Data Protection – Re-Imagined, Inclusive and Business-Savvy

A Perspective by

Jon Toigo

CEO, Toigo Partners International

Chairman, Data Management Institute

24 January 2013

ABSTRACT

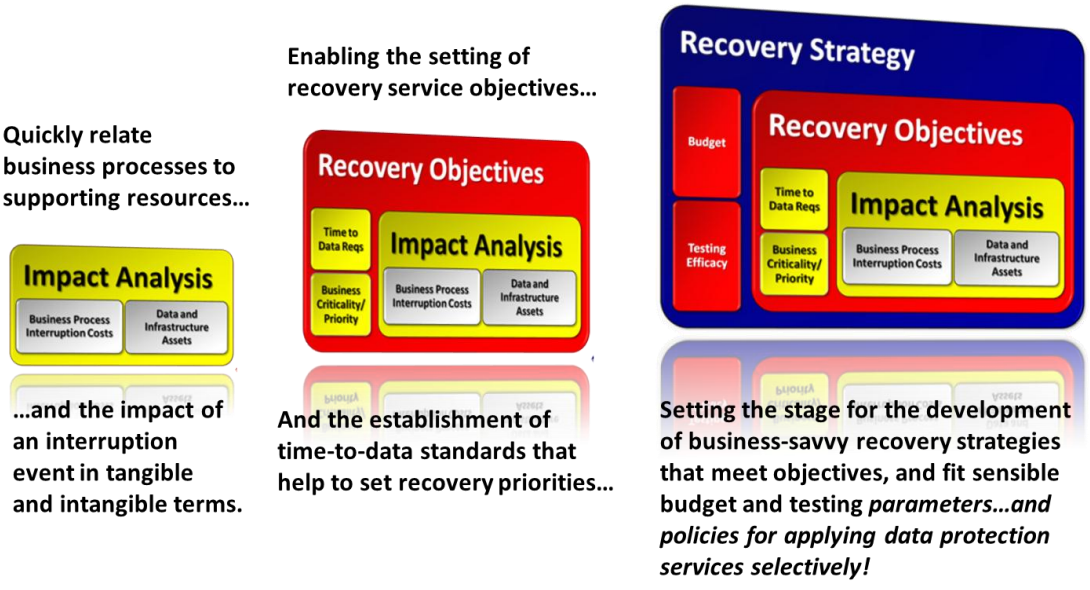
The dependency of business on automated systems to support their critical processes has never been greater than it is today. Even a short term interruption in access to data required by applications and decision-makers can have disastrous consequences. This situation underscores the requirement to develop, implement and maintain a comprehensive data protection strategy. Such a strategy must be guided by the recovery requirements and timeframes of business processes themselves, must be inclusive of the many tools and techniques available, and must apply appropriate protective services to data in a business-savvy way.

INTRODUCTION

Aside from personnel, the most irreplaceable asset of any business organization is its data. Data is both the fuel – and the product – of the automated systems that support virtually all business processes today. Protecting data, and ensuring continuous access to it, are among the most important responsibilities assigned to contemporary business information technologists.

The traditional approach to protecting data is to develop and implement a data protection strategy. Such a strategy typically leverages one or more techniques intended to safeguard data against corruption, loss, unauthorized disclosure, or denial of access. Ideally, protection techniques are assigned to specific data assets supporting specific business processes following careful analysis of the relative criticality of, and requirements associated with, each process. By itself, data is just an anonymous set of 1's and 0's; data inherits its importance like so much DNA from the business process that it supports.

While this paper does not dwell on the analytical effort that must be taken to determine the recovery priorities of business processes (with their associated infrastructure and data), it is clear that this effort must be undertaken prior to developing a data protection strategy. As illustrated in the figure below, a preliminary business impact analysis must be performed to identify what infrastructure and data are associated with a given business process and what the impact would be of an interruption in the services and access to data associated with that process.



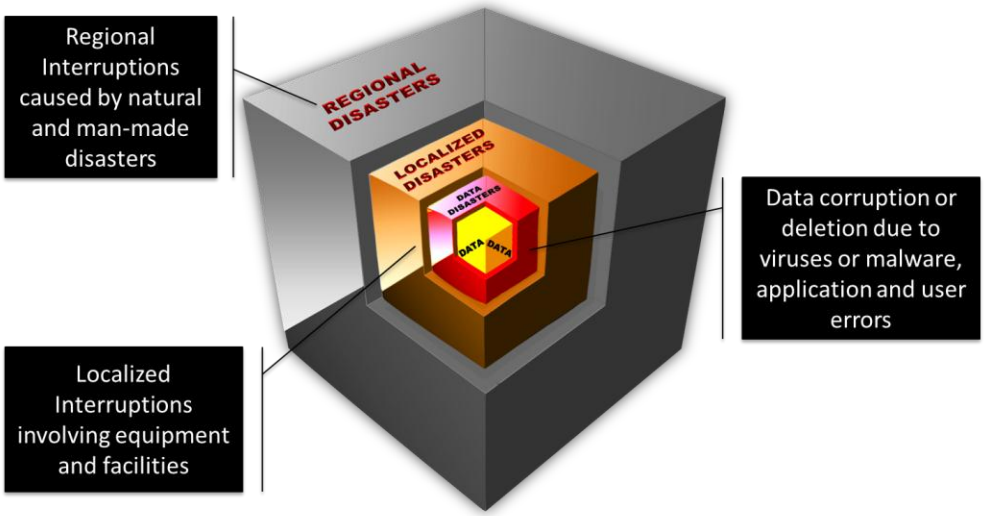
The results of the impact analysis drive the setting of recovery objectives that define the criticality and restoration priority of the subject process and its "time-to-data" requirement (sometimes called a recovery time objective). This in turn provides guidance for the definition of a strategy that applies data protection and recovery services and techniques to facilitate renewed access to data within the required timeframe and in a manner that is "business-savvy" – that is, a manner that fits budgetary realities and minimizes the long tail cost of continuity planning, the cost of testing.

This analytical work is arguably the "heavy lifting" of data protection planning. It answers the first two of three questions that must be addressed by any data protection plan: what data needs to be protected and where is that data located. Unfortunately, most discussions of data protection launch immediately to the third question: what is the best way to protect the data. In practice, that question cannot be addressed effectively without information drawn from answers to the two prior questions.

Of course, there is truth to the assertion that protecting data is a simple matter: make a copy of the data and move the copy off-site so it is not consumed by a disaster that befalls the original. All data protection strategies provide protection as a function of redundancy, representing a contrast with IT infrastructure recovery strategies, which can be based either on redundancy or replacement. Replacing data following a disaster event resulting in its corruption or deletion is not a realistic strategy; you need a redundant copy stored out of harm's way.

While there is general agreement on this concept, different vendors seek to promote different technologies for making the safety copies of the data, each promoting their wares as the one true path to data continuity. The hyperbole around data protection solutions can be frustrating, especially given the fact that more and more planners are finding it necessary to implement a "defense in depth" strategy to protect data assets, which leverages different data

protection tools to provide protection against different types or categories of threats. The following illustration represents defense in depth.



Defense in depth acknowledges that there are different threats to data emanating from different sources. The more critical the data, the more defensive services planners seek to use to protect the data. For example, continuous data protection (CDP) or snap-shots may be used to protect against data corruption or deletion threats that occur close to data creation and data storage processes. Additionally, disk mirroring may be implemented as a hedge against localized interruptions such as a storage array failure or a facility outage. Finally, techniques such as tape backup or some sort of wide area network (WAN) based replication scheme may be implemented to protect against disasters with a broad geographical footprint – such as hurricanes, floods or other natural or man-made disasters.

Defense in depth strategies treat various data protection tools and techniques as "services" that are applied individually or in combination to meet the protection and recovery requirements of data on a business process by business process basis. Policies are defined to associate a subset of protective services with specific data assets.

In an ideal, service-oriented, data protection scheme, data from specific business processes/applications are assigned service policies drawn from a menu of available services, delivered via a variety of hardware and software tools, all in a highly manageable way. This concept is illustrated below.

Unfortunately, vendor marketecture promoting specific hardware or software data protection products too often dominates the discussion of data protection architecture, leaving planners with a *kluge* (a distressing combination of mismatched parts) instead of a coherent, inclusive and business-savvy strategy.

A few vendors, such as Tributary Systems, are stepping up with integrated service models for data protection. Here are our observations about this developing technology space.

NEEDED: A DIRECTOR

As noted above, in most discussions of data protection, vendors usually want to go directly to a debate over the best way to make a safety copy of your data. Things quickly devolve into a discussion of tape backup versus disk mirroring and WAN-based disk replication.

Tape advocates argue that they are less expensive than disk based approaches, less prone to bit errors and silent corruption problems that are increasingly associated with disk storage, and agnostic about the target array that will be used to re-host data. All of this is true.

Recent studies reveal that “silent corruption” (non-recoverable bit errors) impacts data stored to disk media with significantly greater frequency than data stored to tape media. These findings not only characterize tape as a more robust medium for data protection, but also set the stage for future storage architectures that will leverage tape as a production file store – at the very least for infrequently accessed files.

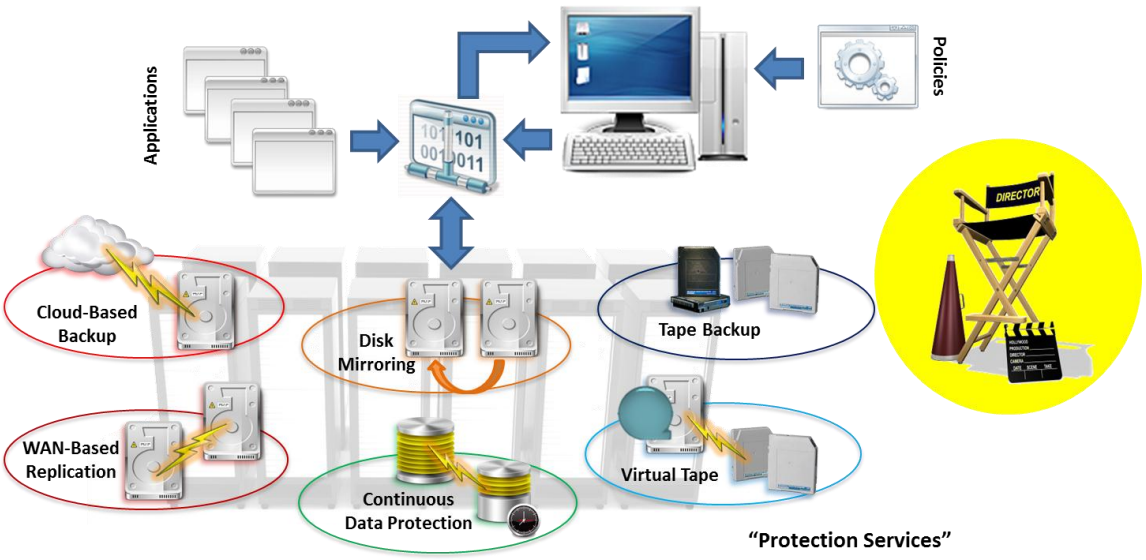
By contrast, disk storage advocates argue that their solutions are simpler to deploy and operate, provide instant recovery of discrete files and work well across modern high speed networks. Again, most of these arguments are true.

However, from the standpoint of data protection planning, these points are largely irrelevant. The selection of tools for data protection should be guided by practical criteria that include:

1. The criticality and restoral timeframe or time-to-data requirements of the data (as "inherited" from the business process that the data serves);
2. The "test-ability" of the strategy (does it avail itself of easy and/or ad hoc testing with minimal preparation and expense); and
3. Available budget and other cost-of-ownership considerations.

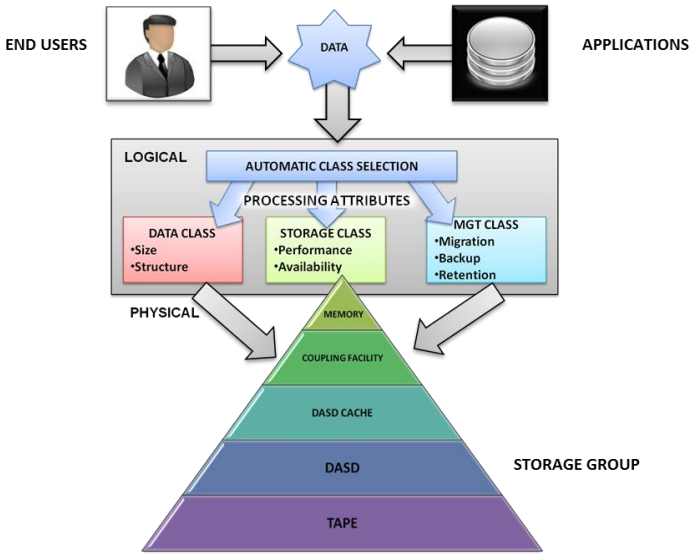
In most organizations today, a diverse assortment of data protection tools and services already exist. Most are underutilized or implemented in a way that isolates the services that they can provide to only specific data.

Part of data protection planning involves the inventorying of tools that are currently deployed, with an eye toward finding a way to leverage them in a more intelligent way. In general, planners find that data protection services delivered by hardware-agnostic software tools are more flexible than functionality delivered on proprietary hardware such as storage array controllers. Hardware-centric tools – on-array mirroring, for example – often have scaling limitations, or they are subject to the vicissitudes of vendor roadmaps, or they are limited only to a certain brand or model of hardware as a replication target.



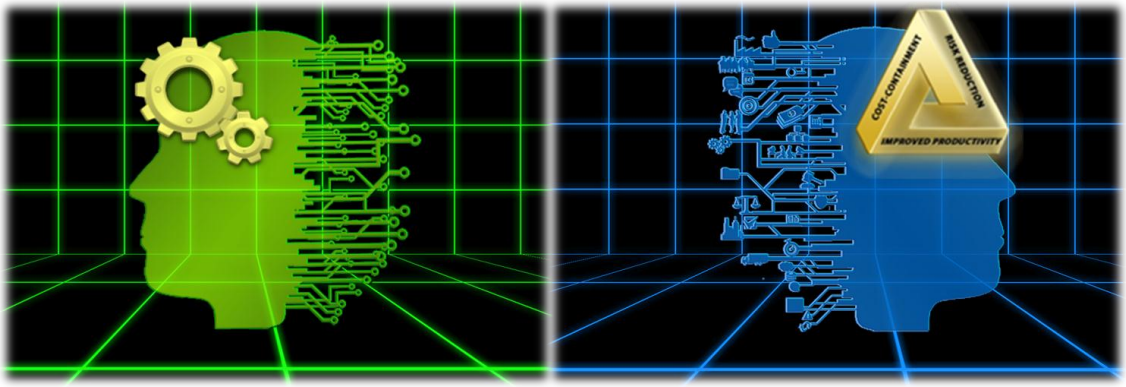
Ultimately, what is needed is a "director" – a software or hardware component that coordinates services to create policies that can be applied to discrete data sets. This is similar to how data management services were delivered in the mainframe data center some 30-odd years ago. Readers may recall that storage management and data management were provided on mainframes using systems managed storage and hierarchical storage management (SMS and HSM) software utilities that worked across storage infrastructure resources.

Systems Managed Storage (SMS) & Hierarchical Storage Management (HSM)



Today, of course, most data centers do not have a single dominant vendor, whose *de facto* technology standards and extensive product portfolio enable a homogeneous and centralized management scheme. Even in IBM mainframe shops, operating systems have evolved, and their data and infrastructure management services have become more "open" to accommodate vendor technology heterogeneity and distributed computing infrastructure. Burgeoning initiatives from IBM like zEnterprise, which seek to unify the management of both the mainframe platform and the distributed computing platforms, have not yet established a centralized, policy-based method to administer and allocate the diverse infrastructure assets and services deployed as part of the corporate IT plant.

Again: what is needed is a "director" – software intelligence that can be used to create defense in depth policies and associate them with selected data so that those data are availed of the protection and recovery services that they require, whether those include cloud-based backup, WAN replication, disk mirroring, CDP, tape backup, or virtual tape.



Such a technology offers a compelling value case from both a technology and a business perspective. From a technology efficiency perspective, a director would streamline data protection service delivery, plus provide a centralized means both to monitor services and to test services in an automated way.

Plus, a director would also provide business value: reducing the labor requirements for data protection and continuity planning, improving the economic efficiency of protection services, reducing or eliminating hardware vendor lock-ins, etc. In short, a data protection director would help contain costs, reduce risks and improve productivity: the three components of a business value case, which is needed these days to justify just about any IT initiative.

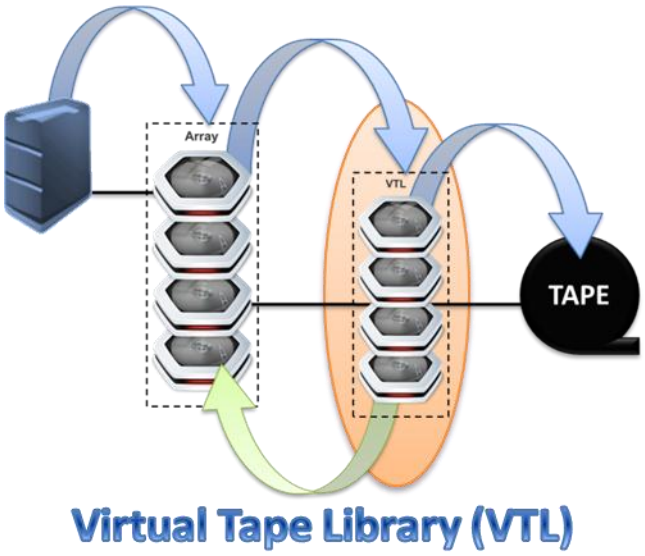
WHERE IS A "DATA PROTECTION SERVICES DIRECTOR" BEST HOSTED?

A "data protection services director" embodies a persuasive business value case that can be leveraged to justify to management its acquisition. As a practical matter, a decision needs to be made about the architecture of the solution, and more specifically, where to locate such a storage director service.

A data protection services director refers to a software-based functionality set, so it could theoretically be placed anywhere. However, since its role is to associate per policy the data from application workflows with an appropriate set of hardware- and software-based data protection services, it is conceptually playing the role of a storage router.

This suggests a location for the director that is in the data path, in the storage infrastructure itself. In many large enterprise shops today, the location in the data path that most data traverses is the virtual tape library, or VTL. VTLs are commonly deployed as the second tier of disk where a data copy is temporarily held pending its write to tape, to a cloud or two another set of disks interconnected to the primary via a Wide Area Network.

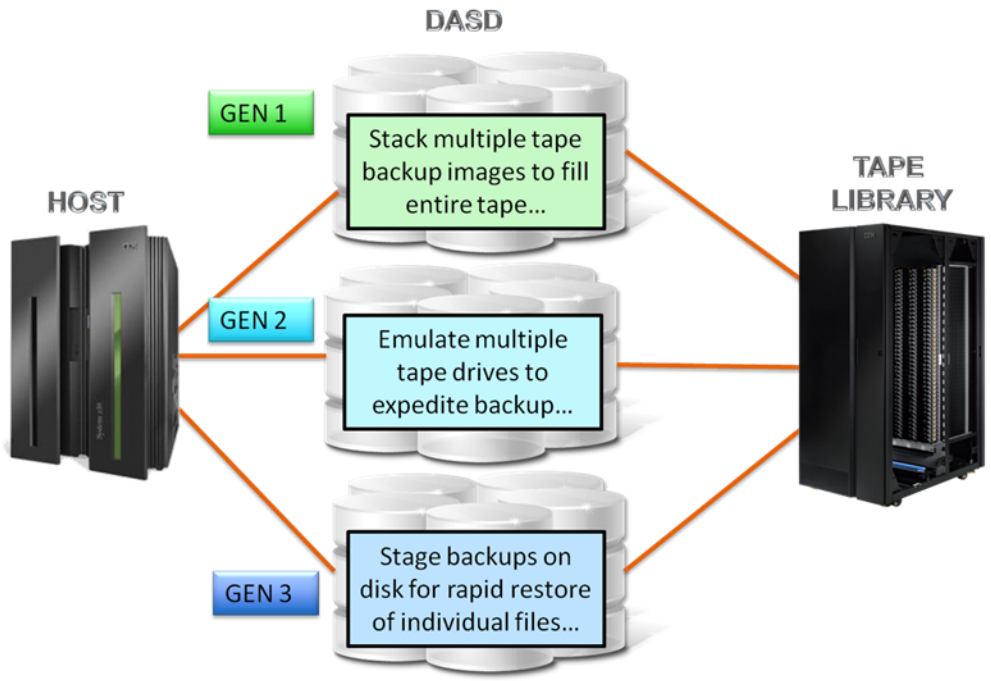
For a growing number of companies, VTLs are "way stations" – the second "D" in "D2D2T" (disk to disk to tape) platforms used to host data copies temporarily for the purposes listed above and to enable the expedited restore of individual files.



Today's VTLs are third generation fixtures in data protection, with over 30 years of history in both mainframe and open systems shops. Originally, VTLs were used to stack tape jobs until sufficient data was accumulated to fill a cartridge – a capability required to compensate for the

inefficient use of tape capacity in early mainframe file systems. Generation 2 VTLs were deployed to provide more tape drives (software-based emulations, actually) with the goal of shortening backup time requirements and to expedite backup operations.

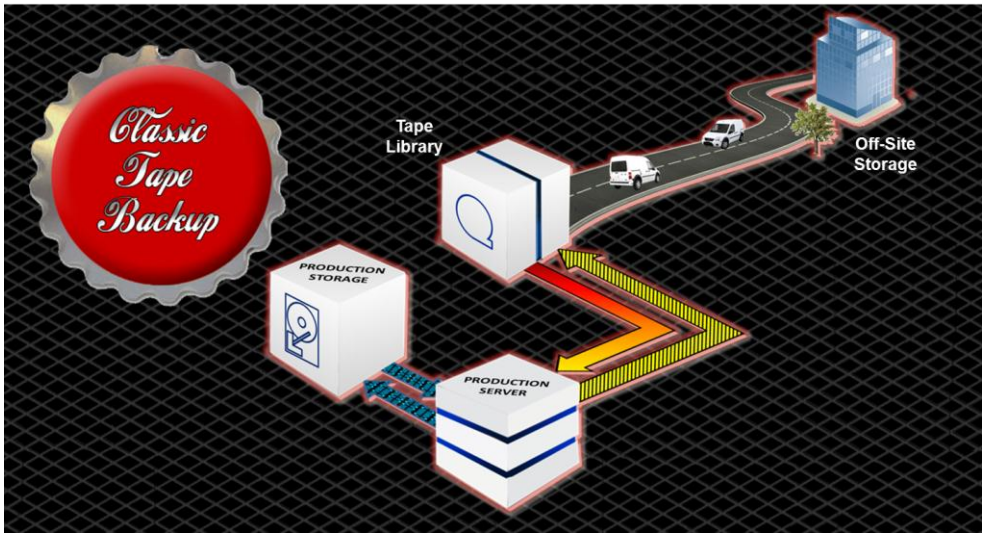
Today, VTLs are used as part of a D2D or D2D2T strategy – or to facilitate any other purpose a vendor chooses. Today, a VTL simply provides a location where "services" can be applied to backup data: encryption, compression, de-duplication, mirroring, CDP, WAN replication, snapshots, and of course tape backup.



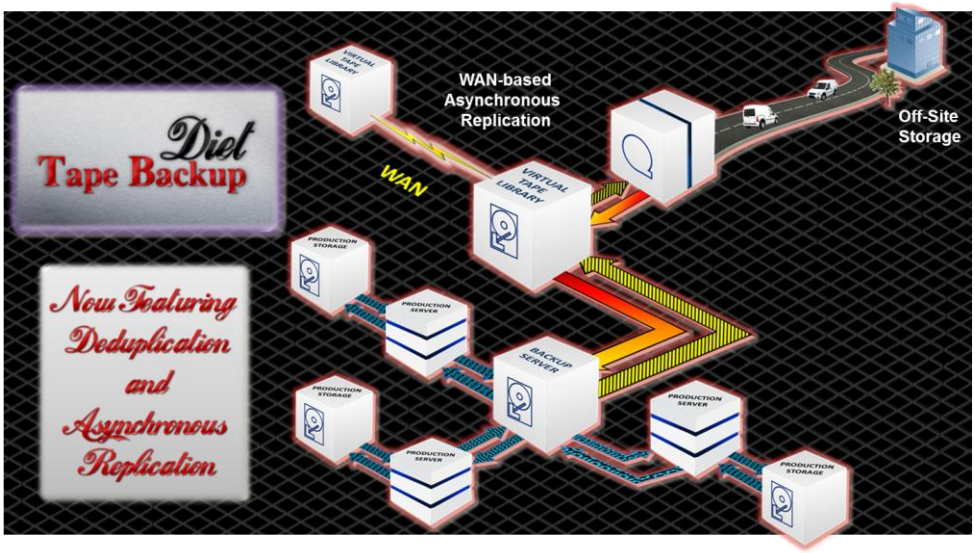
In data centers where a VTL has been established, it is a pretty good candidate for locating a "storage director" – a location where data protection services can be applied to data on a selective basis via predefined policy. Most VTLs (like most storage arrays) are essentially software components running on a commodity server that collectively acts as a "smart controller" acting on commodity disk drives. Given the logical position of the VTL, in the data path, and its capability to "touch" all data, it may provide the appropriate host for director services.

Asserting this architectural case, however, should not be misconstrued as an endorsement of recent marketing campaigns by some VTL vendors suggesting that tape backup is no longer a valid data protection service. The simple truth is that tape is still the most widely used method

for data protection. Virtually all IT operators know how tape works and its technical and economic advantages.

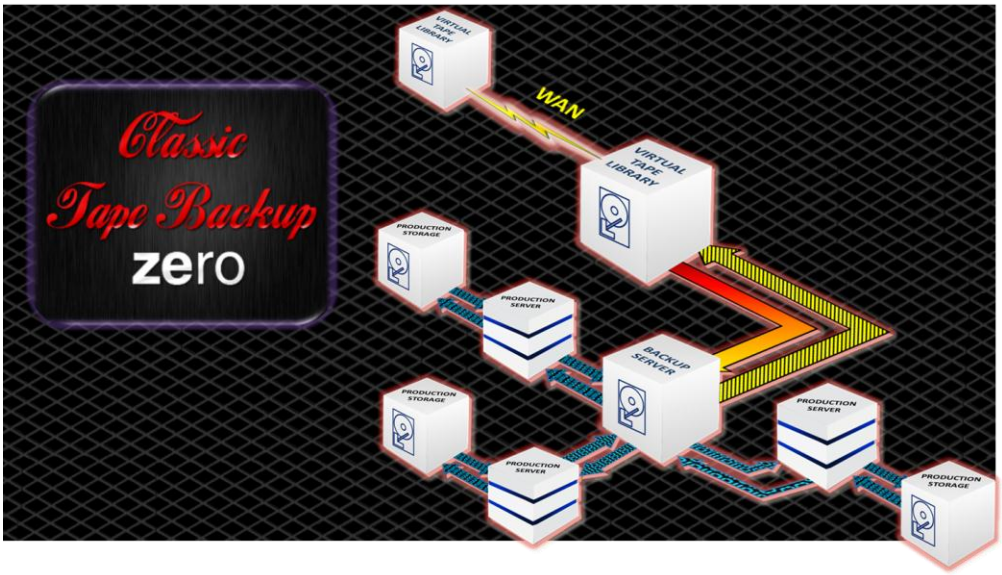


Over time, backup servers, then VTLs, were introduced to add value to tape backup. Placed in the data path, the VTL provided a location for services to be applied selectively to data that required them. Moreover they provided fast restore of individual files and later integrated WAN-based replication that could be used with data supporting applications that required high availability (e.g. “always on” apps).



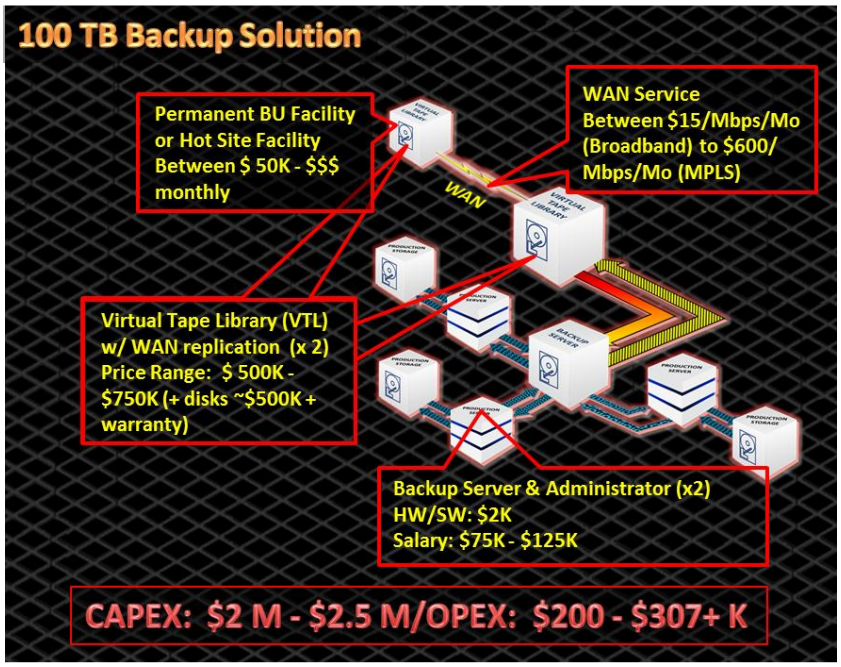
The VTL's WAN based replication features helped to streamline tape backup and restore. Adding features like de-duplication helped to reduce the amount of disk capacity required to store local copies of files. Bottom line: The strategy augmented tape backup; it didn't replace it.

Unfortunately, this strategy has recently been co-opted by some VTL vendors to support a "Tape Zero" strategy, to coin a phrase. Certain vendors of VTL appliances have become vociferous advocates of proprietary disk-only data protection schemes – recommending the replacement of tape backup and other data protection techniques with VTL to VTL replication schemes only (especially when the scheme requires all gear to be purchased from the vendor).

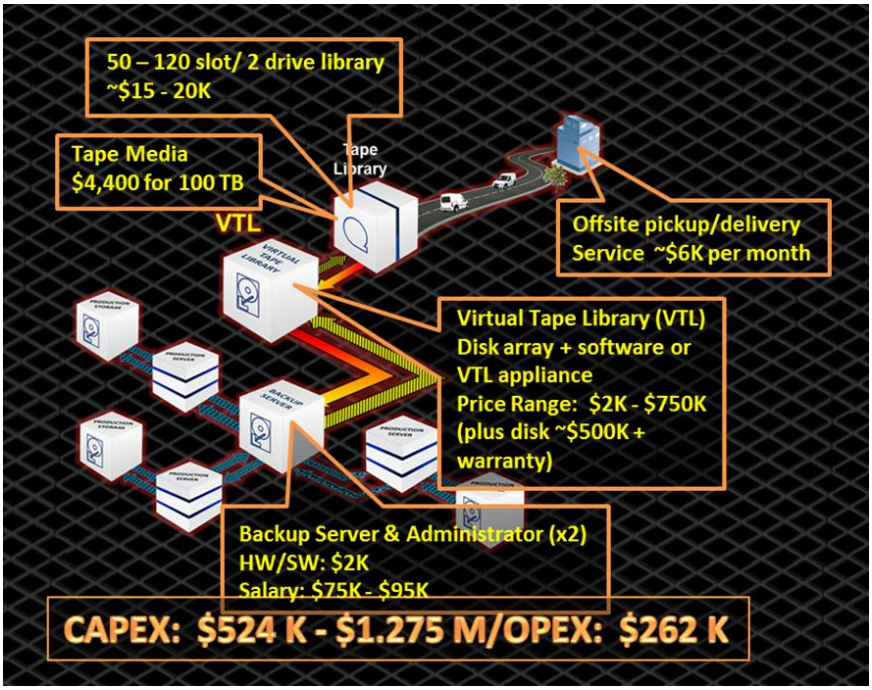


While appealing in terms of its simplicity, such a strategy makes little financial or operational sense. Among its many nontrivial technical issues: unpredictable WAN service levels caused by latency and jitter in a wide area network can cause "data deltas" – differences between local and remote data – that can make the remote data copy useless when recovery is required.

From an economic perspective, the WAN links required to transport data to a remote VTL can be very expensive. Moreover, the remote destination needs to be pre-defined and pre-equipped with suitable gear, which usually entails a lock-in to a particular vendor's products. The illustration below reflects the details of a cost estimate recently prepared on behalf of a Global 2000 firm identifying the CAPEX and OPEX costs of such a "Tape Zero" strategy.



By contrast, a backup solution leveraging tape technology proved significantly less expensive to acquire and operate.



Cost comparisons aside, the important point is that not all data supports mission critical, always on business processes. Hence, this data does not need or merit WAN-based replication as a means for data protection. Where such a data protection service is not consistent with the time-to-data requirements of the protection target, a VTL to VTL replication approach is not justified as a solution, a policy, or an expense.

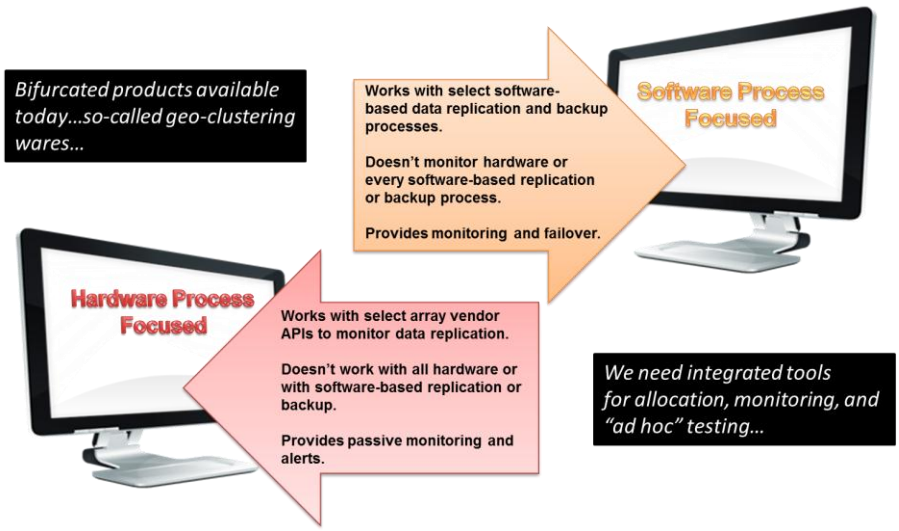
ANOTHER ROLE FOR THE DIRECTOR: STREAMLINE TESTING

There is no one-size-fits-all solution for data protection. A successful strategy typically involves the assignment of a combination of data protection services to the data assets of a given business process based on recovery objectives, technology availability, and budget.

Given this fact, another dilemma confronts planners: that of coordinating the allocation of multiple services to multiple targets, of monitoring the operation of such services, and the challenge of testing and validating strategy performance over time. Ideally, a data protection service director would provide a software dashboard enabling "near real time" monitoring of data protection services irrespective of how (via software or hardware) the services are provided.

While some progress has been made on the development of software with this purpose, most data protection service monitoring products available today are either "software-centric" (providing info only on software-based data protection processes under a given vendor's product umbrella) or they are "hardware-centric" (ignoring software-based data protection activity and reporting only hardware centric replication). In short, our tools for connecting up heterogeneous hardware and software data protection service providers are as limited and bifurcated as our tools for monitoring services coherently.

This situation adds to the cost and inefficiency of data protection overall, and of techniques based on disk to disk mirroring and replication in particular. In the case of mirroring, the only way to validate that the right data is being copied in a mirror is to "break the mirror" – that is, (1) quiesce applications that are writing data, (2) flush cache memories so that all data is written to the production disk, then replicated to the mirrored disk target, (3) and, finally, perform a file by file comparison to ensure that the data that needs to be replicated is actually being replicated. Once the operator has confirmed that the mirror is good, he/she endeavors to restart the mirroring process…usually with fingers crossed. In most cases, mirrors are not checked because of the hassle entailed in breaking and restarting the mirroring process.

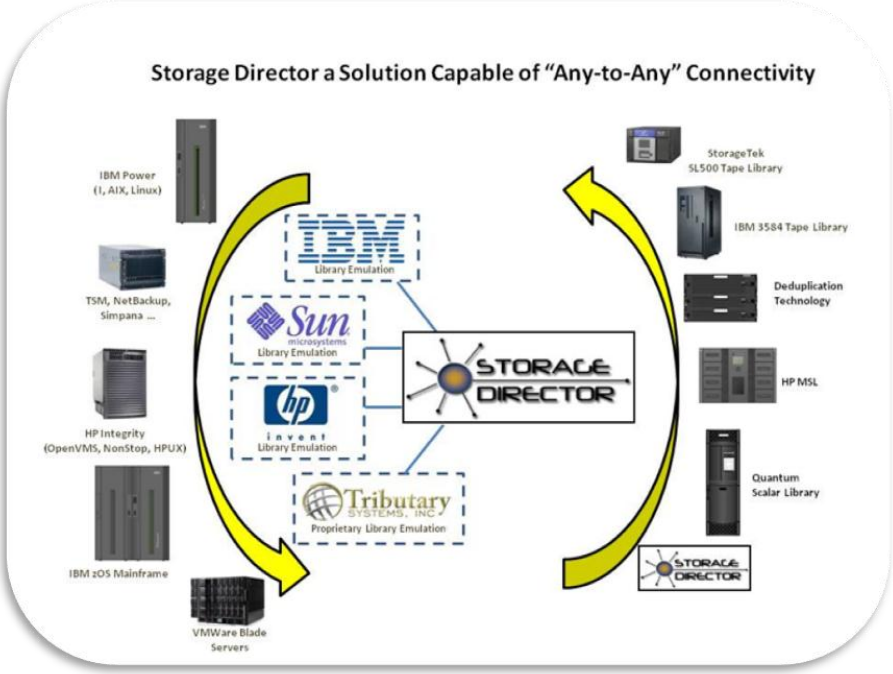


In a perfect world, the data protection service director would provide the ability not just to manage the allocation of services and their on-going monitoring, it would also provide the means to test data protection services on an ad hoc basis and in both simulated and disruptive ways. In fact, the capability of a solution to provide ad hoc testing would dramatically reduce the workload for disaster recovery testing by eliminating data recovery tests. Testing is the long tail cost of continuity planning, and whatever can be done to minimize and streamline testing has a significant impact on over all planning cost.

CONCLUSION: WHITHER THE DATA PROTECTION SERVICES DIRECTOR?

Tributary Systems' Storage Director™ is an innovative hybrid of a traditional VTL and a data protection service engine. It fulfills today the first requirement to serve in the role of a data protection director by providing holistic support for a broad range of data creation platforms and for data storage and data protection service engines.

The platform is also being steadily enhanced to enable it to allocate to data the appropriate data protection services and to monitor the delivery and operation of those services. Going forward, the platform holds out possibilities for use in an expanding data management services universe, providing routes for deep archive and for active archive using the Linear Tape File System in conjunction with tape for mass archival storage of infrequently accessed data.



With Storage Director, customers can realize cost-containment and risk reduction advantages as well as improvements in terms of uptime and productivity. With their Storage Director, Tributary Systems is becoming the breakaway product in the data protection space. We wholeheartedly encourage readers to learn more about the product.

