ストレージフュージョン：ストレージシステムとデータベース管理システムの融合

喜連川 優 *1
茂木 和彦 *2
阿部 淳 *2
鈴木 芳生 *2
小高 俊彦 *2
合田 和生 *1
河村 信男 *2
西川 記史 *2, *1
藤原 真二 *2
星野 喬 *1
土屋 宏嘉 *2
大枝 高 *2
杉江 衛 *3

*1 東京大学
*2 日立製作所
*3 法政大学

はじめに

爆発的に情報が増大する今世紀において，ストレージは，コンピュータシステムの中で最も重要な役割を果たすコンポーネントになりつつあると言えよう．ストレージは，主記憶に対して2次記憶装置と呼ばれたり，あるいは，補助記憶装置，他の外部接続機器とひとくくりにして，周辺機器とも呼ばれてきた．すなわち，明らかにシステムの中で主役ではなかった．しかしながら，膨大な情報からの価値創出がITの本命となりつつある中で，当然のことながら，情報の格納装置としてのストレージの重要性が増し，近年，市場成長率はサーバのそれよりも高い値を維持している[1), 2)]．

奇しくも本プロジェクトが開始された2003年，iSCSIが標準化された．これはディスク装置の読み書きプロトコルであるSCSIを専用のバスではなく，IPネットワーク上でやりとりすることを可能とするものであり，伝送媒体の拡張という視点では取り立てて驚くほどのことでもないが，ストレージにIPアドレスが付与されることとなったという観点では，ストレージがサーバと対等な立場になったことを意味し，大きな変革と言える．そのような背景の中で，これまでほとんど変革が皆無であったストレージシステムのアーキテクチャを改めて考え直すことを試みた．

ストレージシステムのアーキテクチャ

ストレージというデバイスの特性上，当然ではあるが，アドレスを与えて，当該アドレスに格納されたデータを読む，あるいは，当該アドレスに指定データを書くというインタフェースが長く用いられてきた．すなわち，メインフレーム時代にあっては，ディスクは可変長ブロックを許すCKDと呼ばれるフォーマットで編成されていたが，その後，非メインフレームマシンへのディスクの安価な接続を目的として，現状の固定長のブロックデバイスへと進化し，さらに，多様な径のディスク製造から，コモディティ化された限定された径のドライブを多用したディスクアレイへと進化はあるものの，アドレスとデータというサーバとストレージ間のインタフェースは長期にわたり変化していない．ブロック単位ではないデータのやりとりを実現するものとして，1990年頃よりサーバマシンにNFSやCIFS等のファイル共有プロトコルを組み込んだファイルサーバが生まれ，最近では多様なマシンとの接続を可能とするNASと呼ばれる専用ストレージが広く利用されるに至っているが，これとても，ファイルを単位とするアクセスにすぎず比較的単純なデータのやりとりが行われているにすぎない．

一方，組織の保有するデータ量の増大は，その管理コストの増大を招くこととなり，ストレージの集約化（コンソリデーション）の流れが大きく進展している．いわゆるエンタープライズ・ストレージは，大規模化が進展し，キャッシュ容量は64GBを超すと同時に，数十個のマイクロプロセッサを搭載し，数十ポートのファイバチャネルをサポートし多くのサーバとの同時接続を可能とする機種も少なくない．

豊富な計算資源を有するストレージが出現する中で，ストレージシステムには，指定されたアドレスに対するデータの読み書きという機能を超える新しい機能が導入されるようになった．EMC社のSRDF/TimeFinderはストレージレベルでのボリュームコピーの機能であるが，サーバを一切介在させることなく独立に複製を生成することが可能となることから，広く利用されるに至った．新しい情報解析アプリケーションを既存のアプリケーションに影響を与えることなく実現する際などにおいて，ストレージに閉じた複製機能は大きな利用価値を示した．

今後のストレージシステムは，コンピュータシステムからの多様なニーズを受け，さらなる高機能化が進むも

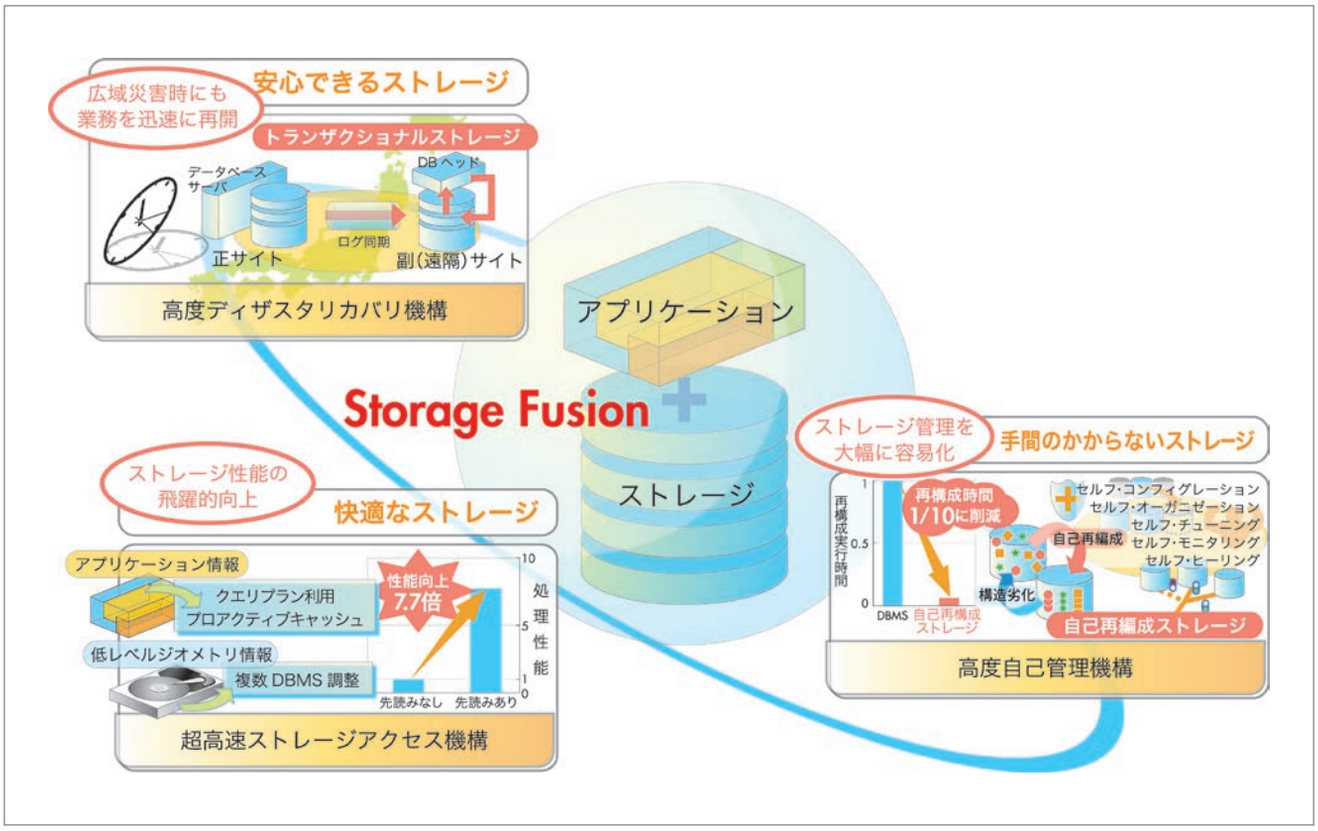


図-1　ストレージフュージョンプロジェクトにおける３つのターゲット

のと予見される．本プロジェクトでは，ストレージに格納される重要なデータを管理するのはサーバ上のデータベース管理システムである点に着目し，データベース管理システムとの密な連携による「ストレージフュージョン」と名付けた次世代ストレージシステム技術の姿について５年間にわたり研究・開発を進めた．

新たな価値を提供するストレージシステム

ストレージとデータベースとの融合という茫漠とした抽象論に終始するのではなく，より具体的な課題をプロジェクト当初に設定した．リーディングプロジェクトは産業活性を目標としていたことから，東大と協力企業である日立は市場価値を有する成果を挙げるべく，現実の課題を取り上げ，いかにストレージフュージョンなる考えが解を与えるかを検討し，高性能化，高信頼化ならびに自己[☆1]管理の３つをターゲットとした．すなわち，**図-1** に示すように，従来システムに比べ著しく高い性能を有する「快適なストレージ」，大規模災害においてもデータを無矛盾な姿で確実に維持する「安心できるストレージ」，ストレージ管理のコストを低減させる「手間のかからないストレージ」を目指した．以下，製品化の現況も含めそれぞれの技術を紹介する．

[☆1] 横田ら[3]はデータのアクセス頻度や容量分布の均等化を自律的に行うストレージシステムを提案している．本研究では後に述べるようにデータベース再編成の自己管理化を取り上げる．

高性能化を目指して：クエリプランに基づくプロアクティブプリフェッチ

ディスクドライブのシーク速度は高々 2.6 倍/10 年の向上率で遅々として改善されないのに対し，プロセッサのいわゆるムーア則は 30 ～ 100 倍/10 年であり，その差はきわめて大きい．トランザクション処理システムの処理レートはさらに 1000 倍/10 年という高いレートで伸びており，ディスクの低性能性を補う技術を継続的に編み出していく必要があると言える．

ストレージとサーバ間のやりとりが，アドレスとデータに限定されていると，やれることには限界があり，これまでも，アクセスパタンをコントローラが解読し，とりわけ連続アクセスがなされる場合には続くアドレスのプリフェッチ（先読み）を投機的に行うなどの手法は採られてきたものの，その効果は限られていた．本プロジェクトではストレージとデータベースが融合し，より多様な情報を共有することが可能となると何が実現できるかを見据え，クエリプランに基づくプロアクティブプリフェッチなる手法を実現した[4]．関係データベースシ

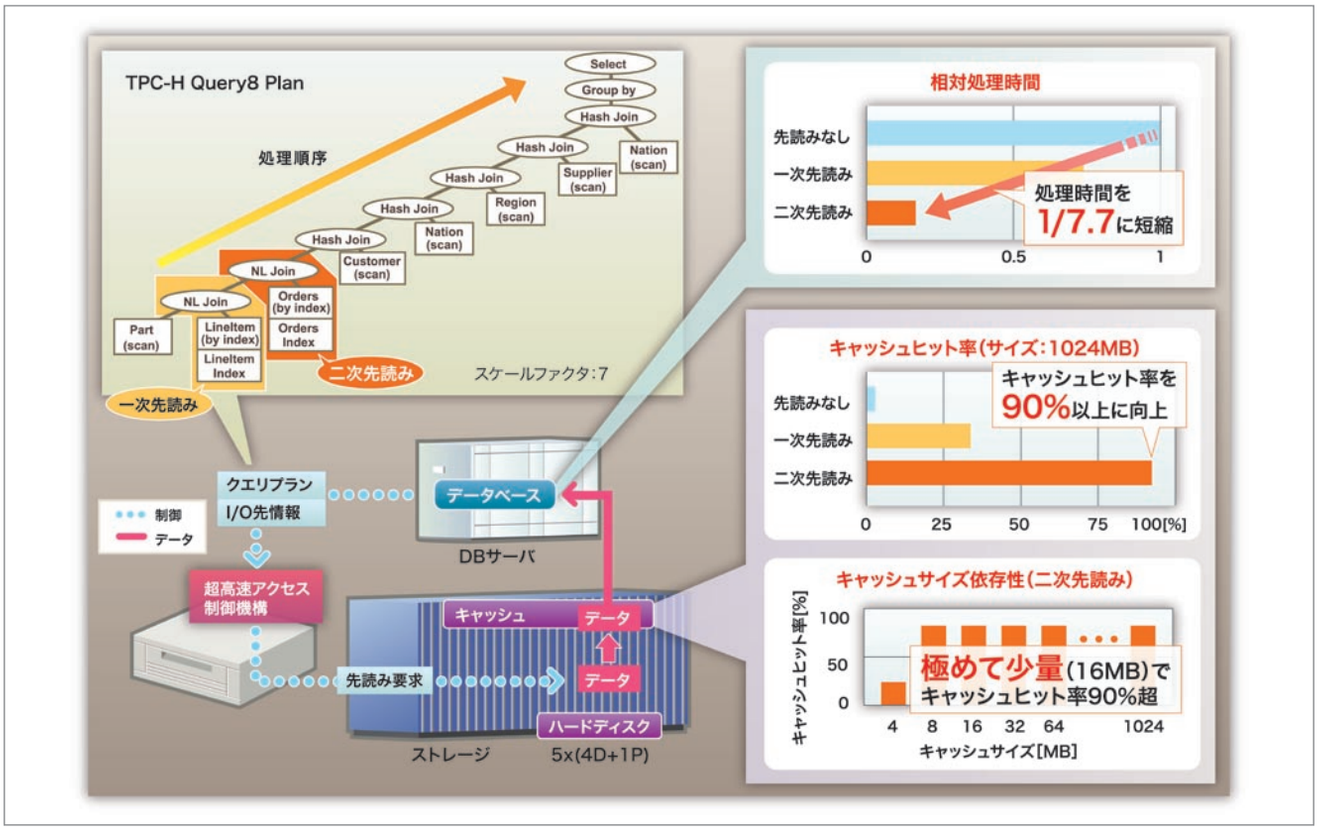


図-2　クエリプランに基づくプロアクティブプリフェッチ

ステムでは一般にクエリプランなる実行計画情報を生成し，それに基づいて実行時ルーチンが処理を進めていく．クエリプランは一般にはデータベース管理システムソフトウェア内で利用されるだけであり，外部には提供されないが，これをストレージシステムに提供し，ストレージシステムが当該情報を基に，より深いプリフェッチを先行的に実施することができたとすると，大きな性能向上が期待される．**図-2** に TPC-H と呼ばれるベンチマークから Query8 というかなり複雑な問合せを取り出し，その有効性について解析した結果を示す．提案するクエリプランに基づくプリフェッチを利用する場合とそうでない場合を比較すると，約 8 倍の性能向上が得られていることが分かる．この際，ストレージのキャッシュヒット率もコールドスタート時において，高々数%から 90%まで向上している．また，必要とされるキャッシュの容量も数 MB 程度ときわめて少なくてすむことが分かっている．このように，データベース管理システムとストレージの融合によりきわめて大きな性能ゲインがもたらされる[5]．

トランザクショナルストレージ

9.11 は世界を震撼させた．倒壊したビル内の企業の中にはデータの喪失により倒産に至ったものもあり，ディザスタリカバリの重要性が強く認識された．社会の基盤を形成する金融業などに対して，米国証券取引委員会（SEC）は，ビジネスの再開までの時間に関する努力目標を設定するなど，企業における具体的な取り組みを促した．稀にしか発生しない事象に対する防衛は，企業にとっては大変大きな負担となるが，その後も，米国ではハリケーンによる被害が続出し，我が国においても豪雨や地震などの大規模な自然災害が頻発しており，災害からのシステムの保全対策は重要な課題となり，昨今，多くの企業がその取り組みを進めている．

このような大規模災害に対してデータを保護するためには，災害地域から離れた地域にデータの複製を保持することしか原理的に方法はなく，遠隔複製（リモートリプリケーション）を行う必要がある．データを複製するには，2 つの手法，すなわち，同期複製と非同期複製が存在し，前者は遠隔になるほど，遅延の影響が大きくなり性能が劣化するが，主サイトと遠隔サイトの間でデータの不整合は発生しない．一方，後者の手法では，性能は距離に影響されないものの，災害時のデータ欠損は避けられず，主サイトと遠隔サイトの間でしばしばデータの不整合が発生する．主サイトと遠隔サイトでデータの不整合が生じると，その回復に複雑な手作業が必要となり，これは一般に多くの時間を要する．データベースは，物理的に大きく分けて，いわゆる本体データと本体デー

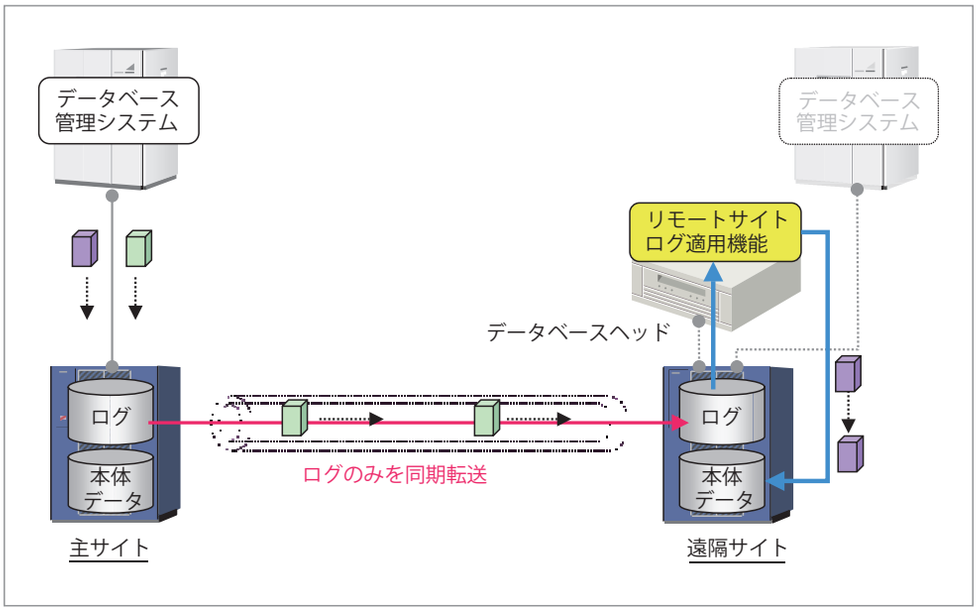


図 -3　ログのみ同期転送によるストレージレベル遠隔複製

タの操作履歴としてのログの 2 種類から構成されている．データベースシステムの回復可能性はログに依存しており，我々はストレージ上でログを同期転送し，本体データを非同期転送する方式を開発した．ストレージだけではどのデータがログでどのデータが本体データかを識別することは不可能であるが，ストレージシステムとデータベース管理システムが融合することにより，このような手法が可能となる．さらに，一歩進め，**図 -3** に示すように，ログのみを同期転送させ，遠隔地においてコミットされたトランザクションのみを適用し，すなわち，遠隔地でログから本体データを再生する方式を開発した．ディザスタリカバリに対する市場からのニーズは大きいため，前者は 2004 年に，後者は 2006 年に日立より製品化された．ログのみを転送する方式はトランザクション欠損がなく，しかも，非同期転送に比べて性能低下を 10% 程度に抑えることができ，また，ログだけを転送すればよいことからネットワーク帯域も少なくてすむという特長が挙げられる [6]．加えて，提案方式は 1 種類のデータのみを転送するため，他の手法と比較して，種々のシステムパラメタの変化に対してきわめて安定な特性を有することを確認している．

本手法は性能という観点以外の側面においても，ストレージシステムにおけるトランザクション特性の実現という大きな特徴を有する．データベースのトランザクション一貫性はホスト上のデータベース管理システムによって実現されているわけであるが，ストレージシステムとしてトランザクション特性を保証することができるならば，すなわち，ストレージがログの管理を司り，常に一貫性を有する界面をストレージ自身で提供することができれば，システムの構築と運用を大きく簡単化できることが期待される．トランザクション特性を有するストレージという意味で，トランザクショナルストレージと名付けてみた．U. C. Berkley の E. Brewer らも Flexible Transactional Storage なる類名の提案を行っているものの，この場合 storage とはホスト上のデータ構造管理機構を意味しており，ロバストなアプリケーションを実現するためのホスト内のソフトウェアフレームワークのあり方に議論がとどまっている．すなわち，トランザクション特性を二次記憶装置としてのストレージシステムへ付与することに踏み込んだものではなく，意味するところは大きく異なる．今後もこの方向性について研究を深めていきたいと考えている．

手間のかからないストレージ

データがアセット（資産）であるという認識の浸透とともに，企業の抱えるデータ量は指数関数的に増大しつつある．ストレージの容量あたりのコストは指数関数的に低価格化していることを合わせて考えると，コスト一定の下で，保持可能な記憶容量を指数関数的に増大させることができ，一見何ら問題はないと感じられる．しかしながら，純粋に格納するだけで済むこととはならず，種々の管理の手間が派生するのが通例である．すなわち，管理者の観点からすると 1 人が管理する容量が指数関数的に増大することとなる．一般に高いスキルのある IT 従事者を雇用することは容易ではなく，また人材養成も容易ではないことから，いかに管理コストを低減させるかは多くの企業の課題となっている．IT システム全体の管理のコストも当然のことながら大きな課題ではあるが，指数関数的に肥大するストレージ空間の管理

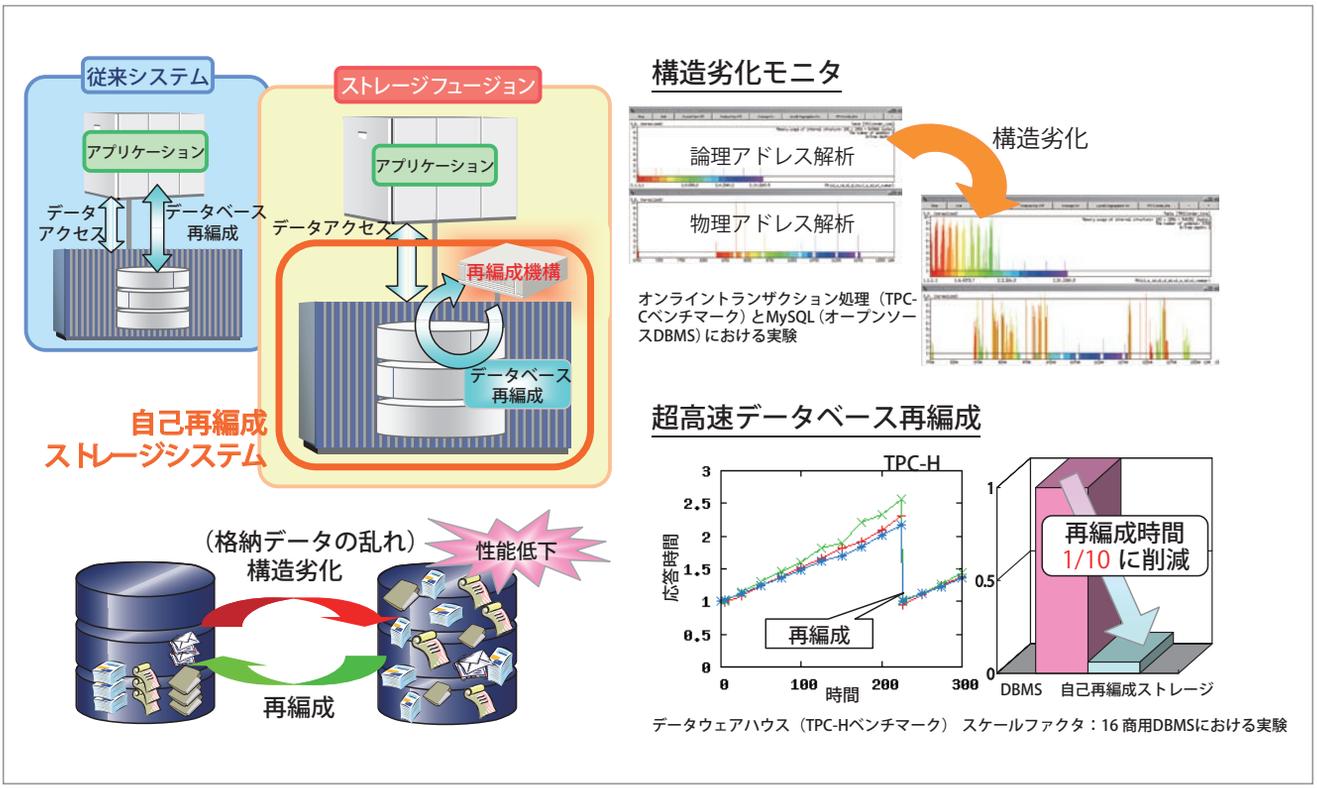


図 -4　自己再編成ストレージシステム

コスト低減は新たな課題とも言える．

すでに述べたように，ストレージシステムは多大な計算資源を有するに至っており，それらの資源を活用して「管理機能」をストレージシステム自身の上で実現することの潜在的な優位性について検討を進めた．本来，データの管理はデータが格納されたストレージから離れたサーバ上で実現するよりも，ストレージ自身で実現することが自然であることは言うまでもなかろう．従来，ストレージシステムは単純な格納装置と見なされてきたが，自身が格納するデータに関してサーバ上のデータベース管理システムと情報を共有することにより，管理機構を部分的にでも，ストレージシステム自身に導入することができないかと考え，その可能性を模索した．

データベースシステムは利用を重ねると，次第に性能が低下してくる．これは，新規データの挿入や，データの更新や削除などが行われるに従い，データの配置が当初のきれいに整列された状態から次第に乱雑になることにより，検索や更新の性能が低下する現象を指し，通常，構造劣化と呼ばれている．性能低下が限界を超えると，もしくはその以前に，通常，データベース再編成と呼ばれるデータベースの大掃除のような処理を施し，性能を回復する必要がある．必要とされる再編成の周期はシステムによって異なり，一般に管理者にとって，そもそもいつ実施すればいいのかの判断は容易ではなく，悩ましいデータベース管理の代表例と言える．我々は独自に構造劣化指標を規定するとともに，当該指標をもとにストレージ自身がデータベース再編成を実施するストレージシステムの研究を進めた[7), 8)]．**図 -4** に概要を示す．また，構造劣化モニタなる可視化ツールを構築し，巨大データベースの中で，どの部分が劣化しているかを同定し，当該部分だけを再編成する部分再編成（partial reorganization）なる手法を提案した[9), 10)]．当然のことながら，再編成はオンライン中に実施され，再編成後に，再編成中に発生した更新を反映し，追いついた段階で，短い中断はあるものの，再編成後のイメージに切り替えられる．

おわりに

ストレージフュージョンなる考えに基づき，いくつかの具体的なシステムを構築することにより，その有効性を明らかにした．すなわち，ストレージシステムとデータベース管理システムの融合により，新たに大きな価値創出が可能であることを明らかにすることができた．サーバは同質化の傾向にあり，ストレージシステムがむしろ高付加価値化のターゲットとなると考えている．文部科学省リーディングプロジェクトは産業活性を第一義的ゴールと設定し，大学と企業が真に連携することの意義に本格的に挑戦したプロジェクトであり，学の先を見た研究実績と産が有する問題を合体することにより，多

くの有意な成果を生み出せたと感じている．e-Society なる産学連携プロジェクトスキームの立ち上げにご尽力いただいた西尾章治郎先生，土居範久先生に深く敬意を表したい．

参考文献

1）IDC. Worldwide Server Market Experiences Modest Growth in Fourth Quarter as Market Revenues Reach Seven-Year High in 2007. Press Release. 27 Feb. 2008.
2）IDC. Worldwide Disk Storage Market Experiences Strong Fourth Quarter Growth as 2007 Revenues Surge Higher. Press Release. 06 Mar. 2008.
3）Watanabe, A. and Yokota, H. : Adaptive Lapped Declustering : A Highly Available Data-Placement Method Balancing Access Load and Space Utilization, Proc. of 21st International Conference on Data Engineering, pp.828-839 (2005).
4）向井景洋，根本利弘，喜連川優：高機能ディスクにおけるアクセスプランを用いたプリフェッチ機構に関する評価，電子情報通信学会第 11 回データ工学ワークショップ（DEWS2000），3B-2 (2000).
5）出射英臣，茂木和彦，西川記史：クエリプラン利用先読み技術における多重処理実行時の性能モデル検討，電子情報通信学会第 18 回データ工学ワークショップ／第 5 回日本データベース学会年次大会（DEWS2007），E2-4 (2007).
6）鈴木芳生，渡辺　聡，水野和彦，藤原真二，河村信男，喜連川優：同期リモートコピーを用いたデータ欠損のない DR システムの性能安定性の評価，情報処理学会論文誌データベース（TOD），Vol.1, No.1, pp.61-78 (2005).
7）合田和生，喜連川優：データベース再編成機構を有するストレージシステム，情報処理学会論文誌データベース，Vol.46, No.SIG 8 (TOD 26), pp.130-147 (2005).
8）Kitsuregawa, M., Goda, K. and Hoshino, T. : Storage Fusion, Proc. of 2nd International Conference on Ubiquitous Information Management and Communication, pp.287-294 (2008).
9）合田和生，喜連川優：構造劣化の局所性を活かしたデータベース部分再編成の提案，日本データベース学会 Letters，Vol.5, No.1, pp.109-112 (2006).
10）星野　喬，合田和生，喜連川優：関係データベースにおける構造劣化監視機構を用いた再編成スケジューラの提案，電子情報通信学会第 17 回データ工学ワークショップ／第 4 回日本データベース学会年次大会（DEWS2006），4C-o1 (2006).

（平成 20 年 9 月 27 日受付）

喜連川　優（正会員） kitsure@tkl.iis.u-tokyo.ac.jp
東京大学生産技術研究所教授，戦略情報融合国際研究センター長，本会副会長，本会フェロー，電子情報通信学会フェロー，ACM SIGMOD Advisory Board Member, IEEE TCDE Asian Coordinator．10 年間にわたる大規模ウェブアーカイブの構築，次世代超高速データベースエンジンの研究開発に従事．

合田　和生（正会員） kgoda@tkl.iis.u-tokyo.ac.jp
東京大学生産技術研究所特任助教，同大博士（情報理工学）．超高機能ストレージシステム，超高速データベースシステムの研究開発に従事．

星野　　喬（正会員） hoshino@tkl.iis.u-tokyo.ac.jp
東京大学情報理工学系研究科博士課程満期退学（2008）．同大生産技術研究所技術補佐員．データベースシステムに関する研究に従事．

茂木　和彦（正会員） kazuhiko.mogi.uv@hitachi.com
（株）日立製作所ソフトウェア事業部先端情報システム研究開発部主任技師，東大博士（工学）．データベース，ストレージシステムの研究開発に従事．

河村　信男（正会員） nobuo.kawamura.df@hitachi.com
（株）日立製作所ソフトウェア事業部 DB 設計部主管技師．データベースシステムの開発に従事．

土屋　宏嘉（正会員） hiroyoshi.tsuchiya.ds@hitachi.com
（株）日立製作所ソフトウェア事業部 DB 設計部部長．データベースシステムの開発に従事．

阿部　　淳（正会員） jun.abe.tt@hitachi.com
（株）日立製作所ソフトウェア事業部事業部員．データベース，ストレージシステムの開発に従事．

西川　記史（正会員） norifumi.nishikawa.mn@hitachi.com
（株）日立製作所システム開発研究所主任研究員，東京大学大学院情報理工学系研究科電子情報学専攻．ストレージ管理ソフトウェアの研究に従事．

大枝　　高（正会員） takashi.oeda.gm@hitachi.com
（株）日立製作所 SAN ソリューション事業部事業企画部部長．ストレージシステムの開発に従事．

鈴木　芳生（正会員） yoshio.suzuki.rf@hitachi.com
（株）日立製作所中央研究所主任研究員．データ処理システムの高速化・高信頼化にかかわる研究に従事．

藤原　真二（正会員） shinji.fujiwara.yc@hitachi.com
（株）日立製作所中央研究所主任研究員．データベースシステム，リアルタイムデータ処理システムの研究開発に従事．

杉江　　衛　sugie@hosei.ac.jp
法政大学大学院イノベーション・マネジメント研究科教授，東京大学博士（工学）．サーバファームのワークロード管理方式の研究開発に従事．

小高　俊彦（正会員） toshihiko.odaka.bb@hitachi.com
（株）日立製作所フェロー，東京大学客員教授，早稲田大学客員教授．超高性能計算技術，高機能ストレージの研究開発に従事．