단일 디스크 입출력을 위한 커널 모듈 프로토타입의 설계 및 구현

황인철° 김동환 김호진 맹승렬 조정완
한국과학기술원 전자전산학과 전산학전공
{ichwang°, dhkim, hojin, maeng, jwcho}@calab.kaist.ac.kr

Design and Implementation of the Kernel Module Prototype for the Single Disk I/O

In-chul Hwang° Dong-hwan Kim Hojin Ghim Seung-Ryoul Maeng Jung-Wan Cho
Division of Computer Science, Dept. of Electrical Engineering & Computer Science, KAIST

요 약

요즘 값싼 PC를 빠른 네트웍으로 묶어 높은 성능을 얻고자하는 클러스터 컴퓨팅에 대한 연구가 활발히 이루어지고 있다. 이러한 연구 중 클러스터 컴퓨팅 환경을 구성하는 여러 컴퓨터들을 하나의 컴퓨터처럼 보이게 해주는 단일 시스템 이미지 서비스는 사용자에게 쓰기 편리한 환경과 높은 가용성을 제공하여 준다. 단일 시스템 이미지 서비스를 제공하기 위해서는 단일 프로세스 공간, 단일 메모리 공간 및 단일 디스크 입출력을 제공하여 주어야 한다.

단일 디스크 입출력은 여러 컴퓨터에 나누어져 있는 디스크들을 하나의 큰 디스크로 보여주고 여러 디스크들을 효율적으로 사용할 수 있도록 서비스를 제공한다. 이러한 단일 디스크 입출력을 사용자 수준이나 파일 시스템 수준에서 제공하여 주는 것은 성능 측면이나 투명성의 측면에서 커널 모듈로 제공하여 주는 것 보다 좋지 않다. 따라서 본 논문에서는 단일 디스크 입출력을 위하여 커널 모듈 프로토타입을 설계하고 구현한다. 그리고 네트웍 파일 시스템과 단일 디스크를 이용하여 단일 디스크 입출력을 위한 커널 모듈의 성능과 비교, 분석한다.

1. 서론

요즘 값싼 PC들을 빠른 네트웍으로 묶어 높은 성능을 얻고자하는 클러스터 컴퓨팅에 대한 연구가 활발히 이루어지고 있다. 이러한 연구 중 클러스터를 구성하는 여러 컴퓨터들을 하나의 컴퓨터처럼 보이게 해주는 단일 시스템 이미지 서비스는 사용자에게 쓰기 편리한 환경과 높은 가용성을 제공하여 준다. 단일 시스템 이미지 서비스를 제공하기 위해서는 단일 프로세스 공간, 단일 메모리 공간 및 단일 디스크 입출력이 제공되어야 한다.

단일 시스템 이미지 서비스 중 단일 디스크 입출력은 분산된 디스크들을 하나의 큰 디스크로 보여주고 이들을 효율적으로 관리할 수 있는 방법이다. 단일 디스크의 입출력은 크게 세 가지 수준에서 제공하여 줄 수 있다. 첫째는 사용자 수준의 라이브러리를 이용하여 사용자들이 사용하는 디스크를 하나의 디스크인 것처럼 보이게 해주는 방법이 있다. 이 방법은 사용자 수준에서 구현되기 때문에 구현의 용이성이 있으나 사용자가 직접 라이브러리를 이용하여 프로그램을 재구성하여야 하고 사용자 수준에서 구현되기 때문에 추가적 부하가 많다는 단점이 있다. 둘째는 파일 시스템 수준에서 단일 디스크 입출력을 제공하여 주는 방법으로 이 방법은 사용자에게 투명성을 제공해줄 수 있고 성능도 좋다는 장점이 있지만 여러 디스크를 파일 시스템 수준에서 관리하여 주어야 하기 때문에 파일 시스템 제작에 많은 비용이 든다는 단점이 있다. 마지막으로 커널 모듈에서 단일 디스크 입출력을 제공하여 주는 방법으로 파일 시스템의 수정이나 사용자 프로그램 수정 없이 투명성 있게 단일 디스크 입출력을 제공하여 주고 성능도 좋은 장점이 있다. 따라서 본 논문에서는 파일 시스템이나 사용자에게 단일 디스크 입출력을 투명하게 제공해 줄 수 있는 단일 디스크 입출력을 위한 커널 모듈 프로토타입의 설계와 구현에 대하여 기술한다. 그리고 커널 모듈 프로토타입의 성능을 기존 NFS와 단일 디스크를 사용할 때와 성능을 비교한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서 기존 단일 디스크 입출력에 대한 연구인 Petal과 SIOS에 대하여 살펴본다. 3장에서는 단일 디스크 입출력을 위한 커널 모듈 프로토타입의 설계와 구현에 대하여 설명한다. 4장에서는 단일 디스크 및 NFS, 그리고 단일 디스크 입출력을 위한 커널 모듈 프로토타입의 성능을 측정하고 비교 분석한다. 5장에서는 향후 연구 방향과 결론을 맺는다.

2. 관련 연구

2.1 Petal

Petal[1]은 DEC(Digital Equipment Cooperation)에서 제작된 클러스터 환경을 위한 자료 저장 시스템이다. Petal은 여러 개의 디스크를 하나의 가상 디스크로 보여주는 역할을 한다. Petal에 의해 생성된 가상 디스크는 실제 디스크와 같은 API를 제공하므로 기존 응용 프로그램이나 파일 시스템에서 투명하게 사용될 수 있다.

Petal에서는 블록의 저장 위치를 전체 지도를 두고서 사상을 시킨다. 따라서 저장 위치가 바뀔 때 마다 전체 지도를 고쳐서 쉽게 블록을 찾을 수 있는 기법을 제공한다. 그리고 Chained Declustering이라는 기법을 이용하여 효율적인 고장 허용을 지원한다.

2.2 SIOS(Single I/O Space)

SIOS[2]는 USC(University of Southern California)에

서 개발된 것으로 분산 RAID[3]를 구성하는데 그 목적을 두었다. SIOS에서는 클러스터 내에 모든 블록에 대하여 단일 주소 공간을 제공한다. 그리고 RAID형태의 데이터 분산을 통하여 높은 가용성과 고장 허용, 그리고 높은 대역폭을 제공한다. SIOS에서는 가용성을 위하여 RAID-0, RAID-1, RAID-5 및 RAID-x와 같은 다양한 종류의 RAID 수준을 제공한다.

3. 단일 디스크 입출력을 위한 커널 모듈 프로토타입의 설계 및 구현

3.1 Overview

다음 그림 1은 단일 디스크 입출력(SDIO : Single Disk I/O)을 지원하는 커널 모듈을 적재한 클러스터의 구조를 나타낸다.

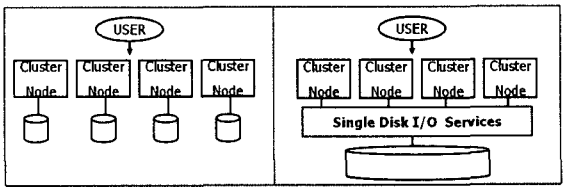

그림 1. 단일 디스크 입출력을 지원하는 커널 모듈의 적재한 클러스터의 구조

기존 각 노드의 디스크를 따로 사용하던 기존의 방법과 달리 단일 디스크 입출력을 지원하는 커널 모듈을 적재한 클러스터는 어느 노드에서나 모든 디스크가 하나의 디스크처럼 보이게 해 준다. 따라서 사용자와 파일 시스템 모두가 다른 노드의 디스크들도 자신의 노드에 있는 디스크인 것처럼 사용할 수 있는 투명한 디스크 인터페이스를 제공할 수 있게 된다.

3.2 단일 디스크 입출력을 지원하는 커널 모듈 프로토타입의 설계

다음 그림 2는 단일 디스크 입출력을 지원하는 커널 모듈 프로토타입의 구조를 나타낸다.

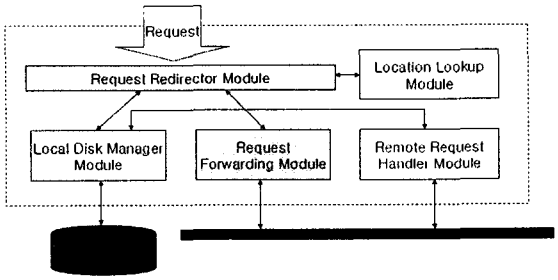

그림 2. 단일 디스크 입출력을 지원하는 커널 모듈 프로토타입의 구조

그림에서 각 모듈은 다음과 같은 역할을 수행한다.

· 요청 분산 모듈(Request Redirector Module) : 커널 모듈에 요청이 들어오면 이 모듈에서는 요청된 블록의 위치를 위치 찾기 모듈에 요청하여 위치를 찾은 후 위치에 따라서 자신의 디스크에서 처리할 수 있는 요구이면 지역 디스크 관리자에게 전달하고 다른 노드의 디스크에서 처리해야 되는 요구이면 요청 전달 모듈에게 전달하여 요구를 처리하는 역할을 한다.

· 위치 찾기 모듈(Location Lookup Module) : 요청된 블록의 위치를 알려주는 역할을 한다.

· 지역 디스크 관리자 모듈(Local Disk Manager Module) : 자신의 디스크에서 처리할 수 있는 요구를 디스크에서 처리하는 역할을 수행한다.

· 요청 전달 모듈(Request Forwarding Module) : 다른 디스크에 있는 블록에 대한 요청일 경우 다른 노드의 요청 처리 모듈에게 블록에 대한 요청을 전달하여 요청을 처리하는 역할을 수행한다.

· 외부 요청 처리 모듈(Remote Request Handler Module) : 자신의 디스크의 블록을 외부 다른 노드에서 요청되어 왔을 때 지역 디스크 관리자 모듈을 통하여 처리해 주는 역할을 수행한다.

3.3 단일 디스크 입출력을 지원하는 커널 모듈 프로토타입의 구현

모든 모듈은 각자 인터페이스의 호출을 통하여 통신을 한다. 하나의 요청이 처리된 후 그 다음 요청을 처리한다.

요청 전달 모듈과 외부 요청 처리 모듈은 TCP 소켓을 이용하여 통신을 구현하였고, 데이터 분산은 RAID-0과 같은 데이터 분산을 통하여 사상 지도 필요 없이 간단한 계산으로 블록의 저장 장소를 찾기 쉽게 하였다.

외부 요청 처리 모듈은 요청 처리 하는 다른 모듈들과 달리 하나의 독립적인 커널 쓰레드[8]를 이용하여 구현하였다. 따라서 처음 커널 모듈이 적재될 때 독립적인 쓰레드를 만들어서 외부 요청이 있는지 살펴보고 외부 요청이 왔을 때 순차적으로 처리해 주게 구현하였다.

지역 디스크 관리자 모듈은 실제 디스크 블록 주소가 넘어오면 디스크에서 디스크 블록을 읽거나 쓰는 역할을 수행한다.

4. 성능 평가

성능 평가를 위해 사용된 환경은 다음 표 1과 같다.

CPU	Pentium IV 1.8GHz
Memory	512MByte 266MHz DDR Memory
Disk	IBM 60G 7200rpm
Network	3c996B-T(Gigabit Ethernet) 3c17701-ME (24port Gigabit Ethernet Switch)
OS	Linux 2.4.18
NFS	Version 2

표 1. 성능 평가 환경

단일 디스크와 단일 디스크 입출력을 위한 커널 모듈 프로토타입에 EXT2 [5] 파일 시스템을 적재한 시스템과 NFS [4]를 IOzone 벤치마크 [6]와 Andrew 벤치마크 [7]을 이용하여 성능을 측정하고 비교하였다. 4개의 노드를 이용하여 단일 디스크 입출력을 위한 커널 모듈을 적재하여 단일 디스크 입출력을 수행하였다. 그리고

NFS는 하나의 서버와 하나의 클라이언트에서 수행하였고 EXT2는 한 노드에서 단일 디스크 입출력을 위한 커널 모듈과 디스크에서 그 성능을 측정하였다.

4.1 IOzone 벤치마크 성능

IOzone 벤치마크는 단일 파일 시스템의 성능을 측정하기위한 벤치마크이다. 2MB의 파일을 128KB씩 읽기/쓰기를 하여 성능을 측정한 결과는 다음 그림 3과 같다.

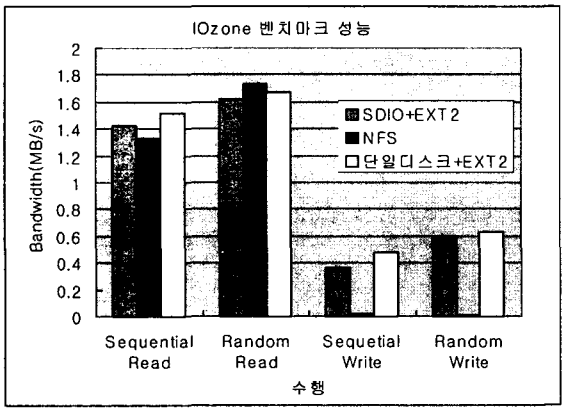

그림 3. IOzone 벤치마크 성능

그림에서와 같이 단일 디스크 입출력을 위한 커널 모듈 프로토타입을 사용하는 것은 단일디스크를 사용하는 것보다 커널 모듈을 통해 자신의 혹은 다른 노드의 디스크로 쓰기/읽기를 하기 때문에 단일 디스크를 이용하는 것보다 성능이 좋지 않다. 그렇지만 네트웍을 이용하는 NFS와 읽기 성능은 비슷하고 쓰기 성능의 경우는 월등히 좋음을 알 수 있다.

4.2 Andrew 벤치마크 성능

다음 표 2는 단일 디스크와 단일 디스크 입출력을 위한 커널 모듈 프로토타입을 이용하였을 때 EXT2와 NFS의 Andrew 벤치마크 성능을 나타낸다.

	SDIO + EXT2	단일 디스크 + EXT2	NFS
phase1(create directory)	0.022sec	0.018sec	0.055sec
phase2(copy files)	0.240sec	0.154sec	0.284sec
phase3(recursive dir stats)	1.069sec	1.058sec	1.963sec
phase4(scanning each file)	1.560sec	1.996sec	2.310sec
phase5(compilation)	8.458sec	1.413sec	0.949sec

표 2. Andrew 벤치마크 성능

읽기가 주로 많은 1~4단계에서는 세 개의 성능 차이가 크지 않다. 하지만 쓰기가 많은 컴파일 단계(5단계)에서는 단일 디스크보다 단일 디스크 입출력을 지원하는 프로토타입 커널 모듈을 이용하는 것이 훨씬 나쁜 성능을 나타낸다. 그리고 NFS는 읽기가 많은 단계에서는 단일 디스크나 단일 디스크 입출력을 지원하는 커널 모듈 프로토타입에 대해 나쁜 성능을 나타내지만 쓰기가 많은 단계에서는 더 좋은 성능을 나타낸다. 앞으로 성능에 대하여 자세히 분석하고 단일 디스크 입출력을 위한 커널 모듈의 성능 개선을 통하여 단일 디스크와 거의 유사한 읽기 성능 및 NFS와 같은 쓰기 성능을 얻을 것이다.

5. 향후 연구 방향 및 결론

값싼 PC를 빠른 네트웍으로 묶어 높은 성능을 얻고자 하는 클러스터 시스템에서 여러 노드들을 단일한 노드처럼 보여주는 단일 이미지 시스템에 대한 연구가 필요하다. 단일 이미지 시스템은 사용자가 하나의 노드처럼 여러 노드들이 모인 클러스터를 이용할 수 있게 함으로서 사용자의 편리성을 제공하고 하나의 노드가 실패되었을 때 다른 클러스터 노드들에게 영향을 미치지 않고 복구를 할 수 있는 가용성 측면에서 연구가 되어지고 있다.

본 논문에서는 단일 시스템 이미지 서비스 중 여러 노드에 분산되어 있는 디스크들을 하나의 입출력 공간으로 제공하여 주는 단일 디스크 입출력을 위한 커널 모듈 프로토타입의 설계와 구현에 대하여 설명하였다. 여러 노드에 커널 모듈만을 로딩 시킴으로서 여러 노드의 디스크들을 마치 자신의 디스크를 사용하는 것처럼 파일 시스템과 사용자에게 투명한 인터페이스를 제공할 수 있다. 그리고 성능 평가를 통하여 단일 디스크를 사용하는 것 보다 많은 양의 데이터를 저장할 수 있지만 아직은 그 성능에서 단일 디스크에 비해 좋지 않은 성능을 나타내었다. 이는 구현된 커널 모듈에서 모든 디스크 입출력에 대한 요구를 직렬적으로 처리하기 때문으로 해석되며 앞으로 커널 모듈에서 병렬 입출력에 대한 기능을 지원할 예정이다. 또한 자세한 성능 분석을 통하여 단일 디스크 입출력에서 성능을 향상 시킬 수 있는 기능을 지원할 예정이다.

6. 참고 문헌

[1] Edward K. Lee, Chandramohan A. Thekkath, "Petal: Distributed Virtual Disks (1996)", Proceedings of the Seventh International Conference on Architectural Support for Programming Languages and Operating Systems 1996.

[2] Roy S.C.Ho, Kai Hwang, Hai Jin, "Design and Analysis of Clusters with Single I/O Space", ICDCS 2000

[3] Hai Jin, Toni Cortes, Rajkumar Buyya, "High Performance Mass Storage and Parallel I/O", IEEE Press

[4] "Linux NFS faq", http://nfs.sourceforge.net/

[5] EXT2, http://e2fsprogs.sourceforge.net/ext2.html

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

[7] J. H. Howard, M. L. Kazar, S. G. Menees, D. A. Nichols, M. Satyanarayanan, R. N. Sidebotham, and M. J. West, "Scale and performance in a distributed file system," *Transactions on Computer Systems*, vol. 6, pp. 51-81, February 1988.

[8] "Linux Kernel Threads in Device Drivers", http://www.scs.ch/~frey/linux/kernelthreads.html