解 説 4

IBM Active Cloud Engine

ー 必要な時に必要な場所でデータをアクセスするための自動最適化されたストレージ技術 ー

携帯型デジタル機器やデジタル・センサーの急速な普及に伴い、保管、処理しなければならないデジタル・データは増加の一途をたどっています。これらのデータを迅速に処理し、新しい価値を生み出すことが新しい時代を乗り切るための重要な経営課題と考えられるようになってきています。

本稿では、今後ストレージ装置に求められる機能を整理し、過去のストレージ技術の変遷を振り返りつつ、注目されているストレージ最新技術を解説します。

Article 4

IBM Active Cloud Engine

- Delivering the Right Data to the Right Place at the Right Time -

Due to the rapid and widespread growth of portable digital devices and sensors, the amount of unstructured digital data stored and processed continues to increase. Businesses are faced with the challenge of leveraging data-driven strategies in order to innovate and to compete against their competitors.

In this article, I will describe the evolution of our high performance parallel file system and summarize the functionality that will be implemented in storage systems of the future. I will also describe the latest in storage technology, which can deliver the right data to the right place at the right time.

❶ ビッグデータ時代到来に向けたストレージの課題

近年、ビッグデータと呼ばれるさまざまなソースから生成されるデータは、従来のIT 機器が処理の対象としていたデータに比較して想像をはるかに超えるペースで増え続けています。特に増加し続けているデータは非構造化データ（テキスト、映像、音声など）と呼ばれるデータです。音声、ビデオのように１つのデータ当たりの容量が大きく、読み書きの速度が重要なものもあれば、センサーから生成されるデータのように1つのデータ当たりの容量は大きくないが、莫大なデータ数に上るものもあります。もはや、データ保護、データ管理を人手に頼ることは不可能との認識が広まっています。

今後、有効活用が望まれるビッグデータはその生成個所が地球規模で地理的に分散しており、データを保管する場所、データを利用する場所を柔軟に選択できることが重要となります。データが生成され、保管される場所が多様化し、さらにはデータを利用する場所がモバイル環境ということも当たり前になってきている状況では、特定の場所にだけ高性能なストレージ・システムを構築するという手法では不十分で、ネットワーク通信による帯域、遅延の影響などを受けてしまう場合もあります。利用者がデータが保管されている場所を気にすることなく利用できなくてはなりません。

保管、管理しなければならないデータがペタバイトを超える場合、従来のようにバックアップの対象データを特定することも困難になってきます。しかし、全データのバックアップを定期的に行うことも非現実的です。そのため、バックアップ対象となるデータの重要度に応じてバックアップの方法を適切に選択する必要があります。データが生成されてからまったくアクセスされていないデータと、毎日更新があり、その更新分のデータ損失が莫大な被害を及ぼすデータとでは、データ保護の方法を変えるべきです。そもそもバックアップが必要なデータと一時的に保管できればよいデータを区別する必要があります。本番サイト、バックアップ・サイトでデータを同期する、コピーを取得しておくなどの方法も検討する必要があります。

つまり、データの分散配置の最適化、データの保護、複数サイトから構成した災害対策などの要件を個別に解決するのではなく、必要な要件を満たせるようにシステム全体を考慮して設計することが重要になります。今までのように、個別にサーバー装置、ストレージ装置を購入し、お客様ごとにシステムを設計、実装する手法は、導入構築のコストや期間の観点からも変革が求められています。

❷ 今ストレージ装置に求められること

まず、増え続けるデータに対応して、容量および性能を動的に拡張できることが重要です。初期投資を抑え、かつランニング・コストの低減の観点から、最初から高性能な装置を購入することは得策とはいえません。保管されるデータの種類により、求められるストレージの性能が大きく異なりますが、どのようなデータが格納され、どのような性能が求められ

るかをあらかじめ要件として定義しておき、ストレージ装置の構成を決めておくことは困難です。性能とコストに見合うストレージ階層を上手に利用することで全体のコストを抑えることができるので、異なる性能のハードディスク間で自動的にデータを移動させるだけでなく、テープ装置も含めて自動的に配置する自動階層化の機能を最初から検討しておくことが重要です。テープ装置は、保管に掛かるバイト単価を低く押さえられることが大きな特長ですが、消費電力の削減という観点からも近年注目されているストレージ装置です。テープ装置を上手に利用することで総容量当たりの総消費電力を削減することが可能となります。

次に、データのライフサイクルに合わせて、日々刻々変化するデータの価値に応じて自動的にデータの保護、配置などを行えることが重要です。そのためには、膨大な数に及ぶファイルの中から更新されているファイルを特定するための高速スキャンの実行が必要となります。

複数のストレージ階層を利用する場合であっても、ユーザーから見えるファイルの場所（ネーム・スペース）が変化しないことが必要です。さらに、地理的に異なる複数のファイル・ストレージ装置全体で単一のファイル名空間がサポートされ、物理的に異なるサイト間でのファイル・レベルでの自動的な配置も可能であることが重要となります。

このローカル・サイト内での自動階層と複数サイト間での自動配置機能の両方を活用することで、これまで問題となっていた運用コストの削減が実現でき、またストレージ装置の使用率の向上、投資対効果の向上にもつながります。

データが想像を超えるペースで増え続ける今の時代は、非構造化データを効率よく保管・管理するために、ファイル用のストレージ装置のアーキテクチャーを根本的に変革する時だといえるでしょう。

❸ ファイル用のストレージ技術の変遷

非構造化データを格納するストレージ装置として Network Attached Storage（以下、NAS）があります。もともと、ファイル・データをディスク装置に格納する方法、フォーマットは、サーバー上で稼働するオペレーティング・システム固有のものでした。UNIX オペレーティング・システムが稼働するオープン・システムが普及を始めた 1970 年から 1980 年代初頭になると、ローカル・エリア・ネットワーク（LAN）が急速に普及し始め、ネットワーク経由でファイル・システムを共有するためのファイル・サーバーが研究・開発されました。広く利用されるようになった最初のネットワーク・ファイル・システムは Sun Microsystems 社（当時）によって開発された Network File System（以下、NFS）です。複数のサーバーから1台のファイル・サーバーに同時にアクセスでき、ファイル・レベルで複数のサーバー間でアクセスすることが可能になりました。ファイルに対する書き込み競合を回避するための排他制御の仕組みなども導入され、ファイル・レベルで読み書きを特定のユーザーに限定するためのアクセス制御の機能も実装されました。Windows 環境でも同様のファイル共有サービスが開発されました。当初、NetBIOS 上で稼働する SMB（Server Message Block）プロトコルで実装されていたファイル共有プロトコルは、現在では TCP/IP 上の CIFS（Common Internet File System）として広く利用されています。ファイル･サーバーは汎用サーバー上の汎用オペレーティング・システム上で動作させることができましたが、多くのクライアントからのファイル入出力を処理するためには、当時のファイル・サーバーでは処理能力が不足していることが普及を妨げる要因の 1 つになっていました。

このようなファイル・サーバーの課題を解決するためにファイル・サービスに特化したアプライアンス製品が登場し、NAS という新しいストレージの市場が生まれました。NAS 装置は、ファイル・サービスに特化したハードウェアやソフトウェアを搭載することで汎用のサーバーとは比較にならない性能を発揮し、しかも簡単に導入・運用できることが特長でした。

登場以来約 20 年にわたり広く普及してきた NAS 装置ですが、近年では新たな課題が認識されてきています。複数のクライアントに対して 1 台の NAS 装置を配置して、ファイル共有サービスを提供することがそもそもの NAS のアーキテクチャーです。ところが、爆発的に増加を始めたデータを格納するためには、柔軟な容量拡張、性能拡張が求められるようになりました。従来の NAS 装置では容量や性能が不足すると、より上位機種の NAS 装置に移行するか、あるいは別の NAS 装置を追加導入しなければなりません。これは、NAS 装置を運用・管理する観点からは、サービスが利用できるまでの時間、投資に必要なコスト、運用コストを考慮すると歓迎できる方法ではありません。複数の NAS 装置にデータを配置すると、NAS 装置固有の共有ファイルの名前が必要なため、どの NAS 装置にデータが格納されているかをユーザーが意識する必要があり、システムの柔軟な構成変更を妨げる要因の 1 つになります。

1 つの組織であっても複数の拠点が存在する場合、拠点ごとに複数の NAS 装置が利用されています。組織全体での NAS 装置内のデータの管理運用が課題になっているだけでなく、拠点ごとに分散しているデータを組織全体で有効に利用できないことも課題となっています。ファイル・サーバーをセンターに統合するという解決手法もありますが、データが生成され、主に利用される拠点にデータを配置することが望まれます。しかし、この方法では拠点ごとにデータ保護を行わなければならず、運用コストを引き上げている大きな要因となっています。

NAS 装置というファイル・サービスに特化した従来のアプライアンスは、本来投資対効果が高い製品ですが、地球規模で分散する非構造データを地球規模で利用するような用途に対しては、データの保護、複数拠点でのデータの複製、データの同期を行う仕組みをお客様ごとに設計・実装しなければならないという課題があります。これらの機能を実現する仕組みを持ち、あらかじめ設計・統合された知的なストレージ装置を利用することで、新たな経営課題の解決により多くの投資と時間をかける時代がやってきたのです。

❹ 進化を続ける IBM General Parallel File System

IBM では、1990 年代にファイル・サーバーの課題を解決するために高性能で高機能な並列ファイル・システムの研究が行われていました。そこで開発されたのが、IBM General Parallel File System（以下、GPFS）[1] で、これは複数ノードから同時アクセス可能な分散共有型の並列ファイル・システムです。デジタル・ビデオ配信のために IBM アルマデン研究所で開発された Tiger Shark マルチメディア・ファイル・システム [2] がその起源となっています。IBM RS/6000 Scalable POWERparallel Systems（以下、RS/6000 SP）用のファイル・システムとして、1998 年に最初の IBM GPFS が発表されました。RS/6000 SP は、RS/6000 をベースとした並列スーパー・コンピューターで、プロセッサー・ノード間を高速に接続するための専用の高速スイッチ機構 High Performance Switch（以下、SP スイッチ）を備えていました。GPFS は、この SP スイッチを経由して、各 SP ノードより透過的なファイルのアクセスを実現するためのソフトウェアとして提供されました。当時は、ハードディスク・ドライブ（以下、HDD）の容量およびアクセス性能が現在に比べて著しく低く、通常の RAID を利用するだけでは、デジタル・ビデオ・データを期待するビット・レートで複数同時に読み出すことができませんでした。GPFS では、複数のサーバーを並列に動作させ、1 本のデジタル・データを複数の HDD に分散させることで、複数のデジタル・ビデオ・データを高速に読み出し、配信することを可能にしました。その後数々の改良が加えられ、現在では TOP 500（処理速度による世界のコンピューター・ランキング）に含まれる多くのスーパー・コンピューターに採用されており、10 年以上の稼働実績があります。

GPFS の構成例を**図 1** に示します。GPFS では、Network Shared Disk（以下、NSD）という機能を使用します。RS/6000 SP 上では、VSD（Virtual Shared Disk）という仕組みが SP スイッチを使って実装されていましたが、現在の GPFS では InfiniBand や Ethernet を利用した NSD が主に利用されます。NSD は、iSCSI のように IP ネットワークを経由した SAN（Storage Area Network）のブロック・アクセスをソフトウェアで実現したものです。ディスク装置が直接接続されていない GPFS ノードであっても、あたかもローカルにストレージ装置が接続されているように動作します。GPFS のノードは、ストレージ装置のブロック入出力を行うための機能、GPFS としてのファイル入出力を行い、アプリケーションを稼働させる機能を持っていますが、どのノードでどの機能を稼働させるかは柔軟に構成できます。**図 1** では、ディスク装置のブロック入出力に特化した、GPFS I/O ノードと GPFS 計算ノードで構成した例です。GPFS で提供するファイル・システムは、POSIX API、POSIX ACL、NFS V4 ACL に準拠していますので、計算ノードからは、通常のファイル・システムと同様のインターフェースでファイルにアクセスすることが可能です。ファイル・システムは、通常の UNIX のファイル・システムと同様に inode とデータ・ブロックから構成されています。通常の NFS を利用したファイル・システムと比較して非常に高速にアクセスできることが特長となっています。GPFS クラスターを構成するノードが GPFS を NFS エクスポートすることで、NFS クライアントから GPFS にアクセスさせることも可能です。

現在では、スーパー・コンピューター市場以外でも GPFS は広く採用されています [3]。GPFS は、ファイルへの高速な入出力のほかに、グローバル・ネーム・スペースのサポート、高いデータ可用性の実現、複数サイト間でのデータの複製、多種多様なワークロードに対してのスケールアウト可能な性能、災害対策機能、ILM（情報ライフサイクル管理）機能など多様な特長を備えています。

スケールアウト型のファイル・サーバーとしては、IBM Scale out File Services（以下、SoFS）[4] が GPFS をベースとして開発されました。今までのファイル・サーバーは、スケールアップ型のアーキテクチャーで設計されており、想定した容量あるいは性能に達すると、上位機種に置き換

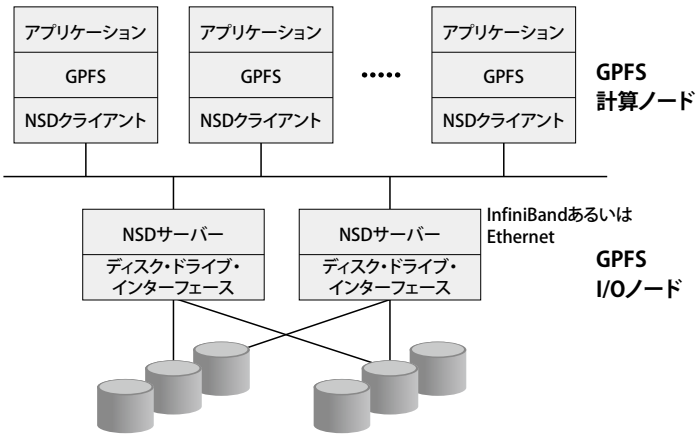


図1. GPFSの基本構成

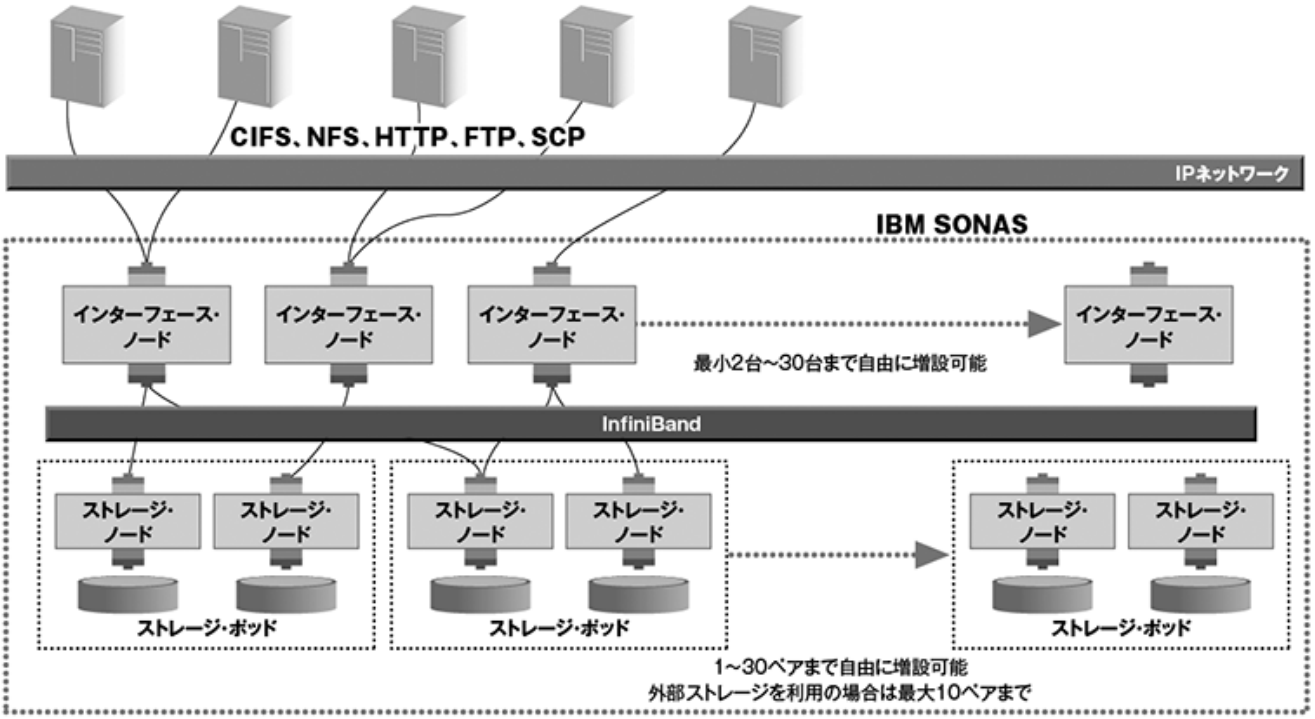


図2. IBM SONASの構成図

えるか、別のファイル・サーバーを追加導入するしかありませんでした。SoFS は、スケールアウト型のアーキテクチャーを採用し、容量の拡張と性能の向上を独立して実現できるクラスター構成に対応したソフトウェアです。その後、SoFS をベースに開発されたファイル・ストレージ製品が IBM Scale Out Network Attached Storage（以下、SONAS）です。SONAS は、出荷前にあらかじめソフトウェアが導入された統合ハードウェア製品です。SONAS では、Windows クライアントから CIFS プロトコルで接続するために Samba のクラスター機能（CTDB）を採用しています。Samba のオープンソースでの開発には、IBM も開発者として参画しています。また、GPFS はコマンドライン・インターフェース（CLI）に特化したユーザー・インターフェースしか提供していないため、SONAS 用に直感的に操作できる GUI を新たに開発しています。災害対策ソリューションのために、複数の SONAS 間でデータの非同期コピーを高速に実現する機能も新たに開発され、実装されています［5］。

図 2 に SONAS のシステム構成を示します。ストレージ・ノードとインターフェース・ノードを InfiniBand 経由でクラスター構成することで、ストレージの拡張性とファイル・サービスおよびユーザーの接続能力の拡張性を独立して行えることが特長です。また、汎用のサーバーと汎用の InfiniBand 製品を利用しているので、最新のハードウェア技術を短期間で SONAS 製品に取り入れることができます。システムの可用性向上のため各ノード、接続ネットワークを多重化しています。

❺ IBM Active Cloud Engine

SONAS は、データの使われ方に応じて、あらかじめ指定したポリシーに基づきシステムが自動的にデータの配置やバックアップを実行する Active Cloud Engine［6］と呼ばれる機能を実装しています。**図 3** に、ファイル・システムの構造を示します。ファイル・システムは、単一あるいは複数のストレージ・プール上に作成可能です。ストレージ・プールは、主に性能に応じてグループ分けされます。例えば、高性能な SAS HDD から構成されるストレージ・プールと大容量の NL-SAS（Near line SAS）HDD から構成されるストレージ・プールにグループ分けするということが可能です。複数のストレージ階層プールに対してファイル・システムを作成すると、ファイル・システム内で自動的なデータの再配置が可

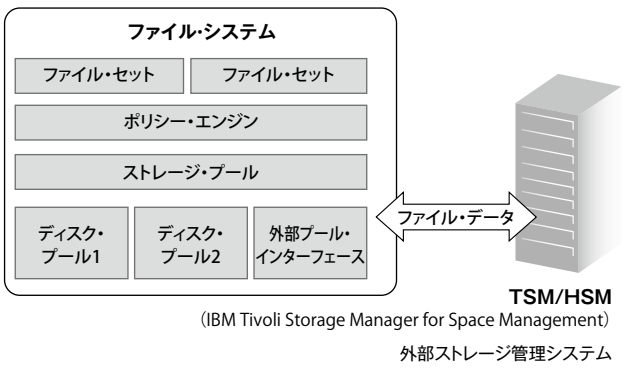


図3. IBM Active Cloud Engineの構成要素

能になります。ストレージ階層には、外部インターフェース経由でテープ装置を含めることが可能です。ファイル・システムは、1つあるいは複数のファイル・セットに分割することができます。ファイルの複製やスナップショットの作成は、ファイル・セット単位で行われます。これらの機能は、GPFS が提供するユーザーによって定義可能な File Placement（ファイルの配置）と File Management（ファイルの管理）という 2 つのポリシーを利用します。ファイルの配置ポリシーでは、ファイルが作成されたときに配置されるストレージ・プールをポリシーにより決定します。例えば、ファイル名やユーザー名に応じて配置するストレージ・プールを指定することが可能です。ファイル管理ポリシーでは、最後にアクセスされた日時などのファイルの属性に応じてファイルを移動させたり、ファイルの削除を行うことができます。**図 4** にファイルの自動配置の例を示します。作成されたファイルで 30 日間アクセスがない場合は、高性能なストレージ・プールから NL-SAS プールに自動的にファイルが移動され、さらに 160 日間アクセスがなかった場合は、外部のテープ装置に移動されます。このファイルへのアクセス方法はユーザーからは認識されません。また、重要なファイルに関してはファイルの複製を自動的に行うことも可能です。このポリシーは、ファイル・システムと密に連携するポリシー・エンジンにより実装されていますので、ファイルが膨大な数に増加しても高い性能を維持でき、1 つの巨大なファイル・システムを構成した場合でも効率よくファイルを管理・運用できます。この機能は、不要なファイルを特定した自動削除、バックアップすべきファイルを特定した自動バックアップ、複製すべきファイルを特定した自動複製を行うためには重要となります。これが、単一サイト内で実現される IBM Active Cloud Engine の特長です。

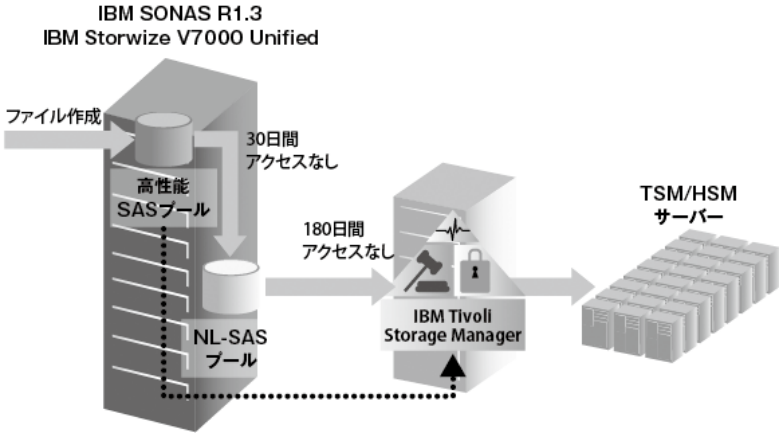


図4. ファイルの自動階層管理例

SONAS は、複数の SONAS 装置を IP 接続し、物理的に分散した複数のシステムを単一の名前空間のファイル・サーバーとして仮想化する機能を提供します（**図 5**）。ファイル・システムは複数の SONAS に分散していますが、それぞれの SONAS からはすべてがローカル・ファイル・システムであるかのように構成することが可能です。SONAS R1.3 では、書き込み権限は 1 つのサイト、読み出しは複数のサイトで可能であるように構成されます。この機能は、Panache［7］と呼ばれる分散ファイル・システムの WAN キャッシュ技術を採用することで実現されています。**図 6** と**図 7** に WAN キャッシ

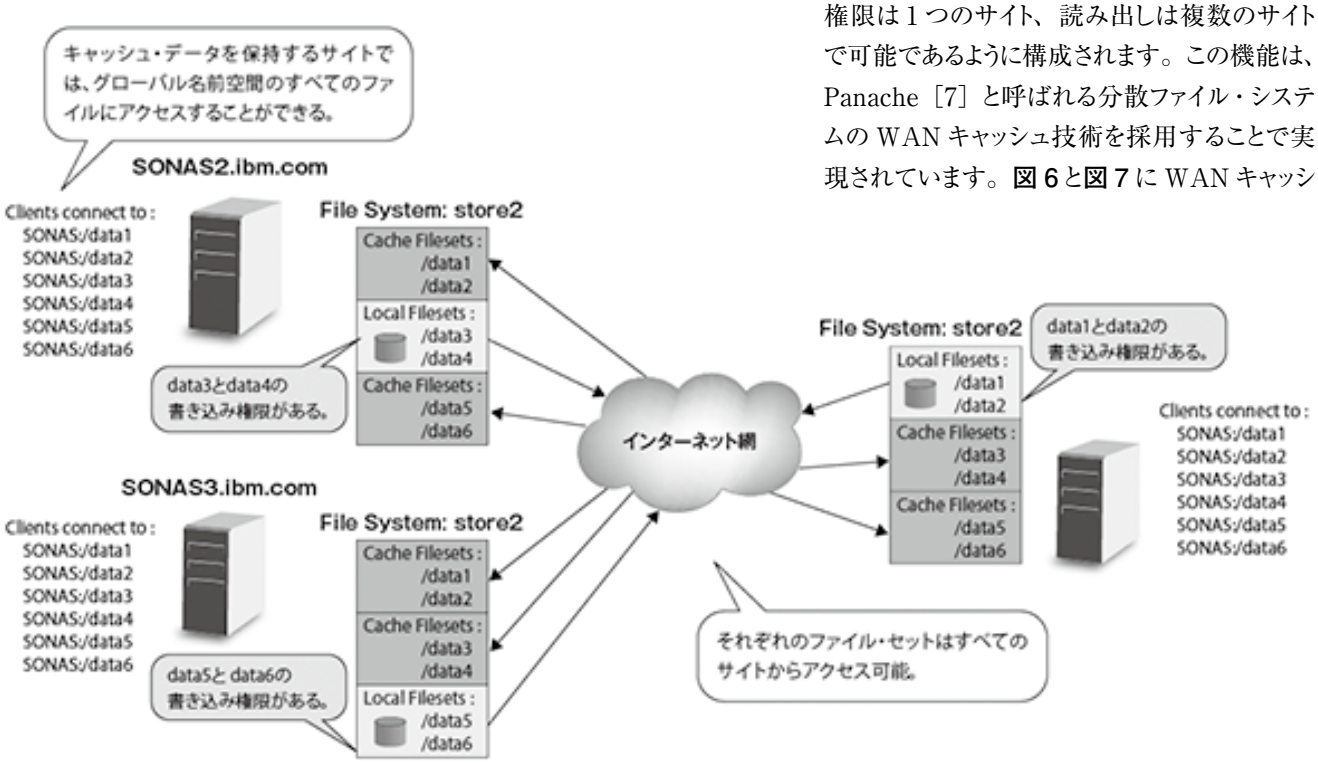


図5. 複数サイトから構成される単一ファイル空間の例

ングの機能を利用したシステムの利用方法の例を示します。今後、同一ファイル・システムに対する複数のサイトからの書き込みについてもサポートする予定です。リモート・サイト上のファイル・システム内のファイルは、ローカル・サイトにキャッシュすることが可能ですので、リモート・サイト上のファイルであっても高速にアクセスすることができます。

❻ まとめ

大容量の分散環境において、非構造化データを高速かつ低コストで管理・運用するためには、ファイル管理の徹底的な自動化が不可欠となります。IBMは、その方向性を見据え、今後もさらなる機能や製品の開発を促進していきます。

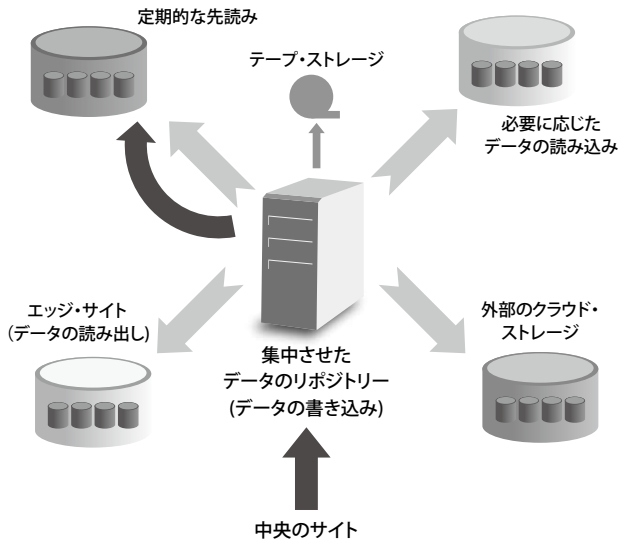


図6. WANキャッシングの利用例(1)

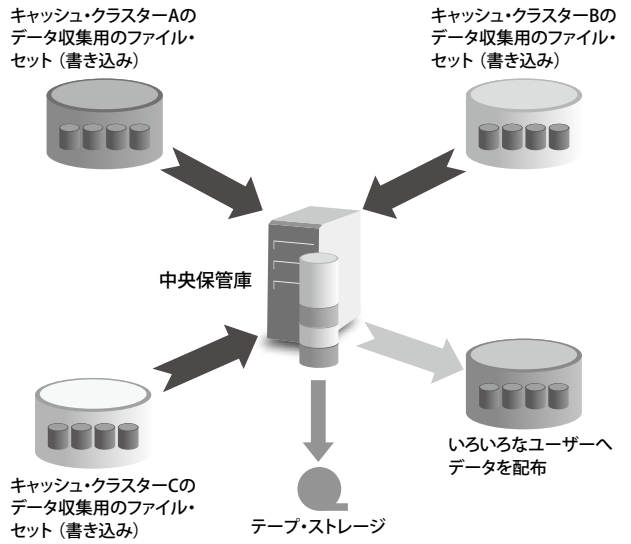


図7. WANキャッシングの利用例(2)

[参考文献]

[1] Frank B. Schmuck and Roger L. Haskin: GPFS: A Shared Disk File System for Large Computing Clusters, Proceedings of the Conference on File and Storage Technologies, p.231-244, (2002.1.28-30).

[2] Roger L. Haskin: Tiger Shark, a scalable file system for multimedia, IBM Journal of Research and Development, Volume 42, Number 2, pp. 185-197, (1998.3).

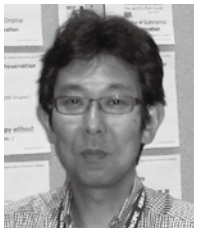
[3] Srini Chari: Optimizing for Higher Productivity and Performance: How IBM' s General Parallel File System (GPFS) Bridges Growing Islands of Enterprise Storage and Data April, (2011).

[4] S. Oehme, J. Deicke, J.-P. Akelbein, R. Sahlberg, A. Tridgell and R. L. Haskin: IBM Scale out File Services: Reinventing network-attached storage, IBM Journal of Research and Development, Volume 52, Number 4/5, (2008).

[5] 三好, 萩原, 松井, 岩崎:SONAS 非同期コピーのパフォーマンス改善 IBM PROVISION No.70, http://www.ibm.com/ibm/jp/provision/no70/ibm_paper2.html (2011-Summer).

[6] IBM, Automated file management with IBM Active Cloud Engine, IBM Systems and Technology Solution Brief 2011.

[7] M. Eshel, R. Haskin, D. Hildebrand, M. Naik, F. Schmuck and R. Tewari: "Panache: A Parallel File System Cache for Global File Access," in Proceedings of the Eighth USENIX Conference on File and Storage Technologies (FAST '10) , San Jose, CA, (2010.2).



日本アイ・ビー・エム株式会社
システム製品事業
テクニカル・セールス
システム・ソフトウェア・ソリューション担当

緒方 正暢 **Masanobu Ogata**

[プロフィール]
1986年、日本IBM入社。東京基礎研究所にて分散リアルタイム・システム、分散マルチメディア・システムのためのシステム・ソフトウェアの研究に従事。ストレージ製品のテクニカル・セールスを担当後、現在、IBM SONASなどのシステム・ソフトウェア製品のテクニカル・セールスを担当。