White Paper

Global Server Load Balancing

Scalability, High Availability and Performance for Distributed Networks

APV Series Application Delivery Controllers

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 presents 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 on 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 that 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 or service access points 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' APV Series Application Delivery Controllers (ADCs) provide an enterprise-proven GSLB solution to meet the performance and availability needs of both enterprise and cloud deployments.

The Array Networks Solution

Global Server Load Balancing (GSLB) is one of the main features of Array Networks application delivery controllers. In addition, Array Networks APV Series appliances also provide local server load balancing (SLB), link load balancing (LLB), SSL acceleration (encryption/decryption), compression, reverse proxy, caching, security 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 (Content Distribution Network) and cloud applications and services.

How GSLB Works

Two or more mirrored sites (referred to as GSLB sites) need to be deployed at geographically separate locations for business continuity, for example Tokyo, London and Las Vegas. Each site has Service IP(s) to serve the Web domain and sub-domain names.

Array Networks APV Series Smart DNS (SDNS) provides two major functions for global server load balancing:

1. APV Series continuously monitors each Service IP's availability (i.e. up/down) for all sites. The APV Series also collects load information, e.g. current connections, for the GSLB sites.
2. The APV Series serves as the authoritative domain name server (DNS) to resolve Web site domain and sub-domain names. In a GSLB deployment, the APV Series handles DNS queries for Web domain names and return the Service IP(s) that serve the domain. For each query, GSLB selects the best site, as well as the most suitable Service IP(s) within the selected site, based on any one of a variety of load balancing algorithms.

When users enter the domain name to be resolved via the APV Series DNS, as the delegated DNS server the APV Series will resolve the IP(s) that belong to the target site. If that site is down or Service IP(s) are under maintenance, the APV Series GSLB will resolve with IP(s) that belong to an alternate site without service interruption for the user.

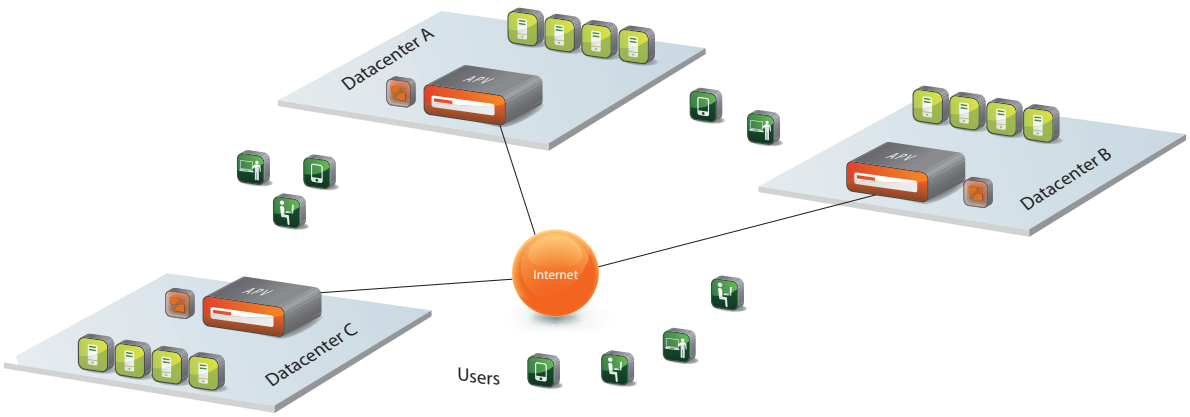

Figure 1: Typical GSLB Deployment Architecture

The process of domain name resolution provided by the APV Series' SDNS is detailed as follows:

1. The client sends the DNS query to the local DNS to resolve the domain name "www.example.com"
2. After a few rounds of lookups, the local DNS finds that Array SDNS is the authoritative DNS for the domain name and forwards the DNS query to SDNS
3. Array SDNS makes the decision to select healthy Service IPs and return them to the local DNS
4. The local DNS returns the Service IPs to the client

The process of the SDNS decision process is detailed as follows:

1. Based on the domain name in the DNS query, SDNS finds the hit SDNS policy.
2. Based on the SDNS policy, SDNS finds the hit SDNS service pool.
3. Based on the method of the hit SDNS service pool, SDNS picks the correct service IPs.
4. SDNS returns the picked Service IPs to the local DNS.

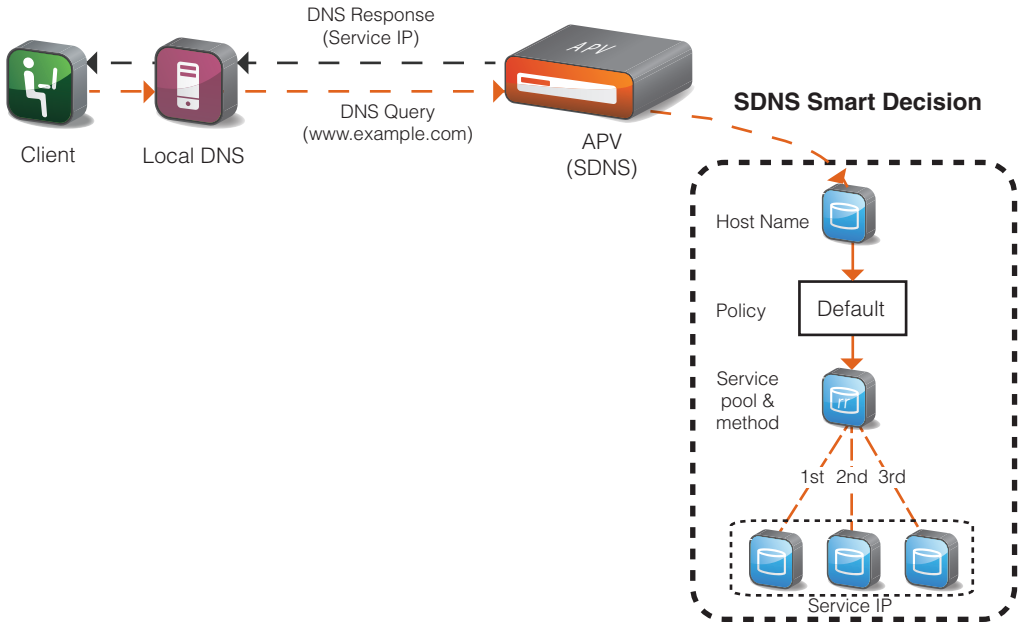

Figure 2: SDNS Smart Decision Process

Further Detail

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.

GSLB Techniques

GSLB is an integral part of the Array Networks APV Series application delivery controllers. The GSLB function can be run as a standalone or in conjunction with other functions such as server load balancing (SLB) or link load balancing (LLB). Because the Array Networks GSLB will be a part of the network DNS infrastructure, typically a pair of APV Series appliances is run on different sites to support DNS redundancy and to form an Array Networks GSLB network. All APV Series appliances in the GSLB network are set up to automatically sync dynamic information, such as Dynamic Proximity System (DPS) information.

Array Networks APV Series appliances support a wide range of global load balancing policies and algorithms to provide the flexibility to choose the policy and method most appropriate for the enterprise's needs. When an APV Series GSLB receives a DNS query (either IPv4 or IPv6), it returns the IP address of an APV Series appliance serving the requested domain. The GSLB function first selects a GSLB site or pool based on the chosen policy (dynamic, region or default), then according to the pool method the GSLB selects the Service IP (server) within that site/pool serving the requested domain. The following GSLB pool methods are available:

1. Round Robin
2. Weighted Round Robin
3. IP Overflow – The healthy Service IP with the highest priority will be returned
4. Hash IP – GSLB will compute the hash value based on the source IP of the DNS request. For DNS queries with the same source IP address GSLB returns the same Service IP
5. Drop – Just drop the DNS query. No DNS response is returned to the local DNS

In a typical deployment, the GSLB pool is designed with high availability and flexibility – for example, the same GSLB pool can have a primary and secondary method. In addition, a backup pool is used in the event that no Service IP is available.

For the region policy, the APV Series allows the network manager to define static region rules and IP maps, which can be the ISP IP allocation information as well. The IP information can be IPv4 and IPv6, and can be predefined in a data file that is imported to the appliance.

Array Networks GSLB also supports a Dynamic Proximity System (DPS) which automatically generates dynamic GSLB proximity rules by distributed network probes. The following DPS network information, collected by the distributed network probes, can be used by the APV Series GSLB to dynamically distribute client traffic to different GSLB sites/pools:

- RTT: Metric = round trip time
- PLR: Metric = packet loss rate
- Hops: Metric = number of hops between local DNS and the proximity site
- Mix: Metric = (RTT x weight) + (PLR x weight) + (Hops x weight)

Disaster Recovery

Array Networks GSLB includes purpose built Disaster Recovery functions. Multiple sites (IPs) 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 over or fail back automatically or manually.

Full-DNS Resolution

In the APV appliance, GSLB processes DNS queries of the A, AAAA and CNAME types, and the Full-DNS function option can process DNS queries of all other types, such as MX and PTR. Also, if the Recursive Query function is enabled, the APV Series' GSLB will forward DNS queries for domain names that are not configured on the APV itself.

In addition, APV Series application delivery controllers are IPv6 Ready Gold Certified, and the GSLB functionality provides comprehensive IPv6 DNS support.

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 Service IP in the system. If the services associated with a Service IP become unavailable, the Service IP is removed

from the system and subsequent requests for the affected domain are redirected to an available Service IP on the same site or on a different site. The site masters continuously communicate with the GSLB members at other sites for monitoring the Service IP 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 an 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. Array Networks GSLB functionality can be distributed across multiple APV Series appliances and locations, and the memory-based design helps assure availability as well. In addition, due to the distributed nature of APV Series GSLB high performance and built-in security safeguards, DNS DDoS flooding attacks can be easily mitigated.

Additionally, the Array Networks family of ADC products' internal cache 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 delivery networking with over 5000 worldwide customer deployments. Powered by award-winning SpeedCore® software, Array application delivery, WAN optimization and secure access 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 250 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, IDC and Frost & Sullivan have recognized Array Networks for its technical innovation, operational excellence and market opportunity.

Mar-2017 rev. a