Owner of the content within this article is www.isaserver.org
Written by Marc Grote – www.it-training-grote.de

Microsoft Forefront TMG – Webserver Load Balancing

Abstract

In this article I will show you how to configure Forefront TMG Server webserver Load Balancing capabilities to balance the load to multiple internal web servers. I will also cover some NLB basics of Forefront TMG and Windows Server 2008 R2 to complete the overview of the load balancing capabilities of Forefront TMG and Windows Server 2008 R2.

Let’s begin

Forefront TMG can distribute Web traffic to identical configured web servers that are normally a special function of a Hardware load balancer. Web server load balancing distributes network traffic to different hosts in the internal network without using classic NLB functions of the Windows operating system.

It is possible to publish a hardware load balancing device to balance web traffic to internal web server but Forefront TMG web farm load balancing has a number of advantages (but also disadvantages):

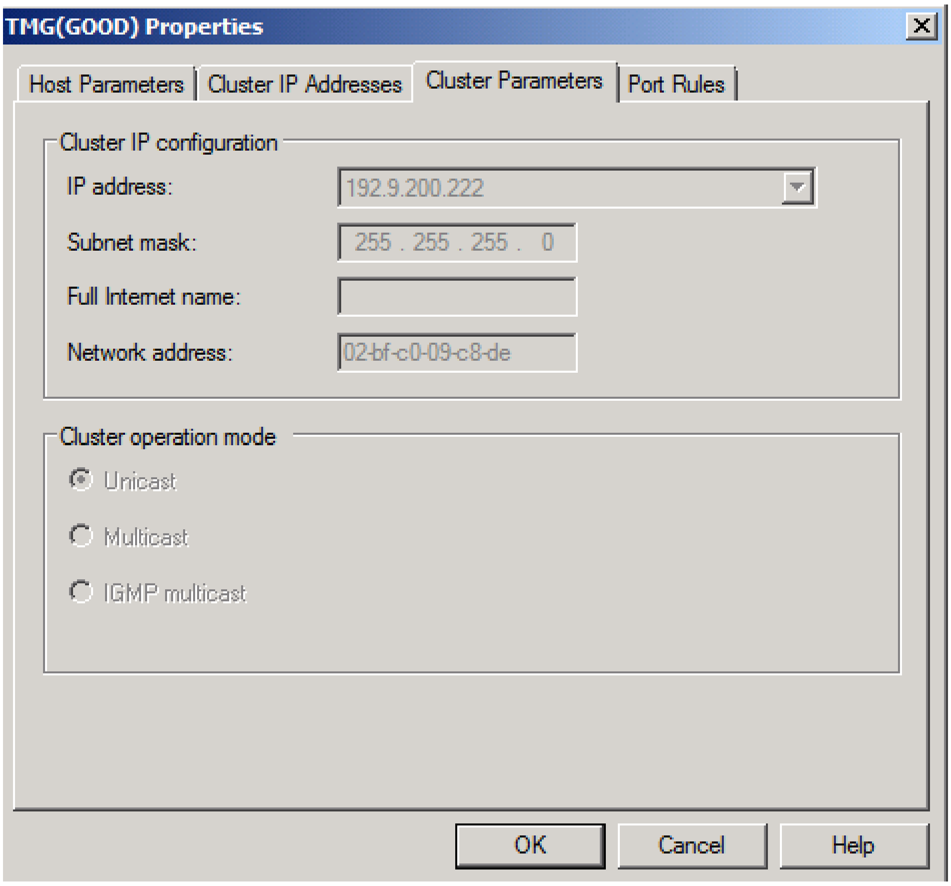
Some Hardware load balancer uses source IP addresses to balance web requests, but this solution might only be suitable in environments where these servers are not behind a NAT device. Forefront TMG doesn’t forward the original IP address in a standard web server publishing scenario. The IP address from the external client will always be masked with the IP address of the TMG Server. If you want to be able to forward the original client IP from the external requesting client, the published web servers has to set its Default Gateway to Forefront TMG which is not suitable in some environments.

Another way to distribute traffic to web servers is to use the Windows integrated Network Load Balancing (NLB) mechanism. NLB allows distributing network traffic based on port rules. All nodes in an NLB cluster use one Virtual IP address (VIP) which is used by Forefront TMG to forward traffic. The NLB algorithm distributes traffic across the NLB cluster members.

Network Load balancing basics

Very briefly; NLB is a kind of cluster technology which is not exclusive to Microsoft Windows. NLB is part of the Windows Server 200x operating system family and is used to distribute network traffic for up to 32 hosts in the network. NLB uses a distributed algorithm that load-balances incoming traffic to all nodes in a Windows NLB cluster. So, NLB can be used to provide failover and Load balancing capabilities

It is possible to enable the Network Load Balancing feature on every Windows Server 2008 version. The following figure shows the Windows Server 2008 R2 Network Load Balancing Manager with only one NLB node.



Cluster IP configuration

IP address: 192.9.200.222

Subnet mask: 255 . 255 . 255 . 0

Full Internet name:

Network address: 02-bf-c0-09-c8-de

Cluster operation mode

● Unicast

○ Multicast

○ IGMP multicast

NLB with Forefront TMG

If you plan to load balance internal Web Servers with the Forefront TMG Web Server Farm Load Balancing feature, you should also keep in mind, that Forefront TMG Server might be the single point of failure (SPOF) when TMG is not load balanced. Forefront TMG Enterprise uses NLB to load balance TMG Server. It is possible to use NLB in integrated mode, the preferred and recommended mode in Forefront TMG. It is also possible to use NLB with Forefront TMG Standard but this is not official supported by Microsoft and has some limitations.

Load balancing mechanism

Round-robin

Webserver requests from different IP addresses will be distributed among the Web farm members. The round-robin mechanism ensures that the request of the user to a Web application serviced by a Web farm is distributed evenly among farm members that are online. When failover occurs, servers that are not responding are detected, and the load is distributed among the available servers.

Cookie based affinity

Session (Cookie) based affinity is normally used to publish Outlook Web Access (OWA) from Exchange Server 200x or Microsoft SharePoint services/Servers sites. You should not use Session affinity if you want to publish RPC over HTTP(S) services or Outlook Anywhere in Exchange Server 2007 and higher. RPC over HTTP(S) is used to give Outlook clients full access to Exchange Server from the Internet. RPC traffic will be tunneled through HTTPS. With Outlook it is not possible to use Cookie based affinity.

IP affinity

With IP affinity, the web server traffic is distributed based on IP to all members of the Web farm. If one Server fails to respond, the traffic will be send to another member of the Web farm.

You should not use IP based affinity, if remote clients are located behind a NAT server, because the web server farm will only see the IP address of the TMG Server. If this is the case you should use Session affinity, if it is possible.

IP affinity is useful in an Exchange RPC over HTTP(S) also called Outlook Anywhere scenario, where session affinity cannot be used or in Exchange Active Sync publishing scenarios where the client does not fully understand HTTP 1.1 (which is needed for cookie based affinity).

To create a webserver load balancing publishing rule start the TMG management console and navigate to the Firewall policy node and create a Web Site Publishing rule.

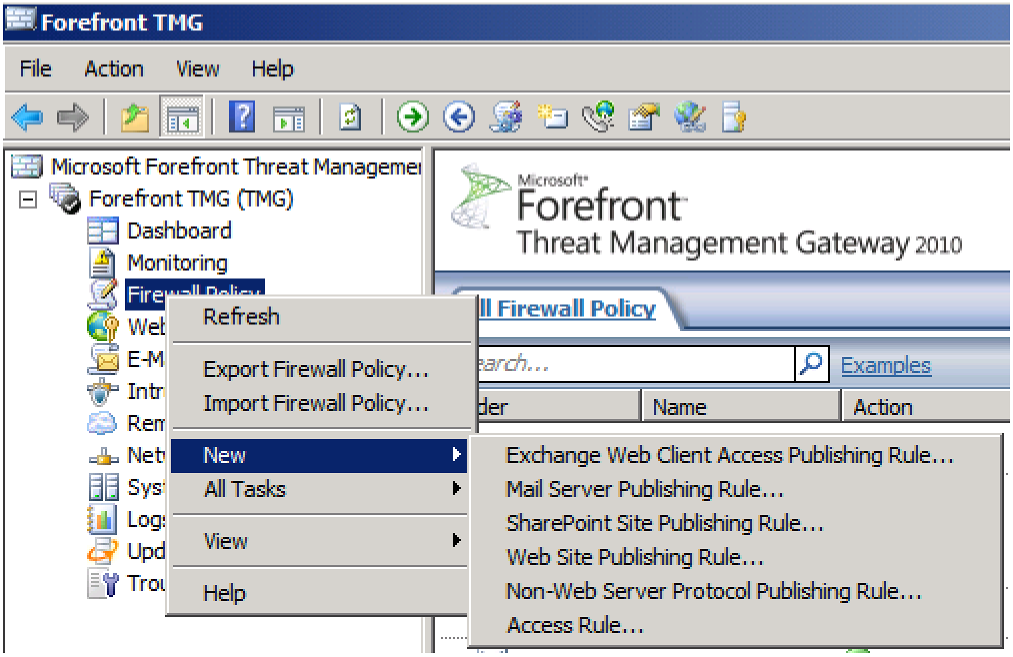
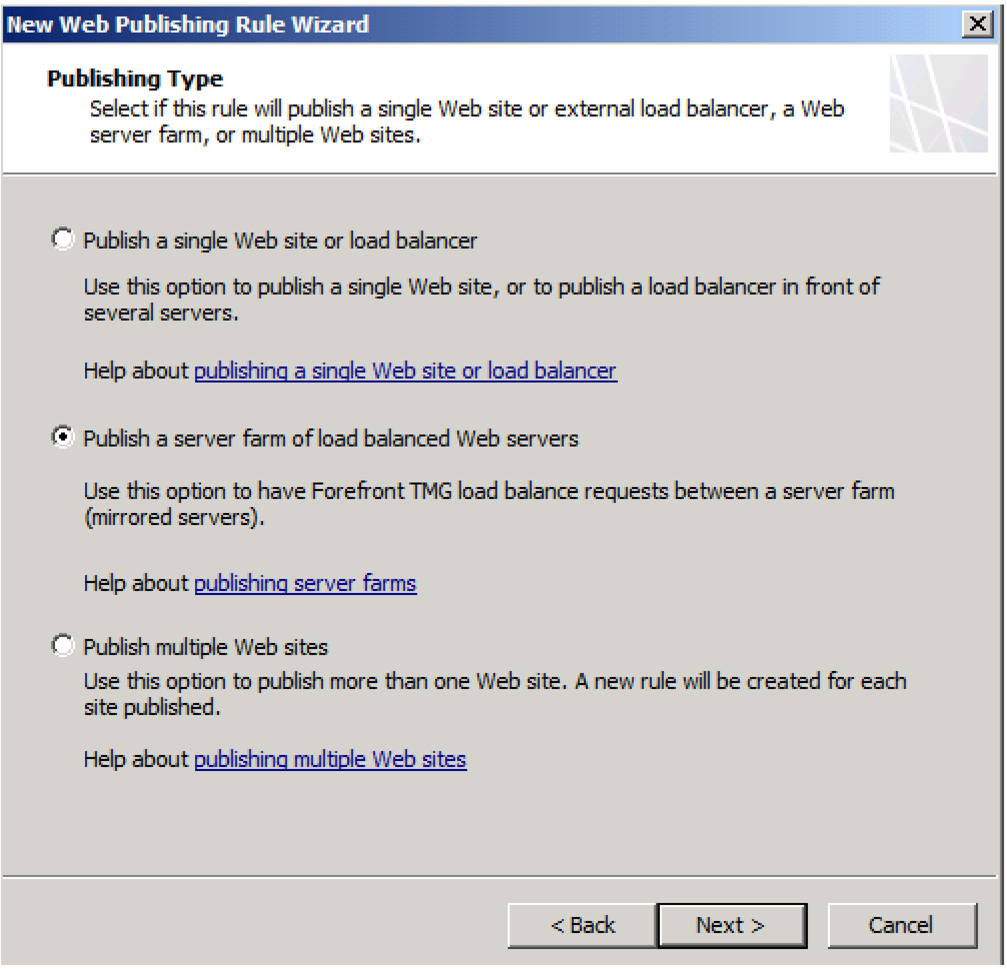

Forefront TMG
File
Action
View
Help
Microsoft Forefront Threat Managemer
Forefront TMG (TMG)
Dashboard
Monitoring
Firewall Policy
Wel
E-M
Intr
Rem
Net
Sys
Log
Upd
Trou
Microsoft
Forefront
Threat Management Gateway 2010
ll Firewall Policy
earch...
Examples
der
Name
Action
Refresh
Export Firewall Policy...
Import Firewall Policy...
New
All Tasks
View
Help
Exchange Web Client Access Publishing Rule...
Mail Server Publishing Rule...
SharePoint Site Publishing Rule...
Web Site Publishing Rule...
Non-Web Server Protocol Publishing Rule...
Access Rule...

Figure 1: Start the Web publishing wizard

Name the new policy rule and Allow the traffic.

Click publish a server farm of load balanced Web servers



Publishing Type

Select if this rule will publish a single Web site or external load balancer, a Web server farm, or multiple Web sites.

○ Publish a single Web site or load balancer

Use this option to publish a single Web site, or to publish a load balancer in front of several servers.

Help about publishing a single Web site or load balancer

● Publish a server farm of load balanced Web servers

Use this option to have Forefront TMG load balance requests between a server farm (mirrored servers).

Help about publishing server farms

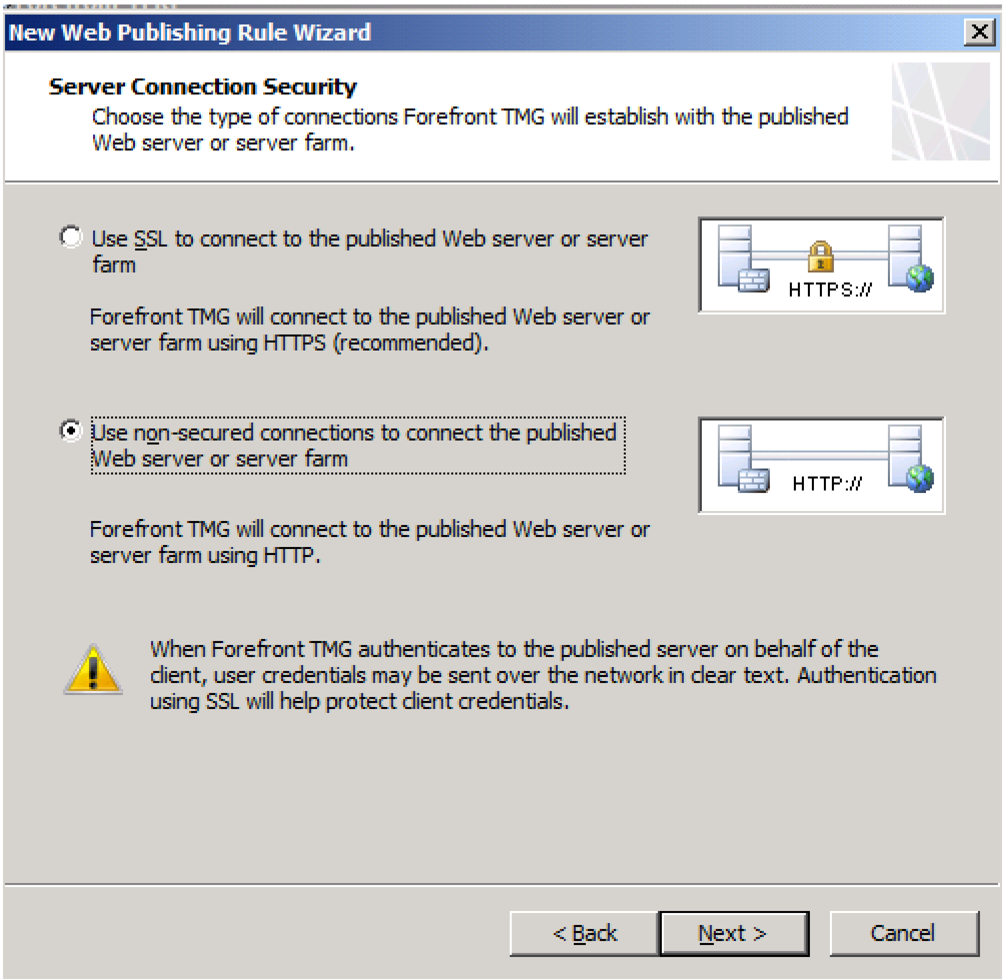
○ Publish multiple Web sites

Use this option to publish more than one Web site. A new rule will be created for each site published.

Help about publishing multiple Web sites

Figure 2: Publish a server farm …

Because we are publishing an internal Web server without HTTPS, specify the appropriate option.



New Web Publishing Rule Wizard

Server Connection Security

Choose the type of connections Forefront TMG will establish with the published Web server or server farm.

○ Use SSL to connect to the published Web server or server farm

Forefront TMG will connect to the published Web server or server farm using HTTPS (recommended).

● Use non-secured connections to connect the published Web server or server farm

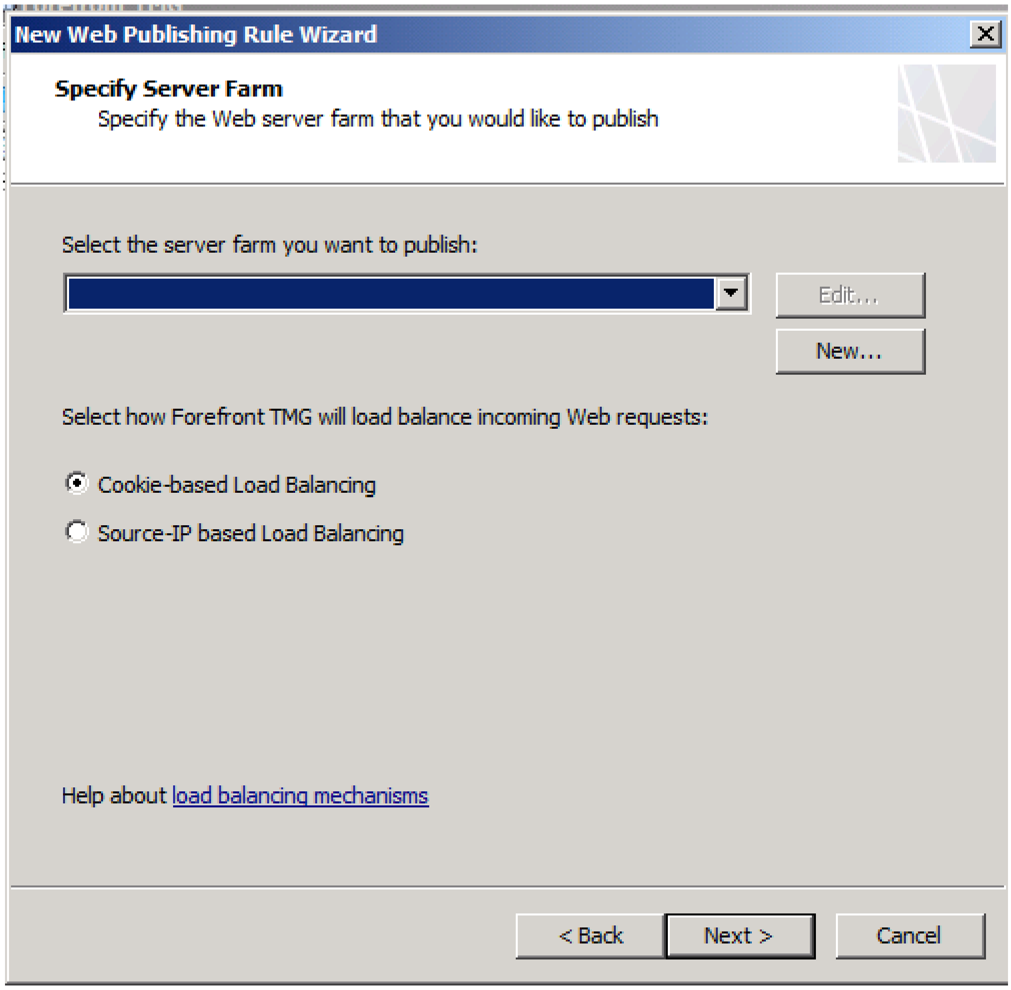
Forefront TMG will connect to the published Web server or server farm using HTTP.

When Forefront TMG authenticates to the published server on behalf of the client, user credentials may be sent over the network in clear text. Authentication using SSL will help protect client credentials.

Figure 3: Use only HTTP

Enter the internal Site name and specify a path if you want to publish the web server only to a specific path.

As the next step create a new Farm, enter the name of the farm and add the internal web servers to the Web Server farm, as you can see in the following screenshot and specify how Forefront TMG should load balance incoming web requests.



New Web Publishing Rule Wizard

Specify Server Farm

Specify the Web server farm that you would like to publish

Select the server farm you want to publish:

Select how Forefront TMG will load balance incoming Web requests:

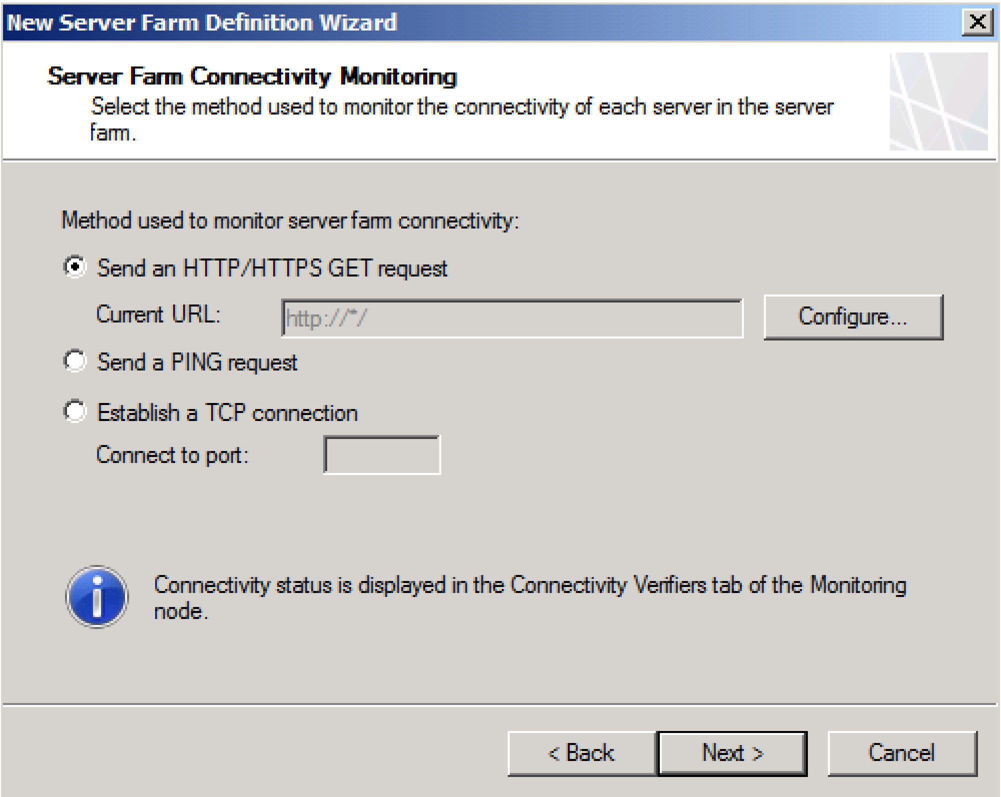
● Cookie-based Load Balancing

○ Source-IP based Load Balancing

Help about load balancing mechanisms

Figure 4: Specify Farm member

Forefront TMG creates a connection verifier to monitor the availability of the farm members. If one server is not reachable an alert will be created. You can customize the alert actions and I will show you later how to do this.


New Server Farm Definition Wizard

Server Farm Connectivity Monitoring

Select the method used to monitor the connectivity of each server in the server farm.

Method used to monitor server farm connectivity:

● Send an HTTP/HTTPS GET request

Current URL: http://*/

○ Send a PING request

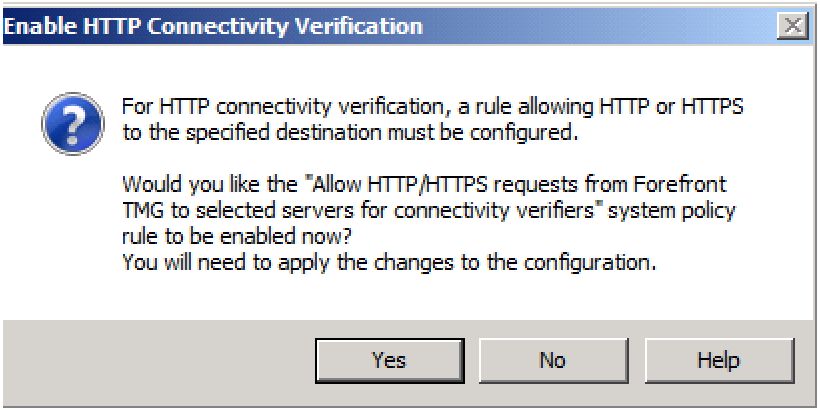
○ Establish a TCP connection

Connect to port:

Connectivity status is displayed in the Connectivity Verifiers tab of the Monitoring node.

Figure 5: Connection verifier

A new popup window appears and will ask you if you want to activate a system policy rule to allow HTTP requests from Forefront TMG to the published web servers. Click *Yes* if you want to do this.


Enable HTTP Connectivity Verification

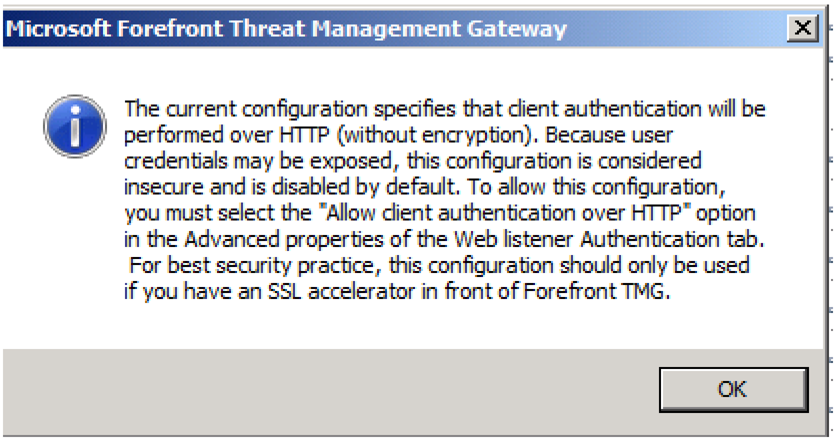
For HTTP connectivity verification, a rule allowing HTTP or HTTPS to the specified destination must be configured.

Would you like the "Allow HTTP/HTTPS requests from Forefront TMG to selected servers for connectivity verifiers" system policy rule to be enabled now?
You will need to apply the changes to the configuration.

Figure 6: System policy rule

The next step is to create the specify Listener which Forefront TMG uses to listen to incoming traffic. Because my article only focuses the Web Server farm load balancing functionality, I will only give you some more information when you publish a web server over HTTP.

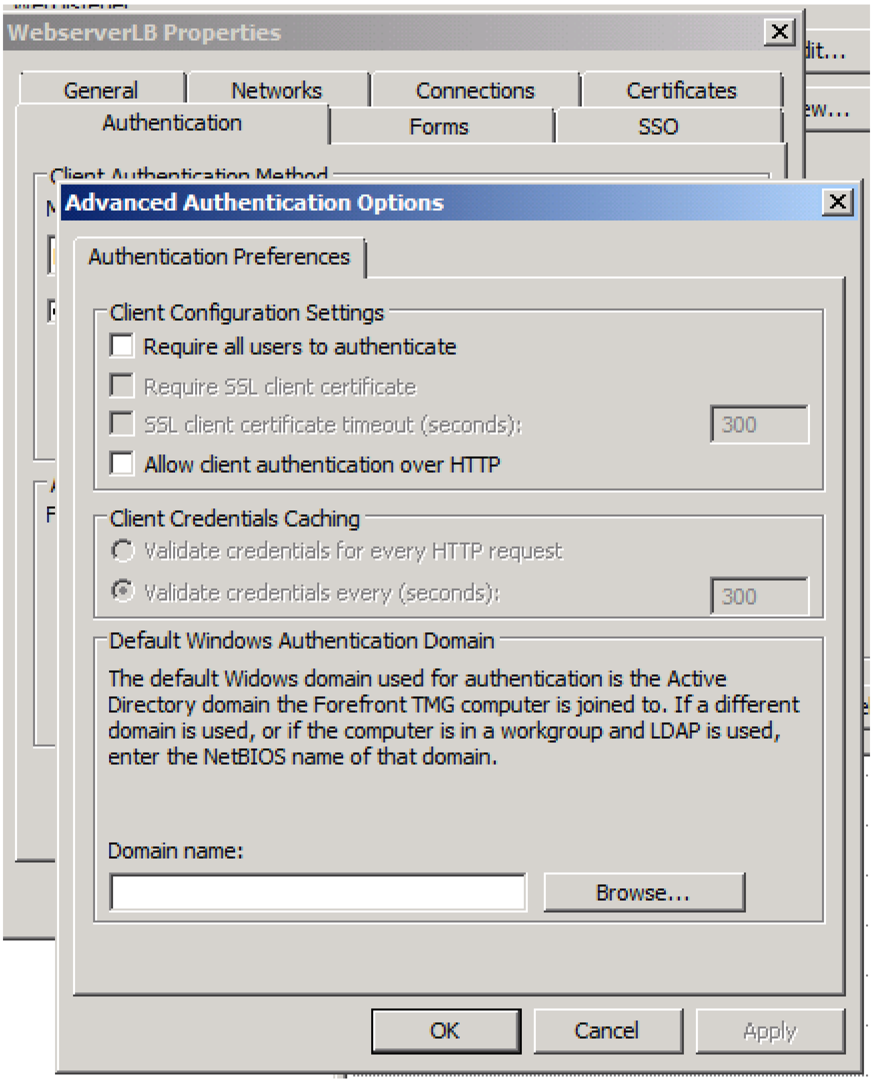
Forefront TMG warns the user that the current configuration may be unsecure when authentication requests are send over HTTP.



The current configuration specifies that client authentication will be performed over HTTP (without encryption). Because user credentials may be exposed, this configuration is considered insecure and is disabled by default. To allow this configuration, you must select the "Allow client authentication over HTTP" option in the Advanced properties of the Web listener Authentication tab. For best security practice, this configuration should only be used if you have an SSL accelerator in front of Forefront TMG.

Figure 7: Allow a system policy rule

To allow client authentication over HTTP you must allow this in the Advanced Authentication options window in the Listener properties as you can see in the following screenshot.



Client Configuration Settings

☐ Require all users to authenticate

☐ Require SSL client certificate

☐ SSL client certificate timeout (seconds): 300

☐ Allow client authentication over HTTP

Client Credentials Caching

○ Validate credentials for every HTTP request

● Validate credentials every (seconds): 300

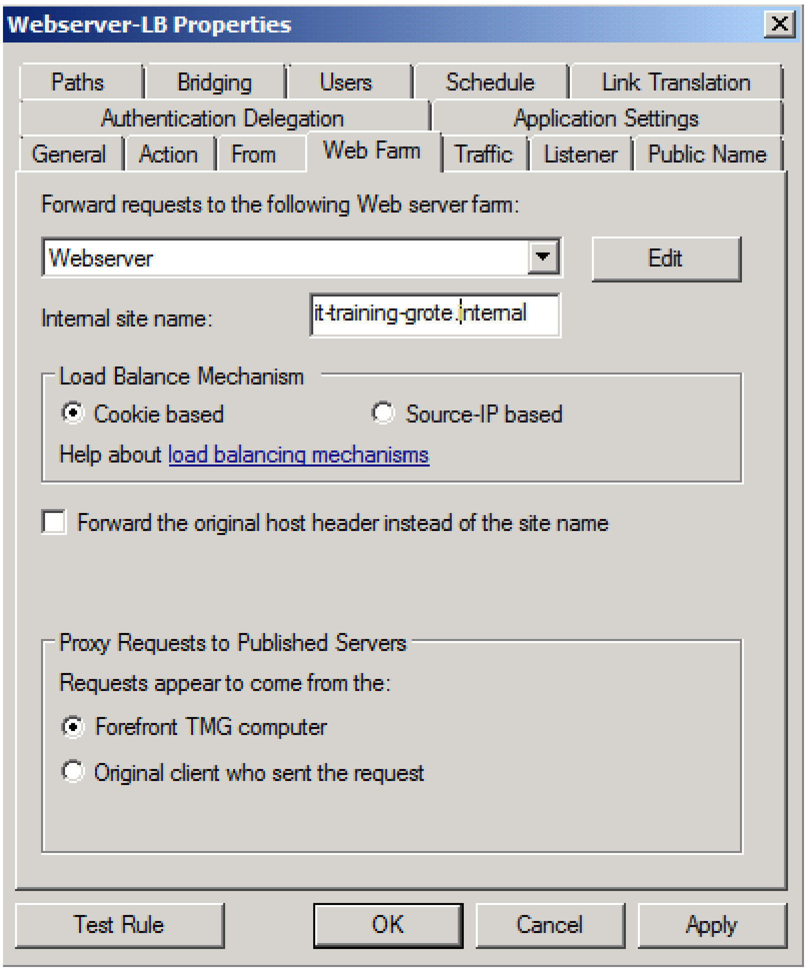
Default Windows Authentication Domain

The default Windows domain used for authentication is the Active Directory domain the Forefront TMG computer is joined to. If a different domain is used, or if the computer is in a workgroup and LDAP is used, enter the NetBIOS name of that domain.

Domain name:

Figure 8: Allow client Authentication over HTTP

After the Webserver publishing rule has been created, navigate to the properties of the rule and click the Web Farm tab to verify the correct configuration.



Forward requests to the following Web server farm:

Webserver

Internal site name: it-training-grote.internal

Load Balance Mechanism

● Cookie based ○ Source-IP based

Help about load balancing mechanisms

□ Forward the original host header instead of the site name

Proxy Requests to Published Servers

Requests appear to come from the:

● Forefront TMG computer

○ Original client who sent the request

Figure 9: Web Farm properties

Monitoring Web Server Farm Status

If you want to know, which web server farm member is available or unavailable, Forefront TMG automatically creates connection verifiers when you create the Web Server farm. A connection verifier detects the status of farm member and reports this event to the alert configuration in TMG Server, which creates notifications like e-mail messages, entries in the event log and more.

Servers in a web server farm can have five different states:

Active

This is the normal state of a web server in the farm and indicates that the server is reachable and able to accept requests.

Out-of-service

This state indicates that the web server didn’t respond to the connection verifier within the specified timeout. No requests are sent to this farm member.

Draining

This state indicates that the web server is in the process of being drained. Existing connections will be finished but new requests will not be send to this server. This feature is useful if you want to place one Server of the Web Server Farm in maintenance mode.

Removed

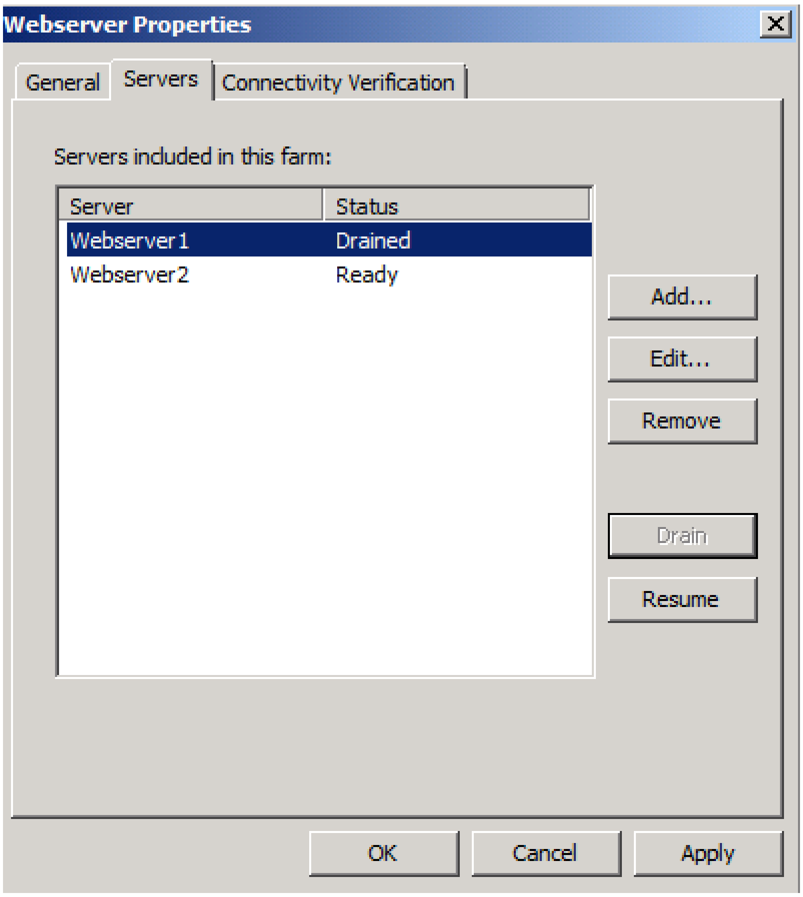
This state indicates that the web server has been removed from the farm, and is not accepting requests.

Unable to verify

This indicates that the server state cannot be verified.

Web Server maintenance

If you want to place one web server manually into maintenance mode navigate to the Servers tab select the server and click the *Drain* button to place the server in maintenance mode so that Forefront TMG knows that this node is not available for load balancing requests. For session based affinity, the server will continue to handle current sessions but will not accept new connections. If you are using IP based affinity a drained server stops receiving requests, but existing connections to that server are still handled.



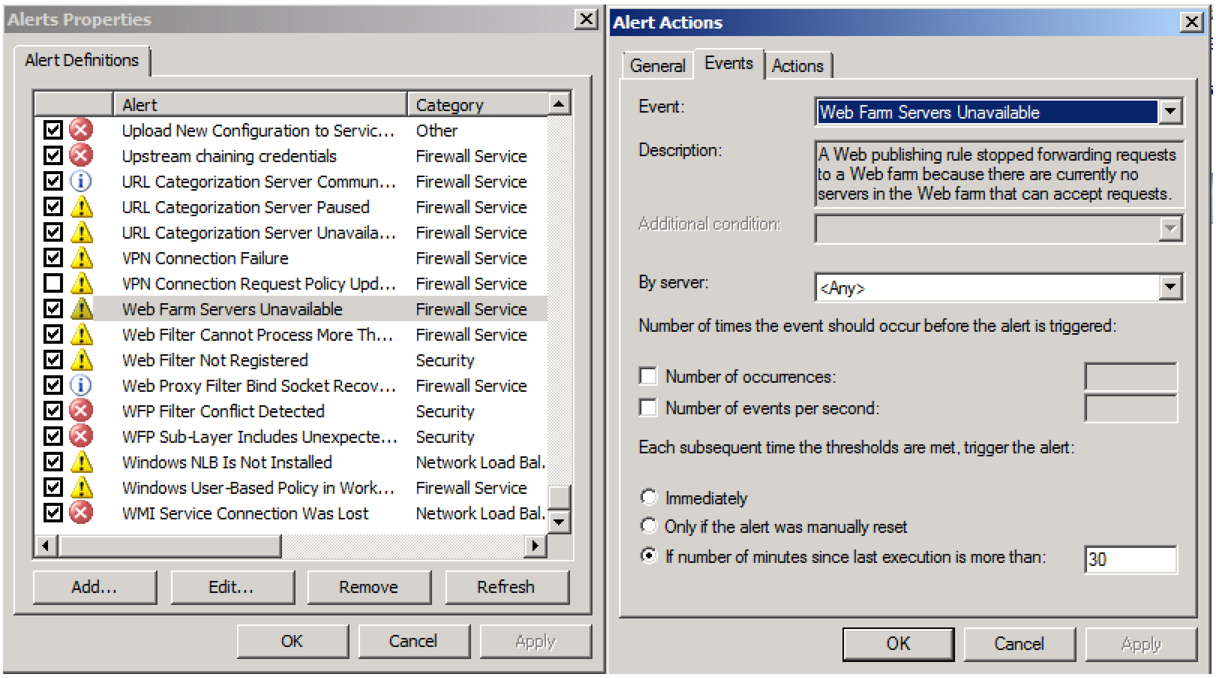
Servers included in this farm:

Server	Status
Webserver1	Drained
Webserver2	Ready

Figure 10: Servers included in the Farm

Alert actions

To configure alert actions when the Web server farm servers are not available, navigate to the monitoring node and in the task pane select Alerts properties and specify the actions you want to perform when a Server in the Web farm is unavailable.



		Alert	Category
☑	⊗	Upload New Configuration to Servic...	Other
☑	⊗	Upstream chaining credentials	Firewall Service
☑	ⓘ	URL Categorization Server Commun...	Firewall Service
☑	⚠	URL Categorization Server Paused	Firewall Service
☑	⚠	URL Categorization Server Unavaila...	Firewall Service
☑	⚠	VPN Connection Failure	Firewall Service
☐	⚠	VPN Connection Request Policy Upd...	Firewall Service
☑	⚠	Web Farm Servers Unavailable	Firewall Service
☑	⚠	Web Filter Cannot Process More Th...	Firewall Service
☑	⚠	Web Filter Not Registered	Security
☑	ⓘ	Web Proxy Filter Bind Socket Recov...	Firewall Service
☑	⊗	WFP Filter Conflict Detected	Security
☑	⊗	WFP Sub-Layer Includes Unexpecte...	Security
☑	⚠	Windows NLB Is Not Installed	Network Load Bal.
☑	⚠	Windows User-Based Policy in Work...	Firewall Service
☑	⊗	WMI Service Connection Was Lost	Network Load Bal.

Event: Web Farm Servers Unavailable

Description: A Web publishing rule stopped forwarding requests to a Web farm because there are currently no servers in the Web farm that can accept requests.

Additional condition:

By server: <Any>

Number of times the event should occur before the alert is triggered:

☐ Number of occurrences:

☐ Number of events per second:

Each subsequent time the thresholds are met, trigger the alert:

○ Immediately

○ Only if the alert was manually reset

● If number of minutes since last execution is more than: 30

Figure 11: Web Farm monitoring and alerting

Conclusion

In this article, I tried to give you an overview how Microsoft Forefront TMG enables web server load balancing to load balance web traffic to several internal web servers without using a classic hardware Load Balancer or a NLB (Network Load Balancing) solution based on Windows Server 2008 R2. In my opinion, the Web Server Load Balancing feature of Forefront TMG is a nice feature for a limited number of Web Servers with basic functionality. A traditional Hardware Load Balancer might have a few more advanced features.

Related links

Explaining ISA Server 2006 Web Server load balancing
http://www.isaserver.org/tutorials/Explaining-ISA-Server-2006-Web-Server-load-balancing.html
Web Farm IP Affinity Load Balancing Algorithm
http://blogs.technet.com/isablog/archive/2009/04/30/web-farm-ip-affinity-load-balancing-algorithm.aspx
Publishing a single Web site or load balancer over HTTP
http://technet.microsoft.com/en-us/library/cc984433.aspx
High availability and scalability design guide for Forefront TMG
http://technet.microsoft.com/en-us/library/dd896997.aspx