



High Availability using Network Load **Balancers.**

Load Balancing with F5 Big IP - Destination NAT (In-line)

 Nuance AutoStore and  Ricoh PCCv5.1

Contents

1.	Version Management	3
2.	Document Summary	3
3.	Disclaimer	3
4.	General Information in F5 Big IP ADCs	3
5.	Testing Environment	3
6.	Backend Servers	4
7.	DNS Host “A” Record	5
8.	F5 Big IP LTM’s Configuration	6
	8.1 (Health) Monitors	6
	8.2 Nodes (Servers)	7
	8.3 Pools	8
	8.4 Persistence Profile	9
	8.5 Virtual Servers	10
9.	Testing Virtual Servers	12
	9.1 Web Services	12
	9.2 Network Traces	13
10.	AutoStore Installation and Configuration	14
	10.1 Caveats	15
	10.2 Installing AutoStore	15
	10.3 Configuring AutoStore	16
	10.4 Configuring the Ricoh SOP device	16
	10.5 Testing	17

1. Version Management

Date	Version	Author	Comments
2018-12-13	v1.0	Javier Gonzalez	Initial Release

2. Document Summary

This document describes the steps to configure the F5 Big IP LTM as In-line appliance to load balance AutoStore from Nuance Document Imaging, providing scalable and highly available services.

3. Disclaimer

Although the greatest care has been taken in the preparation and compilation of this document, no liability or responsibility of any kind (to extent permitted by law), including responsibility for negligence is accepted by the Nuance, its servants or agents. All information gathered is believed correct at the time of publishing.

© Nuance Communications, Inc. All rights reserved.

4. General Information in F5 Big IP ADCs

The F5 Big IP is an Application Delivery Controller (ADC). According to the [Wikipedia](#), an ADC is a network device in a datacenter, often part of an application delivery network (ADN), which helps perform common tasks, such as those done by web sites to remove load from the backend servers themselves. Many also provide load balancing.

A common misconception is that an ADC is an advanced load-balancer. This is not an adequate description. In fact, an ADC includes many OSI layers 3-7 services including ALSO load-balancing. Other features commonly found in most ADCs include SSL offload, Web Application Firewall, NAT64, DNS64, and proxy/ reverse proxy to name a few.

Note that ADC & ADN are marketing terms invented by F5 Networks and other vendors to imply that business applications require front-end intelligence to supplement and enhance application flows from client to server back to client.

5. Testing Environment

For the testing, MSFT Windows Server has been used as Server OS for the backend servers and F5 Big IP and its LTM features as load balancer to manage traffic.

The environment has been setup to mimic, as closely as possible, a production environment. The following IP addresses are used throughout the document to make it easier to understand the configuration steps:

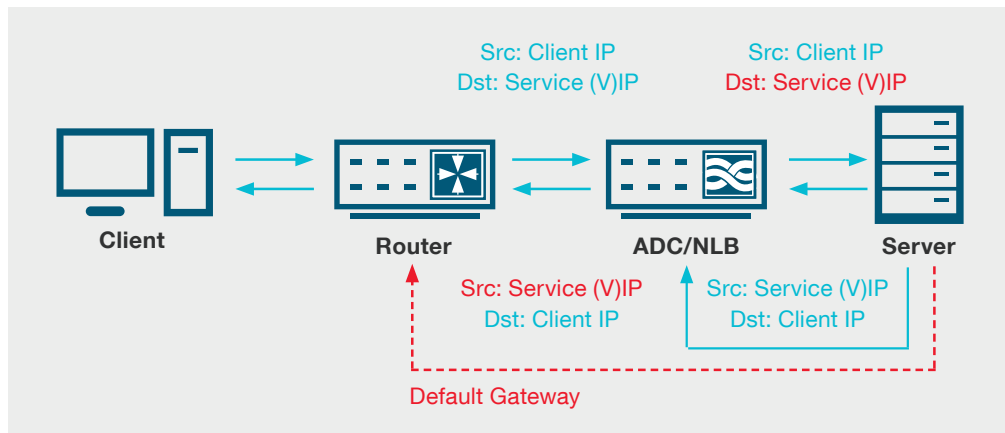
- **VIP: (13.219.3.57)** This would be the (virtual) IP address that would represent all backend servers. Clients would only know about this IP address or, better, the hostname assigned to it. Clients should be able to resolve this hostname to point to the VIP and therefore changes on the DNS infrastructure would be required.
- **Backend servers IP address: (13.219.3.55 and 13.219.3.56).** Two backend servers on the IP addresses already mentioned.

NOTE: Backend Servers are all placed in subnet **13.219.4.0/24** whilst the VIP is on 13.219.3.0/24. This is just one of the many possible setups.

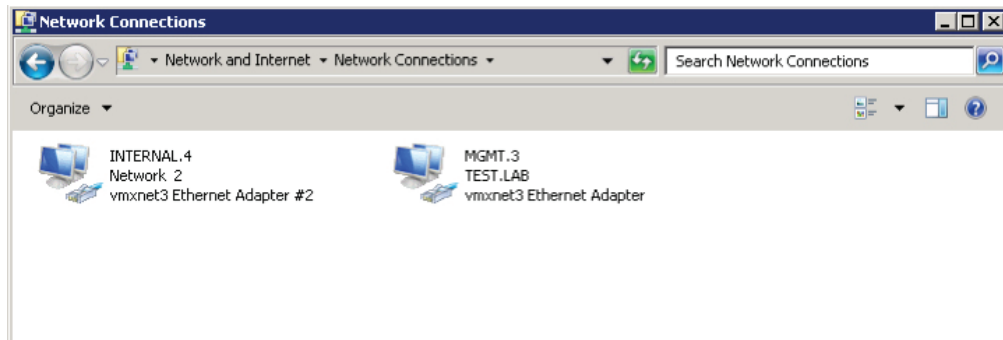
6. Backend Servers

Virtual Server configured as Destination NAT (aka In-line) works by preserving the source IP address and instead changing the destination IP address (DNAT) of the packet with that on the backend server it is trying to reach before forwarding the packet through one of the internal interfaces to one of the nodes in the pool.

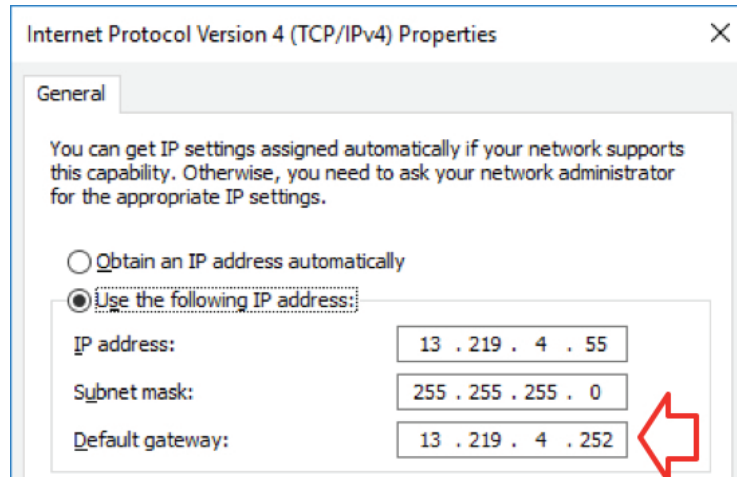
IMPORTANT: To avoid asynchronous routing, the backend servers would need to be configured with a default route pointing back to the ADC instead of the router, as seen below.



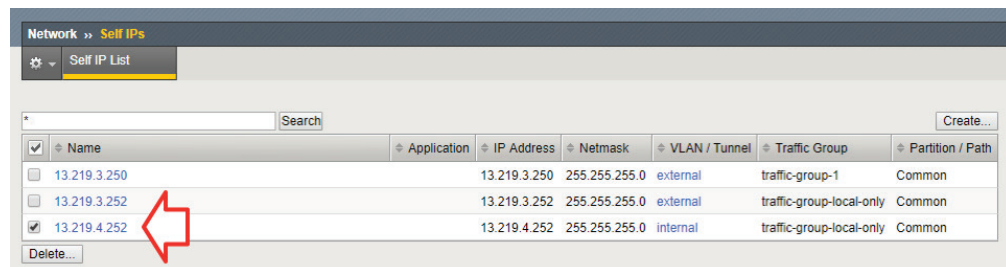
In this configuration, the servers would require having a second NIC so that they can be accessed directly and not through the ADC to perform updates and general maintenance.



The screenshot above shows a sample taken from a backend server. The Internal.4 adapter has 13.219.4.55 IP address, server #1, and a default gateway pointing to a Self-IP address on the ADC. The MGMT.3 adapter is used to remote onto the server to perform general maintenance. IPv4 settings can be seen below.



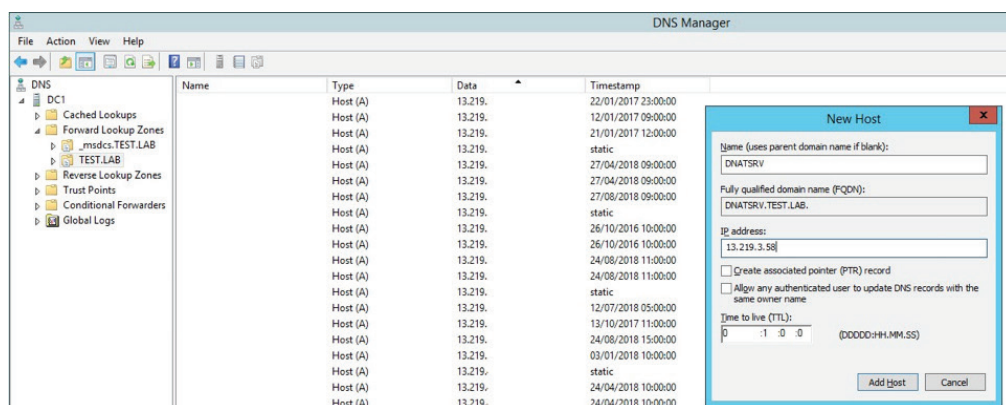
The screenshot below has a list of Self-IP addresses, including a Self-IP address 13.219.4.252 that would be used to forward requests to servers (nodes) in the pool.



7. DNS Host “A” Record

As mentioned before, it is **optional but strongly recommended** that clients are configured with the hostname of the (virtual) print server as opposed to its (virtual) IP address.

In the screenshot below, a Host “A” record have been added to the DNS server that would resolve to the VIP on the F5 (13.219.3.58).



NOTE: At this point the F5 has not yet been configured and even a ping command would fail (there’s no virtual IP address yet). To verify that the host record has been successfully added, use the **nslookup** command instead.

8. F5 Big IP LTM's Configuration

The screenshots that follow have been taken from a F5 Big IP v13.1.0.2. Other newer/older versions might have a slightly different interface. The configuration steps below would describe the creation of the following LTM objects:

- Health Monitors
- Nodes (Backend Servers)
- Pools (Services)
- FastL4 Profile (for nPath)
- Persistence Profile
- Virtual Server (nPath)

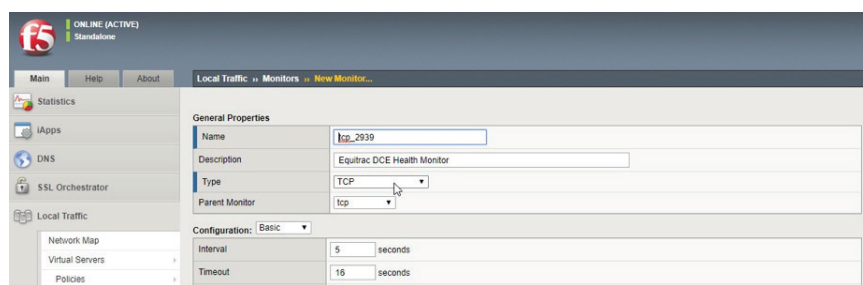
8.1 (Health) Monitors

A monitor is an object that would determine whether a backend service is (still) available. It defines where and what to look for, to determine if a service is down. It would also define how long to wait before taking the backend server off the pool of servers.

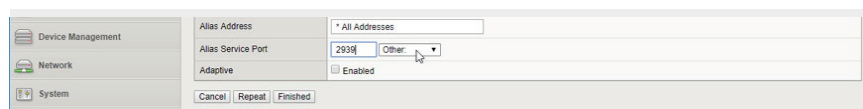
Basic health monitors are already included and could be configured at the node level (e.g. icmp/ping). At a pool level however we recommend to monitor the service(s) the pool is providing. A common approach is to use telnet to connect to the TCP port a **Nuance** Imaging Solution or Service is listening at (i.e. Equitrac DRE on port 2938, Equitrac DCE on port 2939, AutoStore ports such as 3350 or 3310).

To add a new monitor, to check on one of the ports mentioned above, at the ADC's web interface:

1. On the left-hand menu, go to **Local Traffic – Monitors**
2. Click on the **Create** button to create a custom monitor
3. Enter a meaningful name that represents the purpose of the custom monitor
4. Select **TCP** as type, this will drop down further settings



5. Towards the bottom, at the **Alias Service Port**, select **Other** and enter the TCP port to be monitored. In the example below, it would be port 2939 for the Equitrac DCE Service. To monitor **Ricoch Capture** in AutoStore, port 3350 would be the default port for that service.



6. Click on **Finished** to accept the changes.

NOTE: The above is just an example of how to create a health monitor. To monitor other services or solutions replace the TCP port with a relevant one.

8.2 Nodes (Servers)

A node is an object representing backend servers, one node/object per server is required. In this example, we are balancing between two different backend servers with IP addresses 13.219.3.55 and 13.219.3.56.

At the ADC's web interface:

1. On the left-hand menu, go to **Local Traffic – Nodes – Node List**.
2. Click on the **Create** button to create a new real server.
3. Enter a name. Whilst this could be any meaningful name, it is considered best practice to use the hostname of the backend server. This time, a generic name has been used instead.
4. Enter the hostname or IP address of the backend server.

5. Click **Finished** to create and enable the node.

Repeat the steps above for every node. At the end of the process double check the list of nodes.

Status	Name	Description	Application	Address	FQDN	Ephemeral	Partition / Path
Green	Nuance_Server_1	Nuance Imaging Server 1		13.219.4.55		No	Common
Green	Nuance_Server_2	Nuance Imaging Server 2		13.219.4.56		No	Common

NOTE: Notice that a simple icmp health monitor has been selected when the node was created. If no health monitor was selected the server/node status would appear as blue in the list above, red if there is a health monitor that fails to contact the server/s and green if the monitor is successful.

NOTE: ICMP traffic is by default blocked on modern MSFT Server OS versions. Firewall changes might be required to show the backend servers as on-line.

8.3 Pools

The pool is an object that represents a group of backend servers (nodes) and the service it is providing. It also allows admins to set the load balancing method.

At the ADC's web interface:

1. On the left-hand side menu, go to **Local Traffic – Pools – Pool List**.
2. Click on the **Create** button to add a new pool.
3. Enter a meaningful name for the pool and, under **Health Monitors**, select the health monitor added before, to help monitor status of the pool members.
4. As **Load Balancing Method** select **Round Robin**. This is the preferred method for testing, but it is highly recommended to switch to a method that considers server's specification, actual traffic or number of connections before moving onto production.
5. In **New Members**, select **Node List** to have access to the list of nodes previously added. Select a node, select ***All Services** as service port and click on Add ... repeat the above for as many nodes as needed.

Node Name	Address/FQDN	Service Port	Auto Populate	Priority
Nuance_Server_1	13.219.4.55	*	0	
Nuance_Server_2	13.219.4.56	*	0	

6. Click on **Finished** once all the settings have been entered.

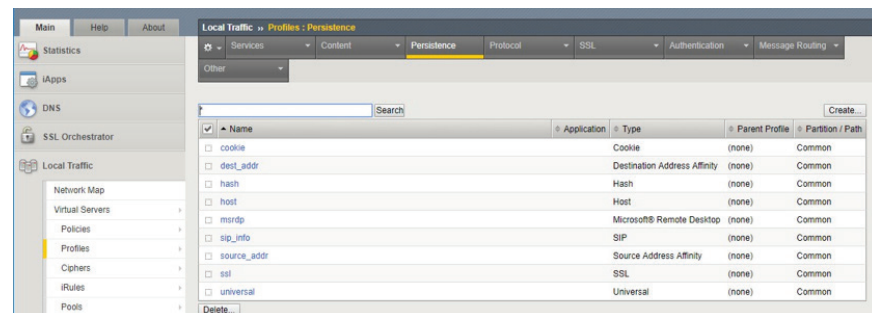
The pool will be created, and the status will go green so long there's at least one node in the pool with a green status and providing the backend servers are online and have been configured as shown in this paper.

Status	Name	Description	Application	Members	Partition / Path
	Nuance_Pool	Pool of Nuance Imaging Servers		2	Common

8.4 Persistence Profile

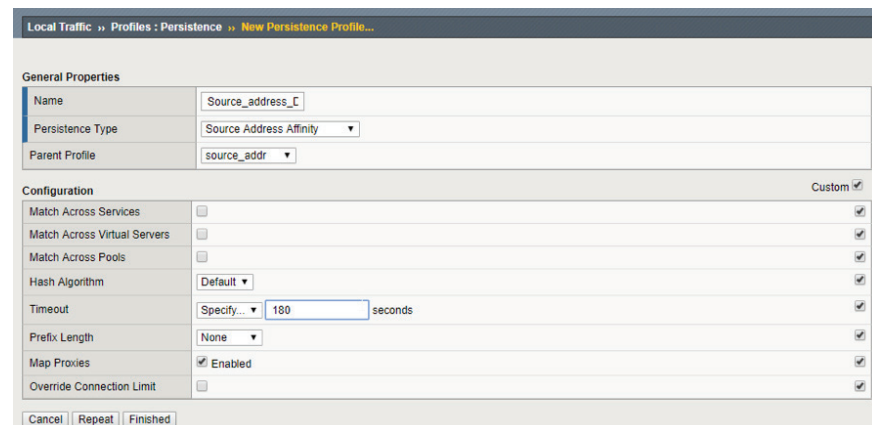
Persistence is a feature that allows the Local Traffic Manager to track and store session data, such as the specific pool member (node) that serviced a client request. The primary reason for tracking and storing session data is to ensure that client requests are directed to the same node throughout the life of a session or even during subsequent sessions. Persistence would allow an MFP to send a large scan job to the same backend server or a workstation client to send a large print job to the same backend (print) server.

Out of all persistence methods, we'd strongly recommend using **Source Address**. This works well with clients being either MFPs or workstations. Review the existing persistence profile and see if changes would be necessary, e.g. adjust timeout. If so, it is recommended to create a new profile based on the existing one.

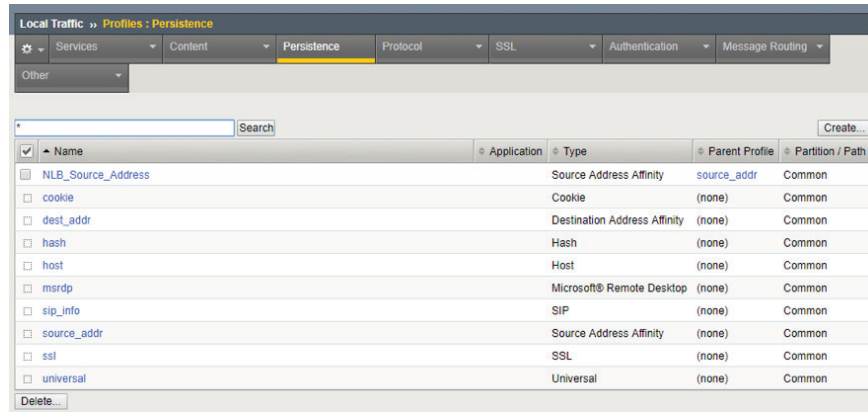


To add a new persistence profile, at the ADC's web interface:

1. On the left-hand side menu, go to **Local Traffic – Profiles – Persistence**
2. Click on the **Create** button to add a persistence profile
3. Enter a meaningful name for the profile and make sure you select **source_addr** as the parent profile
4. Enable Custom Settings by ticking on the tick box on the right-hand side.
5. Adjust the Timeout to a new value. This value will depend on the service being provided. Once the time expires, the entry in the mapping table would be deleted and the client might end up being redirected to a different backend server.



The new profile would be added to the existing profiles.



Name	Application	Type	Parent Profile	Partition / Path
NLB_Source_Address	Source Address Affinity	source_addr	Common	Common
cookie	Cookie	(none)	Common	Common
dest_addr	Destination Address Affinity	(none)	Common	Common
hash	Hash	(none)	Common	Common
host	Host	(none)	Common	Common
msrdp	Microsoft® Remote Desktop	(none)	Common	Common
sip_info	SIP	(none)	Common	Common
source_addr	Source Address Affinity	(none)	Common	Common
ssl	SSL	(none)	Common	Common
universal	Universal	(none)	Common	Common

8.5 Virtual Servers

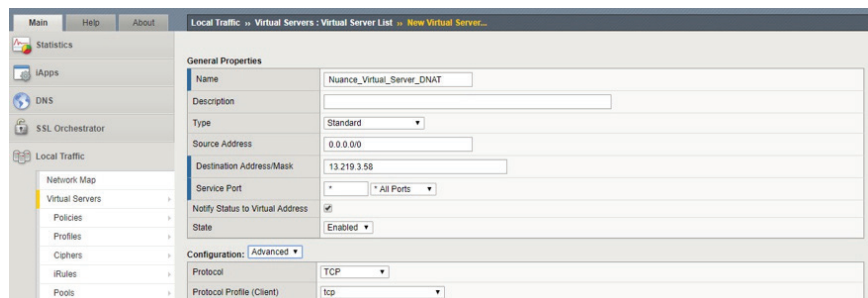
This is the last step in the process. Once the virtual server is online, clients would need to be (re)configured to point to this virtual server instead of the real backend one(s). The server would need an addressable (virtual) IP address and, optionally but highly recommended, a hostname. The latter would have been already in place if the chapter on how configure a static “A record” in the DNS servers has been already followed.

To add a Virtual Server, at the ADC’s web interface:

1. On the left-hand side menu, go to **Local Traffic – Virtual Servers – Virtual Servers List**.
2. Click on the **Create** button to add a virtual server.
3. Enter a meaningful name, the **IP address** of the virtual servers under **Destination Address/Mask** and select ***All Ports** as Service Ports in **General Properties**.

NOTE: Source Address 0.0.0.0/0 means that the virtual server will accept request from any client regardless of the source address. If that field is left empty, it would be automatically populated with the “all” address.

4. Select **Standard** as **Type** under **General Properties**. Make sure TCP is selected for both **Protocol** and **Protocol Profile** under **Configuration**.



Local Traffic » Virtual Servers - Virtual Server List » New Virtual Server...

General Properties

Name: Nuanca_Virtual_Server_DNAT

Description:

Type: Standard

Source Address: 0.0.0.0/0

Destination Address/Mask: 13.219.3.58

Service Port: * All Ports

Notify Status to Virtual Address:

State: Enabled

Configuration: Advanced

Protocol: TCP

Protocol Profile (Client): tcp

5. In **Configuration**, make sure **Address Translation** is enabled and **Port Translation** is disabled.

NOTE: As a reminder, with Destination NAT (or In-Line as it is known on the F5s) the NLB appliance would work as a router and would need to be configured as the default gateway on the server. Client's address would be preserved but destination address would be replaced by that on the backend server.

Address Translation	<input checked="" type="checkbox"/> Enabled
Port Translation	<input type="checkbox"/> Enabled

6. In **Configuration**, make sure **Source Address Translation** is disabled.

NOTE: This setting should be disabled by default but we recommend to double check this.

Source Address Translation	None
----------------------------	------

7. And finally, down below in **Resources**, select the previously created pool as **Default Pool** and the previously created persistence profile as **Persistence Profile**.

Resources							
IRules	<table border="1"> <tr> <td>Enabled</td> <td>Available</td> </tr> <tr> <td></td> <td> /Common _sys_APM_ExchangeSupport_OA_BasicAuth _sys_APM_ExchangeSupport_OA_NtlmAuth _sys_APM_ExchangeSupport_helper _sys_APM_ExchangeSupport_main </td> </tr> <tr> <td>Up</td> <td>Down</td> </tr> </table>	Enabled	Available		/Common _sys_APM_ExchangeSupport_OA_BasicAuth _sys_APM_ExchangeSupport_OA_NtlmAuth _sys_APM_ExchangeSupport_helper _sys_APM_ExchangeSupport_main	Up	Down
Enabled	Available						
	/Common _sys_APM_ExchangeSupport_OA_BasicAuth _sys_APM_ExchangeSupport_OA_NtlmAuth _sys_APM_ExchangeSupport_helper _sys_APM_ExchangeSupport_main						
Up	Down						
Policies	<table border="1"> <tr> <td>Enabled</td> <td>Available</td> </tr> <tr> <td></td> <td></td> </tr> </table>	Enabled	Available				
Enabled	Available						
Default Pool	Nuance_Pool						
Default Persistence Profile	None						
Fallback Persistence Profile	None						

Cancel Repeat Finished

8. Click on **Finished** to save the settings.

The virtual server will be created, and the status will go green so long there's at least one node in the pool the virtual server is using with a green status and providing the backend servers are online and have been configured as shown in this document.

Local Traffic > Virtual Servers : Virtual Server List							
Virtual Server List							
Status	Name	Description	Application	Destination	Service Port	Type	Resources
<input checked="" type="checkbox"/>	Nuance_Virtual_Server_DNAT			13.219.3.58	0 (Any)	Standard	Common

Enable Disable Delete...

9. Testing Virtual Servers

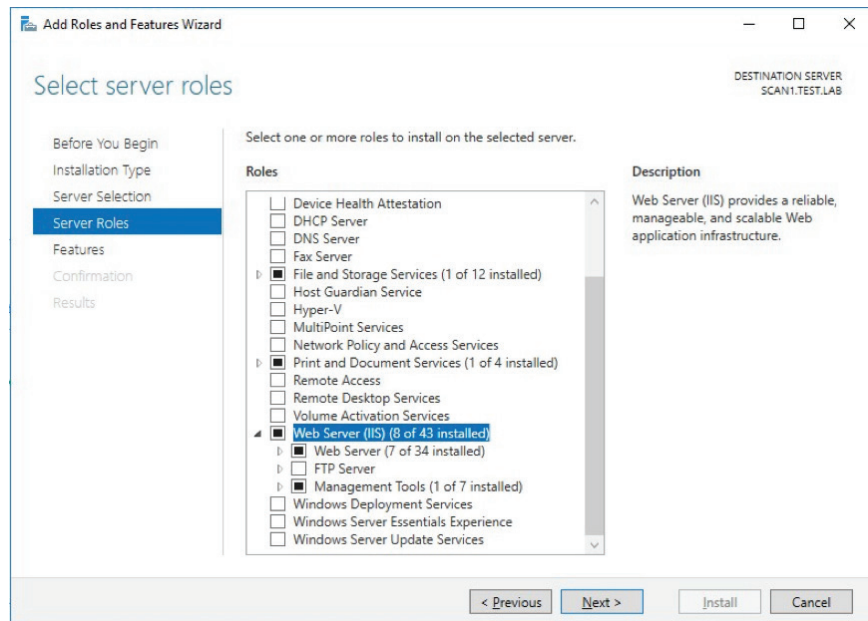
The easiest way to test whether the configuration so far works is to add a Web Server (IIS) role to the backend servers. The configuration of the ADC would currently allow traffic from any/all ports to be forwarded to the backend server. A web server would help test and make sure all the steps taken so far are correct without the need of any other software or solution and therefore no need for any especial client(s), just the backend (web) servers and an internet browser.

NOTE: Even if the ADC has been configured for Round Robin, if persistence has been enabled, the requests coming from the same client are likely to point always to the same backend server, at least until such time the persistence profile times out. To test both/all nodes, just stop the service the health monitor is pointing at. This will, in due time, force the node out of the pool and new requests to be sent to the remaining online nodes.

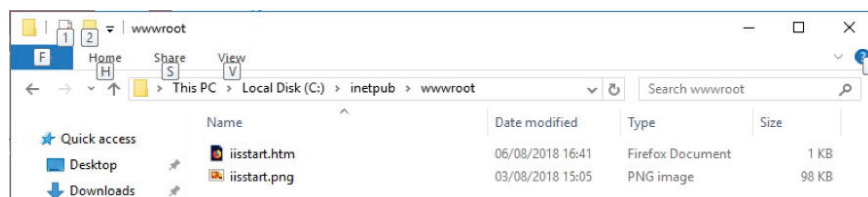
9.1 Web Servers

1. Add the Web Server role to the backend servers. Whilst the steps to add this role are beyond this document's scope, a sample can be seen below.

NOTE: Just basic defaults settings would be sufficient.

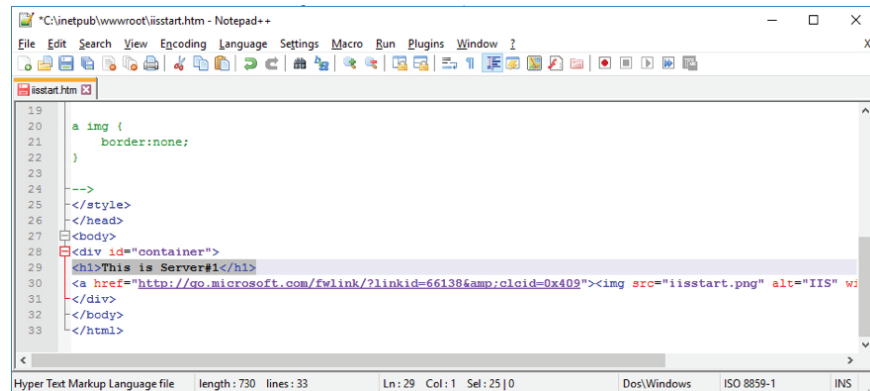


2. Edit the default web page to make sure this page is easily identified and unique on a per backend server basis. The default root path is c:\inetpub\wwwroot.



3. Add the following line to the iisstart.htm file.

NOTE: The path and the htm file might be different depending on the version of the OS and IIS.

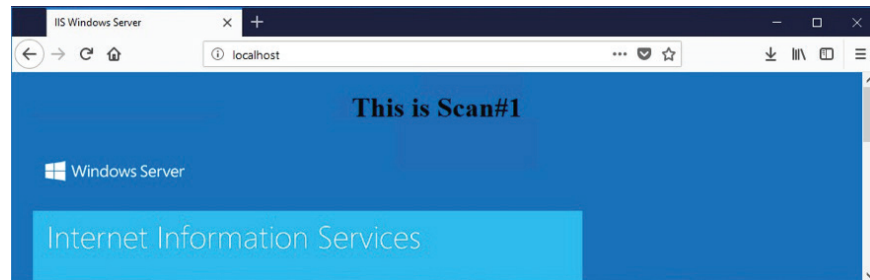


```

19
20 a img {
21     border:none;
22 }
23
24 -->
25 </style>
26 </head>
27 <body>
28 <div id="container">
29 <h1>This is Server#1</h1>
30 <a href="http://go.microsoft.com/fwlink/?linkid=66138&amp;clcid=0x409">
32 </body>
33 </html>

```

4. Open a local browser and verify the web server is indeed working.



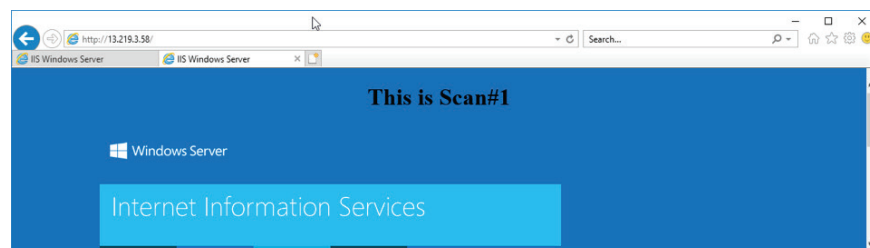
9.2 Network Traces

Wireshark or equivalent can be used to capture network traces and verify the traffic from the ADC towards to backend server behaves as expected. As a reminder, the ADC has been configured as Local Traffic Manager, Destination NAT (In-line).

Reminder of the IP addresses being used:

- 13.219.3.58: Virtual IP
- 13.219.3.55: Backend Server #1
- 13.219.3.56: Backend Server #2
- 13.219.9.153: **Client (browser) IP address**

From the client's browser, point to the virtual IP address (13.219.3.58) and make sure you get a response.



The Wireshark trace will confirm the following:

- The request comes **from the client's subnet and IP address** (13.219.9.153) and not from the ADC. There is **destination address translation** but the **client's address is preserved**.
- The request has been sent to the Virtual IP (13.219.3.58) **but the target address is 13.219.4.55**. The latter is the address of the adapter on the backend server. This adapter is configured with the address of the ADC as the gateway and therefore the server **replies to the ADC**.

No.	Time	Source	Destination	Protocol	Length	Info
500	32.412265	13.219.9.153	13.219.3.58	TCP	62	80 → 36150 [SYN, Seq=0, win=4380, Len=0, MSS=1460, SACK_PERM=1]
501	32.412304	13.219.4.55	13.219.9.153	TCP	62	80 → 36150 [ACK, Seq=0, Ack=1, win=8192, Len=0, MSS=1460, SACK_PERM=1]
502	32.412400	13.219.9.153	13.219.4.55	TCP	60	36150 → 80 [ACK, Seq=1, Ack=1, win=4380, Len=0]
503	32.412401	13.219.9.153	13.219.4.55	HTTP	405	GET / HTTP/1.1
504	32.412105	13.219.4.55	13.219.9.153	HTTP	197	HTTP/1.1 304 Not Modified
505	32.413243	13.219.9.153	13.219.4.55	TCP	60	36150 → 80 [ACK, Seq=352, Ack=144, win=4523, Len=0]
506	32.451728	13.219.9.153	13.219.4.55	HTTP	459	GET /iisstart.png HTTP/1.1
507	32.452007	13.219.4.55	13.219.9.153	HTTP	196	HTTP/1.1 304 Not Modified
508	32.452130	13.219.9.153	13.219.4.55	TCP	60	36150 → 80 [ACK, Seq=757, Ack=286, win=4665, Len=0]

If all the above is correct, the next step is to install and configure the Nuance Imaging Solution.

10. AutoStore Installation and Configuration

Ricoh Unified Client is one of the embedded solutions that leverages the NEUF framework. The NEUF client code is included in an application that runs in the device. Load balancing and high availability between the Ricoh SOP devices and the AutoStore server can be achieved by placing an F5 Big IP LTM, configured as explained in previous chapters, between the Ricoh SOP devices and a group of two or more AutoStore servers.

The following are high level configuration considerations observed throughout this document:

Load Balancer, F5 Big IP LTM.

- Source IP must be available to the AutoStore destination server. This is achieved by configuring LTM as Destination NAT (Inline) as described in previous chapters.
- Persistence based on source IP must be configured. Recommended time would be the maximum time it may take to complete a job from the time the workflow is started until the file is transferred to the AutoStore server, e.g. 5 minutes.
- Port rules must specify the port being used by the client to contact the AutoStore worker service (e.g. 3350 for Ricoh SOP devices) or make sure all ports for a given (V)IP address are forwarded. This document describes the latter.

AutoStore Backend Servers.

- **IMPORTANT:** All servers MUST run the exact same configuration, from the same CFG file. Workflows are identified based on GUIDs and therefore a similar or even identical workflow created on different backend servers will be different and are likely to fail. The best approach would be to only create/update CFG files from one of the backend servers and store it in a shared folder that the other servers have access to, so that they are all reading the exact same file. This way if the configuration is changed, all servers would have access to those without the need to copy and paste the file from server to server.

- Custom Scripts may not store static state outside of the form.
- Output file type must be set to multi-page format such as multi-page PDF as single page format may cause some pages at the beginning of the scan to be entirely dropped from the scan as a failover occurs while the scan is incorrectly marked as successful.

10.1 Caveats

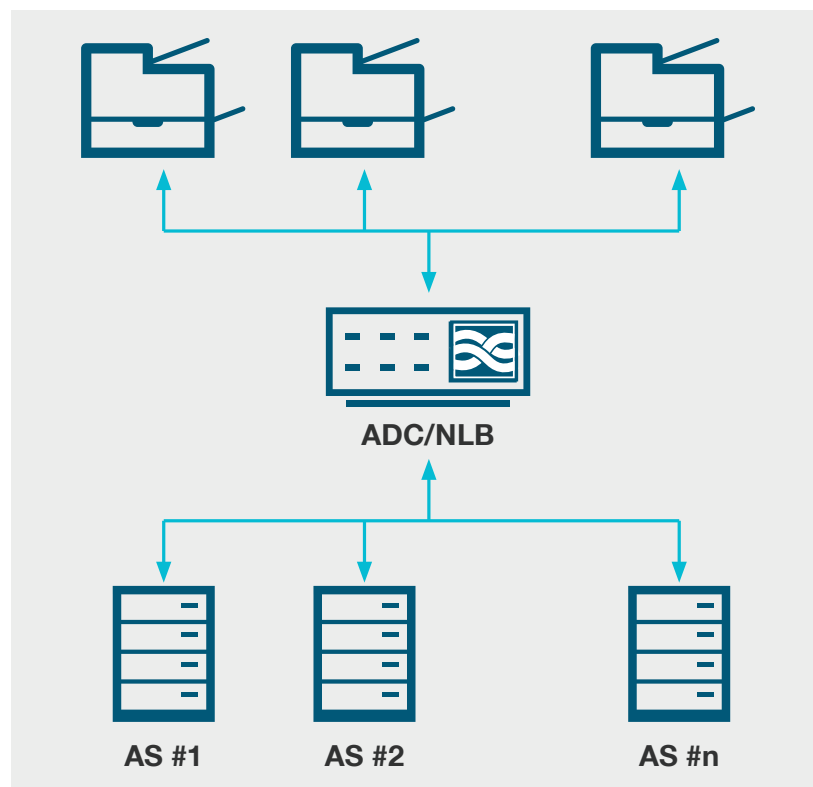
If an AutoStore backend server crashes in the middle of uploading a scan file there would be no way to recover that job and the job will fail. Also, if a node is taken offline by the F5 Big IP LTM in the middle of uploading the scan file, that job will fail.

10.2 Installing AutoStore

At the time this document is written, latest version of AutoStore is v7 SP5. Installing AutoStore is beyond the scope of this document and the process is sufficiently covered by the [Installation Guide](#) (login required). This document also assumes that all different AutoStore roles are installed onto the same server in a share- nothing configuration.

IMPORTANT: Even when separating frontend from backend roles in AutoStore, the use of load balancers is still possible although it would require the appliance to be configured slightly different to the configuration covered in this document.

At the end of the installation process, it is expected to have AutoStore 7 SP5 installed, and licensed, on all nodes part of the F5 Big IP pool dedicated to this service.

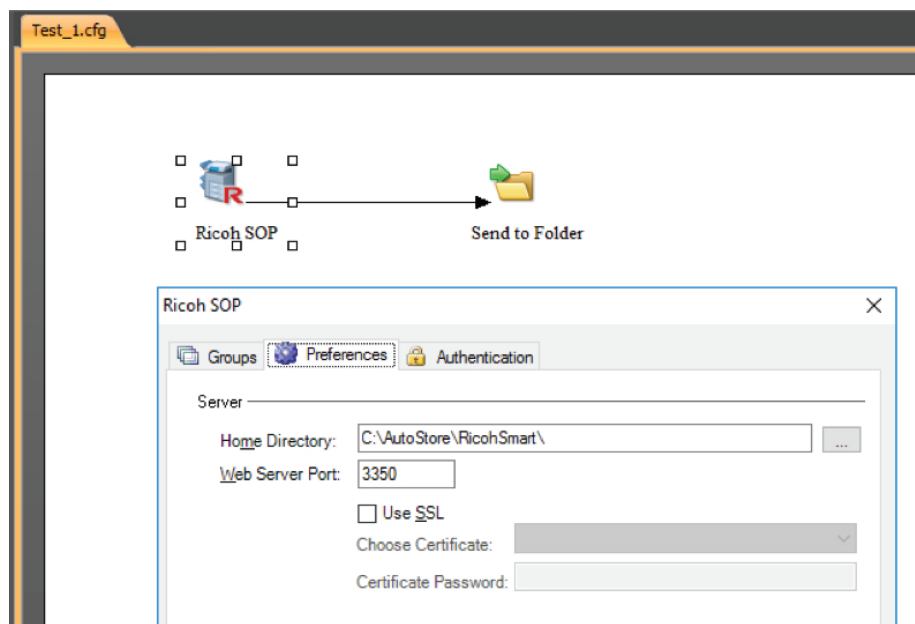


10.3 Configuring AutoStore

Advanced configuration of AutoStore is beyond the scope of this document. It would be safe to say that the workflow complexity would be completely transparent to the fact the scan job is delivered through a load balancer.

As mentioned before, AutoStore backend servers would need to be configured as if they were all standalone servers, making sure the exact same configuration steps are followed on all servers and that they have access to the same configuration and script files.

Below is a workflow example of a basic Scan to Folder workflow. The load balancer's job will finish at the "capture" component, in this case a Ricoh SOP. Network ports associated to the capture component should be used by the load balancer's monitor to determine whether the service is online or not. For the Ricoh SOP capture component, TCP/3350 should be monitored.



NOTE: Whenever possible, it is strongly recommended to test each AutoStore server independently and bypassing the load balancer to make sure the configuration works as intended. In any case it would be safe to assume the load balance is properly configured if jobs are arriving to the capture component.

As mentioned in Chapter 5, Testing Environment, AutoStore v7 SP5 has been installed on 2 nodes with IP addresses 13.219.4.55 and 13.219.4.56.

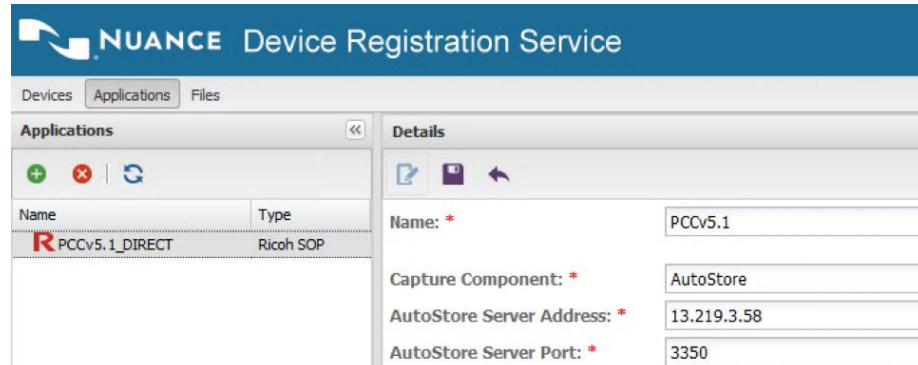
10.4 Configuring the Ricoh SOP device

At the time this document is written, latest version of Ricoh's client is Unified Client v1.1. This client is pushed to the Ricoh SOP device using Nuance DRS v7.12.

NOTE: As with the steps above, the configuration of Ricoh SOP devices for the Unified Client v1.1 using the Nuance Device Registration Service (DRS) v7.12 is beyond the scope of this document and only relevant information would be highlighted moving forward.

The below screenshot shows the "applications" configuration step in Nuance DRS. AutoStore has been selected as Capture Component and server address and port are required.

The server's IP address, or host name if this has been configured, would be the virtual IP address of the Virtual Server in the F5 Big IP appliance.



Name	Type
PCCv5.1_DIRECT	Ricoh SOP

Name: *	PCCv5.1
Capture Component: *	AutoStore
AutoStore Server Address: *	13.219.3.58
AutoStore Server Port: *	3350

NOTE: Ideally there would be no need to change the port number. Changing the default port number might require adjustments in the configuration of the load balancer and the Ricoh SOP Capture component in AutoStore.

10.5 Testing

It is recommended that basic testing is performed on the setup before it is taken into production. Basic items to test can be seen below.

Scenario	Expected Behavior
A user is presented one of more scan workflows taken from one of the backend servers, e.g. node #1. Node #1 fails before the users selects a job and it is taken out of the pool of backend servers by the load balancer.	The user is expected to be able to keep navigating scan workflows, pick one and submit it to be processed by an online node, e.g. node #2.
A form for a given scan workflow is validated on a backend node, e.g. node #1. Node #1 fails in between form validation and scan job submission and the resulting scan job is therefore sent to be processed at node #2 as traffic is redirected by the load balancer.	The scan job is expected to succeed and therefore for be processed by node #2 for as long as node #1 goes down before the file transfer to the server has started and the full scan job is sent to one of the online nodes, e.g. node #2.
A scan job has already started, all pages scanned. A given node, e.g. node #1, starts receiving the scan job when something goes wrong and it goes suddenly offline.	As mentioned in the caveats, the job is expected to fail even if it is seen as successful at the printer's front panel. AS has no way to recover the job and therefore it'd need to be started again in full.

About Nuance Communications, Inc.

Nuance Communications is reinventing the relationship between people and technology. Through its voice and language offerings, the company is creating a more human conversation with the many systems, devices, electronics, apps and services around us. Every day, millions of people and thousands of businesses experience Nuance through intelligent systems that can listen, understand, learn and adapt to your life and your work. For more information, please visit nuance.com.