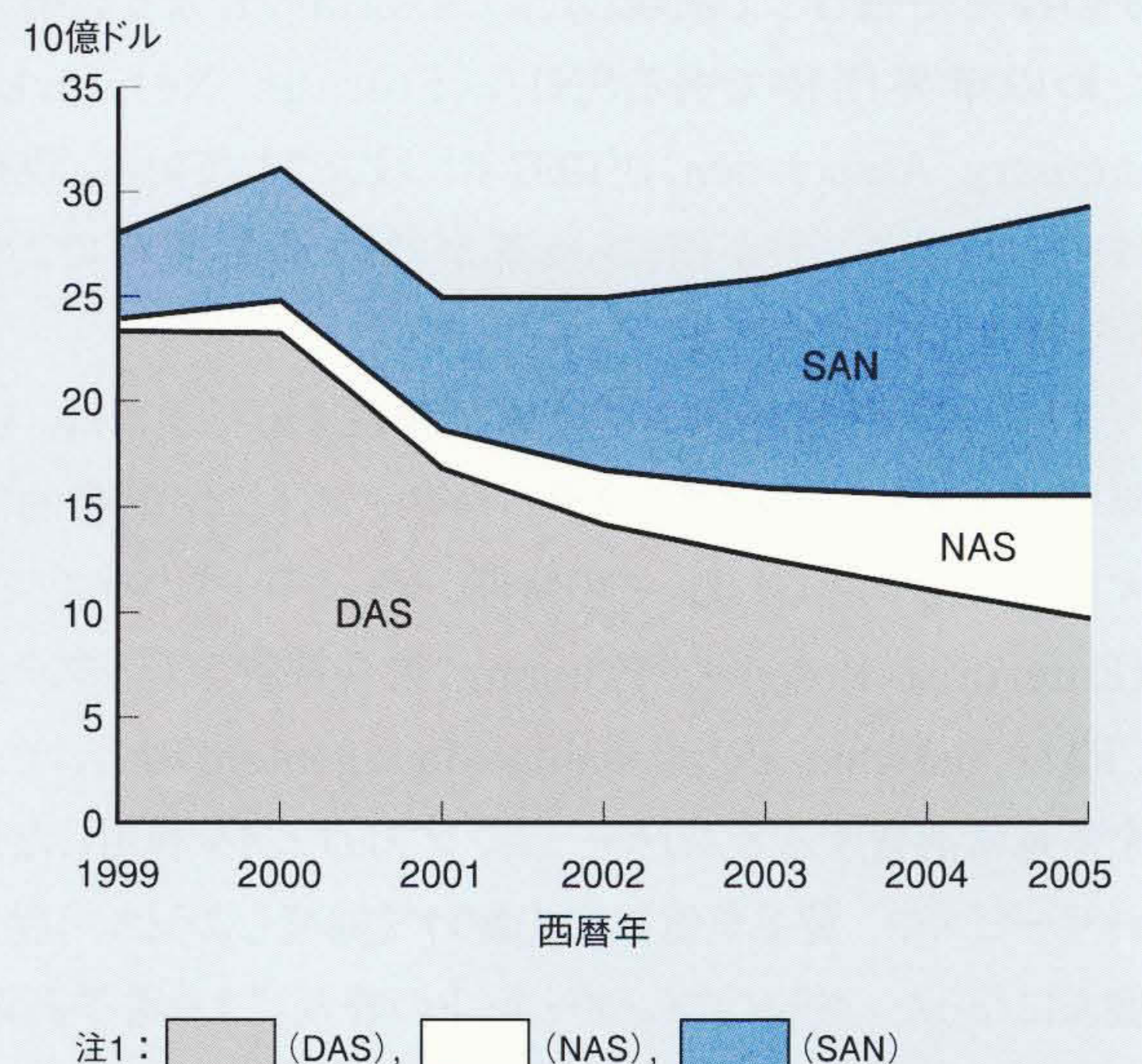
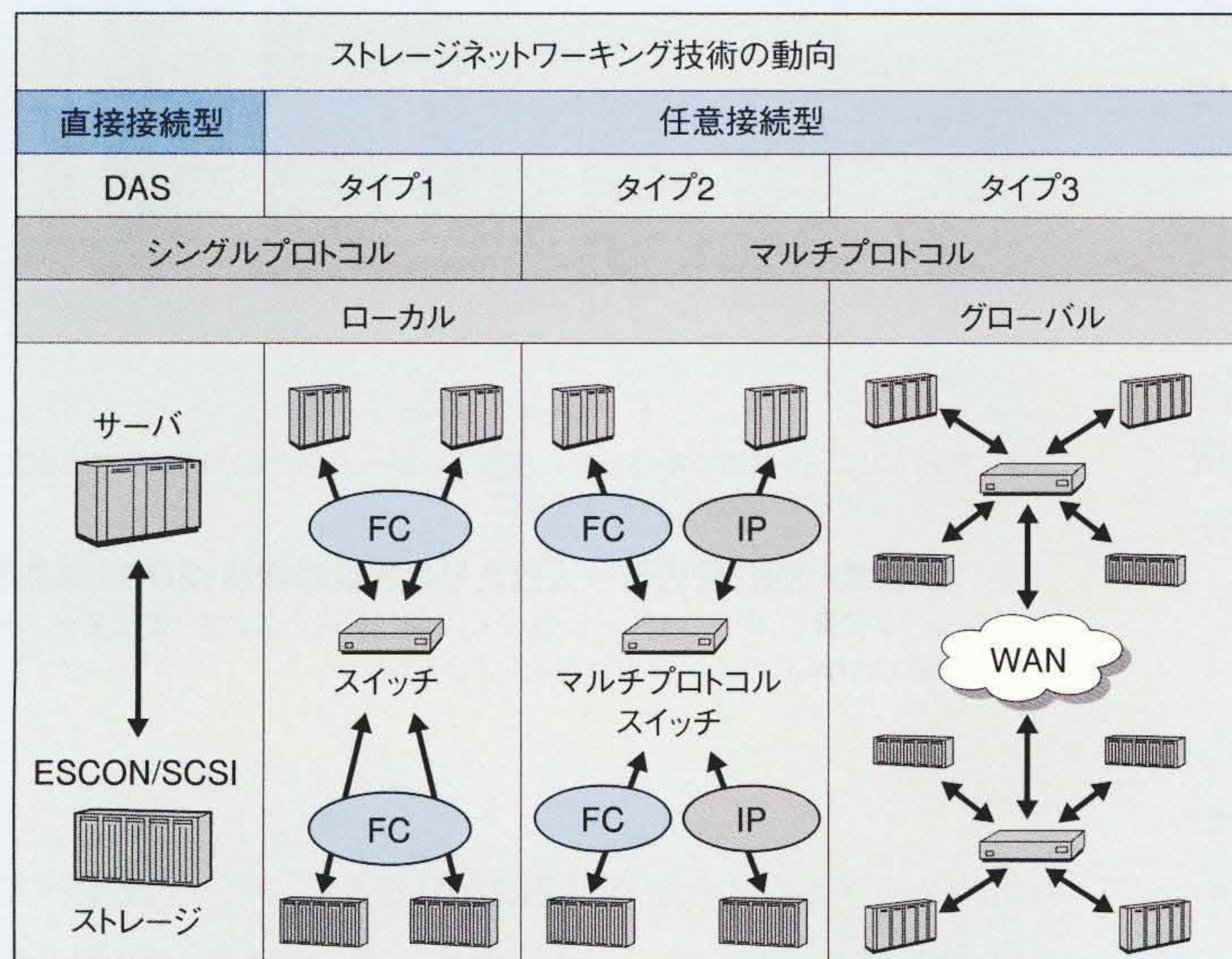


ストレージシステム技術の将来動向

Future Trends of Storage System Technology

大枝 高 Takashi Ôeda



注2: 略語説明

DAS(Direct Attached Storage), SAN(Storage Area Network), NAS(Network Attached Storage), FC(Fibre Channel), IP(Internet Protocol), WAN(Wide Area Network)
ESCON(Enterprise System Connection), SCSI(Small Computer System Interface)

ストレージネットワーキング技術の動向

1998年ころから普及が始まったストレージネットワーキング技術は、当初のファイバチャネルベースからIP技術との融合、さらに広域への対応と進展している。市場シェアでも、SANやNASなどのネットワークストレージが着実に拡大している。

SAN(Storage Area Network)の普及により、コンピュータシステムのストレージセントリック化が進展し、ストレージ技術の重要性がいっそう増大している。ストレージの集約が進むとともに、バックアップなどのデータ運用についての機能をサーバからストレージへ移し、すべてのデータを一元的に管理することで、爆発的に増大するデータの管理コストを削減している。

SANは、これまでファイバチャネルベースのネットワークで構築されてきた。しかし、LAN(Local Area Network)やインターネットなどに広く用いられているIP(Internet Protocol)ネットワークとの融合により、さらに大規模、広域化する可能性がある。2001年米国

ニューヨーク市での「9.11テロ」によって必要性が再評価されているディザスタリカバリへの応用が期待されており、その実現性を確認するための実験が米国では東海岸―西海岸間で、わが国では東京―大阪間などでそれぞれ行われている。

さらに、データベース管理ソフトウェアやアプリケーションソフトウェアとストレージが連携し、従来の限界を超えた性能や信頼性を実現する試みも注目されている。日立製作所は、このようなストレージシステムの進展に貢献するため、実験への参加や技術開発に取り組んでいる。

1 はじめに

1990年代におけるストレージシステム技術での大きなトピックスは、RAID(Redundant Array of Independent Disks)

技術の台頭である。RAIDという用語はカリフォルニア大学バークリー校の技術レポート¹⁾で最初に定義された。この論文の著者はD.Patterson, R.Katz, およびG.Gibsonの3氏である。彼らはRAID技術での貢献を認められ、1999年にIEEE(Institute of Electrical and Electronics Engineers:米国

電気電子技術者協会) Reynold B. Johnson 情報ストレージ賞を受けている。

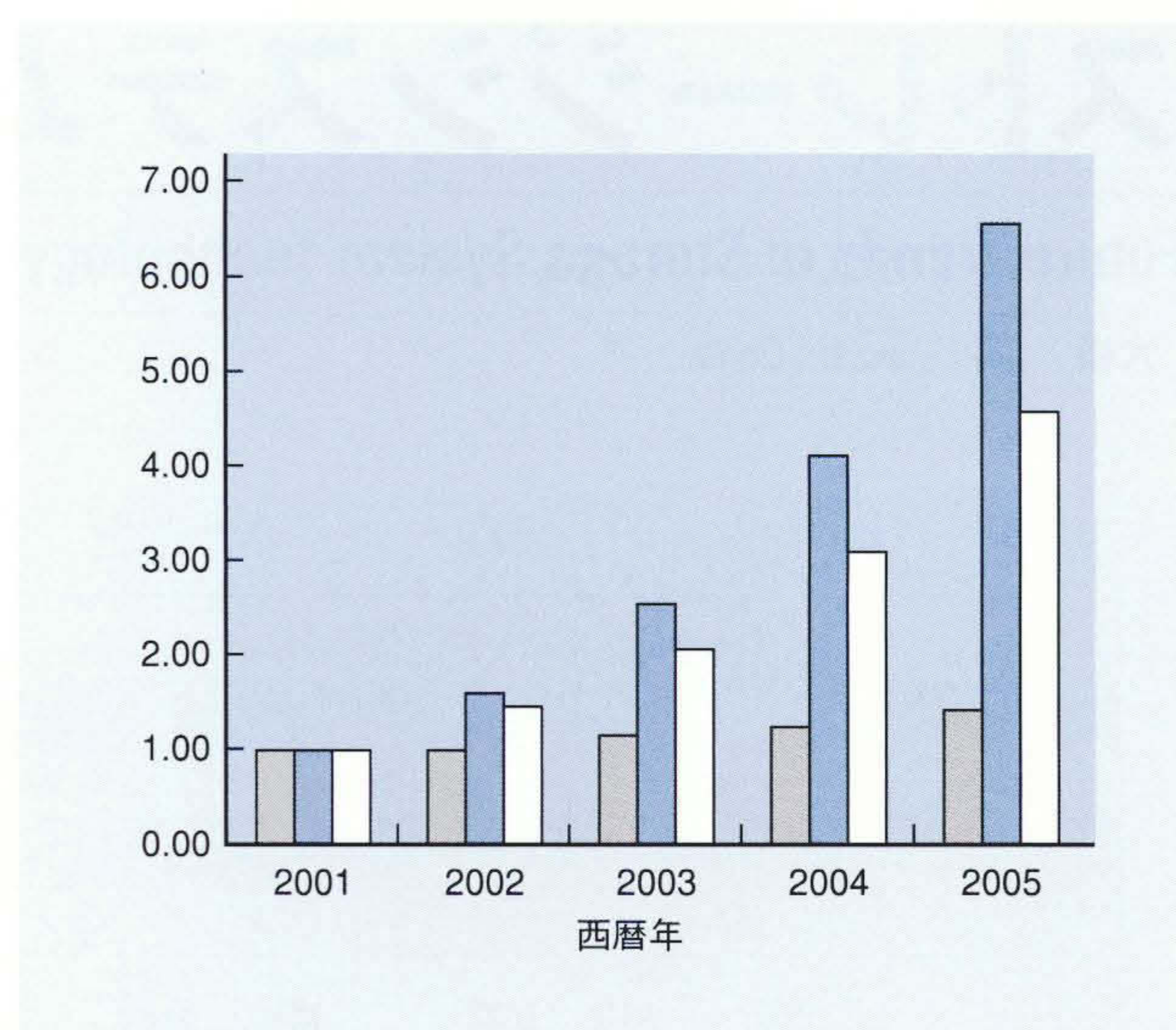
RAIDの次の大きなストレージシステム技術は、SAN (Storage Area Network), NAS (Network Attached Storage) を含むストレージネットワークング技術である。この技術は、メインフレームで行われていたストレージ管理手法、大規模ストレージベンダーの技術、ファイバチャネルなどのいくつかの技術が融合して実現した。この技術の普及と標準化のための業界団体であるSNIA (Storage Networking Industry Association) が1997年に設立されている。現在ではヨーロッパや、わが国にも支部ができ、世界規模でこの技術を推進している²⁾。

ストレージネットワークング技術が普及する背景には、1990年代にエンタープライズシステムのデータ容量が爆発的に増大したことがあげられる。この時期、データウェアハウス、ERP (Enterprise Resource Planning: 統合基幹業務システム)、CRM (Customer Relationship Management) など、データを多量に消費するアプリケーションソフトウェアを利用した新しいサービスや、製品を他社に先駆けて開発しようという動きが盛んになった。さらに、インターネットの普及による電子商取引などが、その後のデータ容量の増加を加速させている。

従来、特にオープン系のソフトウェアでは、サーバにストレージを直接接続するダイレクト接続ストレージが主流であった。しかし、ストレージ容量が急速に増大すると、ダイレクト接続ストレージではサーバプラットフォームごとに管理手法が異なるので、管理コストがストレージの購入コストをはるかに上回るようになった。また、ストレージ容量の伸びに管理者数の増加が追いつかず、1人当たりの管理容量が急速に増大している(図1参照)。そのため、分散していたストレージを集中して、ストレージ専用のネットワークによってサーバ群とストレージを接続するSANが普及してきた。サーバからストレージ管理を切り離し、集中的にバックアップなどの運用を行うことができるので、ダイレクト接続に比べて管理コストを削減することができる。

SANの普及とともに、ストレージの高機能化が進展している。従来、サーバで行っていたデータ運用処理がストレージに移行してきたからである。SANの高速性を生かしたバックアップ機能や、災害対策のためのリモートコピー機能、異種サーバ間のデータ共有機能などが着実に普及している。また、高機能化するストレージを効率的に運用するためのストレージ管理ソフトウェアも重要性を増している。

このように、SANは着実に普及が進んでいるが、現状ではいくつか課題もある。その一つがSANのアイランド化である。規模が数台以下のファブリックでデータセンター内に閉じたSANが島のように点在している例が多く、大規模・広域化のニーズには十分な対応をすることができていない。さらに、サーバと異なり、一般的なSANストレージではファイルやデータベー



注：■ (ストレージ管理者増加指数)、■ (ストレージシステム出荷容量増加指数)
□ (管理者1人の管理容量増加指数)

図1 管理者・容量・管理者一人当たりの容量の増加(2001年基準)
ストレージ容量の増加にストレージ管理者の伸びが追いつかず、管理者一人当たりで管理しなければならない容量が増大している。

スのテーブルなどを識別することができず、データはすべてブロックレベルで扱われる。このため、データ運用に限界を生じる場合がある。

ここでは、これらの課題に対応するIPストレージ、ディザスタリカバリ技術、DBMS (Database Management System)-ストレージ連携新技術などの概要と最近のトピックスについて述べる。

2 IPストレージの可能性

2.1 ファイバチャネルとIP

ファイバチャネルをベースに普及してきたSANには、1章で述べたように広域への対応などの課題もある。IP技術のストレージへの応用によってこの課題を解決しようという動きがあり、プロトコルの標準化や対応製品の普及が進んでいる。

IPストレージが期待される理由を、ファイバチャネルとの比較で考えてみる。まず、IPネットワークがすでにデータセンターのインフラストラクチャーとして定着していることがあげられる。ほとんどのサーバがIPネットワークのコネクションを持っているうえ、スイッチ・ルータなどのネットワーク機器もすでに導入しているので、初期コストがファイバチャネルに比べて小さく、世界中を網羅するネットワークも存在する。また、ファイバチャネルに比べて、習熟した管理者も豊富である。

ストレージへの適用には課題もある。ファイバチャネルをストレージへのIO (Input/Output) 専用用いるのに比べ、ウェブやメールをはじめ、さまざまなアプリケーションソフトウェアに使用されるTCP/IP (Transmission Control Protocol/

Internet Protocol) 処理の負荷は重く、サーバのCPU (Central Processing Unit) などのリソースを消費するからである。ある実験の結果では、420 Mビット/sのデータ転送のためにPCサーバのCPU (Pentium^{※)} III Xeon 1.4 GHz) が100%になっている³⁾。また、ストレージ管理ツールの充実も重要である。管理者もIPネットワークだけでなくストレージ管理に習熟する必要がある。

2.2 IPストレージのプロトコル

SNIAのIPストレージフォーラムは、iSCSI (Internet Small Computer Systems Interface Protocol), iFCP (Internet Fibre Channel Protocol), およびFCIP (Fibre Channel over IP) のIPストレージのプロトコルに対応した三つのサブグループに分かれて標準化を推進している(図2参照)。

(1) iSCSI

iSCSIは、サーバとストレージをIPネットワークによって直接接続することを意図して設計されたプロトコルである。このため、TCP/IPパケットにSCSIコマンドをカプセル化している。サーバは通常のNIC (Network Interface Card) をiSCSI対応デバイスドライバ、もしくはiSCSI専用のホストバスアダプタを用いて、iSCSIをサポートするストレージにアクセスする。TCP/IPパケットにカプセル化されているために、サーバとストレージ間の接続には通常のIPネットワーク機器を用いることができる。

現在普及しているSANは、ファイバチャネルの上位層[FC (Fibre Channel)-4]プロトコルとしてSCSIを用いている。これ

に対し、iSCSIはTCP/IPの上位プロトコルとしてSCSIを用いたものである。iSCSIは2003年2月に、IETF (Internet Engineering Task Force) で正式に承認された⁴⁾。

iSCSIをサポートするサーバがファイバチャネルベースのSANに接続するためのゲートウェイを計画しているベンダーもある。このため、既存のFC-SANにサーバを手軽に接続する手段としても利用することができる。

(2) iFCP

iSCSIとは異なり、iFCPはFC-SANの存在を前提にしたプロトコルである。複数のFC-SANをIPネットワークを利用して接続することにより、遠隔地の各データセンターを結んだ広域SANが実現する。

iFCPではIPパケットの中にFCフレームをカプセル化する。また、iFCPゲートウェイではFCアドレスとIPアドレスのマッピングを行い、パケットをルーティングする。このため、装置間のコネクションはそれぞれ独立であり、障害が発生してもほかの装置間の通信に影響を与えることはない⁵⁾。

iSCSIとiFCPでストレージ装置を発見するためのプロトコルとして、iSNS (internet Storage Name Service) がIETFで検討されている⁶⁾。

(3) FCIP

FCIPもiFCPと同様にFC-SANの存在を前提にしたプロトコルである。複数のFC-SANを、IPネットワークを利用して接続することにより、遠隔地の各データセンターを結んだ広域SANを実現することができる。

IPパケットにFCフレームをカプセル化する方法も同じであるが、FCIPゲートウェイは個々の装置間のコネクションを作るのではなく、FC-SAN間を接続するIPTンネリングを行う。SAN間接続は、FCスイッチからは標準のスイッチ間結合として認識される。コネクションはFCIPゲートウェイ間で張られるので、障害が発生するとすべての通信に影響がある⁷⁾。

2.3 広域接続実証実験「Promontoryプロジェクト」

IPストレージ技術の検証のために、米国の東海岸と西海岸のSANを接続する実験が行われている。このプロジェクトは初めて東西の鉄道がつながった町の名前にちなんで「Promontoryプロジェクト」⁸⁾と呼ばれ、2本の2.5 Gビット/sのリンクを用いてiSCSIとiFCPで性能を検証した(図3参照)。結果は、大陸を横断する距離でもワイヤースピードが達成できることを示した。この実験には日立製作所のストレージが使用されている。

※) Pentiumは、Intel Corporationのアメリカ合衆国およびその他の国における登録商標である。



注：略語説明

SCSI (Small Computer System Interface)

TCP/IP (Transmission Control Protocol/Internet Protocol)

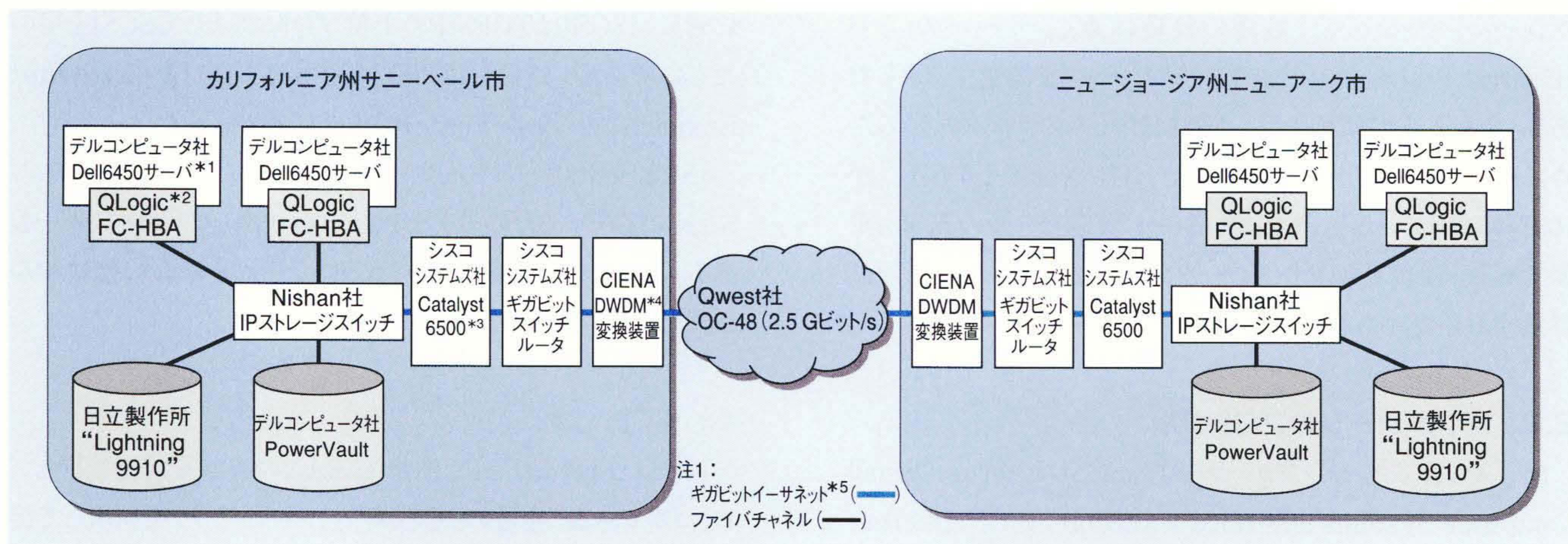
iSCSI (Internet Small Computer Systems Interface Protocol)

FCP (Fibre Channel Protocol), FC (Fibre Channel)

iFCP (Internet Fibre Channel Protocol), FCIP (Fibre Channel over IP)

図2 iSCSI, iFCP, FCIPの比較

iSCSIはTCP/IPにSCSIコマンドを、iFCPはFCフレームをそれぞれカプセル化している。FCIPはFCフレームのIPTンネリングを行う。



注2：*1 Dell6450サーバは、米国Dell Computerの商品名称である。*2 QLogic FC-HBAは、米国QLogic Corporationの商品名称である。

*3 Catalyst6500は、米国Cisco Systems, Inc.の米国および他の国々における登録商標である。*4 CIENA DWDMは、株式会社シエナコーポレーションの商品名称である。

*5 イーサネットは、富士ゼロックス株式会社の商品名称である。

図3 「Promontoryプロジェクト」性能測定用機器の構成

米国の東海岸と西海岸のデータセンターをOC-48回線（2.5 Gビット/s）を使って接続した。IP Storage SwitchでFCとIPの変換を行っている。プロトコルはiSCSIとiFCPを用いている。1本のギガビットイーサネットを双方向で用いて、215 Mバイト/sの平均スループットを得た。

3 ディザスタリカバリ技術の進展

3.1 ビジネス継続の必要性

2001年米国ニューヨーク市での「9.11テロ」は、各企業に災害対策の必要性をあらためて認識させた。テロや地震、火災、風水害などによるビジネス停止の危険性が確かに存在し、その対策が企業の存続にとって緊急の課題になった。災害対策について行政の指導や規制の強化、対策が十分でない企業への保険料の増加などの外的圧力も強められている。

各企業が災害対策を検討するにあたり、まずはビジネス停止によるコストの算出が必要である。災害対策のコストはリカバリ時間によって大きく変動する。適切な手法の選択には、サービスのダウンタイムコストを知らなければならない。時間当たりのダウンタイムコストは、典型的な例では、e (Electronic) コマース、金融では7億8,000万円、POS (Point of Sales) のバックボーンでは3億1,000万円、通信販売では2,000万円などという調査結果がある。

実際のダウンタイムについては、ニューヨーク商品取引所では「9.11テロ」でコンピュータシステムが完全に破壊されたが、ロードアイランドのディザスタリカバリサイトで、約11時間後には取引再開の準備が終了していたといわれている⁹⁾。世界貿易センターは、1993年にも爆弾テロの被害を受けている。ニューヨーク商品取引所もその教訓を生かし、綿密なリカバリプランを策定していたので、このような迅速な復旧が可能となった。

3.2 リモートコピー技術

災害対策のためには、安全な遠隔地にデータを複製する必要がある。データのリモートコピーを実現する手段には、(1) ホストコンピュータのソフトウェアで行うもの、(2) ストレージシ

テムで行うもの、(3) SAN上のアプライアンスで行うものなどがある。

ホストベースリモートコピーでは、ホストコンピュータ上のソフトウェアがアプリケーションソフトウェアやOSのデータ入出力 (IO) をフックし、ネットワーク経由でリモートサイトのホストコンピュータに送る。ストレージシステムベースの場合は、ホストコンピュータから書き込まれたデータをストレージが自律的にリモートサイトのストレージにコピーする。これまでは専用線が使われていたが、最近では2章で述べたように、SAN間接続などが用いられるようになってきた。また、SAN上のインテリジェントスイッチなどで、ストレージシステムと同様のリモートコピー機能を持つものもある。

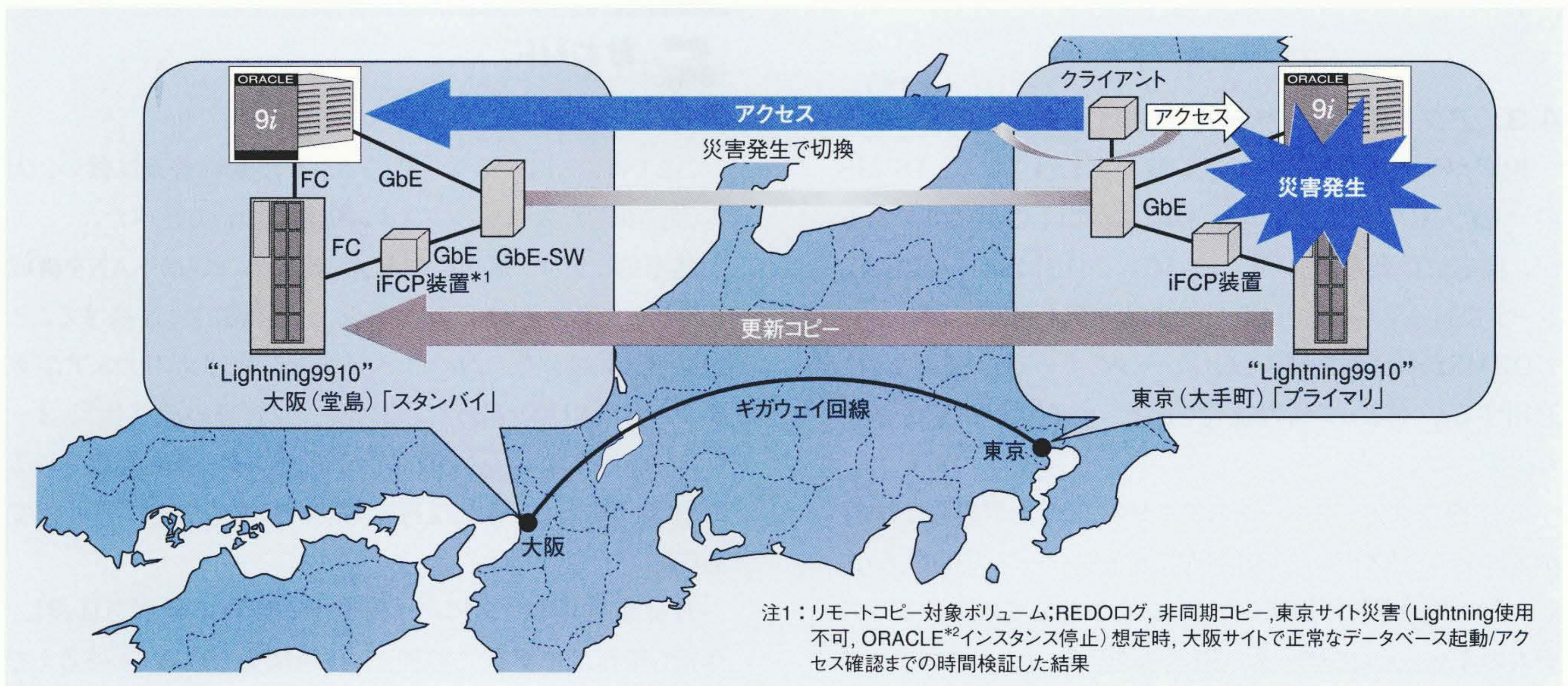
これらのうち、ホストコンピュータのアプリケーションソフトウェア実行に性能的影響が少なく、比較的実績のあるストレージベースのリモートコピーを用いた実証実験について次に述べる。

3.3 東京-大阪間ディザスタリカバリ実証実験

この実験はストレージベースのリモートコピーを利用し、超広帯域IPネットワークと最新のデータベース技術の組み合わせにより、東京-大阪間でのディザスタリカバリが実用的であることを実証したものである(図4参照)。

SAN間接続にはエヌ・ティ・ティ・コミュニケーションズ株式会社のギガウェイ・サービス¹⁰⁾を用い、日本オラクル社のデータベースOracle 9iとそのDataGuard機能でOLTP (On-Line Transaction Processing) 系の処理をフェイルオーバーする。ストレージシステムは日立製作所の「SANRISE」で、そのリモートコピー機能である「TrueCopy」を用い、iFCPゲートウェイでIPネットワークに接続している。

実験はクライアントからのトランザクション実行中に東京サイ



注2：略語説明ほか

GbE-SW (Gigabit Ethernet Switch), GbE (Gigabit Ethernet)

*1 iFCP, IPS3300 (Nishan Systems, Inc), *2 ORACLEは, ORACLE Corporationの登録商標である。

図4 東京-大阪間のディザスタリカバリ実験

東京と大阪のデータセンターを超広帯域IPネットワークで接続し、ストレージベースリモートコピーとサーバベースリモートコピーを組み合わせORACLEデータベースのフェイルオーバー実験を行った。プロトコルはiFCPを用い、5分以内にフェイルオーバーが完了し、サービスが継続できることを確認した。

トが全滅したという想定で行い、まったくトランザクションを損失することなく、5分以内に大阪サイトへの切り替えが終了するという結果が得られた。この実験では、高速な切換のためにホストベースのリモートコピー機能である“Data Guard”も併用している。ホストベースのリモートコピーでは、データベーステーブルやファイルを認識することができる利点がある。ここでは、アーカイブログの転送に用い、リモートサイトでテーブルスペースへ順次反映させている。これにより、災害発生時の切換処理が高速になる。一方、ストレージベースリモートコピーは堅牢でホスト性能への影響が少ない方式である。今回の実験ではこれらをうまく組み合わせて、信頼性が高く、高速なフェイルオーバーを実現した。

このように、ストレージとデータベースなどのミドルソフトウェアやアプリケーションソフトウェアが連携し、従来よりも効率のよいデータ運用を実現することは、これからのストレージ技術の一つの方向性を示している。DBMS-ストレージ連携技術について次章で述べる。

ルは、このような目的で策定された。

シェアードストレージモデルでは、ストレージドメインをファイル・レコード層とブロック層に分割している。ファイル・レコード層はファイルシステムやDBMSであり、例えばNASは、この層を持つストレージシステムである。ブロック層にはデータを蓄積するストレージデバイス部分と、ボリュームのマッピングやRAIDのデータストライピングなど、ブロック構造の変換を行うブロックアグリゲーション部分がある。一般的なSANストレージはブロック層だけを持つ¹¹⁾。

3.3で述べたリモートコピー機能は、ファイル・レコード層での機能 (Oracle Data Guard) とブロック層での機能 (Hitachi TrueCopy) との組み合わせにより、高速で高信頼なフェイルオーバーを実現している。同じ機能でも実現する層によって特徴があり、さらにそれらの組み合わせで従来のSANストレージの限界を超えた機能を提供することができる。この章では、特にデータベースとの連携機能の動向について述べる。

4.2 アクティブストレージ

データベースとストレージの連携という点では、1970年代に行われたデータベースマシンの研究でも、高性能化ストレージ側にデータベースの一部の機能をオフロードすることが提案されている。最近では1998年に、ディスクなどに用いられるプロセッサの高性能化に着目し、ディスク上で検索やソートを行うアクティブストレージがCMU (カーネギー・メロン大学) やUCB (カリフォルニア大学バークリー校) から提案されている。また、自律的負荷バランスやログを利用した自動バックアップなどの機能を用いた自律ディスクが東京工業大学から提案されて

4 DBMS-ストレージ連携技術

4.1 SNIAシェアードストレージモデル

ストレージネットワーク技術の普及により、SAN, NAS, ストレージ仮想化など、各社から種々の方式が提案されている。これらのアプローチの違いを整理するには、ネットワークでのOSI (Open Systems Interconnection) 7層モデルのような定義があることが望ましい。SNIAシェアードストレージモデ

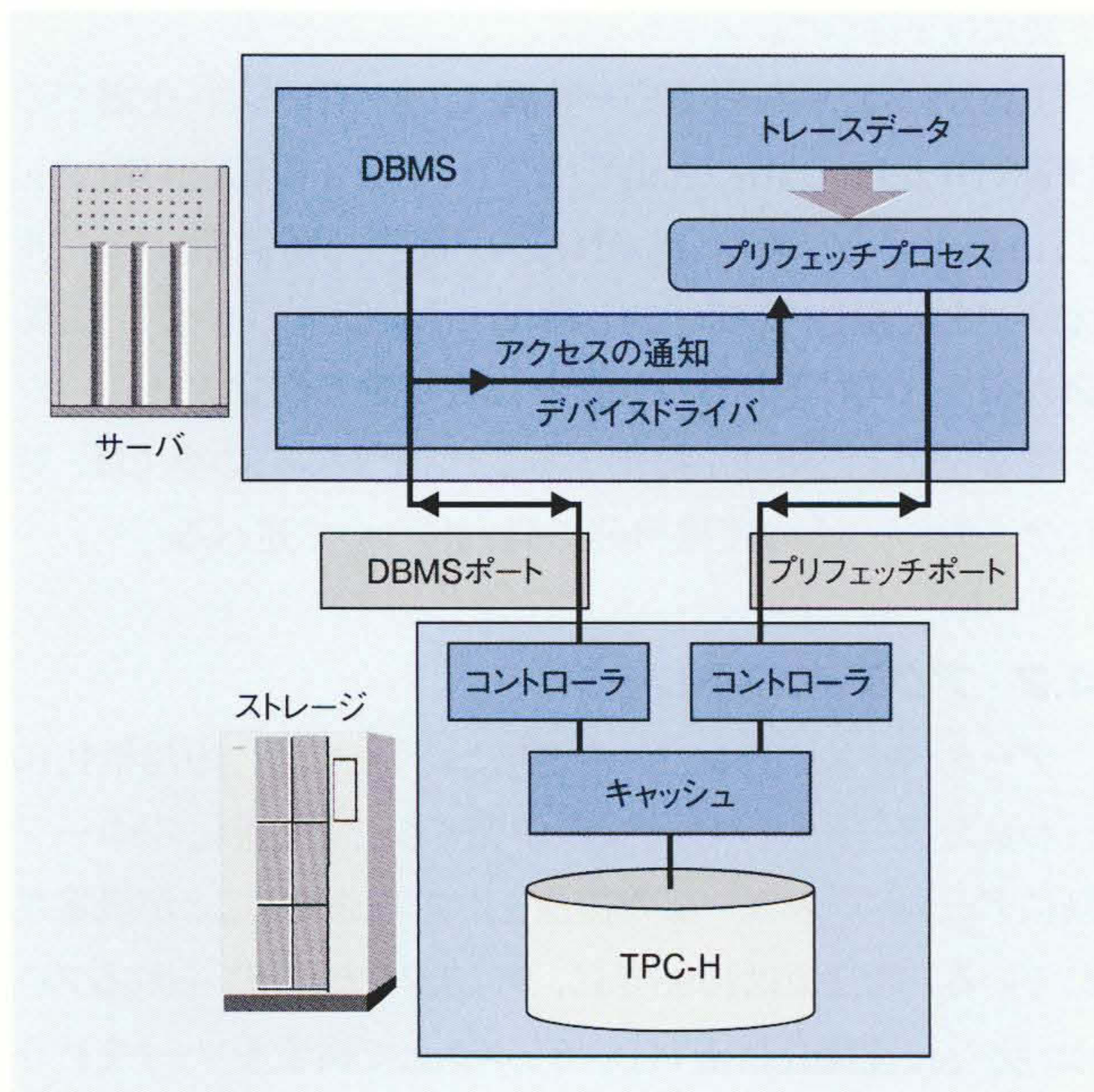
いる。

4.3 アクセスプランベースプリフェッチ

検索やサーチなどをストレージ側で実行するには、DBMSの一部がストレージ上で実行されるような改造が必要であるが、現実には製品化が進んでいない。DBMSの一部をストレージに持ってくるのではなく、これから実行する処理についてDBMSから情報をもらい、ストレージのキャッシュ制御などに適用することによって高性能化を図るアイデアが東京大学から提案されている¹²⁾。

試作機ではデータベースのアクセスプランを取得しておき、ホストのデバイスドライバのレイヤでデータベースのIOをフックし、その情報を基に将来のアクセスを予想してストレージシステムのキャッシュにデータの先読みを行わせた。このアプリケーションの知識を利用した知的なキャッシュ制御により、従来のLRU (Least Recently Used) アルゴリズムの限界を超えた高性能化が可能になる。試作ではTPC (Transaction Processing Performance Council)-Hのあるクエリー処理で約5倍の高性能化を実現している(図5参照)。

冒頭に述べたデータ量の爆発的増大に対応するために、今後このようなデータベースとストレージの連携技術が重要になると思われる。日立製作所のストレージ「SANRISEシリーズ」では、負荷分散による自動性能最適化機能“Performance Management-Cruise Control”や3章で述べたりモートコピー機能 (TrueCopy)、オンラインバックアップに用いられるスナップショット機能“ShadowImage”などが、データベースとの連携をいっそう強化し、効率のよいデータ運用を実現していく。



注：略語説明 DBMS (Database Management System)

図5 アクセスプランベースプリフェッチの試作機(東京大学:喜連川研究室)

DBMSのアクセスプランと実行時のアクセス情報から知的な先読み制御をストレージレベルで行う方式である。

5 おわりに

ここでは、ストレージネットワーク技術の普及に伴い、大きく動き出したストレージシステム技術について述べた。

効率的なデータ運用を実現するには、これらがSANを構成するサーバ、スイッチ、ストレージ上で有機的に連携することが必要である。そのために、ストレージ管理ソフトウェアが重要であり、SNIAが進めているストレージ管理の標準化と、オープンな標準にのっとった製品群が各ベンダーから提供されることが、新しくでてきた技術を実効的にするために不可欠である。

日立製作所は、ストレージシステム技術の発展を目指し、今後も積極的に実験に参加し、技術開発を進めていく考えである。

参考文献など

- 1) D. A. Patterson, et al.: A Case for Redundant Arrays of Inexpensive Disks (RAID), Proceedings of the International Conference on Management of Data (SIGMOD) (1988.6)
- 2) <http://www.snia.org/>, <http://www.snia-j.org/> (日本支部)
- 3) ZDLABS, Alacritech 100x4 Port Server Adapter Performance test, ZDLABS white Paper (2000.3)
- 4) <http://www.ietf.org/internet-drafts/draft-ietf-ips-iscsi-19.txt>
- 5) <http://www.ietf.org/internet-drafts/draft-ietf-ips-ifcp-14.txt>
- 6) <http://www.ietf.org/internet-drafts/draft-ietf-ips-isns-15.txt>
- 7) <http://www.ietf.org/internet-drafts/draft-ietf-ips-fcovertcpip-12.txt>
- 8) http://www.hds.com/pdf/Nishan_Promontory.pdf
- 9) ポール・マシグリア, 外: エンタープライズディザスタリカバリ, ベリタスソフトウェア株式会社編 (2002.7)
- 10) www.ntt.com/gigaway/index.html
- 11) <http://www.snia-j.org/resource/ronbun/index.html> (和訳)
http://www.snia.org/tech_activities/shared_storage_model/ (原文)
- 12) 向井, 外: 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, DEWS' (2000)

執筆者紹介



大枝 高

1985年日立製作所入社, システム開発研究所 第八部 所属
現在, ストレージシステムの研究開発に従事
情報処理学会会員
E-mail: oeda @ sdl.hitachi.co.jp