

WHITE PAPER

The Promise of Virtual Storage: Transforming IT into a Service

Sponsored by: EMC

Benjamin Woo
May 2010

EXECUTIVE SUMMARY

Information Technology (IT) is being redefined as a service. This requires, among other things, a more flexible, virtualized infrastructure in which virtual storage complements virtual servers to liberate information from its physical systems. It is well established that virtual servers bring agility, efficiency and numerous other benefits. EMC's virtual storage strategy and initial execution (VPLEX) extends many of the server/processing benefits to storage.

At the core of virtual storage is EMC's distributed cache coherency, which integrates knowledge of the data across all platforms – local, metro, global, and in the private cloud. Information can be distributed over distance, enabling storage to be federated. Organizations no longer have to anticipate every eventuality by pre-positioning all the information to another location. Virtual storage enables a more dynamic and flexible infrastructure that in a great many use cases, offers profound positive changes for IT and business.

SITUATION OVERVIEW

As the concept of IT resource abstraction took hold, IT professionals saw that virtualization could be adapted to many computer resources, ranging from hardware to operating systems to applications. Information, however, was in virtually all situations, bound to the physical storage hardware. Now it is possible to federate data over distance, to abstract it. In this context, data can be stored at a multitude of physical locations, yet be seen and accessed across the IT platform as a single virtual storage resource.

Federation extends the server/processing metaphor to storage. Data is abstracted from the physical hardware system, making time, location and distance virtually irrelevant for users. Even more important, the combination of virtualization and federation allows IT to fully automate the placement of data to suit any business configuration or usage scenario, making access to data seamless, agnostic, coherent and transparent between storage arrays. In so doing, the use of information becomes realized in a value proposition that has, up until this time, been a context only imagined.

The Evolution of Storage Virtualization

Unquestionably, virtualization has reduced IT costs and created a new model for resource allocation: a virtual server meets and exceeds every tenet of Moore's Law of computing performance per unit. Data virtualization was considered for some time as the final piece of the virtual map, but then storage professionals began to contemplate if it were possible to abstract the logical storage from physical storage, which would free information, making it not just data-independent but virtually independent as well.

If processing resources can be distributed and placed where specific needs and/or applications are most needed, then it follows storage should be deployed to make data instantly available, updated and kept in sync in real time, and completely time- and distance-agnostic. Virtual Storage would mean the use of data would transcend bandwidth, cache coherency, and geographic latency.

The user and data can be anywhere, and the data is current, consistent, and up-to-date. The user immediately has the information needed without having to think about the system or the data at all.

Distributed Cache Coherency

Since the introduction of fixed disk drives in the late 1950s, storage management has been a pretty straightforward activity: store data democratically, uniformly, logically, without much regard for frequent access or use or, for that matter, prompt access. For the user, the issue in a word was latency. Caching was a quick step forward, improving data access speeds and making older, slower fixed disk drives perform better. Solid-state cache doesn't involve moving parts, such as a spinning disk and read/write heads, and serves many useful storage purposes, chief among them to hold frequently accessed data at the ready, without discrimination or latency.

Caching came to the fore as a solution to a new problem once client-server and distributed computing came into broad use. File management and versioning needed to be managed, and the cache was ideal for this kind of change management known as cache coherency. Caching is being redefined again as *distributed cache coherency* as we move into the virtual environment. Storage performance is enhanced by the use of flash, to be sure, but more importantly by dynamic distribution across all types of storage devices. In both cases data is closer to the user - and the closer the better. Closer means data is information. Most aspects of working with information - high availability, fast recovery, version control, backup and security - are improved because the virtual cache is so close to the user.

It has been established that cache addresses two important issues. One, of course, is access speed. The other is the ability to prioritize the data access request. A fixed disk has no such intelligence, while commonly used or reused data stored in cache is always more readily accessed. With virtual storage, the virtual cache becomes a distributed cache as well, and both speed and access are greatly improved. Moreover, distributed cache coherence, which is by its very nature distance-agnostic, requires new algorithms to ensure data integrity and versioning control. The following

have been identified as the key elements of distributed cache coherence that are essential to its functionality:

- ☒ Transparency of storage arrays across the system and from any distance
- ☒ Virtual access anytime, anywhere, any system
- ☒ Data movement tracking for file management and versioning
- ☒ Scalability to suit enterprise, geographic location, workgroup, or individual needs
- ☒ Aggregated capacity management across both physical and virtual devices
- ☒ Application and data mirroring regardless of platform.

Each element, in and of itself, does not seem a particularly new or difficult problem to solve. However, taken in the aggregate, and given the design for implementing global storage federation, the problem-solving becomes more complex. The objective - to create a global dynamic data pool of information resources and the applications required to work with that information that anyone, anywhere, can access with the speed and agility of accessing the machine at their desktop - is a necessary goal if we are to achieve truly virtual IT systems.

The benefits accrue at every level: time, money, productivity and a sharper competitive edge. From a technological point of view, federation produces a globally consistent resource for information access, employing both cache and storage resources at high performance levels. Distributed cache coherency makes just about any task faster and more intelligent, while assuring high data integrity. The most practical approach seems to be to begin locally, then move toward global and beyond.

Considering Diverse Enterprise Workloads

Differing workgroups and applications have different data workloads across time and space. Distributed servers began addressing these issues, but could not fully address data stored in various remote locations. An early solution was for users to store the data they used most frequently on their own desktop machines. Over time, local disk farms or storage devices attempted to manage data for geographically specific groups. Virtual servers made it possible to address data management issues by making remote data available and creating more workload adaptability, but nevertheless users, more often than not, were still stuck using the same limited local data set. If others needed that data, the data set was replicated for another localized data center. Now redundancy, file management and version control problems were magnified.

IT organizations that have accomplished the transition from standalone to virtual servers agree it has been transformational, and now the same evolution is possible with storage. Early on, EMC's VMware virtualization and Symmetrix VMAX demonstrated what could be accomplished:

- ☒ VMware employs technology that enables multiple operating systems to run simultaneously and independently on the same industry-standard server or workstation as virtual machines. Live applications can be accessed or deployed across diverse systems without business disruption. VMware virtual servers allowed users to see distant information resources as if they were local, and represented a critical step forward in understanding the need for virtual storage.
- ☒ EMC Symmetric VMAX, taking advantage of the virtual data center, provides high levels of storage – up to 2 petabytes – for workload consolidation and scalable application performance. Based upon the Symmetrix storage platform, this product addressed the need for higher storage availability and performance for critical applications across physical and virtual servers, facilitated by VMware.

Another revelation in understanding virtual systems was that servers may, and often do, fail, but when that occurs storage devices are usually still up. Therefore, in addition to improving system response and making information more available, federating storage virtualization and integrating distributed cache coherency could redirect data to an online server without interruption. Not only is such a system more reliable and productive, but it significantly reduces the recovery point objective (RPO) and the recovery time objective (RTO). While caching data may sound simple and practical, moving it at a distance has remained one of the most perplexing problems for IT.

VPLEX: The Emergence of Federated Data Storage

Server and application caching was once simple and practical, but it was dependent on local proximity. It has always been much easier to move servers, or disperse applications, than to distribute data over distance. Solving the problem meant rethinking the fundamental computing model.

As EMC explored the concept of federation and began developing a deeper understanding of the problems surrounding bandwidth, latency, and distributed cache coherency, several new perspectives emerged:

- ☒ Think of IT as a service, in much the same way as other distributed services
- ☒ Build common datacenters that replicate identical machines and processes
- ☒ Create virtual storage models based on virtual server models
- ☒ Solve the problem of distributing Data at a Distance (DaaD)
- ☒ Free information from physical devices
- ☒ Federate systems to be utterly consistent from the local to the global scale

As these solutions took shape, it became clear that the old virtualization terminology, “storage virtualization,” did not match the new model for federated data storage. However, by terming it Virtual Storage, a more accurate definition emerged:

- ☒ *Virtual Storage*: the management of distributed cache coherency over distance.

EMC has now introduced its comprehensive vision of Virtual Storage. The product is VPLEX. It is a hardware and software platform for use with data farms using EMC, Hitachi, IBM and other arrays. Just as servers can be federated, so VPLEX can federate storage devices to create cache coherency, or Data at a Distance. The data, whether in megabytes or petabytes, is presented as local data to the user.

Federated data storage, from local to global to the private cloud, means a reduction and often complete elimination of issues with latency and bandwidth. In its place, it creates distributed cache coherency. Multiple computing environments, regardless of their location on earth, see a large, shared pool of data/information that can be accessed from any server and across applications. Users at diverse geographic global locations can access and work with identical information simultaneously while the master file remains in sync.

VPLEX makes it possible to conceptualize a new model of a virtualized IT environment by applying the tenets of virtual server evolution to virtual storage. From the user's screen, the IT environment is viewed as consistent and device-agnostic, even if it isn't. Just as the user has learned not to be concerned about the location of processing resources or the host application, now he or she can work with information as if it were stored on their local machine, regardless of where it is physically stored. VPLEX builds on the dynamic flexibility of FAST (Fully Automated Storage Tiering), which exemplifies the ideal of having the information/data in the right place at the right time – in this case, in the coherent cache.

In short, VPLEX is the fundamental toolset for building a virtual, federated information system (hardware, software, and network) as a private cloud, capable of spanning any distance great or small. In the process, it delivers these benefits:

- ☒ Mirroring a master environment globally means redundant physical resources can be eliminated and replaced with virtual devices
- ☒ Data replication to overcome distance constraints becomes a thing of the past
- ☒ Economies of scale-down mean IT runs smoother, faster, and less expensively
- ☒ Since physical resources are now aggregated and based on a virtual model, it no longer matters where a datacenter is located, so moving operations to less costly environments becomes practical

Future Considerations

VPLEX is a solution designed for current and anticipated requirements. Many IT operations have yet to build out virtual capabilities, and VPLEX is a sight down the road for them. Yet seeing that it is there, on the horizon, and that a business case can be made for moving toward it as a comprehensive solution to many disparate, nagging issues, could create a “build it and they will come” momentum.

VPLEX is a block-based solution initially delivered in two products; VPLEX Local and Metro. Metro stretches the functionality noted above across a synchronous distance of up to 100km. One can imagine the potential use cases when data is liberated, below are simply a few that immediately come to mind:

- Transparent sharing and balancing of resources within and across physical data centers
 - Workload balancing
 - Move workloads to low energy cost locations
 - Support for VMware VMotion
- Real-time data access for remote users
 - Concurrent read and write access to data from multiple hosts
 - Data access to remote physical data centers without local storage.
- Distributed mirroring across mixed platforms
 - Increased resiliency and availability
 - Run applications continuously
- Pooling and aggregating capacity
 - Improve storage utilization

Initially, VPLEX may be best suited for read-intensive applications. For multi writer intensive application, server technologies and application environments, such as transaction processing, may need to catch up. Even so, the prospect of a federated, global datacenter is intriguing and makes the vision of private cloud computing, now so effervescent, seem much more practical. One fact is clear: private or public cloud computing demands a virtualization management layer to manage all the technology resources as one integrated system. Within the private cloud, users can now use any and all resources – servers, applications, information, and the various network connections – as if they were all contained within their personal machine. This is the end-user computing solution that has gone wanting for the past forty years.

The cost-saving benefits that accrue to IT are even greater. Federation in a virtual data center is vendor-agnostic, and unifies platforms across servers and storage farms. It means data is utterly transparent and available “24xforever.” Storage capacities and loads are balanced across the cloud by the utilization of cache coherency. Server and storage device redundancy is either reduced or eliminated altogether, especially as the prime data center model is cloned and propagated throughout the cloud. An average order of efficiency improvement of 6X is not unrealistic.

In this process, IT makes a large stride forward in becoming a true service organization. Federated virtual storage promises a major step in IT evolution. Seasoned professionals will find many ways and means of implementation and varied applications of this disruptive new technology. In conservation, a popular and true axiom is to think locally and act globally. At the local level, the user will see a simple solution to their workflow. At the cloud level, federation changes everything because

all disparate resources either are, or have the power to become, totally integrated and thus aggregated.

This, then, is the VPLEX objective: to implement a federated IT environment that affords a transparent global view of storage that can be distributed at the cache, with consistent coherency across the private cloud environment. Now it is possible for multiple writers to work with the same information with the integrity levels consistent with an enterprise environment. When this is implemented at the right level of abstraction, nearly everyone in the organization, regardless of their perspective, can see the rich benefits that can accrue.

With VM and VPLEX, EMC is creating a new definition of virtualization. Together, they create a 21st century competitive tool and a service-oriented IT environment.

CHALLENGES/OPPORTUNITIES

The opportunities are significant for EMC, given its footprint worldwide and its pervasive acceptance across major enterprises. Organizations large and small have been seeking to solve the distance problem for a very long time.

With the globalization, mobility and the always available expectations of both individual users, organizations and their clients, the availability of consistent, up-to-date data anywhere, any place will provide EMC a level of technology leadership that will force its competitors to find solutions (or better yet, develop) to match the promise that EMC is making.

However, although EMC believes that VPLEX will solve the distance issue, given that problem has been around for such a long time, and so many companies in the past have attempted to address this, it is likely to receive a great deal of skepticism.

Additionally, VPLEX is not a silver bullet to the distance problems. It will be a solution only to certain environments. End-users should not expect that VPLEX will solve every issue they have related to distance.

For example, transaction intensive, latency sensitive applications are not candidates for VPLEX. (To be fair, EMC also will not promote VPLEX to be a solution for such applications).

VPLEX will require associated services to bring its promise to reality – this entails pre-sales consulting, post-sales support and long term maintenance to ensure that the solution operates optimally.

Finally, EMC will need to assist users in differentiating between the global data distribution features of EMC Atmos and the features and function offered by VPLEX.

CONCLUSION

VPLEX, and the vision and roadmaps it brings, provides relief to the issue of distance and a necessary solution in distributed cache coherency. IDC expects EMC to continue to expand on this roadmap with additional solutions over time.

End-users and potential customers for VPLEX need to work closely with the EMC team to leverage the most from VPLEX by thinking outside of the traditional four walls of the datacenter, and be creative about various opportunities for extracting the most return from an investment in VPLEX.

Overall, VPLEX represents a unique (and importantly, integrated) approach to addressing the persistent problem of integrating distributed data centers and making the (IT) world smaller.

Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2010 IDC. Reproduction without written permission is completely forbidden.