オブジェクトストレージ向けIPドライブ KVDrive

"KVDrive" Internet Protocol Drive for Object Storage Systems

田中　信吾　■TANAKA Shingo
後藤　真孝　■GOTO Masataka
フィリップ　クフェルト　■Philip KUFELDT

昨今の世界のデジタルデータ量の飛躍的な増大に対応するため，大規模ストレージを低コストで構築・運用可能にするオブジェクトストレージが普及しつつある。しかし，現在のオブジェクトストレージは既存のハードウェアやソフトウェアを用いて構築されているため，大規模化するには多台数のサーバが必要になるなどの問題がある。

そこで東芝は，この問題を解決するIP（Internet Protocol）ドライブ型のストレージ“KVDrive”を開発した。クライアントからの直接アクセスを可能にする構成やKey-Value（以下，KVと略記）型API（Application Programming Interface）により，大規模ストレージを実現するために従来必要であった多台数のサーバの省略やソフトウェア構成の簡略化を可能にし，システムの低TCO（Total Cost of Ownership）化及び高性能化を実現した。

Object storage, a new storage system that enables large-scale storage systems to be constructed and managed at low cost, has recently been spreading in response to the explosive growth of digital data. Current object storage systems have several issues, however, including the need for a large number of servers when the system is enlarged due to its immature architecture, which relies on currently existing hardware and software.

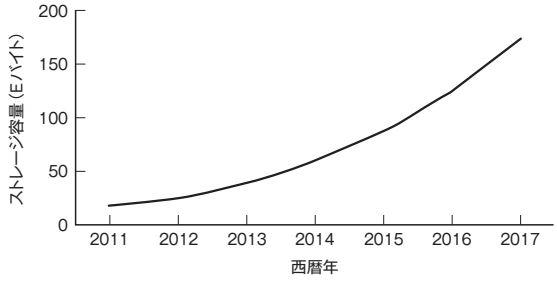
Toshiba has developed the "KVDrive," an Internet Protocol (IP) drive storage system, in order to solve these issues. With the KVDrive, the need for a large number of servers is eliminated and the software architecture is simplified due to (1) the adoption of an architecture that allows direct access from clients, and (2) a key-value (KV) application programming interface (API). These features make it possible to realize a high-performance system with lower total cost of ownership (TCO) using the KVDrive.

1　まえがき

スマートフォンや，SNS（Social Networking Service），クラウドコンピューティングなどの普及により，世の中のデジタルデータ量が飛躍的に増大している。更に，ビッグデータ解析などによる大量のデータ処理技術が普及の兆しを見せており，世界のエンタープライズ領域とクラウドサービス領域で使用されるストレージ容量は2017年には2013年の約5倍と爆発的に増大することが予想されている（**図1**）[1]。

一方，IT（情報技術）システムに対する投資額は世界でも年率3％程度の伸長と予想されており[2]，4年間でもわずか10％強で前述のストレージ容量の成長率との間には大きな差がある。そのため，容量が年々増大するストレージシステムを低コストで構築及び運用する技術の開発が重要な課題となっている。これに応えるため，オブジェクトストレージと呼ばれる新たなストレージシステム技術が普及しつつある。

ここでは，オブジェクトストレージの現状と問題の詳細，及びこれを解決するために東芝が開発したIPドライブ型のストレージ“KVDrive”の特長について述べる。



E：エクサ（10^{18}）

＊IDC（International Data Corporation）「Worldwide File- and Object-Based Storage 2013-2017 Forecast」[1]に基づいて作成

図1．世界のエンタープライズ及びクラウドサービス向けストレージ容量の推移 ― 世界のエンタープライズ領域及びクラウドサービス領域で使用されるストレージ容量は爆発的に増大し続け，2017年には2013年の5倍になると予想されている。

Trend in global storage capacity for enterprise and cloud service use

2　オブジェクトストレージの概要

オブジェクトストレージとは，Webサービスなどのバックエンドで使われているいわゆるクラウドストレージなどの大規模システムを主なターゲットとして設計されたストレージシステムである。

ストレージシステムに使われるハードディスクドライブ（HDD）

はデータをブロック単位で扱い，各ブロックにはLBA（Logical Block Address）と呼ばれる方式を用いてアドレスが与えられている。従来のストレージシステムは，HDDをRAID（Redundant Arrays of Independent（Inexpensive）Disks）で冗長化し，ファイルシステムを介してクライアントが持つアプリケーションへのインタフェース（I/F）を提供するという組合せが基本となっている（**図2**(a)）。RAIDは中央のコントローラによって制御されるシステムであるため，多くとも数十台程度のHDDを収容するのが限界であり，また，ファイルシステムも単一のコンピュータで動作するように設計されたもの（スケールアップ型）である。したがって，システム規模を更に大きくするには，複数のサーバを何らかの方法で分散制御する技術が別途必要になり，一般にはストレージベンダーが提供している高価な専用ストレージシステムを利用する必要があった。

一方，一部の先進的なIT企業は，大規模な仮想ストレージを低コストで実現する技術を独自に開発してきた。この技術が少しずつ一般に開示され，技術の本質がよく知られるようになり，デファクト技術としてまとめられてきたものがオブジェクトストレージである。

オブジェクトストレージの概要を図2(b)に示す。オブジェクトストレージの特長としては，大きく分けて次の三つが挙げられる。

(1) シンプルなAPI　オブジェクトストレージが一般に扱うデータはWebコンテンツ（Webページや，画像，音声，動画など）などであり，基本的にはこれらのデータを塊（オブジェクト）として読み（GET），書き（PUT），及び削除（DELETE）できればよく，従来のファイルシステムが持つディレクトリ構造，並びにファイルのオープン及びクローズやアクセス権の制御などの高度な機能はいっさい不要である。これらを省略し，階層がないフラットな空間へ任意長のデータオブジェクトをGET，PUT，及びDELETEする機能を中心とした，Webコンテンツをやりとりするプロトコルと同様のREST（Representational State Transfer）ベースのAPIにI/Fが単純化されている。

(2) スケールアウト型　オブジェクトストレージでは，一般にオブジェクトの識別子を所定のハッシュ関数に乗じ，得られた値に基づいて特定されたストレージサーバにデータを格納する。クライアントは同じ計算をすることで，他の管理サーバなどにアクセスすることなく分散配置されたサーバ群からターゲットとなるストレージサーバを特定し，直接アクセスしてオブジェクトのGET，PUT，及びDELETEを行うことができる。これにより，基本的にストレージサーバを並列に接続することでシステムを容易に大規模化することが可能な構造（スケールアウト型）になっている。ストレージサーバの追加や削除など構成が変わったときの同期処理のためなど，一般には管理サーバも別途使われるが，クライアントがストレージサーバのデータにアクセスするたびに管理サーバに問い合わせる必要はなく，管理サーバが性能のボトルネックにならないように設計上の配慮がされている。

(3) 耐障害性　各ストレージサーバにデータを振り分ける際，一般に一つのデータは三つのストレージサーバに格納される形で冗長構成が取られている。例えばあるストレージサーバにHDDなどの障害が発生しても他のストレージサーバから自動でデータが読み込まれてクライアントからは障害が見えないように動作し，優れた耐障害性を実現する。また，一般に各ストレージサーバに様々な形である程度のインテリジェンスを持たせ，管理サーバを介することなくストレージサーバどうしでデータをやりとりして効率よく動作することも可能になっている。

更に，これらの特長の発展形として，オブジェクトストレージの利点を活用しながらクライアントに対し従来のI/Fも提供するシステムである，オープンソースソフトウェアのCeph[3]が普及しつつある。Cephの構成を**図3**に示す。

Cephは前述したオブジェクトストレージの技術をベースに，RADOS（Reliable Autonomic Distributed Object Store）と呼ばれるオブジェクトストレージの共通クラスタに対し，クライアント側に複数の異なるI/Fソフトウェアを用意することで，クライアント上のアプリケーションに対して，オブジェクトI/Fだけでなくブロック I/FやファイルI/Fを提供できる。これにより，従来のアプリケーションに対しても，単純にストレージサーバを並列に接続することで大規模ストレージシステムを提供することが容易になり，オブジェクトストレージの特長であるスケールアウト性と耐障害性を幅広いアプリケーション領域に展開可能である。特に同じオープンソースソフトウェアのIaaS（Infrastructure as a Service）クラウドコンピューティング基盤であるOpenstack[4]との親和性が高く，これと組み合わせて

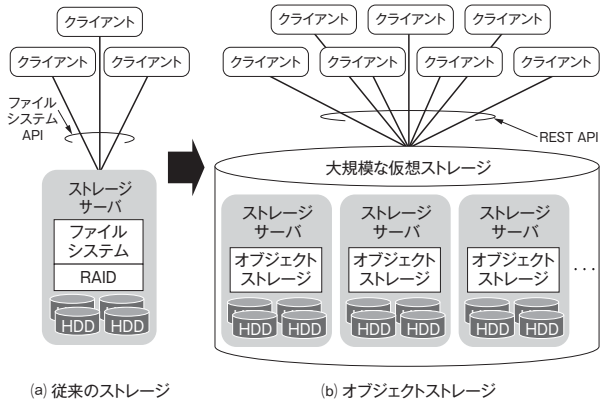


図2．従来のストレージとオブジェクトストレージ ― 従来のストレージシステムはスケールアップ型であるのに対し，オブジェクトストレージシステムはスケールアウト型である。

Configuration of conventional storage and object storage systems

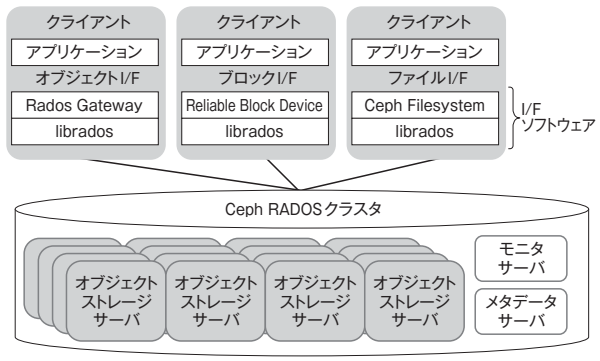


図3. Cephの構成 ― オブジェクト，ブロック，及びファイルの三つのストレージI/Fを構成できる。

System configuration of Ceph distributed object store

使用されるケースが多い。

3 オブジェクトストレージの課題

前述のように大規模ストレージシステムを容易に実現することが可能なオブジェクトストレージだが，現状では次のような課題がある。

(1) 多台数のサーバが必要　　オブジェクトストレージでは，末端のストレージとして一般に汎用サーバベースのストレージサーバを構成する必要がある。このため，システムを大規模化するには，ストレージを増設するだけでなく，増設したストレージを収容してネットワークに接続するためのサーバもハードウェアとして増設する必要があり，その分コストが掛かる。

(2) 依然としてファイルシステムに依存　　オブジェクトストレージのストレージサーバ上に搭載されるソフトウェアは，一般にファイルシステム上に構成される。本質的には前述のとおり可変長のオブジェクトをデータの塊としてGET，PUT，及びDELETEができれば十分なので，ファイルシステムのような複雑なデータ構造を用いる必要はないが，可変長データを扱うのが容易なために依然として用いられているのが現状である。これにより性能の低下が生じている。

(3) ストレージデバイスの使いこなしが未成熟　　オブジェクトストレージは，現状ではアクセス速度がそれほど要求されない領域（アーカイブやコールドストレージなど）で主に使われており，あくまでHDDを使うことだけを目的として実装されるのが一般的である。Cephなどによってその適用領域が高性能領域にも広がってきていると言えるが，それに必要なソリッドステートドライブ（SSD）や，バッテリーバックアップ付きのDRAMのような不揮発性メモリなどを積極的に活用したものはまだほとんどない。

4 KVDrive

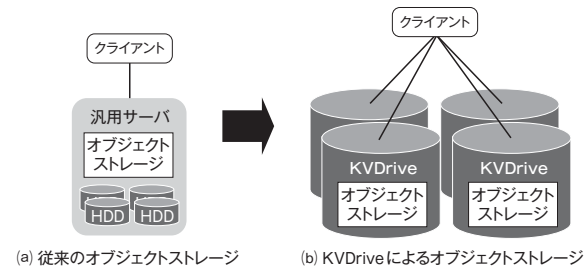
前述のオブジェクトストレージの課題を解決するストレージとして，当社は“KVDrive”（**図4**）を開発した。KVDriveの主な特長について，以下に述べる。

(1) 3.5型形状IPドライブ　　KVDriveは3.5型HDDと同じ形状でありながら，ストレージ（SSDやHDD）とSoC（System on Chip）を内部に持つマイクロサーバである。物理I/FはGigabit Ethernet(†)で，TCP/IP（Transmission Control Protocol/Internet Protocol）通信をベースとしたネットワーク機能を持つ“IPドライブ”としてクライアントから直接データのGET，PUT，及びDELETEのアクセスを受け付けることが可能である。更に，オブジェクトストレージのサーバ上に搭載されるソフトウェアをこのKVDrive内に搭載し，完全なオブジェクトストレージのサーバとして動作させることもできる。これにより，従来必要であったストレージを収容する汎用サーバを省略でき，システムの低TCO化に貢献する（**図5**）。

(2) KV型API　　KVDriveのデータI/FとしてKV型のデータI/Fを提供する。KVとは，任意長の識別子Key

図4. KVDrive ― 3.5型HDDと同じ形状でありながら，内部にストレージとSoCを持つマイクロサーバである。

KVDrive



(a) 従来のオブジェクトストレージ　　(b) KVDriveによるオブジェクトストレージ

図5. 従来のオブジェクトストレージとKVDrive によるオブジェクトストレージ ― KVDriveでは，従来のオブジェクトストレージで必要だった汎用サーバを省略できる。

Configuration of conventional object storage system and object storage system with KVDrive

と任意長のデータValueを関連付けて記憶し，Keyを指定してValueをGET，PUT，及びDELETE可能にするデータ構造であり，KV型APIは同様のアクセス方式で任意長のデータのGET，PUT，及びDELETEに特化したオブジェクトストレージともっとも親和性が高いI/Fである。ユーザーはこのAPIを活用することで，ファイルシステムを使用する必要がなくなるため，ソフトウェア階層をシンプルにして性能を向上させることができる。

(3) SSDとHDDの最適制御　(2)で述べたKV型APIで提供する機能にストレージ制御機能を備え，SSDとHDDを最適に制御する機能を提供する。従来は，これらの異なるストレージを使いこなすにはストレージを利用するユーザー側で技術開発を行う必要があり，更に新たなストレージが登場するたびにそれに対応しなければならなかった。これに対してKVDriveは，SSDやHDDのストレージベンダーである当社の強みを生かし，これらの活用技術をKV型APIという従来のLBAより一段抽象化されたAPIで提供する。これにより，ユーザーは新たなストレージメディアの活用技術を開発し続ける必要がなくなる。特に今後は瓦記録（SMR：Shingled Magnetic Recording）HDDやSCM（Storage Class Memory）などの新ストレージメディアの登場も期待されており，これらの活用によるシステムの高性能化のメリットをタイムリーに提供できる。データ転送性能としては，最大でほぼGigabit Ethernet(†)限界帯域となる約110 Mビット/sのスループットを実現する。これはSSDを活用することによって実現されており，既存の類似のIPドライブの約2倍の性能である。

上記特長(1)と(2)の具体的な仕様は，Kinetic Open Storage Platform[5]に準拠している。

5　あとがき

オブジェクトストレージ向けのIPドライブKVDriveを開発した。ドライブ型でありながら，オブジェクトストレージとしてクライアントからの直接のアクセスを可能にする構成により，従来必要であったストレージを収容する多台数のサーバを省略でき，システムの低TCO化を実現する。また，KV型APIにより従来のファイルシステムへの依存を解消し，更にSSDやHDDなど各種ストレージメディアの活用技術も提供することで，新ストレージメディアの早期活用とそれによるユーザーのシステムの高性能化に貢献する。

今後はAPI仕様やハードウェア外部仕様などの標準化を進めながら，近い将来に普及が見込まれるCephを第一のターゲットシステムとして開発を行い，KVDriveの早期実用化を目指す。

文　献

(1) IDC (International Data Corporation). Worldwide File- and Object-Based Storage 2013-2017 Forecast. USA, IDC, 2013, 46p.
(2) Gartner. "Gartner Worldwide IT Spending Forecast". Gartner Homepage. <http://www.gartner.com/technology/research/it-spending-forecast/>, (accessed 2015-07-19).
(3) Inktank Storage. "ceph". Ceph Homepage. <http://ceph.com/>, (accessed 2015-07-19).
(4) OpenStack Foundation. "OpenStack: The Open Source Cloud Operating System". OpenStack Homepage. <https://www.openstack.org/software/>, (accessed 2015-07-19).
(5) Seagate Technology. “Seagate Kinetic Open Storage プラットフォーム”. Seagate Technologyホームページ. <http://www.seagate.com/jp/ja/solutions/cloud/data-center-cloud/platforms/>,（参照 2015-07-19).

田中　信吾　TANAKA Shingo
セミコンダクター＆ストレージ社 ストレージプロダクツ事業部 ストレージソリューション推進部参事。ストレージ応用製品の研究・開発に従事。
Storage Products Div.

後藤　真孝　GOTO Masataka
セミコンダクター＆ストレージ社 ストレージプロダクツ事業部 ストレージソリューション推進部グループ長。ストレージ応用製品の研究・開発に従事。情報処理学会会員。
Storage Products Div.

フィリップ　クフェルト　Philip KUFELDT
東芝アメリカ電子部品社 ストレージプロダクツ部シニアマネージャー。ストレージ応用製品の研究・開発に従事。
Toshiba America Electronic Components, Inc.