Think Global, Act Local!

From VTLs to Hyperconverged Secondary Data Storage

Prepared for Cohesity
May 2016

Think Global, Act Local!

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In the early 2000s, with the rise of the web and the new economy, we witnessed a radical shift in data access patterns, both in quantity and workloads. As a consequence, data protection changed drastically. At that time the major issues which needed to be urgently addressed were: shrinking backup windows and a growing amount of data to protect.

VTLs can no longer sustain continuous capacity growth as well as the flexibility required by businesses, users and new operational models.

Virtual Tape Libraries (VTLs), or in general disk-based backups, became very popular and they eventually displaced tape libraries as primary backup targets. One of the major benefits, among others, was that they could ingest large quantities of data quickly and store it efficiently for short time periods. The $/GB of this type of device never did shine, but this was considered an acceptable trade-off.

With Enterprises quickly moving towards virtualization, private cloud and DevOps, things have drastically changed again and VTLs can no longer sustain continuous capacity growth as well as the flexibility required by businesses, users and new operational models. In fact, users are demanding much more now than in the past, and data has been recognized as one of the most important assets for any organization. Consequently, data protection must change and become part of an active process to maximize its value. Once again we have two major problems to solve: on one hand we have complexity and lack of efficiency, especially at scale, and on the other hand there are pressing requests from business for timely information to help speed up internal decision making processes. Traditional VTLs simply don't have the answer.

Thanks to a holistic view of all ingested data, it is possible to build innovative features and offload the rest of the infrastructure on tasks for producing more information for decision making

From the infrastructure point of view, lower TCO is now fundamental. No-compromise scalability is a must and with it comes the opportunity of a higher consolidation ratio to take full advantage of deduplication. But better overall storage efficiency is only the first step for a simplified infrastructure with 1:N replication, cloud tiering, DR, and archiving requirements coming immediately after. Even more so, thanks to a holistic view of all ingested data, it is possible to build innovative analytics features and offload existing application infrastructure to produce more information for decision making.

Hyperconverged secondary storage systems are designed to meet these needs: strong data consolidation along with features for search, discovery and big data analytics. This allows the end user to understand how data is produced and consumed, improves security and makes data much more accessible and re-usable. With the growing success of hyperconverged compute infrastructures, with storage and compute resources collapsed in scale-out appliances, data protection must also evolve to become hyperconverged. In fact, a similar concept can be applied to a storage platform capable of integrating data protection with a comprehensive set of innovative data management services. The result is a tremendous simplification and improvement of operations while value and re-usability of data are maximized.

THE PROBLEM WITH VTL

Most common data protection appliances built their success on a single feature: deduplication. But deduplication went mainstream more than 15 years ago (thanks to companies like Data Domain). Besides the fact that this is just a feature, the very outdated design of these appliances is now demonstrating all its limits and constraints:

Even today, data protection is a pain for the vast majority of end users and a single purpose appliance, like a VTL, is no longer sufficient to cope with demanding data management needs.

- **The lack of scale-out capabilities** is truly dissonant with modern datacenters. With the introduction of VMware clusters and, more recently, hyperconverged architectures, you would expect every other component of your infrastructure to follow this trend. But it hasn't happened. In this particular case, scalability means more consolidation, thus better overall efficiency thanks to improved deduplication ratios. Newer scale-out architectures have the huge advantage of incrementing performance and capacity simultaneously when new nodes are added to the cluster.
- **Limited replica functionality** is another major problem with traditional VTLs. 1:1 replicas, once good for electronic vaulting, is no longer sufficient to cover the needs of distributed data protection infrastructures. And it gets only worse if the end user wants to leverage cloud storage, or on premises object storage, as a second tier or for long term retention backups.
- **No cloud integration.** Cloud can be much more than a second tier and can add flexibility to the data protection infrastructure. Unfortunately, most VTLs are "dumb" storage repositories and they don't have the resources or the intelligence to leverage cloud for disaster recovery nor the extensive search capabilities needed to implement archiving applications. Even though some VTLs can leverage external appliances to enable cloud-based long term retention, this approach introduces additional components to manage and cover only a few aspects of the whole data life cycle.
- **No fast recovery options.** Restoring data is the most important part of the backup process and the faster, the better. The obsolete design of some VTLs and their lack of integration with hypervisors, make it impossible to implement fast recovery operations, such as mounting backed up VMs in read/write mode directly from a virtual data store to speed up restores of entire virtual machines or single files.

Still today, data protection is a pain for the vast majority of end users and a single purpose appliance, like a VTL, is no longer sufficient to cope with current high demanding data management needs. End users are aiming for simpler and much more efficient infrastructures with the flexibility added by public and private cloud services. At the same time, they expect more from their data than in the past and want tools that will help them to do more with their data, and make it both faster and easier to administer.

MORE THAN JUST A DATA REPOSITORY

Today it is possible to rethink the entire data protection process. Instead of looking at backup as the end of the process you can realize many more advantages by moving it to center stage and building a comprehensive secondary data platform around it instead.

By rethinking backup, it is possible to re-use data efficiently without making additional physical copies. Also, by making it searchable, it is possible to enable many more applications and find additional information on the stored data.

By rethinking backup, it is possible to re-use data efficiently without making additional physical copies. Also, by making it searchable, it is possible to enable many more applications and find additional information on the stored data. In this scenario a deduplication-based VTL becomes a feature of a much more extensive data platform. It is simply an ingestion interface for a hyperconverged data storage system.

The characteristics of a hyperconverged storage system are essential to take full advantage of this approach:

- **Scale-out architecture** is the first step. It allows data deduplication and CPU resources to grow alongside capacity. This simply means linear scalability in terms of throughput to support backup and restore jobs, but also provides a pool of compute resources aligned with the amount of data under management to enable additional functionalities.
- **Multiple and sophisticated backup options.** End users want freedom of choice and it is highly likely that one or more backup software applications are already in place. A hyperconverged data platform, to be called as such, must support existing backup options as well as a native backup engine to leverage direct integration with hypervisors, and specific applications such as databases or primary storage systems.
- **Multiple protocol support.** NFS is the standard protocol exposed by many VTLs but now SMB3 can be even faster and it's a better choice for Windows servers and clients. Having the ability to expose NFS and SMB volumes enables instant restores of backup images, as well as providing native file sharing. This effectively increases the number of use cases, consolidation ratio and overall data reduction efficiency with the data center.
- **Efficient file system and snapshots.** Contrary to a VTL, which is just a mere repository, one of the key advantages of a hyperconverged storage platform is data reusability. An efficient and scalable file system is the keystone for consolidation but it's thanks to a modern redirect-on-write snapshot mechanism that it is possible to generate a large amount of data copies without impacting the overall performance of the system.

When secondary data (not only backups) is ingested and consolidated onto a single platform it is possible to implement a new set of features which can unleash much more value to the whole organization. For example, it could be possible to implement organization-wide search and e-discovery functionalities, run specific Big Data applications to look for specific patterns (helpful in discovering security and compliance issues of perform quicker audits) and much more. But again, these are only examples. Hyperconverged storage has a huge potential to offload many more data-driven tasks from the primary storage and the rest of the infrastructure.

COHESITY HYPERCONVERGED STORAGE

It is estimated that secondary storage counts for around 80% of the entire data footprint and 50% of total expenditure in most organizations[1], and these numbers don't include data copies for backup and disaster recovery. This is a major problem which is hard to solve, because most of this data resides in different silos, like traditional VTLs, and it is quite complex to manage as a whole or to take adavantage of the benefits made possible by consolidation.

Cohesity, thanks to the experience of its founder Mohit Aron (former CTO and founder of Nutanix and co-creator of the Google File System) and his team, have designed a hyperconverged system based on a web-scale file system. Contrary to the vast majority of the hyperconverged systems in the market, Cohesity has chosen to build a software solution to optimize all secondary storage needs. For obvious reasons of ease of use and support, it is sold as a physical or virtual scale-out appliance (also available on Amazon AWS).

Cohesity C2000 series can easily replace a traditional deduplication-based VTL thanks to its global deduplication capabilities, the native file system protocols and the integration with the most common backup software applications. At the same time this is just one way to ingest data in the Cohesity appliance; other methods (including an innovative backup scheduler integrated in the system) allows backup of VMware-based infrastructures, physical servers, as well as major database applications. Once all backups and other data are consolidated in a single platform it is possible to efficiently share multiple copies of data volumes to external applications thanks to a clever snapshot technology.

Contrary to what usually happens with a VTL, all data saved in a Cohesity cluster can be shared, used for test/dev, searched and analyzed.

Contrary to what usually happens with a VTL, all data saved in a Cohesity cluster can be shared, used for test/dev, searched and analyzed. Basic reporting and extended Google-like search capabilities help to find data quickly, while the large amount of CPU resources available on the system allow you to run custom Big Data applications with ease. At the end of the day, Cohesity C2000 can be considered an enabler to build an enterprise data lake, avoiding dark data issues and helping to exploit maximum value from the data available in the organization.

Traditional VTLs impose too many limits not only because of the lack of consolidation or the absence of advanced functionalities but also for their rigidity when it comes to distributed infrastructures and cloud integration. While both Cohesity and VTLs can run as VMs on public cloud; Cohesity architecture also allows you to leverage external object storage repositories (like Amazon S3 for example), as well as tape storage, to let the end user choose the best media for backup and archiving needs depending on retention policies, security and costs.

Hyperconverged storage also addresses the cost issues of unsustainable data growth in secondary storage, thanks to its efficiency, high consolidation ratio and data re-usability.

[1] Juku.it: The future of storage is now. Consider a two tier storage strategy. - March 2016 - http://goo.gl/kWsSmd

BOTTOM LINE

Similar to hyperconvergence sweeping through the computing and primary storage infrastructure, similar benefits can be obtained by adopting a hyperconverged secondary storage platform.

By collapsing many components of the stack in a single scale-out appliance it is possible to do much more while simplifying processes and operations.

By collapsing many components of the secondary stack into a single scale-out system, it is possible to do much more with your data, while simplifying processes and operations. It's not only about driving down TCO thanks to better overall efficiency, but there is also the opportunity to improve many other aspects of data and business processes. In fact, data consolidation and efficiency, coupled with analytics tools and data management will be the future trend in the data center.

A hyperconverged storage system, like Cohesity, provides the same functionality of a VTL allowing a fast and seamless migration process. The difference is that once data has been saved on the hyperconverged appliance it can be reused, analyzed and searched, helping to fully exploit all its potential.

JUKU

WHY JUKU

Jukus are Japanese specialized cram schools and our philosophy is the same. Not to replace the traditional information channels, but to help decision makers in their IT environments, to inform and to discuss the technological side that we know better: IT infrastructure virtualization, cloud computing and storage.

Unlike the past, today those who live in the IT environment need to be aware of their surroundings: things are changing rapidly and there is a need to be constantly updated, to learn to adapt quickly and to support important decisions - but how? Through our support, our ideas, the result of our daily global interaction on the web and social networking with vendors, analysts, bloggers, journalists and consultants. But our work doesn't stop there - the comparison and the search is global, but the sharing and application of our ideas must be local and that is where our daily experience, with companies rooted in local areas, becomes essential in providing an honest and productive vision. That's why we have chosen: "think global, act local" as a payoff for Juku.

AUTHOR

Enrico Signoretti is an analyst, trusted advisor and passionate blogger (not necessarily in that order). He has been immersed in IT environments for over 20 years. His career began with Assembler in the second half of the 80′s before moving on to UNIX platforms until now when he joined the "Cloudland". During these years his job has changed from highly technical roles to management and customer relationship management. In 2012 he founded Juku consulting SRL, a new consultancy and advisory firm deeply focused on supporting end users, vendors and third parties in the development of their IT infrastructure strategies. He keeps a vigil eye on how the market evolves and is constantly on the lookout for new ideas and innovative solutions. You can find Enrico's social profiles here: http://about.me/esignoretti

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources Juku Consulting srl (Juku) considers to be reliable but is not warranted by Juku. This publication may contain opinions of Juku, which are subject to change from time to time. This publication is covered by Creative Commons License (CC BY 4.0): Licensees may cite, copy, distribute, display and perform the work and make derivative works based on this paper only if Enrico Signoretti and Juku consulting are credited. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Juku consulting srl has a consulting relationship with Cohesity. This paper was commissioned by Cohesity. No employees at the firm hold any equity positions with Cohesity. Should you have any questions, please contact Juku consulting srl (info@juku.it - http://jukuconsulting.com).