

Web Server Load Balancing with Genesys Voice Platform and Radware AppDirector

Documentation Supplement to Genesys Voice Platform: Network Edition 7.2 Deployment Guide Radware: AppDirector User Guide, Software Version: 1.06.09DL

Version 1.1 March 10, 2009 Document Status: Final

CREATED BY:



NOTICE

Copyright © 2006 Genesys Telecommunications Laboratories, Inc. All rights reserved.

Genesys Telecommunications Laboratories, Inc. 2001 Junipero Serra Blvd., Daly City, CA 94014

Phone: 1-888-GENESYS Fax: 1-650-466-1260

The information contained herein is proprietary and confidential and cannot be disclosed or duplicated without the prior written consent of Genesys Telecommunications Laboratories, Inc.

ABOUT GENESYS

Genesys Telecommunications Laboratories, Inc., an independent, wholly owned subsidiary of Paris-based Alcatel, pioneered the field of Computer-Telephony Integration (CTI) and today is the leading provider of infrastructure independent contact center solutions for the enterprise, service provider, and e-business markets. With its ability to integrate interactions across all media types, including the Web and traditional voice, Genesys software helps businesses provide a consistent customer interaction experience.

TRADEMARKS

Genesys and the Genesys logo are registered trademarks of Genesys Telecommunications Laboratories, Inc. UNIX is a registered trademark of The Open Group in the United States and other countries. Microsoft, Windows, Windows 2000, Windows NT and Windows XP are registered trademarks of Microsoft Corporation. All other trademarks and trade names referred to in this document are the property of other companies.

Table of Contents

1.	1. INTRODUCTION	4
		4
		4
	1.4. GLOSSARY AND ACRONYMS	4
2.	2. LOAD BALANCING OVERVIEW	5
2	2.1. METHODS OF LOAD BALANCING	5
	2.1.1. Software Load Balancing	
		5
		5
2	2.2. PERSISTENCE	5
3.	3. GVP AND LOAD BALANCING	7
2	3.1. PIPELINING	
-		7
-	3.3. SESSION PERSISTENCE	
4.	4. TEST ENVIRONMENT CONFIGURATIO	N FOR LOAD BALANCING9
2	4.1. ENVIRONMENT CONFIGURATION	9
4	4.2. GVP ELEMENT MANAGEMENT PROVISION	ING SYSTEM9
4	4.3. Web Server Configuration	
5.	5. RADWARE APPDIRECTOR CONFIGURA	ATION11
	5.1.1. Initial AppDirector Configuration	
	5.1.3. Farm Configuration	
	0 10	
6.	6. RESULTS AND OBSERVATIONS	
(6.1. FUNCTIONALITY	
(6.4. BENEFITS	
7.	7. APPENDIX A – TEST ENVIRONMENT D	ESCRIPTION29
,		Т
	7.3. SOFTWARE VERSIONS LIST	
8.	8. APPENDIX B – DIAGRAM OF TEST ENV	IRONMENT

Documentation Supplement

Web Server Load Balancing with GVP

1. Introduction

This document provides a description of the integration of the Radware AppDirector with Genesys Voice Platform. An overview of load balancing principles and implementation details are included. This document is intended as a guide only and is not a substitute for the Radware and Genesys deployment guides. It is a supplement to be used in conjunction with the deployment guides.

1.1. Purpose

This document is a guide to configuring web server load balancing for GVP using the Radware AppDirector.

1.2. Intended Audience

The document is primarily intended for technical teams tasked with deploying web server load balancing for GVP. This includes, but is not limited to: systems integrators, customers, and Genesys and Radware field personnel. This document assumes that you have a basic understanding of:

- Computer-telephony integration (CTI) concepts, processes, terminology, and applications.
- Web server concepts
- Network design and operation
- Your own network configurations

1.3. References

All related resources listed in the GVP deployment guide and Radware AppDirector User Guide should be referenced as necessary. Consult the third party manuals for the web servers as needed.

Term	Definition
GVP	Genesys Voice Platform
AD	AppDirector
IPCS	Internet Protocol Communication Server
IPCM	Internet Protocol Call Manager
EMPS	Element Management Provisioning System
IIS	Internet Information Services
IVR	Interactive Voice Response
DNS	Domain Name Server
VIP	Virtual IP
ASP	Microsoft Active Server Pages
JSP	JavaServer Pages
EMS	Element Management System
ASR	Automatic Speech Recognition
TTS	Text to Speech
CCAS	Contact Center Activity Simulator (Call Generator)
SIP	Session Initiated Protocol

1.4. Glossary and Acronyms

2. Load Balancing Overview

Load Balancing is used to optimize resource utilization and to decrease processing time. For web servers in particular, the number of web pages that can be served concurrently is limited by the network bandwidth and the performance of the web server itself.

In addition, load balancing provides a method of failover support not available when using only one web server. A load balancer monitors the server farm, thus when server failure or planned downtime occurs it will continue to balance the load across the remaining web servers left in the group. Failure of one server does not result in downtime, although it may result in longer processing times for the entire system.

2.1. Methods of Load Balancing

A load balancing solution manages the selection of the appropriate physical server in a group. This group of servers may be called a "server farm" or a "server pool". There are three types of load balancing solutions available: software, hardware, and clustering or session replication.

2.1.1. Software Load Balancing

A software load balancing solution can be loaded on hardware of the user's choice. A very simple example of a software solution would be "DNS Round Robin". Many web servers and web application servers have load balancing functionality included. Software solutions tend to be less expensive and typically work best for smaller, less complex applications with a lower network complexity. Some software products can provide high complexity configuration and customization; however, they usually require a significant hardware base to run on.

2.1.2. Hardware Load Balancing

Hardware load balancers use virtual IP addresses for a group of servers and rewrite source and destination IP's as they route traffic. There are two types of load balancing solutions that could be classified as hardware load balancers. The first is switch or router based which balances at the network level using layer 2 and 3 functionality. It is typically the most robust; however, it does not provide the ability to direct traffic based on cookies or URLs. The second type is really a software solution that is packaged with specific hardware and sold as a unit. This type of solution provides more functionality and flexibility with switching based on layers 4 through 7, but is more complicated to configure.

2.1.3. Clustering or Session Replication

Clustering or Session Replication can be used as an alternative to hardware load balancing when sticky sessions or persistence is required. It consists of a farm or cluster of web application servers which have the ability to replicate and maintain session state in memory on all servers or using a common database. Traffic can then be directed to any server in the cluster even if persistence is required since all servers have current session state for all sessions. Clustering offers the advantage of high availability and failover support. When one server fails or is taken out of service, the other servers are still aware of session states for sessions on the failed server and are therefore able to seamlessly take over. Clustering does not in actuality replace load balancing. The traffic still has to be directed to the servers in an ordered manner by some process. However it does solve the problem of maintaining session stickiness.

2.2. Persistence

Depending on how the state of a client session is managed, persistence of dynamic content may be required. This can be accomplished using a shared database, session replication in memory on the web servers as described above, or by sending all dynamic requests for a session to the same web server. If the latter option is used, it can be realized either by using source IP persistence where all requests from

one client are always serviced by the same web server. Another method of achieving persistence is session persistence, where all parts of a session are identified by a unique identifier and forwarded to the same web server. There are no persistence requirements for static content.

3. GVP and Load Balancing

In GVP, the web servers host the web applications that generate VXML pages. The IPCS acts as the client, invoking the VXML pages and receiving content from the web servers. The state of the client transaction is managed on the web servers. Therefore, when software or hardware load balancing is implemented, the same web server must handle all transactions that are part of the same call session. Otherwise, the caller will be unable to progress through the IVR, resulting in an unpredictable response from GVP from the end user's perspective. The caller could be disconnected without warning at any point within the IVR.

GVP can benefit from the implementation of web server load balancing. Without load balancing, one or multiple applications are tied to individual web servers. The load across the servers can be widely varied over time, limiting the performance of GVP.

If load balancing is used, the traffic across the web servers can be optimized. This results in consistently equal utilization of the servers' resources, independent of an individual application's volume of traffic and therefore increased overall system throughput.

3.1. Pipelining

When information is required from the web servers, the IPCS opens a TCP connection to a web server. Instead of tearing down the connection after every request, GVP leaves the connection open, allowing for fewer delays due to the setting up and tearing down of connections and therefore resulting in better performance. Additional TCP connections are opened as dictated by load – a new connection is opened once all current connections are full. For continuous and uniform traffic, this improves efficiency. However, it introduces a level of complexity in the area of load balancing.

Due to this feature, a "many-to-many" ratio exists between call sessions and TCP connections. The requests belonging to one call session can be spread across many TCP connections, and one TCP connection can contain requests from many calls sessions. Therefore, you are never guaranteed to have only one call session or even a complete set of requests for a number of calls sessions, within the same TCP connection.

A common assumption of many load balancing devices is that a "one-to-one" or "one-to-many" ratio of call sessions to TCP connections exists. This is an incorrect supposition with respect to GVP. The intricacies of GVP's method of communication via TCP connections results in specific requirements for load balancing.

The pipelining of requests between the IPCS' and web servers makes it impossible to implement session persistence without breaking the "pipes". The load balancing device must either be configured for IP persistence or it must persist at a session level by intercepting every request between the IPCS and web server and examining every packet for a unique ID. Assumptions cannot be made based on the TCP socket that is transmitting the request.

3.2. IP Persistence

IP persistence is realized when the load balancer assigns each IPCS to one web server. All TCP connections from one IPCS will be persisted to a single web server. Since all requests for a call session originate from the same IPCS, this ensures that all parts of a call session are served from the same web server where the state of the call is maintained.

One downside to this method of load balancing is that uneven usage of the web servers may occur. If there are three IPCS' and two web servers, one web server will serve two of the ICPS' and the other web server will only serve one IPCS. This results in one web server having twice the traffic of the other one. As well, it will never be possible, using this persistence method, to have more active web servers than IPCS', as each IPCS must be persisted to a single web server and additional web servers would not be

used. Additional web servers could exist as backups, but would not be utilized until a web server failed or was removed from service.

Another negative result of this method of implementing load balancing is that a web server cannot be added dynamically to the farm or pool. Since all IPCS' are already persisted to a web server, there are none to persist to a new web server when it is added. Unless one of the clients ceases to have traffic and becomes unstuck from its web server, allowing it to select a new web server with which to persist when traffic resumes, the web server will remain unused. Removing a web server for maintenance also poses a problem as once a web server is disabled, the load balancer will no longer send new traffic to it, but any TCP connections from the IPCS will remain open and as long as traffic continues to flow, they will not close.

If calls are reasonably balanced across the IPCS', load is relatively small, and there are a proportionally higher number of IPCS' than web servers, then balancing based on client IP address may be sufficient. This is highly dependent on individual GVP configurations.

3.3. Session Persistence

The optimal load balancing mode for GVP 7.2 involves session persistence. Instead of balancing the IPCS' as discrete units by assigning them a web server, each individual client session, or call, should be load balanced. This will result in the most even load distribution across web servers and will allow there to be more web servers than IPCS' if the need should arise.

Session persistence is executed as follows. As each web application request reaches the load balancing interface, it is examined for a unique ID. In the case of IIS, this unique ID is ASPSessionID and for Apache Tomcat it will be JSessionID. The packet is then routed to a web server based on this information. If the packet is the first in a string of requests from a call it will not have an identifier and a web server is chosen according to the load balancing solution's algorithm. The web server will assign a unique identifier to the session and on the return from the web server, the load balancer will store this unique identifier exists, the dynamic table is accessed to determine the web server that it should be sent to and the packet is forwarded to that web server. All web application requests for a single call session are sent to the same web server. The static requests will not contain a unique identifier and will always be load balanced according to the selected load balancing algorithm.

The downside to this method of load balancing is that opening and examining every packet increases overhead and can cause performance issues and delays. This latency that is introduced may be able to be minimized with proper optimization of the load balancing solution.

Session persistence is however superior to IP persistence when it comes to adding and removing a web server dynamically. When a web server is added the load balancer will immediately begin forwarding call sessions to it according to the load balancing algorithm in place. The load on each web server will quickly become even. When a web server is removed from service, the load balancer will allow all sessions that are stuck to that web server to complete, but will not forward any new sessions to it. This results in the time to remove a web server being equal to or less than the length of one session.

4. Test Environment Configuration for Load Balancing

The environment configured for functionality testing contained three IPCS', three web servers and an AppDirector. The full architecture of the test environment is found in Appendix A and a diagram is found in Appendix B.

The choice of three IPCS' and three web servers allowed the testing of the functionality with different combinations of clients and servers.

Multiple AppDirectors could be deployed for redundancy purposes. However, that functionality is not covered in the scope of this document.

4.1. Environment Configuration

The load balancing test environment consisted of a GVP NE 7.2 installation. This installation included three IPCS' and three web servers. The web servers were running both Microsoft IIS and Apache Tomcat. All GVP servers in the installation were positioned on one subnet with the web servers situated on a separate subnet. AppDirector was configured with one interface on each of the two subnets.

The traffic flow involving the load balancing components begins when an HTTP request is made through an IPCS to advip.lsst.genesys.ca (192.168.43.105), an IP address which is configured on AppDirector to point to the farm/pool of web servers.

AppDirector intercepts the request and inspects the header for a unique ID. If there is no unique ID, the request is load balanced to a web server according to the configured load balancing algorithm, which in this case was Cyclic (Round Robin) and on the return of the request, an entry is made in the dynamic table. If a unique ID is found, the dynamic table is accessed with the unique ID to determine the web server to send the request to.

The default gateways of the web servers were set to be AppDirector's interface on that subnet. This directed the traffic back through AppDirector. Since one web server also housed a database that was used by GVP and the applications, some minor configuration changes had to be made to allow the database traffic to reach the database through the AppDirector. These changes involved the EMPS server and consisted of changing the default gateway of that server to be AppDirector's console interface and adding the IP of the server hosting the database to the Hosts file on the EMPS server.

4.2. GVP Element Management Provisioning System

There is only one consideration within GVP's EMPS when provisioning an application on a system that is load balanced. When setting the Primary IVR URL, instead of pointing the application to a specific web server, the application should be provisioned using a Virtual IP address. This virtual address points to the farm/pool of web servers which each contain the application. In the test environment described above, this corresponds to advip.lsst.genesys.ca.

4.3. Web Server Configuration

The test environment was configured with both Microsoft IIS and Apache Tomcat web servers. This allowed for the testing of both ASP and JSP applications. Both web servers were running on port 80, which meant that only one could be running while testing was taking place. This simplified the management of the web servers as no changes needed to be made within GVP or the AppDirector to accommodate the changing of a port. If different ports were used, the AppDirector would need to perform port multiplexing to transform each request from port 80 to the correct port for the web server.

No specific web server configuration changes need to be performed for load balancing with AppDirector in either IIS or Apache Tomcat. As explained in section 3.3, the implementation of session persistence is dependent on the availability of a unique session identifier. The identifiers are ASPSessionID and JSessionID for IIS (ASP applications) and Apache Tomcat (JSP applications) respectively. The default configuration of both web servers enables session state persistence for applications, meaning that the

appropriate unique identifier is set as a cookie. It is possible to disable session state persistence at an individual ASP or JSP page level. Session state persistence should not be disabled at the web server level or the ASP/JSP page level. This would result in AppDirector losing the ability to load balance using session persistence. One further note for consideration is that if an application calls an external web application, the unique identifiers will not be the same for the calls to the external application.

5. Radware AppDirector Configuration

5.1.1. Initial AppDirector Configuration

Using a serial cable and a terminal emulation program, connect to the AppDirector.

The default console port settings are:

- Bits per Second: 19200
- Data Bits: 8
- Parity: None
- Stop Bits: 1
- Flow Control: None
- 1. Assign the following IP address to interface 1 of the AppDirector:
 - 192.168.43.106 /25 IP Address from the GVP subnet
- 2. Create a default gateway route entry on the AppDirector pointing to 192.168.43.1.
- 3. Using a browser, connect to the IP Address of the AppDirector (192.168.43.106) via HTTP or HTTPS. The default username and password are "radware" and "radware".

Failure to establish a connection may be due to the following:

- Incorrect IP Address in the browser
- Incorrect IP Address or default route configuration in the AppDirector
- Failure to enable Web Based Management or Secure Web Based Management in the AppDirector
- If the AppDirector can be successfully pinged, attempt to connect to it via Telnet or SSH. If the pinging or the Telnet/SSH connection are unsuccessful, reconnect to the AppDirector via its console port. Once connected, verify and correct the AppDirector's configuration as needed.¹

: radv	• 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	• • • • • • • • • • • • • • • • • • •	A	pDirector	Global	agemen with Cookie PS and DoS		0 00 0 0 0 000 00 0 00 00 00 00	Back Forward Refresh
				isistericy,	B 111 II, II	S and D03			
File •									
Device >									
Bridge 🔸									
Router •							F1 F2 F3 F4	F9 F10 F11 F12	
AppDirector +	radware		_	Application Delivery	जनाजना 🕈	AppDirector			
Health Monitoring 🕨	••	G			G4	GS	F5 F6 F7 F8	F13 F14 F15 F16	
Security >									
BWM >									
Classes +									
Performance +									
Services +									
Help									

¹ To enable web-based management from the console command line interface, use "manage web status set enable"

5.1.2. IP Configuration

1. From the menu, select **Router** ⇒ **IP Router** ⇒ **Interface Parameters** to display the **IP Interface Parameters** page similar to the one shown below:

IP Interface Parameters								
Routing Table ARP Table								
IP Interface Parameters								
IP Address	Network Mask	lf Number	VlanTag	×				
<u>192.168.43.106</u>	255.255.255.128	1	0					
Delete Create								

- 2. Click the **Create** button.
- 3. On the **IP Interface Parameters Create** page, enter the necessary parameters as shown below:² This will create the Web Server interface.

IP Interface	e Parameters (Jpdate	
	Routing Tab	ile <u>ARP Table</u>	
(IP Address:	192.168.43.203)	lotwork Mask:	255.255.255.224
(If Number:		wd Broadcast:	enable 💌
Broadcast Addr:	ONE FILL 🕑 🚺	/lanTag:	0
	Set	Cancel	

4. Verify that the new entry was created on the IP Interface Parameters page:

² Items circled in red indicate settings that need to be entered or changed. Items not circled should be left to default settings.

IP Interface Parameters

```
Routing Table ARP Table
```

IP Interface Parameters

IP Address	Network Mask	lf Number	VlanTag	×					
<u>192.168.43.106</u>	255.255.255.128	1	0						
<u>192.168.43.203</u>	255.255.255.224	2	0						
Delete Create									

5.1.3. Farm Configuration

1. From the menu, select **AppDirector** ⇒ **Farms** ⇒ **Farm Table** to display the **Farm Table** page similar to the one shown below:

	Help						
Farm Name	Admin Status	Aging Time [sec]	Dispatch Method	Connectivity Check Method	Sessions Mode	Operational Status	×
			Delet	e Create			

- 2. Click the Create button.
- 3. On the Farm Table Create page, enter the necessary parameters as shown below:³

³ Items circled in red indicate settings that need to be entered or changed. Items not circled should be left to default settings.

	Farm	Table Update		
	Extended Farm Parameters	Laver 4 Policy Table	Server Table DNS Persistency Parar	neters Table
Farm Name:	WebServers		Admin Status:	Enabled V
Operational Status:	Active		Aging Time [sec]:	60
Dispatch Method:	Cyclic	*	Connectivity Check Method:	No Checks 💙
Sessions Mode:	ServerPerSession	~	Bandwidth Limit:	No Limit 💌
Connectivity Check Port:	HTTP	*	Connectivity Check Interval [sec]:	10
Connectivity Check Retries:	5		Extended Check Frequency:	10
Home Page:			Authorized Username:	
Authorized Password:			Connection Denials:	0
Distribution Threshold:	2500		Capacity Threshold:	5000
Redirection Mode:	No Redirection	~	DNS Redirection Fallback:	DNS Only 🕑
HTTP Redirection Mode:	IP Mode 🛛 👻			
		Set	Cancel	

- 4. Click the Set button to save parameters.
- 5. Verify that the new entry was created on the **Farm Table** page:

Farm Table								
Extended Farm Parameters Layer 4 Policy Table Server Table DNS Persistency Parameters Table								
Farm Name	Admin Status	Aging Time [sec]	Dispatch Method	Connectivity Check Method	Sessions Mode	Operational Status		
WebServers	Enabled	60	Cyclic	No Checks	ServerPerSession	Active		
Delete Create								

- 6. Click the Extended Farm Parameters URI at the top of the Farm Table page.
- 7. On the Extended Farm Parameters Table page, click on the Farm Name WebServers.
- 8. On the **Extended Farm Parameters Update** page, enter the necessary parameters as shown below:

Extended Farm Parameters Update							
		Farm Table					
Farm Name	WebServers	Radius Secret					
Connection Limit Exception	Disabled 💌	Client NAT Address Range	0.0.0.0 💌				
Transparent Server Support	Disabled 💌	SSL ID Tracking	Disabled 💌				
Close Session At Aging	Disabled 💙	RADIUS Attribute	0				
Reset Client on Server Failure	Disabled 💌	RADIUS Proxy Attribute	0				
Add X-Forwarded-For to HTTP requests	Disabled 💌	Insert Cookie for HTTP Persistency	Disabled	v			
Hash Parameter For SIP	Call-ID 💌	SSL ID Aging	120				
Select Server Per Transaction	Enabled 💌						
		T ×					
		Set Cancel					

9. Click the **Set** button to save parameters.

5.1.4. Create Layer 4 Policy

1. From the menu, select AppDirector ⇒ Layer 4 Farm Selection ⇒ Layer 4 Policy Table to display the Layer 4 Policy Table page similar to the one shown below:

	Layer 4 Policy Table							
Farm Table Layer 7 Policy Table Layer 4 Policy Statistics								
Virtual IP	L4 Protocol	L4 Port	Source IP From	L4 Policy Name	L7 Policy	Farm Name	×	
Delete Create								

- 2. Click the **Create** button.
- 3. On the Layer 4 Policy Table Create page, enter the necessary parameters as shown below.

La	yer 4 Policy	Table Cre	ate			Help
	Farm Table	Layer 7 Policy 1	<u>Fable</u>	Layer 4 Policy Statistics		
Virtual IP	192.168.43.105	\supset		L4 Protocol	TCP 💌	
L4 Port	Any		*	Source IP From	0.0.0.0	
L4 Policy Name	Web Servers)		Source IP To	0.0.0.0	
Farm Name	WebServers)	*	L7 Policy Name	None	*
Application	HTTP) 💌		Redundancy Status	Primary 🚩	
Backend Encryption Port	0			Bytes of Request to Read	3584	
POST Classification Input	Header	*		HTTP Normalization	Disabled 💌	
L7 Persistent Switching Mode	Overwrite y			Segment Name	*	
		Set	X Cancel)		

- 4. Click the **Set** button to save the parameters.
- 5. Verify that the new entry was created on the Layer 4 Policy Table page:

Layer 4 Policy Table									
		<u>Farm Table</u>	Laver 7 Policy Table	Layer 4 Policy Statistics					

Virtual IP	L4 Protocol	L4 Port	Source IP From	L4 Policy Name	L7 Policy	Farm Name			
<u>192.168.43.105</u>	ТСР	Any	0.0.0.0	WebServers	None	WebServers			
Delete Create									

5.1.5. Configure Dynamic L7 Persistency

1. From the menu, select AppDirector ⇒ L7 Server Persistency ⇒ Text Match to display the Text Match Session ID Persistency page similar to the one shown

	Text	Match Sessi	on ID Persistency		Help
		Pattern Match Sessio	n ID Persistency Static Session ID F	tersistency <u>Farm Table</u>	
Farm Name	Application Port	L4 Protocol	Persistency Identifier	Learning Direction	Ignore Server Reply
	·	<i></i>	Delete Create	, 	· · · · · · · · · · · · · · · · · · ·

2. Click the Create button.

3. On the **Text Match Session ID Persistency Create** page, enter the necessary parameters as shown below.

Farm Name:	WebServers	\supset	*	Application Port:	0	
L4 Protocol:	TCP 💌			Persistency Identifier:		
Lookup Mode:	Cookie	*		Identifier Match:	Prefix 💙	
Learning Direction:	Server Reply	*		Ignore Server Reply:	Never	*
Value Max Length:	256			Value Offset:	24	
Stop Chars:	;			Inactivity Timeout:	900	
Ignore Source IP:	Enabled 💌					
		S	et Ca	ancel		

- 4. Click the **Set** button to save parameters.
- 5. Verify that the new entry was created on the **Text Match Session ID Persistency** page:

	Pel;				
		Pattern Match Sessio	n ID Persistency Static Session ID P	ersistenov Farm Table	
Farm Name	Application Port	I 4 Protocol	Persistency Identifier	Learning Direction	lanore Server Rer

Farm Name	Application Port	L4 Protocol	Persistency Identifier	Learning Direction	Ignore Server Rep
<u>WebServers</u>	0	ТСР	SESSIONID	Server Reply	Never
			Delete Create		

5.1.6. Adding Servers to the Farm

1. From the menu, select **AppDirector** ⇒ **Servers** ⇒ **Application Servers** to display the **Server Table** page similar to the one shown below:

	Server Table									
				Farm Table Physi	ical Servers <u>Stat</u>	ic Session ID F	°ersisten ov			
Farm Name	Server Address	Server Port	Server Name	Operational Status	Operation Mode	Admin Status	Attached Users	Peak Load	Frames Load	Туре
									Delete Cr	reate

2. On the Server Table Create page, enter the necessary parameters as shown below:

	Server Table U	Jpdat	e		Hel
	Farm Table Physic	al Serv	ers Static Session ID Persistency		
(Farm Name:	WebServers		Server Address:	192.168.43.199	
Server Port:	None		Server Name:	Eagle]
Server Description:	GVP WebServers		Admin Status:	Enable 💌	
Operational Status:	Not In Service		Weight:	1]
Operation Mode:	Regular 💌		Туре:	Regular	*
Connection Limit:	0		Response Threshold [ms]:	0]
Client NAT:	Disabled 🚩		Backup Server Address:	0.0.0.0	
Redirect To:			Bandwidth Limit:	No Limit	~
Attached Users:	0		Peak Load:	0	
Frames Rate:	0		Backup Preemption:	Enable 💌	
Client NAT Address Range:	0.0.0.0 💌				
		Set	Cancel		

- 3. Click the **Set** button to save parameters.
- 4. Verify that the new entry was created on the Server Table page:

	Server	Table		Help
<u>F</u>	arm Table	Physical Servers	Static Session ID Persistency	

Farm Name	Server Address	Server Port	Server Name	Operational Status	Operation Mode	Admin Status	Attached Users	Peak Load
<u>WebServers</u>	192.168.43.199	None	Eagle	Not In Service	Regular	Enable	0	0

5. Create the second server using the information below:

Server Table Update										
	<u>Farm Table</u>	Physical S	ervers	Static Session ID Persistency						
(Farm Name:	WebServers		(Server Address:	192.168.43.200)					
Server Port:	None		(Server Name:	Hornet					
Server Description:	GVP WebServer	s		Admin Status:	Enable 💌					
Operational Status:	Not In Service			Weight:	1					
Operation Mode:	Regular 💌			Туре:	Regular	*				
Connection Limit:	0		I	Response Threshold [ms]:	0					
Client NAT:	Disabled 💌		I	Backup Server Address:	0.0.0.0					
Redirect To:				Bandwidth Limit:	No Limit		*			
Attached Users:	0		I	Peak Load:	0					
Frames Rate:	0		I	Backup Preemption:	Enable 🚩					
Client NAT Address Range:	0.0.0.0 💌									
		F	Ł	×						
		Se	et	Cancel						

- 6. Click the Set button to save parameters.
- 7. Create the third server using the information below:

Server Table Update									
	Farm Table Physica	IServe	ers Static Session ID Persistency						
(Farm Name:	WebServers)		Server Address:	192.168.43.204					
Server Port:	None		Server Name:	Tomcat]				
Server Description:	GVP WebServers		Admin Status:	Enable 💌	-				
Operational Status:	Not In Service	-	Weight:	1]				
Operation Mode:	Regular 💌		Туре:	Regular	*				
Connection Limit:	0		Response Threshold [ms]:	0]				
Client NAT:	Disabled 💌		Backup Server Address:	0.0.0.0					
Redirect To:			Bandwidth Limit:	No Limit	~				
Attached Users:	0		Peak Load:	0					
Frames Rate:	0		Backup Preemption:	Enable 💌					
Client NAT Address Range:	0.0.0.0 💌								
		Set	Cancel						

 $8. \ \ {\rm Click \ the \ } {\rm Set \ button \ to \ save \ parameters.}$

5.1.7. Health Monitoring Configuration

- 1. From the menu, select **Health Monitoring** ⇒ **Global Parameters** to display the **Health Monitoring Global Parameters** page.
- 2. On the **Health Monitoring Global Parameters** page, change the parameters as shown below:

lealth Monitoring Global Pa	arameters
Check Table Binding	g Table HM Server Table
(Health Monitoring Status:	Use Health Monitoring 💌)
Response Level Samples:	
SSL Certificate File:	rdwrhmm.cert
SSL Private Key File:	rdwrhmm.key
Ģ	Set

- 3. Click the **Set** button to save parameters.
- 4. Create the Health Monitoring Check for the First Server.
- 5. From the menu, select **Health Monitoring** ⇒ **Check Table** to display the **Health Monitoring Check Table** page similar to the one shown below:

Health Monitoring Check Table							
Binding Table Packet Sequence Table Health Monitoring Global Parameters							
Check Name Check ID Method Status Dest Host 🗙							
Delete Create							

- 6. Click the **Create** button.
- 7. On the HM Check Table Create page, enter the necessary parameters as shown below:

HM Check Table Update				
Bindir	ng Table Packet Sequence 1	Table Health Monitoring Global P	<u>arameters</u>	
Check Name:	Eagle_WebServers)	(Method:	НТТР	
Dest Host:	192.168.43.199	Next Hop:	0.0.0.0	
Destination Port:	0	Arguments:	PATH=GVP HOST=192	
Interval:	20	Retries:	3	
Timeout:	1	No New Session Timeout:	0	
Measure Response Time:	Disabled 💌	Response Level:	0	
Check ID:	0	Status:	Failed	
Dest IP:	0.0.0.0	Reverse Check Result:	disable 💌	
Uptime %:	0	Success Counter:	0	
Failure Counter:	4218	Average Duration:	0	
	Set	Cancel		

- 8. Before clicking the Set button, choose the button next to Arguments to populate the specific settings for the rest of this check.
- 9. Enter the information below:

Path:	GVP
Hostname:	192.168.43.199
HTTP Method:	GET 💌
Proxy HTTP:	No 💌
Pragma Nocache:	Yes 💙
Username:	
Password:	
Match search	
string:	
Match mode:	Lookup stri
HTTP return code:	200
HTTP return code:	
HTTP return code:	
HTTP return code:	
Set	Cancel

- 10. Click the **Set** button for the Method Arguments and click the **Set** button again in the HM Check Table Create window.
- 11. You should have a single entry in the Health Monitoring Check Table:

Health Monitoring Check Table

Binding Table Packet Sequence Table Health Monitoring Global Parameters

Check Name	Check ID	Method	Status	Dest Host	×
Eagle WebServers	0	HTTP	Failed	192.168.43.199	

Note: The status of this check may display "Unknown" until the server replies successfully to the AppDirector's check.

- 12. Create the Health Monitoring Check for the Second Server
- 13. If the **Health Monitoring Check Table** page is not already displayed from the previous step, select **Health Monitoring** ⇒ **Check Table** from the menu.
- 14. Click the Create button.
- 15. On the HM Check Table Create page, enter the necessary parameters as shown below:

HM Check Table Update Health Monitoring Global Parameters Binding Table Packet Sequence Table Check Name: Hornet WebServers) Method: HTTP 192.168.43.200 0.0.0.0 Dest Host: Next Hop: Destination Port: 0 PATH=GVP|HOST=192 Arguments: 3 Interval: 20 Retries: 1 No New Session Timeout: 0 Timeout: Measure Response Time: Disabled 🔽 0 Response Level: Check ID: 1 Status: Failed Dest IP: 0.0.0.0 Reverse Check Result: disable 🔽 Uptime %: Π Success Counter: 0 Failure Counter: 4238 0 Average Duration: × Cancel

16. Click the button next to the *Arguments* text box to configure the check specific arguments:

Path:	
Hostname:	192.168.43.200
HTTP Method:	GET 💌
Proxy HTTP:	No 💌
Pragma Nocache:	Yes 🖌
Username:	
Password:	
Match search	
string:	
Match mode:	*
HTTP return code:	200
HTTP return code:	
HTTP return code:	
HTTP return code:	
Set	Cancel

Arguments for HTTP Method

- 17. Click the **Set** button to save the Method Argument parameters.
- 18. Verify that the Arguments text box has been populated.
- 19. Click the **Set** button to save the Health Check.
- 20. Create the Health Monitoring Check for the Third Server
- 21. If the **Health Monitoring Check Table** page is not already displayed from the previous step, select **Health Monitoring** ⇒ **Check Table** from the menu.
- 22. Click the Create button.
- 23. On the HM Check Table Create page, enter the necessary parameters as shown below:

HM Check Table Update

Binding Table Packet Sequence Table Health Monitoring Global Paramet
--

Check Name:	Tomcat_WebServers	(Method:	НТТР	
Dest Host:	192.168.43.204	Next Hop:	0.0.0.0	
Destination Port:	0	Arguments:	PATH=GVP HOST=192	
Interval:	20	Retries:	3	
Timeout:	1	No New Session Timeout:	0	
Measure Response Time:	Disabled 💌	Response Level:	0	
Check ID:	2	Status:	Failed	
Dest IP:	0.0.0.0	Reverse Check Result:	disable 💌	
Uptime %:	0	Success Counter:	0	
Failure Counter:	4249	Average Duration:	0	
	Set	Cancel		

24. Click the button next to the *Arguments* text box to configure the check specific arguments:

Arguments for HTTP Method

Path:	GVP
Hostname:	(192.168.43.204)
HTTP Method:	GET 💌
Proxy HTTP:	No 💌
Pragma Nocache:	Yes 💙
Username:	
Password:	
Match search	
string:	
Match mode:	×
HTTP return code:	(200
HTTP return code:	
HTTP return code:	
HTTP return code:	
Set	Cancel

- 25. Click the **Set** button to save the Method Argument parameters.
- 26. Verify that the Arguments text box has been populated.
- 27. Click the **Set** button to save the Health Check.
- 28. Binding Health Checks to Servers
- 29. Create the Health Monitoring Binding for the First Server
- 30. From the menu, select **Health Monitoring** ⇒ **Binding Table** to display the **Health Monitoring Binding Table** page similar to the one shown below:

Health Monitoring Binding Table						
Check Table HM Server Table Health Monitoring Global Parameters						
Check	Ser	ver/NHR/Repo	rt	Group	Mandatory	\times
Delete Create						

- 31. Click the Create button.
- 32. On the HM Binding Table Create page, enter the necessary parameters as shown below:

	HM Bindin	g Table Upd	ate	
	<u>Check Table</u>	HM Server Table	Health Monitoring Global Parameter	Ξ
Check:	Eagle_WebServers)	Server/NHR/	Report: Farm WebServers -	192.168.43.199 - 0)
Group:	oļ	Mandatory:	Mandatory 💌	
		Set	Cancel	

33. Click the Set button to save parameters.

34. Verify that the new entry was created on the Health Monitoring Table page:

Health Monitoring Binding Table

Check Table HM Server Table Health Monitoring Global Parameters

Check	Server/NHR/Report	Group	Mandatory	×
Eagle WebServers	Farm WebServers - 19 2.168.43.199 - 0	0	Mandatory	

35. Create the Health Monitoring Binding for the Second Server

- 36. If the **Health Monitoring Binding Table** page is not already displayed from the previous step, select **Health Monitoring** ⇒ **Binding Table** from the menu.
- 37. Click the Create button.
- 38. On the HM Binding Table Create page, enter the necessary parameters as shown below:

HM Binding Table Update				
	<u>Check Table</u> <u>H</u>	M Server Table	Health Monitoring Global P	arameters
Check: Hornet_ Group: 0	WebServers	(Server/NHR/I Mandatory:	Report: Farm WebS Mandatory	ervers - 192.168.43.200 - 0)
		Set	Cancel	

- 39. Click the Set button to save parameters.
- 40. Create the Health Monitoring Binding for the Third Server
- 41. If the **Health Monitoring Binding Table** page is not already displayed from the previous step, select **Health Monitoring** ⇒ **Binding Table** from the menu.
- 42. Click the Create button.
- 43. On the HM Binding Table Create page, enter the necessary parameters as shown below:

HM Binding Table Upd	ate
Check Table HM Server Table	Health Monitoring Global Parameters
Check: Tomcat_WebServers)	
Group: 0 Mandatory:	Cancel

44. Click the **Set** button to save parameters.

This completes the AppDirector Configuration.

6. **Results and Observations**

6.1. Functionality

Extensive testing of AppDirector with GVP applications utilizing DTMF and voice inputs resulted in confirmation that load balancing with session persistence functions correctly. All requests belonging to a single session are load balanced to the same web server and load is evenly divided among web servers. The dynamic addition and removal of web servers and IPCS' can be achieved with no interruption to service. Failure of a web server does result in some calls being lost, as expected, but recovery time is short and recovery is complete. A moderate load was directed to the GVP applications in order that pipelining would be initiated and session persistence proven. However, further testing would be warranted to expose any issues that might arise under a heavy load of calls.

6.2. Latency

One downside of persisting at the session level is that inevitably, a certain amount of latency is introduced. This is the result of the load balancer having to inspect every packet and make a decision based on a lookup in the dynamic table containing sessions. The overhead introduced in the testing was noticeable, but did not seem to have a detrimental effect on the sessions. Again, further testing is necessary to determine the level of latency introduced when the system is under load.

6.3. Management of Dynamic Table

Careful consideration needs to be given to the management of the size of the dynamic session table. This table contains an entry for every session detailing the unique session id and the server that is servicing that call session. An inactivity timeout exists whereby if no traffic for a session is received during the length of the timeout period, the session entry is removed from the table. This timeout value should be set greater than the longest holding time possible within the GVP applications. It should not however be set to an arbitrarily large value as this may cause the table to grow to a point where performance would be impacted either by causing longer lookup times or by the table running out of space.

The entries in this table could be kept to a minimum by having the server set a cookie on each session which contains the server name. This cookie would be set by the web server on the first request of a session. GVP then propagates the cookie through all dynamic requests for the session. If the load balancer were to use this identifier to establish and maintain persistence for a session, then the dynamic table would consist of only as many entries as there are web servers. The size of the table would not be an issue and aging of the table could be set at a maximum value. However, this method of implementation requires some maintenance at the web server or application level that is not required when using the unique session identifiers.

6.4. Benefits

There are additional benefits that can be realized from load balancing aside from the obvious benefit of having more evenly balanced load on the web servers. Health monitoring can be implemented to detect problems with the web servers. As described above, how the load balancer deals with any detected failures will depend on the persistence method implemented. Session persistence will provide the most efficient handling of failures. As well, without load balancing, GVP allows a primary and backup web server per application. Load Balancing can provide redundancy capabilities that are more flexible than are possible without its implementation. To eliminate a single point of failure however, more than one load balancer would need to be deployed.

7. Appendix A – Test Environment Description

7.1. Hardware and Software Environment

Please see the table below for the distribution of processes on the LSST lab hardware. For each machine there is a list of the software configured and the OS installed. Sections 7.2 and 7.3 in Appendix A provide detailed information about the machine and software versions respectively.

	(See Section 7.3 for software versions)	(See Section 7.2 for HW specs)
Host Name	Software Configuration	Operating System
Blackbird	1 IPCS	Windows 2003 Server SP1
Chengdu	1 IPCS	Windows 2003 Server SP1
Draken	1 IPCS	Windows 2003 Server SP1
Eagle	Microsoft IIS web server, Apache Tomcat web server, SQL Server (EMPS and Application dB)	Windows 2003 Server SP1
Falcon	EMPS	Windows 2003 Server SP1
Flanker	EMS: Policy Manager and Bandwidth Manager	Windows 2003 Server SP1
Fulcrum	IPCM: Resource Manager and SIP Session	Windows 2003 Server SP1
	Manager	
Gripen	ASR/TTS Server	Windows 2003 Server SP1
Hornet	Microsoft IIS web server, Apache Tomcat web server	Windows 2003 Server SP1
Nighthawk	StreamManager, CCAS	Windows 2003 Server SP1
Thunderbollt	StreamManager, CCAS	Windows 2003 Server SP1
Tomcat	Microsoft IIS web server, Apache Tomcat web server	Windows 2003 Server SP1
AppDirector	Software Version – 1.02.05DL	APSolute OS

7.2. Environment Hardware List

Description	Host Name	RAM	CPU	Hard Drives
Sun Sunfire V60x	Blackbird	2GB	Dual 3.0 GHz Xeon	Dual 73 GB
Sun Sunfire V60x	Chengdu	2GB	Dual 3.0 GHz Xeon	Dual 73 GB
Sun Sunfire V60x	Draken	2GB	Dual 3.0 GHz Xeon	Dual 73 GB
Sun Sunfire V60x	Eagle	2GB	Dual 3.0 GHz Xeon	73 GB
Sun Sunfire V60x	Falcon	2GB	Dual 3.0 GHz Xeon	Dual 73 GB
Sun Sunfire V60x	Flanker	2GB	Dual 3.0 GHz Xeon	73 GB
Sun Sunfire V60x	Fulcrum	2GB	Dual 3.0 GHz Xeon	Dual 73 GB
Sun Sunfire V60x	Gripen	2GB	Dual 3.0 GHz Xeon	Dual 73 GB
Sun Sunfire V60x	Hornet	2GB	Dual 3.0 GHz Xeon	73 GB
Sun Sunfire V60x	Nighthawk	2GB	Dual 3.0 GHz Xeon	73 GB
Sun Sunfire V60x	Thunderbolt	2GB	Dual 3.0 GHz Xeon	Dual 73 GB
Sun Sunfire V60x	Tomcat	2GB	Dual 3.0 GHz Xeon	Dual 73 GB
Application Server 2	AppDirector			

7.3. Software Versions List

Software	Windows 2003 Server SP1	APSolute OS
GVP Network Edition	7.2.001.00	-
Common Components	7.2.015.23	-
IPCS	7.2.015.20	-
Bandwidth Manager	7.2.015.09	-
EMPS	7.2.001.05	-
Policy Manager	7.2.015.04	-

GENESYS LSST AND TECHNICAL SUPPORT

WEB SERVER LOAD BALANCING

Scripts for LDAP	7.2.000.07	-
Resource Manager	7.2.015.12	-
SIP Session Manager	7.2.015.09	-
Text To Speech Server	7.2.015.04	-
Studio	7.2.012.10	-
Radware Application Switch 2		AppDirector 1.06.09DL(56)
Apache Tomcat	5.0	-
Microsoft Internet Information Services Manager	6.0	-
Nuance OSR		
Apache HTTP Server	1.3.33	-
Open Speech Recognizer	3.0.8	-
RealSpeak	4.0.8	-
SpeechWorks Media Server	3.1.9	-

8. Appendix B – Diagram of Test Environment

