



**Caching and Compression – Key Complementary  
Technologies for Application Acceleration**

White Paper

---

## EXECUTIVE SUMMARY

More business than ever is done in branch offices and at remote sites, but fewer and fewer IT resources (applications, personnel, servers, etc.) are hosted there. Applications and data resources are being consolidated (sometimes centralized, sometimes outsourced) – cost is a driver, but compliance is the catalyst.

In many cases, a consolidated application translates to a poorly performing application. Long physical distances between users and applications, skinny/latency-prone network pipes, and inefficient/chatty protocols result in sluggish and often unusable applications for users at remote sites. These issues are exacerbated by the increasing use of bandwidth-hungry, latency-sensitive applications, such as voice and video.

To address these performance problems, caching and compression are two complimentary technologies that combine to minimize the amount of data that traverses the enterprise WAN. Working together, caching minimizes the transmission of duplicate data over the WAN while compression reduces the total amount of data to be transmitted. Caching and compression are two of the five technologies that are part of the Blue Coat's MACH5 application acceleration framework.

This paper discusses two approaches to caching, object caching and byte caching, that combine to deliver both caching and compression. (Byte caching is technically “caching,” but it is commonly considered compression. For this paper it will be referred to as a caching technology.) This paper further discusses how these two technologies play a pivotal role in delivering a comprehensive solution to accelerate any and all application traffic across a WAN.

## CONSOLIDATION SAVES MONEY...BUT CAN SLOW BUSINESS

### How consolidation can lead to slow and congested WANs

An organization's wide-area network (WAN) is the central nervous system of the enterprise – keeping geographically dispersed locations connected with a flow of information. Increasing use of network-based applications and a move to consolidate data centers has increased the traffic on enterprise WAN links several-fold. More users than ever need to use the WAN in order to do their job. Increasing WAN traffic has caused congestion in the enterprise network, resulting in poor performance for critical enterprise applications. Unacceptable performance of enterprise applications may force organizations to give up the benefits of data center consolidation in order to keep the business processes functioning.

The following example illustrates how data center consolidation may increase WAN bandwidth usage several times.

Edge Corporation has one headquarters and one branch office site. Both of the locations are connected over a WAN link. Each of the locations maintains its own file servers, mail servers and other application servers.

Bob Kent, a user at the branch office, edits a Word document in his home directory on the file server and appends a paragraph to the original 1 MB file. During the modification, the Microsoft Word application creates a temporary copy of the file and saves the file five times. The user then sends the modified file to his team consisting of ten colleagues (who are also in the branch site) and a manager at headquarters. The total traffic on the WAN link generated due to this activity is the sending of 1.1 MB file to the manager at headquarters.

Now assume that the servers of the company have been consolidated at headquarters for various cost, compliance, and administration reasons. The same act of opening a file on the file server, editing it and sending it to eleven recipients will result in approximately 30 MB of traffic over the WAN link (see table 1).

We can see that a simple activity can cause the WAN traffic to increase about thirty times once the servers have been consolidated. In more complex examples e.g. when recipients start to respond to this email message the increase in bandwidth use can be significantly higher than that.

Bandwidth required for initial download – 1 MB
Bandwidth required for intermediate saves – $1.1 * 5 = 5.5$ MB
Bandwidth required to send email to eleven recipients – $1.1 * 11 = 12.1$ MB
Bandwidth required for emails to come back to ten recipients at branch office – $1.1 * 10 = 11$
Total bandwidth used = 29.6 MB

Table 1 – Impact of server consolidation on the WAN bandwidth utilization.

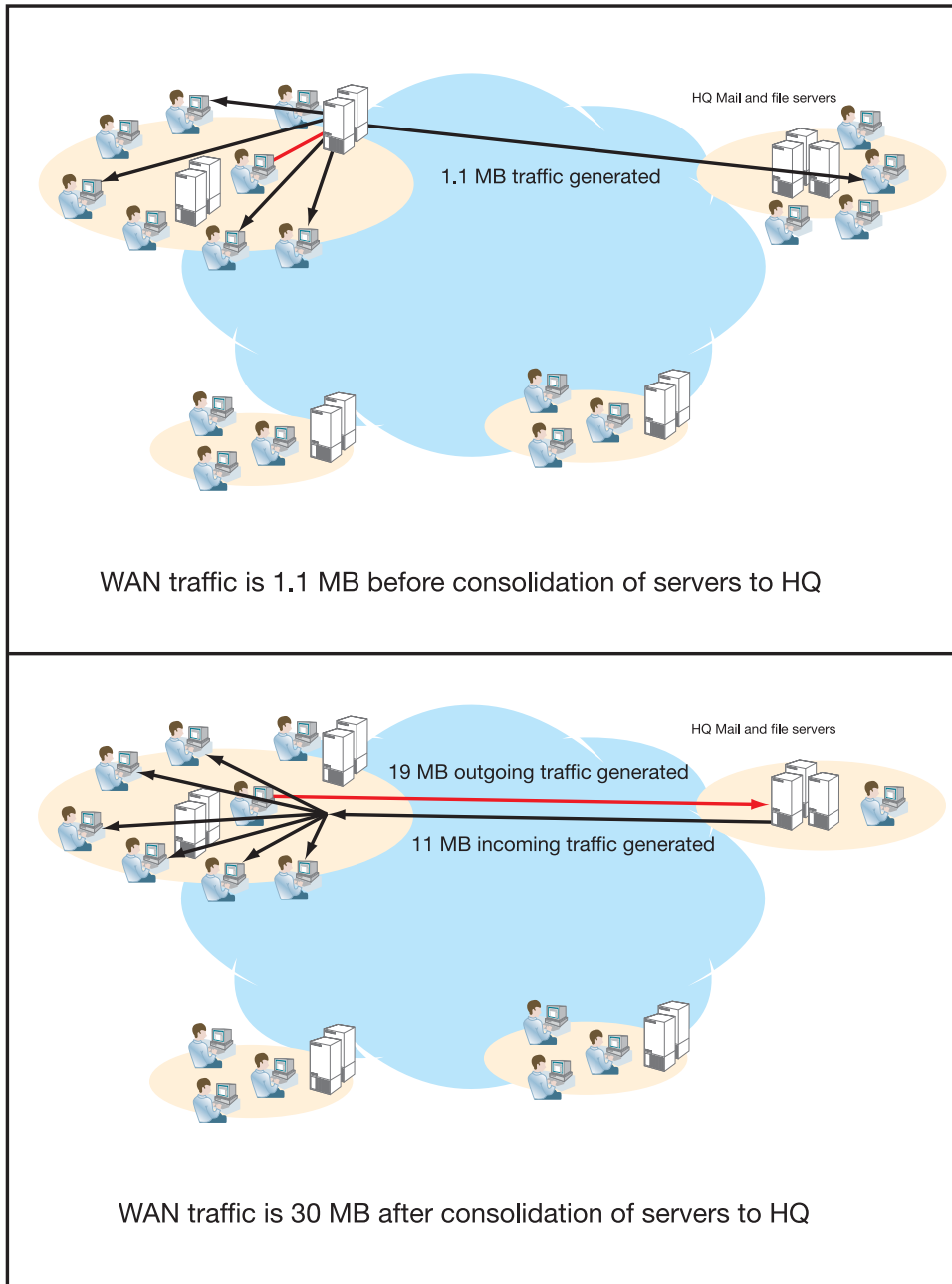


Figure 1 – Impact of server consolidation on WAN bandwidth utilization.

## CACHING CAN SPEED UP BUSINESS

Caching technologies improve response times and WAN utilization by keeping copies of data locally at the remote office locations and serving data locally when a request for the data is seen. Blue Coat's MACH5 technology features two different and complementary caching mechanisms that work together to provide superior application acceleration – object caching and byte caching.

### Object Caching

Object caching has been around for several years and was used mainly to accelerate access to web content. Blue Coat has extended this paradigm to include other kinds of content (e.g. CIFS-accessed files, FTP-accessed files, HTTPS, and video or streaming media objects). Occasionally, object caching is referred to as “proxy caching” since it is implemented using a proxy for the given protocol (e.g. HTTP, HTTPS, FTP, CIFS, or RTSP/RTP).

## How object caching works

The mechanism of object caching is simple. The client sends the request for an object (file, document, image etc.) to the server; this request is intercepted by the proxy that is pretending to be the server itself. Upon receiving the request, the proxy checks to see if it has a fresh copy of the object requested in its cache. If a copy is available, the proxy responds by sending the cached object to the client, else it relays the request to the server. The response from the server is cached by the proxy for responding to future requests from clients. Although the mechanism of the proxy is simple, it poses several challenges to the implementers – most significantly, content freshness and storage.

## Content Freshness

The copies of the cached object have to be kept fresh, or serious data integrity issues ensue. Because content on servers changes, an object cache must keep its temporary store of content up to date. Traditionally, for a proxy to deliver content to the end user with the confidence that the data is fresh, it must send a "refresh check" to the origin server. However, to serve the content quickly, it must not wait until a user requests the content before it performs this refreshing activity. If the refresh checks are performed only at the moment the user requests the content, the user will endure the round-trip delays that cause the Internet to be slow in the first place and application response time will not be substantially improved. Blue Coat's object caching technology uses intelligent adaptive refresh algorithms to guarantee freshness of the content – without negatively impacting the network with unnecessary refresh checks.

## Storage

Another challenge is storage, since storing millions of application objects on a general purpose file system may not be efficient and may result in added latency due to disk reads. The method of storing objects on disk is critical for achieving both high performance and high scalability. It determines (1) how quickly a cached object can be accessed when a client requests it, (2) how rapidly new objects can be acquired and stored on disk, and (3) the rate at which client requests can be serviced per disk drive.

The Blue Coat's SGOS object storage system is not a file system. It is an object cache. There is no directory in the OS. Object access is through a hash table in RAM, ensuring that any object can be obtained in a single disk read. File systems run poorly when they are full, while a cache achieves its highest performance when it is full. Blue Coat's proxy appliance normally runs with its disks full of objects. Old, seldom-used objects are continually removed to make room for new incoming objects. The disk layout and replacement algorithms in the OS facilitate this process to optimize the speed of writing new objects to disk.

## What is it good for?

Object caching is a perfect technology for the following applications

- Content does not change very often such as images, logos and some documents.
- Content can be pre-populated on the remote devices before any user tries to access it e.g. E-Learning, multimedia applications.
- Several users need access to same files and the files do not change often.

When object caching is appropriate, it results in zero WAN roundtrips for the content that is cached locally. In this scenario, users experience vast improvements in application performance, and WAN link is utilized much better as compared to other technologies. However, dynamic content generated by applications may not see any performance improvement. If the two objects differ by only one byte, the whole object still needs to be downloaded over the WAN link.

## Byte Caching

Where object caching falls short, byte caching takes over. Byte Caching caches TCP traffic irrespective of application level protocol, port, or IP address, on both ends of the WAN link.

### How byte caching works

Two Blue Coat SG appliances are deployed, one on each end of the WAN link. Both the SG appliances maintain a cache of all TCP traffic over the WAN link. Each time data needs to be sent over the WAN link, it is scanned for duplicate segments in the cache. If any duplicates are found, the duplicate data (as much as 64KB), is removed from the byte sequence, and a token and a reference to the cache location is inserted (typically a 12 byte package). On the receiving end, the token and the reference are removed from the byte sequence and original data is inserted by reading from the cache thus creating a byte sequence identical to the original byte sequence.

### How does it help?

Up to 90% of WAN traffic is repetitive. The reason for this is that most of the enterprise traffic comprises of the following

1. Web application traffic – All the users at the branch use same or similar web applications. Each interaction with these applications results in WAN traffic that is marginally different than the traffic for previous interactions resulting in resending of same bytes as before.
2. File server traffic – File traffic makes several round trips over the WAN while the user is working on a file. The typical office applications save copies of the file at small intervals resending only slightly modified versions of the same document over the WAN link.
3. Email traffic – Enterprise emails are frequently addressed to several people. For each recipient in the branch a copy of email travels over the WAN. Replies to emails contain repetitive data resulting in further redundant traffic over the WAN.

By eliminating redundant traffic we can expect to effectively increase WAN capacity 10 to 20 times depending on the application.

Blue Coat byte caching works at the TCP layer and does not depend on any knowledge of the application to cache the traffic. This allows byte caching to handle traffic for all applications.

Since byte caching works at the TCP layer, its deployment does not require any changes to the applications themselves or application configuration. This satisfies the requirement that it should be transparent to applications and users. At the same time, Blue Coat's policy engine provides the capability to associate the data with specific applications and users providing control over what data gets cached and what gets blocked. While byte caching accelerates all TCP traffic, some of the specific application protocols that can be accelerated using byte caching are –

Web – HTTP, HTTPS (SSL)

Streaming media – Video on demand

Email – MAPI, POP and any other email protocols

File services – CIFS, NFS and any other file services

Byte caching is designed to work in a mesh, hierarchical or any arbitrary network design and imposes no limitations on the design of corporate networks. Byte caching works well with all of the other MACH5 technologies: object caching, bandwidth management, protocol optimization, and compression.

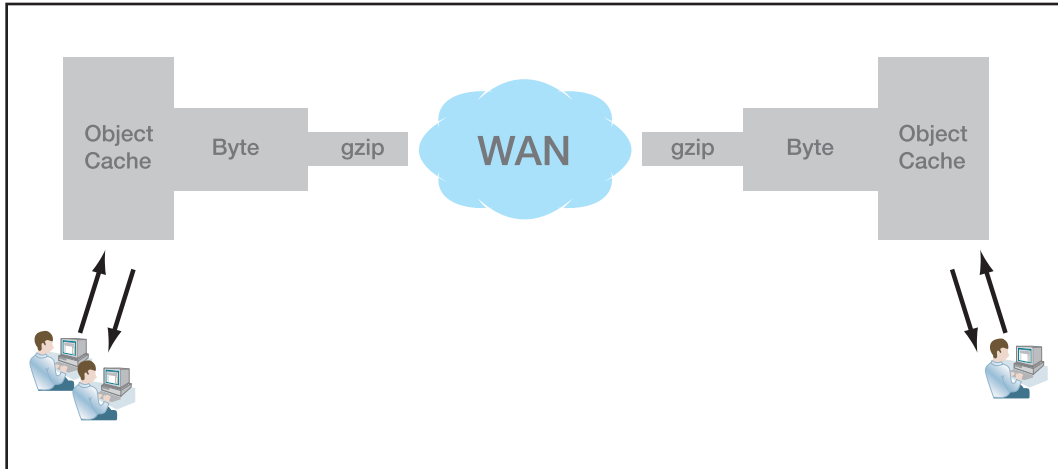
## A comparison of object and byte caching

Object Caching	Byte Caching
Provides bandwidth savings for static content e.g. images, streaming media audio and video, file server objects (files) that do not change often.	Provides bandwidth savings for dynamic content e.g. dynamic web pages, frequently updated files, multi recipient email conversations etc.
Maximum reduction of latency that can be achieved using any technology for the cached content.	First response latency reduction is minimal. Congestion-induced latency for further data is reduced if the data is served from the cache.
Works for only some applications Web - HTTP, HTTPS, FTP Streaming Media - RTSP, MMS File services – CIFS Email services - MAPI	Works at TCP layer so all application level protocols are accelerated.
Uses object name (file name, URL etc.) to reference an object. Same object with a different name will not be recognized in the cache and will be fetched over the WAN.	Works across protocols, objects and if there is common data across different files (e.g. and HTML that was generated from a word document) the data will be served from cache irrespective of the application level protocol requesting it and irrespective of the object name. Does not require any changes to existing applications

Both object and byte caching technologies have their strong points and both are suitable in different situations.

### Object and byte caching together decongest the WAN and reduce latency

Object and byte caching, when applied to enterprise WAN traffic, result in acceleration of enterprise applications and reduction of WAN bandwidth requirements. Together, these technologies cover a wide set of applications and form the backbone of a solution for WAN challenges that enterprises face with the consolidation of their branch office server infrastructure. In addition to providing coverage for a wide set of applications, synergy between the two approaches to caching makes the overall system better than the sum of its parts e.g. existence of byte caching results in transfer of fewer bytes if the object gets changed/modified at the source. This property helps object caching subsystem to be aggressive on adaptive refresh algorithms for keeping the content fresh in the local cache and thus improving the response latency in the branch office. Similarly, existence of fresh content in the object cache results in fewer TCP round trips by the byte caching subsystem since the request can be served locally.



*Figure 2 – Object caching, byte caching and gzip compression when applied to the WAN traffic reduces bandwidth requirements and latency significantly*

### Edge Corp Revisited

Now that Edge Corporation has implemented Blue Coat SG appliances in its consolidated server environment, the activity of opening the file may not initiate any WAN traffic, since the data for the whole file is in the branch SG object cache. Further, since the file is opened locally, Bob Kent does not experience any of the delays users have come to expect when opening files over a WAN link. On each save operation only the incremental data, plus a handful of references (say 100 KB) will be sent across the WAN link. Finally, the act of sending the file to all of Bob's co-workers over email will not result in any WAN traffic, since byte caches on both ends already has all the file data. Thus, the total traffic over the WAN will be less than one MB – a thirty-fold improvement. Moreover, with MACH5, all communications will be compressed using the gzip algorithm, which, on average, results in 20%-30% additional bandwidth savings.

## **CONCLUSION**

Blue Coat's MACH5 technologies and specifically, the caching and compression technologies provide the foundation for achieving several fold bandwidth savings required to keep up with the exponential increase in enterprise WAN traffic. These technologies work well for all applications and complement the other MACH5 acceleration technologies – bandwidth management, protocol optimization. Like all of MACH5 technologies, Blue Coat's caching and compression technologies do not require changes to any applications and can be deployed in a variety of corporate network architectures.



420 North Mary Ave.  
Sunnyvale, CA 94085  
[www.bluecoat.com](http://www.bluecoat.com)

1.866.30.BCOAT  
408.220.2200 Direct  
408.220.2250 Fax

Copyright ©2006 Blue Coat Systems, Inc. All rights reserved worldwide. No part of this document may be reproduced by any means nor translated to any electronic medium without the written consent of Blue Coat Systems, Inc. Specifications are subject to change without notice. Information contained in this document is believed to be accurate and reliable, however, Blue Coat Systems, Inc. assumes no responsibility for its use, Blue Coat is a registered trademark of Blue Coat Systems, Inc. in the U.S. and worldwide. All other trademarks mentioned in this document are the property of their respective owners. Version 1.0

Blue Coat secures Web communications and accelerates business applications across the distributed enterprise. Blue Coat's family of appliances and client-based solutions – deployed in branch offices, Internet gateways, end points, and data centers – provide intelligent points of policy-based control enabling IT organizations to optimize security and accelerate performance for all users and applications. Blue Coat is headquartered in Sunnyvale, California, and can be reached at 408.220.2200 or [www.bluecoat.com](http://www.bluecoat.com).