RAID 시스템의 Log 관리에 의한 서버효율성 향상기법연구

박우영*, 김재중**, 곽계달***, 신승중****

*한양대학교 공학대학원 컴퓨터공학과
**한양대학교 신뢰성분석연구센터
***한양대학교 전기전자컴퓨터공학부
****한세대학교 IT 학부

e-mail : expersin@hansei.ac.kr

Scheme for Raising Effectiveness of Servers by Log Management of RAID System

Woo-Young Park*, Jae-Jung Kim**, Kae-Dal Kwack*** and Seung-Jung Shin****

*Division of Electronics and Computer Engineering, Hanyang University
**Reliability Analysis Research Center, Hanyang University
*** Dept. of Electronics and Computer Engineering, Hanyang University
****Dept. of Information Technology, Hansei University

요 약

이 Paper 는 Server 솔루션에서 사용되는 RAID 컨트롤러가 Disk 상의 에러(Error)를 처리하는 behavior 를 분석하고 이를 개선할 수 있는 방법을 제안한다. RAID Level 중에 가장 기본인 RAID 0 는 주로 Internet portal site 의 Search Engines 이나 Media streaming Servers, 그리고 HD 영상 편집 시스템에서 사용된다. 그러나 Fault tolerance 가 지원되지 않기 때문에 Data Volume(logical Disk)의 보존성이 약한 취약점을 가지고 있다. 따라서 이것을 개선하고 Data 보존성을 높일 수 있는 방안을 제시한다.

1. Introduction

RAID 는 다수의 Disk 들을 연결하여 Large capacity 를 구성하기 위해 처음 디자인되었다. 그러나 근래에는 I/O Performance 와 stability 에보다 중점을 두고 있다. RAID 는 목적에 따라 여러 Level 이 존재한다. 성능에 중점을 둔 RAID 0 외에도 Disk Mirroring 을 지원하는 RAID 1, Parity 기반의 Stripe set 을 지원하는 Level 5 와 6 이 있다. 또한 복합 Level 인 10, 50, 60 이 존재한다. 가장 기본 Level 인 RAID 0 을 살펴보면 **Fig. 1** 처럼 구성된다.

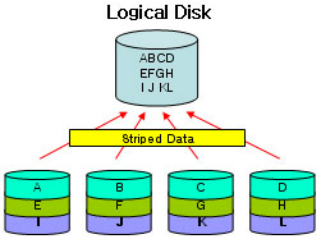

Fig.1 RAID 0 Array. Striped Disks

2 개 이상의 Disk 로 구성 가능한 RAID 0 는 각각의 디스크(Disk)에 Data 를 분산하여 저장하고 읽어들인다. 멤버 디스크의 숫자가 증가할수록 Logical Disk 의 Capacity 와 Performance 역시 증가하게 된다. 하지만 멤버 디스크가 n 개로 증가하면 n 배의 위험율이 증가하는 단점이 있다. 멤버 Disks 중에 1 개라도 문제가 발생하면 전체 Logical Disk 가 손상을 입게되어 Data 를 잃게된다.

2. BASIC STRUCTURE : ERROR 처리 BEHAVIOR

전통적인 RAID 0 의 디스크 에러 처리 behavior 를 살펴보면 **Fig. 2** 와 같은 로직으로 되어 있다.

OS 에서 Read 요청이 RAID 컨트롤러에 보내지면, 컨트롤러는 Logical Disk 의 특정 영역에 있는 데이터를 읽기 위해 각각의 디스크에 Read 명령을 보내게 된다. Data 는 물리적인 디스크에 분산되어 저장되어 있으므로 각각의 member disks 에서 모두 동시에 읽기가 발생한다.

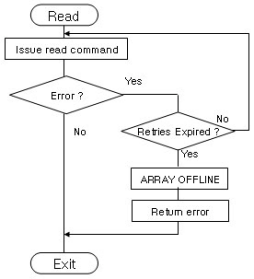

Fig. 2 RAID 0 에서 디스크 에러처리 Behavior (Previous)

Fig. 2 에서는 멤버 디스크에 에러가 있어서 특정 Sector 가 읽혀지지 않는 상황을 가정한다. 전통적인 RAID 0 behavior 에서 하나의 에러라도 발견되면 RAID controller 는 해당 디스크를 Fail 처리한다. 결국 Logical Disk 는 파괴되고 사용자는 모든 데이터를 잃게된다. 결국 디스크내의 1 개의 Sector(512Byte)만 문제가 있어도 전체 데이터를 잃게되는 상황이 발생한다.

Disk Fail condition	Uncorrectable Error	Command Timeout	I/O error	Etc
Rate	50%	30%	15%	5%
Remark	Bad sector	Non-response from HDD	PHY error, PCB 불량	Others

Table 1. RAID Controller 의 Disk fail 컨디션

Table 1.을 보면, RAID 컨트롤러에서 Disk 가 Fail 되는 가장 큰 원인은 Uncorrectable error(Bad sector; Medium error)이다.

Fig. 3 를 보게 되면, Medium error (Read/write error)의 처리 Behavior 가 일부 변경된 것을 알수 있다. 이것은 Disk fail 컨디션중 가장 큰 문제가 되는 bad sector 에 대한 처리 방법의 변경을 의미한다.

수정된 behavior 에 따르면, 일부의 Sector 가 문제가 있더라도 RAID Controller 는 멤더 Disk 를 Fail 처리하지 않아 Logical Disk 를 보존한다. 일부의 데이터는 손실되겠지만, 대부분의 데이터는 아직 읽기가 가능한 상태가된다. 따라서, RAID 0 의 취약점의 일부를 해결할 수 있다.

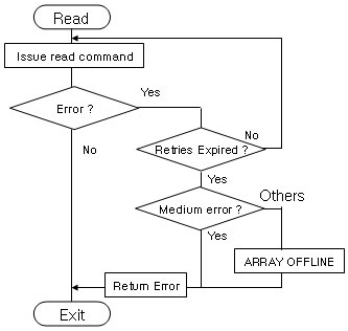

Fig. 3 Medium error 처리 루틴을 변경한 Behavior (current)

현재 일부의 RAID Controller 들은 **Fig. 3** 과 같은 Behavior 를 지원한다. 이로인해 RAID 0 의 데이터 보존성이 상당히 증가하였음이 이미 Field sites 에서 검증되었다. 하지만, 이 behavior 에도 큰 약점이 있다. 그것은 Command timeout 에 대한 방안이 고려되어 있지 않다는 것이다.

3. ENHANCED NEW BEHAVIOR FOR TIMEOUT

Disk 에서 Uncorrectable error 가 발생할 때, Disk 는 내부에서 에러 복구를 위한 Retry 를 반복한다. HDD 에 따라 대략 100-133 번 정도의 내부적인 Retry 가 있다. 이것은 디스크내부에서 Delay time 을 발생시키며, RAID controller 관점에서는 Command Timeout 을 유발할 수 있다. RAID Controller 는 하나의 명령을 보낼때마다 Timeframe 을 함께 설정한다. 따라서 정해진 시간이 지나도 Disk 에서 응답이 없으면 몇번의 Retry 를 반복 후, 해당 Disk 를 Fail 처리한다. 따라서 **Fig.3** 의 Medium error 의 판단조건에서 Other error 에 대한 부분의 개선이 필요하다. (Table 1. 참조)

Medium error 로 인해 유발된 Command timeout 문제를 해결하기 위해서는 H/W 기반의 보조가 필요하다. 현재 서버 시장에 공급되고있는 H/W RAID controller 들은 Memory 와 IOP(IO Processor)를 탑재하고 있으며, 에러로그 핸들링을 위한 NVRAM 을 장착한 제품들이 다수 있다. 또한 차후 출시될 제품들도 이런 기반이 고려되고 있다.

Fig. 4 는 RAID controller 상의 NVRAM 에 저장할 log data 의 형식이며, Disks 의 오류정보와 LBA 를 기록한다. Error log 를 무한정으로 저장할 수 없으므로 NVRAM 의 Size 를 고려하여 한 디스크당 10-20 개 정도의 LBA 정보를 저장한다. 이보다 많은 에러가 발생한다면, Server Idle Time 에 새 Disk 들을 사용하여 Logical Disk 를 재구성하는 것이 바람직하다.

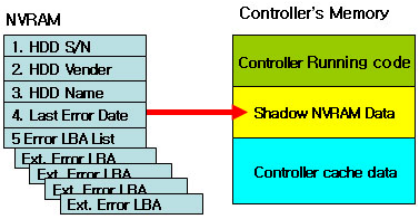

Fig.4 NVRAM mapping for error LBA list

RAID I/O Performance 를 감소시키지 않기 위해 **Fig. 4** 처럼 NVRAM 은 Controller 의 Memroy 에 Shadow 되어야 한다. log 를 읽을때는 Memory 의 Shadow 된 Data 를 읽으며, log 저장시에만 실제의 NVRAM 에 접근한다.

Fig. 5 의 behavior 는 디스크에 Medium error 가 발생시 이를 NVRAM 의 특정 영역에 기록한다. 가 OS 가 해당 LBA 에 Read 요청이 있으면, 해당 디스크의 LBA 를 액세스하지 않고 Error Code 를 OS 에 Return 한다. 따라서, 오류가 있는 Sector 를 액세스하지 않아도 됨으로 Timeout 이 유발될 수 있는 상황을 배제하게 된다.

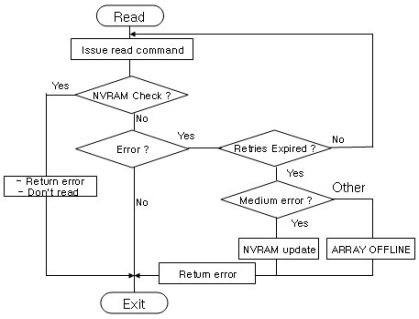

Fig. 5 NVRAM based new behavior for medium error

4. EXPECTED RESULTS

Table 2. 는 NVRAM 기반의 behavior 가 기존의 behavior 와 어떻게 다른지를 보여준다.

Behavior	Processing	Result	Remark
Previous Behavior	1 개의 Disk fail 로 Logical Disk 파괴	전체 Data 를 잃게됨	F/W
Current Behavior	Medium error 무시	Logical Disk 보존성 향상 Timeout 가능성	F/W
Enhanced Behavior	Medium error 가 있는 LBA 를 NVRAM 에 기록후 액세스하지 않음	Medium error 로 인한 Timeout 발생 빈도 감소	F/W &H/W

Table 2. 각각의 Behavior 비교

5. CONCLUSIONS

NVRAM 기반의 Enhanced Behavior 는 RAID 0 외에도 Fault tolerance 가 지원되는 1, 5, 6, 10, 50 , 60 에 확대 적용이 가능하다. Enhanced Behavior 는 Logical Disk 의 보존성에 초점을 맞춘것이며, 물리적인 Disk 가 Ideal 하지 않고 항상 Medium error 의 소지가 있음을 고려한 것이다. RAID 멤버 Disk 의 모든 Fail 컨디션을 해결할 수는 없지만 주요 Fail 원인을 배제함으로써 과거보다 향상된 보존성의 증가가 예상되며, 이로인한 Server Downtime 의 감소로 인해 서버의 가용성 및 효율성 향상을 기대할 수 있다.