Real World Capacity Assessment

Caw Networks' Introduction to Real World Capacity Assessment

Philip Joung

Informational Brief 01
August 2001

Abstract

Capacity assessment has become a necessity in today's complex network environments. Indeed, conducting rigorous assessments can not only bring peace of mind and improved performance and stability to the final system, but also saves time and money. Conducting these assessments requires tools that properly address the realities found in environments such as the Internet or enterprise network infrastructures. In fact, conducting assessments without these realistic factors can be risky, often not resulting in a real world picture of the performance or capacity of the system. We introduce CawUsers as the definition of a real world user—assessments using CawUsers will vastly improve the understanding and confidence in a system to meet the demands and rigors found in the real world.

Document 702-009112, V1.1.2.

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 Caw Networks' Introduction to Capacity Assessment

Businesses have become increasingly dependent on the performance and reliability of their network systems. Because of this greater dependency, understanding and assessing system capacity has become a crucial part in ensuring system availability and performance. As these network systems become increasingly complex, gaining that understanding requires a more robust set of products, many of which are just now becoming available. Using the proper capacity assessment products not only brings understanding and improvements to system performance and capacity, but ultimately saves substantial time and money. In other words, capacity assessment delivers a much improved final product while saving valuable resources—a rare combination.

Conducting realistic assessments requires having a certain amount of expertise, skill and understanding. This document introduces some of the important concepts related to capacity assessment, including why it's crucial, a look into real world capacity assessment, and an introduction of what's available to ensure rigorous and reliable capacity assessment.

2 Importance of Capacity Assessment

During the design and creation of a high performance network system, many design decisions must be made, often with significant impacts on future scalability and performance. Capacity assessment can help right from the start—the ability to quickly discover the impact that these decisions can have results in a final system with higher and more reliable performance. It is also important to understand the final system's performance and capacity so that when nearing its limitations, work on increasing capacity begins before failures ever occur.

Systems that do not meet intended performance goals often end up having costly and embarrassing failures. There have been several highly publicized failures of various Internet Web sites in the past several years. Many of these companies to not only lose market capitalization following such failures, but are also forced to quickly spend millions of dollars in network improvements in an attempt to avert future failures. Outages of internal network systems can also cost companies significant time and money. Employees have become so dependent on the availability of network systems that often an outage prevents any work from being done.

It is expensive to conduct capacity assessments, right? Actually, it turns out that not doing so is much more costly. While acquisition and setup costs associated with capacity assessment products are often hard to justify given cramped IT budgets, studies suggest that not using test tools drastically increases both the time and dollar cost associated with system deployment. Among a survey of 172 IT professionals [NEWPORT99], 52% of the respondents failed to meet capacity expectations for their Web site. It was discovered that 60% of those users did not conduct any performance tests at all as compared to only 6% for respondents where capacity expectations were met. The costs in both time and money were much higher for the group that had failed expectations, as shown in Figure 1.

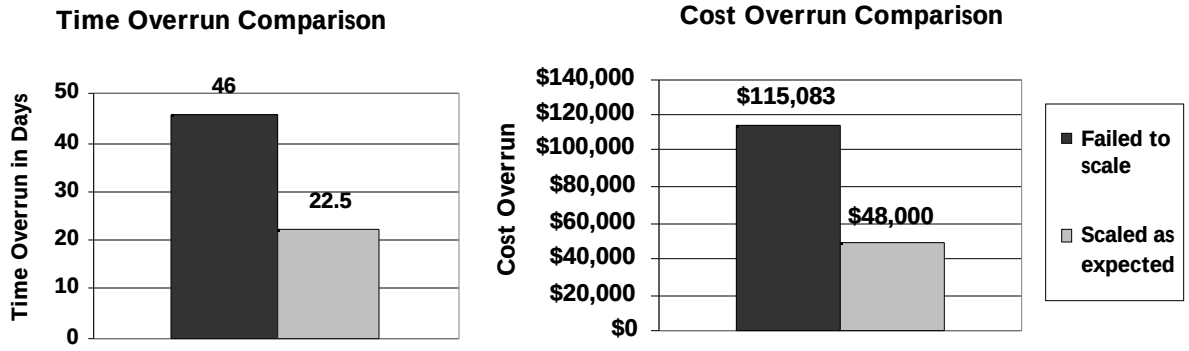

Figure 1 A comparison of the costs associated with failing to run rigorous capacity assessments. Companies that failed to run capacity assessments tend to take much more time and spend more money. Data from a 1999 Newport Group study.

Starting capacity assessments early in the development cycle can maximize time and money savings. By ruling out bad design decisions through capacity assessment, one can avoid proceeding down paths that fail to meet requirements. This avoids redesigns near the end of the design cycle, when changes cost the most.

Finally, gaining confidence in the final design, performance and stability of the system is valuable in itself. Having the knowledge that a system will perform even under a large onslaught of traffic brings not only peace of mind, but ensures that systems are not over-provisioned, thereby achieving vastly improved ROI.

3 Design Lifecycle Performance Management

Designing and maintaining a network system's performance, stability and reliability requires several steps. These steps include modeling, rapid prototyping, development, deployment/manufacturing, monitoring and maintenance. Each of these steps involves various decisions and strategies that often have a significant impact on the final system. Often, one will have to make tradeoffs between performance, budget, stability and functionality in the path towards a successfully completed system. The more information available while making these tradeoffs, the better the final result.

What's the most reliable path to deciding upon various tradeoffs? Certainly experience plays an important role in making these decisions, but today's network systems continue to change rapidly, often making it difficult to apply prior experiences to current designs. As a result, real world capacity assessment is the most reliable way to understand system performance while it is still safely contained in own company and/or network.

Capacity assessment should play a role in each step of a system's lifecycle, from modeling all the way through development and maintenance—the insights gained through these assessments will help create a final product that meets or exceeds the expected goals. Using a product that can support capacity assessments throughout this lifecycle is invaluable—it saves money by reducing acquisition costs while saving time by requiring users to learn only one tool.

4 Real-World Assessment

Why is network realism of assessment traffic so important? The Internet, based on its design, has proven effective in handling millions of simultaneous users, giving its users access to unprecedented amounts of information and a means of communicating faster and more easily than ever thought possible. However, the internets' inherent flexibility and current size and growth rates also introduce a certain amount of "messiness" to the traffic that arrives at a connected system. This messiness can have a surprising effect on performance and scalability. Deploying a system without an understanding of this effect can be very risky, with these systems often failing to meet expected performance levels. This can often be true even when the systems are assessed by certain products that only offer laboratory-clean traffic.

Testing the capacity of a system requires products that fit the task. The product should obviously have enough performance and capacity to exceed the limits of the system under test. The product should have the ability to generate loads that will mimic traffic patterns expected by the system, including important parameters such as user click-paths, think time and browser emulations (HTTP versions, cookies, headers, etc.). Finally, the product should create traffic that accurately reflects the real-world network traffic that arrives at a Web site, including millions of users, connection speeds, packet loss, millions of IP addresses, and user aborts (which all contribute to "messiness"). Assessing the performance of a system using only a subset of these parameters will give a very skewed and unrealistic picture of how the system will perform under real-world scenarios.

4.1 Millions of Users

Busy network systems can encounter millions of users daily, some even hourly. Each user takes a certain amount of resource to maintain, regardless of whether the user generates activity or not. As a result, gaining the confidence that a system can effectively handle these users requires a product that can generate and maintain this number of users.

Each User will generally have a particular behavior on the network. Some of them might generate lots of activity, others momentary spurts of activity, and others almost none at all. For example, on an eCommerce site, some users may simply browse while others have activity that involves a purchase. Once the network users are categorized, these users can be entered into a capacity assessment product, allowing it to simulate a heavy load of realistic users.

4.2 Connection Speed

Connection speed is often counterintuitive in its effect on performance. Many people expect higher connection speeds to cause greater system load, using more resources to deliver the information quickly. While this may sometimes be true, slower connection speeds almost always take more system resources than high-speed connections. Every connection to a system requires both CPU and memory to keep track of the connection. Fast connections will quickly enter, grab data and leave the system, allowing resources to be freed up to handle subsequent connections. However, slow connections take longer to complete data transfer, meaning that the system has to maintain that slow connection until it finishes. The end result is that slower connections reduce the number of users that a system can handle per second.

Still not clear? Take an example of a busy fast food restaurant. It will have two general categories of diners: take-out and sit-down. In this example, the take-out diners are the high-speed connections and the sit-down diners are the low-speed connections. The fast food restaurant can easily serve hundreds of take-out diners an hour, but is severely limited in the number of sit-down diners it can handle. Once the seats fill up, new diners will have to wait in line for the next available seat, thereby using up space in the restaurant for long periods of time.

The realism of a connection speed is very important. Many products often claim to support various connection speeds, but in reality, the traffic being generated becomes wire-speed. Due to limitations of the general-purpose operating systems (e.g., Linux, Windows NT, and UNIX) that these products run on, traffic that is purposely slowed down is often queued by the operating system until the entire message can be sent along the network at full wire-speeds.

Many network systems will encounter highly varying levels of connection speed. These systems often exhibit surprising drops in user handling capacity when submitted to low connection speeds. As a result, a capacity assessment product must have robust support for realistic connection speeds to ensure confidence in the final assessment results.

4.3 Packet Loss

Packet loss can be one of the most devastating conditions on the Internet. Various studies have shown packet loss to cause significant decreases in performance [NASA98, CAL98, SLAC00, UMD97]. The NASA study has shown that FTP performance drops by more than 80% with only 4% packet loss. Going even further, the Stanford Linear Accelerator Center study predicts network performance to drop by 95% with 4% packet loss.

Why is packet loss so problematic? Let's try to tie this to another real-world situation, faxing. Let's assume that you are trying to send a document via fax to a customer. You first call and say that you are sending the fax (which would equate to the first network packet). Your customer agrees on the phone to be ready to receive the fax (second network packet). You walk over to the fax machine, load the documents and send the fax (third network packet). You call the customer again to see if the fax went through successfully (fourth packet), but the customer tells you that the pages got stuck together and therefore didn't come through (fifth packet). You walk over to the machine again to send the document again, this time making sure the pages do not stick (sixth packet). You call again, and finally the pages have gone through correctly (seventh packet). The customer thanks you (eighth packet). Note that the sending of a fax involves several complete communications to make the entire transaction successful. Delays and extra work occur when one of the messages fails to be sent, just like it does for network systems.

How much packet loss exists on the Internet and the World Wide Web? One company has created a site that reports latency, packet loss and reachability for various parts of the Internet [MATRIX01]. Packet loss on the Web typically hovers somewhere between 3-4%, with peaks often exceeding 10%.

As a result, assessing a system's behavior with packet loss becomes crucial to understanding its performance when running on the Internet. Assessments that ignore this important real-world condition on the Internet may fail to accurately determine a system's production capacity and performance.

4.4 IP Addresses

The Internet handles the traffic for millions of users every day, and almost every user generates traffic using a unique IP address. A well-known system connected to the Internet can encounter hundreds of thousands of unique IP addresses at any given time. As traffic ramps up, it becomes important to ensure complete compliance with the real-world conditions that exist on the Internet, and being able to simulate a unique IP address for millions of users becomes an important aspect of that realism.

Many network devices do their work by maintaining a record of the IP addresses they handle. These include addresses handled in the recent past (some must maintain for 30 minutes or more) along with the new ones that continue arriving, often adding up to hundreds of thousands or even millions of addresses. As a result, capacity assessments conducted using only one or a few IP addresses will consume much fewer resources than realistic assessments using hundreds of thousands of IP addresses. Running assessments with large numbers of IP addresses helps ensure that one will understand the behavior of a system in real world situations.

4.5 User Aborts

Everyone is familiar with visiting a slow Web site. Conducting a search, bringing up the home page, or displaying the shopping cart takes many agonizing seconds. After a few seconds, many users will often press the stop button (called a user abort) and try again or just leave. Once the stop button is pressed, the browser will no longer accept data pertaining to that original request even if it finally arrives. However, in many cases, the original request continues processing on the servers, tying up valuable resources that could have been used for new requests.

Heavily loaded systems that encounter this situation often fall into a vicious cycle—the systems continue processing requests that have been stopped while also encountering duplicate requests from users trying again. Simulating this user behavior during an assessment helps improve confidence that a system will behave as expected when live on the Internet.

To clarify the concept of a user abort, let's use another real world analogy: a popular furniture store. Most furniture stores do not stock large items such as sofas or dining tables. Therefore, when a customer wants a particular sofa, the furniture store sends the order to their manufacturer, and the sofa is then built to order, often taking weeks or months. Let's say this particular furniture store has a generous return policy that allows customers to change their mind up until the completed furniture is delivered. Anyone who changes their mind after say 3 months of waiting (a user abort) would cost the store much wasted time and money. The situation might be even worse if the store doesn't tell the manufacturer to stop immediately after a cancellation. A furniture store encountering many "user aborts" would probably not be in business for very long, likewise a network infrastructure encountering high levels of user aborts will struggle as well.

5 CawUsers™: Tying Realism and Performance Together for Capacity Assessment

There are many products available on the market that attempt to assess the capacity of a system. On the surface, they may all seem quite similar, with comparable features such as browser emulation, connection speed, and the ability to support hundreds, thousands or even millions of users. Number of users often becomes one of the more important metrics used by these programs (many even being priced according to number of users) because it is often equated with the performance ability of the product. However, a closer look reveals the individual differences in each product along with the effect that these differences can have in conducting an effective capacity assessment.

In order to create loads that accurately reflect realistic network conditions, a tool must combine the many different elements mentioned above. Successfully tying these different elements together creates traffic that will stress systems in a way that closely reproduces live system performance.

Until now, no solution that brought all these elements together into a robust and reliable definition of realistic capacity assessment—CawUser is that definition. CawUsers tie these realistic elements together to serve as the benchmark standard for real world capacity assessment, allowing assessments to be conducted ensuring a high degree of confidence.

CawUser Feature	Explanation
Connection Speed	Having the ability to realistically create connections at different speeds improves the realism of an assessment. Slow connections can decrease performance.
Think Time	A user at a Web site, for example, will have a delay between viewing each page in order to read the page or to type in information or a search. This is important in ensuring that a CawUser behaves similarly to a real user.
User Abort	User aborts happen on the Internet and Intranets all the time, causing the most pain when the system is already slow to begin with.
Packet Loss	A system's performance often degrades significantly when trying to deal with packet loss. Understanding this effect is crucial to deploying a system that will meet performance objectives.
Click Path	A real network system user looks at various pages on the site and performs different activities such as searches, password logins, and purchases. Each of these activities can create different loads on the system. It is important to test the impact that these different user activities will have on system robustness and performance.
IP Addresses	Because large numbers of IP addresses take system resources, testing performance using this metric will help bring confidence that the system will function as expected on the Internet.
Browser Emulation	A system often customizes its response based on the browser being used. Different browsers support different versions of HTTP, causing different effects on network traffic. This processing overhead makes it important to include this feature while testing.
SSL support	SSL helps ensure the safety of the data being transferred between a user and a Web system. However, data encryption takes an immense amount of processing, and system performance expectedly drops in response. For systems that plan to use SSL, understanding the performance under SSL is crucial.

So what is a CawUser? It represents a whole new realism in capacity assessment that was previously unavailable. By conducting assessments with CawUsers, a complete understanding of system performance can be gained before the system ever serves a real-world user, ultimately bringing confidence, cost/time savings, and reliability to the system.

6 The Caw Networks Product Family

What products are available to conduct real world capacity assessment using CawUsers? In fact, there is only one set of product available that can reliably deliver this level of realism and performance. Caw Networks' family of capacity assessment products form a powerful combination in high-capacity performance testing. Detailed information can be found on the Web at http://www.caw.com/product/product.shtml.

6.1 WebAvalanche

WebAvalanche is an appliance for stressing network systems, with enough performance and capacity to challenge any computing infrastructure to withstand high volumes of realistic Internet/Intranet user traffic. It gives its user high performance, realism and ease-of-use all in one compact appliance. WebAvalanche's versatile design allows it to be used in any stage of a system's design, deployment, or maintenance.

6.2 WebReflector

WebReflector is an appliance that simulates the behavior of a large cluster of Web servers, generating HTTP responses to user requests simulated by WebAvalanche. It is the only commercially available product that can withstand the traffic levels generated by WebAvalanche. As a result, any system placed in the middle of these two products can be assessed for their impact on system performance.

7 The Future of Capacity Assessment

Testing with CawUsers provides new, real world results, ensuring that today's systems meet the challenges found in Internet and Intranet traffic. Of course, as the Internet and networks continue to evolve, today's CawUser will surely look different in the future. While it may be difficult to accurately determine what the Internet will evolve into, it is our commitment to ensure that a CawUser will accurately reflect the realities that make up the Internet. We expect CawUsers to set the benchmark by which rigorous and reliable capacity assessments gets done.

8 Conclusions

Every once in a while, a new concept emerges to challenge previous ideas and methodologies. We believe CawUsers brings a new level of realism sorely missing in current capacity assessment methods. Companies can now conduct capacity assessments and have confidence that their efforts will not only result in an improved system, but save time and money as well. The realism provided by WebAvalanche and WebReflector allows companies to confidently find network problems, discover system capacity, and simulate realistic conditions before the system is deployed in a live environment. Caw Networks is setting the standard by which reliable capacity assessments are done. Real World Capacity Assessment—the only way to understand network performance, and only available from Caw Networks.

9 Glossary

Cookies	When talking about the World Wide Web, cookies refer to small amounts of information that are transferred and stored on a user's computer. This information is used by Web sites to remember the user and present the correct information to the user from page to page.
FTP	Acronym for file transfer protocol. A widely used standard for transferring files on the Internet.
HTTP	Acronym for hypertext transfer protocol. It is the standard used on the Internet for transferring and viewing documents using a Web browser.
IP Address	A numerical value used on a network to describe the network location of a particular machine or device on that network. Each device on a network must have a unique address.
SSL	Acronym for secure socket layer. Web sites use SSL to encrypt information between themselves and a user, helping to ensure the information remains secure.
URL	Acronym for uniform resource locator. URLs are used by Web browsers to make Web sites easier to locate. Instead of typing a numerical value to get to a Web site, one can type a much cleaner looking and easy to remember URL, e.g. "www.caw.com" instead of "216.200.40.68".

10 References

[CAW00] Caw Networks Inc.
High-Volume Web site Capacity Assessment.
Technical Report 1, Caw Networks Inc., November 2000.

[NEWPORT99] Newport Group, Inc.,
Making E-Business Work, Early Adoption and Frequent Use of Load Testing Tools Combat Web Application Scalability Surprises
http://www.newport-group-inc.com/, September 1999

[NASA98] National Aeronautics and Space Administration, Integrated Service Network
The Effect of Packet Loss on TCP Application Performance
http://nileweb.gsfc.nasa.gov/advtech/packet-loss/study.html, December 1998

[CAL98] University of California, Berkeley and IBM T.J. Watson Research Center, H. Balakrishnan, V. Padmanabhan, S. Seshan, M. Stemm, and R. Katz
TCP Behavior of a Busy Internet Server: Analysis and Improvements
Proc. IEEE Infocom '98, San Francisco, CA, USA, March 1998.

[SLAC00] Stanford Linear Accelerator Center, Les Cotrell
Throughput Versus Loss
http://www.slac.stanford.edu/comp/net/wan-mon/thru-vs-loss.html, February 2000

[UMD97] University of Maryland, M. Beynon, R. Ferreira, A. Afework, and G. Mohan
Performance Evaluation of Client Server Architectures for Large-Scale Image-Processing Applications
http://www.cs.umd.edu/~rich/courses/cmsc710-f97/projects/microscope/paper/paper.html, Fall 1997

[MATRIX01] Matrix Information and Directory Services, Inc., Joi L. Chevalier, Editorial Director
Internet Average
http://average.miq.net/