savvius

WHITE PAPER

NPMD Requires the Four Golden Signals of Network Health

Imagine that you're at a football game and you're curious: are any of your friends also at the game? That is a hard question to answer even though you know exactly how many people are in attendance. You've got binoculars, but it would take too long to scan the audience for a familiar face. You could hope to see your friend's face flash across the Jumbotron, but that isn't likely. Sending a group text to everyone you know would be disturbing a lot of people (and letting them know you're playing hooky!) You probably end up living with the irony that you may be sitting in close proximity to someone you would want to see, but will never know who or where they are sitting.

Unfortunately, this situation is familiar to IT professionals charged with monitoring the performance of today's enterprise networks. There are things you want to know, but the tools that allow you to engage with the amount of traffic on today's networks are either too general, so anything you learn turns into an investigation, or too detailed, so you have to know exactly where to look, often because someone complains. You need something in between. You need actionable visibility.

The Solution

What matters to you about network performance? For an answer, we look to the Google SRE (Site Reliability Engineering) team and their formulation of the four "Golden Signals" for network health: latency, traffic, errors, and saturation. If you know these measures about your network traffic, you have a pretty good idea of the user experience. You have visibility.

The Four Golden Signals

Latency describes the time it takes to service a request. High latency leads to unresponsive systems and generates more user complaints than any other performance issue. But a single latency measure isn't enough: both network latency and application latency must be measured for complete visibility (and to reduce finger-pointing between the network and application teams!) Furthermore, those latency measures must weigh a number of factors correctly. For example, an HTTP 500 error that has been triggered due to loss of connection to a database or other critical backend might be served very quickly; however, because this type of error indicates a failed request, factoring 500 errors into overall latency may result in misleading calculations. Alternatively, a slow error is even worse than a fast error. Therefore, it's important to track error latency, as opposed to just filtering out errors.

Traffic is simply a measure of how much demand is being placed on the system. Understanding traffic at an aggregate level is useful. Knowing how various components of the traffic contribute to overall traffic can be vital. RTP spiking during a sporting event is quite different than Citrix ICA dominating traffic at the end of the quarter.

Errors are a measure of quality. Poor quality transactions generally have high error rates and can be an early indicator of other issues. Until you know that errors are occurring, you cannot know why error rates are occurring, yet that may be exactly what you need to know. While some quality measures are well known, such as MOS scores for VoIP, other quality indicators are equally important, such as TCP/IP quality, based on the kinds of errors most likely to affect the user experience or indicate a systemic issue. Assembling errors into a measure of quality requires more than just counts. Algorithms that appropriately weigh errors by significance and through time are the only way to have a quality score that you can trust.

Saturation represents a measure of how "full" your network is, combining traffic with capacity. Once you know traffic levels across your entire network or in a particular segment, you can easily determine overall saturation. But effective capacity, and therefore saturation, varies by the kind of traffic. For example, VoIP degrades in performance well before the network reaches 100 percent utilization. And as the network approaches a saturation point, errors can be introduced which themselves add more traffic, leading to escalating issues. Increases in latency and decreases in quality are often leading indicators.

Actionable Visibility

The Golden Signals provide visibility into network performance, but they are not automatically actionable. Applying them to your entire network and getting a summary may tell you overall health, but that doesn't help you solve (or anticipate) specific problems. You may realize that there was a spike in a certain kind of traffic that led to an overall increase in latency, but what, exactly, happened and where, exactly, did it happen? Time to investigate!

The opposite, drilling down and applying these signals to a specific interaction on the network, may be very useful in a particular case, but it doesn't go beyond solving that one problem. How can we best make our visibility immediately useful?

The answer starts with a technology that has been around for decades: flows. In the 1990s, Cisco created what they called NetFlow to assist router operation. It turns out that assembling packets into specific interactions is a useful higher-level way of understanding network traffic. As a result, many network performance tools today are based either on NetFlow, or one of its successors such as IPFIX and sFlow.

The problem is that these flow descriptions don't provide enough information to measure, or even contain, all of the Golden Signals. Plus, they are often generated by the network infrastructure itself, which introduces flaws that impact data accuracy and reliability. We need flows with a difference; flows created from network traffic itself that have been enhanced at their creation with quantifiable measurements that can be applied to all four signals.

To be actionable, information must be current and complete. That leads us to the definitive method for gaining actionable visibility into network performance: continuously apply the four Golden Signals to every flow in the network, in real time. When networks can have over a million active flows at any single moment, this isn't an easy task. However, clever architecture combined with the abundant amount of compute power available today makes it achievable.

The results are impressive. When every flow on the network is evaluated using the four Golden Signals, simply ranking them from worst to best highlights areas of interest. IT professionals can navigate directly to an area of concern within seconds of it being identified. Each flow contains enough specific information to enable the network professional to take further action or initiate an investigation. In all cases, the ability to see problem areas in real time dramatically accelerates the process to reach a resolution.

Conclusion

Those responsible for network performance have a broad set of tasks: troubleshoot issues, anticipate problems, optimize Quality of Experience and other usability metrics, assist in audit compliance based on established use policies, help isolate the cause and location of security issues, and be equipped for capacity planning. Those tasks are increasingly difficult as budgets tighten, network traffic increases, and tools rooted in previous generation technology remain in use. By utilizing the four Golden Signals of latency, traffic, errors, and saturation on all network traffic in real time, the IT team can achieve unparalleled visibility into the performance of the network and quickly gain the information they need to solve problems and maintain reliability. This is an enormous step toward solving one of the hardest problems in network operations; enabling users to take reliable, high-performance network functioning entirely for granted.