리눅스 클러스터 화일 시스템 SANique™의 오류 탐지 기법*

임화정0, 이규웅
상지대학교 컴퓨터정보공학부
{eastseas0,leekw}@mail.sangji.ac.kr

이장선, 오상규
매크로임팩트㈜
{sunny, sgoh}@macroimpact.com

Failure Detection in the Linux Cluster File System SANique™

Hwa-Jung Lim, Kyu-Woong Lee
School of Computer Information, Sangji Univ.
{eastseas0,kwool}@cizeta.sangji.ac.kr

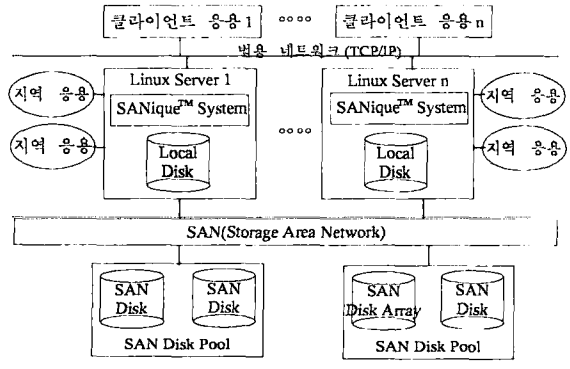
Jang-Sun Lee, Sang-Gyu Oh
Macroimpact co., Ltd.
{sunny, sgoh}@macroimpact.com

요 약

본 논문은 SAN(storage area network)상에 네트워크-부착형(network-attached) 저장 장치들을 직접 연결하여 화일 서버 없이 직접 데이터 전송이 가능한 SAN 기반의 리눅스 클러스터 공유 화일 시스템인 SANique™의 오류 탐지 기법 및 회복 기법에 대하여 기술한다. 클러스터 내의 노드 오류에 의해 발생하는 "split-brain" 오류 상황 및 문제점을 공유 화일 시스템 환경 하에서 정의하고, 이 문제를 해결할 수 있는 오류 탐지 기법을 제시한다.

1. 서 론

최근 화이버 채널 인터페이스 기술 발전으로 인한 네트워크-부착형(network-attached) 저장 장치들이 등장함에 따라, 네트워크 프로토콜 스택을 갖춰야 하는 NFS와 같은 전형적인 분산 화일 시스템의 구조가 SAN 기반의 공유 화일 시스템 구조로 변화되고 있다. SAN 기반 공유 화일 시스템은 <그림 1>과 같이 분산 화일 시스템의 기능을 모두 제공하며, 또한 네트워크 부착형 저장 장치를 서버 없이 직접 저장 장치 전용 네트워크(SAN)에 접속시켜 사용하므로 가용성 및 확장성에 있어서 기존 분산 화일 시스템 보다 우수하다. SANique™ 시스템은 SAN 공유 화일 시스템으로서 기존 분산 시스템에서 서버가 모든 화일 공유의 제어를 담당해야 하는 단점을 극복할 수 있는 새로운 구조의 분산 공유 화일 시스템이다[1].



<그림 1> SAN 공유 화일 시스템

이러한 시스템에서는 중앙집중적인 서버 관리체제의 병목현상을 제거하기 위해 한 노드가 SAN 저장장치에 대한 접근제어를 관할하지 않고 클러스터 노드의 대부분이 전역적인 화일 시스템 운영에 참여하게 된다.

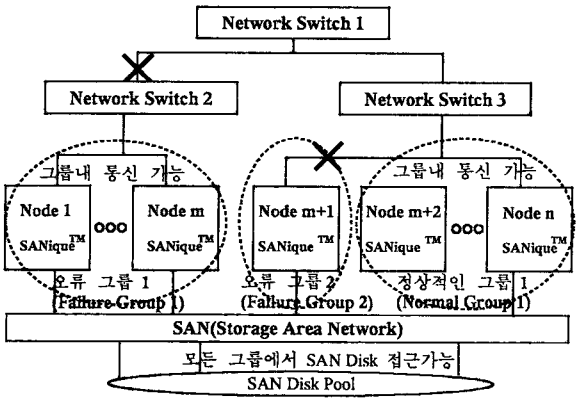
SAN 공유 화일 시스템의 오류의 종류는 일반 분산화일 시스템에서 발생할 수 있는 오류의 종류를 모두 포함한다. 그 중에서 노드간의 네트워크 오류에 의해 서로 통신이 단절된 상태에서 발생하는 오류는 일반 분산 화일 시스템이나 중앙집중적 서버를 갖는 네트워크 공유 시스템과 오류 상황과 매우 다르다. 본 논문에서는 이러한 "split-brain" 문제를 SAN 공유 화일 시스템상에서 정의하고, 이 문제를 해결하기 위한 오류 탐지 기법과 오류 회복 기법을 제안한다. 본 논문의 구성은 다음과 같다. 제 2장에서 SAN 공유 화일 시스템상의 "split-brain"문제를 정의하고, 제 3장에서 이에 대한 오류 탐지 기법을 정의한다. 끝으로 제 4절에서 본 논문의 결론을 맺는다.

2. SAN 공유 화일 시스템의 오류

SAN 공유 화일 시스템 상에서 발생하는 오류는 클라이언트 프로세스의 비정상적 종료에 의한 프로세스 오류, 무 전원 상태, 시스템 하드웨어의 고장등에 의한 시스템 오류, 디스크 및 네트워크 장애로 인한 장치 오류 등으로 분류 될 수 있다. 클러스터 시스템의 네트워크 및 시스템 오류는 정확히 어떠한 이유에 의한 오류인지 탐지하기 힘들다. 특히, 네트워크 오류에 의해 클러스터의 노드들이 그룹화 되어 있을 때 서로의 그룹이 상대방 그룹에 대하여 오류 처리를 하려고 하는 "split-brain"문제가 발생하게 된다[1,2,3,6]. 클러스터 화일 시스템의 오류 탐지는 일반적으로 토큰(token)기반의 주기적 메시지 전달(heartbeat) 방법을 통하여 탐지 된다. 이 방법을 통하여 위에서 언급한 오류 종류들의 대부분은 정확히 탐지될 수 있다.

* 본 논문은 매크로임팩트(주) 연구비 지원에 의한 것임

그러나 클러스터 화일 시스템 상에서 네트워크 오류 발생시, 노드 1은 노드2를, 노드 2는 노드 1을 서로 오류로 판단하는 "split-brain"문제가 발생할 수 있다. 상호 협조 체제(cooperative operation)로 수행되는 환경에서, 각 그룹이 기능적으로 완벽하지만, 서로 다른 그룹끼리 통신이 단절된 상태를 일반적인 "split-brain"문제라 정의한다[2].



<그림 2> SAN 공유 화일 시스템의
다양한 "split-brain" 발생 상황

범용 네트워크의 오류가 발생하면 주기적 상태점검 메시지가 서로 전달되지 않게 되므로 각 노드 1과 노드 N에서는 서로 상대방 노드가 오류인 상태인 것으로 판단하게 된다. 이 때, 서로가 상대 노드를 오류상태로 판단하고 회복한 후 재수행된다면, 각 지역에서 수행되는 지역 응용들은 상대노드들과의 통신 없이 즉, 화일에 대한 로킹이나 제어 정보의 교환 없이 지역응용이 수행되어 불일치성 데이터가 SAN을 통해 저장 또는 판독되게 된다.

SAN 공유 화일 시스템에서 "split-brain"문제는 각 노드들이 네트워크에 연결된 상태에 따라 발생하는 상황이 다양해 질 수 있다. <그림 2>는 다양한 발생상황을 혼재하여 나타내고 있다. <그림 2>에서 정상 그룹 1은 일반 네트워크 상태가 정상이며 그룹 내에서도 통신이 가능하다. 하지만 오류 그룹 1은 그룹 내에서는 통신이 가능하지만, 다른 그룹들 및 외부 네트워크이 단절된 상태이다. 오류 그룹 2는 독자적으로 고립된 상태이다. 세 개의 모든 그룹은 SAN과의 접근이 가능하므로 SAN 디스크에 기록 및 판독연산이 가능한 상태이다. 세 개의 그룹 중에서 단 하나만의 그룹이 선택되어 다른 그룹의 노드를 회복시키고 공유 화일 시스템 서비스를 재개해야 한다.

만약, 두개 이상의 그룹이 살아 남게 되어 서비스를 개시하게 되면 그 그룹들 간에 범용 네트워크를 통한 제어 정보 교환 없이, 즉 공유 화일에 대한 로킹 정보나 화일에 대한 메타정보(inode block) 수정 없이 화일 시스템이 운영되어 불일치한 데이터를 기록 또는 판독하게 된다. 따라서 여러 그룹 중 하나가 선택되어 공유 화일 시스템 서비스를 개시해야 하지만, 각 그룹간의 범용 네트워크 단절로 인해 각 그룹 중 "Winner" 그룹을 선택하는데 어려움이 발생한다. SAN 공유 화일 시스템에서 발생하는 "split-brain"문제를 해결하기 위해 다음과 같은 어려움이 발생함을 알 수 있다.

- 오류 그룹 내부에서 자신의 그룹이 오류 그룹인지 정상 그룹인지 판단 할 수 없다.
- 정상 그룹과 오류 그룹 모두 현재 오류 상황이 몇 개의 그룹으로 형성되었는지 파악할 수 없다.
- 네트워크 통신이 되지 않는 타 그룹들이 현재 네트워크 오류 상태인지 시스템 오류 상태 인지 구별할 수 없다.
- 정상 그룹인 경우, 현재 자신 이외의 정상 그룹이 또 존재하는지 알 수 없으며, 정상 그룹들 중에서 어떤 그룹이 가장 많은 노드를 포함하고 있는지 판단 할 수 없으므로, 최적의 정상 그룹으로 공유 화일 시스템 서비스를 재개할 수 없다.

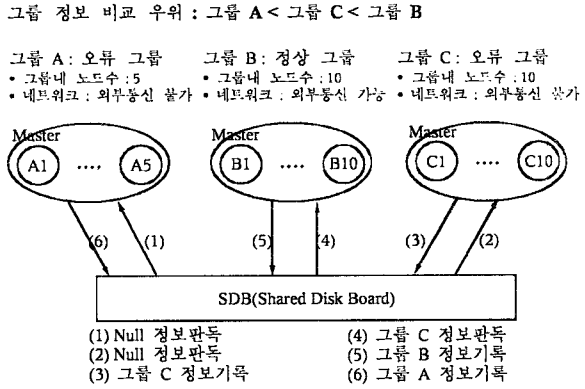
3. SAN 디스크를 활용한 오류 탐지 기법

"split-brain"문제를 해결하기 위해 여러 방법이 제시되었으나 기존의 분산환경에서 고 가용성을 목적으로 데이터 미러링과 같은 시스템에 적합한 방법[2,3]으로 SAN 공유 화일 시스템상에 적용하기에는 위에서 제시한 어려움을 그대로 갖고 있다.

본 논문에서는 SAN 공유 화일 시스템의 "split-brain"문제를 해결하기 위하여 네트워크 오류발생시에도 SAN을 통하여 접근할 수 있는 SAN 디스크를 활용한다. 이 공간은 공유 디스크 보드(SDB: Shared Disk Board)라는 이름으로 모든 노드에서 접근이 가능하다. 제시하는 오류탐지 기법은 다음과 같다. 먼저 각 노드는 주기적 점검 메시지를 이용하여 자신과 통신 가능한 노드 리스트 집합인 현재 노드-뷰를 주기적으로 갱신하고, 초기 노드-뷰와 현재 노드-뷰가 달라지면, 현재 그룹 내에서 마스터 노드를 선정한다. 마스터 노드는 자신의 그룹에 대한 그룹 정보를 작성하여 SDB에 기록한다. 그룹 정보에는 외부(예, 게이트웨이)로의 네트워크 통신 상태와 현재 그룹의 노드 개수 등을 포함시킨다. SDB에 그룹 정보를 기록할 때 이미 다른 마스터 노드가 기록한 정보가 존재하면 이를 읽어온다. 읽어온 정보와 자신의 그룹정보를 비교하여 비교우위에 있지 않으면 자신의 그룹 전체는 클러스터에서 제거(I/O Fence Out)된다. 그렇지 않으면 주어진 횟수만큼 기록과 판독을 재 수행한 후 최종적으로 자신의 그룹 정보가 우세하면 승자로 판정되고, 그 그룹이 다른 모든 그룹의 노드들을 오류처리 하게 된다. 그룹 정보의 비교는 다음과 같이 수행된다. 먼저 정상 그룹이 오류 그룹보다 우세하다. 이 비교는 외부 네트워크 상태를 비교하여 외부로의 통신이 가능한 그룹이 정상 그룹이고, 그렇지 않은 그룹이 오류 그룹이 된다. 같은 종류의 그룹은 통신 가능한 그룹내의 노드 개수가 많은 그룹이 비교 우위에 있게 된다.

SDB에 기록과 판독단계에서 또 다른 고려사항이 발생한다. 각 그룹의 마스터 노드들은 서로 통신할 수 없는 상태이므로 공통 접근 지역은 SDB에 기록 및 판독연산을 수행할 때 로킹 메커니즘이 제공될 수 없다. 참고문헌[4, 5]에서는 디스크 자체에 로킹

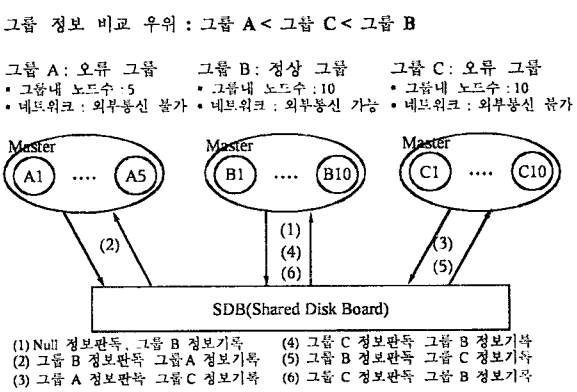
메커니즘을 부여하여 장치 로크(device-lock)을 사용하는 방법을 제시하였으나, 디스크 자체에 부가한다는 단점으로 인해 일반 SAN 환경에 적용이 어렵다. 로킹 메커니즘이 적용되지 않는다면 일반적 데이터베이스 트랜잭션 처리에서 발생할 수 있는 직렬화 가능하지 않는 수행으로 인해 판단결과가 올바르지 않게 나타날 수 있다. 예를 들어, 세 개의 마스터 노드가 있는 <그림 3>의 상황을 고려해 보자.



<그림 3> SDB 기록 수행의 문제

<그림 3>에서 그룹 B의 상황이 가장 좋으므로 다른 그룹들을 오류 처리하여 서비스를 재개하는 것이 최적의 해결방법이다. 그러나 그룹 간 통신이 안되므로, SDB를 활용하여 최적의 그룹을 선택하려고 하였으나, 기록 연산의 불일치 수행 문제 때문에 여전히 어려움이 있다. 즉, <그림 3>에서 그룹 A의 조건이 가장 나쁘지만, 기록연산을 마지막에 수행함으로써 승자로 판단되는 문제가 있다.

따라서, 본 방법에서는 그룹 정보의 기록과 판독을 test_and_set 연산으로 구현하여 기록과 판독이 준 원자적(semi-atomic)으로 수행될 수 있도록 구현하였다. 즉, 그룹정보를 판독하여 결정한 후 기록하는데 까지 의 원자적 수행을 로킹 메커니즘 없이 구현할 수 없으므로, 판독-기록을 원자적으로 수행한 후, 판독된 그룹 정보에 대한 판단 결정은 다음 수행 시 반영되도록 구현하였다. 판독-기록을 하는 test_and_set 연산을 여러 차례 수행케 하고 판독된 정보의 판단은 다음 번 기록시에 판단근거로 사용하게 된다. 즉, 자신의 그룹정보를 기록하고 기존 그룹 정보는 판독한 후에, 다음 번 그룹 정보 기록시에 이전 판독 시에 가져온 기존 그룹 정보가 자신의 정보보다 열세인 경우에만 기록하게 된다. 제안된 방법을 적용하여 수행한 결과가 <그림 4>에 나타나 있다. <그림 4>의 (2) 순서에서 그룹 A는 그룹 B의 정보를 판독하므로 더 이상 test_and_set을 수행할 수 없다. 마찬가지로 (5) 순서에서도 그룹 C는 그룹 B의 정보를 판독하므로 더 이상 기록 연산을 수행할 수 없게 되어 최종적으로 그룹 B가 승자로 판정되고, 그룹 A 및 C를 오류 회복처리하게 된다. 따라서, 가장 좋은 환경의 그룹이 범용 네트워크의 통신 없이 선정될 수 있으며 다른 그룹도 자신의 그룹이 클러스터 화일 시스템에 참여할 수 없다는 것을 알 수 있게 된다.



<그림 4> test_and_set를 적용한 SDB 기록

4. 결론 및 향후 연구

본 논문에서는 클러스터 내의 노드 오류에 의해 발생하는 "split-brain" 오류 상황 및 문제점을 공유 파일 시스템 환경 하에서 정의하고, 이를 해결하기 위한 오류 탐지 기법을 제안하였다. 범용 네트워크 통신이 단절된 상태에서 서로 다른 노드들이 상대 노드를 오류처리 하려고 하는 혼란 된 상태 즉, "split-brain" 상황을 없애기 위해 SAN 디스크인 공유 디스크 보드를 활용하여 최적의 서비스 그룹을 선택할 수 있는 오류 탐지 기법을 제시하였다. 그러나 판독-기록 연산의 수행 횟수와 분할된 그룹의 개수에 따라, 최적의 그룹이 선택될 수 없다는 단점을 갖고 있다. 현재 분할 그룹의 수와 판독-기록 연산횟수가 적은 상황에서도 최적의 그룹이 선택될 수 있는 오류 탐지 기법을 연구중이다.

참고문헌

[1] Sang G. Oh, and Jang S. Lee, "SANique™ : A SAN File system for Linux Cluster", Technical White Paper - Draft, MacroImpact. Co. Ltd., 2001

[2] C. C. Fan and J. Bruck, "The Raincore Distributed Session Service for Networking Elements", Proc. Of the International Parallel and Distributed Processing Symposium, 2001.

[3] P. S Weygant, "Primer on Clusters for High Availability", Technical Paper at Hewlett-Packard Labs, CA, 2000

[4] M. D. Dahlin, "Severless Network File Systems", Ph. D. Thesis at Computer Science Graduate Division of University of California at Berkely, 1995

[5] S. R. Soltis, T. M. Ruwart, and M. T. O'keefe, "The Global File Systems", Proc. Of the 5th NASA Goddard Conference on Mass Storage Systems and Technologies, 1996.

[6] K. W. Preslan, A. Barry, J. Brassow, R. Cattelan, A. Manthei, E. Nygaard, S. Oort, D. Teigland, M. Tilstra, and M. O'keefe, "Implementing Journaling in a Linux Shared Disk File System", Proc. Of the 8th NASA Goddard Conference on Mass Storage Systems and Technologies, 1999.