主メモリ活用型ストレージシステム

進博正† 藤原 睦† 坂本英夫†

この研究では主メモリ利用を前提とした新たなストレージシステムを検討して提案する．本ストレージのデータ管理サーバは，ストレージ媒体となる主メモリ内容を複数のノード間で複製することで相互バックアップと負荷分散を両立する．本ストレージは，データ検索時にアクセスパスとなるディレクトリ木を，コンテンツに基づきリアルタイムに生成する．クライアントの一例として，スプレッドシートをUIとし透過的なストレージアクセスを実現するプログラムを試作した．本ストレージは，従来の共有ファイルによるデータ一元管理はもちろん，グループウェア相当のデータ共有目的の利用にも益する．この発表ではストレージ構成の紹介と試作ソフトのデモを実施する．

A Possible Storage System based on Main Memory

Hiromasa Shin† Mutsumi Fujiwara† Hideo Sakamoto†

1. はじめに

大容量の主メモリを搭載した計算機を HA(High Availability)クラスタ構成とすることで「大容量かつ永続的な記憶メディア」としての主メモリ利用が可能となる．本研究では，上記 HA 構成を前提とする主メモリ活用型ストレージシステムを提案する．主メモリ利用に適したデータモデルを導入し，ランダムアクセス高速性を活かしたアクセスパスとコンテンツの整合性をシステムレベルで保障する新しい機構を提案する．本ストレージのデータ管理サーバは，ユーザ認証に基づくアクセス制御機構とインターネット対応のアクセスプロトコルを有し，広域に分散した多数のユーザに，即時性の高いデータ共有手段を提供する．クライアントの一例としてスプレッドシート UI(User Interface)を利用したプログラムを紹介する．具体的な利用例として，共有ファイルに替わるデータ一元管理やグループウェアとしてのデータ共有目的の利用を想定している．

2. ストレージの構成

2.1. デザイン原理

アクセスパスにディレクトリ木を持つファイルシステム[1]は，汎用的なデータストレージとして広く利用されている．ファイルシステムの構成はランダムアクセスコストの高い磁気ディスクに最適化されており，ファイル内容とディレクトリ木は独立となる．結果，ファイル内容(の一部)から目的ファイルを検索するには適さない．コンテンツによるファイル検索を可能とする後付のインデックスファイル生成機構[2]も存在するが，定期的なバッチ実行を前提としておりリアルタイム性は乏しい．コンテンツとアクセスパスの整合性を保障する機構としては DBMS(Database Management System)が該当[3]するが，データ格納時にスキーマ定義が必要なため，汎用的なストレージとしては活用しにくい．

本ストレージは，上述の短所を解消するため，1．アクセスパスとコンテンツの整合性をシステムが保障すること，2．ストレージ利便性を増すためスキーマ導入を極力控えること，3．広域分散ユーザのシームレスなストレージ利用を実現するインターネット対応プロトコルを標準で備えること，をデザイン原理として掲げた．

2.2. データモデル

本ストレージのデータモデルは，ファイルに相当するストレージユニットと検索ディレクトリ木に相当するファサードから成る(図 1)．ストレージユニットは，コンテンツ格納セルの配列と，アクセスパス上のセル出現パタンを指定するローから成る．ファサードは上記ローのソート表であり，各ローがストレージユニットへのアクセスパスに対応する．単一のストレージユニットは複数のローを持つことが可能であり，ファイルシステムのリンクに相当する多重のアクセスパスを許す．セル内容が変化するたびに関連するローのファサード内の順位を更新する．この単純な仕組みによって，本ストレージはコンテンツとアクセスパスの整合性をリアルタイムに維持する．

†(株)東芝 研究開発センター
Corporate R&D Center , Toshiba Corporation

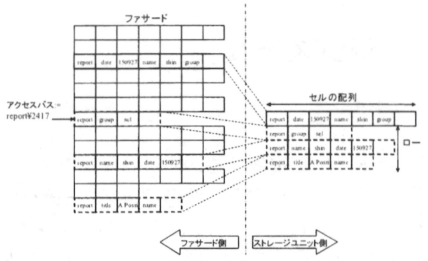

図 1

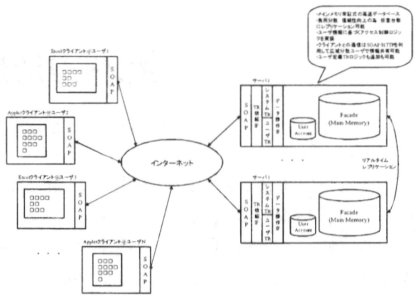

図 2

2.3. アクセス制御

本ストレージのアクセス制御機構はUNIXのファイルシステムを参考にした．パスワード認証を通過したユーザに対してファサードを許可する．ストレージユニット毎に所有者および公開グループを，セルおよびロー毎に参照および更新の許可ビットを与える．セルおよびローの所有者および公開グループはストレージユニットの属性に従う．セルの許可ビットを利用するとファイル内フィールドに相当する細粒度なアクセス制御が可能となる．ローの許可ビットを利用するとディレクトリ単位のアクセスパス制御(制限)が可能となる．

2.4. システム構成

本ストレージは，データ管理サーバとデータ操作クライアントから構成される(図 2)．データ管理サーバはLAN(Local Area Network)で接続された複数ノードでクラスタを組み，データ内容を互いに複製することで相互バックアップを図る．参照アクセスは全ノードで処理することで負荷を分散し，更新アクセスは一台のマスタで処理し他ノードを追随させることでデータ一貫性を保障する．マスタがクラッシュするとスレーブが処理を引き継ぐ．マスタからスレーブへのレプリケーションログ送信には，高信頼マルチキャストを利用するためスレーブノード数に制限はない．このようなHA構成で主記憶データの複製を維持管理するミドルウェアとして東芝製のGigaBase[4]を利用した．

データ管理サーバへのアクセスは，広域に分散したユーザのストレージへの透過的なアクセスを可能とするためSOAP(Simple Object Access Protocol, [5])を利用した．

2.5. クライアント例

本ストレージへアクセスするデータ操作クライアントとしてスプレッドシートUIを持つプログラムを試作した．スプレッドシート上に複数のデータ表示枠(リージョン枠，図 3)を割り当てることが可能である．リージョンの表示モードとして，ファサード表示とストレージユニット表示の二種が存在する．ファサード表示はディレクトリ木のブラウズ表示に相当し，リージョン枠はファサード上の矩形領域にマップされる．ストレージユニット表示はファイル内容の表示に相当し，リージョン枠はストレージユニット上の矩形領域にマップされる．リージョンは互いに独立であり，各々が表示モードと表示位置を指定するアクセスパスまたはストレージユニット識別子を保持する．

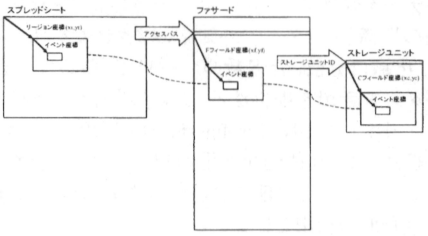

図 3

スプレッドシート上には各セルの値が表示され，そのまま編集できる．もちろん，ユーザ権限とコンテンツの公開範囲に基づき表示範囲や更新可能な範囲はユーザごとに異なる．各セルを選択してメニューを表示させることでアクセス制御リストを編集できる．以下(図 4)は

Microsoft 社 Excel2002 を利用した実装例のスクリーンショットである.

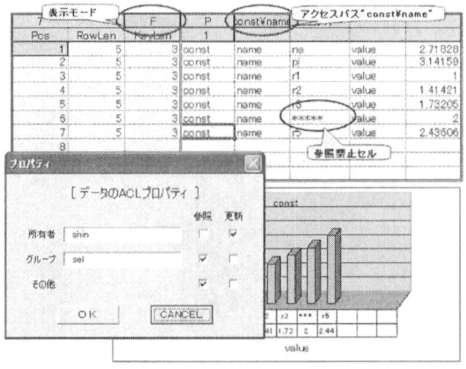

図 4

2.6. 利用イメージ

従来のファイルシステムを利用してテキストファイルによりメモを整理するケースを考える. 次のような内容を持つテキストファイル(図 5)を作成して, ファイル名を日付から"H150927.txt"のように名付ける. このとき, ファイル名とファイル内容の報告日が一致することについて, ファイルシステムは何ら保障しない.

```
文書種別: report
グループ名: sel
報告者: shin
報告日: H15.09.27
本文: 次世代ストレージの GUI クライアント環境は、
インターネット上に広域に分散することを想定する。
したがって、GUI クライアントとストレージサーバの通
信には SOAP Over HTTP を利用する。　(続く)
```

図 5

作成したテキストファイルは, 上司を含む部内関係者が自由に参照できなくては意味が無いため, 自分のホームディレクトリ"/home/shin/report"で作成したファイルを共有フォルダの所定位置"/pub/sel/report/shin/"へコピーする. 期末になり特定の個人やプログラムの単位で参照するか, 特定のキーワードを含む週報を一括して参照するには全文検索に頼るしかない. この様な方法は少人数の組織では有効でも, 組織が大きくなると破綻する。より大きな組織では「週報公開システム」などのカスタムアプリケーションを構築する破目になる. その結果, 業務に応じて複数のカスタムアプリケーションが乱立することもあり, ユーザは個々のアプリケーションの利用方法に精通しなくてはならない.

主メモリストレージを利用すると, テキストファイル(図 5)に対応するストレージユニットを作成してファサードに登録するだけで, タイトルやグループごとに検索して週報を参照することが可能となる. たとえば, アクセスパス"report¥name¥shin"を指定すると個人別日付順の週報一覧が, アクセスパス"report¥group¥sel"を指定するとグループ単位の週報の一覧が取得できる. 前述のアクセス制御機能を利用すると, この週報を参照専用で公開することも容易である. 複製を作成しないため, 訂正や他のディレクトリへ登録する必要がある場合は, オリジナル週報のストレージユニットを修正するだけでよい. 従来であれば, データベースとカスタムアプリケーションを組合せて実現するような機能が, このストレージでは基本機能を利用するだけで実現できる. つまり, ファイルマネージャに相当する基本ソフトの操作を覚えるだけで, 簡易なデータ共有アプリケーションを開発し利用できる.

このように DBMS を導入するには及ばないが, 共有ファイルを利用すると融通が利かないというレベルの言わば「B 級データ」の管理には適したプラットフォームと考えられる. たとえば, オフィスにおいて日々生産される「B 級データ」の量は膨大だが現在のところ死蔵されるケースも多く, 将来において有効活用の道が開けると新たな価値を生む. その為のプラットフォームは本ストレージが役立つのではないか？と考えている.

3. まとめ

この研究では, 主メモリ活用型のストレージシステムを提案した. 本ストレージの特徴として, コンテンツから検索可能な簡易なアクセスパス機構の提供, 広域分散ユーザでデータ共有可能なインターネット対応のアクセスプロトコル, スプレッドシートによる判り易いインターフェース, 負荷分散と相互バックアップを兼ねた柔軟なサーバ構成などを挙げることができる. これらの特徴は, どこに居ても自分のデータへの透過的なアクセス環境を提供すると言う意味で, ユビキタス社会のデータ共有プラットフォームとしても良い性質を有する.

参考文献

[1] Andrew S. Tanenbaum, Operating Systems: Design and Implementation, Prentice Hall, 1987

[2] 高林 哲ほか, 日本語全文検索システムNamazu, http://www.namazu.org/

[3] Jim Gray and Andreas Reuter, Transaction Processing: Concepts and Techniques, Morgan Kaufmann, 1993
[4] (株)東芝, GigaBase(主記憶データ管理 MW), http://www.toshiba.co.jp/efort/market/GIGASOLUTION/g_base.htm
[5] XML Protocol Working Group, SOAP Version1.2, http://www.w3.org/2000/xp/Group/