클러스터 파일 시스템 SANique™ 기반의 DBMS 표준 성능 평가

황진호[0], 백은주, 이규웅
상지대학교 컴퓨터정보공학부
{jhwang[0], ejbaek, leekw}@sangji.ac.kr

이장선
매크로임팩트㈜
sunny@macroimpact.com

TPC Benchmark Test for DBMS based on Cluster File System SANique™

Jin-Ho Hwang[0], En-Ju Baek , Kyu-Woong Lee
School of Computer Information, Sangji Univ.

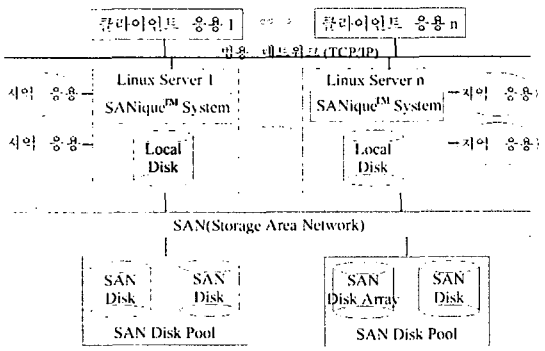
Jang-Sun Lee
Macroimpact co., Ltd.

요 약

본 논문은 SAN(storage area network)상에 네트워크-부착형(network-attached) 저장 장치들을 직접 연결하여 파일 서버 없이 직접 데이터 전송이 가능한 SAN 기반의 클러스터 공유 파일 시스템인 SANique™의 설계 기법을 설명하고 SANique™ 시스템의 기본 파일 입·출력 연산 성능을 집중적으로 평가하기 위해 상용 DBMS의 기본 연산 트랜잭션을 구성하여 기존 파일 시스템 기반의 DBMS와 비교 분석한다. 또한 표준 성능 평가 도구인 TPC-C 표준 벤치마크 테스트 도구를 활용하여 리눅스 시스템 및 Solaris 시스템 환경의 파일 시스템과 비교한다.

1. 서 론

최근 화이버 채널 인터페이스 기술 발전으로 인한 네트워크-부착 형 저장 장치들이 등장함에 따라, 네트워크 프로토콜 스택을 갖춰야 하는 NFS와 같은 전형적인 분산 파일 시스템의 구조가 SAN 기반의 공유 파일 시스템 구조로 변화되고 있다[1,2]. SAN 기반 공유 파일 시스템은 <그림 1>과 같이 분산 파일 시스템의 기능을 모두 제공하며, 또한 네트워크 부착형 저장 장치를 서버 없이 직접 저장 장치 전용 네트워크(SAN)에 접속시켜 사용하므로 가용성 및 확장성에 있어서 기존 분산 파일 시스템보다 우수하다. SANique™ 시스템[(1)]은 SAN 공유 파일 시스템으로서 기존 분산 시스템에서 서버가 모든 파일 공유의 제어를 담당해야 하는 단점을 극복할 수 있는 새로운 구조의 분산 공유 파일 시스템이다[3].



<그림 1> SAN 공유 파일 시스템

(1) 매크로임팩트㈜의 클러스터 파일 시스템

이러한 시스템에서는 중앙집중적인 서버 관리체제의 병목현상을 제거하기 위해 한 노드가 SAN 저장장치에 대한 접근제어를 관할하지 않고 클러스터 노드의 대부분이 전역적인 파일 시스템 운영에 참여하게 된다.

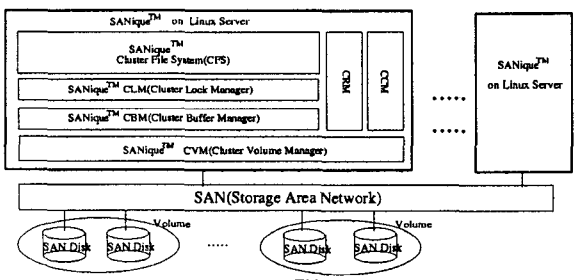
SAN 공유 파일 시스템은 클러스터 시스템에서 운영되는 파일 시스템으로 기존 파일 시스템의 역할을 대체할 수 있어야 하며, 기존 시스템 보다 우수한 성능을 보여야 한다. 본 논문에서는 SAN 공유 파일 시스템인 SANique™을 활용하여 기존 파일 시스템인 리눅스 시스템의 Ext3 파일 시스템과 Solaris 파일 시스템과의 기본 입·출력 연산에 대한 성능 평가를 수행한다. 파일 시스템에 대한 기본 입·출력 연산의 수행시간 측정을 위하여 각 파일 시스템 위에 상용 DBMS를 탑재하고, DBMS에 대한 기본 트랜잭션, 즉 검색, 삽입, 갱신, 삭제 연산을 집중적으로 포함하는 4개의 트랜잭션을 수행하여 각 파일 시스템간의 성능 평가를 보였다. 또한 집중적인 워크로드 발생과 표준 성능 측정을 위하여 DBMS의 표준 성능 측정 도구인 TPC-C 워크로드를 측정하여 각 파일 시스템의 기본 입출력 연산의 단위를 비교한다. 본 논문의 구성은 다음과 같다. 제 2장에서 SAN 공유 파일 시스템의 설계 및 구조에 대하여 설명하고, 제 3장에서 각 파일 시스템을 기반으로 한 상용 DBMS의 성능 측정 결과를 분석한다. 끝으로 제 4절에서 본 논문의 결론을 맺는다.

2. SANique™ 공유 파일 시스템 구조

SANiqueTM은 전형적인 분산 파일 서버 시스템 구조의 병목현상을 제거하고 입·출력 대역폭을 늘려 성능을 최대화한다. 또한 저장 장치로부터 응용에까지 파일 서버의 경유 없이 직접적인 데이터 전송을 하므로 데이터 전송 최단 경로를 보장하며 병렬 데이터 접근을 보장하여 응용 별 입·출력 대역폭을 증가시킬 수 있다[2,4]. SANiqueTM은 클러스터 파일 관리기(CFM), 클러스터 볼륨 관리기(CVM), 클러스터 로크 관리기(CLM), 클러스터 버퍼 관리기(CBM), 클러스터 회복 관리기(CRM)로 구성

되며, 이 서브 시스템 간의 구성은 <그림 2>와 같다.

SAN 상에 연결된 서버 클러스터들은 저장 장치 자원을 노드 간에 공유한다. 이 공유는 다른 노드에 대한 데이터 서비스를 지원하지 않는 점에서 NFS의 공유와 구별될 수 있다. 즉, SANique™ CFS에 의해 클러스터의 모든 단일 노드들은 SAN에 직접 연결된 디스크 장치들에 대해 병행적 공유 접근이 가능하다. SANique™의 CFS는 64비트 주소 공간을 갖는 저널링(journaling) 파일 시스템이다. 단일 파일의 크기에 제한을 두지 않으므로, 최대 2^{64} 바이트 크기의 단일 파일을 생성할 수 있다. SAN에 직접 연결된 논리적 디스크들을 기반으로 형성되는 SANique™ CFS는 슈퍼블록 영역, 비트맵 영역, 데이터 블록 영역으로 구성된다. 슈퍼블록 영역은 파일 시스템의 데이터를 관리하기 위한 메타 데이터를 저장하고, 비트맵 영역은 디스크 공간의 할당 영역과 가용 영역을 관리하기 위한 영역이다. 전형적인 분산 파일 시스템에서는 파일 서버가 비트맵 영역을 관리하지만, 공유 파일 시스템상에서는 여러 노드에 의해 관리 및 접근되므로 상호 배타적인 접근 방법이 필요하고, 비트맵 수정 연산에 대한 일관성 유지 방법이 필요하게 된다. SANique™ CFS에서는 비트맵 접근 단위를 세분화하여 다중 노드에 분산 시키므로 병렬성을 제공한 채 동시 접근 및 수정이 가능하다.



<그림 2> SANique™ 의 시스템 구성도

3. SANique™기반의 상용 DBMS 성능 평가

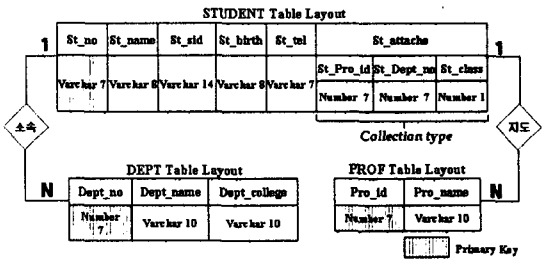
성능평가를 위해 사용된 환경은 다음 <표 1>과 같은 세가지 환경에서 질의 유형별 성능평가와 TPC-C를 이용한 성능평가를 수행하였다. Sanique™ 파일시스템과 Solaris 및 Ext3 파일시스템의 성능비교는 기복적인 파일 입·출력연산의 성능비교를 위해 Sanique™ 파일시스템을 클러스터 환경이 아닌 독립시스템으로 구성하여 동일한 환경하에서 성능 실험을 하였다.

CPU	1x Pentium(R) 1133Mhz		6 x 400Mhz
Memory	512M(512 Cache SRAM)		18 x 256MB
Disk	SCSI 40G 7200rpm		4 x SCSI 6G
OS	RedHat 9.0 (kernel :2.4.2-8)		Solaris 8
DBMS	Oracle 9i for Linux		Oracle 9i for Solaris
File System	Ext3	SANique™	Solaris File System

<표 1> Ext3, SANnique™, Solaris 파일시스템의 성능평가 환경

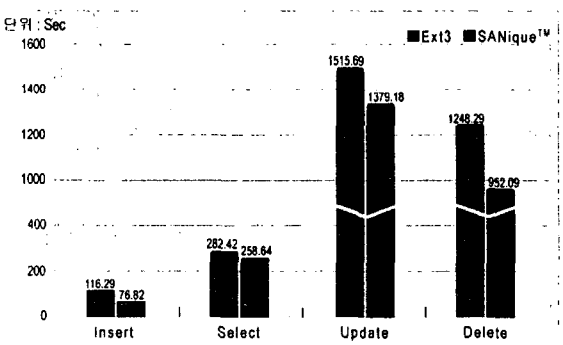
3.1 질의 유형별 성능평가

다음 <그림 3>과 같은 스키마의 데이터베이스를 생성하여 검색, 삽입, 갱신, 삭제의 네가지 유형의 연산을 집중적으로 수행한다. STUDENT 테이블은 디스크 입·출력의 부하를 극대화하기 위해 *Collection* 타입의 St_attache 속성을 포함하고 있다. 삽입에서는 각 테이블에 50000 건의 자료를 입력하고, 검색에서는 STUDENT 테이블과 PROF 테이블을 조인하여 5000건의 자료를 검색한다. 갱신에서는 *Collection* 타입 속성을 포함하여 모든 테이블을 갱신한다. 삭제는 모든 테이블의 모든 자료를 삭제한다. 이러한 네가지 유형의 질의를 수행하여 해당환경에서의 처리시간을 분석하였다.

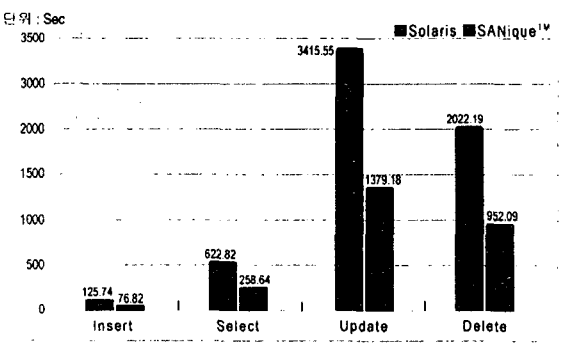


<그림 3> E-R Diagram

다음 <그림 4>, <그림 5>는 질의 유형별 성능측정을 각 파일 시스템에서 수행결과이다.



<그림 4> Ext3 vs SANique™ 질의 유형별 트랜잭션 성능측정



<그림 5> Solaris 파일시스템 vs Sanique™ 질의 유형별 트랜잭션 성능측정

모든 질의처리 시간이 Ext3[5]에 비해 SANique™가 빠른 것을 확인 할 수 있다. Ext3 와 SANique™, Solaris 파일시스템과 SANique™의 수행결과에서 빈번한 디스크

입·출력을 발생시키는 질의 수행 시간이 비교적 큰 폭의 차이를 보임을 확인 할 수 있다.

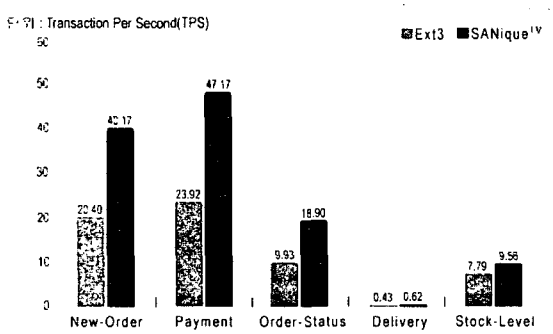
3.2 TPC-C 성능평가

TPC-C 는 신뢰도 높은 온라인 트랜잭션 프로세싱(OLTP) 작업의 성능과 확장성 측정 벤치마크(산업 표준)로 널리 쓰이고 있다. TPC-C 테스트는 약 0.6Gbyte 사이즈의 데이터베이스 스키마상에서 아래 <표 2>에서 5 개의 트랜잭션을 수행 하여 초당 처리 가능한 트랜잭션(TPS)의 수를 측정한 결과로 성능평가를 하였다. TPC-C 의 성능측정은 분당 처리한 New-Order 트랜잭션의 수(tpmC)이나, 본 성능측정에서 5 개의 트랜잭션을 수행결과를 모두 성능평가에 사용한다[6,7]. 아래 <표 2>는 TPC-C 에서 수행되는 트랜잭션의 간결한 내역이다.

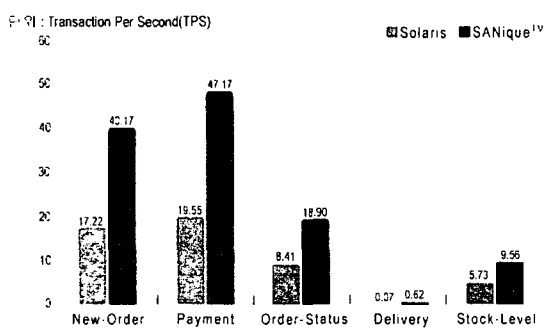
	New Order	Paym-ent	Order Status	Delve-ry	Stock Level
Select	22	4.2	3.6	30	2
Insert	12	1	0	0	0
Update	11	3	0	30	0
Delete	0	0	0	10	0
Non-uinque Select	0	0.6	0.6	0	0
Join	1	0	0	0	0

<표 2> 트랜잭션 질의 유형

다음 <그림 6>, <그림 7>은 질의 TPC-C 성능측정을 각 파일 시스템에서 수행결과이다.



<그림 6> Ext3 vs SANique™ TPC-C 성능측정



<그림 7> Solaris 파일시스템 vs SANique™ TPC-C 성능측정

SANique™ 파일시스템에서 초당 처리 한 트랜잭션의 수가 Ext3 파일시스템에서 보다 월등히 많음을 확인 할 수 있다. 또한 SANique™의 결과는 <그림 7>에서 알 수 있듯이 Solaris 파일시스템에 비해서도 우월한 성능을 보이고 있다. 파일시스템의 차이와 DBMS 성능측정 결과의 연관성을 확인해 보면 디스크 입·출력의 발생을 비교적 적게 유도하는 검색, 삭제 질의를 많이 포함하는 트랜잭션에서는 그 차이가 미비하다. 반면 디스크 입·출력을 빈번하게 유도하는 갱신, 삽입 질의를 다수 포함하는 트랜잭션의 경우 성능 차이가 현저함을 알 수 있어, SANique™ 파일 시스템의 디스크 접근 관련 시스템 호출 구현의 우수성을 입증하였다. 또한 본 SANique™파일 시스템은 클러스터내의 타 서버들에게도 동일한 뷰를 제공하는 공유 기능을 제공하면서 동시에 우수한 디스크 접근 성능을 보이고 있음을 알 수 있다.

4. 결론 및 향후 연구

본 논문에서는 SAN 공유 파일 시스템인 SANique™을 활용하여 기존 파일 시스템인 리눅스 시스템의 Ext3 파일 시스템과 Solaris 파일 시스템과의 기본 입·출력 연산에 대한 성능 평가를 수행하였다. 파일 시스템에 대한 기본 입·출력 연산의 수행시간 측정을 위하여 각 파일 시스템 위에 상용 DBMS를 탑재하고, DBMS에 대한 기본 트랜잭션, 즉 검색, 삽입, 갱신, 삭제 연산을 집중적으로 포함하는 4개의 트랜잭션을 수행하여 각 파일 시스템간의 성능 평가를 보였으며, 표준 성능 측정을 위해 DBMS의 표준 성능 측정 도구인 TPC-C 워크로드를 측정하여 각 파일 시스템의 기본 입·출력 연산의 단위를 비교하였다. 현재 클러스터 파일 시스템 기반의 클러스터 DBMS 성능평가를 진행 중이며, 상용 클러스터 파일 시스템간의 비교를 계획 중이다.

참고문헌

[1] C. C. Fan and J. Bruck, "The Raincore Distributed Session Service for Networking Elements", Proc. Of the International Parallel and Distributed Processing Symposium, 2001.

[2] P. S Weygant, "Primer on Clusters for High Availability", Technical Paper at Hewlett-Packard Labs, CA, 2000

[3] Sang G. Oh, and Jang S. Lee, "SANique™ : A SAN File system for Linux Cluster", Technical White Paper - Draft, MacroImpact. Co. Ltd., 2004

[4] M. D. Dahlin, "Severless Network File Systems", Ph. D. Thesis at Computer Science Graduate Division of University of California at Berkely, 1995

[5] Bouchra Bouqata, Christopher D. Carothers, Boleslaw K. Szymanski, and Mohammed J. "Understanding Filesystem Performance for Data Mining Applications", Proc. 6th International Workshop on High Performance Data Mining: Pervasive and Data Stream Mining (HPDM:PDS'03) at the Third International SIAM Conference on Data Mining, San Francisco, CA, 2003

[6] Transaction Processing Performance Council(TPC), "TPC BENCHMARK™ C Standard Speccfication", 2005

[7] S. Leutenegger and D. Diaz, "A Modeling Study of the TPC-C Benchmark", Proc. ACM SIGMOD International Conference on Management of Data, pages 22-31, Washington D.C., 1993