# Benefits of BIG-IP Application Optimization Over the WAN

## *Overview*

Web-enabled applications have become commonplace in today's enterprise organizations. It is not unusual for a single company to have hundreds of different web-based applications on the network. As deploying these applications becomes easier, the focus is shifting from the applications to security and access concerns. One of the main challenges with web-based applications is trying to keep up with the exponential increase of users accessing these applications remotely. Not only does this introduce new security concerns, but remote users demand the same level of performance as they experience when connecting to applications via the LAN. Solving performance and security issues through customizing each application is not only extremely expensive, but an inefficient use of time and resources. With its strategic location in the network infrastructure, the new BIG-IP system provides a quick, easy and much more cost-effective solution to these challenges.

The BIG-IP system version 9.x includes a number of features that facilitate the optimization and acceleration of application traffic. The F5 Networks Solution Center conducted extensive testing to measure the performance benefit that the BIG-IP system brings to applications and web sites over the Internet. These tests demonstrate the unique capabilities of the BIG-IP device's application acceleration technology, which differentiate it from the rest of industry. While other vendors have made claims about the ability of their products to improve application performance, their testing has been done in a self-contained lab environment. This leaves users the difficult task of trying to determine how well those results translate into real world performance gains. That is the question F5 set out to answer. And we came to the conclusion that the only way to test the real performance seen by users over the Internet was to actually use the Internet.

In order to carry out these types of real tests, F5 chose the Gomez® Performance Network (described in detail in the following section) to test end-to-end transactions. On average, the BIG-IP system improved the end-user response time by two times or more, reduced bandwidth utilization by greater than **75%**, reduced the occurrence of browser timeouts over slow links by more than **80%**, and offloaded up to **98%** of connections from the servers.

## *Solution*

### Speeding Application and End User Response Times

The BIG-IP device contains targeted and specialized optimizations which leverage F5's unique WAN, LAN and data acceleration technologies. This allows the BIG-IP system to deliver unmatched optimization, packet loss recovery, and a more intelligent intermediation between suboptimal servers and clients. Table 1 shows response time improvement for some typical applications.

| Application | Not Optimized | BIG-IP Optimized | Improvement |
|---|---|---|---|
| BEA WebLogic Portal 8.1 | 38.209 seconds | 17.291 seconds | 121% (2.2x) |
| Microsoft IIS 6.0 | 21.09 seconds | 12.39 seconds | 70% (1.7x) |
| Microsoft Outlook Web Access 2003 | 32.88 seconds | 21.23 seconds | 55% (1.3x) |
| Microsoft SharePoint Portal Services 2003 | 40.6 seconds | 18.06 seconds | 125% (2.2x) |
| Siebel Business Applications 7.7 | 88.157 seconds | 33.989 seconds | 126% (2.3x) |

***Table 1:*** *Sample Improvements for Full Application Transactions*

### Improving Server Farm Capacity

In addition to providing better user experiences, the BIG-IP device was also shown to significantly improve the scalability of the surrounding infrastructure.

Using the new **Fast Cache** feature, the BIG-IP system offloaded an average of **36%** of the content serving duties and **95%** of the TCP connections from backend servers; providing dramatic performance and scalability gains for an existing infrastructure.

Through these and other optimization techniques, we found that with the BIG-IP system in the network, servers were generally able to handle twice the workload.

**Increasing Bandwidth Efficiency**
Without optimizations, organizations can often pass only a fraction of the bandwidth they have purchased because of lower protocol and WAN inefficiencies. Through the Traffic Management Operating System (TM/OS) and TCP Express feature set, testing revealed that the BIG-IP device greatly improves the bandwidth efficiency for a site. For example, using the BIG-IP system, tests showed a **224% average increase of data placed on the wire (3.2 times)** and a **50% average packet reduction** on the wire. Overall, the BIG-IP system provided a **322% (4 times) average increase in bandwidth utilization efficiency**.

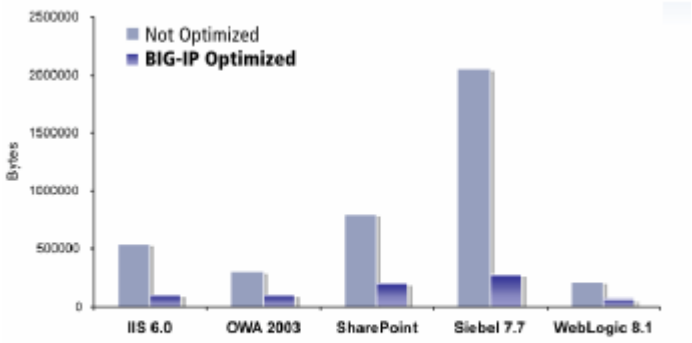


**Figure 1**: *Sample bandwidth reductions using Intelligent Compression and TCP Optimizations*

**Improving the Reliability of WAN Connections**
Using the TCP Express feature set, tests showed that the BIG-IP system reduced the number of TCP timeouts and resets seen by clients, by up to 50%. This was especially beneficial for transmissions over high-loss networks or for clients using low bandwidth connections, such as dial-up.

## *Real World Testing Using the Gomez Performance Network*

When F5 set out to test the performance of our application optimization capabilities in a way that would provide value to our customers, we had to find a trusted 3rd party service that would allow us to access a diverse global population of users with varying degrees of network bandwidth and different operating environments. After a review of the available solutions, we determined the Gomez Performance Network (GPN) represented our best option.

The award-winning Gomez Performance Network is made up of over 10,000 computers around the world using actual end-user Internet connections. This tool is used by many of our enterprise customers and has won numerous awards, including Editor's Choice from Network Computing. Their network was the test tool that provided F5 with the means to gather a fair sampling of realistic end-user performance gains over the Internet.

By using GPN, the F5 Solution Center test results are much more realistic than the marketing materials available by others in the industry. It's very easy to create tests that show ten-times, twenty-times or more improvement by carefully selecting the pages tested and data used within a dynamic application in lab environments. But what good are these numbers if they do not represent a scenario likely to be encountered by the overwhelming majority of users? We concluded these lab-generated results are not what our current and prospective customers would find most useful when trying to select an application acceleration product – people want to see what performance improvement their end-users will likely experience.

F5 decided the best way to demonstrate improved application performance would be to choose applications users would commonly see over the WAN: we chose Portal and Collaboration applications such as BEA® WebLogic®, Siebel® Business Applications and Microsoft® SharePoint®. We also included Microsoft Outlook® Web Access as a key application users

would be using over the WAN, and a generic web site (www.f5.com) served using Microsoft Internet Information Services (IIS) to get an idea of general performance gains that can be seen for mostly static content. While organizations may be running a myriad of applications, the results seen here are good general indicators of application gains that can be realized through this technology.

For each test, we measured the response time of an end-user's full experience with an application, averaged over several days and thousands of repeated tests. For example, the script we employed for testing Microsoft Outlook Web Access included login, message retrieval, calendar viewing, and so on. As you can see, this is completely different than single-page download times and represents end-to-end application response time for a full application transaction, as a user would actually experience it.

### *Test Methodology*

Gomez offers a service that allows customers to monitor the performance of a web site as seen by Gomez monitoring agents that are running on real user's computers throughout the Internet and around the world. Gomez pays businesses and end-users to loan their computer and bandwidth resources to Gomez. Gomez in turn installs an agent on the computer that allows them to instruct it to connect to designated sites. At regular intervals, Gomez instructs a certain number of computers to request a particular web site and measure every aspect of its performance. This information is then returned to the Gomez Performance Network so that the customer can view how well the target web site performed for that end-user.

In addition to choosing the requested web site, Gomez allows customers to select which user population (or "peer population") is used and gives the ability to specify the type of client connections they expect to service for performance metric collection. For example, F5 selected a variety of regions the clients originated in, and the bandwidth category they were a part of. Gomez supplies a recorder to facilitate the configuration of the example customer transaction used by the clients, which allows clients to specify which links are selected on each web site, what login credentials are used (if applicable), and other transaction settings needed to simulate a real user experience. Once the sample user session is configured using their provided recording tool, this session is replayed by the agents chosen on the Gomez Performance Network.

While testing the real end-user response time of applications over the Internet, we found the following to be the most important aspects of controlling the Gomez Performance Network:

- Last Mile tests – these test use real clients around the world.
- Defining bandwidth categories that represent the target audience and geography.
- The key metrics "End-to-End" response time and "Errors"– the other statistics helped narrow down any problems.
- Recording example user sessions instead of doing single page downloads – capturing a typical user session instead of a simple page download.
- Averaging the resulting data over several days to trend what average really is, and working to improve that average by modifying the application or acquiring performance enhancing technology.
- The Gomez agents support all standard browser behaviors and features, but are limited in their ability to accept 3rd party plug-ins, just like many end-users – the application must be capable of handling this type of secure client which may not necessarily be able accept plug-ins.

Our test parameters were fairly simple; we chose users from Global Dialup, US Dialup, and US Low Broadband. For each application, we used a sample user session that involved steps similar to the following:

1. Go to the main page.
2. Click the button to enter login credentials.
3. Enter login credentials.
4. Click link A.
5. Click link B.

This is only a high-level example; in practice the steps were varied to meet a typical user experience for each application.

Once Gomez was configured with our example user population and we had recorded a sample user experience, we then used that sample user experience to request performance metrics for two application instances simultaneously. One application instance was accelerated using the BIG-IP system, and the other application instance was run through a typical server load balancer (SLB) doing only basic switching with no acceleration. Both instances were housed at the same location with identical configuration and hardware. With all settings in place, the Gomez agent was left to run for about 2,000 iterations over several days to get a representative sample.

## *Features and Benefits of the BIG-IP Optimization Technology*

The tests and results described in this document apply to the BIG-IP Application Accelerator products and to the BIG-IP Local Traffic Management device with the appropriate acceleration and optimization features enabled. The following is a list of these features and the benefits they provide.

***Fast Cache: An in-memory RAM web cache***

- Offloads requests from servers by serving common web pages / included objects / images on their behalf.
- Speeds page download times even further when combined with Intelligent Compression by allowing compressed objects to be cached and served without the latency of repeatedly recompressing the same object.
- Stores both compressed and uncompressed content at the same time, and intelligently serves up the correct page based on what the client will accept.
- Fully RFC2616 compliant.
- Caches a variety of server response codes: 200, 203, 206, 300, 301 or 410.
- Can respond to conditional GET and HEAD requests on the server's behalf.
- Allows for the setup of multiple dedicated cache repositories on a single system ("Multi-Store") to direct caching resources to priority applications on a shared system.
- Supports iRules, F5's advanced programming language, for superior control and management over cacheable content.

***Intelligent Compression: Agent-less compression offload***

- Reduces page download times – fewer packets, fewer round trip times. Reduces bandwidth consumption – serves the same number of users with less bandwidth.
- Ability to target compression only to dial-up clients or those users coming in over high latency (long distance) connections (Patent Pending).
- Natively supported by every web browser created in the last five years – no plug-ins or additional software of any kind is required.
- Offloads server cycles from the server tier and centralizes management for compression processing, providing a more lower cost, safe (via granular client targeting) and manageable solution.

***TCP Express: Industry-leading, state-of-the-art TCP optimizations***

- Numerous WAN and LAN optimizations techniques that speed transmission of data according to various client and network conditions.
- Hundreds of real-world TCP interoperability improvements between commercially available stacks (Windows 98, XP, 2000, IBM AIX, Sun Solaris and more).
- Centralized / Clientless WAN Optimization includes both symmetrical and asymmetrical optimizations without client downloads or branch device.
- Based on open standard TCP optimizations:
  - Delayed and Selective Acknowledgments (RFC 2018): Increases performance when dealing with lost and reordered packets on WANs.
  - Explicit Congestion Notification (RFC 3168): Allows the BIG-IP system to proactively signal peers that the intermediate routers are being overloaded so they can back-off and avoid packet loss.
  - Limited and Fast Re-Transmits (RFC 3042 and RFC 2582): Allows for efficient re-transmission of lost data to eliminate the effects of timeouts from packet loss.
  - Adaptive Initial Congestion Windows (RFC 3390): Studies show a 30% gain for HTTP transfers over satellite links and 10% improvement for 28.8 bps dial up, with no accompanying increase in drop rate.
  - Slow Start with Congestion Avoidance (RFC 2581).
  - TCP Slow Start (RFC 3390): Allows for increased bandwidth utilization across links for higher throughput rates on existing public Internet connections and leased lines.
  - Bandwidth Delay Control: Improved and expanded bandwidth delay calculation estimates the optimal load to be placed on the network without exceeding it.
  - TimeStamps and Windows Scaling (RFC 1323): BIG-IP allows for selective use of timestamps which add data to the TCP segment to aid with other optimizations.

***L7 Rate Shaping: Ability to prioritize bandwidth usage***

- Can classify (select) traffic with L2 – L7 information using iRules – this means traffic can be tracked based on MAC address, IP information, or L7 information such as HTTP cookies.
- Contains hierarchical Rate Classes which allow for parent-child bandwidth borrowing relationship – for example, multiple FTP customers might have unique rate policies, but if resources are available it might be desirable to allow borrowing.
- Can classify traffic independently for inbound and outbound.
- Supports Priority FIFO and Stochastic Fair Queue queuing disciplines.
- Configurable bursting, borrowing, and ceiling rates.
- Can be dynamically applied to virtual servers, within iRules.

***SSL Offload: Decrypts SSL so your servers don't have to***

- Increases server capacity – by offloading resource-intensive SSL processing from your servers, they will have more resources to handle their true business-specific objective.

- Faster page download times – F5 has the latest in high-performance cryptographic ASICs that are able to handle encrypted data many times faster than general purpose server CPU's.
- The BIG-IP Application Accelerator device handles up to 5,000 TPS, and the powerful BIG-IP LTM 6800 platform can perform up to 20,000 TPS. Provides unique support for new connection and bulk encryption of data within specialized hardware offload engines.
- Can intelligently rewrite HTTP redirects from HTTP to HTTPS to help seamlessly integrate SSL into your existing HTTP applications.
- Tight integration with iRules allows flexible policy decisions based on encryption strength, client certificates, and any other SSL information.
- Allows for deep application traffic inspection and modification of previously encrypted traffic.

***OneConnect™ TCP Offload: Reduces TCP overhead by multiplexing HTTP requests***

- Aggregates multiple client-side connections into fewer server side connections.
- Does not delay or queue requests – keeps enough server-side connections open to handle all connections simultaneously.
- Transforms HTTP headers to encourage long-lived server-side connections.
- Handles client and server connections independently – this allows for incredible efficiency that can turn millions of connections into only a few hundred that your back-end servers have to handle.

***Content Spooling - Server Buffering***

- Able to read responses from the server as fast as they can transmit, eliminating the burden of having to directly communicate with slow clients.
- Offloads retransmit processing, and optimizes individual flows to get the best performance for each end-user, sending data as quickly as they can receive it.
- With data more quickly read and spooled from the server, the server is then free to process more connections, which increases server capacity.