INTERNATIONAL
HELLENIC
UNIVERSITY

Development of application for selecting an ideal data migration solution in a heterogeneous storage environment

Maja Petrovska

SID: 3301100011

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in ICT Systems

SEPTEMBER 2011

THESSALONIKI – GREECE

INTERNATIONAL
HELLENIC
UNIVERSITY

Development of application for selecting an ideal data migration solution in a heterogeneous storage environment

Maja Petrovska

SID: 3301100011

Supervisor: Prof. I. Vlahavas

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in ICT Systems

SEPTEMBER 2011

THESSALONIKI – GREECE

DISCLAIMER

This dissertation is submitted in part candidacy for the degree of Master of Science in ICT Systems, from the School of Science and Technology of the International Hellenic University, Thessaloniki, Greece. The views expressed in the dissertation are those of the author entirely and no endorsement of these views is implied by the said University or its staff.

This work has not been submitted either in whole or in part, for any other degree at this or any other university.

Signed: .....

Name: .....

Date: .....

Abstract

This dissertation was written as a part of the MSc in ICT Systems at the International Hellenic University. The study was conducted in close collaboration with representative staff from SAGA D.O.O Company in the office located in Skopje.

The increasing criticality, business' growing dependency on digital information and explosive generation of data are leading to larger and more complex information storage environments that are increasingly challenging to manage. In these environments, whether performance improvement, technical refresh, server or storage consolidation, data center relocation have to be made, it is inevitable for businesses to deal with data migration. Migration, as moving data from one device to another and then redirecting all I/O to the new device, is an inherently disruptive process due to the different hardware and software environments that exist. Multiple storage vendors have unique requirements for migration and different migration techniques, adding complexity to the process itself. The requirement for information continuous availability, a widening storage technology knowledge gap across the industry and resources demands are making data migration even more intricate mission-critical task.

This dissertation studies the complexity of data migration and identifies all of the factors that may influence the selection of an ideal migration solution. By examining the concept of data migration, its methodology and strategies, a tool that addresses the issues for selecting ideal migration solution in a given heterogeneous storage environment is proposed. This application is specifically designed to be used as a starting point in any undertaken migration project, having in mind product and services from the top storage vendors currently in the market EMC, NetApp and IBM.

The main goal of this study is to enable readers to utilize this selection application coupled with the theory presented here in order to find the best solutions for their own migration challenges.

Student Name: Maja Petrovska
Date: 26.09.2011

Contents

Table of figures

1 Introduction

"The chaotic volume of information that continues growing relentlessly presents an endless amount of opportunity—driving transformational societal, technological, scientific, and economic changes,"

- Jeremy Burton, Chief Marketing Officer, EMC Corporation

Information has become an intrinsic part of our daily lives, living in an on-command, on-demand world. Like never before, it became increasingly important for businesses and individuals to have the right information in the right time.

In today's digital society businesses of all sizes depend on data to carry out daily business processes. Some of their operations include managing customer relationships, keeping business related records, performing e-commerce, billing systems and scores of other applications. Whether it is finance, production, sales and marketing, or any other functional area, it is critical for business survival to have continuous access to that data. Additionally, businesses, in order to derive economic benefits by extracting meaningful information from the data they have generated, need to maintain and to protect data and ensure its availability over a longer period of time.

As a result of the decrease in the cost of storage digital devices, significant increase in data processing and storage capabilities, and the affordable and faster communication technologies coupled with increasing individual and business needs, have led to accelerated data growth, popularly termed *data explosion* [1]. Since data have different purposes, both individuals and businesses have contributed to varied proportions in this data explosion. Despite the global recession and the slowdown in economic environment of recent years, data growth has still experienced an average 50% to 60% compound annual growth rate (CAGR) [2]. The new IDC Digital Universe Study sponsored by EMC estimates that data is growing faster than the Moore's Law to 1.8 zettabytes of data to be generated in 2011 which means that the enterprises have to manage 50 times more data and 75 times more files to grow in next decade [3]. This explosive growth means that by 2020 our Digital Universe will be 44 times as big as it was in 2009, nearly 800,000 petabytes, illustrated in Figure 1.1.

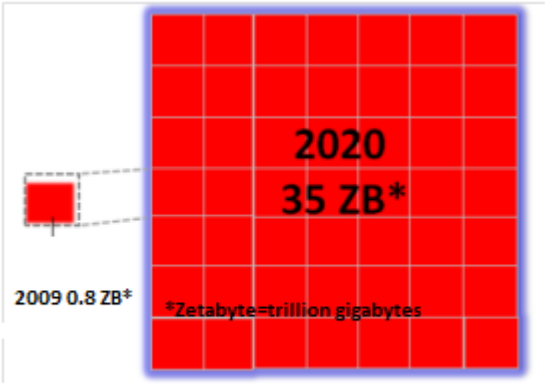

Figure 1.1 The Digital Universe 2009-2020 growing by factor 44 *(*Source: *IDC Digital Universe study, sponsored by EMC, May 2010)*

Nevertheless, companies have faced a challenging decision of replacing systems and working to blend current legacy systems with new technology posed by the explosion of data and increasing criticality of digitized information. Whether acquiring another business, upgrading a legacy system or outsourcing a business function, businesses inevitably need to cope with migrating data from one platform to another. Since it deals with moving business strategic data, data migration on business operations should be considered as a mission-critical task. Thus, this thesis is focused on the process of data migration, namely, in the examination of data migration technologies and strategies that can help organizations minimize the risk and complexity involved in the process of migrating data.

One of the most challenging responsibilities of any migration project is the selection of the best migration tool to successfully move the data. There are many issues to be discussed associated to choosing a potential solution. Special requirements come from the storage infrastructure itself, systems and application in use, and interoperability in each part of the migration environment. Lastly, continuous information availability must be ensured. In the past migration projects were scheduled to be performed during off-hours but today's 24x7 environment requires non-disruptive migration. This project will develop such a tool that will ease the task of selection migration tool and increase the time to its implementation.

1.1 Motivation

There are several reasons for working on developing a tool that will identify a potential migration solution for a given storage environment and application/server requirements. In close collaboration with SAGA D.O.O, one of the leading company in ICT solutions and system integrations in the Macedonian market and broader, we came up with the idea to design such a service that will help small to medium size businesses and storage professionals to determine the appropriate migration solution for a given storage environment without spending their time in doing an extensive research for one in the existing market. This tool can be used as a starting point in every migration project to avoid the complexity hidden behind choosing the right product.

In fact, the main idea grew up from the similar tool develop by EMC, namely, Data Migration Selection Tool which is a Microsoft Excel spreadsheet, that gives up to three results in preferred priority order based on the storage environments. However, two important limitations can be drawn down. Firstly, this tool is built up by EMC professionals and takes into consideration products and services developed by this company. A migration solution between non-EMC platforms cannot be proposed. Secondly, this tool is closed to the public use. It is dedicated to internal use, only within the company. There are many other tools developed for planning and assessment of data migration project in the market, but they are vendor lock in as well.

Our motivation is to design a service that will cover migration tools and platforms not just from one vendor and moreover will be open to any person concerned with migrations between different storage arrays. Personally, by working on development of such a tool will give me a rewarding opportunity to study in depth about data migration process and to achieve a broader understanding of current technologies in use. It is a big challenge for me to deal with the complexity involved in this process, but I strongly believe that it will open the doors for the future prospect in my carrier.

1.2 Report organization

This dissertation analyzes the concept of data migration and identifies the key drivers for choosing the data migration solutions that encompasses the tool proposed. The report is organized into five chapters:

Chapter 1

"Introduction" introduces to the reader the motivation for developing of this thesis. It describes the problem of data storage infrastructure management.

Chapter 2

"Literature review" gives an overview of the data market up till now. This chapter covers the theoretical background related to data migration. It defines and discusses the primary reasons for moving data, data migration formal methodology and strategies and identifies the information that must be collected to assay the storage environment. At the end, current trends, challenges and options in managing information storage are presented.

Chapter 3

"Data migration tool design" describes the main reason for development of data migration selection tool, it purposes, main objectives and delimitations.

Chapter 4

"Storage Systems specifications and associated software"- In this chapter an overview of the storage systems that are implemented in this data migration selection tool is provided. The associated software that can be used for data migration purposes is explained in details

Chapter 5

"Conclusions" describes the functionality of the developed tool together with consideration about the possible solutions that may be used for data migration in heterogeneous environments. Directions for future work and development are given at the end of this chapter.

All details covered in this dissertation, namely data migration technologies, strategies and the key factors and features in selecting data migration tools are up to date. For the future use, it is recommended to review if any changes have been made in the market.

2. Literature Review

1.2 Data migration definition

Although migration literally means movement from one place to another, data migration has a wide range of definitions and it can mean a number of things [4]. As it is explained in [5] and [6] the process of data migration can be defined by its characteristics:

- Data migration is a one-time movement of data. This is not to say that it may not take place over an extended time period, but means that once it has been completed, it stops. This in particular differentiates data migration from data mobility or any data integration tasks which continue on an on-going basis.
- Data migration involves the re-structuring of data in some way. This may mean fields being merged, formats being changed, or the data being transformed in various other ways.
- Data migration maintains consistency of usage. Here, usage is defined as being either operational or analytic. Data migration never takes data from one environment and puts it into the other without any further usage.
- Data migration may or may not involve contextual change. For example, if it happens due to ERP applications consolidation, then the same context is maintaining, or if there is an implementation of a CRM system for the first time, then that will mean contextual change for the source data. Data migrations involving contextual change typically involve more complexity than those that retain the same context.

In other words, data migration does not mean simply copying or replicating the data, where applications continue to access the source data after the target copy is created. In fact, it relocates the data but after the migration, applications that access that data must refer to its new location. Therefore, a part of the migration solution is also the methodology used to point applications to the new data location which is known as application cutover [7].

For the purpose of this study the following definition will be considered: ”*data migration is the process of making an exact copy of an organization’s current data from one device to another device —preferably without disrupting or disabling active applications —and then redirecting all input/output (I/O) activity to the new device”*, as it is given in [8]. It must be noted that data will be accessed only at the target after the migration has been performed. Re-structuring of data may or may not occur but it is out of our scope due to the introduction of further complexity and additional tools that have to be used.

It can be concluded that, the process of migrating data involves a lot more than ripping out one data storage cabinet and plugging in another. In order to properly move data to another storage system, it must be understood how data is being migrated and what steps need to be taken to avoid any potential risks or complications.

2.2 Reasons for moving data

The reasons behind a particular migration will primary drive the requirements of potential migration solutions. There are a variety of circumstances that might cause businesses to undertake a migration of data, which may involve multiple elements, either separately or in a conjunction. An extend list of reasons for moving data is created, that includes the reasons given in [9] and [10]:

- **Data growth.** When data grows beyond its current storage location, it must be expended, which often leads to a data migration need. This can be done by adding additional capacity to the current array or by using capacity in additional array.
- **Technology replacement or upgrade.** The aging of technology itself, which is commonly called technical refresh, may force data migration. As the system becomes older, manufacturer may do not support any products’ upgrade and software needed to scale the system to current requirements. The most common reason is the hardware going off lease which cannot be replaced or it is not cost effective to do it.

- **Platform migration.** Any change in hardware platform, especially, migration from one operating system to another, may also require a significant migration effort.
- **Performance improvements.** If an application needs performance improvement, the total window for completing an operation or the transaction time in the data storage I/O path have to be reduced. Application performance improvement can be achieved by one of two factors: either by spreading I/O more evenly across existing I/O resources, or by actually adding more resources or replacing existing with better performing resources. Spreading I/O more evenly across existing I/O resources can result in improving both transaction time and throughput.

  Storage performance can be improved by using logical disk volumes to give high levels of performance to set of applications and lower but still acceptable level to other applications. The concept of tired storage can be used or the transaction time in the data storage I/O path as well. For instance, the physical disk devices within array may have widely varying disk performance characteristics from the highest level flash memory to the lowest level large capacity ATA disk drive. Migrating existing data from one tier to another can dramatically change performance characteristics. Tiring of data storage can be obtained by having different performance characteristics of each storage array. In this case the migration will be performed from one storage array to another, rather than between disks within a single array.
- **Information Lifecycle Management.** The information lifecycle is defined as "change in the value of information" over time in [1]. In reality, when data is first created, it often has the highest value and is used more frequently. As data ages, it is accessed less frequently and it is of less value to the organization. This brings about the need to migrate data after its original prioritization or placement. Based on its changing value, information requires different levels of accessibility and protection. The goal of ILM is to move data continuously to the best cost resources.
- **Database migration.** Migrating from one database to another, for example, Oracle to DB2, or upgrading from one version of a database to another, will require a significant migration effort. Vendors will often offer migration assistance from one version to the next (and this may also apply in the case of application migration)

but, if fear of migration failure has deferred an upgrade decision, then suitable upgrade tools may not be available.

- **Migration from one version of an application to another**. For example, this may be a major upgrade from one version of an ERP or CRM application to another or, where the application is home grown; it may involve a re-write of the application.
- **Migration from one application to another.** For instance, replacing a SAP application with one from Oracle. It should be noted that the source data for the new application may derive from other locations in addition to the application being migrated from. This applies to other migration environments as well.
- **Application consolidation.** When multiple ERP, CRM or other applications are being consolidated into fewer instances of such an application. In some cases, the consolidated application may not be the same version, or even the same product as some or all of the original applications.
- **Risk reduction.** Lowering the probabilities of data unavailability and data loss conditions are considered as risk reduction. The process of data migration can reduce risk by migrations to the storage platforms where data replication strategies can be implemented and to more highly available storage platforms.
- **Consolidation, localization and regulatory compliance.** Changes in the business processes may drive data migrations. Storage consolidation will result from a merger, centralization, a data center migration or heterogeneous storage platform migration. Localization is needed since distributed operations may reverse data centralization. Regulatory compliance requirements may mandate fixed period of time accessibility and searchability on archived data.

As it is stated in [10] the best practice for migrations is to only change one thing at a time. However, the most migrations will include more than one reason for migration. It depends on the complexity of the main element of the migration.

2.3 Business issues

In the book "Practical Data Migration" the author John Morris has highlighted that the data migration must be treated as a business issue, not as an IT problem [11]. Truly, data migration projects are undertaken because they will support business objectives. It may be to reduce costs or to provide business users with new functionalities that will help to drive the business forward. Whatever the reason are, implementing a new solution, consolidating multiple databases or applications onto a single platform or technical refresh, data migrations should be focused on:

- realizing all the benefits promised by the new system,
- improving the enterprise performance that was the driver for the migration,
- being in compliance with all regulatory, legal and governance criteria,
- increasing the quality of data in order to enhance the business intelligence and
- providing high level of security and protection during the project.

From this, it can be seen that the business needs to be fully involved in the migration prior the start and on an on-going basis. Thus, whether the migration will be treated as a business project with support from IT or the other way around is still debatable, but it is clear that the close collaboration between the business and IT is the most important factor for the success of the migration project.

Notwithstanding, the data migration projects are not simple. They involve risks. If the migration goes wrong or if the project is delayed, the costs for the businesses might be considerable. Some of the costs associated with overrunning projects, from the results of the Bloor survey in 2011 [12], are shown in figure 2.1. All of these have direct impact on the reputation of the business itself.

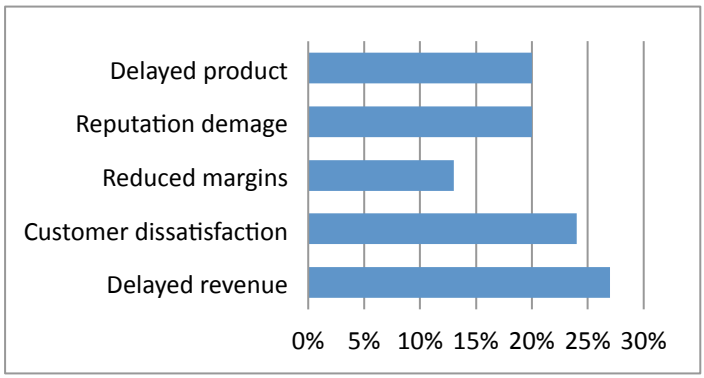

Figure 2.1 Costs to the business of overrunning project (Source: "*Managing Information Storage: Trends, Challenges and options"* [25])

2.4 The state of data migration market

"Data migration as a market worthy of consideration in its own right"

-Phill Howard

Results from the survey conducted by Bloor Research in 2007 regarding the state of the market for data migration showed that data migration, historically, has been under-valued, under-resourced and not treated with the attention it deserves [13]. At that time, data migration has been accomplished by using conventional data integration and data quality tools, or by means of hand coding. In fact, data migration had not been an area of focus for most vendors because only a few specialized tools for migration had been developed. Still, seven years later, after the research undertaken by the Standish Group in 1999, the success rate of the migration projects were the same [26]. Since 2007, more than 80% of data migration projects have run over time, over budget, or both. Briefly, at that period of time, the data migration market was immature in terms of technology.

Another important conclusion that can be drawn from this survey [13] is that traditional tools on their own are not enough. Vendors should be focused on data migration as a sector with its own right. Data migration methodologies, special-purpose solutions or at least add-ons to conventional tools need to be specifically designed to support data migration. In addition, the data migration should not be treated by the IT departments as a

dead end job with no prospects but still experinced and well trained professionals are required.

The second important finding, in this survey, is the size of the data migration market. As it can be seen from the figure 2.2, there is a huge annual spend on data migration. Since only the Global 2000 companies were targeted in this survey [13], where the whole project was budgeted at one million dollars or more, the total budget spent for migration must be higher. This is again a great deal of finance for something that was not considered as a market in its own right.

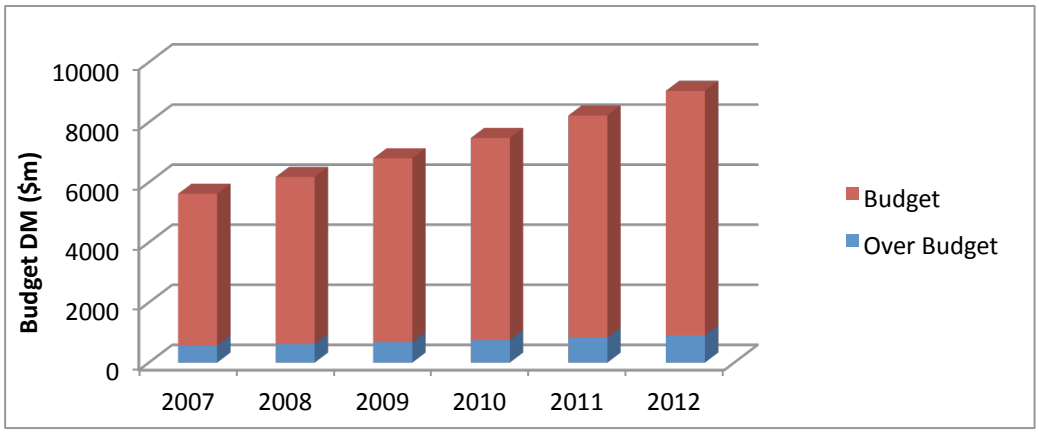

Figure 2.2 Global data migration budget and overruns 2007-20112 (Source: "*Managing Information Storage: Trends, Challenges and options"* [25])

However, this situation is changing. As previously noted, the data migration is a subject that has traditionally been ignored by both analysts and the market. However, there are a number of areas that have significantly changed since 2007 based on the Bloor research from 2011 [12]. The adoptions of appropriate tools together with a formal methodology have led over the last four years to a significant increase in the successful deployment of timely on-cost migration projects. With respect to the methodology, there has been a significant move towards formalised approaches that are supplied by vendors and systems integrators rather than developed in-house. As the results indicate, nowadays, nearly 62% of projects are approached in the proper manner.

The drivers to the success, the data migration methodology and strategies, the initial considerations that organisations need to bear in mind before selecting a migration solution

and best practices for addressing these issues are discussed in the text that follows. A critical factor in selecting relevant tool will be the degree to which it enables collaboration between relevant business objectives and technology.

2.5 Data migration methodology

As data migration was defined in 2.1, in this study will be focused on the process of moving data from one storage device, which is called source, to another referred as target storage. Whether the migration is performed in-house by the IT staff of a company, or the work is given to some external third- party provider, they might follow a formal, proven migration methodology. Not going into details of the different approaches such as the one given in [14] and [9], it is easy to see that they are undertaking the same general steps in any methodology, illustrated in figure 2.3.

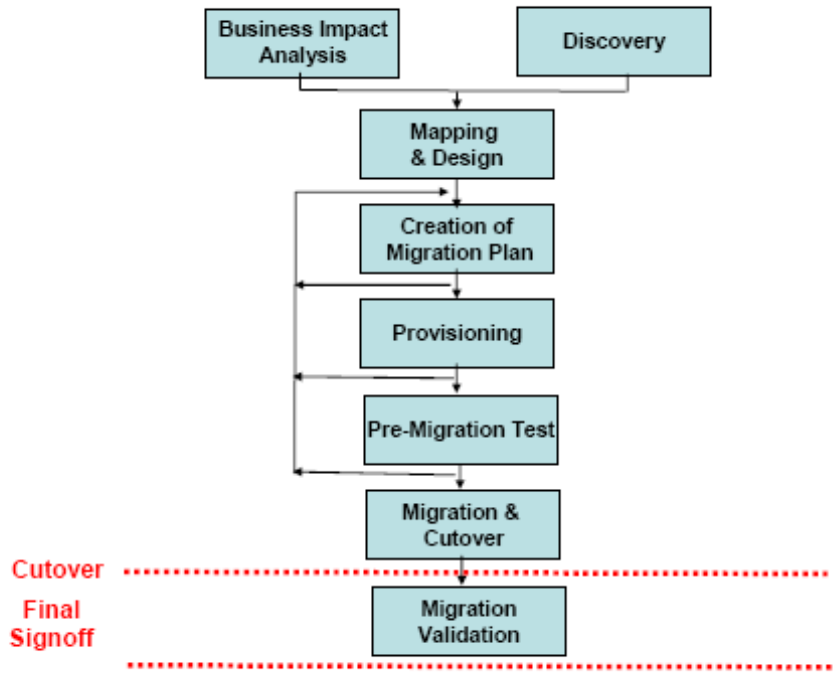

Figure 2.3 Migration methodology (Source: "*Data Migration Best Practices",* NetApp [14])

In this picture, the major steps of any data migration projects are shown. A lot of upfront planning must be made before the actual movement of data in order to ensure a successful migration. The first phase, as for any other project type, is planning which is the number-one success factor. Upfront planning can help to shorten the duration of migration process, to reduce business impact and its associated risk such as application downtime, performance degradation, technical incompatibilities, and data corruption or loss. Migration plan has to define what data type is going to be moved, where it will be moved, how it will be performed, when it is moved, and to estimate how long migration process will take. Moreover, in the process of planning it is important to be involved not just the IT staff but the business owners of the applications and data being migrated as well, which will allow to determine how important a given application or set of data is to the business.

In the text that follows, each step of data migration methodology is described in a concise manner, in order to better understand the formal methodology in use.

The objective of the business impact analysis is to identify the business and operational requirements that impact the migration process. More precisely, it determinates the reasons for moving data explained in details in 2.2.

Information about the migration hardware and software environment needs to be collected during the discovery stage at both the excising and the target. The decisive criteria for selecting the best data migration solution may be the result of the environmental factors. The physical infrastructure that has to be inventory includes hosts, storage networks and storage platforms. The applications and software functionalities available for each platform or device, together with the existing release versions and licenses must be considered. Moreover, capacity levels have to be determined because they might limit the migration process.

Collection of environment information also occurs again in the design phase where much more detailed data is needed. In this more detailed phase interoperability concerns at every level of the infrastructure must be accounted for. An initial check for interoperability is necessary to ensure that a chosen solution is indeed viable. Mapping, on the other side, directly depends on the storage layout. There are two basic mapping layouts. If the source and destination have the same storage layout than one-to-one mapping can be done, otherwise, relayout must be performed. Although a one-to-one mapping enables a much

simpler migration, migration is often seen as an optimal opportunity to consolidate and optimize performance or capacity utilization, so relayout is a very common scenario. A combination of the migration goals and best practices drives the layout of the new storage environment.

During provisioning, the destination storage environment is prepared for the data move. LUNs, volumes, directories, and so on are allocated, security attributes are set, and shares/exports are created. Provisioning for a one-to-one mapping is simple butfor a relayout, it is a more complex task. However, by using information generated from the mapping and design step, it is possible to automate many of the provisioning tasks.

Pre-migration test or migration pilot is important to be made, before any data is moved to the new environment. This is an opportunity to determine any flaws in the plan that were not previously discovered and whether the migration plan has to be modified. For example, if the testing shows that the specified downtime would probably be exceeded, and then the migration tools used, amount of data migrated per session and timeline have to be revised.

Migration and cutover are the points when actually the source data is being moved to the destination devices. There are many options for moving data such as whether the migration solutions will run on the host, storage device or network device. Also, different cutover strategies exist that are explained in the next section 2.5. Lastly, the data can be moved in two different ways: out-of band or inband by means of which path is used to transfer the data.

The last step of data migration project is to validate the new working environment and confirm that all expectations have been met. It is important to validate that the data and functionality of the application after the migration are the same. At last, network access, file permissions, directory structure, and database/applications need to be validated, which is often done via non-production testing.

2.6 Strategies for data migration

Nowadays, data migration tasks are routine, which are performed even on a daily or weekly basis, but still can be considered as risky events. As it is stated in [15], less than

20% of migrations succeed on their objectives. In the recent years, a myriad number of technologies for data migration have been developed, and the strategies that might be used for data migration implementation and similar projects have emerged as well. Here, some of the different data migration go-live strategies and what they mean from a technical and business perspective.

In truth, businesses are most afraid of the switch from old system to new one. The business cutover is actually defined as the way of changing from one environment into another. There is a little room for error during the migration. For example, if the new system comes online and service is impacted then this bad news can spread rapidly and customer confidence can be seriously impacted.

Essentially, there are three possible approaches for the business cutover strategy [5], [15] and [16]: big bang, parallel running and incremental which are described below.

2.6.1 The big bang

The concept of a big bang data migration is a simple one. First, a migration architecture should be built, and then, the physically movement of data is performed in one process. Using the big bang strategy, throughout the duration of the migration project only the original system is running. There is no need to run two systems simultaneously at any stage. At more or less the last moment, a sufficient time window is needed to successfully move the data from the source storage to the target and this typically involves downtime to the business. This process takes place at e pre-specified time and date. When all of the data is load onto the new system, the old system should be turned off and then the new environment starts running. This method relates to the cosmological theory where the start of the cosmos happened at one moment in time. In this case, the new fully tested storage environment turn out to be functional.

However, the big bang adoption has many advantages and disadvantages. The big advantages of this approach are, first, that it eliminates the problems of having dependencies between systems. When using non-big bang approaches some of the old systems have to run side-by-side with the new ones. Second, since the only time when two systems are running is during the changeover period, this approach minimises the hardware

and systems requirement. This, at least, is the theory, though it is often not realised in practice [5]. Third, with a big bang migration there is no need to keep the old environment up to speed with any record updates, since the business has effectively moved on and is now being driven entirely from the target platform. In comparison to the incremental data migration strategy, that is described afterwards, because the big bang strategy cuts over all parts of the system at the same time the errors that may occur during the phasing in of the system can be eliminated.

One of the drawbacks of this approach is the duration of the changeover process. Typically, it takes for several days, hours or a weekend to transfer data. That means that the system must be shut down. But, today's business run in a continuous manner, so, it is of a high cost to take part of them down at all. The first point to make is that this approach obviously is not suitable for web-based applications. For example, it has to be running 24 hours a day, 7 days a week. For any on-line services the cost to the business is likely to overwhelm any savings to be made by taking this strategy. Also, using such an approach for mission critical applications that do not have to be available on a non-stop basis it is not recommended.

The second downside comes from having a limited-time frame to perform the migration of data which can have a major impact if the migration overruns. Although, the synchronisation may not be an issue but fallback strategies can be challenging particularly if issues are found some time after the migration event.

There is a risk associated with the new environment, whether the new system does work well. If it does not, then either the duration of the process have to be extended (with all the costs to the business that that implies) or the business should go back to the original system and try against next time and so on. In both cases, serious damage will be made to the reputation of the IT department with the user community and to the business itself, such as some product launches may have to be delayed. Moreover, the company will start to lose the cost advantages of not having to run two systems at the same time.

In conclusion, the big bang migrations should really be reserved for non-mission critical, non-24x7 applications, where the size of the data is manageable. It is, at least on the face of it, less expensive than an incremental approach, because it requires having parallel systems over a prolonged period of time whereas big bang migrations only need

duplicated systems for a brief period. However, the big bang approach is still the most used and trusted by the companies, perhaps because the more expertise of this kind are floating around [15].

2.6.2 The parallel running

Another very popular approach for business cutover strategy is the parallel running because it can greatly reduce the risk of the successfulness of data migration. The key point of the parallel running strategy is that the data in both systems must be synchronized.

Parallel running starts the same way as the big bang approach, but the difference is that the old system is not cut over once the data is loaded into the new environment. Instead, both systems run in parallel, with the original system continuing to serve the business while the new system is thoroughly tested until everything is ready for a final cutover.

The benefits of this approach are that the both systema can work in parallel for a period of time so the business can fully validate and sign-off the platform and safe in the knowledge that it completely meets their needs. That means the new system is more or less guaranteed to work correctly when finally move to it. Another advantage is that the downtime, between turning off the old system and turning on the new one, it is expected to be shorter compared to big bang approach.

The one downside is the cost of the implementation. Running two systems in parallel is clearly more expensive, not just in computing resources but also in terms of maintenance. More people are required to run it, and the monitoring that needs to go on to ensure that the new system is doing what it is supposed to.

When it is compared to the big bang approach, as an environment, parallel running is also more complex. Incoming data is entered into the old system and this data need to be transferred to the new system, but ensure the data in each system is current. This means that some sort of synchronisation software to keep the two systems in parallel should be involved.

A parallel running migration does not necessarily require data to be moved in one big bang, it can also support incremental migrations.

2.6.3 The incremental cutover

Incremental migration works in a similar fashion to parallel running. During the migration process the two systems are running in parallel, but the difference is in that the data from the old system to the new one is gradually moved rather than turning it on all at once. In addition, this approach makes possible to move only some discrete parts of the business data and thus reduce the risk of the project. By the time that everything has been moved, it will be more or less time to turn off the old system. In some cases, it will be very little, if any full-scale parallel running of the systems.

Historically, incremental migration was performed by on a table-by-table basis by gradually moving, for example, all the customers, and then all the orders, and the order details and so on. As it is stated in [16], it is better to work at the level of business entities. That is, it is better to move a customer with their orders and relevant order details altogether, thus maintaining the relationships, rather than moving on a table by table basis.

As it is described above, when using incremental processing only relatively small datasets are moved at a time. This can be considered as an advantage in the case where some actions do not work and has to be backed out. So, it is far easier to roll back a small portion of data than an entire store when using the big bang or parallel processing strategies. Also, this makes testing easier because it, too, can be done incrementally on small sizes of data sets.

Right through the process of migration, transactions for some customers can be processed by the old system and transactions for others can be processed by the new system. Similarly, some functions might be executed by the old system and some functions by the new one, though this will be as a result of the new application rather than data migration per se [5]. For this reason, businesses can start to get a return on their investment much sooner, which can be considered as one of the advantages of incremental migration.

Notwithstanding, when the data is not fully migrated from one system to another, there is an issue that an application need to know where the data resides, and from where to get the data that is required. The traditional method of supporting incremental migration is to place a flag against the relevant data fields in the source database, which tells the application to look for that data in the new system. Alternatively, a state model that keeps record of the current state of the data, such as, in which system the data is live and where it

is located, can be implemented. This method is to be preferred over the flag-based method because there is no need to amend any data table. Furthermore, the use of a state model can assist in the process of backing out an increment because it maintains details about the state of that data.

Another consideration, when it comes to the incremental migration, is that the systems where the data details are moved on may be different with different database schemas and different metadata. So, in order to enable this changeover process some amends has to be made on data sources. This may lead to degradation of performance when the data is in the new system. Also, the relationships between data in the separate systems can get broken.

In particular, if both systems are to be active simultaneously then bi-directional synchronisation it is required and special facilities to detect collisions and ensure consistency.

2.6.4 Zero-downtime migration

Finally, zero-downtime migration strategy, as the name implies in theory, refers to migrations with no downtime. The main idea is to enable migrations with continuous availability so the application is never taken off-line. The big bang approach is opposed to this strategy regarding the down time.

The zero-downtime approach is more or less implicit within the incremental and parallel strategies. The data is gradually moved from the old system to the new one in a seamless fashion and without ever taking the system down. In most cases, the two systems are running in parallel for some period of time after the migration is complete. During the parallel running stage it may also be helpful to implement failback capabilities that automatically revert to the old system in case of a failure of the new one. Also, some tools to measure the data quality and ensure the data consistency in the both systems should be used.

Zero-downtime is in contrast to the traditional big bang approach that has been most widely used on an historic basis, whereby there is a cut over to the new system, typically over a weekend. An excellent example of exactly what can go wrong with such an approach was by the opening of Terminal 5 at Heathrow huge direct (£16m and counting) and

indirect (loss of prestige, bad press, share price falls and so on) costs [965]. At the Heathrow Airport Terminal 5 in the UK on that day all flights were running just fine, but baggage handling was not. So it is not just a question of whether an application is working but whether it is offering all the functionality that it is expected to provide.

The decision about whether to adopt such a strategy will depend on the importance of the application and data that is migrating. If this supports a web-based application or some other mission critical application that needs to be up 24x7 then a zero downtime approach should be mandated. Zero-downtime migrations should also be considered whenever there is a failure to go live accurately and on time which may be detrimental to staff morale.

2.7 The key decision factors for data migration

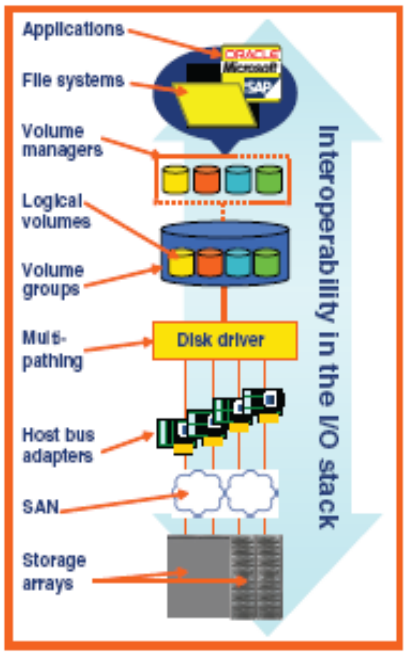

Figure 2.4 Interoperability in the I/O stack

(Source: *Choosing a Data Migration Solution for EMC Symmetrix Arrays,* EMC [9])

In order to determine the best technology to use for a particular environment, a list of all the components is recommended to be made, which is going to be involved in the migration process. Those may directly affect the success of the outcome. Of course, not all of them may have to be applicable to the environment under consideration. These factors

are related to the application performance requirements, hosts (servers), storage networks, storage array information and software, which are described in details in this section. In figure 2.4, all of the levels in which information may need to be collected are given.

2.7.1 Applications information

Since migration relocates the original data of some application, first, it is necessary to determine exactly where the original data resides. The data file can reside directly on a raw host physical device, or can be layered first on either a host file system, a host Logical Volume Manager (LVM), or both as it shown in figure 2.5. This mapping view can be generated by using different file level analytical solutions or reports according to the systems in used.

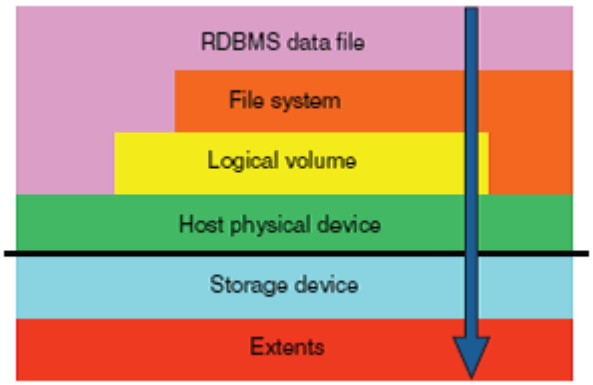

Figure 2.5 Storage management levels

(Source: *Choosing a Data Migration Solution for EMC Symmetrix Arrays,* EMC [9])

After the migration, application need to know from where to get the data that is required in the new systems. So, different mapping mechanism can be implemented to point to the new location. There is a one-to-one relationship between the hoist physical device and storage device, with exception from Directed Attached Storage (DAS) where the storage device is a logical construct made up of extents. Any of the logical representations in this mapping could potentially be used to make the migration transparent to the application. That mapping can occur at one or more levels including the File system or LVM of the host, or both [9].

2.7.2 Hosts/servers

Host is the physical computer on which applications run, and serves their needs or requests. Some of the host software components that have to be taken into consideration were mentioned above including: Database Management Systems (DBMS), File Systems, Logical Volume Managers, and Multipath I/O (MPIO) Drivers. Other host information that must be collected includes: I/O bandwidth, CPU capacity, Host Bus Adaptor (HBA), and Operating System (OS) versions. The version of HBA and OS is critical to insure compatibility with the target environment and also the data migration solution itself.

If the data migration is in band, meaning that uses the host I/O path, then the I/O bandwidth capacity must be sufficient to support this migration. Not just the total capacity, but also for each port the utilizations level has to be estimated. A multipathing disk driver may automatically balance I/Os across multiple paths to achieve equal utilization and best performance.

The CPU capacity must be adequate when conducting a host-based migration. The ability to limit the data migration tool to only use both the I/O and CPU capacity available is very important to avoid too great an impact on application performance [9].

In some cases, there may be the ability to increase the bandwidth for data migration between a source and a target location by connecting additional HBAs ports. Certain data migration solutions may require an additional host or repurpose a host to act as a dedicated data migration application appliance. However, the investment in the hardware may not fit the business factors around the migration.

2.7.3 Storage network

When it comes to the storage network information about the network type and topology, network bridging technology, I/O bandwidth, switch port capacity, and switch software capability are need to be investigated.

The storage network environment may be diverse consisting of DAS, NAS, SAN or the mix of them communicating through different technologies: Fibre Channel (FC), IP, FICON and ESCON.

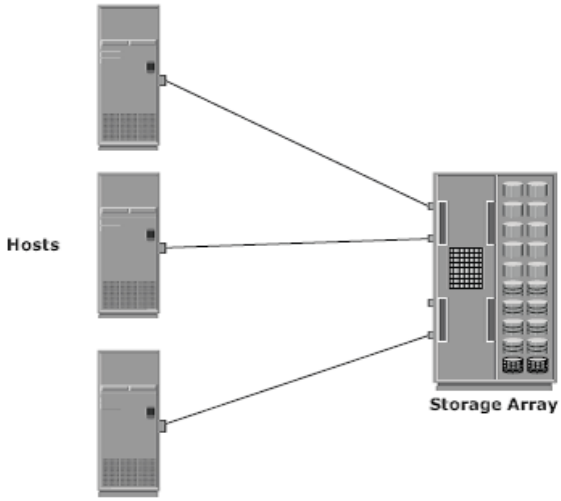

Figure 2.6 D*irect-Attached Storage* (*source:"Information and Storage Management"* [1])

D*irect-Attached Storage (DAS)* is an architecture where storage connects directly to servers, in figure 2.6. Because the storage is dedicated to the hosts it is difficult to manage and share resources on these isolated storage devices. Many efforts to overcome this problem of dispersed data have led to the emergence of the storage area network (SAN). SAN is a dedicated network of servers and shared storage devices, traditionally connected over Fibre Channel (FC) networks and recently over IP, in figure 2.7.

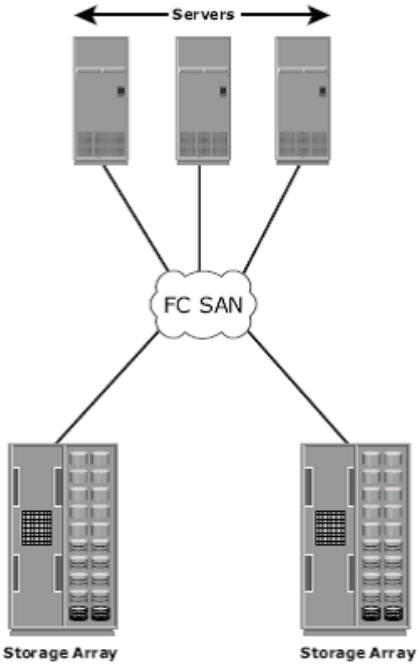

Figure 2.7 Storage Area Network (*source:"Information and Storage Management"* [1])

As well as SAN, *Network-attached storage (NAS)* enables data centralization and consolidation but it eliminates the need for multiple file servers. NAS is IP-based file-sharing device attached to a local area network. To perform filling and storage functions, NAS typically uses NFS for UNIX, CIFS for Windows, and TCP/IP, File Transfer Protocol (FTP) and other protocols for both environments, figure 2.8. Recent advancements in networking technology have enabled NAS to scale up to enterprise requirements for improved performance and reliability in accessing data [1].

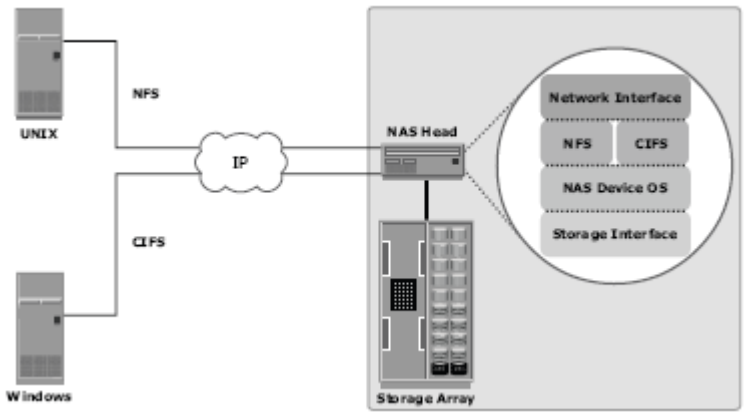

Figure 2.8 Network-attached storage (*source:"Information and Storage Management"* [1])

Although the implementation of storage networking technologies are gaining popularity, DAS has remained ideal for localized data access and sharing in small businesses, while medium-size companies use DAS for file serving and e‑mail, while larger enterprises leverage DAS inconjunction with SAN and NAS.

As it was case with servers, the storage network bandwidth and utilization and also the storage network port capacity are important factors. The latter, relates to the number of hosts and servers that can be connected to the switches or hubs building up the network. The needs are to have enough available ports to connect the source and target storage arrays simultaneously as well as for ports for the migration itself. Depending on the type of migration solution, it may not be necessary to have full connectivity to both the source and target at all times while implementing the data migration solution, but in other cases additional capacity may be needed only for the time of the migration itself. Given the data

growth patterns, in the near future, it is likely to expect that any extra capacity for the migration will be required.

In recent times, the storage technology has introduced new migration solutions that run in the storage network itself but require specific hardware switches to support these software applications. In addition, sufficient capacities must be considered to support the migration process. Storage virtualization, that is becoming very popular technique these days, introduces a logical storage entity within the storage network. This virtual entity enables seamless movement of data between the source and target, completely transparent to the application. Information about the number, size and composition of virtual devices that can support both the migration and target environment to be obtained is necessary.

2.7.4 Storage array information

Consideration related to the storage platform infrastructure covers data capacity and utilization, I/O bandwidth, director port capacity, tiers of storage, homogeneous or heterogeneous storage elements and the number and versions of each storage element.

Storage data capacity does not only mean how much data is stored but the amount of data that can be stored for the future business needs. The goal is to maintain all of the data at a satisfactory performance levels. When measuring capacity, utilization and allocation, are two important things that might be thought of. Storage that is unallocated is free for new uses; however, storage that is allocated but not used is limited to be used within its existing allocation until it is reallocated. Some of the migration solutions may require temporary capacity to store the intermediate copy of data they make during the migration process. This temporary capacity may be located either on the existing environment or on the target side, but it has to be considered prior to the migration. Also, the possibility to expand the storage arrays by upgrading/ replacing to higher capacity drives or by adding new drives must be reviewed. It has to be noted that the measurement of the storage capacity have to be taken in each tier of the storage so the performance objectives to be met.

One of the drivers for data migration may be also the constraints of I/O bandwidth capacity. Like data capacity, I/O bandwidth needs to be measured not just at the total bandwidth level, but also for each port evaluating utilization levels.

A number of hosts that can be directly connected to a storage array port are important for some migration solutions that use the I/O data path. There must be enough excess director port configuration capacity to make this migration possible.

When considering the storage array, at last, the storage element homogeneity is crucial. The systems may be from one vendor, homogeneous, or provided by multiple vendors, heterogeneous Also, it may involve a difference of vendors in storage arrays, or even different types of storage arrays within single vendors, or different elements within each storage array. Looking for a migration solution it is important to support the full diversity of storage elements in the current or in the target environments. The current version of each storage element adds more complexity to the data migration. Some elements may not support or may not be upgraded to the level required for a particular migration solution. Depending on the element, there
may not be a nondisruptive upgrade methodology, requiring the upgrade to occur during a scheduled maintenance windowand thus to extend the time it takes to complete the entire migration.

2.7.5 Software attributes

Besides the critical factors of the hardware infrastructure, the software current licenses and the ability for any upgrades will define the options available for potential data migration solutions. Some software versions may enable number of simultaneous migrations. Support for additional software products in the existing environment or in the target environment or even temporary support in the migration environment can expand the number of potential data migration solutions. Business factors will likely limit the choice of a data migration solution, but it is first necessary to fully understand the existing and target environments in order to identify all the potential data migration solutions.

2.8 Selecting the ideal data migration solution

As discussed earlier, because of the need to transform data from a wide variety of legacy data sources into a new environment that may support different schema and data types, and all of the environmental factors that have to be investigated and variety of

available data migration solution choosing the right solution that can address all of the requirements is very difficult and complex task.

Traditionally, businesses have an opportunity to choose between developing custom code or relying on a vendor specific data migration tool. In-house developed migration tool may offer more flexibility, but on the other hand, there are many hidden costs that organizations cannot be realized upfront. It requires additional IT staff responsible for creation and management of the code, time to be developed and tested. Another weakness of the custom code is that it cannot easily adapt to changes and in most cases it cannot be reused. Alternatively, vendor's migration solutions are fully tested and can provide some cots benefits, they are less flexible and may not be compatible with all vendors' products and technologies involve in the migration process. However, the current trends, as it was noticed in 2.1, are towards using vendor specific technologies because of the features they provide. If the proposed solution includes a technology that was not previously used in production environment, best practice is to validate this approach first following standard procedure in a non-production environment [17].

Organizations need to select the data migration solution that meets the full scope of the project. It is important to bear in mind the budget allocated and the costs for the full project lifecycle (upgrades, maintenance and training). Also, the type of the migration that is most suitable such as real-time, parallel, big-bang or incremental.

In the book "Choosing data migration Solution for EMC Symmetrixs Array" a full model for selecting the ideal data migration solution for a particular environment is proposed. This model consists of six well defined steps that have to be considered when making decisions:

1. Define the reasons for the data migration, clearly noting mandatory and optional objectives.

2. Inventory the existing environment identifying storage elements that must participate with the chosen data migration solution, and resources available to support the migration itself.

3. Inventory the target environment identifying storage elements that must participate with the chosen data migration solution, and resources available to support the migration itself. This step should include scalability considerations to ensure the target environment is

not obsolete too soon. Additionally, the solution might include adding in semi-permanent infrastructure to support future data migrations.

4. Identify potential data migration solutions that can successfully move the data from the existing environment to the target environment.

5. Identify business factors that limit potential data migration solutions due to budget, human resources, and application outage and verification requirements.

6. Compare and evaluate the potential data migration solutions including the criteria identified in steps 1–5.

The first three steps of this model were described in details in previous sections. Here, we are focused on identifying the right migration solution to successfully move the data. Some of the necessary features to be considered in migration software beyond that the software chosen should support the operating system and source hardware platform on which the data resides, as well as support the target hardware.

One of these key attributes is performance which relates to how quickly the data can be copied from the source to the target environment. However, performance must be balanced against network bandwidth and system overhead. When the data is copied at a high speed it consumes too much bandwidth or I/O, then production applications or systems can be severely affected. On the other hand, unless data is copied too slowly, the migration may take longer than anticipated, potentially prolonging downtime. Some migration software products, include a throttling or pacing capability that minimizes impact on production applications, thus enabling faster data movement when systems allow, and slowing down movement when I/O is required for other purposes. This capability helps IT organizations to more easily balance migration versus other system demands [18].

Another important factor is the ability to roll back the migration if something goes wrong. The roll back allows the migration to be terminated and restarted or application processing can continue to run on the source data device. This can be problematic with some technologies such as volume managers.

When selecting data migration solution it is important to take a look of the volume sizes. About 40% of data migrations are due to increase of the volume capacity from the source to the target. Not all of the available technologies can provide this capability. However, the migration target must be equal to or greater than the source volume capacity

A very common problem that occurs in the multi-vendor environment is that the migration solution may not be hardware independent. While host-based products support unlike storage devices, most array-based products require that the source and the target come from the same vendor, and may require that they be the same type or generation, and/or the same firmware version on the storage device.

As a fact that today, both customers and internal users want access to relevant applications at any time of the day or night on any day of the year, companies must ensure continuous availability rather than the merely high availability of applications that has previously been the norm. Application downtime is very critical factor for some business operations. Depending on the type of data and applications being migrated, only a narrow downtime window may be available in order of minutes, or in some cases it may be unacceptable. Nowadays, companies are more interested in software that enables nondisruptive migration, meaning that application can stay online and continue to process data and transaction throughout the migration process. However, online migrations introduce additional risk that must be compared to the impact of application availability. A JPMorgan Chase best practices suggests that whenever is possible offline migration to be done [17]. Also, it is strongly recommend to be considered selecting the technique that is the best compromise between efficiency and the least impact to system users [19].

Overall, all data migration technologies can be classified in three general groups depending on the utilities they used during the migration host-based, array-based and network- based. There are many benefits and drawbacks to running migrations from any of these platforms. Because of their impact on the migration process they are described separately in the next section.

2.8.1 Data migration approaches

Data migration solutions can run on the host, the array or in the network on appliances or intelligent switches. Before choosing any of this approaches, it is important to first understand the context in which the data relocation will be done. The decision process should be also based on the strategy of the data migration and all the constraints that may exist. The advantages and disadvantages coupled with these three approaches are

summarized in this part based on the information given in the articles [20], [21], [22] and [23].

Host-based migrations rely on functions of the host operating system (OS) that includes native tools to perform this type of migration. For most UNIX and Linux migrations, Logical Volume Manager (LVM) is used, while for Windows environments Logical Disk Manager (LDM) or any vendor add-on software or other native volume management tools. Volume management software provides specific tools that directly control and manage disks and storage devices attached to the system. Environments running in a basic disk configuration must rely on copy tools to perform the data relocation [17].

Host-based data migration requires root server level access and thus is most commonly perform by the server administrator. The benefits to this technique are that does not require additional licenses or technologies to be installed, and the current expertise skills of the server administrator are enough to perform this task. It is storage agnostic, which means that can work with different vendor's storage arrays and databases. Host-based orientated data migration is inexpensive when used with OS utilities, but the downside is that these utilities should be used only when the application or files are offline. For instance, because Windows basic disk configurations require advanced copy utility to move data they must be performed offline and, in such situation it uses network resources. Another problem that may occur, related to the UNIX platform is that not all native tools support logical volumes configurations. Most of the third party solutions maintain data availability so it might be taken into consideration. Host-based technology is Very attractive for a small number of hosts, but it is unfavorable for a large number of servers and multiple OS infrastructures because of its complexity.

Array-based or disk-based approach use storage system resources to physically move the data. The advantage is that the work is offloaded from the server resources and these resources can be used for other purposes during the online migration. It is suitable for small or large numbers of servers and multi-OS environments, but it is limited to homogeneous storage infrastructure or requires storage virtualization to be able to reach out to other array. Most storage vendors do not even support replication between their own array families. With array-based replication, only one central data storage unit, SAN or

NAS, is needed to control, process and validate the data being migrated. Throughout the migration, it is not required hosts to be attached at the second site or to the second SAN/NAS. The disadvantage of this migration is that requires a second SAN or NAS unit to be used which increases the cost of the solution. There could be compatibility problems of replication technology/software between SAN/NAS hardware and vendors [24].

The last approach is the *network-based* migration which is most suitable in heterogeneous storage environments because it works with anyone's array and supports any host platform. However, it requires one additional storage component to be managed. This device is situated in the network, between hosts and arrays, and the splitting of I/Os is performed in either an inline appliance or in a Fibre Channel (FC) fabric. The I/O splitter looks at the destination address of an incoming write I/O and, if it is part of a replication volume, forwards a copy of the I/O to the replication target.

The network-based data migration has both positive and negative aspects. In many ways, it combines the benefits of array-based and host-based approaches. Since, all the work is done in the network, the servers and storage arrays are offloaded, and so, it can support a large number of server platforms and storage arrays. Most network-based replication products also offer storage virtualization as an option or as part of the core product.

They are basically two offerings: inline appliances or fabric based products. Some issues related to the inline appliances are their performance and scalability because all I/Os need to pass through the device. The appliances terminate all incoming I/Os and initiate new I/Os that are forwarded to the primary and, in case of write I/Os, they are also forwarded to replicated storage targets. While in fabric-based implementations, the splitting and forwarding of I/Os is performed within a Fibre Channel fabric. By taking advantage of FC switching and separating the data and control path, it becomes the best performing and most scalable approach [23].

2.9 Trends, challenges and options in storage industry for 2011-2012

In this part, the direction in which the fast-growing storage industry is moving is reviewed. The current trends, challenges and options in managing the information storage

are highlighted based on the results from a global survey of more than 1000 IT professionals conducted by EMC for the period of 2011-2012 [25]. There are three major findings to be discussed regardless explosive growth of data, transformation of the storage infrastructure and a widening knowledge and skills gap in the field of storage management. This information might be very useful for the IT managers in the process of planning and decision making.

The information storage infrastructure has become the most critical component of an overall IT infrastructure, from the perspective of data availability and protection. As it was mentioned, in the post-dot-com era business from all sizes depend on digital information to perform day to day operations. The increasing importance of data together with the extraordinary growth, it is leading to larger and more complex information storage environments that are becoming more challenging to manage. With the increasing complexity and criticality of storage, highly skilled and focused storage groups are as mission-critical as the technology being deployed.

Although, the industry segments vary in their data storage size and activities they do, it was uncovered strong consistency across companies in terms of the technology deployed, storage management practices, and challenges. For safety reasons and in compliance with regulations, almost all critical data that organizations poses and mange is now stored on external storage drivers or different storage subsystems. The average usable capacity, that was reported, is approximately 1.3 PB (up 35 percent year over year) which is typically spread across multiple sites. Due to the growth in storage requirements, larger capacity disks and subsystems, and affordable pricing have all tend to deploy large storage configurations with increase complexity.

Up till now, it was found that the SAN storage networks and backup/recovery technologies are most commonly implemented followed by NAS, DAS, and replication technologies. In spite of that, it is indicated in the companies a strong move towards using technologies such as storage virtualization and cloud (private and public). Currently about 53 percent of storage capacities are in traditional/classic IT environments. In the next 24 months, it is expected that about 30 percent of data in classic environments is to be moved either to a virtualized or cloud environment. The higest growth of 69%, is expected to be in internal/private cloud as it is shown on the figure 2.9 below.

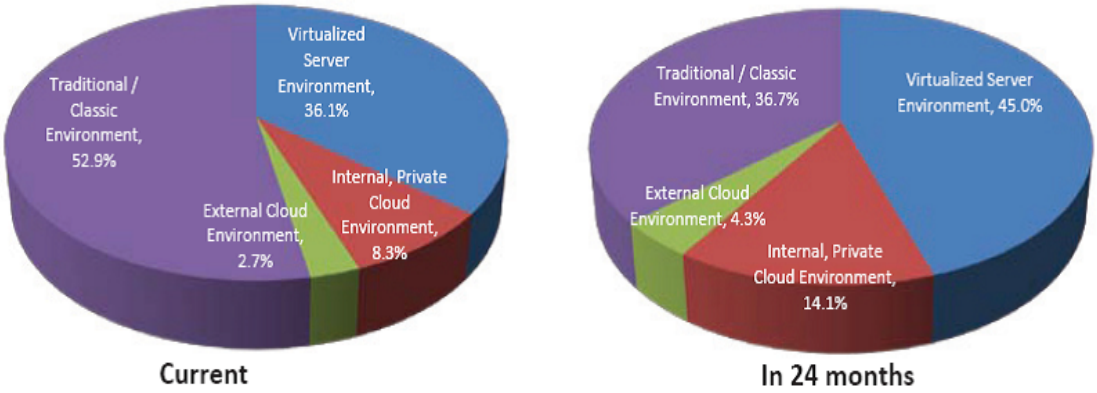

Figure 2.9 Data movement to cloud and virtualized environments in 24 months (Source: "*Managing Information Storage: Trends, Challenges and options"* [25])

In reality, these different storage technology segments are unique by themselves. They offer specific business functionalities and add values to the operational services. In order to provide the business objective requirements, storage infrastructures need different set of skills to be effectively design and manage. The key challenges indentified by this study for the storage managers and professionals in small and in large businesses are:

- Managing storage growth
- Designing, deploying, and managing storage in a virtualized server environment
- Designing, deploying, and managing backup, recovery and archive solutions
- Storage consolidation
- Making strategic decisions
- Designing, deploying, and managing backup, disaster recovery solutions
- Lack of skill storage professionals
- Designing, deploying, and managing storage in a cloud computing environment

It can be concluded that the activities such as backup and recovery, storage consolidation which have been in practice for decades are still concerned by the

professionals that they are not doing enough or not performing well. Primary reasons for not executing many of these activities to the desired level besides the explosion of data and storage requirements and emerging technologies such as virtualization and cloud is the lack of knowledge and expertise in a specific segment. The shortage of broad and deep knowledge directly impacts the ability to make informed strategic decisions and to proactively plan, design, and manage storage infrastructure. This skill gap in the industry continues to widen as organizations adopt virtualization and cloud computing.

In conclusion, due to the lack of comprehensive storage technology education in the industry, most of the storage professionals have been trained on-the-job to be able to perform day to day operations specific to the vendor products provider and self development. To overcome this narrow orientation of storage professionals, trained well to work only with a few number of vendor solutions, well-constructed, comprehensive, and strategic plans must be efficiently implemented to meet the challenges of managing multi-site, multi-vendor environments. This problem together with the issues mentioned before is well defined in the next chapter.

3 Conceptual design of data migration tool

3.1 Introduction

The previous chapters of this thesis have described into details all of the factors and features that must be identified in the process of choosing the best solution for a particular data migration. Based on these factors, selecting the ideal migration tool for a specific storage environment can be considered as a very complex task.

First, the volume of data to be migrated is increasing at astonishing rate. The size of the files that has to be moved from one system to another is becoming larger and larger thus directly affecting the time for the migration process to be performed. Controversially, the down time window for system maintenance is decreasing in size, converging to zero-downtime. These days, people want access to relevant applications at any time of the day which leads to the second migration issue. Continuous data availability has become must for the success of all companies that depend on digital information to conduct business processes. Another issue related to the application performance is the utilization of network bandwidth and system overhead that must be maintained to avoid bottlenecks during the process of migration.

Next, the importance and increasing dependency on data have lead to development of larger and more complex storage environments. Storage infrastructures usually support multiple operating systems such as mainframe, UNIX, Microsoft Windows, or Linux technology, with the mix being determined by the specific applications in use; and different storage arrays provided by EMC, IBM, HP or NetApp and many more. Moreover, multiple storage vendors have unique requirements and different migration techniques, further increasing complexity to the process.

In the context of multi vendor storage environments, interoperability becomes increasingly important concern. Data migration products must operate across every platform involved in the storage migration initiative, including the target storage system, the existing systems being migrated and any other system that may take part in the migration. However, some data migration products are dedicated for specific storage platforms, even more between particular versions; others favor specific application types,

such as transactional database, still others are more suitable for SAN or LAN environments.

Having all of these factors in mind, the problem occurs when it has to be identified potential data migration solution that can successfully move the data from the existing environment to the target environment. As it is stated in the white paper "Managing Information Storage: Trends, Challenges and options 2011-2012" provided by EMC, the storage teams 5.9% of the time in the last 12 months have spent in evaluating storage technologies from different vendors. In reality, there are myriad migration solutions available in the market that can help users address their migration challenges, but many of them are still not aware of these tools.

As information continues to grow at exponential rates, the importance of a sound migration strategy is critical. The increased complexity and scope of businesses storage environments requirements contributes to the need for a quick reference of a possible migration solution for IT staff.

3.2 Data migration selection tool purposes

The main goal of this thesis is the development of a data migration selection tool for heterogeneous storage environments. This application is design to serve as a starting point when considering possible migration solutions, based on the source storage array and the target devices, the host operating system and special user requirements. This will help an organization or any knowledgeable delivery resource in a need of a quick reference to minimize research and evaluation of different storage vendor technologies and maximize planning efficiency. It provides as a result possible migration solutions that are compatible with the given environmental requirements. In addition, some recommendations of using that solution are given, such as the expected downtime or limitations if possible.

This tool is most suitable for people who are familiar with the technologies in use and have already some experience in data migration.

3.3 Main objectives

This migration tool has to meet the following objectives:

- quickly to provide a possible migration solution between storage arrays taking into consideration solutions from multiple providers, not only from a single vendor specific products;
- identify the interoperability problems in multi-vendor, heterogeneous storage environments and;
- to speed up the time to implementation.

3.4 Delimitations

The core of data migration process is to move data from a source system to a new target environment. As a result of that, for data migration can be used also all mechanisms for backup and restore, storage virtualization and replication. However, in this study the scope will be limited only to the data migration tools which have specially developed features to meet the purposes of migrating data, storage virtualization techniques and replication. In a case when there is not such software suggestion to alternative techniques will be provided but not into details.

Although, the traditional method for migrating data is by using a backup and restore mechanism, this often results in taking production applications offline for extended periods of time and it can be very tedious in a case when tape drives and libraries are used to backup data due to limited throughput rates. Also, if disk drivers are used as storage devices the time needed first to backup and restore data and then to migrate is not comparable with the needed just to replicate the original data and to migrate to the new storage platform.

Also, there are a great number of possible migration solutions that exist in the market and because of the different requirements and storage platforms that are involved; only the solutions from the widely used storage platforms in the market will be covered. Those storage arrays are mainly used in small to medium size businesses and large enterprises to manage data. Small (home) network storage devices are not taken into consideration. The data migration to/from this system can be made by simple copy because

we assume the data that has to be move is in a small volume and is not critical for running the business.

Since the main purpose of this selection tool is to give a quick reference to the possible migration tool that can be used mostly between different vendor storage platforms and is focused on the compatibility between different storage arrays, more detailed factors about storage environment including: patch level, software versions, network performance, port capacity, I/O bandwidth and the number of hosts that network switches can support is omitted.

This application gives only a potential migration solution between two storage platforms. The migration between multiple storage systems is out of the scope of this study. Also, migration between different databases or applications is not considered because they are different type of migrations with its own characteristics, requirements and technologies.

Based on the worldwide external controller-based disk storage vendor revenue estimates for the first quarter of 2011 [17] the top-tier vendors are EMC, NetApp and IBM given in a descending order. In the first version of the data migration selection application will be covered products and services from these companies as a result of their storage platforms presence in the market.

In the next chapter, will be exemined the storage platforms hardware specification and associated software that can facilitate the migration of data between those systems. The interoperablility and compatability relationships between the given storage products and services must be properly established.

3.5 Application design

For development of the application for selecting an ideal data migration solution in a heterogeneous storage environment had been used Microsoft Visual Studio 2005 Forms Designer. The source code is written in C# and the database is created in Microsoft Access 2003

In figure 3.1 below is given the database diagram and all relations established between those entitites.

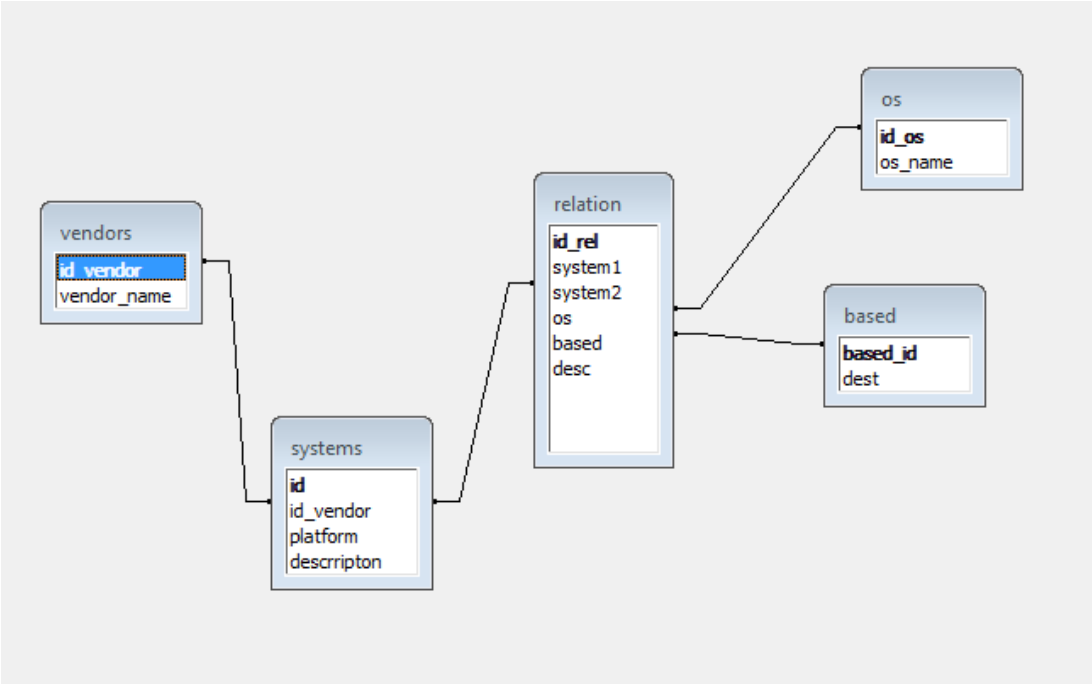

Figure 3.1 Database diagram

The database consists of the following tables:

- Vendors – takes in the storage vendors: IBM, NetApp and EMC
- Systems – comprises of all storage systems where id_vendor is foreign key
- OS – list of all operating systems
- Based – List of all migration approaches (array-based, host-based and network based)
- Relation – includes all possible relations between system1- *the source system* and system2 -*the target array*. These relations depen on the selected operating system and migration approach.

The database and the application are connected with the build in control in Visual Studio 2005 – DataSet. The relation between the applicaton and Microsoft access database is db1DataSet.xsd illustrated in figure 3.2. Dataset is using Microsoft Jet engine to

communicate with Microsoft access. Dataset is importing all data as XML objects and after that Visual Studio is using XML classes for all queries.

db1DataSet.xsd is splitet in 4 objects.

- db1DataSet.cs – automatic generated class containing functions and procedures for connecting databse with xml dataset
- db1DataSet.Designer.cs – This class is generateing GUI for easier access with XML objects, and creating queries with graphic interface.
- db1DataSet.xsc – all data from database presented as XML.
- db1DataSet.xss – XML schema.

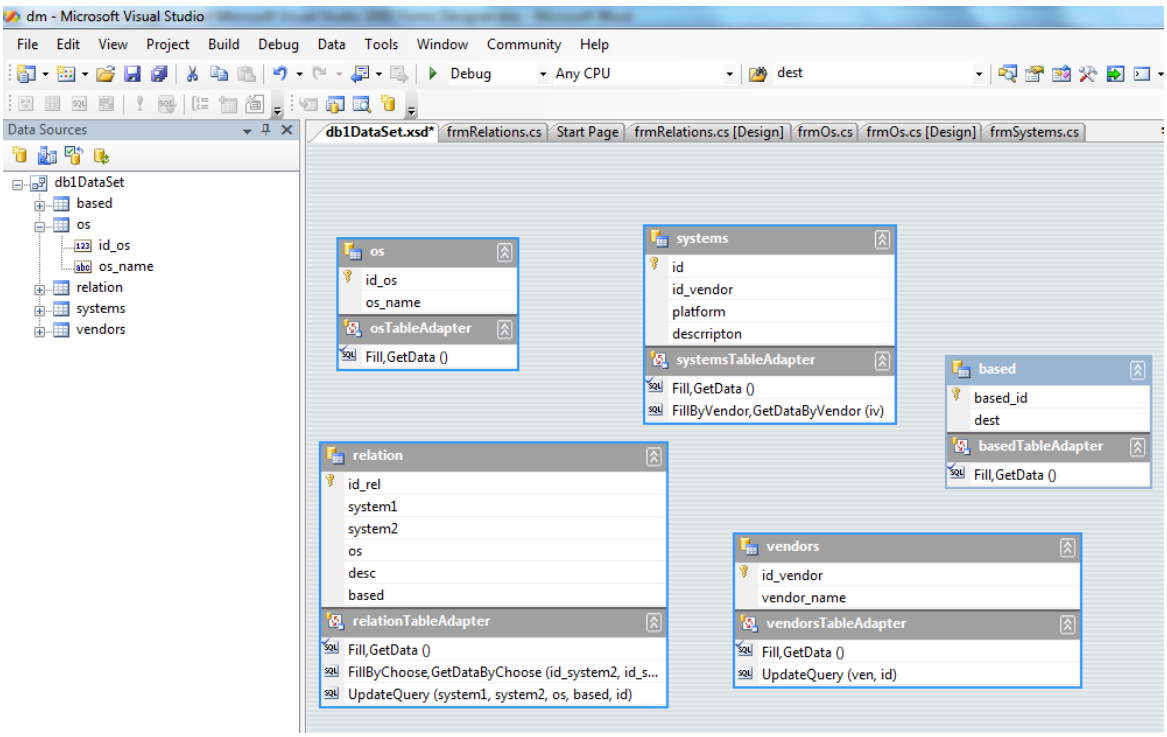

Figure 3.2 db1DataSet.xsd connects the database and the application

In the figure 3.3 are listed all classes used in this application.

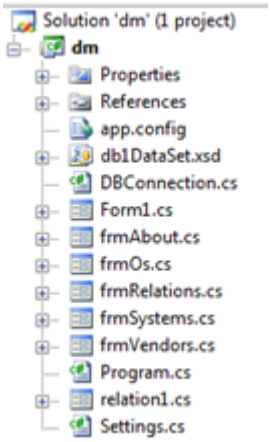

Figure 3.3 Application classes

Form1.cs is the main form class. It is further devided in two subclasses:

- Form1.cs – it is user developed code which provides all functions and procedures.
- Form1.Designer.cs – is visual studio automatic generated code for application design.

The components of the form1 are:

- Menu item – contains the main menu to enable access for users in the other forms in this application. Menu item is placed on the top part of the application.
- 7 listboxes – these boxes are connected with dataset which gather filtered data for diferent puposes. For example, vendorListBox1, located in the top left corner, is listing all vendros from database table Vendors. SystemListBox1 is listing systems filtered by the vendor choosen in vendorListBox1. Filtering is defined as SelectIndexChange event that is using queries defined in DataSet1. The same principle goes for vendorListBox2 and SystemListBox2.
- descListBox1 is the result listbox. The final result depends on the selected information from all other Listboxes. When the user clicks on the button Get List from the user application interface, in the background button1.click event determinies the fill method to disply the possible results on the screen. Button1.click event is using parametered queries defined in relation

TableAdapter.FillByChoose in db1DataSet. In figure 3.4 is given the screenshot of the main form Form1.cs.

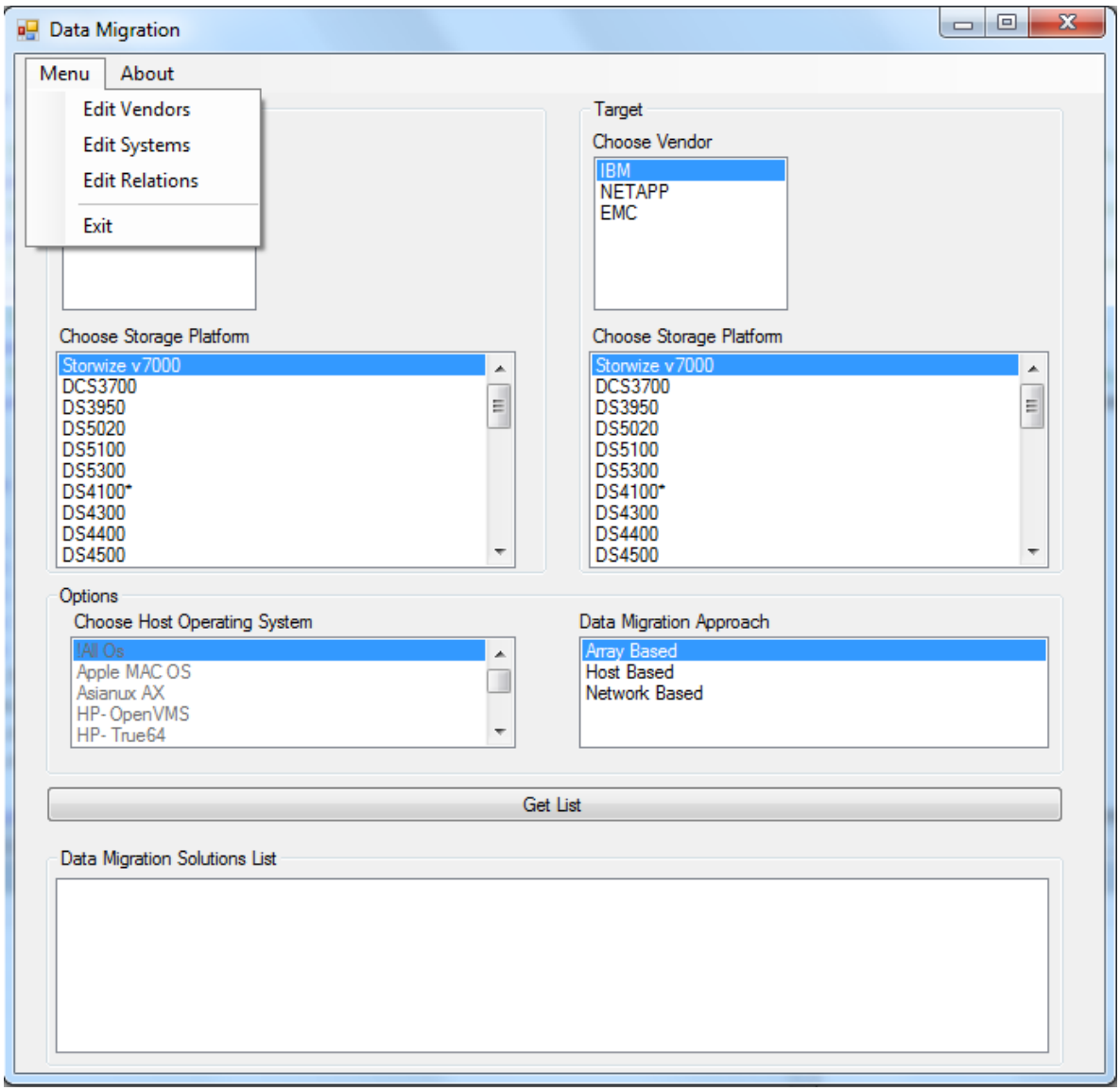

Figure 3.4 Form1.cs

- frmOs.cs, frmSystems.cs, frmVendors.cs are three form classes used for adding, deleting and modifying data in adequate tables.
- frmOs.cs is used for adding, deleting and modifying Operating Systems and it is connected with table “os” over db1DataSet.
- frmSystems.cs is used for adding, deleting and modifying data storage Systems in table “systems” and it’s connected with table “systems” over db1DataSet.

- frmVendors.cs is used for adding, deleting and modifying data storage Vendors in table “vendors” and it’s connected with table “vendors” over db1DataSet.

All these forms are using DataGridView control for listing data in database and BindingNavigator control for navigating trough data. These forms are connected with database over db1DataSet, illustrated in the figure 3.5 below:

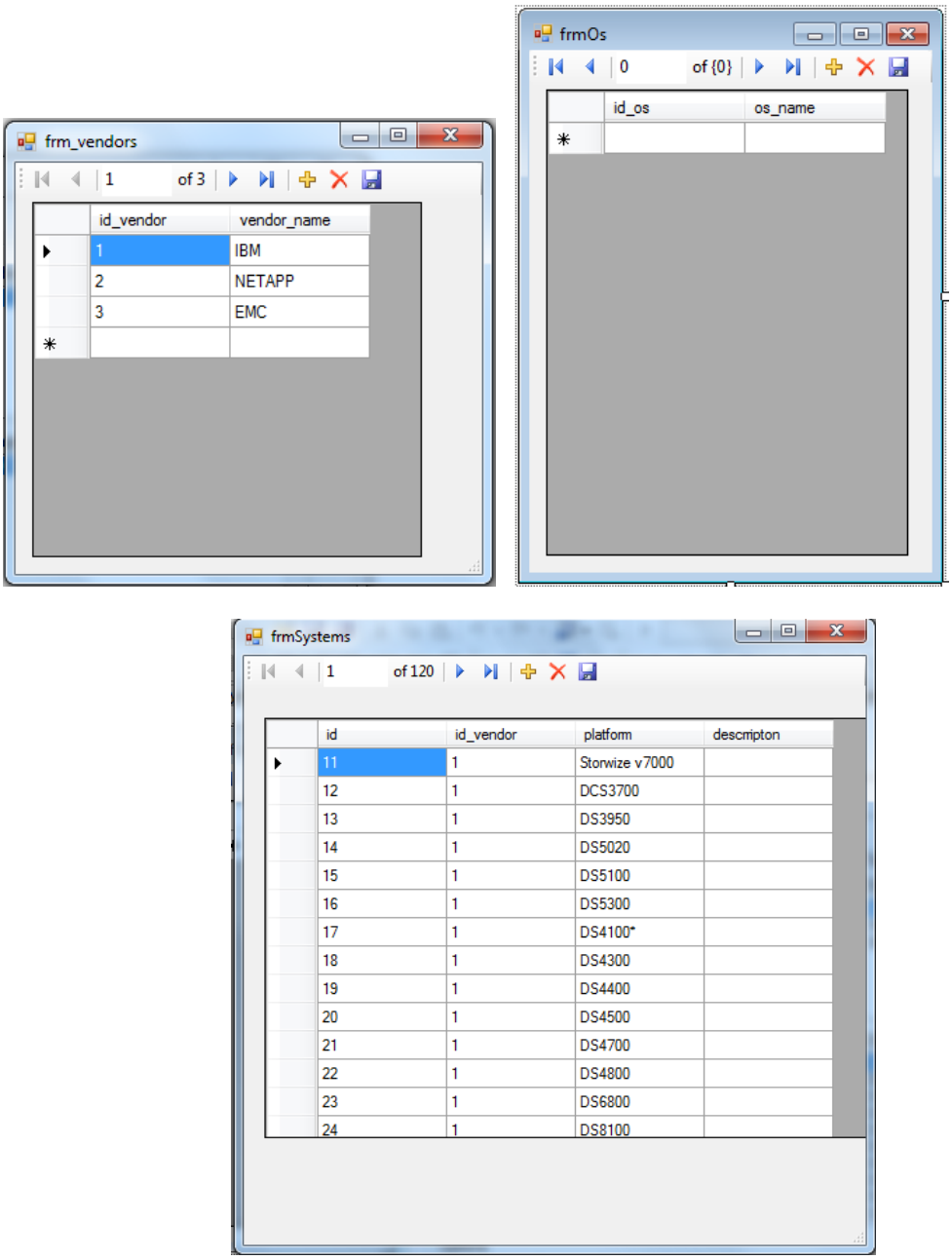

Figure 3.5 Screenshots of the forms: frmVendors, frmOs, frmSystems

The last and the most complex class is frmRelations.cs used for creating relations between tables in the database. Listbox controls are used for choosing data for the relation. PlatformListBox1 and platformListBox2 are displaying data from table systems, whereas osListBox and basedListBox are displaying data from tables: os and based. After these choices have been made the result for the selected relation is displayed in DataGridView control. This control has options for adding, deleting and modifying relations for specific choice. Different queries are used for these functions defined in db1DataSet and any change is imediately applied in the database with relationTableAdapter.Update parametered query.

Screenshot from the form is given in the figure 3.6:

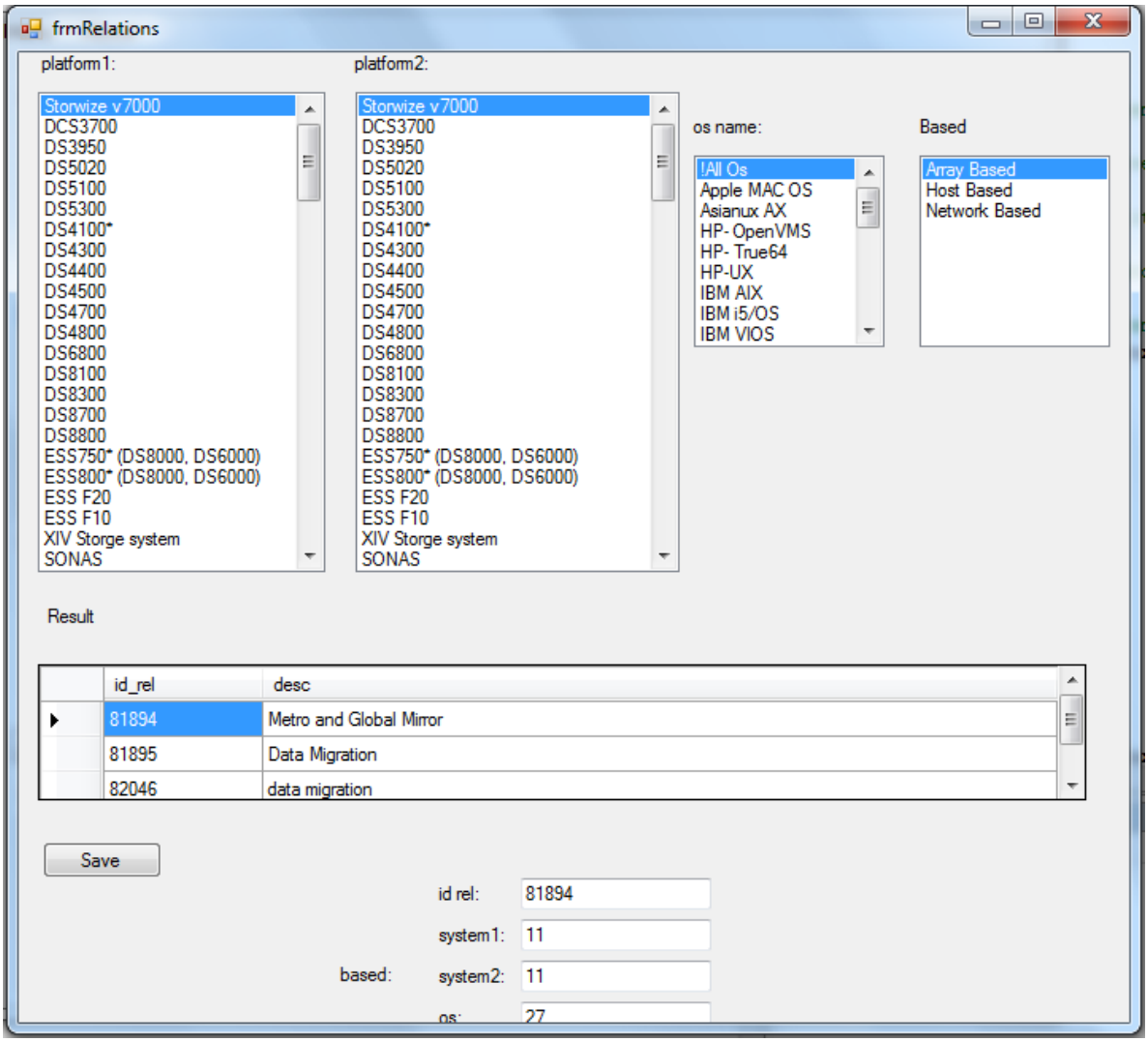

Figure 3.6 Screenshot of frmRelations

4 Storage systems specifications and associated software

4.1 Introduction

In this chapter are discovered the storage systems and its associated software that will be implemented in the data migration selection tool. For more detail information about product models, versions, supported host operating systems, disk types, network protocols and RAID configurations please refer to Appendix A of this document. These storage arrays, as it was stated in 3.4, are part of the product lines of the world's leading vendors in storage area: EMC, NetApp and IBM. As best of the breed of storage technology these products are built in with intelligent.

The so-called intelligent storage systems have some familiar characteristics among which are: having an operating storage environment that controls the management, allocation, and utilization of storage resources, RAID arrays that provide highly optimized I/O processing capabilities and sophisticated algorithms to meet the I/O requirements of performance sensitive applications. These storage systems are configured with *front-end* ports and controllers that are connected to large amounts of *cache* memory to the *back- end* ports and the *physical disks*. In figure 4.1 is shown architecture of an intelligent storage system with its components and their interconnections.

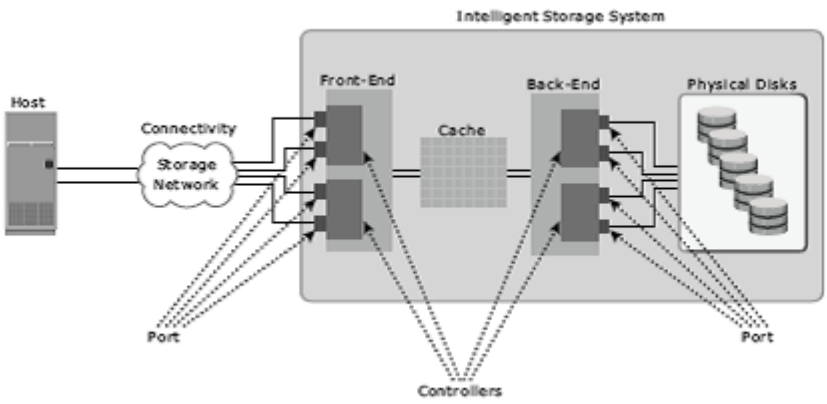

Figure 4.1 Components of intelligent storage systems (*source:"Information and Storage Management"* [1])

- *Front-end*- is the interface between host and storage array. It enables communication between these entities through one of supported transportation protocols: FC, SCSI, iSCSI, FICON, ESCON and TCP/IP. The data that recievs on the front-end port is routed to the cache by utilizing internal data bus.
- *Cache*- The main role of the cache is to speed up the time for retrival of data from the physical disk. It is a semiconductor memory which stores data temporally. Unless the requested data is found in the cache then it should be served from there.
- *Back end*- communicates with the destination physical disks when performing reads and writes that come from the cache. Additionally, on back-end controllers are implemented algorithms that may provide RAID functionality, error detection and protection.
- *Physical disks*- store data persistently. Intelligent storage systems may support different disk types in one enclosure.The hard disk drivers are known by the name of the predefined protocols they support such as IDE/ATA, SATA, SAS, SCSI, FC or SSD.

Nowadays, the most widely used disk types in storage arrays are SATA, FC and SAS. IDE/ATA are becoming obsolete due to its performance characteristics and has been replaced by serial technologies such as SATA and SAS. The new generation of storage devices is towards using of solid state flash drives (SSD) that utilize semiconductor memory to support persistent storage. Contrary to the FC or SATA drives, which are electromagnetically devices with spinning disks and movable heads, flash drives have no moving parts resulting in lower access time and less power requirements to run. However, the cost per GB is higher for them.

4.2 Storage array categories

The intelligent storage systems belong to one of the three main categories: entry level storage systems, midrange storage systems and high-end storage systems known as enterprise storage systems (ESS). They are configured in a manner to meet the business requirements for storage capacity, connectivity, time to service I/Os requests, scalability, data availability and many other features.

The high-end storage arrays, traditionally called active-active arrays, may serve I/Os requests across any of the available paths as shown in the figure 4.2. These systems are absolutely essential for organizations that require high performance and huge storage capacity. High-end storage devices are configured with large number of controllers and cache memory and support connectivity to other high-end or midrange storage platforms via FICON, ESCON or FCoCEE protocols. Also, they provide unique features, such as array based, local or remote replication, virtualization and support for multiple storage tiers. High-end storage systems are most appropriate for large enterprises to centralized data and to fulfill mission-critical applications requirements.

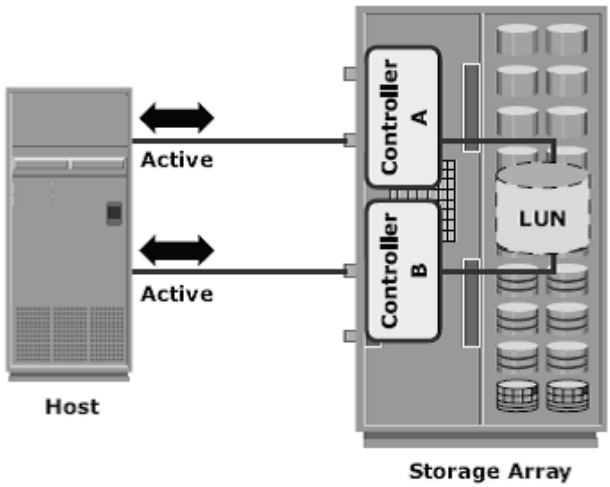

Figure 4.2 High end active-active array configuration (*source:"Information and Storage Management"* [1])

Midrange arrays are suitable for small to medium size enterprises since they come with lower total cost of ownership but provide some of the capabilities of high-end storage systems. Midrange storage arrays are designed with active-passive path where only active paths can be used to carry reads and writes, figure 4.3. They host less storage capacity and global cache and have fewer front-end ports for connection to servers compared to ESS systems. However, they ensure high redundancy and high performance for applications with predictable workloads. As high-end systems do, they also support array-based local and remote replication.

Entry level storage systems are designed to deliver advanced functionality at a affordable price. These systems provide an excellent solution for workgroup storage

applications such as email, file, print and Web servers, as well as collaborative databases and remote boot for diskless servers. Entry level storage systems are most suitable for small business and are less expencive to be maintained. They are usually built with SATA drives and support iSCSI or NFS/CIFS protocols.

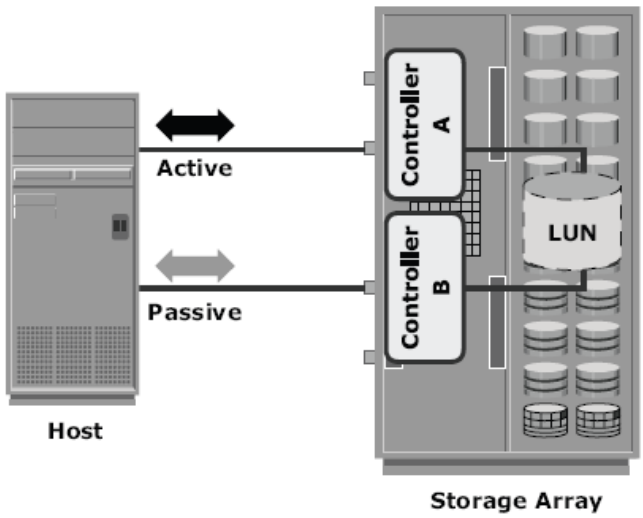

Figure 4.3 Midrange active-passive array configuration (*source:"Information and Storage Management"* [1])

4.3 SAN/NAS storage arrays

In 2.7 "Storage network" the terms of SAN and NAS have been introduced, however here we go back to them from a different point of view. Storage arrays can be classified based on the level of data access. Typically, they can be divided in two groups NAS or SAN, figure 4.4.

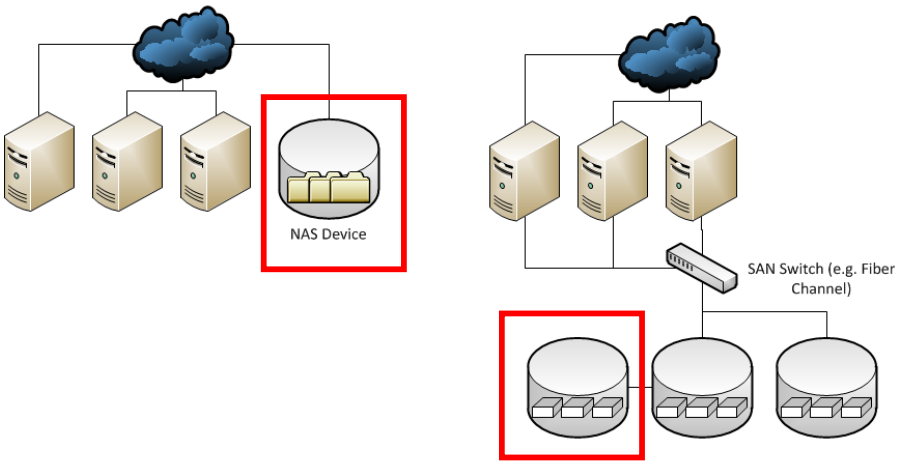

Figure 4.4 NAS file level access vs. SAN block level acces

NAS is a file level computer storage designed specifically for serving files. File sharing protocols such as CIFS, NIFS, NDMP enable users to share file data across different operating environments transparently. Unlike NAS, which is visible from the host side as a shared storage system, SAN appears to the server as a logical disk drives and enables server to have block level access to the specific parts of the storage.

Unified storage array architecture has been introduced in the market as well. This architecture provides capabilities from both NAS and SAN systems, figure 4.5. These systems do support both block and file level storage networking protocols. Thus, unified storage platforms can be easily implemented in any storage environment and can simplify the process of data migration.

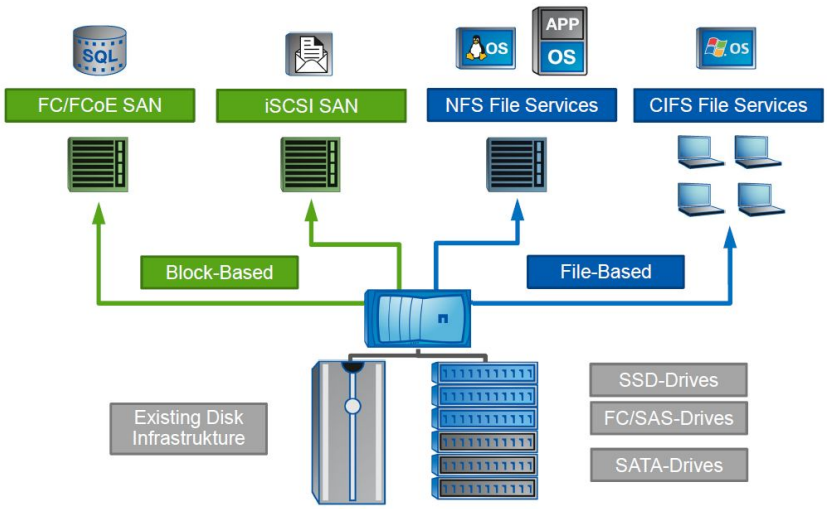

Figure 4.5 Unified storage access

4.4 EMC Storage Systems and associated software

EMC is the leading company in the market offering a broad range of storage products and services. This study is concentrated on specific storage hardware and related software that can be used in the process of data migration. To illustrate the concepts just discussed above, this section covers the EMC intelligent storage arrays classified in the following product families:

- *Symmetrix Family*- high end storage systems
- *Clariion*- midrange storage systems
- *Celerra*- entry level storage system
- *Centera*- content addressed storage system (CAS)
- *VNX*- unified platform
- *VPLEX*- midrange/enterprise

4.4.1 Symmetrix Family

Symmetrix is the family of EMC's high-end storage devices designed to meet the needs of large size enterprises in management and control of mission critical businesss operations. The Symmetrix family of storage arrays is characteristic by providing high level of performances, broad connectivity, consolidation of massive amounts of data and easy of management software. Symmetrix offers an extensive product line currently represented by Symmetrix VMAX series preceded by the Symmetrix DMX- Direct Matrix Architecture (DMX-4, DMX-3, DMX-2 and DMX-1) and earlier Symmetrix storage models such as Symmetrix 8000 series, 3000/5000 and 4400, figure 4.6 [31].

In the core of all Symetrix systems can be found the same basic architecture which consists of three functional areas: front-end, cache and back-end. The communication between the front and the back-end goes through the global cache memory. The primary differentiation between all Symmetrix generations is exactly in the technology used to interconnect the front and the back-end with the cache. This communication between controllers or directors has evolved through years from bus-based or switchch-based that can be found in Symmetrix8000 series and earlier, to all point-to-point connections introduced in Symmetrix Direct Matrix Architecture, and lastly, to the Virtual Matrix Architecture built in Symmetrix VMAX models. Also, differences between system's generations can be found in the system configuration which ismade of different number, type, and speed of processors and disks.

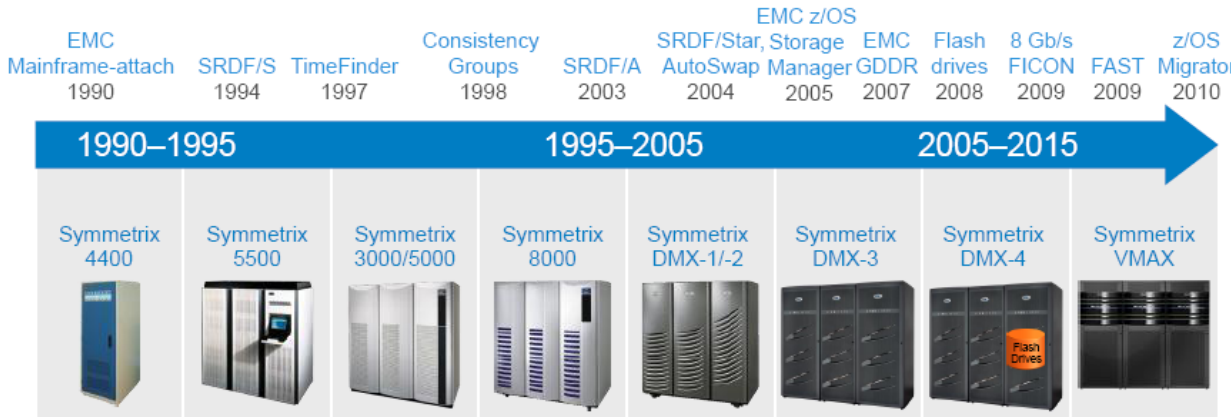

Figure 4.6 Symetrix Family models development through years

EMC Enginuity is a native operating system specifically developed for Symmetrix storage arrayes in order to provide the basic system functionalities and features among which are local and remote replication and data migration facilities. Most precisely, Enginuity is an emulation code which is loaded in the processor of each controller and coordinates the work of independent directors. This means that Engimuity reside in the Symmetrix array. All functionalitites this operating system enviroment provides are closely tied to the level of Enginuity code. Also, it enables simultaneous connections to virtually all mainframe systems, UNIX, Windows, iSeries, and Linux platforms.

In addition, Symmetrix management software can be used on top of EMC Solutions Enabler to discover, monitor and configure the physical disk drives in logical units that can be further presented to other host systems as fixed block architecture. The management software has the ability to automate many of the replication tasks like Open Replicator and SRDF.

Open Replicator

EMC Open Replicator is software that provides capabilities for array-based data migration between qualified systems given in Appendix B. It was firstly introduced in Enginuity code levels 5671 and 5771 [9] as an application that runs in the Fibre Director (FA) of the Symmetrix VMAX or Symmetrix DMX storage arrays. Thus, these systems must be present in all Open Replicator configurations because they take the control over the

replication. The Open Replicator software causes the control arrays to appear as an open systems host to the *remote* storage array, while it continues to simultaneously function as the host front-end to the Symmetrix. Since earlier models of Symmetrix do not support the required 5x71 code level are not able to *control* the Open Replicator process. However, since they are configured with front-end Fibre Channel controller, can be remote (target) array for Open Replicator.

In general, EMC Open Replicator for Symmetrix enables remote point-in-time copies to be used for data migration between EMC Symmetrix VMAX or DMX and qualified storage arrays with full or incremental copy capabilities [29]. They are four possible scenarios when Open Replicator is in usage depending on the state of control array wheter it is online or offline; and the type of copy which can be push or pull. If the conrol array is offline this mode is called *cold*, in opposite, when it is online it is known as *hot* mode. In both cases the remote arrays should be offline. Furthermore, *push* type of copy is when data is copied from the control device to the remote device. In contrast, *pull* operation copies data to the control device from the remote device [9].

The first scenario is when data is pulled from source volumes on qualified remote arrays to a Symmetrix VMAX or DMX volume shown in figure 4.7. In pull operations, the Symmetrix volume can be in a live state during the copy process. A *Copy on First Access* is technology that is used here to ensure that the data is already copied and thus it is available to be accessed from local hosts. Open Replicator hot pull can support the Federated Live Migration (FLM) technology.

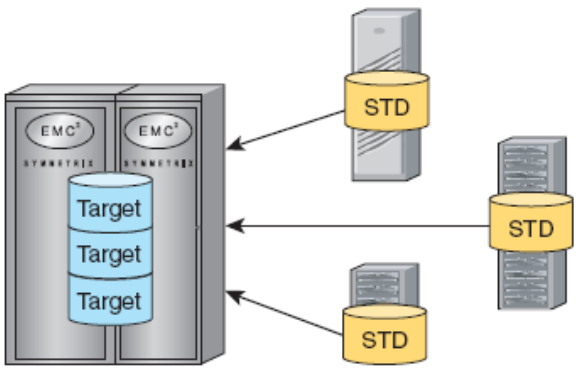

Figure 4.7 Open Replicator hot pull scenario (source: "*Choosing a Data Migration Solution for EMC Symmetrix Arrays"*[9])

Secondly, in a similar manner the Open Replicator cold pull can be performed. The difference is in that the target SymmetrixVMAX or DMX volume is in static state (STD) and cannot be accessed during the migration. This is illustrated in the figure 4.8.

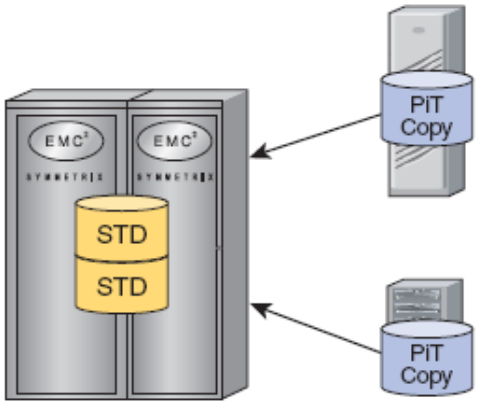

Figure 4.8 Open replicator cold pull scenario (source: "*Choosing a Data Migration Solution for EMC Symmetrix Arrays"* [9])

Open Replicator hot push scenario where the data is pushed from any online source volume, Symmetrix VMAX or DMX, to a target volume on a qualified array with incremental updates. For hot push operation Symmetrix creates logical point-in-time copies of source volumes, but during the transfer to the remote site the I/Os to the source volume are disabled. This is illustrated in figure 4.9 below.

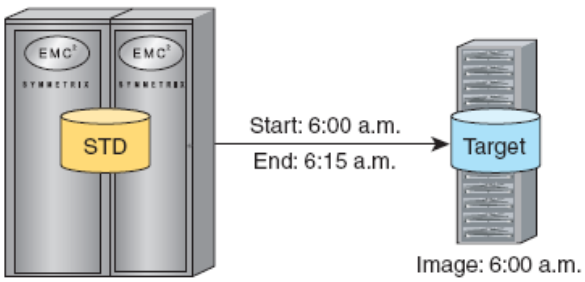

Figure 4.9 Open replicator hot (live) push scenario (source: "*Choosing a Data Migration Solution for EMC Symmetrix Arrays"* [9])

At last, Open Replicator cold push enables data migrations from Symmetrix VMAX or DMX to qualified storage with minimal disruption to the host application. In this scenario, usually a point-in-time copy of production devices is created so that the migration can proceed while the production devices are still in use. This is static copy of production data and is called business countinous volume (BCV). The copies of production data can be

made by using TimeFinder/Clone or TimeFinder/Mirror, Solutions Enabler 7.0, and SMC 7.0 or a TimeFinder/Snap VDEV (Virtual Device) [9]. The Open Replicator cold push is shown in the following figure 4.10.

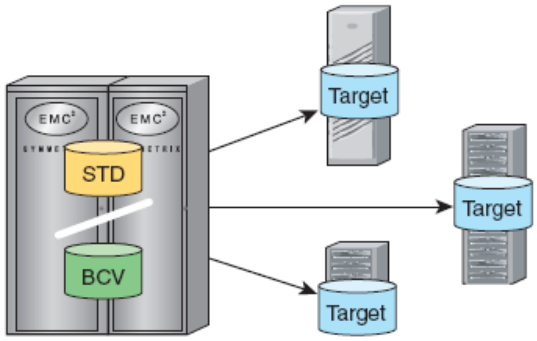

Figure 4.10 Open replicator cold push scenario (source: "*Choosing a Data Migration Solution for EMC Symmetrix Arrays"* [9])

In heterogeneous storage environments Open Replicator hot pull scenario is usually used for data migration in conjuction with other techhologies in order to increase the overall performance of the process. As illustrated in the figure 4.11, during the data migration all I/O requests go through the Symmetrix system which enables the host application to be online. EMC Open Replicator operations can be used together with Federated Live Migration and Power Path Migration Enabler to avoid an application outage when redirecting the application I/O to the new migrated devices. This is so called zero time data migration process.

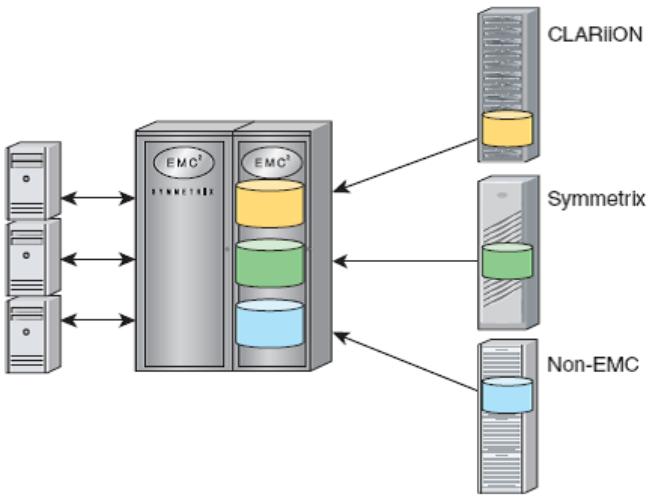

Figure 4.11 Open replicator hot pull scenario in heterogeneous storage environment (source: "*Choosing a Data Migration Solution for EMC Symmetrix Arrays"*[9])

Symmetrix Remote Data Facility (SRDF) Family

The EMC Symmetrix Remote Data Facility (SRDF) family of replication software can be used to facilitate operations regarding data replication. SRDF products offer the capability to maintain copies of data in different physical locations. SRDF is an array based data migration which is embedded in Engiunity operating system code and can facilitate data migration between Symmetrix systems via Fibre Channel, Gigabit Ethernet (GigE), and ESCON [9].

The baseline of SRDF Family consists of SRDF/Synchronous (SRDF/S) replication, SRDF/Asynchronous (SRDF/A) replication and SRDF/Data Mobility (SRDF/DM) which are host independent functionalities. There are supplementary options and features that can be added to the base solutions to solve specific service level requirements for business continuance (SRDF/S), data consistency (SRDF/CG), countinuous protection (SRDF/Star) and many more, but here the focused is on the main functionalities that can be used for data migration purposes.

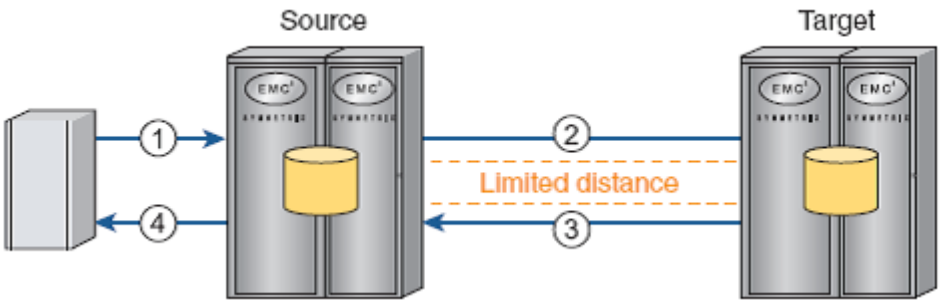

Figure 4.12 Synhronous data replication facility (source: "*Choosing a Data Migration Solution for EMC Symmetrix Arrays"* [9])

SRDF/S synchronous data replication has an ability to maintain a logical copy of data in real time between two Symmetrix systems locally or remotely. This replication of data is performed in the backround of the host applications and it duplicates the data independently of the host operating software. As it is depicted in the figure 4.12, the source system does not indicate that the write request is made to the host since the target Symmetrix has acknowledged it. This directly affetcs the host application and the distance

between the source and the target array, which is limited to maximum 200 km, in the figure 4.12.

In the figure 4.13, the flow of asynchronous mode of SRDF/A is illustrated. It is easy to understand that the distance between source array and target is not constrained. In asynchronous data replication, the write request is first stored on the local system and then the consistent data is sent from the primary to the secondary site in predefined timed cycles, in order of seconds, of delta sets. This eliminates the redunducy of multiple same blocks of data changes to be sent to the target and thus reduce the traffic send over the link. This increases the performance of data replication.

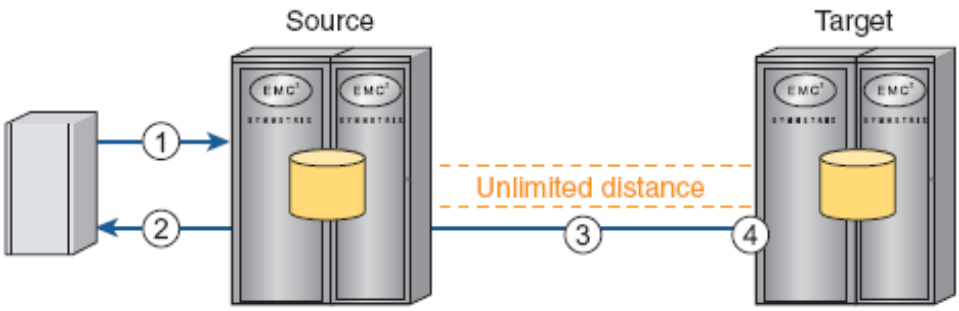

Figure 4.13 Asynhronous data replication facility (source: "*Choosing a Data Migration Solution for EMC Symmetrix Arrays"*[9])

The SRDF/DM product offering permits operation in SRDF adaptive copy mode only and is designed for data replication or migration between two or more Symmetrix systems. These copy modes allow the primary and secondary volumes to be more than one I/O out of synchronization. The flow of the process is the same as the one shown in figure 4.13.

Two additional features of SRDF may have to be considered when SRDF is used for data migration. At first, SRDF allows source and target storage arrays to change their roles. This swap means that the data application has to be restarted at the new site which results in a short application outage. Secondly, data will be remotely accessed by the host from the target device if it is not available on the source array. It is not necessary the synchronization between both systems to be completed before moving to the new source, because any data not yet in the new system can be accessed remotely from the old. However, in order to avoid loss of data from multiple failures the best practice is to wait for synchronization before the swap [9].

Other Symmetrix software

In this part will be disscussed additional Symmetrix features which may play a role in data migration: *Federated Live Migration (FLM)* and *TimeFinder* technology [9].

As the name reffered by itself the Federated Live Migration (FLM) technology is used when the old storage array DMX has to be replced with new VMAX platform [30]. The traditional cutover with FLM is eliminated during a migration. The host application I/O to the new VMAX storage is redirected in such a way that these paths are presented to the host as additional paths to the old system. FLM can handle the multipaths via MPIO driver or may use the PowerPath as application MPIO driver.

FLM uses the Open Replicator as underlying technology. Copying of data between arrays through the SAN is supported by the hot pull scenario, and the donor update messages keeps the source and the target devices synchronized during the migration. This is presented in the figure 4.30.

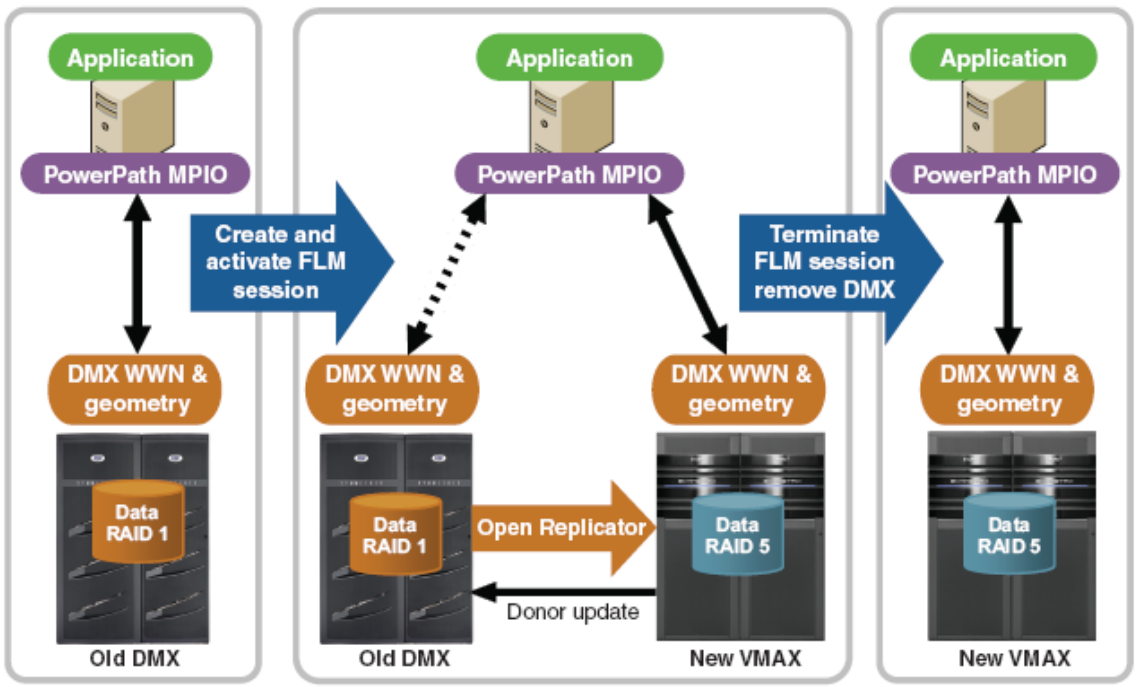

Figure 4.14 Federated live migration process (FLM) flow (source: "*Choosing a Data Migration Solution for EMC Symmetrix Arrays"*[9])

TimeFinder Family

The EMC TimeFinder family of local replication allows users to nondisruptively create and manage point-in-time copies of data within a single Symmetrix array. It is array-based solution which means that the host resources are not used. Symmetrix devices remain online for regular I/O operations while the the data is copying. The TimeFinder family consists of two base replication products: TimeFinder/Clone and TimeFinder/Mirror or TimeFinder/Snap.

TimeFinder/Clone provides clone copy sessions that create point-in-time copies of full volumes or individual datasets. In a similar manner, TimeFinder/Snap provides snap copy sessions that create economical, pointer-based replicas simultaneously on multiple target devices where only the pre-images of changed data are written.

TimeFinder family may take place in data migration process. TimeFinder/Clone or TimeFinder/Mirror can be used to create an independent full point-in-time copy within a single storage array. Redirecting an application to point to the replicated data in its new location is effectively a data migration. However, the storage array in this case is not replaced with a new one. Second, TimeFinder is used to create a local point-in-time copy, which is used as a staged copy for migration to a remote array, as seen in SRDF.

TimeFinder/Snap can also be used together with Open Replicator in two ways. First, TimeFinder may create a local point-in-time copy for an Open Replicator *cold* push migration to a *remote* array. Second, if the *remote* array is a Symmetrix, TimeFinder can be used to ensure a backup copy of the production data is available at the remote location in case there is a disruption during an Open Replicator incremental push. This is necessary, because when Open Replicator performs an incremental update from the production volumes, the data on the remote devices is not in a consistent state until the update completes. Enginuity 5874 and later no longer support native TimeFinder/Mirror operations, but they are emulated via TimeFinder/Clone [9].

4.4.2 Clarrion Family

The first child of Clarrion family namely AVIION was born under Data General production, but then it was abonden by EMC as a base for future developments.

Nowadays, the CLARiiON product line up starts with A4 and AX150 entry level storage devices and CX series of midrange platforms. The CX series has evolved from FC in more capable, extendable and faster storage arrays. The CX has developed through three generations of storage platforms from the first CX200, CX400 and CX600 to the second CX300, CX500 and CX700 characterized by processor and bandwidth improvements. The third generation of CX3Ultra scale was first in the market that has been providing 4 Gbps FC connections. The last generation is Clariion CX4 family with high level of performance [31].

The Clarrion family runs on native operating system FLARE (Fibre Logic Array Runtime Environment) which is self upgradable. It is very intresting that the FLARE Operating Environment gets loaded onto the service processor which is a PC running Microsoft Windows [33].

SAN Copy

SAN Copy software is specially designed for CLARiiON systems to provide very similar functionality as EMC Open Replicator for Symmetrix does. SAN Copy is used to copy blocks of data between multi-vendor storage systems over a high-speed SAN infrastructure. It is array-based application where CLARiiON storage processor acts as a host to the remote storage array so that no special software is needed in the remote site. Before performing SAN Copy of data Solution Enabler must be installed. CLARiiON storage systems that are qualified to host SAN Copy software are: CX4 systems, CX3 systems,CX400/500/600/700, AX4 while CX300 and AX100/150 support only SAN Copy/E [35].

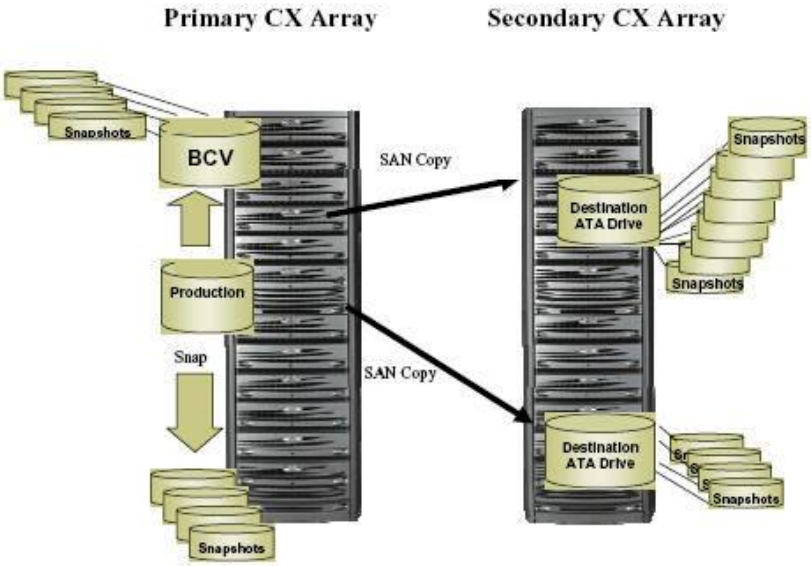

Figure 4.15 SAN Copy with full and incremental copy of data

SAN Copy, as well as EMC Open Replicator may operate in two modes push and pull as was explained in details in previous section. The push method is usually faster than the pull method. Also, unlike a pull, the push supports “incremental” copies. This allows the LUN(s) to remain online during most of the migration; the migrating LUN only needs to be brought offline for the final synchronization and cutover from the source LUN to the new destination LUN. Because of these advantages, EMC recommends to use the push method to perform a migration to your new platform [36].

Like Open Replicator, SAN Copy supports incremental copy capabilities for local array-based sources. SAN Copy would likely be used instead of Open Replicator when there is a need for incremental support while copying data from a CLARiiON source to a Symmetrix target [35]. While the software executes on a CLARiiON storage system, it can copy data from and send data to other supported storage systems on the SAN. Figure 4.15 illustrates some of the ways SAN Copy can be used to move data.

SnapView

SnapView is technology that can create local point-in-time snapshots and complete data clones for testing, backup, recovery and migration operations. A snapshot it is an image of how the source LUN looks at a particular point in time. Any writes to the source LUN result in SnapView storing a copy of the original data on a reserved LUN. A clone, also referred to as a business continuous volume (BCV), is an actual copy of a LUN. The clone is the same size as the source LUN and it takes more time to be created. SnapView clone provides users the ability to create fully copies of LUNs within a single storage system. The synchronize feature of mirroring is used to populate the clones. Once populated, clones can be used as point-in-time replicas of the source. When the clone´s association to the source is fractured, host writes will be to the source LUN only and SnapView software will be used to keep track of the changed regions of the source LUN.

In the same manner, SnapView can be used in association with Open Replicator and SAN Copy. In addition, when the Open Replicator remote array is CLARiiON, SnapView can be used to ensure a backup copy of the production data is available at the remote location in case there is a disruption during an Open Replicator incremental push.

EMC MirrorView

EMC MirrorView provides synchronous and asynchronous data replication between CLARiiON arrays across IP and Fibre Channel networks and can also be used for data migration purposes. Basically, MirrorView is array-based data replication software that provides end-to-end data protection.When MirrorView performs replication to a target volume, all server access to the secondary volume must be initiated through MirrorView. The advantage of using MirrorView over SAN Copy is that SAN copy does not provide the complete end-to-end protection.

Navisphere Management Suite

Navisphere provides management for Clariion's software functionality including SnapView, SAN Copy, and MirrorView. Navisphere also provides mechanism to set LUN masking required to make Clariion LUNs accessible to the Symmetrix VMAX (or

Symmetrix DMX) FA Open Replicator "host." Navisphere Management Suite also can be used to set up the Clariion to work as a *remote* array in Open Replicator operations. It makes the Clarrion LUNs to be accessible to the Symmetrix systems.

Recently, Unisphere management software has been introduced as the replacement for Navisphere Manager. Unisphere is unified manager that has capabilities to mange Clariion, Celerra, and complementary products such as RecoverPoint and Replication Manager through its interface.

Invista and RecoverPoint

EMC Invista is a network-based storage virtualization solution that runs on top of intelligent SAN switches hardware [37]. To perform the basic operation of virtualization Invista takes advantage of the processing power in the switch. Virtualization introduces a hardware abstraction layer to storage infrastructure, decoupling storage from operating systems. The storage device that is visible to a host is no longer array specific, but is a virtual entity that allows the arrays to move, expand, or change the storage infrastructure while the application remains online. Storage Virtualization provides the means for storage management across heterogeneous storage platforms in a way that the data can be migrated between storage systems while reducing or eliminating planned downtime. This is illustrated in figure 4.16, an intelligent switch on which Invista runs is attached to the SAN. Host and storage are physically connected to the fabric and can access their *Virtual Initiators* (VI) and *Virtual Targets* (VT) entities created by Invista.

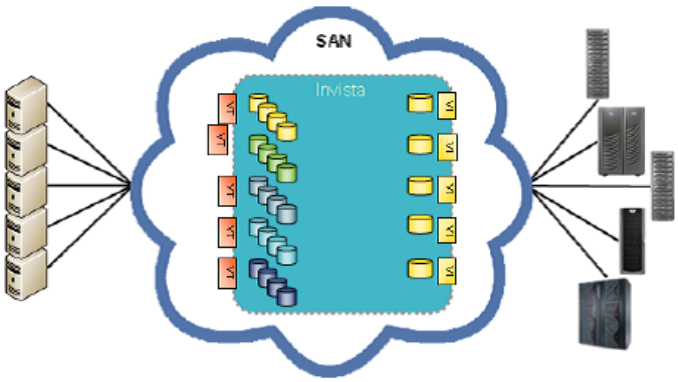

Figure 4.16 Logical topology of Invista implemented in SAN [37] (source: *Networking for Storage Virtualization and EMC RecoverPoint).*

Invista is built with *split-path* architecture to provide virtualization services to a host. Split-path combines intelligent Fibre Channel switches (using purpose-built ASICs) and EMC Invista software to redirect I/O to the appropriate physical storage device based on the mapping operation (association of virtual volume to physical volumes) which is created by the administrator when virtual volumes are created [37].

However, Invista by itself does not deliver functionality for data migration. In real environments, Invista is typically deployed together with other solutions such as RecoverPoint to enable data migration in virtualized storage enviroment. RecoverPoint provides the capability to locally and remotely replicate data at either the host or network level. The RecoverPoint consistes of RecovePoint Appliance (RPA) configured out of the direct data path and a splitter driver that resides on a Fiber channel intelligent switch to replicate data from the designated source volume [37]. The Recover point splitter driver redirects replicated data to the RecoverPoint Appliance which transfers this data to a remote location. The distance from the source to the target will be only limited by the physical limitations of TCP/IP connection between RPA and target device. The simple scenario is shown in the figure 4.17.

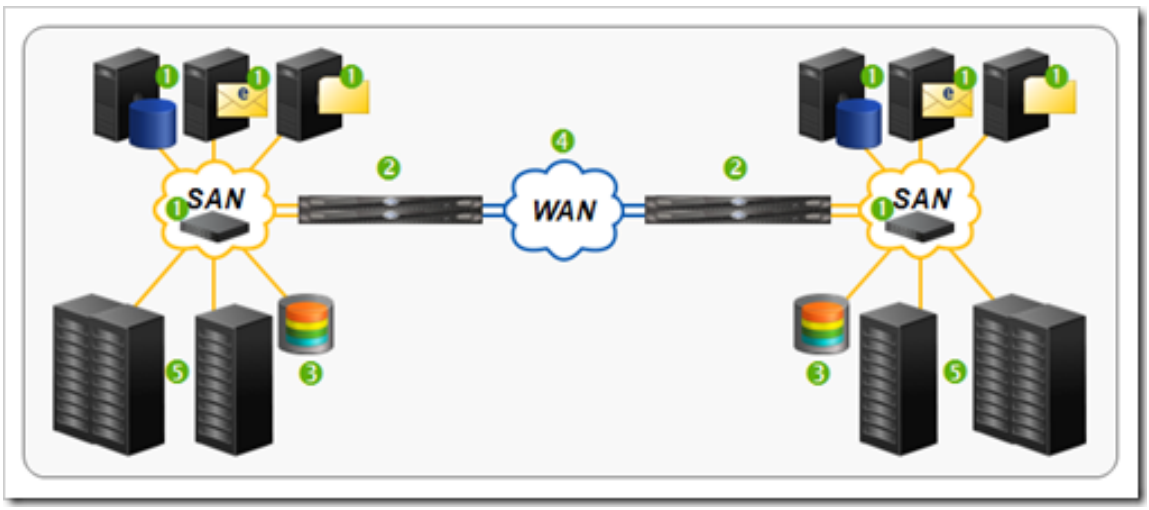

Figure 4.17 RecovePoint splitter driver may reside on the fabric switch as well on the hosts which is marked with number 1. Number 2 is the RPA. Storage arrays are numerated with number 5.

PowerPath Family

The PowerPath family belongd to host based solutions developed by EMC. PowerPath Family mainly is designed to manage the paths between storage devices and hosts to increase the data availability, to provide automatic load balancing and path failover functionalities during the data migrations processes. The PowerMulthipathing technology is used together with FLM to support the active/passive paths during the data migration.

The other member of this family is PowerPath Migration Enabler (PPME) which enables other technologies, like array-based replication, virtualization and Host Copy, to eliminate application downtime during data migrations or virtualization implementations. During the migration copy, PPME directs all application reads from the source device, and can transparently manage the redirection of the application to the target devices after the data movement completes.

The operating system creates *native* devices or specific path to provide access to logical devices. After migratioin is completed the application must be configured to use the new volume of data which is a disruptive process. In order to avoid this, the so called pseudo names can be used. These are actually names independent from the location and represent a logical volume and the path to be accessed. The data migration can be nondisruptive when pseudo names are used. This can be done by using PowerPath Enabler. Also, PPME can be used for smooth deployment of Invista virtualized environments by encapsulating (or bringing under the control of) the volumes that will be virtualized.

Replication Manager

Replication Manager is a host based software which is used for administration of replication technologies from the discovery and configuration to the operation of multiple disk-based replicas. Replication Manager provides support for TimeFinder alone or integrated with SRDF, SnapView alone or integrated with MirrorView, SAN Copy, RecoverPoint, Celerra SnapSure, Celerra Replicator, and Invista Clones.

4.4.3 Celerra and Centera Family

What is in common for both families of systems Cellera and Centera is that they provide file level storage acces or they are implemented as NAS devices.

EMC Centera family consists of platforms used for archiving enterprise documents, images, emails which are stored in order to meet government compliance and require only having access to them if needed it. Centera architecture is based on a no-single-points-of-failure Redundant Array of Independent Nodes (RAIN) architecture [38]. It provides basic capability for asynchronous replication whereby data is automatically copied from a source cluster to a replica cluster.

The Celerra product line consists of NX4 and NS series file systems mainly used to store and share files across small to medium size businesses. Celerra systems can be also integrated with other EMC platforms such as Clarrion. They are configured with data movers which take the responsibility to move data between the system and the host.

Celerra runs on real-time operating system called as Data Access in Real Time (DART). DART OS is a modified UNIX kernel with additional functionality added to operate as a file server. Cellera storage systems support specially designed technologies for file management such as FileMover, SnapSure, Celerra Replicator for asynchronous data replication and SRDF/S for synchronous replication to Symmetrix [31].

EMC Celerra FileMover is a Celerra feature that has the ability to automate the archiving of files in a network-attached storage environment across a hierarchy of storage platforms. Celerra FileMover is an API that allows the automatic migration of Celerra files to a secondary tier of storage such as Centera. It as well used by the File Management Appliance to archive data from Celerra to Centera [39].

EMC Celerra Multi-Path File System (MPFS) is used to support NFS file sharing when in use with EMC Celerra SnapSure or EMC Celerra ReplicatorV2 to protect file system data. A host with implemented MPFS driver interacts with the Celerra file server through a special File Mapping Protocol (FMP) to split the file content data flow from the NFS metadata flow, permitting file content data to move directly between MPFS clients and EMC storage arrays by using an FC or iSCSI link.

Celerra Replicator enables an asynchronous remote data replication for Celerra and supports all types of replications: File Systems, Virtual Data Movers, and iSCSI LUNS. It produces a read-only, point-in-time copy of a source file system, and iSCSI LUN or a

Virtual Data Mover and periodically updates this copy, making it consistent with the source object.

The Rainfinity File Management Appliance (FMA) is a policy-based file system archiving appliance for the EMC Celerra NS Series [38]. Based on established FMA policies, inactive or infrequently accessed data is archived to the Centera. Archived files are replaced on the Celerra with stub files. Stub files contain all the necessary metadata required for file recall by the users. The validated solution archives data residing on a Celerra to a Centera as the secondary storage. Source data on a Celerra can alternatively be archived to another Celerra.

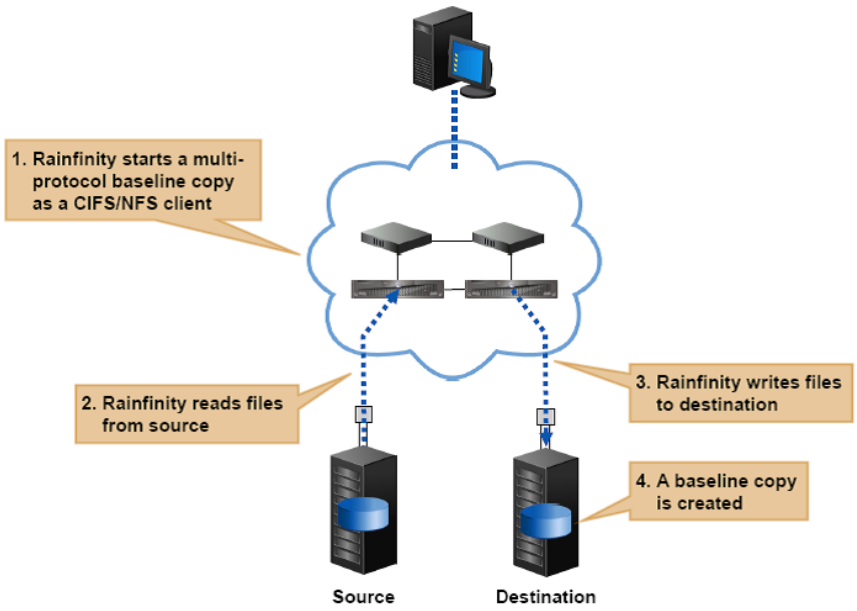

Figure 4.18 A simple scenario where Celerra Replicator is in use

This solution uses Celerra Replicator (V2) for asynchronously replicating the source data on the production Celerra to the target at the remote site, figure 4.18. Additionally, FMA may use the Centera replication to replicate a copy of the archived data from the source Centera to the target Centera.

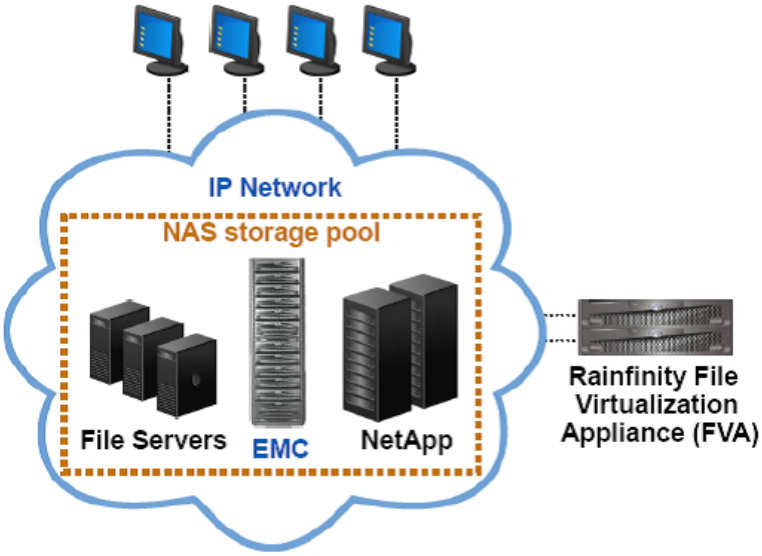

Figure 4.19 Rainfinity File Virtualization Appliance (FVA)

Moreover, EMC does provide a Rainfinity File Virtualization Appliance that dramatically simplifies the management and the data migration in multiprotocol NAS environments. It transparently moved data including active and open files independently of the file system [31]. This is shown in the figure 4.19.

4.4.4 VNX Family

VNX family is a unified storage family of products that consist of VNX midrange storage array, VNXe series of entry level and VNX Gateways that can be attached to Clariion and Celerra family systems. The VNX Gateways provide file access to existing VNX, Symmetrix and Clariion storage systems. VNX runs the Clariion FLARE and Celerra DART operating systems in one box. These products support virtual provisioning and EMC's SAN Copy data copying technology.

The VNX product line including VNXe, VNX and VG use EMC's Unisphere unified management software [31]. Unisphere is a storage management software specifically design to deliver a single interface for managing file, block, object and replication across VNX Series, CLARiiON and Celerra systems.

This family of systems is optimized for virtualization applications. VNX systems support file and block level protocols are given in the Appendix A.

4.4.5 VPLEX Family

EMC VPLEX is a hardware and software platform that resides in the SAN, between hosts and storage, and enables data mobility across different sites within the same data center, across a campus, or over distance. VPLEX Geosyncrony is the operating system that allows single copy of data to be shared, accessed among the systems that are connected to the VPLEX. This system provides tools for data migration across heterogeneous systems that can be found in the support matrix. Also, it provides Local and distributed platform federation. Local federation provides transparent cooperation of physical elements within a site, while distributed federation extends access between two locations across distance [49]. This is illustrated in the figure 4.20. VPLEX is a *Distributed Federation* solution, which means that we can enable a single copy of data to be shared, accessed and relocated over distance. In a traditional replication solution, there is a primary site and a secondary site for failover or recovery of data. In contrast, with VPLEX, the data is active at both sites at the same time. Rather than putting together complex failover scenarios, having the flexibility to have active data between sites will allow for maintenance without downtime and workload balancing across multiple data sites.

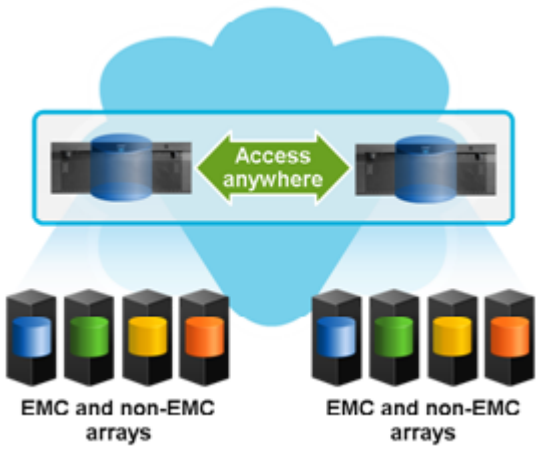

Figure 4.20 VPLEX virtualization storage

4.2 IBM Storage systems and associated software

IBM is one of the leading companies offering a wide range of storage products in the market [40]. The disk systems product line is very long consisting of DS family, Enterprise Storage Systems (ESS), Storwize v7000, XIV and N series systems. As it can be seen from the overall positioning of IBM Storage Systems offerings on figure 4.21 may support the needs of a small to medium size businesses and large enterprises as well.

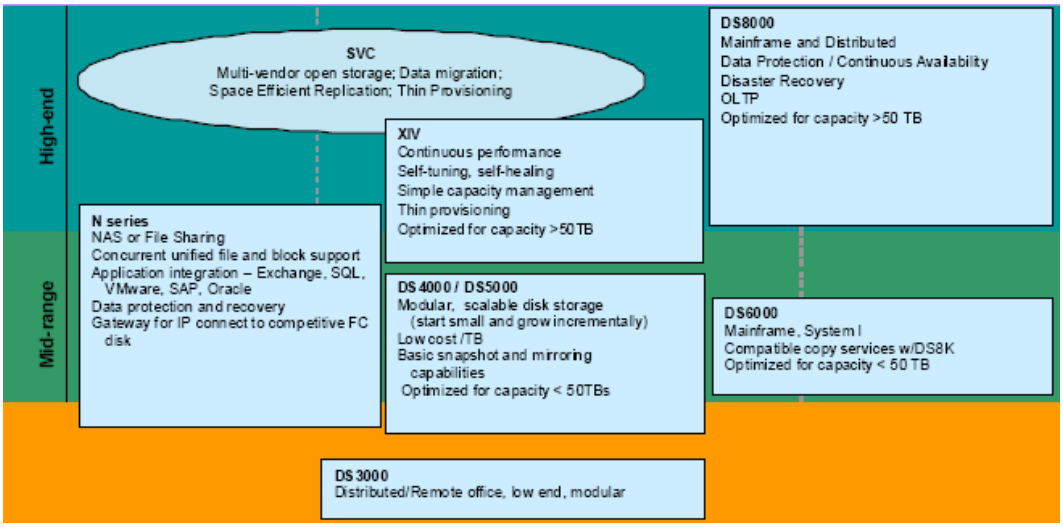

Figure 4.21 IBM Total Storage systems positioning (source: "*IBM Midrange System Storage Implementation and Best Practices Guide"[41]*)

Since, the storage technology is rapidly changing to meet the business requirements for increasing data capacity, availability, disaster recovery and continuity the storage vendors follow this trend by producing new storage systems with higher performance, upgrades the existing models with new capabilities or add new features to the current system versions. Based on this, many of the systems that will be used in this project are already withdrawn from the market. However, since the time they have been in production is not that long, usually less than an year, and most probably they are still widely used in storage environments they are of big interest and will be discussed in the next part. Here, the list of currently available IBM disk storage systems in the market is given from the IBM official web site [32]:

- High-end and Enterprise Systems: DS8000, XIV, SONAS
- Midrange and high performance computing systems: Storwize V7000, DCS3700, DS5000 series, DS4000, N series
- Entry level systems: DS3500, N3000 express

4.2.1 DS Storage systems family

IBM has brought together a broad range of storage disk systems under DS family in order to provide the right solutions which will fulfill business requirements of a small to a large-size enterprises. DS family combines the high-performance IBM System Storage DS6000 and IBM System Storage DS8000 series of enterprise servers that inherit from the ESS (E10, E20, F10 and F20), with the DS4000 and DS5000 series of mid-range systems, and DS3000 line-of-entry systems.

However, recently the lower end of the DS4000 family from DS4100 to DS46000 have been withdrawn from the market and replaced by midrange DS3750 and DS3950, while DS4800 have been upgrade to the DS5100 and DS5300 system performance. DS6800 has converged to the DS8000 high-end storage platform. DS8700 hardware has replaced previous version DS8100 and DS8300. DS3400 entry level was upgrade to DS3524 [ibm withdrawal documents].

DS Storage Manager

The DS Storage Manager software is the primary tool for managing, configuring, monitoring, and updating firmware, support data collection for the DS3000, DS4000, and DS5000 series of storage subsystems, and repair procedures. Various types of work that can be performed are configuration of RAID arrays and logical drives, assigning logical drives to a host, expanding the size of the arrays and logical drives, and converting from one RAID level to another. Finally, it offers implementation and management capabilities for advanced premium feature functions such as FlashCopy, Volume Copy, and Enhanced Remote Mirroring [42]. The Storage Manager software package also includes the required host software components for the specific host environments that are planned to be supported.

The Storage Manager software level is closely tied to the features of the level of the firmware code level that is being run on the subsystem. Newer Storage Manager level are designed to be backward compatible with current firmware levels for previous generations of products as well as earlier versions of firmware for the current product line. The Storage Manager software level is closely tied to the features of the level of the firmware code level that is being run on the subsystem. Newer Storage Manager level are designed to be backward compatible with current firmware levels for previous generations of products as well as earlier versions of firmware for the current product line. Newer firmware levels might require a newer version of the Storage Manager Software.

IBM Tivoli Storage Productivity Center

IBM Tivoli Storage Productivity Center (TPC) is designed to provide various components for a specific needs of services management for DS family systems, ESS, IBM Sun Volume controller, N series storage systems and heterogeneous storage environments. In the focus of this study is the Total Storage Productivity Center for Replication component which is a very useful for managing the replication and copy services: Global Mirror, Global Copy and FlashCopy [43]. It provides tool for monitoring, control and managing of copy services tasks.

In the environments where this module has being used for controlling the process of data mirroring and copy this tool can be used to facilitate the data migration as well. However, due to the complexity and high cost of Storage Productivity Center for Replication in storage environments where it was not yet implemented the most practical will be to use a data migration technique managed by one of the storage interfaces.

Enhanced Remote Mirroring

In this part the concept of remote mirroring of data is described in details. Enhanced Remote Mirroring (ERM) or Remote Mirroring from our point of view is is an important premium feature which can be used as a data migration technique between supported IBM storage systems given in the Appendix B. The ERM technology enables real time data replication between system storages either locally or over a remote distance as it is shown in the figure 4.22. In this context, ERM plays an important role to provide business

continuity in a case of disaster since it has functionally to redirect normal I/O operations to the second system storage which will take over the primary responsibility for them.

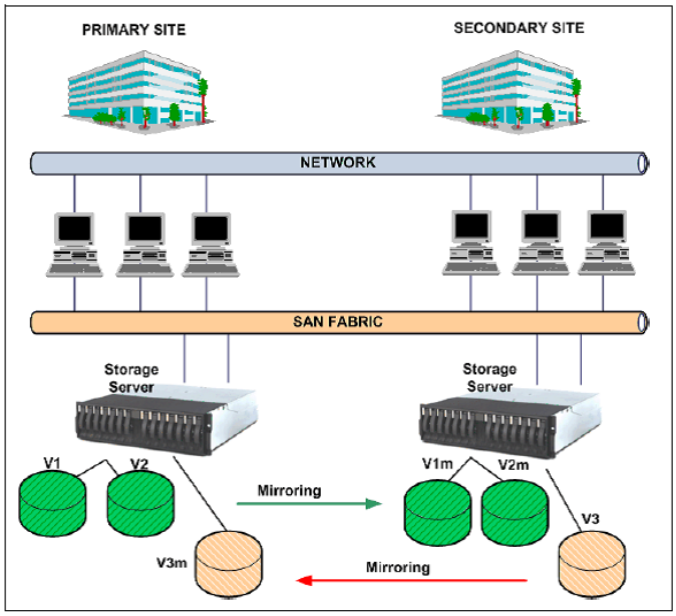

Figure 4.22 Enhanced Remote Mirroring in real storage environment

ERM is a system storage logical drive (LUN) based mirroring capability that is independent of and transparent to host application servers. ERM supports three standard mirroring modes which are known as: Metro Mirroring, Global Mirroring, Global Copy.

Based on the best practices given in [46] IBM Global Mirror is not considered as a practical approach for data migration, but it more intended to be a long distance business continuity solution. The cost and setup complexity of IBM Global Mirror are not typically justified for a one-time data migration project. In this case, Global Copy is much better suited for this task. Also, the IBM Enterprise Storage Server Model F20 is not supported for RMC functions to the DS8000. If migrating from the Model F20, an intermediate step is required to move the data to the Model 800 or 750 before migrating to the DS8000 [46].

Metro Mirroring

Metro mirroring provides synchronous data replication between primary and secondary logical drive of storage systems. This means that the host system when initiates a

write request will not get an acknowledgment that the write has been made to the remote system. The sequence of mirroring process is illustrated in the figure 4.23. When primary controller receives a write request sends this information to the so called mirror repository logical drive and in the same time writes the data to the source logical drive. After that, the write block of data is being copied to the remote site. At the successful completion of this operation, the primary controller removes the log record from the mirror repository logical drive and sends I/O completion information to the host system.

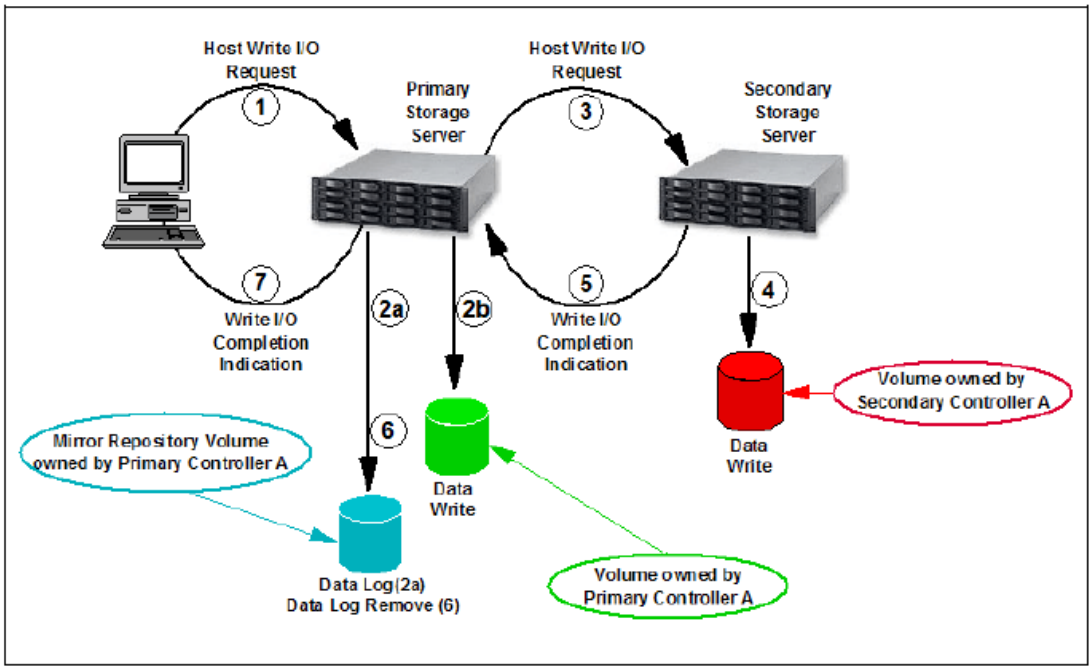

Figure 4.23 Illustrates the process of metro mirroring

Global Mirroring

Unlike Metro Mirroring, Global Mirroring technique performs data replication in an *asynchronous* manner. It this scenario, the primary controller indicates that the write is completed when the write is created on the primary logical disk drive but still it is not sent to the remote site. This mode must be used when the host application has dependent data spread across several logical drives which are being mirrored to ensure that the dependent write requests are carried out in the same order at the secondary remote site. For this purpose, Global Mirror through uses a write consistency group. It is very important for the success of this solution to keep the consistency of logical drives which belong to the group

and are linked to the remote mirrors. When the write request is stored on the primary logical drive and remote copy has been made then the log record from the mirror repository logical drive can be removed. This is shown in the figure 4.24.

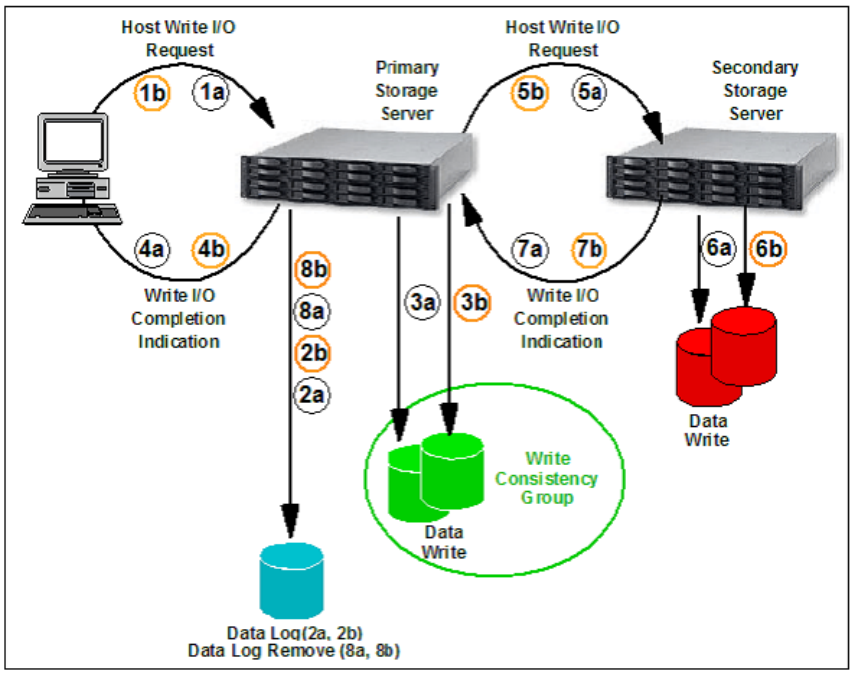

Figure 4.24 Global Mirroring logical data flow

Global Copy

Global Copy as well as Global Mirroring that was explained above performs asynchronous data mirroring to the secondary site but without consistency group. It is called asynchronous because when the write requests received from the host are stored to the primary logical disk drive an indication that the process has been completed is directly sent back to the server. Then the primary controller performs remote copy of the corresponding data blocks to the secondary logical drive. After the data has been copied at the remote site or cached, the primary controller removes the log record on the mirror repository logical drive, figure 4.25.

In a scenario where multiple mirror relationships are established between the primary and secondary controller the background process of coping data blocks are conducted in parallel. Unlike Global Mirroring, Global Copy does not ensure that write

requests at the primary site will be made in the same order. That is the reason why this method is called as *asynchronous mirroring without write consistency group*.

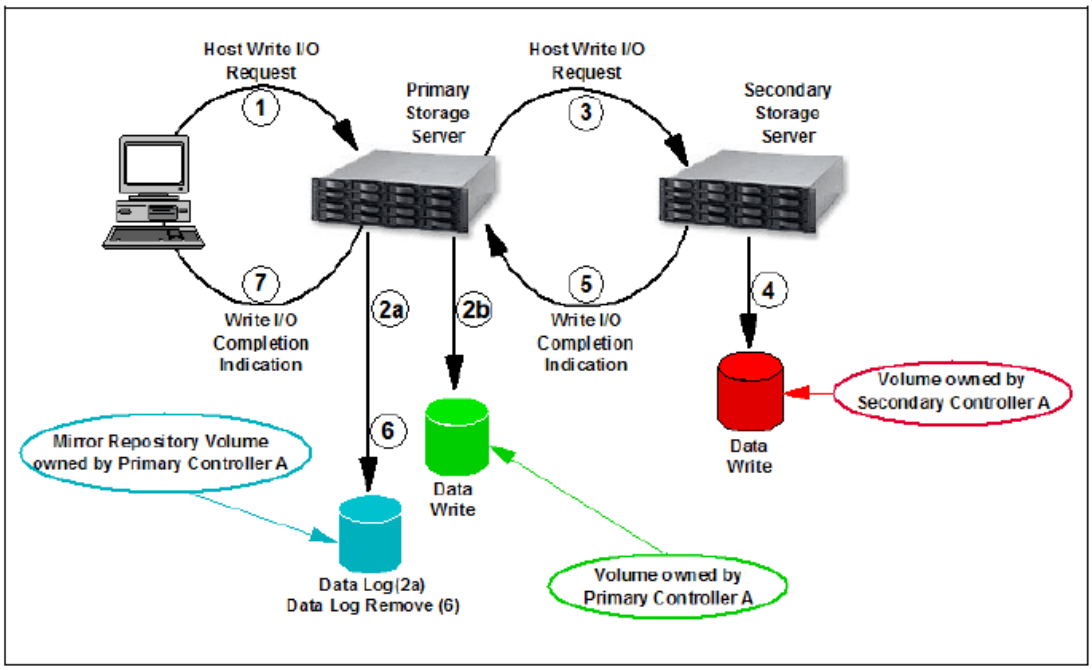

Figure 4.25 Global Copy mode (asynchronous mirroring) data flow

When write caching is enabled on either the primary or the secondary logical drive, the I/O completion is sent when data is in the cache on the site (primary or secondary) where write caching is enabled. When write caching is disabled on either the primary or the secondary logical drive, the I/O completion is not sent until the data has been stored to physical media on that site.

4.2.4 XIV storage system and data migration functionality

IBM XIV Storage System is one of the latest products of high-end disk storage systems designed to meet large enterprise requirements. It is a storage area network system that can be connected via FC or iSCSI protocols to IBM and non-IBM qualified storage systems. Apart from the Remote Mirroring and Copy services that can be used to migrate data between IBM storage systems provided, it offers a new data migration functionality which is embedded in the base of IBM XIV Storage System software [45].

The IBM XIV Data Migration solution offers data transfer to be transparent to a host during the data migration, and data to be available for immediate access. However, it

requires a single short downtime to switch LUN ownership while the connection to the host server is established. From then on, the XIV storage system acts as a host and manages the data migration process. After the connection is established the background copy process is facilitate while the users can access the data. The XIV Storage System is doing a block-by-block copy of data by sending read and write requests to the source storage device, which the XIV is then writing onto an XIV volume. It is important that the connections between the two storage systems remain intact during the entire migration process because XIV system handles the I/O requests and ensures that both storage systems (XIV and non-XIV storage) are updated when a write I/O is issued to the LUN being migrated. By doing this the source system remains updated during the migration process, and the two storage systems remain in sync after the background copy process completes. Similar to synchronous Remote Mirroring, the write commands are only acknowledged by the XIV storage system to the host after writing the new data to the local XIV volume, then writing to the source storage device, and then receiving an acknowledgement from the non-XIV storage device. If at any time during the migration process the communication between the storage systems fails, the process also fails. The simple view of storage environment between IBM XIV and any other storage system is shown on the figure 4.26.

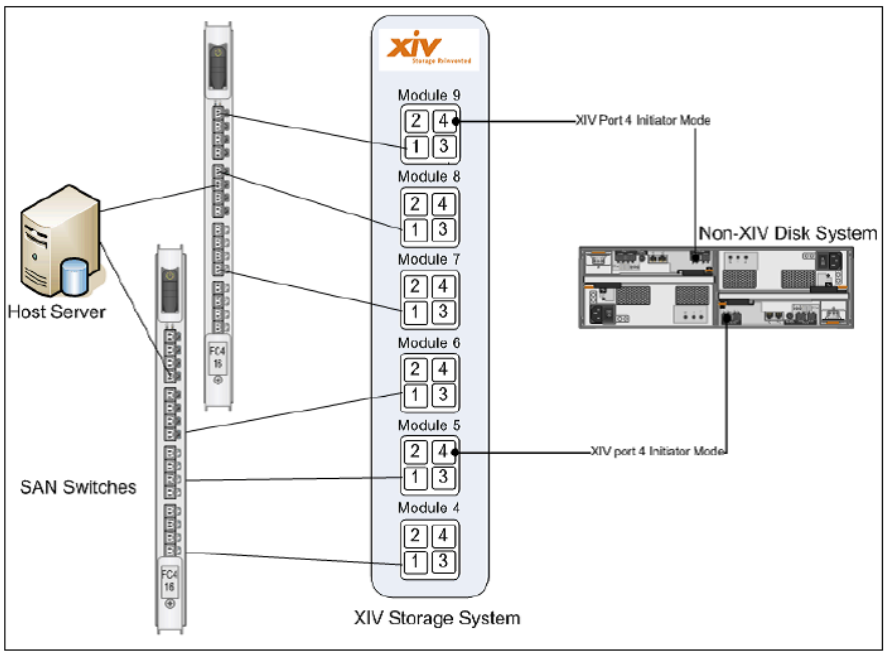

Figure 4.26 Data migration simple view

The data migration facility in XIV firmware revisions 10.1 and later supports up to four migration targets can be configured on an XIV (where a target is either one controller in an active/passive storage device or one active/active storage device). XIV firmware revision 10.2.2 increased the number of targets to eight [45].

The migration process depends on the multi pathing support from the target system. If the non-XIV disk system supports active-active LUN access then multiple paths can be established from XIV to the non-XIV disk system. The XIV includes functionality for balancing the traffic load during the migration across these paths. This might lead to the temptation to configure more than two connections or to increase the initialization speed to a large value to speed up the migration. IBM products that are active-active storage servers are the DS6000, DS8000, ESS F20, ESS 800, and SVC. Special considerations must be made when migrating data from an active-passive storage device such as DS3400 to XIV. In this context, only a single path can be configured between any given non-XIV storage device controller and the XIV system. Due to the single path available for data movement it is commonly used to perform migrations while the host applications are offline which increases the outage.

4.2.5 SAN Volume Controller and Storwize v7000

IBM SAN volume Controller (SVC) and Storwize v7000 will be discussed together under the same subtitle since IBM developers has get together the both systems to share a code. The recently introduced Storwize V7000 storage platform and SVC version 6.1 are products built from a code base of SVC 5.1, with a user interface inherited from XIV, and Easy Tier and a robust RAID implementation inherited from DS8000 [46].

Storwize v7000 provides high storage capabilities which can be compared to the class of Clariion CX series by incorporating SSD drives and provide SAS connectivity. The IBM Storwize v7000 solution provides a modular storage system that includes the capability to virtualize external SAN-attached storage and its own internal storage. Storwize V7000 is a clustered, scalable, and midrange storage system and can be used as an external virtualization device. From the SAN Volume Controller it inherits all functionalities associated to the class of enterprise systems including point-in-time Flash

Copy, Metro Mirror, Global Mirror and Global copy, in-band migration functionality and long distance synchronous and asynchronous replication. Metro Mirror and Global Mirror source volumes can be copied to target volumes on a dissimilar storage subsystem as was described above.

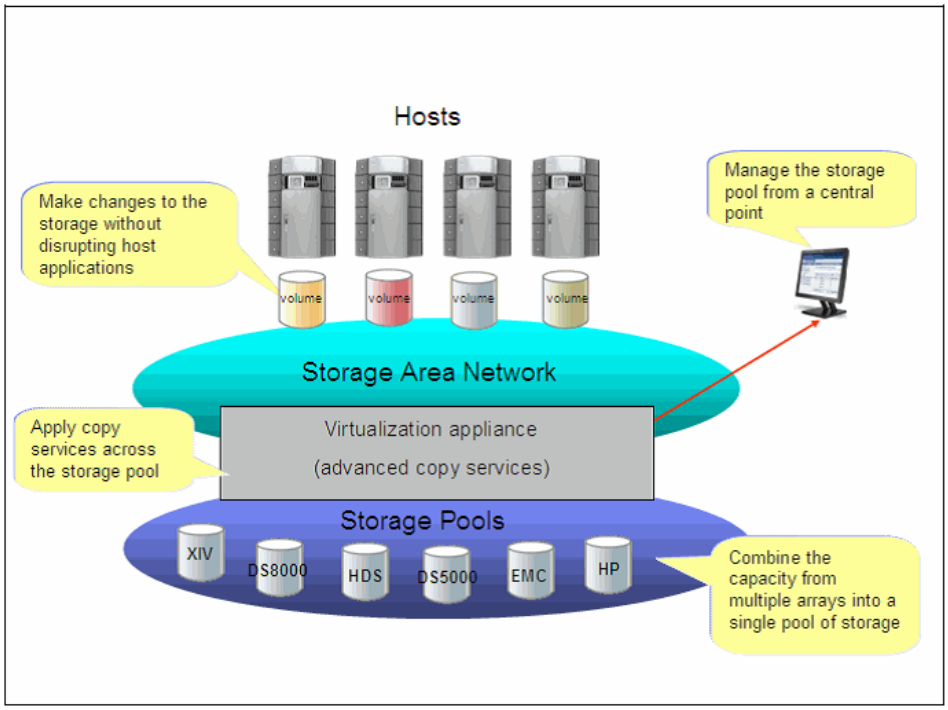

Figure 4.27 Storage virtualization

In the figure 4.27 is depicted a storage virtualization environment where as virtualization appliance apart from Storwize v7000 can be used SVC as well. They actually deliver a single view of the storage attached to the SAN. The physical disks can be managed and migrated nondisruptively even between different storage systems that built the heterogeneous environment. The SAN is zoned in such a way that the application servers cannot see the back-end storage, preventing any possible conflict between the virtualization appliance and the application servers who are both trying to manage the back-end storage.

SAN Volume Controller provides storage virtualization by creating a pool of managed disks from attached back-end disk storage subsystems. These managed disks are mapped to a set of virtual disks (VDisks) and then they are presented to the application servers as LUNs. SVC simplifies device driver configuration on hosts in a way that all hosts within a network use the same IBM device driver to access all storage systems

through the SAN Volume Controller. SVC supports a wide variety of disk storage and host operating system platforms given in the Appendix B.

Data migration using VDisk migration

In this part it is explained how to perform block-level data migration (LUN) in a heterogeneous storage environment including storage platforms from different vendors. VDisk migration uses the SVC ability to copy virtualized extents between source storage and target storage attached to the same SVC where the source for the migration is an image mode VDisk. It must be considered that the destination MDisk must is greater than or equal to the size of the Vdisk so the migration can run.

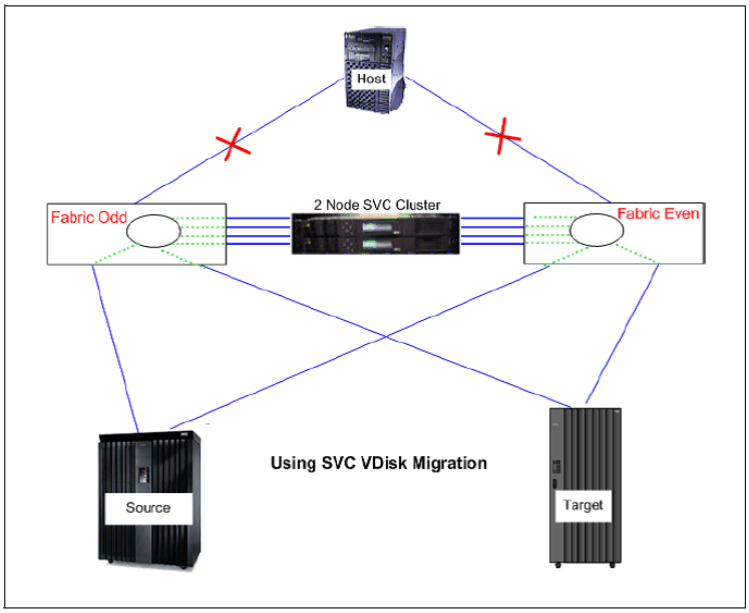

Figure 4.28 Data migration using SVC vDISK migration

The data migration process is shown on the figure 4.28. Data is moving in units of 16 MB so called chunks [46]. First, all new I/O requests on the source MDisk have to be queued in the virtualization layer in SVC and wait for all outstanding requests to be completed. Then, all I/Os of the source MDisk are allowed apart from writes to the specific chunk that is being migrated. Writes to the extent are mirrored to the source and destination synchronously in chunks of 256K.Once the entire chunk has been copied to the destination,

the process is repeating for the next chunk in the extent. When the entire extent has been migrated, all I/Os to the extent being migrated are paused. It has to be checked whether all data has been successfully migrated and then redirect all further reads to the destination, and stop mirroring writes.

Data migration using SVC Metro Mirror

Data migrations between different storage platforms over the SVC can be performed by using Metro Mirror functionality. However, this method is more complicated then data migration when using vDisks. In this case, it is recommended to be used two SVC clusters (minimum four nodes) together with Metro Mirror. The use of a Single Cluster Metro Mirror (intracluster MetroMirror) for a data migration is possible, but this configuration introduces a single point of failure. Any problems occurring in this single SVC cluster or the fabric connecting the environment could potentially affect both the source and target data, especially before the source and target are fully synchronized. This failure scenario can potentially be a very difficult situation to recover from while at the same time maintaining data integrity.

By using Dual Cluster Metro Mirroring (intercluster Metro Mirror), the source volumes will receive the host updates at all times during the data migration. If a problem occurs between the host and its data volumes, a normal recovery is possible. In the background the SVC will be copying to data over to the target volumes, giving the user, in the event of any failure, the option to restart from the source or, if necessary, to restart from the target volumes. In addition, the target LUN should be exactly the same size as the source LUN.

4.2.6 IBM N series products and solutions

IBM N series are product of the OEM relationship established between IBM System Technology Group and NetApp since 2005. Through this agreement, IBM offers the NetApp FAS series products as the IBM N series. By combining the leading storage technologies of IBM N series platforms and the NetApp software portfolio, IBM helps clients to improve their information availability and efficiently manage their storage

infrastructure. The risks involved with data migration, replication, disaster recovery and backup are minimized by supporting solutions from both vendors.

Most specifically IBM N series storage products are designed as stand-alone storage systems. They have support for both the file level access protocolssuch as Network File System (NFS), Common Internet File System (CIFS), HTTP, and iSCSI, and storage area network technologies, such as Fibre Channel (FC). Data protection is achieved by using built-in Redundant Array of Inexpensive Disks (RAID) technologies [48].

As well as the NetApp storage systems, IBM N series run DataONTAP operating systems and supports its futures for data management and control. For the purpose of data migration between N series storages and related NetApp FAS systems can be used SnapMirror that is explained in details in the section "NetApp storage systems and associated software". N series can be attached to the NetApp V series systems and use the mirroring services to perform data migration to any other storage system.

4.3 NETApp Storage systems and associated software

4.3.1 Unified storage systems and associated software

NetApp is the pioneer among storage vendors by offering multiprotocol storage systems in the market since 2002. It has introduced first in the market the well known unified storage platform as an attempt to increase its market share. Today, all NetApp storage systems belonging to the FAS series allow both file access to the storage via CIFS or NIFS and block storage through Fibre Chanel, iSCSI or FCoE. In such a way they can be easily adopt in complex storage environments and thus make the management of the storage infrastructure simpler [33].

NetApp Fabric Attached Storage Systems (FAS) or commonly known as NetApp fillers run native operating system Data ONTAP which provides management and control functionalities from the lower to the highest level described later in this part. All fillers are configured with cache memory called NVRAM which allows a FAS system to commit writes to stable storage quickly, without waiting on disks.

The FAS product line scales from entry level storage systems represented by FAS2000 series, FAS3000 family of midrange storage device up to the high end FAS6000 series systems.

Under the unified storage systems their way has found the V-series storage systems which provide storage virtual tier. The V-Series have been developed by making extension to the FAS controller to support storage LUNs from a storage array connected through a SAN, while at the back-end they aggregates LUNs that are RAID protected by the storage arrays. These systems work as volume controllers and can be used to scale up the management of data across heterogeneous storage environments by supporting storage systems from other vendors such as EMC CX series and IBM DS8000 and DS4000 [50]. As well as FAS series, V-series run Data ONTAP operating system which enables data replication through SnapMirror, to the systems attached to them.

When NetApp V-Series storage system is connected to a qualified storage array via Fibre Channel SAN then LUNs can be provisioned to the V-Series controller as if it were an application host. These array LUNs are collected into a storage pool from which NetApp volumes and LUNs are created and, in turn, allocated from the V-Series controller to application hosts. In other way around, data can be migrated from storage array LUNs to NetApp LUNs and all capacity can be manage behind the V-Series. It should be noted that V-series replicates data between logical volumes. The data migration between physical volumes and logical volumes cannot be performed. A simple storage environment in which V series can operate is given on the figure 4.29.

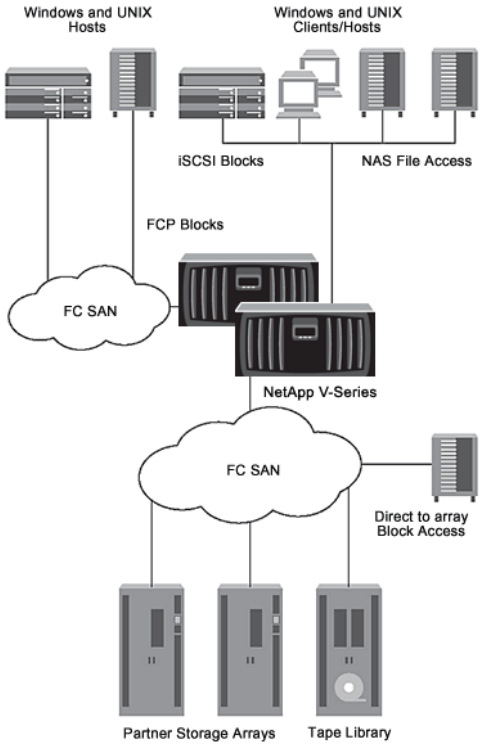

Figure 4.29 NetApp V-series attached to a heterogeneous storage environment (sorce: [50])

4.3.2 Data ONTAP overview

Data ONTAP is originally developed by NetApp and has a look and feel similar to UNIX because it includes code from it. Data ONTAP code is developed in mind to support the unified storage concept model.

In the core of the Data ONTAP operating system is the Write Anywhere File Layout (WAFL) which is a file layout that supports RAID arrays to ensure the highest level of reliability [48]. In contrast to the other approaches which incorporate volume managers, RAID is integrated into the WAFL file system and thus reduces operation errors, operating system and application software release mismatches and patch level mismatches. In addition, WAFL provides functionalities which allow different file systems such as Network File System (NFS) of UNIX and Common Internet File System (CIFS) for Windows operating system and various technologies iSCSI, FC, HTTP or FTP to have an access to the physical disks. It also supports different files on the same volume to have

different security attributes attached to them. For example, UNIX may use either access control lists or a simple bitmask, whereas the more recent Windows model is based on access control lists. These two features make it possible to write a file to a CIFS type of networked filesystem and access it later via NFS from a Unix workstation. The Data ONTAP approach can help to improve the application availability, in that the file system I/O operations are independent of the server's operating system and decrease the downtime of the application host system.

Data ONTAP network interface drivers supports file and block based requests which can be executed in one process opposed to the traditional file servers which employ separate processes for handling network protocol stack, remote file system semantics, local file system, and disk subsystem.

SnapMirror

SnapMirror is one of the Data ONTAP offered functionalities which leverages the unified storage architecture by the easy of management of data replication. In this manner, SnapMirror can be used as a single solution for mirroring data across all NetApp storage arrays and protocols for any application in both virtual and traditional environments in a variety of configurations [48]. Additionally, it enables data replication to the IBM N series as it was mentioned before.

In this study, SnapMirror is considered as technique for data migration which allows volumes or qtree (file system) to be replicated between two qualified storage systems. These systems can be virtually placed in any distance apart as long as the network can provide sufficient bandwidth to transfer the replication traffic. The SnapMirror technology is based on Snapshot technology. Snapshot technology makes extremely efficient use of storage by storing only block-level changes between each successive Snapshot. Only changed blocks are copied after the initial mirror is established.

The Data ONTAP SnapMirror feature can be used in combination with FlexClone volumes to perform migration faster and more efficiently. Briefly, a FlexClone volume is a writable point-in-time image of a logical volume while Snapshot creates statistical point-in-time image.

To replicate data for the first time, the storage system transfers the active file system and all Snapshots from the source volume to the mirror. After the storage system finishes transferring the data, it brings the mirror online. This version of the mirror is the baseline for future incremental changes and, like any other volume, after you finish you can export the mirror for Network File System (NFS) mounting or add a share corresponding to this volume for Common Internet File System (CIFS) sharing. To make incremental changes on the mirror, the storage system takes regular Snapshots on the source volume according to the schedule specified in the configuration file as illustrated in figure 4.30. By comparing the current Snapshot with the previous Snapshot, the storage system determines what changes it must make to synchronize the data in the source volume and the data in the mirror. The destination volume is available for read-only access, or the mirror can be *broken* to enable writes to occur on the destination. After breaking the mirror, it can be re-established by synchronizing the changes made to the destination back onto the source file system.

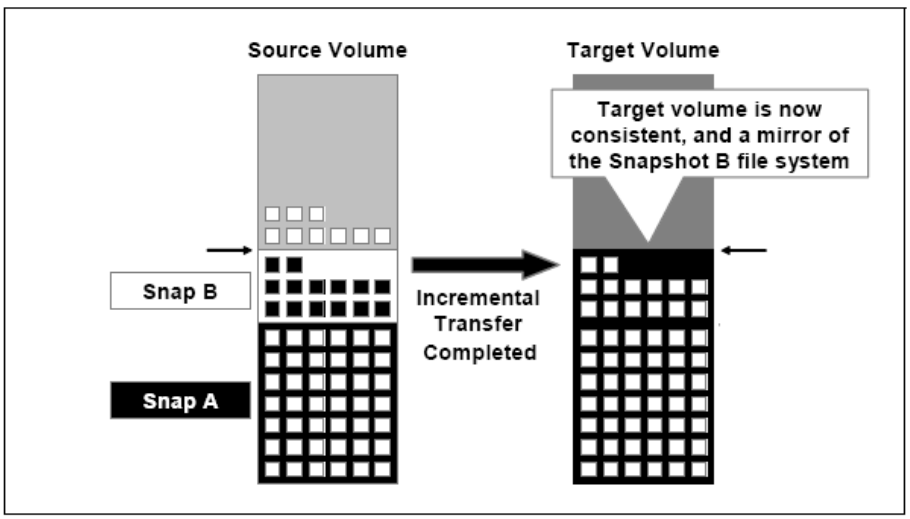

Figure 4.30 SnapMirror architecture

SnapMirror can be used in three different modes:

Asynchronous mode: In the traditional asynchronous mode of operation, updates of new and changed data from the source to the mirror volume occur on a schedule defined by the

storage administrator. These updates could be as frequent as once per minute or as infrequent as once per week, depending on user needs.

Synchronous mode: This mode is also available, which sends updates from the source to the destination as they occur, rather than on a schedule. If configured correctly, this can guarantee that data written on the source system is protected on the mirror volume, even if the entire source system fails due to natural or human-caused disaster. In addition to a standard SnapMirror license, the synchronous feature requires a special license key.

Semi-synchronous mode: This can minimize loss of data in a disaster while also minimizing the performance impact of replication on the source volume. In order to maintain consistency and ease of use, the asynchronous and synchronous interfaces are identical with the exception of a few additional parameters in the configuration file.

More than one physical path might be required for a synchronous mirror. Synchronous SnapMirror supports up to two paths for a particular relationship. These paths can be Ethernet, Fibre Channel, or a combination of the two. Multipath support allows synchronous and semi-synchronous traffic to be load-balanced between these paths and provides for failover in the event of a network outage.

4.4. E-series storage system

As IBM has N series storage system abandon from NetApp and EMC has bought Islion storage devices this attitude continues with NetApp who has acquire Enginio storage system this spring [33]. As a result the new OEM products- the Enginio external storage systems has been renamed in the E-series platforms. With this acquisition NetApp has left their well known operating philosophy for offering unified storage systems in a single production line.

In particular, each family of storage devices offers unique functionalities and different architecture developed with a specific purpose in mind. The E-series product line consisting of E7900, E5400, and E2600 storage systems which are optimized for high-performance computing applications such as genomics sequencing and scientific research, provide large volume sizes of nearly 14 exabytes and deliver I/O burst rates of nearly 800,000 IOPS [51] .

However, currently the E-Series platform does not support the NetApp native Data ONTAP operating system and its rich management features. The management of E-series storage devices is conducted via SANtricity software which is the LSI external RAID SAN management software. In fact this software can be found in the core of others management software's in the industry such as 'IBM DS Storage Manager'. Among data availability, protection and problem resolution it does support NetApp Snapshot technology to create point-in-time copy of data, Volume Copy that creates a complete physical copy (clone) of a volume in a storage system and Remote Volume Mirroring for data replication in synchronous or asynchronous mode.

NetApp E-Series systems can be attached to V-Series storage devices in a SAN architecture and thus gain access to Data ONTAP's management and efficiency features.

6. Conclusions

The main goal of this thesis is to developed application for selecting an ideal solution for data migration in a heterogeneous storage environment. In this chapter a conclusion is given based on the analysis performed on the currently available storage products and services followed by additional guidelines for future.

6.1 Conclusion

This thesis is focusing on exploring the data migration process between storage arrays coming from the same or different vendors, identifying the key factors that directly affect the migration of data, and successively developing an application to quickly provide a migration solution for a given storage environment.

From analysis performed can be concluded that the data migration process is not straightforward. Indeed, it is very *complex task* that requires careful preparation and planning. All iformation regarding the storage array, network performance and application or host environment must be collected prior the migration.

By deaply exemaning the storage systems and associated software that provide functionalities and futures for data migration from the three leading vendors: EMC, IBM and NetApp can be observed a *lack of interoperability* between storage products. Storage vendors try to *lock-in their customers* by providing solution related only to their product lines by making the moving of data very tedious and risky process to be undertaken. It is very important before migrating to verify the support with the storage vendor. Vendor's storage support matrix must be reviewed carefully. These matrixes are usually listed on the storage vendor's website such EMC, or may be only available to the vendor employess, representatives or owners of their platforms which was a case with NetApp . This support matrixes would highlight which systems are supported with their storage device.

As was explained in second chapter, there are three possible data migration approaches based on the utilities they use for data migration which are: array-based, host-based and network-based. Each of them has their own adventages and disadventiges. Array- based data migration come together as a special functionality of storage operating environment and thus not use the CPU of the host system. They are host operating system

independent and perfom the whole migration transparently from the host. Hence, they provide capabilities for nondistruptive data migration and are appropriate for scenarios where mission critical data has to be moved from one platform to another. Nevertheless, the big disadventage of an array-based replication is its lack of support of heterogeneous storage systems. For example, IBM DS3000/DS400/DS5000 and DS800 series does not provide array-based replication to non-IBM systems. NetApp has not even developed a special tool to support migration to third party storage arrays. In this context, EMC offers array-based solutions to facilitate data migration to IBM storage devices through Open Replicator and SAN copy. However, these solutions are limited in a way that the conrol array must be from Clarion family or Symmetrix DMX, VNX and VMAX series. Moreover, array-based replication does not always support data migration between storage systems from the same vendor product line. For instance, it is not possible to migrate data from Symmetrix to Centera by using Open Replicator.

In the family of array-based data migration approaches also belong the mirroring technologies such as SRDF and MirrorView from EMC, SnapMirror from NetApp, and Metro/Global Mirror from IBM. Almost, all vendors of midsized to high-end arrays provide a synchronous and an asynronous replication feature. Even though these replication products are similar in many aspects, a close technical analysis reveals subtle differences. For instance, the efficiency of the handshake between primary and target storage systems used during synchronous replication greatly impacts the distance a replication product can support. Also, they come with some other constraints. Replication technologies can be used only between systems which belong to the same family or the same series. For example, mirroring of data between IBM midrange storage system DS4000 and high-end storage array DS8000 which is from the same IBM DS family is not supported. The benefits of using replication technologies are: they are independent form the host operating system and do not consume host resources.

Furthermore, the network-based or appliance-based products are most commonly used in heterogeneous storage environments. When array-bassed solutions cannot be applied, network-appliance based technologies may find their way. IBM SVC, EMC Rainfinity and NetApp V-series do provide facilities by storage virtualization to migrate data between qualified systems that can be attached to them.

Another common migration approach is host-based solution which is most suitable for heterogeneous storage environments. Almost all operating systems provide native volume mangers capabilities for data migration. However, the host-based solutions may be limited in which device drivers can coexist on the host operating system. There are three important things tha have to be considered before migration: the operating system drivers, microcode levels and multhipathing dependencies.

At last, if the storage environment cannot match any of the exsisting migration solutions the following host operating system methods can be used: direct copy, backup and restore and dump and restore commands. The downside of these approaches is the time that takes to perform the migration, esspecialy when large files are coping, and thus require application outage.

6.2 Support Matrix

All relationships between storage systems, host operating systems, network appliance and services for data migration, replication and copy are based on the support matrixes for compatibility and interoperability specifically created for this application to cover what have been mentioned in previous chapters to be involved in this study. The interoperability matrix identifies important support and compatibility considerations with respect to various host systems and adapters. They cover a wide spectrum of support checks such as supported operating systems, patch levels, required Fibre Channel (FC) adapter firmware, supported SAN switch types/firmware, and much more.

These support matrixes are based on the interoperability matrixes found on the official storage vendors web sites and are given in Appendix B of this document.

6.3. Future work

Further work in this area encompasses several aspects, outlined as following:

This application for selecting a data migration solution is localy based and needs an appropriate installation to be followed. For the future use, it is more sutable to be upgraded to Web based application with centralazied data base where all information are stored and

managed. This will provide access to all knowledgable personen or any person looking for an ideal data migration solution from every place and in any time in the world.

The current version of the application includes storage systems and associated software from the top-tier vendors in the world namely EMC, IBM and NetApp. It is a great opportunity this list of storage vendors to be extended with products from other vendors such as Hitachi, Hewlet Packkward, Dell, Futjitcy and many more. There are numbers of other third-party solutions or open source tools that can be considered as a possible migration technology which are not included in the first release.

The application for selecting an ideal data migration solution considers a basic heterogeneous storage environment. More pricesily, it has been considered a data migration only between two storage platforms and the operating system on which the host work. However, since the real storage environments are more complex, usually configured with more than two storage platforms, storage libraries and swiches to connect them in a dedicated storage area should be added into the application as well. The performance of the switches and their capabilities may additionaly improve the overall data migration process. Furthermore, one of the most important pieces of the puzzle is the required multi-path I/O device drivers and recommended FC adapter firmware or microcode levels that can be taken into consideration as well.

The migration software and operating system environments under control come in different releases. Each version of the software offeres functionalities which may not be supportive by its previous version or it may have been upgraded with new features in order to improve its capabilities. The native code of the software is closely tight to the specific functionalities it provides for data migration. Bassically, the versions of the operating systems of both storage platform and host, as well as the migration solution release should be involved to give more accurately results.

In order to provide more valuable information to the end user, the result may be exported in other document type such as pdf along with information about the installation process of the selected tool and the most critical points while migrating data.

References:

[1] Sumasundaram, G. and Shrivastava, A. (2009). *Information and Storage Management.* Indianapolis (Indiana): Wiley Publishing

[2] Goodwin, P. (2011). *Solving Storage Headaches: Assessing and Benchmarking for Best Practices.* Cognizant

[3] Hopkinton, M. (28.06.2011). *World's Data More Than Doubling Every Two Years—Driving Big Data Opportunity, New IT Roles*. EMC Press Release, available at: http://www.emc.com/about/news/press/2011/20110628-01.htm

[4] EMC Education Services (2006). *Data Migration Solutions Design Concepts*

[5] Howard, P. (2007). *Data Migration*. A White Paper by Bloor Research

[6] Howard, P. and Potter, C. (2007). *Data Migration in the Global 2000*. A Survey Paper by Bloor Research

[7] Fried-Tanzer, D. (2010). *Data Migration EMC: Open Replicator for Symmetrix, Power Path Migration Enabler and Federated Live Migration.* Version 2.0: EMC Corporation

[8] IBM Global Technology Services. (2007). *Best Practices for Data Migration*. USA

[9] Fried-Tanzer, D. (2010). *Choosing a Data Migration Solution for EMC Symmetrix Arrays.* Version 2.1: EMC Corporation

[10] Howard, P. (2008). *Zero-downtime Migration*. A White Paper by Bloor Research

[11] Morris, J. *Practical Data Migration*. Chapter 1

[12] Howard, P. (2011). *Data Migration*. A White Paper by Bloor Research

[13] Howard, P. and Potter, C. (2007). *Data Migration in the Global 2000*. A Survey Paper by Bloor Research

[14] NetApp Global Services. (2006). *Data Migration Best Practices*. USA

[15] Data Migration pro. (26.03.2009). *The data migration go-live strategy – what is it and why does it matter?*. Available at: http://www.datamigrationpro.com/data-migration-articles/2009/3/26/the-data-migration-go-live-strategy-what-is-it-and-why-does.html

[16] Howard, P. (2009). *Business Centric Data migration*. A White Paper by Bloor Research

[17] JPMorgan Chase Research. (2007). *Data Migration Considerations: A Customer Engineering Residency.* EMC Corporation

[18] IBM Global Technology Services. (2007). *Hidden Costs of Data Migration*. USA

[19] Dufrasne B., Seiwert, C., Hazzard, P., Laing, C., Martinez,L., Sedgwick ,J., Strubel E., Werley, T. and Drumm H.P.(2007). *Migrating to IBM System Storage DS800.* IBM Redbooks

[20] Wendt, J. (06.2004). *Pros, cons of host-based technology for data migration*. SearchStorage. Available at: http://searchstorage.techtarget.com/tip/Pros-cons-of-host-based-technology-for-data-migration

[21] Gsoedl, J. *Array-based and network-based replication*: Storage Search. Available at: http://searchdisasterrecovery.techtarget.com/tip/Data-replication-strategies-Array-based-and-network-based-replication

[22] Arian A. (10.11.2008). *Host-based replication vs. array-based replication for backup and disaster recovery*: Storage search. Available at: http://searchstorage.techtarget.co.uk/news/1338351/Host-based-replication-vs-array-based-replication-for-backup-and-disaster-recovery

[23] Gsoedl, J. *The pros and cons of network-based data replication*: Storage search. Available at: http://searchdisasterrecovery.techtarget.com/tip/The-pros-and-cons-of-network-based-data-replication

[24] Howell, D. (12.09.2007). *A primer on array-based and network-based replication*: Tech Republic. Available at: http://www.techrepublic.com/blog/datacenter/a-primer-on-array-based-and-network-based-replication/175

[25] Shrivastava, A. and Sumasundaram, G. (2011). *Managing Information Storage: Trends, Challenges and options 2011-2012* .USA: EMC Corporation

[26] Howard, P. (2008). *Market Update*. Bloor Research

[27] Biztech2 staff. (08.06.2011). *ECB Disk Storage Market Grew 14.1% In Q1FY11.* Available at: http://biztech2.in.com/news/storage/ecb-disk-storage-market-grew-141-in-q1fy11/110042/0

[28] Symmetri presentation picture

[29] (09.2011).EMC Simple Support Matrix EMC Open Replicator EMC. Available at: Powerlink

[30] (09.2011).EMC Simple Support Matrix EMC Federated Live Migration EMC. Available at: Powerlink

[31] EMC Storage products. *Technology specifications*. Available at: http://www.emc.com/products/category/storage.htm. Last visited: 10.09.2011

[32] IBM System Storage products. *Technology specifications*. Available at: Last visited: http://www-03.ibm.com/systems/storage/?cm_re=masthead-_-products-_-stg-allstorage. Last visited: 10.09.2011

[33] NetApp Data Storage Systems. *Technology specifications*. Available at: http://www.netapp.com/us/products/storage-systems/. Last visited: 10.09.2011

[34] Panchigar, D. (06.01.2006). EMC Clarrion FLARE Code Operating Environment. *Available at: http://storagenerve.com/2009/01/06/emc-clariion-flare-code-operating-environment/*

[35] EMC White paper. (10.2010).*EMC SAN Copy: A Detailed Review*

[36]EMC White paper.(01.2011). *Migrating data from an EMC Clariion array to a VNX platform using SAN Copy*

[37] Dharma, R. Lane, D. Hughes, D. *Networking for Storage Virtualization and EMC RecoverPoint.* Version 2.0. EMC Techbook

[38] EMC White paper. (11.2010). *File Archiving from EMC Celerra to data domain with EMC File Managament Appliance*

[39] EMC White paper.(03.2011). *Migrating data from an EMC Cellera array to a VNX platform using Celerra Replicator*

[40] IBM Systems and Technology Group. (07.2011). *IBM System Storage Product Guide:* USA. Available at: ibm.com/storage

[41] Racherla S., Allworth B., Bagnaresi A., Bogdanowicz C., Lottering C., Pedrazas P., Schubert F., Sexton J., Watson A., (03.2010). *IBM Midrange System Storage Implementation and Best Practices Guide:* Redbooks. Available at: ibm.com/redbooks

[42] Racherla S., Aghdam R.F., Lonzer H., O'Neill L.G., Rodriguez M.,Sindelar V.,Watson A., (21.12.2011). *IBM System Storage DS Storage Manager Copy Services Guide:* Redbooks. Available at: ibm.com/redbooks

[43] IBM System Storage DS4000 Storage Manager Version 9.23. (04.2007). *Copy Services User's Guide: Redbooks, Sixth Edition.*

[44].Dufrasne B., Lunardon M., Watson A., Youngs B. (01.2006). *DS4000 Best Practices and Performance Tuning Guide:* Redbooks. Available at: ibm.com/redbooks

[45] Bert Dufrasne B., Eriksson R., Gard t W., Jamsek J., Nause N., Oscheka M., Saba C., Tsypin E., Wagner K., Warmuth A., Westphal A., Wohlfarth R. (10.2011). *IBM XIV Storage System Copy Services and Migration:* Redbooks. Available at: ibm.com/redbooks

[46] Dufrasne B., Seiwert C., Drumm H., Hazzard P., Laing C., Martinez L., Sedgwick J., Strubel E., Werley T.(2007). *Migrating to IBM System Storage DS8000:* Redbooks. Available at: ibm.com/redbooks

[47] Cartwright Brian, Hruby R., Koeck D., Liu X., Rosati M., Vogel T., Wiegand B., Tate J. (05.2011). *Implementing the IBM Storwize V7000:* Redbooks. Available at: ibm.com/redbooks

[48] Osuna A., Javier R. F.(06.2010). *IBM System Storage N series Software Guide:* Redbooks. Available at: ibm.com/redbooks

[49] EMC Data Sheet. *EMC VPLEX Family.* Available at: www.emc.com

[50] Lusnia S. (04.2007). *Simplifying Data Management for EMC, HP, IBM, Fujitsu, & Hitachi Environmentt: Tech OnTap. Available at:* http://partners.netapp.com/go/techontap/matl/v-series.html

[51] http://www.ntapgeek.com/2011/05/netapp-reveals-new-e-series-platform.html http://communities.netapp.com/docs/DOC-9649#comment-4745 data motion

Support matrixes used:

[Appendix B] IBM Systems Support. *System Storage Interoperation Center (SSIC).* Available at: http://www-03.ibm.com/systems/support/storage/ssic/interoperability.wss. Last visited: 10.09.2011

[Appendix B] (09.2011).EMC Simple Support Matrix EMC VPLEX and GeoSynchrony: EMC. Available at: Powerlink

[Appendix B](09.2011).EMC Simple Support Matrix EMC Symmetrix VMAXe: EMC. Available at: Powerlink

[Appendix B] (06.2011) EMC Support Matrix esm by os: EMC. Available at: Powerlink

[Appendix B] (09.2011).EMC Simple Support Matrix EMC Symmetrix Invista: EMC. Available at: Powerlink