JAIST Repository
https://dspace.jaist.ac.jp/

Title	ファイルサーバの運用と課題について
Author(s)	小坂，秀一
Citation	国立大学法人北陸先端科学技術大学院大学技術サービス部業務報告集 ： 平成22年度: 11-14
Issue Date	2011-08
Type	Presentation
Text version	publisher
URL	http://hdl.handle.net/10119/10025
Rights	
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

ファイルサーバの運用と課題について

小坂 秀一

情報社会基盤研究センター

概要

情報社会基盤研究センター(旧情報科学センター)は 1990 年の開学より、利用者である学生教職員に対して世界最高水準の情報環境を提供し，教員の教育研究活動や学生の学習活動に資するため，等質かつ高レベルの情報サービスを展開する基盤の整備を進めている。

ファイルサーバシステムは情報環境の中でも、ユーザの全てのファイルを集中的に保存、管理する JAIST の情報環境の根幹に位置するシステムである。24 時間 364 日動作する可用性と共に最先端の環境を目指しできる限りトライアルなシステムを採用してきた。

それらのシステム中で fs1 は最もトライアルであると共に、運用上でファイルシステムが破損するという障害が重大な問題が発生するなど課題も多い。fs1 の特徴を紹介すると共に、障害の原因調査や再発防止への取り組みなどを紹介する。

1　ファイルサーバ群の概要

現在以下の 5 つのファイルサーバシステムを運用している。fs1 はこれらの中でも学生教職員(主に M1, M2 の学生)にディスク領域をサービスしている最も重要な位置づけに当たるファイルサーバである。

表 1.　運用中のファイルサーバ一覧

システム	サービス実効容量	用途	利用プロトコル	ファイルシステム
fs1	150TB	学生教職員のホームディレクトリ	NFSv4, CIFS	ZFS
fs2	100TB	学生教職員のホームディレクトリ	NFSv3, CIFS	StorFS
fs4	908TB	プロジェクト、大容量ファイル領域	NFSv3, CIFS	GPFS
fs7	266TB	ディスク領域の追加	iSCSI	NTFS, ZFS 等
fs8	12TB	事務職員のホームディレクトリ、共有フォルダ	CIFS, NFSv3	CFS

2　高速大容量ファイルサーバシステム fs1 の特徴について

高速大容量ファイルサーバシステム fs1 は学生、教職員用のホームディレクトリをサービスすることを目的に 2009 年 3 月から運用を開始したシステムである。fs1 は特に可用性が求められるファイルサーバシステムとしては、技術的にトライアルな部分が多いシステムである。

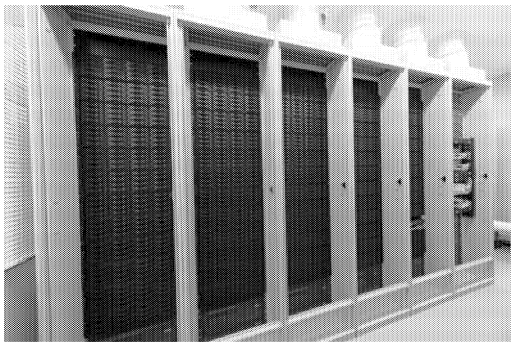

図 1.高速大容量ファイルサーバシステム fs1

2.1　仮想化による柔軟なボリューム構築

fs1 では従来までの lun 単位で構成されるボリュームではなく、Oracle Solaris ZFS および DELL EaqulLogic の仮想ボリュームの採用によりボリュームの仮想化を行っている。

ZFS ではストレージプールのサイズの範囲の中で、ボリュームのサイズ(=quota)を運用中に自由に変更できる。エンドユーザにはこの quota サイズがファイルシステムのサイズとして見えているため、運用中に quota サイズを変更すると一瞬でファイルシステム自体が増えたように見える。

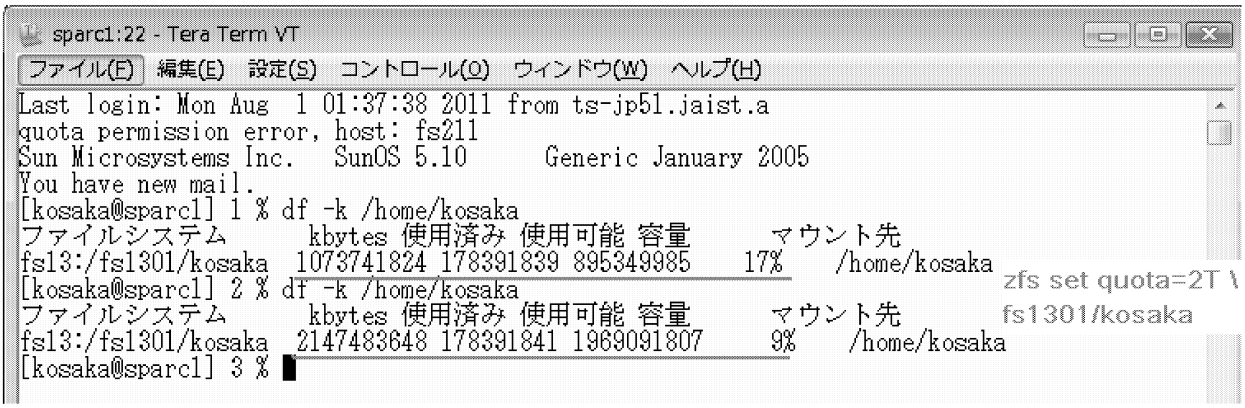

図 2. Quota サイズを 1TB から 2TB に変更した場合の df の出力結果の変化

fs1 では事故防止の観点から quota サイズを 1TB に設定しているが、ユーザからの領域の追加のリクエストにすぐに答えられるようになっている。

また、シン・プロビジョニングという仮想ボリュームの技術によりディスクの容量の仮想化を行っている。このシン・プロビジョニングを利用すると実際に割り当てる物理容量よりも大きなディスク容量を仮想的に設定できる。

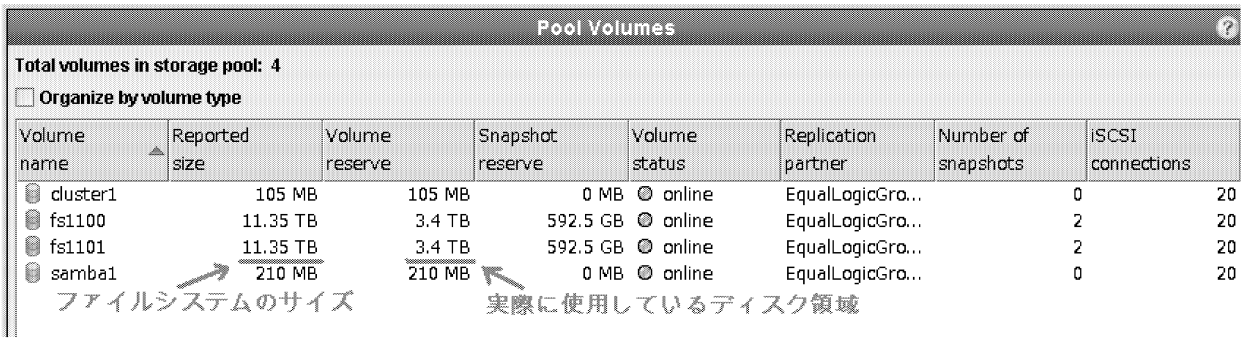

図 3. ファイルシステムのサイズと実際に使用しているディスク領域

fs1 ではディスク装置のファームウェアのバージョンアップに活用されている。ディスク装置を運用中にサービスから一旦外し、ファームウェアをアップグレード後に再びサービスに戻す手法を取っている。この時一時的にではあるがディスクの物理容量よりもファイルシステムのサイズが大きい状態になっている。

2.2　NFSv4, Kerberized NFS 対応

NFSv4 および Kerberized NFS に対応することにより本学が NFS をサービスを継続するうえで重要な 2 点のセキュリティ上の課題を解決できた。

- ACL(Access Control List)により CIFS/NFS 間で透過的アクセス権限設定が可能になった。本学では Windows のシステムからも Unix のシステムからも同じボリューム・ファイルを参照させているため Windows システム上のアクセス権と Unix システム上でのアクセス権の整合性が課題であった。ACL が利用できるようになることでほぼ Windows でのアクセス権=Unix でのアクセス権を実現できた。
- Kerberized NFS により NFS サービスのセキュリティが協力になりました。データ通信が暗号化されると共に、Kerberos による認証で適切なアクセス権を確保できるようになりました。

2.3 ストレージエリアネットワークに iSCSI を採用

一般的に SAN(Storage Area Network)はファイバチャネルを使われて組まれることが多いが、fs1 ではファイバチャネルの代わりに iSCSI を採用した。これにより一般的なイーサネット用のスイッチングハブが利用でき、別途運用している JAIST ネットワークと同様に扱うことができるため管理運用コストの削減が期待できる。

3 ファイルサーバ fs1 の問題点と改善について

fs1 は先進的なシステムであるが、一方で問題点もいくつかあり、4 月にはファイルシステムの 1 つが破損し過去のバックアップからデータを復旧したという重大な障害が発生した。その障害の原因を調査するとともにいくつか対策や運用の改善を行った。

3.1 障害の発生

2011 年 4 月 14 日(木)19:05 頃知識科学研究科の学生のデータを収容しているグループ 3 がフェイルオーバーした。通常は移動した先でサービスが起動する設計になっているが、M2 の学生のデータを収容しているボリューム fs1300 が破損したためサービスが再開できない状態になった。また、M1 の学生のデータを収容しているボリューム fs1301 も設計上 fs1300/fs1301 の両方が online にならないとサービスしない設計になっているためこちらも参照できない状態になった。

3.2 サービスの仮復旧について

サービスの復旧は破損したボリューム fs1300 の復旧の可否や時期が不透明であったため、まずバックアップデータを利用してサービスを仮復旧することとなった。しかし新たに fs1300/fs1301 のボリュームのバックアップは 3 月 18 日から止まっていたことが新たに判明した。停止していた原因は 3 月 18 日に保守業者が行ったメンテナンスの際に一旦停止させていた設定戻し忘れが原因だったが、そのため 4 月 16 日(土)16:30 頃に 3 月 18 日時点のバックアップデータでのサービスの再開することとなった。

3.3 破損したファイルシステムからのデータの復旧

破損したファイルシステムからのデータの復旧はサポート業者への解析依頼と並行して、Solaris 10 以外で ZFS をサポートしている OS(FreeBSD や Solaris 11 等)でのインポートができないか試みた。その結果、ReadOnly だが Solaris 11 でインポートし、ファイルシステム内のユーザのファイルを読めるようになった。

3.4 最終的な復旧

最終的な復旧作業を 4 月 25 日(月)に行った。この時点でユーザのデータは以下の 2 つにわかれて保存されている。

- 元々のファイルシステム上の 4 月 14 日 19 時までのデータ (以下、データ B とする)
- 3 月 18 日時点のバックアップデータをベースに 4 月 16 日から 4 月 25 日までにユーザが作成した

データが加わったデータ(以下、データ B とする)

どのように 2 つのデータをユーザに公開するか検討した結果、ファイルシステムのデータを再度障害発生した 4 月 14 日 19 時のデータ A に戻し、データ B から仮復旧期間中に生成されたデータのみを抽出し、データ A 上にあるユーザのデスクトップフォルダにコピーすることにした。作業手順を以下の通り行った。

1. ファイルシステムを障害発生時(4 月 14 日 19 時)のデータ A に戻す
2. 仮復旧中のデータ B の中から 4 月 16 日から 4 月 25 日の間に更新があったファイルのみを抽出する
3. データ B から抽出したデータのうち Windows で利用しているデータ(~/.windows 以下)を Windows 環境のユーザのデスクトップフォルダにコピーする
4. データ B から 3 の手順でコピーした Windows で利用しているデータを間引く
5. 4 で生成したデータを Unix 環境のユーザのデスクトップフォルダにコピーする

3.5 ファイルシステムの破損の原因

ファイルシステムの破壊の原因を調査するために別システムで再現試験を行った。現在の設定ではフェイルオーバー時にフェイルオーバーした先のホストの活性時にファイルシステムのインポートや強制インポートを行った際にエラーが発生した場合には再度フェイルオーバーを試みる設定になっている。その際にフェイルオーバーした元のシステムは panic リブートすることでインポート処理が停止し、2 重インポートが防げていると考えていたが、再現試験ではファイルシステムの破損を確認することができた。

3.6 システムや運用の改善

今回の障害を受けてシステムの設計の再見直しを行い、下記の項目の改善を実施や検討を行っている。

- これまではフェイルオーバーした先での活性時にエラーが発生した場合には、再度フェイルオーバーを試みたが、ファイルシステムの破壊を招く可能性があるため、活性時に ZFS のインポートエラーが発生した場合にはフェイルオーバーせずに停止する設定に変更した
- フェイルオーバーした際にボリュームの破壊が発生しても復旧できるよう、フェイルオーバーした際に ZFS 上で snapshot を実施するように変更をした
- メンテナンスの実施手順を事前に作成してもらい大学側でも作業内容の確認したり、作業後の確認作業を行うようにした
- レプリケーションによるバックアップの設定が無効になっていないか確認するチェックスクリプトを定期的に実行するようにした
- レプリケーションによるバックアップの日時がすべてのグループで一斉に行われていたが、30 分ずつずらすことで SAN 部分のネットワークの流量が分散するようスケジュールを見直した
- NAS 全体の統計情報が取れるよう Dell EqualLogic SAN HeadQuoters や Zabbix などの運用やその準備を行っている

4 まとめ

今回の障害を受けて前述のような改善を行ったが、これらの多くは運用の開始前に対策されるべきだった内容である。運用開始前に JAIST と納入業者の両者がシステムの構成の確認や運用テストを行い、見直しや修正を充分行っていれば今回の様なファイルシステムの破損という重大な障害は防げたと考えている。それには納入業者の言うことを鵜呑みにするのではなく、我々運用する側である大学の職員が関連ドキュメントを深く読み、システムの構成や動作をしっかり理解し、保守業者と協調して管理運用を行う必要がある。