클러스터형 VOD 서버에서 장애 복구의 설계 및 구현

이좌형○ 서동만 방철석 김병길 박총명 정인범
강원대학교 컴퓨터 정보통신공학과 시스템 소프트웨어 연구실
{ jhlee○, csbang, dmseo, bgkim, cmpark }@snslab.kangwon.ac.kr, ibjung@kangwon.ac.kr

Design and Implementation of the Failure Recovery in Clustered VOD Server

Joa Hyoung Lee ○, Dongmahn Seo, Cheolseok Bang,
Byounggil Kim, ChongMyung Park, Inbum Jung
Dept. Computer Information & Telecommunication Engineering, Kangwon Univ.

요 약

최근 VOD 서버 모델로 제안되는 클러스터형 VOD 서버는 확장이 용이하여 서버의 가용성을 높일 수 있다는 장점이 있지만 서버에 노드수가 증가하면서 서버에 장애가 발생할 가능성이 높아진다는 문제점을 가지고 있다. 본 논문에서는 클러스터형 VOD서버에서 노드 장애 발생시 이를 복구하기 위한 방안으로서 RAID-3와 RAID-4의 특성을 취합, 보완하는 복구 시스템을 제시하고자 한다.본 복구 시스템은 RAID-4 개념을 도입하여 디스크로의 접근을 큰 사이즈의 블록단위로 함으로써 디스크의 효율성을 증가시키며, 네트웍에는 RAID-3 개념을 적용하여 작은 사이즈의 블록으로 나누어 전송함으로써 네트웍을 효율적으로 사용하고 백업서버의 메모리 부하를 줄일 수 있도록 한다.

1. Introduction

최근 컴퓨터 분야와 네트웍 분야의 기술이 급속히 발전하고 널리 보급되면서 멀티미디어 서비스의 한 분야인 VOD 서비스가 활발하게 연구되고 있으며, 데이터를 병렬 처리하는 클러스터형 서버가 저렴한 가격으로 높은 확장성과 고가용성을 지원하는 서버 모델로 제안되어져 왔다. 그러나 클러스터형 서버는 확장을 위해 노드수가 증가하면서 서버에 장애가 발생할 가능성이 높아진다는 문제점을 가지고 있다[1][2][3][4].

이러한 문제점들을 해결하기 위한 방안으로서 RAID [10][11][12][13] 기술을 이용하는 장애복구 기술들이 활발히 연구되고 있다. Bolosky et al.[5]은 Tiger 비디오 서버에서 RAID-1의 Mirror-Based 개념[4][14]을 기반으로 하는 복구방식을 사용하였으며, CD(Chained Declustering)[8][9], RMD(Rotational Mirrored Declustering)[1]과 같은 기술들이 Mirror-Based 개념을 기반으로 연구되었다. 그러나 Mirror-Based 방식의 경우 디스크 공간 사용의 효율성이 떨어지며, 백업노드에 과부하가 발생한다는 단점이 있다. Mirror-Based 방식의 단점을 해결하기 위해 Parity-Based 개념[14]을 사용하는 RAID-3,5 기술을 이용하여 장애를 복구하는 연구들이 활발히 진행되고 있다[7]. RAID-3 기술을 사용할 경우 디스크에서 읽혀지는 데이터의 사이즈가 작아 디스크 성능이 저하되는 단점이 있다. RAID-5 기술의 경우, 큰 사이즈의 블록단위의 데이터 처리로 인한 메모리와 네트웍 사용량 증가가 문제시 되고 있으며, 각 노드에서 복구작업을 수행해야 함으로 시스템이 복잡해 지는 단점이 있다[1][4][14].

본 논문에서는 클러스터형 VOD서버에서 노드 장애 발생시 이를 복구하기 위한 방안으로서 RAID-3와 4의 특성을 취합, 보완하는 복구 시스템을 설계하여 기존 연구된 VOD 시스템인 VODCA에서 [15] 구현한 결과를 제시하고자 한다. 본 복구 시스템은 RAID-4 개념을 도입하여 디스크로의 접근을 큰 사이즈의 블록단위로 함으로써 디스크의 효율성을 증가시켰으며, 네트웍에는 RAID-3 개념을 적용하여 작은 사이즈의 블록으로 나누어 전송함으로써 네트웍을 효율적으로 사용하고 백업서버의 메모리 부하를 줄일 수 있도록 하였다.

* 본 연구는 한국과학재단 목적기초연구(R05-2003-000-12146-0) 지원으로 수행되었습니다.

2. 복구시스템의 구조 및 동작

본 논문에서 제안된 복구 시스템은 VOD시스템인 VODCA[15][16]를 기반으로 설계 및 구현되어 졌으며, 시스템의 전체적인 구성과 동작을 그림 1에 나타내었다. 전체적인 시스템 구성은 Parity-Based방식을 사용하는 RAID-4 개념을 적용하여 MMS(Media Management Server)[15][16]와 별도로 복구작업을 전담하는 백업서버를 설치하여 시스템을 단순화하였으며 Parity 데이터 전송을 위한 분리된 내부 네트웍을 구성하여 기존 네트웍에 대한 오버헤드가 발생하지 않도록 하여 전체적인 성능저하가 없도록 하였다.

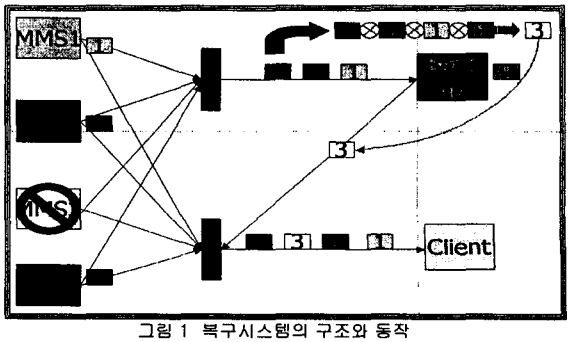

그림 1 복구시스템의 구조와 동작

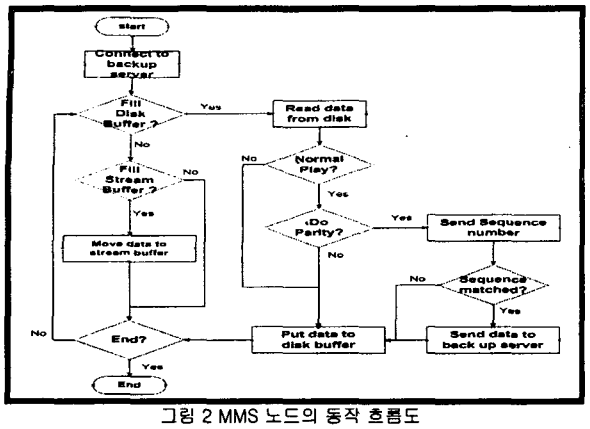

그림 2 MMS 노드의 동작 흐름도

MMS 노드들은 장애가 없는 상황에서는 디스크에서 영화 데이터를 읽어 클라이언트로 전송하며, 다른 MMS 노드에 장애 발생시에는 클라이언트와 백업서버로 동시에 영화데이터를 전송하여 백업서버에서 복구작업이 이루어지도록 한다. 백업서버는 MMS 노드에 장애발생 여부를 검사하여 장애발생시 이를 정상 MMS 노드들에게 통보하고, 영화 데이터를 수신하여 복구작업을 수행한 후 결과를 클라이언트로 전송하게 된다.

그림 2는 MMS 노드의 동작과정을 나타내는 순서도이다. MMS 노드에는 클라이언트별로 두개의 쓰레드가 생성되어 영화 스트림을 전송하며 공유버퍼를 사용한다[15][16]. 디스크에서 읽혀진 영화데이터는 우선 디스크 버퍼에 저장되며, 이후에 스트림 버퍼로 옮겨진 후 클라이언트로 전송되게 된다. 장애발생시 데이터를 복구하기 위하여 별도로 디스크에서 새로운 데이터를 읽어 들이는 것이 아니라 클라이언트로 전송할 데이터를 백업서버로 전송하여 복구작업을 수행하도록 한다. 따라서 추가적인 디스크 입출력작업 없이 복구 작업이 가능하도록 하여 RAID-5 의 문제점 중 하나인 디스크에 대한 오버헤드가 줄어들게 하였다. MMS 노드에서 백업서버로 영화데이터를 전송하기 전에 MMS 노드간에 블록번호를 일치시킴으로써 노드간에 동기를 맞추는 과정이 있으며, 현재 영화상영 상태가 Fast Forward 이거나 Fast Rewind 인 경우 복구작업을 수행하지 않아도 정상적인 서비스가 가능하므로 일반적인 재생일 경우에만 복구작업을 수행하도록 하였다.

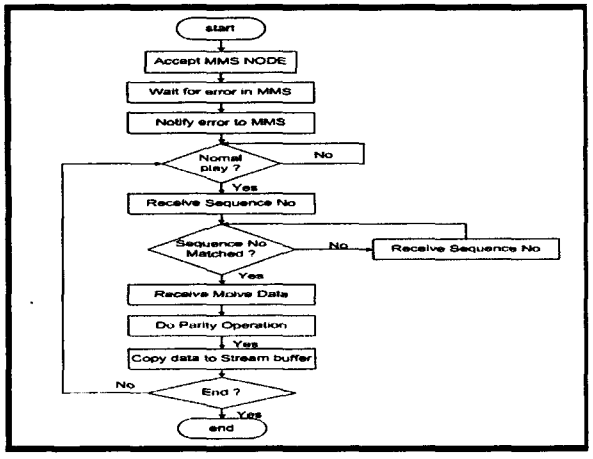

그림 3 백업서버의 동작 흐름도

그림 3은 백업서버의 동작흐름을 보여준다. 백업서버는 MMS 노드에 장애가 발생할 때까지 대기하며, 장애발생시 이를 정상 MMS 노드들에게 통보하고 MMS 노드로부터 블록번호를 수신하여 노드간에 동기를 맞추어 주는 과정을 거친 후, 수신한 데이터와 디스크에서 읽어 들인 Parity 데이터를 이용하여 복구작업을 수행한 후 복원된 데이터를 클라이언트로 전송하게 된다. 백업서버에는 Parity 데이터 외에 Parity 데이터를 관리하기 위해 Parity 데이터 블럭의 번호, 블록의 크기, 그리고 저장위치와 같은 Parity 데이터의 메타 데이터와, 각 MMS 노드에 저장되어져 있는 영화 데이터들의 메타 데이터를 저장하고 있어 어떠한 MMS 노드에서 장애가 발생하더라도 이를 복구할 수 있도록 하였다.

3. 성능측정 및 분석

본 논문에서 제안된 복구시스템을 VOD 시스템인 VODCA에서 구현한 후 성능을 측정하였다. VODCA의 경우 MMS 노드에 수가 4대 일 경우에 최대의 성능을 나타내기 때문에 [15][16], 복구시스템에 대한 성능측정도 4대의 노드로 MMS를 구성한 환경에서 이루어 졌다. VODCA 시스템의 기본적 구성에[15][16], 100Mbps Ethernet 허브를 추가하여 MMS노드와 백업서버 간에 분리된 별도의 내부 네트웍을 구성하였다.

본 실험에서 사용한 Yardstick Program[16]을 사용하여 복구 시스템의 성능을 측정한 결과 MMS 노드와 백업서버 간의 네트웍이 Bottleneck이 되어 전반적인 성능에 저하가 발생한 것으로 판단되어졌으며 이를 확인하기 위하여 부하 변화에 따른 MMS 노드와 백업서버의 네트웍 사용량을 측정하여 보았다.

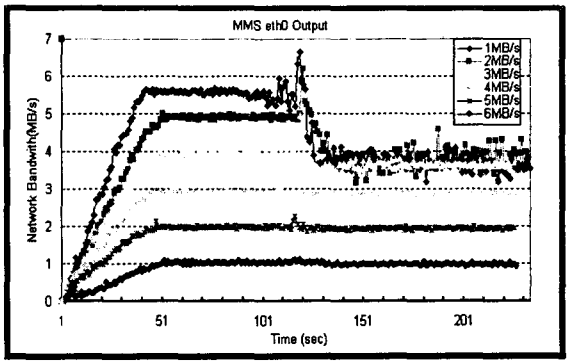

그림 4 부하량 변화에 따른 MMS 노드의 네트웍 사용량 변화 추이

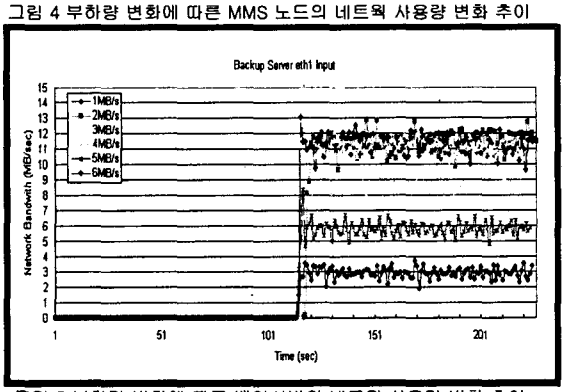

그림 5 부하량 변화에 따른 백업서버의 네트웍 사용량 변화 추이

그림 4는 MMS 노드에서 클라이언트로의 전송량의 변화를 보여주며, 그림 5는 MMS 노드에서 백업서버로의 전송량의 변화를 나타낸다. 그림 4와 5를 살펴보면 백업서버로의 전송량이 클라이언트로의 전송량에 3배가 되는 것을 알 수 있는데, 이는 4대로 구성된 MMS 서버 중에서 하나의 MMS 노드에서 장애가 발생할 시, 정상적인 3대의 노드에서 복구작업을 위한 데이터를 집중적으로 백업서버로 전송하기 때문에 클라이언트로의 전송량에 3배에 달하는 데이터가 백업서버로 전송되기 때문인 것으로 분석된다. 그림 4에서 전송량이 4MB/s 보다 많은 상황에서 장애가 발생할 경우 백업서버로의 입력량이 네트웍 인터페이스의 처리량인 12MB/s를 초과하게 되어 정상적인 서비스가 힘든 것으로 나타났다. MMS 노드에 4MB/s의 작업부하가 있는 상황에서 장애가 발생한 경우에 MMS노드와 백업서버의 네트웍 사용량 변화 추이를 나타낸 그림 6을 보면 MMS노드와 백업서버 간의 관계를 확실히 확인할 수 있다.

그림 7은 MMS 노드의 부하량 변화에 따라 클라이언트에서 기본 영화 데이터 전송 단위인 GOP를 수신하는데 소요되는 시간의 변화를 보여준다. 그래프에 값은 4개의 GOP를 수신하는 시간을 나타내는데 장애가 발생하였을 경우, 데이터의 손실과 MMS 노드와 백업서버의 초기화 과정으로 인해 수신하는 간격이 매우 불규칙적이지만 초기화가 끝나고 복구작업이 시작되면 다시 정상적으로 수신하는 것을 확인할 수 있어 복구시스템이 정상적으로 동작하는 것을 알 수 있었다.

4. 결론 및 향후 연구과제

본 논문에서는 클러스터형 VOD 서버에서 장애발생시 이를 복구하기 위한 시스템을 설계 및 구현하였다. 본 연구에서 구현한 복구 시스템의 성능 측정하여 분석한 결과 네트웍이 성능저하 원인으로 나타났다. 이를 해결하기 위한 방법으로 백업서버에 네트웍 인터페이스를 추가하면 네트웍의 오버로드는 줄어들 것으로 생각되나 MMS노드의 수가 증가할 경우 확장성이 문제가 될 것으로 생각된다.

장애가 발생한 MMS 노드에 접속한 모든 클라이언트에 대해 복구 작업이 동시에 이루어지기 힘들 경우, 인기 있는 특정 영화에 대해서만 복구 작업을 수행하는 방안을 복구 시스템에 적용해 볼 수도 있으며, BIBD(Balanced Incomplete Block Design)[1][6]개념 을 적용하여 각 클라이언트마다 일정한 주기로 복구작업을 수행함으로써 클라이언트간 균등한 복구작업이 이루어지도록 하는 방안도 적용가능 하겠다.

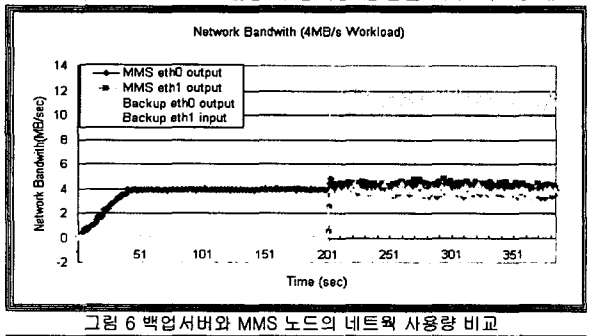

그림 6 백업서버와 MMS 노드의 네트웍 사용량 비교

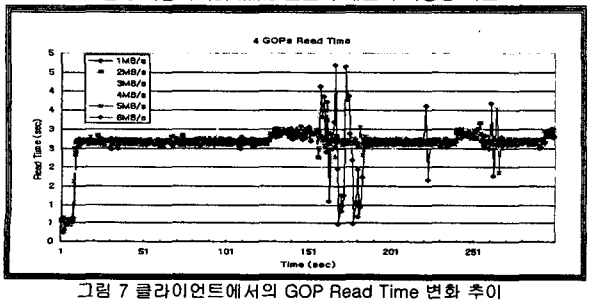

그림 7 클라이언트에서의 GOP Read Time 변화 추이

추후에는 본 논문에서 제시된 복구시스템의 성능저하 원인으로 나타난 네트웍 문제를 해결하기 위해, 백업서버로 집중되는 네트웍 부하를 각 MMS 노드로 분산하는 방안을 설계 및 구현하여 본 시스템과 비교 분석해볼 예정이다. 또한 최근 이슈로 부각되고있는 분야인 다중 장애의 복구에 대한 연구도 진행할 계획이다.

참고 문헌

[1] David H.C. Du, Yuewei Wang and Simon Shim," High Availability in Fault-Tolerant Video Servers" , Proceedings of IEEE International Conference on Multimedia Computing and Systems, 1998.

[2] Prashant J. Shenoy, Harrick M. Vin," Failure recovery algorithms for multimedia servers" , Multimedia Systems 8: 1-19 (2000)

[3] Jack Y.B. Lee, " Supporting Server-Level Fault Tolerance in Concurrent-Push-Based Parallel Video Servers" , IEEE transactions on Circuits and Systems for Video Technology, Vol. 11, No. 1, January 2001.

[4] Jamel Gafsi, Ernst W. Biersack, " Modeling and Performance Comparison of Reliability Strategies for Distributed Video Servers" , IEEE Transactions on Parallel and Distributed Systems April 2000 (Vol. 11, No. 4)

[5] w.J. Bolosky, J.S. Barrera, III, R.P. Draves, R.P Fitzgerald, G.A. Gibson, M.B. Jones, S.P. Levi, N.P. Myhyvold, and R.F. Rashid, " The tiger video server fileserver" , in Proc. 6th Int. Workshop Network and Operating System Support for Digital Audio and Video, Zushi,Japan, Apr. 1996.

[6] H. Hanani, " Balanced Incomplete Block Designs and Related Designs." , Discrete Mathematics, (11): 255-369, 1975

[7] F.A. Tobagi, J Pang, R. Baird, and M. Gang, " Streaming raid™-a disk array management system for video files" , in Proceedings of the 1st ACM Internatinal Conference on Multimedia, August, 1993.

[8] L. Golubchik, J.C.S. Lui, and R.R. Muntz, " Chained Declustering: Load Balancing and Robustness to Skew and Failure." , Proceedings of 2nd Intl Workshop on Research Issues in Data Engineering: Transaction and Query Processing, pages 89-95, February, 1992.

[9] H.I. Hsiao and D.J. Dewitt, " chained Declustering: A New Availability Strategy for Multiprocessor Database Machines." Proceedings of 6th Intl Conference on Data Engineering, pages 456-465, 1990

[10] D.A. Patterson, G. Gibson, and R. H. Katz, " A Case for Redundant Arrays of Inexpensive Disks(RAID)" , in Proceedings of the 1988 ACM Conferences on Management of Data, pp. 109-116. June, 1988.

[11] A. Merchant and P.S. Yu, " Analytic modeling and comparisons of striping strategies for replicated disk arrays" , IEEE Transactions on Computers, vol.44, pp.419-433, Mar, 1995.

[12] P.M. chen, E.K. Lee, G.A. Gibson, R.H. Katz, and D.A. Patterson, " Raid: High Performance, reliable secondary storage" , ACM computing Surveys. 1994.

[13] M. Holland, G.Gibson, and D. Siewiorek, " Architectures and algorithms for on-line failure recovery in redundant disk arrays" , Journal of Distributed and Parallel Databases, vol.2, July 1994.

[14] J. Gafsi and E.W. Biersack, " Performance and cost comparison of mirroring and parity based reliability schemes for video servers." , in Proceedings of KiVS' 99, March 1999.

[15] 서동만, 방철석, 김병길, 이좌형, 박춘명, 정인범 " 클러스터 VOD 시스템에서의 내장형 클라이언트 플랫폼 설계 및 구현" , 제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호

[16] 서동만, 방철석, 이좌형, 김병길, 정인범 "QoS를 지원하기 위한 리눅스 클러스터 VOD 서버의 성능 분석." , 한국정보과학회 2003 봄 학술발표논문집, 제30권 제1호(C), pp.301~303,