The BIG-IP System With Intelligent Compression: Cutting Application Delivery Time and Optimizing Bandwidth

Overview

The number and complexity of applications delivered over the Internet continues to grow. Available bandwidth is stretched to capacity by the sheer volume and chattiness of Internet protocols. The results are long delays, increased latency, unsatisfactory end-user experiences, and unacceptable application performance that costs organizations real dollars in terms of customer reputation and productivity losses.

Compression technology can provide dramatic application performance improvements. This white paper discusses the need for compression, different approaches for compressing web traffic available in the market today, and how F5's BIG-IP® system provides organizations with a powerful way to optimize their bandwidth intelligently while accelerating the delivery of their applications.

What is Compression?

Compression is an optimization technique used to remove redundant patterns from a data stream so that it has fewer packets and consumes less bandwidth, significantly improving application performance. Since there are fewer packets traversing the network from the server to the end user, application data gets delivered faster. HTTP compression is commonly used for web applications which helps reduce the bandwidth consumed by web objects and significantly improves end user response times.

GZIP is a popular HTTP compression technique applied to web traffic and is supported by standard browsers. Most browsers have been equipped to support the HTTP 1.1 standard known as "content-encoding." With GZIP, the client browser essentially negotiates with the server indicating that it can accept encoded data. Upon successful negotiation, GZIP compresses data being sent out from the web server using the encoding format accepted by the client. In the figures below, Message 1 shows the client request for compression and Message 2 shows the server accepting the GZIP compression request:

```
GET / HTTP/1.1
Host: www.f5.com
User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.5)
  Gecko/20031007 Firebird/0.7
Accept: text/xml,application/xml,application/xhtml+xml,text/html;q=0.9,
text/plain;q=0.8,video/x-mng,image/png,image/jpeg,image/gif;q=0.2,*/*;q=0.1
Accept-Language: en-us,en;q=0.5
Accept-Encoding: gzip, deflate
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
Keep-Alive: 300
Connection: keep-alive
```

Message 1: Client Request for Compression

```
HTTP/1.1 200 OK
Date: Thu, 04 Dec 2003 16:15:12 GMT
Server: Apache/2.0
Vary: Accept-Encoding
Content-Encoding: gzip
Cache-Control: max-age=300
Expires: Thu, 04 Dec 2003 16:20:12 GMT
X-Guru: basic-knowledge=0, general-knowledge=0.2, complete-omnipotence=0.99
Content-Length: 1533
Content-Type: text/html; charset=ISO-8859-1
```

Message 2: Server Response Acknowledging Compression Use

Once the compression type is established, the session will begin conversing using the agreed upon compression technique. The browser will decompress data received by the server on the fly, reducing the amount of data sent and increasing the page display speed.

There are two ways to compress data coming from a web server -- dynamically and pre-compressed. Dynamic Content Acceleration typically compresses transmission data such as HTML, XML, CSS, Java, JavaScript, and WML on the fly. Dynamic Content Acceleration is particularly useful in e-commerce applications, database-driven sites, and many other applications. Pre-compressed data is text based data that is generated beforehand and stored on the server in html.gz files or other formats. Since the data is pre-compressed, CPU load is reduced as compression/decompression is not done on the fly.

Challenge

Why is Compression Needed?

The volume of web traffic triples almost every year as applications become increasingly 'webified'. With this exponential increase in demand for bandwidth, combined with low bandwidth client connection types, organizations are experiencing high network latency and poor application response types. To solve application performance problems, the traditional solution for network managers has been to throw more bandwidth at the problem. This approach does not scale well as it requires network managers to repeat the process as application needs grow or when the next 'big' application gets deployed. Additionally, bandwidth prices have not declined as rapidly as expected and adding network capacity is more expensive than implementing compression. Let's look in more detail at some of the drivers for compression:

- Client access speeds and the last mile: Due to the different connectivity access methods, varying bandwidth availability and complexity of the routing protocols, application traffic is subject to many constraints which cause its performance to degrade over time. The cascading and multiple effects of all these variables makes the application of "smart compression" extremely useful. For instance, because of low-bandwidth availability and high latency, dial-up or 'satellite clients' experience the worst response times which can literally render an application unusable. The ability to only compress dial-up and satellite traffic will dramatically reduce CPU loading requirements on the switch, allowing for greater efficiency at a lower price. Organizations that inadvertently compress traffic from broadband clients may prevent optimizing the benefits of compression.

- Network throughput and bandwidth limitation: Organizations are faced with the challenge of using their existing bandwidth intelligently. Web application object sizes have been steadily growing and organizations find themselves encountering bandwidth bottlenecks on a regular basis. This has the effect of slowing down other applications as they get starved for bandwidth which introduces very high latency. Delay-sensitive applications like VoIP cannot tolerate such high latency and become unusable.

Where Do I Compress My Application Traffic?

Although the benefits of compression are obvious, the location where compression is applied to application traffic plays a very important role in determining the efficacy of the compression technique and the overall improvement in application delivery time. There are two mainstream approaches to compressing web application traffic:

- ***Compression on the server***
  This commonly used approach involves compressing all application traffic before it leaves the server. The server can be used to compress static as well as dynamic content. Many popular web servers such as Microsoft's IIS and Apache support server side compression. This approach to compression works well for a small number of applications but does not scale well as the number of applications and the size of application objects grow. Also, this approach makes it harder to change or add compression parameters as this has to be done across multiple servers, increasing management overhead. Compression is a very processor

intensive function. Compressing application traffic on the server consumes valuable CPU cycles and degrades server performance by introducing latency in application response times. To exacerbate the problem further, more components of application delivery such as SSL encryption, client authentication, security, etc. are also being run on the servers, adding even more overhead on the CPU. As the servers become a choke point, the benefits of compression are not fully realized and application delivery times will fail to show a dramatic improvement.

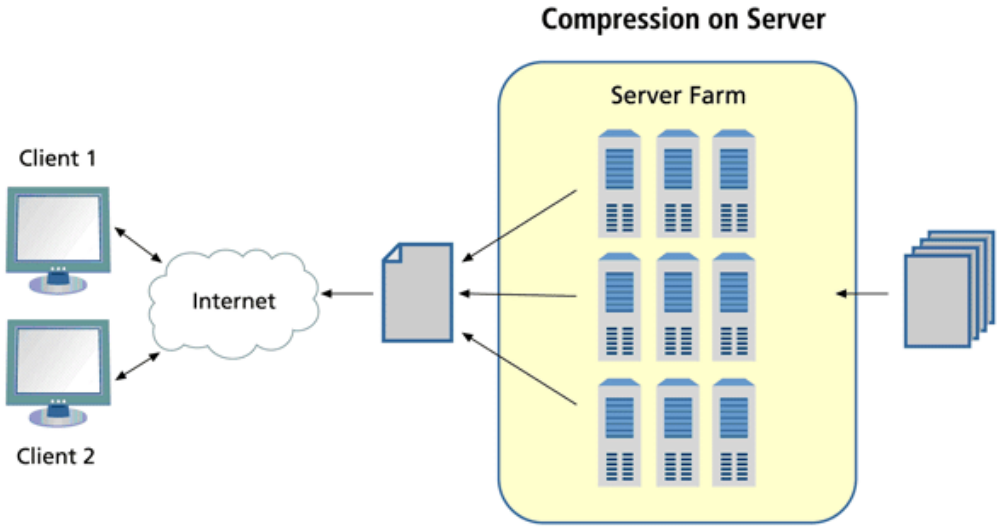

Figure 1: *Compression on the server*

- ***Compression on traffic management devices***
  A new means to compression involves offloading compression onto an application traffic management device that front-ends the servers. Application traffic can now be compressed by the devices on behalf of the web server, thus removing the CPU bottleneck imposed by compression on those servers. This approach is gaining in popularity since organizations can now realize the full benefits of compression. There are two ways to achieve this:

  - Symmetric: This approach involves offloading the compression function to a device deployed in front of the web servers with a second device or a software component at the client end. This end-to-end proprietary approach is suited for a branch office deployment but does not work well for web applications that are accessed by millions of clients on the Internet. Deployment and administration of this type of solution is cumbersome, as downloading or pushing a piece of software from the server to the client is intrusive and may compromise security.

  - Asymmetric: This approach involves offloading the compression function to a traffic management device that can be deployed in front of web servers. The decompression is done by browsers at the client end. This approach takes advantage of the existing decompression capabilities that are a part of all standard browsers and requires no changes at the client end. It also eliminates the majority of browser compatibility issues since the device now acts as a mediator and translates between the client browser and server. This approach is gaining popularity as organizations are able to realize the full benefits of compression by saving valuable CPU cycles on their servers as well as obviating the need for any changes to the client infrastructure or intrusive downloads. This approach also allows organizations to centralize their management of all compression-related configurations and eliminates overhead that is associated with managing multiple servers.

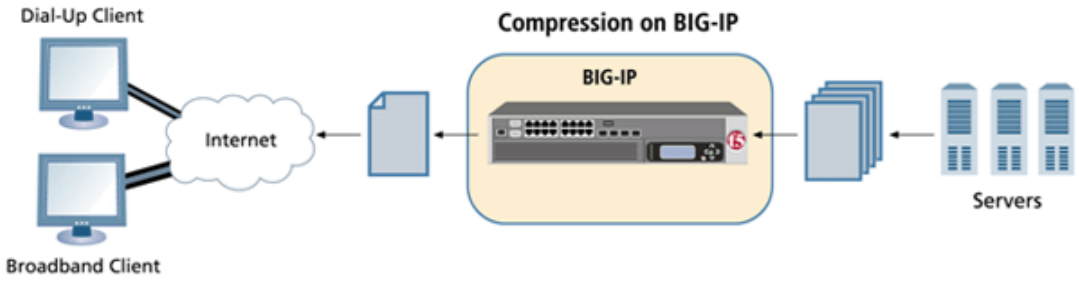

Figure 2: *Compression on the BIG-IP device*

Solution **F5's BIG-IP Cuts Application Delivery Time and Optimizes Bandwidth**

F5's BIG-IP® application traffic management system provides the industry's most scaleable, intelligent and flexible compression solution. By asymmetrically offloading HTTP compression from servers, the BIG-IP system reduces server overhead and decreases total cost of ownership for servers up to 65% by enabling server consolidation. The BIG-IP system takes advantage of existing browser decompression capabilities and obviates the need for any client side changes or intrusive downloads. BIG-IP's Intelligent Compression uses a patent pending approach to determine client connection latency, decreasing bandwidth usage by 60-80% while improving end user response times by over 200%. The BIG-IP solution is the first in the industry to provide organizations with a scaleable compression solution with the option of compressing web traffic through optimized hardware with its Adaptive Compression Offload feature.

What Is Intelligent Compression?

The BIG-IP system's *Intelligent Compression* provides organizations a way to target users for compression. Compressing all types of traffic does not necessarily yield a reduction in bandwidth usage. The challenge with compression is to know how to target it most efficiently, so users receive the most benefit. For example, a dial-up client will benefit most from the compression since it has higher latency while the benefit to a broadband client would be insignificant. This is because broadband clients have larger receive-window sizes. Compression causes response times to degrade as clients now have to wait longer to receive data, negating some of the benefits of compression. BIG-IP uses a patent pending technique to dynamically determine the client connection latency. The BIG-IP system monitors the TCP RTT (Round Trip Times) to dynamically calculate user latency, allowing BIG-IP to devote more power to compressing traffic to users who need it most.

Tunable Compression Engine

BIG-IP provides organizations the capability to fine tune their compression parameters to get the maximum benefit. They can target system resources for compression where they matter most (achieving a higher ROI) as well as achieving better control when compression is invoked. The following compression parameters can be configured:

- ***Minimum content-length for compression***
  This parameter specifies the minimum length of the server response (in bytes) to be considered acceptable for compression.

- ***Tunable compression***
  This parameter allows specification of the desired compression ratio, CPU and memory utilization.

F5's *Adaptive Compression Offload* enables organizations to "tune" compression which boosts bandwidth savings by freeing up valuable system resources. By adaptively offloading HTTP compression from the system processor to the optimized hardware when the system CPU reaches a certain threshold, BIG-IP can deliver unprecedented levels of compression throughput (up to 2 Gbps) and minimize system processor overhead up to 80%.

Granular L7 Policy Based Compression

The BIG-IP system provides the industry's most granular solution to control how and what type of traffic is compressed, delivering better performance and improved ROI. Organizations can configure compression per virtual server, source IP, destination IP, file type or protocol, or based on any Layer 7 variable. By leveraging the BIG-IP system's unique iRules capability, customers can choose to enable compression based on these granular L7 inspection criteria, allowing organizations to turn compression on or off for an individual HTTP request/response pair.

Content Filters and Exception Handling

BIG-IP also provides organizations predefined filters that they can use to target desired content types and also handle exceptions. In order for a server response to be compressed, users may define "include" and "exclude" lists to better target compression or quickly handle exceptions. Such predefined filtering capability includes:

- **URI (from the client request)**
  This is a list of regular expressions used to match the Request-URI part of the client request line. For example, to include requests ending in ".txt", ".htm" and ".html", one would use the following in the URI include field: ".*\.txt" ".*\.htm" ".*\.html".

- **Content Types (from the server response)**
  This is a list of regular expressions representing MIME types. The regular expressions will be checked against the value of the server's "Content-Type:" header. For example, to disable compressing PDFs and all image files, one would use the following in the content type exclude field: "application/pdf" "image/.*". To include all text types, one would use "text/.*" in the content type include field. To include all non-CSS text types, you would use "text/(?!css$)" in the content type include field.

Visibility Into Compression Performance

The BIG-IP device provides rich statistics to monitor compression performance and to demonstrate key benefits to the organization. These statistics help organizations tune their compression policies and measure the benefits of compression depending on the type of object being compressed and provide visibility into bandwidth savings. These statistics include:

- Size of the object before compression
- Size of the object after compression
- Object-type being compressed

Content Type Compression	Pre-Compress	Post-Compress
HTML	1.8G	199.9M
CSS	709.9M	132.9M
JS	777.9M	231.3M
XML	885.2M	273.6M
SGML	0	0
Plain	835.3M	327.0M
Image	0	0
Video	0	0
Other	0	0
Total	5.0G	1.1G

Figure 3: *BIG-IP compression statistics*

ROI Example

Compression ROI has a direct impact on cost that is relatively easy to calculate. Let us consider an organization that has the following bandwidth pricing structure:

Connectivity	Recurring Cost/Month	Bandwidth
Frame Relay – Tier 1	*$1300*	*0-2 Mbps*
Frame Relay – Tier 2	*$2200*	*2-3 Mbps*
Total	*$3500*	

Figure 4: *Example bandwidth pricing structure*

Assuming the organization experiences traffic spikes and bandwidth usage of 2.5 Mbps that goes over into Tier 2, the monthly charges are $3500. If the organization is to be able to reduce the bandwidth usage below 2 Mbps, however, the organization won't get penalized every month for bandwidth oversubscription. Compression can help the company reach this objective. BIG-IP provides various compression modules at different bandwidth rates (5 Mbps, 100Mbps, 500 Mbps, 1000Mbps) to match the customer needs. In the case above, deployment of a BIG-IP compression solution resulted in the following ROI:

Bandwidth Costs Before Compression	**$3500**
Reduction in Bandwidth (assuming 2:1)	**1.25 Mbps**
Bandwidth Costs with Compression (assuming 50%, 1.25 Mbps)	**$1300**
Monthly Savings	**$2200**
BIG-IP 5 Mbps Compression Module	**$2995**
Time to Return on Investment	**Less than 2 months**

Figure 5: *ROI results using compression on the BIG-IP device*

Conclusion

The BIG-IP device's Intelligent Compression capability provides a market leading approach to optimize bandwidth and accelerate the delivery of applications from the server to the end-user. With increasing application performance challenges (bandwidth bottlenecks, delays, timeouts and outages), loss of revenue and customer dissatisfaction is becoming commonplace and the need to solve these problems has become imperative. BIG-IP's compression solution enables organizations to solve their application performance problems by delivering the following benefits:

- The BIG-IP system uses sophisticated bandwidth optimization techniques such as Intelligent Compression to reduce latency and improve end-user access and page download times by over 200% and improve performance by decreasing bandwidth usage by 60-80%.
- Unlike the legacy server-side compression model, BIG-IP offloads server overhead and decreases total cost of ownership of servers up to 65%.
- Unlike the legacy server-side compression model, BIG-IP centralizes compression management and eliminates browser incompatibilities by acting as a mediator between the client and the server.
- Unlike symmetric compression devices, BIG-IP's HTTP compression inherently takes advantage of the decompression capabilities on the client browser and obviates the need for any changes to the infrastructure on the client side.
- BIG-IP provide optional compression offload ASICs which work to further scale and offload compression cycles from the infrastructure.
- Used in conjunction with BIG-IP's multiplexing, caching and TCP optimization features, BIG-IP's compression delivers added end user performance improvement and bandwidth availability.

About F5

F5 enables organizations to successfully deliver business-critical applications and gives them the greatest level of agility to stay ahead of growing business demands. As the pioneer and global leader in Application Traffic Management, F5 continues to lead the industry by driving more intelligence into the network to deliver advanced application agility. F5 products ensure the secure and optimized delivery of applications to any user - anywhere. Through its flexible and cohesive architecture, F5 delivers unmatched value by dramatically improving the way organizations serve their employees, customers and constituents, while lowering operational costs. Over 9,000 organizations and service providers worldwide trust F5 to keep their businesses running. The company is headquartered in Seattle, Washington with offices worldwide. For more information go to www.f5.com.