解 説 3

ビッグデータ活用を支えるシステム・テクノロジー – スケーラビリティーと最適化 –

「ビッグデータ」とは、文字通り膨大なデータの集まりです。さまざまな情報源から得られるデータを分析して企業活動や社会システムの運営での意思決定を支援するシステム、いわゆるビジネス・インテリジェンス (BI) の分析対象となる巨大なデータを指しています。

このビッグデータの活用を支えるITインフラストラクチャーには、高速なデータ処理能力と大容量データの保存・管理能力、そして刻々と変化するアプリケーション要求に迅速に対応できる柔軟性が求められています。さらに、事業継続性やコンプライアンスなどのビジネス・プロセスにおける要件と整合させて実現するには、全体として最適化・統合化できるアーキテクチャーの選択が重要になります。

本稿では、ビッグデータ活用においてシステムを構成するサーバーやストレージが直面する課題を挙げるとともに、これに対応したIBM Systemsのテクノロジーやソリューションを解説します。

Article 3

System technology for Big Data
- Workload Scalability and Efficiency -

Today, huge amounts of information are born-digital or are converted from non-digital forms, and the amount of data worldwide is getting increasingly massive as a result. Turning such unstructured information into business insights is critical for business optimization and competitive advantage.

The IT infrastructure needed to handle such “Big data” requires scalability in processing speed and storage capacity, and also requires flexibility in new applications deployment. It also needs to be a fully integrated system that requires only simplified management, in order to meet requirements including business continuity, security and compliance.

In this article, we will list the challenges faced by server and storage systems managing Big data, and will describe how the IBM’s vision of Smarter Computing accommodates these requirements.

❶ はじめに

1.1 増え続けるデジタル・データ

インターネットや無線通信網の発達によってどこでも手軽にネットワークに接続できるようになり、PCやスマートフォンを介した情報検索やオンライン商取引、掲示板やブログなどのソーシャル・メディアの利用などは、もはや社会生活に不可欠なものになっています。また、写真や動画を高解像度で撮影できるカメラや、気象変化、交通量などの状況をモニターするセンサー技術、荷物管理用のICタグなど、デジタル・データを刻々と生成するデバイスも身の回りに増えています。このようなネットワーク技術の発展、ネットワーク・サービスの普及、そして多くのデバイスからのデータの蓄積の結果、データ量は爆発的に増えています。調査会社の発表によれば、1年間に世界中で新たに生成あるいは複製されたデジタル・データの総量はこの5年間で9倍に増えており、これは年率平均50%を超える増加に当たります［1］。このような新たに生成されるデータの多くは、これまでリレーショナル・データベースなどで体系立てて管理されてきたような構造化データではなく、文書ファイルや電子メールのような多様性に富んだ非構造化データです。

この増大するデータを単に後の参照のために記録・保管し続けるのではなく、データの内容を総合的に分析することによって潜在的な価値を抽出し、得られた知見を積極的にビジネスに生かす動きがあります。例えば、ブログやTwitterなどの利用者の投稿から特定の商品セグメントにおける消費者要求や市場トレンドを読み取り、それを製品企画や販売計画に反映させる試みが挙げられます。また、交通やエネルギー供給などの社会基盤に関連する情報を一元的に集めて解析することにより社会全体での最適化を行い、最終的にCO_2削減を実現させる取り組みも一例です。

1.2 データの価値と鮮度

あるデータが最初に作り出された時刻を起源としてその情報が利用・活用されるまでの時間を考えると、大きく分けて2つのグループのデータがあると考えられます。第1のグループは、刻々と変化する状況を監視し続け、その変化から何かを発見しようとするものです。このグループには金融トレーディングにおける売買注文状況や金融ニュース、気

象情報や交通情報、工場生産ラインの監視センサーなどの情報が含まれ、これらは連続したデータ・ストリームとしてアプリケーションに取り込まれてリアルタイムに処理されていきます。このようなデータの多くは恒久的に蓄える必要性が低いトランジェントなものなのでデータの賞味期間が短いといえます。

第 2 のグループは、逆に継続的に蓄積されることで価値が高まってくるデータで、例えばインターネットの検索サービスや、自然環境変化や事故、疾病などの因果関係の追跡調査のように、過去の情報を大量に収容していることによって新たな利用価値を生むものです。さらに、検索サービスの例でサービス利用者の検索行動の傾向や検索結果に対する満足度などによって情報のランキングを行えば、情報の価値は時間と共に増大していきます。このグループのデータの価値は存続期間が長く、大量のデータの管理方法やその中から最も有用なデータに最短でアクセスするアルゴリズムが重要になります（**図 1**）。

1.3 ビッグデータを扱うシステムのスケーラビリティー

スケーラビリティーとは、コンピューター・システムを設計する際の前提となるデータ処理量や入出力量などのシステム負荷が増大した場合に、設計を大幅に変更することなく透過的に拡張させることができる能力です。ビッグデータを扱うシステムを設計するには、データの増大に伴ってどのようなスケーラビリティー要件を持つのかを把握することが必要となります。そして、技術的にも経済的にも持続可能なスケーラビリティーを実現するには、サーバーやストレージ、ネットワーク、アプリケーションを含むシステム全体での最適化を行い、いかにコストと要件をバランスさせるかが鍵となります。

以降では、スケーラビリティー要件の尺度として 3 つの軸を取り上げます。1 番目はデータをアプリケーションで処理するために増大する「パフォーマンス」に対する要件、2 番目は取り扱うデータがストレージやメモリー上で占める総量の増加に対応する「キャパシティー」に対する要件、そしてデータをそのライフサイクルを通じて保全し管理・維持し続けるための「保管期間」に対する要件です。この 3 つの軸に従って、対象となるデータが現在どのような特性を持つのか、また将来においてどのように変化していくのかを予測しておくことが大切です。

❷ パフォーマンスのスケーラビリティー

2.1 ハードウェアの階層化と最適配置

コンピューターのハードウェア構成を単純化して模式図にすると、**図 2** のようにプロセッサーを頂点としたピラミッドを描き、下位部分をメモリーやストレージなどの幾つかの層（レイヤー）に分けた図を作ることができます。この図では、プロセッサーに近い上位層ほどプロセッサーまでのデータ移動時間が短く、最も遠い最下層からのデータ移動には時間がかかることを表しています。さらに、図中でそれぞれの層が占める面積が容量や密度をも表しており、上位層は高速で少量の記憶領域、下位層は低速ながら大容量の記憶領域ととらえることができます。この階層化は、コストと容量あるいはコストとスピードという、相反する要求を満たすために広くコンピューターの中で取り入れられている技法です。

ストレージ・システムにおいても複数のレイヤー（あるいは Tier) が採用されており、これを実現する仕組みを特に階層ストレージ管理（以下、HSM）と呼びます。ストレージ・システムの階層化は、従来は磁気記録装置であるハードディスク・ドライブ（以下、HDD）とテープ・メディアの組み合わせで実現されていましたが、回転数の異なる HDD

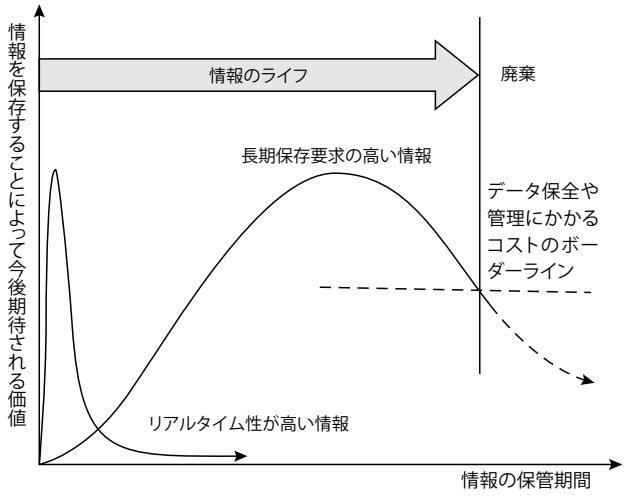

図1. 情報のライフサイクル

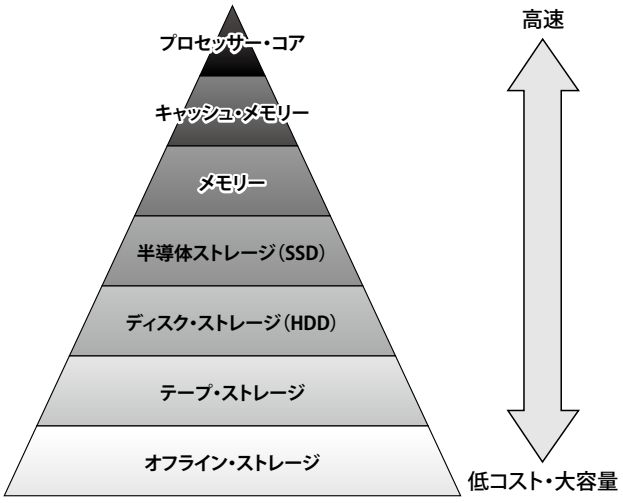

図2. コンピューターとストレージのレイヤー

を組み合わせたり、さらに近年ではフラッシュ・メモリー技術を利用した半導体ストレージ（Solid State Drive：以下、SSD）を HSM に追加する事例が増えてきています。参照頻度が高いデータをより上位層に蓄えることで見かけのパフォーマンスを向上させ、さらにコストの最適化を行うことができます。

IBM のブロック・ストレージ製品が持つ Easy Tier 機能では、ストレージ内部で参照頻度の高い領域（すなわちホットスポット）を SSD に割り当てられるように、SSD と HDD の 2 つの Tier 間でデータを自動的に入れ替えます。この入れ替えはストレージ・システム外部からは透過的に行われるため、このストレージを使用するアプリケーションに影響を与えることなくストレージ全体の性能効率を高めることができます。この検出はデータへのアクセス頻度を統計的に収集することによって行われ、ヒートマップという情報でホット、ウォーム、コールドというクラス分けを行います。このヒートマップによって現在のワークロードの局所性を可視化できるので、SSD を追加することによって得られる投資効果を事前に知ることができます。得られる性能向上の度合いはワークロードに依存しますが、全容量の 2% に相当するストレージを SSD に変更するだけで、再配置の結果最大 3 倍の性能向上が見られたワークロードもあります。この Easy Tier 機能は、現在 IBM System Storage DS8000 シリーズ、IBM System Storage SAN ボリューム・コントローラー、IBM Storwize V7000 に搭載されています（**図 3**）［2］。

このように、サーバー・プロセッサーにおいてもストレージにおいても、パフォーマンスやコスト、容量の異なる 2 つ以上のテクノロジーを組み合わせることによって全体の最適化を行っており、その鍵となっているのがデータを最適な場所に配置するアルゴリズムであることが分かります。

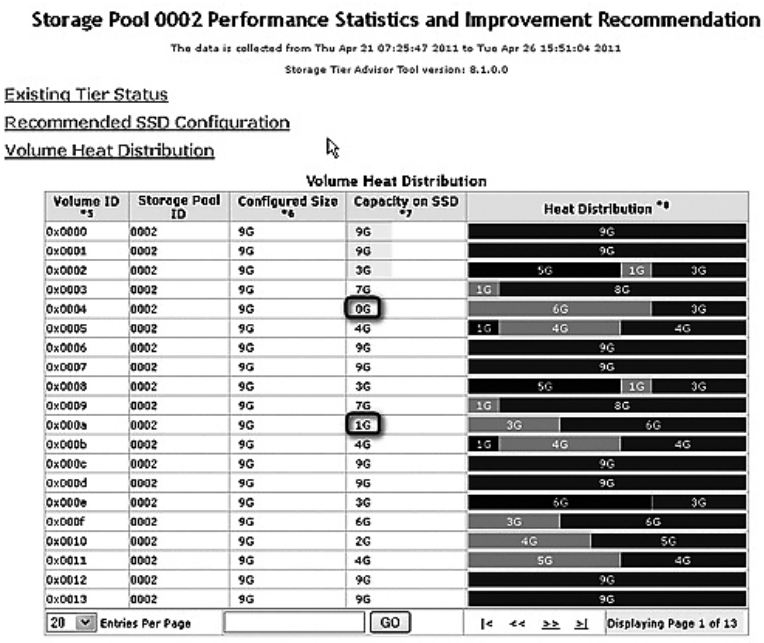

図3. Easy Tierのアドバイザー・ツール画面例

2.2 スケールアウト・アプローチとハイブリッド・コンピューティング

一般にシステムの高速化手法にはスケールアップ型とスケールアウト型の 2 つの手法があります。スケールアップ型は、システムの構成部品をより高速なものに置き換えることによってシステム全体のパフォーマンスを上げる手法です。これに対して、スケールアウト型はシステムのサーバー単体を速くするのではなく、複数台のサーバー（ノード）を相互接続してあたかも 1 台のサーバーであるかのように仮想化することによって処理能力を向上させる手法です。スケールアウト型は状況に応じてノード数を増やすことによってさらなる性能向上を実現できることが利点であり、ノード間で協調して動作させるためのオーバーヘッドをいかに最小化してリニアな性能向上を生むかが技術的な課題となります。IBM がこれまでにアプリケーションごとのシステム要件や特性の調査を行った結果では、アプリケーションは 4 種類のワークロード特性に分類できると考えてられています［3］。ビッグデータの情報解析は 4 つのタイプのうち「アナリティクス」型に当たり、高い計算処理能力、高速・広帯域のメモリーとネットワークやストレージの入出力性能を必要とするため、高速なプロセッサーを搭載したスケールアウト型のシステムが適していると考えられています。

さらに、データ解析の流れをプロファイリングしてみると、高頻度で行われる処理、あるいはプロセッサーや入出力の負荷が高い処理が存在することが分かります。このようなホットスポットの処理を汎用プロセッサーから専用のアクセラレーター・エンジンにオフロードすることによってシステム全体の効率を上げる手法もあります。このようなアクセラレーターには、浮動小数点演算処理、XML 処理、データ圧縮、暗号化などの処理を行うものが用意されています。そのため、システム全体の運用効率を最適にするには、ワークロードをスケールアウトのアプローチで複数のサーバーに分散し、個々のサーバーでの処理に適切なプロセッサーとアクセラレーターを組み合わせて適用することが有効であると考えられます。

スケールアウト・アプローチの例としては、新世代のメインフレーム・サーバーである IBM zEnterprise 196（以下、z196）や同 114（以下、z114）を挙げることができます。z196/z114 は、BladeCenter Extension 拡張フレーム（以下、zBX）を介して POWER7 や x86 プロセッサーのブレード・コンピューターを

追加できるハイブリッド型のサーバーです。IBM Smart Analytics Optimizer（以下、SAO）は zBX に追加できるアクセラレーター・エンジンの一種で、データベースのクエリー処理の一部をオフロード処理します。この SAO は搭載数を増やすことでスケールアウトのメリットが得られます。さらにソフトウェア（IBM DB2 Analytics Accelerator for z/OS）の追加により、外部の IBM Netezza アプライアンス（以下、Netezza）の高速データベース検索エンジンを利用することも可能です。SAO も Netezza も、データの絞り込みを分散化・並列化することでデータベース検索を高速化しています。

このように、z196/z114 では、3 種類の異なる特性を持つプロセッサー群をワークロードによって使い分けて処理を最適化するとともに、さまざまなハードウェア・リソースを体系的に統合しシステム管理を一元化することによって、運用管理の課題を解決しています。共通基盤が整備されることにより、システム管理者はアプリケーションの迅速な展開により注力できるようになります［4］。

2.3 ストレージの仮想化とスケールアウト

HDD のパフォーマンスは、回転数の高速化と記録トラックの長さ当たりの記録密度（線記録密度）の高密度化によりある程度の向上を続けているものの、回転する磁気ディスク表面に沿ってヘッドを移動させることでデータへアクセスするという物理的な基本構造に制約を受けています。そのため、CPU のスピードと HDD のスピードのパフォーマンス・ギャップは年々広がる一方です。また、HDD 1台当たりの記憶容量が倍増してもその HDD のスループットはさほど変化しないことから、容量の条件を同一にして比較すると HDD ヘッド数の減少によってスループットは半減していることになります。すなわち、データ量ごとのアクセス要求が一定であると仮定すると、従来 2 台の HDD で処理をしていたワークロードは新世代の HDD 1台で置き換えてしまうとパフォーマンス面では要件を満たさなくなってしまう可能性があります。また HDD 1台当たりの容量の拡大に伴い全データを読み出す処理時間は長くなり、レプリケーションや RAID のリビルド、ヘルスチェックなどの処理の長時間化として影響が現れています。今後は、HDD の小型化によってストレージ・システムの設置面積当たりの HDD の台数を増やし、さらに仮想化層での HDD の並列アクセスによって高速化することが必須となってきます。

そして、この並列性を高めるには、同一時刻でのアクセス頻度の高いデータが特定の HDD に集中しないように、データを事前に分散配置させておく必要があります。

IBM XIV Storage System（以下、XIV）はブロック・アクセスの仮想化を行うストレージ・システムで、その仮想化においてデータの分散配置を実現しています。XIV では書き込まれたデータを 1MB という細かなパーティションに分割し、搭載されている HDD 全数（1 ラック最大 180 台）にパーティションを分散させて記憶しています。1 つのパーティションを 2 台の HDD に多重記録する際に、パーティションの保存場所を独自のアルゴリズムによって擬似ランダムに選択することが XIV の特長で、これによりサーバーからの I/O ワークロードにかかわらず HDD へのアクセス量は平準化することができます。ストレージ機器内部で自動的にホットスポットの発生を回避できるようにデータの並びが決定されるため、従来手動で RAID 構成を設計することで対処してきたストレージ管理者の負担を軽くすることが可能になり、また不意な I/O ワークロードの変動へも対応できます［5］。

❸ キャパシティーのスケーラビリティー

3.1 ストレージ統合と量の抑制

現在入手できる HDD 1台やテープ 1 巻の容量は最大でも数 TB であり、これ以上の容量のストレージ空間が必要な場合は、RAID システムなどの仮想化によってストレージ・システムを統合することになります。この統合化の目的は、パフォーマンスのスケールアウトに加えて、運用管理の一元化と無駄なスペースをなくすことにあります。ストレージ・システムが統合されていない状況、いわゆるアイランド化した状況では、1 つのストレージ・システムの不足分をほかのストレージ・システムの空き容量で補うことができないため結果的に全体のコストを上げてしまう可能性があります。ストレージを共通のリソース・プールとして統合し、この中から必要量に応じて割り当てると空間全体で有効に活用することができます。さらに、ブロック・ストレージには ThinProvisioning 機能が備わっており、ユーザーへのストレージ・スペースの初期割り当てを仮想化し、実際には利用されていない空間をストレージ・プールの一部として還元することによってスペースを有効活用することができます。また統合化において HSM のアプローチを取り入れてビット単価が低いストレージへの移行を進めることによって、コストと容量のバランスを取りながら増え続けるデータを収容することができます。

一方、増え続けるデータの中から無駄な領域を探して、ストレージ容量の増加を抑止することも必要です。データ重複排除（De-duplication）と呼ばれるストレージ機能では、データの中で重複している同一データを見つけ出し、2 番目以降のデータを最初のデータへのポインターで置き換えることによってストレージ容量を抑止することができます。

同じようにデータ量を小さくする技術としては、データ圧縮（Compression）技術があります。一般にデータ圧縮には写真データで使われる JPEG や動画の MPEG フォーマッ

トのように人間の目では気付きにくい程度にデータを劣化させて非可逆圧縮をするものもありますが、ストレージ・システムの中で自動的に圧縮する場合には可逆圧縮のアルゴリズムが用いられます。データを圧縮することによってストレージ・システムのキャパシティーに対する要求が下がり、さらにデータを転送するネットワークに対する負荷も軽減できるために、災害対策の通信コストを含めた全体のコストの削減にも貢献します。また圧縮・伸長をメモリー上でオンザフライで実行するとストレージへのアクセス量が減ることになり、そのデータ縮小分だけI/Oパフォーマンスを向上させることができます。

3.2 データ管理のスケーラビリティー

アクセスの対象となるデータを特定して読み書きを実行するには、ブロック・ストレージはブロック番号を用い、ファイル・ストレージはディレクトリーやファイルの名前を用います。ブロック・ストレージであってもブロック番号をそのまま使うのではなく、サーバー側でファイル・システムを構成することが一般的です。ファイル・システムのストレージ空間はディレクトリーの中にディレクトリーをネストすることができるツリー状の構造をしており、ユーザーはこの構造に認識しやすい名前を付けることによって情報を整理・分類しています。例えば、個人の写真データであれば、フォルダーが年、月、日のような3層構造で、特定の日付の写真を簡単に照会できるようにします。企業内の大量のデータをファイルの集まりとして扱う場合には、データを一元的に扱えるようにファイル・システムの名前空間は単一（Global Namespace）であることが望ましく、またこのディレクトリー・ツリーで扱うことができるファイルの総数が増えるため、このアドレス空間もスケーラブルである必要があります。

IBM Scale Out Network Attached Storage（SONAS）で採用されているIBM General Parallel File System（GPFS：本誌58ページ以下：解説④参照）は、1兆個のファイルの管理を目標とするプロジェクトでも使われています。このように大量のファイルが存在する場合に特定の条件に合致するファイルを見つけ出すためには、メタデータ（管理情報）の高速検索が必要であり、GPFSはそのための高速なMetadata Scan Engineが備わっています。これによりGPFSはシングル・システム上の100億個のファイルをわずか43分でスキャンするという記録を達成し、将来の大規模ストレージ・システムとしての有用性を実証しました[6][7][8]。この記録は、3時間で10億個のファイルをスキャンするというこれまでの記録の37倍の処理能力に相当します。この実証検証ではHDDとSSDを利用したハイブリッド型のストレージが採用され、ホットスポットとなるメタデータを6.8テラバイトのSSDに割り当てることにより記録的なスキャン速度が実現されました。この検証では、**図4**のように、10台のGPFSクライアントによってメタデータの並列処理が行われて、そのうちの4台にSSDが接続されていました。ここでもスケールアウトのアプローチと、データの最適配置が活用されてハイパフォーマンスを実現しています

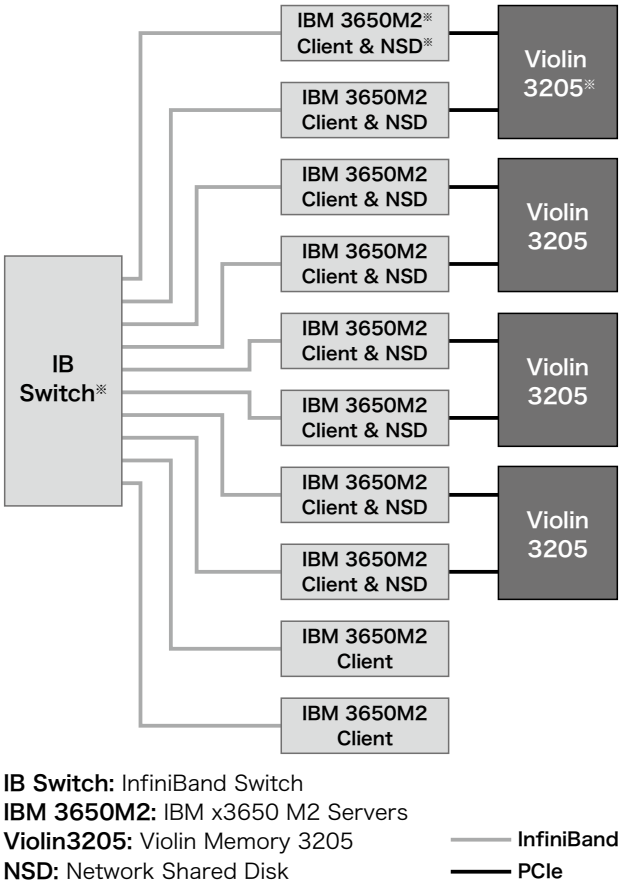

図4. 大量ファイルの検索性能の検証システム

❹ 保管期間のスケーラビリティー

データは生成、利用、廃棄のライフサイクルをたどり、ストレージ・システムはこのライフサイクルを通してデータを確実に維持・保管していくことが必要です。今後は、これまで紙の文書やマイクロフィルムが担っていた長期保管の役割にもITシステムが入っていくと考えられています。この場合、データの寿命はそれを保管するストレージ・システムやそのデータを参照していたアプリケーションの寿命よりも長くなり、データをアクセス可能な状態のまま確実にそして安価にマイグレーションしていく方法が必要となってきます。

ここで、アクセス可能な状態とは、以下のように幾つかの段階があります。

(1) データがストレージ上でどのように整理・分類されているのかが判断でき、データ同士の関連性が分かる。
(2) 個別のデータをビット列としてストレージのメディアから読み出せる。
(3) 1つのデータの中身の構造を解釈して、情報を理解できる。

サーバーの介在なしに機械的に転写・変換していくためには、(1)と(3)のデータ構造の表現方式が長期的に利用可能である必要があり、標準化された形式を採用することで手動でのデータ移行に掛かるコストを下げることができます。現在、データ構造やファイルのパッケージは、データ移行やデータ交換をスムーズに行うために、業界ごとに特有の構造が標準化されている場合があり、そうでない場合でも今後標準化が進んでいくものと思われます。

また、ストレージの磁気媒体そのものの長期保存性を活用して、転写を繰り返さずにメディアだけを永続的引き継ぐ方法もあります。IBMが開発したLinear Tape File System（LTFS）は非圧縮記録時1.5TBの大容量LTOテープをUSBメモリーなどと同じような操作感で扱えるようにしたもので、自己記述型のフォーマットを採用したテープ記録方式です。テープ・メディアがもともと備える長期保存特性や可搬性、経済性を生かし、次世代の放送業界のストレージ媒体として映像遺産の継承とワークフローのデジタル化を推進することが期待されています。

❺ まとめ

ビッグデータを有効に活用するには、「パフォーマンス」「キャパシティー」「保管期間」のスケーラブルな性能要求に対して柔軟に追従できるシステムが不可欠であり、IBMは3つの要件を幅広い製品ポートフォリオとテクノロジーでカバーしています。システム全体の運用効率を最適化するには、アプリケーションごとにシステムをアイランド化することなく、全体を俯瞰（ふかん）的に把握してアーキテクチャーを選択することが重要です。

IBMが提唱する“Smarter Computing”という考え方は、SOA（Service Oriented Architecture）のように統合化のための体系的なアプローチによってITシステムの効率化を目指したものです。この統合されたシステム環境によって新たなITサービスを迅速に提供する基盤が整えられ、ビッグデータをより有効にビジネス戦略に生かすことが可能になってくるでしょう。

［参考文献］

[1] IDC: “Digital Universe Study.”, http://www.emc.com/leadership/programs/digital-universe.htm（2011）.
[2] IBM: IBM Redbooks “IBM System Storage DS8000 Easy Tier,” IBM,（2011）.
[3] 佐貫俊幸：“ 解説記事1：IBM Systems の新しい潮流,” PROVISION NO.67, http://www.ibm.com/ibm/jp/provision/no67/pdf/67_article1.pdf（2010-Fall）.
[4] IBM Redbooks “IBM zEnterprise System Technical Introduction,” IBM,（2011）.
[5] IBM: IBM Redbooks “IBM XIV Storage System: Architecture, Implementation, and Usage,” IBM,（2011）.
[6] IBM: IBM Redbooks “IBM Scale Out Network Attached Storage: Architecture, Planning, and Implementation Basics,” IBM,（2011）.
[7] IBM：プレス発表「ビッグデータ・アプリケーション向けの画期的なストレージ処理能力を実証」, IBM, http://www.ibm.com/jp/press/2011/07/2701.html（2011）.
[8] Richard Freitas, Joseph Slember: Wayne Sawdon, Lawrence Chiu, “GPFS Scans 10 Billion Files in 43 Minutes,” IBM,（2011）.

日本アイ・ビー・エム株式会社
開発製造　大和システム開発研究所
シニア開発マネージャー

石本 健志 **Takeshi Ishimoto**

［プロフィール］
1987年、日本IBM入社。グローバル市場向けIBMハードウェア製品の設計・開発に従事。PC・ワークステーションの開発を経て2000年よりストレージ製品の開発を携わり、これまでにRAID装置、メインフレーム用ハイエンド・ストレージ製品、ファイル・システム製品の開発プロジェクトに参画。2009年よりLTFSの開発を担当。