RAID 시스템에서 자율적 네트웍 조합에 의한 읽기/쓰기 성능 개선

최귀열*

Autonomous Network Combination of RAID System to read/write Performance Improvement

Gwi-yoel Choi*

요 약

다중 디스크 드라이브가 포함된 디스크 배열 시스템에서 디스크의 수가 증가 될 때 시스템 성능은 컨트롤러의 집중화 또는 버스로 사용되는 전송 경로의 병목현상에 의해 제한되어진다. 이러한 단점을 보완하기 위해 고성능 대용량의 RAID가 등장하였으며 RAID 시스템에서 컨트롤러 기능은 모든 디스크 드라이브에 분산되고 각 디스크는 그들의 임무를 수행하는 자율성을 가진 자율적 네트웍이 일반적 계층 시스템 보다 확장성이 좋고 시스템 자원을 보다 효율적으로 이용할 수 있어 디스크 수의 증가율에 따라 높은 읽기/쓰기 처리율의 성능을 제공한다.

ABSTRACT

When the number of disks array systems that contain multiple disk drives, system performance is limited by a bottleneck at a centralized controller at a communication path than uses a bus. A redundant array of inexpensive disks(RAID) consists of many disks to enable high performance and large capacity. We evaluate a scalable architecture called Autonomous network, in which the controller functions are distributed to all disk drives and each disk has autonomy in processing its tasks. Disks drives enable better scalability and more effective utilization of system resources than with a hierarchical system. Autonomous network provided high read/write performance throughput in proportion to the number of disks.

키워드
RAID, 자율적 네트웍, 성능, 처리율

Ⅰ. 서 론

최근의 고성능 디스크 시스템은 프로세서와 2차 저장장치 사이의 성능 갭이 좁은 것을 요구한다. RAID[1]는 다수의 디스크 배열로 구성되어 고성능과 대용량을 갖고 또한 중복 정보로 높은 신뢰성을 유지할 수 있어 많은 상업용이 생산되어 활용되고 있다. 일반적 RAID의 특징은 첫째 디스크 접근은 분산된 많은 디스크 드라이브들에 의하여 이루어지고 그러나 배열 컨트롤러와 디스크 드라이브들 사이의 전송은 버스에 집중되어진다. 만약 버스 폭이 충분히 넓지 않으면 전체 시스템의 성능은 제한되어질 것이다. 둘째 RAID-5에서 다중 디스크 손실의 가능성은 디스크 증가의 수 처럼 증가된다. 만약 그러면 시스템에서 많은 디스크 또는 개개 디스크의 신뢰성은 매우 높지 않고 다중 디스크가 손상되고 데이터를 잃을 것이다. 셋째 배열 컨트롤러에서 많은 접근 요구가 동시 수행되어 질 때 그 시스템 성능은 버스

*재능대학 정보통신계열

접수일자 : 2002. 11. 1

가 충분히 빠르더라도 CPU 성능에 의해 제한될 것이다.

자율적 네트웍은 모든 드라이브와 인터페이스가 외부 환경에 상호 연결망에 의해 결합되어지며 이 구조는 RAID-5내의 버스 또는 배열 컨트롤러에 의한 중앙 전송 경로 형태를 제거하여 시스템 확장성이 제공되어진다. 자율적 네트웍 배열 컨트롤러의 기능은 중복 데이터의 유지, 데이터 캐싱 그리고 디스크 손상으로부터 데이터 회복하기 위한 재구성은 분산된 모든 드라이브에서 한다. [2]에서 동작 디스크는 NADS(network-attached secure disks) [3]의 성공을 결정한다. 이 목적을 달성하기 위해 디스크 드라이브와 응용 서버 사이 데이터를 전송율을 줄이고 디스크 드라이브 응용 프로그램 부분을 이동시켜야 한다. [4]에서 시스템 성능은 패리티 계산의 분산에 의해 향상된 것을 나타낸다.

이 논문은 II장에선 자율적 네트웍의 구조와 작용에 관하여 기술하고 III장에선 새로운 I/O 버스의 종류에 대해서 설명하고 IV장에는 모의실험에 사용된 디스크 시스템의 모델 및 결과에 대해 기술하고 V장에 결론을 나타낸다.

II. 자율적 네트웍 개요

자율적 네트웍은 디스크 노드, 인터페이스 노드 그리고 상호 연결 네트웍으로 구성되어 있다. 이것은 버스 대신 상호 연결 네트웍을 갖는 노드의 접속과 분산 배열 컨트롤러 기능의 디스크 드라이브에 의해 RAID-5에 존재하는 병목현상의 성능을 제거한다.

2.1 디스크 노드

디스크 노드는 디스크 드라이브와 디스크 컨트롤러로 구성되어 있다. 컨트롤러는 자신의 CPU와 메모리를 가지고 있고 그것은 디스크 접근 실행뿐만 아니라 패리티를 계산 중복 데이터를 유지하며 그리고 전송 링크 통해 다른 노드들로 전송도 한다. 디스크 노드가 인터페이스 노드 또는 다른 디스크 노드로부터 요청을 받을 때 그 디스크 드라이브는 읽기/쓰기 데이터에 접근한다. 만약 드라이브가 손상되었다면 그 데이터를 재구성하거나 또는 내부 모드 전송을 통해 인접 주위 노드와 협력하여 패리티를 업데이트 시켜야 한다.

2.2 인터페이스 노드

인터페이스 노드는 또한 디스크 노드와 그리고 노드로부터 결과를 되돌려 보내려는 외부 환경으로부터 접근 요구를 네트웍으로 보내는 것으로 구성되어 있다. 자율적 네트웍의 인터페이스 노드의 수는 무한하다. 왜냐하면 네트웍의 어느 위치에도 있을 수 있기 때문이다. 예컨대 망로 또는 원형망을 사용하다면 디스크 노드에 접속할 수 있고 그 인터페이스 노드는 디스크 노드와 어떤 링크 사이에도 위치(그림 1)할 수 있다. 이와같이 자율적 네트웍은 네트웍을 통한 전송 처리를 할 수 있고 그리고 많은 인터페이스 노드를 통한 외부 환경의 넓은 대역폭을 제공한다.

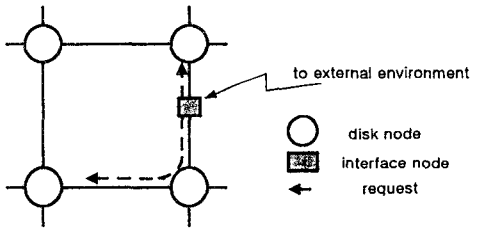

Figure 1. Inteface node in Auto-net

그림 1. 자율적 네트웍 인터페이스 노드

2.3 패리티 그룹

패리티 그룹은 한 단위이고 디스크 노드의 수와 중복 패리티 데이터 유지로 구성되어 있다. 패리티 그룹에 있는 패리티 노드의 위치는 RAID-5의 패리티 블럭의 위치처럼 즉시 변환 할 수 있고 또 각 패리티 스트라이프도 변환시킬 수 있다. 즉 패리티 블럭들은 패리티를 업데이트 시키는 동안 부하 집중을 피하기 위하여 분산되어 진다. [5].

III. 새로운 I/O 기술

스토리지 시스템의 성능을 개선하기 위해서는 저장 매체 자체의 성능 개선과 I/O 경로에 있는 모든 요소들의 하드웨어와 소프트웨어의 성능을 개선하여

야 한다. 지금까지 호스트 I/O 버스로는 S-bus, EISA, VME, PCI 버스 등이 사용되었다.

3.1 PCI-X

PCI-X는 Compaq. HP, IBM, Intel 등의 회사가 주축으로 사양을 PCI의 가격대비 성능을 높이기 위해 개발한 것으로RRS(register-to-register)를 구현한 것이 PCI와 차이점이다.

RRS는 신호를 보다 정확하게 처리 할 수 있고 RRS는 PCI와 비교할 때 쓰기 I/O 성능은 동일하나 읽기 I/O는 속도를 두 배로 높일 수 있다. 또한 지연시간을 감소시킨다. PCI-X는 현재 초당 1GB 전송률을 가지나 미래는 2GB또는 4GB의 전송률을 가질 것으로 예상되며 고속의 네트워크 기술과의 보다 연동이 잘 될 것으로 예상된다.

3.2 Infini Band

Infini Band는 Cisco, Compaq. HP, IBM 등의 회사가 주축으로 진행한 Future I/O와 Intel에 의해 시작되고 Dell, Sun 등이 지원하는 차세대 I/O 프로젝트를 통합하여 나온 기술이다. Infini Band는 통신장치들을 연결하는 케이블에 따라 초당 2.5Gbits, 10Gbits, 30Gbits의 데이터 전송을 제공하며 최대 64,000개의 CPU, 저장시스템과 다른 구성요소를 연결할 수 있는 스위치 기반 구조를 가지고 있다. 이 기술은 구리선을 사용할 경우에 17미터 단일 모드 광케이블을 사용할 경우에 10Km까지 작동범위가 다양하다. Infini Band는 IPv6 주소 방식을 기반으로 하여 각 노드는 128비트의 IP 주소를 사용하여 고속의 인터넷 연결이 가능한 구조를 개발할 수 있다.

3.3 Rapid IO

Rapid IO는 Motorola, IBM, Cisco, Nortel Networks을 포함한 약 40여개의 회사들이 개발하는 기술로 각 구성요소의 데이터 전송률을 높이고 지연을 줄여서 스위치 구조를 가지는 내장형 시스템의 성능을 개선하는데 사용된다.

Rapid IO는 LVDS(low voltage differential signal)을 사용하며 모든 처리가 하드웨어로 이루어짐으로써 I/O소프트웨어를 작성, CPU 또는 전용 프로세서에서 그 소프트웨어를 수행할 필요성을 제거함으로써 CPU의 부담을 줄이고 시스템의 가격을 줄일 수 있다. RapidIO는 높은 대역폭과 신뢰성이 필요한 전화 교환기와 라우터와 같은 네트워크 장치 응용의 요구에 가장 적합하다.

IV. 디스크 시스템 모델링 및 결과

분산된 기능과 디스크 노드의 자율성 효과와 자율적 네트웍을 실험 평가하고 일반적 디스크 시스템인 즉 다중 버스를 갖는 계층 구조와 비교한다.

4.1 계층적 디스크 시스템

많은 디스크 배열 시스템 즉 RAID는 시스템 확장성을 고려해서 계층적 구조를 사용한다. (그림2)에 시스템의 모델링 구조가 묘사되었다. 그것은 다중 RAID 서브 시스템으로 형성되었고 서브 시스템의 수 증가에 의해 규격화된다.

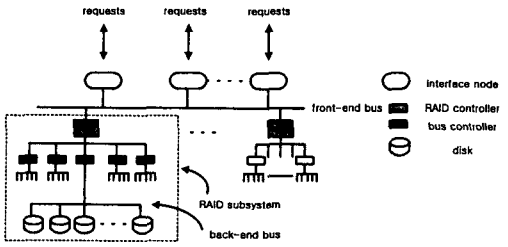

Figure 2. Hierarchical RAID system

그림 2. 계층적 RAID 시스템

외부 환경 인터페이스는 외부 환경으로 부터 읽기/쓰기 요청을 수신하여 전단(front-end) 버스를 통해 배열 컨트롤러로 그들을 송신하고 수신 결과와 송신 결과를 다시 외부 환경에 보낸다. 인터페이스에 의해 생성된 각 디스크 접근 요청의 목적지는 불규칙적으로 선택되어지고 모든 접근은 64-KB 블럭의 읽기/쓰기이다.

시스템은 3종류의 버스를 가지고 있다. 전단 버스는 단지 버스일 뿐이고 모든 인터페이스와 모든 배열 컨트롤러를 연결한다. 중간 버스는 종단(back-end) 버스들의 컨트롤러와 배열 컨트롤러를 연결한다. 종단 버스는 디스크 수들과 연결한다. 인

터페이스에 의해 접근 요청이 발생될 때 전단 버스를 통해 배열 컨트롤러로 송신되고 배열 컨트롤러는 중간 버스를 통해 종단의 버스 컨트롤러에 요청을 송신하고 그리고 버스 컨트롤러는 종단 버스를 통해 목적 디스크에 그것을 송신한다. 결과 데이터 혹은 상태 흐름은 같은 통로의 역순이다.

배열 컨트롤러는 인터페이스로부터 요청을 수신하고 그리고 디스크 접근 요청을 디스크로 송신한다. 만약 그것이 다중 요청 수신일 때 그것은 동시에 그들을 수행하고 만약 한 요구의 수행이 어떤한 이유로 중지된다면 그 컨트롤러는 다른 요청을 수행하기 시작할 것이다. 컨트롤러는 또한 디스크 손상에 대한 예방으로 중복 데이터 유지를 위해 패리티 데이터를 계산한다.

4.2 자율적 네트웍

자율적 네트웍의 각 인터페이스 노드는 계층적 디스크 시스템 예측의 인터페이스와 비슷하고 그것은 두 링크에 의한 2차원적 원형 네트웍으로 연결되어 졌다. 모든 인터페이스 노드는 전체 네트웍을 통한 분산 전송일지라도 그들의 주위로부터 같은 거리에 위치한다. 각 디스크 노드는 4개 링크로 연결되어지고 그리고 전송 방향과 일치한다. 디스크 노드가 다른 노드와 링크로 메시지를 송신 할 때 전송 경로의 부분이고 다른 전송을 위해 사용된다. 그 메시지는 다른 전송이 끝날 때까지 통과 될 수 없다. 디스크 노드는 계층적 디스크 시스템의 배열 컨트롤러 처럼 다중 요구를 조정할 수 있다. 그것은 메모리가 거의 없다든지 요구의 수가 매우 작기 때문이다.

4.3 변수

(표1)에 모의실험에 사용한 변수들을 정리하였다. 이 변수들은 high-end 디스크 시스템 매치를 위해 조정되었다.

전단 버스를 위해 1GB/s 대역폭을 사용했다. 이 같은 대역폭은 PCI-X때 보통이고 10Gbps 광 채널은 가까운 장래에 사용될 것이다.

동시 버퍼 된 요구의 수 그리고 디스크 노드내의 배열 컨트롤러에서 수행된 회수는 50번이다.

표-1. 모의실험에 사용된 변수

대역폭	자율적 네트웍 전단 버스 중간 버스 종단 버스	100 MB/s 1 GB/s 400 MB/s 100 MB/s
셀업 시간	자율적 네트웍 모든 버스	50 μs 1 μs
XOR 성능	디스크 노드 배열 컨트롤러	10 MB/s 200 MB/s
요구 버퍼의 수	디스크 노드 배열 컨트롤러	5 250
비율	인터페이스/디스크 배열 컨트롤러/디스크 패리티 그룹/디스크	1/5 1/50 1/5
디스크	디스크 억세스 단위 평균 탐색 시간 전송율	64 KB 8.2 ms 50 MB/s

4.4 모의실험 결과

(그림3)과 (그림4)는 디스크 수가50에서 500으로 상승할 대 읽기/쓰기 처리율 성능을 나타내었다. 자율적 네트웍은 디스크 수 증가에서 접근 처리율의 비가 읽기에서 높다는 것을 제공한다. 한 편 디스크 수가 200보다 클 때 약 초당 14,000 I/Os(900MB/S)의 처리율로 계층적 시스템 구조에서는 포화됨을 제공한다. 전단 버스는 병목 현상이 나타내는데 그것은 대역폭이 단지 1GB/s이기 때문이다. 쓰기 접근 처리율은 읽기 접근 처리율보다 낮다.

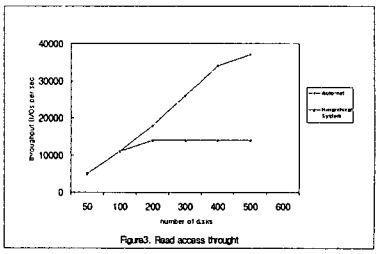

그림 3. 읽기 처리율

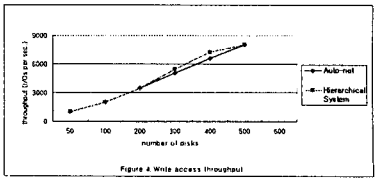

그림 4. 쓰기 처리율

(그림5)는 400장의 디스크를 갖는 계층적 시스템에서 셋업 시간의 영향을 나타낸 것이고 전단 버스는 고 대역폭(3GB/s)를 사용하였다. 전단 버스에서 셋업 시간이 1[μs]보다 짧을 때 시스템 처리율은 자율적 네트워보다 높다. 셋업 시간이 증가하면 그러나 처리율은 감소한다. 셋업 시간이 1에서 5[μs]로 증가할 때 처리율은 30[%] 감소한다.

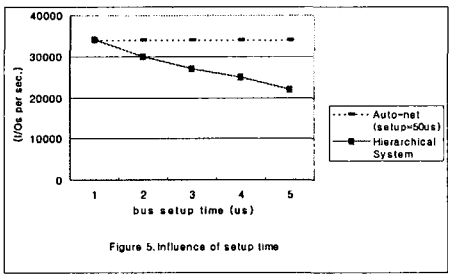

그림 5. 셉업시간의 영향

(그림6)는 각 배열 컨트롤러는 자율적 네트워처럼 쓰기 처리율이 높은 것을 제공하고 높은 패러디 처리율을 필요로 한다.

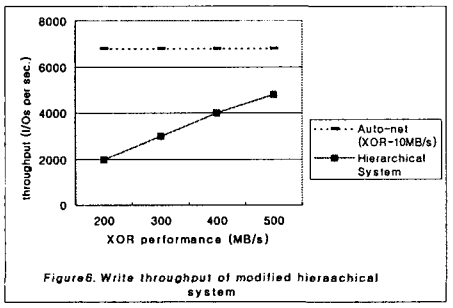

그림 . 계층적 구조 시스템의 쓰기 처리율

패리티 계산 처리율은 500MB/s 일 때 계층적 시스템의 처리율은 70[%]이다. CPU 파워가 지속적으로 급격히 증가해도 패리티 계산 처리율은 메모리 접근 속도에 강하게 의존한다. 이유는 큰 데이터 양이 계산되어야 하고 또 CPU 캐쉬에 전체 데이터를 저장할 수 없기 때문이다.

V. 결 론

이 논문은 자율적 네트웍의 디스크 시스템에서 분산된 기능의 효과와 디스크 드라이브의 자율성과 계층적 구조를 갖는 일반적인 디스크 시스템과 모의실험과 비교를 통해 평가하였다. 자율적 네트웍은 동일 구조와 디스크 수의 비율에서 높은 읽기/쓰기 처리율을 제공한다. 반대로 말하면 계층적 시스템으로 구성된 것은 전단 버스의 병목현상으로 인한 낮은 처리율은 제공한다. 만약 전단 버스가 높은 대역폭을 갖고 포트의 수가 많은 것을 사용한다면 전체 시스템의 성능은 감소하고 셋업 시간은 증가될 것이다. 셋업 시간이 1[μs]에서 5[μs]로 증가되면 처리율은 30[%] 감소된다. 계층적 시스템은 자율적 네트웍 처리율의 70[%]를 제공하고 패리티 계산율이 500MB/s 일 때이다.

결국 자율적 디스크 [6],[7]를 연구하고 자율적 네트웍을 완성하여 LAN 혹은 SAN 환경에서 중복 데이터의 유지 읽기/쓰기 성능을 향상하기 위한 것이다.

참고문헌

[1] D. A patterson, G. Gibson and R. H Katz. A Case for Redundant Arrays of Inexpensive Disks(RAID). In Pro. of ACM SIGMOD Conference. pages 109-116, Jun 1988.

[2] A. Acharya, M. Uysal, and J, Saltz. Active disks: Programming model, algorithms and evalution. In Pro. of Architectural Support for Programming Languages and Operating System, October 4-7, 1998, San Jose, CA USA, pages 81-91, 1998.

[3] G. A. Gibson, D. F. Nagle, K. Amiri, J. Butler, F. W. Chang, H. Gobioff, C. Hardin, E. Riedel, D. Rochberg, and J. Zelenka. File server scaling with network-attached secure disks. In Pro. of the ACM Int'l Conf. on management and Modeling of Computer System. 1997.

[4] P. Cao, S. B. Lim, S. Venkataraman, and J. Wilkes. The Ticker TAIP parallel RAID architecture. In Pro. of the 20th ISCA. pages 52-63, 1993.

[5] S. Tomonaga and H. Yokota. An Implementation of a Highly Reliable Parallel-Disk System using Trandputers. In Pro. of the 6th Transputer/Occam inrn'l Conf, pages 241-254. IOS Press, Jun 1994.
[6] H. Yokota. Autonomous disks for advanced database applications. In Pro. of International Symposium on Database Applications in Non-Traditional Environments, pages 441-448, Nov 1999.
[7] H. Yokota. Performance and reliability of secindary storage systems. In Pro. of World multiconference on Systemics, cybernetics and Informatics, invited paper, pages 668-673, Jul 2000.

저자 소개

최귀열(Gwi-Yoei Choi)
1980년 숭실대학교 전자공학과 졸업
1987년 숭실대학교 대학원 졸업 공학석사
1999년 동대학원 박사 수료
1991년-현재 재능대학 정보통신계열 부교수
※관심분야: 디지털 신호처리, VOD, 컴퓨터 구조