WILS: Network Load Balancing versus Application Load Balancing

Lori MacVittie, 2009-15-09

Are you load balancing servers or applications? Network traffic or application requests? If your strategy to application availability is network-based you might need a change in direction (up the stack).

Can you see the application now?

Network load balancing is the distribution of traffic based on network variables, such as IP address and destination ports. It is layer 4 (TCP) and below and is not designed to take into consideration anything at the application layer such as content type, cookie data, custom headers, user location, or the application behavior. It is *context-less*, caring only about the network-layer information contained within the packets it is directing this way and that.

Application load balancing is the distribution of requests based on multiple variables, from the network layer to the application layer. It is *context-aware* and can direct requests based on any single variable as easily as it can a combination of variables. Applications are load balanced based on their peculiar behavior and not solely on server (operating system or virtualization layer) information.

The difference between the two is important because network load balancing cannot assure availability of the *application.* This is because it bases its decisions solely on network and TCP-layer variables and has no awareness of the application at all. Generally a network load balancer will determine "availability" based on the ability of a server to respond to ICMP ping, or to correctly complete the three-way TCP handshake. An application load balancer goes much deeper, and is capable of determining availability based on not only a successful HTTP GET of a particular page but also the verification that the *content* is as was expected based on the input parameters.

This is also important to note when considering the deployment of multiple applications on the same host sharing IP addresses (virtual hosts in old skool speak). A network load balancer will not differentiate between Application A and Application B when checking availability (indeed it cannot unless ports are different) but an application load balancer will differentiate between the two applications by examining the application layer data available to it. This difference means that a network load balancer may end up sending requests to an application that has crashed or is offline, but an application load balancer will never make that same mistake.

WILS: *WRITE IT LIKE SETH*. SETH GODIN ALWAYS GETS HIS POINT ACROSS WITH BREVITY AND WIT. WILS IS AN ATTEMPT TO BE CONCISE ABOUT APPLICATION DELIVERY TOPICS AND JUST GET STRAIGHT TO THE POINT. NO DILLY DALLYING AROUND.

F5 Networks, Inc. | 401 Elliot Avenue West, Seattle, WA 98119 | 888-882-4447 | f5.com

F5 Networks, Inc.	F5 Networks	F5 Networks Ltd.	F5 Networks
Corporate Headquarters	Asia-Pacific	Europe/Middle-East/Africa	Japan K.K.
info@f5.com	apacinfo@f5.com	emeainfo@f5.com	f5j-info@f5.com