Office Communications Server 2007 R2 Site Resiliency

Published: July 2009

Contents

Introduction 1

The Solution 2

Overview 2

Prerequisites 3

Test methodology 4

Test Topology 4

Enterprise Edition Pool (User) 5

Enterprise Edition pool (Director) 6

Edge Servers 6

Group Chat 6

Archiving and Monitoring Servers 6

File Server Cluster 7

Reverse Proxy 7

Hardware Load Balancers 7

WAN/SAN Latency Simulator 8

DNS 8

Expected Behavior 10

Normal operation 10

Failover 11

Failback 13

Test results 14

Datacenter link latency 14

Failover 14

Failback 15

Findings and Recommendations 15

Findings 15

Recommendations 16

Acknowledgments 16

References 17

Appendices 17

A. Scope of Testing 17

In-Scope Workloads 17

Out-of-Scope Workloads 17

In-Scope Server Roles 17
Out-of-Scope Server Roles 18
B. Stress Testing 18
C. Performance Monitoring Counters And Numbers 18
D. Two Nodes HLB Solution 19
E. Third-Party Vendor Configuration Details 19
HP 19
F5 20
Shunra 20

Introduction

The Backup and Restoration documentation for Microsoft Office Communications Server 2007 R2 (http://technet.microsoft.com/en-us/library/dd572319(office.13).aspx) includes guidelines and best practices for Microsoft Office Communications Server 2007 R2 disaster recovery. The basic recommendations in this Backup and Restoration documentation are as follows:

- Deploy a secondary site that mirrors the primary site.
- Move users from the failed pool on the primary site to the still-functioning pool on the secondary site.

However, executing backup and restore procedures and moving users from one pool to another, can entail some downtime. Customers who require Office Communications Server workloads to be always available have requested Microsoft support for a topology where the Microsoft Office Communications Server Enterprise pool spans two geographically separate locations. In such a topology, even catastrophic server failure in one location would not seriously disrupt usage, because all connection requests would automatically be directed to Front End Servers in the same pool but at the second location. The site resiliency solution described in this white paper specifically targets this split-pool topology and is supported by Microsoft subject to the constraints mentioned in the Findings and Recommendations section.

Unless specifically stated otherwise, all server roles have been installed according to the product documentation. For details, see Office Communications Server 2007 R2 product documentation at http://technet.microsoft.com/en-us/library/dd250572(office.13).aspx.

> *Note: If your failover requirements are not addressed in the product documentation or in this white paper, please give us your feedback at http://social.microsoft.com/Forums/en-US/communicationsserversetup/threads.*

This white paper is divided into three main sections:

- The Solution provides an overview of the tested and supported site resiliency solution described in this paper.
- Test Methodology describes the testing topology, expected behavior, and test results.
- Findings and Recommendations, provides practical guidance for deploying your own failover solution.

This white paper does not include specific procedures for deploying the products that are used in the solution. Specific deployment requirements are likely to vary so much among different customers that step-by-step instructions are likely to be incomplete or misleading.. If you need step-by-step instructions, see the product documentation for the various software and hardware used in this solution.

To successfully follow this paper, you should have a thorough understanding of Office Communications Server 2007 R2 and Windows Server 2008 Failover Clustering.

The Solution

This section describes the tested and supported failover solution, including prerequisites, topology, and individual components. For details about planning and deploying Windows Server 2008 and Office Communications Server 2007 R2, see the documentation for these products. For details about third-party components, see Third-Party Vendor Configuration Details at the end of this document and the product documentation from HP, F5, and Shunra for those components.

Overview

The solution described in this whitepaper entails the following:

- Splitting the Enterprise pool between two physical sites, hereafter called “North” and “South”
- Creating separate geoclusters (physically separated Windows Server 2008 failover clusters) for the following:
  - Enterprise pool that homes users
  - Enterprise pool acting as a Director
  - Group Chat Server pool
- Deploying a Windows Server 2008 file share witness to which all server clusters are connected
- Enabling synchronous data replication between the geoclusters.
- Mirroring all other components that are essential to the deployment, such as Archiving and Monitoring Servers and Edge Servers.

  ***Note**: Mirroring in this context means that some roles were deployed in both sites.*

Figure 1 provides an overview of the resulting topology.

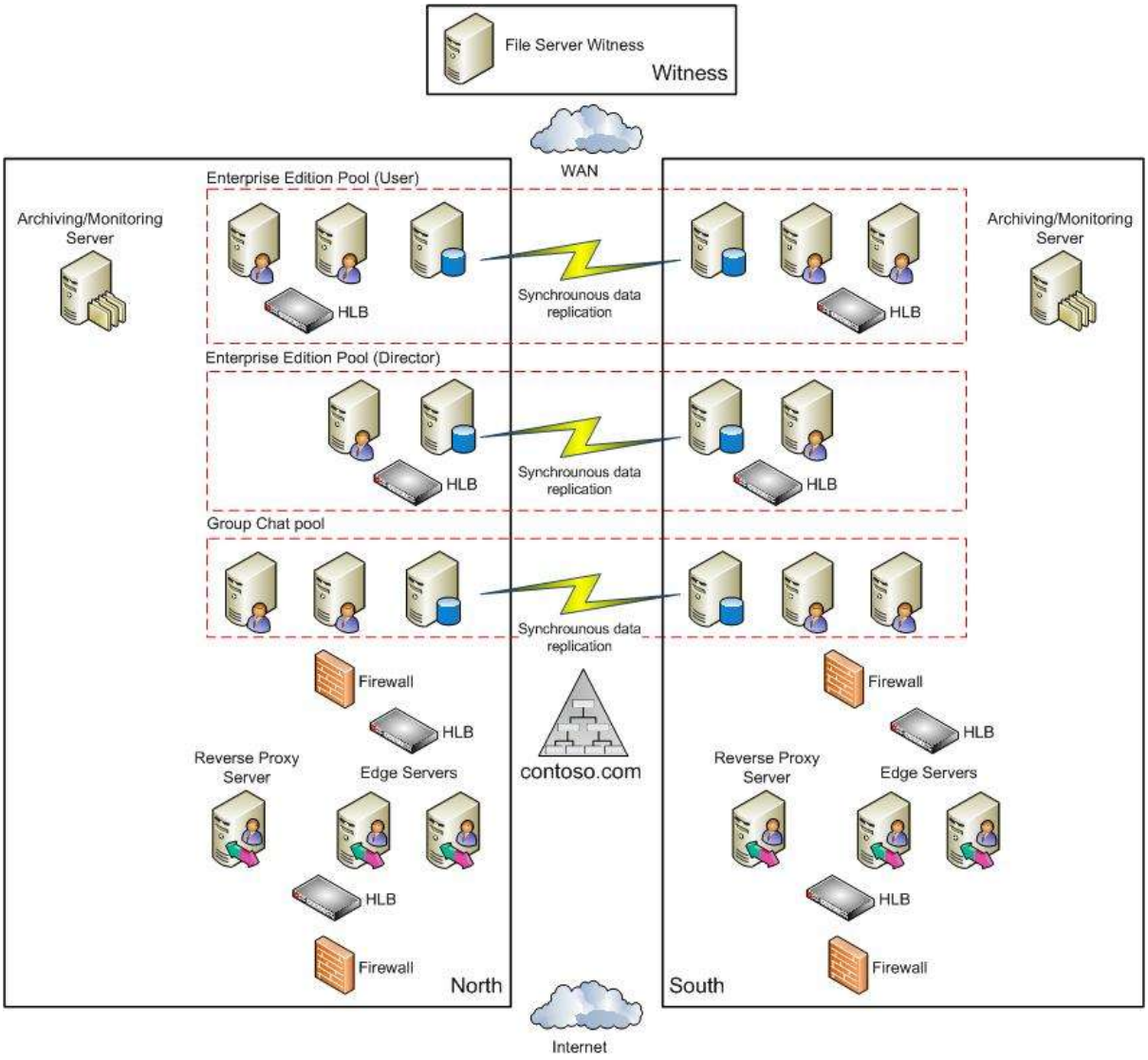

Figure 1: Supported Topology Overview

With the topology depicted in Figure 1, a single site could go down, for whatever reason, and users would be still able to access supported unified communications services within minutes rather than hours. For a detailed depiction of the topology used to test the solution described in this white paper, see Test Topology.

The solution depicted in Figure 1 has been tested and is supported by Microsoft for IM, presence, peer-to-peer, conferencing, and group chat. Other Office Communications Server 2007 R2 workloads and servers are out of scope. For a complete list of workloads and servers that are in scope and out of scope for this solution, see Appendix A, Scope of Testing.

Prerequisites

The solution described in this whitepaper assumes that your Office Communications Server 2007 R2 deployment meets both the core requirements described in the product documentation and all the prerequisites listed below.

- All internal servers must be part of the same stretched VLAN, using the same Layer-2 broadcast domain. Edge Servers must be in the perimeter network, and can be different VLAN from the internal one. Also, the perimeter network need not be stretched between sites.
- Synchronous data replication must be enabled between the primary and secondary sites, and the vendor solution that you employ must be supported by Microsoft.
- Round-trip latency between the two sites must not be greater than 15 ms.
- Available bandwidth between the sites must be at least 1 Gbps.
- A geoclustering solution based on Windows Server 2008 Failover Clustering must be in place. That solution must be certified and supported by Microsoft, and it must pass cluster validation as described in the Windows Server 2008 documentation (http://technet.microsoft.com/en-us/library/cc732035(WS.10).aspx#BKMK_what_validation).
- All geocluster servers must be running Windows Server 2008 64-Bit.
- All Office Communications Servers must be running Office Communications Server 2007 R2.
- All database servers must be running Microsoft SQL Server 2008 64-Bit.
- All servers in the topology must be running on physical computers. Virtualization, as described in http://www.microsoft.com/downloadS/details.aspx?familyid=0A45D921-3B48-44E4-B42B-19704A2B81B0&displaylang=en, is not supported.

To qualify for Microsoft support, your failover solution must meet all the above prerequisites.

Test methodology

The two major goals of our testing were as follows:

- Prove that failover and failback work as expected.
- Identify the maximum acceptable latency between the North and South sites before the user experience starts to deteriorate

We performed both functionality testing and limited-load testing.

Functionality testing means that a real person was sitting in front of the client computer and performed a series of tests while the system was under limited-stress load. Functionality testing allowed us to get pass/fail results and provided useful perspective on the whole user experience. System health was also monitored using performance monitor counters. This data was used to assess the results of the test.

Limited-stress testing means that we did not push the topology to its limits. We simulated 30,000 concurrent users accessing different resources in the topology. All stress testing used the Office Communications Server 2007 R2 Capacity Planning Tool (http://www.microsoft.com/downloads/details.aspx?FamilyID=F8CBDDD6-7608-4BBE-9246-16E96C62BEF4&displaylang=en). For details about how the Capacity Planning Tool was used in the tests, see Appendix B. Stress Testing.

Test Topology

Figure 2 shows the topology that was used to test the solution proposed in this white paper. This diagram will serve as a reference for the remainder of this paper.

The solution shown in Figure 2 has been created from “off the shelf” Microsoft products combined with third-party hardware and software. The solution does not require specific products from any particular vendor, so long as those products meet the prerequisites and requirements set forth in this whitepaper and supporting Microsoft product documentation. Depending on the mix of components you choose for your particular implementation of this solution, you might need help from your vendor of choice to deploy this solution.

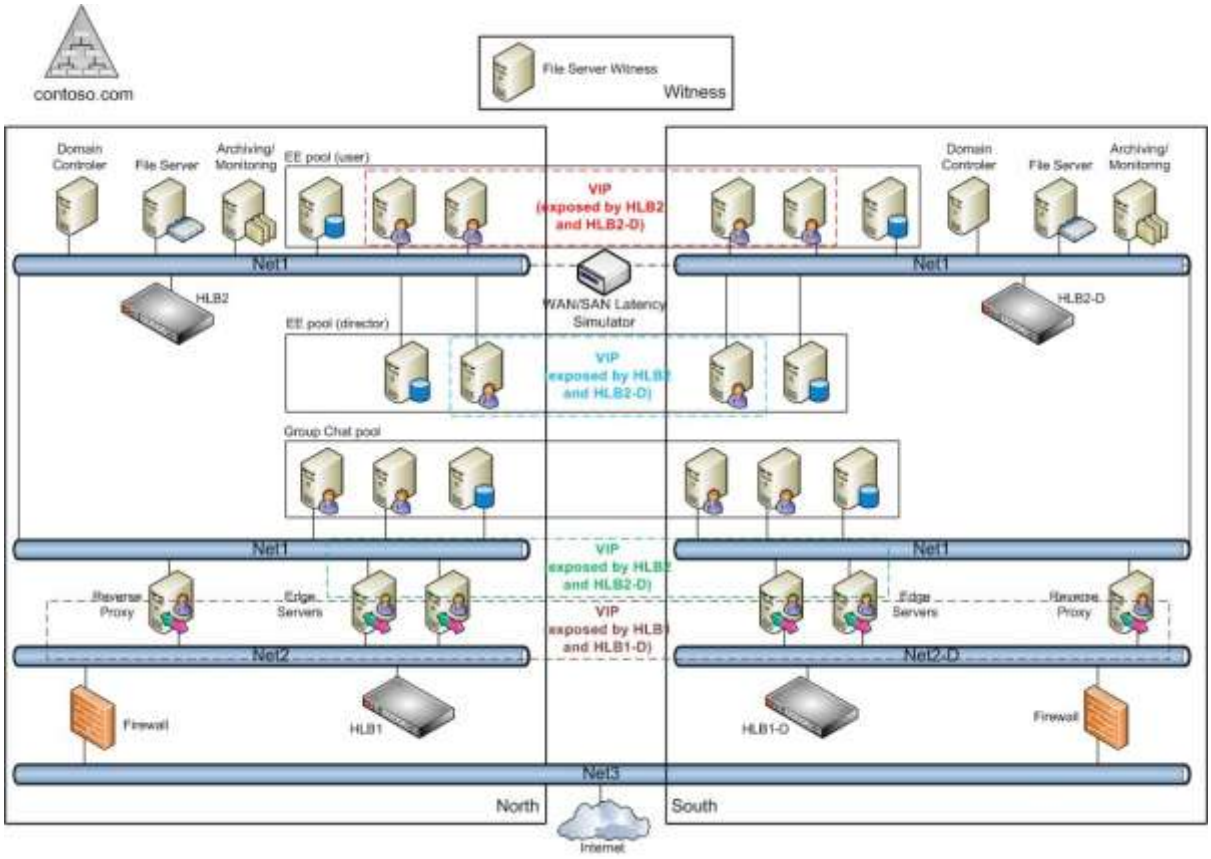

Figure 2: Supported Topology Details

As shown in Figure 2, the tested topology deployed two data centers, along with a third location that hosted a file server functioning as a Windows Server 2008 Failover Clustering Service File Share Witness[1]. The File Share Witness is accessible to all Windows Server 2008 Failover Cluster nodes in both data centers and is accessible to all cluster quorums. All Windows Server 2008 Failover Clusters used in this solution use the Node and File Share Majority quorum mode.

The remainder of this section discusses each of the solution components shown in Figure 2.

Enterprise Edition Pool (User)

This pool hosts all users in the environment that are enabled for Office Communication Server. We split the pool between the North and South sites. Each site included two identically configured Front End Servers. We deployed the back-end database as two Active/Passive SQL Server 2008

[1] For details about witness location, see the Windows Server 2008 Failover Clustering reference documentation in the Appendix.

geoclustering nodes running on top of the Windows Server 2008 Failover Clustering service. Synchronous data replication was required between the two Back-End Database Servers.

We fronted the pool with a redundant pair of hardware load balancers (HLBs). The HLBs communicate with each other and balance traffic across all Front End Servers, exposing supported server roles through Virtual IP addresses (VIPs). For a list of supported server roles, see In Scope Server Roles in Appendix A.

Enterprise Edition pool (Director)

We deployed two redundant Office Communications Server 2007 R2 Directors, one for each site. We disabled the following unnecessary server roles on each Director:

- Web Conferencing Server
- A/V Conferencing Server
- Unified Communications Application Server Components
- Address Book Service

We deployed the back-end database as two Active/Passive SQL Server 2008 geoclustering nodes running on top of Windows Server 2008 Failover Clustering.

Edge Servers

We installed Edge Servers with all server roles, but we tested them only for remote-user scenarios. (Federation and public IM connectivity are beyond the scope of this paper.) External network interfaces were exposed through external HLBs, and internal interfaces were exposed through internal HLBs. Each site has its own IP subnet; perimeter networks were not stretched across the North and South sites.

Group Chat

Each site hosts both a Channel service and a Lookup service, but these services can be active in only one of the sites at a time. The Channel service and the Lookup service in the other site must be stopped or disabled. In the event of site failover, manual intervention is required to start these services at the failover site.

Each site also hosts a Compliance Server, but only one of these servers can be active at a time. In the event of site failover and failback, manual intervention is required to restore the service. For details, see http://go.microsoft.com/fwlink/?LinkId=155785.

We deployed the back-end database as two Active/Passive SQL Server 2008 geoclustering nodes running on top of Windows Server 2008 Failover Clustering. Data replication between the two back-end database servers must be synchronous. A single database instance is used for both Group Chat and compliance data.

Archiving and Monitoring Servers

We deployed Archiving and Monitoring servers in a three-tier topology consisting of agent, service, and database.

Each site has a single server on which both the archiving and monitoring roles are collocated. All Archiving and Monitoring servers point to same back-end database instance.

Each Front End Server in the user pool points to its local Archiving and Monitoring server. For example, all Front End Servers in the North site point to the Archiving and Monitoring servers located in the North site. In the same way, all Front End Servers in South site point to the Archiving and Monitoring Servers located in the South site.

The back-end database has been deployed as two Active/Passive SQL Server 2008 geoclustering nodes running on top of the Windows Server 2008 Failover Clustering service. Data replication between the two back-end database servers must be synchronous.

The Archiving and Monitoring servers each used a separate database instance. One instance hosted the archiving database; a second instance hosted the monitoring database.

File Server Cluster

We deployed a file server as a two-node geoclustering resource on top of Windows Server 2008 Failover Clustering. Synchronous data replication was required.

All relevant Enterprise Edition pool and Group Chat pool data that require a file server were hosted on this file server cluster, including the following:

- Meeting content location
- Meeting metadata location
- Meeting archive location
- Address Book Server file store
- Application data store
- Client Update data store
- Group Chat compliance file repository
- Group Chat upload files location

Reverse Proxy

These were the only servers in the test topology that ran Windows Server 2003 32-Bit and Internet Security and Acceleration (ISA) Server 2006 Service Pack 1. Reverse proxies were used in remote-user access scenarios to publish different internal resources to the Web. External network interfaces were exposed through external HLBs.

Hardware Load Balancers

We deployed two HLBs at each site:

- An internal HLB to distribute traffic among various internal servers
- An external HLB to distribute traffic among Edge Servers and the reverse proxy

The two internal HLBs were configured as a redundant pair and were in constant communication with each other. The internal HLB on one site acted as the primary HLB, exposing the VIP for internal resources (anything connected to the internal network) on both sites. If the primary HLB were to go offline, the secondary HLB would take over and respond with the same VIP (same FQDN and IP address). This is possible because all internal resources in two separate sites were part of the same stretched VLAN in the same IP subnet.

The two external HLBs were also paired and were in constant communication with each other. The primary external HLB responded with VIP of the relevant service. If the primary HLB were to go offline, the secondary HLB would take over and respond with the same VIP (same FQDN but different IP address). Although the FQDN remains the same, the IP address changes because external network interfaces at each site are in different IP subnets.

Deploying four HLBs in the topology provided a clean separation of internal and external resources. Some customers prefer this separation for security reasons. This added security, however, has two disadvantages: more HLBs to buy and manage and somewhat more complex failover and failback procedures because those operations must happen with both load balancers in a pair. For example, if an internal HLB goes down, both internal and external HLBs (where the external is potentially still healthy) would have to failover and failback at the same time.

A two-node HLB solution was also tested and is fully supported. For details about a support topology that requires only two HLBs, see Appendix D. Two Nodes HLB Solution.

WAN/SAN Latency Simulator

In order to see impact of network latency between two sites, we deployed a network latency simulator. The simulator allowed us to test different latencies and come up with a recommendation for maximum acceptable and supported latency.

Besides testing network latency, we also wanted to test the impact of latency on storage (data) replication. In order to test storage latency, we connected two storage nodes (one at each site) by means of a fiber channel to the IP gateway. This connection enabled data replication over the IP network, which made it possible to use the network latency simulator to test latency along the data path.

> ***Note**: The WAN/SAN latency simulator was used for testing purposes only. The simulator is not a requirement for the solution described in this paper and is not required for Microsoft support.*

DNS

Our test topology used split DNS configuration; that is, the internal and external DNS namespaces were both called *contoso.com*, but each namespaces had a separate DNS designation. DNS was deployed according to Microsoft best practices. The only interesting and relevant details for this paper are delegation of VIPs to geo-redundant HLBs. Both internal and external DNS had delegation configured for *vip.contoso.com*. The following example shows how a DNS query worked for a sample VIP. All other VIPs were configured for DNS in the same way:

- Remote user joe@contoso.com starts the Office Communicator client. As a part of the sign-in process, Communicator queries the external DNS server for _sip._tls.contoso.com
- The authoritative DNS server for the contoso.com domain has been configured with following records
  - SRV record for _sip._tls.contoso.com points to ap.contoso.com.
  - CNAME record for ap.contoso.com points to ap.vip.contoso.com.
  - vip.contoso.com domain has been delegated to geo-redundant HLB.

- The geo-redundant HLB was authoritative for vip.contoso.com domain and configured with an A-record for ap.vip.contoso.com is a VIP for Access Edge Service
- Based on service availability, Communicator gets back the IP address of one of the VIPs.

  ***Note**: For internal services, the IP address would be always identical. For external services, the IP address could be different as IP subnets in two different sites are different.*

- Communicator uses the VIP address to connect to Access Edge service and ap.contoso.com for TLS authentication.

The following lists provide a configuration snapshot of both the internal and external DNS servers that were used in our testing.

External Windows DNS

- Authoritative zone is contoso.com.
- vip.contoso.com zone delegated to external geo-redundant HLB.
- CNAME for ap.contoso.com points to ap.vip.contoso.com (ap.contoso.com points to external network interface of Access Edge server).
- CNAME for webconf.contoso.com points to webconf.vip.contoso.com. (webconf.contoso.com points to external network interface of Web Conferencing Edge server.)
- CNAME for avedge.contoso.com points to avedge.vip.contoso.com. (avedge.contoso.com points to external network interface of A/V Edge server.)
- CNAME for proxy.contoso.com points to proxy.vip.contoso.com. (proxy.contoso.com points to external network interface of reverse HTTP proxy server.)

Internal Windows DNS

- Authoritative zone is contoso.com.
- internal.contoso.com zone delegated to internal geo-redundant HLB.
- CNAME for pool1.contoso.com points to pool1.internal.contoso.com. (pool1.contoso.com points to network interface of user pool.)
- CNAME for pool2.contoso.com points to pool2.internal.contoso.com. (pool2.contoso.com points to network interface of Director pool.)
- CNAME for internaledge.contoso.com points to internaledge.internal.contoso.com. (internaledge.contoso.com points to internal network interface of Edge Server.)

Expected Behavior

This section describes the behavior that we expected from our test topology.

Normal operation

When both sites are fully operational, Figure 3 shows a typical call flow. Note that the diagram has been simplified to highlight most important aspect of the topology.

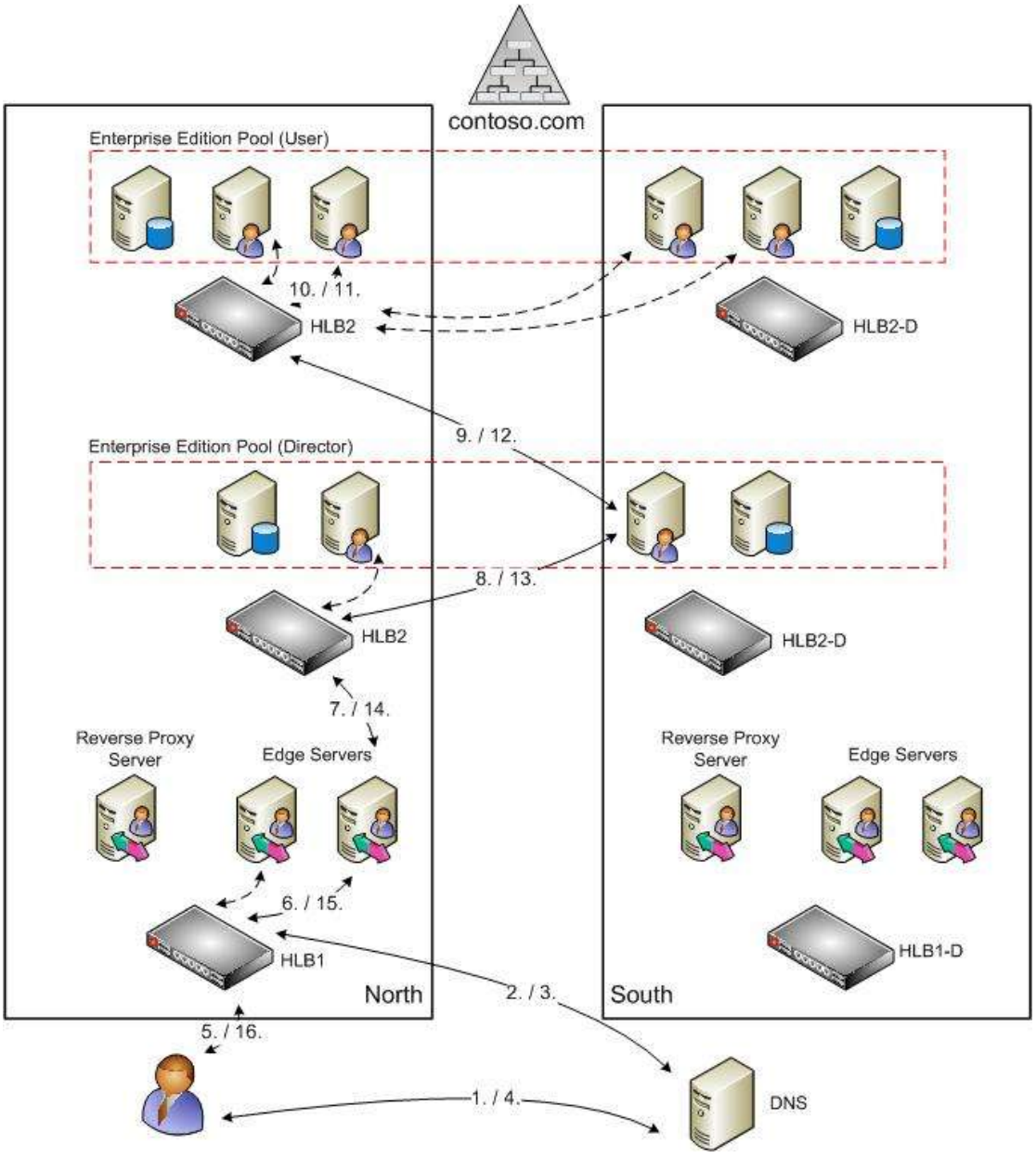

Figure 3: Normal operation. The numbers in the diagram correspond to the numbered steps in the following call-flow description.

1. Remote user joe@contoso.com signs in to Office Communicator. Communicator queries DNS server for its connection endpoint (the Edge Server in this specific instance). To do so, the client queries for several different SRV records. For purposes of this discussion, assume that the DNS server has the following characteristics:
    - Is authoritative for the contoso.com domain.
    - Has _sip._tls.contoso.com defined.

- Has _sip._tls.contoso.com SRV records defined that point to the ap.contoso.com CNAME record, which in turn points to ap.vip.contoso.com. (vip.contoso.com is delegated zone to geo-redundant HLB.)

2. The DNS server makes a recursive query to the geo-redundant HLB that is authoritative for vip.contoso.com. When everything is operational, the geo-redundant HLB is configured to return the VIP of the Access Edge in the North site.
3. The geo-redundant HLB returns the ap.vip.contoso.com A-record to DNS server with short TTL (30 seconds). A DNS server caches that information for subsequent request. TTL is honored so that subsequent requests are served either from DNS cache or by making a new recursive DNS query.
4. Communicator gets back the ap.vip.contoso.com A record and its corresponding IP address, as well as the ap.contoso.com CNAME record.
5. Communicator uses ap.vip.contoso.com to connect to the VIP of the North HLB.
6. Communicator connects by using TLS to one of the Edge servers at the North site. (The geo-redundant HLB will pick one Edge Server based on selected load-balancing logic.)

   Note: All dotted lines with arrows at the end in Figure 3 represent possible and alternative traffic flow.

7. The Edge Server forwards the request to the VIP of the Director. This VIP is managed by pair of internal geo-redundant HLBs, and under normal conditions the VIP will resolve to an IP address that is assigned to the geo-redundant HLB at the North site.

   Note: Figure 3 has been simplified for clarity. The internal geo-redundant HLB that is used to load-balance incoming Director traffic is also used to load-balance traffic to the internal Edge Server interfaces.

8. The geo-redundant HLB load-balances incoming request across all Directors. In this example the request is forwarded to the Director at the South site.
9. The Director determines the user's home pool and forwards the request to the VIP of the user's pool. This VIP is managed by pair of internal geo-redundant HLBs. Under normal conditions, the VIP will resolve to the IP address assigned to the geo-redundant HLB at the North site.
10. The geo-redundant HLB load-balances the incoming request across all Front End servers in the Enterprise Pool. In this example, the request is forwarded to a Front End server at the North site.
11. – The response is returned to Communicator.

Failover

Figure 4 shows typical call flow in the event the North site fails. Note that diagram has been simplified to highlight most important aspect of topology.

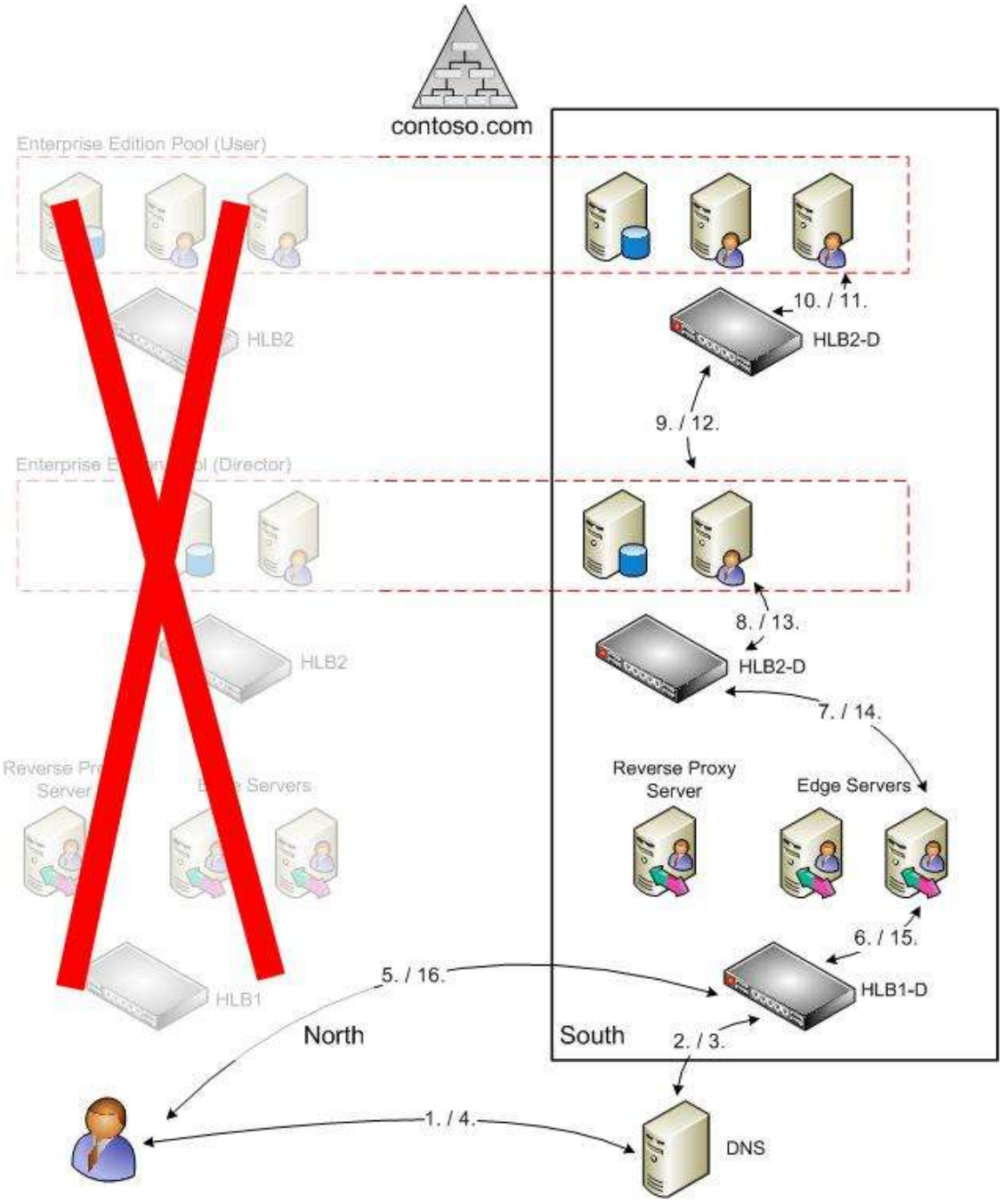

Figure 4: Call flow in the event of North site failure. The numbers in the diagram correspond to the numbered steps in the following call-flow description.

1. Remote user joe@contoso.com logs in to Communicator. As in the normal operation, Communicator queries the DNS server to find its connection endpoint. This time, because the North site is down, the call expires.

   ***Note**: If the DNS server has cached DNS information pointing to the North site, the client will attempt to use that information and will fail because the North site is down. To prevent such failure, set the TTL on the geo-redundant HLBs to be brief (for example, 30 seconds).*

2. The DNS server, which has DNS records for both North and South, makes a recursive query to the geo-redundant HLB that is authoritative for vip.contoso.com. Because the North site is not

reachable, the DNS server makes a recursive query to the geo-redundant HLB at the South site, which returns the VIP of a South Edge Server.

3. The South HLB returns the ap.vip.contoso.com A record to the DNS server with short TTL (30 seconds). The DNS server caches that information for subsequent requests. (The short TTL means that subsequent requests are served either from the DNS cache or by making a new recursive DNS query.)
4. Communicator gets back the ap.vip.contoso.com A record and its corresponding IP address. It also gets back the ap.contoso.com CNAME record.
5. Communicator uses ap.vip.contoso.com to connect to the VIP of the South HLB.
6. Communicator uses TLS to connect to one of the Edge Servers at the North site. (The geo-redundant HLB will pick one Edge Server based on selected load-balancing logic.) The Edge Server forwards the request to the VIP of the next-hop Director. Because the North site is down, this VIP resolves to an IP address assigned to the South HLB.
7. Because the North site is down, the South HLB load-balances the request only to the Director at the South site.
8. The Director receiving the request determines the user's home pool and forwards the request to the VIP of that pool. Because the North site is down, that VIP resolves to an IP address assigned to the South HLB .
9. Because the North site is down, the South HLB load balances the request only to Front End servers at the South site.
10. The response is returned to Communicator.

Failback

The goal of a failback operation is restoration of normal operation. Therefore, once failback is complete, users can once again connect to their home pool as shown in Figure 3. To failback and resume normal operation at the North site, the following steps are necessary:

1. Restore network connection between two sites. Quality attributes of the network connection (for example, bandwidth, latency, and loss) should be comparable to the quality prior to failover.
2. Ensure that geo-redundant HLBs at the South site can communicate with their redundant counterparts at the North site. Also, ensure that the HLBs at the North site resume their normal operation. (Reverting to normal operation should happen automatically after the connection between paired HLBs is restored.) Existing connections to the South site will continue, but new connections will go to the North side. Users will not be interrupted by future connections that fail back to the North HLB.
3. Resynchronize storage so data in North is in full sync with data in South.
4. Bring online all servers and relevant infrastructure in North. Depending on the severity of the North site's failure, it might be necessary to build everything from scratch. On the other hand, if North had suffered from, say, an extended power failure, all equipment would probably come online automatically (or at least under managed supervision). After the services are running, the geo-redundant HLBs will resume sending traffic to the Front End Servers on both sites. Existing connections will continue to the South side, and the load will eventually be redistributed evenly.
5. Failback server clusters from South to North so cluster resources are owned by servers at the North site. Only at this point might users be affected. If they try to do something new, such as publish presence or schedule a conference, the operation will fail for the duration of the failback, but users should remain logged on.

Test results

This section describes the results of Microsoft's testing of the failover solution proposed in this whitepaper.

Datacenter link latency

We used network latency simulator to introduce latency on the simulated WAN link between North and South.

10 ms - We established a baseline by introducing 10 ms round-trip latency into both the network path between two sites and the data path used for data replication between the two sites. The topology continued to operate without problem under these conditions and under load.

15 ms – Once we had established a baseline we began to increase latency. At 15 ms round-trip latency for both network and data traffic, the topology continued to operate without problem. 15 ms is the maximum supported round-trip latency for this topology.

> ***Important:*** *Microsoft will not support solutions whose network and data latency exceeds 15 ms.*

20 ms – At 20 ms round-trip latency, we started to see degradation in performance. In particular, message queues for archiving and monitoring databases started to grow. The Performance Monitor counter "LC:Usrv – 00 – DBStore\USrv – 002 – Queue Latency (msec)" also began to increase. A healthy server should have fewer than 100 ms DBStore queue latencies at steady state. As a result of these increased latencies, user experience also deteriorated. Sign-in time and conference creation time both increased, and the A/V experience degraded significantly. For these reasons, Microsoft does not support a solution where round-trip latency has exceeded 15 ms.

Failover

As previously mentioned, all Windows Server 2008 clusters in the topology used a Node and File Share Majority quorum. As a result, in order to simulate site failover, we had to isolate all servers and clusters by losing connectivity to both the South site and the witness site. That could be achieved either by disconnecting all remote network connectivity or by forcing a "dirty" shutdown of the North site. We tried both approaches. In the first test, we tried a "dirty" shutdown of all servers at the North site. In the second test, we disabled all remote network connectivity to the North site.

The Group Chat Channel service and Lookup service in the other site, which were normally stopped or disabled, had to be started manually.

Listed below are results and observations following failure of the North site:

- Users connected to the North site, if in a peer-to-peer call, were not dropped from the call because of media resiliency.
- Communicator users connected to North site were disconnected and automatically reconnected to South site.
- Live Meeting clients were disconnected. They were able to reconnect manually after failover was complete and the conference had rolled over to a Front End Server on the South site.
- In order to reconnect, Group Chat client users had to sign out and sign back in.

- Internal users connected to the South site were not signed out. New actions, such as new IM sessions or conferences, performed during the failover failed with appropriate errors, but no more errors occurred after the failover was complete.
- After the North site went offline, clusters in the South site came online in less than two minutes
- Site failover duration as observed in our testing was under one minute.

Failback

For the purposes of our testing, we defined failback as restoring all functionality to the North site such that users can reconnect to servers at that site. After the North site was restored, all cluster resources were moved back to their nodes at the North site. Listed below are results and observations following failback of the North site:

- Before cluster resources can be moved back to their nodes at the North site, storage had to be fully resynchronized. If storage has not been resynchronized, clusters will fail to come online.
- To ensure minimal user impact, the clusters were set not to automatically fail back. Our recommendation is to postpone failback until the next maintenance window after ensuring storage has fully resynchronized.
- By default, the North HLB will mark the North site servers as online as soon as the HLB Monitor (TCP connection) succeeds. The Front End Servers will come online once they are able to connect to Active Directory and the back-end database. After the Front End Servers are online and the North HLB marks them as up, new connections will be routed to them. No existing connections to the South Front End Servers will be dropped. If you want to prevent the Front End Servers at the North site from automatically coming back online—for example, if you want better control over the whole process or if latency between the two sites has not been restored to acceptable levels—we recommend either shutting down the Front End Servers or marking the Front End Servers as offline on the geo-redundant HLBs at both the North and South sites.
- After the HLBs have reestablished connection, North HLBs will start handing out VIPs. All new connections will be directed to North HLBs, which will begin distributing the load across internal servers at both sites. Existing connections will not be dropped on the South HLBs.
- Site failback duration as observed in our testing was under one minute.

Findings and Recommendations

The failover solution described in this whitepaper has been tested and is officially supported by Microsoft; however, before deploying this solution, it is important to consider the following findings and recommendations:

Findings

- Cluster failover worked as expected. No manual steps were required. Front Ends were able to reconnect to the back-end after the failover and resume normal service. Office Communicator clients reconnected automatically. Live Meeting clients required manual reconnection, which is expected and by design.
- Cluster failback worked as expected. Bringing the Front End Servers online in the North site did not cause any interruption in service; however, although failback of the cluster to the primary site will not sign out connected users, it may result in errors during the brief failover period.It is therefore crucial to ensure that storage has resynchronized before failback begins.

-
- When failover occurred, the Group Chat Channel service Lookup service at the failover site had to be started manually. Additionally, the Group Chat Compliance Server setting had to be updated manually. For details, see “Configuring Compliance” at http://go.microsoft.com/fwlink/?LinkId=155785.

Recommendations

- Although testing used two nodes (one per site) in each SQL Server cluster, Microsoft recommends deploying additional nodes to achieve in-site redundancy for all components in the topology. For example, if the active SQL Server node goes down, a backup SQL Server node in the same site and part of the same cluster can assume the workload until the failed server is brought back on line or replaced.
- Although testing used components provided by certain third-party vendors, the solution does not depend on or stipulate any particular vendors. As long as components are certified and supported by Microsoft, any qualifying vendor will do.
- All individual components of the solution (for example, geoclustering components) must be supported and, where appropriate, certified by Microsoft. This does not mean, however, that Microsoft will directly support individual third-party components. For component support, contact the appropriate third-party vendor.
- Although a full-scale deployment was not tested, we expect published scale numbers for Office Communications Server 2007 R2 to hold true. With that in mind, you should plan for enough capacity that sufficient capacity remains to continue operation in the event of failover.
- This white paper should be used only as guidance. Before deploying this solution in a production environment, you should build and test it using your own topology.

Note

Microsoft does not support implementations of this solution where network and data-replication latency between the primary and secondary sites exceeds 15 ms. When latency exceeds 15 ms, the end-user experience rapidly deteriorates. In addition, archiving and Group Chat compliance servers are likely to start falling behind, which may in turn cause Front End Servers and Group Chat lookup servers to shut down.

Acknowledgments

We would like to thank our partners without whom this whitepaper would not exist today:

- Microsoft Enterprise Engineering Center (http://www.microsoft.com/eec) for providing facilities, equipment and know-how.
- F5 (http://www.f5.com) for providing Hardware Load Balancers and help configuring them.
- Hewlett-Packard Development Company (http://www.hp.com/go/clxeva) for providing the geoclustering solution.
- Shunra Software Ltd. (http://www.shunra.com) for providing network latency simulator.

References

- For details about Windows Server 2008 Failover Clustering, see http://www.microsoft.com/Windowsserver2008/en/us/failover-clustering-main.aspx
- To ask questions and provide feedback, use the Office Communications Server forums at http://social.microsoft.com/Forums/en-US/category/officecommunicationsserver
- To learn more about Microsoft Office Communications Server 2007 R2, go to http://technet.microsoft.com/en-us/library/dd250572(office.13).aspx
- To learn more about th eWindows Server 2008 Failover Cluster Configuration Program, go to http://www.microsoft.com/windowsserver2008/en/us/failover-clustering-program-partners.aspx
- To learn more about SQL Server Always On partners, go to http://www.microsoft.com/sqlserver/2008/en/us/high-availability.aspx#sqlalwaysonpartners.

Appendices

A. Scope of Testing

The following lists specify which Office Communications Server 2007 R2 workloads and server roles were either in-scope or out-of-scope for testing the failover solution described in this whitepaper. Out-of-scope workloads and server roles are not supported.

In-Scope Workloads

- IM and Presence
- Peer to peer scenarios; for example, peer-to-peer audio/video sessions
- IM Conferencing
- Web Conferencing
- A/V Conferencing
- Application Sharing
- Group Chat

Out-of-Scope Workloads

- Remote Call Control
- Enterprise Voice and Telephony Integration
- Communicator Web Access
- Unified Communications Application Server Components
  - Conferencing Attendant
  - Conferencing Announcement Service
  - Outside Voice Control
  - Response Group Service
- Federation and public IM connectivity

In-Scope Server Roles

- Front End server roles when deployed in an Enterprise Edition Consolidated Pool
  - IM Conferencing Server
  - A/V Conferencing Server

    - Web Conferencing Server
    - Application Sharing Server
    - Web Services (address book download and DL expansion)
- Back-End Database Server in Enterprise Edition Consolidated Pool
- Edge Server
    - Access Edge service
    - Web Conferencing Edge service
    - A/V Conferencing Edge service
- Monitoring and Archiving Servers
- Group Chat Servers

Out-of-Scope Server Roles

- Mediation Server
- Unified Communications Application Server Applications
- Communicator Web Access Server
- Web Services (update server)

B. Stress Testing

All stress testing was done using Office Communication Server 2007 R2 Capacity Planning Tools. Stress testing assumed:

- 30,000 concurrent users :
    - 30,000 were configured for IM/Presence
    - 2,000 were configured for A/V
    - 200 were configured for Application Sharing

While under load, each Front End Server (four in total) hosted the following:

- 5 concurrent application-sharing conferences
- 100 concurrent A/V conferences

For details about using Office Communications Server 2007 R2 Capacity Planning Tools, see http://www.microsoft.com/downloads/details.aspx?FamilyID=F8CBDDD6-7608-4BBE-9246-16E96C62BEF4&displaylang=en.

C. Performance Monitoring Counters And Numbers

To ensure the quality of your failover solution, we recommend that you monitor the following performance statistics:

- **On Front End Servers, monitor the "LC:USrv – 00 – DBStore\Usrv – 002 – Queue Latency (msec)" counter.** This counter represents the time that a request spends in the queue to the Back-End Database Server. If the topology is healthy, this counter averages below 100 ms. Occasional spikes are acceptable. The value will be higher on Front End Servers that are located at the site opposite the location of the Back-End Database Servers. This counter can increase if the Back-End Database Server is having performance problems or if network latency is too high. If this counter is high, check both network latency and the health of the Back-End Database Server.

- **On Front End, Archiving and Monitoring servers, monitor the "MSMQ Service\Total Messages in all Queues" counter.** The size of the queue will vary depending on load. Verify that the queue is not increasing unbounded. Establish a baseline for the counter, and monitor the counter to ensure that it does not exceed that baseline.
- **On Group Chat Channel and Compliance servers, monitor the "MSMQ Service\Total Messages in all Queues" counter.** The size of the queue will vary depending on load. Verify that the queue is not increasing unbounded. Establish a baseline for the counter, and monitor the counter to make sure that it does not exceed that baseline.
- **On the Directors, Edge servers, and Front End Servers, monitor the "LC:SIP – 04 – Responses object\ SIP – 051 – Local 503 Responses/sec" counter.** This counter indicates if any server is returning errors indicating that the server is unavailable. At steady state, this counter should be ~0. Occasional spikes are acceptable.
- **On all servers monitor the "LC:SIP – 04 – Responses \SIP – 053 – Local 504 Responses/sec" counter.** This counter can indicate connection delays or failures with other servers. At steady state, this counter should be approximately 0. Occasional spikes are acceptable. If you see 504 error messages, check the "LC:SIP – 01 – Peers\SIP – 017 - Sends Outstanding" counter. This counter records the number of requests and responses in the outbound queue, which will indicate which servers are having problems.

D. Two Nodes HLB Solution

This solution is an alternative to four-node HLB solution that is the subject of this document.

From a logical point of view, the two-node and four-node solution are identical. If each HLB can be connected to a different subnet and has enough capacity to serve both internal and external requests, a single HLB can replace the two HLBs at each site described for the four-node solution. All load balancing logic remains the same.

Like the four-node solution, the two-node solution has both advantages and disadvantages. The two node solution is less complex and less expensive to deploy and manage, but some customers may not want a single HLB exposed to both internal and external network traffic.

> ***Note:*** *Both four and two nodes HLB solutions have been successfully tested and are supported.*

E. Third-Party Vendor Configuration Details

Microsoft does not stipulate any particular third-party vendors for the purpose of implementing the solution described in this paper. However, in order to perform our testing of this solution at the Enterprise Engineering Center, we used hardware supplied by HP, F5, and Shunra.

> ***Note:*** *The following descriptions of vendor components used for testing the solution described in this paper are included to provide complete technical information. These descriptions do not constitute either an endorsement of the listed vendors or their products, or a requirement, explicit or otherwise, that their products be used.*

HP

In order to implement a geographically dispersed Windows Server 2008 Failover Clustering solution, we have used two HP StorageWorks Enterprise Virtual Array (EVA) Disk Enclosure (one per site) as

database storage. Storage was carved into disk groups, which in turn were associated with their respective clusters. All disk groups utilized synchronous data replication. HP StorageWorks Cluster Extension EVA was used as Windows Server 2008 Failover Clustering resource to facilitate storage failover and failback.

One of the scenarios we wanted to test was the impact of latency on storage data replication between two sites. One problem we encountered was that HP StorageWorks has fiber channel interfaces but the Shunra network latency simulator does not support those interfaces. In order to connect the two, we used a Fiber Channel to IP gateway that HP provided.

F5

We used F5 GTM-enabled HLBs for testing both the four-HLB and two-HLB failover solutions described in this paper.

Shunra

We used Shunra latency simulator to test network and data-replication latency between the two sites. That allowed us to introduce different latencies and come up with recommendation for maximum acceptable and supported latency.