

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

As e-business quickly becomes the norm, virtually every enterprise must establish its 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 products help enterprises to create high-availability presence within a single location. Geographic load balancing 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's Envoy to ensure 24x7 availability and fast connections for web content deployed at more than one geographic location

## ***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 appliance sits between the Internet and a server cluster, acting as a virtual server.

As each new client request arrives, the load balancing appliance makes near-instantaneous 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's Equalizer is a premier high-volume, low-cost load balancer that starts at less than \$4,000. The entry-level Equalizer 250 balances up to 64 clusters of 8 servers each, supporting 64,000 simultaneous connections at T1 rates. The high-end Equalizer 450 can handle an unlimited number of 64-server clusters and 4 million connections, with an aggregate bandwidth of 200 Mbps. 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, the Equalizer can handle over 130,000 HTTP GET operations per minute. The Equalizer can also balance email, news, and FTP traffic, supporting "sticky connections" required to efficiently handle Active Server Pages and SSL. The Equalizer's active content verification ensures that target applications are fully operational, circumventing failures that might go undetected by other load balancers.

## ***Why Geographic 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 load balancing*.

Geographic load balancing 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? Geographic load balancing 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 Load Balancing Works

Coyote Point's Envoy is a full-featured, low-cost geographic load balancing add-on for the Equalizer, priced at just \$2,995 per site. 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.

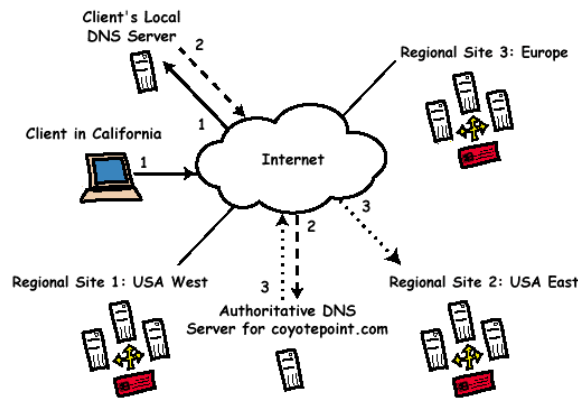


Figure 1: DNS Resolution for Geographic Cluster [www.coyotepoint.com](http://www.coyotepoint.com)

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].

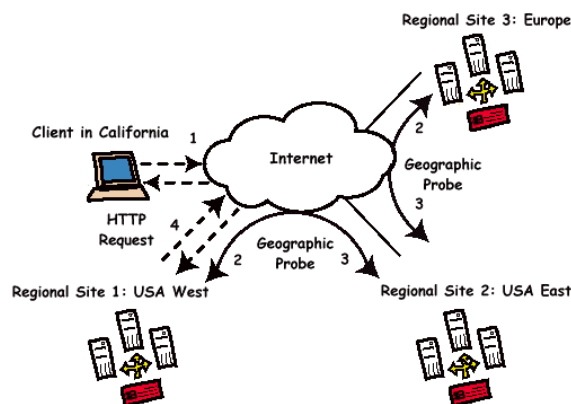


Figure 2: Envoy Probes for Geographic Cluster [www.coyotepoint.com](http://www.coyotepoint.com)

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 delegates authority for www.coyotepoint.com to east, west, and europe.coyotepoint.com. When any client looks up 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].

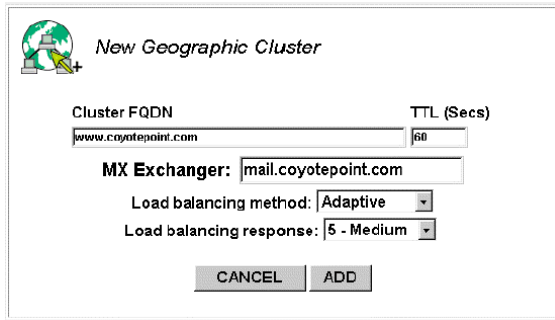


Figure 3: Geographic Cluster [www.coyotepoint.com](http://www.coyotepoint.com)



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 regional sites within them. Using the Equalizer GUI, an administrator can view configured parameters and instantaneous statistics for each regional site, including:

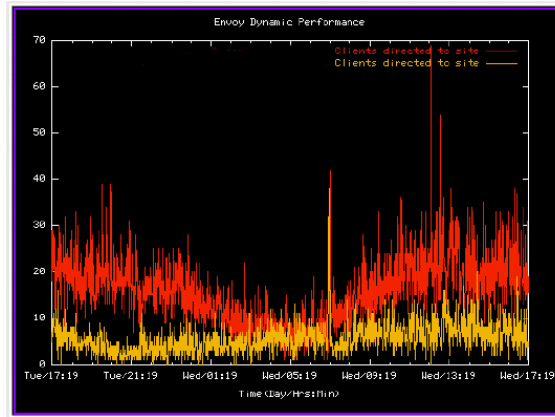
<u>Missed:</u>	Envoy probes with no response, indicating network or site failure
<u>Triangulation Timeouts:</u>	Envoy agent to client pings with no response
<u>Retries:</u>	Envoy probes that were resent
<u>Resource Errors:</u>	Envoy probes that returned resource unavailable error
<u>Returned:</u>	Client requests directed to this site
<u>Default:</u>	Client requests directed to the default site for this cluster
<u>Average Time:</u>	Average successful client ping response time
<u>Resource Load:</u>	Calculated instantaneous load supported by this resource

Site statistics 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.

For a quick view of overall performance, plot geographic cluster statistics with a single click [figure 5]. Four overall values are available:

<u>Site Summary:</u>	Client requests directed to all sites in this cluster
<u>Request Rate:</u>	Client requests received per minute by this cluster
<u>Active Requests:</u>	Client requests being routed by Envoy
<u>Network Latency:</u>	Average triangulation time when at least one site responded

Values can be plotted over the past hour, five hours, twelve hours, or day, and the administrator can zoom in on any area of interest for a closer look. With a single glance, administrators can spot performance degradation signaling network or site failure, or see improvements caused by tuning changes or addition of sites or resources. While many other products charge extra for monitoring and performance tools, Coyote Point includes this essential "dashboard" capability at no extra charge.



Click graph to zoom in

Display:	For Previous:	
Site Summary	1 Hour	PLOT
Request Rate	5 Hours	
Active Requests	12 Hours	
Network Latency	24 Hours	

Figure 5: Keeping Tabs on Your Geographic Cluster

## Conclusion

With Coyote Point's 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."

**Prepared by:**  
Lisa Phifer  
Core Competence, Inc.  
lisa@corecom.com