Technical report:

Microsoft Exchange Server 2007 and IBM System Storage N series with RAID-DP

Best practices

Document NS3574-0

April 11, 2008

Table of contents

Abstract

Microsoft Exchange Server 2007 is a mission-critical messaging application that places several strict demands on underlying storage systems. This paper compares four RAID types from three different angles: protection, performance and price. Based on test results, internal expertise and data in the public domain, it is clear that the IBM System Storage N series RAID 6 implementation, RAID-DP, is the best and most reliable technology for enterprise Exchange Server 2007 environments.

Background

Electronic messaging systems have become recognized in recent years as a mission-critical application. With the introduction of Microsoft® Exchange Server 2007, deployments of Exchange Server utilizing larger low-cost mailboxes together with large-capacity hard drives (greater than 250 GB) is not only desirable but also necessary in order to keep disk spindle count and operation costs manageable. With the increase in the amount of stored Exchange data comes the increased risk of data loss due to a variety of factors. In order to protect Exchange data, many large enterprise customers have invested significant resources in High Availability (HA), Disaster Recovery (DR) and Business Continuance (BC) solutions. Usually these solutions involve the deployment of some form of RAID (Redundant Array of Inexpensive Disks). A solid RAID layer will protect against disk failures and unrecoverable media errors, and at the same time provide a solid foundation for HA, DR, and BC solutions.

This paper examines the probabilities of data loss when using RAID technology and provides guidance in the use of RAID—and, particularly, IBM® System Storage™ N series with RAID-DP™—for deployments involving Exchange Server 2007. This paper does not examine the factors associated with drive failure or data corruption but rather to contrast the way in which selected different RAID types impact data availability.

Motivation

While many papers and marketing collateral describe and highlight RAID-DP technology and its benefits in general, additional application-specific information is helpful. For example, under an Exchange (especially an Exchange 2007) workload, how efficiently and effectively does IBM N series RAID-DP perform? Are there any performance penalties for using RAID-DP with Exchange? How long does it take to rebuild a RAID-DP group while Exchange is still running? This paper attempts to answer these questions.

Additionally, there have been questions about how to best size Exchange Server 2007 taking into account both Microsoft and IBM N series best practices. These issues include:

- On the Microsoft TechNet website, the Microsoft Exchange Server 2007 *Planning Your Deployment* document [TechNet07] recommends RAID 10 as a best practice, while the IBM N series best practice is RAID-DP.
- On the same website, under *Planning Disk Storage* [TechNet07], Microsoft states, ”*RAID6 adds an additional parity block and provides approximately double the data protection over RAID5, but at a cost of even lower write performance.”*

- Additionally, in the same section, Microsoft also states that RAID 6 has poor rebuild performance, poor disk failure performance and poor I/O performance.
- And, the Microsoft Exchange Server 2007 storage calculator does not have an option for RAID-DP. If the RAID 6 option is used to estimate disk spindle count for RAID-DP, the calculator grossly overestimates spindle requirements, under the assumption that all RAID 6 implementations have a write penalty of 6 disk I/Os per host write.

This paper discusses IBM N series with RAID-DP and explains why it is a Microsoft Exchange Server 2007 storage best practice and why it does not suffer from the poor performance factors of other RAID types. When choosing a RAID type, three characteristics of the RAID implementation are important to understand: protection provided, performance required and cost of deployment. The following three sections discuss these three subjects, respectively.

Protection

Since the inception of RAID in the 1980s [Patterson88], RAID technology has primarily been used for providing fault-tolerance to disk failures. Additionally, with the increasing data density of disk drives, RAID also provides the ability to handle the greater likelihood of encountering an unrecoverable read error. Today, the most commonly used RAID types are RAID 5 and RAID 10 [SNIA07, RAID07]. A new RAID type, RAID 6 [SNIA07], capable of protecting against both double-disk failure and undiscovered, latent data corruptions during rebuild, is gaining popularity, especially with the increased use of large-capacity disks. It is important to note that even under the same RAID type (e.g., RAID 5) there exist many different storage vendor implementations. All provide similar levels of protection, yet have different performance characteristics and rebuild mechanisms.

A quick search in the United States Patent and Trademark Office website, www.uspto.gov, reveals that over one thousand RAID-related patents have been issued. Many of these patents involve incremental improvements to RAID technology. However, only a few of these patents represent major innovations. RAID-DP falls into this category. RAID-DP is a high performance implementation of RAID 6. It meets the RAID 6 definition and standards as defined by the Storage Networking Industry Association (SNIA) [SNIA07]. In an IBM N series storage system, RAID-DP minimizes the performance penalties associated with RAID 6 (as seen in several implementations) through integration with IBM System Storage N series with Write Anywhere File Layout (WAFL®) [Corbett04].

The premise of RAID is that by providing additional redundant information in each RAID group, the circumstance of a single-disk failure in any RAID[1] group will result in no data loss regardless of whether the type of RAID being used is RAID 10, RAID 5, RAID 6 or RAID-DP. All of the four RAID types[2] mentioned will try to recreate the lost disk on an unused “spare” drive using the redundant data – this process is known as a rebuild. However, what is the probability of data loss during rebuild? What is the probability of data loss due to double-disk failure?

[1] Disk striping is commonly called RAID 0 [SNIA07]. RAID 0 does not protect against disk failure, and is excluded from this discussion.
[2] RAID 5 and RAID 4 have the same failure scenarios, protection and rebuild characteristics. Therefore, the discussion of RAID 5 in this section equally applies to RAID 4.

Errors and error handling

Hard Disk Drive (HDD) errors and failures can be divided into three classes, each having a different consequence. The first level of data protection—against the first type of error—is built into the HDD itself. When data is recorded, error correcting codes (ECC) are interleaved with the data. These are then read back whenever the data is read. If the ECC does not match the data, the ECC is used to reconstruct the data in less than a revolution of the disk. Generally, these recoverable errors are never seen by the user and are not recorded by the disk drive itself. Sometimes additional time is required while the drive does additional "off-track" reads to recover the data. ECC is used at the drive level regardless of the RAID configuration. The other two failure modes include errors that cannot be corrected with ECC and complete hard disk-drive failures.

Latent defects and corrupted data

A second class of data corruption occurs when either the data is written poorly and has too many errors to be corrected with ECC, or when the data is corrupted (erased) after being successfully written. Drive manufacturers publish a bit error rate (BER), which is typically one bit error per 10^{15} bits read [Seagate04, Corbett04]. However, the BER only accounts for the process of writing data, not for media defects and data corruption after data is written. There are a number of reasons for post-writing corruption [Elerath07a]. Since the disk media can be scratched any time the disks are spinning, good data can become corrupted unbeknownst to the user, resulting in latent data corruption. When the corrupted (erased) data is read, parity across all the other disks in the RAID group allows for the reconstruction of the missing data, usually a relatively small number of sectors. It is then retrieved, passed to the requesting application and resaved in a new location on the disk drive. Thus, the read error rate (RER) is more important to consider than the BER.

By knowing the number of discovered defects per GB read and the average number of GBs read, we can estimate the RER as a function of time for use in the model. In late 2004, a study was completed on 282,000 HDDs used in RAID architecture. The RER, averaged over three months, was $8x10^{-14}$ errors per byte read. At the same time, another analysis of 66,800 HDDs showed a RER of approximately $3.2x10^{-13}$ errors per byte. A more recent analysis of 63,000 HDDs over five months showed a much improved $8x10^{-15}$ errors per byte read. In these studies, data corruption is verified by the HDD manufacturer as an HDD problem and not as a result of the operating system controlling the RAID group.

Conversations with engineers from four of the world's leading HDD manufacturers support the contention that HDD failure rates are usage-dependent, but the exact transfer function of reliability as a function of usage (number of reads and writes, lengths of reads and writes, sequential versus random) is not known. Based on the study of 63,000 HDDs reading $7.3x10^{17}$ bytes of data in five months, we can estimate the read rate to be $2.7x10^{11}$ bytes/day/HDD. Multiplying the errors/byte times the bytes/hour leaves the estimated errors per hour as shown in Table 1. The following analyses stay away from the extremes shown in the table and use 1.08^{-4} per hour as the RER.

Read Errors per Byte per HDD		Bytes Read per Hour: Low Rate $1.35x10^{9}$	Bytes Read per Hour: High Rate $1.35x10^{10}$	
Low	$8.0x10^{-15}$	$1.08x10^{-5}$	$1.08x10^{-4}$	Err/hr
Med	$8.0x10^{-14}$	$1.08x10^{-4}$	$1.08x10^{-3}$	Err/hr
High	$3.2x10^{-13}$	$4.32x10^{-4}$	$4.32x10^{-3}$	Err/hr

Table 1) Read error-rate determination.

Notice that since the RER is dependent on the number of bytes read, the probability of data corruption during a reconstruction increases as the capacity of the drives increases (assuming a constant percent usage).

Operational failures

Drive manufacturers report failure rates on their drives as either annualized failure rates (AFR) or mean time between failure (MTBF). These numbers are estimates often derived from accelerated laboratory tests and unfortunately are not in line with the numbers as seen in actual field deployments. Whether a drive actually meets criteria for failure set by the drive or array vendor, the actual removal of a drive from a RAID group for whatever reason will result in a RAID rebuild or reconstruction.

On average, HDDs have an annualized return rate (ARR) of between 2% to 4% [Schroeder07] and this is in line with published data [Elerath07a]. This means that with an ARR of 2-4% a drive population of 100 drives would expect between two and four reconstruction events in a 12-month period. Extensive data shows that disk drive failure rates are not constant, but change over time [Elerath07a]. However, for the comparative analyses in this paper we will use a constant ARR of 1.9% as the operational failure rate in our calculations.

Restoration from operational failures

A constant restoration rate implies the probability of completing the restoration in any time interval is equally as likely as any other interval of equal length. Therefore, it is just as likely to complete restoration in the interval 0 to 48 hours as it is in the interval 1,000 to 1,048 hours. This is clearly unrealistic for two reasons. First, there is a finite amount of time required for the HDD to reconstruct all the data on the HDD. It is a function of the HDD capacity, the data rate of the HDD, the data rate of the data-bus, the number of HDDs on the data-bus and the amount of I/O transferred as a foreground process. Reconstruction is performed on a high priority basis but does not stop all other I/Os to accelerate completion. However, for comparative purposes, the models used herein all assume constant restoration rates.

Fibre Channel (FC) HDDs can sustain up to 100MB/second data transfer rates, although 50MB/second is more common. The data-bus to which the RAID group is attached has only a 2-gigabits-per-second capability. Later, test data will show that reconstructing a single 15 KB RPM, 144 GB disk in a 16-disk, RAID-DP group required only 1.5 hours. For these relative reliability comparisons, a mean reconstruct time of 12 hours is assumed to allow for the installation of a spare HDD and the added length of time for the larger ATA drives that run at 7200 RPM, and the system is assumed to have a very large amount of foreground activity, thereby slowing reconstruction. These are very conservative assumptions but completely adequate for relative reliability analyses comparing different system configurations.

Data scrubbing

IBM N series systems proactively detect and correct latent data corruptions by continuously performing a "background" scrub operation. During times of low usage, data is read, checked against parity, and corrected if necessary. If not corrected using background scrubbing these corruptions remain on the disk as latent defects for the rest of the life of the data. The significance of latent defects will become apparent in a subsequent section. Since excessive scrubbing impacts performance, it is assumed that the mean time to complete scrubs is 168 hours.

Data loss during RAID reconstruction

A RAID rebuild or RAID reconstruction occurs when a drive in a RAID group is determined by the containing array to have failed and there is an available spare drive to use for reconstructing the data held on the failed drive. The determination of failure conditions is usually up to the manufacturers of the storage arrays but a drive can also be failed as a result of a determination by the administration staff. Regardless of the reason for marking a drive as failed, the fact that a RAID group is running without redundant information available means that the RAID group is potentially at risk for data loss. The probability of data loss during a reconstruction depends on a number of factors:

- Number of disks in a RAID group
- Hard drive’s operational failure rate
- Hard drive's latent defect rate
- Rate of restoration for operational failures
- Rate of discovery and correction for latent defects.

The latent defect rate is, in turn, dependent on:

- Disk capacity
- Amount of data written
- Amount of data read
- The BER (bit error rate)
- The data corruption rate.

In order for data loss to occur, a combination of latent defects and operational failures must exist on multiple disks simultaneously. For each RAID type considered in this paper this unrecoverable error results from a different sequence of events.

RAID 10 With RAID 10, the most likely sequence of events is a latent data corruption in either of the two drives in the pair followed by the operational failure of the second disk drive. Another likely sequence is an operational failure of the surviving drive of the mirror pair prior to the completion of the reconstruction operation to remedy the first failure.

RAID 5 With RAID 5, the most likely sequence of events is a latent data corruption in any of the drives in the RAID group followed by an operational failure of any other drive in the same RAID 5 group. Another likely sequence is a second operational failure in any other drive in the same RAID 5 group prior to the completion of the reconstruction operation to remedy the first failure.

RAID 6 For RAID 6, the most likely sequence is a latent data corruption in any of the drives in the RAID 6 group, followed by an operational failure of a second drive in the same RAID 6 group during reconstruction of the first drive, followed by the operational failure of a third drive in the same RAID 6 group during the reconstruction period. Since RAID-DP is an implementation of RAID 6, the “n+2” model in Section 0 applies to RAID-DP as well as RAID 6.

Probability of data loss

The three RAID configurations mentioned in the prior section require different models to assess the probability of failure. All of the models include two failure rates (latent defects and operational failures) and two restoration rates (operational restoration and scrubbing). The difference in overall probability of failure results from the different RAID configurations.

The models used for these assessments are too complex to write as a set of equations, although many try. The best method, which accounts for nonconstant failure and repair rates, is by “Monte Carlo simulation” [Elerath07b]. However, for comparative purposes of this paper we will assume all transition rates (failure, latent defect, scrub and restoration) occur at constant rates. Even so, a simple set of equations is not possible, so a "Markov model" [Elerath07b] is employed to assess the probabilities of failure. Each of the models is depicted by "state diagram," transition rates and quantities. The failure rates are represented by a quantity, "λ" and a subscript, Ld for latent defect and Op for operational failure. The restoration transitions are "μ" and a subscript, Op for restoration of an operational failure and Scrub for repair of corrupted data through a scrub action.

Both RAID 10 and RAID 5 are "n+1" models, but the number of data disks in the RAID group is different. Thus, the general models for RAID 10 and RAID 5 are exactly the same, but the transition rate multipliers change according to the number of data disks. The RAID 6 model is more complex and is treated separately. In RAID 10 and RAID 5 there are "n" data disks and 1 parity disk (or equivalent), whereas RAID 6 has "n" data disks and 2 parity disks.

The "n+1" model (RAID 10 and RAID 5)

Often, the process of double-disk failures first concerns the probability of an operational failure followed by a latent defect (read error) while attempting to recover. If you realize that latent defects occur at 10-100 times the frequency of operational failures, a better way to view the problem is as follows:

- Latent defects are constantly occurring and being corrected via scrubbing.

If a drive in a RAID group experiences an operational failure, what is the probability that any *other* drive in the RAID group already has an undiscovered latent defect, preventing reconstruction of all data?

Since reconstruction of the data on the failed drive requires all the data on all other drives in the RAID group, any single latent defect on any disk in the RAID group will result in failure to complete the reconstruction. Alternatively, if all the data is uncorrupted on all other drives in the RAID group, then a second process to arrive at a double disk failure (DDF) is to have a second operational failure during the relatively short time to reconstruct.

A diagram depicting the (n+1) models is shown in Figure1. The model assumes that at time=0 hours, all drives are operational and have no latent defects (State 1). State 2 represents the condition in which one or more latent defects have occurred. The transition to this state occurs with the rate $(n+1)\lambda_{Ld}$. Once in this state, two things can happen: scrubbing can remove the latent defect (with rate μ_{Scrub}) and the RAID group returns to the pristine state of no failures and no latent defects (State 1), or a *different* drive suffers an operational failure, which sends the system into State 4, a failure state with one latent defect and one operational failure. Since the operational failure must be in a drive other than the one with the latent defect, only *n* drives are at risk for operational failure, so the transition rate from State 2 to State 4 has a multiplier of *n*.

An alternative path from State 1 is to State 3, in which one of the drives experienced an operational failure. Since any of the (n+1) drives can fail, the rate multiplier is (n+1). From State 3, a repair of the operational failure will return the RAID group to State 1. The last possible path is from State 3 to State 4 by having a second operational failure, but since one drive has already failed for this scenario, there are only *n* possible combinations of HDDs that can fail, so the multiplier is *n*.

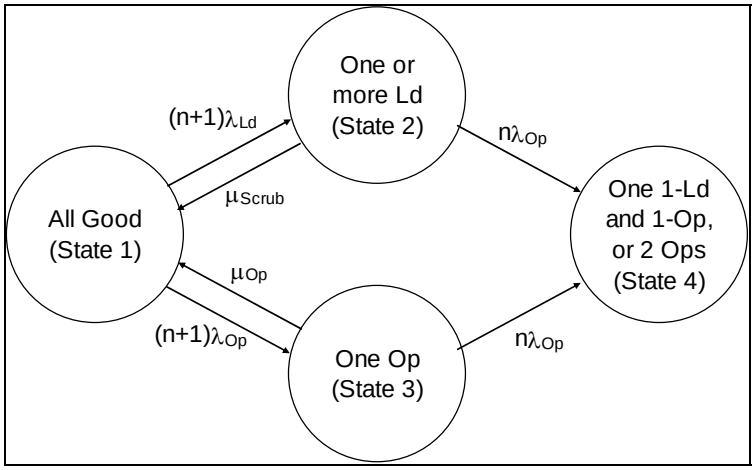

Figure 1) Markov model for RAID 10 and RAID 5.

In the RAID 10 system, there are 2 drives in the group, one data drive and one parity drive. Either can sustain a latent defect or an operational failure, so the transition rate (failure rate) has a multiplier of 2. The plot in Figure 2 shows the number of double-disk failures (DDFs) on the left vertical axis and the probability of failure on the right vertical axis, for a 5-year period, as a function of RAID group size between 2 (RAID 10) and 22 disks. It is apparent that as the number of disks in the group increases, the probability of failure (the number of expected DDFs) increases nonlinearly.

Another concept to consider is an *aggregate* which is a group of RAID groups. For example, suppose one has six RAID groups with eight drives in each RAID group. There are seven data drives in each RAID group for a total of 42 data drives in the aggregate. Table 2 shows seven combinations of n+1 RAID that yield aggregates with 42 data disks including the number of "overhead" drives used inherent to the particular configuration.

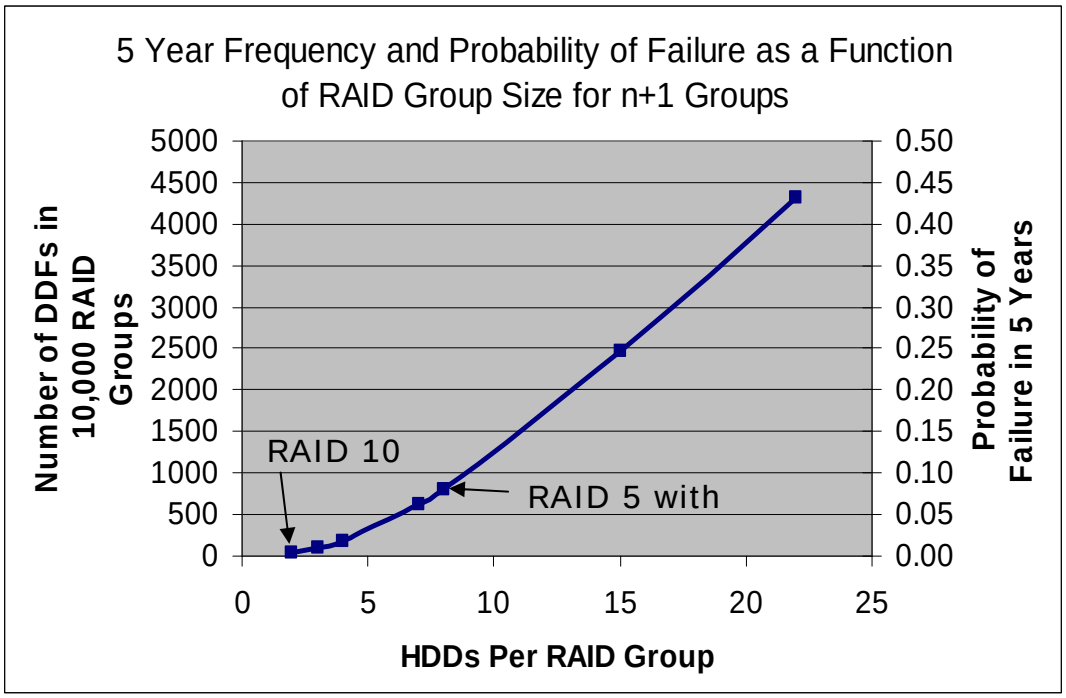

Figure 2) RAID group size comparison.

HDD per RAID group	RAID groups in aggregate	HDD per aggregate	Number of "overhead" HDDs
2	42	84	42
3	21	63	21
4	14	56	14
7	7	49	7
8	6	48	6
15	3	45	3
22	2	44	2

Table 2) Aggregates of "n+1" RAID groups for 42 data disks.

Assuming that aggregate failure occurs when any RAID group in the aggregate sustains data loss adds another dimension to the model. Since all RAID groups in the aggregate must succeed, the reliability is the reliability of the RAID group raised to the power of the number of RAID groups in the aggregate. So in the example of eight drives in the RAID group and six RAID groups, the aggregate reliability is as follows:

$$R_{Aggregate} = \left(R_{RAIDGroup}\right)^{NumberofRAIDGroups} = 0.92^6 = 0.60$$

Knowing that the probability of failure is the complement of the reliability results in

$$F_{Aggregate} = 1 - 0.60 = 0.40$$

This means there is a 40 percent chance of the aggregate losing data sometime in the five-year period. The probability of failure and the number of DDFs for five years is plotted in Figure 3. The horizontal axis shows the effect of the RAID group sizes, from two to 22. Remember that RAID 10 is just a special case of RAID 5 in that there is only one data disk and one disk for recovery, so its reliability is the data point farthest to the left in the plot.

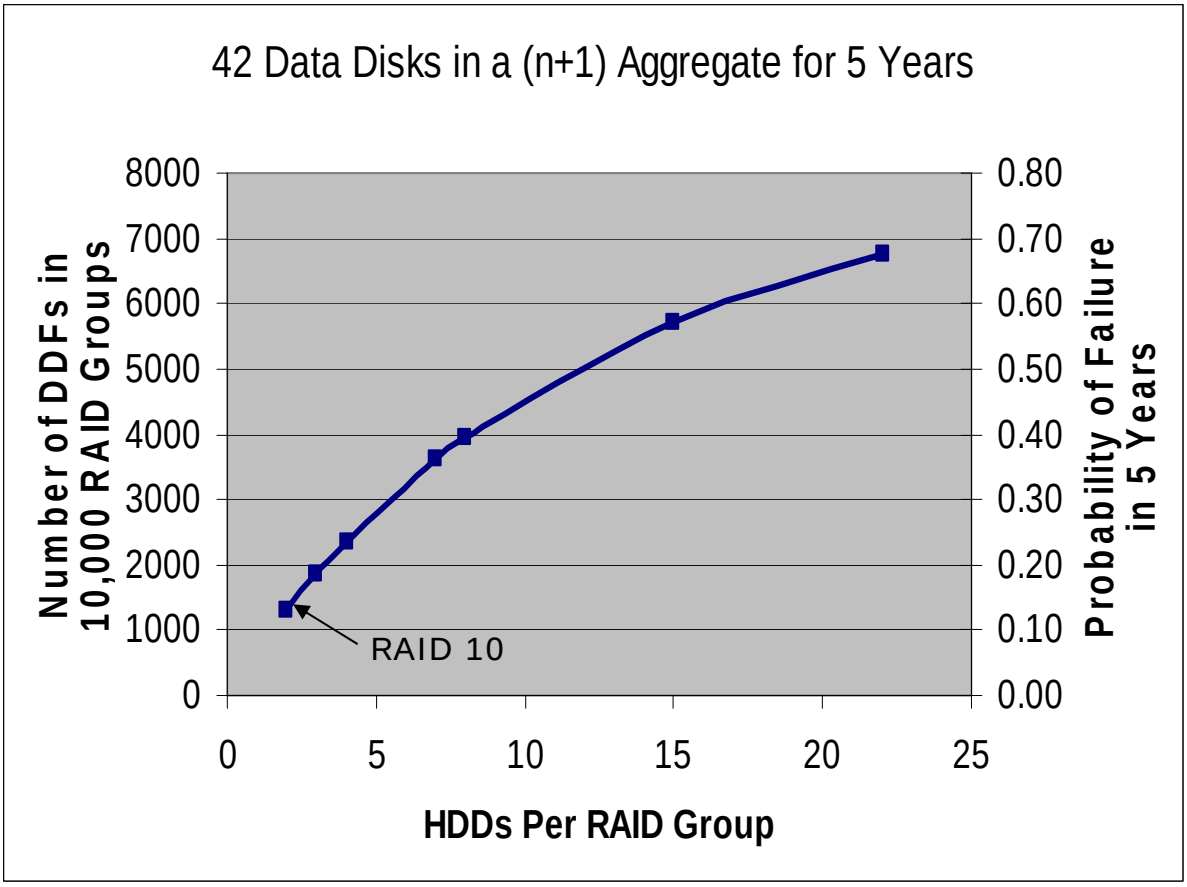

Figure 3) "n+1" aggregate reliability for 42 data disks.

The "n+2" Model (RAID-DP)

The model for RAID-DP, which has two parity disks so it has (n+2) redundancy, is more complex than that of the n+1 configuration. A third failure must occur for the RAID group to fail. In the model, shown in Figure 4, the right-most state (State 6) is the failure state. The left-most state (State 1) is the "good" state and the four states in the middle are degraded states. Multiple latent defects (two or three) are not valid combinations for data loss unless they happen to be in the same data stripe on different disks of the same RAID group. This is very remote, and less likely than the other combinations of operational failures and a single latent defect, so that combination is not included in the model.

The failure conditions for n+1 (State 4 in Figure 1) must be separated into two states (State 4 and State 5 in Figure 4), and a new state (State 6) added. Looking at the notation, it is clear that, unlike the n+1 model that needed only 2 simultaneous events for data loss, the RAID-DP requires 3 simultaneous events. In this model only the combinations of Ld-Op-Op and Op-Op-Op (State 6) result in failure. Order of occurrence is important in that the latent defect must be the 1st or 2nd event, but cannot be the 3rd. Moving from left to right, the number of disks at risk decreases from n+2, to n+1, to n. The reliability for RAID-DP groups is far greater than for RAID 10, 4 or 5. This is evident in Figure 5, in which only three to four failures in 10,000 RAID groups are expected over five years for a RAID group of 20 data disks (22 disks in total). In contrast, Figure 2 indicates that we would expect 33 failures in 10,000 RAID groups for RAID 10 (with only one data disk per group).

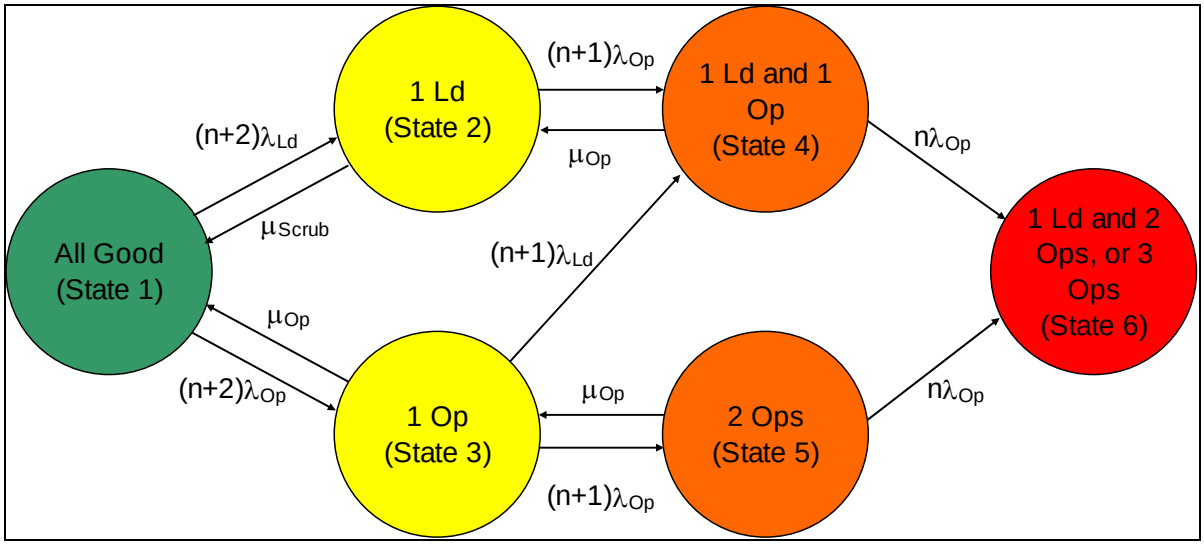

Figure 4) "n+1" aggregate reliability for 42 data disks.

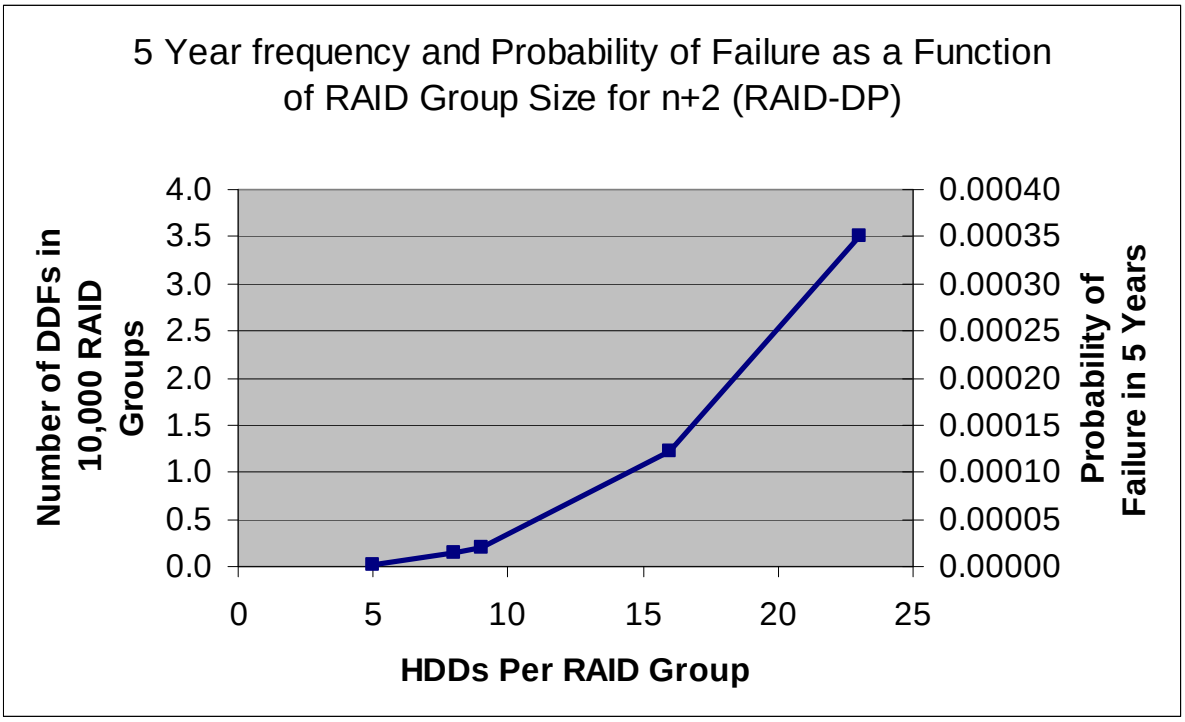

Figure 5) Reliability for RAID-DP (n+2).

Again, going to the concept of aggregates, Table 3 shows five different RAID group sizes that easily form aggregates of 42 data disks for RAID-DP (n+2). The number of triple disk failures and the probability of a triple disk failure are shown in Figure 6. The reliability for n+2 aggregates is calculated the same as for n+1, but the reliability of the n+2 RAID group is used:

$$R_{Aggregate} = \left(R_{RAIDGroup}\right)^{NumberofRAIDGroups}$$

HDD per RAID group	RAID groups in aggregate	HDD per aggregate	Number of "overhead" HDDs
5	14	70	28
8	7	56	14
9	6	54	12
16	3	48	6
23	2	46	4

Table 3) Aggregates of "n+2" RAID groups for 42 data disks

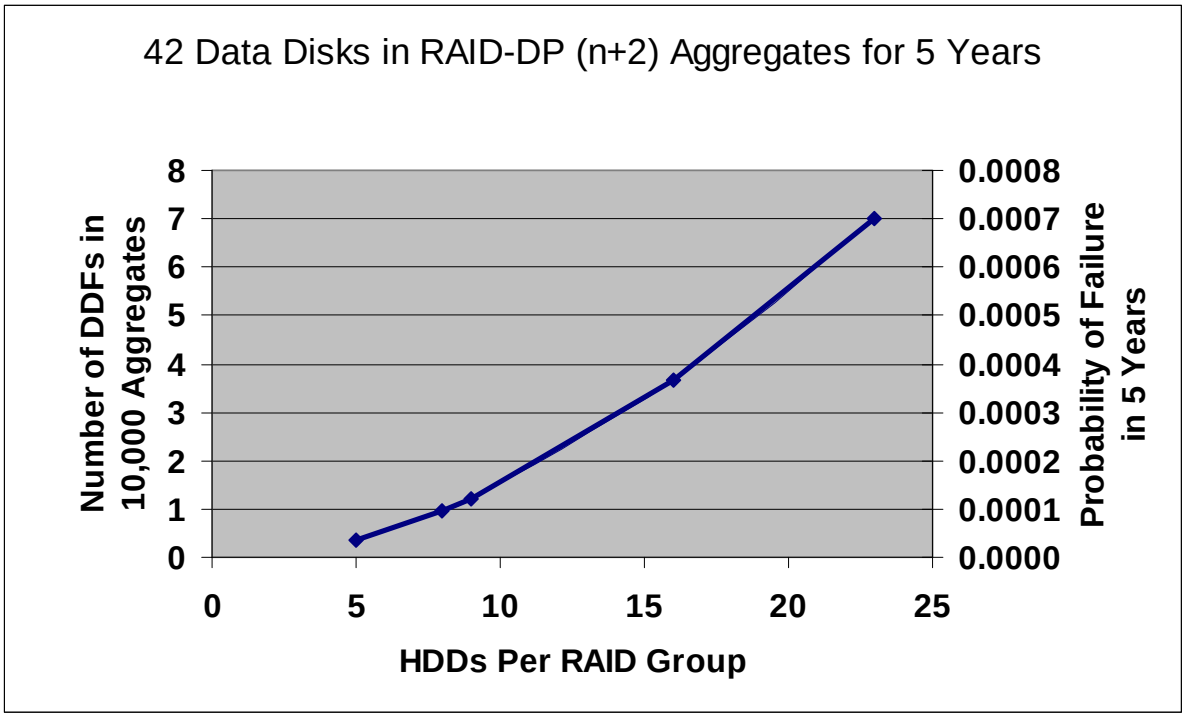

Figure 6) RAID-DP (n+2) aggregate reliability for 42 data disks.

Relative reliabilities for RAID 10, RAID 5 and RAID-DP

When selecting RAID types, the first and foremost question customers need to answer is: What kind of data loss risks can I tolerate? Table 4 shows that the probability of data loss with RAID-DP is 0.002% in five years for RAID groups with seven data disks. For RAID 10, with only one data disk, the probability of data loss in five years is 0.33%, or 163 times as likely as RAID-DP, even though the RAID-DP group has seven times the capacity of the RAID 10 group. With RAID 5, the probability of data loss is approximately 6% in five years for seven data disks, approximately 4,000 times as likely to occur as in RAID-DP. Only RAID-DP and RAID 6 can best protect against data loss for large configurations.

RAID type	Probability of data loss in 5 years	Risk of data loss relative to RAID-DP
RAID 10 (1 data disk)	0.33%	163
RAID 5 (7 data disks)	6.0%	3955
RAID 6 (7 data disks)	0.002%	1.0
RAID-DP (7 data disks)	0.002%	1.0

Table 4) RAID type vs. probability of data loss in five years.

Performance

The performance of RAID controllers can vary widely from one to the other. Even for a given RAID controller, the performance characteristics are different for different RAID levels. Typically, there are four performance aspects of any RAID controller that matter most to customers:

- Overall performance
- Write performance
- Performance in degraded mode[3] and during rebuild
- Rebuild time.

The following subsections discuss each area with respect to Exchange Server 2007.

Overall performance: Jetstress results

A good way to measure the overall storage system performance under the Exchange workload is to use the Jetstress[4] tool. This is mainly because Jetstress can accurately simulate Exchange 2007 I/O workload [Quimbey07]. The acceptable Exchange I/O performance is that the average read and average write latencies are below 20 ms as measured from the server.

Table 5 shows the Exchange 2007 Jetstress performance results for 10,000 users, with 0.5 IOPS per user. The results were taken directly from Jetstress HTML reports. The test was successful, as both the average database read and average database write latencies are below 20ms, and the test achieved 5320 IOPS, which exceeds the planned IOPS of 5000.

Database LUNs	Read Latency (ms)	Write Latency (ms)
G:	10.0	7.0
H:	9.0	4.0
I:	9.0	4.0
J:	10.0	6.0
K:	9.0	4.0
L:	10.0	6.0
M:	10.0	7.0
N:	10.0	7.0
O:	10.0	7.0
P:	10.0	7.0

Table 5) Jetstress-performance test results.

[3]The SNIA definition of *degraded mode* is: a mode of RAID array operation in which not all of the array's member disks are functioning, but the array as a whole is able to respond to application read and write requests to its virtual disks [SNIA07].

[4] The Jetstress tool used in this work is the Microsoft released version of Jetstress for Exchange 2007, v08.01.0038.

The test was performed on a single IBM System Storage N series N5600 controller using IBM N series RAID-DP and 15 KB RPM FC drives. The storage configuration and disk count are shown in Table 6.

Server	Exchange storage group	NTFS volume	Number of disk drives	Storage system
HP Proliant DL-385 (x64, 32 GB RAM) 10,000 users 200MB/mailbox 0.5 IOPS	ESG1-DB1	G:	63 drives	N5600 Single controller
	ESG2-DB1	H:		
	ESG3-DB1	I:		
	ESG4-DB1	J:		
	ESG5-DB1	K:		
	ESG6-DB1	L:		
	ESG7-DB1	M:		
	ESG8-DB1	N:		
	ESG9-DB1	O:		
	ESG10-DB1	P:	11 drives	
	ESG1-Logs	Q:		
	ESG2-Logs	R:		
	ESG3-Logs	S:		
	ESG4-Logs	T:		
	ESG5-Logs	U:		
	ESG6-Logs	V:		
	ESG7-Logs	W:		
	ESG8-Logs	X:		
	ESG9-Logs	Y:		
	ESG10-Logs	Z:		

Table 6) Exchange 2007 Jetstress storage configuration for 10,000 users.

The test demonstrates that RAID-DP provides excellent performance for Exchange Server 2007.

Write performance

The write performance of RAID implementations has always been a subject of great debate and discussion. Without optimization, RAID 10 could have a write penalty of two disk I/Os: one write to the first disk and another write to the mirrored disk. RAID 5 could have a write penalty of four disk I/Os: read data, read parity, write data and write parity. By the same token, RAID 6 could have a write penalty of six disk I/Os: read data, read the 1st parity, read the 2nd parity, write data, write the 1st parity, and write the 2nd parity.

What about IBM N series RAID-DP? Since RAID-DP is officially recognized by SNIA as a valid RAID 6 implementation, those who are unfamiliar with this technology often associate RAID-DP with a write penalty of 6 disk I/Os. However, with RAID-DP this is not the case at all! In fact, RAID-DP write performance is excellent. This is because of the integration of RAID-DP with IBM N series WAFL [Corbett04].

Figure 7 shows the disk write operations to host write operations ratio during the 10,000-user Jetstress test using RAID-DP. The disk_writes is calculated by aggregating the number of write operations to the 63 database disks (see Table 6). The host_writes is arrived by adding the number of host write operations to the 10 database LUNs. Then, the disk to host write operations ratio is computed and plotted. Note that the disk to host write operations ratio is substantially below 1.0.

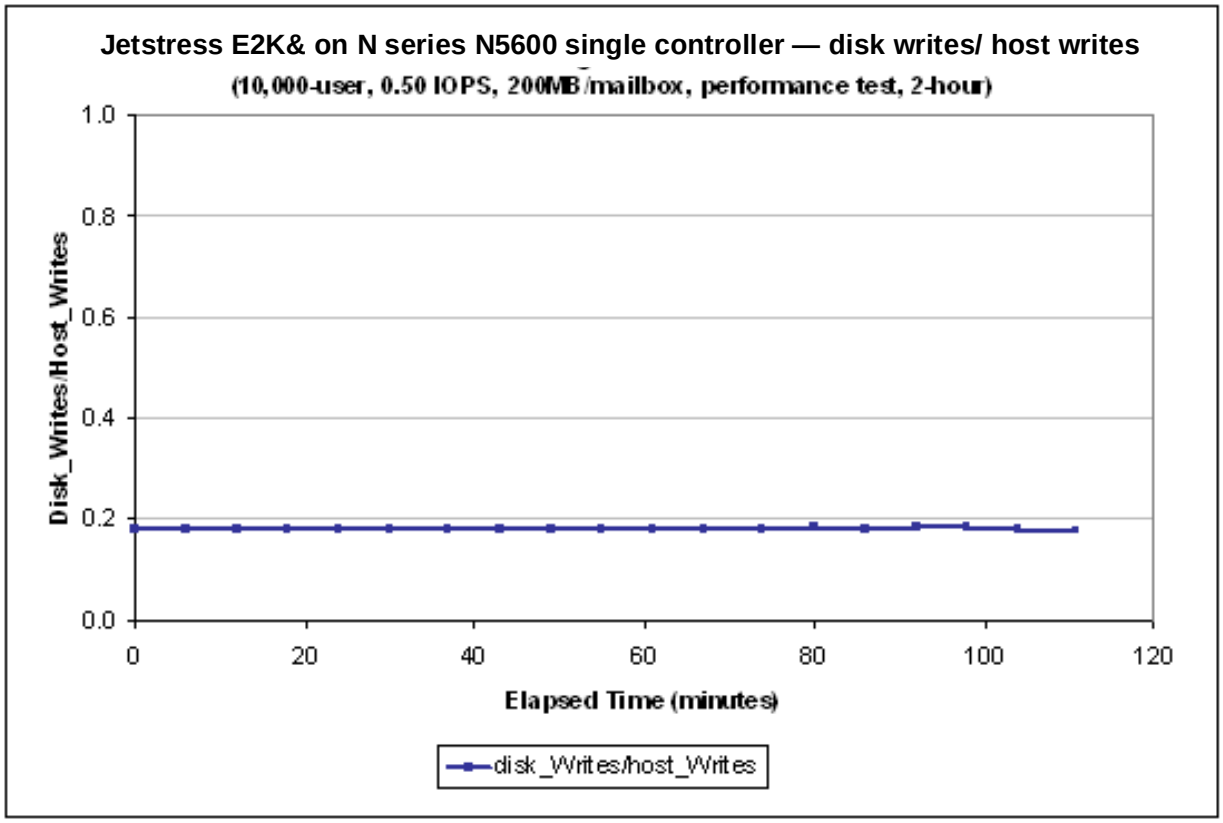

Figure 7) Disk-to-host write operations ratio (measured on the N5600 during the Jetstress test).

On the other hand, some RAID 6 implementations do have a write penalty [R6Perf07] as discussed above. In order to compensate for the additional I/O required when doing write operations, vendors may use more disk spindles thus increasing the initial upfront cost as well as the on-going operations cost, because more disk spindles lead to more data center space, power consumption and cooling costs.

Degraded-mode performance

After a disk failure event, a RAID group will go into the so-called *degraded mode* [SNIA07], where both its protection and performance are degraded. For IBM N series RAID-DP, the rebuild process[5] starts automatically. During the rebuild process, regardless of RAID types, the performance of servicing host I/Os is further reduced because some resources are used to rebuild the failed RAID group. Since the rebuild of RAID-DP is streamlined and automatic, the degraded mode and rebuild stage virtually overlap completely. This shortens the time a RAID group is in the degraded mode.

Table 7 shows the performance impact of a RAID-DP rebuild process due to a single disk failure on an Exchange 2007 workload (from the same 10,000-user Jetstress test). Compared to Table 5, the average database read latency degraded to 15 ms from 10 ms. The average database write latency more or less stayed the same. Note that both average read and average write latencies are still well below 20 ms.

Database LUNs	Read Latency (ms)	Write Latency (ms)
G:	15.0	7.0
H:	14.0	4.0
I:	14.0	4.0
J:	14.0	7.0
K:	14.0	4.0
L:	15.0	6.0
M:	14.0	7.0
N:	15.0	7.0
O:	15.0	7.0
P:	14.0	7.0

Table 7) Jetstress performance during RAID-DP reconstruction.

In general, during the RAID rebuild process (regardless of RAID 5 or RAID 6), both read and write performance would likely suffer. Only IBM N series RAID-DP is capable of limiting the performance hit to read only. Again, this is accomplished by integrating RAID-DP with WAFL [Corbett04].

[5] The rebuild process automatically activates on all IBM N series storage systems provided a hot-spare disk drive of the appropriate type is available.

Rebuild time

How long does it take to rebuild a RAID group after a disk failure? This is a very important factor to consider when choosing a RAID system. The longer the rebuild time, the longer end users will suffer from poor performance and the bigger the risk of data loss. If the rebuild requires human intervention, then that adds to the length of time the user data is in danger.

Figure 8 shows the rebuild time of a RAID-DP group consisting of 16 FC drives (15 KB RPM, 144 GB). From the instant of a disk failure, it only took one second for the rebuild process to start automatically. And it took one hour and 30 minutes to complete the rebuild. This translates to a reconstruction rate of ~27 MB/sec. During the entire period of the RAID-DP reconstruction, the system was under active load from the 10,000-user Exchange 2007 Jetstress test.

The rebuild time of one hour and 30 minutes under active Exchange 2007 workload is excellent. Some RAID 5 or RAID 6 arrays on the market would take eight to 10 hours to rebuild [Treadway05].

```
Fri May 11 17:39:00 GMT [raid.config.filesystem.disk.missing:info]: File system Disk
/aggr_100ku_sgdbs/plex0/rg0/2a.19 Shelf 1 Bay 3 [NETAPP   X275_S15K4146F15 NA01] S/N
[3KN1TX4400007648ANJJ] is missing.
Fri May 11 17:39:00 GMT [raid.rg.recons.missing:notice]: RAID group /aggr_100ku_sgdbs/plex0/rg0 is
missing 1 disk(s).
Fri May 11 17:39:00 GMT [raid.rg.recons.info:notice]: Spare disk 2a.27 will be used to reconstruct one
missing disk in RAID group /aggr_100ku_sgdbs/plex0/rg0.
Fri May 11 17:39:01 GMT [raid.rg.recons.start:notice]: /aggr_100ku_sgdbs/plex0/rg0: starting
reconstruction, using disk 2a.27
Fri May 11 18:00:00 GMT [monitor.raid.reconstruct:warning]: Reconstructing broken data disk in RAID
group /aggr_100ku_sgdbs/plex0/rg0.
Fri May 11 19:09:33 GMT [raid.rg.recons.done:notice]: /aggr_100ku_sgdbs/plex0/rg0: reconstruction
completed for 3b.27 in 1:30:32.76
```

Figure 8) RAID-DP rebuild time, from N5600 log.

In summary, in all four performance categories, IBM N series RAID-DP is the best or among the best. Table 8 summarizes the performance aspects of the four different RAID types.

RAID type	Overall performance	Write performance	Degraded performance	Rebuild time
RAID 10	Best	Best	Best	Best
RAID 5	Good	Poor	Poor	Poor
RAID 6	Poor	Poor	Good	Poor
RAID-DP	Best	Best	Best	Good

Table 8) RAID type qualitative performance comparison.

Price (cost of ownership)

Besides data protection and performance, another important consideration is the total cost of ownership (TCO). This is, to some degree tied, to the number of disks required in the storage solution.

Table 9 uses one example to demonstrate how IBM N series RAID-DP is the most cost-effective solution when compared against RAID 10, RAID 5 and other RAID 6 implementations. The example assumes 4,000 users, 0.33 IOPS, 250 MB and 1,024 MB per mailbox, respectively, and using 10K, 300 GB FC drives. The Microsoft Exchange 2007 storage calculator is used to compute the disk counts for RAID 10, RAID 5, RAID 6, and RAID 6 (no write penalty). For RAID 6, the Microsoft Exchange 2007 storage calculator assumes six disk writes for each host write by default. However, this assumption is inaccurate for IBM N series RAID-DP (see Figure 7). When this assumption is adjusted to one disk write per host write in the computation and all else kept the same, the calculator gives the disk counts shown in the *RAID 6 (no write penalty)* row.

The RAID-DP disk counts were determined by following best practices for sizing Exchange 2007 environments, assuming an IBM N series N5600 storage system.

RAID type	250 MB/mailbox: DB+log disk count	1024 MB/mailbox: DB+log disk count
RAID 10	40	-
RAID 5	-	76
RAID 6	100	104
RAID 6 (no write penalty)	36	80
RAID-DP	24	70

Table 9) Disk count needed for four different RAID types.

When the mailbox size is 250 MB, the Microsoft Exchange 2007 storage calculator recommends RAID 10 and 40 disks to support both Exchange databases and logs. It does not show a disk count for RAID 5, presumably because the disk count for RAID 5 would be greater than 40. If RAID 6 is selected, then the calculator states that 100 disks are required, mainly due to the assumption of the write penalty of six disk I/Os per host write. This assumption may be true for some RAID 6 implementations, but it is absolutely false for IBM N series RAID-DP (see Figure 7). In fact, only 24 disks are needed. ***This is four times better than generic RAID 6, and it also beats RAID 10 by 40 percent!***

When the mailbox size is increased to 1,024 MB, the Microsoft Exchange 2007 storage calculator recommends RAID 5 and 76 disks to support both databases and logs. It does not show a disk count for RAID 10, presumably because the disk count for RAID 10 would be greater than 76. If RAID 6 is selected, then the calculator states that 104 disks are required. Using IBM N series RAID-DP, only 70 disks are needed. ***This is about 30 percent better than generic RAID 6, and it still beats RAID 5 by 9 percent.***

Table 10 translates the disk count to price or cost of ownership. More disks often lead to more operations cost, more frequent disk failures and more complexity.

RAID type	Cost of ownership
RAID 10	Poor
RAID 5	Good
RAID 6	Poor
RAID-DP	Best

Table 10) Costs associated with the 4 different RAID types.

Conclusion

This work demonstrates that for Exchange Server 2007 deployments, IBM System Storage N series with RAID-DP is by far the best RAID technology available in the marketplace today.
RAID-DP is superior not only in protection, but also in performance and cost of ownership.

When fault tolerance is concerned, at the scale of today's enterprise Exchange Server 2007 environment, RAID 5 and RAID 10 provide substantially less data protection than IBM N series RAID-DP. RAID 5 is the most vulnerable to data loss.

When performance is considered, it is important to realize that *not all* RAID 6 implementations have the same write penalty. In fact, IBM N series RAID-DP does not have the write penalty often associated with generic RAID 6, and instead its write performance is excellent. Other performance aspects of RAID-DP, such as rebuild and overall performance, are also among the best of the available RAID types.

In terms of price and cost of ownership, it is well known that RAID 10 is the most inefficient in terms of disk spindle count, leading to poor capacity utilization and high cost. While RAID 5 and RAID 6 are more efficient in space utilization than RAID 10, that efficiency may be offset by the additional disks (and their associated costs) required to compensate for the poor write performance of RAID 5 and RAID 6.

Only IBM N series RAID-DP, through its integration with IBM N series WAFL, is optimized for both efficient space utilization and high performance. Together with the high level of protection it provides, RAID-DP represents the best value and is the most cost-effective and most reliable solution for Exchange Server 2007 enterprise customers.

Source References

These sources (and associated websites) provide useful references to supplement the information contained in this document:

[TechNet07] Microsoft TechNet
Microsoft Exchange Server 2007 Planning Your Deployment
http://technet.microsoft.com/en-us/library/bb124518.aspx

[Patterson88] David Patterson et al.
A Case for Redundant Arrays of Inexpensive Disks (RAID)
ACM SIGMOD International Conference on Management of Data, 1988
http://www.eecs.berkeley.edu/Pubs/TechRpts/1987/CSD-87-391.pdf

[SNIA07] SINA
Dictionary
http://www.snia.org/education/dictionary/r/

[RAID07] Wikipedia
Nested RAID Levels
http://en.wikipedia.org/wiki/Nested_RAID_levels

[Corbett04] Peter Corbett et al.
Row-Diagonal Parity for Double Disk Failure Correction

FAST'04: 3rd USENIX Conference on File and Storage Technologies, 2004
http://www.usenix.org/events/fast04/tech/corbett/corbett_html/

[Seagate04] Seagate
Cheetah 10K.7 Spec
http://www.seagate.com/docs/pdf/datasheet/disc/ds_cheetah10k.7.pdf

[Elerath07a] Jon G. Elerath and Michael Pecht
Enhanced Reliability Modeling of RAID Storage Systems
Dependable Systems and Networks, DSN-2007, Edinburgh, Scotland, 2007

[Schroeder07] Bianca Schroeder et al.
What does an MTTF of 1,000,000 hours mean to you?
FAST'07: 5th USENIX Conference on File and Storage Technologies, 2007
http://www.cs.cmu.edu/~bianca/fast07.pdf

[Elerath07b] Jon G. Elerath
Reliability Model and Assessment of Redundant Arrays of Independent Disks (RAID) Incorporating Latent Defects and Non-homogeneous Poisson Process Events
Ph.D. Dissertation, University of Maryland, College Park, 2007

[Quimbey07] Robert Quimbey
Microsoft configuring, validating and monitoring your Exchange 2007 storage
http://msexchangeteam.com/archive/2007/01/15/432199.aspx

[R6Perf07] Wikipedia
RAID 6 Performance
http://en.wikipedia.org/wiki/Standard_RAID_levels#RAID_6_performance

[Treadway05] Tom Treadway
RAID Reliability Calculations
http://storageadvisors.adaptec.com/2005/11/01/raid-reliability-calculations/

Trademarks and special notices

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

Network Appliance, the Network Appliance logo, RAID-DP and WAFL are trademarks or registered trademarks of Network Appliance, Inc., in the U.S. and other countries.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.

Information concerning non-IBM products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by IBM. Sources for non-IBM list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. IBM has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-IBM products. Questions on the capability of non-IBM products should be addressed to the supplier of those products.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.