



courtesy of
F5 NETWORKS

f.y.i. F5 | guide 3

Improving Web Application Response Time
for Remote and Mobile Users



Improving Web Application Response Time with Server Offload Technologies

This guide provides an overview on how adding web acceleration technologies such as server offload technologies can improve application performance and user experience, and what to look for when considering such options.

Server Offload Technologies – What To Look For

One of the main problems affecting application performance for remote users is that the servers are being overwhelmed by computational tasks they weren't designed to handle. Servers are too busy serving the same data over and over again, there are too many connections to back-end servers, and there's an overflow of connections to the back-end servers that adversely affect their performance.

One of the primary reasons for this slow performance is that serving up commonly used web page objects burdens the servers and slows down performance. One of F5's web acceleration technologies called Dynamic Caching serves up objects and commonly used images from the cache at wire speed, thus unburdening the servers. Dynamic Caching caches unchanging data that may seem dynamic (containing query parameters, etags, and session ids) but is actually static data or changes in an identifiable pattern. F5 web acceleration technology can cache a higher percentage of data from dynamic web applications while maintaining proper application behavior. It accomplishes this by fully inspecting every aspect of HTTP requests, controlling caching behavior, and invalidating cached data.

Also, make sure the web acceleration device you choose can recognize what part of that object is static and what part of that object is dynamic. This gives you the ability to understand what data is specific to a user, and what data can be stored for all users. From the standpoint, your organization can cache and compress data that's deterministically static and then choose to serve it up when you need to.

Offloading Compression

Compression computations drain a lot of CPU on servers. Being able to offload those computations onto another box at wire speeds is another important factor to consider when selecting web acceleration solutions. By taking a document as it comes back from the server and reducing it to a smaller size, it also reduces the amount of bandwidth. But more important than bandwidth reduction is this simple fact: if you're sending less data across the wire, **it takes less time to get there**. From that standpoint, compression is a double win. It both works in the server offload, and also in the network and application offload, making it so the user response time is faster by getting the data across faster because there is less to transfer.

With data more quickly read and spooled from the server, the server is then free to process more connections, which increases server capacity.

In addition, F5 web acceleration technologies are able to read responses from the server as fast as they can transmit, eliminating the burden of having to directly communicate with slow clients. They work by offloading re-transmit processing, and optimizing 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.

Eliminating Slow Page Load Times

In addition, look for a web acceleration product that includes extensive connection management as well as TCP and content offloading capabilities to optimize server performance and speed page load times. F5 connection pooling (known as OneConnect) works by aggregating millions of requests into hundreds of server-side connections, ensuring requests are handled efficiently by the backend system. The result is increased server capacity, with intelligent load balancing maintained to dedicated content servers.

Rate Shaping

F5 web acceleration technology can classify (select) traffic with L2 – L7 information using rules, enabling traffic to be tracked based on MAC address, IP information, or L7 information such as HTTP cookies. Hierarchical Rate Classes 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.

F5 web acceleration technology also supports Priority FIFO and Stochastic Fair Queue queuing disciplines, and provides configurable bursting, borrowing, and ceiling rates all important for controlling and prioritizing bandwidth usage while reducing costs.



For additional guides from our User Experience/Acceleration series, go to <http://learn.f5.com>, email us at resources@f5.com, or call 888-88BIGIP (888-882-4447).

Guide 1 – Overview | Guide 2 – Network Application | Guide 4 – Application Acceleration