(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2012 년 6 월 28 일 (28.06.2012)

WIPO | PCT

(10) 국제공개번호
WO 2012/086919 A2

(51) 국제특허분류:
G06F 15/16 (2006.01) *G06F 9/06* (2006.01)
H04L 12/24 (2006.01)

(21) 국제출원번호: PCT/KR2011/008223

(22) 국제출원일: 2011 년 10 월 31 일 (31.10.2011)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2010-0134841 2010 년 12 월 24 일 (24.12.2010) KR

(71) 출원인 (US 을(를) 제외한 모든 지정국에 대하여): **(주) 케이티 (KT CORPORATION)** [KR/KR]; 경기도 성남시 분당구 정자동 206, 463-711 Gyeonggi-do (KR).

(72) 발명자; 겸
(75) 발명자/출원인 (US 에 한하여): **김미점 (KIM, Mi-Jeom)** [KR/KR]; 경기도 안양시 동안구 관양동 1608-2, 현대 아이스페이스 606 호, 431-060 Gyeonggi-do (KR). **김효민 (KIM, Hyo-Min)** [KR/KR]; 서울시 서초구 우면동 40-1 103 호, 137-900 Seoul (KR). **박창식 (PARK, Chang-Sik)** [KR/KR]; 서울시 강남구 일원본동 가람아파트 108 동 407 호, 135-947 Seoul (KR). **황진경 (HWANG, Jin-Kyung)** [KR/KR]; 경기도 과천시 중앙동 삼성레미안 APT 1108 동 705 호, 427-010 Gyeonggi-do (KR).

(74) 대리인: **리앤목특허법인 (Y.P.LEE, MOCK & PARTNERS)**; 서울시 서초구 서초동 1575-1, 137-875 Seoul (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

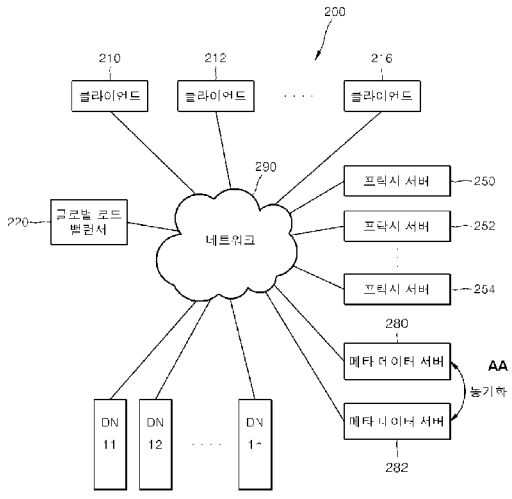
(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[다음 쪽 계속]

(54) Title: DISTRIBUTED STORAGE SYSTEM INCLUDING PLURALITY OF PROXY SERVERS, METHOD FOR MANAGING OBJECT THEREOF, AND STORAGE MEDIUM READABLE BY COMPUTER

(54) 발명의 명칭 : 복수 개의 프락시 서버를 포함하는 분산 저장 시스템 및 그 오브젝트 관리 방법 및 컴퓨터에 의하여 독출가능한 저장 매체

[Fig. 2]



AA ... Synchronization
210, 212, 216 ... Client
220 ... Authentication server
250, 252, 254 ... Proxy server
280, 282 ... Metadata server
290 ... Network

(57) Abstract: Disclosed are a distributed storage system including a plurality of proxy servers, and a method for managing an object thereof. The distributed storage system of the present invention comprises: a metadata server for storing metadata which includes the unique information of the object and the unique information of data nodes onto which the object is stored; at least two proxy servers for relaying an action request by the client for performing a managing action to a target data node, which is related to a target object that is the object of the managing action performed by the client, by referencing the metadata server when the action request is received; and a global load balancer for authenticating the client, and for selecting a target proxy server to process the action request by the client from the proxy servers, wherein the target data node performs an action according to the action request that is received from the target proxy server and transmits the result of the performance to the target proxy server or to the client. According to the present invention, scalability and resilience, which are advantages of cloud computing, can be provided by introducing multiple proxy servers, and the distributed storage system can have a flexible structure.

(57) 요약서:

[다음 쪽 계속]

WO 2012/086919 A2

공개:

— 국제조사보고서 없이 공개하며 보고서 접수 후 이를 별도 공개함 (규칙 48.2(g))

복수 개의 프락시 서버를 포함하는 분산 저장 시스템 및 그 오브젝트 관리 방법이 개시된다. 본 발명에 의한 분산 저장 시스템은 오브젝트의 고유 정보 및 오브젝트가 저장된 데이터 노드의 고유 정보를 포함하는 메타데이터를 저장하는 메타데이터 서버, 관리 동작을 수행하고자 하는 클라이언트의 동작 요청이 수신되면, 메타데이터 서버를 참조하여 상기 클라이언트가 수행하는 관리 동작의 객체인 대상 오브젝트와 관련되는 대상 데이터 노드에 상기 동작 요청을 전달하는 적어도 두 개의 프락시 서버들, 및 클라이언트를 인증하며, 프락시 서버들 중에서 클라이언트의 동작 요청을 처리할 대상 프락시 서버를 결정하는 글로벌 로드 밸런서를 포함하며, 대상 데이터 노드는 대상 프락시 서버로부터 수신되는 동작 요청에 따른 동작을 수행하고 수행 결과를 대상 프락시 서버 또는 클라이언트로 전송하도록 구성된다. 본 발명에 의하여, 여러 개의 프락시 서버를 도입함으로써 클라우드 컴퓨팅의 장점인 확장성과 탄력성이 제공되며, 분산 저장 시스템이 유연한 구조를 가지게 된다.

명세서

발명의 명칭: 복수 개의 프락시 서버를 포함하는 분산 저장 시스템 및 그 오브젝트 관리 방법 및 컴퓨터에 의하여 독출가능한 저장 매체

기술분야

[1] 본 발명은 클라우드 컴퓨팅에서 분산 저장 서비스를 효율적으로 제공하기 위하여 다중 프락시 서버를 이용한 오브젝트 분산 저장 시스템에 관한 것으로서, 특히, 오브젝트의 중복 제거(Deduplication) 및 복제(replication)를 효과적으로 수행하기 위해 오브젝트의 물리적 위치를 저장하는 메타데이터를 중앙에 이중화하여 두고 모든 프락시 서버에서 해당 메타데이터를 공동 참조하는 모델에 관한 것이다.

배경기술

[2] 클라우드 컴퓨팅은 인터넷 망을 통해 여러 가지 IT(Information Technology) 자원들을 분배하여 서비스하는 개념으로 가장 일반적인 서비스 분류로는 첫째, 하드웨어 인프라를 서비스로 제공하는 IaaS(Infrastructure As A Service), 둘째, 응용 개발 및 실행 플랫폼을 서비스로 제공하는 PaaS(Platform As A Service), 그리고 마지막으로 어플리케이션을 서비스로 제공하는 SaaS(Software As A Service)로 나눈다.

[3] IaaS에는 많은 서비스 카테고리가 있는데, 대표적으로 연산 자원을 가상 머신 형태로 제공하는 연산 서비스와 저장 서비스가 있다. 이러한 분산 저장 시스템은 클라우드 스토리지 서비스를 제공하는데, 이를 이용하면 저사양의 하드웨어를 활용하여 공용의 저장소 풀(storage pool)을 만들어 탄력적이고 유연한 사용량을 적시에 만족하는 특징을 가진다. 이를 위해 널리 이용되는 단순하고 강력한 오브젝트 기반 스토리지 기법에서는 물리적인 저장 공간 관리 기능을 저장 장치 자체에서 직접 수행하게 한다. 그러므로, 저장 장치의 성능이 향상되고 손쉽게 저장 장치의 용량을 확장할 수 있다. 또한, 플랫폼과 독립적으로 데이터를 안전하게 공유할 수 있는 특징을 가진다.

[4] 도 1은 종래 기술에 의한 분산 저장 시스템을 개념적으로 나타내는 도면이다.

[5] 도 1에 도시되는 오브젝트 스토리지 시스템은 클라이언트의 인증을 처리하는 인증 서버, 클라이언트의 요구사항을 처리하는 프락시 서버(proxy server)(또는 마스터 서버), 오브젝트들의 물리적인 위치를 포함하는 메타데이터를 저장하는 메타데이터 데이터베이스, 실제 오브젝트 저장 및 관리를 담당하는 데이터 노드, 그리고 데이터의 복제 등을 관리하는 복제 서버(replicator server) 등을 포함한다. 클라이언트는 초기에 인증 서버를 통해 인증을 받으며, 인증이 완료된 후에는 프락시 서버에게 원하는 오브젝트를 관리하는 데이터 노드의 정보를 요청한다. 클라이언트의 요청에 대해 프락시 서버는 메타데이터를 참조하여 해당하는

데이터 노드에게 원하는 동작 요청을 전달하고, 데이터 노드는 동작을 수행한 결과를 프락시 서버를 통해 클라이언트에게 전달한다. 또는, 데이터 노드는 프락시 서버를 거치지 않고 직접 클라이언트에게 응답을 제공할 수도 있다. 이 경우 지연이나 데이터 트래픽이 감소하는 효과를 기대할 수 있으나 모든 데이터 노드가 클라이언트 인터페이스를 가져야 함으로 데이터 노드의 복잡도가 증가할 수 있다.

[6] 오브젝트 저장소는 데이터의 안전성과 높은 가용성을 위해 데이터를 복제(replication)하며 이런 복제본을 레플리카(replica)라고 한다. 널리 이용되는 분산 저장 시스템은 일반적으로 2개 내지 3개의 복제본을 가지지만, 오브젝트의 중요도에 따라 더 많은 복제본을 가지기도 한다. 오브젝트의 복제본은 서로 동기를 유지해야 하며 이는 보통 별도의 복제 서버 (replicator server)에서 처리한다.

[7] 그런데, 종래 기술에 의한 분산 저장 시스템의 경우, 많은 개수의 클라이언트들이 모두 동일한 프락시 서버를 참조하여 복수 개의 데이터 노드에 접근한다. 그 결과, 분산 저장 시스템의 클라이언트 및 데이터 노드의 개수가 증가할수록 해당 프락시 서버의 처리 부담이 현저히 증가한다. 특히, 프락시 서버는 클라이언트의 요구를 받아 요청을 처리할 수 있는 데이터 노드로 요구사항을 전달하고 응답을 다시 데이터 노드에서 받아 클라이언트에 전달하기 때문에 많은 로드가 집중된다. 결국, 종래 기술에 의한 분산 저장 시스템에서는, 시스템 자원의 효율적인 이용을 위해 분산 저장 개념을 도입했으면서도, 프락시 서버에 의하여 처리 속도가 지연되는 치명적인 약점이 존재한다.

[8] 더 나아가, 단일 프락시 서버가 오작동을 하거나 다운된 경우, 해당 프락시 서버가 담당했던 분산 저장 시스템 전체의 동작이 불가능한 치명적인 집중 실패 문제(소위, point of failure problem)도 존재한다. 또한, 종래의 분산 저장 시스템에서는 단일 프락시 서버의 처리 용량 이상으로 시스템을 확장하는 것이 불가능하다.

[9] 그런데, 클라우드 컴퓨팅의 장점인 확장성과 탄력성 등을 제공하기 위해 유연한 구조를 가져야 함으로 프락시 서버의 다중화 구조를 지원할 필요성이 있다. 또한, 프락시 서버별로 메타데이터를 가지는 경우 메타데이터간 데이터 동기화 문제가 이슈화됨으로 메타데이터를 공유하는 구조가 역시 유리하다.

[10] 따라서, 클라우드 스토리지 서비스에서 요구되는 확장성과 유연성을 확보하는 것은 물론, 단일 프락시 서버의 단점인 집중 실패 문제점을 극복하면서도, 메타데이터를 안전하게 관리할 수 있는 기술이 절실히 요구된다.

발명의 상세한 설명

기술적 과제

[11] 본 발명의 목적은 클라우드 스토리지 서비스를 위해 유연하고 효율적인 멀티

프락시 오브젝트 스토리지 구조를 제공하는 것이다.

[12] 또한, 본 발명의 목적은, 클라우드 컴퓨팅의 장점인 확장성과 탄력성 등을 제공하기 위해 여러 개의 프락시 서버를 도입함으로써, 분산 저장 시스템이 유연한 구조를 가지도록 하는 것이다.

[13] 또한, 본 발명의 목적은 멀티 프락시 서버 분산 저장 시스템에서 메타데이터를 공유하도록 함으로써 모든 프락시 서버가 안정적으로 동일한 메타데이터를 참조하도록 하는 것이다.

과제 해결 수단

[14] 상기와 같은 목적들을 달성하기 위한 본 발명의 일면은, 복수 개의 클라이언트들로부터 네트워크를 통해 전송되는 오브젝트를 복수 개의 데이터 노드들에 분산 저장하는 분산 저장 시스템(distribution storage system)에 관한 것이다. 본 발명의 일면에 의한 분산 저장 시스템은 오브젝트의 고유 정보 및 오브젝트가 저장된 데이터 노드의 고유 정보를 포함하는 메타데이터(metadata)를 저장하는 메타데이터 서버, 관리 동작을 수행하고자 하는 클라이언트의 동작 요청이 수신되면, 메타데이터 서버를 참조하여 상기 클라이언트가 수행하는 관리 동작의 객체인 대상 오브젝트(target object)와 관련되는 대상 데이터 노드(target data node)에 상기 동작 요청을 전달하는 적어도 두 개의 프락시 서버들, 및 클라이언트를 인증하며, 프락시 서버들 중에서 클라이언트의 동작 요청을 처리할 대상 프락시 서버(target proxy server)를 결정하는 글로벌 로드 밸런서(global load balancer)를 포함하며, 대상 데이터 노드는 대상 프락시 서버로부터 수신되는 동작 요청에 따른 동작을 수행하고 수행 결과를 대상 프락시 서버 또는 클라이언트로 전송하도록 구성된다. 특히, 글로벌 로드 밸런서는 클라이언트 및 프락시 서버들의 위치 정보 및 프락시 서버들의 로드 중 적어도 하나를 고려하여 대상 프락시 서버를 결정하도록 구성된다. 더 나아가, 분산 저장 시스템은 메타데이터를 각각 저장하는 복수 개의 메타데이터 서버들을 포함하며, 메타데이터 서버들은 메타데이터를 서로 동기화하도록 구성되고, 프락시 서버들 각각에는, 메타데이터 서버들 중 어느 하나가 디폴트(default) 메타데이터 서버로서 할당된다. 또한, 프락시 서버 각각은 자신에게 할당된 디폴트 메타데이터 서버가 정상 상태일 경우 디폴트 메타데이터 서버를 참조하여 대상 오브젝트와 관련되는 대상 데이터 노드의 고유 정보의 목록을 클라이언트에게 제공하고, 디폴트 메타데이터 서버가 다운된 경우 다른 메타데이터 서버들 중 하나를 참조하도록 구성된다. 또는, 본 발명의 일면에 의한 분산 저장 시스템은 메타데이터를 각각 저장하는 두 개의 메타데이터 서버들을 포함하며, 메타데이터 서버들은 메타데이터를 서로 동기화하도록 구성되고, 프락시 서버들 각각에는 두 개의 메타데이터 서버들 중 어느 하나가 디폴트 메타데이터 서버로서 할당된다. 뿐만 아니라, 본 발명의 일면에 의한 분산 저장 시스템은 지역(zone)별로 그루핑되며 각각 적어도 하나의

오브젝트들을 저장하는 복수 개의 데이터 노드들을 포함하며, 데이터 노드들은 동일한 오브젝트는 동일한 지역 그룹(zone group)에 속한 데이터 노드들 중 오직 하나에만 저장되도록 구성된다. 특히, 동작이 상기 오브젝트의 저장 동작인 경우, 분산 저장 시스템은 데이터 노드들 및 클라이언트의 위치 관계에 기반하여 대상 오브젝트를 저장할 대상 데이터 노드가 속하는 지역 그룹을 선택하고 선택된 지역 그룹 및 클라이언트 간의 거리에 기반하여 지역 그룹별 우선 순위를 결정하는 위치 인식 서버(location-aware server) 및 동일한 지역 그룹에 포함되는 데이터 노드들의 가용 저장 용량 및 오브젝트 저장 내역을 고려하여 동일한 지역 그룹에 포함되는 데이터 노드들에게 우선 순위를 부여하기 위한 로컬 로드 밸런서(local load balancer)를 포함하며, 대상 프락시 서버는, 가장 높은 우선 순위를 가지는 데이터 노드를 대상 데이터 노드로서 결정하도록 더욱 구성된다.

[15] 상기와 같은 목적들을 달성하기 위한 본 발명의 다른 면은 복수 개의 클라이언트들로부터 네트워크를 통해 전송되는 오브젝트를 복수 개의 데이터 노드들에 분산 저장하는 분산 저장 시스템에서 오브젝트를 관리하기 위한 방법에 관한 것이다. 본 발명의 다른 면에 의한 오브젝트 관리 방법은 글로벌 로드 밸런서가 클라이언트를 인증하는 단계, 인증된 클라이언트로부터 대상 오브젝트에 대한 관리 동작을 수행하고자 하는 동작 요청이 수신되면, 글로벌 로드 밸런서가 복수 개의 프락시 서버들 중 동작 요청을 처리할 대상 프락시 서버를 결정하는 대상 프락시 서버 결정 단계, 대상 프락시 서버가 오브젝트의 고유 정보 및 오브젝트가 저장된 데이터 노드의 고유 정보를 포함하는 메타데이터를 참조하여 대상 오브젝트와 관련되는 대상 데이터 노드를 결정하고, 동작 요청을 대상 데이터 노드로 전달하는 단계, 대상 데이터 노드가 대상 프락시 서버로부터 수신되는 동작 요청에 따른 관리 동작을 수행하고 수행 결과를 대상 프락시 서버 또는 클라이언트로 전송하는 대상 오브젝트 관리 단계, 및 메타데이터가 변경되었을 경우 메타데이터를 갱신하는 메타데이터 갱신 단계를 포함한다. 특히, 대상 프락시 서버 결정 단계는 클라이언트 및 프락시 서버들의 위치 정보 및 프락시 서버들의 로드 중 적어도 하나를 고려하여 대상 프락시 서버를 결정하는 단계를 포함한다. 또한, 메타데이터 갱신 단계는 메타데이터를 각각 저장하는 복수 개의 메타데이터 서버들이 갱신된 정보를 이용하여 메타데이터를 서로 동기화하는 단계를 포함한다. 더 나아가, 메타데이터 갱신 단계는 프락시 서버들 각각에, 메타데이터 서버들 중 어느 하나를 디폴트(default) 메타데이터 서버로서 할당하는 단계 및 프락시 서버 각각이 자신에게 할당된 디폴트 메타데이터 서버가 정상 상태일 경우 디폴트 메타데이터 서버를 참조하여 대상 오브젝트와 관련되는 대상 데이터 노드의 고유 정보의 목록을 클라이언트에게 제공하는 단계, 및 디폴트 메타데이터 서버가 다운된 경우 다른 메타데이터 서버들 중 하나를 참조하여 대상 데이터 노드의 고유 정보의 목록을 클라이언트에게 제공하는 단계를 포함한다. 뿐만 아니라, 관리 동작은 대상 오브젝트를 대상 데이터 노드에 저장하기 위한 저장

동작이며, 대상 오브젝트 관리 단계는 데이터 노드들을 각 데이터 노드들의 위치 정보를 기반하여 지역별로 그루핑하여 지역 그룹을 생성하는 단계, 저장할 대상 오브젝트가 존재할 경우, 대상 오브젝트가 저장될 대상 데이터 노드가 속하는 지역 그룹을 결정하는 지역 그룹 결정 단계, 동일한 지역 그룹에 속하는 데이터 노드들 중 하나의 데이터 노드를 대상 데이터 노드로서 결정하는 대상 데이터 노드 결정 단계, 및 대상 데이터 노드에 대상 오브젝트를 저장하는 저장 단계를 포함한다. 특히, 지역 그룹 결정 단계는 위치 인식 서버가 데이터 노드들 및 클라이언트의 위치 관계에 기반하여 대상 오브젝트를 저장할 대상 데이터 노드가 속하는 지역 그룹을 선택하는 단계 및 위치 인식 서버가 선택된 지역 그룹 및 클라이언트 간의 거리에 기반하여 지역 그룹별 우선 순위를 결정하는 단계를 포함하고, 대상 데이터 노드 결정 단계는 대상 프락시 서버가 위치 인식 서버가 선택한 지역 그룹 당 하나의 대상 데이터 노드를 결정하는 단계, 및 대상 프락시 서버가 대상 데이터 노드들의 목록 및 지역 그룹별 우선 순위를 클라이언트에게 전송하는 단계를 포함하며, 메타데이터 갱신 단계는 대상 프락시 서버가 결정된 대상 데이터 노드들의 목록을 이용하여 메타데이터 데이터베이스를 갱신하는 단계를 포함하고, 저장 단계는 클라이언트가 가장 높은 우선 순위를 가지는 지역 그룹에 속하는 대상 데이터 노드에 대상 오브젝트를 저장하는 단계, 및 우선 순위에 따라 순차적으로 더 낮은 우선 순위를 가지는 지역 그룹에 속하는 대상 데이터 노드들에 대상 오브젝트를 복제하는 복제 단계를 포함한다. 특히, 오브젝트의 고유 정보는 오브젝트의 ID, 크기, 데이터 타입, 및 작성자 중 적어도 하나를 포함하고, 데이터 노드의 고유 정보는 데이터 노드의 ID, IP 주소, 및 물리적 위치 중 적어도 하나를 포함한다.

발명의 효과

[16] 본 발명에 의하여, 클라우드 스토리지 서비스에서 요구되는 확장성과 유연성 그리고 높은 탄력성을 동시에 만족시킬 수 있다.

[17] 또한, 본 발명에 의하면 많은 클라이언트를 지원하기 위해 단일 프락시 서버의 단점인 집중 실패 문제점을 극복하면서도 프락시 서버간 적절한 부하 분산 효과를 기대할 수 있다.

[18] 더 나아가, 본 발명에 의하면 메타 데이터를 이중화하기 때문에 모든 프락시 서버에서 참조할 수 있도록 함으로써 메타 데이터의 안전성과 효율성을 동시에 확보할 수 있다.

도면의 간단한 설명

[19] 도 1은 종래 기술에 의한 분산 저장 시스템을 개념적으로 나타내는 도면이다.

[20] 도 2는 본 발명의 일면에 의한 복수 개의 프락시 서버를 포함하는 분산 저장 시스템의 일 실시예를 개념적으로 나타내는 도면이다.

[21] 도 3은 본 발명의 다른 면에 의한 오브젝트 관리 방법을 개념적으로 나타내는 흐름도이다.

[22] 도 4는 본 발명의 일면에 의한 복수 개의 프락시 서버를 포함하는 분산 저장 시스템의 다른 실시예를 개념적으로 나타내는 도면이다.

발명의 실시를 위한 최선의 형태

[23] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시에 의하여 달성되는 목적을 충분히 이해하기 위해서는 본 발명의 바람직한 실시예를 예시하는 첨부 도면 및 첨부 도면에 기재된 내용을 참조하여야만 한다.

[24] 이하, 첨부한 도면을 참조하여 본 발명의 바람직한 실시예를 설명함으로서, 본 발명을 상세히 설명한다. 그러나, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 설명하는 실시예에 한정되는 것이 아니다. 그리고, 본 발명을 명확하게 설명하기 위하여 설명과 관계없는 부분은 생략되며, 도면의 동일한 참조부호는 동일한 부재임을 나타낸다.

[25] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라, 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "...부", "...기", "모듈", "블록" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[26] 도 2는 본 발명의 일면에 의한 복수 개의 프락시 서버를 포함하는 분산 저장 시스템의 일 실시예를 개념적으로 나타내는 도면이다. 도 2에 도시된 분산 저장 시스템(200)은 네트워크(290)에 연결되는 복수 개의 클라이언트들(210, 212, 216), 데이터 노드들(DN1-DNn), 복수 개의 프락시 서버들(250, 252, 254), 및 글로벌 로드 밸런서(220)를 포함한다. 또한, 분산 저장 시스템(200)은 복수 개의 메타데이터 서버(280, 282)를 포함한다.

[27] 글로벌 로드 밸런서(220)는 멀티 프락시 환경을 지원하기 위해 프락시 서버들(250, 252, 254)의 지역적 위치 및 각 프락시 서버들(250, 252, 254)의 로드를 고려하여 프락시 서버들 중 하나의 프락시 서버를 특정 클라이언트에 대한 대상 프락시 서버로 설정한다. 또한, 글로벌 로드 밸런서(220)는 클라이언트를 인증하는 인증 서버의 역할을 겸할 수 있으며, 더 나아가 프락시 서버의 위치와 프락시 서버별 할당 클라이언트의 수 및 정보를 관리한다.

[28] 클라이언트들(210, 212, 216)로부터 관리 동작을 수행하겠다는 요청이 수신되면, 글로벌 로드 밸런서(220)는 우선 해당 클라이언트를 인증하고, 인증된 클라이언트의 위치와 프락시 서버들(250, 252, 254)들의 로드를 참조하여 위치적으로 클라이언트에 가까우면서도 로드가 낮은, 즉 할당 클라이언트의 수가 적은 프락시 서버를 해당 클라이언트에 대한 대상 프락시 서버로서 결정한다.

[29] 그러면, 요청한 클라이언트는 글로벌 로드 밸런서(220)가 결정한 대상 프락시 서버의 주소와 인증시 받은 키(또는 토큰)를 이용하여 대상 프락시 서버에

접속하여 대상 오브젝트의 업로드, 다운로드, 수정 등의 동작을 수행하도록 요청한다. 동작 수행 요청을 수신한 대상 프락시 서버는 메타데이터 서버(280, 282)들을 참조하여 해당 오브젝트를 관리하는 데이터 노드에 클라이언트의 요청을 전달한다. 그러면, 이러한 요청을 수신한 대상 데이터 노드는 해당 동작을 수행하고 그 결과를 다시 대상 프락시 서버에 전달한다. 그러면, 대상 프락시 서버는 동작 수행 결과를 요청한 클라이언트에게 전달하게 된다. 물론, 대상 데이터 노드는 관리 동작을 수행한 결과를 대상 프락시 서버를 거치지 않고 직접 클라이언트로 전송할 수 있다.

[30] 도 2에 도시된 실시예에서, 메타데이터는 두 개의 메타데이터 서버(280, 282)에 의하여 이중으로 관리된다. 또한, 메타데이터 서버(280, 282)들이 저장하는 메타데이터는 서로 동기화된다. 만일, 각각의 메타데이터 서버(280, 282)들이 참조하는 메타데이터들이 서로 동기화되지 않는다면, 클라이언트에 어떤 대상 프락시 서버가 할당되었는지에 따라서 동작 결과가 상이하게 나타날 수 있다. 예를 들어, 동기화에 실패해 제1 메타데이터 서버(280)는 제1 데이터 노드(DN1)에 대한 정보를 가지고 있으나, 이 정보를 제2 메타데이터 서버(282)는 가지고 있지 못하다고 가정한다. 또한, 제1 프락시 서버(250)와 제2 프락시 서버(252)에는 각각 제1 및 제2 메타데이터 서버(280, 282)가 디폴트 메타데이터 서버로서 할당된다고 가정한다. 이러한 가정은 설명의 편의를 위한 것으로서 본 발명을 한정하는 것이 아님은 명백하다. 클라이언트가 제1 데이터 노드(DN1)에 저장된 대상 오브젝트를 독출하려는 동작 요청을 하고, 해당 클라이언트에 제1 및 제2 프락시 서버들(250, 252)이 각각 할당된다고 가정한다. 이 경우, 클라이언트에 제1 프락시 서버(250)가 대상 프락시 서버로서 할당되면, 제1 프락시 서버(250)는 제1 메타데이터 서버(280)를 참조하기 때문에 대상 오브젝트가 제1 데이터 노드(DN1)에 저장되어 있다는 것을 알 수 있다. 따라서, 제1 프락시 서버(250)는 동작 요청을 제1 데이터 노드(DN1)에 전달하고, 제1 데이터 노드(DN1)는 전달된 동작 요청에 따라 해당 대상 오브젝트를 독출하여 클라이언트에게 전송한다. 그러나, 클라이언트에 제2 프락시 서버(252)가 대상 프락시 서버로서 할당되면, 제2 프락시 서버(252)는 제2 메타데이터 서버(282)를 참조하기 때문에 대상 오브젝트가 제1 데이터 노드(DN1)에 저장되어 있다는 것을 알 수 없다. 따라서, 제2 프락시 서버(252)는 동작 요청에 대해서 실행 불가능하다는 응답을 클라이언트에게 반환하게 될 것이다. 그런데, 클라이언트에게 어떤 프락시 서버가 대상 프락시 서버로서 할당되는지에 따라 관리 동작의 실행 결과가 달라진다면 분산 저장 시스템(200) 전체의 동작 신뢰성이 현저히 떨어진다. 그러므로, 메타데이터 서버(280, 282)들은 서로 동기화되어야 한다. 또는, 메타데이터 서버(280, 282)들은 모두 동일한 메타데이터 데이터베이스(미도시)를 참조할 수 있다. 이 경우, 모든 메타데이터 서버(280, 282)들이 동일한 메타데이터 데이터베이스를 참조하기 때문에 병목 현상이 발생할 수 있으나, 메타데이터 서버들의 동기화가 불필요하게 되는

장점이 있다. 그러므로, 메타데이터 서버들의 동기화가 필요 없이 다중 프락시 서버들을 효과적으로 지원할 수 있다.

[31] 본 명세서에서 '대상 오브젝트(target object)'란 클라이언트가 저장하고자 하는 오브젝트나 데이터 노드로부터 조회하고자 하는 관심 대상인 오브젝트를 의미한다. 또한, '대상 데이터 노드(target data node)'란 여러 개의 데이터 노드 중에서 대상 오브젝트가 저장된 데이터 노드를 의미한다.

[32] 또한, 본 명세서에서 대상 오브젝트의 관리 동작이란, 대상 오브젝트를 대상 데이터 노드에 저장하는 저장 동작, 저장된 대상 오브젝트를 대상 데이터 노드로부터 독출하는 독출 동작, 대상 오브젝트가 저장된 지역 그룹 또는 대상 데이터 노드를 변경하려는 변경 동작, 및 저장된 대상 오브젝트를 수정하려는 동작 등을 모두 포함하는 것으로 이해되어야 한다. 즉, 클라이언트가 오브젝트를 분산 저장하고, 저장된 오브젝트를 독출하여 수정하는 모든 동작이 관리 동작에 포함되는 것으로 이해된다. 또한, 본 명세서에서 분산 저장 시스템이란 클라우드 컴퓨팅을 지원하는 클라우드 스토리지 시스템과 같이, 동일한 오브젝트를 두 개 이상의 저장소에 기록하여 데이터의 손실을 방지하고 저장 부담을 경감시키려는 모든 시스템을 총칭하여 가리키는 용어로 이해되어야 한다.

[33] 대상 프락시 서버 메타데이터 서버(280, 282)들은 모두 동작 상태가 액티브 상태인 것이 바람직하다. 프락시 서버들(250, 252, 254)들은 메타데이터 서버(280, 282)들 중 하나의 메타데이터 서버를 자신의 디폴트 메타 데이터 서버로서 가지고 있으며, 이러한 디폴트 메타데이터 서버에 대한 정보를 최초 설정값으로 알고 있다. 그러므로, 프락시 서버는 자신에게 할당된 디폴트 메타 데이터 서버가 다운되지 않는 한 항상 디폴트 메타 데이터 서버를 이용해 대상 오브젝트의 물리적 위치를 파악한다. 프락시 서버들을 최초 설정할 때 메타데이터 서버(280, 282)들을 번갈아가면서 할당함으로써 메타데이터 서버(280, 282)들 각각이 비슷한 개수의 프락시 서버들에 할당되도록 하는 것이 바람직하다. 그러면, 메타데이터 서버(280, 282)들 간에 부하를 비슷하게 유지할 수 있기 때문에, 어느 하나의 메타데이터 서버에만 많은 부하가 가해지는 현상을 방지할 수 있다.

발명의 실시를 위한 형태

[34] 도 3은 본 발명의 다른 면에 의한 오브젝트 관리 방법을 개념적으로 나타내는 흐름도이다.

[35] 도 3에 도시된 오브젝트 관리 방법에서는, 우선 글로벌 로드 밸런서가 클라이언트를 인증한다(S310). 글로벌 로드 밸런서는 인증을 성공한 클라이언트의 동작 요청은 무시하고 클라이언트 인증 동작을 다시 수행한다(S320). 인증이 성공하면, 글로벌 로드 밸런서는 인증된 클라이언트로부터 대상 오브젝트에 대한 관리 동작을 수행하고자 하는 동작 요청을 수신한다(S330). 동작 요청이 수신되면, 글로벌 로드 밸런서는 복수 개의

프락시 서버들 중 해당 동작 요청을 처리할 대상 프락시 서버를 결정한다(S340). 글로벌 로드 밸런서는 클라이언트 및 프락시 서버들의 위치 정보 및 프락시 서버들의 로드 등을 고려하여 대상 프락시 서버를 선택할 수 있다.

[36] 대상 프락시 서버가 결정되면, 대상 프락시 서버는 클라이언트로부터의 동작 요청을 해당 동작 요청의 객체인 대상 오브젝트가 관련되는 대상 데이터 노드를 결정하고, 동작 요청을 대상 데이터 노드로 전달한다(S350). 본 명세서에서 A라는 오브젝트가 B라는 데이터 노드와 '관련'된다는 의미는 다음과 같이 이해된다. 우선, 관리 동작이 오브젝트 저장 동작일 경우에는, A 오브젝트가 저장될 대상 데이터 노드로서 데이터 노드 B가 결정된다는 것을 의미할 수 있다. 또는, 관리 동작이 오브젝트 독출 동작일 경우에는, A 오브젝트가 이미 데이터 노드 B에 저장되어 있다는 것을 의미할 수 있다.

[37] 이와 같이 동작 요청이 전달되면, 이를 수신한 대상 데이터 노드는 요청된 관리 동작을 수행하고 그 결과를 대상 프락시 서버 또는 클라이언트로 전송한다(S360). 마지막으로, 메타데이터가 변경되었을 경우 메타데이터 서버가 메타데이터를 갱신한다(S370). 전술된 바와 같이, 메타데이터 갱신 단계에서, 메타데이터를 각각 저장하는 복수 개의 메타데이터 서버들이 갱신된 정보를 이용하여 메타데이터를 서로 동기화하는 것이 바람직하다. 또는, 모든 메타데이터 서버들이 하나의 메타데이터 데이터베이스만을 참조할 수도 있다. 다만, 여러 개의 메타데이터 데이터베이스를 두고 이들 메타데이터 데이터베이스들을 서로 동기화하는 경우, 하나의 메타데이터 데이터베이스에 고장이 발생하더라도 다른 메타데이터 데이터베이스를 통하여 원하는 메타데이터를 얻을 수 있기 때문에, 분산 저장 시스템의 동작 신뢰성이 향상된다.

[38] 도 4는 본 발명의 일면에 의한 복수 개의 프락시 서버를 포함하는 분산 저장 시스템의 다른 실시예를 개념적으로 나타내는 도면이다.

[39] 도 4에 도시된 분산 저장 시스템(400)은 네트워크(490)에 연결되는 복수 개의 클라이언트들(410, 412, 416) 및 데이터 노드들(DN11-DN1n, DNm1-DNmn)을 포함한다. 또한, 도 4에 도시된 분산 저장 시스템(400)은 글로벌 로드 밸런서(420), 제1 및 제n 프락시 서버(450, 452), 위치 인식 서버(460), 복제 서버(470) 및 메타데이터 서버(480)를 더 포함한다.

[40] 도 4에 도시된 클라이언트들(410, 412, 416), 글로벌 로드 밸런서(420) 및 메타데이터 서버(480)의 구성 및 동작은 도 2에 도시되는 대응 구성 요소들의 그것과 유사하다. 그러므로, 명세서의 간략화를 위하여 반복적인 설명이 생략된다.

[41] 도 4를 참조하면, 복수 개의 데이터 노드들(DN11-DN1n, DNm1-DNmn)이 각각 다른 지역 그룹(ZG1, ZGm)에 속한다는 것을 알 수 있다. 예를 들어, 제1 지역 그룹(ZG1)에는 데이터 노드들(DN11-DN1n)이 포함되고, 제2 지역 그룹(ZG2)에는 데이터 노드(DNm1-DNmn)가 포함된다. 도 4에 도시된 도시되는

지역 그룹(ZG1, ZGm)들은 효과적인 복제본의 분산 저장을 위하여, 각각 지역적으로 인접한 데이터 노드들을 그루핑함으로써 정의된다. 또한, 동일한 지역 그룹에 속하는 데이터 노드들은 동일한 오브젝트를 저장하지 않도록 구성된다. 즉, 하나의 오브젝트의 복제본은 다른 지역 그룹에 속하는 데이터 노드들에 분산되어 저장되기 때문에 두 개의 복제본이 어느 하나의 지역 그룹에 속한 두 개의 데이터 노드에 공통적으로 저장되지 않는다. 이를 메타데이터의 관점에서 보면, 오브젝트의 물리적인 위치를 나타내는 메타데이터에서 하나의 오브젝트의 복제본들은 다른 지역 그룹에 속한 데이터 노드들에 매핑된다. 그러므로, 어느 특정한 지역 그룹이 전체 네트워크 상에 문제가 생기는 등의 물리적인 피해를 입은 경우라도 복제본이 다른 지역 그룹에 속한 데이터 노드에 분산되어 저장되어 있으므로 신뢰성을 높일 수 있다.

[42] 도 4에 도시된 분산 저장 시스템(400)에서 지역 그룹은 하나의 데이터 센터가 될 수도 있고 좀 더 좁은 지역으로는 하나의 서버 랙이 될 수도 있다. 지역 그룹이 설정되면, 해당 지역 그룹 내에 속한 데이터 노드들을 그 지역 그룹에 속하는 것으로 메타데이터에 등록된다. 그러면, 오브젝트의 복제본들은 다른 지역 그룹에 속한 데이터 노드에 복제된다.

[43] 데이터 노드를 지역 그룹으로 그루핑함으로써 얻어지는 장점은 다음과 같다.

[44] 1) 본 발명에서, 모든 클라이언트들(410, 412, 416) 및 데이터 노드들(DN11-DN1n, DNm1-DNmn)들은 네트워크(490)를 통하여 서로 통신한다. 즉, 클라이언트 각각 및 데이터 각각 간에는 가상 채널(virtual channel)이 존재한다. 그러나, 이러한 가상 채널은 클라이언트 및 데이터 노드의 모든 쌍에 대해 반드시 동일한 조건을 가지는 것은 아니다. 예를 들어, 가상 채널의 통신 환경은 클라이언트 및 데이터 노드 간의 물리적 거리에 따라 달라질 수 있다. 클라이언트 및 데이터 노드 간의 물리적 거리가 멀수록 오브젝트는 더 많은 릴레이 노드 또는 게이트웨이를 통하여 전달되기 때문에 송수신 시간이 오래 걸린다. 또한, 가상 채널의 통신 환경은 네트워크 트래픽의 양 및 가상 채널을 구성하는 네트워크 자원의 성능에 따라서도 달라질 수 있다. 가상 채널을 통해 전달되는 트래픽의 양이 많을수록 가상 채널 상에서 전송 충돌(transmission collision)이 발생될 가능성이 높아지며, 네트워크 자원의 성능이 높을수록 가상 채널의 송수신 속도는 빨라진다. 그러므로, 본 발명에서는 이와 같은 가상 채널의 통신 환경을 고려하여 클라이언트 및 데이터 노드 간에 가장 최적의 가상 채널을 선택한다. 최적의 가상 채널을 선택하기 위하여 본 발명에 의한 분산 저장 시스템에서는 클라이언트 및 지역 그룹 간의 물리적 거리를 참조할 수 있다. 따라서, 저장한 오브젝트를 포함하는 클라이언트로부터 가장 가까운 거리에 위치한 지역 그룹에 속한 데이터 노드에 오브젝트를 저장함으로써 오브젝트의 업로드 시간을 최소화할 수 있다.

[45] 2) 또한, 본 발명에 의한 분산 저장 시스템은 오브젝트를 복제할 때 동일한 지역 그룹에 속하는 데이터 노드로 복제되지 않도록 한다. 따라서, 저장할 대상

오브젝트는 여러 개의 지역 그룹에 분산 저장된다. 일반적으로, 네트워크 장애가 발생하면 인접 지역의 데이터 노드들의 동작도 불가능한 경우가 많다. 예를 들어, 어느 데이터 센터에 여러 개의 데이터 노드들이 존재하고, 이 데이터 센터가 하나의 지역 그룹으로 설정된다고 가정한다. 이러한 가정은 본 발명을 용이하게 설명하기 위한 것으로 본 발명을 한정하는 것이 아님은 명백하다. 갑작스런 정전 등의 사고로 해당 데이터 센터가 동작 불능의 상태에 빠지는 경우가 발생할 수 있다. 이 경우, 본 발명에 의한 분산 저장 시스템은 대상 오브젝트를 해당 데이터 센터의 데이터 노드들 중 오직 하나의 대상 데이터 노드에만 저장하고, 그 복제본은 다른 지역 그룹에 속하는 대상 데이터 노드에 저장한다. 그러므로, 데이터 센터의 모든 데이터 노드들이 장애를 일으키더라도, 다른 지역 그룹에 속하는 대상 데이터 노드로부터 원하는 대상 오브젝트를 용이하게 조회할 수 있다.

[46] 다시 도 4를 참조하면, 분산 저장 시스템(400)에서 데이터 노드를 할당하는 방식으로는 수동 선택 방식과 자동 선택 방식의 두 가지 방식이 제안될 수 있다. 수동 선택 방법은 클라이어트가 최초로 계정을 만들 때 프락시 서버에서 선호하는 지역 그룹을 클라이언트에게 질의하여 클라이언트에게 선택권을 주는 방법이다. 복제본의 개수에 따라 지역 그룹의 개수가 정해지기 때문에, 클라이언트는 허용되는 가장 많은 복제본 개수만큼 선호하는 지역 그룹을 우선 순위에 따라 선택한다. 즉, 클라이언트는 오브젝트에 따라 복제본 개수를 다르게 선택할 수 있는 경우를 고려해 허용되는 최대 복제본 개수만큼 선호 지역 그룹을 선택한다. 이 후 선호하는 지역 그룹을 바꾸고자 할 경우에는 프락시 서버에게 지역 그룹 변경을 요청하고, 이 요청이 받아들여지면 그 이 후에는 대상 오브젝트를 변경된 선호 지역 그룹에 속하는 대상 데이터 노드에 저장한다. 본 명세서에서 '우선 순위'란 특정 대상 오브젝트를 저장하기에 어떤 지역 그룹 또는 데이터 노드가 다른 지역 그룹 또는 데이터 노드에 비하여 더 적합한지 판단하여 매긴 순위를 나타낸다. 우선 순위에는 특정 지역 그룹이 다른 지역 그룹에 비하여 가지는 우선 순위 및 동일한 지역 그룹 내에 속한 데이터 노드들 간의 우선 순위가 포함될 수 있다. 또한, 우선 순위는 클라이언트가 직접 어느 대상 오브젝트에 관련한 특정 지역 및 데이터 노드에 대한 선호도에 기반하여 매길 수도 있고, 또는 프락시 서버 또는 위치 인식 서버에 의하여 자동으로 결정될 수도 있다.

[47] 자동 선택 방법은 위치 인식 서버(location-aware server)(460)를 이용하여 지역 그룹을 선택하는 방식이다.

[48] 인증된 클라이언트가 대상 오브젝트를 저장할 대상 데이터 노드에 대해서 대상 프락시 서버(예를 들어, 450)에 질의하면, 대상 프락시 서버(450)는 가장 유리한 지역 그룹에 대해서 위치 인식 서버(460)에 문의한다.

[49] 위치 인식 서버(460)는 여러 가지 방법으로 클라이언트의 위치를 파악할 수 있는데, 일반적으로는 클라이언트의 IP 주소로 클라이언트의 물리적 위치를

파악할 수 있다. 위치 인식 서버(460)는 대상 프락시 서버(450)의 요청에 따라 클라이언트의 기본 복제본 개수만큼의 지역 그룹을 선정한 후 선정된 지역 그룹 리스트를 대상 프락시 서버(450)에게 전송한다. 위치 인식 서버(460)는 물리적으로 대상 프락시 서버(450) 내에 통합되어 구현될 수 있다.

[50] 위치 인식 서버(460)에 의하여 결정된 지역 그룹 각각에 속하는 대상 데이터 노드를 결정하는 동작은 대상 프락시 서버(450) 또는 위치 인식 서버(460)에 의하여 수행될 수 있다. 위치 인식 서버(460)가 대상 데이터 노드까지 결정하는 경우, 위치 인식 서버(460)는 메타데이터 데이터베이스(480)를 참조하여 선택된 지역 그룹 내에서 대상 오브젝트를 가지고 있는 클라이언트와 가장 근접한 대상 데이터 노드를 선택할 수 있다. 반면에, 대상 프락시 서버(450)가 대상 데이터 노드를 선택한다면, 대상 프락시 서버(450)는 로컬 로드 밸런서(load balancer, 455)를 이용하여 각 지역 그룹에 속하는 데이터 노드들의 상태를 점검하고, 이 중에서 최적 조건을 가지는 데이터 노드를 대상 데이터 노드로서 선택할 수 있다. 로컬 로드 밸런서(455)는 대상 프락시 서버(450)에 포함되는 것으로 도시되었으나, 이는 본 발명을 한정하는 것이 아님이 이해되어야 한다.

[51] 또한, 대상 프락시 서버(450)는 각 지역 그룹 내 데이터 노드들의 정보를 메타데이터에 관리하고, 데이터 노드들의 로드 밸런싱을 위해 각 데이터 노드의 스토리지 용량을 고려하여 각 데이터 노드의 가중치 값을 사전에 결정한다. 현재까지 각 데이터 노드의 오브젝트 저장 내역과 데이터 노드의 가중치 값을 고려해 요청 클라이언트의 데이터 노드를 선정함으로써 지역 그룹 내 데이터 노드간 로드 밸런싱을 유지한다.

[52] 도 4에 도시된 분산 저장 시스템(400)은 복수 개의 프락시 서버들을 포함함으로써 단일 프락시 서버를 이용하는 경우에 비하여 분산 저장 시스템의 안정성을 도모할 수 있으며, 프락시 서버의 확장이 자유롭기 때문에 분산 저장 시스템을 확장하는 것도 용이하다. 또한, 대상 오브젝트를 저장할 때 클라이언트와 데이터 노드들 간의 거리를 참조하여 대상 데이터 노드를 결정하기 때문에, 오브젝트의 업로드 및 다운로드 시간을 절감할 수 있다. 또한, 대상 오브젝트를 다른 지역 그룹에 속하는 대상 데이터 노드에 저장하기 때문에, 어느 한 지역 그룹의 네트워크 환경이 악화되더라도, 손쉽게 다른 지역 그룹에 속한 대상 데이터 노드로부터 대상 오브젝트를 독출할 수 있으므로, 동작 신뢰성이 현저히 향상된다.

[53] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다.

[54] 또한, 본 발명에 따르는 방법은 컴퓨터로 읽을 수 있는 기록 매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록 매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함할 수 있다. 컴퓨터가 읽을 수 있는 기록 매체의 예로는 ROM,

RAM, CD-ROM, 자기 테이프, 플로피 디스크, 광 데이터 저장 장치 등이 있으며, 또한 캐리어 웨이브(예를 들어 인터넷을 통한 전송)의 형태로 구현되는 것도 포함한다. 또한 컴퓨터가 읽을 수 있는 기록 매체는 네트워크로 연결된 분산 컴퓨터 시스템에 의하여 분산 방식으로 실행될 수 있는 컴퓨터가 읽을 수 있는 코드를 저장할 수 있다.

[55] 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

산업상 이용가능성

[56] 본 발명은 복수 개의 프락시 서버를 이용하여 분산 저장 시스템의 동작 신뢰도 및 확장성을 향상시키기 위한 기술에 적용될 수 있다.

청구범위

[청구항 1] 복수 개의 클라이언트들로부터 네트워크를 통해 전송되는 오브젝트를 복수 개의 데이터 노드들에 분산 저장하는 분산 저장 시스템(distribution storage system)에 있어서,
상기 오브젝트의 고유 정보 및 상기 오브젝트가 저장된 데이터 노드의 고유 정보를 포함하는 메타데이터(metadata)를 저장하는 메타데이터 서버;
관리 동작을 수행하고자 하는 클라이언트의 동작 요청이 수신되면, 상기 메타데이터 서버를 참조하여 상기 클라이언트가 수행하는 관리 동작의 객체인 대상 오브젝트(target object)와 관련되는 대상 데이터 노드(target data node)에 상기 동작 요청을 전달하는 적어도 두 개의 프락시 서버들; 및
상기 클라이언트를 인증하며, 상기 프락시 서버들 중에서 상기 클라이언트의 동작 요청을 처리할 대상 프락시 서버(target proxy server)를 결정하는 글로벌 로드 밸런서(global load balancer)를 포함하며,
상기 대상 데이터 노드는 상기 대상 프락시 서버로부터 수신되는 상기 동작 요청에 따른 관리 동작을 수행하고 수행 결과를 상기 대상 프락시 서버 또는 상기 클라이언트로 전송하도록 구성되는 것을 특징으로 하는 복수 개의 프락시 서버들을 가지는 분산 저장 시스템.

[청구항 2] 제1항에 있어서, 상기 글로벌 로드 밸런서는,
상기 클라이언트 및 상기 프락시 서버들의 위치 정보 및 상기 프락시 서버들의 로드 중 적어도 하나를 고려하여 상기 대상 프락시 서버를 결정하도록 구성되는 것을 특징으로 하는 복수 개의 프락시 서버들을 가지는 분산 저장 시스템.

[청구항 3] 제2항에 있어서, 상기 분산 저장 시스템은,
상기 메타데이터를 각각 저장하는 복수 개의 메타데이터 서버들을 포함하며,
상기 메타데이터 서버들은 상기 메타데이터를 서로 동기화하도록 구성되고,
상기 프락시 서버들 각각에는, 상기 메타데이터 서버들 중 어느 하나가 디폴트(default) 메타데이터 서버로서 할당되는 것을 특징으로 하는 복수 개의 프락시 서버들을 가지는 분산 저장 시스템.

[청구항 4] 제3항에 있어서, 상기 프락시 서버 각각은,
자신에게 할당된 디폴트 메타데이터 서버가 정상 상태일 경우

상기 디폴트 메타데이터 서버를 참조하여 상기 대상 오브젝트와 관련되는 대상 데이터 노드의 고유 정보의 목록을 상기 클라이언트에게 제공하고,
상기 디폴트 메타데이터 서버가 다운된 경우 다른 메타데이터 서버들 중 하나를 참조하도록 구성되는 것을 특징으로 하는 복수 개의 프락시 서버들을 가지는 분산 저장 시스템.

[청구항 5] 제4항에 있어서, 상기 분산 저장 시스템은,
상기 메타데이터를 각각 저장하는 두 개의 메타데이터 서버들을 포함하며,
상기 메타데이터 서버들은, 상기 메타데이터를 서로 동기화하도록 구성되고,
상기 프락시 서버들 각각에는, 두 개의 메타데이터 서버들 중 어느 하나가 디폴트 메타데이터 서버로서 할당되는 것을 특징으로 하는 복수 개의 프락시 서버들을 가지는 분산 저장 시스템.

[청구항 6] 제1항에 있어서, 상기 분산 저장 시스템은,
지역(zone)별로 그루핑되며 각각 적어도 하나의 오브젝트들을 저장하는 복수 개의 데이터 노드들을 포함하며,
상기 데이터 노드들은, 동일한 오브젝트는 동일한 지역 그룹(zone group)에 속한 데이터 노드들 중 오직 하나에만 저장되도록 구성되는 것을 특징으로 하는 복수 개의 프락시 서버들을 가지는 분산 저장 시스템.

[청구항 7] 제6항에 있어서, 상기 분산 저장 시스템은,
상기 관리 동작이 상기 오브젝트의 저장 동작인 경우, 상기 데이터 노드들 및 상기 클라이언트의 위치 관계에 기반하여 상기 대상 오브젝트를 저장할 대상 데이터 노드가 속하는 지역 그룹을 선택하고 선택된 지역 그룹 및 상기 클라이언트 간의 거리에 기반하여 상기 지역 그룹별 우선 순위를 결정하는 위치 인식 서버(location-aware server) 및
동일한 지역 그룹에 포함되는 데이터 노드들의 가용 저장 용량 및 오브젝트 저장 내역을 고려하여 동일한 지역 그룹에 포함되는 데이터 노드들에게 우선 순위를 부여하기 위한 로컬 로드 밸런서(local load balancer)를 포함하며,
상기 대상 프락시 서버는, 가장 높은 우선 순위를 가지는 데이터 노드를 상기 대상 데이터 노드로서 결정하도록 더욱 구성되는 것을 특징으로 하는 분산 저장 시스템.

[청구항 8] 제1항에 있어서,
상기 오브젝트의 고유 정보는 상기 오브젝트의 ID, 크기, 데이터 타입, 및 작성자 중 적어도 하나를 포함하고,

상기 데이터 노드의 고유 정보는 상기 데이터 노드의 ID, IP(Internet Protocol) 주소, 및 물리적 위치 중 적어도 하나를 포함하는 것을 특징으로 하는 분산 저장 시스템.

[청구항 9] 복수 개의 클라이언트들로부터 네트워크를 통해 전송되는 오브젝트를 복수 개의 데이터 노드들에 분산 저장하는 분산 저장 시스템에서 오브젝트를 관리하기 위한 방법에 있어서,
글로벌 로드 밸런서가 상기 클라이언트를 인증하는 단계;
인증된 클라이언트로부터 대상 오브젝트에 대한 관리 동작을 수행하고자 하는 동작 요청이 수신되면, 상기 글로벌 로드 밸런서가 복수 개의 프락시 서버들 중 상기 동작 요청을 처리할 대상 프락시 서버를 결정하는 대상 프락시 서버 결정 단계;
상기 대상 프락시 서버가 상기 오브젝트의 고유 정보 및 상기 오브젝트가 저장된 데이터 노드의 고유 정보를 포함하는 메타데이터를 참조하여 상기 대상 오브젝트와 관련되는 대상 데이터 노드를 결정하고, 상기 동작 요청을 상기 대상 데이터 노드로 전달하는 단계;
상기 대상 데이터 노드가 상기 대상 프락시 서버로부터 수신되는 상기 동작 요청에 따른 관리 동작을 수행하고 수행 결과를 상기 대상 프락시 서버 또는 상기 클라이언트로 전송하는 대상 오브젝트 관리 단계; 및
상기 메타데이터가 변경되었을 경우 상기 메타데이터를 갱신하는 메타데이터 갱신 단계를 포함하는 것을 특징으로 하는 복수 개의 프락시 서버들을 가지는 분산 저장 시스템에서 대상 오브젝트를 관리하기 위한 방법.

[청구항 10] 제9항에 있어서, 상기 대상 프락시 서버 결정 단계는,
상기 클라이언트 및 상기 프락시 서버들의 위치 정보 및 상기 프락시 서버들의 로드 중 적어도 하나를 고려하여 상기 대상 프락시 서버를 결정하는 단계를 포함하는 것을 특징으로 하는 복수 개의 프락시 서버들을 가지는 분산 저장 시스템에서 대상 오브젝트를 관리하기 위한 방법.

[청구항 11] 제10항에 있어서, 상기 메타데이터 갱신 단계는,
상기 메타데이터를 각각 저장하는 복수 개의 메타데이터 서버들이 갱신된 정보를 이용하여 상기 메타데이터를 서로 동기화하는 단계를 포함하는 것을 특징으로 하는 복수 개의 프락시 서버들을 가지는 분산 저장 시스템에서 대상 오브젝트를 관리하기 위한 방법.

[청구항 12] 제11항에 있어서, 상기 메타데이터 갱신 단계는,
상기 프락시 서버들 각각에, 상기 메타데이터 서버들 중 어느

하나를 디폴트(default) 메타데이터 서버로서 할당하는 단계 및
상기 프락시 서버 각각이 자신에게 할당된 디폴트 메타데이터 서버가 정상 상태일 경우 상기 디폴트 메타데이터 서버를 참조하여 상기 대상 오브젝트와 관련되는 대상 데이터 노드의 고유 정보의 목록을 상기 클라이언트에게 제공하는 단계; 및
상기 디폴트 메타데이터 서버가 다운된 경우 다른 메타데이터 서버들 중 하나를 참조하여 상기 대상 데이터 노드의 고유 정보의 목록을 상기 클라이언트에게 제공하는 단계를 포함하는 것을 특징으로 하는 복수 개의 프락시 서버들을 가지는 분산 저장 시스템에서 대상 오브젝트를 관리하기 위한 방법.

[청구항 13] 제9항에 있어서,
상기 관리 동작은 상기 대상 오브젝트를 상기 대상 데이터 노드에 저장하기 위한 저장 동작이며,
상기 대상 오브젝트 관리 단계는,
상기 데이터 노드들을 각 데이터 노드들의 위치 정보를 기반하여 지역별로 그루핑하여 지역 그룹을 생성하는 단계;
저장할 대상 오브젝트가 존재할 경우, 상기 대상 오브젝트가 저장될 대상 데이터 노드가 속하는 지역 그룹을 결정하는 지역 그룹 결정 단계;
동일한 지역 그룹에 속하는 데이터 노드들 중 하나의 데이터 노드를 상기 대상 데이터 노드로서 결정하는 대상 데이터 노드 결정 단계; 및
상기 대상 데이터 노드에 상기 대상 오브젝트를 저장하는 저장 단계를 포함하는 것을 특징으로 하는 분산 저장 시스템에서 대상 오브젝트를 관리하기 위한 방법.

[청구항 14] 제13항에 있어서,
상기 지역 그룹 결정 단계는,
위치 인식 서버가 상기 데이터 노드들 및 상기 클라이언트의 위치 관계에 기반하여 상기 대상 오브젝트를 저장할 대상 데이터 노드가 속하는 지역 그룹을 선택하는 단계 및
위치 인식 서버가 선택된 지역 그룹 및 상기 클라이언트 간의 거리에 기반하여 상기 지역 그룹별 우선 순위를 결정하는 단계를 포함하고,
상기 대상 데이터 노드 결정 단계는,
상기 대상 프락시 서버가 상기 위치 인식 서버가 선택한 지역 그룹 당 하나의 대상 데이터 노드를 결정하는 단계; 및
상기 대상 프락시 서버가 상기 대상 데이터 노드들의 목록 및 상기 지역 그룹별 우선 순위를 상기 클라이언트에게 전송하는 단계를

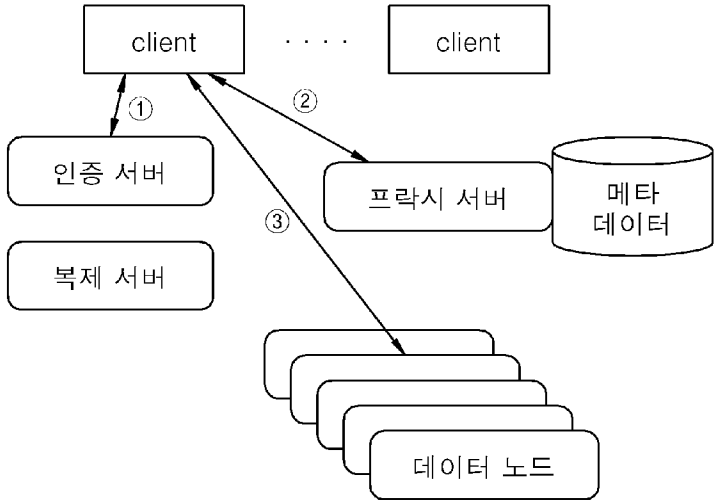
포함하며,
상기 메타데이터 갱신 단계는, 상기 대상 프락시 서버가 결정된 대상 데이터 노드들의 목록을 이용하여 상기 메타데이터 데이터베이스를 갱신하는 단계를 포함하고,
상기 저장 단계는,
상기 클라이언트가 가장 높은 우선 순위를 가지는 지역 그룹에 속하는 대상 데이터 노드에 상기 대상 오브젝트를 저장하는 단계; 및
상기 우선 순위에 따라 순차적으로 더 낮은 우선 순위를 가지는 지역 그룹에 속하는 대상 데이터 노드들에 상기 대상 오브젝트를 복제하는 복제 단계를 포함하는 것을 특징으로 하는 분산 저장 시스템에서 대상 오브젝트를 관리하기 위한 방법.

[청구항 15] 제9항에 있어서,
상기 오브젝트의 고유 정보는 상기 오브젝트의 ID, 크기, 데이터 타입, 및 작성자 중 적어도 하나를 포함하고,
상기 데이터 노드의 고유 정보는 상기 데이터 노드의 ID, IP 주소, 및 물리적 위치 중 적어도 하나를 포함하는 것을 특징으로 하는 분산 저장 시스템에서 대상 오브젝트를 관리하기 위한 방법.

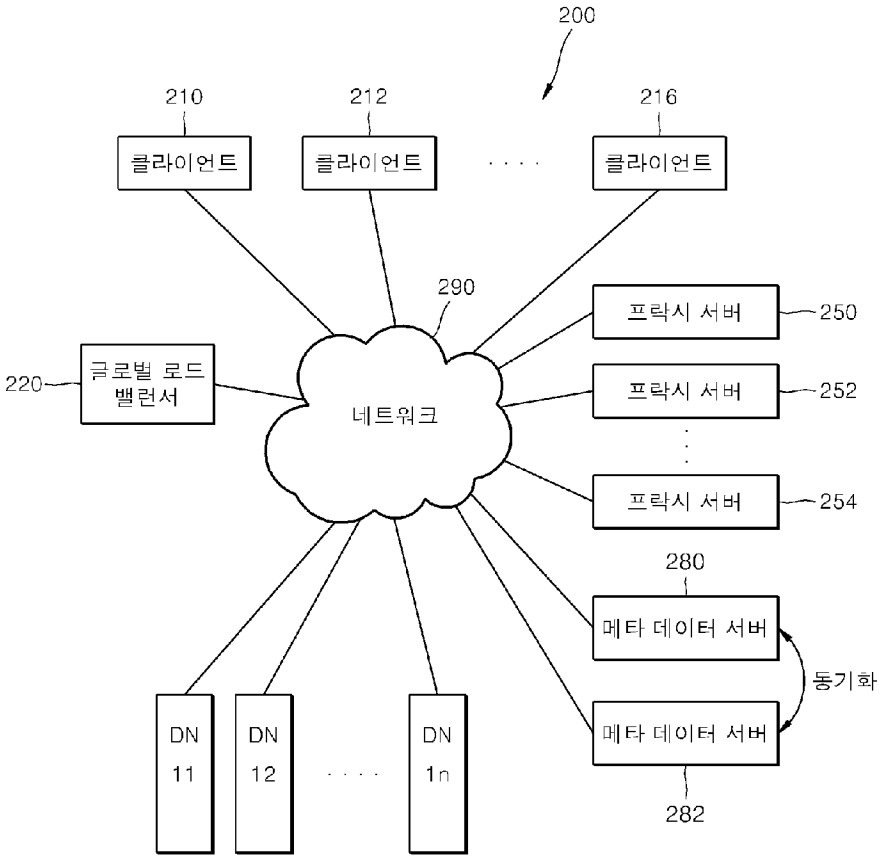
[청구항 16] 제9항 내지 제15항 중 어느 한 항에 따르는 방법을 구현하기 위하여 컴퓨터에 의하여 실행될 수 있는 컴퓨터 프로그램 명령어들을 저장하는 컴퓨터에 의하여 독출가능한 저장 매체.

WO 2012/086919 PCT/KR2011/008223

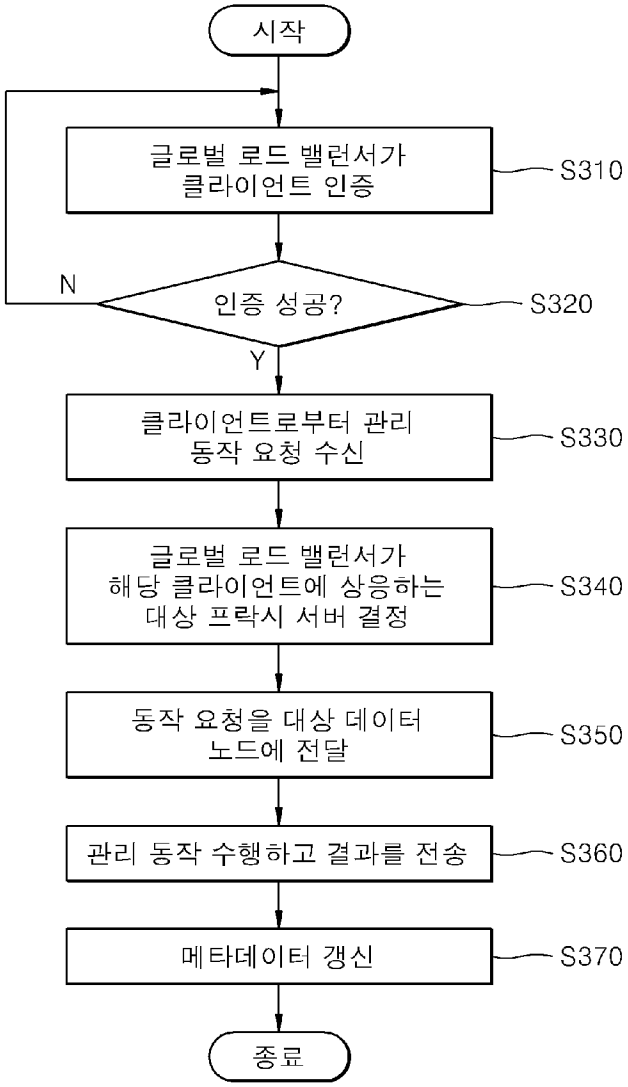
[Fig. 1]



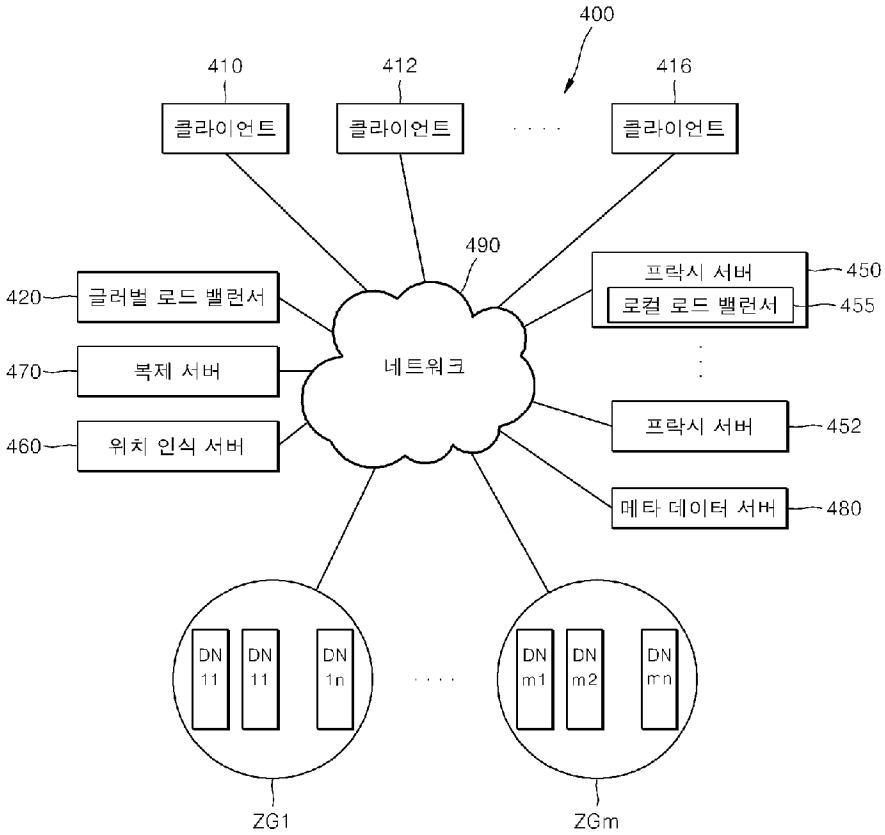
[Fig. 2]



[Fig. 3]


시작
글로벌 로드 밸런서가 클라이언트 인증
S310
인증 성공?
S320
N
Y
클라이언트로부터 관리 동작 요청 수신
S330
글로벌 로드 밸런서가 해당 클라이언트에 상응하는 대상 프락시 서버 결정
S340
동작 요청을 대상 데이터 노드에 전달
S350
관리 동작 수행하고 결과를 전송
S360
메타데이터 갱신
S370
종료

[Fig. 4]


400
410
412
416
클라이언트
클라이언트
클라이언트
490
420
글러벌 로드 밸런서
470
복제 서버
460
위치 인식 서버
네트워크
프락시 서버
450
로컬 로드 밸런서
455
프락시 서버
452
메타 데이터 서버
480
DN 11
DN 11
DN 1n
DN m1
DN m2
DN mn
ZG1
ZGm