 하였다.

4.2 실행시간의 비교

실행시간의 비교는 다른 컴퓨터에서의 성능 비교를 가장 쉽게 보여 준다. 그림 4는 두 클러스터를 이용한 병렬 MM5와 IBM SP-1[1], Cray T3E[2]등 상업용 병렬 컴퓨터와의 성능을 비교한다. 두 클러스터는 다른 상업용 병렬 컴퓨터들보다 훨씬 빠른 결과를 보여 준다.

1) IBM SP-1은 미국 Argonne국립연구소에서 운영하였던 병렬 컴퓨터로서 이 실행 결과는 1995년에 실행하였음.
2) Cray T3E는 한국과학기술정보연구원에서 운영중인 병렬 컴퓨터임.

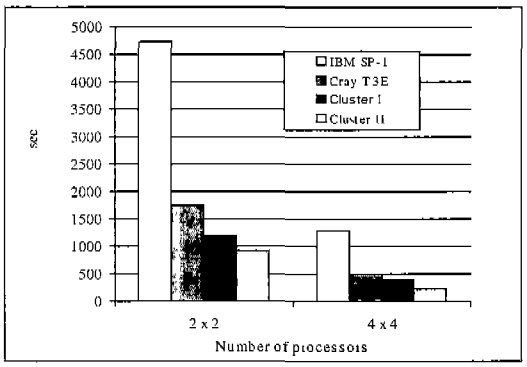


그림 4 Runtime results from IBM SP-1, Cray T3E and Clusters

4.3 부하균형 (Load balance)

부하균형은 데이타의 값들에 따라서 물리과정의 계산을 많이 하는 프로세서와 그렇지 않은 프로세서들간의 통신의 동기점을 찾기 위해 발생한다. 일반적으로 1 time-step당 여러 개의 통신 동기점들이 발생하는데 병렬 MM5의 경우에는 3개의 통신 동기점들이 1 time-step당 발생한다[7,8]. 다음 예는 부하균형을 어떻게 계산하는 지를 보여 준다.

그림 5에서는 4개의 프로세서가 1 time-step당 3개의 통신 동기점을 가진다. P_0는 첫 번째의 동기점에서 계산을 가장 많이 하였으며 P_0와의 통신을 위하여 나머지 프로세서들은 기다리게 된다(idle time). 두 번째 동기점에서는 P_3이 가장 계산을 많이 하였으며 세 번째 동기점에서는 P_1이 계산을 가장 많이 한다. 이를 전체에 대한 비례를 구하면 하면 (1)식과 같이 계산을 할 수 있다.

$$Load\ balance = \frac{T_{mean}}{T_{\max}} = \frac{\sum t_{ij}/\#proc}{\sum T_{\max}\text{'s in each sync. points}} = \frac{\sum t_{ij}/4}{t_{00}+t_{13}+t_{21}} \quad (1)$$

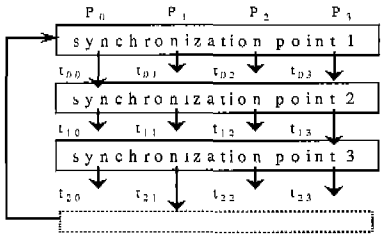


그림 5 An example of load balancing analysis

그림 6에서는 두 클러스터와 다른 두 병렬 컴퓨터와의 부하균형을 비교한다. 전체 계산에서는 다들 비슷한 결과를 보여 주는데 이는 클러스터에서의 계산부분의 안정성을 의미한다.

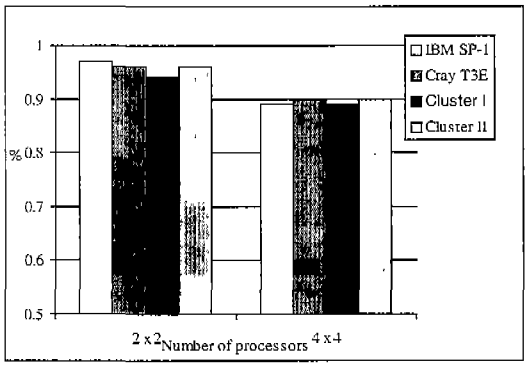


그림 6 Load balance of 2×2 and 4×4 case for IBM SP-1, Cray T3e and the clusters

4.4 통신 시간의 비교

분산 병렬 처리에서는 1개의 CPU를 사용하는 경우와는 달리 프로세서들간에 서로 자료를 주고 받는 통신 부하가 발생한다. 따라서 병렬 프로그램에서는 이러한 통신 부하를 줄이는 것이 프로그램의 성능에 가장 큰 영향을 미치게 된다. 또한 병렬 컴퓨터에서도 안정되고 빠른 통신을 제공하는 것이 가장 중요하다. 여기에서는 병렬 MM5의 실행시 통신에 소요된 시간만을 따로 분석하여 클러스터에서의 네크워크의 성능을 분석한다.

일반적으로 분산 병렬 프로그램은 통신과 계산을 반복하는데 여기서는 통신부분만을 따로 측정하였다. 병렬 컴퓨터들은 통신을 위한 Buffering을 하는 동안에도 계산을 수행하여 통신과 계산의 중복이 발생하므로 통신만을 측정하기 위하여 통신이 끝난 후 각 프로세서들이 같이 계산을 시작할 수 있도록 강제적으로 동기점을 프로그램에 삽입하였다(이 경우 프로그램의 결과에는 영향이 없음). 그림 7에서 통신에 경과되는 시간을 클러스터와 Cray T3E와 2×2 및 4×4 프로세서들을 이용하여 비교를 하였다. 상업용 컴퓨터인 Cray T3E에서는 통신 시간이 비교적 고르게 진행되는 반면 범용 Switch인 인텔 510T를 사용하는 클러스터에서는 통신 시간의 변화가 많기 때문에 표준 편차와 분산을 그림에 나타내었다. 그림에서 (1)과 (2)는 각각 동기점 사이의 통신 시간을 나타내며 전용의 통신 장비를 사용하는 Cray

T3E에 비교하여 범용 Hub를 사용하는 클러스터에서는 많은 시간의 편차가 발생하는 것을 보여 준다. 이 편차를 줄이기 위해서 1세대 클러스터에서 Hub 장비를 인텔 460T로 교체하였으며, 2세대에서는 Myrinet을 사용한 결과 편차가 거의 없는 것으로 나타났다.

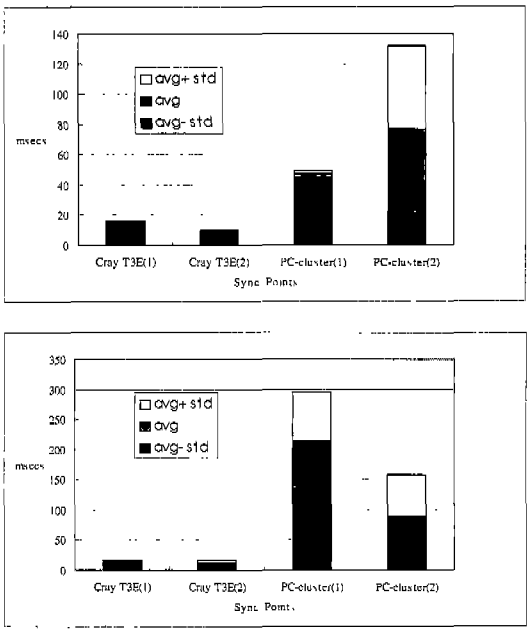


그림 7 Communication variance of 2×2 (upper) and 4×4 (lower) for Cray T3E and the 1st generation cluster with Intel 510T

4.5 2세대 클러스터의 확장성(scalability)

2세대 클러스터에서는 성능분석을 위하여 100Mbps의 Ethernet과 Myrinet을 동시에 사용하기 때문에 각각의 의 성능을 비교와 더불어 'MPICH over GM'에서 지원하는 directcopy의 성능을 알아보기 위하여 각 노드내의 2개의 CPU를 가지고 있는 공유메모리 방식(SMP, Shared Memory Process)도 함께 비교하였다.

그림 8은 Ethernet 클러스터와 Myrinet 클러스터의 분산 메모리 방식(Distributed Memory Processor, DMP)과 공유메모리방식에 대한 성능평가 결과를 수행시간과 확장성(scalability)로 나누어 각각 나타내었다. 여기서 32개의 CPU에 대한 결과는 SMP의 결과를 나타낸다. 그림 8a에서 16개의 CPU를 사용하였을 경우 Ethernet은 1.92시간, Myrinet SMP는 1.7시간, Myrinet DMP의 경우에는 1.43시간이 각각 걸리는 것으로 나타났다. 4개의 CPU를 사용할 경우에는 Myrinet SMP가 시간이 가장 많이 걸렸으나 8개 CPU이후부터는 Myrinet DMP, Myrinet SMP, Ethernet 순으로 실행시간이 적게 걸리는 것으로 나타났다. Fig. 8b에는 각각의 클러스터 구성에 대해 CPU 4개를 사용하였을 경우를 기준으로 하여 상대적인 확장성을 나타내었다. 16개의 CPU를 사용한 경우 Ethernet의 확장성은 2.14이며, Myrinet DMP의 확장성은 2.54로 Myrinet이 우수하게 나타났다. 특히 주목할 만한 것은 Myrinet SMP의 경우 32개의 CPU를 사용할 경우 확장성이 4.15로 16개를 사용할 때와 비교하여 확장성이 현저하게 증가함을 알 수 있다.

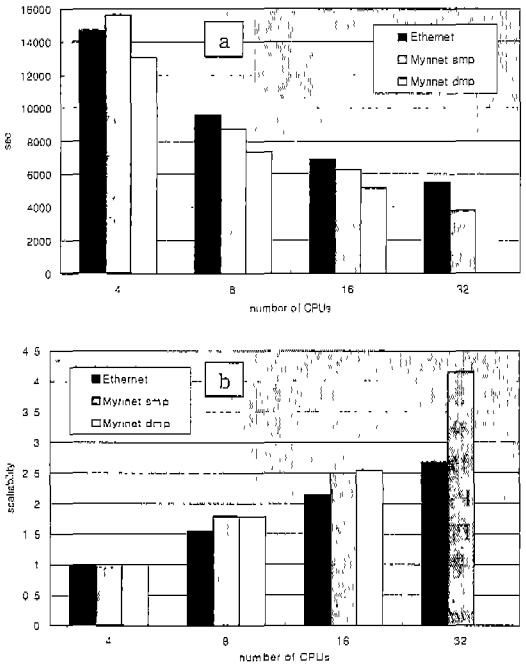


그림 8 Benchmark of a) Wall clock time and b) scalability using Ethernet Myrinet DMP Cluster and Myrinet SMP Cluster (dual processors on each node) with MPICH and 'MPICH over GM'.

표 3에는 각각의 CPU 수에 대한 Ethernet에 대한 상대적인 실행시간비를 나타내었다. 16개의 CPU를 사용할 경우 Myrinet DMP가 Ethernet의 경우에 비해 34% 정도 실행속도의 향상이 있었음을 보여주었다. Myrinet SMP의 경우에는 16개의 CPU를 사용할 경우에는 10%의 증가를 예상하였으나 32개의 CPU를 사용

할 경우 약 46%로 현저하게 증가함을 보인다. 이는 32개의 CPU를 사용할 경우 SMP의 특성상 dual CPU상에서 공유하는 메모리량이 반으로 줄어듬에 따라 'MPICH over GM'에서 사용한 directcopy에 의한 SMP의 효율이 높아졌음을 의미한다.

표 3 Wall-clock ratio of Myrinet cluster(SMP, DMP) versus Ethernet of each number of CPUs

#CPUs	Ethernet	Myrinet SMP	Myrinet DMP
4	1	0.94	1.13
8	1	1.09	1.29
16	1	1.10	1.34
32	1	1.46	

그림 9에는 Myrinet 16 port switch를 사용할 경우 Myrinet SMP의 효율을 알아보기 위하여 Myrinet DMP의 경우를 32 ports를 사용하였을 경우의 예측결과를 나타내었다. 가장 좋게 예측한 경우는 8개의 CPU에서 16개의 CPU로 증가할 때의 확장성을 선형적으로 예측하였을 경우이며, 나쁘게 예측한 경우는 4개의 CPU에서 16개의 CPU로 증가했을 때의 확장성의 감쇄를 고려한 경우를 나타낸다. 그 결과에서 Myrinet SMP의 경우의 Myrinet DMP를 32개의 노드로 확장하였을 경우 예측된 이상적인 값에 매우 근접하여 Myrinet SAN의 대역폭을 충분히 활용하고 있음을 알 수 있다. 이 결과는 Myrinet을 사용한 SMP는 중규모 예측모형의 적분영역의 크기를 적절히 조절하여 공유메모리 영역에서의 부하를 충분히 감안한다면 Myrinet DMP를 사용할 경우 노드 확장으로 인한 경비를 최소화 할 수 있음을 의미한다[9]. 이 연구 결과는 향 후 예측의 목적과 예산에 맞는 적절한 규모의 클러스터 구성에 좋은 지침으로 사용할 수 있을 것이다.

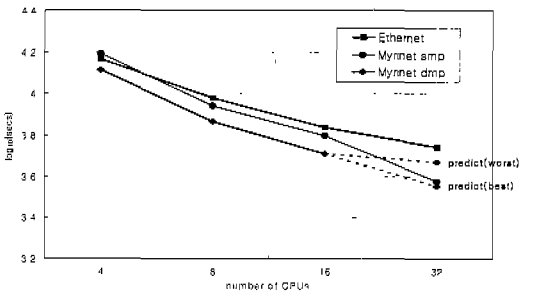


그림 9 Same as Fig. 8a except wall clock time prediction of Myrinet DMP using 32 CPUs.

5. 결 론

본 연구에서는 먼저 PC용 UNIX 운영체제인 Linux를 운영체제로 하는 16대의 PC를 이용하여 1세대 클러스터를 개발하였다. 이 후 1세대의 성능을 대폭 개선한 2세대 클러스터를 구축하여 훨씬 향상된 성능을 보여주었다. 두 클러스터의 성능을 KISTI 보유의 분산형 슈퍼컴퓨터인 Cray T3E와 비교함으로써 클러스터의 계산능력을 보여 주었다.

본 연구에서는 클러스터를 현업에서 초고속 병렬 기상 모델의 운영에 실시간으로 활용함으로써 클러스터가 슈퍼컴퓨터의 대안으로 가격대비 성능의 경제성을 가진 컴퓨터로 운영됨을 보여주었다. 그리고 중규모 기상재해 예측시스템의 개발을 위해 클러스터 기술과 병렬화 기술을 접목하여 초고속 기상재해 예측기술의 개발에 성공함으로써 기상재해예측에 있어서의 새로운 패러다임을 제시하였으며, 클러스터의 구성 및 운영에 특별한 해법을 제시하지 못하고 있는 관련분야에서 높은 평가를 받을 수 있을 것으로 기대된다.

참 고 문 헌

[1] Bernaschi, M., and Richelli, G., "Development and results of PVMe on the IBM 9076 SP1," *Parallel and Distributed Computing*, 26, pp. 75-83, 1995.

[2] Pacheco P., *Parallel Programming with MPI*, San Francisco, CA: Morgan Kaufmann, 1997.

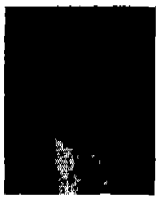
[3] Kim, Y., Pan, Z., Takle, E. and Kothari, S., "Parallel Implementation of Hydrostatic MM5 (Mesoscale Model)," *The 8th SIAM Conference on Parallel Processing for Scientific Computing*, Minneapolis USA, Mar. 14-17, 1997.

[4] Michalakes, J., Canfielf, T., Nanjundiah, R. and Hammond, S., "Parallel implementation, validation, and performance of MM5," *Proc. 6th Workshop on the use of Parallel Processors in Meteorology*, Reading, U. K., European Center for Medium Range Weather Forecasting, 1994.

[5] Grell, G. O., Dudhia, J. and Stauffer, D. R., "A description of the Fifth-Generation Penn State/NCAR Mesoscale Model (MM5)," *Tech. Rep.*, NCAR/TN-398+STR, National Center for Atmosphere Research, Boulder, Colorado, June 1994.

[6] Lee, Y., Choi, J., Kim, J. Y. and Kim, Y., "The Effect of Hi-resolution SST on Storm Scale Prediction in Point of Operational Prediction System," *The Tenth PSU/NCAR Mesoscale Model User's Workshop*, Boulder USA, June

21-22, 2000.
[7] Kim, Y., Kothari, S., Takle, E. and Pan, Z., "A run-time library and load balance analysis for parallel atmospheric models." *Symposium on Regional Weather Prediction on Parallel Computer Environments*, Athens, Greece, 1997.
[8] Foster, I. and Toonen, B., "Load-Balancing Algorithms for Climate Models," Proc. *Scalable High-Performance Computing Conf. IEEE*, pp. 674-681, 1994.
[9] Hsieh, J., Leng T., Mashayekhi V. and Rooholamini R., "Architectural and Performance Evaluation of GigaNet and Myrinet Interconnects on Clusters of Small-Scale SMP Servers," *Proc. Supercomputing Conference*, Dallas, USA, 2000.
[10] Lin, W., Lau, R., Hwang, K., Lin, X. and Cheung, P., "Adaptive Parallel Rendering on Multiprocessors and Workstation Clusters," *IEEE Transactions on Parallel and Distributed Systems*, 12:3, pp. 241-258, March 2001.



김 영 태
1986sus 연세대학교 수학과 졸업(학사). 1992년 Iowa State University, Computer Science, M.S.(석사). 1996년 Iowa State University, Computer Science, Ph.D.(이학박사). 1997년 Iowa State University, Research Assiciate (연구원). 1998년 ~ 현재 국립강릉대학교, 컴퓨터과학과 교수. 관심분야는 Distributed parallel system, Parallel compiler

이 용 희
1992년 경북대학교 천문기상학 학사. 1997년 경북대학교 대기과학 석사. 1998년 ~ 현재 기상연구소 예보연구실 기상연구사.

최 준 태
1990년 연세대학교 천문기상학 학사. 1998년 연세대학교 대기과학 석사. 1998년 ~ 현재 기상연구소 예보연구실 기상연구사.

오 재 호
1976년 서울대학교 대기과학 학사. 1983년 미국 Oregon State Univ. M.S. 1989년 미국 Oregon State Univ. Ph.D. 1994년 ~ 2001년 기상연구소 예보연구실장. 2001년 ~ 현재 부경대학교 환경대기과학과 교수.