解 説 2

ストレージの性能を支える技術

― お客様をお待たせしないために ―

Web の検索サイトで文字列を入力すると、一瞬で結果が返ってきます。また、銀行が提供するネットバンキングは、自宅での金融商品購入や決済処理を即座に実施してくれます。こうしたサービスを提供するシステムで取り扱うデータ量や処理内容がたとえ増えたとしても、お客様の感覚的な「待てる時間」は決して長くはなりません。むしろ待てる時間内に処理を終えるためにシステムを設計・増強することが必要な時代になってきています。ここでは、システム全体の処理速度に大きな影響を持つストレージ・システムの性能に関する基本事項を押さえ、お客様の要求に応えるための技術動向を解説します。

❶ はじめに

今、皆さんの前にある電話が鳴ったとします。あなたは何回目の呼び出し音で電話に出ることができますか。

大和ハウス工業株式会社の樋口武男氏は、日本経済新聞に連載されている「私の履歴書」の中で、経営哲学として凡事徹底を旨とし、電話には 1 回の呼び出し音で出るよう社員に指導していたと記しています［1］。

筆者が日本アイ・ビー・エム株式会社に入社した 1997 年当時、新人ビデオ研修で社長自らが同じことを伝えていたのを思い出します。加えて 3 回以上の呼び出し音で待たせてしまった場合は「お待たせしました」と添えるようにと述べていました。

この 2 つの事例に共通するのは、社員の誰もが共通に理解できる電話対応を例に、業務全般におけるお客様に対するべき姿勢を示したことと同時に、どれだけ良い製品、良いサービスを持っていても、お客様窓口の対応次第でそれを生かしきれない場合があることが示されています。

同じように、システムにおける応答性能も、出力される結果を人間が待っている以上、待ち時間への配慮を避けて通ることはできません。IT 予算圧縮の圧力が強い昨今であっても、人の感覚は技術の進歩に伴って「待ち時間は少なくなるだろう」と感覚的に期待しています。その期待に応えるためのアプローチを、ストレージの観点で見てみましょう。

❷ 応答時間

語学を学習する上で、辞書は欠かせない存在です。単語を調べる手順は次の通りです。

(1) 辞書を本棚から取り出す。
(2) 単語が載っているページまで検索する。
(3) 単語の意味を読む。

同様に、ストレージ製品や PC に搭載されているハードディスク・ドライブ（以下、HDD）も次の動作をします（**図 1**）。

(1) ディスク上のデータが書かれている位置まで読み込みヘッドを動かす（シーク）。
(2) ディスク上のデータが書かれている場所がヘッドの下に来るのを待つ（サーチ）。
(3) データを読み込む（転送）。

この 3 つの動作を行うための時間が HDD に対する 1 回の I/O（Input/Output）にかかる時間となります。

HDD ベンダーは HDD の平均シーク時間および平均サーチ時間を Web サイトで公開しています。転送時間は非常に短いため、1 回の I/O にかかる平均時間は平均シーク時間と平均サーチ時間の和で示すことが多く、

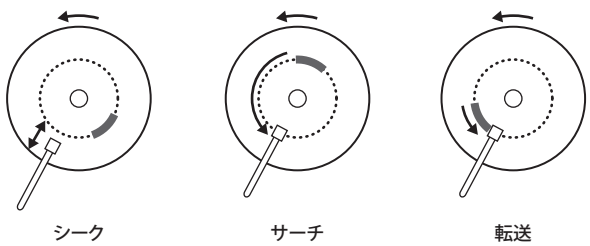


図 1. HDD の動作

例えば前者が 2.6［ms（ミリ秒）］、後者を 2.0［ms］とした場合、その和の 4.6［ms］が平均 I/O 時間となります。

ただし、性能の指標としては、1 秒間に何回 I/O ができるか（I/O per second：以下、IOPS）として示しますので、1 回の I/O にかかる時間を平均 4.6［ms］とした場合、1,000［ms］（1 秒間）÷4.6［ms］（平均 I/O 時間）=217［IOPS］と示します。

これは HDD の持つ性能であり、容量に依存しません。例えば 1 台の HDD に 1 枚 100GB のディスクが 1 枚入っていれば 100GB の HDD ですし、2 枚入っていれば 200GB の HDD となります。I/O の処理としては 1 枚目のディスクに対しても 2 枚目のディスクに対しても同じ軸の上に回転して、読み込みヘッドは同じように動きますから IOPS としては同じになります。

では語学の学習に戻り、1 つの単語を 5 カ国語で調べたいという場合、1 人で調べていると（1）から（3）の順番に処理しますので、1 カ国語の場合に比べて 5 倍の時間がかかります。しかし、5 人で作業を分担し、1 人 1 カ国語を担当すれば、同時に 5 カ国語の処理をすることができるので、調べる時間は 1 カ国語の場合と変わりません。

辞書の検索と同様に HDD1 台で 5 個の I/O を処理すると 5 回の I/O が必要となり、時間は 5 倍の 4.6［ms］×5=23［ms］が必要になります。しかし、5 個の HDD を用意してそれぞれに分散処理をさせた場合、1 回の I/O 時間で 5 倍の処理をすることができます。このため、指標値は 217［IOPS］×5［個］=1,085［IOPS］となります。

このように複数の HDD を並列処理させることはストライピングと呼ばれ、性能を高めるための一般的な手法です（**図 2**）。

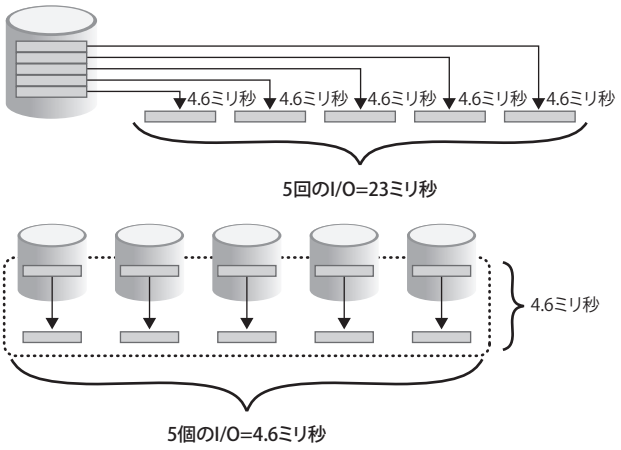


図 2. HDD の処理時間

❸ キャッシュ・メモリー

語学の学習が進むにつれて、何度も辞書を引くとその単語の意味は頭の中に記憶として残ります。記憶は辞書を引く物理的な動作が不要になるだけでなく、単語の意味を解釈する過程が済んでいるため、理解が非常に高速になります。つまり語学の習熟度合いが上がるにつれて辞書を引く頻度が減り、文章を高速に理解することができるようになります。

同様に、コンピューター・システムにおいても何度も参照するデータは HDD ではなく、もっと高速にアクセスできるところに置くことで性能を高められます。I/O 時間は HDD がミリ秒［ms］単位であるのに対し、半導体はナノ秒［ns］と 100 万倍も高速です。従って、この高速な半導体メモリーを記憶域（キャッシュ）として使うことでシステムの性能を上げることができるのです。しかしながら、半導体メモリーは記憶のために通電が必要で、かつ高価なため、HDD に比べて記憶単位当たりのコストが高くなります。コスト面を考慮すると少量の半導体メモリーを効率よく使用することが必要でしょう。

仮に 1 個の HDD の応答時間を 4.6［ms］、キャッシュから読む場合の応答時間を 0.1［ms］として、1,000 回の I/O の平均応答時間を検証してみます。その際、繰り返しの読み込みによってキャッシュ上にデータが載っている場合（キャッシュヒット）を 500 回、データが載っておらず HDD にデータを読みに行った場合（キャッシュミス）を 500 回とすると、キャッシュの利用効率は 50% となり、平均応答時間は 4.6［ms］×500［回］+ 0.1［ms］× 500［回］を 1,000［回］で割って 2.3［ms］となります。このようにキャッシュ・メモリーの効果により、HDD だけ

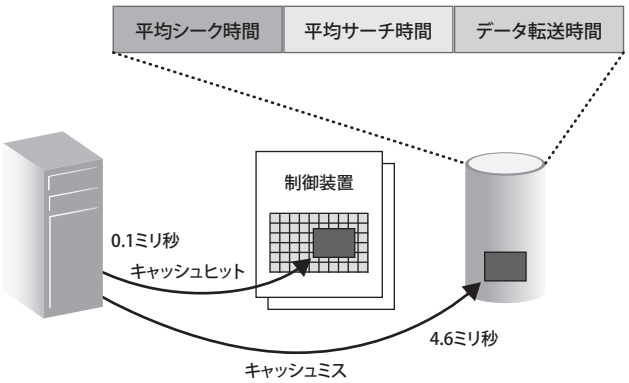


図 3. キャッシュ・メモリーの効果

を利用したときよりも高速に処理したかのように見せることができます（**図 3**）。

このとき、1 個の HDD 当たりの見かけ数値は 1,000［ms］÷2.3［ms］ = 434［IOPS］、5 個 の HDD を並べた場合は 434［IOPS］×5［個］= 2,170［IOPS］となり、単位時間に処理できる I/O の数はキャッシュがない場合に比べて倍の処理が可能になったように見せることができます。

この指標は、オンライン・システムなどで利用するストレージ・システムの性能を示す数値として参考となります。

❹ ストレージ・システム

複数の HDD を束ねてサーバーからの I/O を分散させたり、参照頻度の高いデータを半導体メモリーに載せておいたりすることを制御しているのがストレージ・コントローラーです。多くの場合、このコントローラーをはじめとしてストレージ内部で構成される部品は耐障害性を向上させるために多重化され、冗長性を持たせることで、データ損失を防ぐようになっています（**図 4**）。

HDD やストレージ・コントローラーを含むハードウエア一式を箱に収めたものが、ストレージ・システムとして各社から発売されています。部品としての HDD を生産するベンダーはすでに数社となり、IBM をはじめとするストレージ・システム・ベンダーはほぼ同じ HDD を部品として利用しています。しかしながら、ストレージ・システムを制御するコントローラーのプログラムを工夫することで、各社は特徴的な製品を提供しています。

例えば IBM のハイエンド・ディスク装置である IBM System Storage DS8000（以下、System Storage DS8000）は、IBM の研究所で開発した Sequential Prefetching in Adaptive Replacement Cache（SARC）と呼ぶ独自のアルゴリズムを搭載しています［2］。これは、いつどのデータがキャッシュに読み込まれ、キャッシュがいっぱいになったときにどのデータを破棄するかを学習し、サーバーからのさまざまな I/O 要求の中で効果があると判断したときに適用することで、キャッシュ・メモリーをできるだけ効果的に使い、I/O 要求をより高速に処理するためのアルゴリズムです。これが業界トップクラスの性能を発揮する要素の 1 つとなっています。

また、IBM の特徴的なディスク装置である IBM XIV Storage System（以下、XIV）は、ストレージ・システムに利用されている高速かつ耐久性の高い HDD ではなく、個人用の PC に使われるような、比較的低速だが安価で大容量の HDD を大量に搭載することで容量単価および容量当たりの電力消費を抑えています。XIV は「HDD は低速である」こと、「HDD 障害は発生し得るもの」という前提に設計されており、サーバーからのデータをバラバラに分割し、その断片をコピーして二重化した上で搭載されている別々のディスク上に分散配置することで HDD の速度をカバーし、かつ障害発生時の復旧を迅速化します。

このように、同じ部品としての HDD を利用していても、その性能や信頼性はストレージ製品ごとに異なっており、それが製品の差異につながっています。

❺ ボトルネック

朝の通勤時間帯の駅の混雑は、人の流れがある一定時間に集中することで発生します。例えばある駅で、改札とホームは混雑時の利用者数を想定して十分に確保していても、構造上の問題でホームと改札をつなぐ階段が上下方向にそれぞれ1列ずつしか幅が確保できない場合、混雑が始まると人の流れが悪くなります。このように混雑しているときに狭くて通り抜けるために待ちが発生してしまう場所をワインやビールのビンのように注ぎ口が

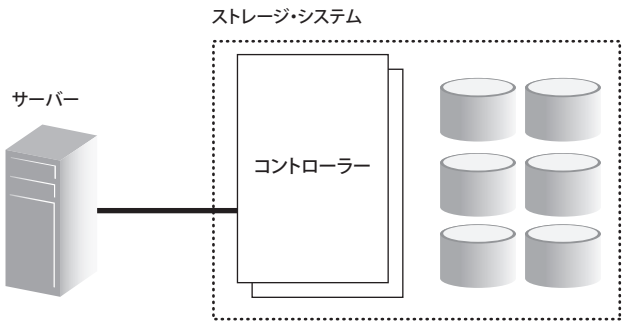


図 4. ストレージ・システム

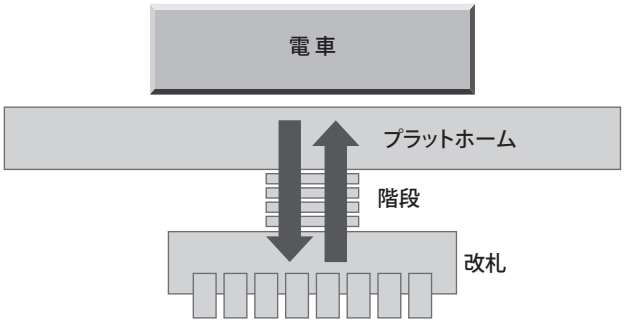


図 5. 駅の構造

細くなっている様子から「ボトルネック」と称し、その点が全体の速度を落とす原因となります（**図 5**）。

コンピューター・システム全体も同様にデータが流れる経路全体を見渡して、どこにボトルネックの要素があるかを把握し、要求される処理能力に見合う十分な幅（帯域）が確保されるよう設計することが重要です。流れるデータ量がボトルネックを超えると、すべての I/O が平等に遅くなります［3］。

❻ ソリッド・ステート・ドライブ（Solid State Drive：以下、SSD）

SSD とは、USB メモリーのような電源供給がなくとも記憶を維持できるフラッシュ・メモリーを用いた半導体ドライブで、HDD のようなシーク時間やサーチ時間がないため、非常に高速に動作します。これを HDD と互換性があるパッケージに詰め、従来の HDD と同様に扱えるようにすることで、ストレージ・システムの部品として利用できます。

SSD の利点と欠点をそれぞれ 3 つ挙げると**表 1** の特徴を持ちます。

表 1. SSD の利点と欠点

利点	欠点
高速	高価
省電力・軽量	小容量
データ保持に電力不要	書き換え回数制限

しかしながらこれらの欠点のうち「高価」と「小容量」については、容量の拡大と価格の低下に伴って解消されつつあり、HDD との差は次第に小さくなっています。「書き換え回数制限」に関しても SSD 内部の書き込みロジックで半導体部品に対して均等に書き換えることで、寿命を延ばすことができ、計算上現在の平均的な書き換え頻度で上限に達するのは 30 年後であることが知られています。実際にスマートフォンやタブレット端末、携帯型音楽プレイヤーで書き換え上限回数を気にして使用する人はいないのがよい例です。

SSD は主にランダムに読み込む際の IOPS が飛び抜けて高速なため、読み込み主体のオンライン処理などでは格段の高速化が期待できます。もちろん書き込み処理や順次読み込み処理も HDD に比べて高速なため、全体の処理速度向上を図ることが可能です。

❼ SSD の有効利用

ストレージ上のデータは、必ずしも高速にアクセスしたいものだけでなく、活用頻度が高くないデータも混在しています。例えば一般的に高速処理が求められる DB サーバーであっても、参照頻度が高く性能要件が厳しいデータと、アーカイブ・ログなど参照頻度がそれほど高くなく、速度が遅くても問題にならないデータが混在しているのが普通です。

システム全体の性能を向上させる目的で、SSD を搭載したストレージ・システムを導入しようとしたとき、すべてのデータを SSD に配置することができれば、何も気にせず高速な処理が可能です。しかし、それでは非常に高価なシステムとなってしまいます。

そこで多くのストレージ・ベンダーは SSD と HDD 両方を組み合わせたストレージ・システムを提供しており、性能要件が厳しいデータだけを SSD に載せる工夫をすることで、性能向上とコスト適正化を両立しようとしています。では、どのデータをどのくらい SSD に配置すれば効果が高いのでしょうか。

SSD を活用するアプローチとして、多くのストレージ・ベンダーが採用しているのは SSD を大きなキャッシュ・メモリーとして利用する方法です。従来からキャッシュとして利用されてきた半導体メモリーよりも容量単価が安く HDD よりも高速なため、主に HDD のシークおよびサーチ時間の割合が高くなるランダム I/O データのうち参照頻度が高いものを SSD 上に記憶することで、あたかもキャッシュ・メモリーが大量に増えたように見せることが可能になります。その結果、オンライン DB サーバーなどのランダム I/O 発生頻度が高いサーバーからの I/O 処理効力が大幅に向上します。

しかし、この場合の SSD 上に存在するデータはあくまでキャッシュ・メモリーと同じ扱いであり、HDD 上に保管されるデータの I/O 速度向上には寄与しますが、同じデータが SSD と HDD の両方に存在することから保管コストの観点では効率的とはいえません。とはいえ、SSD をキャッシュとしてではなく、HDD の代替として高速エリアとして利用する場合には、どのデータをどのように配置したらよいのか悩むことになります。

IBM では保管コストを最適化しデータ配置を自動化するために、System Storage DS8000 や IBM System Storage SAN Volume Controller（以下、SVC）および IBM Storwize V7000（以下、Storwize V7000）

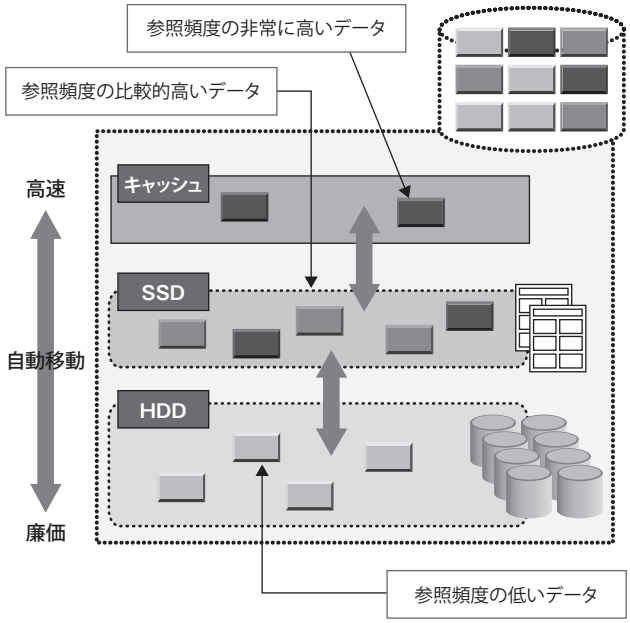


図 6.Easy Tier の構造

で Easy Tier と呼ぶ機能を備えています。利用者は SSD と HDD の両方から構成される「ハイブリッド・プール」を作成しておき、そこにデータを配置するだけでストレージ・システムが内部的に統計を取り、I/O 要求の高いデータを自動的に SSD に配置するので、高速化を図ることができます（**図 6**）。

2012 年から 2013 年にかけてストレージ・システムとしても SSD の高速性を最大限に引き出す高速コントローラーを搭載した製品が続々発表されています。SSD、高速 HDD、大容量 HDD を混在させ、処理速度に応じたデータを自動的に配置する機能によりコスト上昇を抑えながらデータ配置を最適化することが可能になったことから、今後 SSD が本格的に普及するでしょう。

❽ 性能情報の収集

Google 社のサービスに YouTube という動画コンテンツ共有サイトがあります。ネットワークの環境が良ければ、比較的スムーズに動画コンテンツを視聴することができます［4］。

このサイトの人気が高い理由は、比較的高品質な動画をそれほど待たずにすぐに視聴できる点です。コンテンツを送信しながら、速度の統計情報を収集することでサーバーの能力向上を常に見直し、全体最適化を図ることで利用者の「待ち時間」を減らしています。

例えば YouTube 動画を表示して動画上で右クリックし「速度をテストする」を選ぶと、加入している Internet Service Provider（ISP）での平均速度、使用している PC が存在すると想定される市区町村および国での平均速度、世界中のユーザーの平均速度が測定され、それがグラフで時系列に表示されます［5］。

これはコンテンツの人気や社会情勢の変化など、システム外の不確定要素で負荷が変わるため、設計段階では事前に性能要件を立てづらいシステムといえます。その困難を乗り越えるために、実際の性能値を常時取得し、最適化を図ることで満足度を維持しています。このサービスは性能保証をしていませんが、利用者の満足度を向上するために統計情報には非常に神経質になっていることが分かります。

時計メーカーのシチズン時計株式会社が実施した調査では、PC の起動は 1 分、ネットのコンテンツは 10 秒待たされると 7 割以上の人がイライラすると報告されています［6］。このようにネットの向こうでどのような処理がなされているかにかかわらず人の感覚という指標を満足させるためのシステム性能要件は厳しさを増す一方です。

どんなシステムであっても最終的には人が利用することに変わりはなく、人が「待つ」以上は性能の維持管理のために、性能情報を常時取得し、分析することが重要といえます。

メインフレームの時代は 1 台のサーバーに多くのディスク装置が接続され、性能情報はサーバー上で一括確保・管理することが一般的でした。ところが、現在ではストレージ統合が進み、1 台のストレージ・システムに多くのサーバーが接続された形態になっています（**図 7**）。

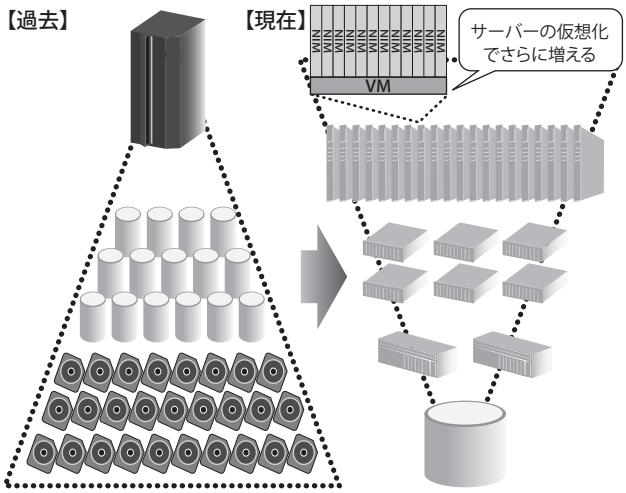


図 7. ストレージ利用の変遷

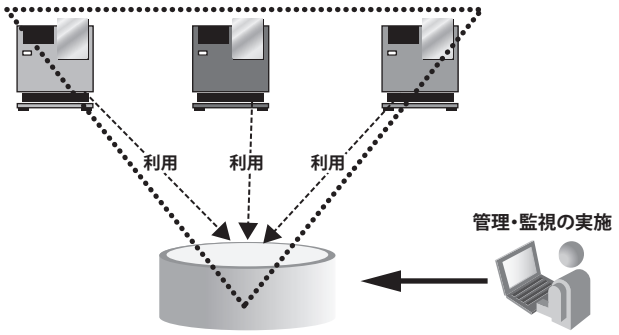


図 8. ストレージ管理の視点

また、最近は物理サーバーの上に仮想化サーバーが搭載されることが多く、そのすべてが共通のストレージ・システムを使うことから、サーバー単位に個別に性能情報を取得・分析することは困難です。従ってストレージ・システムのレベルで性能情報を取得し、最適化を図っていくことが必要なのです（**図 8**）。

❾ まずは健康診断

性能情報は数字として取得できますが、その数字自体を取り出して良し悪しを判断することの困難さはあまり知られていません。

例えば、健康診断で血圧を測定したときに、基準値を超えると「高血圧」として指摘されることがあります。しかし、最近急に血圧が上がったのか、それともずっと同じなのか、意識しないうちに急に上がってきたのか、もともと高いと言われているので健康に気を付けてその数字を維持しているのかといった情報を医師は問診で聞き出すことで、初めてその数字が評価できるわけです。

同様にストレージ・システムの性能評価も、ある一時点の数字だけでは一般論でしか語れません。応答時間 2［ms］というデータは、これまで 1［ms］で応答していたのであれば応答時間が倍に遅くなったと解釈しますが、これまで 4［ms］だったのであれば、むしろ応答時間が半分になり、速くなったと解釈します。このように、差を見て判断するためには、日ごろの状態を時系列で知っておく必要があります。

IBM ではIBM Tivoli Storage Productivity Center（以下、TPC）というストレージ管理ソフトウエアを用意しており、性能情報をストレージの視点から時系列で収集・管理できます。その結果、特定のサーバーに対する性能だけでなくストレージ・システム装置内部のボトルネックまで分析することが可能です［7］。

加えて TPC は利用者が決めた閾値（しきいち）を超えた場合に警告を出すことも可能であり、性能管理も実施することができます。

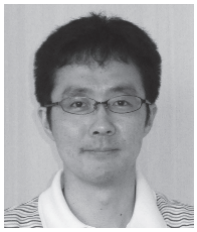
❿ まとめ

IT インフラの性能向上とコスト削減という両立が難しい要求を満たすためには、不要不急のデータをより安価なストレージ上に配置することが必要です。

不要不急データは、Easy Tier の機能で保管コストが安いタイプのディスク装置上に配置する、あるいはテープ・ストレージへのアーカイブ・ソリューションを活用するなど、参照頻度に合わせたストレージを選択して配置するというストレージ管理の方向性が、これからのデータ拡大の時代に必要となるでしょう。

［参考文献］

［1］ 樋口武男:日本経済新聞朝刊,日本経済新聞社.
［2］ Westphal, A., Dufrasne, B., Brandenbur, J. et al.: IBMSystemStorageDS8000:Architecture and Implementation, IBMCorp., http://www.redbooks.ibm.com/abstracts/sg248886.html
［3］ 前田泉:待ち時間革命,日本評論社（2010）.
［4］ 西田圭介:Googleを支える技術,技術評論社（2008）.
［5］ YouTube:動画速度の比較,http://www.youtube.com/my_speed?hl=ja
［6］ シチズン時計株式会社:ビジネスパーソンの「待ち時間」意識調査,http://www.citizen.co.jp/research/time/20030528/index.html
［7］ Orlando, K., Frueh, D., Angelo, P. et al.: SAN Storage Performance Management Using Tivoli Storage Productivity Center, IBM Corp., http://www.redbooks.ibm.com/abstracts/sg247364.html



日本アイ・ビー・エム株式会社
システム製品事業
システム製品テクニカル・セールス
IT スペシャリスト

櫻田 昌己 **Masaki Sakurada**

［プロフィール］
アドバンスド・テクニカル・サポート・グループで IBM 大型ディスク製品を軸にストレージ製品全般に関する IBM 内外のエンジニア支援活動を実施。1997 年入社以来、メインフレームからオープン系サーバーまで多くのお客様のシステム管理に携わった経験から、ストレージのみならずシステム全体の設計を提案できることが強みである。