
Benefits of Web Caching

Joe Cooper

<joe@swelltech.com>

Copyright © 2000, 2001, 2002 Joe Cooper

2 May 2002

Revision History		
Revision v0.01	10 Aug 2000	rjc
First public draft. Comments welcome.		
Revision v0.02	2 May 2002	rjc
Second verse... Comments welcome.		
Revision v0.03	9 Nov 2002	rjc
XML conversion, cleaned up, figures added.		

A guide that answers the question, "Why cache?" Provides a detailed cost benefit analysis of web caching in a real network environment. If you're wondering if caching is a good idea for your network, this document is for you.

Table of Contents

Introduction	1
What is a Web Cache?	1
The Longer Answer	2
How Does it Work?	2
How Does it Save Bandwidth?	2
Number of Clients	3
Type of Clients	3
Size of Cache	3
How Does it Improve User Experience?	3
An Example Cost Benefit Analysis	3
Executive Summary	4
Cost Table of Each Option	4
Explanation of Table 1	4
Risk Factors	4
Benefits	5
Explanation of Table 2	5
Cost Benefit Analysis Conclusion	6
Final Words and Continued Reading	6

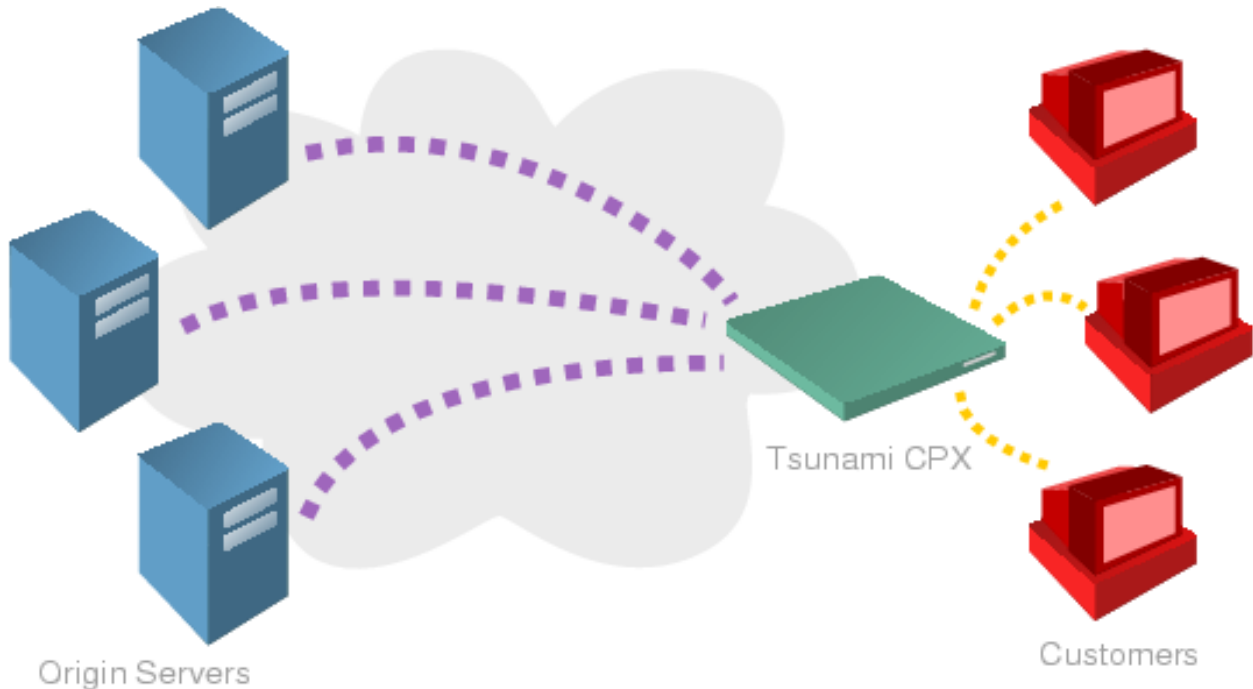
Introduction

This document is an attempt to explain web caching and the benefits of web caching to a general technical and non-technical audience. No previous caching experience is expected. If you already know why caching is a good idea for many networks, then perhaps you should skip this document and go straight to the *Designing A Web Caching Infrastructure* guide located in the documentation section of the Swell Technology Homepage.

What is a Web Cache?

In short, a web cache is a device designed to store local copies of network web objects that are fetched from the internet, so that they can be retrieved quickly and without further bandwidth use the next time they are accessed. Documents retrieved from the cache are more quickly accessible, and a significant amount of bandwidth can be saved. This leads to increased client satisfaction and reduced cost of network connectivity.

Figure 1. A Web Cache Accelerated Network



The Longer Answer

The internet is a big place. Documents, images, and other files, are stored on servers in every corner of the planet. Retrieving those objects can be an expensive endeavour. High speed internet access is, and will for the foreseeable future, be outpaced by the increasing demands users place on the network.

As the internet gets larger and more diverse, network administrators struggle to maintain a high level of service quality without increasing bandwidth expenditures. Web caching is the most cost effective way to solve that problem.

Adding web caching to a network can reduce bandwidth usage by 25%-35% depending on the type and number of users on the network. Quality of service will also be improved significantly when web caching is in place. Request response time will be reduced by more than a factor of ten. Finally, network reliability will increase as the ability of the network to perform during load surges will be enhanced.

How Does it Work?

Web caching works by storing a local copy of data from the internet. For example, if a user on a network without web caching in place visits a website it will be loaded directly from the internet. If she returns to that page later in the day, her internal browser cache may have saved a copy for her. However, if another user on the same network visits that same page five minutes after the first user, the page will still have to be loaded directly from the origin web server. Every other user who requests that page will have to retrieve it from the distant origin server as well. But if the network has web caching in place, that page can be served to one or one hundred users with no more internet bandwidth expended.

How Does it Save Bandwidth?

When web caching is in place on a network, it is not unlikely that a request will be able to be fulfilled by the web

cache rather than the origin server. Because the web cache is on the internal network, the extranet will not be used to retrieve the data. Data will only traverse the local network from the cache to the client's browser. This frees expensive external bandwidth for usage by more users.

The amount of bandwidth saved is usually in the range of 25% to 35%. The savings is variable and is affected by factors such as the number of users on your network, the type of user, and the size of the cache. An explanation of the factors that have a bearing on bandwidth savings follows.

Number of Clients

The more clients there are accessing a cache, up to its rated load, the better it will perform from a bandwidth saving perspective. This is because, given a large enough cache, the larger the client base is the more likely it is that *someone* will have visited any given site. Hit ratios usually climb to their highest levels during the peak load periods of the day, and fall during low use periods.

Type of Clients

The type of users on a network is also an important factor in estimating how much bandwidth can be saved. If you have a tightly focused group of users, bandwidth savings will be higher than if all users have diverse interests and visit a wider array of sites. Also, high demand users, or *power users* may have similar browsing habits which can be exploited to maximize bandwidth savings. For example, if a large number of network users hear about a new browser release and all rush to the website to download it, it can be cached when the first user retrieves it. During this period of time, bandwidth savings may be significantly higher than average. This effect is seen in many segments of network users. For example, new versions of RealPlayer, Windows service packs, video game demos, and even the Linux kernel may cause a spike in bandwidth savings for a few days.

Size of Cache

The size of the web cache in place can also have an impact on the amount of bandwidth that can be saved. Cache storage size, and the web caches ability to perform well under peak network load, are the two most important considerations. More cache storage leads to better bandwidth savings, as more objects can be saved locally, and the chance of a request being a *cache hit* increases. In short, the more web objects that can be stored on the cache hard disk(s) the better bandwidth savings will be. This, of course, assumes that the cache is able to effectively handle the full network load even during peak periods. Buying a cache that is underpowered, but features several large disks will not result in increased bandwidth savings beyond that of an appropriately powerful cache with smaller storage, and in fact will probably result in lower bandwidth savings in addition to increasing latency to uncomfortable levels. Both factors must be balanced in order to maximize bandwidth savings.

How Does it Improve User Experience?

Quality of service is another important reason to consider web caching. The time it takes to receive a document from the web plays a very important role in how enjoyable the users web browsing experience is. A rule of thumb among web developers is that a response time over two seconds is too long for most users to wait for a page to load. Unfortunately, response times over two seconds are very common on the web.

Even on networks where local internet bandwidth is not saturated and does not impose any limits on the speed of the clients browsing, the internet at large is out of the control of local network administrators. There will be temporary network slowdowns, distant web sites will sometimes be overloaded, and web pages will occasionally be large and slow to load. A web cache can alleviate these problems to a significant degree. Only uncached requests will experience any delays that are present. Every subsequent client that accesses a cached site will receive fast access directly from the local network. This often correlates to improving response times by a significant amount, in the case of a local ethernet it is reasonable to expect latency to drop below a tenth of the origin server latency.

An Example Cost Benefit Analysis

What follows is a cost benefit analysis for a fictional internet service provider named CricketNet.

Executive Summary

CricketNet currently rents four T1 lines from their backbone service provider. CricketNet is in a location where there is some competition in the T1 market, so they only pay \$1200 per month for each of their four T1 lines. CricketNet now supports approximately 700 simultaneous users during peak periods including a small number of DSL, several frame clients, and a majority of dialup users. Their total subscriber number is in the thousands and occasionally they experience slowdowns as network usage climbs to 800 or 900 users.

CricketNet has recently upgraded routing infrastructure to support the new DSL customers, but bandwidth has not yet been upgraded for these new users, and it is beginning to become evident in the response time and reliability of the network. The purpose of this cost benefit analysis is to compare and choose between the addition of a fifth T1 line, or the addition of a Swell Technology Tsunami web cache, in order to allow continued expansion of the CricketNet customer base and improvement of service.

Cost Table of Each Option

The cost of each option in nearterm and longterm expenses for maintaining each option is outlined in the chart.

Table 1. T1 vs. Tsunami 20s

Network Upgrade Option	Base Cost	Installation and Other Initial Costs	Monthly Expense Over 18 Months	Total Expenditure For 18 Months
T1	\$1200.00/month	\$350.00	\$1200.00	\$21,950.00
Tsunami 20s	\$4189.00	\$0.00	\$232.72	\$4189.00
Saved	-\$2989.00	\$350.00	\$967.28	\$17761.00

Explanation of Table 1

While Table 1 is quite simple, it does present most of the data needed to make a decision based on cost alone. Other factors such as response times and network surge handling are addressed below. Table 1 presents the costs of installing and renting a T1 line for 18 months as opposed to purchasing and maintaining a Tsunami 20s for the same 18 month period. It is quite clear that the lower cost option is the web cache. The table does not take into account any unexpected additional expenses that may arise related to either product. Those are best discussed as risk factors.

Risk Factors

With any product there is some risk of failure or problems. When choosing a product, the support and methods for handling such failures and problems should be carefully considered. T1 access providers vary wildly in the types and levels of service that are included in the normal monthly cost. In our example, CricketNet has chosen a very nice T1 provider who offer free technical support and free repairs of network lines except for that which is inside the premises. So unless CricketNet damages the lines inside their offices, they will never have to pay to have them repaired.

The only concern with the T1 then is how quickly the T1 provider responds when the line fails. Again, response time for repairs from T1 providers varies wildly. We will again assume that CricketNet has a good provider that usually respond to repair requests within one week, sometimes quicker. Assuming the problem is easily traceable, the fifth T1 may be back online in under one week. And if the problem is line damage in an unknown location, possibly two weeks.

In the case of a Tsunami, the major risks include failed hard disks, or failed system board or processor. Disk failure is by far the most common failure in a web cache, though it is still very rare. Repair or replacement is covered for two years by the Swell Technology warranty. If CricketNet experiences a drive failure, they may simply take the failed drive offline and run at reduced capacity until a new preformatted drive arrives from Swell a few days later. With standard swappable drive bays in both the mid-range 12i and 20s models, disk replacements in the field are

fast and painless. Shipping is also paid for by Swell. So there are no expenses in the event of a Tsunami failure. The only consideration then is the time it takes to get the cache back online in the event of total system failure, requiring a replacement or factory repair. Swell will repair the unit within five business days, usually sooner. Further, technical support is unlimited and free for two years from the purchase of the Tsunami unit. Continued support contracts are available after the end of the two years.

So the two options are quite similar, in this case, with regard to risk factors. Though some problems with the T1 may present longer service outages than the Tsunami, it's quite likely that neither will present service failures during this 18 month period of service.

Benefits

Each option has benefits, which should be considered. While both will allow more web traffic to flow through the network, the Tsunami will improve response times for some data, providing a more pleasant browsing experience for users. The T1 on the other hand will provide additional bandwidth for all services, not just web traffic. Nonetheless, because the Tsunami will offload traffic from the four other T1 lines, similar throughput capability will be the result.

Table 2 is a brief outline of response times and bandwidth availability for each option.

Table 2. Bandwidth and Response Times

Network Upgrade Option	First Visit Reponse Time	Second and Subsequent Visit Response Time	Comparable Bandwidth
T1	0.25 sec	0.25 sec	1.5 Mbps
Tsunami 20s	0.25 sec	0.02 sec	1.26 Mbps
T1 - Tsunami Difference	same	0.23 sec faster response	.24Mbps Less aggregate bandwidth

Explanation of Table 2

It may be a little more complicated to see just what is going on in Table 2. The *First Visit Response Time* is the amount of time a web request takes to complete at the internal router. Because there are so many outside factors, no attempt was made to compare client side response times. It would be very difficult to make an accurate estimation, given the variety of client types (56k modem, DSL, Cable, frame relay, wireless, etc.) and the difficulty of obtaining reliable statistics. *Second and Subsequent Visit Response Time* is the time it takes for requests for cacheable objects that have been visited before by someone on the network. The Tsunami generally exhibits response times for object hits less than 1/10th as long as the T1 or better.

Comparable bandwidth is probably the most difficult value to estimate, and many networks will exhibit differing load types than CricketNet. In our example, CricketNet's network traffic is 60% web browsing traffic. This number is higher on some networks, and lower on others, so one should calculate based on your own network load patterns. So, assuming 60% web traffic and a 35% cache hit ratio, we have a calculation something like this:

Example 1. Bandwidth Savings Calculation

4 T1 lines x 1.5 Mbps = 6 Mbps Total Real Bandwidth

6 Mbps x 60% = 3.6 Mbps Total Web Traffic

3.6 Mbps x 35% = 1.26Mbps Bandwidth Savings given 35% hit rate

Cost Benefit Analysis Conclusion

In conclusion, the most cost effective bandwidth expansion option for CricketNet is purchase of a Tsunami 20s web caching appliance. The overall cost of purchase and operation is substantially less than leasing of a T1, while the overall benefit to their clients is greater. The Tsunami will allow them to continue to operate with their current bandwidth, while improving customer service and increasing their customer base. It doesn't preclude adding another T1 in another 6-12 months, as their customer base expands to create those demands. If CrickNets client population and bandwidth needs continue to grow rapidly, their Tsunami can be upgraded via addition of a third disk and more memory. Or they may wish to purchase a second Tsunami to provide complete redundancy.

Final Words and Continued Reading

It is clear that web caching is a cost effective means of expanding network capabilities when compared to non-caching alternatives. The capital saved by implementing caching in nearly any network will very quickly pay for the expense of the web cache. Further, a web cache will continue to operate for many years, as long as it is needed, whereas many non-caching alternatives will continue to present expenses for as long as the product is in use.

We at Swell feel that we offer web caches that are an excellent value with a solid support agreement. Swell products feature a two year unlimited warranty and two years of unlimited technical support. Software upgrades are free for the life of the product. You may wish to continue your reading with our guide to *Designing a Web Caching Infrastructure for Your Network* or you may contact a Swell technical representative at support@swelltech.com [mailto:support@swelltech.com] to find out what product is right for you, and how you can make your network run faster and more efficiently through web caching.