Meeting Service Level Agreements Using Web Infrastructure Stressing Appliances

John Kenney, PhD

Application Brief
June, 2001

caw networks

Summary

This application brief describes how the unique and powerful technologies employed in the WebAvalanche and WebReflector will help IT service providers attain the highest levels of application-level quality of service from their network infrastructure. Before the introduction of WebAvalanche and WebReflector, credible assessment of application performance has been hindered by an inability to simulate high loads of realistic web traffic and apply those loads to a service provider's network. IT service providers can now validate their SLAs with an extremely powerful yet easy to use appliance form-factor solution as these revolutionary new products can meet the challenge of increasingly complex and higher capacity web-stressing tests.

The key benefits of conducting thorough and high-capacity Web stressing are discussed in detail. WebAvalanche is generating a new set of metrics based upon tests that simulate the highest levels of realistic network behavior during peak loads of Internet traffic. These new and improved metrics will provide substantial credibility to newly written SLAs as they go several steps beyond the hardware MTBF measurement that has been used to measure predicted network hardware uptime.

The WebAvalanche and WebReflector solutions are all-in-one appliance hardware designs that will save tremendous amounts of time and money as they effectively replace racks of costly testing hardware (such as workstations and server farms). They also substantially reduce costly hardware/software set-up time by employing a standard Web browser interface to build, run and generate detailed test reports thereby eliminating the requirement of custom programming and/or writing scripts to create and run the tests.

Best testing practices dictate that capacity stress testing be deployed prior to going live and that the highest load levels be generated along with the highest levels of Internet realism in a test environment to identify bugs and flaws in the network infrastructure and application software. WebAvalanche and WebReflector are the first network appliance testing solutions that can help IT service providers truly assess the capacity of network devices and topologies at layers 4-7.

Document 702-000000, V2.0.3

Caw Networks
67 East Evelyn Avenue
Mountain View, California 94041 USA
Phone: 650.961.7000 · Fax: 650.961.2769
EMail: info@caw.com

Table of Contents

1 A New Generation of SLAs

Customers demand that IT service providers take responsibility for the quality of services they provide. Service Level Agreements (SLAs) developed in the 1990's specify a certain level of network availability and performance. These metrics focused on network availability and layers 1–3 of the TCP/IP protocol stack. SLAs for today's customers focus on applications and layers 4–7 of the TCP/IP stack.

Service providers that adopt these new application-driven SLAs are starting to develop the necessary correlation between the quality of service metrics they promise, and the application-level quality of service that customers require. Yet, credible assessment of expected application performance has been hindered by an inability to generate and apply high loads of realistic traffic to a service provider's network and data center infrastructure. Often, a service provider's own engineers are uncertain they can reliably meet required SLAs since they lack high-powered load testing solutions. In fact, the lack of realistic application load testing tools is felt at all layers of the TCP stack. Service providers simply do not have the right tools to build confidence in their ability to deliver on SLAs. It is not surprising, then, that customers frequently experience skepticism relative to service providers' abilities to meet their needs.

Customers are growing increasingly wary; many service providers' SLAs are identical and many do not address customers' primary concerns. When service providers offer the same metrics and 'five nines' of performance, SLAs are no longer credible differentiators between competitors (or even between one carrier's standard or premium offerings). This is especially true when SLAs focus on the lower network layers of the TCP stack, when customers need to focus on the performance of their applications, i.e., layers 4–7.

What is required is a new paradigm for capturing and assessing layer 4–7 performance. Credible SLAs give service providers a significant advantage over competition. In response to this requirement, Caw Networks has introduced a suite of capacity assessment appliances that improve the character and performance of SLAs throughout the SLA management lifecycle, from network design and implementation to quality assurance, production, sales, and customer retention.

Customers, too, need tools to help them evaluate service providers and to validate the ability of service providers to provide the agreed-upon service quality, especially relative to a customer's specific applications and mix of users.

This application brief discusses how service providers can use Caw Networks' high-volume capacity assessment appliances, WebAvalanche and WebReflector, during each of the critical stages identified above to ensure that the services will satisfy the SLAs under real-world traffic conditions. These appliances also support engineering, design and validation of fast-track deployments while maintaining high confidence in the continued delivery of robust services.

2 A New Set of Metrics

Business analysts agree that current SLAs are not the decisive factor for customers selecting a service provider. A key limitation of SLAs is their failure to specify application-layer, end-to-end performance. Customers need to predict, assess, and monitor the performance of the applications that will run on top of outsourced services like networks, Web servers and XML-based marketplaces.

The focus of SLA-based performance must be end-user experience. Previously unobtainable assessment of the end user application experience can now be harnessed to express performance objectives in SLAs. These metrics include the end-user's connection throughput and response time–and within the server farm, the maximum number of concurrent users and maximum number of new users per second

a service can support. These application metrics are at an abstraction level above network traffic patterns, network errors and protocols and deal with application behaviors like HTTP GETs and POSTs, cookies, aborts and SSL. The metrics must also take into account the end-users network behavior and human factors.

Caw Networks is driving new better metrics into SLAs, specifically for layers 4–7. Caw Networks' objective is to build industry consensus for a series of benchmarking tests that simulate Internet service usage at the application level, tests that will enable service providers to demonstrate the performance of evolving versions of their services. While layers 1–3 benchmarks are important to service providers for partially testing the efficiency of their implementations, the results of such tests are difficult for their customers to interpret relative to the *applications* they are running.

The fundamental motivation for the necessary shift that SLAs must make from a layers 1–3 focus to a layers 4–7 focus is that network devices are providing functionality higher on the TCP protocol stack. As this trend develops, customers expect such functionality to match their corresponding business objectives–with appropriate metrics written into SLAs.

SLAs will therefore become increasingly complicated and require more rigorous lifecycle management. The three main phases of the lifecycle are:

- Design, planning and modeling,
- Quality assurance testing and capacity assessment, and
- Production monitoring and reporting.

While many companies have addressed the first and third phases – and most have focused on the third, Caw Networks is unique in its focus on quality assurance and capacity assessment. While all the phases are important, the lack of capacity assessment is especially damaging for SLAs. Capacity assessment determines how a service will respond to very high traffic loads; Caw Networks performs this in a repeatable, debuggable way. Capacity assessment also aids in the development of plans for scaling networks in an evolutionary manner to meet the ever-increasing demands of new customers.

Caw Networks' products create and measure real application traffic and real network conditions by modeling user behaviors and manipulating TCP level characteristics and behaviors. With Caw Networks, quality assurance testing and capacity assessment can be accomplished easily and efficiently. This affords higher confidence for both service providers and customers in the stability and performance of services. These products are customizable to be used equally well by service providers and their customers. Service providers will use Caw Networks' products to verify and demonstrate their service benchmarks, while customers will use them to evaluate and validate service providers' ability to support the customers' specific mix of end-users and applications.

3 The Challenge of Testing

Service providers face significant challenges in assuring both themselves and their customers that SLAs that define their relationships can be fulfilled. Many challenges center around the difficulties inherent in cost-effectively and consistently assessing and validating the infrastructures upon which such SLAs are based.

Obviously, service providers prefer to avoid testing network performance in live customer environments. They therefore construct test or parallel infrastructures that mimic the behavior their production infrastructure. While this approach is conceptually acceptable, constructing such parallel infrastructures entails complex and expensive tasks. The quality assurance, client project management and product management departments responsible for guaranteeing the integrity and performance of infrastructures (whether system-wide or for client-specific projects) are rarely equipped with the hardware resources or staffing to build and maintain such parallel infrastructures.

These departments must often cobble together temporary test infrastructures using significant amounts of borrowed hardware. The subsequent configuration effort mirrors the full spectrum of IT activities—including installation of hardware, applications and middleware, even troubleshooting—that are required by any normal infrastructure development project.

Yet another problem inherent to this approach is the use of traditional, script-based testing tools. Such tools require complex, expensive and time-consuming script development. Such script development often leads to missed deadlines and cost overruns that aggravate a service provider's already present sense that their testing processes are not in control.

The continual dismantling and reconstruction, with different hardware components, of such test beds raises a second significant challenge faced by service providers—achieving consistency in the testing process. The near impossibility of maintaining a standardized test bed over time under these conditions makes it extremely difficult to achieve consistency, ease of use and repeatability in the testing process. This level of difficulty and complexity places the successful fulfillment of SLAs at risk.

4 WebAvalanche and WebReflector

Caw Networks was founded to deliver high performance network solutions that improve Web and network infrastructures' ability to meet the rigors of the Internet. Caw Networks' first products, WebAvalanche and WebReflector, are network appliances, i.e., high performance hardware built for a single purpose with a custom architecture. These appliances are powerful weapons in a service provider's arsenal, with applicability from pre-sales efforts through production deployment.

To address the market opportunities and overcome the testing limitations discussed above, service providers require an integrated, all-in-one stressing solution that scales to high load volumes and eliminates box proliferation and repeatability problems. Caw Networks provides built-for-purpose appliances that can be effectively used with minimal configuration or training. These appliances produce reports with meaningful test results that can educate customers about the adequacy of their co-located infrastructures. An added benefit is reporting tailored to reflect a specific customer's forecasted load.

Caw Networks appliances deliver other significant advantages, including the provision of a foundation upon which a service provider could develop value-added, layer 4–7 load testing and service assurance products to its co-location and hosting customers. And, in a market environment characterized by indistinguishable SLAs, capacity validation with Caw Networks appliances represents a significant differentiator by enabling a service provider to accurately represent its infrastructure as fully tested—and verified—at levels that satisfy all the conditions of an SLA.

As discussed earlier, industry SLAs currently focus on network availability and reliability, with an emphasis on design redundancy and Mean Time Between Failure (MTBF) of hardware devices. While these are critical metrics, WebAvalanche and WebReflector enable service providers to extend the breadth of their SLAs. Beyond simple equipment failure, service providers can now test how key elements of Internet behavior affect applications. WebAvalanche and WebReflector are configurable to introduce unprecedented realism into tests, from network latencies and TCP behavior to human behavior, thereby enabling a service provider to gauge throughput at layers 4-7.

4.1 WebAvalanche

WebAvalanche is a high-capacity appliance designed to assess the capacity of Web sites (including back-end application, CGI, etc.) and network infrastructure (switches, load balancers, caches, etc.). Configuration of tests is performed via an intuitive Web interface. The appliance performs Web application load tests at volumes that cannot be produced using traditional software stress testing

methods. Tests should be targeted to systems either in development, QA or staging areas where service providers need to test systems at high load.

WebAvalanche achieves its level of performance through an integrated hardware and software architecture designed for processing efficiencies while generating large amounts of application traffic. Traditional software-only solutions suffer performance limitations from the one-size fits all overhead of general-purpose operating systems, while WebAvalanche is tightly integrated with a micro-kernel operating system that takes maximum advantage of its underlying hardware. WebAvalanche also leverages a customized, proprietary TCP/IP stack and high-performance device drivers for maximum performance, functionality and robustness. Caw Networks appliance hold open 1 million active TCP connections across 1 million IP addresses.

Caw Networks developed a custom TCP/IP stack because existing TCP/IP stacks did not offer the level of performance or robustness required to stress the largest Web infrastructures. With this TCP/IP stack, Caw Networks has unique visibility and configurability to manipulate test parameters such as network latency or packet loss in an assessment – and can inject network level TCP errors such as duplicate packets, out of sequence packets and packets with errors.

4.2 WebReflector

WebReflector is a high-capacity appliance that simulates a large cluster of Internet Web servers and serves as a mirror image of WebAvalanche. Through its unique ability to process large numbers of both static and dynamic Web requests, WebReflector accurately emulates the operation and performance of multiple Web, application, and database servers. Using WebReflector, a service provider can easily simulate different Web infrastructure environments without having to reconfigure hardware or reinstall server software.

WebReflector echoes the traffic generated by browser client or WebAvalanche. Unlike simple bit blasters, WebReflector tests an infrastructure's performance by introducing traffic that is directly representative of that generated by Web transactions.

WebReflector facilitates infrastructure testing without the added cost of provisioning servers to support the load. And its simplicity speeds the process of assessing the usability of a new piece of equipment, the performance of a network design, or the impact of configuration changes on an existing network implementation.

4.3 WebAvalanche and WebReflector Deployment

Used separately or in combination to test network infrastructure, WebAvalanche simulates users while WebReflector simulates a large Web site. Their performance is matched evenly so that they can be easily configured to scale to any level of performance.

A single WebAvalanche can simulate one million simultaneous users and hold open 1 million TCP connections. For even greater load, the WebAvalanche and WebReflector were designed to support teaming. A test can be configured that generates traffic from multiple WebAvalanche and WebReflector appliances to generate a unified report.

WebAvalanche and WebReflector are typically be placed in a QA lab or a workbench lab to test data center components on an ongoing basis, including testing all the elements of an infrastructure in combination. Typical users include product managers, engineers, program managers and SLA managers as well as Account Teams, managing the entire deployment process, including design, QA and customer satisfaction.

A key use of the appliances is as a testing or benchmarking solution for infrastructure product comparisons in internal lab testing, frequently in multiple testing labs and staging areas. Service providers

with multiple data centers will generally deploy WebAvalanche and WebReflector in each of those data centers (rather than running tests over WAN circuits).

Once an infrastructure product has been selected based upon its performance in WebAvalanche/WebReflector testing, the appliances can then be used to benchmark that product's performance in a variety of configurations, including the final infrastructure itself.

5 Illustrative Use of WebAvalanche and WebReflector

A common challenge facing a service provider is capacity assessment of its networking infrastructure, including its network of switches, routers and WAN lines. Tests are required for equipment selection 'bake-offs', quality assurance of new designs and implementations or customer driven analysis. For instance a customer may expect a business event that will result in a significant increase in application traffic. Customers may want to prove out an infrastructure and its ability to meet throughput SLAs under degraded conditions. The following example illustrates how WebAvalanche and WebReflector can be used to repeatedly test the performance of different network topologies.

5.1 System Setup

WebAvalanche and WebReflector can be used individually or in combination to assess the capacity of network devices and architectures, see Figure 1. This configuration consists of a WebAvalanche and a WebReflector testing a network device or system. WebAvalanche and WebReflector support both 10/100Base-T (Fast Ethernet) or Gigabit Ethernet interfaces and can deliver wire-speed performance that exceeds high-bandwidth customer connections from Fast Ethernet to oc-3, oc-12 and Gigabit Ethernet.

Because Caw Networks appliances can saturate networks and potentially render the appliances unreachable, a separate, isolated Ethernet network should be used to connect the WebAvalanche and WebReflector to a browser-based control workstation.

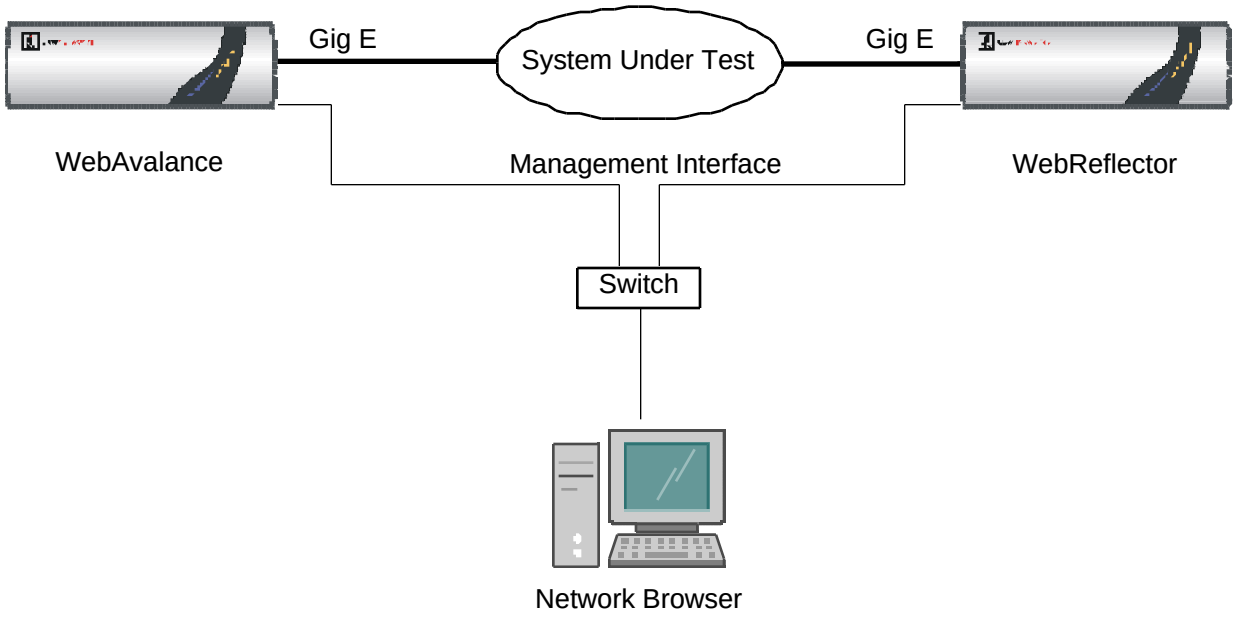

Figure 1

5.2 Test Configuration

WebAvalanche and WebReflector can be used separately or together. WebAvalanche can be used to stress test a single customer's Web site or the service provider's infrastructure or server farm.

Both appliances are configured remotely via a browser-based graphical user interface. Once connected to the appliance, screens for configuring the transactions and tests are available—pull down lists and reusable profiles speed configuration. Tests consist of load generated over time. The load is configured in user profiles that define human and network behavior, servers, URLs accessed (HTTP GET) and forms (HTTP GET and POST). Up to 100 profiles can be selected for a test run and application traffic load and the duration of the test are defined per test. WebAvalanche's user profiles to reflect the mixture of application level traffic to be generated. For each user profile the following attributes are configured:

User Network Behavior Attribute	**Description**	**Examples**
Link Speed	Specifies the speed of the link the simulated user is using.	9600, 14.4k, 28.8k, 56k, 64k ISDN, 128k ISDN, T1, OC3, OC48
Packet Loss Rate	Specifies the rate at which packets are lost. The configuration of packet loss rate is especially important for judging an end-user's experience due to network robustness conditions.	0.01% to 99%
Source IP Addresses	A list of source IP addresses (configured in blocks) from which simulated user accesses will originate. This versatile configuration enables the tester to simulate Web site access originating from a variety of access types, e.g., HTTP access over MPLS or VPN, and browsing methodologies.	128.10.1.1

User Behavior Attribute	**Description**	**Examples**
Wait Time	The amount of time a user thinks or waits before proceeding to another URL	5 seconds
Version	Indicates which version of HTTP the server is using	HTTP 1.0 or HTTP 1.1
HTTP Aborts	How often a user will click to another URL while a page is loading	9 seconds
URL List	The list of URLs to be accessed, including GET, POST and HTTPS (SSL) GETand POST	http://www.caw.com
Forms Population	Population of form data is a key configuration contributing to the interaction between the web-server and associated database/processing resources involved with the web service. Allows interaction with applications and databases.	http://www.caw.com/1?<keyval.$1> http://www.caw.com/1/%20%20

WebReflector can be used to emulate a server farm and can be configured to reflect the transaction characteristics of a Web site. This involves picking a unique numeric transaction identifier, e.g., 1024. Various attributes can be associated with that transaction, including:

Server Attribute	Description	Examples
Version	Indicates which version of HTTP the server is using	HTTP /1.0 or HTTP /1.1
Result Code	The return code of the HTTP GET	200, 404
Header String	An optional message that is included in the status line of the response	OK
Content Type	Describes the type of data being returned	ASCII or binary
Content Length	Describes the length of the data being returned	0 up to hundreds of megabytes

During a test run, multiple User profiles and Server profiles can be selected. Test duration and load targets are specified and TCP can be tuned.

Test Parameters	Description	Examples
User Profiles	The user profiles to generate traffic	'Browsers' 'Buyers' Broadband users 19.2K users
TCP Parameters	Controls TCP behavior of Avalanche and Reflector	Retries, Timeouts, MTU
Load Parameters	Controls Load Levels, progressive load tests, and ramps	Maximum concurrent sessions per second, Maximum new sessions per second, Number of sessions per step, duration of steps

When used with WebReflector, the configuration of the WebAvalanche URL list may be further configured to specify (in the http get or post request) which transaction to return as well as configured to over-write the return packet's attributes. If no attributes are specified, defaults, i.e., the default transaction identifier, its associated default return code, data type, embedded header string and packet body size, will be applied based on the http rfcs.

Running the test involves starting WebReflector and WebAvalanche from the gui. Tests can simulate as many as 1,000,000 simultaneous connections, over 10,000 transactions per second, and bandwidth over 800 Megabits per second from just one appliance.

Measured results contain detailed information about the network, server and application performance:

Throughput, Success and Failure in Transactions

Desired and Current Load (User Sessions)

Cumulative Attempted, Successful and Aborted Transactions

Attempted, Successful and Aborted Transactions per second

Throughput in Bits

Network Response Time

Current and Maximum Open TCP Connections

Minimum and Maximum Time to TCP SYN/ACK (msec)

Current Time to TCP SYN/ACK

Server Response Time

Current, Minimum, Maximum and Estimated Server Processing Time

Current, Minimum and Maximum Time to TCP First Byte

Application Response Time

Current, Minimum and maximum Response Time Per URL and Page

HTTP Return Codes

TCP Error Codes

5.3 Reporting Results

These statistics are reported both at a summary level and with granularity down to the time slice, server, and per User Profile/per URL detail. During a test run, real time statistics are generated on the user GUI and after a test all statistics are written to a comma separated file (CSV). This information is presented in customized report with charts and graphs generated by Caw Networks' *IcePick* plug-in to Microsoft Excelä . Or, the Service Provider may import the CSV file data into other reporting packages.

Using the process described above, the deliverability of SLA-based performance guarantees can be determined. Metrics demonstrate an infrastructure's performance can meet SLA goals like response time and successful transactions even at the highest loads.

Often the same test is performed again with WebAvalanche and WebReflector tightly coupled to take baseline measurements. Subsequent tests with the device or system under test enable the calculation of performance results attributable to the device under evaluation.

Service providers can determine if a configuration change, such as tuning the cache size of a server, offers any benefit relative to the load-handling characteristics of the service. It is therefore possible to justify or reject purchases of equipment upgrades, and clearly justify of that decision.

The same tests can be used to conduct a side-by-side comparison to determine if a new piece of infrastructure offers performance advantages that more cost-effectively meet their SLA commitments. Caw Networks appliances provide a steady baseline while equipment is compared or tuned. This methodology removes guesswork or extrapolation based on low, software-generated load test from the selection of load balancers, network caches, switches, etc., and is critical for right-sizing infrastructure expansion/upgrade with existing back-end systems.

6 Process Applicability

WebAvalanche and WebReflector provide significant value at each stage of a service provider's engagement with a customer.

Step	Service Provider Challenge	Application of WebAvalanche and WebReflector
Network Design	Fast comparison of network design alternatives. Cost avoidance by designing a 'right-sized' solution–no overbuilding or under building.	Stressing routers, load balancers, caches, firewalls and other network devices stressed at high volumes. Assessment of performance of a network exposed to large volumes of traffic during successful stages of deployment with corresponding design changes.
Service Productization	Development of tiered and differentially priced service offerings based upon differing levels of user experience, etc.	Verification of network designs relative to productized service offerings.
Sales	Removal of customer skepticism relative to performance promises embodied in an SLA	Provision of verifiable evidence that proposed services can accommodate high load and transaction levels.
Implementation	Tracking performance characteristics of network design changes.	Reporting to customers on the performance of an implementation in progress.
SLA maintenance	Ongoing assurance that network deployment meets SLA commitments.	Periodic assessment of network performance, enabling service provider to avoid SLA-related problems.

7 Conclusions

An SLA is an important instrument for reaching agreement between a service provide and customer. However, current SLAs have been network-centric, covering only layers 1–3 of the TCP stack. SLAs should include objectives that are customer-centric and cover layers 4–7. Caw Networks is building industry consensus for a series of application-specific metrics and benchmarking tests based upon those metrics. Caw has produced appliances that emphasize network realism as well as the human behavioral impact of applications and that assess the capacity of network devices and topologies at layers 4–7. Service providers use these appliances to verify their ability to provide the quality of service promised in their SLAs, in a way that is natural to the customer.