ストレージ仮想化技術の取り組み

NTTソフトウェアイノベーションセンタでは，ストレージ仮想化に関する研究開発に取り組んでいます．本稿では，任意のファイルシステムから利用可能な分散ブロックストレージSheepdogと，運用性が高く堅牢な分散オブジェクトストレージOpenStack Swiftに関する取り組みについて紹介します．

福本 佳史（ふくもと よしふみ） /内藤 一兵衛（ないとう いちべえ）
日高 東潮（ひたか としお） /白石 正裕（しらいし まさひろ）
NTTソフトウェアイノベーションセンタ

ストレージ仮想化とは

仮想化はクラウド基盤における計算機資源の柔軟な管理やコスト削減を実現する技術として，広く普及しつつあります．中でも複数のストレージ装置をまたいで単一のストレージ装置があるかのように見せたり，単一のストレージに複数の領域があるように見せることのできる「ストレージ仮想化技術」は，仮想マシンイメージの柔軟な管理，アプリケーションデータの保存・共有などのために重要な技術となっています．

本稿では，NTTソフトウェアイノベーションセンタで開発に取り組んでいる，オープンソースのストレージ仮想化技術SheepdogおよびOpenStack Swiftについて紹介します．分散ブロックストレージSheepdogは，PCやサーバに内蔵されるハードディスクと同じようにファイルシステムを通して利用できるストレージで，分散オブジェクトストレージOpenStack Swiftは大量のデータを格納し，アプリケーション間で共有するためREST APIを使ってファイルを読み書きするストレージです．

分散ブロックストレージ Sheepdog

私たちが普段利用しているPCでは，ファイルシステムを介してハードディスクにファイルを読み書きしています．ハードディスクはブロックデバイスの一種として扱われ，それ自体はデータを一定サイズのブロック単位で読み書きする単純な機能しか持ちません．そのためファイルシステムがブロックデバイス上のどの位置にファイルデータがあるか管理することで，ファイルの読み書きを実現しているのです．ファイルシステムはさまざまな種類がありますが，ブロックデバイス上で動作することは共通しています．ブロックストレージはブロックデバイスを提供できるストレージです．つまりブロックストレージは，データの保存・読み書きの基本的な役割を果たす，もっとも汎用的なストレージであるといえます．

仮想マシンの動作には仮想化されたブロックストレージが不可欠で，特にサーバ仮想化環境の構築には，任意のサイズの仮想ブロックデバイス（仮想ディスク）をネットワーク経由で提供できる共有ストレージ製品を導入することが主流です．共有ストレージは，シンプロビジョニング・スナップショット・ライブマイグレーションなどの仮想化機能によって，サーバ仮想化環境の運用性・信頼性を高めることができます．

Sheepdogは，共有ストレージ製品と同様に利用できるブロックストレージを，**図1**のように複数の汎用サーバをクラスタ化することで構築できるオープンソースソフトウェアです．クラスタに属するPCサーバの内蔵ディスクを1つのストレージプールに束ね，仮想ディスクを切り出して提供することができます．仮想化基盤ソフトウェアであるOpenStackやQEMU/KVMに採用されており，一般的なストレージインタフェースであるiSCSIにも対応します．

一般的な共有ストレージ製品は，スケーラビリティの課題（容量・性能の拡張には事前設計が必要で，縮退は原則不可，ベンダロックインも発生）と，信頼性の課題（ハードウェア故障によりサービス停止・一部データへのアクセス不可）があります．それらの課題に対し，Sheepdogはクラスタ内のサーバがすべて同じ役割を果たすいわゆる「完全対称型」のシステムとなってい

るため，その恩恵として，①クラスタへのサーバ増減設が容易で，要求される規模に対して共有ストレージよりも柔軟に容量スケール・負荷分散が可能，②単一障害点がなく一部サーバの故障が発生してもサービス停止・データ損失を回避可能で信頼性が高い，③サーバ追加・離脱に伴うデータリバランシング・冗長性回復処理などが自動化され，手動オペレーション低減によって管理者の負担が小さい，という特長を持っています．

Sheepdogが提供する仮想ディスクは**図2** ⓐのように一定サイズ（初期値は4 MB）のオブジェクトに分割・多重化され，クラスタ内の各サーバに分散配置されます．図2 ⓑはオブジェクトの配置先決定に用いられるコンシステントハッシュ法を示しています．Sheepdogでは，各サーバ（物理ノード）に対して仮想ノードというデータ構造が生成され，それぞれが分散するようにハッシュ値を用いてリング状に並べられます．仮想ディスクに対して書込みが行われると，配置先となる3つの物理ノードに対しオブジェクトの生成・更新が行われます．オブジェクトのIDから決定された仮想ノードを基準にリングに沿って2，3番目が選ばれ，それらの仮想ノードに対応する物理ノードが配置先となります．このように，Sheepdogはコンシステントハッシュ法によって数学的にデータ配置先を決定できるため，集中管理サーバを排して自律性を高めることを実現しており，特長①～③に寄与しています．

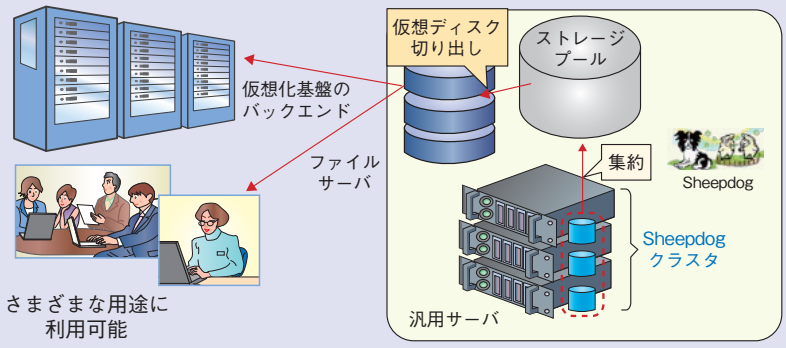
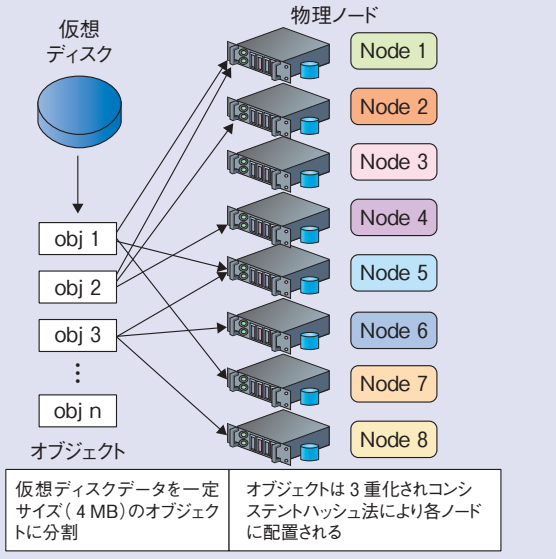


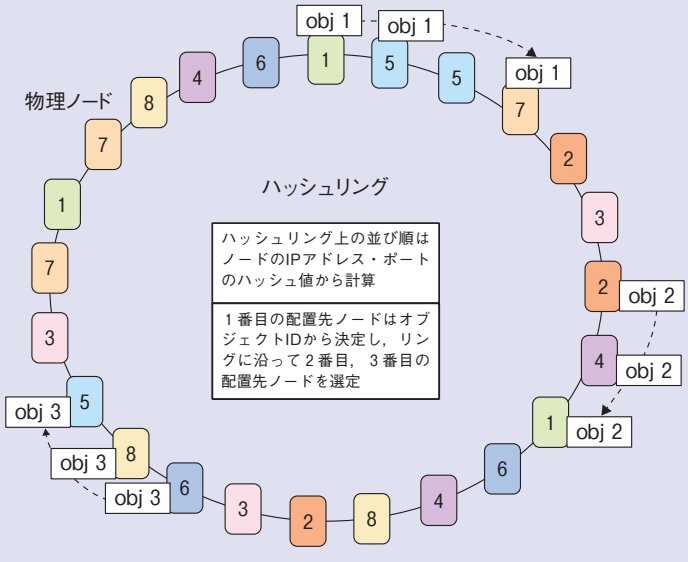
図1　Sheepdogの概要

最近の取り組み

NTTソフトウェアイノベーション



ⓐ　オブジェクトの物理配置



ⓑ　配置決定アルゴリズム

図2　コンシステントハッシュ法

センタでは，商用サービスでも安心してSheepdogを利用できるよう，主に運用性・信頼性の向上に取り組んでいます．

Sheepdogではクラスタに所属するサーバの追加・離脱管理にZookeeperを用いることができます．NTT研究所はZookeeperと組み合わせたSheepdogクラスタの網羅的なテストや長期安定試験によって課題抽出を行い，問題個所の修正をSheepdogコミュニティに提案して品質を高めています．

また，クライアントとSheepdogとの接続をクラスタ内の1台だけでなく，複数台のサーバと接続可能にし，読み書きのための経路を冗長化できるマルチパス機能の実装に取り組んでいます．この機能によって，クライアントとSheepdogクラスタ内のサーバとの接続が断たれても，別のサーバに接続して読み書きを継続することができるようになります．

さらに，激甚災害や停電による拠点単位の故障が発生した場合でも，遠隔拠点を利用してサービス停止・データ損失を防ぐ機能の開発にも取り組んでいます．

Sheepdogのオープンソースコミュニティでは，Erasure Codeと呼ばれる機能が実装されています．これはデータ損失回避のためにオブジェクトの単純複製を行うのでなく，RAID5のように，分割データと復元用データを配置し，一定数がそろえば元のデータを復元可能とする技術で，ディスク容量の消費を削減し，ハードウェアコストを抑えることができます．

オブジェクトストレージ OpenStack Swift

スマートフォンで撮影した写真をクラウド上で保存して，ほかの端末で閲覧するようなケースが一般的になってきました．その結果，クラウド上のデータは膨大な量になり，安価で信頼性の高いクラウドストレージが求められています．このニーズにこたえるため，OpenStackコミュニティで開発されているソフトウェアがオブジェクトストレージOpenStack Swift（Swift）です．NTTグループやRackspace社などでの商用実績もあります．

Swiftの特長として，以下の3点が挙げられます（**図3**）．

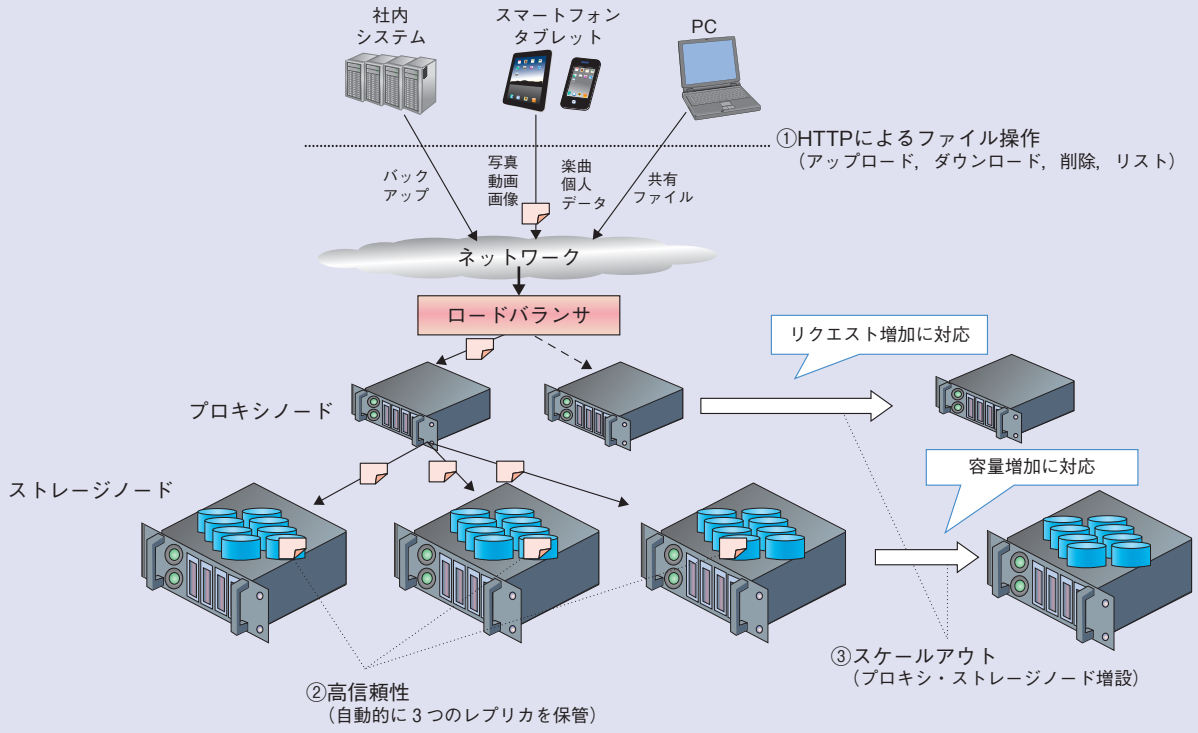


図3　Swiftの特長

⑴　HTTPによるファイル操作（REST API）

HTTPを利用することで，スマートフォン，タブレット，PCなどあらゆる端末からSwift上のデータ操作が可能です．利用用途としてはバックアップデータや写真，動画などのマルチメディアデータなど，書き換えが少ないデータの保存に適しています．

⑵　高信頼性

ストレージシステムにとって保存するデータを失うことはあってはならないことです．Swiftは高信頼性実現のため標準で3つのレプリカをクラスタ内に作成します．また，クラスタ内のオブジェクトを保持する各ノードに，レプリケータと呼ばれるプロセスが動作しており，自ノードで保持するデータがクラスタ内のほかのディスクに2つあるかを確認し続けます．ディスクが故障しアンマウントされたことを確認した場合は新しいレプリカを自動的に復元します．

⑶　スケールアウト

Swiftは自律分散で単一障害点がないため，スモールサイズからスケールアウトしていくことが可能です．典型的なSwiftクラスタの構成例を図3に示します．この例ではクライアントからリクエストを受け付けるプロキシノードと，実際にそのデータが保存されるストレージノードから構成されています．リクエストが多い場合はプロキシノード，ストレージ容量が足りなくなった場合はストレージノードを必要に応じて増設できるため，拡張性の高いクラスタアーキテクチャとなっています．

Swiftの運用性向上

NTT研究所では自立分散された多数のノードからなるクラスタを，よりスムーズに商用導入し，低コストでのサービス提供を行うため，運用の効率化に取り組んでいます．総容量1 PB（ペタバイト）のクラスタの総運用時間を分析するため，実際にPoC環境を構築し，構築，監視，増設，トラブル解析と復旧，ソフトウェアアップデートの稼動時間を定量評価しました．その結果，スケールアウトのための増設対応時間とディスク故障の復旧対応時間の占める割合が高かったため，改善を検討しました．

増設対応時間についてはOS導入・アプリケーション導入部分をPXEネットワークブート利用による自動インストールで行い，Chefによる設定自動化，リリース前の動作確認部分でAPI試験の自動化ツールTempestを利用することによって，ノード増設手順の一部を効率化し，約3分の1に短縮できる見通しを得ました（**図4**）．TempestはOpenStackコミュニティの試験ツールですが，NTT研究所としてSwift部分の試験項目を拡充し，効率化を実現するだけでなく，網羅的な試験を行うことが可能です．

ディスク故障から復旧までの時間は，

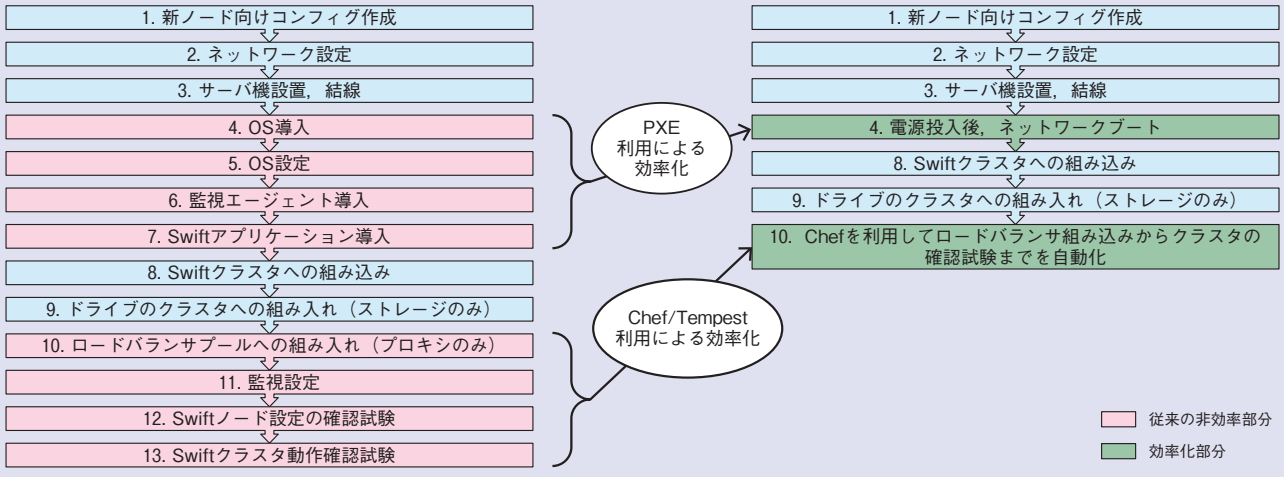


図4　ノード増設時の効率化

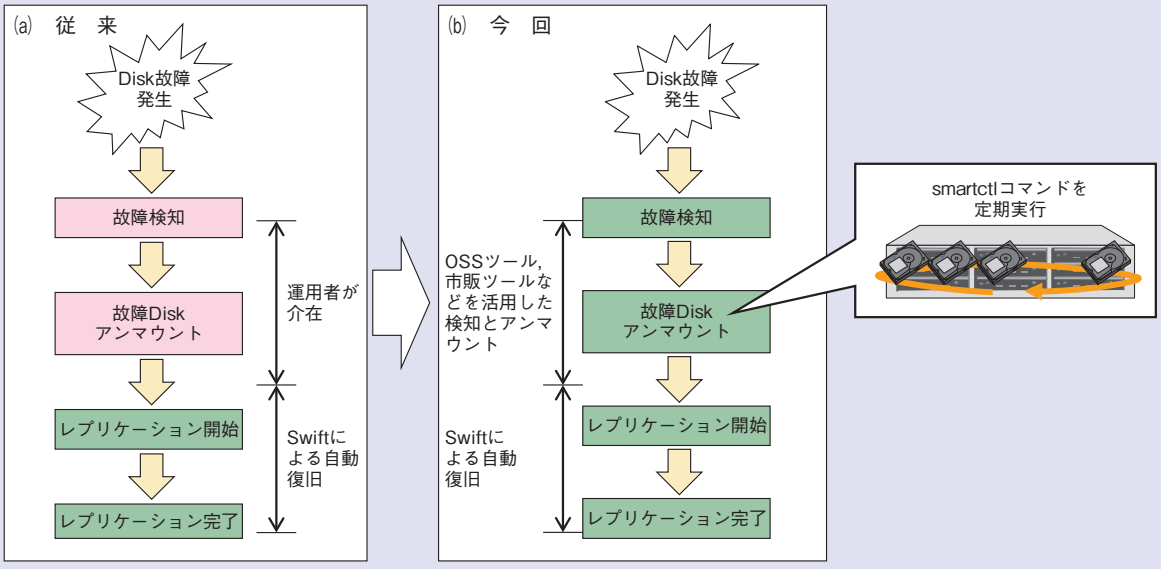


図 5　ディスク故障時の効率化

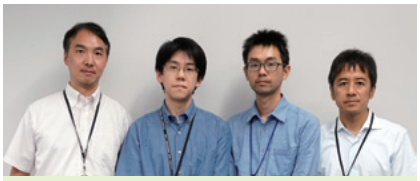
データの高信頼性を確保するうえで最小化する必要があります. 本検討では, HDDに内蔵されている自己診断機能S.M.A.R.T.（Self-Monitoring Analysis and Reporting Technology）より故障ディスクを自動検出し，さらに当該ディスクをアンマウントする自動化ツールを作成することにより，ディスク故障から復旧までのプロセスを自動化し，手動手順に比べ約 5 分の 1 に短縮できる見通しを得ました（**図 5**）.

今後の展開

Sheepdogは高い拡張性・信頼性・運用容易性を備えた完全対称形の分散ブロックストレージであり，国内外での実サービスへの導入が始まっています．今後はより安心してSheepdogを導入していただくために，引き続き品質・信頼性向上に取り組むとともに，ユーザに対する運用手順の共有や共同検証を進めていく予定です.

またSwiftは信頼性が高く，スケーラブルなオブジェクトストレージです．今回紹介した運用の効率化を発展させ, 自動化を進めると同時に, コミュニティで新機能として検討されているErasure Code機能についても取り組んでいきます.

Sheepdog，Swiftともに今後もコミュニティの主要開発者やユーザの皆様と品質向上や機能拡張に取り組み，安定性・性能・運用性の向上を目指します.



（左から）日高　東潮/ 福本　佳史/
内藤　一兵衛/ 白石　正裕

ユーザのさまざまな要求に柔軟にこたえられる仮想化基盤を目指し，今後もストレージ仮想化技術の研究開発に取り組んでいきたいと思います.

◆問い合わせ先
NTTソフトウェアイノベーションセンタ
分散処理基盤技術プロジェクト
第四推進プロジェクト
TEL　0422-59-2207
FAX　0422-59-2072
E-mail　sic@lab.ntt.co.jp