How Network Load Balancing Works

Network Load Balancing (NLB) is a service that runs on each member of an NLB cluster. NLB is typically bound to a single network adapter in each member and load balances traffic received by that adapter. NLB ignores traffic sent to the adapter's private IP address. Instead, it instructs the adapter's driver to respond to an additional virtual MAC address created internally by NLB. NLB associates that MAC address with any IP addresses assigned to the cluster.

Therefore, when traffic is sent to a cluster IP address, the ARP protocol resolves the IP address to the virtual NLB MAC address. The traffic is placed on the network with that MAC address as its destination, and the traffic is received by all cluster members. As described in the main text, NLB simply decides which member will respond to the request based on its load balancing and affinity rules.

It's not unusual for multiple members to respond to what might seem to be a single request. For example, when a Web browser loads a Web page, it opens multiple connections. One connection retrieves the HTML for the page, while additional connections retrieve graphics and other page elements. These requests are sent in parallel and can be handled in parallel by multiple cluster members.

NLB does not communicate between the cluster members at all times. Instead, it builds an algorithm table that allows each member to independently determine which requests to handle. This enables the members to act independently for maximum performance. Cluster members do send a heartbeat signal to one another, effectively letting one another know that all's well.

Members can change status—such as failing or being taken offline for maintenance—from time to time, so NLB periodically performs a process known as *convergence*. Convergence always occurs whenever a cluster member's heartbeat fails; it also happens whenever a member is added to, or removed from, the cluster. In the convergence process, NLB uses a multicast transmission to contact each cluster member. Note that this occurs on each cluster member, meaning that each member is briefly contacting every other member. Convergence enables the cluster to rebuild its load-balancing table and typically takes less than 10 seconds. While convergence is underway, members continue to respond to incoming requests based on the old load-balancing table.

In Windows Server 2003, NLB supports the creation of *virtual clusters*, which are clusters that span multiple disparate sets of physical members. For example, suppose you have four servers, named A, B, C, and D. You also have three Web applications you want to cluster with NLB, and the applications are named 1, 2, and 3. You can create a virtual cluster for application 1 on servers A, B, and C; one for application 2 on servers B, C, and D; and one for application 3 on servers A, C, and D. The servers distinguish between the virtual clusters through their unique IP addresses.