A Recommended New Approach to Improve Digital Application Performance

EMA™

*IT & DATA MANAGEMENT RESEARCH,
INDUSTRY ANALYSIS & CONSULTING*

Table of Contents

Executive Summary

Online applications and services, from ecommerce to government services, increasingly address a global audience. These services require a globally-distributed architecture that is optimized to deliver a high-quality user experience. Traffic optimization solutions based on managed DNS services are a critical component of such distributed architecture. They offer an efficient method for optimizing globally-distributed architectures for both end-user performance and cost.

When these optimization services are able to make traffic management decisions based on real-time performance analytics, they become immensely more valuable and effective. They allow application owners to build traffic optimization policies that are based on real-time network conditions, rather than purely static conditions like geography. Combining real-time performance data with cost metrics enables application services to be optimized for performance and cost.

This white paper examines why operators of online services should leverage managed traffic optimization services that use a combination of real-time data feeds, including real user measurements (RUM), active and passive monitoring, and location-based information. It briefly reviews NS1's Pulsar RUM analytics and Managed DNS traffic management platform.

Performance is Essential to Online Services

Performance can have a significant impact on the success of an application. End users have little tolerance for slow page loads, given the highly competitive nature of online services. If one ecommerce site is slow to return product search results, a user will often click to a competitor. Thus, isolated incidents of latency can directly hurt the bottom line of a business. For example, EMA research found that 51 percent of digital performance management professionals see severe impacts on revenue generation when the Internet performs poorly in service of their critical applications.[1] A user will abandon a search engine because search results don't load fast enough. Ecommerce sites will lose revenue because online transactions won't close. Website conversions are lost when a site doesn't load quickly.

Today's complex application architectures can be highly sensitive to network latency. Given the transactional nature of network communications and web applications, 50 milliseconds of latency can add whole seconds to page load times in a browser.

Even the best coding and content delivery practices cannot overcome the effects of last-mile network performance, intermittent network congestion, and capacity issues. If an application isn't coded properly for a distributed environment, the cascading effect of latency can be even more problematic. The uncertain performance of the Internet and content delivery infrastructure can result in subpar performance for many users, while appearing to perform well on average. However, when a well-provisioned application delivery infrastructure is combined with effective traffic optimization, it can deliver improved performance for all users.

Online Services Require Highly Distributed Application Architectures

Sensitivity to latency is tied directly to the fact that online services leverage highly distributed application architectures to serve a global audience. Because these applications must perform well across multiple geographies, they require multi-cloud and hybrid cloud architectures, with points of presence (PoPs) as close to end-user populations as possible. A software as a service provider often serves its application from dozens of PoPs across the globe, while also employing multiple CDNs to optimize application delivery in various regions.

---

[1] All research in this paper is taken from EMA, "User, Customer, and Digital Experience: Where Service and Business Performance Come Together," March 2017.

These applications typically rely on the public Internet to connect application architectures across multiple global PoPs, which adds to the performance problem. According to EMA research, 85 percent of enterprises say cloud technology, with its inherent reliance on the Internet, made digital experience management essential to their organizations. The Internet is a shared resource with countless first-mile and last-mile providers that vary in performance. Unlike private networks, the Internet offers no service-level agreements. To manage and optimize digital performance, enterprises deploy a combination of CDNs, DNS routing services, and other technologies to manage traffic between end users and distributed applications.

Distributed Application Architectures Have Inherent Latency Budgets

A complex, distributed application architecture is extremely sensitive to latency. The speed of light inevitably adds a minimum amount of latency at every link in the distributed application chain, so there is an inherent latency budget that application owners must manage.

Application owners must optimize for geography, but they must also manage against any other sources of latency. A single problem in the application architecture can break the latency budget. For instance, a cloud provider could hit a peak period, driving up response time for a web application hosted there. A preferred CDN provider might lack cache capacity in a specific geography. A last-mile Internet service provider might experience sudden congestion. A DNS authority might fail to route traffic to the closest PoP. It is extremely difficult to monitor all of these variables in real time. Most enterprises lack the tools to provide the end-to-end visibility needed to assure performance in these situations.

EMA research found that digital experience management professionals are highly aware of this latency budget. When managing against latency, they optimize against multiple layers of metrics, most often including response time per transaction, application availability, response time per application, response time for multiple transactions to support a complete business service, and general browser responsiveness.
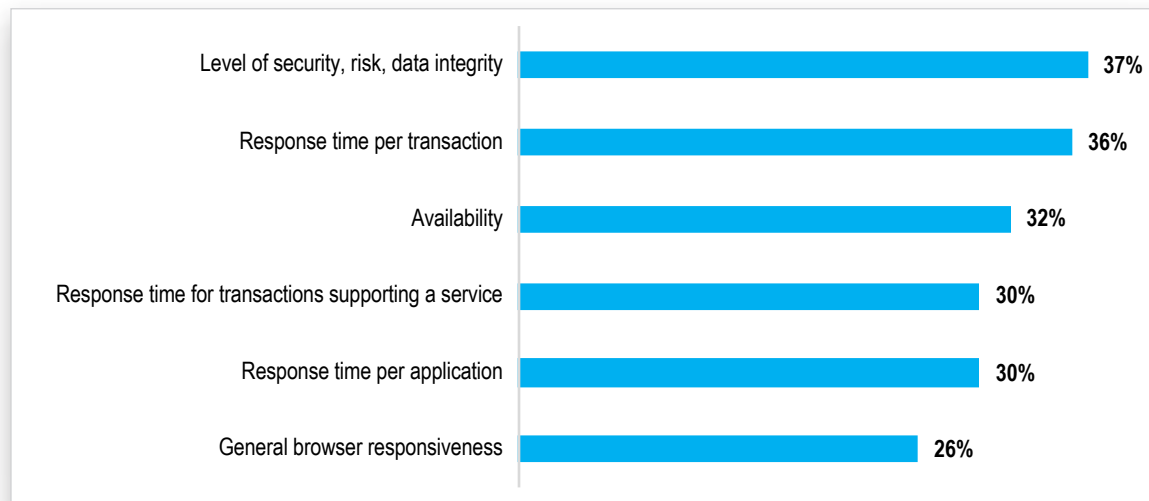

Figure 1. Top Technical Metrics Used for Digital Experience Management

Optimization Strategies for Distributed Applications

The Case for Real-Time Optimization

Distributed applications typically use shared resources, including the public Internet, public clouds, and CDNs. The quality of those resources often varies by geography. They can be volatile, leading to unpredictable levels of availability and performance. They are sensitive to congestion, resource contention, and suboptimal routing.

Real user measurements can track this volatility and help optimization solutions make informed traffic management decisions. In this context, granularity of performance measurements can be extremely important, too. For instance, performance measures based on the average performance over a 24-hour period will not account for daily peak usage and other volatile bursts of activity.

The Cost Optimization Opportunity

Real-time monitoring can also look beyond performance. With the right analytics, traffic management solutions can optimize networks based on transit costs and contract commitments. For instance, cost-related external SLAs and business activity metrics are the two most popular business metrics applied to digital experience management. These metrics can be wrapped into real-time monitoring and analytics to optimize costs. A high-cost CDN may not always be the performance leader in a given geography. Or a low-cost CDN provider might provide the minimum level of performance required to live within an application's latency budget.

EMA research found that organizations also apply user-oriented metrics to digital experience management, most often user productivity (number of processes executed) and user effectiveness (success versus failure ratio). Application owners can combine these metrics with performance monitoring to create business policies that balance cost optimization with end-user experiences. With the right analytics and policies, the infrastructure team can deliver a level of user experience that is cost-effective and high-performing.

The Value of Managed DNS Services with Real-Time Performance Analytics

Many operators of distributed, online services outsource their DNS to managed DNS services. Most of these managed DNS services offer some level of traffic routing that improve application performance. Typically, these capabilities optimize traffic based on geography (geolocation), which routes user traffic to the geographically closest PoP in the distributed application architecture. This basic routing function can reduce latency to the cloud, enterprise data centers, and CDNs, but because network (especially on the Internet) and data center conditions are variable, geolocation is often a poor surrogate for traffic optimization.

EMA research also found that enterprises often struggle to understand the performance impact of third-party services, such as CDNs or cloud providers. In fact, EMA research found that a transaction monitoring capability offered by cloud-based DNS services is the top monitoring metric for digital experience management professionals. DNS services that can incorporate real-time performance insight into their traffic management capabilities will deliver immense value to online service operators. For instance, the DNS service can route user traffic to the cloud data center that is performing best in a given geography, rather than just the closest one.
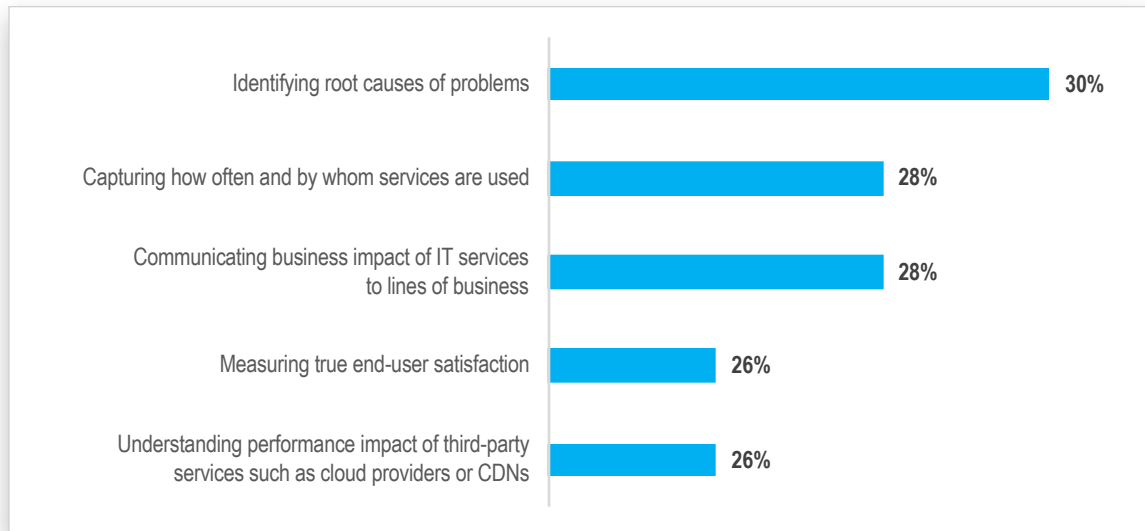
Figure 2. Top Challenges for Digital Experience Management

Pulsar: Powering NS1's Traffic Optimization Services with Real User Measurements

Pulsar is a real user measurement (RUM) solution that is integrated with NS1's Managed DNS service. Pulsar collects and analyzes billions of RUM data from user browsers to track real-time latency between users and CDNs, cloud facilities, and data centers. It uses this data to map the performance of online services from every eyeball network where users are located to the PoPs where online services are hosted. This provides NS1's DNS with the intelligence to factor in latency when choosing an answer. This is analogous to a modern GPS, which has the real-time intelligence to direct drivers to the least congested routes to their destinations.

The NS1 platform combines Pulsar with active and passive synthetic monitoring feeds, geographic data, and business logic to provide complete, optimized traffic management. This 360-degree view of global infrastructure enables NS1 to provide a managed DNS service that routes traffic based on availability, performance, and business logic. This approach gives enterprises the tools to deliver superior application performance while avoiding excess costs by meeting CDN commitments and avoiding overages.

Pulsar also helps enterprises implement and optimize multi-CDN strategies. It can identify which CDNs are performing best at any point in time. Given the variability in CDN performance based on time of day and geography, Pulsar provides an important capability to keep user connections from exceeding the latency budget of the application. Pulsar's analytics engine allows operators of online services to build policies that apply this infrastructure insight in real-time. Thus, they can optimize their applications for performance and cost.

EMA Perspective

Many businesses offer applications and online services that depend on complex distributed architectures. They include government services, online education programs, ecommerce, business-to-business supply chain and logistics organizations, and global media companies. Operators and owners of these globally and regionally distributed applications and online services rely on complex chains of services for application delivery, such as multi-cloud architecture, the public Internet, and CDN. These services vary in quality by time of day, geography, and other factors.

Infrastructure and application architects must build a global application infrastructure based on these components that can deliver optimal performance to users wherever in the world they reside. These architectures need traffic management solutions that can optimize user and application traffic, manage latency across the application delivery chain, and track end-user experiences.

Whether building a new distributed application or optimizing a legacy service, enterprises should start by identifying the real-time performance monitoring capabilities of traffic optimization solutions. They should ask these optimization service providers about the granularity and real-time nature of their visibility into networks and infrastructure.

Enterprises should determine whether traffic optimization services can track real user metrics through browser-based instrumentation. It is also important to determine whether such a service can use real-time analytics to optimize costs in addition to performance. Given the highly competitive nature of online services, performance is critical to business success. All of these capabilities can help the operators of online services deliver a high-quality experience to end users and protect their businesses' bottom line.

About NS1

NS1 is the leader in next-generation DNS solutions that orchestrate the delivery of the world's most critical Internet and enterprise applications. Only NS1's purpose-built platform, which is built on a modern API-first architecture, transforms DNS into an intelligent, efficient, and automated system, driving dramatic gains in reliability, resiliency, security, and performance of application delivery infrastructure. Many of the highest-trafficked sites and largest global enterprises trust NS1, including Salesforce, LinkedIn, Dropbox, Nielsen, Squarespace, Pandora, and *The Guardian*. Visit **www.ns1.com** to learn more.