データセンタ間における伸縮性を持つキーバリューストレージ実現手法

堀 江 光[†1]　浅 原 理 人[†2]
山 田 浩 史[†1]　河 野 健 二[†1]

1. はじめに

近年多くのウェブサービスがクラウド環境へ移行している．クラウド環境では，計算資源のみでなく記憶容量等の資源も需要に応じて提供する．このような環境では，効果的に物理資源を利用するために，多数のノードの資源を動的に管理する必要がある．

現在，主なクラウドサービスにおいて実データを保持する各データノードは Peer-to-Peer (P2P) の仕組みを用いて管理されている．P2P による管理は，各データノードが自律的に動作する Pure-P2P 型，各データノードを管理するための特別なノードを用いる Hybrid-P2P 型に大別される．Pure-P2P 型の代表的なストレージには Amazon S3[1] や Cassandra[2] 等があり，負荷や機能の集中が起こりづらいためスケールアウトが容易であるという点で，Hybrid-P2P 型に対して優位性がある．今後，データセンタで取り扱うデータ量が増加するにつれ，さらに Pure-P2P 型が適した状況となる．

一方で，現在，記憶資源の伸縮性は単一のデータセンタ内に留まっており，サービス稼働中に全部または一部のデータを他のデータセンタに移送することは考慮されていない．しかし，単一のデータセンタにおいて利用可能なストレージは容量・スループットともに限られている．将来的な必要量の増加や短期的な負荷増に対応するために，データセンタ間を跨いだ伸縮性を持つストレージは必要である．

そこで本研究では，データセンタ間を跨いだ伸縮性を持つ Pure-P2P 型のキーバリューストレージ (KVS) の実現手法を提案する．本手法では，各データノードをデータセンタ間を跨ぐひとつのオーバレイネットワーク上で管理し，このオーバレイネットワークへデータノードが leave/join することによってデータセンタ毎に利用する資源の割合を任意に変更することが可能である．

†1 慶應義塾大学
†2 NEC サービスプラットフォーム研究所

2. 提案手法

2.1 概　　要

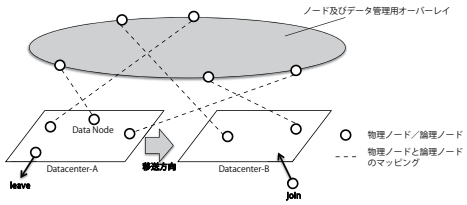

図 1　提案手法によるデータ移送の方法

本研究では，データセンタ間を跨いだ伸縮性を持った Pure-P2P 型のキーバリューストレージ (KVS) の実現手法を提案する．本手法を用いると，あるデータセンタから別のデータセンタへ，サービスを停止することなくデータの全部又は一部を移送することが可能となる．図 1 に本手法によるデータ移送の概要を示す．本手法では，各データノード及びデータをオーバレイネットワークで管理する．データ移送の際には，移送先のデータセンタに存在する物理ノードを新たに join した後に移送元の任意のノードを leave する．保存する各データは一定以上の冗長度で複製しておき，また，各ノード間で定期的なデータのリバランシングを行うことで，leave/join するノード同士が直接データ転送すること無く，各データセンタの利用する資源割合を任意に変更することが可能である．

2.2 探索コストの軽減

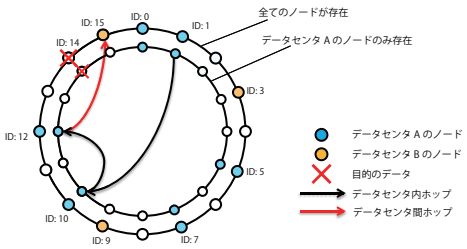

図 2　提案手法によるデータ探索

本手法では，探索の際に同一のデータセンタに含まれるノードを優先的に探索する仕組みを導入すること

で，最高でも 1 回のデータセンタ間ホップにより目的のノードに到達できるようにした．このようにしたのは，既存のルーティングアルゴリズムをそのまま用いると，オーバレイネットワーク上における各ノードが物理ノードの配置を反映していないことにより，探索の際に不必要なデータセンタ間を跨ぐ通信を行ってしまい著しい性能低下が発生するためである．

図 2 に本手法によるルーティングの例を示す．各ノードはデータセンタ内のノードのみを対象とした経路表と，全てのノードを対象とした経路表を保持する．探索の際には前者を優先的に用いて目的の ID へと接近し，適切なノードが存在しない場合に後者を用いる．このようにすることで，最高でも 1 回のデータセンタ間ホップで到達が可能となる．

3. 予 備 実 験

提案手法が，複数のデータセンタ間を跨いで構築したストレージにおいて効率的なルーティングを実現することを確認するために，予備実験を行った．予備実験には Chord[3] を用い，通常のものと同一データセンタ内を優先探索するものを比較した．実装には OverlayWeaver[4] を用い，2 つのデータセンタにそれぞれ 500 ノード存在する状況下でランダム生成したキーを探索する実験を行った．データセンタ内を優先探索する場合は，不必要なデータセンタ間ホップを行わず効率的に目的のノードに到達可能であることがわかった．

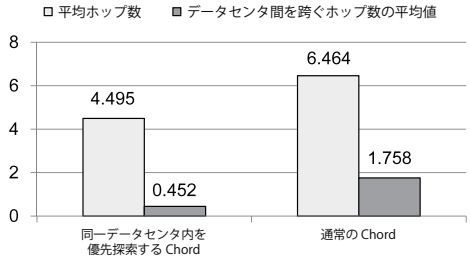

図 3 ノード探索に要したホップ数の比較

4. 関 連 研 究

Walter[5], COPS[6] は，データセンタ間レプリケーションを行った際も性能を大きく低下せずにサービス提供が可能なストレージを実現する手法である．これらはレプリカとの同期を必要最低限とすることで一貫性と可用性を確保しつつ性能の向上を目指しており，データ量や負荷に応じてデータノードを増やすといったことは対象としておらず，本研究の目的とは異なる．また，Zephyr[7] はデータベースとしてのサービスを停止することなくデータノード間でデータの移送を実現する手法である．これはデータセンタ内での移送を対象としておりデータセンタを跨いだ移送については考慮しておらず，本研究の目的とは異なる．

5. まとめと今後の予定

データセンタ間における伸縮性を持つ Pure-P2P 型の KVS を実現する手法を提案した．本手法では，各データノードをオーバレイネットワーク上で管理し，leave/join を行うことでデータセンタ毎に利用する資源の割合を任意に変更することが可能である．この際，データセンタ内を優先的にルーティングを行うことで，大きなオーバヘッドを伴うデータセンタ間ホップを必要最低限に抑える．

今後は提案手法上でウェブアプリケーションを稼働し，ノードの移送に伴う性能低下や目的のデータが同一データセンタ内に存在しない場合の影響等について評価する予定である．

参 考 文 献

1) Amazon Web Services LLC: Amazon Simple Storage Service Cloud. http://aws.amazon.com/s3/.
2) Lakshman, A. and Malik, P.: Cassandra: A Decentralized Structured Storage System, *Proc. of 3rd ACM SIGOPS Int'l Workshop on Large Scale Distributed Systems and Middleware* (2009).
3) Stoica, I., Morris, R., Karger, D., Kaashoek, M.F. and Balakrishnan, H.: Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications, *Proc. of ACM Special Interest Group on Data Communications Conference* (2001).
4) Kazuyuki, S., Yoshio, T. and Satoshi, S.: Overlay Weaver: An overlay construction toolkit, *Computer Communications*, Vol. 31, No.2, pp.402–412 (online), DOI:http://dx.doi.org/10.1016/j.comcom.2007.08.002 (2008).
5) Sovran, Y., Power, R., Aguilera, M.K. and Li, J.: Transactional storage for geo-replicated systems, *Proc. of 23th ACM Symposium on Operating System Principles* (2011).
6) Lloyd, W., Freedman, M.J., Kaminskyy, M. and Andersen, D.G.: Don ŧ Settle for Eventual: Scalable Causal Consistency for Wide-Area Storage with COPS, *Proc. of 23th ACM Symposium on Operating System Principles* (2011).
7) Elmore, A.J., Das, S., Agrawal, D. and Abbadi, A.E.: Zephyr: Live Migration in Shared Nothing Databases for Elastic Cloud Platforms, *Proc. of ACM's Special Interest Group on Management Of Data* (2011).