an analysis of RAID 5DP

a qualitative and quantitative comparison of RAID levels and data protection

hp — white paper

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

Hewlett-Packard Company

Hewlett-Packard Product Information

an analysis of RAID 5DP – a qualitative and quantitative comparison of RAID levels and data protection

Published: July 2001

Revision level 1.1

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

Table of contents

Executive summary

Multiple RAID levels are available to users in most modern disk arrays. Different RAID levels provide different availability and performance results. RAID 5DP, an implementation of RAID 5 and 6 from HP provides a cost effective balance of outstanding availability and low cost, relative to most RAID 1 or traditional RAID 5 configurations. Some configurations of RAID 5 are not practical due to the availability risks they would create.

- RAID 1 is the least efficient and most expensive RAID Level.
- RAID 5, while more efficient offers less data protection than RAID 5DP configurations.
- RAID 5DP offers an optimal combination of availability and cost effectiveness.

Introduction

Fault tolerance and redundancy have been used to increase system reliability. In particular, in data storage the Redundant Array of Independent Disks (RAID) systems have been widely used to protect against data loss and downtime. Among the RAID levels, RAID 1 and RAID 5 have gained increased popularity in the last decade.

RAID 1 and RAID 5 can tolerate only one disk failure per group.

In RAID 1, every piece of data is written to both a data disk and a check disk (mirror disk). In this case, in order for the data in the array to be lost, the failure of one disk followed by the failure of its mirror must occur within the time to recover from the first failure.

In RAID 5, data is written to a group of disks in stripes. Each stripe contains data sectors and one check sector. The check sector is usually the parity of the data sectors on the same stripe. The data and parity are spread over all disks (no single check disk). The failure of a disk in RAID 5 will force the array to a degraded state (some data must be re-computed) without losing data. A second disk failure within the same group that occurs while the first disk failure is still being recovered will cause a data loss.

Because both RAID 1 and 5 use a single check disk per group, only one disk failure per group can be tolerated.

RAID 5DP provides a greater protection against data loss.

A new RAID level is being introduced by Hewlett Packard on the VA 7000 series arrays. This RAID level is similar to RAID 5, however it uses two independent parity calculations. Each stripe contains data sectors and two check sectors. The new RAID level is called RAID 5 Double-Parity (RAID 5DP). We

will show, under most conditions, that RAID 5DP provides higher protection against data loss than either RAID 1 or RAID 5 since failure of two disks in the same group is tolerated by this new RAID level.

Qualitative analysis

In this section, we compare qualitative data loss potential of three RAID levels, RAID 1, 5, and 5DP, and their relative cost.

RAID 5DP presents a lower risk to data loss or corruption.

- RAID 1 and RAID 5 present higher risk to data corruption or loss than RAID 5DP. In the event of a disk failure, the system must rebuild the missing data based on the data on the surviving disk(s). If there is an internal defect in a sector of one of the surviving disks, then the data for the stripe residing on that particular sector cannot be rebuilt (or it is rebuilt incorrectly).

  However, RAID 5DP can tolerate the failure of two disks. Thus, for the scenario described above, the surviving disk (with an internal defect in a sector) can be put in a failed state (ignored) and the data rebuilt correctly.

- In a RAID 1 or RAID 5 configuration, once a disk fails, the data in that group becomes exposed (critical state). Therefore, the array must rebuild the data on the failed disk as quickly as possible to recover the group redundancy and thereby remove the array from the critical state. A fast rebuild of data by the array is generally conducted at the expense of system I/Os.

  In certain applications, dedicating system I/Os for a fast rebuild may cause high performance degradation. However, RAID 5DP can tolerate the failure of two disks. Thus, the failure of a disk in a group will not cause urgency on rebuilding the data since the group is still one failure away from a critical state. For this reason, the system I/O performance level is preserved and the rebuild occurs slowly without incurring any risk.

  The period of time required to rebuild a drive is directly proportional to the size of that drive. Typical rebuild times of one to two days are not uncommon with today's commonly available arrays. As drive sizes continue to increase over the coming years, and drive rebuild times continue to increase accordingly, being able to operate an array effectively during a rebuild will become increasingly important. The VA family's ability to rebuild at low priority in the background because of the two disk failure tolerance provides the VA series of arrays this ability to operate effectively during a rebuild.

Which is most economical?

- RAID 5 is more economical than RAID 5DP for small parity groups since in each group only one disk's capacity is dedicated for parity; whereas RAID 5DP consumes two disks for parity in each group. However, the expense of RAID 5DP can be mitigated by using longer stripes. The longer stripes can make RAID 5DP equivalent to RAID 5 in storage efficiency. For example, two 5+1 RAID 5 stripes have the same storage efficiency as a 10+2 RAID 5DP stripe. In this paper, the term storage efficiency is defined as the ratio of data disks divided by the total of data and check disks. In the case above, a RAID 5+1 stripe would have a storage efficiency of 5/6 or 83%.

Why not implement RAID 5 in longer stripe lengths?

- System designers generally do not implement RAID 5 in longer stripe lengths, such as 10+1 or 20+1 RAID 5 stripes. The rebuild times, and hence the period the array remains in a critical state, after a failure in very long stripes are unacceptably long for RAID 5. By contrast, because RAID 5DP stripes can withstand the loss of two drives without a data loss, configuring long RAID 5DP stripes is acceptable from a risk and performance perspective, and desirable from an economy perspective. These generalizations about stripe length and rebuild times are quantified for the VA 7400 arrays in the calculations later in this paper.

Which is the most expensive?

- RAID 1 is by far the most expensive configuration, since for each data disk a mirror disk is needed for this configuration.

Quantitative analysis

In this section, we compare quantitative data loss potential of three RAID levels, RAID 1, 5, and 5DP, and their relative cost. The two metrics that are used for this comparison are the Mean-Time-To-Data-Loss (MTTDL) and Storage Efficiency (SE).

In this comparison, we will focus on the data loss caused by disk hardware failures *only.* Software, firmware, external events (such as human errors or catastrophic events), and other array component failures that may cause data loss are ignored. In addition, storage efficiency is viewed from disks only. Before reviewing the comparison, we shall review how the two metrics are calculated.

Theory

Let us assume that disk hardware failures are random and independent. Assume that Mean-Time-To-Failure for disk is MTTF(disk), and that rebuild starts instantaneously following a disk failure (that is; either the array has extra disks used as hot spares or a user gets immediate notification of a failed disk and can replace the failed disk with a shelved spare in a very short time). In addition, assume that the Mean-Time-To-Rebuild (MTTR) depends on RAID level and the number of disks in the group. Let us assume that the array consists of **K** groups of the same RAID level, and each group has **N** disks including **M** check disks (for RAID 1 or 5, M = 1; for RAID 5DP, M = 2; and for RAID 1, N=2).

Table 1 shows a set of four equations which are used for the quantitative analysis portion of this paper. We are assuming that MTTR(group) is much smaller than the MTTF(disk).

TABLE 1. Sample equations based on theory scenario

Number	Equation
	For a group of N disks configured in RAID 1 or RAID 5:
1	$MTTDL(group) \sim [MTTF(disk)]^2/[N*(N-1)*MTTR(group)]$
	Also for a group of N disks configured in RAID 5DP:
2	$MTTDL(group) \sim 2*[MTTF(disk)]^3/[N*(N-1)*(N-2)*(MTTR(group))^2]$
	Then for an array consisting of K groups, where the array configuration has equal number of disks in all the groups, and the same RAID level is used for all the groups in the array, then:
3	MTTDL(array) = MTTDL(group)/K
	The storage efficiency is the ratio of all data disks (excluding check disks) to all disks (including check disks) in the array:
4	SE = Number of data disks/Number of data and check disks. (SE is usually expressed as a percentage.)

Note: Equations 1, 3, and 4 in Table 1 on page 4 are similar to calculations of disk array reliability described in "A Case for Redundant Arrays of Inexpensive Disks (RAID)," David A. Patterson, Garth Gibson, and Randy H. Katz, Computer Science Division, University of California, Berkeley, CA 94720.

Equation 2 in Table 1 is an extension of Equation 1 to account for the tolerance of two disks failures.

Analysis

The maximum array capacity consists of 7 disk enclosures containing 15 disks each, for a total of 105 disks. Eight RAID configurations was designed for this comparison (see Table 2 on page 6). Each configuration was designed to be as close as possible to the maximum capacity of a VA 7400 array without violating any of the constraints listed below.

1. A RAID 1 group must have two disks (one data disk and one check disk).
2. A RAID 5 group must have at least three disks (two data disks and one check disk) and a maximum of 6 disks, 15 disks, or 52 disks (see Case 2, 3, and 4 shown in Table 2 on page 6).
3. A RAID 5DP group must have at least five disks (three data disks and two check disks) and a maximum of either 15 disks or 52 disks (see Case 5 through 8 shown in Table 2 on page 6).
4. Only similar RAID groups are supported within an array configuration (that is, all RAID groups in the array have the same RAID level and the same maximum number of disks in a group).
5. A RAID group must attain its maximum number of disks, before creating a new group.
6. A new RAID group cannot be created until the minimum number of disks for the group is reached.

Using the above constraints, Table 2 on page 6 shows the number of disks for each case. With the exception of Case 4, all the other cases have fully populated groups. Case 4 has 17 fully populated groups (6 disks in each group) and group 18 is partially populated (3 disks only). In addition, Table 2 shows that the *rebuild priority* for Cases 1–6 is *high* as compared to Cases 7 and 8 where the *rebuild priority* is *low.*

What do we mean by "rebuild priority high" or "rebuild priority low"?

Rebuild priority high refers to reserving 50% (on the average) of array CPU resources to rebuilding a failed disk, whereas *rebuild priority low* refers to reserving only 10% (on the average) of array CPU for rebuild. The MTTRs listed in Table 2 account for rebuild priority, RAID level, and the number of 73.4 GB disks in a group. The rebuild times (MTTRs) are based on a performance model developed for the VA 7100. For this comparison study, MTTF(disk) of 1,000,000 hours is used, which is typical in today's industry.

While the VA 7400 is used to illustrate the relative MTTDLs in this paper, other arrays or other disk capacities will show similar relative results.

TABLE 2. Case studies and result summary

Case No.	RAID Level	Disks in Group (N)	Groups in Array (K)	Disks in Array	Rebuild Priority	MTTR in hours	MTTDL (array) in hours	Chance of Data Loss (per year)	Storage Efficiency (array)
1	1	2	52	104	High	3.64	2.64 E+09	~ 3/M **	50.0%
2	5	52	2	104	High	30.44	6.19 E+06	~ 1000/M	98.1%
3	5	15	7	105	High	9.90	6.92 E+07	~ 100/M	93.3%
4*	5	6	18*	105	High	7.42	2.64 E+08	~ 30/M	82.9%
5	5DP	52	2	104	High	30.44	8.14 E+09	~ 1/M	96.2%
6	5DP	15	7	105	High	9.90	1.08 E+12	~ 0/M	86.7%
7	5DP	52	2	104	Low	152.22	3.25 E+08	~ 30/M	96.2%
8	5DP	15	7	105	Low	49.48	4.33 E+10	~ 0/M	86.7%

* Case 4 has 17 groups of six (6) disks each. The 18th group has only three (3) disks.

**The chance of data loss per year for Case 1 is 3 per million (3/M).

Table 2 shows the results for a fully configured array (104 or 105 disks maximum). For example, Case 1 refers to the array being configured with 52 groups. If each group has two disks in RAID 1, then:

- N = 2
- MTTF(disk) = 1,000,000 hours
- MTTR(group) = 3.64 hours
- K = 52

Using Equation 1, MTTDL(group) is 1.38E+11 hours and using Equation 3, MTTDL(array) is 2.64E+09 hours. Then, using Equation 4, SE(array) is 50%. (*For a list of equations, see Table 1 on page 4*.) In this paper we will use MTTDL to measure he risk of losing data, however, the chance of data loss per year is also provided in Table 2 as an alternative metric.

Below are the results for a fully configured array (104 or 105 disks), for Table 2.

1. RAID 5 has a higher risk of data loss (lower MTTDL) than RAID 1. However, RAID 1 is *less efficient* in storage than RAID 5.

   For example, Case 3 is approximately 38 times higher in risk of data loss than Case 1; Case 3 storage efficiency is 93.3% compared to only 50% for RAID 1.

2. RAID 5DP has a lower risk of data loss (higher MTTDL) than RAID 1 for equal rebuild priority. In addition, RAID 5DP is *more efficient* in storage than RAID 1.

   For example, case 6 is approximately 400 times lower in risk of data loss than Case 1; Case 6 storage efficiency is 86.7% compared to only 50% for RAID 1.

3. RAID 5DP has a lower risk of data loss than RAID 5 for equal rebuild priority and equal numbers of disks in a group. However, RAID 5 is *slightly more efficient in storage* than RAID 5DP.

   For example, Case 6 is approximately 15,000 times lower in risk of data loss than Case 3; Case 6 storage efficiency is 86.7% compared to 93.3% for Case 3.

4. RAID 5DP with a *low* rebuild priority still has a lower risk of data loss than RAID 5 with a *high* rebuild priority. In addition, RAID 5DP often has *lower risk of data loss* than RAID 5 even for a larger group size than RAID 5.

   For example, Case 7 (the largest group size in this study for RAID 5DP) is approximately 1.23 times (23%) lower in risk of data loss than Case 4 (the smallest group size in this study for RAID 5). However, the Case 7 storage efficiency is 96.2% compared to only 82.9% for Case 4. Although the Case 7 rebuild priority is lower than Case 4, ***RAID 5DP is still better*** (risk to data loss and storage efficiency) than RAID 5.

This study was extended to quantify data loss risk and storage efficiency as a function of number of disks in the array. The results are shown in Figures 1-4.

A comparison of storage efficiency is shown in Figure 1 on page 8. For a given number of disks in the array, RAID 1 has the lowest efficiency, while RAID 5 and RAID 5DP are closer to each other. Figure 2 on page 9 shows that for *high* rebuild priority, RAID 5DP has the lowest risk of data loss, followed by RAID 1, and then RAID 5. Figure 3 on page 10 shows that for a given storage efficiency, RAID 5DP is always better than RAID 5 and RAID 1. This observation holds true even when rebuild priority for RAID 5DP is *low* as shown in Figure 4 on page 11.

Conclusions

Listed below are the findings of this study.

1. RAID 1 is the *least efficient* in storage.
2. RAID 5 is *slightly more efficient* than RAID 5DP in storage.
3. RAID 5DP *can tolerate the failure of two disks in a group* as compared to only one disk for RAID 1 or RAID 5.
4. For a given efficiency, ***RAID 5DP is a clear winner*** in data protection (higher MTTDL) compared to RAID 1 or RAID 5, even when rebuild priority for RAID 5DP is *low*.

Another advantage of RAID 5DP is that one disk failure will **not** cause urgency on rebuilding data; therefore the system I/O's performance level is preserved and the rebuild occurs slowly without incurring any risk.

FIGURE 1. Storage efficiency versus number of disks

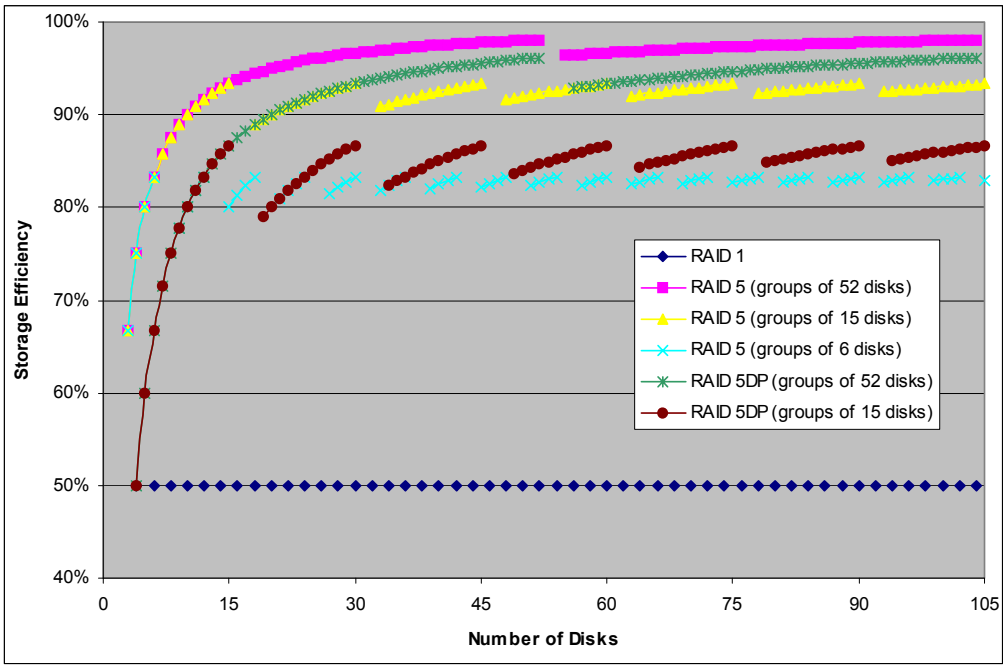
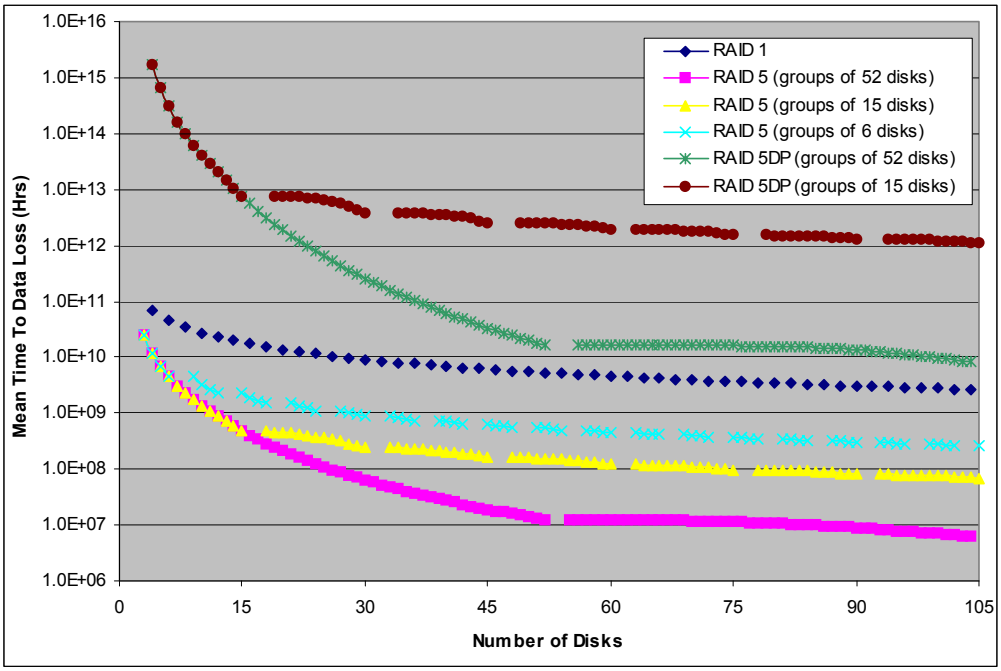


Figure 1 illustrates the storage efficiency of VA 7400 using a variety of RAID levels. The very highest theoretical storage efficiency is obtained with RAID 5. However, as shown earlier in the paper, actually using RAID 5 configured in this fashion is *unacceptably risky* in practice. RAID 5DP delivers nearly the

same level of storage efficiency as RAID 5, but is much *less risky*. RAID 1 configurations have relatively low storage efficiency at all configuration sizes.

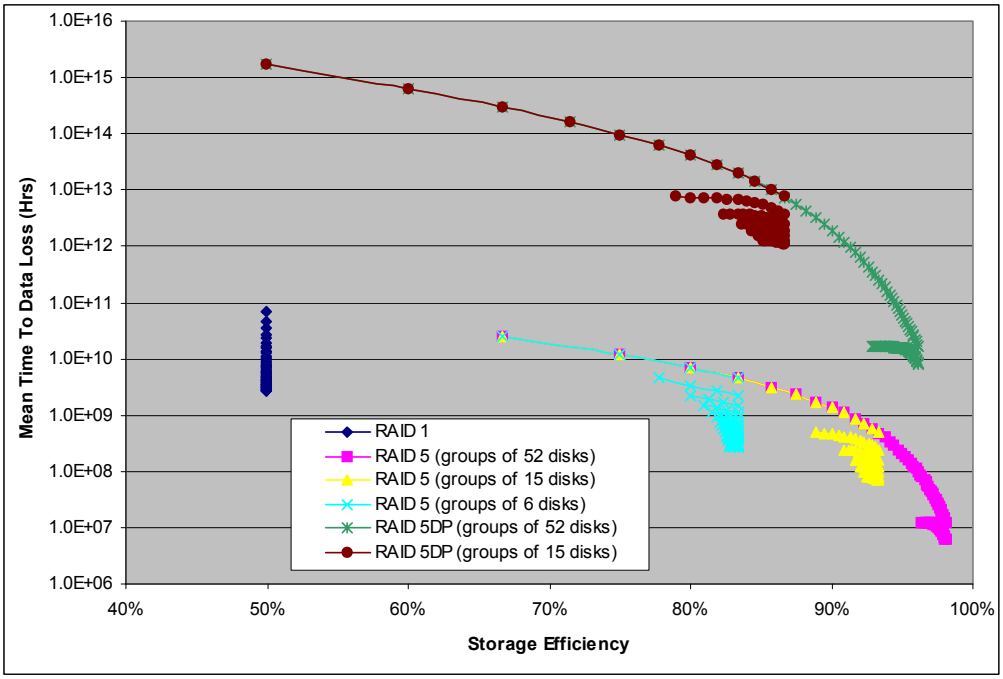
FIGURE 2. Mean-time-to-data-loss versus number of disks



Note: Rebuild priority for RAID 5DP is as *high* as for RAID 1 and RAID 5.

Figure 2 illustrates RAID 5 in very wide stripes presenting the shortest mean-time-to-data-loss. This shows why RAID 5 is rarely used in groups of more than 6 disks. By contrast, RAID 5DP configurations in groups of either 15 or 52 disks maintain a *very long* and *low risk* mean-time-to-data-loss.

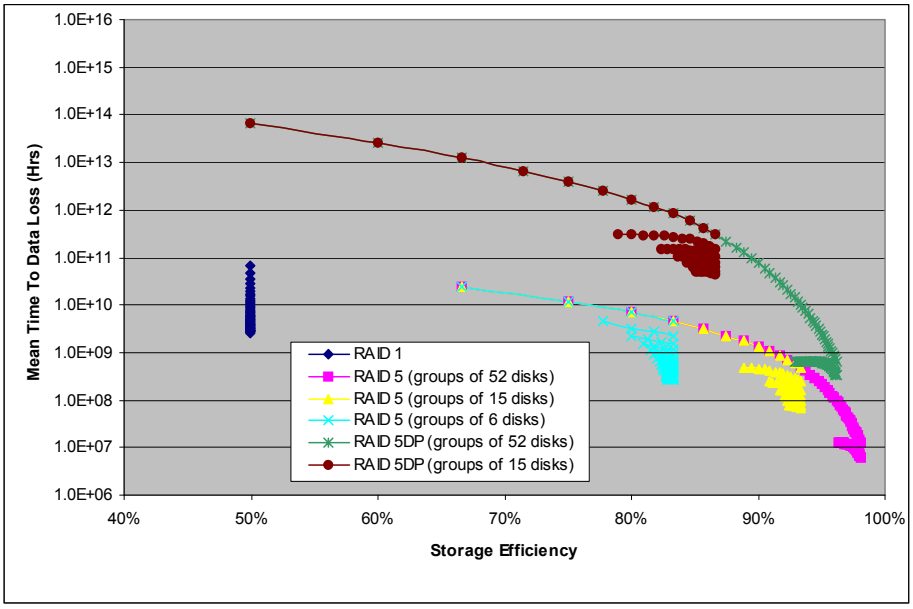
FIGURE 3. Rebuild priority high — mean-time-to-data-loss versus storage efficiency



Note: Rebuild priority for RAID 5DP is as *high* as for RAID 5 and RAID 1.

Figure 3 illustrates mean-time-to-data-loss at a given storage efficiency where rebuild priority is high for all RAID levels. For a certain storage efficiency percentage, RAID 5DP offers much longer mean-time-to-data-loss than does an equivalent RAID 5 configuration. Furthermore, RAID 1 offers neither the best in storage efficiency, nor the best mean-time-to-data-loss at that efficiency.

FIGURE 4. Rebuild priority low — mean-time-to-data-loss versus storage efficiency



Note: Rebuild priority for RAID 5DP is *low*, while rebuild priority for RAID 5 and RAID 1 is *high*.

In Figure 4, we compare the mean-time-to-data loss versus storage efficiency. Again, at any given storage efficiency, RAID 5DP has a longer mean-time-to-data-loss than an equivalent RAID 5 configuration, even with the RAID 5DP being handicapped with a low priority rebuild setting. Even given this setting, RAID 5DP shows a longer mean-time-to-data-loss than an equivalent RAID 5 configuration with *rebuild priority* set ***high***.