PRR_{gp} 기반 결함허용 연속 매체 저장시스템의 설계와 성능평가

오 유 영† · 김 성 수††

요 약

VOD (Video On Demand) · MOD(Multimedia On Demand)와 같은 멀티미디어 시스템에서 동시에 여러 사용자에 의해 임의로 요청되고 저장 매체의 실시간 접근을 요구하는 연속 매체(continuous media) 인산을 지원하기 위한 고성능 저장 시스템으로 여러 개의 디스크들을 인동하는 디스크 배열을 사용하고 있다. 연속 매체를 실시간 접근하는 환경에서 연속 매체의 병렬 및 병행 처리를 위해 저장 시스템은 사용자의 요구를 독립적으로 처리할 수 있어야 한다. 본 논문에서는 기존에 제안된 신뢰도와 쉬의 속도 탐색을 고려한 연속 매체 디스크 배치 기법인 PRR_{gp} (Prime Round Robin with Grouped Parities) 방식에 근거한 결함허용 연속 매체 저장 시스템을 제안하고, 성능 평가를 위해서 RAID 5, Declustered 배열과 저장 공간 이용률, 신뢰도, 디스크 부하 균형, 버피 요구량, 서비스 가능한 사용자 수, 결함 복구 시의 오버헤드(overhead) 등을 비교 · 분석한다. 제안된 결함허용 연속 매체 저장 시스템은 임의 속도 탐색 시에 RAID 5, Declustered 배열보다 전반적으로 디스크 부하 균형을 효과적으로 만족시키고 실시간으로 보다 많은 사용자들에게 서비스를 제공한다.

Design And Performance Evaluation of Fault-Tolerant Continuous Media Storage System Based on PRR_{gp}

Yuyoung Oh† · Sungsoo Kim††

ABSTRACT

Multimedia Systems such as VOD (Video On Demand) and MOD (Multimedia On Demand) need to support continuous media operations which are randomly called by concurrent users and require that stored media be accessed in real-time. To satisfy such requirements, disk arrays consisting of multiple disks are generally used as storage systems. Under the real-time environments to provide users with accessing continuous media in the parallel and concurrent manner, storage systems should be able to deal with user requests independently. In this paper, we present a new fault-tolerant continuous media storage system called PADA (PRR_{gp} bAsed Disk Array), which is based on a PRR_{gp} (Prime Round Robin with Grouped Parities) disk placement scheme with enhanced reliability and load-balancing. We have compared and evaluated the storage space overhead for fault-tolerance, the reliability of disk array systems, the degree of disk load-balancing, the demanded buffer space, the maximum number of users being capable of supporting and the fault recovery overhead for PADA, RAID 5 and Declustered storage systems. According to the results, PADA is the best among them in that PADA satisfies load-balancing more effectively and serves more users in case of arbitrary-rate retrievals.

※ This work is supported in part by the Ministry of Education of Korea (Brain Korea 21 Project supervised by Korea Research Foundation)
† 준 회 원 삼성전자(주) 정보통신총괄
†† 정 회 원 아주대학교 BK21 정보통신전문대학원 교수
논문접수 1999년 4월 14일, 심사완료 2000년 3월 31일

1. 서 론

최근 저장 및 압축 기술과 통신 기술의 발전은 고속 네트워크 상에서 멀티미디어 데이터를 저장하고 접근하는 것을 가능하게 했다. 멀티미디어 데이터 유형 중에서 오디오와 비디오 같은 시간 종속적인 연속 매체는 사용자에게 서비스할 수 있는 상업적인 측면 때문에 많은 연구가 진행중이다. 멀티미디어 서비스를 제공하는 비디오·멀티미디어 서버는 클라이언트/서버 구조를 가지며 클라이언트는 서버의 연속 매체를 대화형으로 실시간 접근하게 된다. 그러나 현재 자기 디스크의 탐색 시간 및 전송 속도의 한계로 비디오 서버에서 동시에 제한된 수의 비디오 처리는 가능하지만 많은 사용자의 병행 접근을 지원하는 데에 한계가 있다. 연속 매체는 기존의 텍스트 위주의 데이터에 대한 탐색 방법과는 달리 서비스를 받는 사용자가 원하는 화면을 검색할 수 있도록 다양한 비디오 연산을 지원해야 한다[1-3]. 또한 연속 매체는 데이터 유형에 따라 저장 시스템의 입출력 대역폭이 다양하고 수백 수천 편의 비디오와 같은 연속 매체를 저장하기 위해서 대량의 저장 공간이 요구된다.

대량의 저장 공간을 제공하고 사용자 연산을 동시에 지원하기 위한 저장 시스템은 개별적인 디스크 전송률을 초과하는 전송률을 제공해야 한다. 디스크 배열 구조의 저장 시스템에서는 디스크 접근을 독립적으로 수행함으로써 통합된 디스크 전송률을 얻을 수 있다. 따라서 RAID(Redundant Array of Inexpensive Disks)[4-8] 기법이 널리 사용되고 있고, 디스크 스트라이핑(disk striping) 및 분할(declustering)[9-11] 기법을 사용하여 비디오 데이터를 처리하는 연구가 진행되었다.

본 연구에서는 임의 속도 탐색 시에 디스크 부하 균형을 만족시키고 실시간으로 보다 많은 사용자들을 서비스할 수 있도록 기존에 제안된 신뢰도(reliability)와 임의 속도 탐색을 고려한 연속 매체 디스크 배치 기법인 PRR_{gp}(Prime Round Robin with Grouped Parities) 방식[12-14]으로 연속 매체를 배치하는 디스크 배열 저장 구조를 갖는 결함허용 연속 매체 저장 시스템에 대해서 제안하고, 성능 평가를 위해서 RAID 5, Declustered 배열과 저장 공간 이용률, 신뢰도, 디스크 부하 균형, 버퍼(buffer) 요구량, 승인된 서비스 가능 사용자 수, 결함 복구 시의 오버헤드 등을 비교·분석한다. 본 논문은 2장에서 관련 연구로 PRR_{gp} 기법에 근거한 디스크 배열, RAID 5, Declustered 배열을 갖는 결함허용 저장 시스템에 대해서 설명하고, 3장에서는 임의 속도 탐색을 지원하는 결함허용 연속 매체 저장 시스템을 제안하고, 4장에서는 제안된 결함허용 연속 매체 저장 시스템과 RAID 5, Declustered 배열 저장 시스템의 성능을 비교·분석하고, 마지막으로 5장에서는 결론을 맺고 향후 계획에 대해서 논한다.

2. 관련 연구

VOD·MOD와 같은 시스템에서 동시에 여러 사용자들의 연속 매체에 대한 연산을 지원하기 위한 고성능 저장 시스템 구조로 디스크 배열을 사용하는 추세에 있지만, 디스크 배열은 다수의 대용량 디스크들로 구성되므로 평균 고장 발생 시간(MTTF : Mean Time To Failure)이 단일 디스크에 비해 짧기 때문에 디스크 배열에 결함허용 특성을 부가한 RAID 기법을 주로 사용하고 있고[4, 16], 디스크 스트라이핑 및 분할 기법을 사용하여 연속 매체를 처리하는 연구가 활발히 수행되었다. 이들 기법은 디스크 부하가 균등할 때 여러 디스크의 통합된 대역폭으로 데이터 접근이 가능하기 때문에 연속 매체를 효율적으로 처리하기 위해서는 부하 균형 문제를 해결해야 한다. 다양한 비디오 VCR 연산을 제공하는 방법으로 고속 탐색을 위해 미리 코딩된 고속 탐색 사본(fast-forward replica)을 사용할 경우에 별도의 저장 공간이 필요하고 제공된 속도 이외의 속도로 검색이 불가능하기 때문에 항등 비트율(CBR : Constant Bit Rate)로 압축된 비디오 데이터(MPEG-1 : 1.5Mbits/s)[1]에 대해서 세그먼트(segment) 단위로 라운드 로빈, 세그먼트 선택 기법으로 연속 매체를 배치시켰다. 이 배치 기법들은 임의 속도 탐색 시에 디스크 부하 균형을 효과적으로 만족시키지 못했다. 이 한계점을 극복하기 위해서 소수 라운드 로빈 기법[1]이 제시되었지만 저장 공간의 낭비와 신뢰도에 대한 고려가 없었다. 이 기법에서 낭비된 저장 공간에 결함허용을 위한 패리티를 배치하는 기법이 PRR_{gp}이다[12-14].

결함허용 연속 매체 저장 시스템 설계에 관한 많은 연구가 있었지만 대부분의 연구가 RAID 배열 구조를 바탕으로 미러링(mirroring), 패리티 인코딩(parity encoding)을 사용했다[16-18]. 미러링은 100%의 저장 공간 오버헤드가 있기 때문에 주로 패리티 인코딩이 RAID 3, 4, 5 배열에 사용되었다. 이 때 패리티 블록

과 결함 시에 패리티 계산을 위해 사용되는 데이터 블록들을 묶어서 패리티 그룹(parity group)이라고 정의한다. RAID 5 배열에서 데이터는 고정된 블록으로 디스크에 인터리빙(interleaving)되고 패리티 블록은 모든 디스크에 균등하게 분산 배치된다. RAID 5 배열에서 한 디스크의 결함 시에 패리티 그룹에 속한 나머지 모든 디스크들이 결함 복구를 위해서 접근되는 오버헤드가 따른다. 이 오버헤드를 줄이기 위한 방법으로 다중 RAID와 분할 패리티 배열 구조가 제안되었다. 다중 RAID는 디스크 배열을 구성하는 디스크들을 여러 개의 클러스터(cluster)로 나누어서 각 클러스터에 대해서 RAID 방법을 적용한 것이고, 분할 패리티 배열은 디스크들을 다중 RAID처럼 클러스터로 나누고 각 클러스터에 대해서 RAID보다 결함 복구를 위한 오버헤드를 줄이기 위해서 패리티 그룹에 속하는 디스크 수를 감소시킨 것이다[19]

3. 임의 속도 탐색을 지원하는 결함허용 연속 매체 저장 시스템

이 장에서는 임의 속도 탐색을 지원하는 새로운 결함허용 연속 매체 저장 시스템을 제안한다. 저장 시스템 수준에서 결함허용 저장 시스템 구조, 디스크의 결함을 허용할 수 있도록 패리티 인코딩에 기초한 디스크 배열의 연속 매체 디스크 배치 정책, 결함 발생 시에 결함복구 오버헤드를 줄이기 위한 결함복구 정책을 살펴본다.

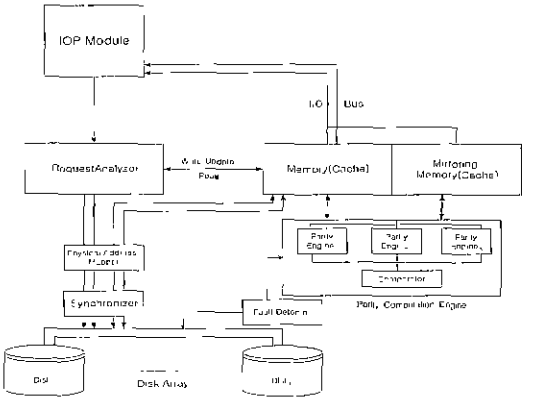
<표 1>은 논문에서 사용된 기호에 대한 설명을 보여준다.

〈표 1〉 논문에서 사용된 기호

기호	설 명
N	디스크 배열을 구성하는 디스크 수
N_p	N과 같거나 적은 최대 소수
M	패리티 배치를 위한 패리티 그룹의 수 $M = N - N_p$
G	Declustered 배열에서 패리티 그룹을 구성하는 디스크 수
d_r	1 배속 탐색 연산에 대한 라운드 거리(rounding distance)

3.1 결함허용 연속 매체 저장 시스템 구조

IOP(I/O Processor)는 클라이언트의 요청을 승인한 프로세서(CPU)로부터 디스크 배열 저장/검색 요청을 받아서 디스크 배열 제어기와 지능적으로 상호 작동하여 DMA(Direct Memory Access)를 사용하여 읽은 데이터를 네트워크 인터페이스로 전송하거나 주메모리로부터 읽은 데이터를 동일한 특성을 갖는 디스크들로 이루어진 디스크 배열에 저장한다. 디스크 배열 제어기는 많은 입출력 요구 처리와 결함허용을 제공하기 위해서 다수의 제어기들로 구성될 수도 있다. (그림 1)은 일반적인 결함허용 연속 매체 저장 시스템의 구조를 개략적으로 보여준다. 디스크 배열 제어기는 PhysicalAddressMapper로 가상 디스크의 주소와 물리적인 디스크의 주소를 매핑(mapping), RequestAnalyzer로 IOP로부터 입출력 요구의 분석 처리, 그리고 Fault Detecter와 Parity Computation Engine으로 결함허용을 위한 영구적 디스크 결함 감지와 결함 복구를 수행한다. 연속 매체의 검색은 실시간 서비스 보장을 위해서 디스크 배열 제어기의 Synchronizer에 의해 동기화가 맞춰진 후에 병렬로 동시에 접근이 이루어진다. 또한 연속 매체 검색 시에 디스크 배열 제어기의 메모리를 선인출(prefetching)에 대한 캐쉬(cache)로 사용함으로써 입출력 응답시간을 더욱 향상시킬 수 있다



〈그림 1〉 결함허용 연속 매체 저장 시스템 구조

3.2 연속 매체 디스크 배치 기법

패리티 인코딩를 기반으로 한 기존에 제안된 임의 속도 탐색과 신뢰도를 고려한 PRR_{gp} 기법[12-14]으로 연속 매체를 배치한다. 이 기법은 일정한 시간 간격으로 분할된 고정 길이의 세그먼트 단위(70KB)[1]로 연속 매체를 데이터 블록으로 나누고 M개의 패리티 그룹에 대해서 결함허용을 위한 패리티 블록을 계산한 후에 사상함수[12, 14]에 의해서 데이터와 패리티 블록을 디스크에 저장한다. N_p 배속을 제외한 임의 속도

탐색 연산 시에 디스크 부하 균형을 만족시킬 수 있고 동시에 발생한 두 개의 영구적(permanent) 디스크 결함 중에서 약 30% 이상의 경우에 대해서 저장된 패리티 정보를 이용한 복구가 가능하다

3.3 결함 복구 정책

한 디스크의 영구적 결함이 감지되는 시점에 패리티 그룹에 속한 나머지 디스크들에 접근함으로써 패리티 정보를 이용해서 복구하는 방법은 중복해서 접근해야 하는 일부 디스크들로 인해서 일시적인 부하가 증가하게 됨으로서 서비스 중인 사용자에 대한 실시간 제약을 만족시키지 못할 수도 있다 따라서 서비스가 승인된 후에 처음 한 라운드(round)의 서비스 주기에 사용자가 요구한 연속 매체에 대해서 미리 선인출한 후에 다음 라운드에 선인출한 데이터를 서비스하게 된다. 서비스가 승인된 후에 초기 지연이 발생하고 결함을 고려하지 않을 경우에 요구되는 버퍼보다 많은 버퍼가 필요하지만 결함이 발생하더라도 항상 결함을 복구할 충분한 시간(한 라운드)을 갖는다.

4. 성능 평가

3장에서 살펴본 본 논문에서 제안된 결함허용 연속 매체 저장 시스템(PADA : PRR$_{gp}$-bAsed Disk Array)과 RAID 5, Declustered 배열을 저장 공간 오버헤드 비율, 디스크 부하 균형, 신뢰도, 버퍼 요구량, 최대 서비스 가능한 사용자 수, 결함 복구 오버헤드 측면에서 비교·분석한다. 특히, 패리티 정보가 다수의 디스크에 분산 배치되는 측면에서 RAID 5와 Declustered 배열을 PADA와 비교되는 저장 시스템으로 선택했다. RAID 5, Declustered 배열에서 클러스터 개수는 한 개, RAID 5 배열은 왼쪽 비대칭(left asymmetric) RAID 5, Declustered 배열에서 패리티 그룹에 속한 디스크 수(G)가 $N/2$보다 크다고 가정한다

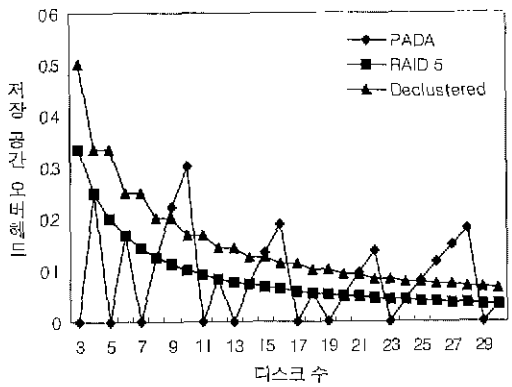
4.1 결함허용을 위한 저장 공간 오버헤드

결함허용을 위한 저장 공간 오버헤드는 결함허용 연속 매체 저장 시스템의 디스크 배열에서 데이터와 패리티 블록이 배치된 전체 저장 공간 중에서 결함허용을 위한 패리티 정보가 저장된 패리티 블록이 차지하는 비율로 정의된다. <표 2>에 PADA, RAID 5, Declustered 배열에 대한 저장 공간 오버헤드 계산식이 있다. (그림 2)는 각 저장 시스템에 대해서 N을 3에서 30까지 변화시키고 G = N/2+1일 경우에는 저장 공간 오버헤드 비율을 보여준다. PADA에서 N이 소수일 경우에 오버헤드가 없다. 즉 결함허용을 위한 패리티 정

〈표 2〉 결함허용 저장 시스템에 대한 성능 분석 계산식

Performance Metric \ Storage Systems		PADA	RAID 5	Declustered
저장 공간 오버헤드		$\frac{M}{N}$	$\frac{1}{N}$	$\frac{1}{G}$
MTTDL		$\frac{MTTF^{M+1}}{N\times(N-1)\times\cdots\times(N-M)\times MTTR^{M}}$, M=0, 1 $\frac{MTTF^{2}}{N\times(N-1)\times MTTR} \le$ MTTDL $\le$ $\frac{MTTF^{M+1}}{N\times(N-1)\times\cdots\times(N-M)\times MTTR^{M}}$, M ≥ 2	$\frac{MTTF^{2}}{N\times(N-1)\times MTTR}$	$\frac{MTTF^{2}}{N\times(N-1)\times MTTR}$
신뢰도		R_P[12, 14]	$\sum_{k=0}^{1}\binom{N}{k}\times Rel^{N-k}\times(1-Rel)^{k}$	$\sum_{k=0}^{1}\binom{N}{k}\times Rel^{N-k}\times(1-Rel)^{k}$
결함 복구 시에 부하 오버헤드		$\frac{\lceil \frac{N}{M} \rceil - 1}{N-1}$	$\frac{N-1}{N-1}$	$\frac{N-G}{N-1}$
버퍼 요구량	결함 고려 안함	$n^{PADA}b(k^{max})$	$n^{RAID\ 5}b(k^{max})$	$n^{Declustered}b(k^{max})$
	결함 고려 함	$n^{PADA}b(k^{max}+\lceil \frac{N}{M} \rceil - 1)$	$n^{RAID\ 5}b(k^{max}+N-1)$	$n^{Declustered}b(k^{max}+G-1)$

보를 저장하고 있지 않다.



(그림 2) 결함허용을 위한 저장 공간 오버헤드
(N : 변화)

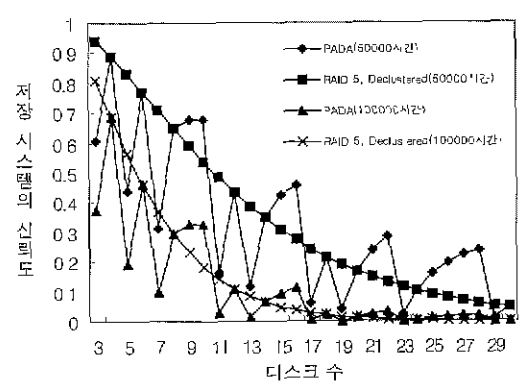
4.2 평균 데이터 손실 시간 및 신뢰도

독립적인 영구적 디스크 결함만을 고려하는 경우에는 결함허용을 위해 저장된 패리티 정보만으로도 디스크 결함으로부터 손실된 데이터를 복구할 수 있다. RAID 5, Declustered 배열에 대한 평균 데이터 손실 시간(MTTDL : Mean Time To Data Loss) 계산식은 [4]를 참조하였다. PADA에 대한 평균 데이터 손실 시간 계산식은 <표 2>에서와 같이 유도될 수 있다. <표 3>은 N을 3에서 30까지 변화시키고 한 디스크에 대한 평균 고장 발생 시간을 300,000 시간, 평균 수리 시간(MTTR : Mean Time To Repair)을 2 시간으로 가정할 경우에 평균 데이터 손실 시간을 보여준다. 디스크 배열을 구성하는 N개의 디스크를 동일한 것으로 간주할 때 한 디스크에 대한 신뢰도(Rel)는 $e^{-\lambda t}$에 의해서 구해진다[20, 21]. 여기서 λ는 일정한 결함발생율이고 t는 시간이다. [12, 14]에 PADA에 대한 조합모델(combinatorial model)을 이용한 신뢰도(R_P) 계산식이 유도되었다 <표 2>에 PADA, RAID 5, Declustered 배열에 대한 신뢰도(R_R, R_D) 계산식이 있다. (그림 3)은 N을 3에서 30까지 변화시킬 경우에 조합모델로 계산된 신뢰도를 보여준다. RAID 5와 Declusstered 배열은 한 개의 영구적 디스크 결함만이 복구 가능하지만, PADA는 PRR_{gp} 디스크 베치 기법에서 패리티 그룹 수만큼의 결함허용이 가능하기 때문에 비교되는 두 개의 저장시스템들보다 높은 신뢰도를 갖는다.

〈표 3〉 평균 데이터 손실 시간

(단위 : 년)

N	PADA	RAID 5, Declustered
3	11	856164
4	428082	428082
5	7	256849
6	171233	171233
7	5	122309
8	91732	91732
9	71347~1528864971	71347
10	57078~22932974559687	57078
11	3	46700
12	38917	38917
13	3	32929
14	28225	28225
15	24462~282251885	24462
16	21404~2646112449195	21404
17	2	18886
18	16788	16788
19	2	15020
20	13518	13518
21	12231~96559893	12231
22	11119~658362906020	11119
23	1	10152
24	9306	9306
25	8562~56836807	8562
26	7903~322135428598	7903
27	7318~17896112669986714	7318
28	6795~958736394357395600	6795
29	1	6326
30	5905	5905

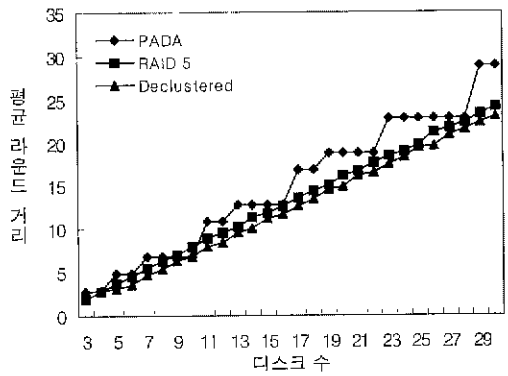


(그림 3) 저장 시스템의 신뢰도
(N : 변화, λ 0.0000033, t : 50000, 100000 hour)

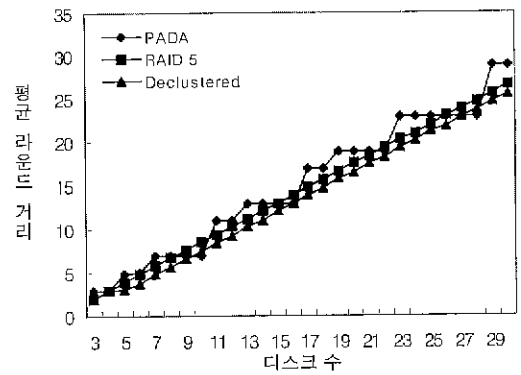
4.3 디스크 부하 균형

본 논문에서는 특정 속도의 탐색에 대한 디스크 접근의 균형 정도를 평가하는 척도로 평균 라운드 거리[1]를 이용한다. 연속 매체가 다수의 디스크에 분산 저장되어 있는 경우에 탐색 연산을 위해 디스크에 저장된 데이터 블록을 접근하게 되는데, 이 때 디스크의 최소 접근 주기를 라운드 거리라고 한다. 디스플레이할 데이터 블록의 간격은 탐색 속도에 따라 변하므로 라운드 거리도 달라진다. 식 (1)은 평균 라운드 거리(d)를 구하는 계산식이다. 이 식에서 w는 검색 연산 중에서 재생이 차지하는 가중치 값이다. (그림 4. 5)는 각각 w가 0.8, 0.9일 경우에, N을 3에서 30으로 변화시킬 때에 평균 라운드 거리를 보여준다.

$$d = wd_1 + (1-w)\sum_{i=2}^{N}\frac{d_i}{(N-1)} \qquad (1)$$



(그림 4) 고속 탐색 연산 시에 평균 라운드 거리 (N : 변화, w : 0.8)



(그림 5) 고속 탐색 연산 시에 평균 라운드 거리 (N : 변화, w 0.9)

4.4 버퍼 요구량

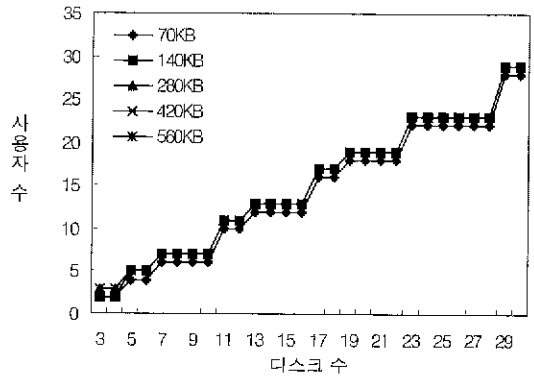
본 논문에서는 각 저장 시스템마다 디스크 배열을 구성하는 디스크 수에 따라서 저장되는 데이터와 패리티 블록의 수가 다르기 때문에 승인되는 사용자 수가 다르지만 서비스 가능한 최대 사용자 수를 각각 n^{PADA}, $n^{RAID\ 5}$, $n^{Declustered}$으로, 한 라운드에서 한 사용자에 대한 최대 데이터 블록 수를 k^{max}로, 한 블록의 크기를 b로 가정한다. 결함 발생을 고려하지 않을 경우와 고려할 경우에 요구되는 최대 버퍼량은 <표 2>의 계산식을 이용하여 구할 수 있다.

4.5 사용자 수용 능력

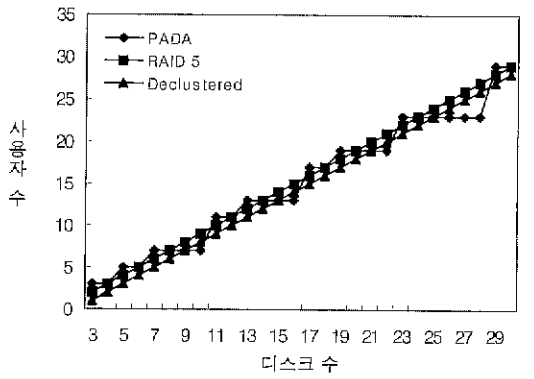
연속 매체는 저장 시스템의 디스크 배열 내의 디스크에 블록 스트라이핑 되어 저장되므로 매 $T_{display}$ 시간(블록당 재생시간)마다 발생하는 데이터 블록의 검색 요청은 여러 디스크에 분산되어, 각 디스크에는 한 블록의 검색 후 $m * T_{display}$ 시간(한 라운드) 이후에 검색 요구가 도착한다. 따라서 각 디스크는 $m * T_{display}$ 시간 이내에 각 블록의 검색을 완료해야 한다[19, 22]. 여기서 m은 검색에 참여한 디스크 개수이다. <표 4>는 사용자 수용 능력에 대한 성능 분석을 위해서 사용된 디스크의 특성에 관한 파라미터를 보여준다. 식 (2), (3)에 의해 최악의 디스크 탐색 시간($l_{seek}(t_1, t_2) = a + b \times |t_1 - t_2|$), 회전 지연 시간($l_{rot}^{max}$), 한 라운드에서 각 사용자마다 한 디스크에서 검색되는 최대 디스크 블록 수(k_i^{max}) 등을 고려하고 $T_{nondisk(n)}$은 영향을 주지 않는 인자로 가정할 경우에 서비스 가능한 사용자 수를 구할 수 있다. 이때 스트라이핑 되는 데이터 블록은 세그먼트의 배수가 되고, 최대의 성능을 발휘할 수 있는 데이터 블록 크기(스트라이핑 유닛)가 결정되어야 한다. (그림 6)은 스트라이핑 유닛의 변화에 따른 PADA의 서비스 가능한 사용자 수를 보여준다. 스트라이핑 유닛의 크기가 커짐에 따라서 점차로 사용자 수

<표 4> 성능 분석에 사용된 디스크 특성 파라미터

디스크 용량	0.5 GB
실린더 수	1024
트랙(면) 수(T)	1024
디스크 블록 크기	4 KB
디스크 회전 속도	3600 RPM
$l_{seek}(t_1, t_2)$	$4 + 0.02 \times \|t_1 - t_2\| ms$
최대 탐색 시간	24.48 ms
최대 회전 지연 시간	16.66 ms



(그림 6) 최대 사용자 수를 산출하는 데이터 블록 크기 결정(N · 변화)



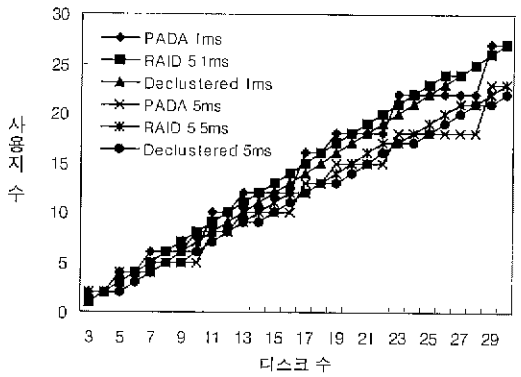
(그림 7) 결함이 없을 때에 서비스 가능한 사용자 수 (N 변화, 데이터 블록 크기 · 420KB)

가 증가하다가 420KB 이후로 전송 시간 오버헤드가 큰 영향을 주기 때문에 더 이상 변화가 없으므로 성능 분석을 위한 최적화된 스트라이핑 유닛을 420KB로 가정한다. (그림 7)은 N을 3에서 30으로 변화시킬 경우에 결함이 없는 정상적인 상태에서 서비스 가능한 사용자 수를 보여준다

$$m \times T_{display} \geq T_{disk(n)} + T_{roundisk(n)} + T_{recovery} \quad (2)$$

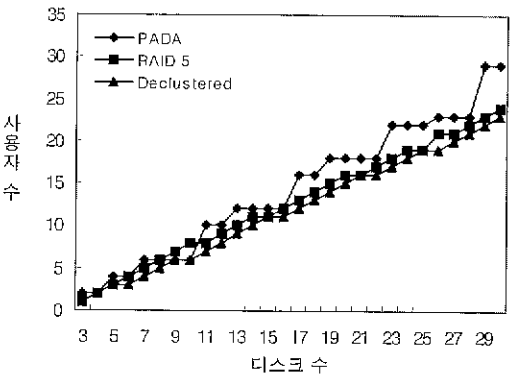
$$T_{disk(n)} = b \times T + (a + l_{tot}^{\max} \times \sum_{i=1}^{n} k_i^{\max}) \quad (3)$$

결함 발생 시에 식 (2)에서 결함 복구에 대한 오버헤드($T_{recovery}$), 즉 패리티 정보를 이용한 패리티 연산에 소요되는 시간을 고려할 경우에 서비스 가능한 사용자 수를 구할 수 있다. 이 때에 라운드마다 각 사용자는 많아야 한번 결함이 발생한 디스크에 접근한다. (그림 8)은 결함 복구를 위한 오버헤드가 1ms, 5ms일 경우에 N을 3에서 30으로 변화시킬 때에 서비스 가능한 사용자 수를 보여준다.



(그림 8) 결함 복구를 고려한 경우에 서비스 가능한 사용자 수(N 변화, 데이터 블록 크기 : 420KB)

식 (2), (3)에서 제생 연산 외에 임의 배속 탐색 연산(고속 탐색)을 고려할 경우에 서비스 가능한 사용자 수를 구할 수 있다 이 때 N과 요청된 임의 배속 탐색 연산에 따라서 일부 디스크에 부하가 집중되므로 디스크에 대한 통합된 대역폭을 제공하는 병렬 접근이 이루어지지 않기 때문에 $T_{disk(n)}$에 대한 값이 증가하고 m은 감소한다. (그림 9)는 w가 0.8일 경우에 N을 3에서 30으로 변화시킬 때에 서비스 가능한 사용자 수를 보여준다.



(그림 9) 임의 배속 탐색을 고려한 경우에 서비스 가능한 사용자수
(N · 변화, 데이터 블록 크기 · 420KB, w · 0.8)

5. 결론 및 향후 계획

본 논문에서는 기존에 제안된 신뢰도와 임의 속도 탐색을 고려한 연속 매체 디스크 배치 기법인 PRR_{gp} 방식에 근거한 결함허용 연속 매체 저장 시스템의 구조, 디스크 배치 기법, 결함 복구 정책을 제안하고, 제시된 결함허용 연속 매체 저장 시스템(PADA)의 성능 평가를 위해서 RAID 5, Declustered 배열 저장 시스템과 저장 공간 이용률, 신뢰도, 디스크 부하 균형, 버퍼 요구량, 서비스 가능한 사용자 수, 결함 복구 시의 오버헤드 등을 비교·분석했다. 제안된 결함허용 연속 매체 저장 시스템은 임의 속도 탐색 시에 비교되는 기존의 저장 시스템들보다 효과적으로 디스크 부하 균형을 만족시켰고 실시간으로 보다 많은 사용자들에게 서비스를 제공했다. 향후에 저장 시스템 용량에 대해서 최악을 가정하지 않은 유동적인 승인 제어(admission control)와 복구 알고리즘을 개발하여 실질적으로 서비스 가능한 사용자 수를 획득할 것이다. 또한 대역폭의 낭비를 줄이면서 가변 비트율(VBR : Variable Bit Rate)로 압축된 연속 매체를 고려한 저장 시스템을 제시할 것이고, 독립적이고 영구적인 디스크 결함 외에도 일시적(transient) 디스크 결함 및 디스크 배열 제어기에서의 결함 등을 수용한 성능 분석을 수행할 예정이다.

참 고 문 헌

[1] 권택근, *연속 매체 저장 시스템에서 디스크 입출력 성능 향상 기법*, 박사학위논문, 서울대학교 컴퓨터 공학과, 1996.

[2] D.J. Gemmel, H.M. Vin, D.D. Kandlur and P.V Rangan, "Multimedia Storage Servers : A Tutorial and Survey," *IEEE Computer*, pp.40-49, May 1995.

[3] M.S. Chen, D Kanulur and P. Yu, "Support for Fully Interactive Playout In Disk-Array-Based Video Server," in *Proc. of ACM Multimedia*, 1994, pp.391-398.

[4] P Chen, E. Lee, G. Gibson and D Patterson, "RAID : High-Performance, Reliable Secondary Storage," *ACM Computing Surveys*, pp.145-186, June 1994.

[5] D. Patterson, G Gibson and R Katz, "A Case for Redundant Array of Inexpensive Disks(RAID)," in *Proc. of ACM SIGMOD'88*, June 1988, pp.109-116.

[6] G. Gibson, *Redundant Disk Arrays : Reliable, Parallel Secondary Storage* PhD thesis Univ. of California at Berkeley, December 1991.

[7] E.K. Lee, *Performance Modeling and Analysis of Disk Arrays* PhD thesis, Univ of California at Berkeley, December 1993.

[8] G.R. Ganger, B.L. Worthington, R.Y. Hou and Y.N. Patt, "Disk Arrays : High Performance, High-Reliability Storage Subsystems," *IEEE Computer*, pp.30-36, March 1994.

[9] P.J. Shenoy and H.M. Vin, "Failure Recovery Algorithms for Multi-Disk Multimedia Servers," *Technical Report 96-06*, Univ. of Texas at Austin, 1996.

[10] M. Holland and G. Gibson, "Parity Declustering for Continuous Operation in Redundant Disk Arrays," in *Proc. of Architectural Support for Programming Languages and Operating Systems(ASPLOS)*, October 1992, pp.23-35.

[11] A. Merchant and P.S. Yu, "Design and Modeling of Clustered RAID," in *Proc. of FTCS-22*, June 1992, pp.140-149.

[12] 오유영, 권원석, 김성수, 강창훈, "신뢰도와 임의 속도 탐색을 고려한 연속 매체 디스크 배치 기법", *1998년 한국정보과학회 춘계학술발표논문집*, 제25권 제1호(A), pp.45-47, 1998. 4.

[13] 오유영, 김성수, "결함복구율을 고려한 연속 매체 디스크 배열의 신뢰도 분석", *1998년 한국정보과학회 추계학술발표논문집*, 제25권 제2호(A), pp.9-11, 1998. 10.

[14] 오유영, 김성수, 김재훈, "결함허용과 임의 속도 탐색을 고려한 연속 매체 디스크 배치 기법", *한국정보과학회, 정보과학회논문지(A)*, 제26권 제9호(A), pp.1166-1176, 1999. 9.

[15] P. Cao, S.B. Lim, S Venkatraman and J. Wilkes,

"The TickerTAIP parallel RAID Architecture." in *Proc. of International Symposium on Computer Architecture(ISCA)*, May 1993, pp.52-63.

[16] 오유영, 김성수, "임의 속도 탐색을 지원하는 결함허용 연속 매체 저장 시스템", *1999년 한국정보과학회 춘계학술발표논문집*, 제26권 제1호(A), 1999. 4.

[17] S. Berson, L. Golubchick and R.R. Muntz. "Fault Tolerant Design of Multimedia Servers," in *Proc. of the ACM SIGMOD Conference*, 1995, pp.364-375.

[18] B. Ozden, R. Rastogi, P.J. Shenoy and A. Silberschatz, "Fault-tolerant Architecture for Continuous Media Servers." in *Proc. of the ACM SIGMOD Conference*, June 1996, pp 79-90.

[19] H.M. Vin, A Goyal and P. Goyal, "Algorithms for Designing Large-Scale Multimedia Servers," *Computer Communications*, pp.192-203, March 1995.

[20] D.K. Pradhan, *Fault-Tolerant Computer System Design*. Englewood Cliffs, NJ : Prentice-Hall, 1996.

[21] B.W. Johnson, *Design and Analysis of Fault-Tolerant Digital Systems*. Addison- Wesley Publishing Company. 1989.

[22] D. Anderson, Y Osawa and R. Govindan, "A File System for Continuous Media." *ACM Transactions on Computer Systems 10(4)*, pp.311-337, November 1992.

오 유 영

e-mail : yyoh2430@hanmail.net
1998년 아주대학교 정보 및 컴퓨터공학부(공학사)
2000년 아주대학교 BK21 정보통신전문대학원(공학석사)
2000년~현재 삼성전자(주) 정보통신총괄
관심 분야 : 정보통신, 결함허용 시스템, 실시간 시스템, 분산 객체 등

김 성 수

e-mail : sskim@madang.ajou.ac.kr
1982년 서강대학교 전자공학과 (공학사)
1984년 서강대학교 전자공학과 (공학석사)
1995년 Texas A&M University, Computer Science Dept. (공학박사)
1983년~1986년 삼성전자(주) 종합연구소 컴퓨터연구실 (주임연구원)
1986년~1996년 삼성종합기술원(수석연구원)
1991년~1992년 Texas Transportation Institute(연구원)
1993년~1995년 Texas A&M University, Computer Science Dept.(T.A. & R.A.)
1997년~1998년 한국정보처리학회, 한국정보과학회 논문지 편집위원
1996년~현재 아주대학교 BK21 정보통신전문대학원 부교수
관심분야 : 멀티미디어, 결함허용, 디지털방송시스템, 이동컴퓨팅, 성능평가 등