Enhancing Application Delivery and Load Balancing on Amazon Web Services with Riverbed Stingray Traffic Manager

Stingray™ Traffic Manager from Riverbed Technology is a high performance software-based application delivery controller (ADC), designed to deliver faster and more reliable access to public web sites and private applications. As a software-based ADC, it frees applications from the constraints of legacy, proprietary hardware-based load balancers and provides unprecedented scale and flexibility to deliver applications across the widest range of environments, from physical and virtual data centers to public and hybrid clouds.

ADCs are a key component of any business-critical application infrastructure and offer several benefits, including:

- **Speed**: Accelerate services, increase capacity, and reduce costs by offloading performance-draining tasks such as SSL and compression onto Stingray Traffic Manager’s optimized implementations. Administrators can also cache commonly requested content and optimize traffic delivery to applications so they’ll run as fast as they would in a perfect benchmark environment.

- **Reliability**: Improve application availability by intelligently load balancing traffic, either locally or globally; avoiding failed or degraded servers; monitoring performance problems; and, shaping traffic spikes.

- **Improved security**: Gain full control over how traffic is internally routed and use high-performance inspection to verify any part of a request or response to apply global filtering or scrubbing policies. Stingray™ Application Firewall (optional) also protects against a broad range of web application attacks such as SQL injection, cross-site scripting, session hijacking, and other threats.

Stingray Traffic Manager also integrates well with applications, meaning that application owners can get products to market faster, and with better performance. Stingray Traffic Manager integrates fully within the Amazon Elastic Compute Cloud (Amazon EC2) environment, bringing several benefits to application developers, owners, and architects. The remainder of this paper looks at the architecture and capabilities of combining Amazon EC2 with Stingray Traffic Manager for load balancing.

**Cross-region load balancing**

Amazon Web Services Elastic Load Balancer can balance loads across instances in one or more Availability Zones (AZs) within a single region. The advanced application delivery capabilities available with Stingray Traffic Manager permit load balancing across more than one region through two approaches: managing dispersed Stingray Traffic Manager instances as a single cluster, and geographically-aware global server load balancing. Additionally, Stingray Traffic Manager cooperates with Amazon EC2 Auto Scaling and can dynamically add or remove application instances as load varies within and across AZs and Regions.
Multi-site cluster management
Within a multi-site configuration (see figure 1), a group of Stingray Traffic Manager instances deployed over one or more AZs or Regions forms a centrally managed cluster to provide the delivery of application services in a fault-tolerant manner.

![Figure 1. Stingray Traffic Manager performs local and global load balancing across Amazon Web Services Regions while providing advanced clustering features for simplified management.](image)

Stingray Traffic Manager service configuration is replicated between all instances in the cluster. As with non-multi-site environments, the majority of the configuration data is shared. However, multi-site management extends this facility by making it possible to set AZ and Region-specific configuration that is active only on the instances marked at that AZ or Region. This provides a form of sub-clustering, or local service delivery.

Multi-site cluster management is commonly used when the services made available in one Region are duplicated in a second Region. Regardless of their locations, the Stingray Traffic Manager instances are managed as a single virtual cluster. The management interfaces place the instances on an interactive world map, which allows administrators to see a broad overview of resource deployment and also narrow the view to specific Stingray Traffic Manager instances, application servers, and users requesting services.

Global server load balancing
Stingray Traffic Manager can be configured to provide powerful and easy-to-use global server load balancing (GLSB) and failover functionality across multiple Regions (see figure 2). It is easy to deploy and gives rich feedback on site performance and traffic distribution.

![Figure 2. Global server load balancing improves user experience with geographic awareness of services and users.](image)

The primary purpose of GSLB is business continuity—to ensure that services are always available, even when one or more service locations become unavailable. A second purpose of GSLB is to improve customer experience—to route each user to the best location from a choice of several that are distributed across the globe.
GSLB manages how clients connect to a particular geographical location when a service is hosted in multiple Regions:

- **Active-passive**: One location is nominated as active for each service. The other locations are idle for that service. If the active location becomes unavailable, one of the passive locations becomes active and all clients are directed to it.
- **Active-active**: All locations are used and clients are load-balanced between them based on location performance and proximity.

To ensure that each client receives the best possible level of service, factors that influence the decision are completely customizable, and include location performance, client proximity, and resource demand and availability.

GSLB offers two mechanisms for load balancing:

- **DNS-based**: When a service is hosted in multiple locations, each instance of the service will be available from a different IP address. So that every client can use the same fully qualified domain name to access the service, Stingray Traffic Manager manipulates the DNS resolution of this domain name to control how each user is directed to a location. It monitors the performance and availability of each location to inform the load-balancing decisions it makes.
- **Stingray Traffic Manager-based**: In this situation, Stingray Traffic Manager operates with a standard round-robin DNS configuration. The DNS servers are configured to return the IP addresses of all of the locations hosting a service, and the software operates as a proxy in front of these DNS servers. It rewrites the round-robin DNS responses, ensuring that the client is directed to the most appropriate location.

### HTTP connection multiplexing

Stingray Traffic Manager manages client-side and server-side connections independently, re-using keepalive connections on the server side whenever possible to reduce the number of established and new TCP connections to the server. Connection multiplexing (see figure 3) pools a large number of incoming connection requests from individual clients into a small number of connections to the application servers. Multiplexing also keeps the connections to the servers open for longer periods. These techniques minimize the number of concurrent connections the servers need to handle, and brings big performance and capacity gains.

![Figure 3. Connection multiplexing consolidates many incoming client connections to reduce load on application instances.](image)

By reducing the overhead required for managing individual client connections, Stingray Traffic Manager enables servers to respond more rapidly to a higher volume of incoming requests. An existing application server instance immediately benefits from the ability to handle a much greater number of simultaneous incoming requests than without the solution.

Unlike some other ADCs, Stingray Traffic Manager implements this feature transparently. No special configuration or scripting rules are required to benefit from the performance gains that come from connection multiplexing. Developers can concentrate their efforts on rules that improve application delivery while allowing the ADC to automatically optimize performance.

### Client-side performance management

#### Bandwidth management

Stingray Traffic Manager can limit the bandwidth used by inbound or outbound traffic. Normally, network bandwidth is provided at the highest rate possible for all connections. This may result in uneven use of an application, possibly with too much bandwidth used by secondary services at the expense of more critical services. Bandwidth management controls this imbalance. For example, a 20 Mbit/s network connection that is over-utilized by anonymous browsers would negatively affect the responsiveness of logged in users. To ensure better performance for logged in users, a rule can limit outbound bandwidth to 2 Mbit/s for all connections that lack an authentication cookie. Bandwidth limits are automatically shared and enforced across all Stingray Traffic Manager instances in a cluster. Depending on the load on each instance, individual instances take different proportions of the total limit and unused bandwidth is equitably allocated.
across the cluster depending on the need of each instance.

**Request rate shaping**

Individual users may dominate the use of a service to the detriment of other users. A back-end application infrastructure with limited scalability can be easily overwhelmed when too many requests are given to it. Request rate shaping restricts the rate at which certain activities can occur. Per-second and per-minute limits can be specified on a wide range of events, with very fine-grained control over how events are identified. Some examples include:

- **Rate-shaping** individual web spiders, stopping them from overwhelming a web site. Each web spider, from each remote IP address, can be given maximum request rates.
- **Throttling** individual connections, or groups of connections from the same client, so that each connection is limited to a maximum number of transactions per second.
- **Globally rate-shaping** the numbers of connections per second that are forwarded to an application.
- **Restricting** each user to a limited number of attempts per minute by identifying an individual user’s attempts to log in to a service and impeding any dictionary-based login attacks.

**Additional capabilities**

In addition to the capabilities described above, Stingray Traffic Manager on Amazon Web Services provides traffic management features that are unavailable with traditional load balancers. These include:

- Advanced session persistence of non-HTTP/HTTPS applications
- Client IP address transparency
- Application availability and latency monitoring
- Content-based routing decisions
- Incoming and outgoing request/response traffic manipulation
- Balancing of TCP traffic on any port
- Instance draining (to prepare it for clean removal)
- Web traffic compression (when necessary)
- Web content caching

Combined with the robust and reliable Amazon Web Services cloud computing infrastructure, Stingray Traffic Manager allows organizations of all sizes to quickly attain global reach and scale.


**About Riverbed**

Riverbed delivers performance for the globally connected enterprise. With Riverbed, enterprises can successfully and intelligently implement strategic initiatives such as virtualization, consolidation, cloud computing, and disaster recovery without fear of compromising performance. By giving enterprises the platform they need to understand, optimize and consolidate their IT, Riverbed helps enterprises to build a fast, fluid and dynamic IT architecture that aligns with the business needs of the organization. Additional information about Riverbed (NASDAQ: RVBD) is available at [www.riverbed.com](http://www.riverbed.com).