소규모 클러스터 시스템에서의 분산 파일 시스템에 대한 성능 평가

조혜영°, 차광호, 김성호
한국과학기술정보연구원 슈퍼컴퓨팅센터
e-mail:{chohy°, khocha, sungho}@kisti.re.kr

Performance evaluation of distributed file systems on a small scale cluster system

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

요 약

고속 네트워크로 연결된 대형 병렬 컴퓨터 및 클러스터 시스템의 사용이 증가되면서, 대용량 스토리지의 효율적인 활용을 위한 분산 및 병렬 파일 시스템에 대한 관심이 증가하고 있다. 특히 다수의 컴퓨터에 장착된 디스크 또는 스토리지를 네트워크로 연결하여 하나의 논리적이 파일 시스템으로 구성하는 분산 및 병렬 파일 시스템은 유휴 자원의 활용, bandwidth 및 throughput의 증대라는 장점으로 많은 연구가 진행 중이다. 본 논문에서는 대표적인 분산 및 병렬 파일 시스템을 대상으로 소규모 클러스터 시스템에서 성능 및 특징을 비교, 분석하였다.

1. 서론

네트워크의 속도가 빨라지고, 고성능 마이크로프로세서의 기술이 발전하면서, 대형 병렬 컴퓨터 및 클러스터 시스템에 대한 관심이 증가하고 있다. 이에 따라 스토리지가 대용량화되고, 대용량 스토리지를 효율적으로 사용하기 위한 파일 시스템에 대한 연구가 활발히 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.

이에 본 논문에서는 대표적인 분산 및 병렬 파일 시스템이라 할 수 있는 PVFS(Parallel Virtual File System), PVFS2, Lustre 및 GFS(Global File System)를 대상으로 소규모 클러스터에서 성능 및 특징을 비교, 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서는 성능 측정에 사용된 4종류의 대표적인 분산 및 병렬 파일 시스템에 대하여 간략히 설명하고, 3장에서는 파일 시스템의 성능 측정을 위한 실험 환경을 설명하였다. 그리고 4장에서는 실험에서 얻은 결과를 분석하였고, 5장에서는 결론에 대하여 기술한다.

2. 병렬 및 분산 파일 시스템

본 장에서는 실험에 사용된 PVFS, PVFS2, Lustre 및

GFS에 대하여 설명한다. 이 4종류의 파일 시스템은 분산 및 병렬 파일 시스템이라는 특징을 가지고 있으며, 메타데이터를 처리하는 메타데이터서버, 데이터 관리를 맡아 서버, 그리고 I/O를 요구하는 클라이언트 등의 공통된 구성 요소를 포함하고 있다. 표 1에 파일 시스템별 구성 요소를 간단히 정리해서 나타내었다.

표 1. 파일 시스템별 구성 요소

File System	PVFS	PVFS2	Lustre	GFS
메타데이터 처리	mgr	pvfs2_server	MDS	GNBD_CCA
데이터 관리	iod	pvfs2_server	OST	GNBD
클라이언트	pvfs_client	pvfs2_client	luster_client	client

2.1 PVFS (Parallel Virtual File System)

파일 시스템의 성능을 높이기 위하여 병렬 파일 시스템이 취하는 전형적인 방식이 RAID 0처럼 파일을 쪼개서(stripe) 서로 다른 저장장치에 저장하는 방식이다. Clemson 대학에서 개발된 PVFS 역시 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].

2.2 PVFS2 (Parallel Virtual File System ver. 2)

PVFS의 기능에 추가하여 설치의 편리성, 이질적인 클러스터 시스템 지원 및 스토리지 및 네트워크를 위한 모듈화 기능 강화 등에 바탕을 두어 개발된 것이 PVFS 버전 2이다. PVFS가 TCP/IP위주의 프로토콜을 사용하는 반면, PVFS2는 클러스터 시스템용 고성능 네트워크인 Myrinet과 In-finiband를 위한 프로토콜의 지원도 포함하고 있다[4,5].

2.3 Lustre

클러스터 시스템의 규모가 급속도로 증가하면서 이를 지원하기 위한 파일 시스템 또한 필요하게 되었는데 이러한 요구사항을 반영하여 개발된 파일 시스템이 lustre이다. 대규모(약 1만)의 계산노드에 대한 서비스와 페타바이트 규모의 스토리지를 구성할 수 있도록 설계되었으며 가장 큰 특징은 객체 기반의 파일 시스템(object-based cluster file system)이라는 점이다[6,7]. 기존의 파일 시스템과는 달리 데이터의 저장 단위가 객체이며 이를 저수준 파일 시스템에서 지원하기 위하여 OSD(Object Storage Device)라는 개념을 도입하였다.

2.4 GFS(Global File System)

GFS은 SAN(Storage Area Network) 환경에 사용되는 Red Hat에서 제공하는 클러스터 파일 시스템이자 볼륨 관리자이다. GFS는 오픈 소스이며, POSIX의 표준을 따르고 있다. 현재 레드햇이 지원하는 모든 서버와 스토리지 플랫폼에서 실행가능하다. GFS는 분산된 메타데이터와 클러스터에서 최적화된 수행을 위한 병렬 저널링 기능을 제공한다[8,9].

3. 성능 측정

본 논문에서는 그림 1과 같은 소규모 클러스터 시스템 환경에서 파일 시스템의 성능을 측정하였다. 각 노드는 Intel Pentium 4 XEON 2.8C CPU 1개, 1GB의 메모리, 80GB STAT 하드디스크, FastEthernet NIC 1개, GigabitEthernet NIC 1개를 갖추었다. OS는 Linux (Kernel 2.4.21)을 사용하였으며, PVFS(version 1.6.3), PVFS2 (version 1.0.1), Lustre(version 1.0.4), GFS(version 6.0.0)에 대하여 성능을 측정하였다. 그림과 같이 9개의 노드 중 4개를 데이터를 관리하는 서버, 1개를 메타데이터를 처리하는 서버로 구성하고, 나머지 노드가 클라이언트 역할을 하도록 구성하여 성능을 측정하였다. 파일 시스템의 성능 측정을 위한 벤치마크 프로그램으로는 대표적으로 가장 많이 사용되고 있는 대표적으로 가장 널리 사용되고 있는 Bonnie[10] 와 IOzone[11]을 사용하였다.

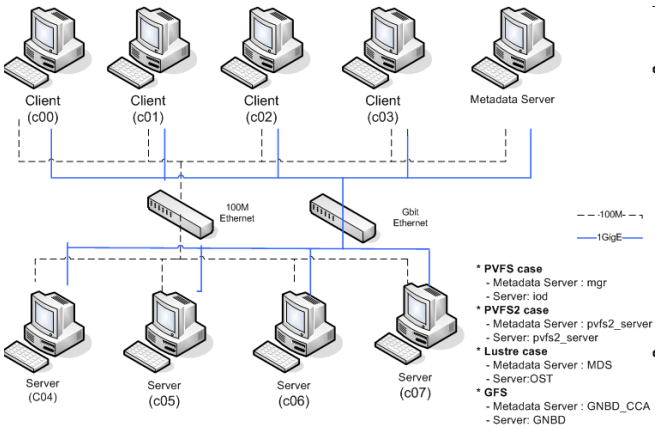

그림 1. 테스트 환경

4. 실험 결과

그림 2는 4종류의 파일 시스템에 대해서 FastEthernet으로 시스템을 구성하고, bonnie를 이용하여 측정한 성능을 보여준다. Bonnie에서 테스트 파일의 크기는 100MB로 하여 측정하였다. Read(block)은 GFS, Lustre, PVFS, PVFS2의 순으로 성능이 좋게 나왔으며, 성능의 차이도 GFS가 Lustre의 2배, Lustre는 PVFS에 66.5배 정도로 큰 차이를 보였으며, PVFS와 PVFS2의 성능은 비슷했다. Write(block)의 경우, PVFS, PVFS2, Lustre가 비슷한 성능을 보였으며, GFS가 다른 3개의 파일 시스템에 비해 우수한 성능을 보였다. Read(character)와 Write(character)의 경우, GFS가 가장 성능이 높게 측정되었으나 4가지 파일 시스템의 성능 차이는 상대적으로 적었다. PVFS와 PVFS2를 비교했을 때는 Read(Block)와 Write(Block)에서는 PVFS가 PVFS2보다 조금 높게 나왔으며, Read(character)와 Write(character)에서는 PVFS2가 PVFS보다 약간 높았다.

그림 3은 GigabitEthernet을 이용한 구성에서 측정된 성능을 보여준다. Lustre의 Read를 제외하고 GigabitEthernet를 이용한 경우 FastEthernet을 이용했을 때보다 3.3~5.3배의 성능 향상을 보여 주었다. 그리고 GFS에서는 GigabitEtherent와 FastEthernet을 이용했을 때 성능의 차이가 적었다.

그림 4는 IOzone을 이용한 성능 측정 결과를 보여준다. 전반적으로 IOzone의 테스트 결과도 Bonnie의 테스트 결과와 마찬가지로 GFS가 대부분 우수한 성능을 보여 주었고, Lustre, PVFS, PVFS2 순의 결과를 보여주었다. 이 때 PVFS, PVFS2는 거의 비슷한 성능을 보였다. Bonnie 테스트시 사용된 파일의 크기가 100MB이라는 점을 고려하면 수치적으로도 Bonnie와 IOzone의 결과가 비슷함을 알 수 있다.

5. 결론

본 논문에서는 대표적인 분산 및 병렬 파일 시스템인 PVFS, PVFS2, Lustre 및 GFS의 성능을 소규모 클러스터에서 비교, 분석하였다. 본 실험 결과에서는 전반적으로 GFS가 우수한 성능을 보였다. 그러나 본 실험에서는 보편화되어 있는 벤치마크 프로그램을 사용하여, 안전성을 측정하기에는 어려움이 있었고 확장성을 고려한 실험에도 제약이 있었다. 즉, 본 실험으로 대규모 클러스터 시스템에서 각 파일 시스템의 정확한 성능을 예측하기에는 미흡한 점이 있어, 이러한 환경을 반영할 수 있는 보다 정확한 성능 분석을 계획 중이다.

참고문헌

[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.parl.clemson.edu/pvfs/desc.html

[4] Rob Latham, Neill Miller, Robert Ross, and Phil Carns, "A Next-Generation Parallel File System for Linux Clusters," LinuxWorld, pp 56~59, Jan. 2004

[5] Parallel Virtual File System 2 Web site, http://www.pvfs.org/pvfs2/

[6] Lustre Web site, http://www.lustre.org/

[7] Richard Hedges, Bill Loewe, Tyce McLarty, and Chris Morrone, "Parallel File System Testing for the Lunatic Fringe: The Care and Feeding of Restless I/O Power Users," Proc. of 22nd IEEE / 13th NASA Goddard Conference on Mass Storage Systems and Technologies, pp 3 ~ 17, 2005

[8] Red Hat GFS 6.0 Administrator's Guide, http://www.redhat.com

[9] Red Hat GFS Web site, http://www.redhat.com/software/rha/gfs/

[10] Bonnie Web site, http://www.textuality.com/bonnie/

[11] IOzone Filesystem Benchmark, http://www.iozone.org

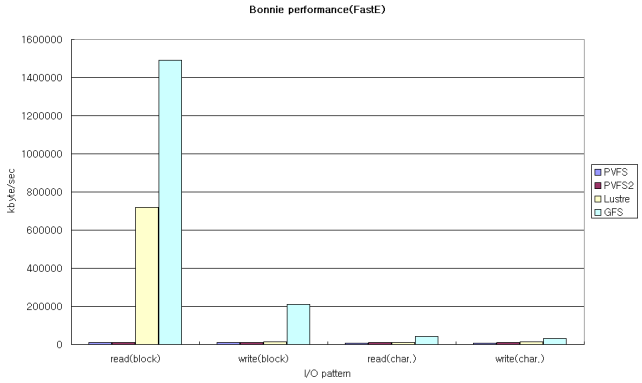

그림 2. Bonnie 테스트 결과(FastEthernet)

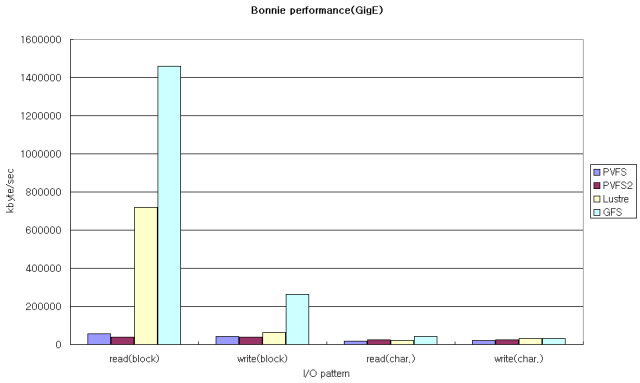

그림 3. Bonnie 테스트 결과(GigabitEthernet)

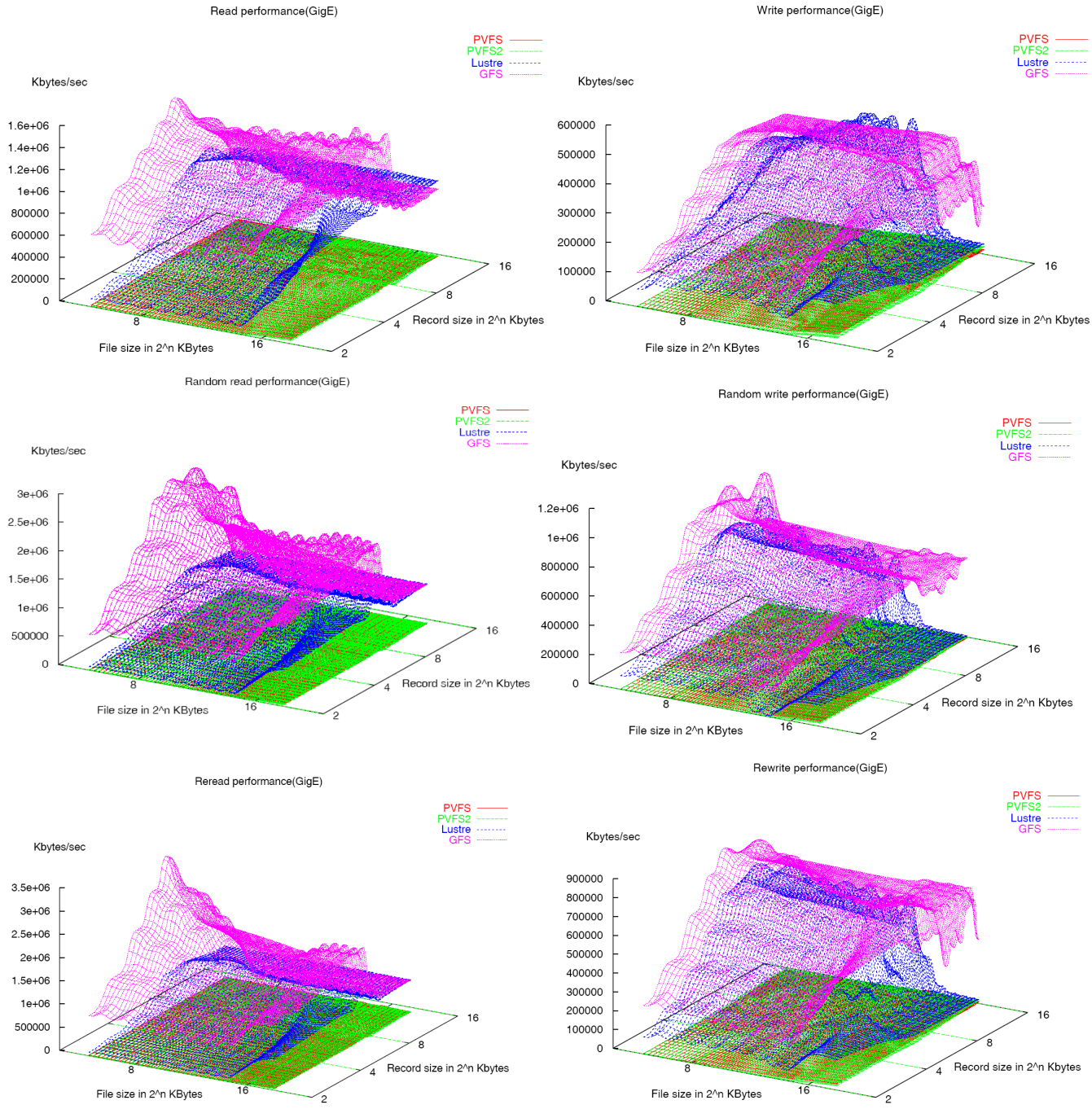

그림 4. IOzone 테스트 결과(GigabitEthernet)