Creating a Highly Available RD Gateway Environment with Azure Multi-Factor Authentication

In our last article about RD Gateway and Azure Multi-Factor Authentication, we showed you how to add Azure Multi-Factor Authentication (Azure MFA) to your on premises RD Gateway deployment to further secure the login process. In that walkthrough we used one RD Gateway server and one MFA server. This configuration works well, as long as the servers never go down. We thought it would be a good idea to plan for a situation more in line with reality.

In this article we will show you how to configure load balanced RD Gateway servers with two on premises Azure MFA servers to create a more highly available environment. There are two ways RD Gateway can be load balanced, and we will walk through the configuration for both of these scenarios:

- Scenario 1: You have two load balanced RD Gateway servers, two on-premises MFA servers, and each RD Gateway uses its own local NPS (the typical configuration).
- Scenario 2: You have two load balanced RD Gateway servers, two on-premises MFA servers, and both RD Gateway servers use a central NPS. (Companies might use a centralized NPS if they already have one in place in their network, and want to keep all things related to NPS in this centralized location.)

A few notes about preparation: This article builds on our previous article "Step By Step – Using Windows Server 2012 R2 RD Gateway with Azure Multi-Factor Authentication". It may be helpful to review it first as a reminder of how to setup on premises Azure MFA servers, how to enable RADIUS authentication on the Azure MFA server(s) and how to add users and test the configuration. Also, we assume you know how to load balance RD Gateway and we start this article with two RD Gateway servers already set up in the typical HA configuration. If you do not know how to load balance RD Gateway, read this: http://redmondmag.com/Articles/2013/12/24/RD-Gateway-in-Windows-Server.aspx?Page=1)

Scenario 1: Load balanced RD Gateways Using a Local NPS, Two MFA Servers

When most companies load balance RD Gateway, they configure each server to use a local NPS and configure the RD CAP settings identically on each server. Follow these steps to add two on-premises Azure MFA servers to your load balanced RD Gateway configuration to create a more highly available solution that includes two factor authentication.

Injecting Azure MFA into the Authentication Sequence

As we explained in our previous article, when you have one RD Gateway server with a locally running NPS service (the default configuration), you have to have some way to get the MFA server into the communication sequence. You do this by tricking RD Gateway to use a centralized NPS server but you point it to the MFA server. The same is true for an HA scenario – you will set up most things twice (once for each RD Gateway and/or MFA server). The communication works as shown in Figure 1:

1. One of the load balanced RD Gateway servers gets the initial user login request.
2. RD Gateway forwards the RADIUS request through NPS to the first MFA server in its list (they are listed as Central NPSs in the RD Gateway Manager).
3. MFA server forwards it right back to NPS on the RD Gateway server
4. RD Gateway validates the user credentials and does the RD CAP check using its local NPS.
5. The local NPS then sends an ACCEPT or REJECT to MFA server.
6. On ACCEPT, MFA will perform the two factor authentication sequence with the user (via phone call, text or mobile app). If the user returns the correct letter / number sequence, it sends an ACCEPT to NPS on RD Gateway.
7. Finally RD Gateway will check the RD RAP and either allow or deny the connection.

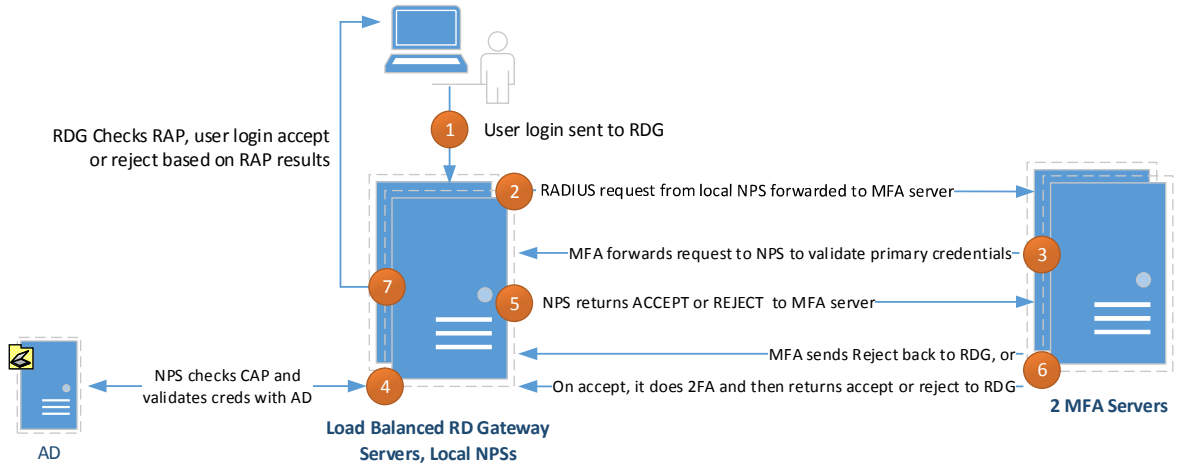

Figure 1: Trick RD Gateway into thinking it is using a centralized NPS.

Configure the RD Gateway Servers

On each RD Gateway server configure RD Gateway to use a Central RD CAP store, but point it to both MFA servers:

1. Open RD Gateway Manager, right click the server name, and select Properties.
2. Select the RD CAP Store tab (shown in Figure 2).
3. Select the Central server running NPS option.
4. Enter the name or IP address of the master MFA server and click Add.
5. Enter a shared secret in the corresponding popup box and click OK.
6. Do this again for the slave MFA server.

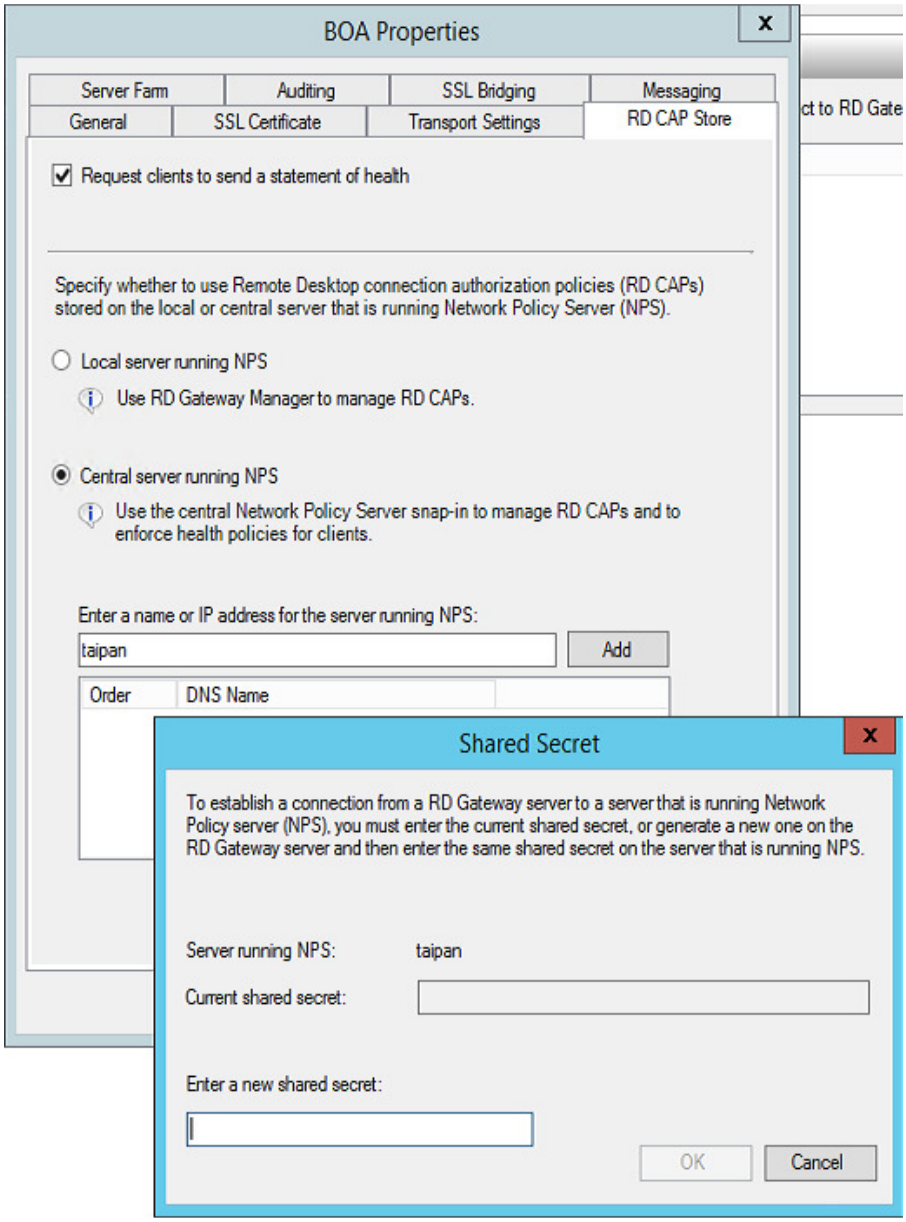

Figure 2: Configuring both RD Gateway servers to use a central NPS

Make NPS and MFA Talk To Each Other

To enable communication between the local NPS on each RD Gateway server and both MFA servers, configure NPS on each RD Gateway server and the MFA server software on each MFA server as shown in Figure 3:

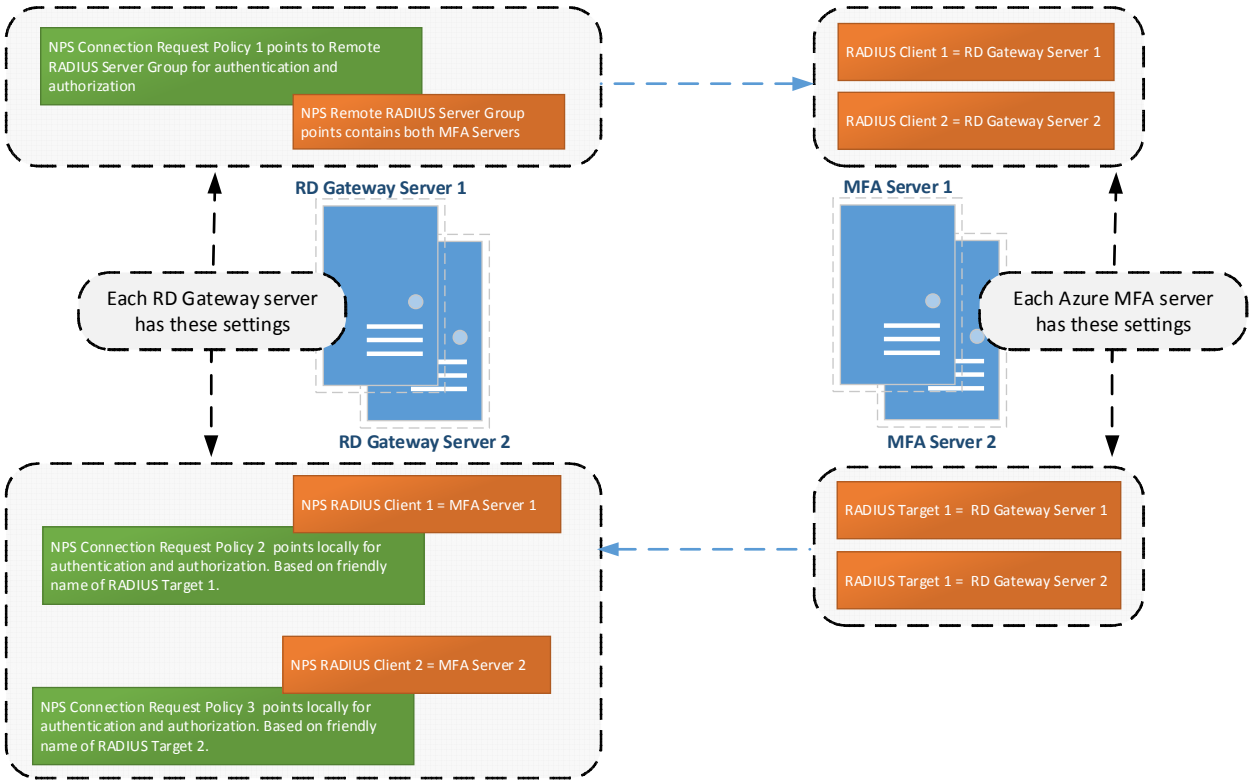

Figure 3: NPS and MFA server use a Remote RADIUS Server Group, RADIUS Targets and RADIUS Clients to communicate with each other.

We will go through the steps in detail, but as an overview, these are the components you create in the RD Gateway server's local NPS, and in MFA Server:

In NPS on each RD Gateway server, configure three Connection Request Policies and a corresponding Remote RADIUS group or RADIUS client:

- The first Connection Request Policy will send communication to either MFA Server via a Remote RADIUS Server Group
- The second and third Connection Requests Policies will receive communication from MFA servers via RADIUS Clients

On each MFA server configure:

- Two RADIUS Clients that will receive communication from each RD Gateway server
- Two RADIUS Targets that will facilitate communication to NPS on each RD Gateway server

Configure NPS on Each RD Gateway Server

First, you need to prevent NPS from timing out before MFA's authentication process completes. Follow these steps (shown in Figure 4):

On each RD Gateway server:

1. In NPS, expand the RADIUS Clients and Servers menu and select Remote RADIUS Server Groups.

2. When you setup RD Gateway with a central NPS, it creates an entry here named “TS GATEWAY SERVER GROUP”. Right click this group and select Properties.
3. Both MFA servers should be listed here because you already added them as a Central NPS in RD Gateway. For each MFA server listed, highlight the server, click Edit and:
    a. Select the Load Balancing tab.
    b. Change the “Number of seconds without response before request is considered dropped” and the “Number of seconds between requests when server is identified as unavailable” to 30-60 seconds.

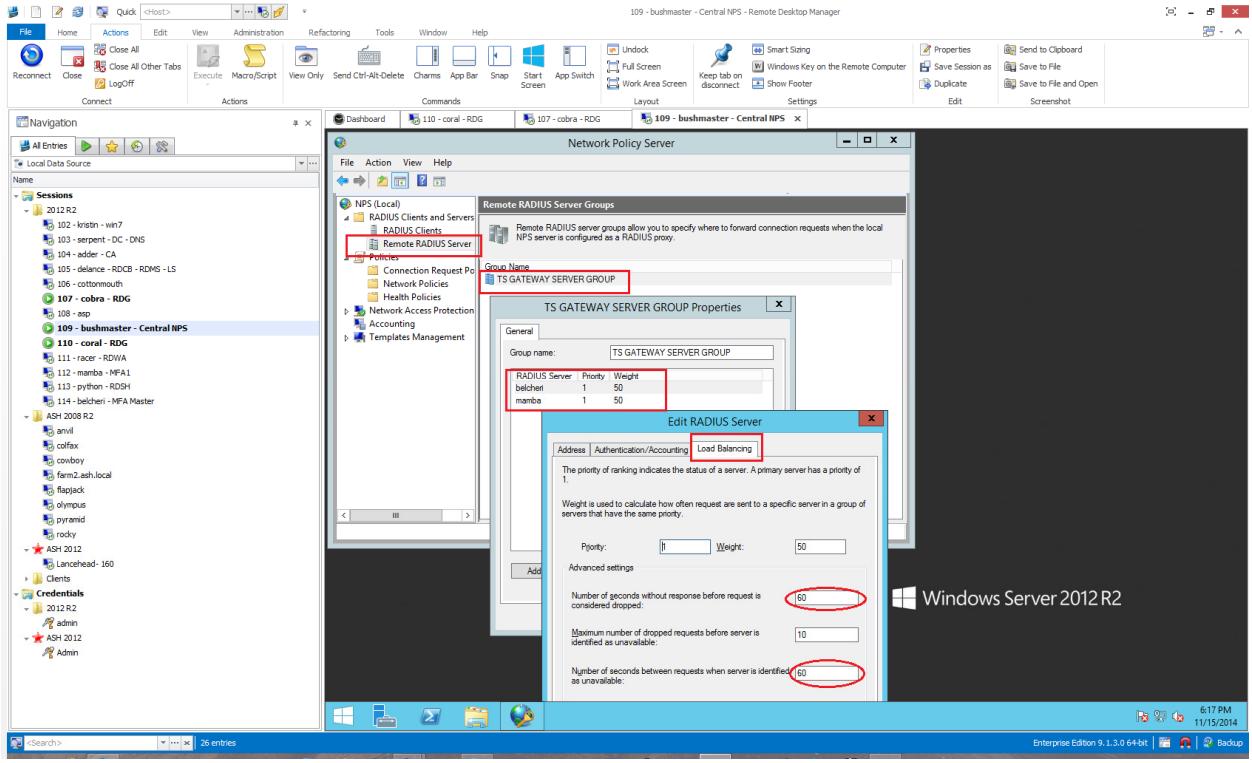

Figure 4: On each RD Gateway server in NPS adjust the Remote RADIUS Server group entries’ load balancing settings.

Next you need to configure the local NPS on each RD Gateway server to receive RADIUS authentications from both MFA servers. Create two RADIUS clients on each RD Gateway server. Follow these steps (twice, once for each MFA server) in NPS on each RD Gateway server:

1. In the left column, right click RADIUS Clients and choose New.
2. Add a Friendly Name and the address of the Azure MFA server as shown in Figure 5.
3. Add a shared secret and click OK.

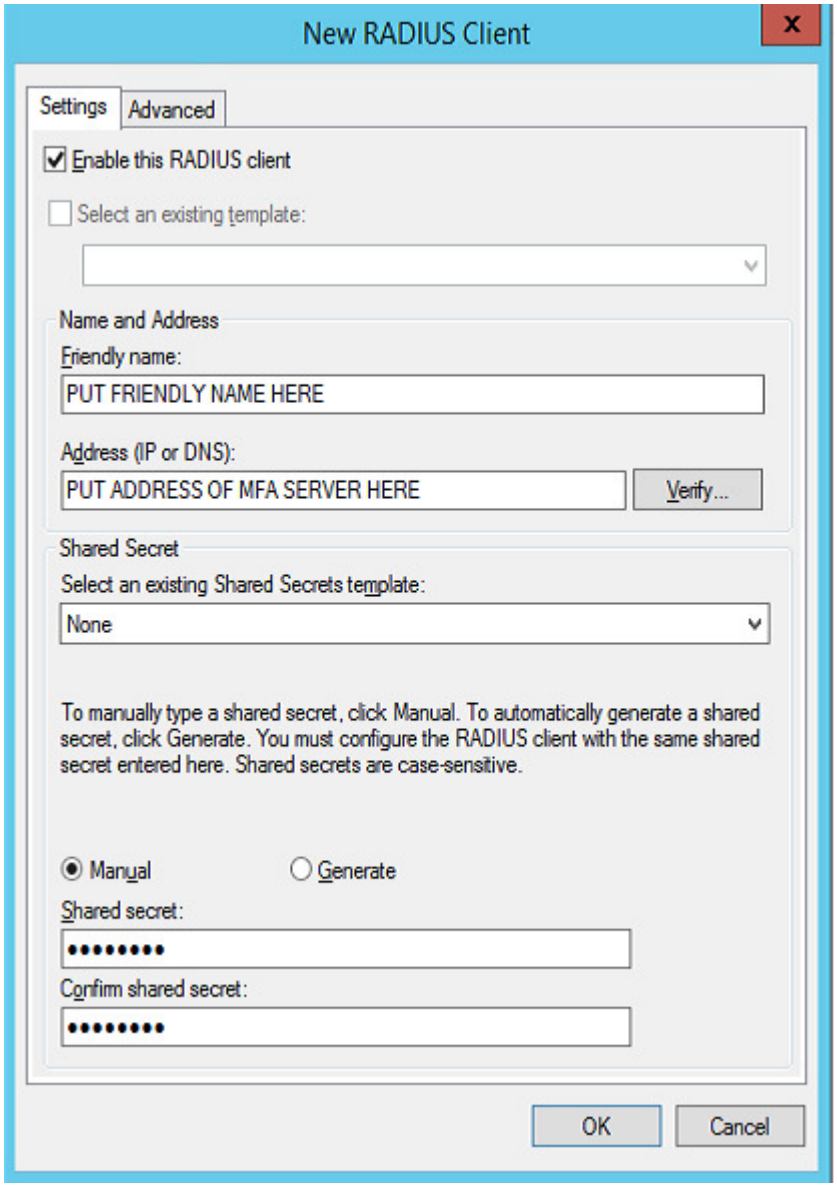

Figure 5: Create two RADIUS clients in NPS on each RD Gateway server.

When you are finished your NPS on each RD Gateway server should show two RADIUS clients, with two different Friendly Names as shown in Figure 6:

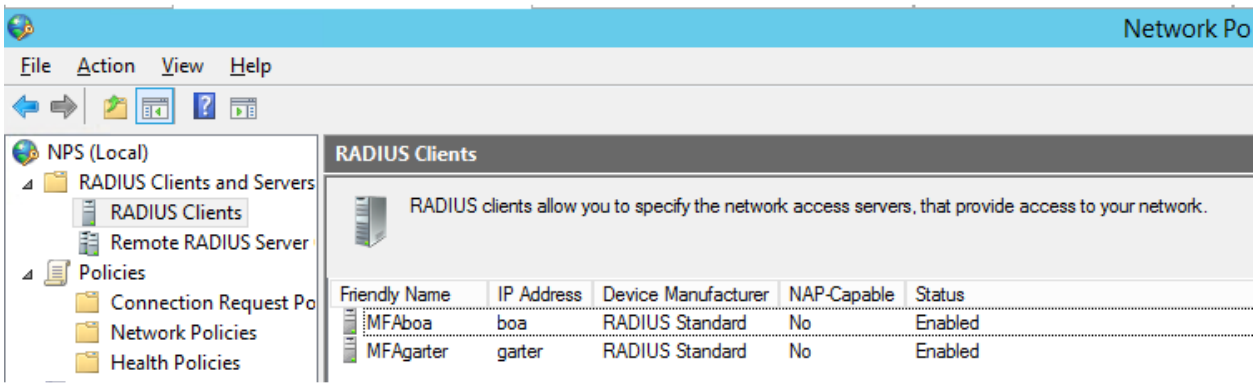

Figure 6: Configure two RADIUS clients in NPS on each RD Gateway server.

Next, on each RD Gateway server you configure three Connection Request Policies in NPS – the first one will forward requests to the Remote RADIUS Server Group (which contains both Azure MFA servers), and the next two will each receive requests coming from the MFA servers.

The easiest way to do this is to use the existing policy created when you made an RD CAP in RD Gateway. Follow these steps on each RD Gateway server:

1. In NPS expand the Policies section in the left side of the screen and then select Connection Request Policies. You should see a policy already created there, called TS GATEWAY AUTHORIZATION POLICY.
2. Right click this policy and select Duplicate Policy.
3. Double click the new duplicate policy and select the Conditions tab.
4. Add a Client Friendly Name as shown in Figure 7. Use the same name you set for the first RADIUS client (pointing to MFA Server 1) you created earlier.

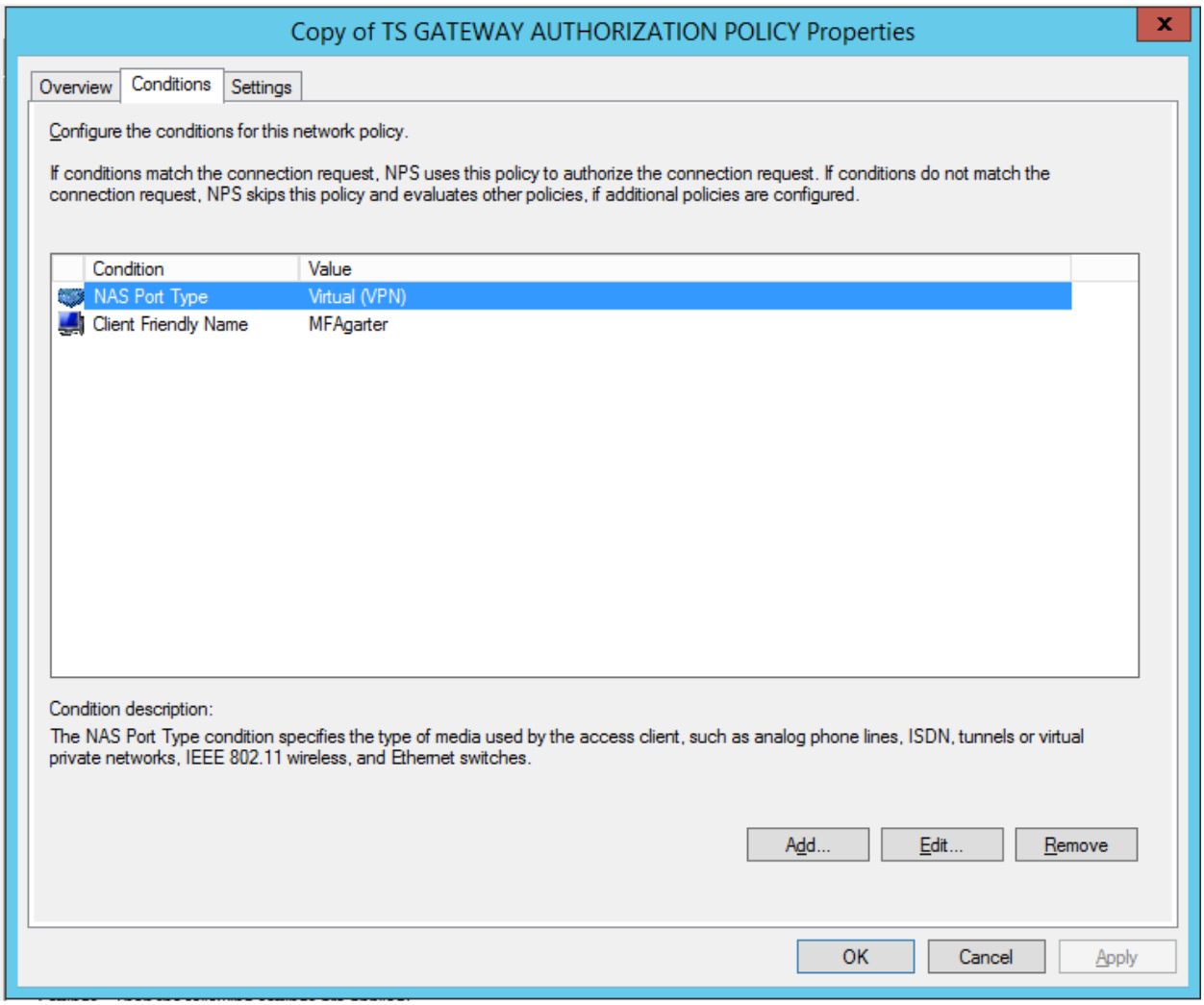

Figure 7: Add a Client Friendly Name to each copy of TS GATEWAY AUTHORIZATION POLICY.

5. Now select the Settings tab and change the Authentication Provider to “Authenticate requests on this server” as shown in Figure 8.

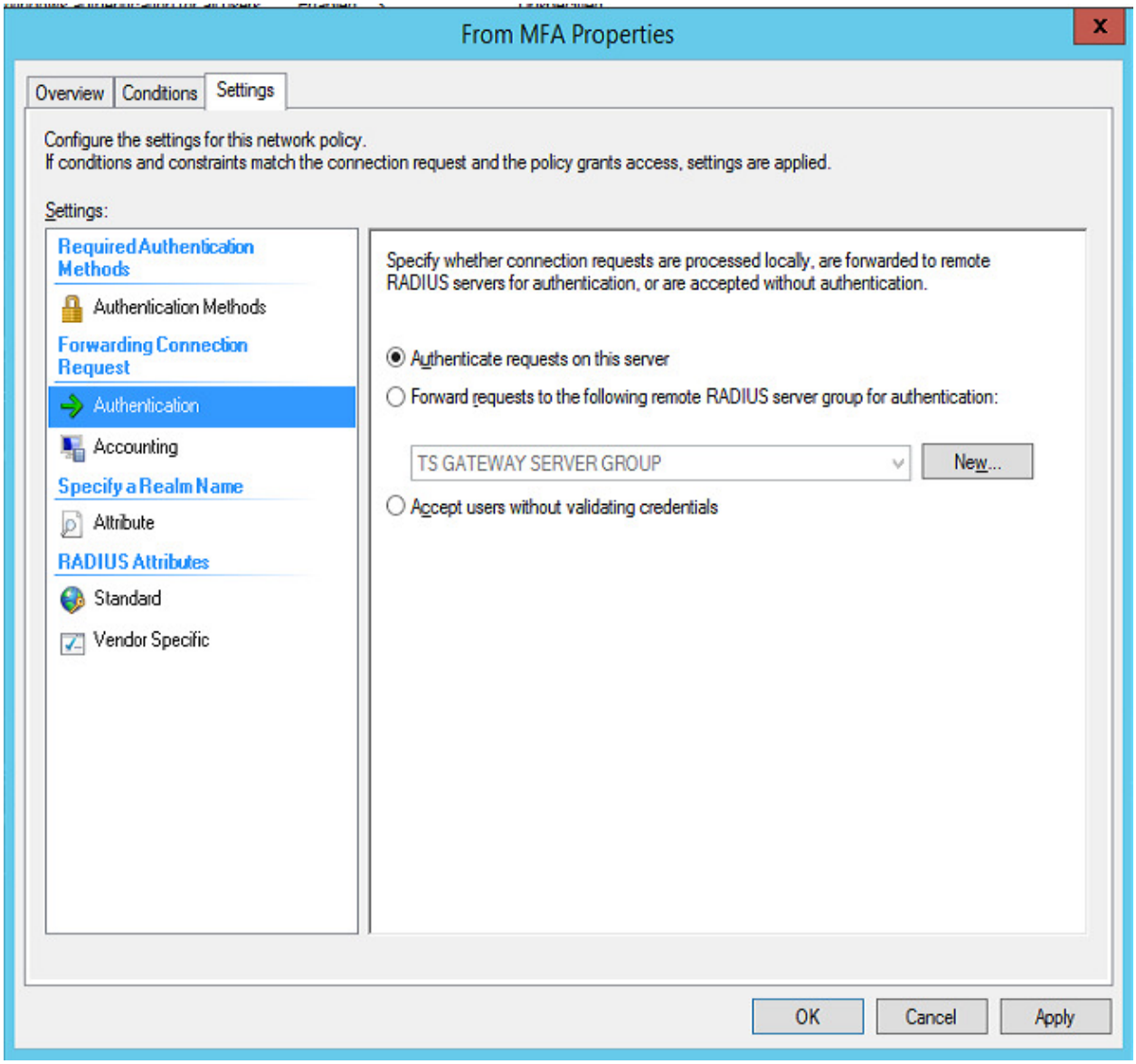

Figure 8: Change each copied policy to authenticate requests locally.

Select Accounting and make sure the “Forward accounting requests…” check box is not checked. Then click OK.

6. Repeat steps 1-5 and set the Client Friendly Name to match the second RADIUS client you created (that points to the second MFA server).
7. IMPORTANT! Make sure that this policy (the copy of the original) as well as the second copy (the second copy of the original) are ordered first, ahead of the original policy.
8. Right click both newly created Connection Request Policies and select Enable.

When you are done, your policy settings should show up on the main interface as shown in Figure 9.

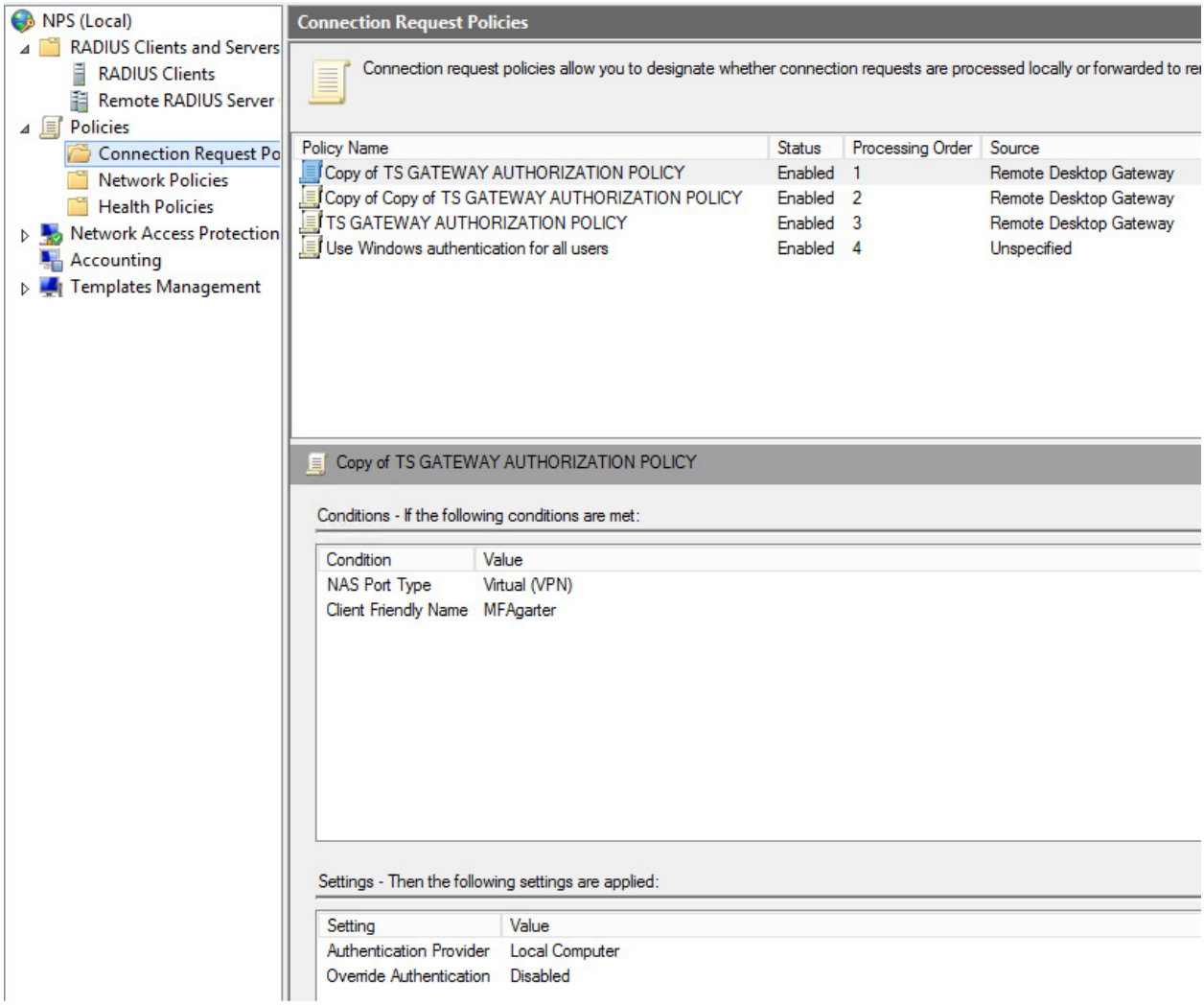

Figure 9: Overview of the NPS policies on each RD Gateway server.

Tip: you should not have to edit the original TS GATEWAY AUTHORIZATION POLICY. Just make sure it is listed below the two copied policies. Also, it may be beneficial to you to rename your policies to something more descriptive. It will not harm the configuration to do so.

Configure the MFA Servers

Now you need to configure the master MFA Server (the first server) with RADIUS targets and clients (one target and one client for each RD Gateway server). These settings are automatically synced to the slave (second) MFA Server.

1. Open the Multi-Factor Authentication Server and click on Status to confirm you are working on the Master MFA Server.
2. Click the RADIUS Authentication icon.
3. Check the Enable RADIUS authentication checkbox.
4. On the Clients tab, click the Add… button.
5. Add an RD Gateway server IP address, and a shared secret. The shared secret needs to match the corresponding Remote RADIUS Server Group entry in NPS. Create two clients, one for each RD Gateway server. To be able to distinguish the clients, provide a unique Application name.
6. Click the Target tab and choose the RADIUS server(s) radio button.

7. Click Add and enter the IP address, shared secret and ports of an RD Gateway server. The shared secret must match the one configured for the RADIUS client created in NPS on the RD Gateway server. Create two targets – one for each RD Gateway server.

Tip: We use the same shared secret for each piece of this configuration that requires one.

Scenario 2: Companies that use a Centralized NPS

For the 3% of companies that use a centralized NPS with RD Gateway, the setup is slightly different. You still point RD Gateway servers to the MFA servers, but instead of pointing the MFA server back to the local NPS on RD Gateway (like Scenario 1), the RADIUS Target on the MFA servers now points to the Central NPS. Some of the steps are repeats of the previous configuration. Follow the diagram shown in Figure 10 and create the components listed on each server.

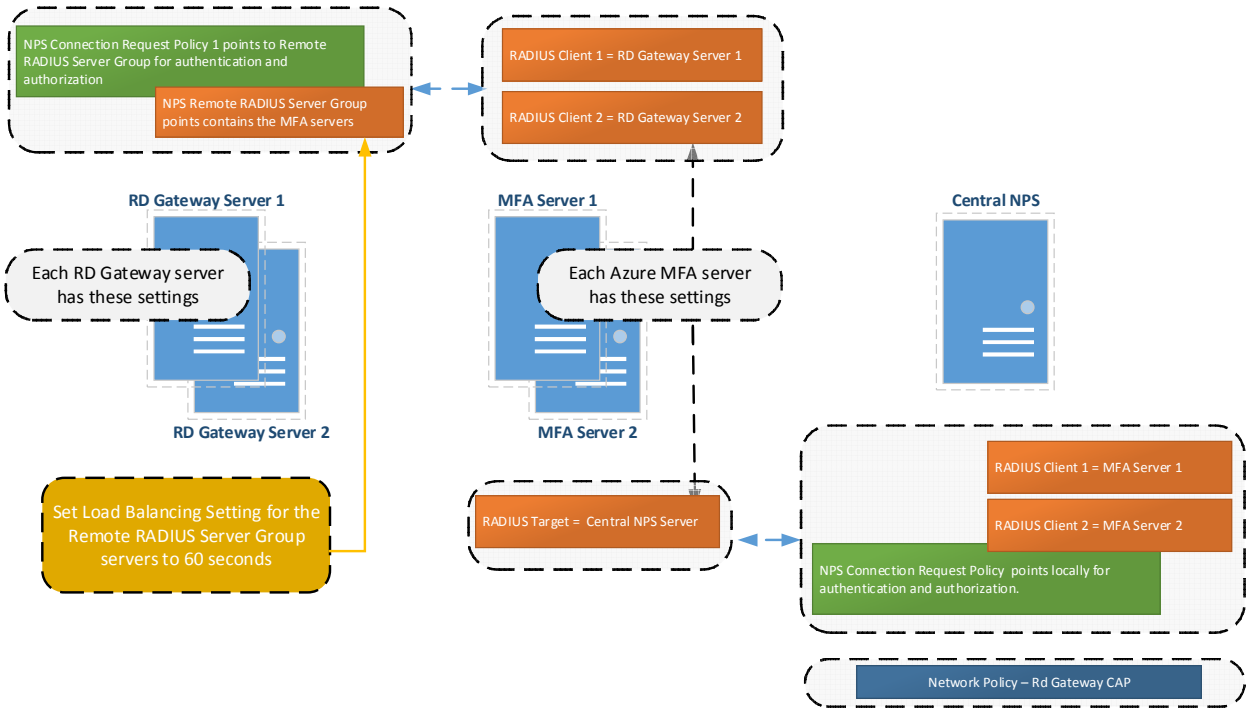

Figure 10: Follow this diagram to setup Azure Multi-Factor Authentication with load balanced RD Gateway servers that use a centralized NPS server.

In NPS on each RD Gateway server you configure:

- 1 NPS Connection Request Policy that points to Remote RADIUS Server Group for authentication and authorization
- 1 NPS Remote RADIUS Server Group containing the MFA servers

In the MFA server UI you configure:

- 2 RADIUS Clients to receive communication from RD Gateway servers
- 1 RADIUS Target to send communication to the Central NPS server

On the Central NPS you configure:

- 2 RADIUS Clients representing the two MFA servers
- 1 Connection Request Policy that is configured to authenticate requests on the server (locally).

- 1 Network Policy (the RD CAP)

Extending Timeout Values

Once you have created the components shown in Figure 10 above, you need to tweak this environment a bit more.

Even when you use a central NPS to house the RD CAPs, the local NPS on each RD Gateway server is still responsible for forwarding requests to the MFA server. So (just like in Scenario 1) you need to prevent the local NPS on each RD Gateway from timing out before the MFA server authentication has completed.

1. In NPS on each RD Gateway server, expand the RADIUS Clients and Servers menu and select Remote RADIUS Server Groups.
2. Right click the "TS GATEWAY SERVER GROUP" and select Properties.
3. Highlight the first MFA server, click Edit and:
    a. Select the Load Balancing tab.
    b. Change the "Number of seconds without response before request is considered dropped" and the "Number of seconds between requests when server is identified as unavailable" to 30-60 seconds. The click OK.
4. Complete step 3 again for the second MFA server.

MFA Quirks

It's best that you understand as fully as possible the software that you implement (and this is especially true for shortcomings). Therefore, here are some "quirks" about Azure MFA:

No Updates without the Master

Any actions that perform a write to the on-premises MFA Server data file are always routed to the master MFA Server which then replicates changes to the slaves. A slave can process authentication whether the master is up or not, but it cannot receive any configuration changes if the master goes offline. So there is the potential for data loss if you made configuration changes on an Azure MFA master server and it goes down before those changes replicate.

No Interface on Slaves If the Master is Down

Also interesting is that the MFA interface will not open on slaves if the master goes down. In order for the user interface (UI) to open on an MFA slave server, you have to elevate a slave to take the place of the downed master or bring the original master online.

> *"When you open a UI on a slave server, it actually connects to the master because all writes have to be written to the master except for a few server-specific settings such as LDAP bind passwords, hooking plug-ins into IIS and log file settings"… "Any changes to the data file are then replicated from the master to any slaves. The slaves all have their own copy of the data file so that they can read from it and process authentications, and so that they can be promoted to be the master if needed."*
>
> *– Shawn Bishop, MSFT*

No Load Balancing for MFA

There is no built-in load balancing mechanism for Azure MFA. So having two Azure MFA servers up and running just means that the first server (the master) will do all the work. The slave will only do work when the master is offline. Once the master comes back online it resumes doing all the work.

That being said, you can place a third party load balancer between the RD Gateway servers and the MFA servers to load balance the RADIUS traffic.

Azure MFA + RD Gateway = Failover (Sort Of)

The on-premises Azure MFA Servers don't do any kind of failover on its own. But when used with RD Gateway, we can rely on RD Gateway to determine which Azure MFA server is usable. This can suffice for failover, but there is a catch: If NPS sends a RADIUS request to an MFA Server that is down, it won't time out for 60 seconds due to timeouts configured for the NPS Remote RADIUS Server Group entries. The higher timeout is required so that we have time to perform 2FA and get a response back to NPS before the request is considered dropped and the server is considered unavailable. However, that also means that NPS won't be able to determine that an MFA Server to which it sent a request is down until that timeout occurs. It's a catch 22 situation. The first user logon request that occurs after an MFA server goes down will always fail. Once the first user fails however, the next login request gets passed to the next MFA server in the list, and the time it takes for 2FA to take place should return to normal.

Questions & Answers

Should you have further questions or run into trouble, here are a few "questions/answers" that may help you further understand your configuration options or help you resolve problems with your implementation.

Q Can I install the on-premises Azure MFA server software on the actual RD Gateway server to cut down on the quantity of servers I need for this configuration?

A No. This is not a supported configuration, and we have not gotten it to work in our testing. Your first two NPS connection request policies state that traffic containing a certain "friendly" name gets handled locally. Those connection requests would be dealing with data from the MFA servers. Then the third connection request policy actually forwards requests to the TS Gateway Server group, which would then send the radius request on to MFA server. If you house MFA together on the same server as RD Gateway, your request just gets sent back to itself on the standard ports. So RD Gateway / NPS deals with the request and MFA is not contacted.

It is possible to configure the MFA Server to bind to other ports and to configure RDG/NPS to forward the RADIUS requests to the non-standard ports. So theoretically it seems possible to make it work with MFA, NPS and RDG on the same server (e.g. RDG/NPS→MFA on non-standard port→NPS on port 1812 or 1645). However, when people have tried this approach, server event log entries state that NPS can't forward RADIUS to the local machine. Keep things simple and install MFA on a separate server.

Q On which MFA server (the master or the slave) should I perform configuration changes?

A You can use either the master or the slave Azure MFA server to perform edits. Either way, the master will always receive the changes. If you make changes on the slave, the slave UI actually communicates to the master MFA Server to write the updates there. Then the updates are synced back out to the slaves.

Q What port(s) do MFA master and slave use to communicate with each other, and can I set this to a non-standard port?

A Azure MFA slaves communicate with the master over RPC. To set the port on which slaves communicate to the master follow these steps:

1. On the master Azure MFA server open the Registry Editor and navigate to:

   HKEY_LOCAL_MACHINE\SOFTWARE\Wow6432Node\Positive Networks\PhoneFactor

2. Create a new DWORD registry key called "Pfsvc_ncan_ip_tcp_port".
3. Set the value to the port number on which you want the MultiFactorAuth service to run.
4. Restart the server.

Q My Azure MFA servers are domain-joined. My Azure MFA slave server is having trouble communicating with the master. The Azure portal shows the slave as connected but the master does not, and the UI on the slave will not open. Instead, I get an error message (shown in Figure 11) saying "Unable to communicate with the master Multi-Factor Authentication service…"

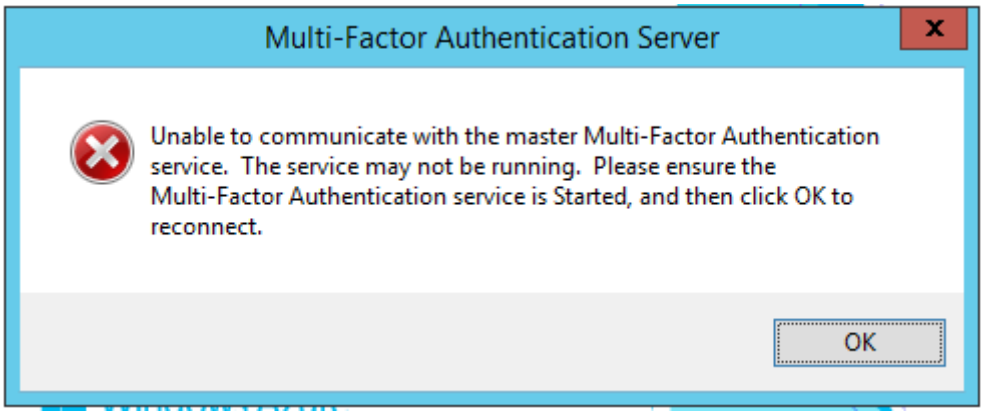

Figure 11: You may get this error message if you're the slave MFA server computer account is not added to the proper Active Directory security group.

How can I remedy this?

A All Azure MFA servers that replicate need to either be joined to the domain or must all be stand-alone. And you must run the Multi-Server Configuration Wizard on each server. Domain-joined MFA servers get added to the PhoneFactor Admins security group in Active Directory.

> "If the servers are joined to the domain, the wizard creates a security group in AD called "PhoneFactor Admins" and puts that machine into the security group. That gives the machines the ability to trust the RPC communications sent between them. If the servers are stand-alone, running the wizard will create client certificates that are used to authenticate and communicate via S-Channel".
>
> - Shawn Bishop, MSFT

In our testing, we have seen this type of error if the Multi-Server Configuration Wizard is run on each server, but for some reason a server does not get added to the PhoneFactor Admins security group in Active Directory. Make sure all MFA servers are listed in this security group, and then restart each MFA server that you had to add manually to this group.

In the Azure MFA UI, it is also important to click on the RADIUS Authentication-->Multi-Factor Auth Servers tab and check the box next to each server. That ensures the RADIUS service is started on each server.

Summary

Our previous article explained the basics of how RD Gateway, NPS and MFA work together to authenticate users. In this article, we're expanding on that to combine load balanced RD Gateway servers with two MFA servers to create a redundant solution that is more suited for live environments that can't afford much downtime. We have also covered some of the quirks you will run into when combining MFA with RD Gateway. Finally we provided some special configuration and troubleshooting tips in this article's Question and Answer section. Got more questions? Ping me at virtualkristin (at) outlook (dot) com.