F5 White Paper

Inband Passive Monitors—Maintain Application Performance and Health

Application health monitors are now a tried and true technology of the Application Delivery Controller, yet traditional monitors require interaction with the application. Inband passive monitors change that requirement and monitor applications transparently.

by Alan Murphy and Ken Salchow
Technical Marketing

Introduction

One of the negative effects of the continued evolution of the load balancer into today's Application Delivery Controller (ADC) is that it is often too easy to forget the basic problem for which these devices where originally created—delivering highly available, scalable, and reliable application services. We get too preoccupied with intelligent application routing, virtualized application services, and shared infrastructure deployments to remember that none of these things are possible without a firm foundation of basic load balancing technology. For example, it is necessary to have a solid base in load balancing to monitor the health of the application servers and to identify and locate when and where there is a problem. Application health monitoring, or the ability to verify that back-end systems are operational **before** traffic is routed to those systems, is a basic yet critical tenant of load balancing and more sophisticated ADCs. If the ADC is unaware of when an application is malfunctioning or unavailable, its ability to route traffic to the best possible destination becomes impaired.

Like almost all aspects of Application Delivery Networking, health monitoring has been plagued by intelligence versus performance issues since day one, although incorrectly so—monitoring the status and availability of an application by the ADC shouldn't impact the performance of either the application or the ADC.

Application Health Monitoring

An Historical View

The original health monitor for back-end applications, still used by many products, is a simple network test—an ICMP ping of the server hosting the application. While this can certainly communicate that the application server is receiving network traffic—the absence of which is a definite sign that the application or the network is unavailable—it doesn't reveal anything about the actual application state of that server. A successful ping response denotes an answer from a particular IP address, but not what is running on that address, how well it is running (if the application is working correctly), or if there is an application handing traffic. For example, Microsoft Windows NT Server is notorious for responding to a ping even though the system itself has "blue screened" and no real application is running or capable of processing those network packets. Ping, along with other network testing tools, is a useful tool for checking the network and verifying something is on the receiving end, but not much beyond that.

The next iteration of the health monitor is the migration from an IP-based (layer 3) health monitor to a TCP (layer 4) health monitor. Instead of relying on lower layer responses, these health checks attempt to interact with the TCP port (by completing a full TCP handshake) associated with the application to verify that a connection can be made, signifying that an application is running and available for users. A typical example of this is an attempt to attach to TCP port 80 of a web server. A successful connection to the appropriate port is a far better indicator than a simple network ping that an application is actually listening to the port on the server. While this instills confidence that the application is answering, available, and able to validate and differentiate multiple application requests on a single server, a positive response is still not a definitive indication that the application is capable of handling user traffic. A successful TCP handshake to a web server doesn't reveal if the web server is actually returning HTTP content, for example, only that something is listening and answering on port 80.

The last major iteration of health monitoring is the application health monitor—a monitor that is capable of interacting with and interpreting responses from the application itself. These monitors do more than just connect to the application port and complete a handshake: they interact with the applications directly. An HTTP monitor, for instance, not only connects to the HTTP server but also evaluates the response from the HTTP server, either by verifying the response code (such as "200 OK") or by requesting a specific piece of content with a known, expected response (parse the response page and look for a particular text string or image). These monitors also solve other problems like being able to differentiate between multiple websites that are partitioned and hosted off of the same web server. By adding application awareness to the health monitor, the ADC can now verify that the application server accepting user traffic was capable of receiving, processing, and replying to real traffic.

Just like the basic load balancer has now become an ADC, health monitors evolved from being network-centric to being application-aware, applying more intelligence along the way.

The Impact of Inband Health Monitoring

Application-aware health monitors can also have a downside: a performance hit. Much like the impact of adding intelligence to load balancing decisions—moving up the stack from layer 3 to layer 7—adding intelligence to health monitoring can negatively impact the applications being monitored and their ability to handle normal, live application traffic. Using application layer health monitors increases

the number of connections and transactions that the applications need to process. Every "200 OK" response from the web server is still a request and response that has to be handled by the server, either to the monitor or to a real client.

By negatively impacting the applications they were designed to monitor, health checks can begin to chew up significant amounts of resources. Indiscriminate use of sufficiently advanced application health monitoring can actually cause unintended network and application failures. From this comes a need to manage health monitoring activities and balance between the benefits of application awareness and the side effects of increased use. In many cases, it becomes a decision to abandon intelligent health monitoring altogether in lieu of less intelligent, but more optimized TCP or network health monitoring. Losing the benefit of intelligent application health monitoring due to the starving application resources is counterintuitive to why intelligent monitors are chosen and used in the first place.

Monitoring Without Starving the App

In keeping with the spirit of the F5® TMOS™ architecture, the first intelligent, full-proxy platform capable of performing at network line speed, F5 has created a solution that addresses the "intelligence" versus "performance" conundrum in health monitors—inband passive monitors. Inband passive monitoring enables new levels of confidence in the availability and delivery of applications without adding the additional overhead of traditional health monitoring. Inband passive monitoring does this through application awareness that has minimal impact on the application.

Because F5 BIG-IP® Local Traffic Manager™ (LTM) leverages the full-proxy software architecture of TMOS, it is capable of not only using real-world, bi-directional application data to intelligently monitor services and route application traffic, but also of examining the real-world application responses as an indication of application failures. For example, if a user request for content made to a website behind BIG-IP LTM results in an error message—such as a "503 Service Unavailable" error—BIG-IP LTM can immediately determine if that application is unavailable and not capable of processing future application requests. BIG-IP LTM can mark that specific service on that particular server (or node) as unavailable. It can then route the failed request and all new application traffic to a service that is available, and invoke an active monitor to probe for current status of the unavailable service. In this way, BIG-IP LTM can determine the true state of the application without adding any further overhead. Only when a service does something unexpected and fails to process real traffic are the active monitors

processed—and then, only for the malfunctioning system, not the dozens of systems that are operating as expected.

Inband passive monitors aren't binary, "yes" or "no" traffic gatekeepers; they support multiple levels of thresholding, enabling a configurable grace period of failures before an application service is marked as unavailable. The inband monitors track the number of failures incrementally and evaluate the current failure rate against a configurable failure interval. If the number of failures in a given time period exceeds the number of allowable failures, BIG-IP LTM marks that service as unavailable and routes traffic to an available service. In addition to failures, inband monitors can also act based on response time. If a service does not respond to a request within the defined time, this is treated as a failure and that value is incremented and monitoring continues. Inband passive monitors enable flexibility without impacting the application; they provide a non-intrusive method for routing traffic based on application awareness. They enable holistic application monitoring without forcing a choice between intelligence and performance.

Conclusion

Health monitors are a critical component of both basic load balancers and more advanced ADCs. Even as they have evolved and moved up the network and application stacks, health monitors have always gathered and provided the intelligence necessary for an ADC to appropriately route real-time application traffic to the best possible destination. While active application-aware monitors serve their purpose, they do carry with them the possible downside of over-burdening the application. Inband passive monitors finally fuse together intelligence and performance; it's no longer an "or" decision, now it becomes an "and," enabling both application intelligence and performance to co-exist. By taking advantage of managing all bi-directional traffic, inband monitors give BIG-IP customers the best of both worlds: complete application awareness and status without burdening the application.

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

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

WP-Inband Passive Monitors 02/09