仮想一元化NASシステムX-NASの同期バックアップ機能の実現と評価

保田淑子[†] 川本真一[†] 江端淳[†]
沖津潤[†] 樋口達雄[†]

クラスタ型NASシステムの信頼性を向上させるため，ユーザがNAS上のファイルを作成および更新するのに同期して，ファイルブロック単位でバックアップNAS上のファイルも作成および更新可能なNASの同期バックアップを提案した．本同期バックアップは，汎用NASをバックアップNASとして利用できるよう，汎用ファイルアクセスプロトコルであるNFSを用いて実現した．さらに，同期バックアップ処理にともなうオーバヘッドを削減するため，レプリケーションキャッシュ，レプリケーションスレッドと部分非同期処理機能を開発した．同期バックアップの有効性を評価するため，NFSv3ベースの仮想一元化NASシステムX-NASのプロトタイプに同期バックアップ機能を搭載し，ファイルサーバ標準ベンチマークとバックアップ用途を想定したファイルコピープログラムを用いて性能評価を行った．評価の結果，同期バックアップにより，ファイルサーバ標準ベンチマークではバックアップ機能のないX-NASの80%の性能を，ファイルコピープログラムではバックアップ機能のないX-NASの70%の性能を維持しつつ，バックアップNASの導入コストを抑えて信頼性を向上させられるクラスタ型NASシステムを提供できることを実証した．

An On-line Backup Function for a Clustered NAS System (X-NAS)

Yoshiko Yasuda,[†] Shinichi Kawamoto,[†] Atsushi Ebata,[†]
Jun Okitsu[†] and Tatsuo Higuchi[†]

An on-line backup function for X-NAS, a clustered NAS system designed for entry-level NAS, has been developed. The on-line backup function can replicate file objects on X-NAS to a remote NAS for each NFS operation in real-time. It makes use of the virtualized global file system of X-NAS, and sends NFS write operations to both X-NAS and the remote backup NAS at the same time. To reduce the overhead of the on-line backup function, a replication cache, a replication thread, and a partially asynchronous method have been developed. The performance of the on-line backup function was evaluated and the evaluation results show that the on-line backup function of X-NAS improves the system reliability while maintaining from 70% to 80% of the throughput of the X-NAS without this function.

1. はじめに

NAS（Network-Attached Storag）は，急増する電子ファイルを効率良く管理するためのIPネットワークに接続可能なストレージ装置である．Windows[☆]やLinuxといった様々なOSの稼動するクライアント計算機間でファイルを共有できることから，ファイルサーバやバックアップ用途として市場で広く使われてきている．なかでもエントリNASは，安価で管理も容易なことから，スモールオフィスや企業部門において導入が進んでいる．

我々は，エントリNASからミッドレンジNAS市場を狙った，安価で使い勝手が良く，かつ管理容易でディスク容量を簡単に拡張可能なクラスタ型NASシステムX-NAS（eXpandable NAS）を提案している[1)~9)]．X-NASは複数の安価なエントリNASをクラスタ化し，それらを仮想一元化ファイルシステムにより一元管理することで，ユーザおよび管理者に対して仮想的に1つのNASに見せることができる．また，新規NASをX-NASに追加することで，X-NAS全体のディスク容量を簡単に拡張できる．

† 株式会社日立製作所中央研究所
Central Research Laboratory, Hitachi Ltd.

☆ Windows および DFS は米国 Microsoft Corporation の登録商標．NSI と Double Take は米国 Network Specialists, Inc の登録商標．その他，記載されている会社名，製品名は，各社の登録商標．

しかしながら，X-NAS ではディスクコントローラ等が故障し，X-NAS の構成要素であるエントリ NAS に障害が発生すると，その NAS 上のファイルが消失する．障害によるファイルの消失を防止するには，各要素 NAS 上のファイルの複製をバックアップとしてあらかじめ作成する必要がある．

一般に，X-NAS の構成要素であるエントリ NAS は，信頼性を向上させるために，2 台から 4 台のハードディスクを備え，RAID 機能[10] をサポートしている．たとえば，RAID5 で構築されたエントリ NAS の場合，1 台のハードディスクが故障しても，ユーザはファイルにアクセスし続けることができる．しかしながら，RAID5 は単体 NAS 内に構築されるため，NAS に搭載される OS や RAID コントローラに障害が発生すると，全ファイルが消失してしまう．このような理由により，エントリ NAS ユーザは通常，RAID 機能による冗長化に加えて，あらかじめ NAS にバンドルされたバックアップソフトウェアを用いて，リモートの別システムやリモートの NAS にバックアップを作成している．

しかしながら，NAS 上のファイルの複製をバックアップ先のシステムや NAS（バックアップ NAS）に作成する処理は，ユーザによるファイルの作成および更新作業とは同期せずにスケジュールベースで行われ，かつアーカイブ形式でバックアップ NAS に保持されるのが一般的である．そのため，最新ファイルの複製がバックアップ先に作成されていない状況で最新ファイルを保持する NAS に障害が発生すると，そのファイルが消失してしまう．また，アーカイブ形式のバックアップの場合，ツールを利用してリストアしない限り個々のファイルにはアクセスできないため，障害発生時に，即座にバックアップ NAS 上のファイルにアクセスできない．

上述の問題に対し，複数のファイルサーバ間でファイルディレクトリツリーを維持したまま，リアルタイムにファイルのバックアップを作成する手法が提案および製品化されているが[11)~13)]，複数の NAS で構成されるクラスタ型 NAS システムには対応していない状況である．また，これらの手法ではファイルサーバ間でプロプライエタリなプロトコルを使用するため，バックアップ NAS として使用できる NAS の種類が限定される．いいかえると，プロプライエタリなプロトコルをサポートしたバックアップ NAS を新たに導入する必要があり，コスト高である．もし，ユーザがすでに保有している汎用 NAS をバックアップ NAS として使用できれば，導入コストを抑えて信頼性の高いシステムを構築できる．

そこで本稿では，汎用のプロトコルを用いて，ユーザがクラスタ型 NAS システム上のファイルを作成および更新するのに同期してバックアップ NAS 上のファイルも作成および更新可能な同期バックアップ方法を提案する．そして，提案した同期バックアップの有効性を検証するため，クラスタ型 NAS システム X-NAS に同期バックアップを実装し，X-NAS プロトタイプを用いて同期バックアップの性能を評価した結果について述べる．

2. NAS の同期バックアップ

2.1 従来のバックアップ手法の問題点

従来の NAS のバックアップ手法では以下に示す問題点があった．

- バックアップの粒度・頻度
  従来の NAS を用いたバックアップは，ユーザによるファイルの作成および更新作業とは同期せずにスケジュールベースで行われていた[14)]．そのため，最新ファイルを保持する NAS に障害が発生すると，そのファイルが消失してしまう．また，バックアップ機能を使用することにより性能が低下する．
- バックアップ NAS の制御方法
  従来の NAS 用バックアップソフトウェアでは，バックアップ NAS に特別なエージェントソフトウェアを搭載し，プロプライエタリなプロトコルで通信する[11),13)]．そのため，バックアップ NAS の種類が限定されるという制約がある．
- バックアップの設定方法
  従来の NAS を用いたバックアップでは，クライアント計算機にもバックアップソフトウェアを搭載する必要があった．そのため，バックアップ機能をシステムに追加する場合にはクライアント計算機の設定を変更する必要がある[14)]．その結果，運用管理手順が複雑化し，管理コストが増大する．

そこで，これらの問題点すべてを解決するバックアップ方法を開発することが課題となる．このバックアップ方法を**同期バックアップ**と呼ぶ．

2.2 同期バックアップの定義

ランダムアクセスやファイルブロック単位での高速読み書き等，NAS の特長を活かすことにより，ユーザがファイルを作成および更新するのに同期してバックアップ NAS 上のファイルも作成および更新するバックアップ方法を NAS の同期バックアップと定義する．

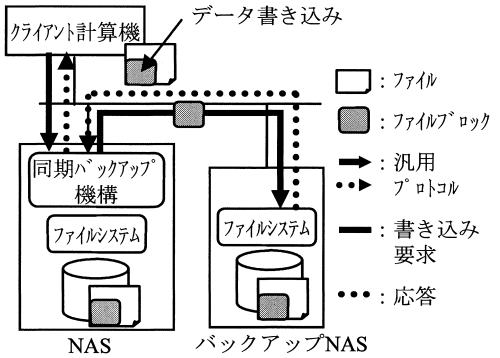

図 1 NAS の同期バックアップアーキテクチャ
Fig. 1 Architecture of on-line backup function.

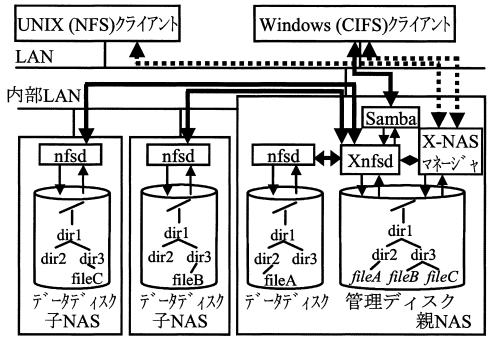

図 2 X-NAS の基本構成
Fig. 2 X-NAS overview.

2.3 同期バックアップのアーキテクチャ

図 1 に，NAS の同期バックアップのアーキテクチャを示す．NAS の同期バックアップでは，NAS 上のファイルの複製をファイルブロック単位でバックアップ NAS に作成する．これにより，最新のファイルの消失を防止し，リアルタイムにファイルの複製を作成できる．また，NAS とバックアップ NAS 間では汎用プロトコルを使用してデータをやりとりする．これにより，システム構成上の制約を排除し，ユーザが自由にバックアップ NAS を選択できる．さらに，バックアップ機能を NAS に追加する場合に，クライアント計算機側の設定を追加変更することなく，バックアップ機能つき NAS へと簡単に移行させる．これにより，システムの運用管理を容易にし，管理コストを低減できる．

3. X-NAS の同期バックアップ機能

2 章で述べた NAS の同期バックアップの有効性を検証するため，この機能を我々が提案する仮想一元化 NAS システム X-NAS に搭載することを検討した．本章では，X-NAS の基本構成と処理方式について簡単に述べた後，X-NAS における同期バックアップ機能について説明する．

3.1 X-NAS の基本構成と処理方式

図 2 に X-NAS の基本構成を示す．X-NAS は，親 NAS と複数の子 NAS で構成する．親 NAS および子 NAS の OS は標準 Linux である．子 NAS は，UNIX 系のファイル共有プロトコル NFS を提供する nfsd [15] とデータディスクを含む．各データディスクは，UNIX 系の標準的なファイルシステムにより管理され，ファイルの実体を保持する．親 NAS は，nfsd およびデータディスクに加えて，X-NAS 特有の構成要素（管理ディスク，Samba [16]，Xnfsd，X-NAS マネージャ）を持つ．

管理ディスクは，クライアント計算機と同一のファイルシステムビューを持つ．ただし，管理ディスク上のファイルはサイズが 0 バイトのダミーファイルである．ダミーファイルは，ファイルの実体が保持されている要素 NAS のデータディスクを特定するために使用する．Xnfsd は，マルチスレッド対応の NFS サーバのラッパーデーモンである．Xnfsd は，nfsd に代わりクライアント計算機が発行したファイルアクセス要求を受け付け，管理ディスクにアクセスし，ダミーファイルの inode 番号を利用して，その要求のアクセス対象ファイルの実体を保持するデータディスクを特定する [1]．そして，特定したデータディスクを搭載した要素 NAS にファイルアクセス要求を転送し，その要素 NAS 上で稼動する nfsd に要求を処理させる．

X-NAS マネージャは，X-NAS の管理機能を提供する．X-NAS の管理機能には，要素 NAS を追加あるいは削除することにより X-NAS のディスク容量を増減するオンライン拡張/縮退機能，要素 NAS 間でディスクの残容量を平準化する自律容量リバランス機能がある．各機能の詳細は文献 2) および 4) を参照していただきたい．

3.2 X-NAS における同期バックアップ機能の検討

X-NAS では，複数 NAS の仮想一元化を NFS サーバのラッパーデーモン Xnfsd のレイヤで実現している．Xnfsd は，NFS クライアントにはシームレスに NFS プロシージャを受け付け，そのプロシージャを他の NFS サーバに転送できる．そこで，X-NAS における同期バックアップ機能を実現するため，バックアップ NAS にも NFS プロシージャを転送できるように Xnfsd の機能を拡張する．これにより，クライアント計算機にはシームレスに，かつ汎用の NFS プロトコルを使用して，プロシージャ単位で X-NAS とバックアップ NAS 上のファイルの内容を一致させることが

できる．

同期バックアップ機能の検討では，本機能をエントリ NAS ユーザがどのように使用するかを想定し，以下に示す 5 つの機能を実現した．

- オンラインレプリケーション
  X-NAS 上のファイルの複製をバックアップ NAS にオンラインで作成する．
- 新規同期化
  X-NAS 単体で運用していたユーザが，新しくバックアップ NAS を追加し，X-NAS 上の全ファイルの複製をバックアップ NAS にコピーして，オンラインレプリケーションを開始する．
- 運用切替え
  バックアップ NAS が故障により使用できなくなった場合あるいは X-NAS とバックアップ NAS 間のネットワークに障害が発生した場合，障害を検出してオンラインレプリケーションを停止し，X-NAS 単体で運用を継続する．また，X-NAS が故障してアクセスできなくなった場合には，障害復旧までの暫定処置として，管理者がバックアップ NAS の設定を変更しバックアップ NAS をユーザに公開する．
- 再同期化
  障害復旧後，X-NAS 上あるいはバックアップ NAS 上の最新ファイルをバックアップ NAS あるいは X-NAS に反映し，オンラインレプリケーションを再開する．
- ヘルスチェック
  X-NAS とバックアップ NAS の状態を定期的に監視し，障害が発生した場合には管理者に通報する．同時に，ディスク容量の異なる X-NAS とバックアップ NAS を効率良く管理する．

これらの機能のうち，オンラインレプリケーションは同期バックアップ機能が有効である間，つねに動作するため，NAS の運用時に最も影響を及ぼす．そこで，本稿ではオンラインレプリケーションにフォーカスする（その他の機能の詳細については文献 5) および 9) を参照していただきたい）．

3.3 オンラインレプリケーション

3.3.1 処　　理

オンラインレプリケーションは，新規同期化あるいは再同期化処理の実行により，X-NAS とバックアップ NAS 上のファイルディレクトリ構造が一致している場合に，書き込み系 NFS プロシージャ単位で，X-NAS とバックアップ NAS のファイルデータの内容を一致させる処理である．NFSv3 における代表的な書き込み系プロシージャとしては，WRITE，CREATE，RENAME，REMOVE，SETATTR，MKDIR，RMDIR，COMMIT，SYMLINK と LINK がある[15]．

オンラインレプリケーションは，X-NAS とバックアップ NAS のファイルの内容が完全に一致することを保証するため，書き込み系プロシージャ単位で X-NAS とバックアップ NAS からの応答を待ち合わせて同期をとる．これにより，バックアップ NAS 上に X-NAS 上の最新ファイルの複製を保持できる．

図 3 および **図 4** (b) にオンラインレプリケーションが有効である場合の WRITE プロシージャの処理の流れを示す．図 4 において，(i) は i 番目のファイルアクセス要求，(i+1) は i+1 番目のファイルアクセス要求を示し，(1)，(2)，(3)，(4) は処理内容を示す．たとえば，(i)-(1) は i 番目のファイルアクセス要求の管理ディスクアクセスを示す．Xnfsd は，NFS クライアントから書き込み系プロシージャを受付けると，以下のステップでオンラインレプリケーションを実現する．

(1) 親 NAS の管理ディスクにアクセスし，管理ディスク上のダミーファイルの inode 番号からアクセス対象ファイルの実体を保持するデータディスクを特定する．
(2) プロシージャをバックアップ NAS に転送し，バックアップ NAS 上の nfsd にそのプロシージャを処理させる．
(3) 特定したデータディスクを搭載した子 NAS にプロシージャを転送し，子 NAS 上の nfsd にそのプロシージャを処理させる．
(4) 子 NAS の nfsd とバックアップ NAS の nfsd からの応答を待ち合わせ，応答がそろった時点で NFS クライアントに返送する．

3.3.2 オーバヘッド削減機能

上述のように，オンラインレプリケーションでは，Xnfsd が NFS プロシージャ単位で，X-NAS とバックアップ NAS からの応答を待ち合わせて同期をとる．しかしながら，同期待合せが性能を大幅に低下させることは明白である．この問題点を解決するため，X-NAS の同期バックアップは，以下に示す 3 つのオーバヘッド削減機能を備える．

(1) レプリケーションキャッシュ
  バックアップ NAS 上のファイルにアクセスするためにはファイルハンドル（ファイルの識別子）が必要である．ファイルハンドルを特定するには，バックアップ NAS に対して LOOKUP

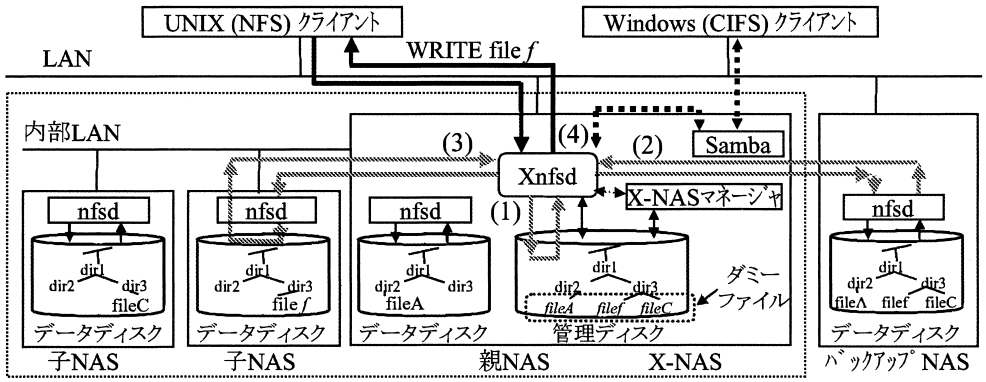

図 3 オンラインレプリケーション処理
Fig. 3 Flow of WRITE operation with on-line replication.

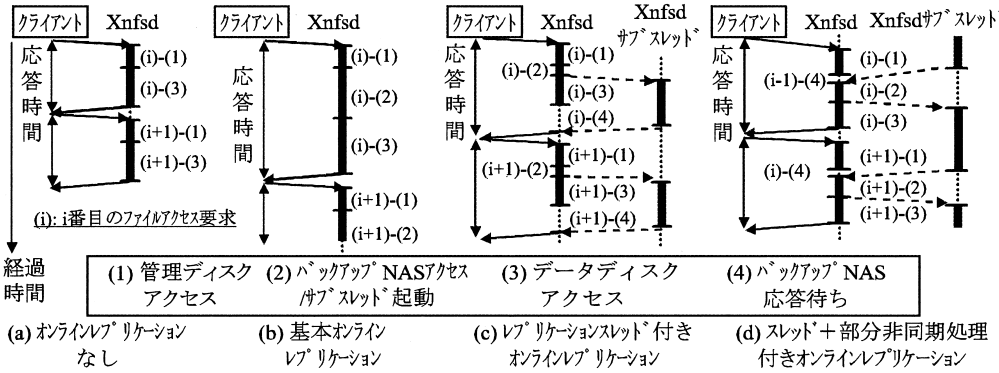

図 4 オンラインレプリケーション処理のタイミングチャート
Fig. 4 Timings of WRITE operations with on-line replication.

プロシージャをネットワーク経由で発行する必要があり，処理コストを無視できない．このコストを削減するため，X-NAS ではレプリケーション専用のファイルハンドルキャッシュを設ける．

ファイルハンドルキャッシュは，管理ディスク上のダミーファイルのファイルハンドル（グローバルファイルハンドル）とバックアップ NAS 上のファイルのファイルハンドルの対応を記録する．レプリケーションキャッシュにより，バックアップ NAS に LOOKUP プロシージャを発行しなくてもファイルハンドルを特定できるため，オーバヘッドを削減できる．

(2) レプリケーションスレッド

Xnfsd は，バックアップ NAS に対してネットワーク経由でファイルアクセス要求を転送し，バックアップ NAS からの応答を待ち合わせた後，次処理を行うためオーバヘッドが大きい．そこで，このオーバヘッドを削減するため，Xnfsd のメインスレッドは管理ディスクにアクセスした後，バックアップ NAS にファイルアクセス要求を処理させるためにレプリケーション専用にサブスレッドを起動する．

図 4 (c) にレプリケーションスレッドを使用した場合の書き込み処理フローを示す．この図に示すように，レプリケーションスレッドにより，X-NAS におけるデータディスクアクセスとバックアップ NAS に対するアクセスを並列に処理できるようになり，応答時間を改善することができる．

(3) 部分非同期処理

先に説明した同期待合せはシンプルな同期バックアップ方法であるが，オーバヘッドが大きい．バックアップ NAS からの応答を待たない完全非同期処理であれば，同期バックアップ機能のない X-NAS と同等の性能が得られる．ログは完全非同期処理を実現する方法の 1 つであるが[17]，エントリ NAS では搭載するメモリ量が数百 MB に制限されるため，ログを十分に確保することは難しい．そこで，X-NAS では，ログ

を使用することなく，同期待合せのオーバヘッドを削減する部分非同期処理を開発した．

図 4 (d) に部分非同期処理を行った場合のタイミングチャートを示す．部分非同期処理では，Xnfsd が，後続（i+1 番目）のファイルアクセス要求に対する管理ディスクアクセス処理後に，先（i 番目）のファイルアクセス要求に対するバックアップ NAS からの応答を待ち合わせる．Xnfsd のメインスレッドは，バックアップ NAS からの応答を待つ間に，次（i+1 番目）のファイルアクセス要求に対する管理ディスクアクセス処理を行うことができるため，メインスレッドの処理効率を向上し，バックアップ NAS からの応答を待ち合わせる時間を短縮できる．一方で，ある時刻において，X-NAS では i+1 番目のファイルアクセス要求が処理され，バックアップ NAS では i 番目のファイルアクセス要求が処理されるため，両 NAS で処理中のファイルアクセス要求は完全には一致しない．しかしながら，部分非同期処理では，i+1 番目のファイルアクセス要求のバックアップ NAS への転送と，i+1 番目のファイルアクセス要求に対する X-NAS のデータディスクアクセスは，先行する i 番目のファイルアクセス要求に対するバックアップ NAS からの応答を待ち合わせた後に行う．それにより，障害が発生した場合に，片側のディスクにのみデータが書き込まれてしまうのを防止できる．

4. 評　　価

X-NAS の同期バックアップ機能の有効性を評価するため，NFSv3 ベースの X-NAS プロトタイプに同期バックアップ機能を搭載し，性能評価を行った．

4.1 評価環境

表 1 に実験で使用した同期バックアップ機能つき X-NAS の評価環境を示す．評価では，X-NAS の実用的な台数が 4 台から 8 台程度[1] であることを考慮し，4 台構成の X-NAS と，1 台のバックアップ NAS を用いた．親 NAS は，nfsd とデータディスクに加えて，Xnfsd と Samba および管理ディスクを搭載する．バックアップ NAS と子 NAS は，標準的な NFS サーバであり，新規ソフトウェアを搭載しない．

表 1 実験環境
Table 1 Experimental environment.

評価機の構成	親 NAS+子 NAS × 3 台+バックアップ NAS
要素 NAS の仕様	OS: Red Hat 7.2 CPU: Pentium III 1 GHz Memory: 1 GB HDD: 36 GB(Ultra160 SCSI 10000 rpm) NIC: 100 Mbps

4.2 評価プログラムと評価指標

X-NAS は，オフィスや企業部門において主にファイルサーバやバックアップストレージとして使用されると考えられる．ファイルサーバとバックアップストレージでは，ファイルアクセス形態が異なるため，これら 2 ケースについて同期バックアップ機能つき X-NAS の性能を評価した．

ファイルサーバの性能評価では，UNIX 系のファイルサーバ標準ベンチマーク SPECsfs97[18] と Windows 系ファイルサーバ標準ベンチマーク NetBench[19] を使用した．これらのベンチマークでは，UNIX 系のサイトあるいは Windows 系のサイトにおけるファイルサーバ実運用時の統計データに基づき，ワークロードミックスがあらかじめ規定されている．SPECsfs97 では，UNIX クライアントにおいてアクセス対象の NFS サーバに与える要求負荷（1 秒あたりに実行する NFS プロシージャ数）を段階的に増大させ，そのときの応答負荷（実効スループット）とプロシージャ全体の平均応答時間を測定する．NetBench では，Windows クライアント数を段階的に増大させた場合のスループットと応答時間を測定する．

X-NAS をバックアップストレージとして使用することを想定した評価では，ファイルの書き込み性能が重要になる．そこで，X-NAS に一定数のファイルをまとめて書き込む時間を測定する．ファイルセットとしては，以下の 2 つを使用した．

- ファイルセット 1
  オフィスで扱うドキュメントの平均サイズ（100〜300 KB）を解析用にモデル化したものであり，256 KB のファイルを 900 個．
- ファイルセット 2
  我々がレポート等のオフィスドキュメント作成用に使用しているクライアント計算機に保持されたドキュメント群を含む 221MB のサイズのフォルダ．コピー対象のフォルダは 4 個のサブフォルダを含みサブフォルダには 42 個のフォルダと 335 個のファイルが格納されている．

これらの評価プログラムを，以下の 4 つの構成で実行し，スループットと応答時間を測定した．また，オンラインレプリケーションの実験では，測定前に新規同期化処理を行い，X-NAS とバックアップ NAS のファイルディレクトリ構造を一致させた．

(1) 同期バックアップなし：
同期バックアップ機能なし X-NAS
(2) 同期バックアップあり（C1）：
同期バックアップ機能つき X-NAS（レプリケーションキャッシュのみ）
(3) 同期バックアップあり（C2）：
同期バックアップ機能つき X-NAS（レプリケーションキャッシュ，レプリケーションスレッドあり）
(4) 同期バックアップあり（C3）：
同期バックアップ機能つき X-NAS（レプリケーションキャッシュ，レプリケーションスレッド，部分非同期処理あり）

4.3 評価結果

4.3.1 ファイルサーバ評価結果

図 5 および**図 6** に上記 4 つの構成（同期バックアップなし/同期バックアップあり（C1）/同期バックアップあり（C2）/同期バックアップあり（C3））について，NetBench および SPECsfs97 を実行した場合のスループットと平均応答時間を示す．NetBench の場合，同期バックアップあり X-NAS のスループットは，C1，C2，C3 の差異によらず，同期バックアップなしケースの 83〜89%である．また同期バックアップあり X-NAS の平均応答時間は，同期バックアップなしケースに比べ 15%から 19%増加する．

SPECsfs97 の場合，C1 ケースのスループットは，同期バックアップなしケースの 77%にとどまる．しかしながら，レプリケーションスレッドを有効にすることで，3%スループットが向上し（図 6 (C2)），オンラインレプリケーションの全特長機能を有効にすることで，さらに 2%スループットが向上し，同期バックアップなしケースの 82%のスループットとなる（図 6 (C3)）．同期バックアップあり X-NAS における平均応答時間は，同期バックアップなしケースの 1.35 倍から 1.47 倍となっており，NetBench よりも同期バックアップの処理オーバヘッドによる性能低下の割合が大きい．

この原因は，ワークロードの差異にあると考える．ファイルサーバのベンチマークでは，読み出し要求と書き込み要求が一定の比率でクライアント計算機から発行されるが，NetBench は SPECsfs97 に比べて，データ読み出し比率が高い．X-NAS では，読み出し要求はバックアップ NAS には転送されない．したがって，読み出し比率の高い NetBench では，同期バックアップの処理オーバヘッドによる性能低下度合いが小さいと考えられる．同時に，同期バックアップの各特長機能による性能差も小さくなる．

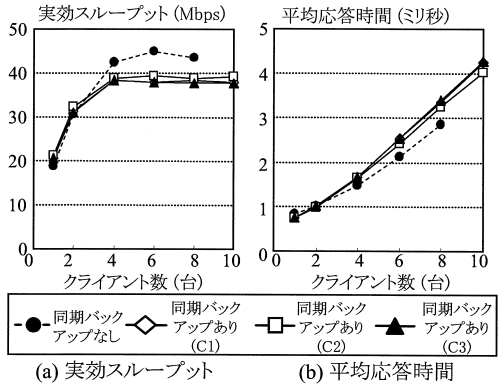

図 5 NetBench のスループットと平均応答時間
Fig. 5 Throughput and average response time of X-NAS with or without on-line backup function in the case of NetBench.

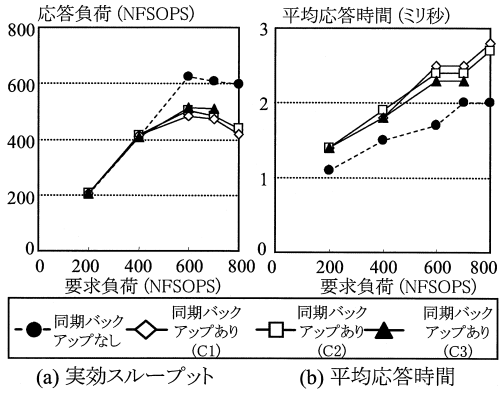

図 6 SPECsfs97 のスループットと平均応答時間
Fig. 6 Throughput and average response time of X-NAS with or without on-line backup function in the case of SPECsfs97.

以上をまとめると，ファイルの読み書きを行うファイルサーバ用途においては，同期バックアップあり X-NAS は同期バックアップなしケースの 80%以上のスループットを達成できることを確認した．また，ファイルサーバ用途でかつ書き込み比率が高い場合，レプリケーションキャッシュおよびレプリケーションスレッド，部分非同期処理の効果により，最大 5%性能を向上させられることを確認した．

4.3.2 バックアップストレージ評価結果

図 7 に，同期バックアップ機能つき X-NAS にファイルセット 1 および 2 をコピーした場合の書き込み時間とスループットを示す．通常バックアップストレージとして運用される場合，ファイル書き込みしか発生しないため，ファイルサーバに比べて同期バックアップの処理オーバヘッドが大きくなる．

同期バックアップあり X-NAS におけるファイルセッ

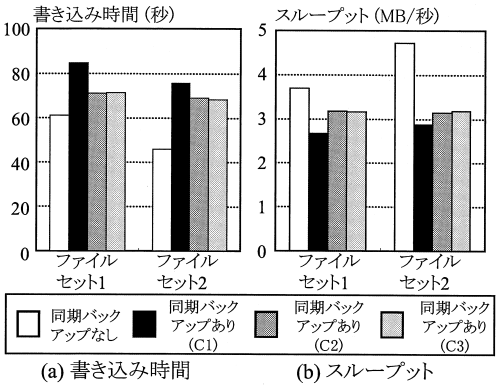

図 7 ファイルセット 1 および 2 の書き込み時間とスループット
Fig. 7 Times and throughput of X-NAS with or without on-line backup function for writing file sets.

ト 1 の書き込み時間は同期バックアップなしケースに比べて約 1.2 倍になっている（図 7 (a)）．また，C2 および C3 の差異による性能差は小さい．一方，ファイルセット 2 の書き込み時間は，レプリケーションキャッシュのみ有効にした同期バックアップ (C1) の場合，同期バックアップなしケースの 1.6 倍以上になっている（図 7 (b)）．よって，ファイルセット 2 は，ファイルセット 1 よりも同期バックアップのオーバヘッドが大きいといえる．この理由については，次節で考察する．同期バックアップあり X-NAS におけるファイルセット 2 の書き込み時間は，レプリケーションスレッドを有効にすることにより，同期バックアップなしケースの 1.54 倍まで短縮され（図 7 (a) の (C2)，部分非同期処理を有効にすると，さらに 1.48 倍に短縮される．

同期バックアップが有効な場合，ファイルセット 1 コピー時のスループットは，同期バックアップなしケースの 72%である（図 7 (b) の (C1)）．ファイルセット 1 では，レプリケーションスレッドの効果により，本機能がない場合に比べて性能が 13%向上する（図 7 (b) の (C2)）．ファイルセット 2 の場合，C1 ケースのスループットは，同期バックアップなしケースの 60%の性能に落ち込むが，レプリケーションスレッドを有効にすることにより，7%向上する（図 7 (b) の (C2)）．一方，部分非同期処理による性能向上はみられない．

4.4 考　　察

同期バックアップ処理のオーバヘッドが大きいファイルセット 1 および 2 のコピーと SPECsfs97 の結果について考察する．

まず，X-NAS 同期バックアップ実行時のファイルセット 1 とファイルセット 2 の書き込み性能の差について考察する．前節で述べたように，同期バックアップあり X-NAS の書き込み時間は同期バックアップなしケースに比べて，ファイルセット 1 の場合 1.2 倍，ファイルセット 2 では 1.6 倍である．**表 2** にファイルセット 1 およびファイルセット 2 の処理プロシージャの内訳を示す．ファイルセット 2 は，ファイルセット 1 よりも WRITE プロシージャの比率が 10%低く，代わりに MKDIR プロシージャや LOOKUP プロシージャの比率が高くなっている．最も比率の高い WRITE プロシージャの平均応答時間を調査したところ，いずれのファイルセットにおいても，同期バックアップなしケースでは約 1 ミリ秒，同期バックアップあり X-

表 2 ファイルセット 1 および 2 におけるプロシージャの内訳
Table 2 Difference of NFS procedures between fileset 1 and file set 2.

プロシージャ	ファイルセット 1	ファイルセット 2
WRITE	84.16%	74.92%
CREATE	2.63%	3.62%
MKDIR	0.00%	0.46%
LOOKUP	5.26%	8.88%
GETATTR	5.26%	6.80%
COMMIT	2.66%	3.90%
READDIR	0.02%	1.33%

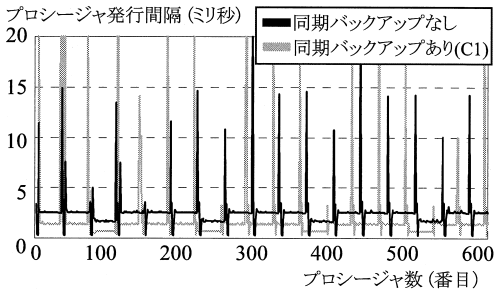

図 8 ファイルセット 1 書き込み時の X-NAS における NFS プロシージャ発行間隔
Fig. 8 Issue intervals of NFS operations on X-NAS with or without on-line backup function in the case of writing the file set 1.

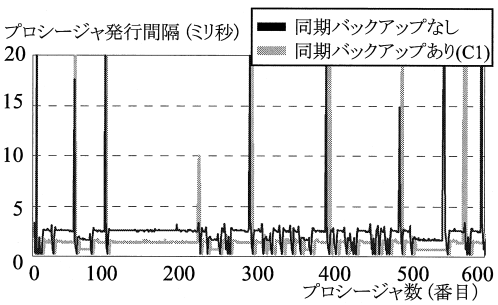

図 9 ファイルセット 2 書き込み時の X-NAS における NFS プロシージャ発行間隔
Fig. 9 Issue intervals of NFS operations on X-NAS with or without on-line backup function in the case of writing the file set 2.

NAS(C1) では 2 ミリ秒であり，ファイルセットの違いによる WRITE プロシージャの応答時間の差は見受けられなかった．

次に，ファイル書き込み時の Xnfsd における処理プロシージャの発行間隔を調査した．**図 8** および **図 9** にファイルセット 1 およびび 2 の結果を示す．ファイルセット 1 では，同期バックアップの有無によらず，一定の間隔でプロシージャの発行間隔が非常に長くなっている．ところが，このプロシージャの直前のプロシージャの応答時間を調査したところ，応答時間は延びていない．つまり，前のプロシージャの処理が完了しているにもかかわらず，後続のプロシージャの処理を開始できないタイミングが定期的に存在している．これは，バッファキャッシュのディスクへの書き込みオーバヘッドであると考えられる．パルス以外の部分では，同期バックアップの有無による WRITE プロシージャの処理時間の差が見えるが，ファイルセット 1 では，定期的なプロシージャ発行待ちが同期バックアップ有無の性能差を決定付けている．ファイルセット 2 では，プロシージャ発行待ち回数がファイルセット 1 よりも大幅に減少しており，同期バックアップの有無による WRITE プロシージャの性能差が，全体的な書き込み性能差の支配要因となっている．ファイルセット 2 は，ファイルセット 1 に比べて WRITE 比率が低いため，バッファキャッシュの書き込み待ちになる可能性が低く，その結果ファイルセット 2 の性能がファイルセット 1 よりも良くなっていると考えられる．

SPECsfs97 では，ワークロードミックスに占める WRITE プロシージャ比率が高いこと，WRITE プロシージャは 8KB のデータを持つことから，WRITE プロシージャの応答時間が全体に占める割合が高い．そこで，WRITE プロシージャの平均応答時間を調査した結果を **図 10** に示す．同期バックアップなし X-NAS の場合，WRITE プロシージャの平均応答時間は 3.7 ミリ秒である（要求負荷 600NFSOPS）．一方，C1 ケースの平均応答時間は同期バックアップなしケースの 2.3 倍となる．レプリケーションスレッドを有効にすることで平均応答時間は同期バックアップなしケースの 2.1 倍に改善され，さらに部分非同期処理の効果により同期バックアップなしケースの 1.5 倍にまで改善される．

最後に，ファイルセット 1 および 2 のコピープログラムと SPECsfs97 における WRITE プロシージャのプロファイルを調査した．プロファイルは，Xnfsd が WRITE プロシージャを受け付け，クライアント計算機に返送するまでの X-NAS 内部での処理の内訳である．

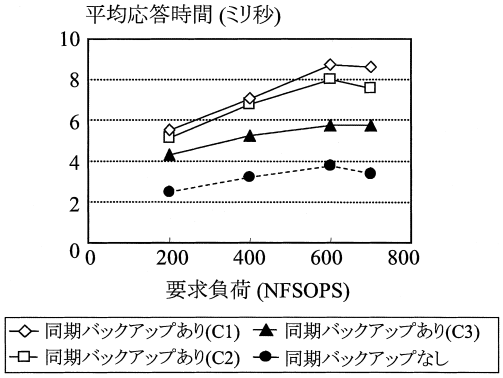

図 10 SPECsfs97 における WRITE プロシージャの平均応答時間

Fig. 10 Average response times for WRITE operations in the case of SPECsfs97.

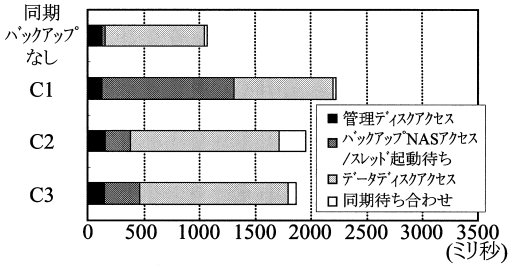

図 11 ファイルセット 1 をコピーした場合の WRITE プロシージャのプロファイル分析

Fig. 11 Profiling result of WRITE operations in the case of writing file set 1.

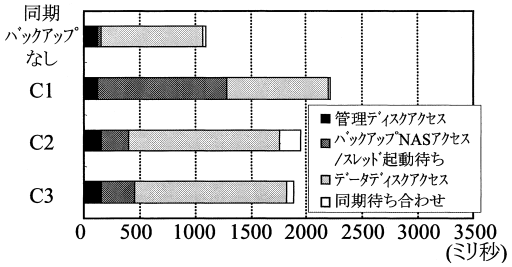

図 12 ファイルセット 2 をコピーした場合の WRITE プロシージャのプロファイル分析

Fig. 12 Profiling result of WRITE operations in the case of writing file set 2.

図 11 および **図 12** にファイルセット 1 および 2 を X-NAS に書き込んだ場合の WRITE プロシージャのプロファイル分析結果を示す．いずれのファイルセットでも，C1 ケースの WRITE プロシージャ処理時間は同期バックアップなしケースの 2 倍となっている．C2 ケースでは，レプリケーションスレッドの効果に

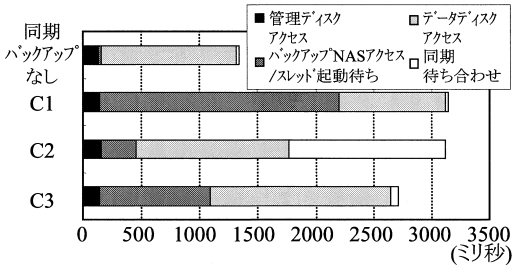

図 13 SPECsfs97 における WRITE プロシージャのプロファイル分析
Fig. 13 Profiling result of WRITE operations in the case of SPECsfs97.

より，WRITE プロシージャの処理時間が C1 ケースに比べて 12%短くなっている．

次に，**図 13** に SPECsfs97 を実行した場合の WRITE プロシージャのプロファイル分析結果を示す．同期バックアップなしケースでは，SPECsfs とファイルコピーの処理時間差は約 250 ミリ秒であるが，同期バックアップ実行中はその差が約 970 ミリ秒に広がっている（図 11 (C1) と図 13 (C1) を参照）．管理ディスクアクセス時間とデータディスクアクセス時間は，SPECsfs97 実行時もファイルコピー時も同等である．一方，SPECsfs97 ではバックアップ NAS に対するアクセス時間がファイルコピー時の 2 倍以上となっており，これが WRITE プロシージャの処理時間の 65%を占めている．このバックアップ NAS へのアクセス時間が長い理由は，ファイルコピーがディスクの連続アドレスに対する書き込みであるのに対して，SPECsfs97 における書き込みはランダムアクセスであることが原因であると考えられる．

またレプリケーションスレッドを有効にすると（図 13 (C2) を参照），バックアップ NAS へのデータ転送とデータディスクアクセスを並列に実行できるようになり，バックアップ NAS へのデータ転送処理時間が短くなる．その一方で，Xnfsd でのバックアップ NAS からの応答待ち時間の比率が WRITE プロシージャ処理時間の 43%となり（ファイルコピー時はこの比率が 10%），WRITE プロシージャの処理時間はほとんど改善されていない．次に部分非同期処理を有効にすると（図 13 (C3) を参照），応答待ち時間の比率が数%に低下し，それにともない，全体の処理時間も 14%短くなる．

以上の結果から，バックアップ NAS に対してディスク領域に連続的にアクセスする場合，レプリケーションスレッドが有効であり，ランダムに書き込みを行う場合には，部分非同期処理が有効に機能するといえる．

5. 関 連 研 究

複数の NAS をクラスタ化することにより，仮想的に大容量の NAS にする技術がすでに提案されている[20)~23)]．しかしながら，これらのクラスタ型 NAS システムではファイルのバックアップ方法については言及していない．一方，分散した複数の NAS あるいはファイルサーバ間でファイルのバックアップを作成する方法については製品化および提案されている．

DFS[14)] は，複数の Windows システム上に簡単にファイルの複製を作成する方法である．DFS は，DFS クライアントと DFS サーバを使用して Windows システム上のファイルやフォルダの複製を一定の間隔で別システムに作成する．

DRBD[11),12)] は，Linux ベースのシステムにおいて 2 つのノード間で HA クラスタを構築するためのカーネルモジュールである．DRBD は，プロプライエタリなプロトコルを使用して，2 つのファイルサーバ間でファイルおよびディレクトリの複製をブロック I/O 単位で作成する．この方法はブロック I/O ベースなので，バックアップサーバ上のファイルやディレクトリはリアルタイムに更新される．DRBD による同期バックアップ実行中のシーケンシャルライト性能は同期バックアップなしの場合に比べて 30%から 88%とかなりばらついている[24)]．

Double Take[13)] はマスタ NAS とスレブ NAS においてファイルの複製を作成するサードパーティのソフトウェアである．Double Take も DRBD と同様，マスタ NAS とスレブ NAS 間でプロプライエタリなプロトコルを使用する．

本稿では，クラスタ型の NAS システムである X-NAS において同期バックアップ機能を実現し，評価した．X-NAS の同期バックアップ機能は，スケジュールベースやファイル単位ではなく，NFS プロシージャ単位で X-NAS とバックアップ NAS 上のファイルのデータを一致させることができる．また，X-NAS とバックアップ NAS 間のデータ転送は汎用の NFS プロトコルを使用して行うため，バックアップ NAS に特別なエージェントを搭載する必要がなく，バックアップ NAS に様々な NAS を適用できる．そのうえ，プロプライエタリなプロトコルを利用する DRBD と比較しても同等のオーバヘッドで同期バックアップを実現できる．本機能により，バックアップ NAS の導入コストを低減させつつ，信頼性を容易に向上できる．

6. おわりに

エントリ NAS ユーザをターゲットとしたクラスタ型 NAS システムの信頼性を向上させるため，ユーザがファイルを作成および更新するのに同期して，ファイルブロック単位でバックアップ NAS 上のファイルも作成および更新可能な NAS の同期バックアップを提案した．

同期バックアップの実現においては，汎用 NAS をバックアップ NAS として使用できるように，標準ファイルアクセスプロトコルである NFS を使用した．さらに，クラスタ型 NAS システムとバックアップ NAS 上のデータ一致保証にともなう処理オーバヘッドを低減させるため，レプリケーションキャッシュ，レプリケーションスレッド，部分非同期処理機能を開発した．

同期バックアップ機能の有効性を検証するため，本機能を NFSv3 ベースの X-NAS プロトタイプに搭載しファイルサーバベンチマークプログラム NetBench および SPECsfs97，バックアップ用途を想定したファイルコピープログラムを用いて性能評価を行った．評価の結果，上記 3 つのオーバヘッド削減機能の効果により，ファイルサーバベンチマークプログラムでは，同期バックアップなし X-NAS の約 80%の性能を達成できることを確認した．また，バックアップ NAS に対する負荷が最も大きく，同期バックアップ性能の最悪ケースと考えられるファイルコピープログラムでは，同期バックアップなし X-NAS の約 70%の性能を達成できることを確認した．

本機能の導入により，エントリ NAS ユーザに対して，同期バックアップなし X-NAS の 70%から 80%の性能を維持しつつ，バックアップ NAS の導入コストを抑えて信頼性を向上できるクラスタ型 NAS システムを提供できる．

謝辞 本稿の執筆にあたり，匿名査読者諸氏から数多くの有益なコメントをいただいた．ここに感謝の意を表する．

参考文献

1) Yasuda, Y., et al.: Concept and Evaluation of X-NAS: A Highly Scalable NAS System, *Proc. IEEE/NASA MSST 2003*, pp.216–224 (2003).
2) 川本真一ほか：ファイル自律配置方式を備えた仮想一元化 NAS システム X-NAS の実現と評価，第 14 回データ工学ワークショップ 4-B-01 (2003).
3) Yasuda, Y., et al.: Scalability of X-NAS: A Clustered NAS System, 情報処理学会論文誌：コンピューティングシステム，Vol.44, No.SIG11, pp.68–78 (2003).
4) 川本真一ほか：仮想一元化 NAS システム X-NAS における自律容量リバランス機能の実現と評価，信学技報，Vol.103, No.248, pp.1–6 (2003).
5) 保田淑子ほか：仮想一元化 NAS システム X-NAS における同期バックアップ機能の実現と評価，信学技報，Vol.103, No.248, pp.7–12 (2003).
6) 江端 淳ほか：仮想一元化 NAS システム X-NAS の同期バックアップ実現に向けた順序制御方式の検討，FIT2003 論文集，pp.57–58 (2003).
7) Yasuda, Y., et al.: An On-line Backup Function for a Clustered NAS System (X-NAS), *Proc. 12th NASA/21th IEEE MSST 2004*, pp.165–170 (2004).
8) 保田淑子ほか：クラスタ型 NAS システム向きディスク使用量制限機能の提案と評価，情報処理学会論文誌：コンピューティングシステム，Vol.45, No.SIG11 (2004).
9) Yasuda, Y., et al.: RX-NAS: A Scalable, Reliable Clustered NAS System, 情報処理学会論文誌，Vol.46, No.1 (2005).
10) Patterson, D.A., et al.: A Case for Redundant Arrays of Inexpensive Disks (RAID), *Proc. SIGMOD'88* (1988).
11) Reisner, P.: DRBD, *Proc. 7th Linux Kongress* (2000).
12) Reisner, P.: DRBD Distributed Replicated Block Device, *Proc. 9th Linux Kongress* (2002).
13) NSI software: Double-Take Theory of Operations (2001). http://www.nsisoftware.com
14) Microsoft Corporation: Deploying Windows Powered NAS Using Dfs with or Without Active Directory (2001). http://www.microsoft.com
15) Callaghan, B.: *NFS Illustrated*, Addison Wesley, Reading, Massachusetts (2000).
16) Eckstein, R., et al.: *Using Samba*, O'Reilly and Associates, Inc (1999).
17) 安部洋平ほか：非同期バックアップにおけるログ制御手法と性能，第 14 回データ工学ワークショップ 1-B-01 (2003).
18) Standard Performance Evaluation Corporation: SFS3.0 Documentation Ver.1.0 (2002). http://www.spec.org
19) etestinglabs: *NetBench 7.0.3* (2002). http://www.etestinglabs.com
20) 山川 聡ほか：NAS スイッチ：NFS サーバの仮想化統合技術の開発，信学技報 CPSY2002-36 (2002).
21) Karamanolis, C., et al.: DiFFS: a Scalable Distributed File System, Technical Report HPL-2001-19, HP Laboratories Palo Alto (2001).
22) Karamanolis, C., et al.: An Architecture for Scalable and Manageable File Services, Tech-

nical Report HPL-2001-173, HP Laboratories Palo Alto (2001).
23) Anderson, D.C., et al.: Interposed Request Routing for Scalable Network Storage, *ACM Trans. Comput. Syst.*, Vol.20, No.1 (2002).
24) DRBD Performance (2004). http://www.drbd.org

(平成 16 年 7 月 20 日受付)
(平成 16 年 10 月 22 日採録)

保田　淑子（正会員）

1991 年早稲田大学理工学部卒業．同年（株）日立製作所入社．入所以来中央研究所に勤務．並列計算機，スーパーコンピュータ，オープンサーバ，ネットワークストレージの研究開発に従事．IEEE/Computer 会員．

川本　真一（正会員）

1991 年東北大学工学部卒業．1996 年同大学大学院博士課程修了．同年同大学院助手．1998 年（株）日立製作所入社．入所以来中央研究所に勤務．現在，主任研究員．オープンサーバ，ネットワークストレージ，高可用アプリケーションサーバの研究開発に従事．博士（情報科学）．

江端　淳

1993 年筑波大学基礎工学部卒業．1995 年同大学大学院修士課程修了．同年（株）日立製作所入社．入所以来中央研究所に勤務．大型計算機，ネットワークストレージの研究開発に従事．

沖津　潤

1999 年東京工業大学工学部卒業．2001 年同大学大学院修士課程修了．同年（株）日立製作所入社．入所以来中央研究所に勤務．オープンサーバ，ネットワークストレージの研究開発に従事．

樋口　達雄

1988 年東京大学工学部卒業．1990 年同大学大学院修士課程修了．同年（株）日立製作所入社．入所以来中央研究所に勤務．現在，主任研究員．並列計算機，オープンサーバ，ネットワークストレージの研究開発等に従事．電子情報通信学会会員．