Optimal Configuration of Enterprise Storage Capacity for a Telecom Company

Mario O. Krastev[1], Ludmila H. Raikovska[2] and Kiril I. Dakov[3]

***Abstract* – A research on enterprise storage systems optimal disk configuration is carried out for maximum performance and disaster recovery purposes of online applications in a Telecommunication company. The described solution is based on high performance, security and storage capacity demands needed for a typical telecommunication operator. The results can be used for planning of similar solutions and developing data protection scenarios for any modern telecommunication company.**

***Keywords* – Storage systems, Telecom applications, Data protection, capacity optimization**

I. Introduction

Telecommunication operators use a variety of on-line applications, running on different HA systems in complex heterogonous IT infrastructure. The aim of the current paper is to present a methodology for the optimal enterprise storage array disk space configuration based on telecom applications requirements. The proposed solution follows the best practices and standards for data protection; it could be successfully implemented for business continuity and disaster recovery purposes [1, 2]. The main purpose of the research is to show how the proposed configuration and design lead to higher performance and storage utilization results that could be achieved without any additional hardware or software investments.

II. Methodology for Optimal Configuration of Enterprise Storage Capacity

A. Enterprise storage systems requirements

The most important business critical enterprise applications and their requirements for performance, capacity and protection level, that are typical for any Telecom operator and are most important for its daily work, are presented in Table I. These systems' data should be hosted on enterprise storage arrays. The requirements for the main application databases and their typical capacity needs, number of IOPS and needs for local and remote replication are summarised in Table I.

Table I. Main applications in telecom operator

Appli-cation	Protec-tion	Size	Max IOPS	I/O avg size	Type I/O	Read/ Write	Rep licat ion	B C V
Billing	high	3 TB	9000	32 k	Ran-dom	3:1	yes	3 pairs
CRM	high	1,5 TB	9000	32 k	Ran-dom	3:1	yes	3 pairs
ERP	high	2 TB	6000	32 k	Ran-dom	3:1	yes	3 pairs
Develop ment	normal	10 TB	3000	32 k	Ran-dom	3:1	yes	2T 20%
Develop ment, Test, Data ware-house	normal	30 TB	2000	32 k	Sequ ential	4:1	no	no

B. Choice of drive types

The proper choice of disk drives is important for meeting the requirement of all business critical applications. Correct choice and configuration of drives could lead to significant improvement of response time and performance of the applications. In Table II the basic disk drives characteristics of EMC DMX3 storage array [3] are presented.

Table II. Drives specifications

Specification	73, 146, 300 GB 15k rpm	73, 146, 300 GB 10k rpm	500 GB 7,2k rpm
Internal data rate (MB/s)	68-114	47-94	47-94
Avg rotation latency (ms)	2	3	4,17
Avg read seek (ms)	3,5	4,7	8,5
Avg write seek (ms)	4	5,4	9,5
Read single track seek (ms)	0,27	0,20	0,8
Write single track seek (ms)	0,45	0,50	1
Read full stroke seek (ms)	7,4	9,7	16
Write full stroke seek (ms)	7,9	10,4	16,55

In most cases the main factor for drive selection is capacity and rotation speed but this approach leads to many performance and configuration issues. For example if we have an equal hyper volume size over FC drives of 300 GB 15k rpm and 146 GB 10k rpm the higher performance will be achieved with the volume with 146 GB drives although drives are slower. This is because of the fact that 4 drives at 300 GB are needed for one hyper volume or 8 drives at 146 GB. In the first case all IO operation will be distributed over 4 disks and

[1] Mario O. Krastev is with the Faculty of Telecommunications, TU-Sofia, 1000 Sofia, Bulgaria, e-mail: mkrastev@tu-sofia.bg

[2] Ludmila H. Raikovska is with the Faculty of Telecommunications, TU-Sofia, 1000 Sofia, 8 Kliment Ohridski blvd., Bulgaria, e-mail: l.raikovska@snt.bg

[3] Kiril I. Dakov is with S&T Bulgaria eood, 1528 Sofia, 7 Iskarsko shosse blvd., Bulgaria, e-mail: k.dakov@snt.bg

over 8 in the second and thus the slower drives would feature better performance [4].

Therefore not only drive size and rotation speed should be taken into account in the layout configuration design of the storage system, but also latency, access times and configuration features. In most cases is not recommended to use the biggest capacity disk because it will cause performance degradation.

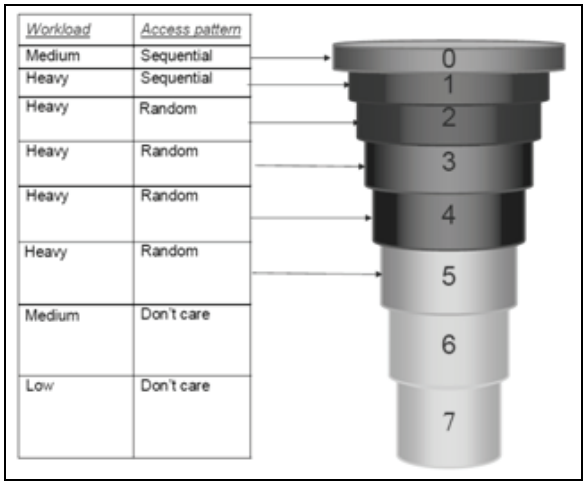

Fig.1. Disk drive workload

The access pattern of a single drive is presented on Fig.1. One of the drive parameters that is most often neglected is the seek time. This is the time for moving drive heads from current position to the proper track. If the volumes are not correctly configured, the drives' heads will move from inner to outer end and vice verse very frequently during random IO operations. It is of major importance to take this into account especially when we have random read/write operations.

For performance oriented applications with a lots of random requests it is recommended that only the middle part of drive plates should be used and not to use about 25% in the inner and outer part [3]. The outer parts are rotating faster than the inner, so they are very suitable for consequential operations.

In the discussed case there are a lot of random queries for most applications so it will be better to use bigger drives, with fast rotation speed and to use only the middle part of the drive. This will minimize the latency of moving of the heads.

The number of maximal theoretical IOPS for each drive can be obtained from the following expression:

$$\frac{1000ms}{AvgRotationLatency + Avr\,\mathrm{Re}\,adSeek} = IOPS \quad (1)$$

Multiplying this IOPS by the number of drives we obtain the maximal theoretical operations number which this part of the array can serve. Maximal IOPS weight from host side is calculated from the formula:

$$\mathrm{ReadOp}.MW + \mathrm{WriteOp}.K.MW = IOPS \quad (2)$$

where

MW – maximal IOPS which a host can handle

ReadOp – percentage of read operations in all IOPS to this host

WriteOp - percentage of write operations in all IOPS to this host

K – RAID protection level coefficient. It shows the number of parallel writes that are done over drives for one single operation. The most common values are K=1 for RAID-0 – no redundancy; K=2 for RAID-10, because all data are written simultaneously on two drives; K=3 for RAID 5 (3+1) – data are written over 3 drives.

Therefore the total number of IOPS which a single host can handle is:

$$MW = \frac{IOPS}{\mathrm{ReadOp} + \mathrm{WriteOp}.K} \quad (3)$$

The proportion between read and write operations for different applications could be obtained from Table I. The above formulas represent only a theoretical evaluation but in a real production system these values should be corrected. This correction is done during the configuration of the storages.

In a configuration process following the best practices of the storage manufacturer EMC [3], one hot spare drive is dedicated for each 30 drives and two drives per director pair are dedicated for service and system information.

C. Optimal drive configuration

In order to obtain a solution, fully compliant to all specific Telco requirements, we propose/choose to divide the storage capacity into different groups. For a Telecommunication company usually at least two groups are necessary. The first group will contain fast volumes for most performance oriented applications with more than 3000 IOPS and the second group will consist of slow volumes for the rest of the applications. Our solution is significantly different from the standard way proposed by the storage vendor. The main purpose of the default configuration is to achieve maximum workload for the whole system and to distribute equally all drives and IOPS between directors, which does not take into account the specific performance needs of Telco applications.

III. TELECOM STORAGE SYSTEM EXAMPLE

The proposed method has been implemented in a real working environment. The infrastructure consists of SAN environment, servers running under IBM AIX operation systems and the storage arrays are EMC Symmetrix DMX3 [3]. During the planning phase of the IT solution 40% increase of the usable capacity and 10% increase of performance needs within the next three years have been taken into account.

The IT infrastructure for information protection and disaster recovery includes three identical storage arrays situated at three different geographical sites, working in concurrent storage based synchronous and asynchronous replication between them [2]. Additional capacity for BCV – business continuity volumes (similar to snapshots and clones) has been

also calculated and the different scenarios for IO workload, needed to achieve storage optimization without additional recourses have been examined.

In the primary data centre the main IT resources are located. Local and remote replications are performed on the main storage array. Local replicas are for tests, backups and statistical needs without impact on the production systems. The proxy site is located about 25 km away from the main data centre (DC) and is synchronized in real time with the primary data centre. In case of failures in the primary DC, all applications except „Development, Test, and Data warehouse" could be started there. The third site is identical with the primary. It is located at about 800 km apart from the central DC and there is asynchronous replication between storages at primary and proxy sites and disaster site. The main idea is that in case of disaster the proxy site should assure synchronous copy of the production data until the disaster site becomes ready to failover all operations and business applications [2]. At the disaster site stand by servers and storages have been located, ready in critical situations to take over all operations. There are secondary independent systems for Data warehouse Test and Development at the disaster site. Connections between primary DC and proxy site are over CWDM, and between primary DC and remote disaster site - over DWDM. All lines are doubled and independent on each other.

In this example the configuration and optimal disk space configuration for the primary DC storage has been examined. This system is the most important in whole Telco IT infrastructure and serves not only as storage for business critical applications but also it takes part in replications, backup and assures business continuity plan for the organization. The configurations of the other two systems are more or less the similar.

A. Fast group

The fast group needs 6,5 TB storage capacity and should handle at least 24000 IOPS at peak workload. There is planned 40% growth of the capacity and 10% performance increase. Due to the fact that there are big requirement for IOPS, two disk directors' pairs would be dedicated for this group, with 16 DAE connected to each, which means that maximum 240 disk drives will be dedicated to the fast group. All drives will be configured in RAID 10, striped meta-volumes with 8 hyper-volumes. Each hype-volume will be distributed on different slice from the directors.

According DMX system recommendation [3] eight drives are dedicated as global hot spares and 4 drives are for service and system information, so for customer data 224 drives will be used. The drive type for the fast group will be 146GB 15k rpm. This implies about 16535 GB usable capacity. To achieve maximum performance we need only about 40% to 70% of disk capacity, which will be used at the middle part of the plates (as shown on Fig. 1). Therefore the real usable capacity will be between 40-60% or from 6000 GB to 9800 GB. These values fully satisfy requirements for the fast group as well as the predicted capacity growth - up to 9100 GB within next three years.

According to (1) the theoretical maximum of IOPS for this drives is evaluated. We have 181,8 IOPS per drive or 40732 IOPS for the whole fast drives group. Now using the expression (3) we obtain the theoretical maximum of IOPS from host side – MW = 32578. This value takes into account the RAID protection and 3:1 relation between read and write operations.

It is very difficult to have such big workflow and achieve these IOPS in practice. Statistics from real storage arrays shows that practically for 146 GB 15k rpm drives there are about 130 IOPS, which means 33600 IOPS for the whole fast group or 26880 IOPS maximum workload from hosts' side.

These values show the maximal number of I/O operations from hosts which storage array could handle. This is only in worst case scenario at maximum workload and it is achieved in very rear situations. Even in this case it is obvious that the proposed configuration of fast group drives will satisfy the performance requirements of the applications and will assure about 2800 IOPS reserve over the desired level. It will comply with the predicted 10% growth in the next three years.

B. Slow group

The 40 TB storage capacities will be used for the slow group. These volumes would handle about 5000 IOPS maximum workload. There is the planned 40% capacity growth and 10% performance increase. In this group not only production volumes will be included but also all volumes needed for replication (local and remote). His group is not so performance critical, so only one director pair will handle all transactions. Having in mind that this group of drives has enormous size - 40TB we will have 16 DAE directly connected to directors' pair and 16 DAE connected into Daisy Chain to the first ones. This scheme will lead to performance degradation about 7-10% for all drives. For the slow group there are also 240 dedicated drives. Protection level for this drives will be RAID 5 (3+1) for Development and Test & Applications and unprotected RAID 0 volumes for data replication. These RAID levels are chosen in order to achieve best price, performance and capacity ratio for the solution. The RAID 5 volumes will be used due to specific application behavior, in this case we lose only 25% of disk capacity for protection, the performance is not a critical issue and volumes workload is not heavy. The RAID 0 groups will be used only for replicas (the source volumes are protected) so they do not need special protection and the performance will be quite satisfactory due to parallel operation over multiple drives.

Following best practice configuration rules [3] for the slow group there is dedicated 8 global hot spare drives and 2 drives for system and service needs. So for user data there are 230 drives available. Due to more effective service and maintenance requirements for the system we propose to use different drives for protected (RAID 5) and unprotected (RAID 0) volumes. Although it is not a default configuration of the storage the main benefit of this approach is that a failed drive in one of groups will not impact the other. For example if on the failed drive there are only one type of volumes - unprotected, it could be replaced immediately without waiting a rebuild process to finish, the information will be lost but it will be renewed immediately after drive change. Capacities of volumes used for replication (BCV) are presented in Table 3:

TABLE III. REPLICATION VOLUMES CAPACITY

Application	Size, TB	BCV	BCV size, TB	Max size within 3 years
Billing	3	Yes, 3 pairs	9	12,6
CRM	1,5	Yes, 3 pairs	4,5	6,3
ERP	2	Yes, 3 pairs	6	8,4
Development	10	Yes, 2T, 20%, 1 pair	2	2,8
Total , TB	16,5	-	21,5	30,1

For the slow group we propose low cost FC drives – 500 GB 7,2k rpm. For better balance of the array we keep the structure of the fast group – the number of drives for hyper and meta volumes should be dividable to 8, and should comply to size requirements, so for unprotected volumes (RAID 0) 64 drives will be needed. This fully answer the requirements for disk space and future growth of 40% within next three years and provide 30100 GB usable capacity.

Having in mind the disk specifications (Table II), we calculate the maximum IOPS per single drive 78,9 IOPS and for the whole unprotected volumes form slow group are 5069 IOPS. In this case we have no loses from protection type (RAID 0) but we have to take into account that back-end is connected in daisy chains, which leads to 7-10% performance degradation. Therefore the maximum theoretical workload from host side should be 4544 IOPS. These values are practically hard to be achieved, in most cases from statistic data the maximal workload which a single 500 GB FC drive handles, is 60 IOPS. This means that the whole unprotected group could receive 3840 IOPS and the maximum host workload should be 3456 IOPS. This numbers fully satisfy the requirements for local and remote replications because they work mainly in background and transparently for host, so after the base synchronization only changed blocks are written to drive tracks.

For the other part of the slow group, which will be RAID 5 (3+1) protected, 166 usable drives remain. The closest number of drives which will meet the capacity and performance need and is dividable to 8 is 160. These 160 drives provide 60 TB maximal usable capacity, which over meet needs for 40% data growth within next three year from currently needed 40 TB to predicted 56 TB. Performance demands for slow group applications show that for Development applications 3300 IOPS would be needed and up to 4950 IOPS during next three years, while Development, Test, Data warehouse application will need 2200 IOPS and up to 2347 IOPS during the next three years. Therefore the whole workload for slow group will be 4950+2347=7297 IOPS. These are theoretical calculations for IOPS from host side. In a real practical situation it couldn't be reached.

A theoretical calculation for 160 drives pointed out that maximum workload is 12624 IOPS. The real number from practical measurements and statistics is 9600 IOPS. Taking into account the specific backend losses due to daisy chain configuration the proposed solution will provide 8640 IOPS in worst case which will assure about 1200 IOPS reserve for these applications.

So finally for the whole slow group will be used 234 drives, from which 64 unprotected for BCV, 160 for production data (RAID 5) and 10 for hot spare and service needs. The summary of the proposed configuration for the primary storage array is presented in Table IV.

TABLE IV. PROPOSED SPACE DISTRIBUTION

Disk group	Drive type	Number of drives	Hot spares	System drives	RAID level	Usable drives
Fast	146 GB 15k rpm	236	8	4	1,0	224
Slow	500 GB 7,2k rpm	234	8	2	0	64
					5 (3+1)	160

In the same manner the capacity in proxy and remote data center have been configured. These configurations lead to significant performance improvement compared to proposed default methods. Using the standard method have to mix protected and unprotected volumes on same disk drive and no performance oriented group are constructed. In this case can obtain more usable capacity but the maximum IOPS which system can handle are about 30% less.

IV. CONCLUSION

The most important requirements for corporate storage systems in Telecom operators have been examined. A new solution for development of reliability, performance and effective use of capacity by choosing proper configuration and design has been proposed. Needs for local and remote replications and configuration of the drives in order to obtain maximum workload of the system have been taken into account.

The storage capacity has been divided into different groups which serve specific application needs. The whole available capacity is divided between the two groups called fast and slow. The result is up to 30% higher utilization IOPS than with standard default configuration recommended by manufacturer.

Proposed solution is based on real statistics data and configurations from Bulgarian telecommunication operator. Although the current research is done for EMC Symmetrix DMX3 storage arrays the same method of configuration could be applied for all other systems due to the fact that the specific Telco application requirements are more or less the same. The described solution is according the latest tendencies, utilize most of the advantages of the enterprise storage systems and follow the best practices and strategies for performance, data protection and reliability.

V. REFERENCES

[1] Raikovska, L., N.Dimchev, M.Saykov, "A modern high performance approach for data storage infrastructure design and data protection", Telecom conference 2004, Varna

[2] Raikovska, L., N.Dimchev, M.Saykov, „Information protection and data transfer methods for disaster recovery solutions", Telecom conference 2004, Varna

[3] EMC Symmetrix DMX-3 Product Guide, Hopkinton, MA, EMC Corporation, November 2006

[4] Economies of Capacity and Speed: Choosing the most cost-effective disc drive size and RPM to meet IT requirements, Seagate Technology Paper, May 2004