R&D ホットコーナー ソリューション

格段に使いやすくなった高可用ファイルシステム NILFSv2

NTTサイバースペース研究所

もりあい　さとし
盛合　敏

NILFSはデータの可用性を高め，システムの異常停止後の復旧時間を短縮させるだけでなく，設定ミスや操作ミスなどによる障害からの復旧を容易にするために開発されたLinux用ファイルシステムです．ここでは，NILFSv2までに強化されたユーザインタフェースなどについて紹介します．

ストレージシステムの課題

ハードディスクの進化は目覚ましく，1台で1テラバイトの記憶容量のものを容易に入手できるようになりました．また，性能や信頼性も高まっています．しかし，ディスクやヘッドを高速に動かす機械式の装置であるため，コンピュータシステムの中で故障しやすい部品の1つとなっており，データセンタの障害の約半数がハードディスクの故障によるものといわれています．

データセンタ向けシステムでは，RAID（Redundant Arrays of Inexpensive Disks）によって信頼性の向上を図っています．しかし，同一ロットの部品でシステムを組み上げた場合，多重障害の起きる率が高いという問題があり，多重障害に対応できるシステムにするためにはかなりの投資が必要です．

コストを惜しまなければ，ハードウェアに起因する障害をほぼ防ぐことができますが，それでもシステム障害は発生してしまいます．重大なシステム障害は，人為的なミス，すなわち，設定ミスや操作ミスによることが少なくありません．こういった障害は防ぐことが難しいばかりでなく，完全な復旧は困難です．さらに，ストレージシステムの容量増大に伴い，データのバックアップ・リストアや（システム異常停止後の）データ整合性チェックの時間が増大し，円滑なオペレーションの阻害要因となってきています．

NILFS（The New Implementation of a Log-structured File System）は，システム異常停止後の復旧時

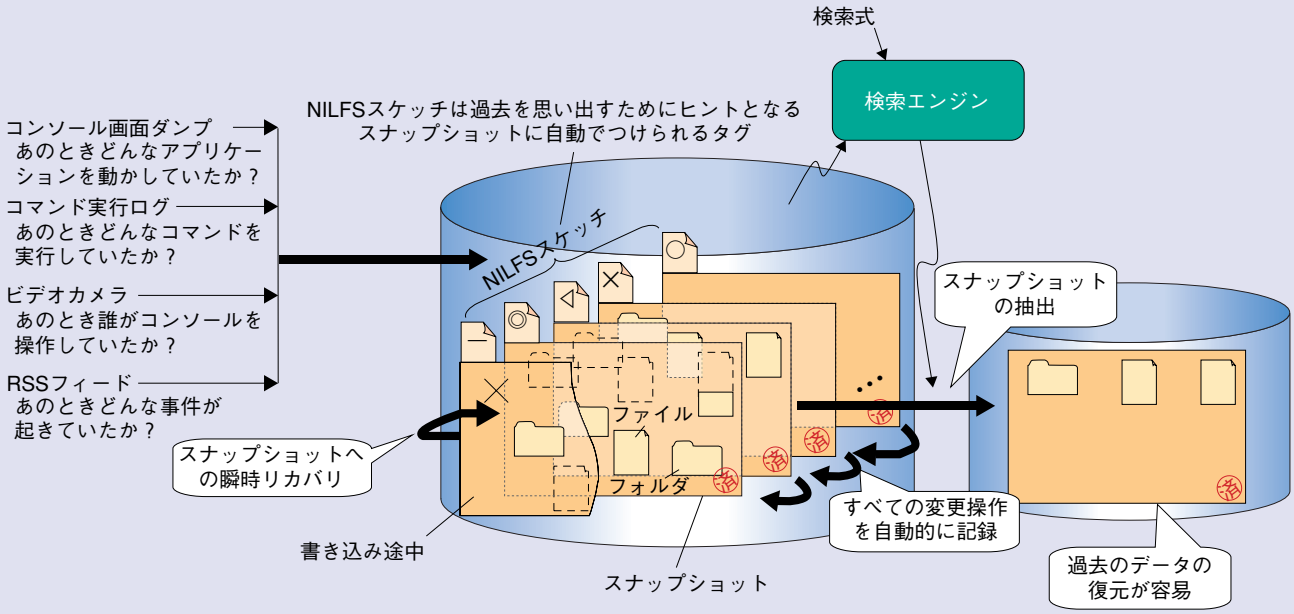


図1　NILFSの概要

間を短縮させるだけでなく，人為的なミスからの復旧を容易にすることを目的に開発されました．

スナップショットの取得と利用

NILFSは，Linux*1オペレーティングシステム用の新しいファイルシステムです（**図1**）．ファイルシステムとは，ハードディスクなどの記憶装置上のデータをファイルやフォルダ（ディレクトリ）といった単位で管理するプログラムで，オペレーティングシステムの一部を成すものです．ファイルやフォルダのデータとこれらを管理するためのデータ（メタデータ）の全体をファイルシステムのイメージと呼びます．NILFSは瞬間的なファイルシステムのイメージをスナップショットとして自動的かつ連続的に保存するファイルシステムです．そして，いつでも簡単に任意のスナップショットをアクセスできます．

スナップショットはそれが生成された時刻のファイルシステムのイメージを完全に矛盾なく再現するものです．つまり，ファイルやフォルダをどんなに作成・変更・削除しても，ある時刻のスナップショットをアクセスすれば，その時刻のファイルやフォルダを見ることができます．面倒なファイルの履歴管理が自動的に行えますし，過去のデータを容易に復元することができます．人為的なミスやソフトウェアのバージョンアップなどによってトラブルが発生したとしても，簡単に元の状態に戻すことができます．

PC上では削除したファイルを復活させるソフトウェアもあります．しかし，復活できるかどうかは運次第だったり，特定の時刻の状態にしか復元できなかったり，ファイルの移動や名前の変更には対応できなかったりと，任意の過去を正確に復元できるものはありません．このため，別媒体に定期的にバックアップをとることは重要なことですが，オンラインで完全なバックアップをとり，しかもバックアップ中にも記憶装置の入出力パフォーマンスに影響を与えないようにするためには，かなりのコストがかかります．

過去の検索

NILFSでは連続的にスナップショットが生成されるため，大量のスナップショットが生成されることがあります．このとき，過去の時刻を指定して，必要なファイルを探し出すのは容易ではありません．そこで，次のような手段を用意しました．

1つはNILFSスケッチです．これは，スナップショットごとに貼り付けられた付箋紙で，任意の情報を書き込めます．NILFSスケッチのデータはスナップショットに取り込まれてディスクへ書き出されます．NILFSスケッチには，探索の手掛かりとなるような情報，例えば，コンソール画面のサムネイル，コマンド投入ログなどを書き込むとよいでしょう．NILFSブラウザ（**図2**）やそれぞれのオペレーティングシステム標準のファイルブラウザ（**図3**）を用いてNILFSスケッチの一覧をブラウズしながら，目的の過去を探し出すことができます．

もう1つはNILFSサーチです．これは，スナップショットを含めてデスクトップ検索を行うもので，ファイルの過去の情報まで検索対象となります．NILFSブラウザの検索フィールドにキーワードを打ち込むことで，指定したキーワードをファイル名やファイルの内容に含むものを過去から探し出すことが可能となります．

また，オペレーティングシステム標準のデスクトップ検索エンジンとの連携も可能です．デスクトップ検索エンジンとしては，LinuxであればBeagleが利用できます．また，NILFSを使ったファイルサーバをWindows*2オペレーティングシステム上のネットワークドライブとして使う場合には，Windowsデスクトップサーチが利用できます（**図4**）．NILFSブラウザと同様に過去を検索することができます．

高可用なファイルシステム

NILFSはファイルシステム上のデータやメタデータの変更部分のみを次々とディスクへ追記します（**図5**）．書き込み中のディスクのヘッドの移動がほとんど不

図2　NILFSブラウザの動作例

*1　Linuxは，Linus Torvalds氏の日本およびその他の国における登録商標または商標です．
*2　Windowsは，米国Microsoft Corporationの米国およびその他の国における登録商標または商標です．

要となるため，高速なデータ書き込みが実現できます．また，データ書き込みの保証を行う同期型書き込みにおいても，従来のファイルシステムに比べて高い書き込み性能を得ることができます．

不意にシステムが異常停止した場合，再起動時には必ずファイルシステムのリカバリ処理が必要となります．リカバリ処理中はユーザにサービスを提供することはできませんので，可能な限りこの時間が短いことが望まれます．

古典的なファイルシステム（Linuxのext2やWindowsのFATなど）では，リカバリ処理としてファイルシステム全体の整合性チェックと矛盾の修復を行います．この処理はディスク全体のランダムアクセスとなり，ディスクの性能が極端に低下してしまいます．さらに，ディスク容量の増大に伴って，整合性チェックに長時間を要することが大きな問題となってきています．

より現代的なファイルシステム（Linuxのext3やWindowsのNTFSなど）では，メタデータの変更記録を「ジャーナル」と呼ばれる領域に保存しながら動作します．このため，リカバリ処理においては，ジャーナルを参照しながら，矛盾のあるデータを破棄または復旧することで，きわめて短時間で整合性のとれた状態とすることができます．しかし，ジャーナルにはファイルのデータの変更記録は保存されていないため，リカバリ処理によって，ファイルシステムとして正常に動作可能な状態とすることはできますが，システムの異常停止直前のデータを復元することはできません．

NILFSは，ログ構造化ファイルシステムとなっており，データとメタデータの変更記録を，適切なタイミングでディスクへ書き出します．1回の書き出しごとにチェックサムを計算し，その値も併せてディスクへ保存しています．このため，データを書き出した順に読み込んでチェックサムの確認を行うだけで，ファイルシステムのチェックが行えます．そして，チェックサムエラーとなるシステムの異常停止時の直前まで復元することができます．これらの処理は短時間に完了することができます．

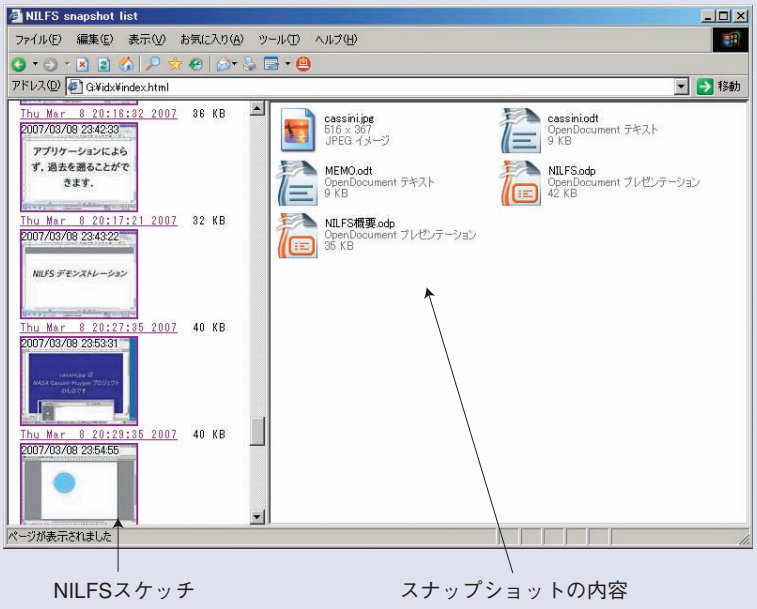


図3　ファイルブラウザを使ったスナップショットの一覧表示

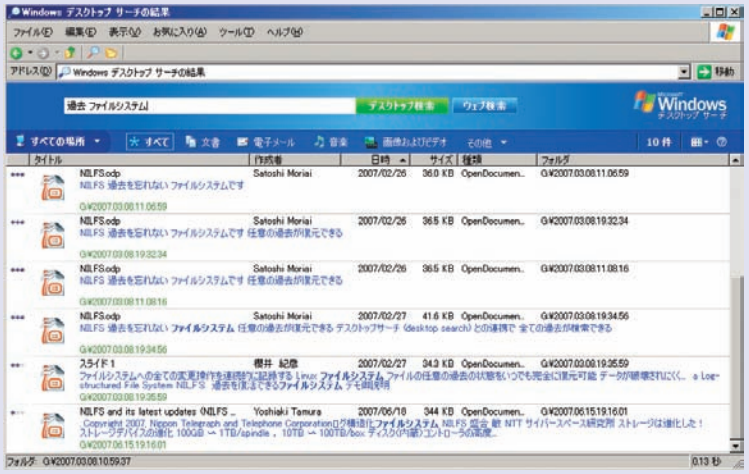
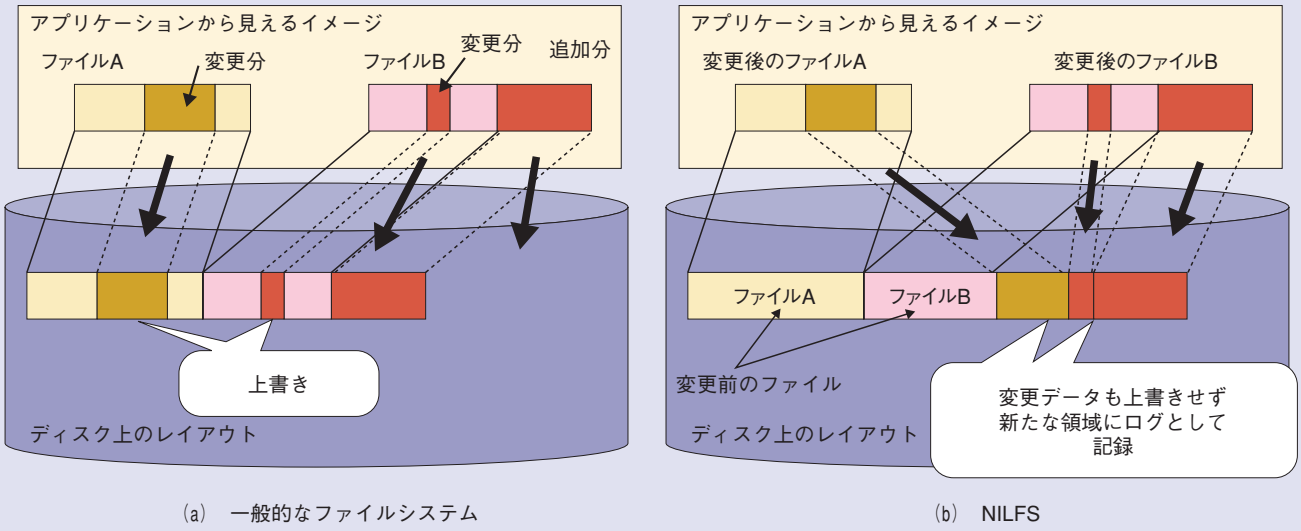


図4　デスクトップサーチを用いた過去の検索の例

マルチスナップショットGC

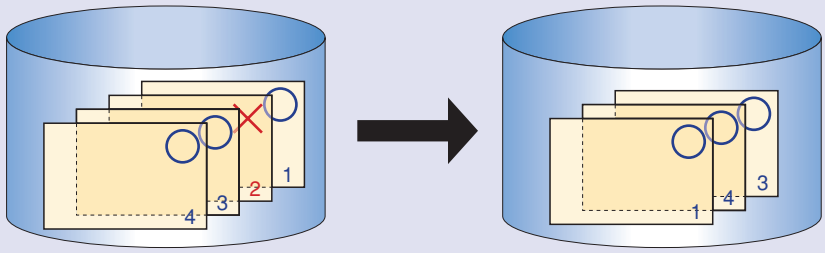
どんなにディスクの容量が大きくなったとしても，ディスクはいつか満杯になってしまいます．古くなればなるほど，スナップショットの間隔は粗くても良いでしょう．また，期限付きライセンスのコンテンツは期限がきたらシステム的にアクセスできないようにしなければなりません．つまり，不要なスナップショットを削除し，どのスナップショットからもアクセスされないデータを消去する機能があると便利です．

NILFSではオンラインで不要領域を回収するガベージコレクション（GC）の機能を備えています（**図6**）．他の類似の機能を持つファイルシステムでは保持で



(a) 一般的なファイルシステム　(b) NILFS

図5　一般的なファイルシステムとNILFSの比較



(a) GC起動前　(b) GC完了後

図6　マルチスナップショットGC

きるスナップショットの数が固定されていますが，NILFSでは，任意の複数のスナップショットをGCの対象から外して，それ以外のスナップショットをGCの対象とすることができます．そして，GCによって，利用中の領域に移動を行い，連続した空き領域を確保することができます．また，GC動作中も高可用ファイルシステムとして振る舞い，GC中にトラブルが発生した場合，GC起動前の状態に復旧することができます．

利用シーン

NILFSはLinuxカーネル2.6.11以降を採用したシステムで動作します．ファイルの世代管理が有用な場合，読み込み性能よりも書き込み性能が重要な場合，高信頼な書き込みが必要な場合であれば，システムの規模によらず，利用可能です．Linuxシステムの設定ファイルや実行ファイルの保存領域，各アプリケーションの領域，ログの領域などでの利用に適しています．また，Linuxを使ったSambaファイルサーバにNILFSを利用すると，Windows用の高可用ファイルサーバが安価に構築できます．

今後の展開

NILFSはオープンソースソフトウェアとして公開しており，Linuxの開発者やユーザの皆様の声をもとに安定化や機能拡充に取り組んでいきます．最新情報はhttp://www.nilfs.org/jaから入手できます．ソースコードだけでなく，メジャーなディストリビューションでそのままお使いいただけるパッケージやWindowsで動作するバーチャルマシン用のディスクイメージなども提供する予定です．

盛合　敏

人に優しい・ハードウェアに優しいシステムをスローガンに，今後も安価で信頼性が高く使いやすい技術の開発に取り組んでいきます．

◆問い合わせ先
NTTサイバースペース研究所
OSSコンピューティングプロジェクト
TEL　046-859-2982
FAX　046-855-1152
E-mail　moriai.satoshi@lab.ntt.co.jp