インターネットを介した拠点間RAIDシステムの試作

宮元　章*・脇山 正博

A prototype of a RAID system between several base points through Internet

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. Install RAID-enabled NAS in each office, so this system has a high 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 can restore data from the another NAS, even if two of every NAS is broken down by disaster.

Keywords : RAID, internet, base-to-base

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，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る。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，各拠点をまとめて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。本試作ではRAID6の仕組みを採用しているため，全拠点のうち2拠点が災害等でダウンしてもその他の拠点のデータから復旧させることが可能である。各拠点で用意する機器はNASとコントローラサーバのみとなり，ランニングコストは僅かとなる。さらに復旧は各拠点で行えるため，データセンタ内のサーバルームへの入室手続きの煩わしさもない。本研究ではこのようなシステムの試作を目的とする。

2. システム

2. 1. 概要

本システムは，図1に示すように，各拠点にRAID対応のNASを設置し，それらをインターネットに接続する。仮に，各拠点のNASをまとめて擬似的なNASとして捉え，RAIDで運用した場合，インターネットを経由することや，専用のコントローラを用いてCPUの負荷をかけない状態でハードウェア的にRAID処理するものではなく，コントロールサーバ上でソフトウェア的に処理することから，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擬似的な1台のNASとして考えた。つまり，各拠点のバックアップ部は一般的なRAIDシステムの1台のHDDとして捉えることとした。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。また，拠点数n，各拠点のNASの実効容量C，NAS内のデータ部の容量をxとすると，

$$C = x + \frac{x}{n-2} \cdot n$$

となり，拠点数nが増えれば増えるほど冗長性が増すと同時にxも増加し，利用効率を向上させることができる。

2. 2. 通常運用時の流れ

図2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2では，拠点数を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により各拠点のNASのデータ領域及びバックアップ領域をマウントする。
3. 元ファイルを(拠点数-2)等分する。
4. 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。
5. 分割したファイルもしくはパリティファイルを各拠点の

* 教育研究支援室 機器分析技術グループ

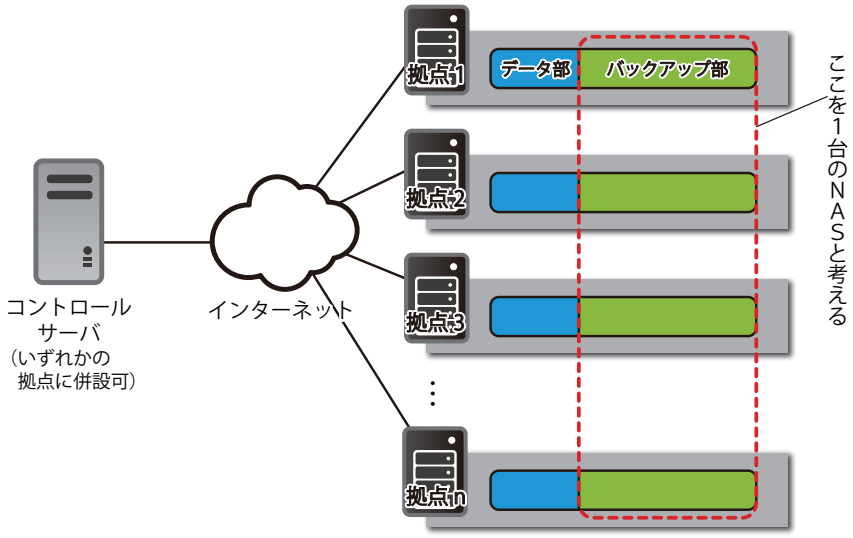

図1　システム概要

NASのバックアップ領域へ保存する。

6. 各拠点のNAS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たNASの拠点名をログとして保存する。
7. コントロールサーバは各拠点のNAS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NAS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う。こうすることで通常運用時のバックアップの時間を大幅に短縮することができた。

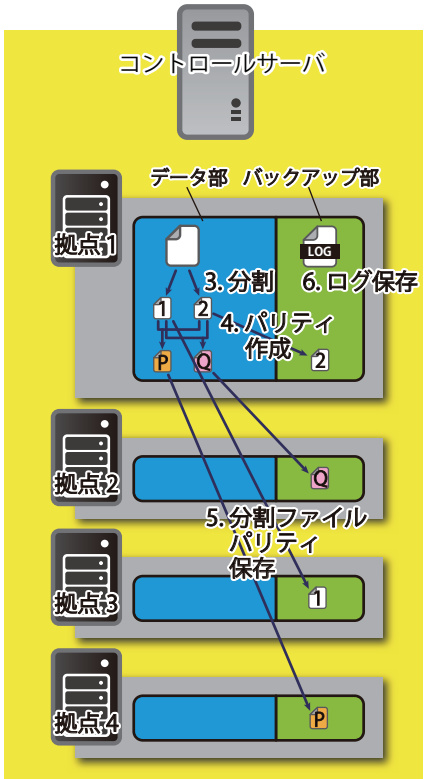

図2　通常運用時の流れ（3. 分割～6. ログ保存）

2. 3. ファイル復元時の流れ

以下に，各拠点から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により各拠点のNASのデータ領域及びバックアップ領域をマウントする。
3. 各拠点のNAS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NASを復元対象のものとして認識する。
4. 復元対象以外のNAS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。
5. 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。
6. 復元されたファイルを対象のNASのデータ領域に保存する。
7. 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，各拠点のNASのバックアップ領域へ保存する。
8. バックアップ時の流れと同様にログを作成し，保存する。
9. コントロールサーバは各拠点のNAS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。

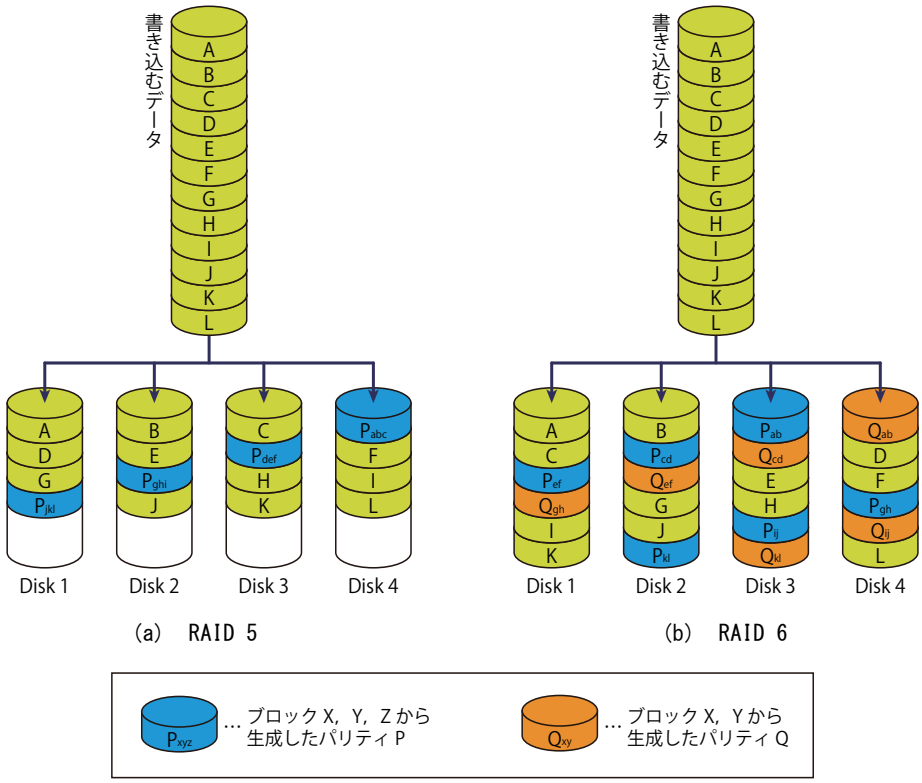

図3　RAID 5およびRAID 6のデータ保存の流れ

3. RAID

3. 1. RAID技術

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[2]。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。

3. 2. RAID 5とRAID 6

RAID 5におけるデータの保存の流れを図3(a)に示す。保存するデータをブロックごとに分割し，それぞれを複数のディスクに書き込む際，同時にそれらの分割したデータからパリティと呼ばれる冗長コードを生成し，そのコードを残りのディスクに保存する方式である。すべてのディスクのうち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。RAID 5のデータ実効容量をC_5，ディスク1本あたりの容量をd，ディスクの本数をnとすると，

$$C_5 = (n-1)\cdot d$$

となる。この際，ディスクは最低3本必要となる。

RAID 6におけるデータ保存の流れを図3(b)に示す。この方式は，RAID 5を応用した仕組みである。RAID 5では，ブロックごとに分割したデータからパリティを1つ作成するのに対し，RAID 6では2つのパリティを作成し，残り2つのディスクに書き込む。このことにより，すべてのディスクのうち2本のディスクが障害を起こしてもデータを復元することが可能である。RAID 6のデータ実効容量をC_6，ディスク1本あたりの容量をd，ディスクの本数をnとすると，

$$C_6 = (n-2)\cdot d$$

となる。この際，ディスクは最低4本必要となる。

3. 3. RAID 6におけるパリティの生成とパリティによる復元

合計n拠点のディスクのブロックのデータをd_1，d_2，$\cdots d_n$，生成するパリティをP，Qとすると，それぞれのパリティは，

$$P = d_1 + d_2 + \cdots + d_n$$
$$Q = 1\cdot d_1 + 2\cdot d_2 + \cdots + n\cdot d_n$$

となる。

次に，パリティによりデータを復元する方法を述べる。障害が発生したブロックのデータをd_x，d_y($1 \leqq x \leqq n$，$1 \leqq y \leqq n$，$x \neq y$)，障害が発生していない拠点番号の最大値をm($m \neq x$，$m \neq y$)とすると，

$$P - (d_1 + d_2 + \cdots + d_m) = d_x + d_y \quad (m \neq x,\ m \neq y,\ m \leqq n)$$
$$P - (d_1 + d_2 + \cdots + d_m) = P' \text{ とすると，}$$
$$P' = d_x + d_y \qquad (1)$$
$$Q - (1\cdot d_1 + 2\cdot d_2 + \cdots + m\cdot d_m) = x\cdot d_x + y\cdot d_y$$
$$(m \neq x,\ m \neq y,\ m \leqq n)$$

$Q-(1\cdot d_1+2\cdot d_2+\cdots+m\cdot d_m)=Q'$とすると，

$$Q'=x\cdot d_x+y\cdot d_y \qquad (2)$$

(1)，(2)式より

$$d_y=\frac{Q'-x\cdot P'}{y-x}$$

$$d_x=P'+d_y$$

となり，障害が発生したブロックのデータをd_x，d_yを復元することができる。

4. システムの検証

4. 1. 検証実験概要

本システムの検証を行うため，拠点数を4と仮定し，学内に設置した4台のNASと1台のコントローラサーバを用いて以下の手順で実験を行い，手順2～6をバックアップに要した時間，手順7～11を復元に要した時間とし，ファイルごとに10回ずつ計測した。

1. 事前に1kBのファイルを用意し，任意のNASのデータ部に保存する。

【バックアップ】

2. コントローラサーバが4台のNASをマウントする。
3. 上記のファイルを2分割する。
4. 3で作成されたファイルよりパリティP，パリティQのファイルを作成する。
5. 3，4で作成されたファイルをそれぞれ4台のNASへ保存する。
6. コントローラサーバが4台のNASをアンマウントする。

【復元】

7. コントローラサーバが4台のNASをマウントする。
8. 4台のNASのうち，2台に障害が発生したと仮定し，そのNASのデータを全て削除し，障害発生状態とする。
9. 障害の発生していないNASから分割ファイルおよびパリティファイルを取り出し，ファイルを復元する。
10. 9で復元されたファイルを障害が発生したNASへ保存する。
11. コントローラサーバが4台のNASをアンマウントする。
12. 事前に用意するファイルを10kB，100kB，1MB，10MB，100MBに変更し，手順2～11を行う。

4. 2. 検証実験結果

4. 1で行った実験の結果を図4に示す。学内（最大2Gbps）のネットワーク環境で実験を行ったため，計測時間のうち，マウント，アンマウント，各NASへのデータ保存にはほとんど時間を要しなかった。このことを考慮すると，実験結果よりファイルサイズとバックアップ，復元に要する時間は比例していると考えられる。10MBの1つのファイルをバックアップおよび復元するだけで20分以上の時間を要し，200MBのファイルに至っては3時間以上の時間を要してしまうという思わしくない結果となってしまった。運用時には多くのファイルが存在すると想定されることからこのことは極めて重大な問題として言わざるをえない。

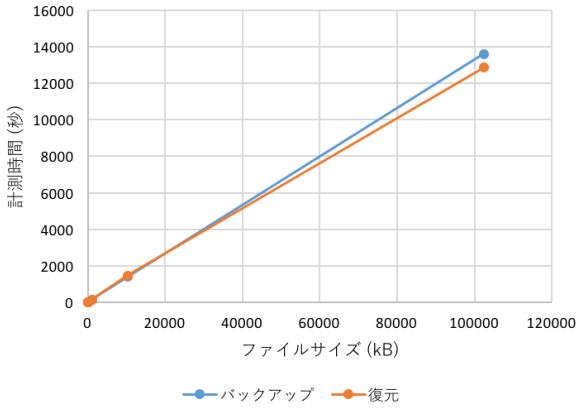

図4　実験結果

5. 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。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ができた。また，障害を想定した実験では，元あるデータをきちんと復元することができた。このことで，冗長化を観点に考えるとデータセンタにデータを保存する場合に比べ，本システムは大きくランニングコストを削減することができる。また，データセンタ内のサーバルームへの入室手続き等の煩雑さもなくすことで逸早い復旧を行うことができる。

今後の課題としては，バックアップ時間，復元時間の短縮が挙げられる。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。

謝辞

本研究はJSPS科研費 JP15H00386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) A Case for Redundant Arrays of Inexpensive Disks (RAID). D. A. Patterson, G. A. Gibson, R. H. Katz. Proceedings of the International Conference on Management of Data (SIGMOD), June 1988.

（2016年11月7日　受理）