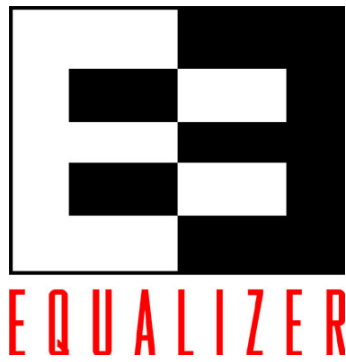


Intelligent Load Balancing SSL Acceleration and Equalizer v7.0



Intelligent Load Balancing: Layer 4 or Layer 7?

In the three years since the dotcom bust, network architecture has sustained a quiet, rapid evolution as businesses move more and more forms of transaction and interaction online. Given the dramatic slowdown in IT spending since the boom years, this continuing development of ebusiness capabilities and network complexity is a minor miracle. Consistently, businesses have demanded – and successfully obtained -- more bang for their IT buck.

During this period, website and intranet quality of service standards have risen rapidly. Website and portal users have come to expect essentially zero site downtime – and very little waiting time while pages are downloading. These raised expectations are both a cause and effect of server load balancing, a technology now employed by all but the most rudimentary networks. Load balancers intelligently distribute network load among clustered servers, thus reducing download times and providing failover by automatically bypassing down servers. Load balancing also provides network scalability, enabling network administrators to add inexpensive commodity servers to a cluster on an as-needed basis.

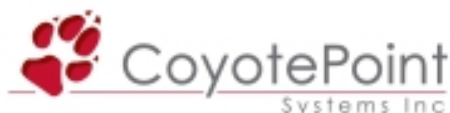
In a cost-conscious time, Coyote Point has established itself as the price/performance leader in Internet Traffic Management (ITM). Since its founding in 1997, Coyote Point's mission has been to provide ITM essentials – intelligent load balancing, reliable failover, intuitive configuration, robust reporting of server and cluster performance – at a fraction of the prices charged by competitors. Eschewing some of the more exotic bells and whistles in the field, Coyote Point has functioned as a 'value filter,' incorporating into Equalizer, the company's flagship load balancing appliance, only those innovations that have proved their value to website administrators.

Perhaps the most important advance in ITM technology in recent years has been the development of Layer 7 (L7) load balancing. Traditionally, load balancing operated at layer 4 (L4), the transport layer. Working at the application level (L7), L7 load balancers enable advanced network architectures by allowing routing decisions based on application information and by examining content beyond the packet header, rather than relying simply information in the Layer 4 (TCP/IP) headers. L7 load balancers also maintain session persistence by examining cookies sent by user browsers, ensuring that users who input information will be returned consistently to the server storing that information.

Not every network requires L7 load balancing capability. If your network needs begin and end with high availability, fail-over protection, and easy cluster management, an intelligent L4 device that can manage multiple server clusters may be the most cost-effective solution. If your needs are more subtle – that is, if you need to differentiate content delivery among different user groups and ensure session persistence for a high volume of users – then L7 capability will be worth the added cost. This paper will enable network administrators to determine whether L4 or L7 load balancing best suits their needs. Coyote Point's solutions at both levels – the L4-enabled E250i and the L7-enabled E350 and E450 models of Equalizer – all offer unsurpassed bang for the buck. The models' differing capabilities are described in more detail below.

Evolution of Load Balancing Technology

Early load balancing schemes provided website visitors with the appearance of one server on a single URL while actually distributing traffic to a cluster of servers offering identical content. Early implementations included crude load-sharing, such as round robin DNS, where a name server would offer one of several IP addresses in response to a hostname query, or inflexible and proprietary solutions such as that employed in early Netscape Navigator versions, which turned client requests for www.netscape.com into www1.netscape.com, www2.netscape.com, and so on, for Netscape's exclusive benefit. The next phase was the first generation of 'intelligent' L4 load balancing appliances that made more sophisticated decisions about which cluster member should receive an incoming request, instead of indiscriminately spreading the load evenly regardless of the respective availability and load of cluster members. (Coyote Point's Equalizer 1.0, released in 1997, was among the first intelligent load balancers.) But these L4 solutions still assume a cluster of servers serving identical,



static content. They can cause problems when faced with dynamic content generation, and situations where session persistence is required, because a client may be directed to a different server in the middle of a session. Today, most websites will need a strategy for ensuring session persistence, whether at the L4 or L7 level.

L7 load balancing enables incoming requests to be directed to different servers based on criteria such as HTTP and SSL versions used, type of web browser, URI pathnames, file names and extensions, and even cookie data. As such, L7 allows administrators to create highly flexible cluster configurations by specifying rules that the appliance will use when directing client traffic to servers.

L4 Functionality

The L4-enabled E250i provides HA and fail-over security at an entry cost comparable to that of a single beefed-up Intel-based server. The E250i can detect a failed cluster node and stop directing traffic to it and an administrator can add or remove servers without disrupting service. The E250i can be set up to direct traffic for multiple sites, distributing requests to the appropriate servers based on IP address and/or TCP port number, all transparent to the end user. For a highly scalable configuration, L4-enabled E250i is also highly reliable: a failed server will be automatically detected, and traffic no longer directed to it. Redundant E250i's may be deployed to eliminate any single point of failure.

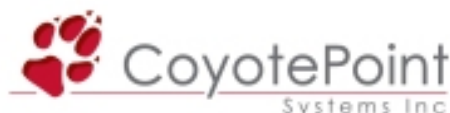
Also, with L4 support, E250i adds an extra layer of security and control for networks, providing conventional traffic filtering based on IP address and port. E250i can be set up to deny requests from a specific IP addresses or address ranges, or according to specific policies, such as permitting only certain protocols. E250i can also protect ports, so virtual clusters are indeed virtual – traffic won't be able to bridge over to an actual server. In addition, NAT support enables administrators to assign non-routable internal IP addresses to servers, making them accessible via the greater Internet only through Equalizers. In addition to these advanced features, E250i can also enable additional security protection such as Denial of Service protection from so-called SYN-floods and port scans.



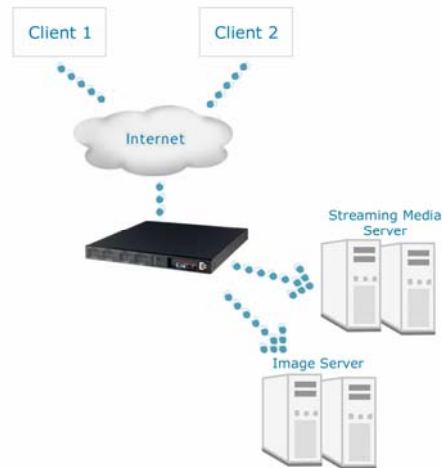
Equalizer provides intelligent virtual clustering capability, monitoring server load in real time, directing traffic appropriately

L7 Functionality

By providing L7 load balancing capability, E350 and E450 offer much deeper functionality than simply distributing Web requests across multiple servers. Because L7-aware appliances will inspect request URIs to dispatch traffic to the appropriate servers, it's easy to construct a cluster of servers specializing in different roles instead of a cluster of servers of identical capability. For example, some servers in the cluster may provide more complex services such as dynamic, database-driven applications, while others optimized for raw speed serve up static data like images and sounds, all while maintaining the appearance of a single web site. Other servers in a cluster can be assigned to serving requests from obsolete clients. These models can identify the exact object being requested and route the request



appropriately. As such, E350 and E450 are able to use the contents of these application layer data elements to enable intelligent, context-aware load balancing decisions in various contexts.



L7-enabled Equalizer allows for content-based load balancing. In this scenario, based on the client request, Equalizer routes the request to the optimal server based on content type and server load.

Equalizer 7.0 is capable of examining the HTTP headers of incoming client requests and using the information contained therein to choose which subset of clustered servers should handle the request. A flexible and comprehensive rule-based configuration permits the administrator to specify these decisions using familiar boolean operators. For example a cluster of image servers might handle traffic when the client's request URI was for file types ending in *.GIF* or *.JPG*.

Cookie Persistence: L7 supports cookie persistence by inspecting cookies sent by user browsers. Based on cookies, users whose ISPs hide the individual's IP address behind a proxy server can access their data from different locations even as their source IP address changes and continue to connect to the same server as before -- a key feature for e-commerce and other sites that keep must keep session data across multiple pages. Various flags can be set to modify the way the cookie persistence is performed. For example, "cookie stuffing" may be enabled, causing the load balancer to insert a server identification cookie into the data stream returned by the server to the client. Thus, when the client connects to the website again, the cookie can be examined and its contents used to select the same server as last time. This makes it possible to granularly control cookie stuffing based on the specific application. Administrators can also specify a numeric value, in seconds, for the "Max-Age" cookie attribute, which will cause the client browser to eliminate the cookie after the specified number of seconds have elapsed. The cookie can also be restricted so that it is presented to only those servers whose host name is within the indicated domain, for security.

Aside from supporting cookie persistence, Equalizer 7.0 can manage other forms of session persistence, ensuring a client communicates with the same server for the length of a given session. This is important for security as it helps prevent session hijacking and frees application developers from worrying about session persistence in an application backend.

SSL Acceleration and Offload: Although currently the most secure client/server communication encryption method available, encrypting traffic with SSL remains highly CPU intensive. Managing SSL work from the load balancer improves performance, while saving money. For sites with significant levels of SSL traffic, Coyote Point offers the XCEL SSL hardware acceleration module for the E350/450. XCEL is based on Netoctave's 2000 TPS SSL acceleration chip and allows the load balancer to quickly process thousands of transactions per second, without overload. Offloading the SSL encryption processing from the servers significantly increases their throughput capacity as well.



L4 or L7?

The range of possibilities that L7 affords, whether to better optimize resources or to load balance new types of traffic, has barely been explored. Disparate machines can be meshed together to present a unified whole, and a web site can be rearchitected without wholesale disruption of URLs already in circulation.

At present, the most fully-realized benefit of L7 load balancing is the assurance of session persistence for website visitors who engage in transactions or other exchanges of data. The ability to reliably identify visitors whose IP addresses are masked by proxy servers and send their repeat requests to the same server – while still distributing overall traffic among servers with maximum efficiency – solves one of the most urgent performance challenges currently faced by network administrators. For sites which have been designed with state sharing in mind (i.e. all servers in a cluster have some means of serving client's stateful requests), L4 load balancing is ideal. It is a quick, resource friendly means of distributing traffic optimally. L4 load balancing is also appropriate for static content clusters (such as image servers) where stateful transactions are not an issue. Layer 4 load balancing is a very efficient and easy to deploy solution for sites which do not have sophisticated content clustering requirements, and which do not need perfectly reliable client-server persistence mechanisms. For example, an image server cluster or cluster of static content servers would probably not require these capabilities.

While the possibilities that L7 support offers are exciting, think carefully about your network needs and architecture before choosing, and buy only the features you need. Most simply, consider the number of virtual clusters your backend servers can support, and how many simultaneous client connections your application may require. For those enterprise networks that can satisfy their session persistence requirements with L4 methods, the E250i will very likely suffice, providing HA, fail-over security, and easy management features whilst allowing use of all network components to maximum capacity. Moreover, upgrading from E250i to its higher end counterparts is quite easy. On the other hand, networks requiring cookie persistence, higher throughput, or a high degree of specialization on backend servers will find that the L7-enabled E350 or E450 meets their needs

Why Equalizer?

All models of Equalizer use intelligent load balancing algorithms to distribute traffic to the appropriate server: simple round robin, administrator-assigned priority values, fastest to respond, fewest connections, server based agents, and block. These algorithms are easy to set up, yet powerful enough for most configurations. The load balancing algorithm responsiveness can be adjusted by specifying a responsiveness value. All models of Equalizer can also prioritize traffic based on TCP or UDP ports, including a combination of destination address and destination port number, ensuring that critical applications can receive priority. All networks benefit from L4-enabled E250i's level of load balancing functionality, which ensures both HA and fail-over security.

All Equalizers also provide a wealth of operational attributes to control network specifics in the way of configuration file settings. For example, "client_timeout", allows the administrator to specify in tenths of a second how long before dropping a client connection due to inactivity during a request. The "connect_timeout" parameter enables specification of a time out for an Equalizer connection to a backend server, which isn't necessarily reflected back to the client. Equalizer provides similar programmatic control over server timeouts and ACV (Active content verification) probes. Again advanced network specific control is available, including NAT (Network Address Translation), support for difficult protocols like passive FTP, and the ability to support non-RFC compliant clients.

Equalizer is easy to manage securely. All Equalizer versions feature a well-designed, intuitive, and award-winning Web based GUI tracking server and cluster performance in graphical format, in addition to support for NAT/Access Lists. As such, the security features are top-notch, particularly compared to competitive products, some of which do not support SSL/SSH-based management GUIs.

Coyote Point has always offered a rich feature set for a low cost and prided itself on providing the best value to its customers. At an entry cost of \$3995, E250i can balance up to 64 virtual server clusters



comprised of up to 8 servers each, supporting up to 64,000 simultaneous connections over a T1 connection. For \$5995, E350 can handle 2,000,000 simultaneous connections on an unlimited number of clusters containing up to 16 servers each and an unlimited number of virtual IP addresses at a recommended T3+ bandwidth. At \$9995, the E450 provides a dual gigabit ethernet interface and handles up to 4,000 simultaneous connections, balancing the load on an unlimited number of clusters with up to 64 servers each. All Equalizer models include graphical, real-time performance monitoring. Deep support packages that are affordable and flexible.

Summary

Coyote Point's L4 and L7 load balancers offer unmatched price/performance to meet the reliability, scalability, and performance requirements of today's Internet-based economy. Time-tested L4 load balancing with E250i offers a feature-rich HA solution at minimal cost. The innovative L7 architecture of E350 and E450 provides dramatic capital cost savings due to improved network resources utilization, reuse of existing external hardware, and operational efficiency. In short, Coyote Point offers a highly efficient, cost-effective traffic management solution for every network infrastructure and budget.

For More Information, visit us at: www.coyotepoint.com or call us at 650.969.6000.

