mathematics behind RAID 5DP

hp technical guide

for the hp va 7000 series disk arrays

for information about the va 7000 series and periodic updates to this guide see the HP SureStore website at http://www.hp.com/go/storage

Hewlett-Packard Company

Hewlett-Packard Product Information

mathematics behind RAID 5DP – for the **hp** va 7000 series disk arrays

Published: April 2001

Revision level 1.0

For the latest updates to this document see
http://www.hp.com/go/storage

Warranty

This document is supplied on an "as is" basis with no warranty and no support. Hewlett-Packard makes no express warranty, whether written or oral with respect to this document. HEWLETT-PACKARD DISCLAIMS ALL IMPLIED WARRANTIES INCLUDING THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE.

LIMITATION OF LIABILITY: IN NO EVENT SHALL HEWLETT-PACKARD BE LIABLE FOR ERRORS CONTAINED HEREIN OR FOR ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES (INCLUDING LOST PROFIT OR LOST DATA) WHETHER BASED ON WARRANTY, CONTRACT, TORT, OR ANY OTHER LEGAL THEORY IN CONNECTION WITH THE FURNISHING, PERFORMANCE, OR USE OF THIS MATERIAL.

The information contained in this document is subject to change without notice.

No trademark, copyright, or patent licenses are expressly or implicitly granted (herein) with this white paper.

Disclaimer

All brand names and product names used in this document are trademarks, registered trademarks, or trade names of their respective holders. Hewlett-Packard is not associated with any other vendors or products mentioned in this document.

Table of Contents

Overview

Understanding the mathematics of RAID 5DP

Hewlett Packard's Virtual Array RAID 5 Double Parity (RAID 5DP) provides superior data availability by allowing the operation to continue after two simultaneous disk failures. This paper describes the mathematical details behind RAID 5DP.

HP's RAID 5DP, while similar to RAID 5, is *superior* because it uses two separate and independent mechanisms for the redundant information used to recreate data during disk failures.

In general, RAID 5DP comprises a system of two equations in two unknowns, with the two unknowns being the two failed data disks. Standard linear algebra techniques can be used to derive a general solution. The two redundancy schemes, **P** and **Q**, are computed as a *sum of products* with different coefficients being applied to each data block for P and Q.

P uses the value one (1) for all coefficients which reduces to a standard exclusive OR (parity) calculation. **Q** uses coefficient values *other than the value one (1)* which makes the calculation of Q more complex than standard exclusive OR. The Q coefficients are chosen so that the Q equation is linearly independent from P allowing the two equations to be solved in all cases.

The RAID 5DP coding-scheme is an instance of the well-known *Reed-Solomon* class of error correction codes. Variations of *Reed-Solomon ECCs* are also used in the design of hard disk drives and are used in conjunction with semiconductor memory (RAM).

This paper use **N**+**M** nomenclature to indicate the width (N+M) of the redundancy group. In each stripe of an N+M redundancy group **N** is the number of storage elements used for data and **M** is the number of storage elements used for redundancy. For RAID 5DP, **M=2**. There is one (**1**) redundant storage element used for **P** and one (**1**) used for **Q** in each stripe.

Data Layout

A set of disks are known as a redundancy group.

The similarity of the data layout for RAID 5DP and the traditional RAID 5 is that the parity information rotates through all the disks. Where it differs, however, is that the equivalent capacity of two disks per redundancy group are used for storing parity information by RAID 5DP, where only a single disk is used by RAID 5. HP Virtual Array uses the term "redundancy group" to refer to a set of disks that are isolated with respect to disk failures. Disk failures in one redundancy group do not effect the operation of other redundancy groups. Each RAID 5DP stripe uses a single segment from each disk in a single redundancy group.

N+2 Mathematics

How does the N+2 redundancy group provide dual correction?

The *N+2 redundancy group* arrangement provides data correction capability for any two failed disks in a group. Thus, an N+2 group can tolerate the loss of any two disks in the group and still maintain data availability. This is in contrast to an N+1 redundancy group, which can only correct a single failed disk in a stripe. To achieve the dual correction capability, two segments of redundant data are stored along with N segments of user data in a RAID 5 stripe. The values for the two redundant segments are computed with a more powerful error-correction coding scheme than the simple parity calculation typically used for N+1 redundancy groups.

Below is a general form of the N+2 computation.

$$P = p_0x_0 + p_1x_1 + p_2x_2 + \ldots + p_{N-1}x_{N-1} = \Sigma p_i x_i (i = 0 \ldots N-1)$$

$$Q = q_0x_0 + q_1x_1 + q_2x_2 + \ldots + q_{N-1}x_{N-1} = \Sigma q_i x_i (i = 0 \ldots N-1)$$

Where

P is the value of one redundancy segment

Q is the value of the other redundancy segment

x_i are the values of the user data segments

p_i and q_i are coefficients of the error correction-coding scheme

Note: The equation for **P** reduces to the simple parity calculation typically used for N+1 groups when all coefficients p_i have the value one (1).

The equations for calculating **P** and **Q** *error-correction terms* form a system of two equations that, by the rules of linear algebra, can potentially be solved for any two unknowns (x_a and $x_{b)}$, which represent any two failed disks in the stripe. In fact, the equations can be solved for any two unknowns if the sets of coefficients p_i and q_i are linearly independent.

An example of linearly independent coefficients is p_0=1, p_1=1, p_2=1, … (simple parity) and q_0=1, q_1=2, q_2=3, and so on. Many other examples are possible.

A general solution for any two unknowns, x_a and x_b, can be derived from a linear algebra matrix manipulation as shown below.

$$R = p_a x_a + p_b x_b = P - \Sigma p_i x_i (i = 0 \ldots N-1, i \neq a, i \neq b)$$

$$S = q_a x_a + q_b x_b = Q - \Sigma q_i x_i (i = 0 \ldots N-1, i \neq a, i \neq b)$$

In matrix form

$$\begin{bmatrix} p_a & p_b \\ q_a & q_b \end{bmatrix} \begin{bmatrix} x_a \\ x_b \end{bmatrix} = \begin{bmatrix} R \\ S \end{bmatrix}$$

The matrix solution, using the standard formula for the inverse of a two by two square matrix, is

$$\begin{bmatrix} x_a \\ x_b \end{bmatrix} = \begin{bmatrix} p_a & p_b \\ q_a & q_b \end{bmatrix}^{-1} \begin{bmatrix} R \\ S \end{bmatrix} = (p_a q_b - p_b q_a)^{-1} \begin{bmatrix} q_b & -p_b \\ -q_a & p_a \end{bmatrix} \begin{bmatrix} R \\ S \end{bmatrix}$$

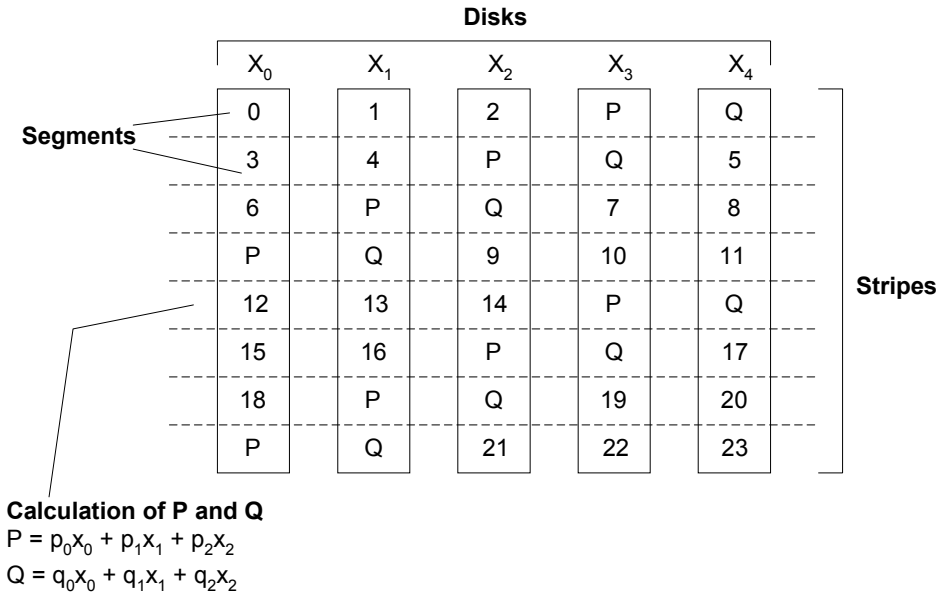
For a solution to exist the term $(p_a q_b - p_b q_a)^{-1}$ must exist. For the inverse to exist the difference must be non-zero and the inverse of the non-zero value must exist in the arithmetic system being used to perform computations. A non-zero difference is ensured by linear independence of the coefficients. A non-zero difference is a formal method of stating the need for linear independence in the coefficients. To ensure that the inverse of any non-zero value exists, the system of arithmetic called ***Galois Field*** or finite field arithmetic is used. One of the properties of finite field arithmetic used, is that addition and subtraction are the same. Therefore, the matrix solution is as follows.

$$\begin{bmatrix} x_a \\ x_b \end{bmatrix} = (p_a q_b + p_b q_a)^{-1} \begin{bmatrix} q_b & p_b \\ q_a & p_a \end{bmatrix} \begin{bmatrix} R \\ S \end{bmatrix}$$

The following figures and explanation gives a description of how data would be written to a set of disks. We show how data can be recalculated or recovered in the event of a loss of a disk or the simultaneous loss of two disks.

Figure 1 shows how data is striped across three data disks and two parity disks. Independent calculations are made to generate the values for P and Q for each stripe. The specific equations for generating the values of P and Q for the row starting with segment 12 are shown in Figure 1. The parity data is rotated to different disks by stripe.

FIGURE 1. Example of Error-correction Coding — Finite Field Arithmetic



In the event one disk fails, the data on that disk can be recalculated using either the P or Q segment and the data stored on the remaining good segments in the stripe. The equations to recalculate the data are as follows in Figure 2.

FIGURE 2. Recalculation of segment on disk X_1

Single correction using P	Single correction using Q
(disk x_1 failed)	(disk x_1 failed)
$P = x_0 + x_1 + x_2$	$Q = q_0x_0 + q_1x_1 + q_2x_2$
$x_1 = P + x_0 + x_2$	$x_1 = q_1^{-1}(Q + q_0x_0 + q_2x_2)$

If two disks were to fail simultaneously, for example disk X_1 and X_2, or if a second disk were to fail before the first failed disk was completely recovered, then using the RAID5DP as described in this paper, the data from both failed disks could be rebuilt using the equations as shown in Figure 3.

FIGURE 3. Recalculation of data on disks X_1 and X_2

Double correction => two equations in two unknowns

(disks x_1 and x_2 failed)

$$P = p_0x_0 + p_1x_1 + p_2x_2$$

$$Q = q_0x_0 + q_1x_1 + q_2x_2$$

$$p_1x_1 + p_2x_2 = P + p_0x_0$$

$$q_1x_1 + q_2x_2 = Q + q_0x_0$$

. . . . . .

Do the matrix inversion to solve for x_1 and x_2

$$x_1 = (p_1q_2 + p_2q_1)\text{-}1(q_2(P+p_0x_0) + p_2(Q+q_0x_0))$$

$$x_2 = (p_1q_2 + p_2q_1)\text{-}1(q_1(P+p_0x_0) + p_1(Q+q_0x_0))$$

Both segments of missing data can be recalculated using the P and Q parity data and the remaining data segment on X_0.

If the data contained in the segment of a lost disk was either P or Q parity data, then the parity data can be recalculated from the remaining user data. If two disks are lost and one disk contains user data and the other parity data, the user data can be recalculated first using the remaining parity segment, as above, then the lost parity segment could be recalculated using the restored user data. Thus the array can survive the simultaneous loss of two disks where RAID5DP has been used to record the data.

Note: Information presented in this paper, including diagrams, is for the purpose of illustrating the mathematical principles and does not necessarily represent the exact implementation in the Virtual Array.

References

A complete description of the finite field arithmetic used in this error-correction coding scheme is found in “Practical Error Correction Design for Engineers,” Neal Glover, Data Systems Technology Corp, Broomfield, CO, 1982.