병렬 파일 시스템의 구성 변경에 대한 성능 평가

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

Performance evaluation of parallel file systems

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

요 약

네트워킹 기술의 발달은 파일 시스템 개념에도 변화를 가져와 병렬 파일 시스템을 구성하는 기법을 고려하게 하였다. 다수의 컴퓨터에 장착된 디스크 또는 스토리지를 네트워크로 연결하여 하나의 논리적인 파일시스템으로 구성하는 기술로서 유휴 자원의 활용, I/O처리 대역폭의 증대라는 장점으로 많은 연구가 진행 중이다. 그러나 이러한 파일시스템을 운영하기 위해서는 기존 파일시스템이 갖는 특징 이외에 네트워크 특성, 각 노드들의 구성 방법 등을 추가로 고려하여야 한다. 본 논문에서는 대표적인 병렬 파일 시스템을 대상으로 네트워크 및 노드의 역할을 변경하면서 병렬 파일 시스템의 성능에 어떤 영향을 미치는가에 대하여 조사하고 분석하였다.

1. 서론

파일시스템의 성능을 개선하고자 하는 노력은 네트워킹 기법을 이용하여 다수의 디스크 내지는 스토리지를 연결하고 I/O 처리를 분산시키는 병렬 파일시스템의 개념을 만들어 내었다. 즉, 다수의 컴퓨터에 장착된 디스크 내지는 스토리지를 네트워크로 연결하여 하나의 논리적인 파일시스템으로 구성함으로써 유휴 자원의 활용, I/O 처리 대역폭 증대 등의 효과를 기대할 수 있어서 고성능 컴퓨팅 분야뿐만 아니라 대규모 데이터 처리를 위한 파일시스템으로 고려 되고 있다.

이러한 파일 시스템을 운영하고자 할 때, 고려 사항으로는 기존 파일시스템이 갖는 특징 이외에 네트워크 특성, 각 노드들의 구성 방법등을 생각하여야 한다. 본 논문에서는 대표적인 병렬 파일 시스템을 대상으로 네트워크 및 노드의 역할을 변경하면서 병렬 파일 시스템에 어떤 영향을 미치는가에 대하여 조사하고 분석하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 성능 측정에 사용된 대표적인 병렬 파일시스템 대하여 간략히 설명하고, 3 장에서는 실험 환경을, 4 장에서는 실험 결과, 5 장에서는 결론에 대하여 기술한다.

2. 병렬 파일 시스템

본 장에서는 실험에 사용된 3 종류의 파일시스템에 대하여 설명한다. 병렬 파일 시스템이라는 특징을 가지고 있으며 특히 표 1 과 같이 공통된 구성요소를 포함하고 있다.

[표 1] 병렬 파일 시스템 별 구성 요소

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

2.1 PVFS (Parallel Virtual File System)

파일시스템의 성능을 높이기 위하여 병렬 파일시스템이 취하는 전형적인 방식이 RAID 0 처럼 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].

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 2 이다. PVFS가 TCP/IP위주의 프로토콜을 사용하는 반면, PVFS2는 클러스터 시스템용 고성능 네트워크인 Myrinet과 Infiniand를 위한 프로토콜의 지원도 포함하고 있다[3,4].

2.3 lustre

클러스터 시스템의 규모가 급속도로 증가하면서 이를 지원하기 위한 파일시스템 또한 필요하게 되었는데 이러한 요

구사항을 반영하여 개발된 파일시스템이 lustre이다. 대규모(약 1 만)의 계산노드에 대한 서비스와 페타바이트 규모의 스토리지를 구성할수 있도록 설계되었으며 가장 큰 특징은 객체 기반의 파일시스템(object-based cluster file system)이라는 점이다[5,6]. 기존의 파일시스템과는 달리 데이터의 저장 단위가 객체이며 이를 저수준 파일시스템에서 지원하기 위하여 OSD(Object Storage Device)라는 개념을 도입하였다.

3. 성능 측정

본 실험의 목적은 각 파일시스템의 구성요소의 변화가 성능에 미치는 정도를 파악하는데 있다. 첫 번째로 수행한 실험은 네트워크의 변화에 따른 성능 변화 정도를 측정하였는데 FastEthernet 과 GigabitEthernet 으로 구성한 파일시스템의 성능 차이를 비교하였다.

두번째 실험은 메타데이터 서버의 단독 수행 여부가 성능에 미치는 정도를 파악한다. 각 파일시스템은 메타데이터의 처리를 위한 프로세스를 보유하고 있는데 이는 별도의 서버에서 운영될 수도 있고 데이터 저장 및 관리를 담당하는 서버에 통합되어 운영될 수도 있다. 이러한 두 경우의 성능을 측정하였다.

실험에서 하드웨어는 인텔 기반의 아키텍쳐를 사용하였고, 4 대의 노드에 부착된 SATA 방식 디스크들을 하나의 논리 디스크로 구성하였다. 성능 측정을 위한 벤치마크 프로그램으로는 Bonnie[7]와 IOzone[8]을 사용하였다.

4. 실험 결과

4.1 네트워크 변경 테스트

그림 1 은 네트워크의 변경에 대한 IOzone 을 이용한 테스트 결과를 보여주고 있다. PVFS 와 PVFS2 는 비슷한 결과를 보여주었다. PVFS 의 경우에는 그림에서 보는 바와 같이 확연한 성능 변화가 있음을 확인할 수 있었다. Lustre 의 경우에는 Read 의 경우 특별한 성능 차를 확인하기 힘들었으나 Write 의 경우 파일의 크기가 증가하면서 Gigabit Ethernet 으로 구성한 경우가 좀더 좋은 성능을 보여주었다.

4.2 메타 데이터 서버 변경 테스트

그림 2 는 메타데이터 서버를 독립적으로 구성한 경우와 일부 구성 요소와 혼용한 경우의 성능을 IOzone 을 이용하여 테스트한 결과이다. 그림에서와 같이 일부 특수한 조합에서 성능의 차가 존재하기는 하나 전반적으로 큰 차이를 발견하기 힘든 상황이라고 할 수 있다.

4.3 Bonnie 테스트

표 2 는 Bonnie를 이용하여 테스트한 결과를 보여두고 있다. 테스트시 사용된 파일의 크기가 100MB이라는 점을 고려하면 IOzone과 같은 결과를 보여준다고 할 수 있다.

5. 결론

본 논문을 통하여 병렬파일시스템을 구성하는데 네트워크와 메타데이타 서버의 위치등이 시스템 성능에 미치는 영향을 분석하였다. Bonnie 를 이용한 결과를 살펴 보면 Lustre 의 Read 를 제외하고 3 개의 파일시스템에서 GigabitEthernet 을 사용한 경우 FastEthernet 을 이용했을 때보다 3.3~5.3 배의 성능향상을 보여 주었다.

메타데이터 서버의 독립 운영 여부는 -9% ~ 13%의 성능 차이를 보였고 IOzone 을 이용한 테스트 결과 역시 파일시스템의 성능 향상에 크게 영향을 미치지 못한다는 사실을 확인할 수 있었다

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 performance distributed applications," Proc. of 5th IEEE International Symposium on High Performance Distributed Computing, pp 471 ~ 480, 1996

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

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

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

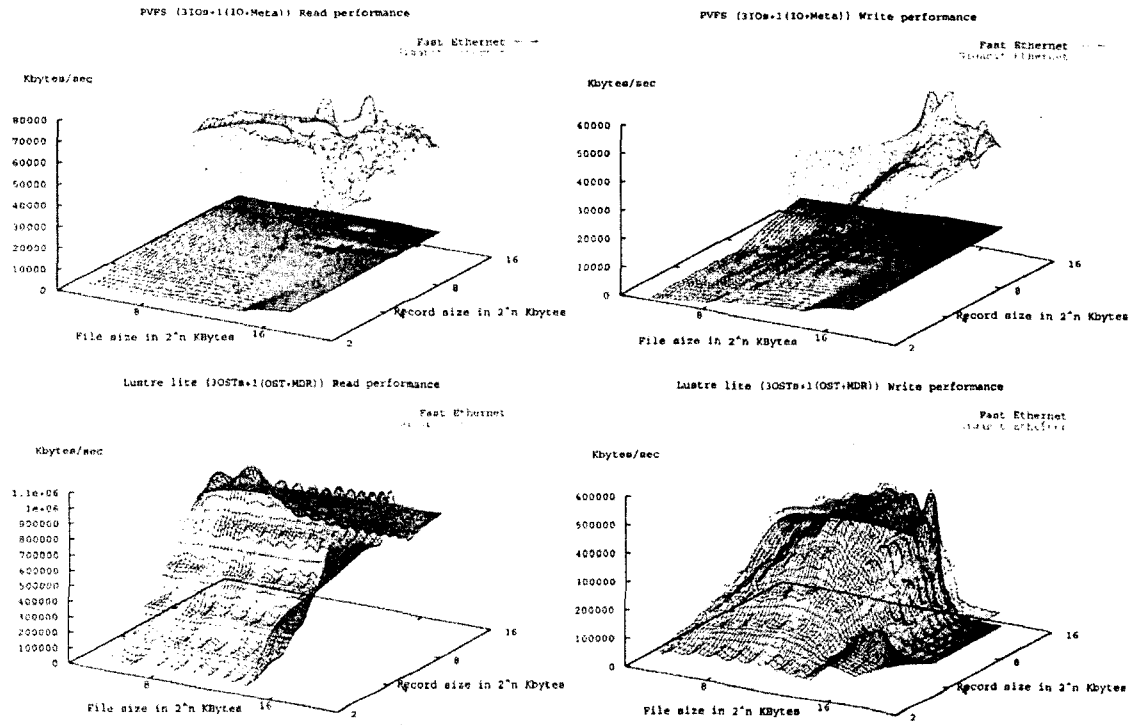
[6] Richard Hedges, Bill Loewe, Tyce McLarty, and Chris Morrone, "Parallel File System Testing for the Lunatic Fringe: The Care and Feeding of Restless I/O Power Users," Proc. of 22nd IEEE / 13th NASA Goddard Conference on Mass Storage Systems and Technologies, pp 3 ~ 17, 2005

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

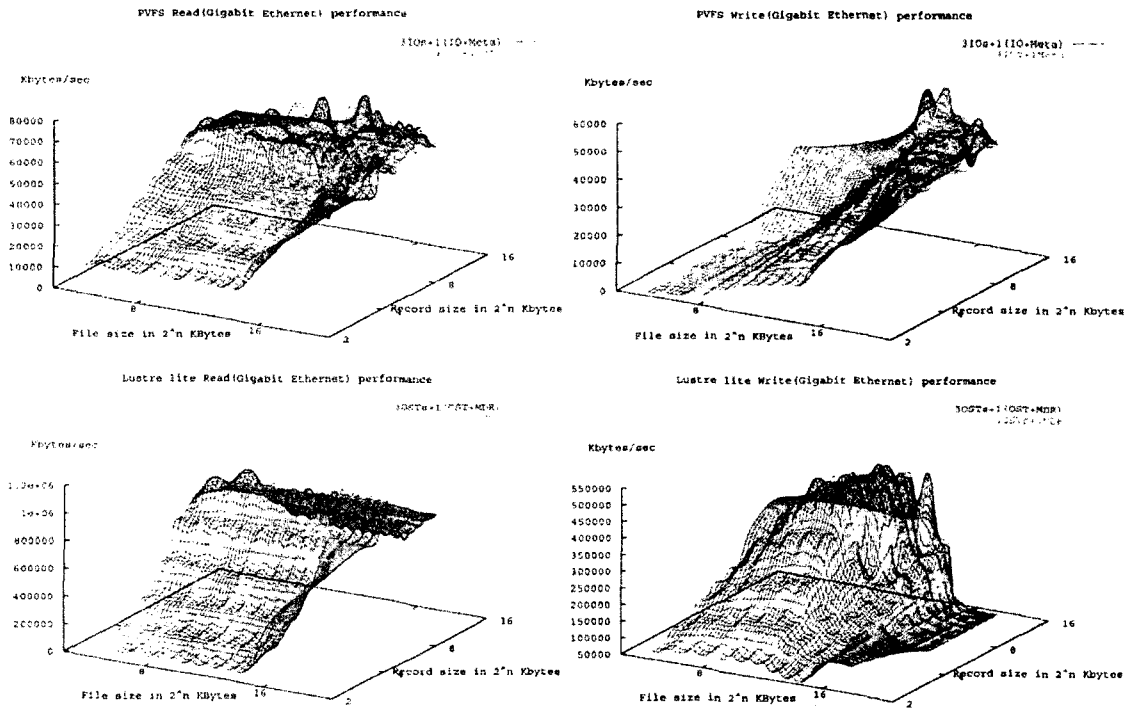
[8] IOzone Filesystem Benchmark, http://www.iozone.org/

[표 2] Bonnie 테스트 결과: 생성 파일 크기 100 MB, 단위 kB/s

	PVFS				PVFS2				lustre			
	3IOs+1(IO+Meta)		4IOs+1Meta		3IOs+1(IO+Meta)		4IOs+1Meta		3OSTs+1(OST+MDR)		4OSTs+1MDR	
	FastE	GigE	FastE	GigE	FastE	GigE	FastE	GigE	FastE	GigE	FastE	GigE
Write(Char.)	8668	20901	8647	21046	8939	23953	8946	23289	12331	31170	12533	31874
Write(Block)	10915	42466	10411	43420	10501	37822	9630	37690	12613	61609	12455	63059
Read(Char.)	8649	21032	8640	18640	8990	25480	9061	26297	10382	22158	10342	22063
Read(Block)	11013	50585	10795	57322	11107	37523	10445	37990	702043	707739	717624	718012



[그림 1] 네트워크 변경에 대한 IOzone 결과



[그림 2] 메타데이터 서버의 위치 변화에 대한 IOzone 결과