拡張性を考慮した拠点間ネットワークRAIDシステムの試作

宮元　章[1]・脇山 正博

A prototype of a Network RAID system considering extensibility

Akira MIYAMOTO, Masahiro WAKIYAMA

Abstract

In recent years, we had been devastated by great earthquake, severe rain storm and so on. Accordingly, it is desirable that we should store our data in a storage server in the data center to prevent data from getting damage and being lost, but requires expensive facilities and high running cost. On the other hand, to install RAID-enabled NAS in each office has equal or higher level of redundancy. It can have more high level of redundancy to operate as one virtual NAS which is gathered every NAS. Therefore, the system will be able to restore data from the another NAS, even if two of every NAS is broken down by disaster.

However, it needs same capacity size of NAS. If it uses different size of NAS, the smallest NAS will be the disk space used by the RAID array and it wastes surplus space of large NAS, so a data block units will be used. Also, there is only one controller server in this system, so to achieve availability and reliability, I use redundant servers. Moreover, I build web service for registering new NAS.

Keywords : RAID, internet, base-to-base, web service

1. 緒言

我が国は，昨今，巨大地震の発生や甚大な集中豪雨等により壊滅的な被害を受けた。今後も南海トラフ巨大地震や各地の活断層による直下型地震の発生が懸念されている。また，火災や水害，落雷等で事業所が大きな被害を受けることも想定される。

これらのことから，事業継続計画(1)（以下，BCPと略す）の重要性が再認識されている。BCPの策定により，災害等の不測の事態においても最小限のダウンタイムで逸早く事業を再開させることができる。

BCPの継続性の観点においてデータやシステムを守ることはとても重要である。そのため，データやシステムは破損や消失の可能性が高い事業所内に保存するのではなく，データセンタ等の災害に強い場所に保存することが望ましい。しかし，データセンタを利用するには高額なランニングコストに加え，一朝有事の際，サーバルームへの入室手続きがとても煩雑である。

そこで，筆者はこれまでに各事業所にRAID対応のネットワーク対応ストレージ（以下，NASと略す）を設置し，データの冗長性を図り，その上で更に各拠点に設置したNASを1台のHDDとして捉え，そのNASをデータ部とバックアップ部に分け，各拠点のバックアップ部を纏めて1台の擬似的なNASとして運用することで更に冗長化を持たせるシステムの試作を行った(2)。そのことでデータセンタ利用のランニングコストの削減・入退室の煩わしさを解消することができた。

しかし，このシステムにはいくつかの問題点がある。1点目は，RAIDの仕組みの関係上n台のNASの容量が一致している必要がある。1つでも容量が少ないNASが存在するとその容量に合わせる必要があるため無駄な領域が存在してしまうことがある。2点目は，各データをRAID形式で保存するためにある拠点に1台のコントローラを要するが，その拠点のシステムダウンと同時にシステム全体がダウンしてしまう。3点目は，一度システムが完成し，稼働してしまった後に新規NAS領域を追加することが困難である。そこで，例えば1口500GBとデータブロックを定め，その容量に合うようなブロック単位でデータを保存することで無駄な領域を削減させる。次に，高可用性と高信頼性を得るためコントローラを冗長化し複数の拠点でデータ処理作業を行うことを可能とする。最後にWeb上操作のみで簡単にNASを追加できるようにする。本研究ではこのようなシステムの構築を目的とする。

2. RAID

RAIDとは安価な複数の磁気ディスクを使って冗長性を確保する技術である。1987年カリフォルニア大学バークレー校のDavid A.Patterson, Garth Gibson, Randy H.Katzによって提唱された(3)。ディスクの組み合わせ方，ディスクへの書き込み方によって大きく1～5にレベル分けされるが，RAID 1とRAID 5を組み合わせたRAID 15のようにそれぞれの特徴を併せ持った方式も存在する。また，現在では，既存のレベル分けに加え，冗長性はないが複数のディスクに分散して書き込むRAID 0，RAID 5の仕組みを応用したRAID 6をまとめてRAIDと呼ぶことが多い。RAID 5の場合，ディスクは最低3本必要で，1本のディスクが障害を起こしても他のディスクからデータを復元することが可能である。また，RAID 6の場合，ディスクは最低4本必要で，2本のディスクが障害を起こしても復旧可能である。

[1] 教育研究支援室 機器分析技術グループ

3. システム

3. 1. 概要

既存システムの基本構造を図1に示す。各拠点にRAID対応のNASとNASのデータ処理を行うコントロールサーバを設置し，それらをインターネットに接続する。その後，各拠点のNASをまとめて擬似的なNASとして捉え，RAIDアレイの構築を考えた。しかし，一般的なハードウェア的にRAID処理する場合と比較すると，本システムはソフトウェア的に処理を行うため遅延が頻発すると予測され，ファイルのリアルタイム書き込みを行うには極めて無理がある。そこで，前提条件として1日1回程度のバックアップを行うと想定し，各拠点NASの一部をデータ部，それ以外をバックアップ部とし，バックアップ部をまとめてRAIDアレイを構築した。本システムではRAID6の仕組み（パリティを2つ用意する）を採用しているため，同時に2つ以内の拠点がダウンした場合でもその他の拠点のデータ及びパリティから消失したデータを復旧することができる。しかし，この仕組みをディスクの容量の無駄のないように構築するためにはNASの容量を一致させる必要がある。容量が一致していない場合，最小容量のNASをベースにRAIDアレイを構築するため，容量の大きなNASは，そこに利用されない領域がデッドスペースとなってしまう。そこで，本システムでは，RAIDアレイを構築する最小単位をNASとして考えるのではなく，500GB，1TBのような比較的大きなデータブロックとして考えた。例えば，図2に示すように500GB，1TB，1TB，2TBの4台のNASが存在する場合，1つのデータブロックを500GBとすると，合計で9個のデータブロックが存在し，ブロック単位でデータのバックアップ，リストアを行うことでNASの容量を有効に利用することができるシステムとなった。さらなる変更点として，冗長化を実現するため複数のコントロールサーバをホットスタンバイの形で冗長化し，いずれかのサーバがダウンしてもそれ以外のサーバにより処理を継続させることを可能とした。また，遠隔地からでもかんたんにNASを追加できるようにするため，Webサーバ上にRuby on Railsというフレームワークを用い，新規NAS登録用のWebアプリケーションを作成し，どこからでもNAS追加が可能となる仕組みをシステムに追加した。

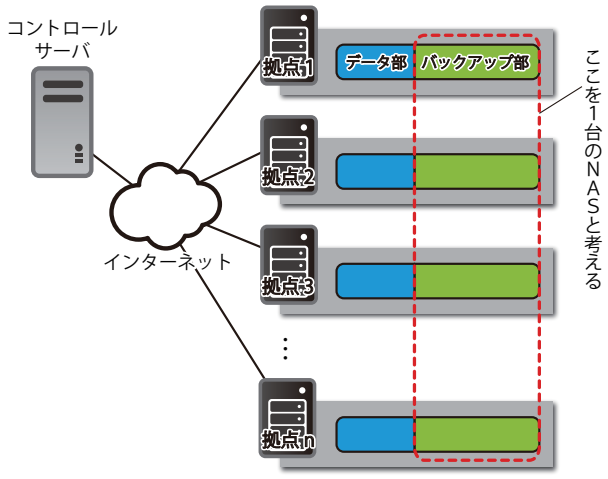

図1　既存システム概要

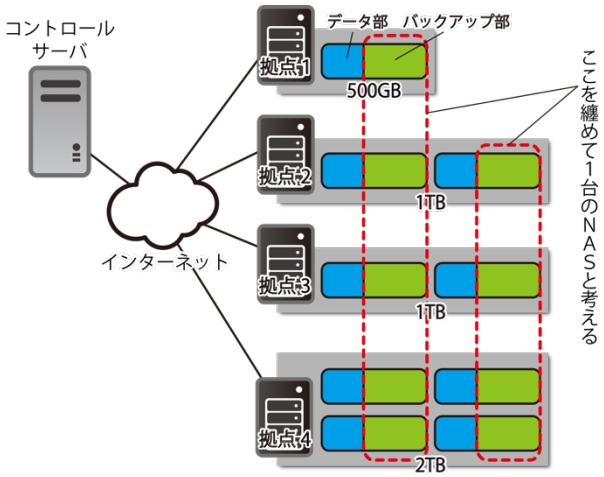

図2　データブロックによるNASの有効利用

3. 2. 通常運用時の流れ

図3に本システムの通常運用時（データの破損や消失が起きていない状態）の流れを示す（図3では，拠点数を3，ブロック数を4と仮定している）。まず始めに，1つのファイル保存する流れを説明する。

1. バックアップ時刻以前に一般ユーザによりデータ領域に1つのファイル（以下，元ファイルと略す）が保存される。
2. バックアップの時刻になると，コントロールサーバはWebサーバ上のデータベースに登録されたNASの情報をダウンロードし，その情報を元にNAS上の各データブロックのデータ領域及びバックアップ領域をマウントする。
3. データブロック容量確保のために作成した，後述する「ダミーデータ」を削除する。
4. 元ファイルを(データブロック数-2)等分する。
5. 分割したファイルからパリティP，パリティQのファイルを作成する。
6. 分割したファイルもしくはパリティファイルを各データブロックのバックアップ領域へ保存する。
7. 各データブロックのバックアップ領域に，分割ファイルもしくはパリティファイルの保存時刻，パス及び元ファイルのファイル名，ハッシュ値，元ファイルの存在していたデータブロック名をログとして保存する。
8. 次に，ダミーデータの保存について説明する。1つのデータブロック内のデータ部の容量をC，全データブロック数をn，データ部の容量をd，バックアップ部の容量をbとすると，

$$C = d + b \qquad (1)$$

となる。また，1ファイルのデータサイズを f とすると，

そのファイルは n-2 分割されると同時に分割されたデータと同じサイズの2つのパリティを作成する。そのため，各データブロックのバックアップ部には，データ容量が $f/(n-2)$ の分割データもしくはパリティのデータが保存される。そのため，バックアップ部の容量 b は，

$$b = \frac{n}{n-2} \cdot d \qquad (2)$$

となる。(1)，(2)より

$$b = \frac{n}{2n-2} \cdot C$$

となり，バックアップ部には上記の容量を確保する必要がある。他のデータブロックに存在する元ファイルの分割やパリティによって作成されたデータがバックアップ部に保存できなくなることを防ぐため，バックアップ部の実効容量をaとすると，データ容量が b-a となるように調整されたダミーデータを保存する。

9. コントロールサーバは各データブロックをアンマウントする。

複数ファイルの場合は，上記の流れ3～6を繰り返すことでファイル分割・パリティ作成を行い各データブロックへ保存することができる。通常運用時には，ログに記載されていない元ファイルのみファイル分割・パリティ作成を行う。ファイル名が同じ場合にはファイルのハッシュ値を比較し，そのファイルが上書き保存等されていないかどうかを確認後，異なる場合のみ処理を行い，通常運用時のバックアップの時間の短縮を図っている。

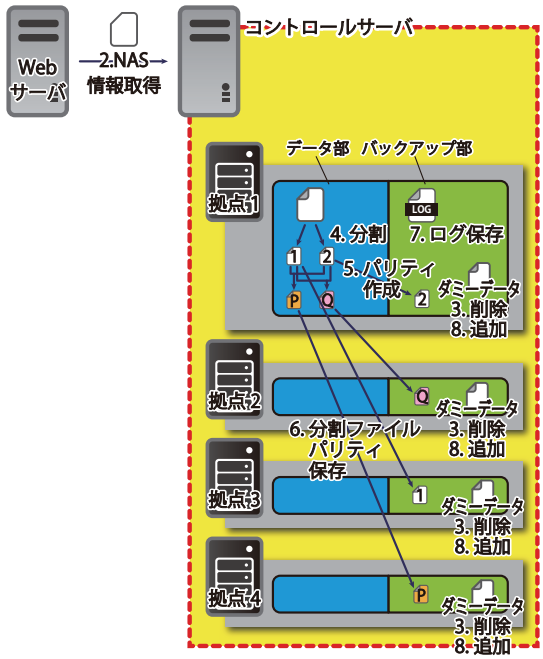

図3　通常運用時の流れ
（2. NAS情報取得～8. ダミーデータ削除）

3. 3. ファイル復元時の流れ

以下に，各データブロックから1つのファイルを復元する流れを示す。

1. 障害のある拠点のNASを新たなものに交換したり中のHDDを交換するなどしてアクセスできる状態に戻す。
2. コントロールサーバはWebサーバ上のデータベースに登録されたNASの情報をダウンロードし，その情報を元にNAS上の各データブロックのデータ領域及びバックアップ領域をマウントする。
3. 各データブロックのバックアップ領域を確認し，そこにログファイルが存在しないデータブロックを復元対象のものとして認識する。
4. 復元対象以外のデータブロック内のダミーデータを削除する。
5. 復元対象以外のデータブロックのバックアップ領域から分割ファイルおよびパリティファイル（分割ファイルのみの場合もあれば，パリティファイルのみの場合もある）からファイルを復元する。
6. 4. で利用した分割ファイルおよびパリティファイルを削除する。
7. 復元されたファイルを対象のデータブロックのデータ領域に保存する。
8. バックアップ時の流れ同様，復元されたファイルを(拠点数-2)等分し，パリティP，パリティQファイルを作成し，各データブロックのバックアップ領域へ保存する。
9. バックアップ時の流れと同様にログを作成し，保存する。
10. コントロールサーバは各データブロックをアンマウントする。

複数ファイルの場合は，上記の流れ4～8を繰り返すことで復元することが可能である。

3. 4. Webによる新規NAS登録時の流れ

本システムにおいて新規にNASを追加申請する際の流れを例とともに示す（この例では，データブロック数が4の状態から新規データブロックを1つ追加する流れを示す）。

1. 新規に登録するNASに1つのデータブロックとしてディレクトリを作成し，そのディレクトリに対して書き込み／読み取り権限のあるユーザを作成する。
2. 図4に示すような登録サイトにアクセスし，「Registration for New RAINAS Storages」ボタンをクリックし，図5に示す画面に移動する。該当欄にNASのFQDNもしくはIPアドレス，共有フォルダ名，手順1で作成したユーザ名およびパスワード等を入力し，「Create Storage」ボタンによりこれらの情報がWebサーバ上のデータベースに格納され，新たにNASが登録されたのを確認する。
3. データベースに登録されることで，コントロールサーバからの要求により登録されたNASのデータを図6のようなJSON形式でダウンロード可能な状態となる。

4. データブロックが追加された直後の段階では，既存の分割データやパリティはデータブロック数が4のときに作成されたものであるため継続して利用することができず，一度RAIDをリセットし，分割データやパリティを再作成する必要がある。そのため，「3.3 ファイル復元時の流れ」のコマンドを実行し，RAIDをリセットする。

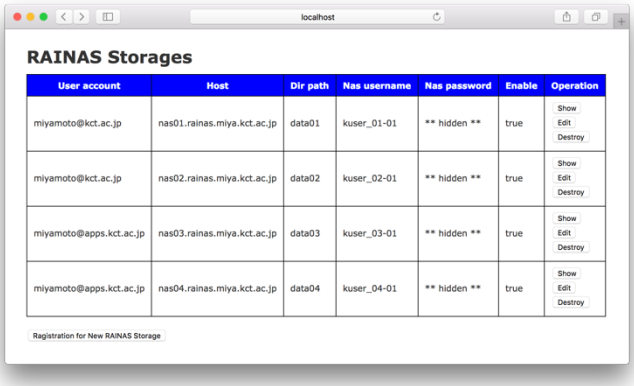

図4　新規NAS登録サイト

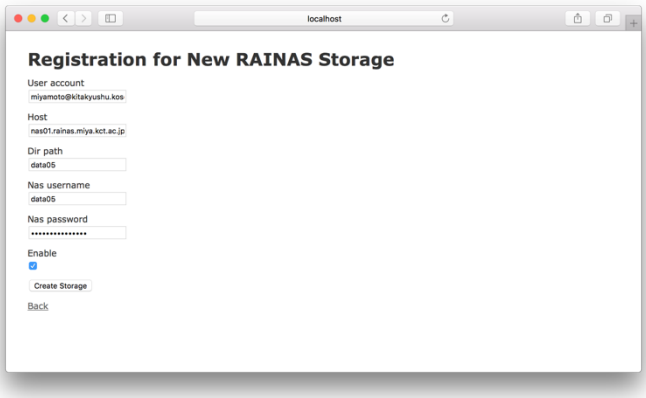

図5　新規NAS登録ページ


```
[{"id":1,"user_account":"miyamoto@kct.ac.jp","h
ost":"nas01.rainas.miya.kct.ac.jp","dir_path":"
data01","nas_username":"kuser_01-
01","nas_password":"PASSWORD","enable":true,"cr
eated_at":"2017-11-
05T16:43:28.149Z","updated_at":"2017-11-
05T16:43:28.149Z","url":"http://localhost:3000/
storages/1.json"},
                    (中略)
{"id":4,"user_account":"miyamoto@apps.kct.ac.jp
","host":"nas04.rainas.miya.kct.ac.jp","dir_pat
h":"data04","nas_username":"kuser_04-
01","nas_password":"PASSWORD","enable":true,"cr
eated_at":"2017-11-
05T17:00:17.815Z","updated_at":"2017-11-
05T17:52:22.446Z","url":"http://localhost:3000/
storages/4.json"}]
```


図6　NAS情報のJSON形式データ
（セキュリティ上パスワードの箇所は変更している）

4. 結言

本研究では，インターネットを介した拠点間RAIDシステムの試作を行った。既存の仕組みによりRAID対応のNASを1台のHDDとして捉え，各拠点のバックアップ部をまとめて1台の擬似的なNASとしてバックアップを行うことに加え，データブロック単位で管理することによりNASを有効に利用することができるシステムとなった。また，冗長化されたコントロールサーバにより高い信頼性と可用性を得ることができた。また，Webサーバに新規NASの登録用Webサービスを構築することで遠隔地からでも簡単にNASの新規登録を行うことが可能となった。

今後の課題としては，まず1つ目に以前からの課題であるバックアップ時間，復元時間の短縮が挙げられる。解決のための具体案としては，コントロールサーバのCPUをコア数が多いものを採用しマルチスレッディングを実現する方法や，複数のコントロールサーバを用意し，同時並行で処理を行う分散コンピューティングを用いる方法等を検討している。2つ目は，データブロック単位で管理することにより若干の冗長性低下が懸念される。データブロック単位で管理していないときにはNASが2拠点破損しても残りの拠点のNASからデータを復旧することができるが，本システムでは，仮に1つのNASに2つのデータブロックが存在する場合，そのNASを含む2拠点のNASが破損してしまった場合，3つ以上のデータブロックが失われるためデータ復旧が不可能となる。この課題の解決案として，1つのNASに搭載するデータブロックを2つ以下にすることや，データブロックの容量は最も容量の小さなNASに依存するため，そのNASを大容量のものに交換し，データブロックのサイズを上げ，全体のデータブロック数を減らす方法や，RAID 61等でさらなる冗長化を行うことを検討している。

謝辞

本研究はJSPS科研費 JP16H00394の助成を受けたものです。

参考文献

(1) BCP（事業継続計画）とは
http://www.chusho.meti.go.jp/bcp/contents/level_c/bcpgl_01_1.html

(2) 宮元 章・脇山正博，インターネットを介した拠点間RAIDシステムの試作，北九州工業高等専門学校研究報告 第50号，2017年1月

(3) D.A.Patterson, G.A.Gibson, R.H.Katz. “A Case for Redundant Arrays of Inexpensive Disks (RAID).” Proceedings of the International Conference on Management of Data (SIGMOD), June 1988.

（2017年11月6日 受理）