CONTINUOUS APPLICATION AVAILABILITY WITH EMC VPLEX

A growing number of organizations are using EMC VPLEX and other off-the-shelf technologies to ‘stretch’ application processing and data access across distance—for continuous application availability that is practical, affordable, and automatic.

ABSTRACT

Enterprises are consolidating HA and DR practices and technologies to achieve Continuous Availability (CA) that provides users with uninterrupted access to data and applications during planned and unplanned outages.

Certified by VMware, and proven to work with Oracle, Microsoft, and other off-the-shelf cluster technologies, EMC Continuous Application Availability Solutions powered by EMC VPLEX provides data coherency across distance for simultaneous access to the same data in multiple locations, enabling active/active CA—without changes to application code or custom integration.

This paper describes how EMC’s deployment models for applications work; how the assets in your existing application sets can be redeployed to implement pure CA and near-line CA services; and how to accelerate putting CA to work to protect mission-critical applications and data for business advantage.
IT TRUST

Increasingly, organizations need more than the ability to minimize or recover from downtime. They need to be able to trust that the data and applications their business and users depend on will continue to be available and operate, through both planned and unplanned outages.

When business is conducted digitally and globally, business operations are by definition 24x7x365. Beyond the traditional risks to revenue and reputation, in an always-connected world, even a minor outage has the potential to ‘go public’ and do damage very quickly.

A recent Forrester Consulting study¹ found ‘increased reliance on technology’ the number one factor driving increased risk in their business. And a 2013 global survey² of business and IT decision makers concludes that IT trust is an opportunity, as well as a risk.

THE END OF APPLICATION DOWNTIME

As anyone who works with business application owners to schedule planned outages will attest, getting approval for downtime to perform tasks such as updates, upgrades, and migrations is increasingly difficult. Users—employees, partners, customers, and consumers—increasingly expect uninterrupted access to data and IT services.

As stakes and expectations rise, high availability (HA) and disaster recovery (DR) are not enough. Increasingly, organizations are recognizing the need for zero downtime—uninterrupted operation through both unplanned and planned outages.

Figure 1. Ratio of planned to unplanned sources of downtime. Scheduled events account for approximately 85 percent of application downtime.

---

¹ How Organizations Are Improving Business Resiliency With Continuous IT Availability, 2013. Forrester Research, Inc.
LACK OF CONFIDENCE IN DR

Companies make a considerable investment in DR because the potential impact of an extended outage would be devastating. Yet in practice, organizations realize little or no return on this investment.

Experience shows that companies do not leverage their DR solutions for planned outages, which account for 85 percent of application unavailability. Less obvious perhaps, is the fact that, in practice, the DR solution or site is typically not invoked for most unplanned outages. Short of a complete and protracted data center site loss—the cause of less than one percent of all downtime—the DR solution is rarely used. Instead, the decision is usually to forgo DR failover, even when the workaround restore will take hours or even days.

Why? Most organizations do not have confidence in their ability to recover using their DR solution. When asked about their current application and data availability strategies, 46 percent of respondents in the latest Independent Oracle Users Group (IOUG) survey were less than satisfied.

For one thing, DR is typically architected as "all or nothing" event. Failover, restart or restore are manual, requiring a decision to act. Many organizations do not do the testing or updating of their DR plan required to keep pace with changes in their production environment. When an unforeseen outage occurs, they lack of confidence in a smooth failover/failback. The fear is that invoking DR will "make it worse" and extend the outage.

In addition to the investment made in a capability that is not used, the vast majority of DR today relies on idle, redundant infrastructure in a recovery site, incurring considerable capital and operational (space, power, maintenance) expense.

CONTINUOUS AVAILABILITY: A NEW APPROACH

For all these reasons, organizations are moving beyond traditional HA and DR to achieve “Continuous Availability.” Continuous Availability (CA) provides users with uninterrupted access to applications and data, enabling operations to continue without interruption or degradation during planned and unplanned outages.

While long the ideal, until recently CA has been prohibitively expensive and complex for all but the most demanding applications. For example, although Tandem and Stratus provide fault-tolerant systems, achieving multi-site CA, able to survive the loss of one site without interruption, has traditionally required custom applications, multi-phased commit logic, and systems integration, making such solutions costly to build, operate, and maintain.

STRETCHING CLUSTERS OVER DISTANCE WITH EMC VPlex

Now, however, organizations can achieve continuous application availability across heterogeneous environments using commercially available off-the-shelf technology available from multiple vendors. That’s because EMC VPlex virtual storage delivers data coherency and performance across distance, enabling organizations to “stretch” their existing clusters across dual sites for active/active CA that is much simpler and more affordable. In many instances, CA with VPlex costs less than the traditional HA and DR solutions most organizations rely on today. What’s more, VPlex Metro provides CA for all of the block data in an organization, regardless of platform.

For more information go to http://www.emc.com/microsites/it-trust/protecting-information-systems-processes.htm

---

MOVING FROM HA/DR N+1 ACTIVE/PASSIVE TO CA 2N ACTIVE/ACTIVE

From a technical perspective, EMC’s CA application deployment architecture is an active/active 2N solution. That is, everything the application or service needs (N) to be available is actively provided, in parallel, by two providers.

In contrast, traditional HA and DR are N+1 solutions, in which N represents the production need—plus (+) a set of redundant, typically passive, backup resources required to be able to restore service. HA +1 solutions in the production site are typically designed to fail over automatically; and DR +1 solutions in the DR site are typically set up to be manually invoked.

For critical applications using traditional HA and DR solutions, organizations effectively have three service providers, the production need (N) and two +1 passive solutions; +1 for HA and +1 for DR, which results in a lot of idle assets.

The primary difference between CA 2N and HA/DR N+1 architectures, is that with CA 2N, two sets of resources in two production sites are actively serving the application (active/active), if one goes down, the other continues to provide services at 1N (one times the need).

CA 2N makes active use of all assets at potentially lower cost by eliminating the idle HA resources at the prod site and requires no recovery or restore.

STRETCHING CLUSTERS OVER DISTANCE

CLUSTER BUILDING BLOCKS

Today most modern applications are deployed in either two or three tiers. Typically, there is a presentation layer, an application layer, and a data or database layer. To provide high and continuous availability, various forms of clustering are deployed.

DATA LAYER CLUSTERS

In general, data layer clusters are used to provide high and continuous availability for database servers or servers that manage dynamic data. There are three basic types of widespread commercially used clusters and each has its own use case.

• Wave 1: OS Clusters—Operating system (OS) based clusters were among the first to gain widespread commercial use. They are typically an active/passive (N+1) technology, with two nodes—one active node representing the need, and a passive node representing the +1 or the backup. The backup node usually waits idle in standby mode.

Both nodes are connected to the cluster’s resources, and both are configured to write to the disk resources. A locking mechanism based on the ownership of a quorum is used to control which node owns the cluster’s resources. By agreement, the node that owns the quorum owns the cluster resources and can write to the cluster’s disks.

In the event of the failure of the active node, the standby automatically takes over and brings up the necessary services to take over the role of the active node.

OS-based clusters are typically used for physical servers, although they can be used in a virtual environment, such as IBM AIX LPAR technology or inside a hypervisor cluster.

• Wave 2: Hypervisor Clusters—Hypervisor clusters are more sophisticated than OS clusters because the hypervisor owns and manages locking for many virtual machines running on many physical servers, and some hypervisors can orchestrate failover and restart.

The hypervisor controls locking of the logical disk resources owned by the virtual machines. The VMware vSphere Hypervisor (based on ESXi), for example, uses a number of locking mechanisms. One of them is a disk-locking mechanism embedded in the VMware Virtual Machine File System (VMFS).

VMFS is a clustered file system that leverages shared storage to allow multiple physical hosts to read and write to the same storage simultaneously. VMFS provides on-disk locking to ensure that the same virtual machine is not powered on by multiple servers at the same time. If a physical host fails, the on-disk lock for each virtual machine is released so that virtual machines can be restarted on other physical hosts.

Compared to an OS cluster, VMFS controls write operations to logical volumes from many virtual servers intermixed on one or several LUNs. Whereas OS clusters typically manage whole LUNs or volumes.

VMware vSphere High Availability (HA) provides an N+1 architecture (active/passive) architecture for virtual machines and VMware vSphere Fault Tolerance (FT) provides a 2N (active/active capable) CA architecture for virtual machines.
Wave 3: Database Clusters—Database clusters are a scale-out architecture where many servers are working on the same database. Locking mechanisms managed by the database allow multiple servers to write to the same database simultaneously. Database clusters are an xN plus active/active architecture, where x is greater than one. Typically, they are deployed in a range of 1.2N (120% of need) to 2N (twice the need) architecture. Some examples of database clusters are Oracle RAC, IBM DB2 PureScale, and Sybase.

It is worthy to note that a database cluster can also reside inside of a hypervisor cluster, allowing deployment options for physical and virtual servers.

Stateless Server Clusters—Stateless servers are different from data servers in that they only hold onto transaction data for the duration of a transaction. They are often deployed in the presentation and application tiers.

Typically, multiple servers are deployed in a farm and local load balancing technologies are used to provide HA by distributing transactions based on policies or load. If a server is unresponsive, transactions are directed to another server in the farm.

In a network load balanced cluster, several servers will usually be banded together to provide the need—and a spare server or two is added to the farm for availability. As a result, this deployment architecture provides availability at a relatively lower cost.

Network load balanced clusters are by nature a continuous availability active/active xN architecture, where x is greater than one. Again, x is generally in the range of 1.2N to 2N, or 120-200% of need.

Network load-based clusters can be used for physical servers or in conjunction with a hypervisor based cluster for virtual environments.

PUTTING IT ALL TOGETHER
The EMC Global Services design and deployment model for CA over VPLEX uses off-the-shelf technology to enable CA applications stretched over two sites. Determining the right deployment architecture depends on existing compute assets, availability objectives, application, and other factors. EMC service professionals have helped companies design hundreds of solutions using the building blocks below. They can help identify crucial decision factors to select the right deployment architectures for each application and a middleware service that is economical and resilient.

At the bottom of the stack is virtually any of the major block disk arrays, one in each site. The next layer up is the VPLEX storage virtualization layer. VPLEX is the component that keeps the arrays in sync of each of the sites in sync.

VPLEX presents the compute layer with a virtualized disk volume that spans the two sites. Think of the virtual disk volume as a RAID-1 mirror where Mirror-A is in one site and Mirror-B is in the other site. A disk-write operation to Mirror-A is mirrored to Mirror-B and a write operation to Mirror-B is mirrored to Mirror-A. Since both sides of the mirror are equal, a read operation can be performed in either site.

To traditional ways of thinking, this can sound both wonderful and dangerous—wonderful, because application stacks can be deployed with a complete set of resources in two-sites, and scary, because the question arises: what prevents a simultaneous update to the same disk sector from each of the stacks? The answer—and appeal of—this deployment model is that it leverages the write locking built into
the clustering mechanisms to prevent this from happening.

At the data layer, using one of the above clustering models, the database cluster is stretched between the sites. Continuing up the processing stack, the model stretches the network load balanced preso and app farms between sites. Above the compute layer at the distribution layer, which routes incoming traffic to one of the processing sites, the model deploys a global load distributor. Global load distributor solutions are available from Cisco, Citrix, F5, and others.

Other requirements include: a data center interconnect or DCI, common name space, and virtual LAN (VLAN) spanning both sites.

- The DCI is a high-speed dual-path link that connects the sites—typically, it uses two 10Gb Ethernet (or faster) links. The DCI is used for VPLEX disk mirroring, inter-application, and inter-cluster communication, and command and control. Usually, user based transactions do not traverse the DCI.

- The common name space requires DNS services in both sites. A replication mechanism that keeps all of the DNS servers in sync is built into most DNS solutions. AD is also required in both sites along with your single-sign-on (SSO) solutions.

- VLAN spanning is required by most of the stretched clustering deployments. Traditionally, VLAN spanning has had a bad reputation and most network architects were taught to 'never do that' because historically network chatter would flood the DCI and little real work could be done. Today, however, modern stretched VLAN technologies such as Cisco's OTV and VPLS from Brocade and other vendors have solved the issues associated with spanned VLANs.

Using the above technologies, an abstracted application stack can be deployed over two sites. If a component fails, redundancy in the compute clusters and other technologies will take over and provide the necessary service. And if a whole site fails, the other site automatically takes over, with transactions automatically routed by the global load balancing technology to the surviving site.

Figure 2. With VPLEX, organizations can extend the data locking provided by database clusters across distance for dual-site active/active CA capability.
CONTINUOUS APPLICATION AVAILABILITY
Using the building blocks described above, organizations can achieve true CA for their applications. But if the existing database clustering technology used only supports an active/passive technology what are the options?

First, CA architectures are really deployment architectures; the underlying application architecture is unchanged. In a CA deployment, the compute and infrastructure assets are deployed in two different sites and the application is abstracted from an individual site. As a result, the number of components for the two active sites is greater than one active site, but can be less than an active site with HA technologies and a backup DR site.

Second, the VPLEX technology presents virtual data disks in both sites in a 2N or CA mode. Network load balanced clusters are CA, whether deployed in one site or two.

Finally, when it comes to clustering technologies for the database servers, database clusters and VMware FT are 2N active/active CA technologies. While stretched OS based clusters and hypervisor clusters are N+1 HA active/passive technologies, it is possible to create deployment models with CA up and down the stack, or a near-line CA model, with a minimum of CA VPLEX provided data, and various combinations of the cluster models.

To be certain, the current partial CA models do not have all of the benefits of a pure CA deployment, but they are foundation to build on as more applications become active/active enabled. Another consideration is that improvements in some OS based clustering models can mask the failover time, maintaining the appearance to the user as if the application is “always-on.” Microsoft Always-On SQL and the EMC deployment architecture for SAP are examples of always-on solutions.

COMMON QUESTIONS AND CONCERNS
VPLEX enables a whole new approach to continuous application and database availability that makes continued operation and access to data and applications during planned and unplanned outages feasible for a large number of organizations. Some of the most frequently asked questions and common concerns about this new approach follow.

HOW DOES VPLEX AFFECT APPLICATION AND DATABASE PERFORMANCE?
The impact depends on application, distance, and data center configuration. A maximum of five milliseconds (5ms) round trip time (RTT) is recommended for most VPLEX implementations. As a rule of thumb, a 5ms RTT translates to a round trip distance of about 60 miles/100km connected by fiber optic cables or 10Gb Ethernet. The ping command can provide a quick snapshot of RTT.4

VMware has certified VPLEX with vSphere HA auto-failover at up to 5ms RTT and allows manual failover with storage vMotion at up to10ms RTT. Microsoft has published a 34-page paper on implementing EMC VPLEX and Microsoft Hyper-V and SQL Server with enhanced failover clustering support over synchronous distances.5

---

4 EMCC offers a service with tools that can precisely measure RTT and check the cleanliness of data center interconnects.
5 http://download.microsoft.com/download/2/6/6/26660F7D-301E-48DE-8200-8CFF64C7BCEA/h7116_vplex_hyper_v_sql_wp.pdf
Other factors to consider:

- **Only write data needs to be mirrored to between sites.** Latency and its impact on application response time is a concern only for writes. Most applications have something like an 80/20 read/write ratio (generally falling between 70/30 and 90/10). In our experience, organizations deploying VPLEX will see read response times improve, because read requests are serviced directly out of the VPLEX cache. Considering also that most applications read several records first and then do a write, overall throughput tends to average out or be only very marginally impacted.

- **Improvements in disk write response time.** Just a few years ago, local SCSI disk response time averaged between 15-18ms. Today, FC response time is in the range of 3-5ms and Flash is in the range of 1-2ms.

- **SAN degradation.** We have also observed SANs, initially rolled out with a 3-5ms average response time, become degraded through changes and additions over time. In some cases, the response time has increased to as much as 15ms for an average write.\(^6\) With a well-tuned FC (Fibre Channel) or FCoE (Fibre Channel over Ethernet) SAN, VPLEX typically adds 1-5ms to each disk write, depending on distance.

- **Tradeoffs.** For many applications, the gain of CA, elimination of RTO, and HA/DR complexity and uncertainty will justify the increase in write time.

If the performance is still unacceptable, write response times can be improved in other ways to offset any latency caused to achieve CA with VPLEX. These include: adding FAST Caching or Flash storage and/or, as noted above, tuning SANs to spec. In some Oracle RAC environments, adding additional servers can increase parallelism and overall throughput. EMC application and storage architects can work with organizations explore options and determine how best to integrate VPLEX across two sites with little or no latency impact.

**WHAT ABOUT MIDDLEWARE?**

Applications need to get data from other sources and need to pass data to other applications. If an upstream app is in one location and downstream in another, the middleware data queues need to be mirrored and protected over VPLEX. Middleware, such as TIBCO, Oracle Enterprise Service Bus, and IBM MQ Series can be deployed over VPLEX with CA availability using an active/active database cluster stretched across sites.

**WHAT ABOUT NAS DATA?**

VPLEX currently supports block data only. For low-to-medium use, NAS mounts and shares can be served by a Linux or Windows server that stores data on block SAN virtualized over VPLEX. The server can then be protected by either a Wave-1 or Wave-2 cluster.

For large-scale NAS deployments on platforms such as EMC Isilon or EMC VNX with EMC Celerra, an active node in one site and a passive node in the second site can be mirrored using the native array replication software. The combination of a global file virtualization solution, such as EMC Rainfinity, and a manual failover trigger can be used to move the NAS service between locations, in the event of site loss.

---

\(^6\) EMC offers a health check service that measures SAN response time and bottlenecks.
In either case, for applications that hit NAS share/mounts heavily in one site only—and where bandwidth is not an issue, NAS services can be hosted in one site for all users and the lighter NAS traffic can traverse the DCI.

**ORACLE HAS AN EXTENDED ORACLE RAC DEPLOYMENT OPTION—CAN I JUST USE THAT?**

A number of years ago, Oracle introduced an active/active dual-site solution called Extended Oracle RAC using Real Application Clusters (RAC) with Automatic Storage Management (ASM) mirroring. However, the complexity and reliability of the host-based ASM mirroring overhead made the 2-site extension unworkable for most organizations and applications. The solution was also limited to only Oracle-based applications and data.

Because ASM Mirroring mirrors Oracle data only, in a dual-site configuration, another technology is needed to sync other types of data. There is no way to provide replication consistency across upper-level applications or multiple dependent databases without a third-party solution.

EMC’s CA solution with Extended Oracle RAC over VPLEX stretches the RAC servers between the two sites and the ASM mirroring role is provided by VPLEX.

**WHAT IF VPLEX GOES DOWN?**

VPLEX is itself based on clustering technology with shared memory cache. One VPLEX cluster is deployed in each site. If storage or a component of the VPLEX cluster fails, the other engines in the cluster automatically take over the load.

For more information, please refer to:

TechBook: [EMC VPLEX Metro Witness Technology and High Availability](#)

**SUMMARY**

EMC VPLEX virtual storage, combined with off-the-shelf clustering and global workload balancing technologies, enables dual active/active production data centers, separated by distance, to deliver continuous availability—typically at a lower cost than today’s HA and DR solutions. CA with VPLEX is different from traditional HA/DR approaches.

Traditional HA/DR requires:

- Separate HA and DR efforts and solutions
- Redundant, passive, standby infrastructure
- Recovery using replicated data—whether from a backup, data replication, and/or snapshot—so there is always a delay and risk in failback and recovery from the last known uncorrupted point for data and applications

In contrast, CA with VPLEX:

- Combines HA and DR into one solution, enabling continuous simultaneous operation in two active/active production data centers, separated by distance
- In the event of a planned or unplanned outage in one data center, CA with VPLEX automatically shifts workloads to the other site and operations continue—providing users with uninterrupted access to applications and data
The advantages of active/active CA enabled by VPLEX include:

- **Zero downtime**—Applications and data continue to be available and continue processing during outages caused by technical failure, human error, site-loss, as well as during planned outages for application testing, migrations, non-data upgrades, and the like.

- **Off-the-shelf choices**—CA is implemented with off-the-shelf technologies available from multiple vendors—with no application coding, customization, or integration required.

- **Increased asset utilization**—Passive redundant recovery infrastructure can be actively used—and resources pooled across sites to share workload. Not only do idle resources become useable, but also the ability to distribute load means organizations can typically reduce total server count by 30-40 percent.

- **Simpler management**—Organizations no longer need to manage separate HA and DR solutions, maintain and update DR runbooks, or test recovery plans; dynamic load balancing within and across data centers automatically optimizes performance; workloads can be easily shifted from site to site so migrations, maintenance, and technology refreshes can occur without disrupting user access or application and data processing.

- **Greater confidence**—Active/active operation and automated failover eliminate the risks associated with orchestrating manual and error-prone recovery tasks.

- **Reduced CAPEX and OPEX**—Better utilization, less infrastructure (licensing, maintenance, etc.) and simplified management frees budget and resources that can be shifted to strategic application innovation and IT transformation initiatives.

In a business case developed by Forrester Research for a composite organization based on interviews with EMC customers already using EMC VPLEX-enabled CA, the analysts found an ROI of more than 250 percent over three years, gained through minimized downtime, improved asset utilization and load balancing, and increased efficiency in data migration and hardware maintenance.7

**REVISIT AND RETHINK AVAILABILITY NOW**

VPLEX-enabled active/active CA implementations are at work today in company data centers, supporting organizations most critical applications, including SAP on Extended Oracle RAC and SAP on VMware HA and vMSC. In a recent Forrester research survey, 44 percent of respondents reported running some production loads in an active/active mode—and 12 percent of this group reported using stretch clusters.8

Organizations should act now to address availability risks and achieve the continuous application and data availability and mobility required in today’s business and digital and cloud-enabled marketplaces.

---

7 The Total Economic Impact of EMC Continuous Availability: A Case Study Focused On Continuous Availability With EMC VPLEX, November 2013. A Forrester Total Economic Impact™ Study commissioned by EMC. Project Director: Reggie Lau, Forrester Research, Inc.
ACCELERATING CA WITH EMC GLOBAL SERVICES

CA success relies not only on technology, but also on the ability of business, application, database, and IT stakeholders to think differently about availability. This can be difficult when concepts like the Recovery Time Objective (RTO) and Recovery Point Objective (RPO), part of the conversation for decades, no longer apply and the business objective becomes zero downtime.

As an embedded aspect of everyday production operations, CA also changes the way application owners, application architects, DBAs, and IT operations work together.

Organizations can accelerate the transition to CA by taking advantage of EMC experience gained in helping hundreds of customers make the technical, process, and skillset changes required to implement active/active data center solutions.

EMC Continuous Availability Advisory Service experts can help organizations to quickly evaluate their CA readiness and determine the best possible approach. Our trained professionals have designed hundreds of solutions using the building blocks described above. They will help you with the nuances of application deployment architectures and identify the crucial decision factors to help you select deployment architectures for each application and middleware service that is economical and resilient. Services include:

- CA Cost/Benefit Analysis and Business Case
- CA Deployment Solutions for Applications and Middleware
- Active/Active Infrastructure Solutions
- Step-by-Step CA Roadmap & Timeline
- CA Implementation
- CA Onsite, Shadowing, and Education Services to supplement, mentor, and train internal staff

Learn more about EMC Continuous Availability Advisory Services.