SMP 슈퍼컴퓨터에서의 집합 통신 성능 평가

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

Performance evaluation of collective communication on an SMP supercomputer

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

요 약

하드웨어 기술이 발전하면서 단위 노드당 집적되는 CPU 수와 메모리량이 증가되어 단위 시스템의 성능이 크게 개선되었다. HPC 시스템도 SMP 시스템과 같은 고성능 노드를 사용하는 클러스터 시스템 위주로 발전하게 되었다. 본 연구에서는 IBM Power 595 기반 슈퍼컴퓨터에서 집합 통신이 갖는 특성을 알아 본다.

1. 서론

MPI 와 같은 메시지 전달에 기반한 병렬처리 시스템에서의 집합통신(Collective Communication)은 작업에 참여하는 모든 계산 노드가 통신을 수행하기 때문에 일대일 통신보다 복잡하면서도 유용하게 이용되는 기능이다.

본 연구에서는 SMP 를 사용하는 슈퍼컴퓨터에서 집합 통신의 성능을 측정하고 그 특성을 살펴 보았다. 서로 다른 통신 비용이 요구되는 SMP 클러스터 시스템의 특성상 CPU 할당 방식에 따른 성능 차이를 확인할 수 있었다.

2. SMP 슈퍼컴퓨터

본 연구에서 사용된 슈퍼컴퓨터는 Power6 기반의 64way 시스템을 단위 노드로 사용하고 있으며, 노드들을 연결하는 주 네트워크는 인피니밴드를 이용하였다. 단위 계산 노드로 사용된 Power6 기반의 P595 시스템은 듀얼 코어로 구성된 CPU 4 개가 하나의 Processor book 을 구성하고 8 개의 Processor book 이 하나의 시스템을 구성하게 된다.

이처럼 SMP 시스템을 계산 노드로 이용하는 경우에는 CPU 내부, CPU 간, 시스템 내부 그리고 시스템 간의 통신 성능에 차이를 갖게 되므로 CPU 를 할당하는 방식에 따라 성능 차이가 발생할 수 있다. 본 연구에서는 이와 같은 사항을 고려하여 CPU 할당에 차이를 두어 집합 통신의 성능을 분석하였다.

3. 실험 환경

집합 통신의 성능을 분석하는데 사용된 하드웨어와 소프트웨어의 구성은 표 1 과 같다.

본 연구에서는 가능한 한 노드 안에 위치하는 프로세서를 사용하는 경우와, 일정 수의 계산 노드에서 균등하게 프로세서를 사용하는 경우의 집합 통신의 성능을 측정하고 분석하였다. 단위 노드 내에 프로세서를 할당하는 경우의 프로세서의 위치에도 차이를 줄 수 있는데 인접한 프로세서를 우선으로 배치하는 방법과 노드내에서 최대한 프로세서를 흩어지게 위치시키는 방법도 함께 적용하여 성능을 분석하였다. 16 개의 단위 시스템을 이용하여 1024 개의 프로세서까지 테스트하였다.

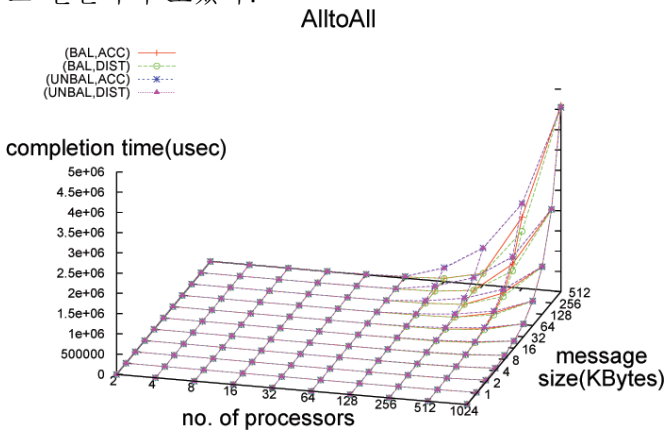
<표 1> 하드웨어 및 소프트웨어 구성

System	Power 595
CPU	IBM 64bit POWER6
CPU Clock Speed	5.0 GHz
Cache / core	L1: 64/64KB(I/D), L2: 8MB, L3: 32MB
OS	AIX Version 6.1
File System	GPFS Version 3.2.1
Programming Language	IBM XL Fortran Version 12.1 IBM XL C/C++ Version 10.1
MPI	PE Version 5.1.1
Scheduler	LoadLeveler Version 3.5.1
Benchmark Program	IMB 3.2

4. 실험 결과

집합 통신 중 주로 사용되는 기능(Broadcast, Allreduce, Allgather, Alltoall, Alltoallv)을 대상으로 실험을 진행하였다. 그림 1 은 실험 결과 중 일부를 보여주고 있다. 지면상 생략된 나머지 결과 역시, 프로세서수와 처리할 메시지의 크기가 증가할수록 성능 저하가 심하게 나타나고 있다. 이처럼 일부 구간에서의 성능저하 폭이 극심하여 각 경우의 성

능 차이를 파악하기가 용이하지 않아, 표 2 와 같이 각 경우 별 성능 차이 정도를 원래 성능에 대한 비율로 환산하여 보았다.



(그림 1) 집합 통신 성능: AlltoAll

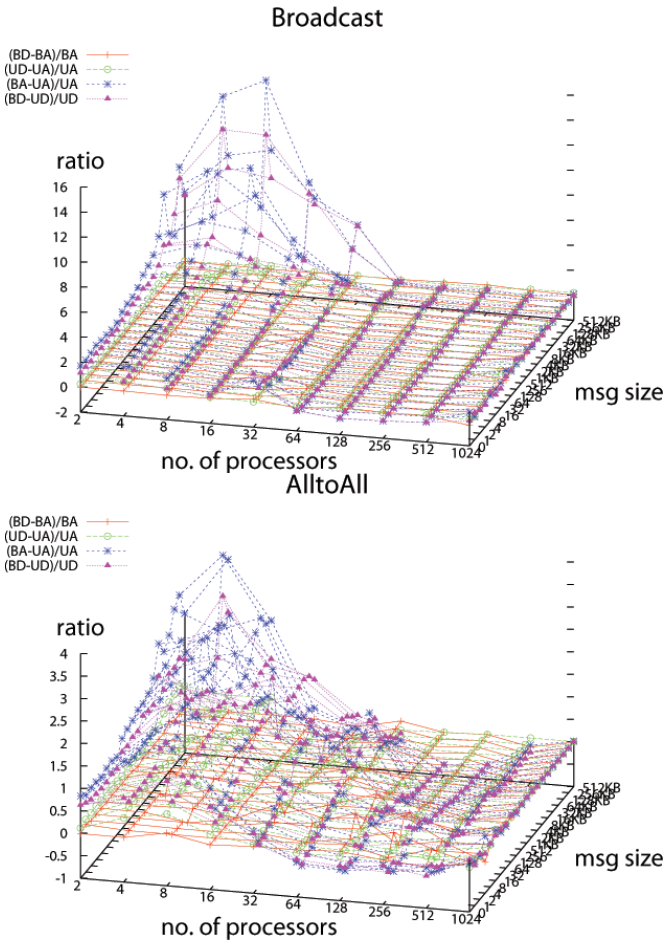
<표 2> 프로세서 할당 방식에 따른 성능 차이 비율

BD: 계산 노드(16 대)에 프로세서를 균등하게 분배, 노드 내에서는 흩어지게 배치
BA: 계산 노드(16 대)에 프로세서를 균등하게 분배, 노드 내에서는 인접하게 배치
UD: 단일 노드 내에 프로세서를 우선으로 배정, 노드 내에서는 흩어지게 배치
UA: 단일 노드 내에 프로세서를 우선으로 배정, 노드 내에서는 인접하게 배치

		(BD-BA)/BA	(UD-UA)/UA	(BA-UA)/UA	(BD-UD)/UD
Broadcast	평균	0.0785	0.2381	1.1605	0.6587
	최소	-0.3558	-0.6596	-0.7358	-0.6994
	최대	0.9495	2.1457	15.5963	11.3582
Allreduce	평균	0.0686	0.5217	1.2899	0.7402
	최소	-0.9331	-0.1187	-0.3235	-0.9434
	최대	1.0072	26.0308	17.2411	4.0272
Allgather	평균	0.0926	0.3418	0.5693	0.2917
	최소	-0.2186	-0.2919	-0.8535	-0.8615
	최대	1.4230	1.9600	3.8398	1.8427
Alltoall	평균	0.0173	0.1331	0.6129	0.3545
	최소	-0.3025	-0.3804	-0.7603	-0.7612
	최대	0.5278	0.8137	3.6256	2.7137
Alltoallv	평균	-0.0038	0.0804	0.3958	0.2533
	최소	-0.2612	-0.4005	-0.6926	-0.6888
	최대	0.2995	0.6594	3.6855	2.7707

표 2 에 표시된 결과값이 실행 시간을 이용하여 구한 값이라는 점을 염두에 두고 각 실험 구간에 대한 평균 값을 살펴보면 대부분의 경우, 노드 내에서는 프로세서 할당을 인접하게 하는 것이(BA, UA) 근소하게 우수한 결과를 보여 주었다. 노드 내부에서 CPU 할당 정책보다는 전체 노드에 프로세서를 어떻게 배치하는가가 성능에 영향을 많이 주고 있었다. 즉 가능하면 한 노드내에 프로세서를 우선 할당하는 것이(UD, UA) 우수한 성능 보이는 것으로 나타났다.

그림 2 는 표 2 의 결과를 구간별로 보다 자세히 표현한 것이다. 지면상 생략된 나머지 결과 역시 유사한 패턴을 보이고 있다. 프로세서 수가 64 개 이하인 경우, 성능 차이가 크게 존재함을 확인할 수 있다. 이는 단위 노드가 최대로 수용할 수 있는 64 개까지는 같은 노드에 작업을 할당하는 것이 집합 통시 비용을 최소화 할 수 있음을 나타내고 있다. 그러나 64 개 이상의 프로세서를 사용하는 경우에는 인피니밴드를 이용한 노드간 통신을 수행하기 시작하므로 성능 차이가 크지 않음을 알 수 있다.



(그림 2) 집합 통신 성능 차이 비율: Broadcast, AlltoAll

5. 결론

본 연구에서는 SMP 를 사용하는 슈퍼컴퓨터에서 집합 통신의 특성을 살펴 보았다. 프로세서 수가 증가함에 따라 급격한 성능 저하를 보이고 있으며, 프로세서의 할당 방법에 따라 성능 차이가 나타났다. 특히 계산 작업에 쓰일 프로세서 수가 단위 노드가 수용할 수 있는 물리적인 프로세서 수 이하인 경우, 단일 노드에 할당하는 것이 집합 통신 비용을 최소화 할 수 있음을 확인하였다. 즉 요구되는 프로세서 규모에 따라 계산 자원의 할당에 다른 기준을 적용하여야 할 것으로 예상된다.

참고문헌

[1] IBM Redbooks, "IBM Power 595 Technical Overview and Introduction," http://www.redbooks.ibm.com/abstracts/redp4440.html?Open

[2] H. Q. Le, et al, "IBM POWER6 microarchitecture," IBM Journal of Research and Development, Vol. 51(6), pp. 639~662, 2007