데이터 가용성(HA)의 효율적인 지원을 위한 RAID 에러 핸들링 정책

강동재*, 박유현*, 김영호*, 김창수*, 신범주*
*한국전자통신연구원 컴퓨터시스템 연구부
e-mail : {djkang, yhbak, kyh05, cskim7, bjshin}@etri.re.kr

The Error-Handling Strategies of RAID for Data Availability

Dong-Jae Kang*, Yu-Hyeon Bak*, Young-Ho Kim*,
Chang-Soo Kim*, Bum-Joo Shin*
*Dept of Computer & System, ETRI

요약

본 논문에서는 SANtopiaVM의 데이터 가용성 및 성능을 고려한 에러 핸들링 정책을 제안한다. 제안하는 에러 핸들링 정책은 RAID1과 RAID5에 대한 정책으로 구분하며 에러 발생시의 복구를 위하여 RAID1에서는 FBB(Failed Block Bitmap)라는 비트맵을 추가하여 데이터의 일관성을 유지하고 디스크의 복구 시에는 읽기 연산에 대한 부하 분산 및 복구 비용을 감소시키는 쓰기 연산을 제안 함으로서 에러 핸들링 시에 입출력 비용을 줄인다. RAID5에서는 추가적인 여분 디스크를 사용한 Sparing Disk 기법을 제안함으로서 디스크 에러 모드에서도 정상 모드의 성능에 근접하는 입출력 성능을 보장하며 빠른 디스크 복구를 지원한다. 제안하는 에러 핸들링 정책은 SANtopiaVM RAID의 오류 발생시, 시스템 성능의 급격한 저하를 완화할 수 있으며 에러로부터 빠른 복구를 지원 함으로서 데이터에 대한 효율적인 고 가용성의 특징을 제공한다.

1. 서 론

볼륨관리자는 여러 개의 물리적 디바이스를 하나의 논리적 디바이스로 인식할 수 있도록 함으로서 스토리지 관리의 효율성을 제공하고 고성능의 데이터 입출력과 결합을 허용하는 고가용성의 특징을 지원한다. 이러한 특징을 제공하기 위해서는 물리적 디바이스들을 논리적인 가상 디바이스로 인식되도록 하는 디바이스의 가상화(virtualization) 기능과 물리적 디바이스 영역과 가상 디바이스 영역 사이의 주소 연결 관리(mapping) 및 디스크 오류 시에 저장된 데이터의 손실이나 부정확성으로부터 보호할 수 있는 기능이 충족되어져야 한다.[1]

본 논문에서는 SANtopiaVM RAID에서 오류 발생시 시스템의 서비스가 중단됨이 없이 데이터의 가용성을 지원하는 에러 핸들링 정책을 제안하며 RAID1과 RAID5의 정책에 대하여 기술한다. 본 논문에서는 RAID에서의 에러 핸들링을 위하여 *RAID1*에서 FBB(Failed Block Bitmap)라는 메타 데이터를 디스크에 유지함으로서 중복 데이터의 일관성을 유지하고 디스크의 복구 시에는 읽기 연산에 대한 부하 분산 및 복구 비용을 감소시키는 쓰기 연산의 정책을 제안 함으로서 에러 핸들링 시에 입출력 비용을 줄인다. RAID5에서는 추가적인 여분 디스크를 사용한 Sparing Disk 기법을 제안함으로서 디스크 에러 모드에서도 정상 모드에서의 성능에 근접하는 입출력 성능을 보장하며 빠른 디스크 복구를 지원한다.

2. SANtopiaVM

SANtopiaVM은 볼륨관리자로서 크게 4부분인 구성 관리자(*Configuration manager*), 입출력 관리자(*I/O manager*), 주소연결관리자(Mapping Manager), 자유공간관리자(Free Space Manager)로 구성된다.구성 관리자는 연결된 독립적인 디스크 장치들을 하나의 볼륨 디바이스로 가상화하며 볼륨의 생성, 추가, 삭제, 변경에 대한 기능등을 지원한다. 입출력 관리자는 볼륨에 대한 입출력 요구를 처리하고 다양한 소프트웨어 RAID 레벨을 지원함으로서 데이터의 가용성 및 고성능 입출력 기능을 제공한다.[10] 현재 Linear 모드, RAID0, 1, 5가 지원가능하며 향후 RAID4와 Hybrid RAID를 지원할 예정이다. SANtopiaVM의 주소연결 관리자는 물리적인 디바이스와 논리적 디바이스 사

이에 주소 공간을 관리하며 SANtopiaVM에서는 주소관리 테이블(mapping table)을 통해서 주소 연결을 관리하기 때문에 동적으로 볼륨 구성을 변경할 수 있는 유연성을 제공한다. 또한 자유공간 관리자는 저장 공간의 사용 유무에 관한 정보를 관리하여 저장 공간의 요구, 철회시에 빠른 디바이스 공간의 할당 및 해지가 가능하다. 이러한 SANtopiaVM은 고성능 입출력 및 데이터의 고가용성이라는 특징을 지원하기 위해서 RAID의 에러 핸들링 정책을 제공하고 있으며 해당 정책은 3절에서 기술하도록 한다.

3. RAID 에러 핸들링 정책

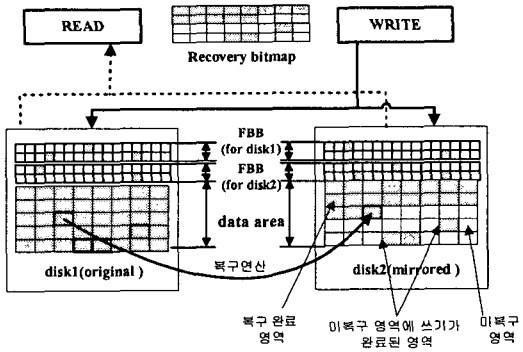
본 장에서는 SANtopiaVM의 RAID 에러 핸들링 정책에 대하여 기술한다. Linear 모드와 RAID0(stripping)의 경우는 중복 데이타나 패리티 데이터와 같은 체크 데이터(check data)를 수용하지 않는 RAID 레벨이므로 RAID 데이터의 에러 발생시에 데이터의 가용성을 지원하지 않는다. 따라서 본 논문에서는 RAID1과 RAID5의 에러 핸들링 정책에 대해서만 다루도록 한다.

3.1 RAID1(mirroring)의 에러 핸들링 정책

본 장에서는 RAID1에서의 에러 발생시, 데이터의 일관성을 유지하기 위한 정책과 영구적 디스크 오류의 복구를 위한 방법을 기술한다.

3.1.1 중복 데이터의 일관성 유지

RAID1의 중복 데이터 사이의 데이터 일관성의 유지를 위한 방법에 대하여 살펴보자 RAID1에서의 중복된 데이터들 사이에 데이터 일관성이 부적절한 상태가 되는 경우는 미러링된 데이터들 중에서 일부의 데이터만이 갱신되는 경우이다. 시스템의 오류나 네트웍 오류로 인한 전체 시스템 접근 불가능 등의 원인으로 원본과 복사본의 디스크에 모두 쓰기를 실패하는 경우는 볼륨 전체의 오류로서 디스크 쓰기가 발생하지 않으므로 데이터의 일관성에 문제가 발생하지 않는다. 상기에서와 같이 데이터 일관성이 부적절한 상태가 되는 것은 일반적인 입출력 연산에서 일부의 디스크에만 쓰기가 진행되는 경우이다. 본 논문에서는 RAID1의 데이터의 일관성 유지 및 일관성이 부적절한 데이터들에 대한 복구를 위해서 [그림 1]와 같이 RAID1을 구성하는 각각의 디스크에 데이터의 일관성을 관리하기 위한 비트맵 FBB(Failed Block Bitmap)을 유지하고 임의의 디스크에서 오류가 발생할 경우에도 비트맵의 정보를 통한 복구 및 데이타 일관성의 유지를 위해서 원본과 복사본들의 디스크에 중복하여 저장하며 초기에 0으로 설정한다. 연산 수행 중에 I/O 장애가 발생하는 경우에 원본 디스크와 복사본의 디스크에 저장된 FBB에서 해당 블록의 비트에 일관성 정보가 표시된다. 또한, FBB는 원본과 복사본의 디스크에 모두 존재하므로 해당 디스크 전체가 오류인 경우에도 RAID1을 구성하는 나머지 정상 디스크에 존재하는 해당 디스크의 FBB에 접근이 가능하다 일관성의 정보를 유지하기 위한 중복된 개개의 FBB들이 가지는 일관성 정보는 서로간에 불일치할 수 있으나, 장애의 발생시, 해당 데이터에 대한 일관성 정보는 중복된 FBB 중에서 적어도 하나는 올바른 정보를 유지하게 된다. 그러므로 이러한 데이터의 불일치를 복구하기 위해서는 중복된FBB의 일관성 정보 사이에 비트OR 연산을 수행하면 FBB들 사이에 일관성이 유지되지 않는 경우에도 불일치가 발생한 데이터를 모두 찾을 수 있다. 따라서 임의의 디스크에 대한 장애의 복구 시에는 각 디스크에 존재하는 해당 FBB의 정보를 비트OR 연산을 수행함으로서 일시적인 접근 불가능 상태의 모든 경우에 대한 올바른 결과를 획득할 수 있다. 데이터 불일치의 복구는 비트OR 연산의 결과인 FBB의 값에서 일관성의 부적절함이 표시된 블록들을 검출하고 RAID1을 구성하는 임의의 정상 디스크 블록으로부터 해당 데이터 블록을 읽은 후 갱신함으로서 수행된다. 본 발명에서 제안하는 중복 데이터 사이의 일관성을 유지하기 위한 방법은 발생 가능한 여러 가지 오류의 경우에 대하여 정확한 데이터 일관성 정보를 유지함으로서 RAID1 의 데이터 신뢰성을 보장한다.



[그림 1] FBB 및 디스크의 복구

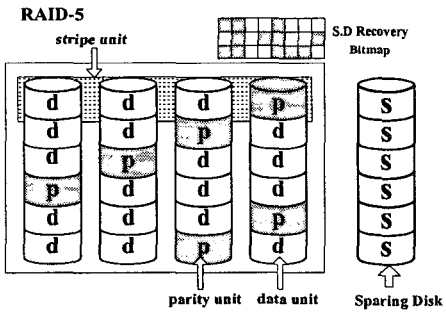
3.1.2 오류 디스크의 복구

영구적인 디스크 오류의 복구는 동일 데이터를 가지고 있는 정상 디스크로부터 디스크 복사를 수행함으로서 진행되며 디스크 에러 모드에서 성능의 급격한 저하를 피할 수 있는 효율적인 에러 핸들링 정책이 요구된다. [그림1]에서 Recovery Bitmap은 디스크의 복구가 진행되는 동안 새로운 디스크의 블록들에 대해서 복구 여부의 정보를 관리하기 위한 것이다. 새로운 디스크의 복구 완료 영역은 정상 디스크로부터 데이터의 복사가 이미 이루어진 영역을 의미하며 미복구 영역은 아직 데이터의 복구가 이루어지지 않은 영역을 의미한다. 제안하는 에러의 복구 방식은 복구가 진행되는 동안의 입출력 요구 수행의 부하를 감소시키기 위한 것이다. 디스크의 복구가 진행되는 동안, 읽기 연산의 처리는 대상 블록의 Recovery Bitmap 정보를 참조하여 복구 완료 영역의 블록인 경우 정상 모드에서와 같은 방식으로 부하 분산(load balancing)을 위한 라운드 로빈(Round Robin) 방식이나 임의 디스크 선택 방식에 의해서 대상 디스크를 선정한 후에 연산을 수행하고 대상 블록이 미복구 영역인 경우, 정상 디스크로부터 데이터를 읽는다. 쓰기 연산의 경우는 새로운 디스크의 복구 여부에 상관없이 정상 디스크와 새로운 디스크에 모두 쓰

기 연산을 수행한다. 복구 영역인 경우는 정상 모드에서의 연산과 동일하고 미복구 영역인 경우, 정상 디스크와 새로운 디스크에 모두 쓰기 연산을 수행한 후에 Recovery Bitmap에 복구 여부를 체크 함으로서 연산을 마친다. 앞에서 기술한 쓰기 연산의 방식은 읽기 연산에서의 부하분산이 가능한 디스크의 영역을 확장해 줄 뿐만 아니라 새로운 디스크에서 복구 해야 할 영역을 감소시켜서 디스크 복구 비용을 줄이는 장점을 제공한다.

3.2 RAID-5의 에러 핸들링 정책

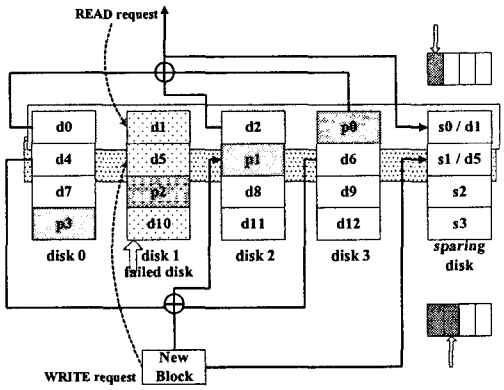
본 장에서는 RAID5의 디스크 오류(degraded mode)시의 에러 핸들링을 위하여 여분 디스크를 사용하는 방식인 Sparing disk 기법을 제안한다. Sparing Disk 기법의 구조는 [그림 2]와 같이 기존의 dedicated sparing기법과 유사한 구조를 가지며 Sparing Disk에는 Sparing 디스크 블록들의 회복 여부를 체크하는 S.D(Sparing Disk) Recovery Bitmap이 존재한다. 디스크 에러가 발생하지 않는 정상 모드에서는 제안하는 Sparing Disk 기법의 읽기 / 쓰기 동작은 일반적인 RAID-5에서의 동작과 동일하므로 본 논문에서는 RAID-5의 디스크 에러 모드에서의 동작에 대해서만 기술하도록 한다.



[그림 2]Sparing Disk 기법을 위한 구조

3.2.1 디스크 오류시(degraded mode)의 정책

RAID5에서의 디스크 오류(degraded mode) 경우, 데이터의 가용성을 지원하기 위하여 대상 블록에 대한 재생성(regeneration) 연산이나 패리티 데이터만의 갱신을 통해서 입출력 요청을 처리해야 한다. 따라서 에러 모드(degraded mode)에서의 읽기/쓰기 연산은 시스템에 과중한 처리 부하를 주게 되어 성능을 저하시키며 요청에 대한 응답 속도(response time)를 지연시키는 문제점을 갖는다. 상기 문제점을 완화하기 위하여 제안하는 Sparing Disk 기법에 대하여 살펴보자. [그림 3]는 제안하는 Sparing Disk 기법이 적용된 시스템에서 에러 디스크에 대한 읽기 연산이 처리되는 과정이다. [그림 3]에서 디스크 1 이 에러(failure)인 경우에 d1 블록에 대한 읽기 연산을 요청한 경우를 가정하자. 제안하는 기법에서는 읽기 연산을 수행하기 전에 해당 블록에 대한 S.D Recovery Bitmap의 회복 정보를 먼저 검사한다. Bitmap에 회복 여부가 표시되지 않은 경우, 기존의 RAID-5에서와 동일하게 에러 디스크의 블록인 d1 데이터의 재생성을 위하여 d1블록이 소속되어 있는 stripe unit의 각 디스크 데이터(d0, d2, p0)를 읽고 패리티 연산을 수행하여 손상된 d1 블록을 재생성함으로서 읽기 연산을 수행한다. 또한 재생성된 d1의 데이터를 Sparing disk의 동일 블록(s0)에 쓰는 추가적인 단계를 포함하며 Sparing Disk에 임의의 데이터가 기록한 후, S.D Recovery Bitmap에 갱신된 블록에 해당하는 Bitmap의 위치에 복구되었음을 표시한다.



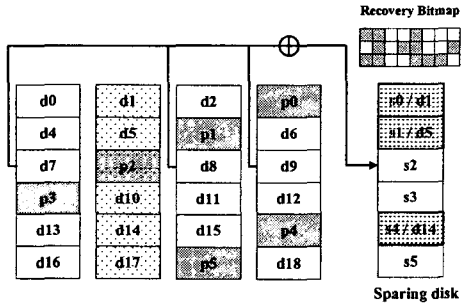
[그림 3] 읽기/쓰기 연산의 에러 핸들링

Bitmap에 회복 여부가 표시된 경우, 해당 블록에 대한 재접근(re-access)임을 의미하며 Sparing Disk의 동일 블록으로부터 읽기 연산을 수행한다. 따라서 오류 디스크의 임의 블록에 대한 두 번째부터의 읽기 요청은 Sparing Disk에서 처리되며 재생성 되어진 데이터 블록이 Sparing Disk의 동일 블록에 존재하므로 정상적인 모드에서의 읽기 연산과 동일한 과정을 거쳐서 처리된다. 다음으로, [그림 3]에서 에러 디스크의 d5 블록에 대한 쓰기 요구가 발생하는 경우를 가정하자. 쓰기의 연산도 읽기 연산과 동일하게 연산의 수행전에 S.D recovery bitmap의 데이터 회복 여부를 검사한다. 데이터가 회복되지 않은 블록의 경우, 기존 RAID-5에서와 동일하게 패리티 블록에 대한 데이터를 계산하기 위해서 동일한 stripe unit에 소속된 디스크들의 블록(d4, d6)을 읽는다. 다음으로 읽혀진 데이터들과 갱신할 데이터(d5) 사이에 패리티 계산을 수행하고 이를 해당 패리티 블록에 기록한다. 이것은 기존의 RAID-5의 WRITE 수행과 동일하지만 본 논문의 방법에서는 갱신할 데이터를 Sparing Disk의 s1 블록에 기록하고 해당 갱신 블록에 해당하는 Bitmap에 복구되었음을 표시한다. 반대로, S.D recovery bitmap의 회복 여부가 표시된 경우는 이전에 손상된 데이터가 회복되었음을 의미하며 Sparing disk의 데이터를 갱신함으로서 쓰기 연산을 수행한다. 따라서 에러 디스크의 임의의 블록에 대한 두 번째 이후의 I/O 요구는 Sparing Disk 에 기록된 데이터에 대해서 I/O 를 수행하므로 쓰기 연산의 수행 시에 d4, d6 블록에 대한 입출력은 수행할 필요가 없으며 Sparing Disk의 s2/d5 블록과 parity 블록에 대한 연산만으로 에러 디스크의 블록에 대한 쓰기 연산의 처리가 가능하다. 이것은 임의의 블록에 대한 첫 번째 접근은 추가적인 연산이

요구되지만 해당 블록의 두 번째 접근부터는 RAID-5의 정상 모드에서의 I/O와 동일하게 처리할 수 있음을 의미한다.

3.2.2 오류 디스크의 복구

[그림 4]는 RAID-5 디스크 에러 모드에서 제안하는 Sparing disk 기법의 적용한 후, 에러 디스크의 데이터 복구를 위한 연산 과정을 도시한 그림이다. [그림 4]에서 Sparing Disk의 s0, s1 및 s4는 에러 디스크에 대해서 이전에 읽기 / 쓰기 입출력 요청이 발생하였던 블록들이며 해당 블록들에는 재생성 되어진 정상 데이터들이 복구(rebuilding)되어 있다.



[그림 4] RAID5의 디스크 복구

디스크의 복구 시에 기존의 RAID-5에서는 RAID-5를 구성하는 모든 디스크에서 동일한 stripe unit에 소속된 각각의 데이터에 대하여 읽기 연산과 stripe unit 블록들의 패리티 연산을 수행함으로서 에러 디스크 블록의 데이터를 순차적으로 복구하며 이러한 기존의 방식은 복구 시간이 길어져 시스템의 가용성을 저하시키는 단점을 가지며 복구를 위한 부하가 커서 시스템의 성능을 저하시킨다. 제안하는 Sparing Disk 기법에서는 에러 디스크의 복구를 시작하는 시점까지 진행되어진 입출력 요청에 의해서 회복되어진 데이터들이 이미 존재하므로 복구 연산에서는 S.D Recovery Bitmap이 체크되지 않은, 입출력 요청이 한번도 발생하지 않은, 블록들에 대해서만 복구를 하게 된다. [그림 4]에서 디스크의 복구를 위해서 S.D Recovery Bitmap을 검색하여 복구되지 않은 블록을 검색한다. 복구되지 않은 블록(s2)이 존재하면 해당 블록이 소속된 stripe unit의 데이터들(d7, d8, d9)과의 연산을 통하여 해당 블록의 데이터를 복구하고 재 생성된 데이터(s2/p2)는 교체되어질 Sparing Disk에 기록한다. 이와 동일한 방식으로 모든 Bitmap을 검색하게 되며 복구되지 않은 데이터들(s3, s5)에 대해서 선별적으로 복구를 한다. 따라서 제안하는 Sparing Disk 기법의 디스크 복구 방식은 기존의 RAID-5 시스템의 디스크 복구 방식에 비해서 복구 시간이 짧아지며 시스템 성능의 저하를 완화시킬 수 있다는 장점을 갖는다.

4. 결론

본 논문에서는 RAID에서의 데이터 가용성 및 성능을 고려한 에러 핸들링 정책을 제안하였다. 본 논문의 정책은 SANtopiaVM RAID의 오류 발생시, 시스템 성능의 급격한 저하를 완화하며 에러로부터 빠른 복구를 지원하고 데이터들 사이에 일관성을 효과적으로 유지함으로서 RAID 데이터에 대한 신뢰성 및 가용성을 제공한다.

참고문헌

[1] Paul Massiglia, "The RAID book" 6th edition, RAID Advisory Board, 1997

[2] Sanghoon Jeon, Byoungchul Ahn "A Cost - Effective Solution to the Single disk Failure in RAID Architecture" , Communications, Computers and Signal Processing, 1997, pp. 285 -288 vol.1

[3] Hai Jin, Kai hwang, Jiangling Zhang, "A RAID Reconfiguration Scheme for Gracefully Degraded Operations" , Parallel and Distributed Processing, 1999. PDP '99. Proceedings of the Seventh Euromicro Workshop, 1999, pp.66-73

[4] M.Holland, G,Gibson, and D.Siewiorek, "Fast, On-line Failure Recovery in Redundant Disk Array" , Proceedings of the Fault Tolerant Computing, 1993, pp.

[5] Muppalaneni, N, Gopinath, K, "A multi-tier I/O RAID storage system with RAID1 and RAID5" , Parallel and Distributed Processing Symposium, 2000. IPDPS 2000. Proceedings. 14th International, 2000, pp. 663-671

[6] David A. Patterson, Garth A. Gibson, Randy H. Katz, "A Case for Redundant Arrays of Inexpensive Disks(RAID)" , ACM SIGMOD Conference Proceedings, 1988, pp. 109-116

[7] Jai Menon, Dick Mattson, "Comparison of Sparing Alternatives for Disk Arrays" , Proceeding of International Symposium on Computer Architecture, 1992, pp. 318-329

[8] Peter M. Chen and Edward K. Lee, "Stripping in a RAID Level 5 Disk Array", In Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems, 1995, pp 136-145

[9] Kai Hwang, Hai Jin, Roy Ho, "RAID-x: A New Distributed Disk Array for I/O-Centric Cluster Computing", High-Performance Distributed Comput ing, 2000. Proceedings, 2000, pp 279-286

[10] 김경배, 김영호, 김창수, 신범주, "SAN을 위한 전역 파일 공유 시스템의 개발 ", 한국 정보 과학회지, VOL.19 NO.3 pp. 24-32, 2001. 03.