Microsoft Office Communications Server 2007 R2 (OCS R2) is designed for enhanced instant messaging, presence, voice and video, conferencing, e-mail, voice mail, and fax communications from mobile workstations. It uses packet-switched communication through Session Initiation Protocol (SIP) and Secure Real-Time Transport Protocol (SRTP) over a data network, and can also integrate with private branch exchange (PBX) phone systems.

In large production deployments of OCS R2, ensuring high availability and scalability can be critical to success. Best practices recommend using hardware load balancing to help meet this requirement for the front-end server pools, directors, Microsoft Office Communicator Web Access (CWA) servers, and edge servers. This article provides an example architecture, configuration details, and best practices for using F5 BIG-IP Local Traffic Manager (LTM) systems to implement hardware load balancing in large OCS R2 deployments.

EXAMPLE MICROSOFT OCS R2 LOAD-BALANCING ARCHITECTURE

Figure 1 shows the primary components of an example Microsoft OCS R2 Enterprise Edition architecture incorporating F5 BIG-IP LTM hardware load balancing. In this architecture, client systems that log on from the internal network register with the front-end servers; these servers also handle instant messaging and call routing along with a number of other colocated services such as OCS R2 application services, audio/video conferencing, and Web conferencing. The CWA servers enable end users to access Communicator features using a supported Web browser over a secure channel. Load-balanced edge servers direct SIP traffic to the director array. A director within that array then authenticates requests made from the external network and routes traffic to the front-end servers.

The BIG-IP LTM load balancers in this architecture help distribute application traffic, such as SIP and HTTP over Secure Sockets Layer (HTTPS), across multiple servers. Load balancing can help increase application capacity, fault tolerance, and overall performance by decreasing the server overhead associated with managing application connections, traffic optimizations, and encryption offload.

F5 BIG-IP LTM CONFIGURATION

Load balancing a Microsoft OCS R2 deployment requires a variety of F5 BIG-IP LTM features, including virtual servers, profiles, Secure Network Address Translation (SNAT), and pools. Administrators can access these features through the BIG-IP LTM management interface using a Web browser over an HTTPS connection.
OCS R2 hardware load balancing requires that administrators allow certain ports and protocols to be load balanced on the BIG-IP LTM systems. For example, the front-end servers, directors, and edge servers must be enabled for TCP or Mutual Transport Layer Security (MTLS) traffic through ports 5060 and 5061, respectively.

Health monitors are used to check the availability of OCS R2 services in the server pools. For example, the health monitors for ports 443, 3478, 5061, and 5062 must be configured for the edge servers. Following the creation of the health monitors, pools are created and assigned with health monitors. A separate health monitor and pool are created for each service port.

The final features to set up are the virtual servers and TCP profiles. TCP profiles allow administrators to customize the TCP idle time-out for each virtual server. The virtual server is essential for signaling using SIP and for Web traffic. For example, the CWA virtual server uses port 443 for load balancing Web traffic from the CWA clients. After the virtual server is configured, administrators must manually update the Domain Name System (DNS) to associate the BIG-IP LTM virtual IP address with the pool’s fully qualified domain name (FQDN) specified in the OCS R2 deployment steps.

Figure 2 shows two Communicator clients sending requests to the front-end server pool; the first client is load balanced to front-end server 1, while the second, which logs on shortly afterward, is forwarded to front-end server 2 using the Least Connections (node) load-balancing method. In this example, the SNAT address of 192.168.1.16 is used for communication with the front-end server pool.

**BEST PRACTICES FOR MICROSOFT OCS R2 DEPLOYMENTS**

Adhering to recommended best practices can help ensure a successful deployment of Microsoft OCS R2 with F5 BIG-IP LTM hardware load balancing:

- Use the BIG-IP LTM high-availability mode to provide application fault tolerance for OCS R2. This mode requires two BIG-IP LTM systems and additional
configuration to complete the setup, which can include Media Access Control (MAC) masquerading, configuration synchronization, and connection mirroring.

- Use the Least Connections (node) load-balancing method, which specifies that the BIG-IP LTM systems should pass new connections to the node that has the fewest number of current connections. Load-balancing calculations can be localized to each pool (a member-based calculation) or can apply to all pools of which a server is a member (a node-based calculation).

- Use BIG-IP LTM Secure Sockets Layer (SSL) acceleration to offload the HTTPS encryption processing from the CWA servers and help reduce processor utilization. Administrators can enable SSL acceleration by first enabling the CWA servers to process HTTP (instead of HTTPS), and then loading a certificate on the BIG-IP LTM system and associating that certificate with the SSL profile.

- For CWA servers, use the recommended TCP idle time-out value of 1,800 seconds; for front-end server load balancing, use a value of 1,200 seconds. This value specifies the number of seconds that a TCP connection to a virtual server can remain inactive before the load balancer closes the connection.

- For CWA servers, use the Least Connections (node) load-balancing method and HTTP cookie insert persistence, which helps ensure that when a client is load balanced to a pool member, future connections from that client are directed to the same member, and also allows session affinity between the clients and servers. Simple or SSL session ID persistence can be used to help ensure that SIP over Transport Layer Security (TLS) connections to a front-end server maintain affinity.

- Use a SNAT pool when handling more than 65,000 simultaneous connections. A SNAT pool allows the grouping of SNAT addresses, each of which can accommodate only 65,000 connections. SNAT is required for load balancing front-end server pools in OCS R2 Enterprise Edition, and can map multiple CWA or Communicator client IP addresses to a translation address defined on a BIG-IP LTM system.

**Figure 2. Load-balancing configuration for two Microsoft Office Communicator clients**

**ROBUST LOAD BALANCING FOR MICROSOFT OCS R2**

For large deployments of Microsoft OCS R2, using hardware load balancers can help maintain high availability and scalability for the front-end server pools, directors, CWA servers, and edge servers. Using the robust features of F5 BIG-IP LTM systems and adhering to recommended best practices for configuring and deploying those systems can help organizations successfully implement hardware load balancing in enterprise environments.

Fred Johnson is a strategic partner engineer with F5 Networks dedicated to Dell Labs, and works closely with Dell development and services groups on proof-of-concept designs, benchmarks, technical publications, and training involving Dell, F5, and third-party software products. He has 20 years of experience in IT and has a B.A. in Psychology from the University of Texas at Austin.

Akshai Parthasarathy is a systems engineer working on Dell Unified Communications and Microsoft Exchange solutions as part of the Dell End-to-End Solutions team. His current interests include performance characterization and sizing for these two Dell solutions. Akshai has a bachelor’s degree and a master’s degree in Electrical and Computer Engineering from the Georgia Institute of Technology.