

클러스터형 VOD 서버에서 고가용성을 고려한 자체 복구 시스템

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

Self Recovery System With High Availability in Clustered VOD Server

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

요약

최근 VOD 서버 모델로 제안되는 클러스터형 VOD 서버는 확장성과 가용성을 높일 수 있다는 장점이 있지만 서버에 노드수가 증가하면서 서버에 장애가 발생할 가능성이 높아지는 문제점을 가지고 있다. 본 논문에서는 클러스터형 VOD 서버에서 노드 장애 발생 시 이를 복구하기 위한 방안으로서 RAID-3, 4 의 특성을 취합하고, 이에 파이프라인 개념을 더한 복구 시스템을 제시하고자 한다. 본 복구 시스템은 RAID-4 개념을 도입하여 디스크로의 접근을 큰 사이즈의 블록단위로 함으로써 디스크의 효율성을 증가시키며, 네트워크에는 RAID-3 개념을 적용하여 작은 사이즈의 블록으로 나누어 전송함으로써 네트워크를 효율적으로 사용하고 메모리 부하를 줄일 수 있도록 한다. 또한 파이프라인 개념을 도입하여 복구과정을 여러 노드에서 분담하여 동시에 처리함으로써 CPU, 네트워크, 메모리 등과 같은 자원에 대한 부하가 모든 노드로 분산되도록 한다.

1. Introduction

최근 컴퓨터 분야와 네트워크 분야의 기술이 급속히 발전하고 널리 보급되면서 멀티미디어 서비스의 한 분야인 VOD 생시 이를 복구하기 위한 방안으로서 RAID-3, 4 의 특성을 서비스가 활발하게 연구되고 있으며, 데이터를 병렬 처리하 취합하고, 이에 파이프라인 개념을 더한 장애 복구 시스템을 는 클러스터형 서버가 저렴한 가격으로 높은 확장성과 고가 제시하고자 한다. 본 복구 시스템은 RAID-4 개념을 도입하는 확장을 위해 노드수가 증가하면서 서버에 장 디스크의 효율성을 증가시키며, 네트워크에는 RAID-3 개념을 적애가 발생할 가능성이 높아진다는 문제점을 가지고 있다. 용하여 작은 사이즈의 블록으로 나누어 전송함으로써 네트워크 [1][2][3][4].

이러한 문제점들을 해결하기 위한 방안으로서 RAID 또한 파이프라인 개념[17]을 도입하여 복구과정을 여러 노드[10][11][12][13] 기술을 이용하는 장애복구 기술들이 활

발히 연구되고 있다. Bolosky [5]은 Tiger 비디오 서버에서 RAID-1 의 Mirror-Based 개념[4][14]을 기반으로 하는 복 구 방식을 사용하여 VOD 서버를 구성하였으며, CD(Chained Declustering)[8][9]나 RMD(Rotational Mirrored Declustering)[1]과 같은 기술들이 Mirror-Based 개념을 기반으로 연구되고 있는데, Mirror-Based 방식의 경 우 디스크 공간 사용의 효율성이 떨어지며, 백업노드에 과부

하가 발생한다는 단점이 있다. 최근에는 Mirror-Based 방식의 단점을 해결하기 위해 Parity-Based 개념[14]을 사용하는 RAID-3,5 기술을 이용하여 장애를 복구하는 연구들이 활 벌히 진행되고 있는데[7], RAID-3 기술을 사용할 경우 디스 템의 전체적인 구성과 동작을 그림 1 에 나타내었다. 전체적 크에서 읽혀지는 데이터의 사이즈가 작아 디스크 성능이 저 인 시스템 구성은 Parity-Based 방식을 사용하는 RAID-4 개

2. 자체 복구 시스템

2.1 자체 복구 시스템의 구조 및 동작

본 논문에서 제안된 자체 복구 시스템은 VOD 시스템인는 RAID-3,5 기술을 이용하여 장애를 복구하는 연구들이 활 VODCA[15][16]를 기반으로 설계 및 구현되어졌으며, 시스 벌도로 Parity Data 를 저장하고, 복구작업을 관리하는 백 가가 문제시 되고 있으며, 각 노드에서 복구작업을 수행해야 업서버를 설치하여 시스템을 단순화하였으며 Parity 데이터 전송을 위한 분리된 별도의 내부 네트워크를 구성하여 기존 네트워크에 대한 오버헤드가 발생하지 않도록 하여 전체적인 성 능자하가 없도록 하였다.

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

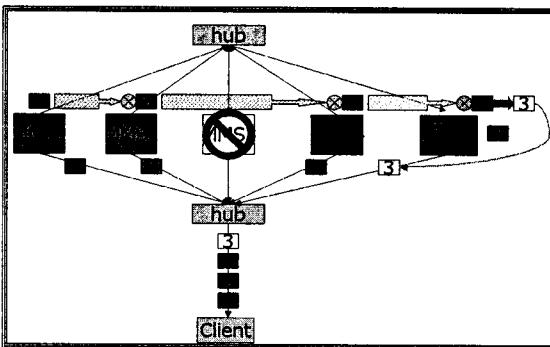


그림 1 복구 시스템의 구조와 동작

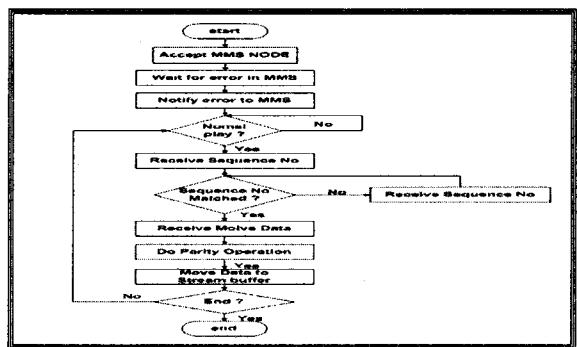


그림 3 백업서버의 동작 흐름도

본 논문에서 제안된 자체 복구 시스템에서 MMS 노드들 상영한다. I Picture는 다른 Picture에 대한 참조 없이 독립은 장애가 없는 상황에서는 디스크에서 영화데이터를 읽어 적으로 복호화 및 상영이 가능하며, 하나의 I Picture를 건너 클라이언트로만 전송한다. 다른 MMS 노드에 장애 발생시에 위더라도 사용자가 인식하기 힘들다. 따라서 영화 재생 상태는 클라이언트로 전송하는 데이터와 이웃하는 MMS 노드에 가 정상재생이 아닌 경우에는 복구작업을 수행하지 않아도 수신한 데이터를 이용하여 복구 작업의 일부를 수행한 후, 정상적인 서비스가 가능하므로 일반적인 정상 상영일 경우에 이웃하는 MMS 노드나 백업서버로 복구용 데이터를 전송 만 복구작업을 수행하도록 현재 플레이 상태를 검사하도록 한다. 복구작업이 MMS 노드와 백업서버에서 분산되어 이루 하였다.

어지도록 하였다. 백업서버는 MMS 노드에 장애발생 여부를 검사하여 장애발생시 이를 정상 MMS 노드들에게 통보하고, 이웃하는 MMS 노드로부터 복구용 데이터를 수신하여 복구 작업을 수행한 후 결과를 클라이언트로 전송하게 된다.

그림 2는 장애 복구 시스템에서 동작하는 MMS 노드의 동작 과정을 나타내는 순서도이다. MMS 노드에는 클라이언트 별로 두개의 쓰레드가 생성되어 영화 스트림을 전송하며 공유보퍼를 사용한다[15][16]. 디스크에서 읽혀진 영화데이터는 우선 디스크 버퍼에 저장되며 이후에 스트림 버퍼로 옮겨 진 후 클라이언트로 전송되게 된다. 장애발생시 데이터를 복구하기 위하여 별도로 디스크에서 새로운 데이터를 읽어 들이는 것이 아니라 클라이언트로 전송할 데이터를 백업서버로 도 전송하여 복구작업을 수행한다. 이런 단계적 절차는 추가적인 디스크 입출력 작업 없이 복구 작업이 가능하도록 하여 디스크에 대한 오버헤드가 줄어들게 하였다.

장애 상황 발생시 MMS 노드에서 이웃하는 MMS 노드나 백업서버로 영화데이터를 전송하기 전에 MMS 노드간에 블록번호를 일치시킴으로써 노드간에 동기를 맞추는 과정이 선행되어 수행된다. 현재 영화상영 상태가 Fast Forward 이거나 Fast Rewind 인 경우 I Picture 만 전송하여 사용자에게

2.2 자체 복구시스템의 특징

본 논문에서 구현된 백업서버는 RAID-4 개념에서의 백업서버와 달리 모든 복구작업을 백업서버에서 수행하는 것이 아니라, 서비스에 참여하는 정상적인 MMS 노드들에 분산되어 수행된 복구작업의 중간결과를 수신하여 저장하고 있는 Parity 데이터와 복구작업을 수행한 후 클라이언트로 전송하게 된다. 이렇게 함으로써 복구작업으로 인한 오버헤드가 백업서버로 집중되어 백업서버가 성능 저하의 원인이 되지 않는다는 특징이 있다. 또한 백업서버에서 여러 MMS 노드로부터 경쟁적으로 영화 데이터를 수신하는 것이 아니라, 이웃하는 하나의 MMS 노드에서만 데이터를 수신함으로써 서비스에 참여하는 MMS 노드의 수에 상관없이 최대한도로 네트워크 자원을 활용할 수 있다는 장점이 있다.

장애복구 작업에 참여하는 MMS 노드는 중에서 첫번째 노드는 수신하는 데이터 없이, 클라이언트로 전송할 영화 데이터 블록을 그대로 이웃하는 MMS 노드로 전송한다. 중간에 위치하는 MMS 노드들은 이웃하는 MMS 노드로부터 데이터 블록을 수신하여 자신이 가지고 있는 데이터 블록과 복구 연산을 수행한 후 이웃하는 MMS 노드나 백업서버로 작업 결과를 전송하게 된다. 이렇게 복구작업을 여러 노드에서 분산하여 처리하는 것은 프로세서를 설계하는데 사용되는 개념인 파이프라인 기술[17]을 활용하여 적용한 것이다. 이렇게 함으로써, 백업서버로 집중될 수 있는 복구작업으로 인한 오버헤드를 여러 노드로 분산시킴으로써 성능향상을 꾀할 수 있다는 장점이 있다.

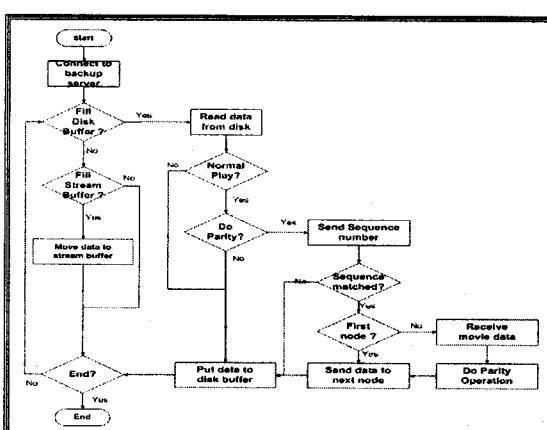


그림 2 MMS 노드의 동작 흐름도

장애 복구를 위해 VOD 시스템에 구현되는 복구 시스템의 경우 시스템 내의 많은 자원을 사용하게 된다. 본 논문에서 제시하는 복구 시스템이 Parity-Based 방식이기 때문에 데이터를 복구하기 위해 정상적인 노드들의 영화 데이터와 백업서버에 있는 Parity 데이터를 이용하여 Parity 연산을 수행하여야 한다. 이는 시스템의 CPU 자원을 많이 사용하여야 하는 작업이다. 그리고 복구 작업을 위해 필요한 데이터들이 여러 노드에 분산되어 저장되어져 있으므로 Parity 연산을 수행하기 위해서는 노드간에 데이터 이동이 필요하며 이로 인해 많은 네트워크 대역폭을 사용하게 된다. 또한 복구 작업에 필요한 데이터를 저장할 별도의 메모리 공간도 복구시스템에서 큰 비중을 차지하게 된다.

이러한 특징을 가지고 있는 복구 시스템을 VOD 시스템에서 어떻게 구현하느냐에 따라 복구 시스템의 전체적인 성능이 달라질 수 있다. 본 논문에서 제시하는 복구 시스템에서 사용하는 RAID-4 개념을 그대로 적용하여 복구 시스템을 설계할 경우 복구 작업을 전담하는 백업서버가 성능저하의 원인이 되게 된다. 이러한 성능저하의 가장 큰 원인으로 네트워크를 들 수 있다. RAID-4 방식에서는 모든 복구 작업을 위한 연산이 백업서버에서 이루어지므로 각 MMS 노드에서 백업서버로 네트워크를 통해 영화 데이터를 전송하게 된다. 하나의 영화데이터 블록을 복구하기 위해서는 모든 정상적인 MMS 노드들의 영화 데이터들이 동시에 필요하게 된다. 이를 위해 영화 데이터 블록 크기에 MMS 노드의 수를 곱한 만큼의 네트워크 대역폭을 필요로 한다. 따라서 서비스에 참여하는 MMS 노드의 수가 많고, 서비스하는 클라이언트 수 또는 많은 경우에는 네트워크가 성능저하의 원인이 되게 된다.

2.3 파이프라인방식의 복구

본 연구에서 도입한 파이프라인의 기본개념[17]은 순차적으로 수행되는 작업을 겹쳐서 동시에 수행함으로써, 가용한 자원을 최대한 활용할 수 있으며, 전체적인 수행시간도 크게 단축시킬 수 있는 기술로써 프로세서 설계에 많이 사용되어진다. 파이프라인 개념의 장점으로 크게 두 가지를 들 수 있는데 하나는 가용한 자원을 최대한 활용함으로써 자원의 효율성을 높일 수 있다는 것이다. 다른 하나는 동시에 여러 작업을 겹쳐서 수행함으로써 전체적인 수행시간을 크게 단축시킬 수 있다는 것이다.

VOD 시스템은 CPU 자원을 사용하는 연산 위주의 작업은 비중이 적으며 디스크과 네트워크 입출력에 대부분의 시간을 소모하고 있다. 구현된 VODCA 시스템에서 관찰된 결과를 보면 서비스 시에 CPU 자원보다는 메모리를 필요로 한다. 그리고 디스크로의 쓰기 작업은 거의 없는 반면에 읽기 작업이 큰 비중을 차지하게 된다.

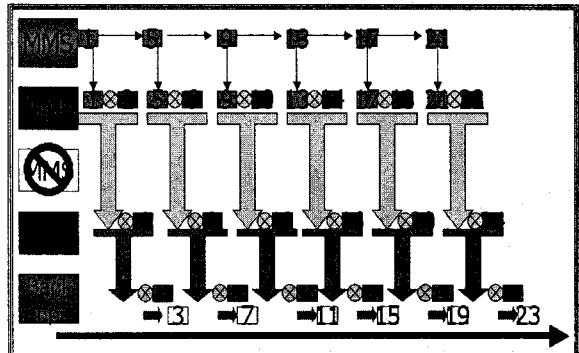


그림 5 RAID-4 개념을 적용한 복구 시스템의 동작 과정

또한 VOD 시스템의 성능향상에서 큰 비중을 차지하고 있는 네트워크를 살펴보면, MMS 노드가 네트워크를 통해 데이터를 수신하는 경우는 거의 없으며, 정상적인 경우 클라이언트로의 전송과 장애 복구시 백업서버로의 전송 작업이 네트워크의 많은 대역폭을 차지한다.

이러한 특징을 가지고 있는 VOD 시스템에 파이프라인 개념을 적용한 복구 시스템을 구현할 경우, MMS 노드의 가능한 자원들을 복구 작업에 사용하게 됨으로 백업서버의 오버헤드를 크게 줄일 수 있으며 시스템의 전체적인 자원 활용도도 높아지게 된다. 우선 복구 작업인 Parity 연산의 경우 CPU 자원을 많이 사용하는 연산위주의 작업인데, 이 작업을 백업서버에서만 수행하는 것이 아니라 MMS 노드에서 활용도가 떨어지는 CPU를 이용하여 작업을 수행하도록 함으로써 자원의 활용도가 높아지게 된다. 그리고 Parity 연산이 백업서버에서만 수행되는 것이 아니라 각 MMS 노드에서 분산되어 수행되어지므로 복구작업에 필요한 영화 데이터들이 백업서버로 집중되지 않고 각 MMS 노드들로 분산되므로 네트워크 자원을 효율적으로 사용할 수 있게 된다. 이는 MMS 노드로 복구용 데이터 외에 별다른 네트워크 입력이 미미하기 때문에 MMS 노드의 성능에는 영향을 미치지 않는다.

그림 4는 RAID-4 방식을 그대로 적용하여 복구 시스템을 구현하였을 때의 동작과정을 보여주며, 그림 5는 파이프라인 개념을 적용하여 구현하였을 때의 동작과정을 나타낸다. 그림 4를 보면 모든 복구 작업이 백업서버에서 수행되므로 MMS 노드들에는 큰 오버헤드가 발생하지는 않지만 다음의 복구작업에 필요한 데이터를 전송하기 전에 이전의 복구작업이 끝날 때까지 대기하여야 하는 단점이 있다. 또한 모든 데이터가 백업서버로 집중됨으로써 백업서버로의 네트워크가 성능저하의 큰 원인이 되게 된다.

이에 반해 파이프라인 개념을 적용한 그림 5를 살펴보면 복구작업이 모든 노드에서 분산되어 수행되므로 여러 데이터 블록에 대한 복구작업을 동시에 수행할 수 있으며 데이터를 전송할 때 대기시간도 크게 줄어들게 된다. 또한 복구작업에 필요한 데이터들이 백업서버로 집중되는 것이 아니라 각 노드로 분산되므로 네트워크 자원의 효율성도 높아지므로 전체적으로 성능이 향상되게 된다. 이렇게 파이프라인 개념을 적용함으로써 보다 짧은 시간에 보다 많은 복구 작업을 수행할 수 있게 된다. 그리고 복구 작업으로 인한 오버헤드가 백업서버에 집중되는 것이 아니라 각 노드에서 사용되지 않는 자원을 활용하여 수행되므로 시스템의 전체적인 자원을 효율적으로 활용할 수 있게 된다. 하지만 이웃하는 MMS 노드에서 수신한 데이터를 저장할 메모리가 추가적으로 사용되어지는 것이 파이프라인 개념을 적용했을 때의 문제점으로 부각된다.

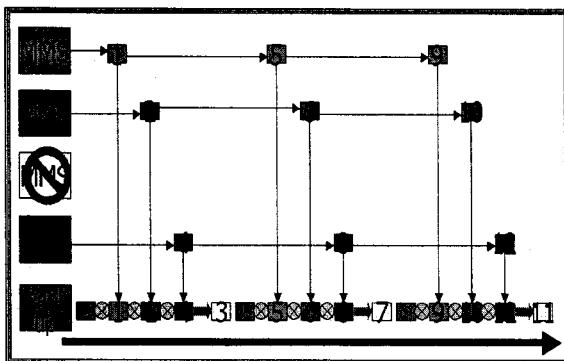


그림 4 RAID-4 개념을 적용한 복구 시스템의 동작 과정

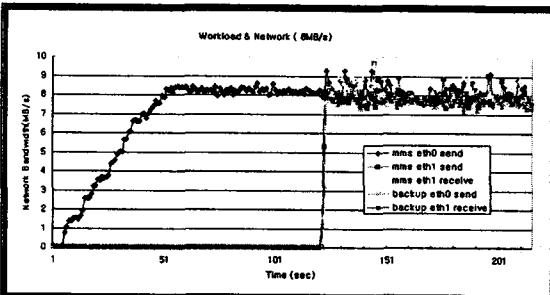


그림 6 MMS 노드 부하량(8MB/s)에 따른 네트워크 사용량 1

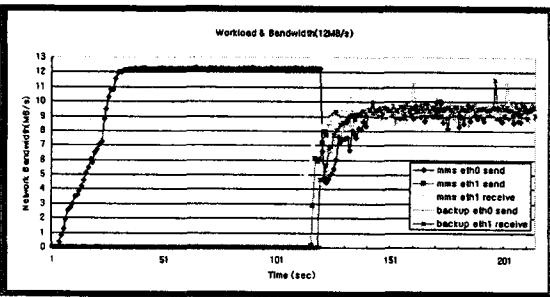


그림 7 MMS 노드 부하량(12MB/s)에 따른 네트워크 사용량 2

3. 성능측정 및 분석

본 논문에서 제안된 복구시스템을 VOD 시스템의 VODCA에서 구현한 후 성능을 측정하였다. VODCA의 경우 MMS 노드수가 4 대일 경우에 최대의 성능을 나타낸 것으로 발표되었다[15][16]. 복구시스템에 대한 성능측정도 4 대의 노드로 MMS를 구성한 환경에서 이루어 졌다. 실험을 위하여 기존 VODCA 시스템 구성에[15][16], 100Mbps Ethernet 하브를 추가하여 MMS 노드와 백업서버 간에 분리된 내부 네트워크를 구성하였다.

성능측정을 위하여 Yardstick Program[16]을 사용하여 복구 시스템의 성능을 측정하였으며, VOD 시스템의 성능 저하의 큰 요인으로 작용하는 네트워크 사용량이 MMS 노드의 부하량에 따라 어떻게 변화하는지를 알아보기 위해 부하 변화에 따른 MMS 노드와 백업서버의 네트워크 사용량을 측정한 결과를 그림 6과 그림 7에 나타내었다. 그림 6을 살펴보면 장애 복구작업 수행시 MMS 노드와 백업서버에서 사용하는 네트워크 대역폭은 같은 것을 알 수 있어 네트워크에 대한 부하가 모든 노드로 분산된 것을 확인할 수 있다. 그림 7은 시스템을 구성하는 네트워크의 최대치인 12MB/s의 작업 부하가 발생하였을 때의 네트워크 사용량을 측정한 것이다. 장애 발생 시 네트워크 대역폭을 최대로 사용하지 못하는 것으로 나타났는데 이는 작업 부하량이 늘어날 경우 네트워크보다 메모리가 성능저하의 원인으로 작용하기 때문인 것으로 판단된다[16].

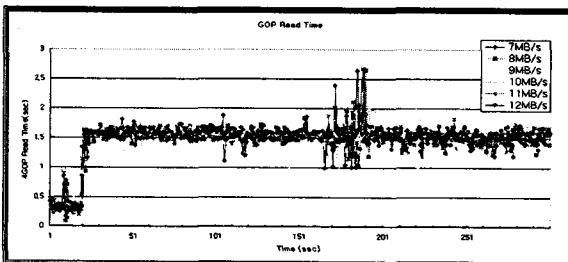


그림 8 작업 부하량 변화에 따른 GOP Read Time 변화 추이

그림 8은 작업 부하량에 따라 클라이언트에서 데이터 전송 단위인 GOP를 읽는데 소요되는 시간을 측정한 것인데, 장애 발생 후 복구 시스템에 의해 서비스가 정상적으로 제공되는 것을 확인할 수 있다.

4. 결론 및 향후 연구과제

본 논문에서는 클러스터형 VOD 서버에서 장애발생시 이를 복구하기 위한 시스템으로 RAID-3, 4의 특성을 취합하고, 이에 파이프라인 개념을 더한 복구 시스템을 설계 및 구현하였다. 본 연구에서 구현한 복구 시스템의 성능을 측정하여 분석한 결과, 파이프라인 개념으로 인해 복구 작업으로 인한 오버헤드가 백업서버로 집중되는 것이 아니라, 모든 노드로 분산되는 것을 확인할 수 있었으며, 네트워크 대역폭이 충분히 제공되는 상황에서 메모리가 성능 저하에 원인으로 나타났다. 추후에는 본 연구실에서 연구중인 스트리밍 가속기에서의 복구 시스템을 설계하고자 하며, 최근 이슈로 부각되고 있는 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 fileserv", 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 International 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 Int'l 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 Int'l 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 KVS' 99, March 1999.
- [15] 서동안, 방월석, 강병길, 이좌형, 박충영, 정인범 "클러스터 VOD 시스템에서의 내장형 클라이언트 플랫폼 설계 및 구현", 제 19 회 한국정보처리학회 추계학술발표대회 논문집 제 10 권 제 1 호
- [16] 서동안, 방월석, 이좌형, 김병길, 정인범 "QoS 를 지원하기 위한 리소스 클러스터 VOD 서버의 성능 분석", 한국정보과학회 2003 통합학술발표논문집, 제 30 권 제 1 호(C), pp.301~303.
- [17] David A. Patterson and John L. Hennessy, "Computer Organization & Design", PP. 392~490, Morgan Kaufmann, 1998