SANfs-VM : 리눅스 클러스터 시스템을 위한 볼륨 관리 기법에 관한 연구

임승호[0] 황주영 박규호
한국과학기술원 전자전산학과 전기 및 전자공학 전공
{shlim[0], jyhwang, park}@core.kaist.ac.kr

SANfs-VM : volume management driver for linux cluster system

Seung-Ho Lim[0] Joo-Young Hwang, Kyu-Ho Park
EECS Dept. Korea Advanced Institute of Science and Technology

요 약

본 논문에서는 대용량 공유 파일시스템의 자원을 효율적으로 관리할 수 있는 볼륨 관리기에 대해서 제안하고 리눅스 상에 구현을 해 보았다. SANfs[5]는 Storage Area Network(SAN)의 대용량 저장장치를 지원할 수 있도록 제안하고 구현된 확장성 있는 공유 파일 시스템이다. SANfs의 자원을 효율적으로 이용하기 위해서 저장장치들을 효율적으로 관리할 수 있는 도구가 필요하게 되었고, 이 논문에서 SANfs의 구조에 적합한 볼륨 관리기인 SANfs-VM을 새롭게 제안하고 구현하였다. SANfs-VM은 SANfs의 엔터프라이즈 컴퓨팅을 위해서 다양한 레벨의 RAID, online resizing/reconfiguration 등의 기능을 제공함으로써 SANfs 저장장치의 확장성, 가용성을 향상시켰다. 또한 SANfs-VM은 저장 장치 시스템의 관리를 쉽게 함으로써 easy management 기능을 증진시켰다.

1. 서 론

폭발적인 인터넷 사용의 증가, e-business, 멀티미디어 데이터의 사용 증가는 저장장치 시스템의 대용량화와 함께 저장장치 시스템 관리의 중요성을 가져오게 되었다. 최근의 Storage Area Network(SAN) 기술의 발전은 저장 장치 시스템의 획기적인 변화를 가져왔으며 SAN을 이용한 공유 파일 시스템을 사용함으로써 이러한 대용량 시스템을 만들 수 있게 되었다. 그러나, 최대의 가용성, 효율성, 대용량의 서비스를 원하는 엔터프라이즈 유저들에게는 예전의 저장 장치 관리 기법으로는 한계가 있었고, 새로운 저장장치의 관리 기법이 필요하게 되었다. 특히, 디스크 제작 기술의 발전으로 저장장치의 성능이 좋아지게 되면서, 저장 장치 시스템의 관리는 사용자들에게 최대의 업타임, 최대의 가용성, 최적의 효율성, 확장성 및 공유성을 제공할 수 있도록 하는 기술을 필요로 하게 되었다.

볼륨 관리기는 이러한 저장장치 시스템을 관리하는 주요한 요소가 되었다. 볼륨 관리기는 물리적으로 독립인 여러 개의 저장요소를 묶어서 하나의 큰 논리 볼륨을 형성함으로써 사용자에게 많은 기능을 제공할 수 있다.

SANfs-VM은 예전에 우리가 구현한 SAN 환경의 공유 파일 시스템인 SANfs의 저장장치를 효율적으로 관리해 줄 수 있도록 제안하고 구현된 볼륨 관리기이다. 이것은 다양한 RAID level(0,1,5), online resizing/reconfiguration 등의 기능을 제공하고 easy management할 수 있도록 구현됨으로써 높은 확장성과 가용성을 제공할 수 있다.

2. 관련 연구

LinuxLVM[1]은 단일 리눅스 시스템에서 사용되는 볼륨 관리기이다. LinuxLVM은 단일 시스템에서는 online resizing/reconfiguration, snapshot 기능 등의 엔터프라이즈 유저를 위한 강력한 기능을 제공하지만 저장장치들을 여러개의 컴퓨터가 공유할 경우 사용될 수 없다는 단점이 있다.

클러스터 시스템의 볼륨관리기로는 GFS의 Pool driver와 SANtopia의 SANtopia 볼륨관리기를 들 수 있다.

GFS의 Pool driver[2,3]는 GFS 공유 파일 시스템의 저장장치를 관리하기 위한 도구이다. GFS는 저장 데이터의 일관성을 유지하기 위해서 device level의 락(Dlock)을 관리하고 있다. 그렇기 때문에 Pool driver의 주요한 특징으로는 Dlock을 지원해주고 있다는 점이다. 그렇지만 Pool driver는 다양한 레벨의 RAID를 제공해 주지 못하고, online resing/reconfiguration을 지원하지 못한다.

SANtopia volume management(SVM)[4]은 SANtopia를 지원하기 위한 볼륨 관리기이다. SVM은 다양한 레벨의 RAID (0,1,5)를 지원하고, 엔터프라이즈 유저에게 유용한 online resizing/reconfiguration 기능을 지원한다.

3. SANfs 공유 파일 시스템

클러스터 시스템에서, 볼륨 관리기의 구조는 그 시스템의 구조에 따라서 많이 달라지게 된다. SANfs-VM은 우리가 예전에 제안하고 구현했던 SANfs 공유파일 시스템을 지원하기 위한 볼륨 관리기이기 때문에, 먼저 SANfs의 구조에 대해서 살펴봐야 할 것이다.

SANfs 공유 파일시스템[5]은 대규모 저장장치를 겨냥하여 탁월한 확장성을 갖도록 설계하였으며, Linux상에서 개발되었다. SANfs의 시스템 구성도는 다음 그림과 같이 나타낼 수 있다. SANfs는 하나의 메타서버와 다른 여러 개의 SANfs client들로 구성되어 있으며, 저장 장치 시스템과 SANfs client들이 SAN을 통해서 연결되어 있다. 저장장치들은 여러 가지 레벨의 configuraion을

가질 수 있으며, 이것들은 SANfs-VM을 통해서 하나의 전체적인 저장장치 공간을 형성하게 된다.

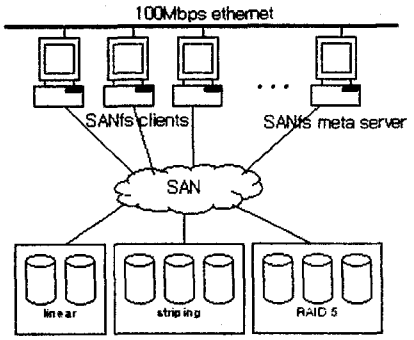

그림 1 : SANfs 시스템 구성도

SANfs의 주요한 특징으로는 다음과 같은 것들이 있다.

3.1 파일 데이터 일관성 유지

클러스터 상황에서 서로 다른 호스트간에 동일한 파일을 공유하면서 쓰는 경우는 거의 일어나지 않고, block 단위의 consistency를 지원하는 것은 복잡하고 불필요하기 때문에, SANfs에서는 파일 데이터의 일관성을 block lock이 아닌 file lock을 통해서 유지한다. file lock의 관리는 중앙의 메타 서버를 통해서 이루어진다. 메타서버는 유저 레벨 프로그램으로 구현되어 있으며, 각 호스트는 자신이 수정하고자 하는 파일을 열 때, 메타서버로부터 그 파일에 대한 권한을 얻은 뒤 해당 파일에 대한 작업을 수행한다. lock의 종류는 읽기 전용 lock과 읽기/쓰기 lock의 두 가지가 있다. 한 호스트에서 액세스한 파일에 대해서는 해당 호스트가 다시 액세스하게 되는 가능성이 높기 때문에, 한번 그 파일에 대한 lock을 얻게 되면 다른 호스트에 의해서 lock 요청이 이루어질 때까지 그 호스트가 계속 lock을 가지고 있는 callback locking 방식을 사용한다.

3.2 free space 관리

SANfs 파일시스템의 free space는 메타서버에서 집중관리하며, 각 호스트는 free space가 필요할 때마다 메타서버로부터 가져온다. 메타서버의 부담 및 통신량을 줄이기 위해서 메타서버에서 한번에 2,048개의 block chunk를 가져온 다음 local process들의 free block요청을 서비스한다. Free block들은 unmount시에 메타서버로 되돌린다.

4. SANfs 볼륨 관리기

SANfs의 주요한 특징들을 지원하면서 저장장치의 가용성, 확장성을 높이기 위한 SANfs-VM의 구조는 다음과 같다.

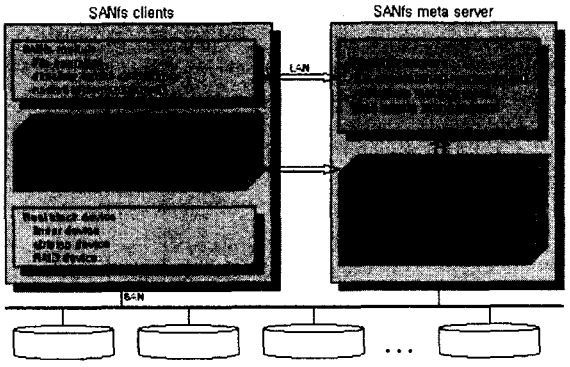

그림 2 : SANfs-VM block diagram

그림에서 보는 바와 같이 SANfs-VM은 메타서버에 볼륨 관리의 주요한 기능을 담당하는 볼륨관리 모듈을 두고, 각 SANfs 클라이언트에는 실제적은 논리 주소와 물리 주소를 매핑 시켜주는 블락 디바이스 드라이버인 logical volume driver를 두는 방식을 사용한다.

메타 서버의 볼륨 관리 모듈은 유저 레벨의 프로세스이며, 다음과 같은 작업을 수행한다. 첫째, 볼륨 관리 모듈은 시스템 관리자의 요청에 따라서 다양한 configuration에 대한 매핑 테이블을 만든다. 그리고 그 매핑 테이블과 configuration에 대한 정보를 각 클라이언트의 logical volume driver에게 알려줌으로써 글로벌한 논리 볼륨을 형성하는 작업을 수행한다. 둘째, 저장장치 시스템의 데이터를 한 곳에서 다른 곳으로 옮길 경우, 옮기는 데이터 블록에 대한 락을 관리하면서 시스템 데이터의 일관성을 유지시켜주고, 시스템의 가용성을 증가시켜준다. 셋째, 논리 볼륨의 크기를 늘이거나 줄일 경우, 메타 서버의 파일 락 서버와 연동하여 전체 논리 볼륨의 크기를 관리하는 역할을 한다. 이와 같은 작업은 시스템의 업타임 동안에 가능하도록 한다.

메타 서버의 볼륨 관리 모듈에 의해서 만들어진 매핑 테이블의 정보를 가지는 logical volume driver는 메타 서버로부터 전해 받은 매핑 테이블의 정보를 이용하여 SANfs가 요구한 논리 볼륨의 주소에 대한 실제적인 저장장치의 주소를 반환하는 일을 한다.

지금부터는 SANfs-VM의 특징들에 대해서 기술한다.

4.1. 다양한 configuration level

SANfs-VM은 다양한 레벨의 configuration을 동시에 지원할 수 있다. 논리 볼륨을 구성하는 가장 작은 단위는 저장장치의 파티션이다. 각 물리 볼륨은 자신이 속한 논리 볼륨의 정보를 디스크 파티션 테이블에 저장한다. 여러 개의 파티션을 이용해서 linear, striping, RAID5와 같은 configuration을 구성할 수 있다. RAID5를 지원하는 방식은, 공유 파일 시스템에서 parity 일관성을 유지시켜주기 위해서는 각 stripe에 대한 lock을 따로 관리를 해주어야 하는데, SANfs-VM시스템에서는 새롭게 제안한 selective stripe lock[6]이라는 방식을 사용해서 parity 일관성을 유지시켜줌으로써 성능의 향상을 보이고 있다. selective stripe locking은 stripe의 종류에 따라서 stripe이 하나의

파일에 속하거나 어떤 데이터도 할당되지 않은 trivial stripe과 stripe에 두 개 이상의 파일이 할당되거나 stripe의 일부만이 할당된 non trivial stripe으로 나누어서 관리를 한다. stripe을 이런 식으로 관리를 할 경우 trivial stripe에 대해서는 stripe lock을 요청할 필요가 없으므로 stripe lock의 횟수가 현저히 줄어들어 시스템의 성능을 향상시켜준다.

4.2 online resizing/reconfiguration

엔터프라이즈 컴퓨팅에서 가장 중요한 요소가 바로 시스템의 업타임이다. 즉 e-business, 멀티미디어 서비스 등을 위해서는 콘텐츠를 제공해주는 서버의 사용자는 저장장치 시스템의 크기를 온라인 상태에서 재조정하기를 원할 것이다. SANfs-VM에서는 다음과 같은 방식으로 온라인 리사이징을 지원해 준다. 시스템 관리자의 요청을 받아들인 볼륨 관리 모듈은 요청에 맞도록 configuration을 재조정해서 다시 만들어 낸다. 이렇게 바뀌어진 configuration 정보는 SANfs 클라이언트의 volume driver에게 전달된다. 디스크의 데이터 블락의 재조정이 필요할 경우, 볼륨관리 모듈은 클라이언트로부터 각 블락에 대한 lock을 얻어온 후 데이터의 이동, 매핑 테이블의 업데이트와 같은 작업을 수행한다. linear mapping같은 경우는 이전의 매핑 테이블에서 사이즈만 늘이면 되지만, RAID level은 매핑 테이블 뿐 아니라 데이터 블락의 재조정까지 이루어져야 하기 때문에 복잡한 일련의 과정들이 이루어져야 한다.

또 시스템을 사용하다 보면, 특정 시간에 , 특정 데이터에 대한 요청이 많아지거나, 데이터의 요청이 집중되어서 저장장치를 효율적으로 사용하지 못할 경우가 있다. 이럴 경우 데이터의 이동 또는 복사본을 두어서 시스템의 성능을 유지시켜줄 수가 있다. 이런 작업 역시 SANfs-VM의 볼륨 관리 모듈에서 담당해서 수행해 줄 수 있다.

이렇게 재조정된 볼륨에 대한 정보, 매핑 된 블락에 대한 정보, free space에 대한 정보는 파일 락 서버와 연동해서 SANfs 클라이언트에게 제공된다.

5. 실험 결과

SANfs 시스템에서 SANfs-VM의 성능을 테스트하는 것은 두가지 의미를 둘 수 있다. 첫째, SANfs 시스템에 볼륨 관리 모듈을 추가할 경우 볼륨 관리의 I/O request에 대한 오버헤드를 이전의 시스템과 비교할 수 있고, 둘째, SANfs 파일 시스템의 Bandwidth를 측정함으로써 시스템의 전체적인 성능을 측정할 수 있다는 점이다.

실험 환경은 4대의 9GB Seagate Fibre-channel SCSI disk와 3대의 PC로 구성되었다. 각 PC는 256MB의 메모리와, 700MHz Pentium CPU를 가지고 있다. OS는 Linux kernel 2.2.12를 사용하였다. PC 한 대는 메타서버와 파일시스템 클라이언트 모두로 동작하도록 구성하고, 나머지 2대는 파일시스템 클라이언트로만 동작하도록 하였다. SANfs-VM 시스템의 성능을 측정하기 위해서 각 파일시스템 클라이언트는 3개의 서로 다른 디스크의 파일시스템을 마운트하여 랜덤하게 액세스하도록 하였다.

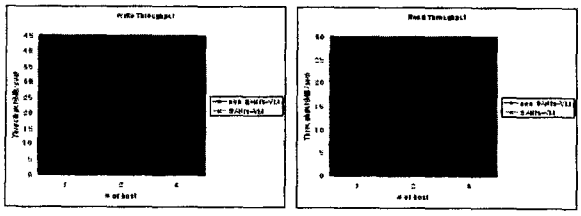

그림 3 : SANfs-VM 시스템의 성능 테스트

위의 그래프에서 보는 바와 같이 호스트의 개수가 증가함에 따라서 시스템의 성능도 역시 증가한다. 이것은 메타 서버가 bottleneck이 일어나지 않음을 말해준다. 그리고 볼륨 관리 기능을 추가한 시스템과 그렇지 않은 시스템의 성능이 차이가 없음으로, SANfs-VM은 이전의 시스템에 영향을 주지 않으면서 다양한 기능을 제공해 줄 수 있다는 것을 확인할 수 있다.

6. 결론 및 추후 과제

본 논문에서는 새로운 SAN 기반의 공유 파일시스템인 SANfs의 저장장치 시스템을 효율적으로 사용할 수 있는 볼륨 관리 기법에 관해서 제안하고 Linux에서 구현해보았다. SANfs-VM은 SANfs의 특징을 이용해서 다양한 기능을 지원해 준다. 그것은 다양한 레벨의 configuration을 지원해 주며, online resizing/reconfiguration 기능을 지원해 줌으로써 시스템의 가용성과 확장성을 향상시켜 준다. 그리고 볼륨 관리 방법이 메타서버의 볼륨관리 모듈에서 집중해서 이루어지기 때문에 관리하기가 쉬워지며 유저에게 다양하고 쉬운 인터페이스를 제공한다.

그렇지만 현재, 저장장치를 관리하고 성능을 유지시켜주는 작업은 시스템 관리자의 판단에 의해서 이루어지기 때문에 어려운 작업이 될 수 있다. 이에 대해서, 현재 볼륨 관리 모듈에 저장장치의 억세스 패턴, 데이터 블락의 억세스 빈도등을 모니터링할 수 있는 볼륨 모니터에 대한 개발과, 이 데이터를 이용한 auto resizing/reconfiguration을 통한 저장장치의 load balancing에 대한 연구를 수행중이며 앞으로 이루어져야 할 일들이 되겠다.

7. 참고문헌

[1] Heinz Maulschagen. Logical Volume Manager for Linux. http://linux.msede.com/lvm

[2] David C. Teigland, Heinz Mauelshagen. Volume Managers for Linux. http://www.sistina.com

[3] David C. Teigland, The Pool Driver: A Volume Driver for SANs. http://www.sistina.com

[4] Chang-Soo Kim, Gyoung-Bae Kim, Bum-Joo Shinm "Volume management in SAN Environment" IEEE

[5] SANfs Shared File System. http://core.kaist.ac.kr

[6] Joo Young Hwang, Chul Woo Ahn, Se Jeong Park, Kyu Ho Park "A Scalable Multi-Host RAID-5 with Parity Consistency" IEICE transactions on Information and Systems VOL.E85-D No.7 JULY 2002