

# Understanding SpeedCache™

## Array Networks Server Side Caching Technology

---

### White Paper

#### Overview

As businesses become dependent on the Internet for additional customers and revenue, it is critical for them to run high performance Web-enabled applications that support maximum availability and minimal download times. SpeedCache is yet another Array Networks innovation helping enterprises optimize application delivery and streamline network infrastructure. Learn about:

- SpeedCache's advanced server side caching features and benefits
- Common server side caching deployment options
- How Array appliances deal with scalability, redundancy, and cache management issues

## Array Networks Server Side Caching Technology

### → Introduction

The purpose of this white paper is to describe the advanced Web caching features and benefits that Array Networks AFE appliances offer. This paper will also outline common server side caching deployment options and how Array appliances deal with scalability, redundancy, and cache management issues.

### → Problem Overview

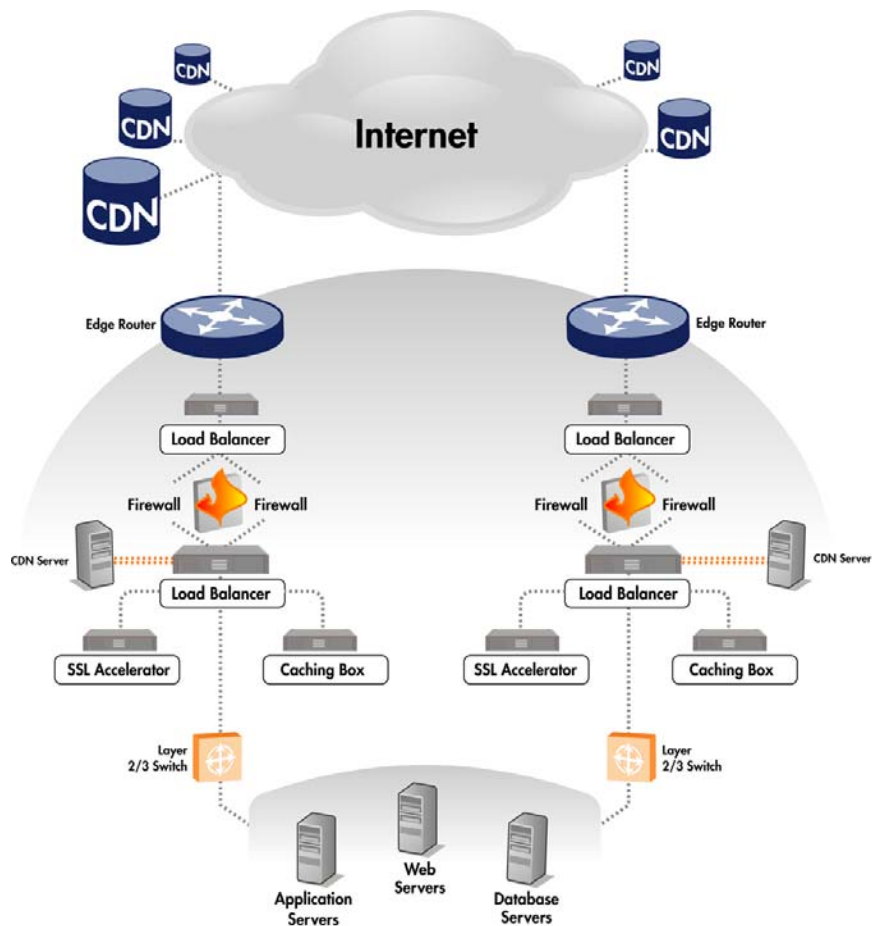
The explosive growth of the Internet has created latency problems for end-users accessing Web sites and Web-based applications. As businesses are becoming more dependent on the Internet for additional customers and revenue, it becomes critical for them to run high performance Websites and enterprise applications that support maximum availability and minimal download times.

To be successful on the Web, businesses must deliver increasingly complex content to a growing number of end users. In many cases, infrastructure has not been properly designed for scalability, performance or redundancy. Web infrastructure can perform poorly due to:

- **Non-scalable, complex networking infrastructure** – Network administrators typically add more web servers and bandwidth in order to improve response times. In addition to Web servers, server load balancers, global server load balancers, firewalls, firewall load balancers, and SSL accelerators; caches are added to further enhance performance (Figure 1). In most cases, these devices are not optimized and not specifically designed for the enterprise's web traffic scenario, resulting in less than ideal performance. Another problem is it is very difficult to scale the infrastructure for performance and redundancy. If one device breaks down in the infrastructure, it could drastically reduce the performance of the entire site or even bring the site down.
- **Non-optimized Web servers** - Web server software runs on general-purpose operating systems and hardware that have not been totally optimized to deliver Web server content.
- **Web server CPU load** - Web servers must perform a number of other CPU-intensive activities in addition to serving Web pages. Many Web servers today must also generate dynamic pages and process SSL connections.

## Array Networks Server Side Caching Technology

### Current Internet Infrastructure



### → Caching Overview

When a Web user requests specific information, such as a web site, the cache feature is used to direct the request to the origin server, then return the data back across the Internet. Web caches process the delivered data and store the content. When another request is made for that same data, there is no need to send the request all the way through the network; instead, the response is sent from the cache memory. This way, duplicated requests are responded to more quickly while also preventing the network from being bogged down by multiple requests for identical information.

A Web cache also keeps track of whether a Web object is fresh or not. If a Web object is no longer fresh or past its expiration date, a Web cache would delete it from its storage or replace it with a newer object from the origin Web server.

There are two kinds of Web caches: client side cache and server side cache. Client side caches are typically deployed between a Web user's browser and the Internet. The idea of client side cache is to reduce the bandwidth consumption on the WAN (Wide Area Network). Client side caches store objects normally being requested by the Web browser.

## Array Networks Server Side Caching Technology

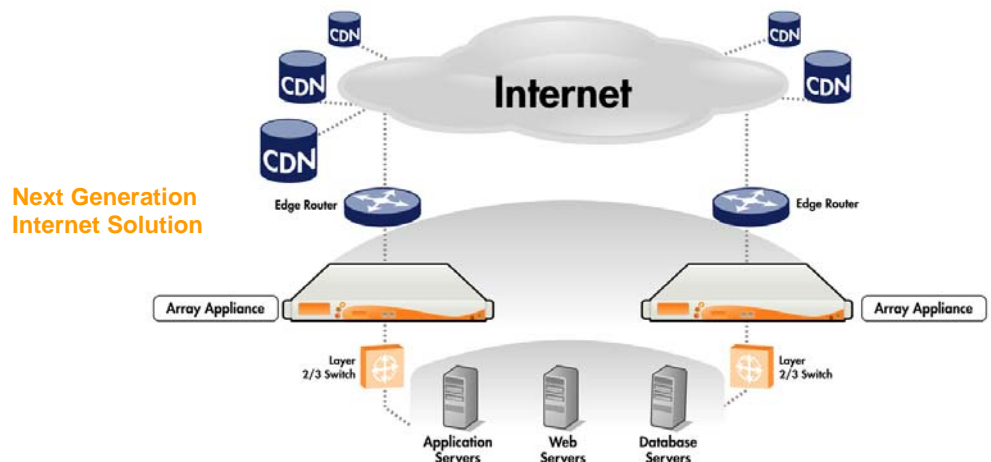
For example, if everyone in a company goes to a financial Web site everyday for the same financial report, a client side cache would store that report in its storage and serve it to all the users. In this scenario, Web users don't have to go across the Internet to get the report from the origin server. As a result, WAN bandwidth usage is reduced.

The second type of cache is the server side cache, which is designed to reduce the number of requests handled by back end Web servers. Server side caches are deployed between the Web server and the client. If a request from a browser is for a Web object that's stored in the server side cache's storage, then the server side cache would serve the object from its storage instead of passing it on the origin Web server. The idea is to reduce the load on the Web server, while providing Web users with faster access to web pages.

Another important feature for caches is the amount of working cache size or amount of space available for storing cacheable Web objects. In most caches, a combination of disk and RAM are used for storing Web objects. A cache can serve Web objects stored on its RAM much faster than if the Web objects are stored in the disk drive.

### → The Array Networks Solution

Array Networks AFE appliances offer a solution that simplifies the management of complex web infrastructures. The Array appliance is designed to sit between routers, switches and the Web server farm (Figure 2). In addition to the high performance server side caching function, the Array appliance also incorporates important data center center functions like server load balancing, global server load balancing, SSL acceleration, clustering and Web firewall into one appliance.



There are several advantages to this design: integration, scalability, and ease of management.

## Array Networks Server Side Caching Technology

### → High Performance Server Side Caching

The Array appliance's built-in server side caching function is geared for performance. Specifically designed for server side caching, the Array appliance can offload a tremendous amount of work from back-end Web servers. Array AFE appliances are capable of delivering industry leading 1.5 Gbps of total throughput, while setting up over 50,000 new requests a second. In addition, Array AFEs can handle up to 2 million concurrent connections depending on physical memory available.

Today, many competitors use disk drives to store cacheable objects; disk access is extremely slow when compared to memory access. The Array appliance uses high performing RAM, instead of disk drive, for all cache storage. By using only memory for cache storage, the Array appliance can get to all the cached Web objects more quickly.

*Array Networks' patented design sends cached Web objects in Ethernet frames, instead of packets, from memory, which makes request processing even faster.*

Cache working size is the amount of space available for caching. Array AFEs are capable of using up to 2 GB of cache working size.

Another important function, which speeds up overall performance, is the ability of the Array appliance to maintain persistent server connections to the back end servers to minimize the TCP setup time. Network administrators can configure the number of persistent server connections to maintain to each back end server. When a request for a new URL for an existing server comes in, the Array appliance will use an existing persistent TCP connection to connect to the back end Web server. This saves time on each new client request, since the entire TCP connection setup procedure is avoided.

A high performance server side cache means fewer Web servers in the back, which translates to greater savings. Since Array appliances are more cost effective than high-end Web servers, and much easier to maintain, it makes even more sense to install more Array appliances instead of additional Web servers.

### → High Performance Server Side Caching

Array Networks' caching implementation is based on HTTP 1.1 specification, RFC 2616. As a result, Array Networks' caching functionality is fully compliant with RFC 2616. For detailed definitions of various header tags described below, please refer to RFC 2616.

## Array Networks Server Side Caching Technology

Network administrators have total control over on what is cacheable and what is not on a Website. For example, Web content is not cached if there is a “Cache-control: no store” header or “Authentication: “ header present in the URL. The Array appliance will not respond to a request from its cache if the following information is in the URL header:

- **“Cookie:” header**  
This is being sent back to the correct Web server using Array Networks’ cookie-persistence feature
- **“Range:” header**
- **“Cache-Control: no-cache” header**
- **If the following headers are in the URL request: “If-Modified-Since:”, “If-Unmodified-Since:”, “If-Match”, and “If-None-Match:”**  
The date and time stamp are being tracked using “Etag:“ and “Date:“ headers from cached responses. As a result, these requests are sent back to the back end Web servers.

The Array will continue to monitor the cached entry to maintain freshness using the following headers in the URL: “Cache-Control: s-max-age”, “Cache-Control: max-age”, “Expires:“, “Cache-Control: min-fresh”, and “Cache-control: max-stale”.

The Array appliance will cache all Web objects that are cacheable. This includes objects on static Web pages and common objects on dynamically generated pages.

### → Cache Preloading

One important feature of the Array Networks appliance is the ability to preload a cache with Web server content. Network administrators can specify a URL, how many levels down from the current URL, and maximum amount of cache working size in an appliance for preloading. The Array appliance will stop the preload if all the specified URLs are preloaded or if the maximum specified cache working size is reached, whichever comes first. This feature can dramatically increase the performance of a Web site if the network administrator knows certain URLs are popular with site visitors. Instead of letting the cache figure out what to cache over time, this can speed up the process tremendously.

### → Scalable Performance and High Availability

One of the many problems network administrators face is the ability to scale infrastructure in both performance and availability. As infrastructure usage increase, the ability to add performance is required in order to serve Web content and applications in a timely manner.

## Array Networks Server Side Caching Technology

In the case of today's infrastructure design (Figure 1), additional load balancers, global load balancers, caches, firewalls, compression appliances, and SSL accelerators are required. This creates a configuration nightmare for network administrators in addition to the pain of troubleshooting and managing a complex infrastructure. In the past, no single vendor provided all the essential performance and high availability functionalities required for today's Web environment; network administrators had to work with multiple vendors who deliver different devices, to achieve a complete solution. This method proved to be very difficult and costly to scale for performance and high availability.

Array Networks' solution is simple to scale and manage. As additional performance is required, a network administrator simply adds additional Array appliances to the cluster. Unlike many competing products today, there are no dedicated wires required for redundancy. Once an Array appliance is in a cluster, it automatically load-balances itself among other Array appliances in the cluster for both performance and redundancy. Currently, up to 32 Array AFE appliances can be clustered together in an N+1 active-active configuration, a feat unrivaled by any other manufacturer today. Performance can be scaled up to 16 Gbps of total throughput, which is more than capable of handling the largest of today's enterprise Web infrastructures. Another important function is the ability to provide additional redundancy as more Array appliances are added to the cluster. Today, most competitors' products can't go beyond more than two appliances for redundancy. Array appliances can scale up to 32 AFEs for redundancy. If one Array fails for any reason within the cluster, other appliances can immediately take over without bringing the network down.

### → Advantages of Integration

In order to increase the performance of Web content and Web-based applications, many other networking functions besides caching are required. For example, many more Web servers are required when higher performance is required. As a result, server load balancers are required in order to distribute the workload among all the servers. If a Website has multiple mirrored sites at different locations, global server load balancers are required to load balance traffic between all those data centers. If security is a critical issue, multiple firewalls are required to protect the network. As secured transactions increase on a Website, dedicated SSL accelerators are needed to offload that functionality from much slower Web servers. Once these networking functions are incorporated into the Web site infrastructure, scalable redundancy and performance become important as the user base of the Web resources grows. In the past, multiple Web devices were required to provide all of those functions. As a result, Web infrastructure became difficult to scale and manage, and required excessive rack space and power.

## Array Networks Server Side Caching Technology

The Array appliance is designed to solve all these problems, and more. It provides server load balancing, global server load balancing, hardware SSL acceleration, an application firewall, server side caching, clustering and CDN-rewrite functions in one single, simple, easy to manage appliance. Instead of managing multiple devices from multiple vendors separately, network administrators can perform and maintain all Web networking functions using a single user interface on one single appliance. Array Networks' built-in CLI and WebUI make management easy and quick. The Array appliance can also be managed using SSL, SSH and SNMP. Another huge factor is limited space and power consumption when dealing with one appliance instead of several.

*In fact, deploying a Array Networks solution when compared to today's traditional multiple server deployments can result in a 64% savings in hardware cost, a 78% savings in operations cost and a 59% savings in maintenance costs.*

### → Summary

Array Networks AFE appliances are next generation Web caches that incorporate all essential data center networking functions into one single, simple, easy to manage device. In addition to industry-leading, high performance server side caching, Array appliances can be used as a server load balancer, a global server load balancer, an application firewall, a compression appliance, an SSL accelerator, and a CDN rewriter. Array's advanced clustering features makes our AFEs extremely useful for networks that require scalable performance and redundancy, and an integrated design saves network administrators time and money while eliminating management headaches.

## Array Networks Server Side Caching Technology

### *About Array Networks*

Array Networks is a world leader in secure application acceleration and deployment appliances for global enterprises. Built upon the Array SpeedStack(TM) technology, Array's unified secure content access solutions enable industry-leading performance, integration, scalability and ease of implementation and management. Headquartered in Campbell, California with sales offices in the U.S., Europe, Asia Pacific and Latin America, Array engineers and manufactures its products in the Silicon Valley and sells them through direct and indirect channels across the globe.

#### *Array Networks, Inc.*

*254 East Hacienda Avenue*

*Campbell, CA 95008*

*Phone: (408) 378-6800*

*Toll Free: 1-866-MY-ARRAY*

*Fax: (408) 874-2753*

*Email: [info@arraynetworks.net](mailto:info@arraynetworks.net)*

*[www.arraynetworks.net](http://www.arraynetworks.net)*