



Global Server Load Balancing

APV Series Application Delivery Controllers

May 2011

Introduction

Scalability, high availability and performance are critical to the success of any large commercial application deployment. While many enterprises attempt to scale capacity by deploying additional servers and infrastructure at a single location, these centralized deployments are subject to a number of inherent limitations. Centralized Web deployment present a single point of failure for application service delivery - if the site loses connectivity to all or part of the public Internet, it will be inaccessible to users and customers which can have significant impact to the business. 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 distant locations may experience large and highly variable delays, which are exacerbated by the large number of round trips that HTTP requires to transfer content. Centralized architectures are also not appropriate for international companies which must serve localized content to users in different parts of the world.

Global Server Load Balancing (GSLB) overcomes these problems by distributing traffic among a collection of servers deployed in multiple geographic locations. By serving content from many different points in the Internet, GSLB alleviates the impact of network bandwidth bottlenecks and provides robustness in case of local server or network failures at a particular server site. Users can be automatically directed to the nearest or least loaded site at the time of the request, minimizing the likelihood of long download delays and/or service disruptions. 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 an application or service.

With the growth of the mobile internet and proliferation of smart personal devices from smart phones to tablets, the demand for “always-on” connectivity to commercial and business applications continues to accelerate. An effective GSLB solution is needed to provide high availability and performance to potentially millions of users across multiple continents or geographies.

Array Networks’ family of Application Delivery Controllers (ADC) provides an enterprise-proven GSLB solution to meet the performance and availability needs of both enterprise and cloud deployments.

The Array Networks Solution

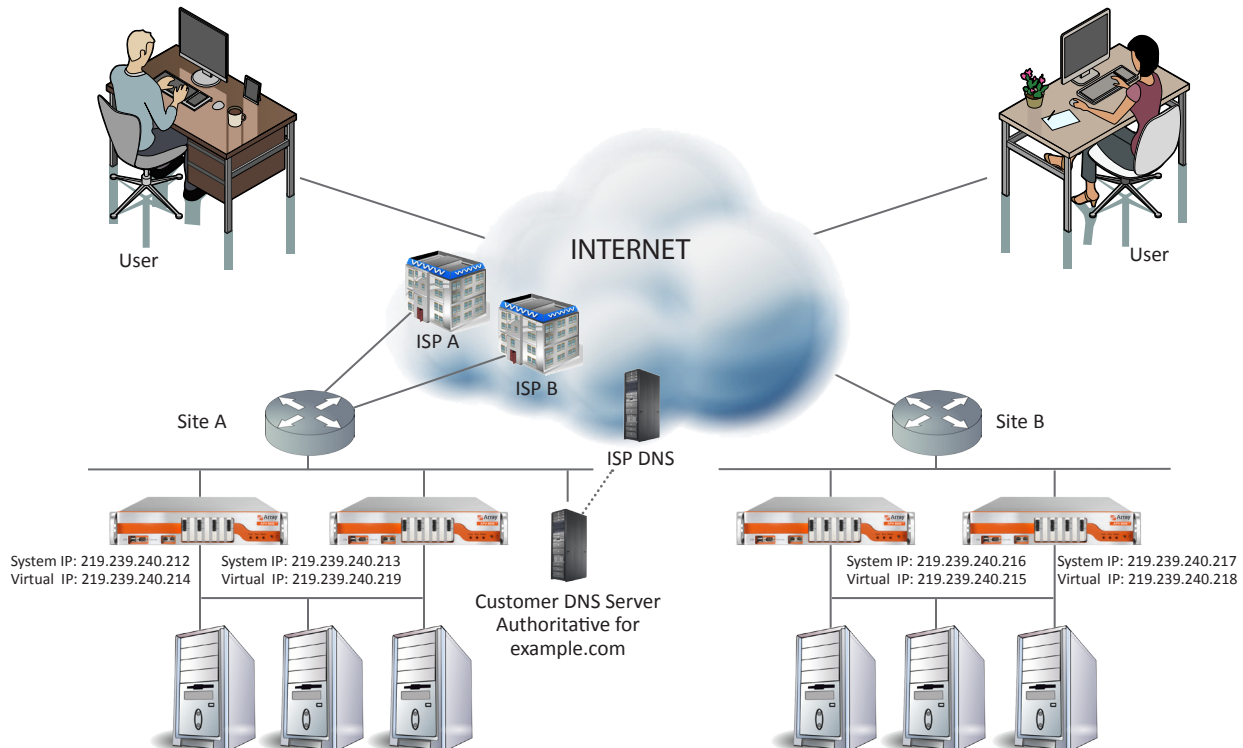
With the Array Networks solution, Global Server Load Balancing (GSLB) is one of the main features of our Application Delivery Controllers. In addition, Array Networks APV series of appliances also provide local server load balancing (SLB), link load balancing (LLB), SSL acceleration (encryption/decryption), compression, reverse proxy, caching and an application firewall, among other features. By combining GSLB with SLB and LLB, Array Networks ADC appliances provide a complete load balancing solution for large distributed Web and cloud applications and services.

How GSLB Works

Two or more mirror sites (referred to as GSLB sites) need to be deployed at geographically separate locations for business continuity, e.g. Tokyo and Las Vegas. At each site, the APV appliances advertise one or more virtual IP addresses (VIPs), each corresponding to one or more of the Web domains or applications served by the site.

GSLB works by having Array Networks ADCs 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.



Further Detail

Each GSLB site contains at least one Array Networks appliance or typically a pair of Array Networks appliances for high availability, which acts as the site master at any given time. When a user requests a Web page, the 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 ADC appliances supports 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 Networks 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.

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 user. The Array Networks 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 Networks atlas or by what was automatically detected. To simplify and reduce the cost of a multi-site GSLB solution, Array Networks supports a standalone software "client agent" which can be installed on any UNIX or Linux server. This eliminates the cost of installing and maintaining appliances at every GSLB member site. This agent can be configured to probe the GSLB site master on a regular basis so the site master is aware of the site's health and proximity.

Global Connection Overflow (GCO): Prevents DNS query overflows and packet drops when overflows occurs. The least loaded server in the query chain is automatically selected to receive the overflow traffic without any packet loss. GCO supports up to 64 APVs in an overflow chain and ensures that servers are never overloaded.

Global Least Connection (GLC): The site master can be configured to select the site with least number of TCP connections for accelerated application delivery.

IP Overflow (IPO): IPO methods can be configured at a member site for resolving domain names to a healthy IP address with the highest priority, increasing availability and optimizing delivery of the content.

Please note that each DNS query may be followed by one or more HTTP requests to the selected site, and 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 and network 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.

Disaster Recovery

Array Networks GSLB includes purpose built Disaster Recovery functions. Multiple sites can be grouped as Primary or Standby sites. Traffic will be directed to Primary sites and only when Primary sites fail will traffic be directed to Standby sites. Customers can also select to fail back automatically or manually when Primary sites are recovered.

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 failover 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 Networks 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 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. The site masters continuously communicate with the GSLB members at other sites for monitoring the VIP availability and build their own dynamic DNS database accordingly.

Summary

Array Networks ADC appliances are a cost effective, high performance, next-generation load balancing and traffic management solution ideally suited for enterprises and cloud services. They conveniently integrate global and local server load balancing and link load balancing in architecture designed for high availability and fault tolerance. Array Networks ADC 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 Networks family of ADC 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.

About Array Networks

Array Networks is a global leader in application, desktop and cloud service delivery with over 5000 worldwide customer deployments. Powered by award-winning SpeedCore™ software, Array solutions are recognized by leading enterprise, service provider and public sector organizations for unmatched performance and total value of ownership. Array is headquartered in Silicon Valley, is backed by over 300 employees worldwide and is a profitable company with strong investors, management and revenue growth. Poised to capitalize on explosive growth in the areas of mobile and cloud computing, analysts and thought leaders including Deloitte, Red Herring and Frost & Sullivan have recognized Array Networks for its technical innovation, operational excellence and market opportunity. To learn more, visit www.arraynetworks.net.

May-2011 rev. a

Corporate Headquarters

Array Networks, Inc.
1371 McCarthy Blvd.
Milpitas, CA 95035
408-240-8700
1 866 MY-ARRAY
arraynetworks.net

ASIA Headquarters Array Networks China (Beijing) Corp., Inc.

Liang Ma Qiao Road,
Chaoyang District,
Beijing, No. 40, the
Twenty-First Century,
10-Story Building,
Room 1001-1017
Post Code: 100016
+010-84446688

EMEA Headquarters

Array Networks UK
4 Cross End
Wavendon
Milton Keynes
MK178AQ
+44 (0) 7717 153 159

To purchase
Array Networks Solutions,
please contact your
Array Networks
representative at
1-866 MY-ARRAY
(692-7729) or
authorized reseller.

Copyright 2011 Array Networks, Inc. All rights reserved. Array Networks, the Array Networks logo, AppVelocity, NetVelocity, ArrayGates, and SpeedCore are all trademarks of Array Networks, Inc. in the United States and other countries. All other trademarks, service marks, registered marks, or registered service marks are the property of their respective owners. Array Networks assumes no responsibility for any inaccuracies in this document. Array Networks reserves the right to change, modify, transfer, or otherwise revise this publication without notice.