'98 춘계학술발표회 논문집
한국원자력학회

PC 클러스터를 이용한 병렬처리컴퓨터 구축과 Benchmark

이 영욱 조영식 조인행 장종화
한국원자력연구소

요 약

고성능 network 장비로 연결된 PC cluster를 이용한 가상병렬처리컴퓨터 구축방법을 소개하였으며 이를 이용한 MCNP Benchmark 계산결과 CPU 수에 비례하는 계산능력을 보여주었다. PC를 이용한 가상병렬처리컴퓨터는 경제적인 장점 뿐 아니라 계산능력면에서도 고가의 대형컴퓨터를 추월하고 있다. 또한 병렬처리컴퓨터에 적합한 형태로 실행파일을 생산하는 compile 와 library 가 꾸준히 개발되는 추세이다. 따라서 많은 전산시간을 요하는 과학/공학계산에서 고가의 대형컴퓨터를 대체할 수 있는, PC cluster를 이용한 경제적인 가상병렬처리컴퓨터의 사용을 권장한다.

Abstract

We introduce our experience in construction and application of the virtual parallel computer using PCs clustered by high speed network cards. MCNP Benchmark calculation using the virtual parallel computer shows that the computing power is proportional to the number of CPUs. The virtual parallel computer with PC cluster has not only economic advantage but also its comparable performace to the high-priced super computer. Current development of compilers and libraries for parallel computing suggests us to consider an ecomonic approach using the virtual parallel computer with PC cluster in scientific and engineering calculations which need long computing time.

1. 배 경

과학 및 공학계산시 상당한 전산시간을 요하는 많은 문제들이 있다. 예를 들어 분자역학 모사, 3차원 영상처리, 장기기상예측, Monte Carlo 모사 등은 방대한 계산시간이 소요되기 때문에 병렬처리컴퓨터와 병렬처리 알고리즘을 사용한다. 현재 응용되고 있는 병렬처리연산용 컴퓨터는 하드웨어적 관점으로 볼때 크게 두가지로 구분할 수 있다. 첫째로는 설계당시부터 한개 이상의 CPU를 장착할 수 있도록 한 CRAY, IBM 등에서 개발한 고가의 공유메모리(distributed-memory) 컴퓨터이다. 이에 반해 최초 설계시에는 단일 CPU만을 제어할수 있는 컴퓨터를 network 장비로 연결하여 병렬처리연산을 가능하게 하는 방법이 있다. 후자의 경우를 가상병렬처리컴퓨터라고 부르며 최근 network 장비의 성능이 크게 향상되고 저가의 고성능 PC의 출현으로 고가의 공유메모리-다중 CPU 컴퓨터에 비해 가격대비 뒤지지 않는 성능을 보여서 주목된다. 이러한 병렬처리연산기를 응용계산에 사용하기 위해서는 기존 원시코드(source code)를 병렬연산에 알맞는 실행파일로 생산하는 compiler 를 이용하는 방법과 기존 원시코드를 약간 수정하고 병렬처리용 library 와 결합하는 방법이 있다.

본 보고서에서는 한국원자력연구소 핵자료평가랩에서 구축한 병렬처리컴퓨터와 PVM[1] library 를 이용한 병렬처리연산의 benchmark 결과를 소개하여 향후 저가의 고성능 병렬처리컴퓨터의 구축에 관심을 가지는 곳에 참고가 되도록 하였다.

2. PC Cluster 구축

그림 1에서와 같이 8 대의 Pentium II PC에 Linux OS를 설치한 후 100 Mbps 급 network card를 이용하여 병렬처리를 위한 LAN(local area network)을 구축하였다. 이 중 하나의 PC에는 10 Mbps급 network card를 추가하여 외부접속을 가능하게 하였다. 사용된 PC 및 network card, 그리고 network hub의 사양은 표 1에 요약하였다.

각각의 8대 PC에는 LAN 구축용으로 유보된 fake IP address 192.168.2.1 - 8 을 부여하고 internet 외부접속을 위하여 2개의 network 카드가 장착된 PC는 network device 를 eth0, eth1 두개를 가지도록 하여 정식 IP adress를 추가 할당하였다. 따라서 병렬처리연산시 외부 network traffic의 간섭을 최소화하고 internet 을 통하여 외부에서도 PC cluster를 접속할 수 있도록 하였다. 각 PC의 시간을 동기화 (synchronize)하기 위하여 timed daemon을 모든 PC에 설치하였고 PC 각각의 local disk는 나머지 7대의 PC가 공유할수 있도록 NFS(network file system)을 설치하였다. PC cluster는 switch board 를 통하여 하나의 monitor 와 keyboard로 조작할 수 있도록 하여 설치공간을 최소화 하였다. 이상으로 8대의 PC 가 고속 network card를 통해 8개의 CPU 를 가지고 8개의 local disk가 NFS로 묶여진 network 컴퓨터가 구축된 것이다. 이 시스템으로부터 가상병렬처리컴퓨터를 구현하기 위해서는 적합한 library 를 설치하는 일과 병렬처리가 가능한 응용프로그램을 작성/실행하는 일이 남아 있다.

3. PVM (Parallel Virtual Machine) 소개

PVM[1] 은 unix 를 탑재한 서로 다른 기종의 컴퓨터를 하나의 가상병렬컴퓨터로 사용할 수 있게하는 소프트웨어 시스템의 하나로 MPI 와 함께 현재 실제적인 표준(standard de facto)으로 자리잡고 있다. 1989년 미국 ORNL(Oak Ridge National Laboratory) 에서 개발이 시작되었으며 현재 미국에너지성(DOE), 미국의 과학재단 (National Science Foundation) 그리고 테네시주의 지원을 받으며 현재 version 3까지 나와 있다. PVM 하에서는 사용자가 정의한 직렬, 병렬 또는 vector 컴퓨터들은 하나의 공유메모리를 가지는 거대한 가상컴퓨터로 작동한다. PVM은 가상컴퓨터에서 태스크룰 기동하게 하며 태스크간의 통신과 동기화를 할 수 있도록 하는 함수들을 제공한다. 여기에서 태스크라 함은 PVM에서의 연산단위를 말하는 것으로 unix 에서의 process 와 같은 개념이다. Fortran 이나 C 로 작성된 응용프로그램은 대부분의 분산메모리컴퓨터에 응용되는 message-passing 구조를 이용하여 병렬화할 수 있으며 이에 따라 생성되는 다중태스크들이 협력하여 문제를 병렬처리 한다.

PVM의 장점중의 하나는 이질성(heterogeneity)를 허용한다는 데 있다. 다른 태스크간의 수치자료교환을 위한 표준 프로토콜인 RPC(remote procedure call)[2] 를 채용함으로써 서로다른 하드웨어를 가진 기종간에도 자료를 교환할 수 있다. 따라서 공유메모리를 가진 다중 CPU 컴퓨터 뿐만 아니라 서로 다른 컴퓨터들이 여러가지 network으로 연결된 가상컴퓨터까지 지원한다. 이런 PVM의 장점으로 저가의 고성능 PC 와 고속 network장비를 이용한 PC cluster에 쉽게 적용할 수 있다. 다음 장에서는 PC cluster 에 설치한 PVM 3.3.11 상에서 PVM version의 MCNP[3]를 이용한 benchmark 결과를 소개한다.

4. MCNP 벤치마크

4.1 MCNP PVM 방법론

MCNP에서는 초기부터 활동하는 주태스크(master task)외에 입력으로 지정한 갯수의 부태스크(subtask 또는 CPU)를 PVM을 통하여 형성한다. 또한 주태스크에서는 매 계산주기마다 부태스크 갯수의 약 20배 정도의 마이크로태스크를 준비한다. 하나의 마이크로 태스크는 약 200개의 선원입자로 구성된다. 한번의 주기에 결정된 수십개의 마이크로 태스크는 수 개

의 서브태스크에 의해 처리되며 서브태스크가 하나의 마이크로태스크를 완료하면 다음번의 마이크로태스크를 할당받는다. 이러한 과정을 총괄하는 주태스크는 각 마이크로태스크를 분배하므로서 빠른 태스크는 더 많은 마이크로태스크를 처리하게 되고 결과적으로 전체 경과시간은 최소화된다. 이 과정에서 응답이 매우 느린 태스크는 제외되며 이후의 작업에 참여하지 못한다. 이러한 과정은 부프로그램 msgcon에서 총괄하고 있으며 개략적인 흐름도를 그림 2 에 도식화하였다.

MCNP에서는 초기 선원입자를 위한 선원난수로서 152917 번씩 건너뛴 난수들을 사용한다. 따라서 선원발생에 사용하는 선원난수열은 독자적으로 난수성을 가지며, 입자 수송과정에 필요한 난수와의 상관관계가 없게 된다. 이러한 선택은 MCNP를 다중 프로세스에 사용할 때도 편리하다. 즉 선원입자를 바꾸는 경우에 선원난수만 증가시키면 되기 때문에 각 마이크로 태스크에서 사용할 초기 선원난수를 미리 결정할 수 있다. 중성자 수송과정에서 사용할 난수는 초기 선원난수를 초기치로 사용하는 난수열을 사용하면 된다. 이러한 전략을 사용하면 다중프로세스 작업의 결과와 단일 프로세스 결과의 난수열이 모두 같아지므로 다중화와 관계없이 동일한 결과를 얻게된다. 수송과정에 필요한 난수의 발생은 부프로그램 rang를 사용하지만 선원난수의 발생은 부프로그램 advijk를 이용한다.

4.2 Benchmark 결과

실험에 사용한 문제는 하나로의 임계도를 구하는 문제로서(KCODE) 2,001,020 history를 계산한다. 본 클러스터의 계산기는 모두 pentium-II PC이나 CPU중 4 개는 266 MHz, 4 개는 300 MHz 속도이다. 표에는 부태스크의 갯수와 사용한 CPU갯수에 따른 실제소요시간을 보여준다. 부태스크의 숫자가 1 개일 경우에는 CPU도 1 개만 사용하나 부태스크의 숫자가 2 개이상이면 주태스크가 사용하는 CPU 1 개가 추가되어 태스크 숫자보다 1개 더 많은 프로세스를 사용한다. 이 프로세스의 숫자가 CPU 숫자보다 많으면 주태스크가 있는 CPU에서 나머지 부태스크를 처리한다.

표 2. 의 소요시간 자료에서 태스크 숫자가 2 - 7 개인 경우에 대한 상관관계를 분석하면 태스크 갯수와 소요시간간에는 다음식이 성립함을 알수있다.

$$T = 37 + \frac{627}{N_T} \ (\pm 4 \text{ min}), \qquad N_T : \text{태스크의 갯수}$$

이러한 결과는 더 많은 CPU를 사용하여 검증할 수 있을 것이며 18 개 정도의 CPU를 사용하면 보다 정확한 상관관계를 알 수 있다. 참고로 표 3. 에서는 동일한 문제를 타기종에서 하나의 태스크로 계산할 경우의 소요시간을 나타내었다.

5. 맺 음 말

이상으로 PC Cluster를 이용한 병렬처리컴퓨터 구축방법을 소개하였으며 이를 이용한 MCNP Benchmark 계산결과 CPU 수에 비례하는 계산능력을 보여주었다. 최근 intel 계열의 고성능 CPU가 저가에 공급되고 있으며 alpha CPU가 장착된 저기의 WS 이 공급되고 있고 Network 장비가 고속화되면서 가상병렬처리컴퓨터는 경제적인 장점 뿐 아니라 계산능력면에서도 고가의 대형컴퓨터를 추월하고 있다. 또한 병렬처리컴퓨터에 적합한 형태로 실행파일을 생산하는 compiler 와 library 가 꾸준히 개발되는 추세이다. 따라서 많은 전산시간을 요하는 과학/공학계산에서 고가의 대형컴퓨터를 대체할 수 있는, PC cluster를 이용한 경제적인 가상병렬처리컴퓨터의 사용을 권장한다.

참고문헌

[1] Al Geist, et al., "PVM 3 User's Guide and Reference Manual", ORNL/TM-12187

(1994).
[2] ISO Remote Procedure Call Specification, ISO/IECCD 11578 N6561, ISO/IEC (1991) 또는 OSF, DCE RPC Internal and Data Structures, revision 1.0 (1993).
[3] J. F. Briesmeister, "MCNP, A General Monte Carlo N-Particle Transport Code", LA-12625-M, Version 4B (1997).

표 1. PC cluster 하드웨어 사양

CPU	Pentium II 266 MHz	4 대
	Pentium II 300 MHz	4 대
Memory	64 MB	8 대
Hard disk	삼성 VA34324A, 4124MB w/478kB Cache	8 대
Network Card	3Com 3c905 Boomerang 100 Tx	8 대
	NE2000 card	1 대 (외부접속용)
Network Hub	3Com Superspec II, 100Tx 12 ports	1 대

표 2 소요시간

Task	CPU	시간 (분)	비고
1	1	780	PC1
1	1	671	PC2
2	3	351	PC2 * 3
3	4	242	PC2 * 4
4	5	197	PC2 *4 + PC1
5	6	161	PC2 * 4 + PC1 * 2
6	7	146	PC2 * 4 + PC1 * 3
7	8	122	PC2 * 4 + PC1 * 4
8	8	135	PC2 * 4 + PC1 * 4

PC1 : Pentium-II 266 MHz
PC2 : Pentium-II 300 MHz

표 3. 기종별 소요시간 (single CPU)

기종	소요시간 (분)
HP715/100 (100MHz)	1882
SGI Origin 2000	507
Sun Ultra Sparc II 200 MHz	586

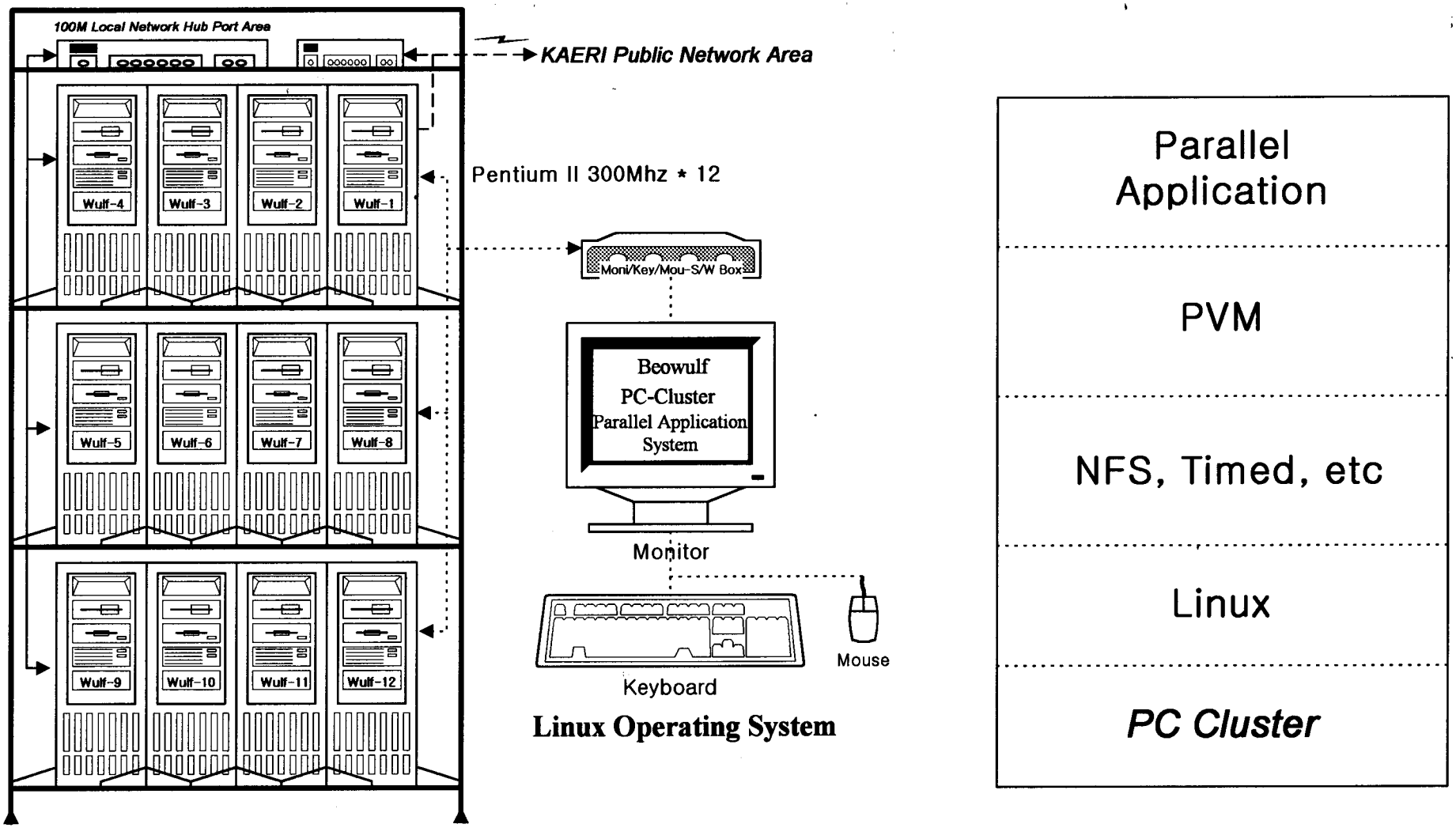

그림 1 PC Cluster 구 성 도

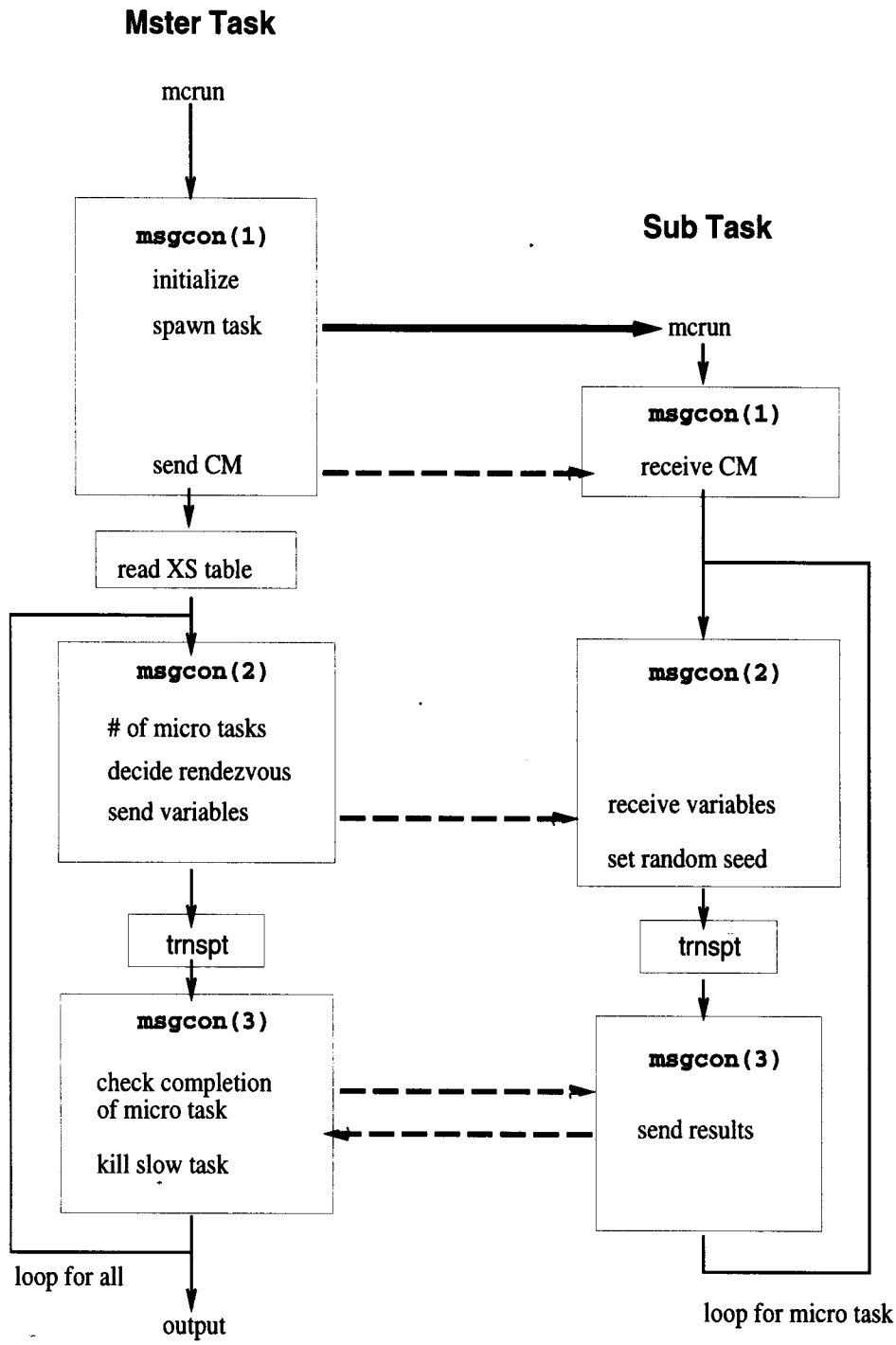

그림 2. MCNP PVM 의 작업흐름도