

Proactive maintenance and adaptive power management using Dell OpenManage Systems Management for VMware DRS Clusters

White Paper

Balasubramanian Chandrasekaran, Puneet Dhawan
Dell Virtualization Solutions Engineering

December 2006

Table of Contents

| | |
|--|----|
| Table of Contents..... | 2 |
| 1 Introduction..... | 3 |
| 2 Background..... | 3 |
| 2.1 Dell OpenManage Systems Management Suite..... | 3 |
| 2.1.1 OpenManage Server Administrator (OMSA)..... | 3 |
| 2.1.2 IT Assistant (ITA)..... | 4 |
| 2.2 Virtualization using VMware Infrastructure..... | 4 |
| 2.2.1 Distributed Resource Scheduling..... | 4 |
| 2.2.2 VMware Infrastructure SDK..... | 5 |
| 3 VMware Infrastructure and Dell OpenManage Integration..... | 5 |
| 3.1 Proactive Response to Hardware Faults..... | 5 |
| 3.2 Adaptive Power Management..... | 7 |
| 3.2.1 User Configurable Parameters..... | 8 |
| 3.2.2 Power Management Algorithm..... | 10 |
| 4 Future Enhancements..... | 10 |
| 5 Conclusion..... | 11 |

1 Introduction

Organizations world wide are deploying virtualization solutions to consolidate existing workloads, to save on datacenter power and cooling costs and to be more responsive to ever changing business needs. 24x7 service and low power consumption are two key requirements in any data center. VMware Infrastructure 3 allows administrators to view several individual server resources as a unified cluster resource. It provides automated load balancing within the cluster and allows dynamic and seamless on-demand provisioning of servers.

This paper describes how administrators can combine Dell's OpenManage Systems Management suite with the cluster resources of VMware Infrastructure 3 to achieve proactive maintenance that enhances service continuity and adaptive power utilization that further drives down power and cooling costs.

The scripts and programs provide a framework to integrate systems management with VMware VirtualCenter by using the VMware SDK. Users can download and modify the code to fit their environment. The scripts and programs are provided "as is", without any implied support or warranty.

2 Background

This section describes main systems management components and resource management features of a Dell|VMware Infrastructure 3 solution.

2.1 Dell OpenManage Systems Management Suite

Dell OpenManage is a suite of system management applications for managing Dell PowerEdge servers. It offers comprehensive set of standards-based and interoperable tools for server deployment, monitoring and change management. Two OpenManage components that are relevant and important in context of this paper are OpenManage Server Administrator and IT Assistant.

2.1.1 OpenManage Server Administrator (OMSA)

OMSA enables easy management of individual servers and respective internal storage arrays. Using OMSA, administrators can perform functions like review server status and inventory, configure BIOS and RAID, set actions based on events, shutdown/restart of a server, etc. OMSA is installed on each Dell PowerEdge server.

OMSA is fully qualified to run within ESX Server environment. More details on installation, usage and support can be found under *Support Documents* at <http://www.dell.com/vmware>.

2.1.2 IT Assistant (ITA)

ITA is a comprehensive and standards-based console for managing all Dell servers, storage, tape libraries, network switches, printers and clients systems. Among many features, ITA provides administrators with ability to:

- Have web-based one-view console for Dell systems with red, yellow and green status indication
- Capture events and alerts generated by Dell servers running OMSA
- Configure actions based on events and alerts
- Monitor server performance statistics such as CPU, memory, I/O, etc.

2.2 Virtualization using VMware Infrastructure

VMware Infrastructure (VI) 3 is the latest suite of enterprise-grade virtualization products to enable consolidation, management, resource optimization and high availability, etc., for IT data centers. New features like iSCSI and NAS storage support, 4-way SMP for virtual machines (VM), support for 64-bit VMs, up to 16GB memory for VMs, VMware Distributed Resource Scheduling (DRS), and VMware High Availability (HA), make Virtual Infrastructure 3 a compelling virtualization solution for any enterprise. VMware Infrastructure is fully qualified to run on Dell PowerEdge servers, including the latest 9th generation servers. These servers offer enhanced processor technologies including quad core, Intel Virtualization Technology (VT), AMD-V, additional memory, more connectivity and enhanced management features.

Two VI features that are more relevant in context of this paper are VMware DRS and VI Software Development Kit (SDK). DRS provide dynamic resource scheduling and optimization of available physical resources. VI SDK enables third party applications to manage and control ESX hosts and virtual machines.

2.2.1 Distributed Resource Scheduling

VI 3 introduces the concept of an ESX server cluster: *a group of loosely tied ESX hosts that can be managed as a single entity*. As the name suggests, DRS brings distributed computing resources on physical servers under one pool (cluster) and schedules virtual machines on servers that can best serve the resource requirements of concerned VMs. It is built on VMotion technology. DRS provides following major functionalities:

- Automatic initial placement of virtual machines on a 'best fit' host in a cluster
- Automatic relocation (if chosen) of virtual machines based on resource requirements
- Automatic resource optimization and relocation based on change in cluster's computing power, e.g. addition or removal of a new host

VMware Infrastructure also introduces a new host status: *maintenance mode* that migrates all virtual machines from the concerned ESX host to other hosts that are part of a DRS cluster. Without any manual intervention, DRS makes sure that all

virtual machines are relocated among remaining hosts in a way to achieve best possible resource optimization.

2.2.2 VMware Infrastructure SDK

VMware Infrastructure SDK allows developers to build custom SOAP-based applications to manage ESX server hosts and virtual machines. This also allows IT administrators to integrate existing management applications with VMware Infrastructure and automate many tasks such as cloning and configuration of VMs, performance reporting, etc.

3 VMware Infrastructure and Dell OpenManage Integration

This section describes how different components of Dell server management and VMware Infrastructure can be tied together to enable comprehensive physical and virtual infrastructure management and task automation. We describe two example scenarios to illustrate benefits of integration of OpenManage with VMware Infrastructure:

1. Proactive Response to Hardware Faults
2. Adaptive Power Management

The scripts and program files can be downloaded from www.dell.com/vmware, under *Support Documents*.

3.1 Proactive Response to Hardware Faults

In this section, we explore how the system monitoring capability of OpenManage can be integrated with the VirtualCenter to improve fault tolerance in VMware clusters and provide a mechanism for proactive maintenance. Servers with hardware faults are automatically transferred into *maintenance mode* and the Virtual Machines are migrated to other healthy servers. This section is based on an earlier white paper by *Dave Jaffe and Todd Muirhead: [Implementing Fault Tolerance through Dell OpenManage and the VMware Software Development Kit](#)*. We extend the same concepts to VMware DRS cluster and take advantage of the fully automated VMotion capabilities for load balancing.

The above mentioned white paper was written for VMware ESX Server 2.x and Virtual Center 1.x. Based on any fault reported by OMSA agent running on an ESX Server, an action can be configured on ITA server layer. For example, based on a fault reported by OMSA, say loss of power redundancy on an ESX server, an action can be taken by ITA to migrate VMs from the faulty server to other servers based on existing load on each healthy server. The paper discussed algorithms to choose target servers based on CPU load on each of candidate servers in the migration pool. This approach adds complexities for developer to build an optimal algorithm to choose target servers and take migration decision for each VM being evacuated from faulty server.

This section outlines how such fault tolerance can be achieved in a simpler manner by using ESX server clusters combined with DRS (with fully automated VMotion setting). New programs written in JAVA to interact with Virtual Infrastructure web services are provided at the end of this white paper.

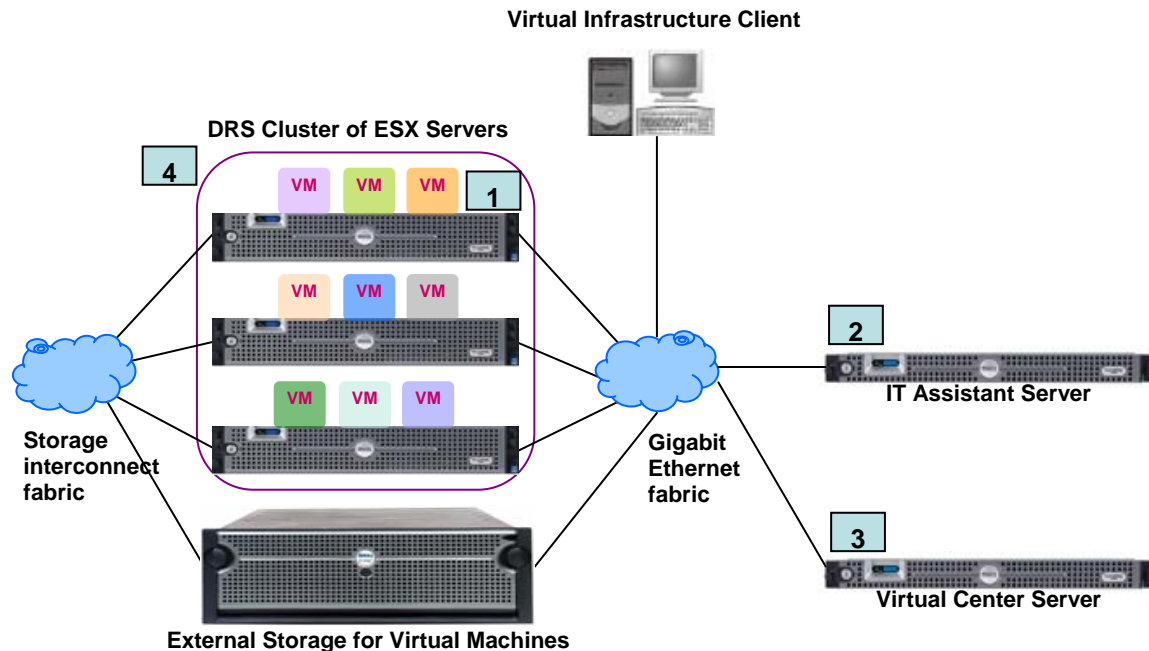


Figure 1: Proactive Response to Hardware Faults

Figure 1 describes the following sequence of actions that take place to perform proactive maintenance on a faulty server:

Action performed upon a hardware fault in a server:

1. In response to a hardware fault, OMSA sends an SNMP trap to the ITA server. The trap contains information about the server and nature of the alert.
2. ITA server filters the traps as configured by the user, and invokes a JAVA program, by passing the server name and severity as arguments.
3. On Warning and Critical Alerts, the JAVA program connects to the VirtualCenter, using VMware SDK and issues a command to put faulty ESX server in maintenance mode.
4. All VMs from the faulty host are evacuated to other hosts in the cluster. Placement of virtual machines on remaining healthy ESX hosts is taken care of by DRS algorithm.
5. The system administrator can now look at the faulty server and perform maintenance action, with zero down time to the running virtual machines.

Action performed when the health of the server is restored:

1. When server is returned to the healthy state (for example, after rebuilding a broken RAID using a hot-spare drive), OpenManage sends an SNMP trap to the ITA server.
2. ITA server filters the traps as configured by the user, and invokes the JAVA program, by passing the server name and severity as arguments.
3. On Normal alerts, the JAVA program performs an SMNP query to the server for global health information. OpenManage exposes this information through a particular OID. The SNMP query can be done either through *snmpwalk* command in Linux on *snmputil* from Windows. The SNMP query is required to make sure that the other subsystems in the server are healthy as well and are without any hardware faults.
4. If the global status is normal, the program then connects to the VirtualCenter, using VMware SDK and issues a command to remove ESX server from maintenance mode.
5. DRS Service discovers the addition of new computing resources to the cluster and redistributes VMs to balance load on the cluster.

3.2 Adaptive Power Management

Power Consumption is quickly becoming one of the most important challenges datacenter administrators face today. As the number of servers and supporting infrastructure increases, it leads to increases in:

- total power consumption and hence the cost of electricity
- air conditioning costs
- space required to house infrastructure
- associated management costs

All this adds to the net costs for an enterprise. Using server virtualization, customers can consolidate a large number of under-utilized servers. This provides some immediate benefits like reduction in number of servers required to do the same work, hence less capital computing infrastructure investment. The fewer the number of servers, the lower the utility costs as well. A decrease in the number of servers hence results in a decrease in Total Cost of Ownership (TCO).

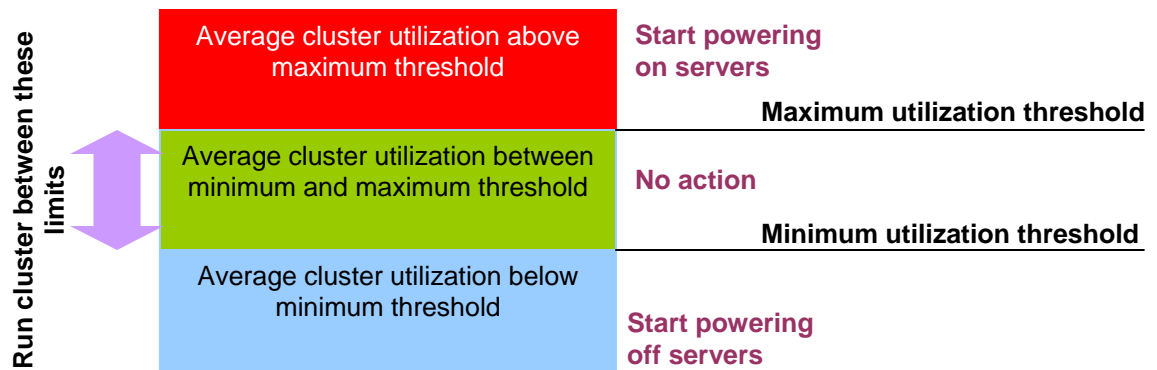
With increasing competition and price pressures, it is very important for any enterprise to keep costs down. More power savings can be extracted from a virtualized infrastructure, resulting in lower overall costs. Typical data center workloads have some characteristic pattern of peaks and troughs in utilization. The current practice of keeping all servers in a cluster powered on all the time is not an optimal use of power. If workloads can be automatically consolidated on fewer servers during off-peak hours, it can result in cost savings from not having to run extra servers. For example, workloads like internal data shares, mail, print and web servers may have peak utilization during office hours but go low on utilization during nights and weekends.

As described earlier in the paper, using VMware Distributed Resource Scheduling (DRS), many physical servers can be aggregated together to form a cluster. Such

a cluster can support multiple virtual machines running respective workloads. Using VMware SDK the cluster performance can be monitored and during periods when the average cluster utilization is low, virtual machines are automatically consolidated to a fewer servers and servers are automatically powered off using Dell Remote Access Controller (DRAC). As the utilization increases, the servers power on automatically to handle the increase in load, using DRAC and VMware SDK.

Steps involved in adaptive power management:

1. Poll the VirtualCenter and measure cluster level CPU and memory utilization using VMware SDK.
2. If the average cluster resource utilization goes below a set threshold for CPU or Memory utilization, start powering off servers until the cluster utilization gets above the minimum limits.
3. If the cluster level utilization goes beyond either maximum CPU or memory thresholds, start powering on servers until the cluster utilization gets below the maximum set thresholds.



3.2.1 User Configurable Parameters

An administrator can define the following set of user-configurable parameters in the configuration file (PowerSave.xml):

Maximum utilization threshold:

Average Maximum CPU Utilization (cpuMAX): The maximum limit on cluster CPU utilization. This is a percentage value of the total CPU resource available in the cluster.

Average Maximum Memory Utilization (memMAX): The maximum limit on cluster memory utilization. This is a percentage value of the total memory available in the cluster.

Servers will be automatically powered on after average cluster utilization goes above any one of the above set values.

Minimum utilization threshold:

Average Minimum CPU Utilization (cpuMIN): The minimum limit on cluster CPU utilization. This is a percentage value of the total CPU resource available in the cluster.

Average Minimum Memory Utilization (memMIN): The minimum limit on cluster memory utilization. This is a percentage value of the total memory available in the cluster.

Servers will be automatically powered off after average cluster utilization goes below both of the above set values.

Active VMs per server:

Minimum VMs per server (VMMin): Minimum number of active virtual machines per physical server.

Maximum VMs per server (VMMax): Maximum number of active virtual machines per physical server.

Administrators who want finer control on how many VMs are running at a given time on a physical server can use *VMMin* and *VMMax* parameters to control the cluster behavior. We make sure that the average number of active VMs is between these two values.

If the average number of active VMs per server is below the *VMMin*, no more physical servers will be powered on. If active number of VMs per server is more than *VMMax*, then no more physical servers will be powered off. This is irrespective of the Memory and CPU threshold. To override *VMMin* and *VMMax* settings, set *VMMin* = 1 and *VMMax* to some high value.

Timing parameters:

Operation Time-Out Period (opTimeout): The time it takes for a server to change state from either:

- Maintenance Mode to Production Mode or vice versa
- Powered On to Powered Off state or vice versa

Steady State Time (steadyState): Time from a threshold breach to any corresponding action. The algorithm will wait for *steadyState* time before taking any action based on a threshold breach. This parameter makes sure that random spikes or troughs in load do not cause servers to be powered on or off.

3.2.2 Power Management Algorithm

As explained in the previous section, adaptive power management aims to minimize the number of powered on physical servers that support the virtual infrastructure workload so that the net utilization of the cluster remains within specified minimum and maximum limits.

Below is the pseudo code for the algorithm to adapt number of active server to cluster load:

1. Get information about the hosts and VMs in the cluster
2. Using SNMP query get DRAC IP address of the ESX hosts (Use *snmpwalk* in Linux or *snmputil* in Windows)
3. Loop:
 - a. Poll the VirtualCenter server to get CPU and memory utilization and active VMs per server
 - b. If maximum threshold is reached for any one criteria and is held steady for user configured time
 - i. Select a powered off server and power it on using DRAC
 - ii. Once the server is connected to Virtual Center, put it in production mode
 - c. If minimum threshold is reached for ALL three criteria and is held steady for user configured time
 - i. Select a server and put server in maintenance mode
 - ii. Perform graceful shutdown using DRAC or VI web service.

4 Future Enhancements

This paper provides a starting point for administrators who want to reap benefits of Dell Systems Management with VMware Infrastructure. The algorithms presented can be enhanced or tweaked according to specific needs. Following are some enhancements that can be easily incorporated into this paper, especially in the adaptive power management algorithm:

- Optimum selection of server to power cycle: Selecting server(s) to power down based on least load and desired power cycles. In the current algorithm the servers are powered off starting with the first server in the cluster. However, selecting the server based on least load and making sure that any server is not getting power cycled beyond permissible limits would be useful.
- Consider historical workload characteristics: Recording workload characteristics over a period of time to understand workload trends would help to predict a peak or trough in load. This information can be used to proactively power on servers before more computing power is needed. This will make sure the desired computing power is available right before it's needed.
- Power-cycle based on time: Adding an option to power-cycle a set of servers based on specific time, e.g. power off 2 out of 4 servers in the cluster at 6:00 PM and power on at 6:00 AM.

- Extension for blade servers: Extending the adaptive power algorithm to include blade servers.

5 Conclusion

This paper provides a framework for integrating Dell OpenManage systems management and VirtualCenter, using VMware SDK. Individual servers are aggregated to form cluster resources. Servers can be dynamically added or removed to be a part of the cluster resource based on either the current utilization of the cluster or health status of the server. By integrating systems management, both proactive maintenance and adaptive power maintenance can be achieved.

6 References

- [*Implementing Fault Tolerance through Dell OpenManage and the VMware Software Development Kit*](#), Dave Jaffe and Todd Muirhead, Dell White paper.
- Dell|VMware alliance home page: www.dell.com/vmware
- Dell Systems Management: www.dell.com/openmanage

THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

Dell, PowerEdge is a trademark of Dell Inc. Intel and Xeon are registered trademarks of Intel Corp. VMware, the VMware “boxes “ logo, ESX Server, VirtualCenter and VMotion are either registered trademarks and/or trademarks of VMware, Inc. in the United States and/or other jurisdictions. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. Dell disclaims proprietary interest in the marks and names of others.

Copyright 2006 Dell Inc. All rights reserved. Reproduction in any manner whatsoever without the express written permission of Dell Inc. is strictly forbidden. For more information, contact Dell. Information in this document is subject to change without notice.