Application Delivery and Load Balancing for VMware View Desktop Infrastructure

A Dell™, F5 Networks® and VMware® Technical White Paper

End-to-End Solutions Team
Dell | Product Group – Enterprise

Global Strategic Alliance – Dell, Inc.
F5 Networks

Desktop Business Unit
VMware | Desktop CTO Office

Contents

Introduction

VMware View Virtual Desktop Infrastructure is designed to rapidly deploy secure and manageable Virtual Machine desktops to end users. A key component in green computing, VMware's consolidated approach to desktop management can reduce power and cooling requirements, simplify I/T processes and decrease risks related to desktop compliance as well as data storage, data security and data loss prevention. Desktop virtualization is becoming more popular in all business segments. In large deployments, network saturation and CPU bottlenecks on connection brokers are some of the headaches that need to be solved. Creating additional connection brokers help this situation, but to really make the return on investment stick a network load balancing solution with SSL off-load needs to be implemented as well. Network load balancing allows administrators to scale both security and connection servers. This paper addresses usage and configuration of F5 Network's BIG-IP© Local Traffic Manager™ (LTM) in combination with VMware's View 3.0 solution on Dell servers and Dell EqualLogic iSCSI storage. The next paragraph introduces LTM and describes some of the key features and benefits.

F5 Networks BIG-IP® Local Traffic Manager™ (LTM)

Application Delivery Controllers (ADCs) are modern hardware load balancers; rapidly evolving and far exceeding the power and functionality of traditional hardware load balancers. Their value extends from the network all the way through the applications at layer 7. F5 BIG-IP® Local Traffic Manager™ is an Application Delivery Controller and full proxy between users and application servers, creating a layer of abstraction to secure, optimize, and load balance VMware View traffic. LTM can make in-depth application decisions without introducing bottlenecks for clients and ensures that only explicitly allowed services can pass through to the View servers. Using the following LTM features, IT staff can intelligently and reliably deliver secure VMware virtual desktops to end-users.

LTM features include:

- **Scalability**, supports multiple Connection servers and multiple Security Servers by distributing client load across the View servers using selected load balancing and persistence methods. Enables load balancing of View using a ***single URL***.
- **High Availability** and **Fault Tolerance**, if any View server or LTM fails the service is maintained with minimal or no interruption
- **SSL Acceleration** or off-load, reduces processor utilization on View servers, accelerates the encryption/decryption processing, enables the ability to inspect and manipulate traffic once it has been decrypted
- **Protocol Optimization** and **Acceleration** improve client, server and network performance and reduce bandwidth requirements for LAN, WAN and web applications
- **Web Application Firewall** provides intrusion prevention and enables regulatory compliance for web applications
- **Security** features provide protection against DoS attacks, SYN floods and other network-based attacks. Unauthorized systems cannot directly connect to the backend VM infrastructure. LTM is a default-deny device; only services that are explicitly allowed in the configuration are allowed to pass through to the application servers. Resource cloaking and content security virtualizes and hides application, server error codes, and real URL references that may provide hackers with clues about infrastructure, services, and their associated vulnerabilities.
- **Scripted Traffic Control** provides the ability to create custom scripts to inspect and manipulate application traffic, on the fly, as it goes through the LTM.
- **Management API** provides access to automate administrative processes and provisioning on the LTM such as creating, deleting, enabling and/or disabling application servers

- **Application-Centric Configuration** allows simple and quick LTM set up for VMware View and many other applications
- **Single Point of SSL Certificate Management** on the LTM. Using SSL acceleration can help avoid managing individual certificates on each application server.
- **Unified Code** simplifies management and support over the lifetime of LTM. All models can run the same operating system code versions.
- **Central Management** using Enterprise Manager™ creates a single point of administration for up to 300 BIG-IP devices

The following sections will present key concepts, features and best practices for a LAN implementation of F5 BIG-IP LTM Application Delivery with VMware View Connection servers.

Load Balancing for VMware View 3.0 Components

Basic load balancing can be achieved using software, hardware or DNS round-robin methodologies. Hardware load-balancing is the most robust mechanism and, unlike software or DNS, it can provide superior reliability, features and performance through the use of specialized hardware and software. As shown in figure 1, there are two components of a VMware View deployment that require hardware load-balancers for a scalable configuration, VMware View Manager Security servers and VMware View Manager Connection servers.

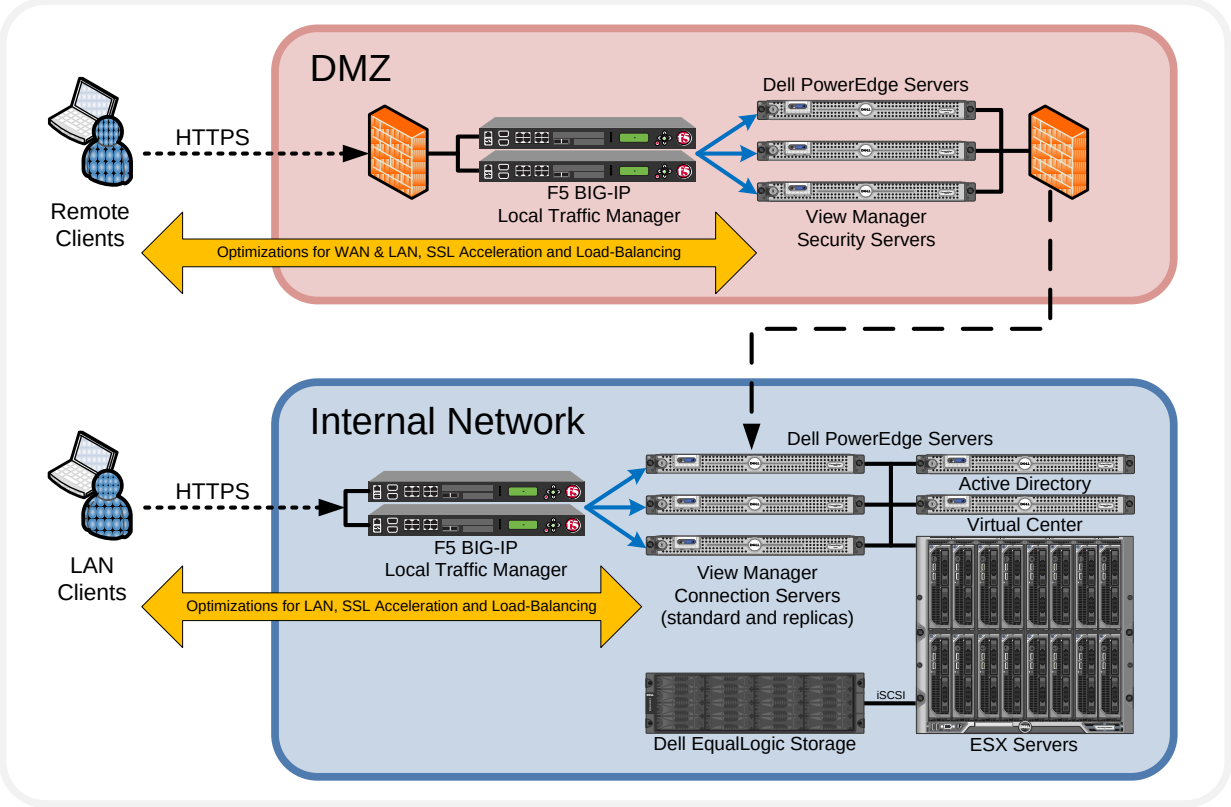

Figure 1: Application Delivery for VMware View

Remote clients accessing VMware View from an external network, such as the Internet, can communicate with the View Manager Security servers over encrypted TLS/SSL tunnels. The load balancer and Security servers are located in a DMZ network, protected by firewalls. The DMZ load balancer provides decryption, encryption and load balancing of the client connections across the Security servers according to the selected load balancing and persistence methods. The Security servers provide VMware View clients with secure logon and private/encrypted access to the VM infrastructure. The Security servers communicate directly with the Connection servers in a one-to-one relationship. Each Security server communicates with a single Connection server and it is not possible to load balance the connections between the Security servers and the Connection Servers. With the connections established, the remote users are able to safely access the scalable VMware View infrastructure.

LAN clients accessing VMware View from an internal network can communicate directly with the View Manager Connection servers over encrypted TLS/SSL tunnels. The internal network is not exposed to the Internet or other external networks. The internal load balancer decrypts/encrypts, inspects and distributes the requests across the Connection servers allowing users to open VMware View desktop virtual machines on the ESX servers. With application delivery and load balancing configured, the VMware View Connection services are scalable and fault tolerant, maximizing application performance and minimizing user downtime in the event of a failure.

LTM Features

Traffic Management Operating System (TMOS) Version 10 (V10) is the most current release of the LTM operating system. Highlighted are five features that can be used in a VMware View deployment, the new V10 features are noted.

Dashboard

For system management, LTM can be can be accessed using a secure web browser (HTTPS), secure shell connection (SSH/CLI) or direct console connection (CLI). V10 includes a new graphical reporting engine accessible from the secure web interface that displays real-time historical statistics by the hour, day, week, or month. This new tool called Dashboard reports statistics on CPU and memory usage, connections and throughput in a graphical view as shown in figure 2. It is possible to monitor VMware View traffic throughput, connections and SSL acceleration using the LTM Dashboard.

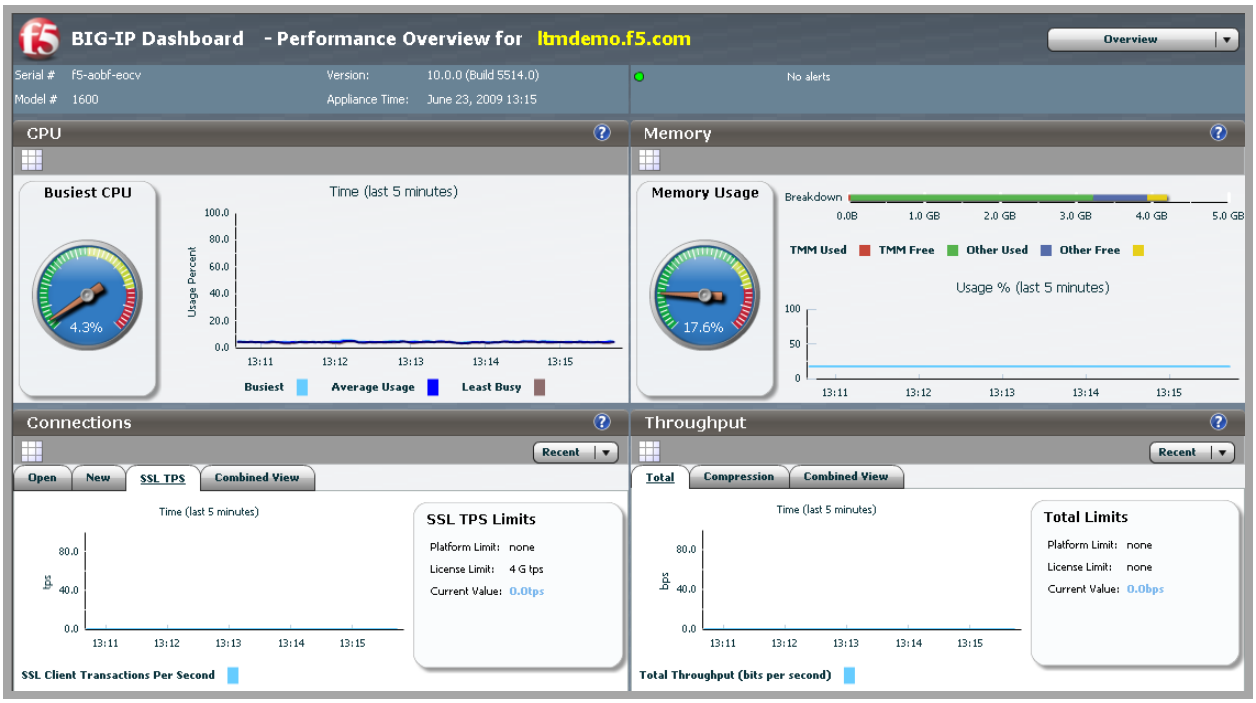

Figure 2: LTM V10 Dashboard

iRules

iRules are an LTM feature that provides detailed controls to manipulate and manage any IP application traffic. iRules are event driven and utilize a scripting syntax based on Tool Command Language (TCL) enabling IT staff to customize how the LTM intercepts, inspects, transforms, and directs inbound or outbound application traffic. When applied to an LTM virtual server, iRule scripts are executed against the application traffic that is associated with that virtual server. For VMware View, iRules can be used within a persistence profile to identify unique information about each View client session and then use this information to build the connection table and load balance the traffic. iRule persistence is included in the V10 VMware VDI application template, both are explained later in this paper. For more information on iRules and other development topics logon to http://devcentral.f5.com/, DevCentral is a free developer forum.

TCP Express

TCP Express is an advanced network stack developed by F5 for the LTM that combines various techniques and improvements in the RFCs to minimize the effects of network congestion, latency, packet loss, and recovery. It can also shield and transparently optimize older or non-compliant TCP stacks that may be running on servers or clients. TCP Express features are enabled by the V10 VMware VDI application template through the use of protocol profiles. LTMs come pre-configured with profiles, such as WAN and LAN optimized profiles, which make use of the LTM full proxy architecture to optimize network communications on both the server-side and the client-side. Using protocol profiles in a VMware View deployment can improve bandwidth efficiency, improve performance and reduce server overhead under a variety of network conditions. Profile features can be enabled, disabled or customized by the administrator. Examples of RFCs implemented in the tcp-wan-optimized profile are: RFC 2018-TCP Selective Acknowledgements SACK, RFC 1323-TCP Extensions for High Performance, Scaled Windows and TimeStamps, RFC 3390-TCP Slow Start, Increasing TCP's Initial Window and RFC 3465-TCP Congestion Control with Appropriate Byte Counting. See RFC web links in the More Information section.

OneConnect
The OneConnect feature can reduce millions of VMware View client TCP connections down to just hundreds of server-side TCP connections, this is also known as TCP re-use or TCP multiplexing. The LTM full proxy allows OneConnect to optimize TCP connections across View clients and servers, eliminating server overhead associated with TCP connections. This improves server utilization by allowing the Security and Connection servers to allocate maximum resources to the VMware processes. OneConnect is enabled in the V10 VMware VDI application template.

iControl
iControl is a web services-enabled open API providing secure access to configuration, management and monitoring of LTM and it can be integrated with business process and workflow applications. Integration with VMware allows for automated provisioning and de-provisioning of applications in computing environments. Some examples of automation using iControl are: adding servers dynamically to an application pool, stopping requests from going to an application server, changing the way requests are routed to servers and influencing the choice of servers based on current application or server load.

These are just a few of the LTM features that can be used for VMware View application delivery and load balancing. Secure web management with Dashboard monitoring, iRules real-time custom scripted traffic control, iControl Management API and advanced network stack optimizations like TCP Express and OneConnect extend the LTM benefits beyond traditional load balancers for a VMware View deployment.

Application Delivery Best Practices for VMware View

The following sections describe best practices for LTM application delivery in a VMware View deployment.

High Availability
High Availability (HA) mode provides application fault tolerance for VMware View or any other application delivered through an LTM. To support HA, the LTMs must be deployed in pairs consisting of an Active unit and a Standby unit. In the event of a failure, the Standby unit detects the problem, performs a sub-second failover and resumes all VMware View requests therefore minimizing interruption of service to the clients. Repairs can then be made during a scheduled maintenance window. Additional LTM HA-specific configurations may include failover cable connections, MAC masquerading, configuration synchronization, connection mirroring and VLAN or network failover. V10 also introduces a device wizard to simplify the configuration of an LTM high availability pair. HA is recommended for all production and other deployments where application fault tolerance is needed.

Application Templates
New to the LTM platform with V10, application templates allow for quick and easy configuration of the LTM for various software applications including VMware View (formerly known as VMware VDI). Templates reduce complexity, human error and resources required to deploy applications. For example, using the LTM VMware VDI 2.1 application template only takes a couple of minutes to complete. From the LTM secure web management interface, simply expand the Templates and Wizards menu, click on Templates, click on VMware VDI and the Template interview screen is displayed. Fill in the requested information such as application prefix name, IP addresses, ports, SSL, WAN/LAN optimization, etc. and then click Finished. The LTM is now ready to deliver VMware View. In addition, V10 with security

partitions and application templates allows for the safe delegation of tasks and responsibilities. It is possible to have people with a limited knowledge of the LTM platform safely configure application delivery for a VMware View or other application deployment. Templates are developed and certified by F5 Networks to provide the optimal settings for the application; however, the LTM application settings can still be customized by the administrator. See the example of HTTP profile settings later in this paper.

As a best practice, use LTM application templates to ensure fast and consistent VMware View deployments with less complexity. The following LTM configuration objects are created by the VMware VDI application template process, explanations and some best practices are given for each.

Virtual Servers
LTM virtual servers are created by the application template and they are a key component for application delivery and load balancing, representing the VMware View application as a name and IP:port combination on the LTM. This configuration object manages a variety of settings such as protocol profiles, SSL acceleration profiles, persistence profiles, server pools, connection mirroring, SNAT, iRules and other traffic management features associated with the application. In this paper, the term virtual server refers to the LTM configuration object and should not be confused with features or functions of VMware. Figure 3 shows an example of a virtual server configuration created using the application template.

During the application template set up process, the administrator supplies information customizing the configuration for the virtual server and related objects. As shown in figure 3, the virtual server has the name Demo_VDI_https_virtual_server and the IP address 172.16.100.100 (VIP) with service port TCP 443 (HTTPS). The state is enabled so LTM listens for application traffic on this IP and port. There are also custom profiles applied, these objects are named using the prefix Demo_VDI_ specified by the administrator at set up. More information about profiles and persistence are presented later in this section.

Upon completion of the LTM application template, DNS should be configured to resolve the application name to the virtual server IP address so that VMware View clients are directed to the LTM for services. When clients attempt to access the virtual desktops they will communicate through the LTM to the backend VMware servers. For example, View clients use a browser to access https://vmwareview.dell.local/, the DNS query gets resolved to the IP address 172.16.100.100 and the clients get directed to the LTM virtual server. ***The View Manager configuration must be modified for each View Server; pointing the External URL to the DNS name and service port of the LTM (ex: https://vmwareview.dell.local:443/).***

Additional Information: *Security for web applications is a critical component in any application deployment. LTM provides application security out of the box because it is a default-deny system. Traffic cannot pass through the LTM until an administrator creates the Virtual Server and enables the specific application traffic to pass. In addition, LTM is a full proxy which protects or shields the backend application servers from direct TCP and UDP connections coming from valid users or others. LTM add-on software modules are also available to help protect web applications from unwanted traffic and attack. The modules can be loaded directly on to LTM without any additional hardware required. The LTM 3600 platform and above have support for add-on software modules. BIG-IP®Application Security Manager (ASM) add-on software module is a Web Application Firewall (WAF) that can reduce and mitigate risks associated with web applications. It can also help to meet regulatory compliance requirements. BIG-IP®*

Protocol Security Manager (PSM) is an add-on software module that provides broad security protection for HTTP, SMTP and FTP. ASM and PSM best practices are not covered in this paper.

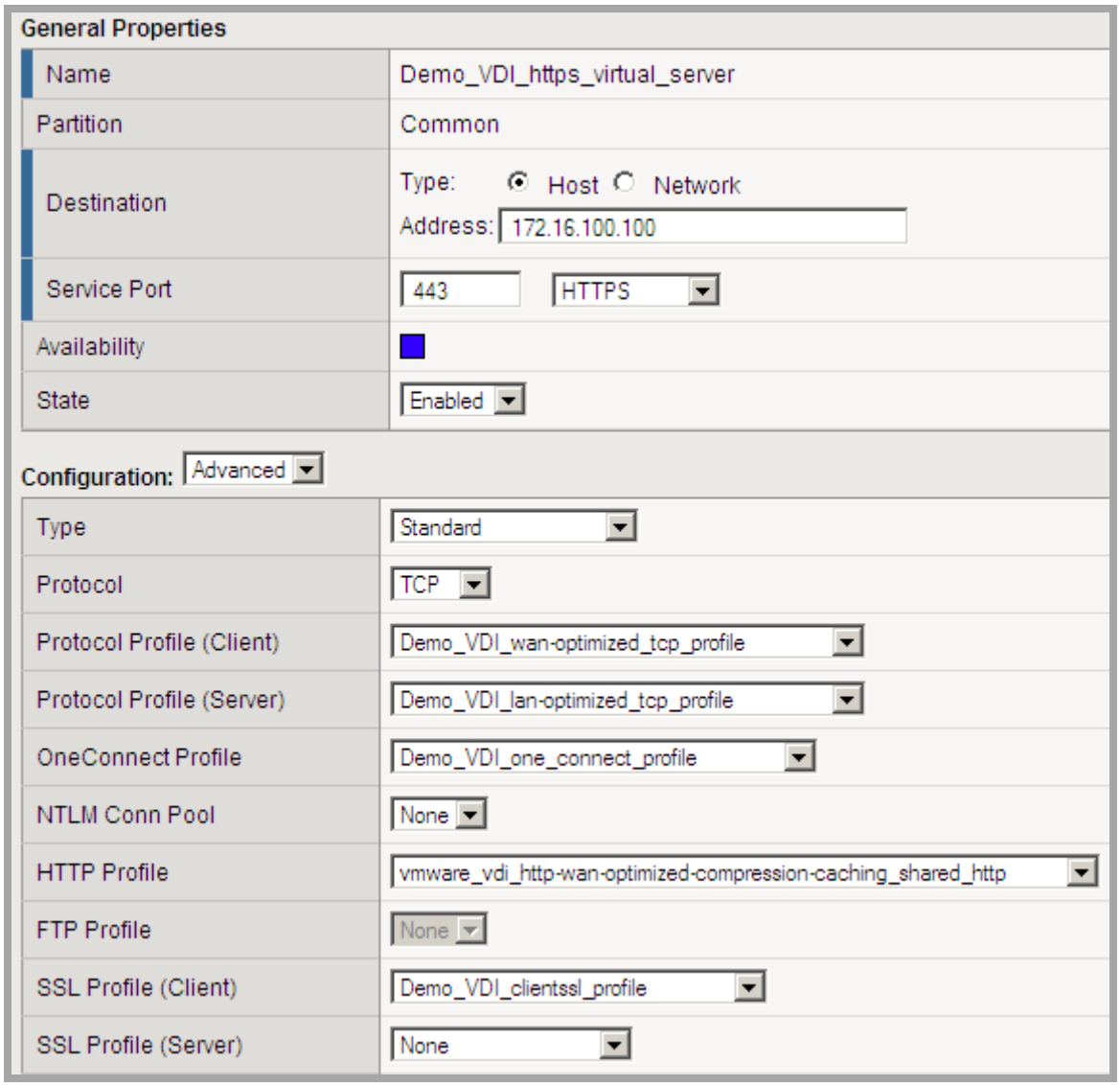

Figure 3: LTM Virtual Server Configuration for VMware View

Profiles

Profiles are commonly used for granular control, acceleration and optimization of the LTM network stack and related application traffic. There are various types of profiles used for LTM configuration including protocol, SSL, OneConnect, HTTP, FTP and others. For each type of profile there may be additional sub-categories. The following profiles pertain to a VMware View infrastructure deployment.

SSL Acceleration Profiles

SSL acceleration profiles should be used to off load encryption processing to the LTM and reduce CPU utilization on the VMware View Security and Connection servers. A certificate and key must be loaded on the LTM and associated with the SSL client profile. Then associate the SSL client profile with the virtual server and when the VMware View clients connect to the virtual server the LTM will decrypt the

incoming requests and load balance them to the VMware servers. In figure 3, SSL Profile (Client) is configured for the customized profile named Demo_VDI_clientssl_profile.

TCP Optimization Profiles

TCP protocol profiles allow administrators detailed control over the TCP stack of the LTM. Taking advantage of the LTM full proxy architecture, these profiles should be applied to virtual servers on the client-side and the server-side to optimize both LAN and WAN connections. While providing granular control, profiles are also easy to use. For example, the VMware VDI application template prompts the administrator to specify whether clients will connect over a LAN or a WAN. Based on the response, the appropriate TCP protocol profiles are automatically created and assigned to the virtual server. As shown in figure 3, Protocol Profile (client) has been assigned the WAN optimized TCP profile and Protocol Profile (server) has been assigned the LAN optimized TCP profile. The result will be better network and application performance for both clients and servers.

OneConnect Profiles

OneConnect, mentioned in the Features section, can significantly reduce the number of server-side TCP connections therefore reducing network overhead on the backend servers. OneConnect is recommended for VMware View deployments and the application template automatically assigns the profile to the new virtual server. Figure 3 shows that the OneConnect Profile is enabled and configured for Demo_VDI_one_connect_profile.

HTTP Optimization Profiles

An HTTP profile is configured for the VMware View virtual server using a WAN optimized HTTP profile, shown in figure 3. Because the LTM is performing SSL acceleration, the HTTP traffic is decrypted (in the clear) and the LTM is able to inspect and accelerate the traffic. When optimizing HTTP, object caching is critical to server performance and bandwidth utilization for both LANs and WANs. Static content such as images, style sheets and documents (PDF, text, etc.) should be cached at the LTM and the client to reduce overhead on the web server and eliminate unnecessary traffic from traversing the network. For WAN connections, object compression also improves performance by reducing the number of bits that cross the WAN link. However, there is client processing overhead associated with compression therefore only HTTP WAN optimized profiles have this feature enabled. It is not recommended to use HTTP WAN optimized profiles or object compression for LAN connections. LANs have plenty of available bandwidth and low latency so it is better to skip the compression and transmit the objects at full size. It is possible to view and modify profiles using the LTM secure web interface. As shown in figures 4, 5 and 6, the HTTP profile for VMware View has numerous settings that can be customized, including values for Compression and RAM Cache.

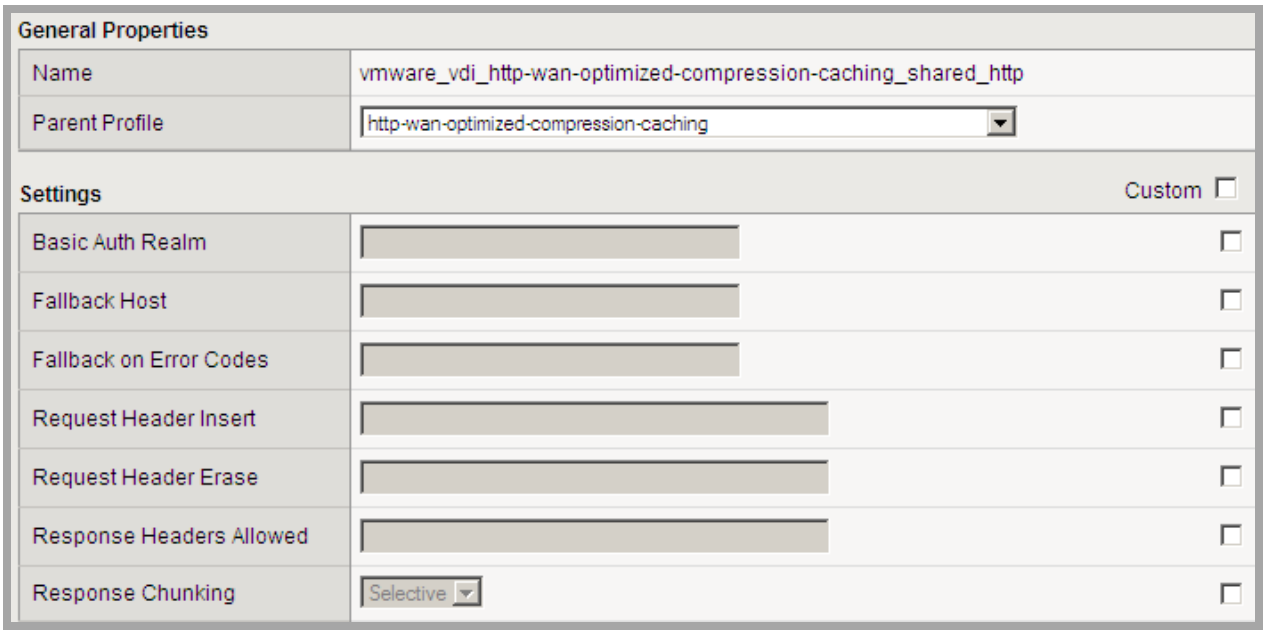

Figure 4: General Properties and Settings, HTTP WAN Optimized Profile for VMware View

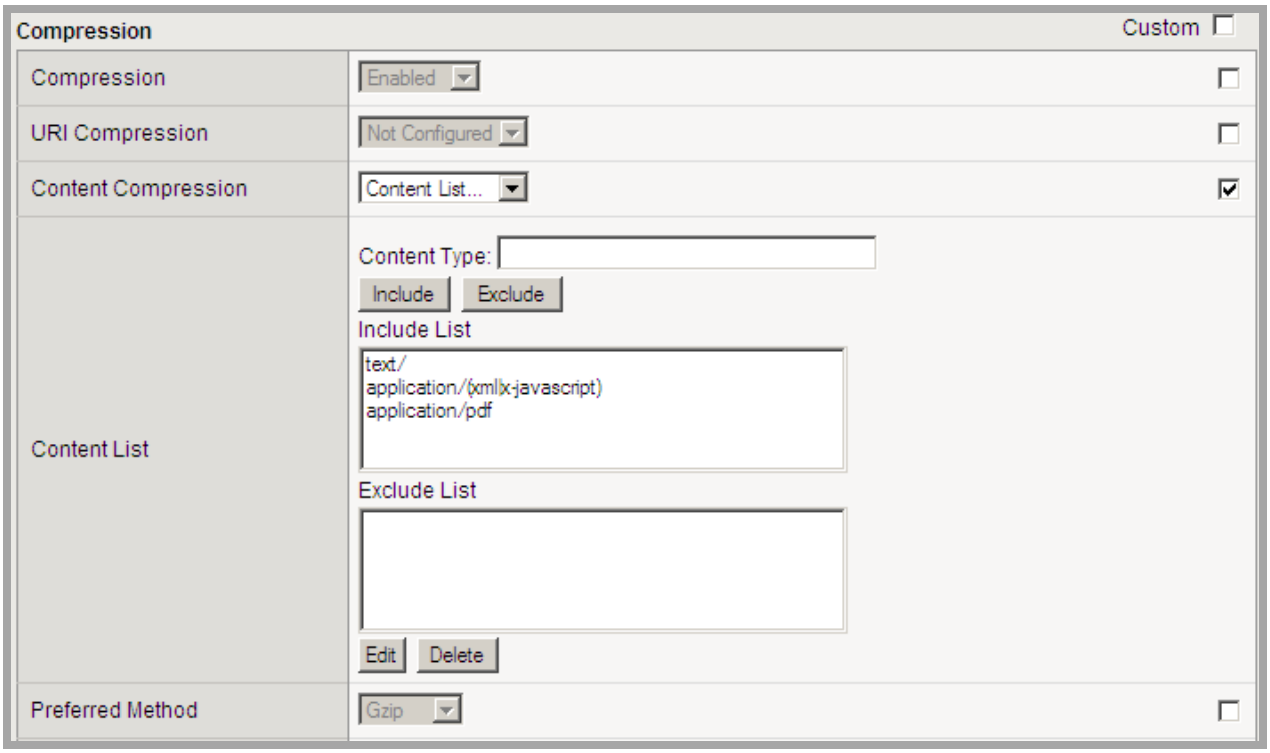

Figure 5: Compression, HTTP WAN Optimized Profile for VMware View

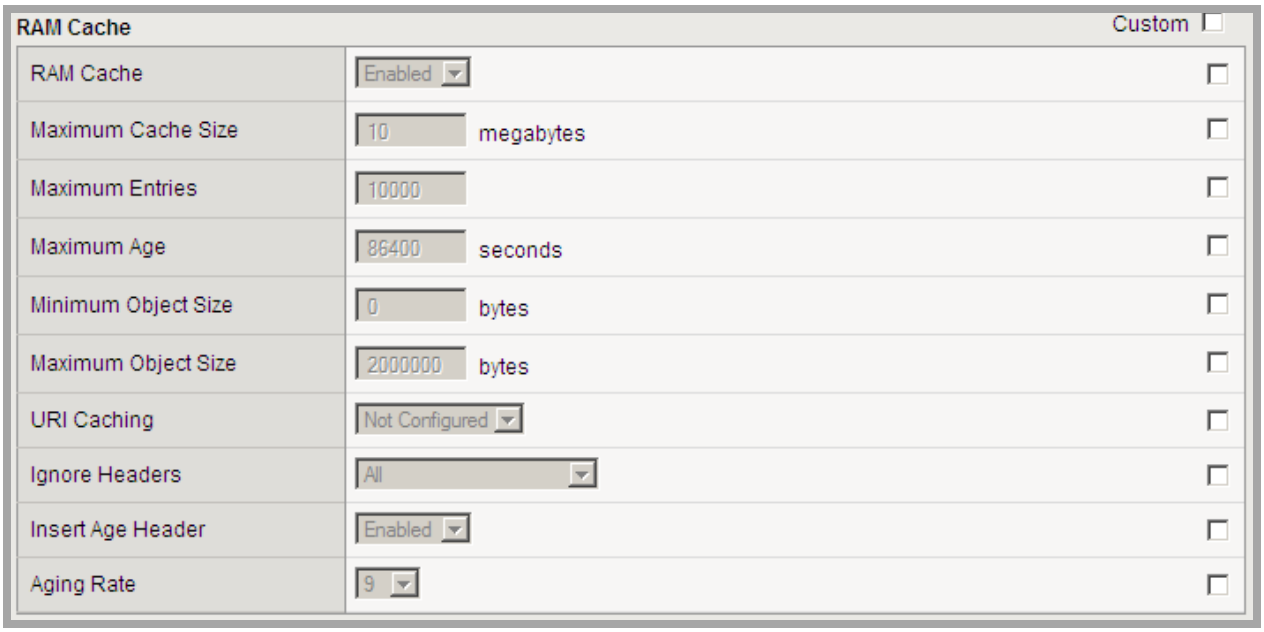

Figure 6: RAM Cache, HTTP WAN Optimized Profile for VMware View

Persistence Profiles using iRules and UIE

iRules, mentioned in the Features section, can be used in various ways to intercept, inspect and modify application traffic as it passes through LTM. In the case of VMware View, an iRule is used within a Universal Persistence Profile to enable granular control over the load balancing and persistence (stickiness) based on client session information. The iRule configuration achieves a more efficient load distribution to the backend servers and allows load balancing of VMware View using a ***single URL***. Clients are directed to the LTM Virtual IP (VIP) via DNS. The LTM Virtual Server and iRule perform persistence and load balancing allowing the View client control channels to be load balanced and the View client secure tunnel channels to be directed to the correct Connection servers. For this iRule to function correctly, the LTM virtual server must be configured for SSL acceleration so that LTM can see the payload of the HTTP traffic. The following paragraphs and figures explain how this works.

Clients launch the View software making requests (HTTP_REQUEST) to the Connection servers through the LTM. The first response from the server (HTTP_RESPONSE) contains a JSESSIONID cookie and the iRule enters that session ID into the LTM connection table. Upon further client requests, the iRule looks for the session specific information in a cookie or in the URI argument to establish persistence for the control and tunnel channels. For more information on the iRule persistence, refer to the script shown in figure 7 and the iRule processing flow shown in figure 8.

The VMware VDI application template automatically creates this iRule and assigns it to the persistence profile. The profile is then assigned to the new virtual server. It is recommended to use this method of persistence with VMware View.

Properties

Name	Demo_VDI_persistence_irule
Partition	Common
Definition	

```
when HTTP_REQUEST {
  if { [HTTP::cookie exists "JSESSIONID"] } {
    set jsess_id [string range [HTTP::cookie "JSESSIONID"] 0 31]
    persist uie $jsess_id
  } else {
    set jsess [findstr [HTTP::uri] "tunnel?" 7]
    if { $jsess != "" } {
      persist uie $jsess
    }
  }
}
when HTTP_RESPONSE {
  if { [HTTP::cookie exists "JSESSIONID"] } {
    set jsess_cookie [HTTP::cookie "JSESSIONID"]
    persist add uie [HTTP::cookie "JSESSIONID"]
  }
}
```

Figure 7: LTM Persistence iRule for VMware View

Universal Inspection Engine (UIE)

UIE and iRules are LTM features that can read any value(s) of an IP-based packet header or payload and direct it to the appropriate resource. Universal persistence takes this iRules feature one step further, by allowing persistence for sessions based on content data, or based on connections to a specific member of a pool. Universal persistence does this by defining some sequence of bytes to use as a session identifier (ex: the JSESSIONID cookie). UIE allows you to correlate connections/TCP streams according to this specified data. As shown in figure 7, line 15 of the iRule issues the command “persist add uie [HTTP::cookie “JESESSIONID”]” which uses the Universal Inspection Engine to analyze the HTTP traffic and add session-specific information in the connection table to be used for persistence. On lines 4 and 8, the command “persist uie …” persists the connection based on the contents of the cookie or the URI argument. In this case, the VMware View control channel and the tunnel channel connections are correlated in the persistence table by this session information. The server that receives the load balanced control channel connection will also receive the secure tunnel channel connection.

When the LTM is processing traffic, persistence always comes before the load balancing. A key function of this iRule is to determine which requests are eligible for load balancing. Those requests that are not eligible for load balancing go through the persistence processing. Only the initial View client HTTP request is load balanced and all other connection requests, including the tunnel request, are routed using persistence.

Figure 8 shows the iRules UIE persistence processing flow for VMware View.

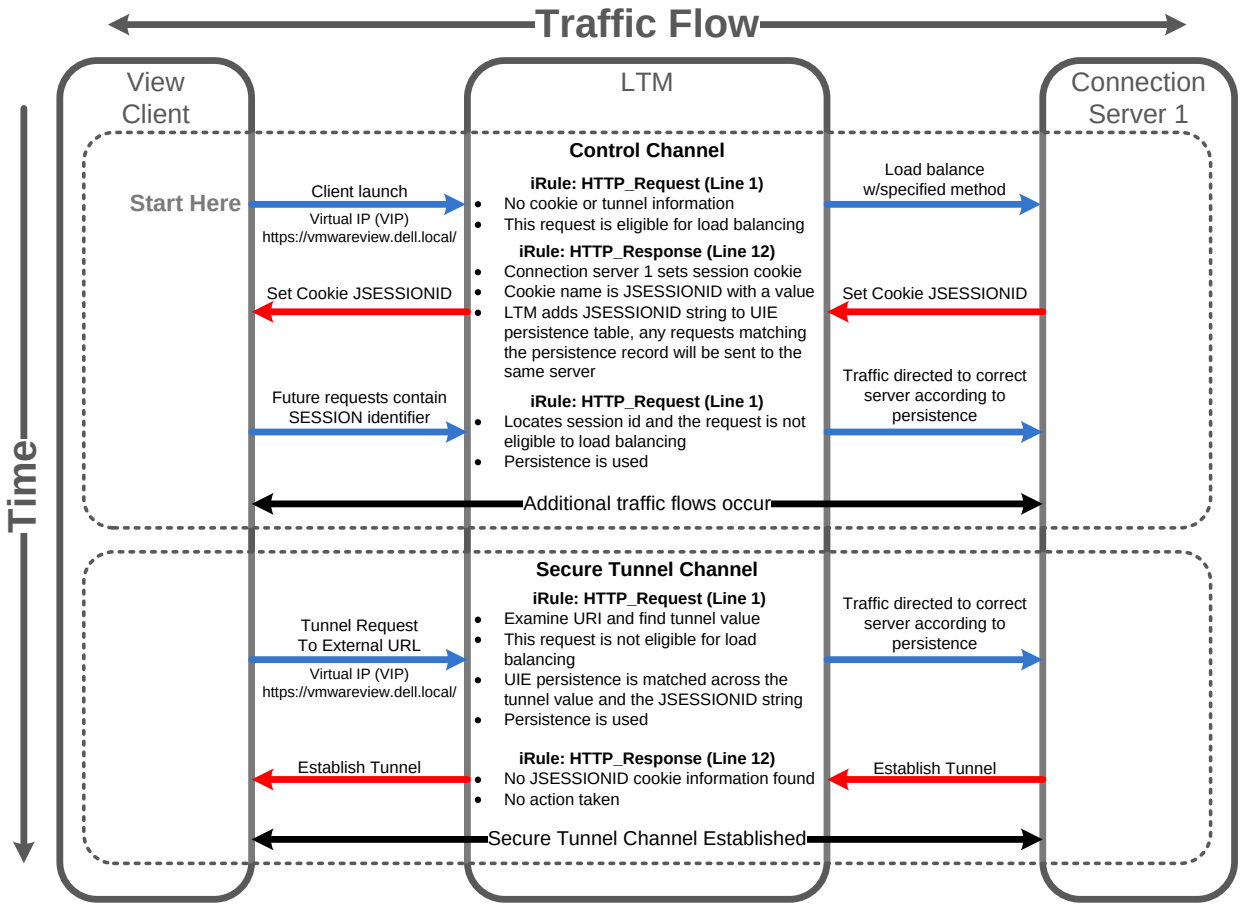

Figure 8: iRules UIE Persistence Processing Flow for VMware View

Load Balancing

Load balancing provides application scaling and fault tolerance to a VMware View deployment. Pools are another key component for LTM application delivery created by the application template; managing health monitoring, load balancing and pool membership. For example, a pool named Demo_VDI_connection-pool is created and customized by the application template and then it is associated with the new virtual server. The pool settings consist of the following:

- **Pool Members** are created using the IP:port combination of backend servers. VMware View deployed with SSL acceleration requires the Connection servers to listen on TCP 80 for HTTP. Examples of the pool members are 172.16.200.101:80, 172.16.200.102:80 and 172.16.200.103:80. Pool members can be added, deleted, enabled or disabled by the administrator.

- **HTTP Health Monitors** actively check the health and availability of the application server web listener. F5 recommends a 1:3+1 interval/timeout ratio therefore a 30 second interval and 91 second timeout would be appropriate for VMware View. The interval determines the frequency of the health checks against each server and the timeout defines how long the health monitor will wait for a server response before marking a pool member down. The values can be adjusted by the administrator according to the needs of the customer. The application template automatically assigns a basic HTTP monitor to the pool.

For compatibility with VDM 2.1 and as a best practice for VMware View, it is recommended to use the Send String "GET /favicon.ico HTTP/1.0". The Receive String contains a regular expression that is used to analyze the Connection server's response. No response or any response other than an HTTP 200 will mark the pool member down. Figure 9 show the Send String and Receive String configurations.

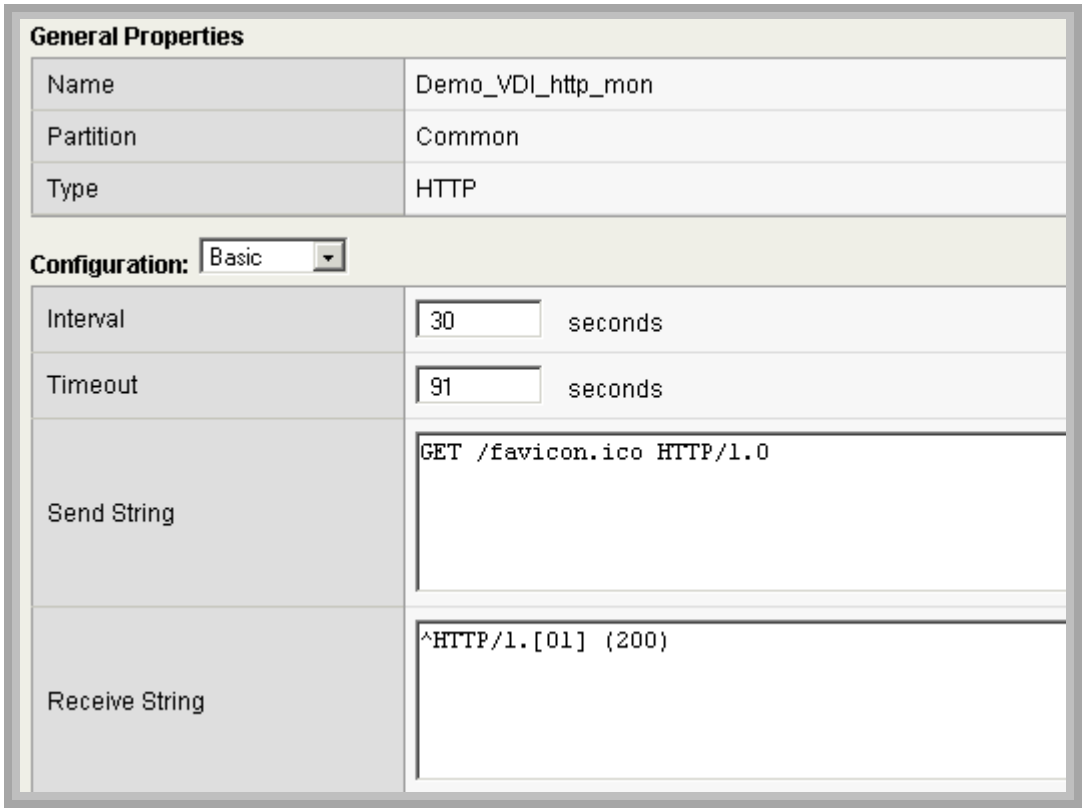

Figure 9: Example HTTP Monitor

***Additional Information**: ECV (Extended Content Verification) health monitors initiate a connection with the server, request data and then compare the returned data with a pre-configured search string. EAV (Extended Application Verification) health monitors use an external program or script to monitor a server. A script or program is required if you use an EAV with a protocol other than those that are pre-configured for BIG-IP.*

- **Least Connections (member)** load balancing method distributes VMware View traffic to the pool member (Connection servers) with the fewest number of open connections for that pool. This method provides a very even load distribution and good use of resources on comparably equipped servers.

- **Slow Ramp Time** ensures that if a pool member becomes available after maintenance or a new member is added, the Least Connections load balancing algorithm does not send all new connections to that member because a newly available member will always have the least number of connections. F5 recommends 300 seconds (5 minutes) for VMware View.

Detailed configuration instructions for pools and other features are provided in the F5 Deployment Guide for VMware Virtual Desktop Infrastructure. See the web link in the More Information section.

Conclusion

Hardware load balancing is recommended to scale a VMware View infrastructure beyond one View Manager Security or one View Manager Connection server. SSL acceleration and load balancing maximize the utilization of the Dell servers and eliminate the need to manually assign View clients to a particular server. F5 BIG-IP Local Traffic Manager is an Application Delivery Controller platform that combines traditional hardware load balancing with advanced security, network and application delivery features. Application-centric configuration allows IT administrators to quickly set up and optimize LTM for VMware View using a single URL. Deploying VMware View with Dell servers and F5 BIG-IP Local Traffic Manager using the features and best practices described in this paper can help ensure that the virtual desktop infrastructure is secure, scalable, flexible and available.

More Information

For more information on Dell's Virtualization Solutions refer to the Dell Virtualization Implementation and Architecture web site.

Dell:
http://www.dell.com/

VMware View:
http://www.vmware.com/products/view/

F5 Networks, Dell Technology Alliance:
http://www.f5.com/dell/

F5 Networks, Products Overview:
http://www.f5.com/products/

F5 Deployment Guide - VMware Virtual Desktop Infrastructure (BIG-IP v10 System):
http://www.f5.com/pdf/deployment-guides/vmware-vdi-big-ip-v10-dg.pdf

F5 Deployment Guide - Tuning the OneConnect Feature on the BIG-IP LTM (BIG-IP v10, 9.x):
http://www.f5.com/pdf/deployment-guides/oneconnect-tuning-dg.pdf

RFC 2018, TCP Selective Acknowledgements (SACK):
http://www.ietf.org/rfc/rfc2018.txt

RFC 1323, TCP Extensions for High Performance, Scaled Windows and Time Stamps:
http://www.ietf.org/rfc/rfc1323.txt

RFC 3390, TCP Slow Start, Increasing TCP's Initial Window:
http://www.ietf.org/rfc/rfc3390.txt

RFC 3465, TCP Congestion Control with Appropriate Byte Counting:
http://www.ietf.org/rfc/rfc3465.txt