LAN환경에서 유휴시간 예약에 기반한 PC Cluster의 설계

김영균° 오길호
금오공과대학교 컴퓨터공학부
{ygkim°, gilho}@cespc1.kumoh.ac.kr

Design of an Idle Time Reservation based PC Cluster on LAN Environment

Young-Gyun Kim° Gil-Ho Oh
School of Computer Engineering, Kumoh National Institute of Technology

요 약

본 논문에서는 TCP/IP프로토콜을 사용하는 LAN환경에서 PC들을 사용한 PC클러스터에 관해 연구하였다. LAN환경의 PC들은 안정된 사용환경을 확보할 수 있고, 특정한 사용 시간에 따라 작업을 배치할 수 있는 특징을 갖는다. 또한 별도의 전용의 클러스터를 설치하기 위한 하드웨어가 필요로 하지 않는 장점도 갖는다. 특히 PC실험실습실의 PC들은 주로 주간 시간대(오전9시~오후6시)에만 실험실습 수업에 사용이 되고, 야간 시간대(오후6시~익일 오전9시)에는 사용되지 않는 유휴 시간을 갖는다. 이러한 사용되지 않는 유휴 시간대의 PC들을 CPU지향적인 복잡한 작업들을 연산하기 위해 사용할 수 있으며, 별도의 클러스터 시스템을 구성하기 위한 하드웨어(클러스터를 위한 컴퓨터들, 스위치장비, 케이블링 등)를 구입할 필요가 없기 때문에 저가격의 고성능을 요구하는 PC클러스터를 쉽게 구성할 수 있다. 공동실험실습실의 PC들마다 유휴시간을 Time table에 예약하고 이에 따라 작업을 배치하고 연산결과를 수집한다. 장기간의 연산작업이 필요로 하는 대규모 연산에 사용함으로써 상당한 경제적 효과를 얻을 수 있다. 또한 동일 실습실의 PC들은 대개 동일한 성능의 동일한 하드웨어와 운영체제를 사용하는 PC들로 구성되기 때문에 Homogeneous node로 구성된 PC 클러스터를 구축하기가 용이하다.

1.서론

최근 Cluster Computing에 대해 활발한 연구가 진행되고, 저비용 고성능의 컴퓨팅 플랫폼에 대한 해결책으로 PC Cluster에 대해 널리 연구되어 왔다[1,2,3]. 특히 PC Cluster는 병렬/분산처리를 위한 특별히 설계된 전용의 하드웨어의 제작 없이 흔히 구입할 수 있는 부품들로 시스템을 구축할 수 있는 특징을 갖는다. 본 논문에서는 LAN환경에서 유휴(Idle) PC들을 연산에 활용하는 PC Cluster 시스템에 대해 연구하였다. LAN환경에는 많은 유휴 PC가 있으며, 특히 PC실험실습실에 있는 PC의 경우 특정 시간대에만 사용되고, 다른 시간대에는 사용되지 않는 특징을 갖는다. 대개 오전 9시에서 오후 6시까지 집중적으로 사용되고, 오후 6시부터 다음날 오전 9시까지는 사용되지 않는 특징을 보인다. 이러한 사용되지 않는 유휴 시간대에 많은 시간을 필요로 하는 연산작업을 각 유휴 PC마다 할당하여 연산함으로써, 장시간의 연산 시간이 요구되는 작업들을 위해, 별도의 전용 하드웨어를 구입할 필요가 없다는 장점을 갖는다. 실험실습실 PC의 경우 공휴일(토, 일요일 포함)과 연중 하계 및 동계 방학기간의 4개월 정도는 항상 유휴 상태로 있다. 이러한 다른 용도로 사용되지 않는 유휴 시간대와 유휴기간을 타임테이블(Time table)에 등록하여 연산지향작업을 배치함으로써 상당한 경제적인 효과를 얻을 수 있다. 특히 실험실습실 PC의 경우 동일한 성능의 동일한 운영체제와 동일한 하드웨어로 구성되어 있고, 고속의 LAN으로 연결된 경우가 많아 PC Cluster시스템을 구성하기에 용이한 특징을 갖는다. 또한 LAN환경의 경우 인터넷기반의 메타컴퓨팅(Meta-computing)과 달리 좀더 신뢰할 수 있고 안정된 사용환경을 확보할 수 있다. 특히, 많은 사용자들이 실험실습용도로 사용되는 주간의 시간대를 피함으로써 노드의 부하를 고려하거나 네트워크 부하를 고려해야 하는 복잡한 스케쥴링(Scheduling)방법이 필요하지 않는다. 또한, 물리적인 위치를 변경할 필요가 없기 때문에, Cluster시스템을 설치하기 위한 별도의 물리적 공간이 필요 없으며, 실험 실습용으로 사용되는 PC들을 그대로 사용하므로 실험실습실 PC들의 활용도를 높일 수 있다는 것이 가장 큰 장점이다.

2.관련연구

2.1 PC Cluster

높은 비용의 전용 대형컴퓨터를 사용하는 것보다 낮은 비용의 PC들을 네트워크로 연결 함으로서 비슷하거나 훨씬 뛰어난 성능을 발휘할 수 있는 클러스터(Cluster)시스템들에 대한 연구가 활발히 진행되어 왔다[1,2,3]. 클러스터 시스템은 크게 전용의 클러스터로 구성된 경우와 비전용의 클러스터로 구성된 경우, 이 두 가지 형태가 혼합된 형태가 있다. 고성능(High-Performance)과 고가용성(High-Availability)을 위해 주로 사용되고 있으며, 네트워크 상의 분산된 컴퓨팅 자원을 효율적으로 사용하기 위해 많이 연구되고 있다.

2.2 LAN으로 연결된 실험실습실 PC들을 활용한 PC Cluster

각 대학이나 연구기관마다 전용의 PC실습실이 갖추어져 있으며, 실습실의 PC들은 실험실습 시간외에는 사용되지 않는 특징을 갖는다.

이러한 사용되지 않는 유휴 시간을 조사해보면 실습실마다 차이가 있지만 24시간 중 대략 60%이상이 유휴하고, 토, 일요일에는 100% 사용되지 않는다. 또한 대학교의 실습실의 경우, 연중 하계 및 동계 방학기간 중인 총4개월 정도도 100%유휴 상태로 있다는 것을 쉽게 알 수 있다. 실험실습실의 PC를 사용할 경우와 기존의 전용 클러스터 시스템의 장단점을 표1과 같이 비교해 볼 수 있다.

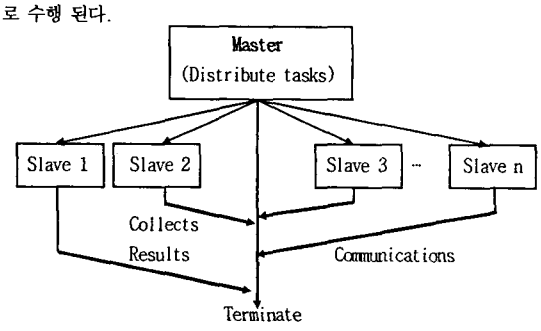
<표1. 기존의 PC클러스터와 제안한 실습실 PC클러스터의 특징 비교>

	기존 PC클러스터	제안한 PC클러스터
설치 공간	별도의 설치 공간 필요	실습실의 PC사용, (설치 공간 불필요)
전용 하드웨어	전용 PC 필요	전용 PC 필요 없음
클러스터 전용 소프트웨어	필요 (PVM,MPI,Java기반)	필요 (PVM,MPI,Java기반)
통신 하드웨어	전용의 통신하드웨어 및 케이블링 작업 필요	불필요
신뢰도	높음	높음, 전용 PC클러스터 보다 낮음
구성 노드	Homogeneous	Homogeneous
전송 속도	전용의 통신 하드웨어 사용시 아주 높음	보통 (실습실 환경에 의존적)
사용 시간	100% 클러스터 연산을 위해 사용	유휴시간 활용 (심야 시간 및 토, 일요일, 하계, 동계 방학 4개월)

3.시스템의 설계

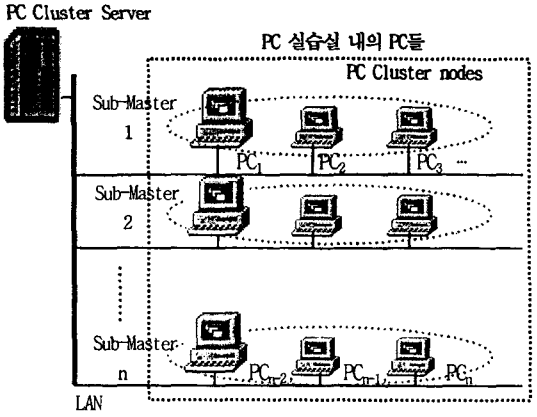
3.1 시스템의 구성

제안한 방법을 사용하는 시스템은 TCP/IP 프로토콜을 사용하는 LAN 환경으로서 그림2와 같이 구성 되고, 그림 1과 같이 Master와 다수의 Slaves로 구성된 Task-Farming(or Master/Slave)형태[1]의 PC Cluster로 수행 된다.



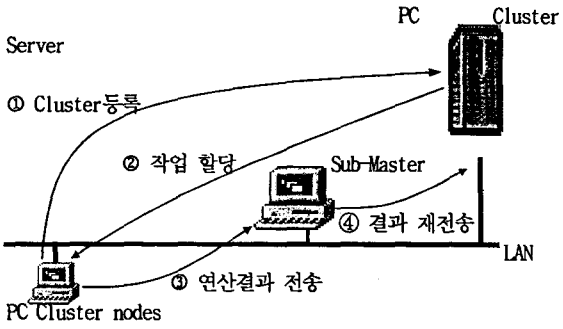
<그림 1. Task-Farming(또는 Master/Slave) 모델>

Master역할을 수행하는 PC Cluster Server가 해결하고자 하는 문제(Problem)를 작은 형태의 작업들(tasks)로 분할된 작업을 가지고 있고, 연산의 최종 결과들을 만들기 위해 Slaves로부터 부분 연산 결과를 수집하는 책임을 진다. 작업의 분배(Distribution)는 연산 초기에 Master(PC Cluster server)로부터 Slaves(PCs)로 수행되며, 연산결과를 수집할 때 단일의 Master에게 다수의 Slaves로부터 연산결과가 동시에 집중될 때 발생할 수 있는 병목현상을 피하기 위해 일정한 크기의 PC를 소규모로 클러스터링 하여 Sub-master를 두는 방법을 사용한다. Sub-master는 지정된 노드들의 집합으로부터의 부분연산 결과를 통합하여 최종적으로 Master에게 전달하는 형태를 취한다.



<그림2. 제안한 PC실습실의 PC를 사용하는 PC Cluster의 구성>

연산에 참여하고자 하는 노드로서 실습실의 각 PC는 PC Cluster Server에 IP주소를 등록하고, 유휴시간에 작업을 할당 받는다. 노드는 할당된 연산을 수행 후 PC Cluster Server에게 연산 결과를 전송한다. PC Cluster Server는 최종적으로 각 노드로부터 수신된 연산 결과를 통합한다.



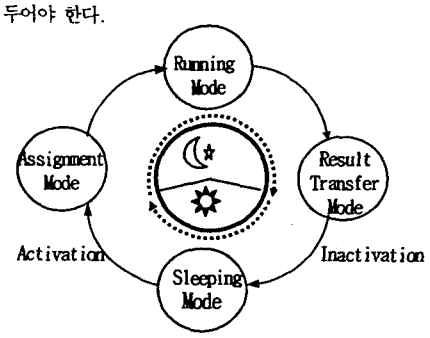
<그림3. 제안한 시스템의 수행 과정>

Server는 각 PC로부터 수신된 결과를 저장하기에 충분한 하드디스크 용량을 갖고 있어야 한다. 각 노드에 전송할 연산 코드와 연산에 필요로 하는 연산 데이터를 저장하고 이를 전송한다. 일정한 시간 간격으로 노드가 연산을 계속 수행하고 있는지 감시한다. 각 노드는 단일의 유휴종료 시간 내에 연산이 완료되지 않으면, 유휴종료시간 이전에 연산 결과를 Server로 전송하고, 연산을 다음 유휴 시작 시간까지 일시 중단하고 Sleeping모드로 설정한다. 이후 유휴시작 시간이 되면, 중단된 연산을 활성화 하여 중단된 연산을 계속 수행한다.

3.2 작업할당 방법

Server는 클러스터에 참여하는 등록된 PC의 IP테이블을 참조하여 각각의 유휴 노드로 등록된 PC들의 연산시작 시간에 연산 코드를 담고 있는 작업을 전송한다. 특히 일반 사용자들이 사용하지 않는 유휴시간을 이용하므로 각 노드의 부하를 동적으로 균등화할 필요가 없으며 복잡한 스케쥴링 기법이 필요 없다는 장점을 갖는다. 각 노드에 배치된 작업을 각 노드의 Client프로그램이 수행하고 예약된 종료시간 전에 연산된 결과를 Server로 전송함으로써 연산 결과를 수집한다. 각 노드에는 지역 하드디스크의 공간 중 일부를 클러스터 연산의 중간 결과를 저장하기 위한 공간으로 고정된 크기의 하드디스크 저장 공간을 미리

확보해 두어야 한다.



<그림 4. 각 노드의 유휴 시간대에 따른 작업 수행 주기(cycle)>

□ PC Cluster Server의 작업 순서

① 각 노드의 유휴 시작 시간 확인 후 각 노드에 연산전송
② 연산 중 중단된 노드가 있는 지 주기적으로 점검
③ 모든 연산 결과가 도착할 때까지 ②번을 반복하며 대기
④ 연산결과 통합

□ 각 노드의 PC Cluster Client 작업 순서

① 서버로부터 연산작업 수신
② 연산작업을 수행
③ 유휴종료시간에 맞추어 서버로 연산 결과 전송
④ 다음 유휴시간까지 Sleeping mode로 진입

3.3 Cluster 등록 테이블의 구조

연산에 참여하는 LAN상에 위치한 PC실습실의 각 노드에 대한 유휴시작 시간과 유휴종료시간을 Cluster Sever에 표2와 같이 정보를 유지함으로써 각 PC들을 등록, 관리 한다.

< 표2. PC Cluster Server 등록 테이블의 구조 >

No.	참여노드의 IP address	유휴시작시간	유휴종료시간	할당된 연산작업
1	202.31.130.91	오후6시	오전8시	Work 1
...	...	...	...	...
n	202.31.130.255	오후6시	오전8시	Work n

4. 제안한 방법의 평가 및 고려 사항

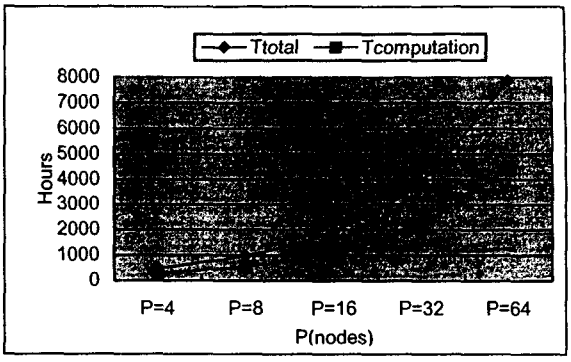
전체 연산소요 시간 T_{total}은 연산시간(유휴시간의)과 비연산 시간(비유휴시간의)이 주기적으로 나타나게 되므로 작업할당이 최초에 1번만 이루어지고, 연산결과는 주기적으로 전송된다고 가정하고 작업 할당 시간 T_{assign}, 연산시간을 $T_{running}$, 비연산시간을 $T_{sleeping}$, 결과수신시간 T_{result}라고 하면 연산일자 n에 대해 P개의 PC를 사용할 경우 식1과 같이 나타내어 진다.

$$T_{total} = (T_{running} \times n + T_{sleeping} \times n + T_{assign} + T_{result} \times n) \times P \quad (식1)$$

P개의 노드가 연산에 참여할 수 있는 총시간 $T_{computation}$ 은 비연산 시간과 송수신시간을 뺀 시간이므로 식2와 같이 나타내어 진다.

$$T_{computation} = (T_{total} - T_{sleeping} \times n - T_{assign} - T_{result} \times n) \times P \quad (식2)$$
$$= (T_{running} \times n) \times P$$

연산에 참여하는 노드의 개수 P가 증가함에 따라 $T_{computation}$은 늘어나기 때문에 분할이 가능한 작업에 효과적임을 알 수가 있다. 또한 각 노드의 부하와 네트워크 부하를 고려할 필요가 없는 정적 스케쥴링 기법을 사용할 수 있기 때문에 PC Cluster의 전체 성능은 연산 노드의 총 개수 P에 크게 의존하게 된다. 그림 5는 유휴시간이 오후 6시부터 다음날 오전 8시까지 14시간이고 비유휴 시간이 10시간, T_{assign}= 0.2, T_{result} =0.5 , n=5일인 경우에 대해 T_{total} 와 $T_{computation}$ 의 관계를 보여주고 있다. P가 증가함에 따라 비연산 시간도 증가함을 알 수 있다.



<그림5. 노드 수 P에 따른 연산에 사용 가능한 시간(Hours)>

중간 연산결과를 연산 노드 역할을 수행하는 각 PC의 지역 하드 디스크에 저장할 수 있는 안정된 공간만 확보되면, 연산 결과를 통합하기 위한 PC Cluster 서버의 저장 공간이 전체 연산결과를 저장하기에 부족할지라도 연산이 가능하다. 특정PC를 백업 서버로 지정하는 것도 가능하기 때문에 별도의 백업 서버가 필요 없다는 장점도 갖는다.

5. 결론 및 향후 연구방향

본 논문에서는 LAN으로 연결된 동일한 성능을 가지는 실습실PC들로 구성된 유휴 PC를 사용한 PC클러스터링 시스템에 대해서 연구하였다. 제안한 방법으로 PC클러스터링 시스템을 구성할 경우 신뢰성을 확보하고, 별도의 설치공간이 필요 없으며, 전용의 PC들을 구입할 필요가 없으므로 경제적으로 PC Cluster시스템을 구축할 수 있다는 장점을 갖는다. 유휴시간만을 사용하기 때문에 전용의 PC클러스터 시스템에 비해 처리 속도가 떨어지는 단점이 있지만, 장기간의 연산이 필요한 작업에 대해 경제적으로 사용할 수 있는 대안이 된다. 차후, LAN에서 유휴 자원을 최대한 확보할 수 있는 개선된 시간예약 방법에 대해 연구해 보겠다.

참고문헌

[1] Rajkumar Buyya, High-Performance Cluster Computing, Vol. II, Programming and Applications, Prentice Hall PTR,1999

[2] Putchong Uthayopas, Surachai Phaisithbenchapol, Krisana Chongbarirux, "Building a Resources Monitoring System for SMILE Beowulf Cluster", Proceeding of High Performance Computing, Asia '99, 1999

[3] Rajkumar Buyya, "PARMON: a portable and scalable monitoring system for clusters", SOFTWARE-PRACTICE AND EXPERIENCE, Softw. Pract. Exper. 2000; 30:1-17

[4] Anthony T. Chronopoulos, Razvan Andonie, "A Class of Loop Self-Scheduling for Heterogeneous Clusters", Proceedings of the 2001 IEEE International Conference on Cluster Computing(CLUSTER'01)

[5] Kam Hong Shum, "Fault Tolerant Cluster Computing through Replication", Proceedings of the 1997 International Conference on Parallel and Distributed Systems(ICPADS '97)

[6] Partha Dasgupta, Zvi M. Kedem, Michael O. Rabin, "Parallel Processing on Networks of Workstations: A Fault-Tolerant, High Performance Approach", 15th Intl. Conference on Distributed Computing Systems, May 1995, Vancouver, BC, Canada