소규모 클러스터 시스템에서의 PVFS 성능 최적화에 관한 연구

An Analysis of PVFS Performance Optimization on Small Cluster System

조혜영, 차광호, 김성호
한국과학기술정보연구원 슈퍼컴퓨팅센터

Hyeyoung Cho, Kwangho Cha, Sungho Kim
Supercomputing Center Korea Institute of Science and Technology Information

요약

고속 네트워크로 연결된 병렬 컴퓨터 및 클러스터 시스템의 응용 분야가 다양화되고 사용자가 증가함에 따라 분산 및 병렬 파일 시스템에 대한 관심이 높아지고 있다. 특히 복잡한 네트워크로 구성된 클러스터 시스템을 보다 효율적으로 사용하기 위해서 분산 및 병렬 파일 시스템의 성능을 최적화하려는 많은 연구가 진행 중이다. 본 논문에서는 소규모 클러스터 시스템에서 널리 사용되고 있는 파일 시스템인 PVFS(Parallel Virtual File System)의 성능을 분석하고, 주어진 네트워크 환경에 따라 성능을 최적화할 수 있는 방법인 FlowBuffer의 따른 변화에 PVFS의 성능을 비교 분석하였다.

Abstract

Recently with increasing the use of parallel computing and cluster system which was connected high speed network, the interest about distributed and parallel file system is increasing. Specially, there are many researches, which focused on optimizing the performance of distributed and parallel file system for the more efficient use of cluster system. In this paper, we analyzed the performance of PVFS(Parallel Virtual File System) in small cluster system. In addition, to improve the PVFS performance we proposed the chancing the size of flow buffer according to the network speed and we optimized the PVFS performance on small cluster system.

1. 서 론

고속 네트워크로 연결된 병렬 컴퓨터 및 클러스터 시스템의 응용 분야가 다양화되고 사용자가 증가함에 따라 분산 및 병렬 파일 시스템에 대한 관심이 높아지고 있다. 특히 복잡한 네트워크로 구성된 클러스터 시스템을 보다 효율적으로 사용하기 위해서 유휴 자원의 활용, bandwidth 및 throughput의 증대라는 측면에서 많은 연구가 진행되고 있다.

파일 시스템의 성능을 개선하고자 하는 노력은 네트워킹 기법을 이용하여 다수의 디스크 내지는 스토리지를 연결하고 I/O처리를 분산시키는 분산 및 병렬 파일 시스템의 개념을 만들어 내었다. 즉, 다수의 컴퓨터에 장착된 디스크 내지는 스토리지를 네트워크로 연결하여 하나의 논리적인 파일 시스템으로 구성함으로써 유휴 자원의 활용, I/O처리 대역폭 증대 등의 효과를 기대할 수 있어서 고성능 컴퓨팅 분야뿐만 아니라 대규모 데이터 처리를 위한 파일 시스템으로 고려되고 있다. 이러한 현상을 반영하듯, 여러 종류의 분산 및 병렬 파일 시스템들이 발표되고 있고, 구성이나 성능 면에서 약간씩 차이를 보이고 있다.

본 논문에서는 소규모 클러스터 시스템에서 널리 사용되고 있는 파일 시스템인 PVFS(Parallel Virtual File System)의 성능을 분석하고, 주어진 네트워크 환경에 따라 성능을 최적화할 수 있는 방법으로 PVFS의 내부 버퍼인 FlowBuffer의 변화에 따른 PVFS의 성능을 비교 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서는 분석 대상인 PVFS(Parallel Virtual File System)에 대해서 살펴보고, 3장에서는 성능 측정 환경을 설명한다. 4장에서는 성능을 최적화할 수 있는 요소인 FlowBuffer의 변화에 따른 PVFS의 성능을 비교분석하고, 마지막으로 5장에서는 결론에 대하여 기술한다.

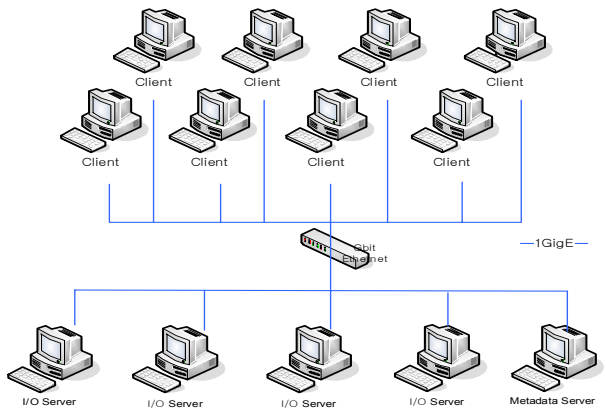
2. PVFS(Parallel Virtual File System)

Clemson 대학에서 개발되어 소형 클러스터 시스템에 많이 이용되고 있는 분산 파일 시스템인 PVFS(Parallel Virtual File System)은 파일 시스템의 성능을 높이기 위하여 병렬 파일 시스템이 취하는 전형적인 방식이 RAID 0처럼 파일을 쪼개서(stripe) 서로 다른 저장장치에 저장하는 방식을 사용한다. I/O를 담당하는 복수 I/O노드에 파일이 분산되어 저장되며 이에 대한 위치 정보를 관리하는 관리 노드가 존재한다. 이때 I/O노드와 관리 노드의 역할 수행을 위한 프로세스는 단일

노드 내에 존재가 가능하다[1,2,3].

3. 실험 환경

본 논문의 성능 측정 환경은 그림 1과 같이 4개의 I/O 서버와 1대 메타서버를 별도로 구성하였으며 4대 서버의 I/O 성능을 측정할 수 있도록 충분한 부하를 주기 위해 서버의 2배인 8대의 클라이언트 노드로 구성하였다.



▶▶ 그림 1. 시스템 구성도

표 1과 2는 성능 측정 시스템의 하드웨어 사양 및 소프트웨어 구성을 보여준다.

[표 1] 단위 노트 사양

CPU	AMD Opteron Processor 240(1.4Ghz)
# CPU / node	2
Memory	2GB
I/O Interface	PCI-X 100MHz
Storage Interface	Ultra SCSI

[표 2] 소프트웨어 구성

OS	Linux 2.6.9-55.0.2.ELsmp
MPI Library	mpich2-1.0.5p4
Parallel File System	pvfs-2.6.1
Benchmark Programs	IOR-2.10.0.1

4. 실험 결과

4.1 Flow Buffer 크기 변화에 따른 PVFS2 성능 비교

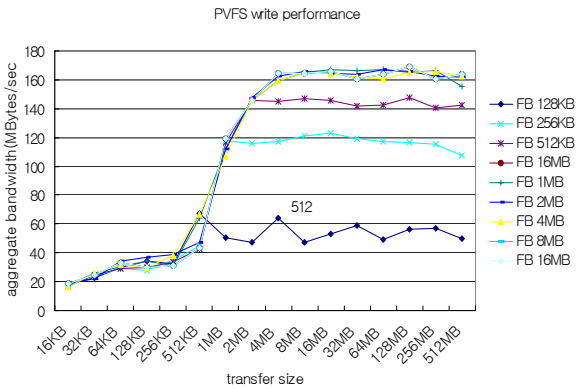
FlowBuffer란 PVFS의 내부에서 사용하는 버퍼로써, 하나의 I/O request가 1개의 flow로 되고, 이 flow 리스트들은 큐로 관리된다[4,5]. Flow Buffer는 이 때 사용되는 버퍼로 BMI(Buffered Message Interface)를 통해 한 번에 보내는 데이터의 크기에 영향을 준다. BMI는 PVFS에서 네트워크 계층 위에 제공되는 추상 계층으로 TCP, Myrinet, Infiniband 등 다양한 네트워크 프로토콜 위에서 동작한다.

Flow Buffer의 크기 변화에 따른 PVFS2의 성능의 테스트하기 위해서 병렬 파일 시스템의 성능을 측정하는데 널리 이용되고 있는 IOR 벤치마크를 이용하였다. IOR 벤치마크는 LLNL(Lawrence Livermore National Laboratory)의 SIOP(Scalable I/O) 프로젝트에서 개발되었다[6]. 표 3에 나타낸 것과 같이 I/O 서버 4대와 메타데이터 서버 1대로 PVFS를 구성하고, 8대의 클라이언트가 512MB파일을 읽고 read, write하여 총 4GB 파일로 테스트하였다. Transfersize는 16KB-512MB 구간에 대해 성능을 측정하였으며, 그 결과를 그림 2와 3에 보였다.

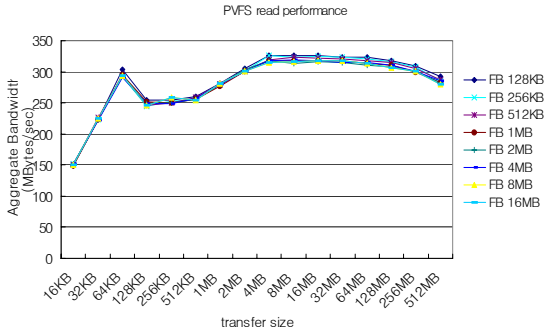
성능 측정 결과에 따르면, FlowBuffer의 크기는 PVFS의 Read 성능에는 거의 영향을 미치지 못 했으나, PVFS의 Write 성능 최적화에는 중요 요소가 될 수 있음을 보여주었다. Write의 경우, PVFS에서 default로 제공되는 256KB에 비해 최대 44.91%까지 성능 향상을 볼 수 있었다. FlowBuffer 크기에 변화에 따른 성능 최적화는 한 번의 I/O 콜에 사용하는 transfer size가 512KB 이상일 때부터 높은 성능 향상 폭을 보였다. 또한 1Gpbs에 네트워크 환경에서 4대의 I/O 서버 환경에는 FlowBuffer 크기가 1MB일 때부터 최적화 된 성능을 보였으며, 그 이상의 FlowBuffer 크기 향상은 성능에 많은 영향을 주지 못했다. 그리고 FlowBuffer의 크기 변화는 주어진 환경에서는 PVFS의 Read 성능은 거의 영향을 주지 못했다.

[표 4] 성능 측정 조건

I/O Server	4
Metadata Sever	1
# of Client	8
Block Size	512MB
Aggregate Filesize	4GB
TransferSize	16KB-512MB



▶▶ 그림 2. PVFS Write 성능



▶▶ 그림 3. PVFS Read 성능

4.2 Client 수의 변화에 따른 성능

그림 4와 5는 표 3과 같은 조건에서 transfer size가 4MB일 때, Client의 수의 변화에 따른 성능을 측정한 결과이다. Client의 수에 따른 시스템의 성능을 보면, 전체적으로 Client 노두 수에 따라 8개까지는 Client 수에 비례해서 PVFS의 전체 성능이 증가함을 볼 수 있다. PVFS의 기본 성능(FlowBuffer 256KB)과 비교할 때, Write의 경우 평균적으로 85.12%의 성능 향상을 보였으며, 특히 2노드와 4노드일 때 2배 이상의 성능 향상을 볼 수 있었다. 주로 FlowBuffer의 크기가 1MB와 2MB일 때 최대 성능을 보였으며, 4MB로 조절했을 때는 성능이 떨어지는 경향도 볼 수 있었다.

▶▶ 그림 4. Client 수 변화에 따른 PVFS Write 성능

▶▶ 그림 5. Client 수 변화에 따른 PVFS Read 성능

Read의 경우, 평균적으로 9.63% 성능 향상을 확인할 수 있었으며, FlowBuffer의 크기가 2MB의 경우 최대 성능을 발휘할 때가 많았고, 1MB와 2MB의 성능 차는 미미하였다. Read에서도 FlowBuffer의 4MB로 조절했을 때는 그 성능이 떨어짐을 확인할 수 있었다.

5. 결 론

한정된 자원을 보다 효율적으로 활용하기 위한 노력으로 클러스터 시스템에서 분산 및 병렬 파일 시스템의 성능 최적화는 중요 연구 분야이다. 본 논문은 소규모 클러스터 시스템에서 대표적으로 사용되고 있는 병렬 파일 시스템인 PVFS의 성능을 분석하고, 네트워크 환경에 따라 성능을 최적화할 수 있는 요소인 PVFS의 내부 버퍼 크기의 변화에 의한 PVFS의 성능 변화를 비교 분석하였다.

PVFS의 내 버퍼인 FlowBuffer의 크기 변화는 Read보다는 Write 연산에서 높은 성능 향상을 보였다. 본 논문의 실험 환경에서는 FlowBuffer 1MB, 2MB에서 최대 성능을 보였으며 Write의 경우 최대 44.91%의 성능 향상을 얻을 수 있었다. FlowBuffer의 크기를 다양한 네트워크 환경, I/O 서버 수, Client 노드 수 등에 따라 최적화함으로써, 주어진 환경에서 보다 높은 PVFS의 성능을 기대할 수 있을 것으로 예상된다. 이에 향후 InfinBand, Myrinet 등 다양한 네트워크 미디어 상에서 FlowBuffer의 크기 변화에 따른 성능을 비교 분석해 볼 계획이다.

▌참 고 문 헌▐

[1] John M. May, "Parallel I/O for High Performance Computing," Morgan Kaufmann, 2000.
[2] W.B. Ligon III, and R.B.Ross, "Implementation and performance of a parallel file system for high per-formance distributed applications," Proc. of 5th IEEE International Symposium on High Performance Distributed Computing, pp 471~480, 1996.
[3] The Parallel Virtual File System Web site http://www.pvfs.org
[4] PVFS Distribution Design Notes, PVFS2 Development Team. 2004.
[5] PVFS-2.6.1 source codes, available on http://www. pvfs.org
[6] The IOR README File, http://www.llnl.gov/asci/platforms/purple/rfp/benchmarks/limited/ior/ior.mpiio.readme.html, Lawrence Livermore National Laboratory, Livermore, CA, 2001.