

# everRun Product and Technologies Technical White Paper

Jerry Melnick, CTO, Marathon Technologies

## COMPANY BACKGROUND

Marathon Technologies has been developing software that provides the highest levels of server availability for over 15 years. The company was founded by hardware and software engineers with deep knowledge in the principals of fault tolerant and high availability computing and experienced as developers of traditional fault tolerant hardware platforms at Digital Equipment Corporation.

Marathon is the only company in the world that provides a software-only solution that delivers fault tolerance for any Windows application. Marathon's everRun software includes a full suite of availability options provided in a single package and managed from a single interface. Availability options range from basic restart to the most sophisticated and highest availability settings including non-stop system fault tolerance. Using everRun, the administrator can simply "dial" a separate level of availability for each application to optimize availability and resources in the computing environment.

The underlying technology uses a patented architecture that manages all dimensions of availability; data, network, and system, in a comprehensive, highly reliable fashion. Redundant resources are actively validated, thereby eliminating the source of many configuration errors and ensuring correct operations through faults and failures. Fault detection, management, and recovery are completely embedded and automatic. No application modification or special scripting is necessary, making everRun an extraordinarily simple yet highly sophisticated availability solution. The product runs on commodity x86/x64 hardware from any server vendor.

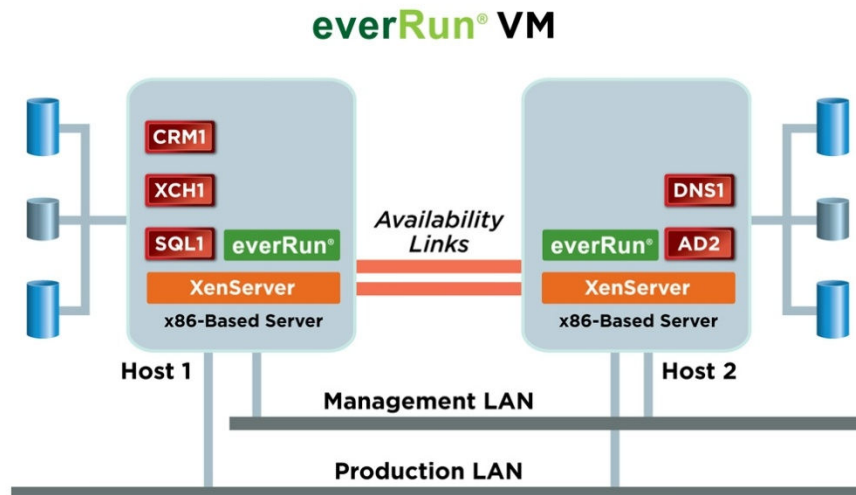
# everRun Product and Technologies White Paper

## MARATHON'S AVAILABILITY SOFTWARE FOR VIRTUAL MACHINES: everRun VM

everRun VM is the fourth generation of Marathon's fault tolerant software. The product incorporates all of the features, characteristics and technology of previous generations of Marathon's fault-tolerant technology and integrates these with Citrix's XenServer availability features. everRun VM operates on XenServer Enterprise and Platinum editions.

In most environments, different applications require different levels of availability depending on their business requirements. Recognizing this need, everRun VM takes a unique approach by defining 3 distinct levels of availability that can be applied individually to VM's running within a XenServer pool. Marathon terms this *Dialable Availability* and describes these as Level 1, Level 2, and Level 3 availability.

An everRun physical configuration consists of a pool of XenServer hosts configured with shared or local storage. Hosts in the pool are interconnected with two additional 1 Gb networks called *Availability Links*. These links are dedicated to synchronization activities that support everRun's availability levels 2 and 3.



Installation of everRun VM is simple. XenServer is first installed on each host. Using the XenCenter management UI, the user establishes the pool and creates and

# everRun Product and Technologies White Paper

configures VMs. everRun is then installed on each host by executing a simple install script.

Once installed, the administrator uses the everRun Availability Center management interface (eAC) to manage and configure availability for any VM in the XenServer pool. The eAC is simply accessed by addressing a URL on the pool primary using a web browser from any client computer connected to the management LAN. The eAC provides a treeview of all VMs in the pool. Protecting a VM is simple. Icons in the treeview use intuitive notations to indicate which VMs are protected and their level of protection. Using the eAC's GUI, the administrator selects a VM for protection and is provided a wizard to select any one of the three levels of availability supported:

## Level 1 – Basic VM Restart

Host node failure detection with VM restart on surviving hosts

## Level 2 – Component-Level Fault Tolerance

Non-stop operations in the event of disk or network losses

## Level 3 – System-Level Fault Tolerance

Non-stop operations in the event of any class of failure

## everRun Management Console (eAC)

The screenshot displays the everRun Management Console (eAC) interface. The main window shows a 'Protected VM Summary' on the left with a treeview of VMs (Test1, Test10, Test3, Test4, Test8, XCH1, XenWin27). The 'Virtual Machine Status' table lists VMs with their status, name, type, power state, hosts, and pool. The 'Test3' VM is highlighted, and its details are shown in the center, including state (Good), host (p27xenhost), and static maximum memory (512.0 MB). A 'Protect a Virtual Machine' wizard is overlaid on the right, showing the 'Select Level of Availability' step. The wizard offers three levels: Level 1 (Basic VM Restart), Level 2 (Component-Level Fault Tolerance), and Level 3 (System-Level Fault Tolerance). Level 2 is selected. The wizard also shows the selected VM (AD1) and host (XenHost2).

Status	Name	Type	Power	Hosts	Pool
Protected	Test10	Protected	On	p27xenhost, sdxxen	eAC Dev Pool
Protected	XCH1	Protected	On	p27xenhost, sdxxen	eAC Dev Pool
Protected	Test3	Protected	On	p27xenhost, sdxxen	eAC Dev Pool
Protected	Test4	Protected	On	p27xenhost	eAC Dev Pool
Candidate	Test1	Candidate	Off	p27xenhost	eAC Dev Pool

# everRun Product and Technologies White Paper

Once the wizard is completed, everRun VM automatically initiates the appropriate actions to configure the VM for the chosen level of availability. No application modification, scripts, or other custom administration is required.

The eAC also provides a monitoring function that provides a comprehensive, simple view of the health and state of all protected VMs. The state of each protected VM is indicated graphically. A protected VM with failing components appears with various status icons indicating the degraded condition. The administrator can quickly browse through detail to identify the degraded component and initiate appropriate repairs.

## everRun LEVELS OF AVAILABILITY

The following sections describe the characteristics and underlying technology of the various levels of availability supported by everRun VM operating on XenServer 5.0.

### LEVEL 1: BASIC FAILOVER

Level 1 availability is implemented using the base capabilities of XenServer HA, the high availability feature of XenServer Version 5.0. XenServer HA was created through a joint development project between Citrix and Marathon engineers. The technology, developed by Marathon engineers, was incorporated into XenServer 5.0 and delivered as a component of XenServer HA with XenServer Enterprise and Platinum editions.

Level 1 provides basic VM restart facilities useful for protecting simple applications such as DNS, web, file or print servers. The environments these applications operate in can typically tolerate some interruption in service or have alternative load balancing or availability capabilities that are desirable to use. Level 1 protection can be applied to any VM configured with 1 or more vCPUs.

XenCenter is used to configure the pool and enable HA pool wide. Level 1 protection of a VM can be accomplished using either XenCenter or the eAC's protection wizard. The resulting protection characteristics for the VM are the same regardless of which approach an administrator chooses.

The Level 1 protection architecture is fundamentally a cluster technology implemented at the host level. Host health is monitored using redundant heartbeats on the management network and across a shared disk holding a state-file. The design detects host node failure and recovers Level 1 protected VMs on surviving

# everRun Product and Technologies White Paper

hosts. Level 1 can only be applied to VMs that configure storage on fibre channel based SAN or iSCSI shared storage.

XenServer HA uses a dynamically computed failover plan to ensure that the pool has sufficient memory resources to restart protected Level 1 VMs on surviving hosts in the event of one or more host failures. When a VM is started, the failover plan is recalculated. If the total number of host failures that can be tolerated falls below an administrator set limit, the VM will not be allowed to start.

Level 1 protection provides basic restart facilities without the assurances or higher levels of availability inherent in Level 2 and Level 3. A number of factors should be considered when protecting an application with Level 1 protection:

#### ***Host Level Failure Detection Only***

A defining characteristic of Level 1 protection is that only host server failure is detected and managed. The failure of a VM, an application network or disk is not detected or repaired. Should any of these components fail, Level 1 protection will take no action and operations of the VM will be degraded. The administrator can configure bonded NICS and employ multi-path shared disk configurations to address network and data redundancy issues.

#### ***Limited Guarantees for VM Restarts***

While the failover plan proactively considers pool-wide memory resources, there are no active validation protocols to completely assure that configured IO components are fully accessible and operational on a surviving host, and that a surviving host is capable of resuming VM operations. Without active validation, it is possible that the protected Level 1 VM may be restarted on a surviving host that has a degraded network or disk pathway and is unable to successfully resume operations.

#### ***Need for Administrative Care in Configuration and Management***

Level 1 protection is reliant on careful configuration, management, and maintenance of disk and network hardware components. While the Level 1 facility handles restart of a VM, it does include protection provisions for data and network. Typical configurations require setup of multi-path SAN configurations using reliable shared storage devices to insure data redundancy and bonded NICs to manage network redundancy. However, these facilities must be separately managed and monitored requiring additional administrative skills and diligence in configuration. These requirements may be a source of operational errors in some IT environments.

# everRun Product and Technologies White Paper

As you will see in the following sections, Level 1 and Level 2 protection address the need to manage the considerations described above.

## Level 2: Component-Level Fault Tolerance

Level 2 protection utilizes Marathon's fault-tolerance technology rather than the XenServer HA underlying heart-beat technology. Level 2 provides significantly greater protection for business-critical applications such as Microsoft Exchange and SQL Server. Level 2 protection can be applied to any Windows VM configured with 1 or more vCPUs.

Applications operating with Level 2 protection compute through any loss of a network or disk pathway without interruption or operator intervention. In addition, Marathon's patented network fault tolerance eliminates the need for NIC bonding and greatly improves the availability and robustness of the network.

Level 2 provides an important advantage over Level 1 protection in that all classes of storage are supported including non-shared locally attached disks and performs synchronous mirroring between hosts to ensure that disk data is redundant and available at all times.

During the Level 2 protection process, everRun examines the target VM in terms of CPU, memory, disk, and network configuration and creates an identical, fully redundant VM configuration. The redundant VMs operate in a primary/secondary relationship on separate hosts. In Level 2 protection, the primary is active and the secondary is in standby. These roles can be changed without interrupting the application, which is useful for managing planned downtime.

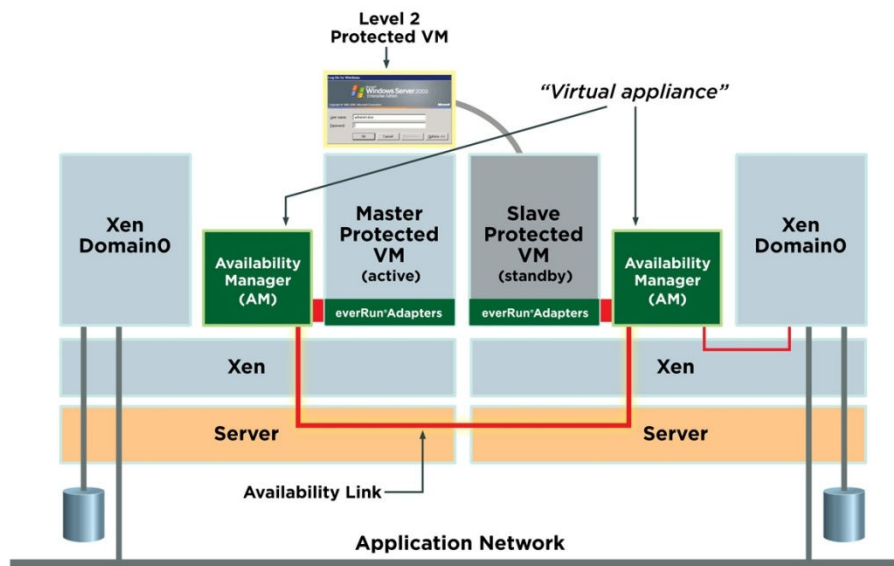
Level 2 protection utilizes a sophisticated architecture to detect and respond to failures in a highly responsive fashion. This architecture provides a significantly greater coverage model for all classes of IO failures that is more reliable and highly available than Level 1 techniques.

At the heart of the design is a *virtual appliance* called the Availability Manager (AM). The AM is a lightweight, service VM with no user or management controls. The AM's responsibility is to protect a VM by monitoring and managing the VM's IO pathways. When a Level 2 protected VM is started, an AM is dynamically created for both the primary and secondary. The AMs establish communication with each other over the dedicated Availability Links. The primary VM is started on one host and a redundant secondary VM is created on a second host. The secondary VM is paused and made

# everRun Product and Technologies White Paper

ready to start should the primary fail. All IO of the protected VM is redirected through and managed by the AMs.

## everRun® VM - Level 2 Architecture



The AMs coordinate and maintain a synchronous mirror copy of all disk data between the two hosts and bind application networks on each host to create a redundant and fault tolerant network pair. Since the AM is in the IO pathway, it can actively monitor these to detect and respond to discrete failures without dependencies on specific NIC or storage hardware or configurations. If an IO failure is detected, the AMs transparently redirect IO to operational components on the other host, thereby permitting the primary VM to continue uninterrupted through the failure. If a failure of the primary VM is detected, the AMs coordinate the restart of the secondary VM. This restart happens automatically and immediately, and the VM maintains the same IP address and identity as when it was running on the primary.

Using this design, Level 2 provides a significantly higher level of availability and reliability beyond Level 1 facilities. Distinguishing capabilities are:

### **Fault Tolerant IO**

Failures in IO components are detected and actively managed. Failures of a disk or network pathway do not interrupt the operations of a Level 2 protected VM. If a disk failure occurs, the AM transparently redirects IO to use the surviving member of the mirror set on the secondary host. Similarly,

# everRun Product and Technologies White Paper

should a network failure occur, the secondary network adapter is activated and used. IP sessions are maintained across this failure with no interruption of client connectivity.

## **All Storage Classes**

Since everRun Level 2 provides full synchronous disk mirroring. Any type of storage can be used. This includes local, direct attached storage, FC SAN, iSCSI, or NFS. Use of local storage is particularly important for environments where the cost and complexity of shared storage is not practical or desirable. The design also allows for storage mirroring between locations without the need to have like storage at each site.

## **Active Component Validation**

everRun's IO architecture solves the problem of *failed-failovers* by actively validating and using all redundant IO components during steady-state operations. If a faulty pathway is identified, the fault is identified, and the administrator is notified. In this way, a degraded system can be repaired before a subsequent failure can impair operations.

## **Hardened Resource Reservation**

All redundant IO devices on the secondary are configured, accessed, and reserved during the startup of a protected Level 2 VM. The secondary AM maintains and monitors these reservations while active. If the redundant devices cannot be reserved during protected VM startup, startup is aborted. As a result, the administrator knows at startup whether all required devices are configurable, working and reserved. Coupled with Active Validation, this design assures that devices are usable when needed and avoids the problem of discovering that a device is inoperable during a failure recovery.

In addition, memory on the secondary is statically reserved during protected VM startup by starting the secondary VM and placing it in a paused state. This assures reliable restart on the secondary should the primary host or VM fail. It also eliminates any potential administrative overrides that might compromise the availability of needed memory resources. The secondary VM memory footprint can be configured to reduce the reservation if desired. This will of course run the application on the secondary with less memory should the primary fail.

## **Full Fault Coverage and No Hardware or Application Dependencies**

Many availability technologies are device, configuration, vendor, or application dependent, requiring certification and validation for specific configurations. Independent tests or probes may be employed at an adapter

# everRun Product and Technologies White Paper

or device to manage specific subsystem monitoring. These techniques require careful assembly and configuration of components and may not fully cover all classes of failures.

Unlike Level 1, the mechanisms for detection and response to IO failures are fully integrated into Level 2 protection without relying on separate hardware or configuration conventions such as bonded NICs or shared storage. Integration of these mechanisms is essential in achieving more reliable operation as it removes opportunities for configuration errors, a common source of availability system failures.

everRun's Level 2 protection tests and protects the entire IO pathway from the application all the way through to the target device ensuring that a usable pathway operates redundantly at all times. Failure detection and recovery are independent of the underlying hardware. There are no port, device, or switch dependencies. Errors in configuration as well as physical device failures are covered equally well thus eliminating administrative error.

## **No Agents**

A popular approach to availability is the use of a software agent or service that operates in the context of the application environment. Busy application environments can freeze or fail an agent. This can negatively impact the agent's operations resulting in false error callouts or failovers. In these cases, the application is incorrectly terminated and restarted when it may in fact be healthy.

The Level 2 AM architecture operates below the VM and outside of the execution context of the application. This approach formalizes the separation between fault-detection processes and the application execution environment. Fault detection and recovery logic runs in the separate AM context with its own resources thereby eliminating false error call outs and unwanted outages.

## **Active-Active Host Based Network Pairing**

A customary technique for protecting networks is the use of network teaming or bonding. Bonding implements an active/passive relationship between two network interfaces on the same host. The passive network is started when it is determined that the active network has lost connection to the network.

Bonding falls short in providing simple, reliable fault and recovery handling for a number of reasons. (1) The passive network is not actively tested and known to be good at all times. As a result, when a failover to the bonded NIC is attempted, communications may fail. (2) Bonding can be complex to

# everRun Product and Technologies White Paper

configure and manage properly. The administrator must ensure that networks are on separate interface cards and properly cabled to the switch network. Some switches require special setup for proper handling of bonding failover. (3) Bonding detects only physical network connectivity rather than the application's ability to connect with clients and devices on the network. More complex switch or network path failures may not be visible to the failure detection capabilities of bonding technology.

In contrast, Level 2 network IO protection uses an active-active host based pairing technology that extensively covers and monitors the network path operations beyond physical connectivity and across redundant networks on two hosts. Network IO is active through the primary host during normal operations. The primary and secondary AMs monitor network activity for each of two networks on the separate hosts. The comparisons are shared across the Availability Links, a network independent from the networks being monitored. Use of a separate network for this comparison is a key reliability feature. The secondary automatically assumes communication responsibility should a joint determination conclude that the primary pathway is degraded. This design gives significantly greater visibility into the application network health beyond simple connectivity tests. Since the redundant networks are operating on separate hosts, network configuration errors are eliminated as well.

### Level 3: System-Level Fault Tolerance

everRun Level 3 protection has an identical architecture to Level 2 and yields all of the same benefits outlined in the previous section. In addition, Level 3 provides an important feature that delivers an incrementally higher level of availability suitable for the most critical applications. This feature is compute fault tolerance. Compute fault tolerance permits a Level 3 protected VM to survive any class of system failure including loss of VM or host. VMs protected by Level 3 continue operations through any fault or failure without interruption or loss of client connectivity to an application.

Level 3 is typically employed to protect specialized applications where any loss of service or in flight transactions may have severe negative consequences. Examples of these types of applications include process control, emergency services, or on-line media or services. Level 3 protection can be applied to any Windows VM configured with 1 vCPU.

Level 3 employs the same AM architecture used in Level 2 environments to accomplish protection of the IO environment. Like Level 2, the Level 3 protection process creates a redundant VM configuration operating between two hosts and



# everRun Product and Technologies White Paper

## FEATURE COMPARISON TABLE FOR everRun's THREE LEVELS OF AVAILABILITY

Feature	Level 1	Level 2	Level 3
Protection from host failure	Yes, through restart on surviving host	Yes, through restart of VM on redundant environment	Yes, VM continues without interruption
Protection from VM failure	No	Yes, through restart of VM on redundant environment	Yes, VM continues without interruption
Protection from IO failures	None integrated NIC bonding can be used to manage some classes of network failure Multi-path and redundant shared storage configurations can be used to handle storage failures	Yes, integrated Transparently protects from IO failures without application interruption or loss of client connectivity No need for NIC bonding or shared storage configuration	Yes, integrated Transparently protects from IO failures without application interruption or loss of client connectivity No need for NIC bonding or shared storage configuration
Shared storage	Yes, SAN and ISCSI	Yes, SAN, ISCSI and NFS	Yes, SAN, ISCSI and NFS
Local storage	No	Yes	Yes
Active component validation	No	Yes	Yes
Notification on loss of redundancy	None integrated Can be accomplished to some extent using various vendor frameworks and agents	Yes, integrated	Yes, integrated
VM agents	None	None	None
Hardened resource reservation	No	Yes	Yes
Full fault coverage for entire IO pathway	No	Yes	Yes
Active-Active Host Based Network Pairing	No	Yes	Yes
Disaster tolerant configurations	No	Yes	Yes

*Want to keep your Windows Server applications up and running through failures and disasters? Contact us for more information or to take test drive, [www.marathontechnologies.com](http://www.marathontechnologies.com)*

**MARATHON**

Run to Infinity

The Marathon logo, SplitSite and everRun are trademarks or registered trademarks of Marathon Technologies Corporation. Microsoft and Windows are registered trademarks of Microsoft Corporation. All other trademarks and registered trademarks are the property of their respective owners. Copyright 2008 Marathon Technologies Corporation. All rights reserved. Marathon Technologies Corporation reserves the right to make changes to this document at any time and without further notice. Marathon Technologies Corporation assumes no responsibility for any errors that may appear in this document.