BACKUP CONSISTENCY– A GAME OF CHALLENGES

Mohamed Sohail
Data Protection & Availability Specialist
EMC

Sameh Gad
Senior Consultant
EMC

Emanuela Caramagna
Advisory Systems Engineer
DPAD

Denis Canty
Principal Architect
xGMO

Table of Contents

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect EMC Corporation’s views, processes or methodologies.

Backup Consistency Prayer

Oh Lord Forgive me, for I have sinned.

I have sacrificed consistency to get better benchmark numbers.

I have written distributed systems in languages prone to race conditions and memory leaks.

I have failed to use model checking when I should have.

I have failed to use static analysis when I should have.

I have failed to write tests that simulate failures properly.

I have tested on too few nodes or threads to get meaningful results.

I have tweaked timeout values to make the tests pass.

I have implemented a thread-per-connection model.

I have failed to monitor or profile my code to find out where the real bottlenecks are.

I know I am not alone in doing these things, but I alone can repent and I alone can try to do better. I pray for the guidance please give me the strength to sin…….. No more.

Amen.

Abstract

One of the modern challenges in Data Centers is The (Backup Consistency) and how to automate the backup validation criteria. It is really a game of challenges.

With nowadays-large data centers and the huge amount of production data that exists on most modern business applications, one of the pains is: will I be able to recover the data easily? Will it be the exact thing I pushed to the backup system? To test and verify their backups, customers build long procedures to restore and ensure the production backup sets are always useable and consistent. While this process is quite healthy, it requires large investments and quite some time. On the other hand it may not - sometimes - fulfil the targeted accuracy level.

In this article we suggest that the traditional techniques of “Data Protection status reporting” should include some automation for backup consistency checking (and possibly verification as well). Thus your backup systems dashboard should be able to show information (backup time, recovery time options, Age, etc.), controls (like start backup consistency checking), and recommendations (based on backup meta-data like the change percentage in backup objects). Those controls may allow launching a series of recovery and verification scripts for each of the protected information assets. The recommendations and tuning options may point to a low change information asset to have its backup frequency or level reconfigured, or to a recommended archival action to move some of the data from active tier to archive tier on a Data Domain® host for example.

The suggested backup consistency checking mechanism includes a basic set of scripted steps that would automate the recovery of the data to some virtualized host system (in cases where virtualization is available) succeeded by a suitable data verification tool (like hash checking) on the recovered data.

Benefits:

- Minimize data loss risks
- Minimize used spaces and maximize backup performance
- CapEx/OpEx will be reduced due to the full backup process optimization

Introduction

In designing any IT solution, backup is normally one piece in the overall design. Many administrators will verify that their technology (both hardware and software) can handle the data load, ensure that the timelines are appropriate to the business, and then move to the next design consideration for the solution. There are cases where the administrator will do some off-site verification for that node. But, it is unusually rare for the admin to simulate a complete disaster, with the loss of a single or multiple nodes.

In a large percentage of data or service loss use cases, it does not affect the business interoperability to a great extent. However, in the small chance of a disaster occurring, the unknown is often too much, with too many questions, and not enough heuristic data to assist in bringing everything back in the correct sequence. Planning is naturally very important, but it is often the case that these simulations for recovery are very time consuming, and return on investment (ROI) of people hours becomes a hurdle in ensuring an effective backup consistency model is known and understood.

There are a number of backup consistency types, and it is important to understand the differences in these when in the planning phase. Let’s start with the oldest. An inconsistent backup is when backup software starts at the beginning of the file system, and copies all the data until the end. If any file changes during this process, then the result is inconsistency in the data. These work OK for file systems that do not change often, but are not acceptable for database type applications.

Building on this, a crash consistent backup is when the backup data is grabbed at exactly the same time, and is very popular for any application that does not rely on a database. When a crash consistent backup is restored, the data is in a mirror state to the time of the backup.

An application consistent backup builds on the features and method of the crash consistent backup, but critically, it will also capture any in memory data and transactions and execute these prior to the backup. Once this backup is then complete, the controller will then notify the database to resume service.

Challenges

Traditional backup is becoming more challenging and less practical for many administrators. With many types of backup and backup consistency models available as a toolset to the user, trying to find a “one size fits all” for their IT applications is becoming more and more complex and time consuming.

The more granular needs of the organization are often overlooked, and the more of these that are considered, increases complexity. Perhaps the user only wants to back up some databases on and off site, or that they want a different strategy for their databases compared to their VMs? Corporate policy and data governance are also adding further requirement and load into ensuring that the correct backup consistency strategy is found.

Many end users would like these issues resolved, and not just in the singular sense, but across the application requirement change landscape that is accelerating today. Not only are their application requirements changing, but also the sheer volume of data is expanding at a near flood rate. As these volumes are increasing, classical backup consistency models are being put under more strain.

Agility within backup consistency is not very apparent in design phases, which has long term implications. As application and database types change rapidly, can our backup consistency strategy handle this?

Clearly there is a big win from introducing some level of automation into the classification, verification, and reporting of backup consistency models to ensure the end goal: consistently recover the exact snapshot that was pushed for backup in a timeline suitable to the business needs.

It is proposed that automation scripting at key stages such as consistency checking, reporting, and tuning could assist in solving some of these challenges. As is the case with many solutions, this alone introduces concern points for consideration. How do we trust this automation? Will it deliver a better accuracy level than more manual methods? How is failover handled?

Case Study: Real Telco User Example

Cluster Hayper-V (Windows Server 2012) Image Level Backup Example

Scenario: “XXX backup product Plug-in for Hyper-V VSS: Hyper-V backup may not complete or complete with exceptions after applying Microsoft Windows Server 2012 R2 Update KB2919355”

Impact Severity Rating: Critical (Data Unavailable / Data Loss)

Impact Description: XXX backups of Hyper-V virtual machines on Microsoft Windows 2012 R2 Cluster Shared Volumes (CSV) may report the status of ‘Failed’ or ‘Completed with Exceptions’ and may result in backup(s) to be not available for restore.

Issue To verify if you are impacted by this issue:

1) Find the problematic backups by navigating to the Activity Monitor in the XXXX Administrator GUI (Graphical User Interface). If you see end status of the backup either as ‘Failed’ OR ‘Completed with Exception’ associated with a Windows Hyper-V VSS Plugin on a Windows Server 2012 R2 CSV Server with Update KB2919355, then you must apply the new Microsoft update(s) as described in the resolution section.

If you need additional verification that the ‘Failed’ or ‘Completed with Exceptions’ backups are associated with this issue:

2) Within the Hyper-V VSS plug-in log associated with the ‘Failure’ OR ‘Completed with Exception’ status, under the "Errors and Exceptions" section, error messages will appear for two main scenarios. The first scenario is when the backup error(s) has occurred on the primary proxy node. The second scenario is when the backup error(s) has occurred on the secondary proxy node.

Scenario 1: Errors on the primary proxy node

Errors on the primary proxy node will contain Hyper-V Plug-in error messages similar to the following : (Double-Click the backup work order entry in the Activity

Monitor):

```
2014-06-10 23:47:14 Error <7553>: Specified source path
"C:\ClusterStorage\<VOLUME>\<VM>\Snapshots\0023D52C-99FC-4314-9E5E-
68E0419A8EEB.xml" does not exist; ignored (Log #1)
2014-06-10 23:47:14 Error <7553>: Specified source path
"C:\ClusterStorage\<VOLUME>\<VM>\Blank Dynamic Data Virtual Hard
Disk_ACBB2BA0-37B6-46CF-A6DF-26C78576EE13.avhdx" does not exist;
ignored (Log #1)
2014-06-10 23:47:14 Error <7553>: Specified source path
"C:\ClusterStorage\<VOLUME>\<VM>\co-ps-w08e64-t01_disk_1_ACBB2BA0-
37B6-46CF-A6DF-26C78576EE13.avhdx" does not exist; ignored (Log #1)
2014-06-10 23:47:14 Error <7553>: Specified source path
"c:\clusterstorage\<VOLUME>\<VM>\Snapshots\ACBB2BA0-37B6-46CF-A6DF-
26C78576EE13.xml" does not exist; ignored (Log #1)
2014-06-11 00:24:38 hypervvss Error <13117>: Unable to successfully process
backup workorder MOD-1402461193445#1, targets
C:\ClusterStorage\Volume4\co-ds-cim-w02\Virtual Machines\4D71445E-94FD-
4D4A-A039-7B1DC1CB8107.xml (Log #2)
2014-06-11 00:24:38 hypervvss Error <0000>: Failed to process local backup
workorder (Log #2)
2014-06-11 00:24:38 hypervvss Error <17206>: Calling backup complete with
failure. (Log #2)
2014-06-11 00:25:20 hypervvss Error <0000>: Multi proxy backup did not complete
successfully. (Log #2)
```

Where:

<VOLUME> - Is the name of a disk volume, for example "Volume4."
<VM> - Is the hostname of a virtual machine being backed up.

Scenario 2: Errors on the secondary proxy node(s)

Errors which occur on the secondary proxy node(s) will exhibit only Hyper-V Plug-in messages which will appear similar to the following (Double-click the backup work order entry in the Activity Monitor).

```
2014-06-10 02:15:18 hypervvss Error <0000>: The plugin on remote client <HOSTNAME_1> (<IP ADDRESS_1>) terminated with code 10020: Completed with errors, client log should be examined (Log #2)
…
2014-06-10 02:23:49 hypervvss Error <0000>: The plugin on remote client <HOSTNAME_N> (<IP ADDRESS_N>) terminated with code 10020: Completed with errors, client log should be examined (Log #2)
2014-06-10 02:43:49 hypervvss Error <0000>: <N> remote client(s) failed to complete backup of the remote file targets (Log #2)
2014-06-10 02:43:49 hypervvss Error <17206>: Calling backup complete with failure. (Log #2)
2014-06-10 02:45:01 hypervvss Error <0000>: Multi proxy backup did not complete successfully. (Log #2)
```

Where:
<HOSTNAME_1> ... <HOSTNAME_N> - Are the respective hostnames associated with the secondary proxy node(s).
<IP ADDRESS_1> .... <IP ADDRESS_N> - Are the respective IP addresses associated with the secondary proxy node(s).
<N> - Is the number of proxy node(s) involved in the backup.

To find the associated log file, directly access the secondary proxy node and navigate to the "<xxx Installation Directory>"\var\, where the default location for <xxx Installation Directory> is "C:\Program Files\avs."

Review the log file which is in the following format: MOD-<WORKORDERID>-3032-Hyper-V_VSS#<NUM>.log.

Where:
<WORKORDERID> - Is the backup work order associated with the backup activity.
<NUM> - Is the identification for primary and secondary proxies. For primary proxy node, the number will be "1," while for secondary node(s) the number will be "101",

"102"

Within the log file, users may find errors similar to those shown in "Scenario 1".

Environment

xxx Software: xxx 7.0.X Plug-in for Hyper-V VSS
System: Microsoft Hyper-V
Operating System: Microsoft Windows 2012 R2
Application Software: Microsoft Volume Shadow Copy Service (VSS)
Feature: Microsoft Cluster Shared Volumes (CSV)

Cause

Hyper-V backups either 'Fail' OR 'Complete with Exceptions' after installing Microsoft update KB2919355.

Resolution

Follow the directions for applying the Microsoft updates as defined in the Best Practices for Hyper-V over CSV Cluster Data Protection Using xxx and tech-note.

Here we can find that without having a backup consistency **methodology** and policy we might see the backups are completed and they are not recoverable. In this case we need to have a predefined policy for the backup consistency and also a method for validation. First of all, we need to have a consistent backup then a way to check the consistency of this backup. This should be proactive and driven by the system administrator and planners in the data protection and availability department.

Challenge 1: Backup consistency policy

Setting up backup infrastructures for large-scale data management systems that can be operated cheaply and accessed with low latency has emerged as a practical problem. As a solution, we present in the next few pages a real example of a typical large telecommunication and ISP provider in Europe. It is an example for a highly scalable and cost-efficient architecture for backup management in a distributed file system. We describe techniques for the creation of consistent backups at runtime, as well as approaches to resource management in connection with an integrated backup architecture.

In recent years, the management of huge data sets has become an important topic for research and industry. Systems of our customer example generate data in the range of multiple petabytes per year, and have a large number of data centers across the globe with enormous amounts of data. Data volumes at such a scale require specific storage, provision, and backup solutions.

A widely-adopted approach to deal with huge data volumes are hierarchical storage management architectures: frequently used data is managed by upper storage tiers that are backed by quickly accessible storage devices like disks, whereas rarely used data is stored by lower tiers, which are generally backed by long-term storage devices with higher access times, such as tape libraries. Backups are generally managed by lower tiers, since they can store large data volumes in a cost-efficient manner.

Although tape libraries can be operated cheaply, long access latencies and seek times limit their applicability to incremental backups. We present the design and research perspective of a disk-based system for data storage that is capable of maintaining snapshots and incremental backups of large data sets at a low cost and will be easier to check the data consistency. The system guarantees immediate and fast access to former versions of the managed data, while ensuring that these versions are in a consistent state.

In the figure below we illustrate a portion of the onsite infrastructure just to give an overview and a way to imagine how the flow of data is travelling through the different locations.

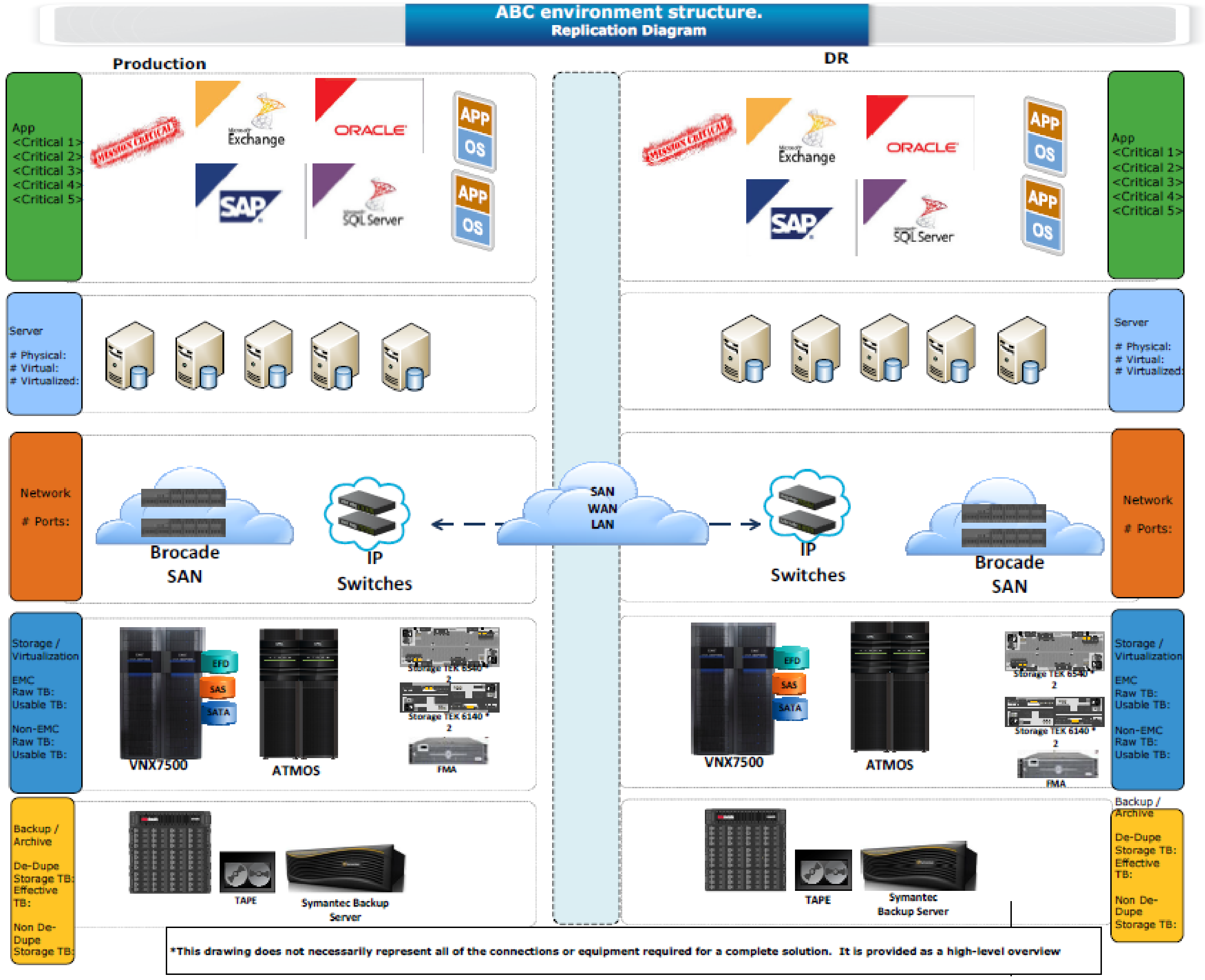


Figure 1: Main Datacenter replicated to another Datacenter in the same area

LI

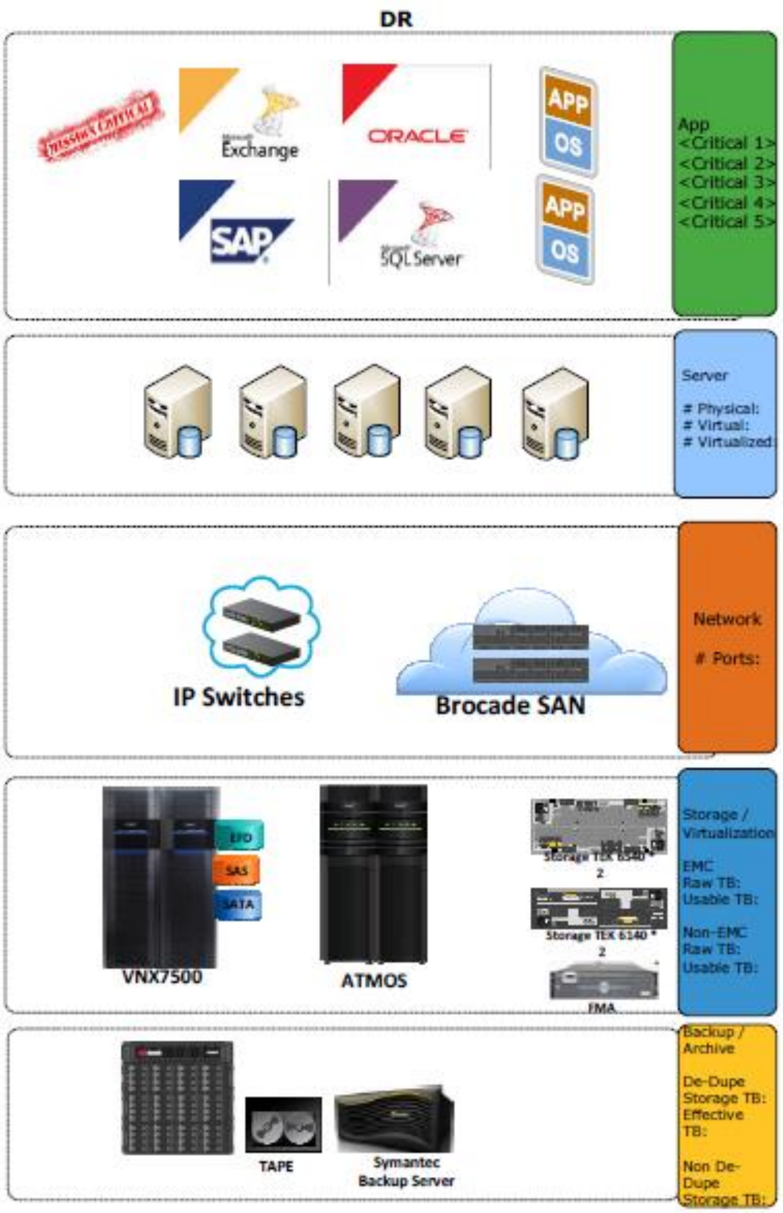


Figure 2: Remote Disaster Recovery Site

Here the customer has 3 copies of the data every day and has an average 8 hours for the backup window. The customer on a monthly basis do the consistency check by restoring specific DBs and applications - upon request - and not a monthly task made by the managed services team. The customer wants to be sure about the consistency of the backed up databases “billing systems, exchange, share point, etc.”

The copy is being restored on the third clone of the data on available and similar machines to be sure that the resident data on the backup systems are consistent and restorable.

This approach is time consuming and requires freeing some administrator from the managed services to perform such tasks. This can be accomplished automatically through an automated tool which can save time in deploying such mechanisms to check the data consistency.

Challenge 2: Efficient management of consistent backups

Here we need to confirm that a backup system based on network attached disk storage can be easily scaled up in terms of capacity and read/write throughput by adding new disks. Disk-based systems can potentially exploit the aggregated capacity of all disks. Most of the world's largest data centers use parallel file systems like Lustre and Panasas. Active Scale is used to store the data. Such systems require external backup systems that rely on dedicated hardware. With a backup solution that is integrated in the file system architecture, we present a novel approach to provide for cost-efficient backups of large-scale file systems.

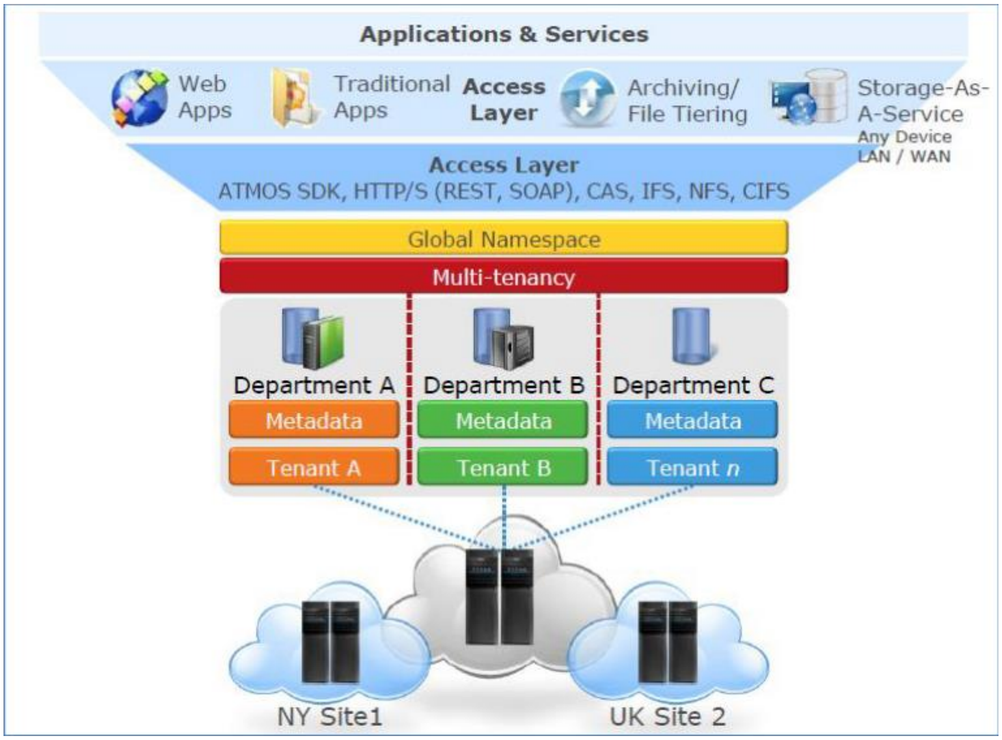


Figure 3: Typical deployment of a cloud object base solution

The design put forth aimed to solve these business challenges:

- Cost Competitiveness
- Highest Levels of Reliability
- Ease of Management
- High Performance
- Compatibility

Cost Competitiveness

Being cost competitive is paramount in order to build and maintain business. Whether facing economic turmoil or economic boom times, we must ensure that the solutions we offer fit our customer‘s budgets.

Highest Levels of Reliability

Offering cost-effective solutions is meaningless if the solution is plagued by outages, which guarantee the consistency of backups and insure an efficient way of management. Backup infrastructures must be capable of performing even after suffering multiple component failures. Customer loyalty will be lost if we fail to meet our availability obligations. We strive to offer products and services that remain operational 24 x Forever. An added benefit of highly reliable systems is the cost savings realized the longer the systems remain operational. Systems capable of remaining in production seven or more years can yield significant long-term savings and/or profit over those capable of running production workloads for 3 to 5 years.

Ease of Management

Hand-in-hand with being cost-effective and reliable, systems need to be as automated and easy to use as possible – being able to do more with less. The more complex the solution, the more resources it takes to maintain and operate over its lifecycle driving overall cost up while driving reliability down. Ensuring staffing levels remain stable in the face of unabated growth is essential in cost containment and is the main reason ease of management remains a key requirement.

High Performance

The overall solution must be capable of delivering during periods of high usage and must be designed to eliminate congestion points. Delivering solutions that suffer from poor performance frustrates customers and wastes precious time and resources tracking down and resolving performance-related issues.

Compatibility

Gone are the days of implementing independent computing silos. It‘s expensive and difficult to maintain solutions designed in isolation. To meet aggressive growth objectives, we need to ensure all of the systems being deployed are compatible and work with one another. Everything needs to work together and scale in order to keep the overall solution as simple and manageable as possible.

Current State of Work and Research Perspectives

In the following section, we give an overview of the current progress on the research topic and outline remaining issues and research perspectives.

A- EMC Atmos

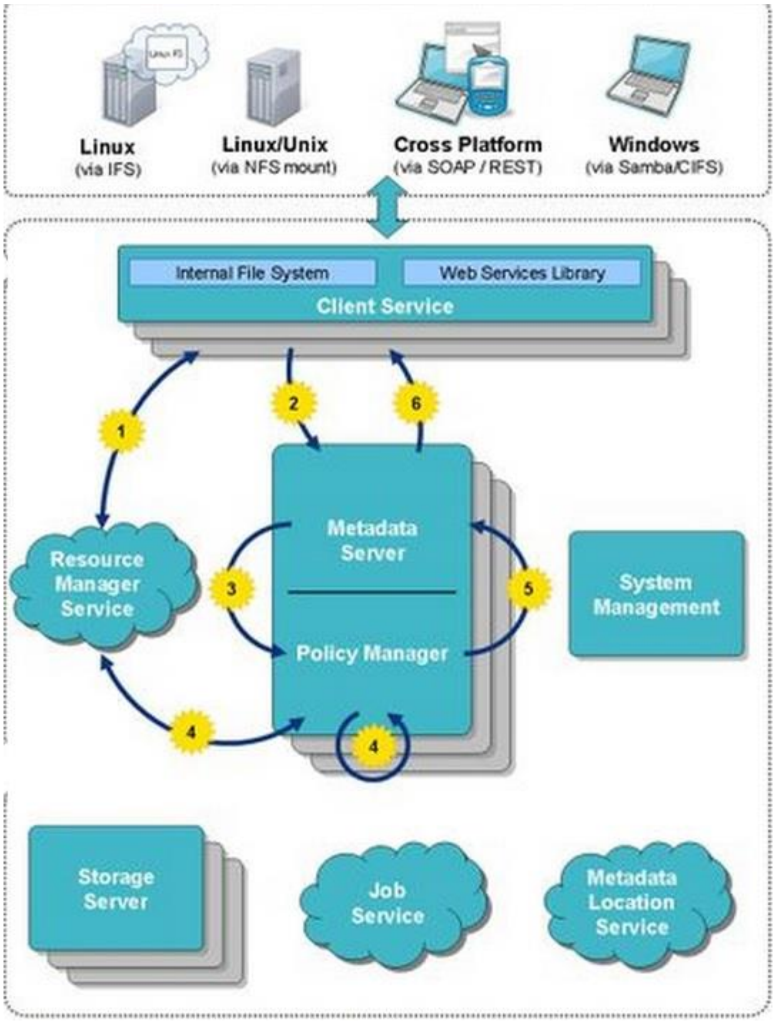


Figure 4: Policy-Driven and information storage and distribution

With EMC Atmos, it has been developed as a distributed file system for cluster (LAN) and Grid (WAN) environments. In accordance with the concept of object-based storage, file metadata is managed by a dedicated metadata server called Metadata and Replica Catalog (MRC), while file content is split into equally sized, rather small objects that are stored across object storage devices (OSDs). EMC Atmos offers various features like striping, replication, X.509-based security, POSIX-compliant access control, and checksums. It has been designed to be offered with hardware, or as a software only version. Atmos is at heart a software storage system. Atmos implementations are available from EMC either already integrated into pre-packaged physical building blocks or as a virtual machine solution for VMware vSphere that can leverage other EMC or 3rd party storage resources. See the figure above.

Atmos is a multi-petabyte offering for information storage and distribution. It combines massive scalability with automated data placement to deliver content efficiently anywhere in the world.

Architecture

Core services

- Metadata server

Stores and manages access to object metadata and namespace.

- Metadata location service (MDLS)

Maintains a distributed mapping from object identifier to MDSs

- Storage server (SS)

Reads and writes user data to disk.

- Resource management service (RMS)

Tracks location and monitors status and properties of service instances in the system in a distributed manner.

- Client service (CL)

Communicates with MDS, SS, RMS, and MDLS to access user's data.

- Job service (JS)

Manages system data maintenance tasks like asynchronous replication and consistency checking

- Policy Manager

  Supports MDS by selecting appropriate policies for objects, instantiates abstract policies into concrete layout descriptions

- System management (SM)

  Accepts and executes administrator operation, reports system status.

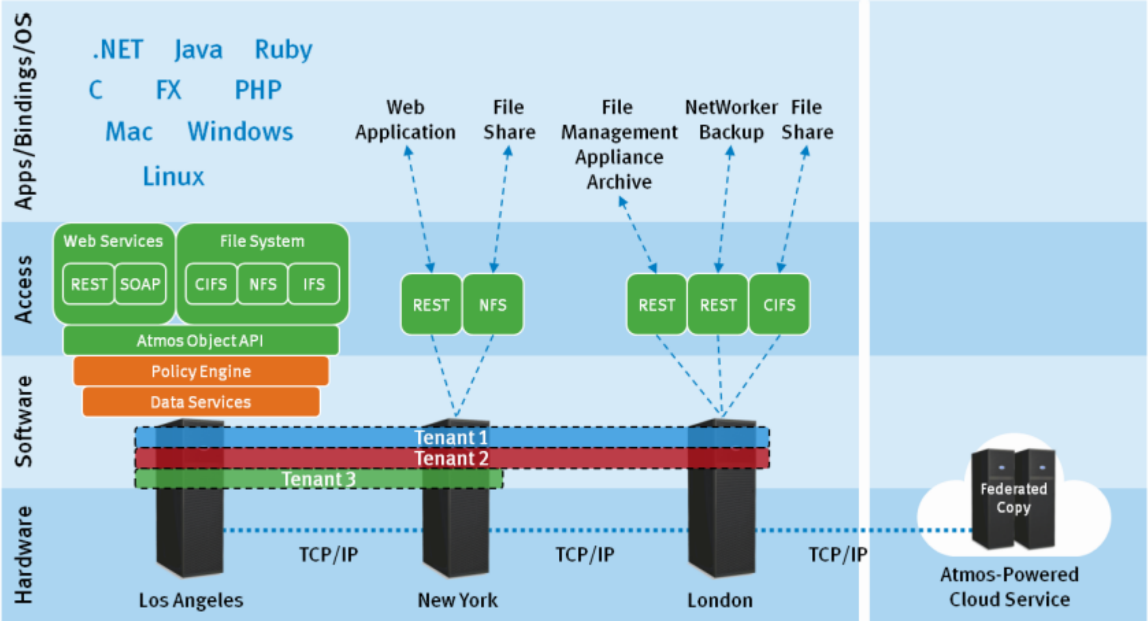


Figure 5: Atmos service architecture in a distributed environment

In addition to full replication, Atmos also provides an erasure coding option called GeoParity. Instead of keeping two or more full 100% copies, “9/12” erasure coding enables storing an “expanded” object containing only 33% additional encoded “redundant” data broken up into 12 segments. By using erasure coding, the original data can be reconstructed dynamically from any 9 of the segments. These segments are cleverly distributed so that the object can survive (and even be accessed during) multiple failures. For greater protection, there is also a “10/16” coding with a 60% capacity overhead. Erasure coding does impact access performance, especially at ingestion, but provides great fault tolerance with much lower capacity utilization. Of course, policies can be written to convert replicated objects to erasure coded schemes as they age appropriately.

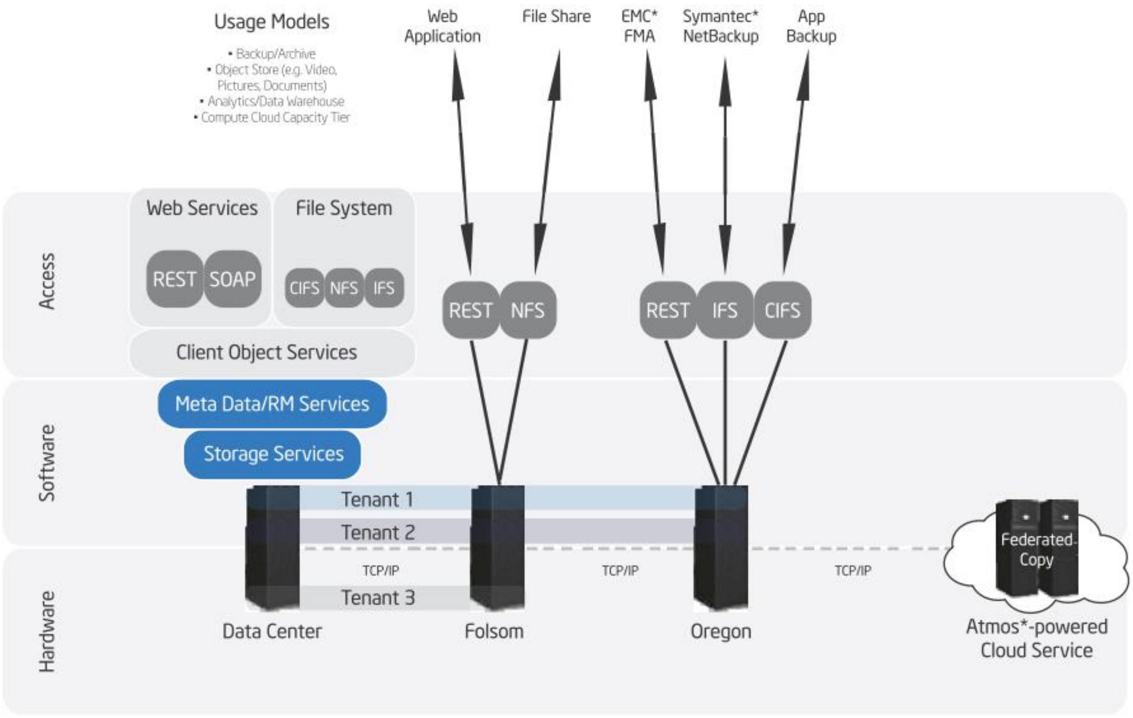


Figure 6: Atmos Service Architecture Stack

How Object is created

1- CL queries RMS to find MDS, RMS responds with available MDS instances.

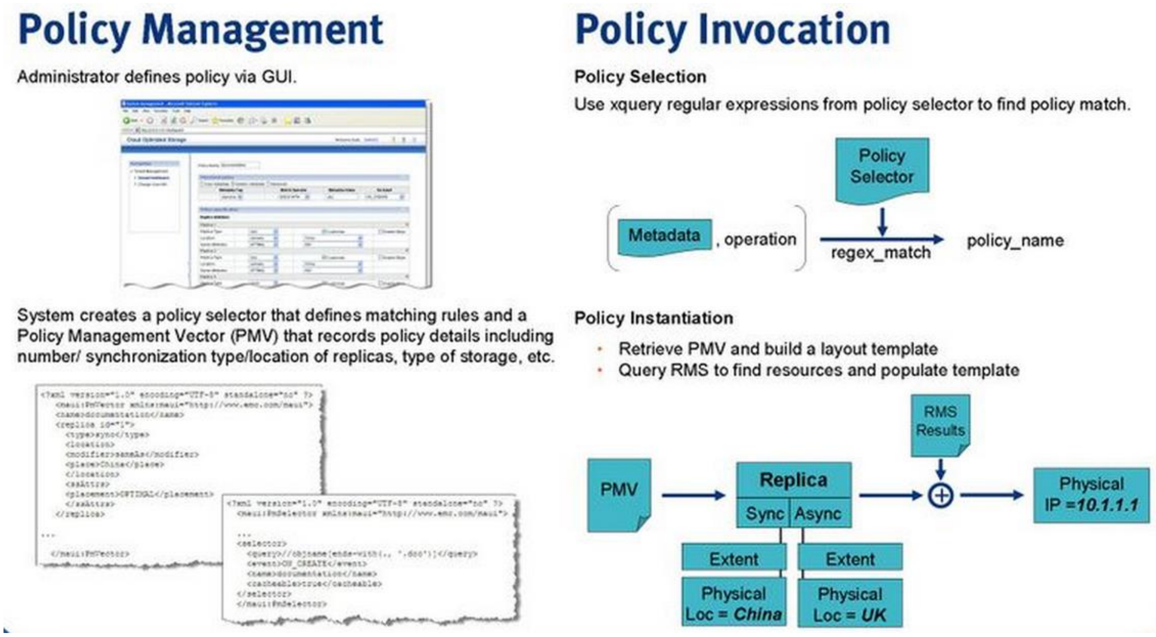
2- CL selects nearby MDS and submits (create) request.

3- MDS generates object metadata; invokes the PM with metadata and operation type.

4- PM performs policy selection and instantiation (see below)

5- PM returns object layout description to MDS.

6- MDS stores object metadata and pass it back to CL.



Policy Management

Administrator defines policy via GUI.

System creates a policy selector that defines matching rules and a Policy Management Vector (PMV) that records policy details including number/ synchronization type/location of replicas, type of storage, etc.

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
  <maui:PmVector xmlns:maui="http://www.emc.com/maui">
  <name>documentation</name>
  <replica id="1">
    <type>sync</type>
    <location>
    <modifier>sameAs</modifier>
    <place>China</place>
    </location>
    <ssAttrs>
    <placement>OPTIMAL</placement>
    </ssAttrs>
  </replica>

...

  </maui:PmVector>
```

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
  <maui:PmSelector xmlns:maui="http://www.emc.com/maui">

  ...
  <selector>
    <query>//objname[ends-with(., '.doc')]</query>
    <event>ON_CREATE</event>
    <name>documentation</name>
    <cacheable>true</cacheable>
  </selector>
  </maui:PmSelector>
```

Policy Invocation

Policy Selection

Use xquery regular expressions from policy selector to find policy match.

Policy Instantiation

- Retrieve PMV and build a layout template
- Query RMS to find resources and populate template

Figure 7: Policy Management vs. Policy Invocation

Turning to the point Atmos was designed by the Cloud Infrastructure and Services Division (CISD) from the ground up with a number of distinct characteristics.

- Information inside the Atmos repository is stored as objects. Policies can be created to act on those objects and this is a key differentiator as it allows Atmos to apply different functionality and different service levels to different types of users and their data. Managing information, which is what we should be doing, as opposed to wrangling blocks and file systems as we tend to do.
- There is no concept of GBs or TBs to EMC Atmos; those units of storage capacity are too small. Atmos is designed for multi-Petabyte deployments. There are no LUNs. There is no RAID. There are only objects and metadata.
- There is a unified namespace. Atmos operates not on individual information silos but as a single repository regardless of how many Petabytes containing how many billions of objects are in use spread across whatever number of locations available to who knows how many users.
- There is a single management console for management regardless of how many locations the object repository is distributed across. This global scale approach means that Atmos had to be an autonomic system, automatically reacting to environmental and workload changes as well as failures to ensure global availability.

Questions on the track

"Fundamentally, this was a distributed systems problem. How do we take a loose collection of services distributed across a wide area and make them operate as we want them to operate?"

Data growth is continuing to explode and not everybody has data which justifies that level of expenditure or has the financial resources to justify spending that much money on storage. So, EMC Atmos was to provide a low cost bulk storage system for these emerging markets, like Web 2.0 companies or other industries with lots of user generated content, or even the evolution of IoT.

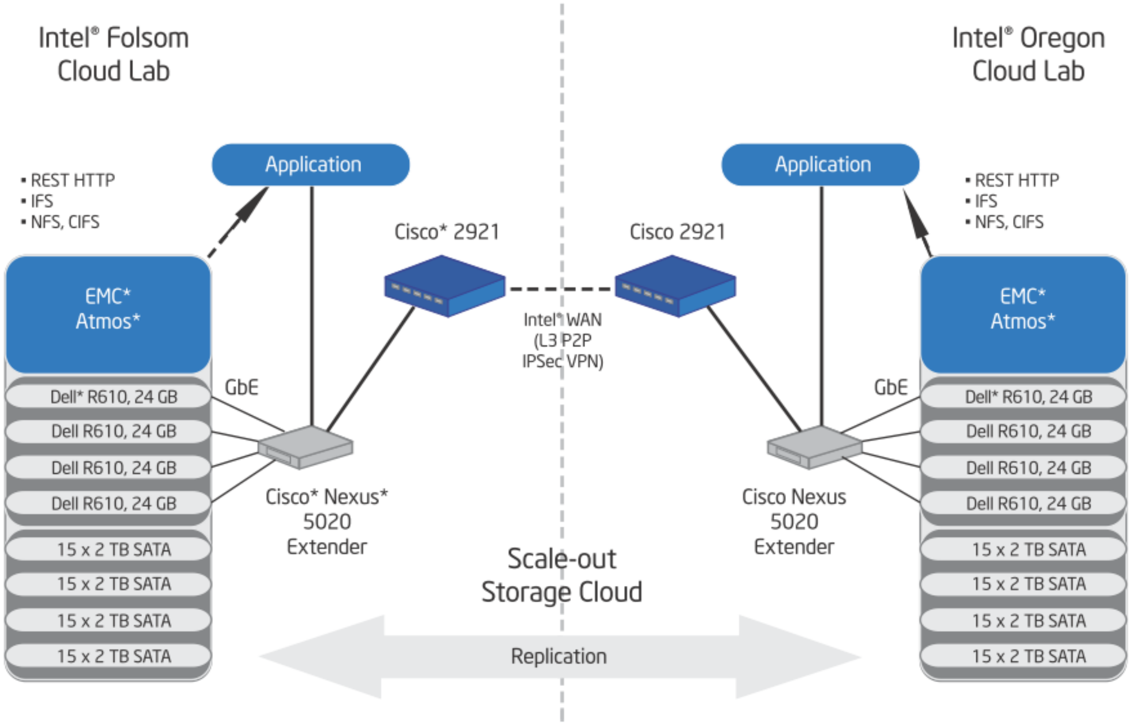


Figure 8: Distributed System Example

Yes, you can put that stuff on regular SAN or NAS systems and that's what customers have been doing as the only other option was to start writing and maintaining their own storage software and build their own storage hardware. That is far from ideal as the value of these companies is in their applications and the services those applications provide.

Atmos can provide a Terabyte approaching ten or more times cheaper than existing SAN or NAS storage systems can offer. That is the problem Atmos was designed to solve and a key part of the product vision comes from the policy driven features of Atmos. The TME and Web

2.0 spaces with those mountains of user generated content, but people want to use that storage in very different ways. Some people want to have one data center; some want many more with the data consistency concern in mind. Some need to support different types of workloads, various types of object sizes, or control where they locate specific objects and how they get them close to their customer regardless of where the customer is located in relation to where the data was first stored.

The core of the Atmos design is how we enable customers to define policies as to how data actually hits disk. There are no administrators saying "Joe's photos should be on this particular piece of spinning rust", rather they write policies to describe how Joe is a subscription customer therefore his files require a certain number of copies associated with them for backup and should have a certain rolling retention policy in case he cancels his account. Thus, they should be in this data center here and not in one thousands of miles away.

But if Joe packs up the family and the dog and moves across country his data may be replicated to the data center now closest to him depending on the policies applied to his files.

Information management is something many IT solutions providers' talk about a lot so providing a storage solution designed with policy based information management at its core is a big thing that can be done with Atmos, for example. You're not just storing information, you're replicating it to where it's needed and putting it as close to the user as possible. You're compressing it, de-duplicating it, or deleting it depending on what policies are applied to it and if it hasn't been accessed in a while you can even spin down the drives inactive objects are stored on to save power. In addition to supporting compliance and retention policies, metadata can be used to drive automated file distribution, access control, and data protection activities optimizing for the appropriate level of data resiliency, performance, and availability. For most applications, thoughtful use of user metadata can remove any need to implement a separate management tracking database for stored objects.

“Multi-tenancy. Could we talk about that a bit more? Could I offer storage as a service to different users or organizations?”

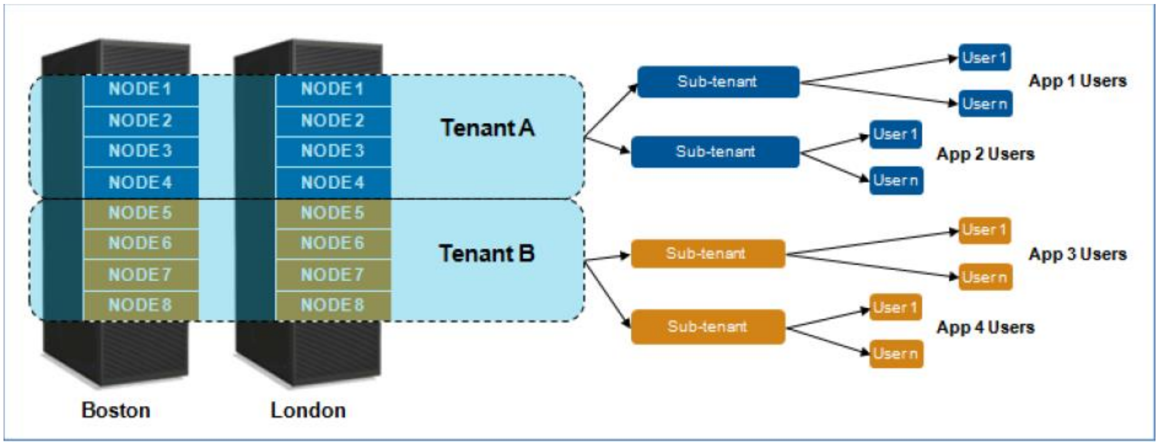


Figure 9: Multi-Tenancy Diagram

Yes, we could. Multi-tenancy means that Atmos can support many different tenants and subtenant features to enable policy and administrative partitioning of the cloud with logical isolation. Each tenant can have their own private namespace under the Atmos namespace but tenants are not aware of other tenants or the objects belonging to those tenants.

You could be providing services to users on the Internet and hosting application test and development as well as providing services to your internal business units, but none of those tenants would know about each other.

The Atmos architecture attempts to offer the broadest range of access mechanisms possible. Most end users will interact with Atmos through pre-integrated applications, custom packaged, so the end user will not be aware of what is storing and managing the information. Application developers at Enterprises and ISVs can best take advantages of Atmos via the Atmos Web Services interfaces (REST and SOAP), which allow ubiquitous, scalable, and full features access to Atmos. Most of the increasing number of EMC and third party integrated packaged applications use the Atmos REST API. For certain use cases, end users can also take advantage of Atmos as a mounted drive via the Atmos installable File System Linux and traditional like NFS or CIFS.

Figure 10: flexibility to add data from multiple sources

“We were talking about this being a low cost solution, what's low cost at the scale we are talking about here? Sure there's capacity cost but it's not just that…..”

Well, not only does the initial cost of delivering the product to the doorstep have to be low but also it has to be something that the customer can maintain very easily and we're talking about the Petabyte range – as an example of the large scale we proposed – when we're talking about deploying this so one of the key design elements was how to provide a customer installable configurable and maintainable implementation.

Going back to the traditional EMC model of "We'll make sure it works but you're going to pay for it", where parts show up at your door with a service engineer attached shoots the entire low cost target out of the water if you have to do that more than a few times a year. That's why a lot of the installation, configuration, and maintenance can be done by the customers themselves.

- Low cost, low touch, incredible scale and density
- Billions of objects globally distributed with policy based information management
- Petabytes of storage which could be in the same room or distributed around the world but with a single point of management

Those were some of the design goals.

B- Backup Architecture

To efficiently maintain incremental file system backups, we suggest a complementary backup management design: for each instance of an Atmos live system service, one or more backup service instances run on across the remote servers. Backup services are lightweight versions of their live system counterparts, with a set of functions limited to backup management. Although backup and live system services share the same hardware pool, backup service implementations rely on different code bases, and backup data is stored on different hard disks, in order to make the system less susceptible to software bugs and disk failures.

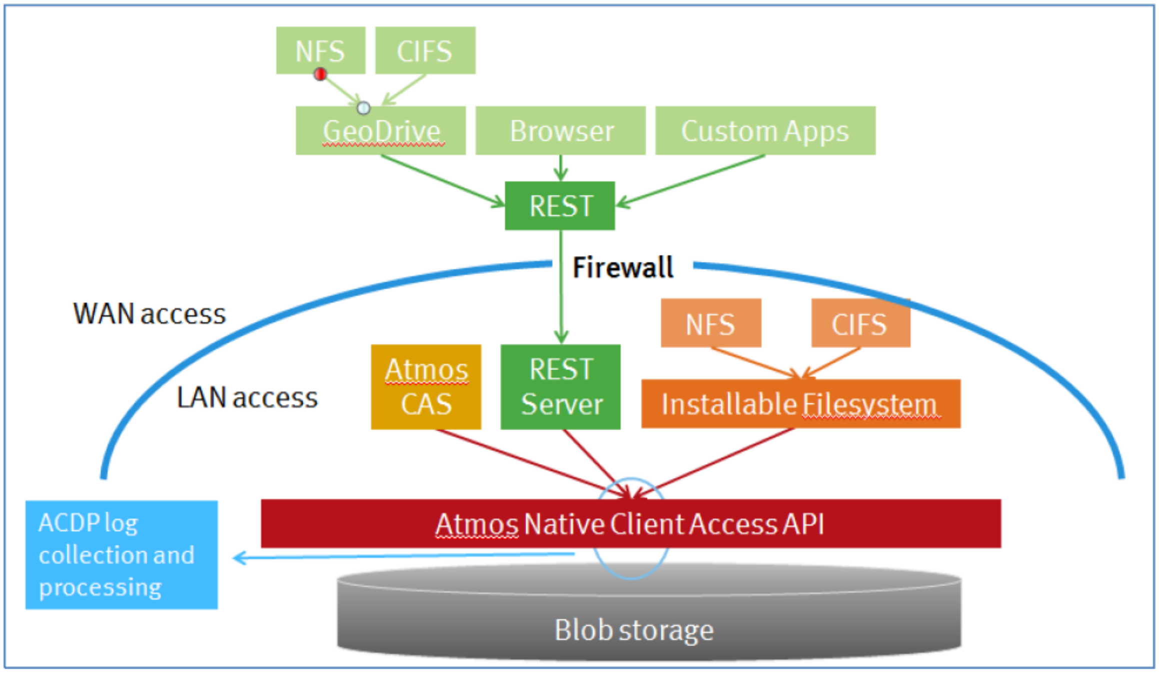


Figure 11: EMC Atmos High level architecture

Such a backup architecture has a major advantage as it implements GeoParity with the Cauchy Reed-Solomon algorithm, and uses two different implementations. The first is a 9/12 configuration, where an object is split into 9 data fragments and 3 coding fragments. This efficiently tolerates up to 3 drive failures, and gives the object 5 nines of data availability. In addition, the storage overhead is only 33% compared with 100% for a RAID-1 configuration, or 25% for a RAID-5 configuration (using k/m to calculate overhead: 3/9, 1/1 or ¼). Such benefits come at a price; in this case some performance overhead is required for the encoding and decoding operations.

The second configuration is a 10/16 option, where there are 10 data fragments and 6 coding fragments. This configuration will tolerate up to 6 drive failures, has a 60% storage overhead, but at a cost of additional performance overhead than the 9/12 configuration.

This saves acquisition, operating, and maintenance costs. Adding new hosts to the live system automatically adds new hosts to the backup system, which ensures that scaling up one of the systems also scales up its counterpart. The scale does not only increase in terms of storage capacity, but also in terms of throughput, since multiple live services (and backup services, respectively) can be accessed in parallel. This eliminates potential bottlenecks in the connection between the live system and the backup system and ensures that backups can be created and accessed fast.

Capacity can be added seamlessly in 2 methods: either by expanding capacity within an existing datacenter by deploying new cabinets of purpose-built, preconfigured nodes and disk enclosures, or by expanding incrementally by simply adding nodes and disk enclosures to existing cabinets. In addition to flexible growth, we can add this new capacity seamlessly to an up and running cloud. Once new capacity is added to the cloud, it is immediately available for use without any provisioning needed. In this context, the solution scales horizontally as rack models can be mixed and matched together, both within a site and between sites.

C- Research Perspectives

Based on this mechanism for backup creation, strategies need to be developed to select suitable sets of backup services. Possible placement criteria for backups are throughput and latency between backup and live services: throughput ought to be maximized, while latency ought to be minimized, so as to be capable of creating and restoring backups fast. Backup placement strategies also have to take failure tolerance aspects into account. A single failure of a hardware unit (e.g. Host machine, network switch, etc.) must not simultaneously affect both a set of live system services and their corresponding backup services. Another research perspective is scheduling mechanisms for automatic backup creation. Such mechanisms could be policy-based; it may, for example, be reasonable to restrict backup creation to times of low system utilization, with the aim of minimizing the impact on normal file system activity. Likewise, assigning priorities to backup-related data transfers may help to reduce the competition for network and compute resources between backup and normal file system workloads.

D- Contribution and proposed work

As the main contribution of this article, the design of a disk-based backup infrastructure for object-based file systems was presented, with a strong focus on the technical aspects of the system, mechanisms for the creation of consistent backups in details, and different techniques to implement such backups. The design of an OSD was elaborated showing that it is capable of managing multiple versions of a file’s content. Approaches to store version vectors for large numbers of files with minimal overhead were also introduced. Based on these mechanisms, and solutions with respect to consistency guarantees, approaches were presented that provide for different consistency guarantees, like time-based consistency or consistency from the point of view of a process group. On the basis of the presented mechanisms for consistent backup creation, a comprehensive, integrated backup infrastructure for Atmos can be found in separate sections in deep detail. Different technical aspects of such an infrastructure can be researched on www.emc.com/emcatmos, like backup creation, recovery, and placement and scheduling. The whole system can be evaluated in a wide range of experiments. There are also global measurements conducted in a real world environment, in order to determine the different characteristics of the system. Such characteristics include backup creation and reconstruction performance, energy consumption, and operating costs.

Challenge 3: Supporting the operation “Data Protection Advisor” DPA employed.

The IT world is changing from a cost-based asset to service approach and this revolution requires that operation can share data about service status, report point in time status and history, manage capacity planning, and chargeback to different internal/external customers.

In order to achieve different level of service customer need to implement a different data protection strategy that includes:

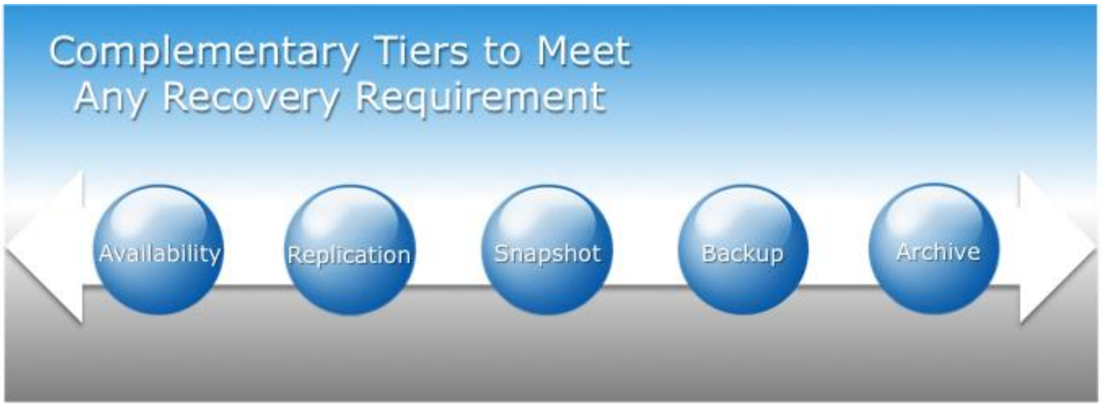


Figure 12: Complementary Tiering Requirements

This means that data protection information is stored in different places without a unique repository. In fact, this situation will have a very high impact on operations because they need to monitor and control different objects in order to give visibility on service status.

Customers can have different level of maturity to support service:

- Basic level is to collect data manually from different source and post-elaborate on data to create reports. This means that the customer spent time and resources to collect information and don’t have a point in time situation.

- Middle level manages some scripts to collect and post-elaborate on data. This means that the customer needs to maintain scripts during time.
- Advanced level, customer has one or more tools to automatically manage data collection and reporting.

In any case, EMC Data Protection Advisor can support customer to have a complete view of the Data Protection scenario covering availability, replication, snapshots, and backup information and can be implemented to have one flexible and powerful tool to manage all aspects of service reporting, monitoring, chargeback, and capacity planning reports.

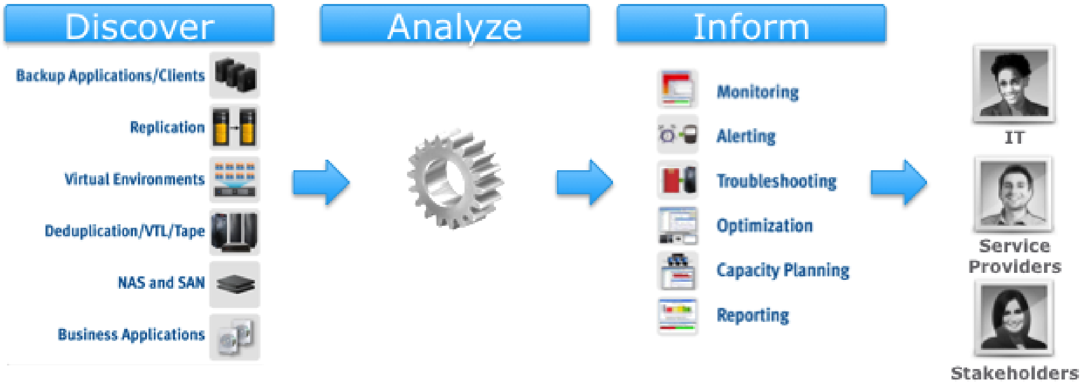


Figure 13: Data Protection Advisor Flowchart

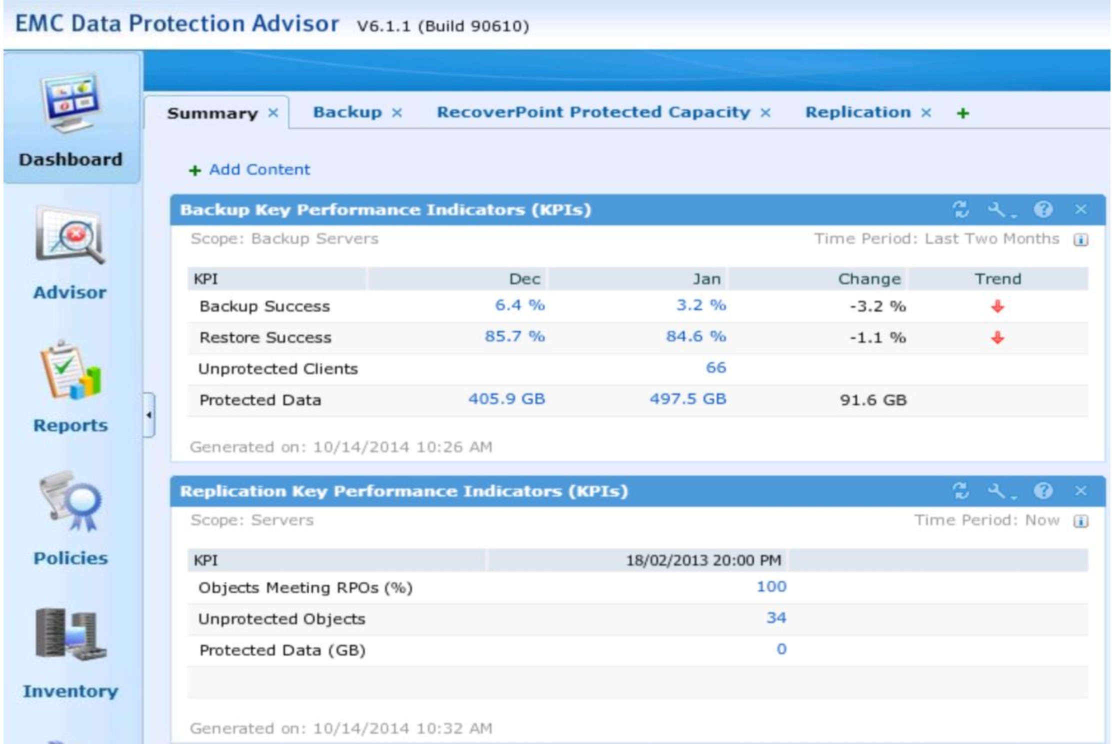
DPA gives the customer all the required information to complete the deployment of data protection services. A customer can manage a complex/multiplatform environment, and analyze all data and inform all internal/external stakeholders with the correct level of data aggregation.

DPA is a complete reporting suite; below are examples to demonstrate major functions.

Custom Dashboard

Data Protection Advisor can be configured to show a custom dashboard when a user logs in to the system and every user/group can have a different view.

For example, customer can create one dashboard with backup and replication information on the same page with info about RPO, backup and restore success rate, and unprotected clients/objects.



Backup Key Performance Indicators (KPIs)

Scope: Backup Servers

Time Period: Last Two Months

KPI	Dec	Jan	Change	Trend
Backup Success	6.4 %	3.2 %	-3.2 %	↓
Restore Success	85.7 %	84.6 %	-1.1 %	↓
Unprotected Clients		66		
Protected Data	405.9 GB	497.5 GB	91.6 GB	

Generated on: 10/14/2014 10:26 AM

Replication Key Performance Indicators (KPIs)

Scope: Servers

Time Period: Now

KPI	18/02/2013 20:00 PM
Objects Meeting RPOs (%)	100
Unprotected Objects	34
Protected Data (GB)	0

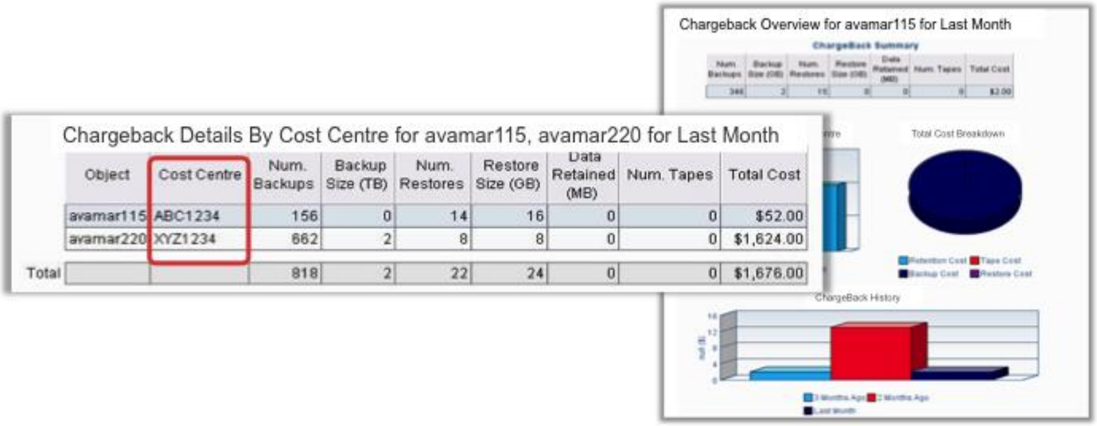
Generated on: 10/14/2014 10:32 AM

Figure 14: EMC Data Protection Advisor GUI

Chargeback

To automatically manage the cost of service (for example TB backed up or replicated), customer can set up customized chargeback policy with different configured costs.

A report can be created with automatic chargeback information divided by different levels of aggregation client/ organization. This can be included inside customer’s process in order to export data to automatic create bill for final internal/external customer.



Chargeback Overview for avamar115 for Last Month

ChargeBack Summary

Num. Backups	Backup Size (GB)	Num. Restores	Restore Size (GB)	Data Retained (MB)	Num. Tapes	Total Cost
346	2	15	0	0	0	$2.00

Chargeback Details By Cost Centre for avamar115, avamar220 for Last Month

	Object	Cost Centre	Num. Backups	Backup Size (TB)	Num. Restores	Restore Size (GB)	Data Retained (MB)	Num. Tapes	Total Cost
	avamar115	ABC1234	156	0	14	16	0	0	$52.00
	avamar220	XYZ1234	662	2	8	8	0	0	$1,624.00
Total			818	2	22	24	0	0	$1,676.00

Figure 15: Chargeback Details for Avamar Example

Capacity Planning

To support provisioning operations, capacity planning functions help customers predict future needs based on historical trends.

In this figure we can see the need for media capacity planning for backup of a specific backup server.

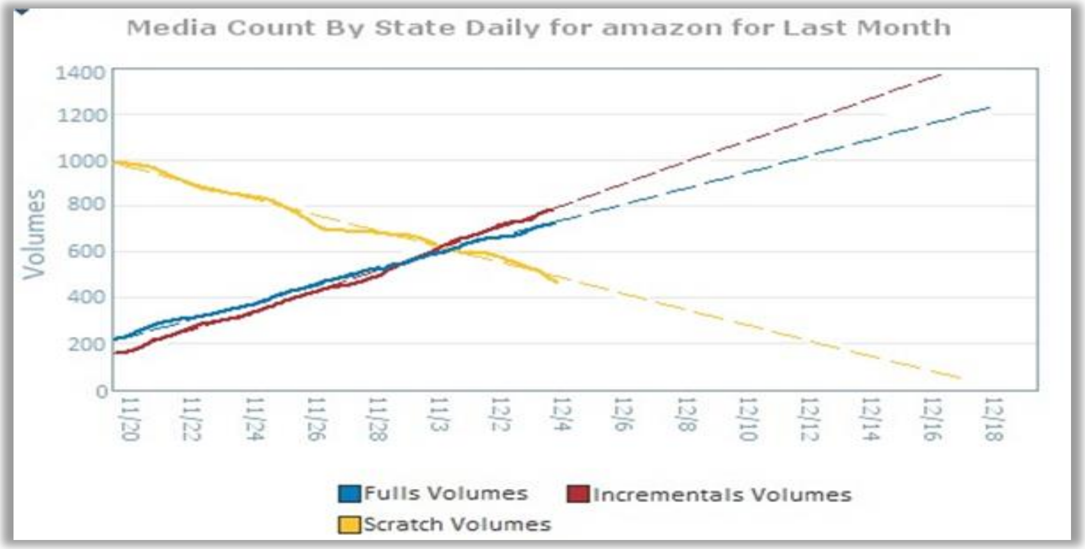


Figure 16: Media Capacity Planning

Reporting

Reports can have different formats and can be scheduled to automatically send periodic reports to different users.

Backup Job Summary for EMC Avamar for Las...
Job Summary
Hourly Number of Jobs
Daily Data Backed Up
Cumulative Jobs Completed
Cumulative Data Backed Up
All Database Changes

Figure 17: Backup Job Summary for EMC Avamar

Optimization

Help customer to identify under-utilized resources or timeframe with less utilization.

Aggregate Throughput
Data Domain Frame Utilization
Filesystem Throughput for EMC Net...
Total Throughput
Backup Server Utilization
Read vs Write Throughput

Figure 18: Optimization Graph Model

Our point of view for future development of DPA

On Third Platform, data protection requirements will change with additional requests to support the new world that is coming with more flexibility, advanced options, and to create business value from backup information.

The goal of this section is to describe some scenarios and required features to simplify the operation at customer site.

Analytics & advanced search

Business wants to have strategic information about IT development in the future and actual TCO for each application.

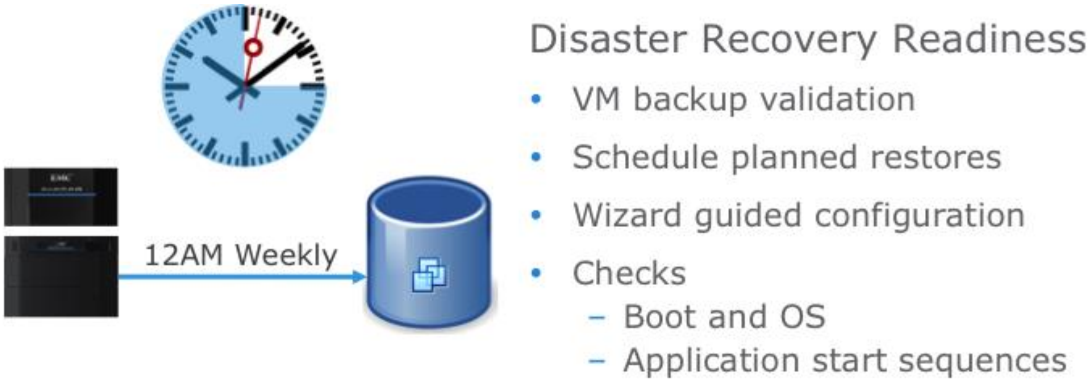
In addition, applications administrators and end users ask for a self-service restore function in order to manage restore operation in a quick and simple mode.

End users want a simple and easy interface to search inside backup information and easily manage a self-service restore. In the future, users will be able to log in to a search interface like Google with advanced searching functions across the backup solutions and search for files, email, SharePoint document, etc. inside backup repository and recover items directly from this GUI. This feature will change the restore path from backup administrator to final user management; final user can directly restore his data without request to backup admin that will take time to be released.

Market requirements are to have one single point of restore for Backup, Archive, and snapshot data on any format and platform with a self-service approach and strong security management.

Cross-check consistency for backup

Due to an internal or external certification procedure, a periodic disaster recovery test starting from backup is required. It is very challenging to beat it with manual operations on a complex environment like we presented that can have a high operational cost to achieve the results needed.



Disaster Recovery Readiness

- VM backup validation
- Schedule planned restores
- Wizard guided configuration
- Checks
  - Boot and OS
  - Application start sequences

Today, many IT vendors like EMC have stepped forward on this track with EMC Avamar or Veeam which can manage a programmed recovery for one or more virtual machine for consistency check with deep integration with VMware.

Figure 19: Disaster Recovery Readiness

Backup server Database consolidation & partial migration

Organizations are growing and changing very fast which creates a dynamic situation on application configuration and backup requests.

This scenario generates frequent backup request changes in order to move a client from one backup server to another, consolidate more backup servers into one, or divide a backup server into different instances.

With the current backup software i.e. NetWorker, Backup exec, NetBackup, it is very difficult to manage every change because there isn't a feature to merge or partially import/export the database.

Today, to manage this operation it is required to re-create objects on the target backup server (manually or with a script) and manage a scanner operation to import the media database; a very complex and long operation to manage.

The future is to have a tool that enables us to merge/import/export configurations of a database and allow the customer to move data across different backup servers and different backup solutions.

Challenge 3: Fulfill consistency needs “CAPEX/OPEX”

Nowadays, every customer needs to justify the cost for backup management and try to move OPEX to CAPEX and reduce global TCO for the backup solutions.

Every customer has different requirements and a different level of maturity and technology adopted and need to evaluate TCO for Backup solution in place.

The best way to support this is to create a path to justify backup consistency needed based on different steps:

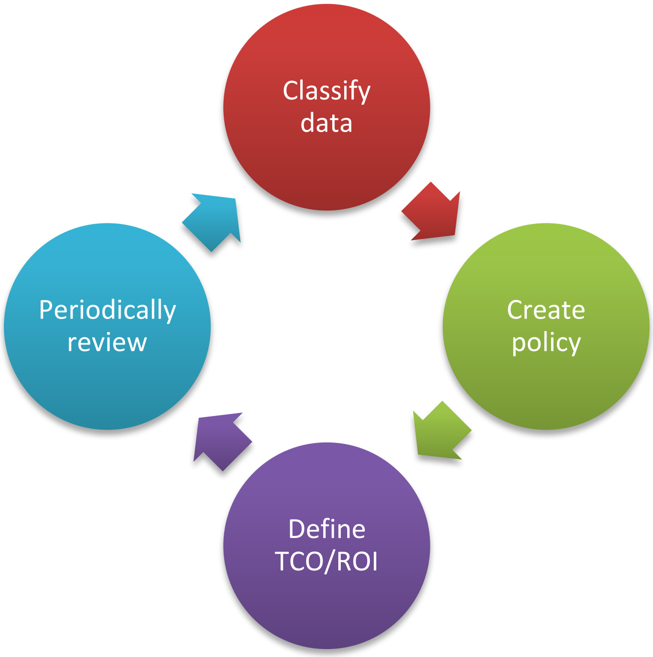


Figure 20: Backup Consistency Model Steps

1. Classify data on the basis of business value
2. Create the correct data protection strategy for every data category
3. Define the TCO for the solution with a ROI Analysis for new investments
4. Periodically review the process and in place policies

Many vendors – including EMC – put a lot of effort into mitigating the many challenges that could face customers regarding the creation of customized Data Classification and Policy Strategy based on customer needs. The output of the first two steps is a service catalog with details on backup policy and reference architecture.

A simple service catalog example can be a table that summarizes policy for every type of data:

TYPE	RTO	RPO	Retention	Replica Site	Backup Strategy
Platinum	4 h	2 h	30 days	2	Backup from Clone/Snapshot or remote storage copy to EMC Data Domain with NetWorker. First replica after saveset end and clone job for DR copy
Gold	8 h	4 h	15 days	1	Backup from production with client direct on Data Domain and copy on DR site at the end of group
Silver	24 h	24 h	7 days	0	Backup from production with direct client on Data Domain. No replica.

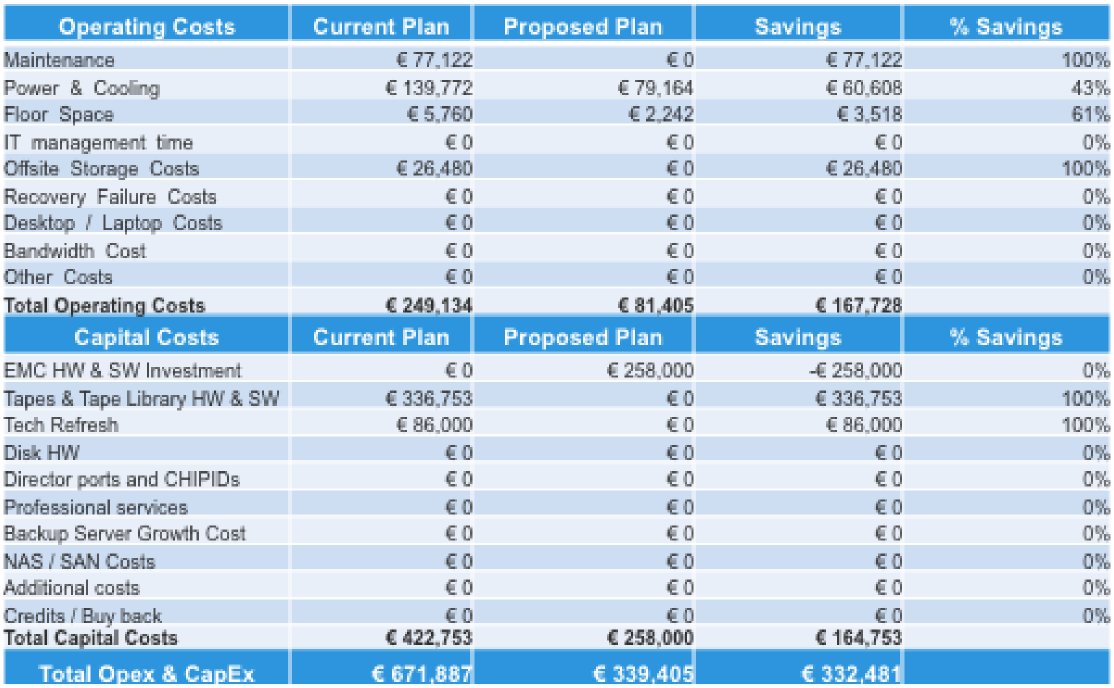
With this approach, the customer will have a different cost for different data type:

- Critical Application will have a Server less backup with three copies in different locations. This architecture requires 3 Datacenter, 3 target backup systems, 2 copies of data inside storage, and will have a very high cost.
- Medium level Application will have 2 copies on 2 different sites and backup will use production disk. This means lower cost in term of disk, target backup systems, and connectivity.
- Low level Application will have only a local copy of data with limited cost on data protection site.

To ensure that this approach will be adopted from application reference suggestion, create a chargeback procedure in order to quantify cost for every application and give this evidence to management. In this way, application reference is aware of the cost related to the protection of data protection, and gives a better understanding of the importance of classifying the data in different security levels.

To justify new investment for tech refresh or new technology it is possible to create a TCO/ROI Analysis that can help the customer justify the importance of a new investment or switching to a new backup technology from tape-based to a disk-based approach.

The goal is to create a customized proposal with all details and pricing justification analysis organized in a presentation with tables and graphics that can be presented to management. This approach is more effective than support cost justification with a poor document or without any data to support the request. Below is one example of detailed cost/benefits table generated with TCO/ROI Tool support.



Operating Costs	Current Plan	Proposed Plan	Savings	% Savings
Maintenance	€ 77,122	€ 0	€ 77,122	100%
Power & Cooling	€ 139,772	€ 79,164	€ 60,608	43%
Floor Space	€ 5,760	€ 2,242	€ 3,518	61%
IT management time	€ 0	€ 0	€ 0	0%
Offsite Storage Costs	€ 26,480	€ 0	€ 26,480	100%
Recovery Failure Costs	€ 0	€ 0	€ 0	0%
Desktop / Laptop Costs	€ 0	€ 0	€ 0	0%
Bandwidth Cost	€ 0	€ 0	€ 0	0%
Other Costs	€ 0	€ 0	€ 0	0%
Total Operating Costs	**€ 249,134**	**€ 81,405**	**€ 167,728**	
Capital Costs	**Current Plan**	**Proposed Plan**	**Savings**	**% Savings**
EMC HW & SW Investment	€ 0	€ 258,000	-€ 258,000	0%
Tapes & Tape Library HW & SW	€ 336,753	€ 0	€ 336,753	100%
Tech Refresh	€ 86,000	€ 0	€ 86,000	100%
Disk HW	€ 0	€ 0	€ 0	0%
Director ports and CHIPIDs	€ 0	€ 0	€ 0	0%
Professional services	€ 0	€ 0	€ 0	0%
Backup Server Growth Cost	€ 0	€ 0	€ 0	0%
NAS / SAN Costs	€ 0	€ 0	€ 0	0%
Additional costs	€ 0	€ 0	€ 0	0%
Credits / Buy back	€ 0	€ 0	€ 0	0%
Total Capital Costs	**€ 422,753**	**€ 258,000**	**€ 164,753**	
Total Opex & CapEx	**€ 671,887**	**€ 339,405**	**€ 332,481**	

Figure 21: ROI Analysis

This is a complete HW/SW TCO analysis with customized environments settings and saving. Every customer can decide to consider different aspects of TCO including FTE, Cooling, Power, Datacenter space, etc.

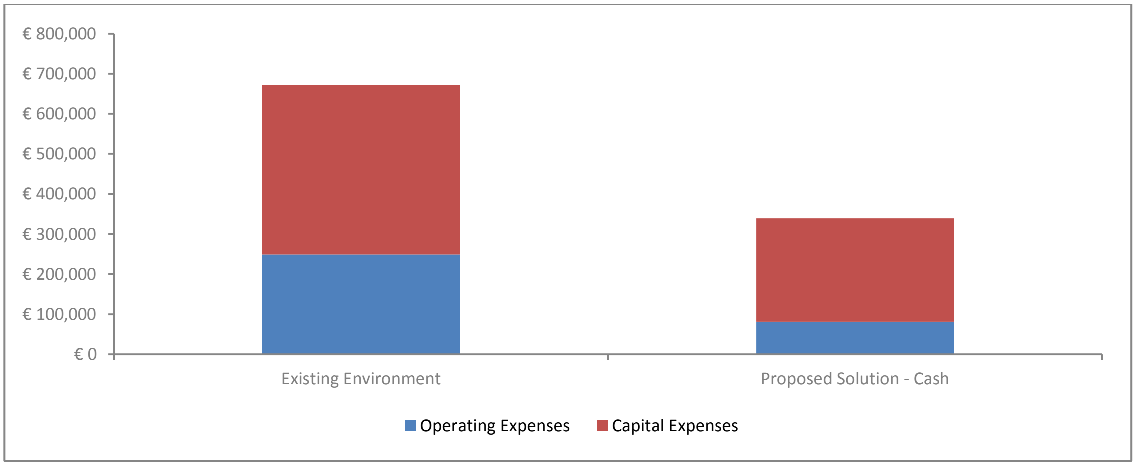


Figure 22: Expense Model Comparison

The graph provides evidence of CAPEX/OPEX costs if customer stays with actual configuration or if they decide to swap tape-based backup to disk-based backup. Here is an EMC Data Domain example.

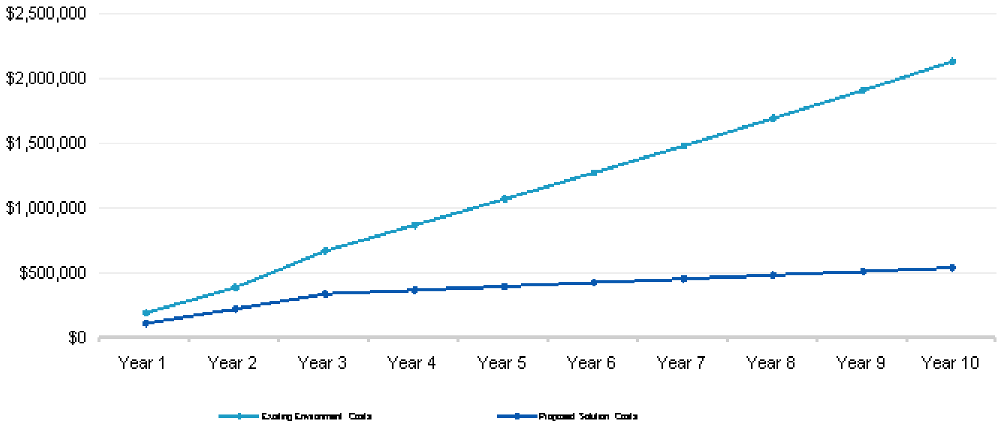


Figure 23: CAPEX/OPEX Cost Comparison

This is a view of cash flow with or without a proposed solution prospected with a 10% per year data growth. Notice that there will be increased savings by switching to the new technology.

The Future: How to mitigate these challenges

Increasing the pressure on Backup/Recovery and the expected data growth will be the key challenge, along with how to maintain backup requirements using traditional approaches.

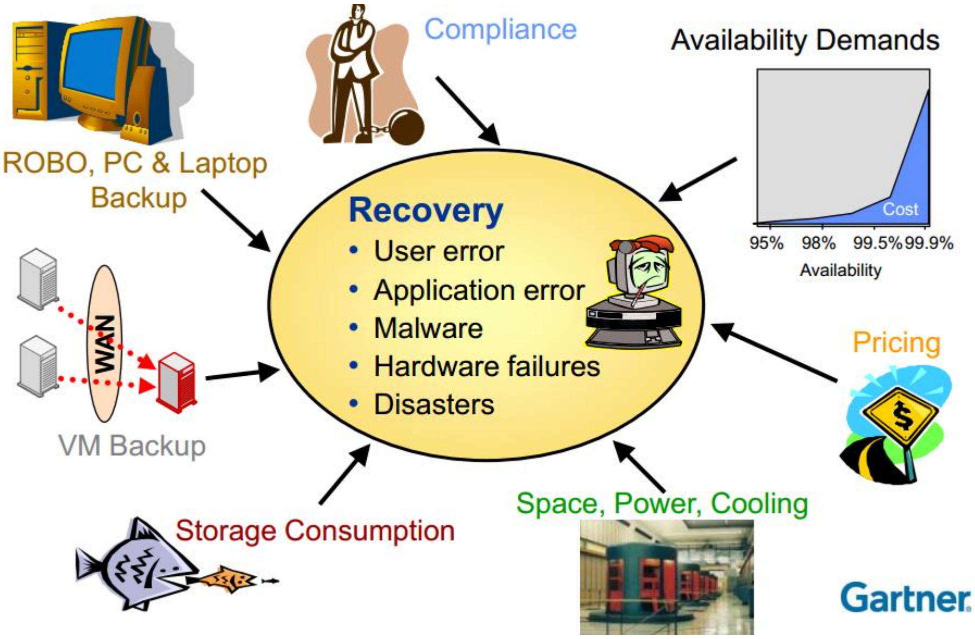


Figure 24: IPO Diagram for Recovery

The backup consistency future with virtualization technologies will become more achievable with higher accuracy improvement and easily can verify the workability of backups and assure recoverability without building a new environment and more hardware to test and verify them.

This can be achieved by providing backup validation and assurance service through simulated recovery to minimize the risks and recovery time, and also maximize backup efficiency and storage resources utilization. The recovery simulation will facilitate measuring the actual time needed to restore backups, and provide automatic recommendations on ways to be improved.

Recovery Simulation will help in gaining immediate testing and validation procedures to insure that the data is consistent and can be recovered. As mentioned previously, EMC DPA can be developed to add real-time monitoring and analysis for all the backup activities.

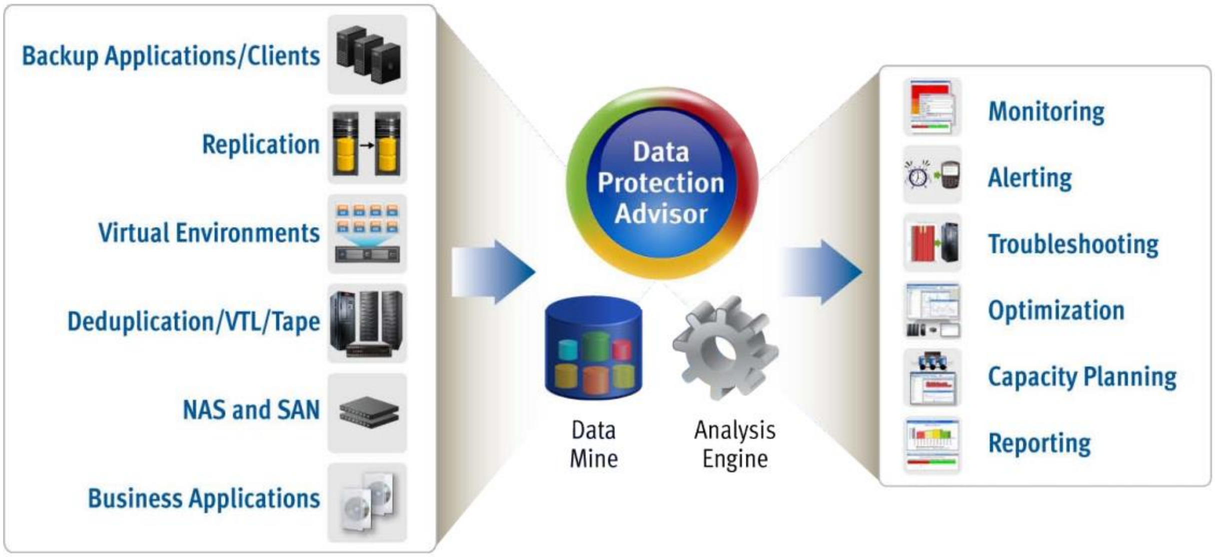


Figure 25: EMC DPA Figure

The Recovery Simulation feature allows customers to test and validate their backup for database, operating system, and applications and insure the data integrity plus the consistency under several backup policies and conditions to provide proactively corrective actions based on the predefined RPO/RTO policies.

Existing Market Research

The market always focuses on the Backup and Recovery technologies and the reliability of the available solutions through surveys. Continues Data Protection (CDP) solutions can fulfill the required future backup and recovery consistency measures in certain levels.

Most CDP solutions are designed to protect data from any application on SAN-attached servers and storage arrays. These solutions protect applications during point in time by minimizing data loss with instant replication of changes to an application, including databases.

Near-continuous data protection (near CDP) is a general term for backup and recovery products that take backup snapshots at set intervals. The term evolved from a need to differentiate those vendor products that take snapshots on a pre-determined schedule from those that take snapshots whenever new data is written (true CDP).

Both near CDP and true CDP support instantaneous recovery, meaning that if the primary image is damaged, a recovery image can be mounted immediately. The difference between near CDP and true CDP is the recovery point objective (RPO) they offer.

A near CDP product is an acceptable backup option when the potential loss of a small amount of data can be tolerated. Near CDP products are usually limited to a specific number of snapshots that the application or storage system can create. Once that limit is met, earlier snapshots are overwritten.

The evolution of near-CDP

We didn't call near-CDP systems "near-CDP" until the CDP market was invented. Companies that built up the "real" CDP market tightly defined CDP so it excluded anything that did snapshots. But then everybody that did snapshots and replication wanted to be known as CDP products, and you had two different groups of products both saying they did the same thing. The truth is that they both are still very different than traditional backup, as they both are block-level-incremental-forever products, so they were more like each other than they weren't. The term "near-CDP" was born. Now we call all products that do snapshots and replication "near-CDP."

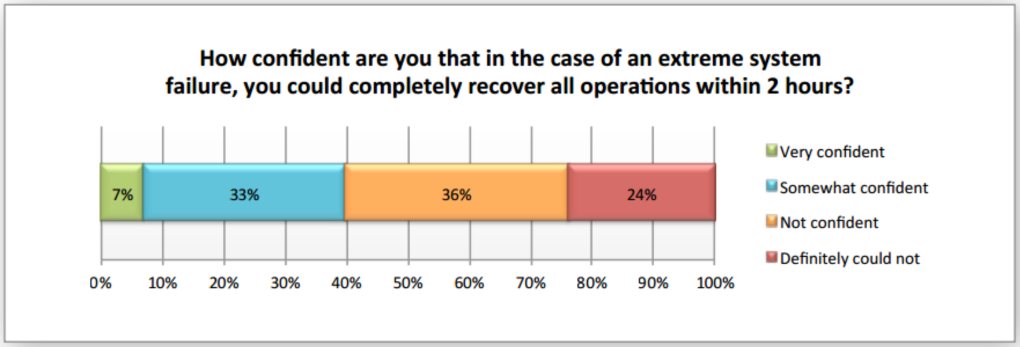


Figure 26: Axcient-SMB-Backup-and-Recovery-Survey-Report-2014

Conclusion

The topic of data consistency remains one of the major challenges in the modern data center as it touches the most critical part of any organization’s assets, which is the data. In this Knowledge Sharing article we tried to shed light on many aspects and show our point of view of the pain of the customers and how to mitigate such pains and concerns. We tried to illustrate our vision of how to achieve the consistency and be ready to the Third Platform with a solid strategy using one of the most successful examples in the market, EMC, the leader in the Data Protection and Availability sector. We also showed our design of Efficient Management of Consistent Backups in a Distributed File System.

Appendix

1- Axcient-SMB-Backup-and-Recovery-Survey-Report-2014.pdf

2- gartner_-_the_broken_state_of_backup_(6-09)_(1).pdf

3- AST-0105542_15_minute_guide_English.pdf

4- Falconstor_sDataBackup_SO#034029_E-Guide_061311.pdf

5- http://www.gfi.com/blog/wp-content/uploads/2013/06/Backup-n-IT-admins-02_RGB72dpi.jpg

6- http://www.gfi.com/pages/research-brief-for-backup-survey-data.pdf

7- http://www.biztechmagazine.com/sites/default/files/tiny-uploads/2013/backup-cdw-infographic-760_0.jpg

8- http://storagezilla.typepad.com/storagezilla/2008/11/building-emc-atmos.html

9- Efficient Management of Consistent Backups in a Distributed File System. Jan Stender

Zuse Institute Berlin.

10- Jeff Darcy quote

11- Figures from Intel cloud builders guide: Cloud Design and Deployment on Intel Platforms.

12- www.emc.com/emcatmos

13- Backup Optimization “Networker inside” KS 2014.

14https://my.syncplicity.com/share/me01dgyznotfegg/6_Getting%20the%20most%20from%20your%20Data (from DPUG 2014 tour)

15- DPA 6_2 what’s New_v4.pptx (from inside EMC)

16- https://mainstayadvisor.com (ROI/TCO Tool output example)

17- Axcient-SMB-Backup-and-Recovery-Survey-Report-2014.pdf

18- gartner_-_the_broken_state_of_backup_(6-09)_(1).pdf

19- AST-0105542_15_minute_guide_English.pdf

20- Falconstor_sDataBackup_SO#034029_E-Guide_061311.pdf

21- http://www.gfi.com/blog/wp-content/uploads/2013/06/Backup-n-IT-admins-02_RGB72dpi.jpg

22- http://www.gfi.com/pages/research-brief-for-backup-survey-data.pdf

23- http://www.biztechmagazine.com/sites/default/files/tiny-uploads/2013/backup-cdw-infographic-760_0.jpg

24- http://searchstorage.techtarget.com

Table of Figures

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED "AS IS." EMC CORPORATION MAKES NO RESPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.