파이프라인 개념을 이용한 VOD 서버의 장애 복구 방법 연구

이좌형* · 박총명* · 정인범**

Design of Pipeline-based Failure Recovery Method for VOD Server

Joa-hyoung Lee* · Chong-myoung Park* · In-bum Jung**

본 연구는 산업자원부와 한국산업기술재단의 지역혁신인력양성사업으로 수행된 연구결과임

요 약

클러스터 서버는 front-end 노드와 여러 backend 노드로 구성된다. backend 노드 수의 증가로 더 많은 클라이언트들에게 QoS(Quality of Service)를 보장하는 스트리밍 서비스를 할 수 있지만, backend 노드의 오류 가능성도 이와 비례하여 증가한다. 서버의 장애는 모든 스트리밍 서비스를 중단시킬 뿐 아니라 현재 재생 위치 정보도 잃어버린다. 본 논문에서는 backend 노드가 오류 상태가 될 때, 끊이지 않는 스트리밍 서비스를 지원하기 위한 복구 방법을 제안한다. 클러스터 기반의 VOD 서버 구조를 고려하지 않고, 기본적인 장애복구 기술을 사용한 애플리케이션은 복구를 위한 내부 네트워크 성능의 병목현상과 backend 노드들의 비효율적인 CPU 사용을 야기시킨다. 본 논문에서는 이러한 문제를 해결하기 위해, 파이프라인 개념을 이용한 새로운 장애 복구 방법을 제안한다.

ABSTRACT

A cluster server usually consists of a front end node and multiple backend nodes. Though increasing the number of backend nodes can result in the more QoS(Quality of Service) streams for clients, the possibility of failures in backend nodes is proportionally increased. The failure causes not only the stop of all streaming service but also the loss of the current playing positions. In this paper, when a backend node becomes a failed state, the recovery mechanisms are studied to support the unceasing streaming service. The basic techniques are known as providing very high speed data transfer rates suitable for the video streaming. However, without considering the architecture of cluster-based VOD server, the application of these basic techniques causes the performance bottleneck of the internal network for recovery and also results in the inefficiency CPU usage of backend nodes. To resolve these problems, we propose a new failure recovery mechanism based on the pipeline computing concept.

키워드

Recovery, Pipeline computing, VOD, Clusters, QoS, Parallel processing

* 강원대학교 컴퓨터정보통신공학과 박사과정

** 강원대학교 컴퓨터정보통신공학과 교수 (교신저자)

접수일자 2008. 01. 16

Ⅰ. 서 론

최근 컴퓨터와 네트워크 기술의 발달로 VOD (Video-On-Demand), 전자 도서관, 원격 교육 같은 멀티미디어 서비스를 경제적으로 제공할 수 있게 되었다. VOD 서비스는 가장 대표적인 멀티미디어 애플리케이션이며, QoS를 보장하는 스트리밍 비디오 데이터를 온라인 사용자에게 제공한다[1, 2].

VOD 서버는 저장장치에서의 저장과 인출 및 네트워크로 영화 데이터를 전송하는 과정이 실시간이어야 하는 제약이 있다. 스트리밍 비디오의 끊김과 지터는 VOD 클라이언트에게 무의미 하므로, 스트리밍 미디어는 각 클라이언트의 QoS 기준을 만족시킬 수 있어야 한다. 서버들 중에 장애가 발생하더라도, 스트리밍 서비스는 사용자가 허용 가능한 MTTR (Mean Time To Repair) 값 안에서 복구되어야 한다 [3, 4].

최근 클러스터 서버 구조는 웹, 데이터베이스, 게임, VOD 서버 등의 분야에 이용되고 있다. 일반적으로 클러스터 서버 구조는 front-end 노드와 여러 backend 노드들로 구성된다. 각 노드의 하드웨어 및 소프트웨어 사양은 같을 수 도 있고 다를 수 도 있다. 비디오 데이터가 여러 backend 노드들에 분산 저장되기 때문에 backend 노드 수의 증가에 따라 저장장치를 비롯한 성능이 향상된다.[5, 6, 7, 8].

노드의 장애는 모든 스트리밍 서비스를 중단시킬 뿐 아니라 모든 상영되는 영화의 서비스되는 위치 정보를 잃어버리게 된다. 노드의 장애에도 VOD 서버가 QoS를 보장해야 하기 때문에, 실제 VOD 서비스를 다루기 위해 장애 복구 기법이 필요하다.

본 논문에서는, 클러스터 기반의 VOD서버에서 backend 노드의 장애 발생시 QoS(Quality of Service)를 보장하기 위해 모든 backend 노드들 간에 파이프라인 컴퓨팅 기반의 새로운 장애 복구 시스템(RS-PCM, Recovery System based on Pipeline Computing Mechanism) 을 제안한다. 제안된 RS-PCM 시스템은 배타적 OR 를 이용하여 컴퓨팅 부하 뿐 아니라 backend 노드 간의 네트워크 트래픽을 분산시킨다. 정상적인 backend 노드들은 모두 복구 과정에 참여하여 클러스터 기반의 VOD 서버에서 backend 노드 장애 발생시 끊기지 않는 스트리밍 서비스를 하는 개선된 성능을 제공한다. 제안하는 장애 복구 시스템을 평가하기 위하여, 클러스터 VOD 서버인 VODCA(Video On Demand on Clustering Architecture)에 파이프라인을 이용한 제안 시스템 RS-PCM과 VODCA에 임의의 부하를 줄 수 있는 부하 발생기를 구현하였다.

본 논문의 구성은 다음과 같다. 2장에서는 클러스터 VOD 서버인 VODCA 와 클러스터 아키텍처에서 비디오 블록 관리에 대하여 설명한다. 3장에서는 파이프라인 개념을 이용한 새로운 복구 방법을 제안한다. 4장에서는 제안 시스템의 성능을 평가하고 5장에서는 본 연구와 관련된 연구를 설명하고 6장에서 본 논문의 결론을 맺는다.

Ⅱ. 클러스터 VOD 서버

2.1. 클러스터 노드에서 MPEG 미디어의 병렬 처리

MPEG-1, 2 비디오는 비디오 시퀀스 레이어, GOP (Group Of Pictures) 레이어와 픽처 레이어로 구성된다. GOP 레이어는 영화를 상영하기 위한 최소 단위이며 보통 임의 접근을 위해 사용된다. 픽처 레이어는 화면에 보여지는 단일 이미지이다. 이 픽처 레이어는 I, B, P, D의 네가지 프레임으로 구성된다. 각 프레임은 디코딩 과정에서 각자 다른 기능을 한다. 각 GOP는 임의 접근이 가능하며 최소 하나 이상의 I 프레임을 포함한다 [2].

병렬 처리에 MPEG 미디어 특성을 이용하기 위해서 스트라이핑을 GOP 단위로 하였다. MPEG 스트림에서 각 GOP는 같은 상영 시간을 가지기 때문에, MPEG 영화는 GOP 단위로 스트라이핑되어 시퀀스 번호, 메타데이터와 함께 각 노드에 분산 저장된다.

2.2. VODCA의 구조

대용량 VOD 서비스를 위해, 우리는 클러스터 VOD 서버인 VODCA를 구현하였다. VODCA는 front-end 서버인 HS (Head-end Server) 와 backend 노드인 여러 MMS (Media Management Server)로 구성된다. 클라이언트는 HS와 MMS 노드들과 상호 작용하며 HS와 MMS 노드 사이에 내부 네트워크를 통해서 동작 상태와 내부 명령을 전달한다.

HS 노드는 클라이언트의 요청을 받아들일 뿐 아니라 QoS를 제공하기 위해 MMS 노드를 관리한다. 새로운 MPEG 영화가 등록되면, HS는 영화를 나누어 각 MMS 노드에 분산 저장한다. MMS는 HS의 관리하에 나누어

저장된 영화 데이터를 클라이언트에게 전송하며 주기적으로 HS 노드에게 현재 동작 상태 정보를 전송한다. 이 메시지는 HS 와 MMS 노드 사이에 Heartbeat 프로토콜로 운용된다.

Ⅲ. 파이프 라인 기반의 복구 시스템

3.1 시스템 구조

그림 1은 RS-PCM의 구조와 VODCA 서버의 비디오 블록의 흐름을 나타낸다. 그림 1에서 RS-PCM은 복구 과정에 필요한 네트워크 트래픽 뿐만 아니라 모든 MMS 노드에 배타적 OR 연산을 분산한다.

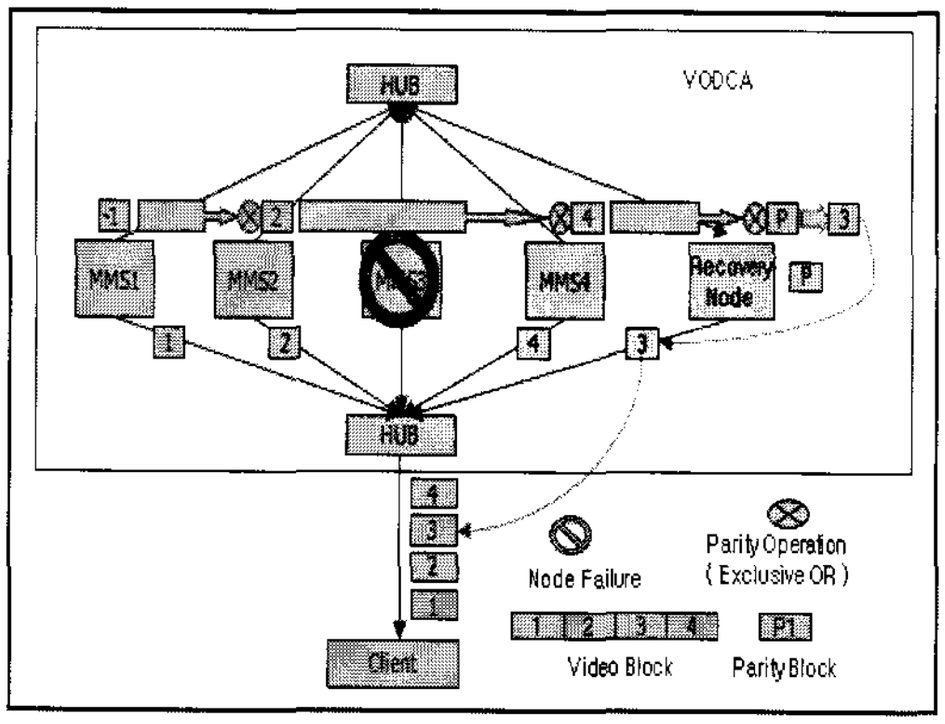


그림 1 RS-PCM 의 비디오 블록의 흐름.
Fig 1. Flow of video block in RS-PCM

MMS 노드 오류 발생시 모든 정상 MMS 노드는 자신의 비디오 블록을 복구 노드에 직접 전송하지 않고 저장하고 있던 비디오 블록이나 배타적 OR 연산의 결과 블록을 이웃 MMS 노드에 전송한다. 각 MMS 노드는 로컬 디스크로부터 인출한 비디오 블록과 이웃 MMS 노드로부터 수신한 블록으로 배타적 OR 연산을 수행한다. 이웃MMS 노드로부터 수신된 블록은 디스크에 저장되어 있는 원본 비디오 블록이거나 다른 이웃 MMS 노드에서 배타적 OR 연산에 의해 생성된 결과이다. 계산된 결과는 명령어 레벨에서 파이프라인 과정을 통해 이웃 MMS 노드에게 성공적으로 보내진다 [16].

마지막으로, 복구 노드는 모든 MMS 노드에서 계산된 결과와 패리티 블록으로 마지막 배타적 OR 연산을 수행하여 오류가 발생한 MMS 노드의 비디오 블록을 복원하여 외부 네트워크를 통해 클라이언트로 전송한다.

3.2 RS-PCM 의 특징

그림 2는 RS-PCM의 파이프라인 개념에 따른 복구 과정을 보여준다. 그림 2에서, 매 사이클 최소 한번 배타적 OR 연산을 수행, 인출, 전송 과정이 발생하는 것은 명령어의 병렬처리에서 파이프라인 기술과 흡사하다 [16]. 오류 블록을 복원하는 이러한 병렬 처리는 제안된 RS-PCM에서 성능을 향상된다.

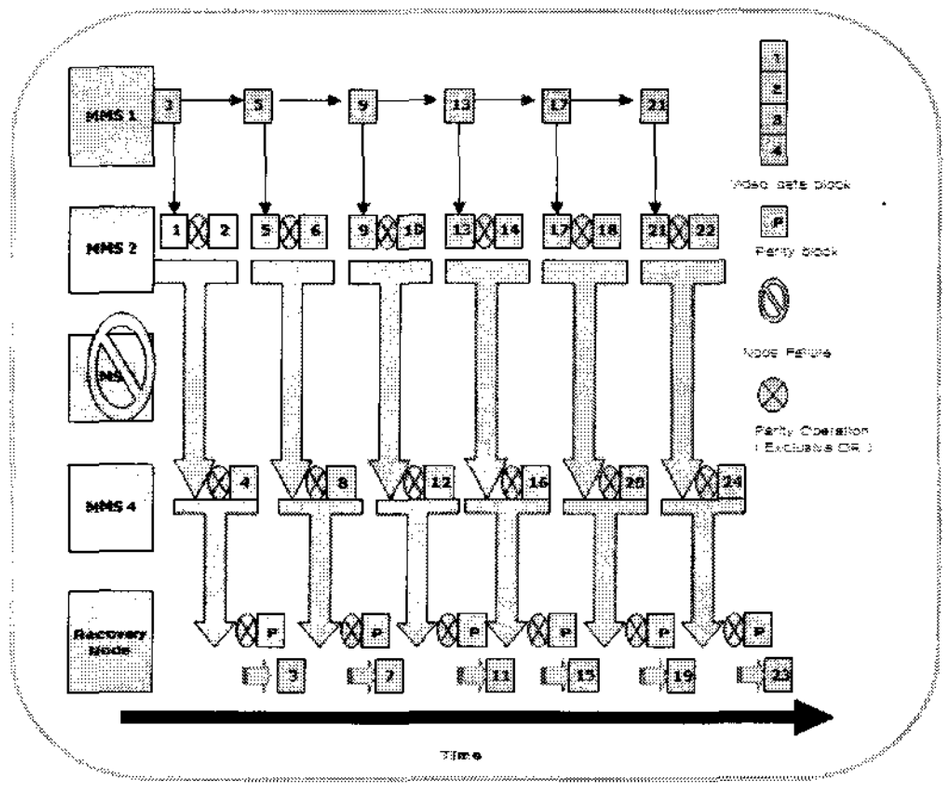


그림 2 파이프라인 개념 기반의 복원 단계
Fig 2. Recovery process based on Pipeline

RS-PCM에서 복구 노드는 각 사이클 동안 한번 배타적 OR 연산을 실행한 후 결과를 클라이언트로 전송한다. 제안된 시스템은 배타적 OR 연산을 위한 컴퓨팅 부하 뿐만 아니라 네트워크 트래픽을 모든 MMS 노드로 분산한다. 복구 노드의 입력 네트워크 트래픽은 MMS 노드 하나의 출력 트래픽과 같다.

그림 3은 RS-PCM에서의 네트워크 트래픽을 보여준다. 각 MMS 노드의 출력 네트워크 트래픽은 m 으로 같다. 복구 노드와 MMS 노드는 내부 네트워크의 성능을 최대로 사용한다. 그림에서와 같이 n-1 MMS 노드가 살아있고 각 출력 트래픽이 m 이라도 복구 노드의 입력 네트워크 트래픽은 m 이 된다. 복구 노드의 입력 네트워크 트래픽은 MMS 노드의 수에 의존적이지 않아서 RS-PCM은 복구 노드 네트워크의 입력 포트에서 병목현상이 일어나지 않는다.

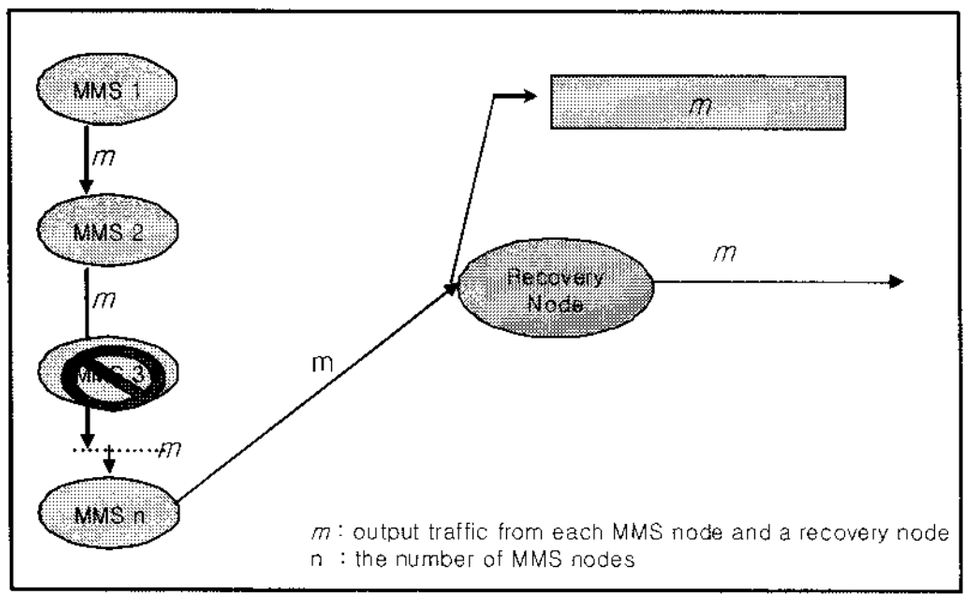


그림 3 RS-PCM 의 네트워크 트래픽 모델.
Fig 3. Network Traffic model in RS-PCM

Ⅳ. RS-PCM의 성능평가

4.1 실험환경

실험을 위한 VODCA 서버는 HS 노드와 4개의 MMS 노드 그리고 복구 노드로 구성되며, 각 노드는 Linux 운영체제로 동작한다. MMS 노드, HS 노드, 클라이언트는 100Mbps 이더넷 스위치를 통해 연결되어 있다. 모든 MMS 노드와 복구 노드 또한 100Mbps 이더넷 스위치를 통한 내부 네트워크로 연결되어 있다. 표 1은 VODCA 시스템에서 각 MMS 노드의 하드웨어 구성을 나타낸다.

표 1 MMS 노드와 복구 노드 사양.
Tab 1. Spec of MMS node and recovery node

CPU	Intel Pentium 4, 1.6 GHz
Memory	256 Mbyte DDR
Disk	Segate 40GB 7200RPM x 2
OS	RedHat 7.3 (Kernel 2.4.18)
Network	100 Mbps Fast Ethernet, 100Mbps Ethernet Switch

4.2 부하 생성기와 성능 평가 요소

클러스터 기반의 VOD 서버의 성능을 측정하기 위해 yardstick program 을 사용하였다 [13]. Yardstick 프로그램은 가상 부하 생성기와 가상 클라이언트 데몬으로 구성된다. 가상 부하 생성기는 HS 노드에 위치하고 λ = 0.25의 포아송 분포에 따라서 클라이언트 요청을 생성한다 [14, 15]. 생성된 요청은 각 MMS 노드에 보내지게 되고 모든 MMS 노드는 클라이언트가 만족할 수 있도록 동시에 미디어 스트리밍 서비스를 시작한다.

MMS 노드의 네트워크 트래픽의 변화, 복구 노드의 네트워크 트래픽, 클라이언트 측면에서 손상된 스트리밍 미디어의 평균 복구 시간을 성능 평가 요소로 한다. 본 실험에서 MMS 노드로부터의 네트워크 트래픽은 서비스 받는 클라이언트의 수를 의미한다. 복구 시스템의 올바른 동작 과정을 확인하기 위해, 많은 네트워크 트래픽 부하를 생성하였으며, 성능 평가 요소에 맞게 MMS 노드와 복구 노드의 반응을 관찰하였다.

4.3 실험결과

그림 4는 12MB/sec의 네트워크 부하가 발생할 때, MMS 노드와 복구 노드에서의 네트워크 트래픽을 나타낸다. MMS 노드의 오류는 시간축의 120초 에서 발생한다. 위 그림에서, 오류가 발생한 후 MMS 노드에서 클라이언트로의 네트워크 트래픽은 12MB/sec 에서 9MB/sec 로 감소한다. 이러한 현상이 발생하는 이유는 이웃 MMS 노드로부터 전송된 비디오 블록이 MMS 노드의 주 메모리를 차지하기 때문이다.

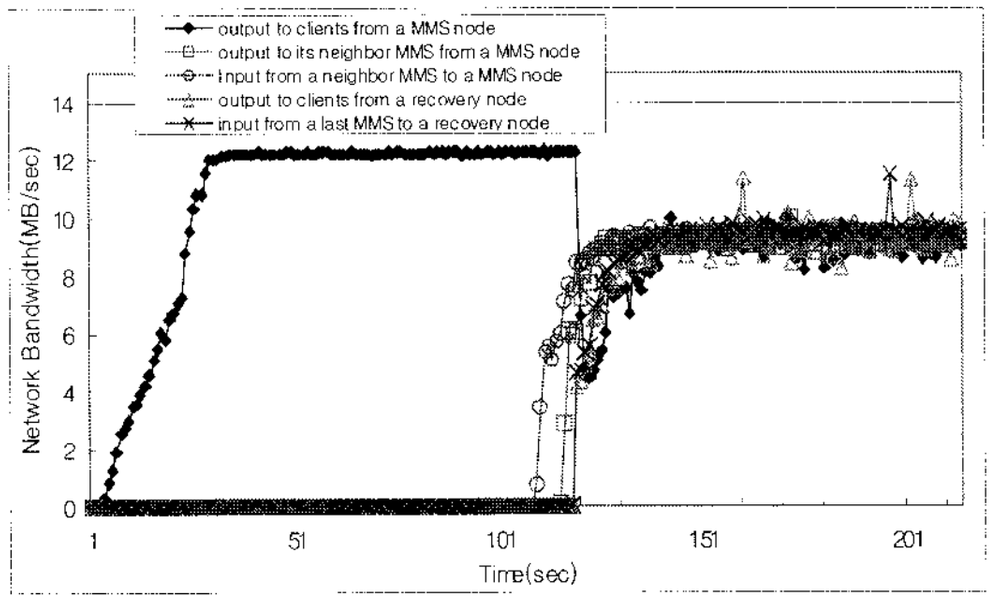


그림 4 노드들의 네트워크 트래픽.
Fig 4. Network traffic of node

그림에서 이웃 MMS 노드로의 출력 트래픽은 사각형으로 나타내었다. RS-PCM에서 현재 MMS 노드가 마지막 MMS 노드가 아니라면 자신의 비디오 블록이나 배타적 OR 연산으로 계산된 결과 블록을 이웃 MMS 노드로 전송한다. 원으로 표시된 부분에서 이웃 MMS 노드로부터의 입력 트래픽이 출력 트래픽과 거의 같음을 알 수 있다. 마지막 MMS 노드로부터의 입력 트래픽은 9MB/sec 에 가까워지며 복구 노드 또한 9MB/sec 의 비율로 비디오 블록을 복구할 수 있다. 그 후 복구 노드는 복구된 비

디오 블록을 클라이언트로 전송한다. 그림에서 삼각형으로 표시된 부분에서 복구 노드의 복구된 블록의 출력 트래픽은 9MB/sec 가 된다.

그림 5는 클라이언트에서 1 GOP를 읽는데 걸리는 시간을 나타낸다. 본 실험은 7MB/sec와 12MB/sec 사이의 부하 환경에서 수행되었다. RS-PCM은 이러한 네트워크 트래픽 부하를 지원할 수 있다. MMS 노드의 오류는 120초 에서 발생하였다. 서버와 클라이언트 사이에 네트워크 지연이 있기 때문에 클라이언트에서 읽는 시간의 변화는 148초와 176초 사이에 발생하였다. 불안정 상태 후에 읽기 시간은 0.65초에서 안정 상태로 변한다. 변화는 28초간 지속되었다.

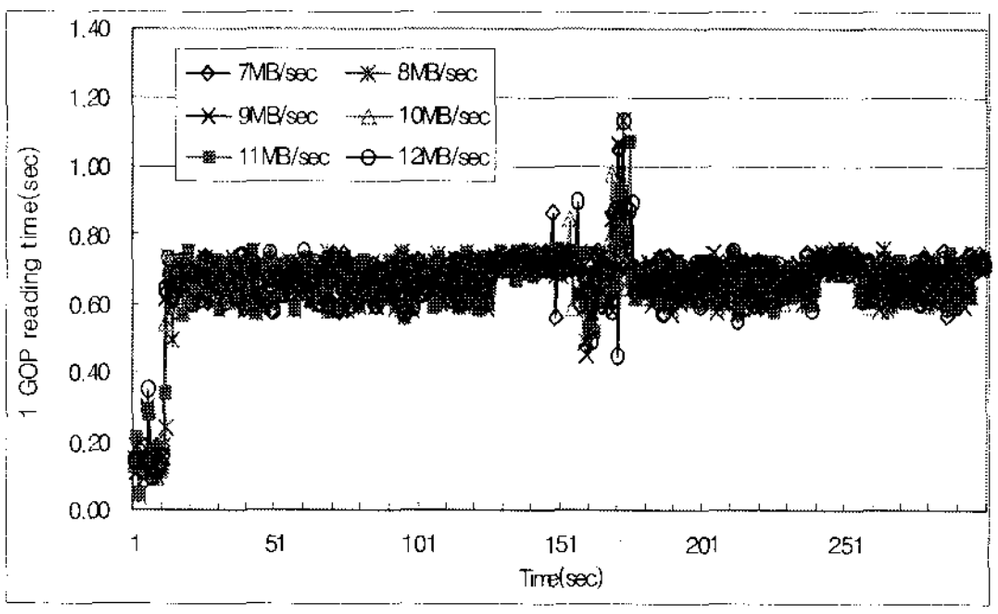


그림 5 클라이언트의 GOP reading time.
Fig 5. GOP reading time of client

V. 관련연구

다양한 클라이언트의 요구와 제한된 자원 하에서 좀 더 많은 클라이언트에게 안정적인 서비스를 제공하기 위한 VOD 시스템에 관한 많은 연구가 이루어지고 있다 [17, 18]. VOD 서버의 부분적인 오류 상태에서 끊김이나 지터 현상 없는 QoS 스트림을 보장하기 위한 충분한 연구가 이루어지고 있지 않다.[8, 5]

디스크나 backend 노드의 오류가 발생하더라도, 불규칙한 끊김이나 지터가 인간이 허용하는 시간 안에서 해결되어야 한다 [3, 4]. 하지만 현재까지 클러스터 기반의 VOD 서버에서 복구 방법에 관한 연구가 깊이있게 이루어지지 않았다. 좀 더 상업적인 VOD 서비스를 위해서 미디어 스트리밍의 특징에 맞게 복구 시스템이 연구되어야만 한다.

VI. 결론 및 향후 연구

VOD (Video-On-Demand)는 미디어 스트리밍 서비스를 위한 대표적인 기술이며 많은 분야에서 연구되어 왔다. 하지만 성공적인 VOD 서비스를 위해 VOD 서버에서 부분적인 오류가 발생하더라도 사용자가 허용하는 MTTR 값 안에서 모든 클라이언트에게 미디어 스트리밍 서비스를 보장해야 한다. 오류 상태에서 VOD 서비스를 제공하기 위해, 미디어 스트리밍의 특징이 복구 방법에 반영되어야 한다. 본 논문에서는 클러스터 기반의 VOD 서버에서 파이프라인 기반의 RS-PCM 을 제안하였다. RS-PCM 에서, 복구 노드는 복구된 비디오 블록을 생성하여 각 사이클에 한번 클라이언트로 전송한다. 이 방법은 명령어 실행단계의 파이프라인 과정과 비슷하다. 이 파이프라인 컴퓨팅을 기반으로 RS-PCM은 배타적 OR 연산을 위한 컴퓨팅 부하뿐만 아니라 모든 MMS 노드의 네트워크 트래픽을 분산시킨다.

참고문헌

[1] Dinkar Sitaram, Asit Dan, "Multimedia Servers: Applications, Environments, and Design,"Morgan Kaufmann Publishers, 200.

[2] http://www.mpeg.org

[3] Armando Fox, David Patterson, "Approaches to Recovery Oriented Computing," IEEE Internet Computing, Vol. 9, no. 2, pp.14-16, 2005.

[4] Dong Tang, Ji Zhu, Roy Andrada, "Automatic Generation of Availability Models in RAScard," IEEE International Conference of Dependable Systems and Networks, June 23-26, pp. 488~494, 2002.

[5] T. Chang, S. Shim, and D. Du, "The Designs of RAID with XOR Engines on Disks for Mass Storage Systems,"IEEE Mass Storage Conference, March 23-26, pp. 181~186, 1998.

[6] Prashant J. Shenoy, Harrick M. Vin, "Failure recovery algorithms for multimedia servers,"Multimedia Systems, 8: pp. 1~19, Springer-Verlag, 2000.

[7] 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, pp. 25~39, January

2001.

[8] Jamel Gafsi, Ernst W. Biersack,"Modeling and Performance Comparison of Reliability Strategies for Distributed Video Servers," IEEE Transactions on Parallel and Distributed Systems, Vol. 11, No. 4, pp.412~430, 2000.

[9] 서동만, 방철석, 이좌형, 김병길, 정인범, "리눅스 기반의 클러스터 VOD 서버와 내장형에 클라이언트의 구현", 정보과학회논문지 제10권 제6호 pp.435~447, 2004

[10] Jung-Min Choi, Seung-Won Lee, Ki-Dong Chung, "A Muticast Delivery Scheme for VCR Operations in a Large VOD System,"8th IEEE International Conference on Parallel and Distributed Systems, pp.555~561, June 26-29, 2001.

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

[12] 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, pp. 295~335, 1994.

[13] Brian K. Schmidt, Monica S. Lam, J. Duane Northcutt, "The interactive performance of SLIM: a stateless, thin-client architecture," ACM SOSP'99, pp. 31~47, 1999.

[14] W.C. Feng and M. Lie, "Critical Bandwidth Allocation Techniques for Stored Video Delivery Across Best-Effort Networks,"20th International Conference on Distributed Computing Systems, pp. 201~207, April, 2000.

[15] Jung-Min Choi,Seung-Won Lee, Ki-Dong Chung, "A Muticast Delivery Scheme for VCR Operations in a Large VOD System,"8th IEEE International Conference on Parallel and Distributed Systems, pp.555~561, June 26-29, 2001.

[16] David A. Patterson and John L. Hennessy, "Computer Organization & Design," PP.392~490, Morgan Kaufmann, 1998.

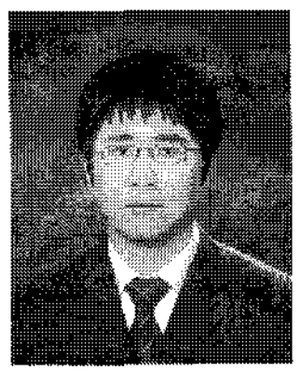
[17] Nabil J. Sarhan, Chita R. Das, "Caching and Scheduling in NAD-Based Multimedia Servers," IEEE Transactions on PARALLEL AND DISTRIBUTED SYSTEMS, Vol.15, No.10, pp.921~933, 2004.

[18] Sang-Ho Lee, Kyu-Young Whang, Yang-Sae Moon, Wook-Shin Han, "Dynamic Buffer Allocation in Video-on-Demand Systems," IEEE Transactions on PARALLEL AND DISTRIBUTED SYSTEMS, Vol.15, No.6 pp.1535~1551, 2003.

저자소개

이좌형 (Joa-Hyoung Lee)

2003년 강원대학교 정보통신공학과 (공학사)
2005년 강원대학교 컴퓨터정보통신공학과(공학석사)
2005년 ~ 현재 강원대학교 컴퓨터정보통신공학과 (박사과정)
※관심 분야: 멀티미디어 시스템, 센서 네트워크



박총명(Chong-Myung Park)

1995년 강원대학교 정보통신공학과 (공학사)
2007년 강원대학교 컴퓨터정보통신공학과 (공학석사)
2007년 ~ 현재 강원대학교 컴퓨터정보통신공학과 (박사과정)
※관심분야 : 센서네트워크, 멀티미디어 시스템

정 인 범(In-Bum Jung)

1985년 고려대학교 전자공학과 학사.
1985년~1995년 ㈜ 삼성전자 컴퓨터 시스템사업부 선임 연구원.
1992년~1994년 한국과학기술원 정보통신공학과 석사
1995년~2000년 8월 한국과학기술원 전산학과 박사
2001년~현재 강원대학교 컴퓨터정보통신공학전공 교수
※관심분야 : 운영체제, 소프트웨어 공학, 멀티미디어 시스템, 센서네트워크