

**Coyote Point Systems White Paper:  
Establishing a Geographically-Distributed Internet  
Presence with Envoy**



## Establishing Geographically-Distributed, High-Availability Internet Presence with Coyote Point Envoy

Today's E-Commerce environment demands that successful enterprises establish an Internet presence. But it can do more harm than good to tie your business to a web site with sluggish response and intermittent downtime. Load balancing and application acceleration products like Equalizer help enterprises create a high-availability presence within a single location. Geographic server load balancing (GSLB) goes one step further by enabling growth across many locations, creating a distributed Internet presence that dramatically improves end-user experience and up-time while reducing overall cost of operation. In this paper, we'll explore how to use Coyote Point Equalizer with the Envoy GSLB software to ensure 24x7 availability and fast connections for web content deployed at multiple geographic locations.

### **What is Load Balancing?**

Today, very few enterprises can afford to host their company's web site on a single, monolithic server. Rather, sites are deployed on server clusters that improve performance and scalability. To provide fault tolerance and hide cluster detail from site visitors, a load balancing and application acceleration appliance sits between the Internet and a server cluster, acting as a virtual server.

As each new client request arrives, an application-aware load balancing appliance makes near-instantaneous and intelligent decisions about the physical server best able to satisfy each incoming request. Load balancing optimizes request distribution based on factors like capacity, availability, response time, current load, historical performance, and administrative weights. A well-tuned adaptive load balancer ensures that customer sites are available 24x7 with the best possible response time and resource utilization.

Coyote Point Equalizer is a full line of premier high-volume, low-cost, application acceleration and load balancing appliances that starts at less than \$2,000. The entry-level Equalizer E250GX provides the basic Layer 4 and Layer 7 features that small businesses can afford. The E350GX provides more sophisticated content-based traffic management and is optimized for small to medium enterprises. The E450GX adds additional processing power and hardware SSL acceleration, for more demanding E-Commerce applications. The E650GX is Coyote Point's Data Center workhorse, augmented with hardware data compression for faster downloads and also comes with the Envoy GSLB software included (an option on the other platforms). For a complete review of the Equalizer product line, including data sheets, please go to our web site at <http://www.coyotepoint.com/products/>.

Equalizers can be deployed in a hot-backup configuration for maximum reliability. Designed to meet the extreme demands imposed by heavily-loaded, mission-critical web sites, Equalizer can handle HTTP, HTTPS, email, news, FTP traffic, and more – and provides the Layer 4 and Layer 7 persistence required to efficiently handle Active Server Pages, SSL, and other applications that require maintenance of session information. Equalizer's active content verification ensures that target applications are fully operational, circumventing failures that might go undetected by other load balancers. And, if you have a virtual server farm running on VMware, Equalizer VLB can provide additional probing and control methods for VMware virtual servers.

### **Why Geographic Server Load Balancing?**

This basic load balancing provides horizontal scalability and fault tolerance for servers at a single location. But many enterprises establish a worldwide Internet presence by deploying servers in many locations. To add this vertical scalability and disaster resistance, these sites employ *geographic server load balancing* (or GSLB).

GSLB increases availability by allowing regional server clusters to share workload transparently, maximizing overall resource utilization. Why let servers sit idle at 5:00 am in Hong Kong when they could be handling afternoon "rush hour" traffic generated by clients in the US?

Furthermore, who can afford to let business grind to a halt if the San Francisco cluster goes down due to earthquake, denial of service attack, or telecommunications failure? GSLB enables disaster recovery on a global scale, bypassing regional interruptions automatically.

Even when everything is running smoothly, load balancing across regional server clusters offers many benefits. Response time can be minimized by directing clients to the closest server cluster, and transmission costs can be reduced by avoiding costly trans-Atlantic or trans-Pacific hops. Regional servers can use local ad insertion and language customization to deliver content appropriate to the client's geographic location. Distributed load balancing provides the benefits of regional server deployment while preserving the high availability and transparency essential to sound e-Business.

## How Geographic Server Load Balancing Works

Coyote Point Envoy is a full-featured, low-cost, GSLB add-on for the Equalizer. Envoy allows any Equalizer to cooperate with its peers, enabling intelligent request distribution across geographically-distributed server clusters.

An Envoy-enabled web site is a geographic server cluster, composed of regional clusters. Each regional cluster is composed of servers that provide a common service, supervised by an Equalizer running Envoy. For example, the web site [www.coyotepoint.com](http://www.coyotepoint.com) might be supported by three regional clusters, located in California, New York City, and London. An Equalizer running Envoy software and web servers with similar content are deployed at each of these locations. Here's how Envoy routes each client request to the "best" server, avoiding regional clusters and servers that are unavailable or overloaded.

When a client browser addresses an HTTP request to <http://www.coyotepoint.com>, this fully-qualified domain name is resolved using Internet standard Domain Name Server (DNS) protocol. A "lookup" query is sent by the client to its local ISP or enterprise DNS [figure 1, step 1]. The local DNS forwards the query to the "authoritative" DNS: in this case, the one responsible for coyotepoint.com [step 2]. The authoritative DNS returns IP addresses for the three Equalizers running Envoy. The client sends its HTTP request to the first IP address, trying other addresses if no response is received [step 3]. In this manner, the client's HTTP request is received by the first reachable Envoy site: in our example, New York.

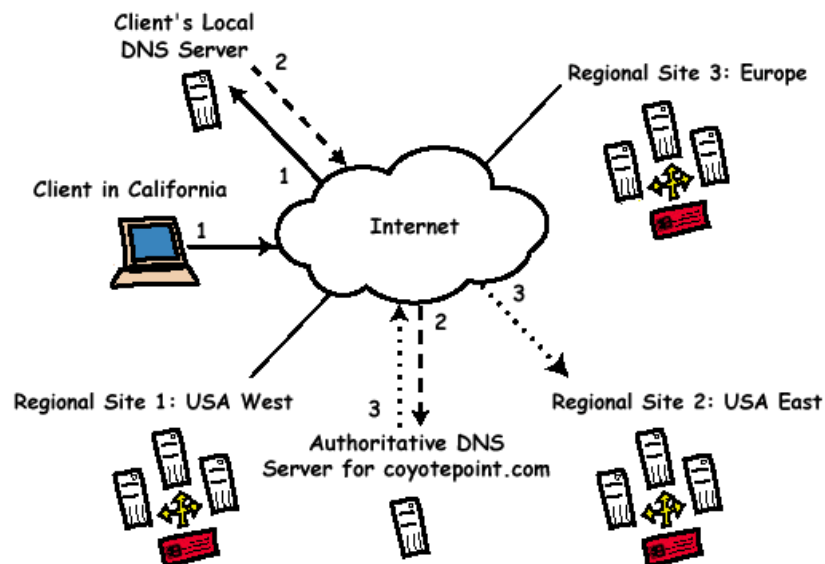


Figure 1: DNS Resolution for Geographic Cluster [www.coyotepoint.com](http://www.coyotepoint.com)

When Envoy receives a client HTTP request [figure 2, step 1], it uses configuration data to identify all regional sites for the geographic cluster [www.coyotepoint.com](http://www.coyotepoint.com). Geographic probes containing information about the client and the requested URL are sent to Envoy agents at each regional site: the New York Envoy probes itself, California, and London [step 2]. Each agent checks local resource availability and responds with an error if the requested URL is unavailable. If the URL is available, the agent "pings" the client to calculate latency, then returns a response to the

probe initiator [step 3]. The initiating Envoy uses all responses to determine the "best" regional site and forwards the request to that site [step 4].

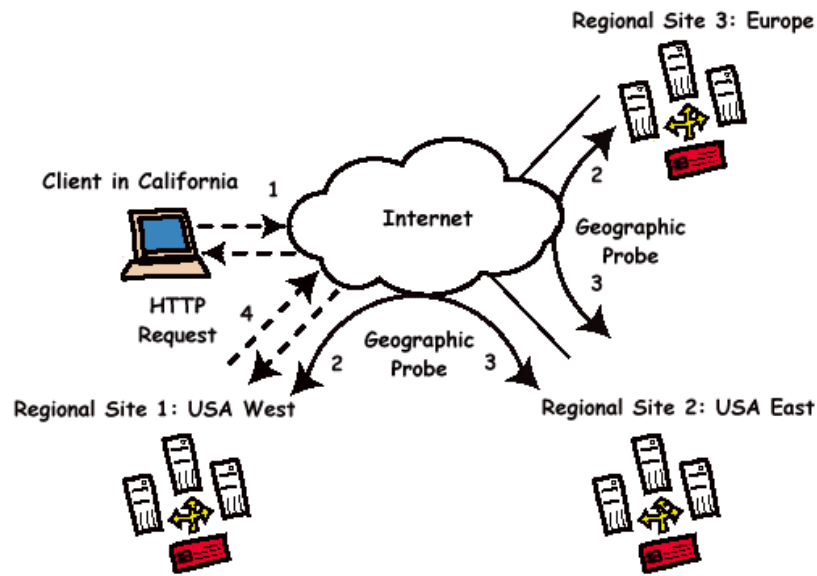


Figure 2: Envoy Probes for Geographic Cluster [www.coyotepoint.com](http://www.coyotepoint.com)

In our example, the California cluster is selected to handle this request. This site may be selected because it is closest to the client (i.e., has the shortest round-trip time). However, if the requested resource at the California were down or heavily loaded, the request would be handled by New York or London, without any client intervention or awareness. Administrator-defined policies can influence Envoy's decision: for example, a cluster with regional ads might favor proximity over load factors. And as conditions change, Envoy probes guarantee that decisions are made based on current data. It is easy to see how Envoy works transparently and adaptively to optimize response time, distribute load across geographically-distributed sites, and bypass failures.

### ***Deploying Distributed High-Availability Clusters with Envoy***

Envoy is a simple software upgrade to Equalizer. To deploy Envoy, first complete normal Equalizer installation and configuration at every regional site. Then install Envoy software on each Equalizer. If your network is firewalled, you'll need to open ports used by Envoy.

Each geographically-distributed, high-availability cluster is configured in three easy steps.

1. **DNS Configuration:** For each geographic cluster to be balanced, the authoritative name server must be configured to return name server and alias records for Envoys at every regional site. In our example, the authoritative DNS for [coyotepoint.com](http://coyotepoint.com) delegates authority for [www.coyotepoint.com](http://www.coyotepoint.com) to east, west, and [europe.coyotepoint.com](http://europe.coyotepoint.com). When any client looks up [www.coyotepoint.com](http://www.coyotepoint.com), the queried delegate identifies all three Envoys in its DNS response.
2. **Add A Geographic Cluster:** Envoy is administered through the Equalizer's graphical user interface. Create a new geographic cluster with the name defined in DNS, then specify a load balancing method and responsiveness factor [see figure 3].

There are four **load balancing methods** supported by Envoy. **Adaptive** lets Envoy take all factors into consideration when selecting a regional site. **Round Trip** emphasizes client proximity, while **Site Load** gives greater weight to server load measured by Equalizer. **Site Weight** specifies a static factor that skews request redirection. Weights might be used to implement primary and backup sites, while load optimizes overall server utilization at the expense of client response. Round trip may be appropriate when delivering regionalized content.

Five **load balancing responsiveness levels** are supported by Envoy, controlling how quickly balancing decisions will be impacted by dynamic changes. **Slowest** causes probe results to be averaged over a longer period of time, while **Fastest** causes Envoy to recalculate balancing criteria more frequently. **Medium** provides rapid response to changing conditions, while smoothing out transient network glitches.

3. **Add Sites To The Geographic Cluster:** Finally, use the Equalizer GUI to add each regional site to the geographic cluster created in step 2 [see figure 4]. Each site identifies the virtual server used to balance requests locally and the resource accessed through this virtual server. A static weight can be used to bias results towards individual resources with greater capacity. One site can be designated as the default for this geographic cluster.

In our example, one geographic cluster would be created for [www.coyotepoint.com](http://www.coyotepoint.com) [figure 3], and three regional sites would be configured for east, west, and europe.coyotepoint.com [figure 4].

**Add New Geo Cluster** [?] [X]

In order to add a new Geo Cluster, please fill out the following required information. You will then be taken to a detailed cluster view, where you can select other options.

**Geo Cluster Parameters**

FQDN Name:

DNS ttl:

Figure 3: Geographic Cluster [www.coyotepoint.com](http://www.coyotepoint.com)

**Add New GeoSite** [?] [X]

In order to add a new site, please fill out the following required information. You will then be taken to a detailed site view, where you can select advanced options.

**GeoSite Parameters**

GeoSite Name:

A Record IP Address:

Agent IP Address:

Site Version:  v8  v7

Resource Name:

Figure 4: Site [west.coyotepoint.com](http://west.coyotepoint.com)

## Keeping Tabs On Your Internet Presence

Of course, any enterprise that is balancing requests for a mission critical resource must have quick, easy, reliable access to status and performance data. Envoy continuously monitors operational statistics for geographic clusters and the regional sites within them. Using Equalizer's browser-based Administrative Interface, an administrator can view configured parameters and instantaneous statistics for each regional site, including:

site statistics	
total requests	50900
number queued	0
timed out	0
Site had zero weight	0
agent retries	50916
agent misses	28
agent errors	54
unavailable	0
site returned	53
returned default	0
resource performance	1

<b>total requests</b>	The number of requests directed to this Site since the last reboot.
<b>number queued</b>	The number of requests queued for this site.
<b>timed out</b>	The number of agent-to-client triangulation probes that timed out before Equalizer received a response.
<b>site had zero weight</b>	The number of times the server was chosen but had a zero weight.
<b>agent retries</b>	The number of probes Equalizer re-sent to its agent.
<b>agent misses</b>	The number of Equalizer-to-agent probes that received no response. Interruptions in network connectivity between the Equalizer server and site agents and site failures can result in missed probes.
<b>agent errors</b>	The number of Equalizer-to-agent probes that returned a resource-unavailable error -- that is, Envoy on the remote site determined that the requested resource is unavailable.
<b>unavailable</b>	The number of times the server was chosen but was unavailable.
<b>site returned</b>	The number of clients directed to this site. You can compare this number with the values for other sites to determine the relative number of users sent to each site. If a value for one site is zero and the others are non-zero, consider why the zero site has no traffic.
<b>returned default</b>	The number of clients directed to the default site.
<b>resource performance</b>	The load on the above resource that the Equalizer agent calculates. The load incorporates data on resource response time, number of active requests, and load-balancing variables.

Envoy also support statistics plotting over time, as shown in the screen shots on the following page.

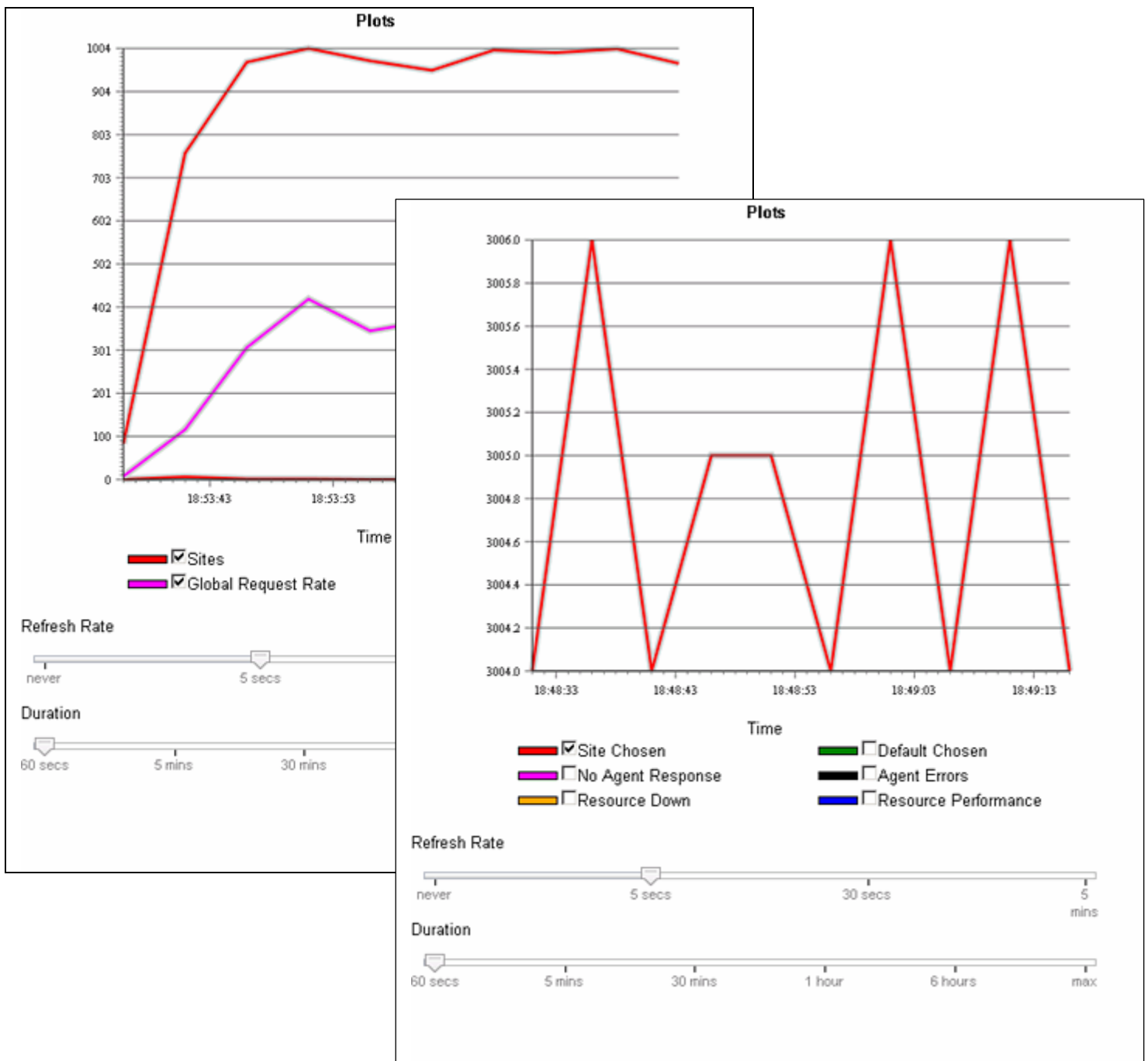


Figure 5: Geographic Cluster Plot at left. Site plot at right.

For a quick view of overall performance, plot geographic cluster statistics with a single click [figure 5]. Values can be plotted over a range of durations and refresh rates. With a single glance, administrators can spot performance degradation, discern trends indicating network or site configuration issues, as well as see traffic pattern changes caused by tuning parameters or the addition of sites and resources.

Site statistics and plots can help identify network configuration errors. For example, a large number of triangulation timeouts may indicate an incorrectly-configured firewall. Site weights can be tuned to establish desired load distribution, based on the number of requests directed to each site. And key performance metrics like average client response time can be gathered without requiring any further network instrumentation. Both graphical and text views for are available for site statistics.

## **Conclusion**

With Coyote Point Envoy, balancing load across geographically-distributed server clusters isn't rocket science. Envoy's adaptive load balancing algorithms are easy to understand and configure. Results are easily quantified. Tuning parameters are available for administrators who need them, without the added complexity and cost that can discourage deployment with some other load balancing products.

Coyote Point customers are quick to vouch for Envoy. "Our sites had no way of covering for each other until Envoy came along," says Jake Dias, Systems Manager for IMDb, an Internet movie data provider with regional clusters in the US and UK. With Envoy, "we are now able to offer quick service to all users, wherever they are. Any site can go down and nobody will even notice."