

Global Server Load Balancing

White Paper

Overview

Many enterprises attempt to scale Web and network capacity by deploying additional servers and increased infrastructure at a single location, but centralized architectures are subject to a number of inherent limitations. Learn about:

- Issues associated with centralized Web architectures and how GSLB overcomes these issues
- How Array's GSLB implementation works and why it stands out from our competitors
- Additional Array load balancing and optimization features designed to enhance Web content and application delivery

Array GSLB Technology White Paper

→ Introduction

Scalability, high availability, and performance are critical to the success of any large commercial Web deployment. While many enterprises attempt to scale capacity by deploying additional servers and infrastructure at a single location, these centralized architectures are subject to a number of inherent limitations. Centralized Web infrastructures present a single point of failure in the Internet topology - if the site loses connectivity to all or part of the public Internet, it will be inaccessible to end users. Quality of service is also highly sensitive to bandwidth bottlenecks and congestion in the vicinity of the site. Furthermore, users accessing the site from geographically or topologically distant locations may experience large and highly variable delays, which are exacerbated by the large number of round trips that HTTP requires to transfer some Web content. The centralized architecture is also not appropriate for international companies that must serve localized content to users in different parts of the world.

Global Server Load Balancing (GSLB) overcomes these problems by distributing Web traffic among a collection of servers deployed in multiple geographic locations. By serving content from several different points in the Internet topology, GSLB alleviates the impact of network bandwidth bottlenecks and provides robustness in case of local server or network failures at particular server sites. Users can be directed to the nearest or least loaded server site at the time of the request, minimizing the likelihood of long download delays and service disruptions. Web usage studies have shown that fast and reliable access to content and applications is critical for online businesses to succeed, being that end users are notoriously impatient, and failure to respond within seven seconds can cause at least 30 percent of users to abandon the site.

Although the rate of growth of the Internet has slowed somewhat within the United States, demand for Web content continues to increase sharply in other countries and the traffic distribution at large sites is becoming increasingly global. An effective GSLB solution is needed to provide high availability and performance to potentially millions of Web users across multiple continents. Array Networks provides a compact and elegant GSLB solution integrated into our Application Front End next-generation traffic management and load balancing appliances.

Array GSLB Technology White Paper

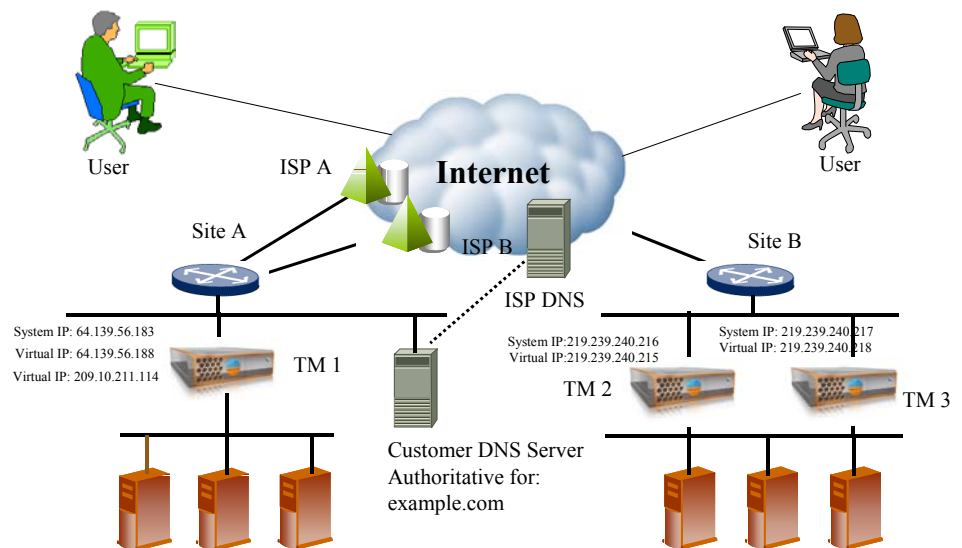
→ Array Networks Solution

With the Array Networks solution, Global Server Load Balancing (GSLB) is an integrated part of our Application Front End appliances. The AFE also provides local server load balancing (SLB), link load balancing (LLB), SSL acceleration, compression, reverse proxy caching, and an application firewall, among other features. By combining GSLB with SLB and LLB, Array AFE appliances provide a complete load balancing solution for large distributed Web infrastructures.

→ How GSLB Works

Several mirror sites (referred to as GSLB sites) need to be deployed at geographically distant sites. At each site, Array appliances advertise one or more virtual IP addresses (VIPs), each corresponding to one or more of the Web domains served by the site.

GSLB works by having Array AFEs handle DNS queries for domains corresponding to virtual sites and returning for each query a VIP serving the requested domain on one of the GSLB sites.



For each query, GSLB selects the best site, as well as the most suitable VIP within the selected site, based on any one of a variety of load balancing algorithms.

Array GSLB Technology White Paper

→ Further Detail

Each GSLB site contains one Array appliance that is acting as the *site master* at any given time. When a user requests a Web page, their Web browser first queries their local DNS server to resolve the domain name to an IP address. The local DNS server recursively queries a root DNS server, which responds with the address of an authoritative DNS server for the requested domain.

Depending on the DNS delegation arrangements used for the requested domain, this authoritative DNS server may either be one of the GSLB site masters, or another DNS server which in turn delegates the domain to one of the site masters.

After a response to a DNS query is generated, it is usually cached at intermediate DNS servers on the response path, as well as on the user's Web browser. Web browsers typically cache DNS entries for several minutes, while most DNS servers cache each entry for a duration indicated by the time-to-live (TTL) value specified by the authoritative DNS server. Cached responses are used for subsequent queries for the same domain name; thus, the GSLB system generally does not receive additional DNS queries when additional URLs are requested from a domain that was recently visited by the same user, or by another user sharing a common DNS server.

While caching reduces end-user latency as well as the load on the distributed site's DNS servers, it also reduces the site's control over incoming HTTP traffic. For these and other reasons, effective global server load balancing requires more sophisticated techniques than the standard methods used by most local SLB products.

→ GSLB Techniques

Array Networks AFE appliances support a wide range of global load balancing algorithms and provide the flexibility to choose the methods most appropriate for the enterprise's needs. When a site master receives a DNS query, it returns the IP address of an Array appliance serving the requested domain. The site master first selects a GSLB site using the configured global load balancing method. It then selects the least loaded appliance within that site serving the requested domain. The following global load balancing methods are available:

- **Weighted Round Robin:** the site master selects sites in a fixed ratio defined by the site weights
- **Weighted Least Connections:** the site master selects the site having the smallest number of open connections relative to the site weight

Array GSLB Technology White Paper

- **Administrative Priority:** the order of preference of the sites is manually specified. For example, one site might be configured as the primary site, with another site serving as a backup in case the primary site becomes unavailable or overloaded.
- **Geography:** the site master selects the most appropriate site based on the geographic location of the client. The Array atlas resolves the geographic location of IP addresses at the continent, country, and state/region levels.
- **Proximity:** the site master selects the site closest to the client, according to network metrics collected by the Array atlas.

Please note that each DNS query may be followed by one or more HTTP requests to the selected site, and naïve techniques for load balancing DNS queries do not necessarily result in effective load balancing of HTTP traffic. Thus, adaptive dynamic mechanisms are needed for stable and robust global load balancing. Load conditions on each GSLB site are continuously monitored by the site masters. Site overload is automatically detected if the number of open connections exceeds a threshold, or if the site's response time exceeds a maximum duration. If any site becomes overloaded, the site masters will automatically divert new traffic away from the site until the load has decreased to acceptable levels.

→ High Availability and Redundancy

Array Networks' GSLB solution is based on an architecture designed to support a high level of availability and redundancy. This infrastructure provides automatic fault detection and transparent fail-over of each component, and contains no single points of failure. Since each GSLB site is capable of answering DNS queries, loss of connectivity to one site does not render the entire distributed Web site inaccessible. This design is significantly more robust than many existing GSLB solutions, where the authoritative DNS for the distributed site runs on a special appliance deployed at one of the GSLB sites. In the event that a GSLB site master goes down, the system will detect the event and assign the site master's IP address to another Array appliance in the same site. This appliance will then assume responsibility for responding to incoming DNS queries and communicating with the other site masters.

GSLB also uses application-level health checks to monitor the end-to-end availability of each VIP in the system. If the services associated with a VIP become unavailable, the VIP is removed from the system and subsequent requests for the affected domain are redirected to an available VIP on the same site or on a different site.

Array GSLB Technology White Paper

→ Content Management for Distributed Sites

Finding the best site for each requesting user is not the only problem that must be solved in order to operate a distributed Web infrastructure. In addition, large enterprises must distribute and synchronize content between servers at the various GSLB sites in order to provide a consistent, unified global Web presence. For most commercial applications, the content replication mechanism must be reliable, secure, and highly scalable. Modern commercial Web sites often contain gigabytes of data with frequent updates, making manual and naïve automated solutions all but impractical. Array AFE appliances provide the capability to manage and schedule the replication and synchronization of content among all origin servers at all GSLB sites.

→ Summary

Array AFE appliances are a cost effective, high performance, next-generation load balancing and traffic management solution ideally suited to large enterprises. They conveniently integrate global and local server load balancing and link load balancing in an architecture designed for high availability and fault tolerance. Array AFE appliances enhance GSLB functionality with advanced geographic and network intelligence to select the best site for each user based on location as well as load. Additionally, the Array family of AFE products provides the ability to securely and efficiently manage and distribute large collections of content mirrored at multiple distributed sites. Array Networks provides a truly complete solution for running a large distributed Web and network infrastructure - along with the simplicity and ease of administration offered by a highly integrated design.

Array GSLB Technology White Paper

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.

1371 McCarthy Blvd.

Milpitas, CA 95035

Phone: (408) 240-8700

Toll Free: 1-866-MY-ARRAY

Fax: (408) 240-8752

Email: info@arraynetworks.net

www.arraynetworks.net