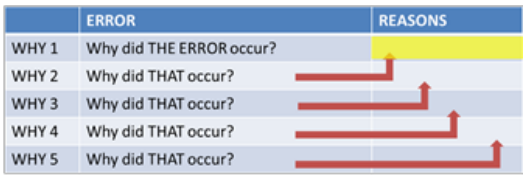
Visibility: Keystone to Rapid Recovery Everywhere

Lori MacVittie, 2012-29-10

#caim Because knowing is half the battle in application performance management



If you can't even identify why #1 you are unlikely able to find the root cause

One of the processes used to determine the root cause of a problem by Six Sigma is called The Five Whys. In a nutshell, it's a method of continually asking "why" something happened to eventually arrive at an answer that is actually the real root cause.

This method is probably derived from philosophy and its study of causality, something most often associated with Aristotle and the desire to trace all of existence back to a "first cause".

The Five Whys doesn't require that level of philosophical debate, but it does require digging deeper at every step. It's not enough to answer "Why did you run out of gas" with "Because I failed to fill the car up this morning" because there's a reason you failed to do so, and perhaps a reason for that reason, and so on. At some point you will find the root cause, a cause that if not addressed will only end up with the same result some time in the future.

Thus, questions like "Why did the web application become available" can't simply be answered with "because the database stopped responding" because that isn't the root cause of the actual incident. The actual root cause might end up being a failed NIC on the database server, or a misconfiguration in a switch that caused responses from the database to be routed somewhere else. Simply knowing "the database stopped responding" isn't enough to definitely resolve the *real* problem.

Thus, the ability to find the real problem, the root cause, is paramount to successfully not only resolving an incident but also preventing it from happening in the future, if possible. Finding it fast is essential to keeping the associated costs under control.

> The study, completed in 2011, uncovered a number of key findings related to the cost of downtime. Based on cost estimates provided by survey respondents, the average cost of data center downtime was approximately $5,600 per minute.
>
> Based on **an average reported incident length of 90 minutes**, the average cost of a single downtime event was approximately $505,500.
>
> -- Understanding the Cost of Data Center Downtime: An Analysis of the Financial Impact on Infrastructure Vulnerability

Unfortunately, when an outage or incident occurs, you likely don't have to time to get philosophical on the problem and sit around asking questions. You still have to ***ask*** the same questions, but generally speaking it's a lost more fast and furious than one might expect Aristotle was used to. One of the most helpful tools in the IT toolbox to being able to rapidly walk through the five (or fifteen in some cases) whys is monitoring and the alerts/information that they often provide.

VISIBILITY is NUMBER ONE

There are any number of ways in which visibility is achieved. Agents are an age-old method of achieving visibility across tiers of architecture and are becoming just as commonplace for solving visibility challenges in cloud environments where use of other mechanisms might not be feasible. Monitoring up and down the full stack is vital to ensuring availability of all services and is used for a variety of purposes across the data center. You may recall that Axiom #1 for application delivery is: "Applications are not servers, hypervisors, or operating systems." Monitoring of the OS tells you very little about the application's performance or availability, and it can't indicate any kind of status with respect to whether or not the application is actually working correctly.

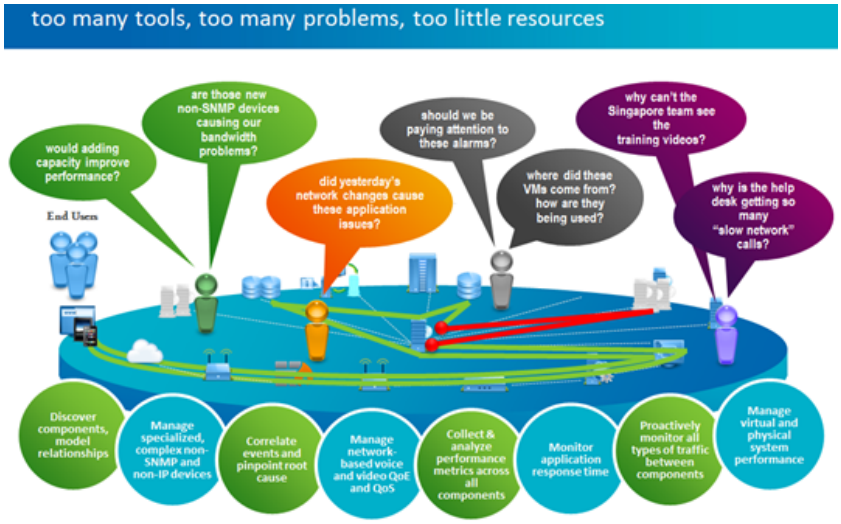
For example, load balancing and application delivery services can monitor applications for availability, for correctness, and for performance. doing so enables the use of more robust load distribution algorithms as well as more immediate response to failure. This information is also critical for cloud delivery brokers which must be able to decide which environment is best able to meet requirements for a given application. Without performance and availability data, there's no way to compare two environments against what are very basic SLA metrics used by business and operations alike. And with increasing devices and applications and always-connected applications, there's more and more traffic to be analyzed and more possible points of failure in the application delivery chain.

> Through 2015 at least 50% of hyperconverged solutions will suffer from poor performance, end-user dissatisfaction and failure to meet business needs.
>
> -- New Innovations – New Pain Points (CA + Gartner)

And while some infrastructure makes use of monitoring data in real-time for operational decision making, other solutions use the information to manage performance and resolve problems. Most APM (Application Performance Management) solutions have matured beyond aggregating data and presenting aesthetically appealing dashboards to providing comprehensive tools to monitor, manage, and investigate performance-related issues.

CA Technologies has been playing in this game for as long as I've been in technology and recently announced its latest initiative aimed at improving the ability to aggregate and make use of the massive volume of monitoring data for maximum impact.

CA: Converged Infrastructure Management



The latest version of CA Infrastructure Management (IM) adds some significant enhancements around scalability and support for cloud environments with multi-tenancy. Offering both agent (required for cloud environments) and agent-less options, CA IM is about to aggregate performance data from data center infrastructure – including non-IP devices (its past acquisitions such as that of Torokina have provided alternative technology to enable data collection from a broader set of devices).

But what's important about a modern solution is not just the ability to collect from everything but to aggregate and allow rapid analysis across network and application domains. This is the kind of flow analysis that enables operations to very quickly cycle through the "Five Whys" and get to the root cause of a problem – and resolve it. This kind of analysis is as important if not more important than the data collected in the first place. The sheer volume of information collected by devices and systems across environments is simply overwhelming and one could not hope to manually browse through it in a reasonable amount of time to find any given issue.

CA IM is heavily invested in the notion of convergence; convergence of functional domains across environments, convergence of networks to include voice, video, and data, and convergence of the analytics required by operations to determine baselines, calculate trends, establish thresholds, recognize anomalies and trigger investigations.

A primary focus for CA remains end-to-end application response time monitoring, enabling operations to understand application response time between infrastructure devices. That's increasingly important in a virtualized data center, where some 80% of traffic is estimated to be between servers (both physical and virtual) on the LAN. East-west traffic patterns are dominating the network, and being able to monitor them and understand the relationship it has on end-user performance is critical. Being able to identify choke points between services that commonly rely on one another for proper functioning, capacity and network planning activities is an important part of not only ensuring the network is optimally configured but that provisioning of capacity is appropriate for the application and its performance requirements.

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

F5 Networks, Inc.
Corporate Headquarters
info@f5.com

F5 Networks
Asia-Pacific
apacinfo@f5.com

F5 Networks Ltd.
Europe/Middle-East/Africa
emeainfo@f5.com

F5 Networks
Japan K.K.
f5j-info@f5.com