

분산 파일 시스템 GLORY-FS - 1 : 초고속 대용량 분산 파일시스템 GLORY-FS 소개

김영균 한국전자통신연구원 스토리지시스템 연구실장

개요

최근 클라우드, 빅데이터, 사물인터넷으로 이어지는 IT 트렌드의 공통점은 이들이 다루는 데이터의 양이 많다는 것이며, 이렇게 막대한 양의 데이터를 저장하고 관리하기 위해 지불해야 하는 비용을 최대한 줄이는 것이 가장 큰 이슈로 부각되고 있다는 것이다. 따라서 이들이 활용하는 IT 인프라를 저렴하게 구축하고 효율적으로 활용할 수 있는 기술들에 대한 연구 및 실용화가 활발히 진행되고 있으며, 막대한 양의 데이터를 수많은 서버들에 분산 저장하고 관리하는 분산 파일시스템 기술은 그러한 기술 중의 하나이다. 상기의 IT 트렌드들에서 요구하는 분산 파일시스템은 단순히 데이터의 저장과 관리만을 하는 것이 아니라 하드웨어의 장애에 유연하게 대처하여 서비스가 중단되지 않도록 하고 적절한 병렬 처리를 통해 서비스가 요구하는 성능도 만족시킬 수 있어야 한다. 대표적으로 구글의 Google 파일시스템, Apache의 Hadoop 파일시스템, 레드햇 Gluster 파일시스템, 그리고 인텔 Lustre 파일시스템과 같은 분산 파일시스템들이 존재하지만 최근까지 클라우드 분야에 활용되고 있는 분산 파일시스템은 그리 많지 않다. 한국전자통신연구원(ETRI)에서 개발한 GLORY-FS 파일시스템은 UCC, IPTV, 스토리지 호스팅과 같은 클라우드 서비스에 최적화된 파일시스템으로 1천 개 이상의 서버들로 이루어진 스토리지 공간을 단일 스토리지 공간으로 가상화할 수 있다. 이러한 가상화를 통하여 수십 페타바이트 이상으로 스토리지 공간 확장이 가능하며, 서버의 추가에 따른 선형적인 입출력 성능 향상을 보장한다. 또한, 일반적인 저비용 하드웨어 환경에서 높은 가용성과 신뢰성을 보장하는 파일 서비스를 제공할 수 있다.

특징 및 장점

GLORY-FS 파일시스템은 수천 대의 저비용 서버들을 이용하여 저장 공간 구축에 드는 비용을 최소화하면서도 장애에 대한 효율적인 통제 능력과 높은 입출력 처리 성능을 갖춘 대규모 클라우드 서비스를 위한 분산 파일시스템이다. GLORY-FS는 다음과 같은 특징들을 가진다.

저렴하고 합리적인 비용

일반적인 저비용 하드웨어를 사용하여 소유 총비용(TCO : Total Cost of Ownership)를 절감한다.

대규모 스토리지 제공을 위한 온라인 서버 확장

온라인으로 디스크나 데이터 서버를 추가함으로써 스토리지 공간의 총 크기를 PetaByte 규모로까지 확장 가능하다. 서비스할 데이터의 증가에 따라 스토리지 공간이 부족할 경우 60초 이내에 신규 서버를 추가함으로써 유연한 서비스 운용 환경을 제공한다.

선형적으로 확장 가능한 최적화된 성능

클라이언트의 데이터 I/O 요청을 서로 다른 데이터 서버에 효과적으로 분배하여, 모든 데이터 서버의 활용률을 높임으로써 최대의 통합 I/O 성능을 제공한다. 메타데이터 서비스를 처리하는 다수의 서버를 활용함으로써 최대의 통합 메타데이터 성능을 제공한다.

자동 복구를 통한 중단 없는 서비스

효과적인 데이터 복제(Replication) 및 복구 정책을 사용하여 365일 무정지 서비스를 제공한다.

지능적이고 유연한 스토리지 관리

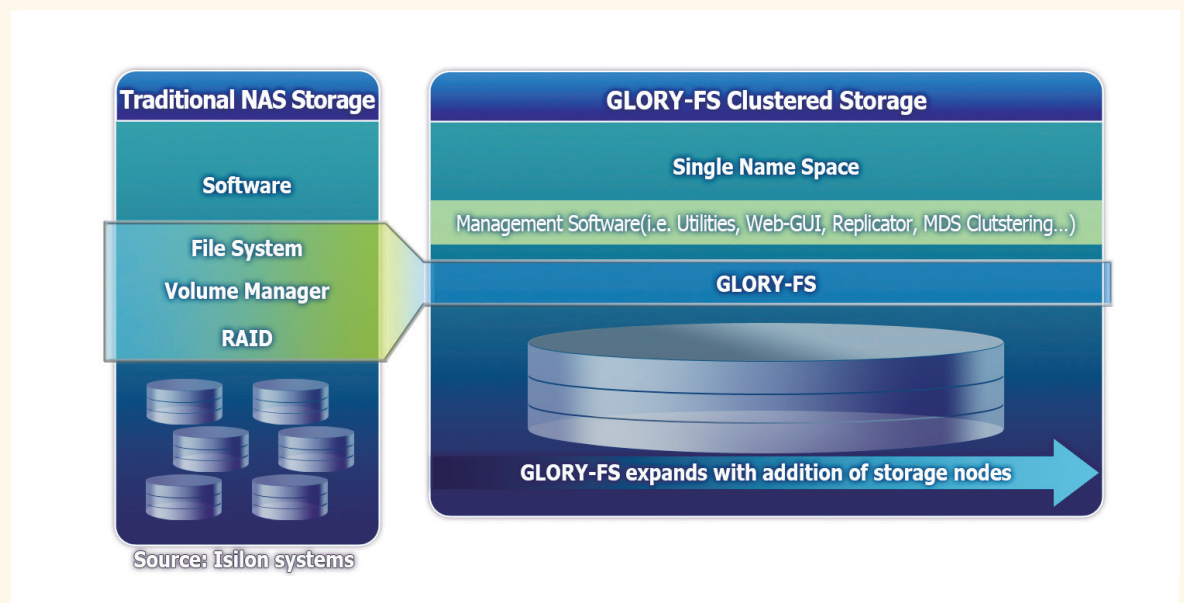
자가 감시(Self-Tuning), 자가 치유(Self-Healing & Self-Diagnosis), 스토리지 불균형 해소(DS Rebalance), 효율적인 스토리지 활용과 같은 지능적인 스토리지 관리 기능을 제공한다.

강력하고 편리한 관리 도구

셸기반 유틸리티 및 웹 관리 도구를 통해 편리한 시스템 운영을 지원한다.

기존 응용들을 지원하는 표준화된 API

POSIX(Portable Operating System Interface) 호환 API를 제공함으로써 기존 응용 서비스를 그대로 적용할 수 있다. 또한, 윈도우 파일시스템 호환 API를 제공함으로써 윈도우 응용 서비스를 적용할 수 있다.



[그림 1] GLORY-FS 파일시스템 개요

시스템 구성

GLORY-FS는 메타데이터를 저장하고 관리하는 서버와 데이터를 저장하고 관리하는 서버를 독립적으로 분리하는 out-of-band 구조를 지원한다. 시스템은 관리 서버(Management Server(MGT)), 메타데이터 서버(Metadata Server(MDS)), 데이터 서버(Data Server(DS)), 클라이언트 파일시스템(Client Filesystem(CLIENT))으로 구성된다.

Management Server(MGT)

관리 서버 블록은 파일시스템을 구성하고 관리하는 주 서버로서 메타데이터 서버나 데이터 서버를 등록하고 주기적으로 상태 정보를 수집하여 시스템 관리에 활용한다.

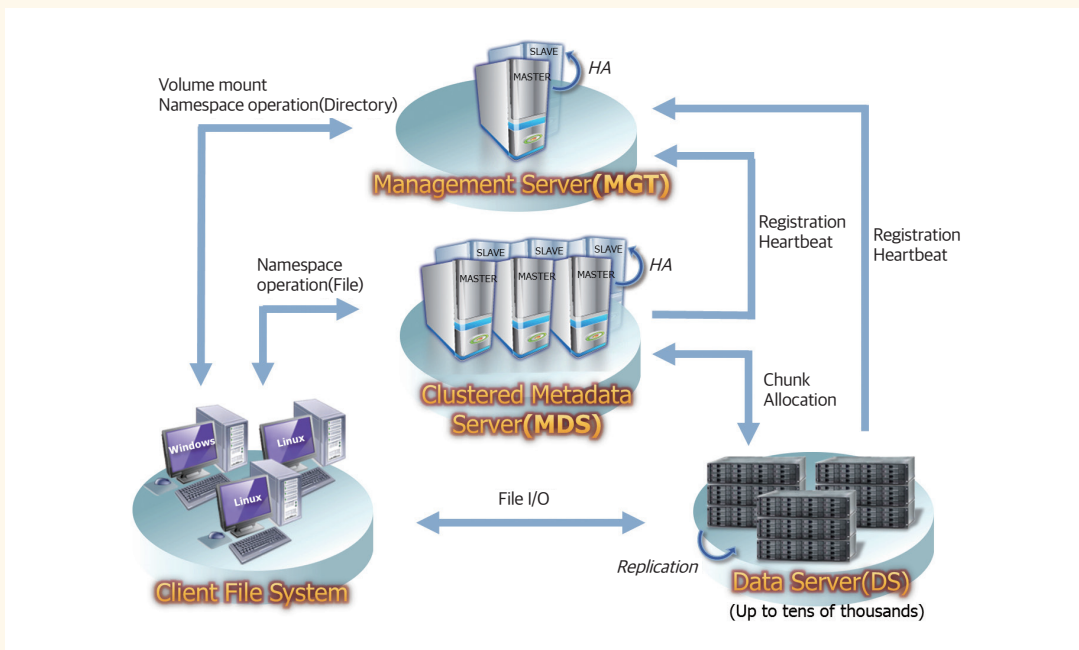
GLORY-FS에서는 파일시스템 운영에 필요한 디렉터리와 파일의 메타데이터를 분리하여 관리한다. MGT에서는 디렉터리 메타데이터를 DB에 저장 및 관리하며, 디렉터리에 대한 namespace 연산을 처리한다. 등록된 모든 데이터 서버들을 가상화시켜 대규모 단일 스토리지로 제공하거나 다수의 논리적 볼륨들을 생성하여 사용자마다 볼륨 단위로 제공한다.

MGT에서는 시스템 운용에 필요한 다양한 정보를 수집하여 관리한다. MDS나 DS 노드의 상태 정보(CPU 활용률, 메모리 사용량, 네트워크 대역폭, 디스크 용량 및 성능 등)와 파일시스템 API 통계 정보 등으로 관리자는 모니터링 유틸리티를 활용하여 확인할 수 있다.

Metadata Server(MDS)

메타데이터 서버 블록은 디렉터리 단위로 저장된 파일들의 메타데이터를 DB에 저장 및 관리하고, 파일에 대한 namespace 연산을 처리한다. 또한, 파일시스템에서 보다 많은 파일을 저장 관리하고 메타데이터 연산 처리 성능을 높이기 위하여 다수의 MDS를 클러스터링하여 구성할 수 있다.

단일 혹은 연속적인 데이터 서버 고장이 발생하더라도 지속적인 파일시스템 서비스를 제공할 수 있도록 파일 데이터를 여러 데이터 서버에 복제하거나 장애가 발생한 데이터 서버에 저장된 파일 데이터를 다른 데이터 서버에 복제하여 복구하는 기능을 수행한다.



[그림 2] GLORY-FS 시스템 구성과 요소들간 인터액션

Data Server(DS)

데이터 서버 블록은 클라이언트로부터의 파일 데이터 입출력 요청을 처리하고, 메타데이터 서버로부터의 파일 생성, 삭제, 크기 변경 등의 요청을 처리한다. 저장되는 청크 데이터의 유효성을 보장하기 위하여 버전 정보를 저장 및 관리하고, 요청에 따라 해당 청크의 유효성 검사를 위해 CRC 값을 산출하는 기능을 수행한다. 또한, 관리 서버에 클러스터로 참여 및 탈퇴, 주기적인 상태 정보를 보고하고, 클라이언트로부터의 실시간 복제 요청 또는 메타데이터 서버의 지시를 따라 다른 데이터 서버로의 청크 복제 작업을 수행하는 기능을 수행한다.

시스템을 구성할 때 복제 전용의 네트워크를 설정하는 경우 복제 트래픽과 입출력 처리 트래픽을 분리할 수 있으면, 안전한 데이터 복제를 보장할 수 있다.

Client File System(CLIENT)

클라이언트 파일시스템 블록은 리눅스와 윈도우 플랫폼에서 운용할 수 있는 사용자 수준의 파일시스템이다.

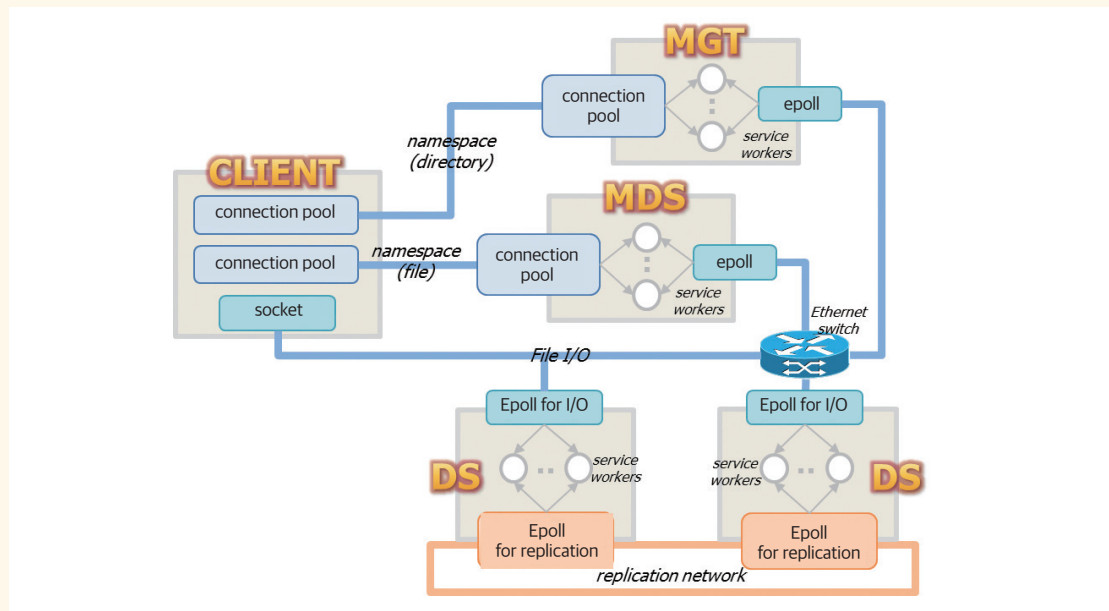
리눅스 플랫폼에서는 오픈 소스인 FUSE(Filesystem in Userspace)를 기반으로 하며, 클라이언트 응용에 POSIX에 준하는 파일시스템 API를 제공한다. 또한, 32비트 또는 64비트 시스템에서 운용될 수 있으며, 리눅스 커널 버전에 독립적이고 커널 패치를 필요로 하지 않는다.

윈도우즈 플랫폼에서는 FUSE와 유사한 Eldos 사의 Callback Filesystem(CBFS)을 기반으로 하며, 윈도우즈 파일시스템 API를 처리하는데 필요한 CBFS API를 제공한다. 윈도우즈 8에서 윈도우즈 2012 Server까지 다양한 OS를 지원하며, 네트워크 드라이브로 마운트하여 가상 디스크를 통해 입출력 요청을 처리한다.

통신 프로토콜 및 파일 공유

GLORY-FS에서 블록간 요청은 네트워크 기반의 RPC(Remote Procedure Call) 형태로 처리한다. 주로 TCP(Transmission Control Protocol)로 통신하지만 DS의 상태 정보(DS Heartbeat)는 UDP(User Datagram Protocol)로 통신한다.

MGT, MDS, DS 통신 처리 모듈에서는 여러 클라이언트나 DS의 요청을 감지하고 처리하기 위하여 리눅스 epoll을 사용하고, epoll에서 감지되어 읽어 들인 요청은 독립적인 서비스 쓰레드를 생성하여 처리한다. MGT 서버는 안전한 시스템 운용을 위하여 MDS와 DS로부터 주기적으로 상태 정보를 수집하는데 MDS로부터는 TCP 소켓으로, DS로부터는 UDP 소켓으로 상태 정보를 전달받는다. 클라이언트 파일시스템은 MGT나 MDS와의 효율적인 통신을 위하여 마운트할 때 지정된 수의 소켓 연결 풀을 생성하고, 이를 이용하여 통신 처리를 수행한다. DS와는 파일 데이터에 대한 순차 접근을 최적화하기 위하여 한 번의 읽기 또는 쓰기 요청 이후에 데이터를 계속적으로 전송 또는 수신하는 형태로 통신한다.



[그림 3] GLORY-FS 통신 프로토콜 정의

GLORY-FS는 대규모 인터넷 서비스에 적합하도록 설계된 파일시스템으로 파일 공유 방식은 한 노드가 일정한 시간이 지나면 다른 노드에 의해서 변경된 정보를 얻을 수 있는 NFS 공유 시맨틱을 따른다. 이는 대부분의 인터넷 서비스에서 파일을 생성한 후 변경이 거의 없이 순차적인 읽기만 하는 워크 로드를 주로 요청하기 때문에 동일 파일에 대해 여러 노드간 일관성을 지원하는 잠금기반의 엄격한 노드간 동기화가 불필요하기 때문이다.

성과와 용량의 선형 확장

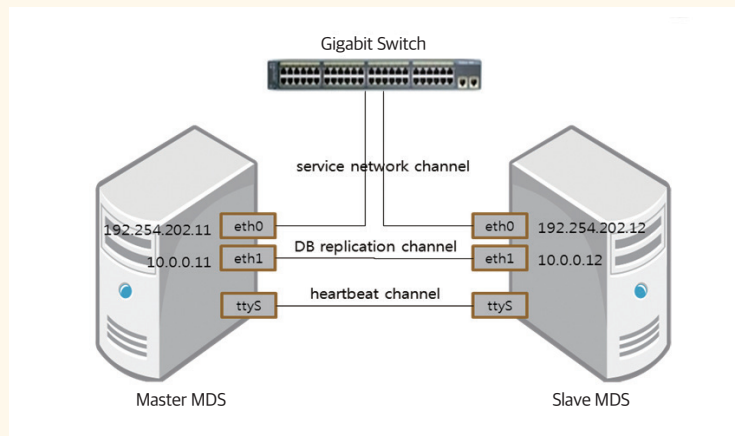
GLORY-FS는 데이터 및 메타데이터 측면에서 성과와 저장 공간을 선형적으로 확장할 수 있다. 데이터 측면에서는 파일시스템이 운용되는 중에 온라인으로 DS를 추가하거나 DS 프로세스가 운용 중인 상태에서 디스크를 추가시킴으로써 전체 스토리지 공간을 확장시킬 수 있다. 또한, 추가한 DS 만큼 네트워크 대역폭이 증가하기 때문에 전체 데이터 입출력 성능도 증가한다. 신규 디스크가 추가됨으로써 발생하는 기존 디스크와의 사용량 불균형 문제는 DS Rebalance 기능으로 해결할 수 있다.

메타데이터 측면에서는 여러 MDS를 클러스터로 구성하고, MGT에서 디렉터리 단위로 MDS를 할당하여 해당 디렉터리에 저장되는 파일들을 관리하도록 한다. 그리고 온라인으로 MDS를 클러스터에 추가함으로써 메타데이터 저장 공간이 증가하고, 단일 파일시스템에서 관리할 수 있는 파일의 수는 선형적으로 증가한다. 또한, 시스템 수준에서 메타데이터의 처리 성능도 비례하여 증가한다.

고가용성 메커니즘

GLORY-FS는 인터넷 서비스 특성상 무정지 서비스를 제공할 수 있도록 높은 수준의 장애 대처 기능을 제공한다.

메타데이터 가용성을 지원하기 위하여 MGT 및 MDS는 가능한 HA(High Availability) 모드로 운용한다. 이 모드에서는 Master 노드와 Slave 노드를 설정하고, 리눅스 heartbeat 패키지를 이용하여 상호 상태를 확인한다. 그리고 DBMS의 비동기적인 복제 기능을 활용하여 Master 노드의 DB에 저장된 메타데이터를 Slave 노드의 DB로 복제함으로써 메타데이터 동기화를 수행한다. Master 노드에 장애가 발생하면 상호 감시를 통해 Slave 노드가 이를 감지하여 자신을 Master 노드로 승격한 후 시스템을 운용한다.



[그림 4] 고가용성 보장을 위한 MDS 이중화 구성

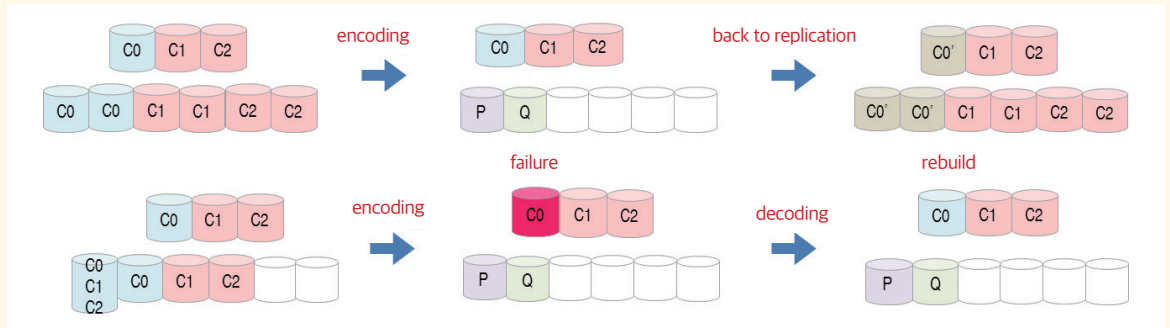
데이터 가용성을 지원하기 위하여 파일의 데이터는 복제를 기반으로 저장한다. DS나 디스크로 접근이 되지 않거나 디스크 입출력 처리 중에 오류가 발생하면, 이를 감지하고 자동으로 복제를 통해 복구함으로써 데이터 손실을 방지한다. 그리고 데이터 중요도, 스토리지 효율성 등을 고려하여 저장하는 복제본의 수도 디렉터리 단위로 조정할 수 있다.

효율적인 스토리지 공간 활용

GLORY-FS는 복제 방식으로 데이터를 저장함으로써 데이터 가용성을 보장한다. 복제 방식을 선호하는 이유는 여러 복제본을 저장함으로써 접근 요청을 분산 처리하는 효과와 함께 장애가 발생하는 경우 복제본을 이용할 수 있어 장애로 인한 성능 저하를 최소화할 수 있기 때문이다. 그러나 저장하는 복제본 배수만큼 저장 공간이 낭비되는 단점이 있다. 데이터에 따라 어느 정도 시간이 지나면 접근이 현저하게 줄어들거나 백업의 용도로 저장되는 경우가 있다.

GLORY-FS는 ILM(Information Lifecycle Management) 측면에서 오랫동안 접근되지 않는 파일을 복제 방식이 아닌 데이터를 인코딩(패리티)하여 데이터와 함께 저장하는 기능을 제공한다. 일반적인 RAID와 유사한 방식으로 사용자에게 의해 설정된 데이터 단위로 Reed-Solomon 코딩 알고리즘으로 인코딩하는 방식이다. 데이터에 장애가 발생하면 저장된 패리티를 이용하여 복구하며 이를

디코딩이라 한다. 복구 가능한 데이터 수는 저장된 패리티 수와 동일하다. 인코딩하는 데이터 크기와 패리티 저장에 필요한 스토리지 용량은 반비례한다. 즉, 인코딩하는 데이터 수가 증가할수록 필요한 스토리지 용량은 감소한다. 그에 반해 인코딩이나 디코딩 처리 시간은 길어지기 때문에 전체 성능은 저하된다. 따라서 패리티 수에 따라 인코딩하는 데이터 수를 적절하게 설정하는 것이 중요하다.

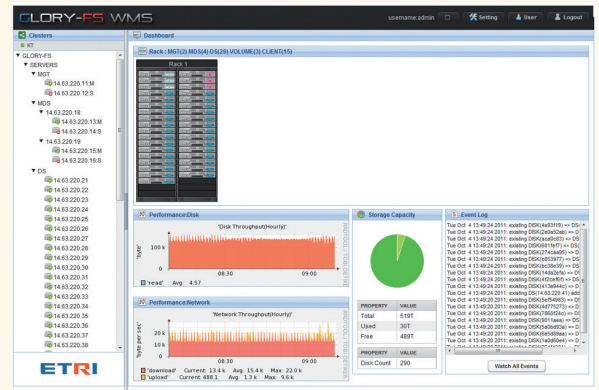


[그림 5] 파일 기반 RAID 관리 기술

GLORY-FS 파일시스템에 데이터 갱신은 복제 방식으로 저장된 파일에 대해서만 지원한다. 따라서 인코딩된 파일의 데이터를 갱신하려면 해당 파일을 다시 복제 방식으로 전환한 후 처리해야 한다.

지능적인 모니터링 및 관리

GLORY-FS는 시스템 규모에 따라 소규모부터 Petabyte 급까지 폭넓은 구성이 가능하다. 사용자(관리자 포함)에게 사용의 편의성을 제공하기 위하여 다양한 관리 및 모니터링 유틸리티를 제공한다. 또한, 언제 어디서나 시스템의 운용 상황을 모니터링할 수 있는 웹기반의 관리 도구를 지원한다.



[그림 6] 웹 기반 모니터링 및 관리 도구

결론

이번 호에서는 2007년 우리 기술로 자체 개발을 시작하여 현재까지 지속적인 기능과 성능 개선을 통해 다양한 활용분야에서 그 우수성을 인정받고 있는 GLORY-FS 파일시스템에 대해 살펴보았다.

GLORY-FS 파일시스템은 비용적인 측면에서의 효율성, 지속해서 증가하는 데이터의 수용, 빈번하게 발생하는 고장에 대한 대처, 관리의 편리성, 신속한 입출력 성능, 데이터 최적 배치, 효과적인 캐시 사용, 부하 집중에 대한 유연한 대처, 데이터에 대한 보안 등과 같이 분산 파일시스템이 갖추어야 할 사항들을 집약하고 있다. 현재는 국내 통신사와 방송사들로부터 그 우수성을 인정받아 실제 클라우드 서비스는 물론 데이터 아카이빙 및 스트리밍 서비스를 지원하는 분산 파일시스템으로 운용되고 있다. 그동안 글로벌 제품 일색이던 스토리지 시장에 변화의 바람이 불고 있는 것이다.

다음 호에서는 그동안 여러 분야에 적용된 사례들을 통해서 각 분야의 다양한 데이터 워크로드 특성을 수용하는 스토리지 솔루션으로서의 모습을 소개하도록 하겠다. ☺

※ 본 기술은 미래창조과학부 SW컴퓨팅산업원천기술개발사업의 일환으로 지난 2012년부터 ETRI가 수행한 개발한 결과물이다.