Pulse Secure®

WHITE PAPER

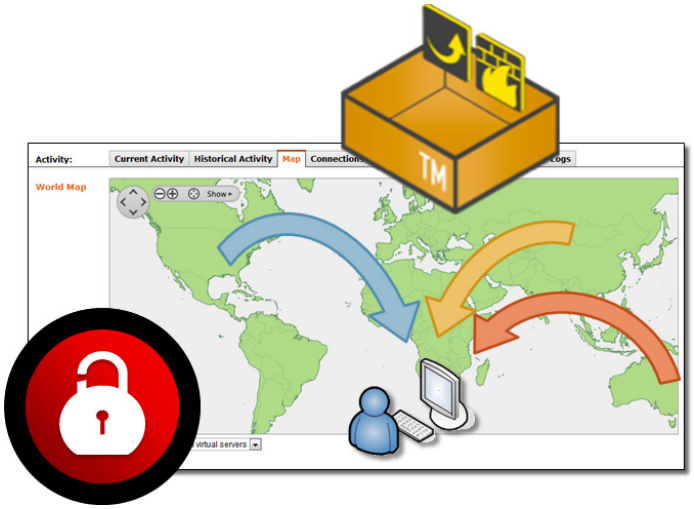
# Global Load Balancing with Pulse Secure Virtual Traffic Manager

## Introduction

Every year, global enterprises suffer application downtime due to failures in software or infrastructure, whether the applications are in secure data centers, or in public clouds. However you measure it, the cost of application downtime can be very high for many organizations. For enterprises that provide online services and applications, the cost of downtime impacts revenue, sales and customer loyalty.

Global Load Balancing is often used to improve application response and service continuity for geographically-distributed customers. Pulse Secure Virtual Traffic Manager can help to optimize service response times, as well as simplify failover and recovery, giving a real-time response to application workloads or service outages.

## Server Load Balancing Within a Data Center




There are two commonly used techniques to minimize the chance of a failure causing downtime in network-based applications. These are Server Load Balancing and Global Load Balancing.

Techniques like server load balancing and clustering are often used within a data center to build clusters of fault-tolerant, scalable applications. These clusters are resilient to isolated failures—for example, when a server develops a hardware fault. Clusters also make it easy to add more capacity to an application to meet growing demands. However, a clustered, fault-tolerant application running in a single data center is still vulnerable to downtime:

- The application may fail because of a single, critical point of failure such as a database or SAN, or it may fail because of administrator error
- The data center may be disrupted due to a catastrophic natural or man-made disaster, such as power failure because of rolling blackouts, maintenance errors or even malicious attack
- The data center may become unavailable because of a denial-of-service attack mounted against a different service running in that data center, or because of a failure in local network connectivity.

Organizations who wish to protect against these risks often choose to deploy a Global Load Balancing solution which routes application traffic to multiple distinct data centers and removes the single point of failure.

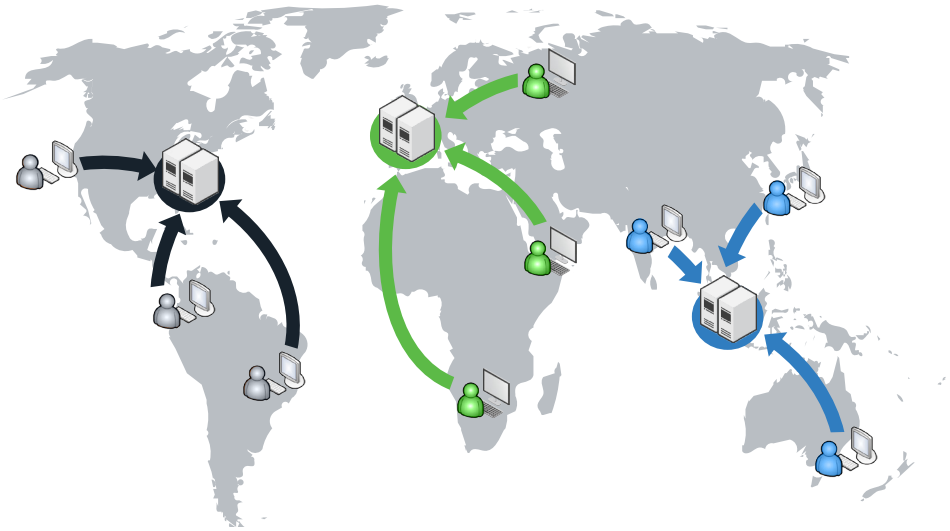
## Global Load Balancing Between Data Centers

Global Load Balancing (GLB) systems manage how clients are connected to a data center, when a service is hosted in multiple distinct data centers.

- In an Active-Passive configuration, one data center is nominated the active one for each service. The other data centers are idle for that service. If the active data center becomes unavailable, one of the passive data centers becomes active and all clients are directed to it.
- In an Active-Active configuration, all data centers are used and clients are load-balanced between them based on data center performance and proximity

The primary purpose of a GLB system is **Business Continuity**—to ensure that services are always available, even when one or more service locations (data centers) becomes unavailable.

A second purpose of GLB is to Improve **Customer Experience**—to load-balance each user to the best data center from a choice of several. The choice can be based on data center performance and proximity, so that clients are directed to the data center that is closest and is performing the best, so that clients gets the best possible level of service.



A GLB device can direct users to the closest data center for best levels of service

## Who Might Use a Global Load Balancing Solution?

A GLB solution is relevant to:

1. Organizations which provide or depend on an online service, such as a public-facing Web site, or a network-based application for internal use
2. Organizations which cannot tolerate service failure, whether this results in lost productivity, lost revenue, or lost customers
3. Organizations which need to improve the SLA (Service Level Agreement) with its users or customers, providing them with a superior and competitive level of service

# Case Studies

## Disaster Recovery

A specialist music and book retailer processes a large number of orders every day. Any period where users could not access the online shop would result in significant loss of revenue and reputation.

The retailer hosts their primary Web site in a hosting facility in New York and replicates all database transactions to a second backup Web site in Boston. During normal operation, users are directed to the New York Web site, but if that Web site becomes unavailable, a GLB service directs all users to the backup site in Boston.

When a contractor severed a fiber optic cable in the New York hosting facility, the GLB service detected that the site was no longer accessible and immediately started directing users to the backup site in Boston instead. As the database was continually replicated, users were able to continue with their transactions and complete their purchases. Furthermore, since database replication is often one-way, when the cable repair occurs, the system should not fail back until the customer is ready, and has ensured the databases are back in sync.

## Providing High Levels of Service

A UK-based publishing company publishes several prestigious scientific journals. Universities and research institutions across the world pay a subscription to access the content of these journals electronically.

A disaster recovery solution is required because the paid subscribers will not tolerate downtime. In addition, many of the subscribers in the U.S., Far East and Australasia report that the Web site is slow, and it can take too long to download premium PDF content.

The publishing company establishes mirror sites in the US and Japan and uses a GLB service to direct each user to the site that is geographically closest to them. Download times for many customers drop by up to 75 percent.

## Upselling Services to Hosting Customers

An innovative ISP was seeking additional services they could provide to their hosting customers. Using data replication to a server platform located in a different data center, the ISP was able to synchronize customers' Web content between two locations.

With a GLB service, they were able to direct traffic for some customer sites to the City North data center, and other sites to the City South data center, and thus control and manage the bandwidth used by each data center.

The ISP's customer's SLA contracts contained exclusions for major data center failure caused by elements outside the ISP's control. For an additional fee, the ISP was able to upsell a premium hosting package that included a data center failover service to minimize the risk of a data center failure rendering a customer's site inaccessible.

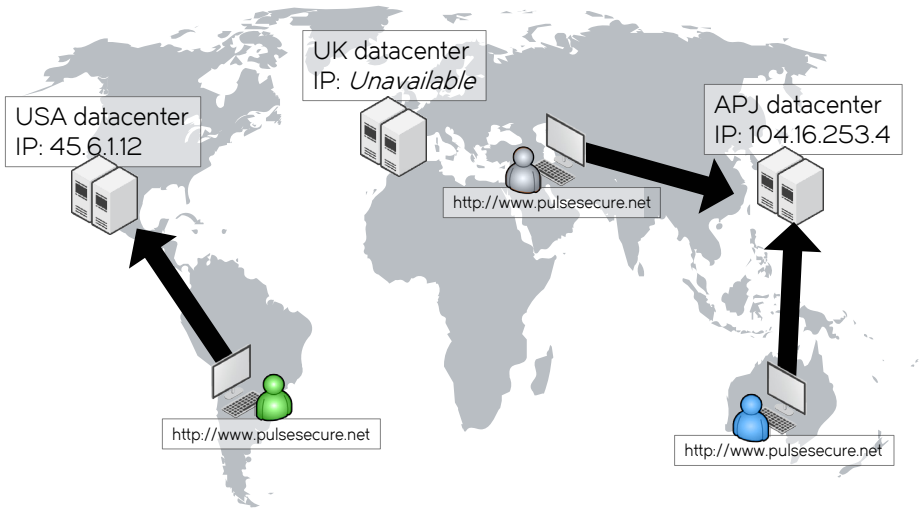
# How Does Global Load Balancing Work?

## DNS-based Global Load Balancing

The majority of GLB devices function by manipulating the Domain Name System (DNS) resolution process. An application such as a Web browser needs to locate a service on the intranet before it can use it. Services are published using a Domain Name, such as www.pulsesecure.net.

Behind the scenes, the application uses a process called “DNS Resolution” to find out the IP Address of the Internet server that provides the service with the given domain name. The DNS system is very much like a global Internet phone book—you may know an individual by their full name (for example, “Tim Berners Lee”), but you need to look up their phone number before you can get in touch with them.

Different servers in different locations will have different IP addresses. A GLB device controls how domain names are resolved to IP addresses, and thus controls which data center clients are directed to.




Several users access the same Web site, but they are each directed to different data centers: When users in the US try to access www.Pulse Secure.com, they are directed to IP address 45.6.1.12—while users in other locations are directed to IP address 104.16.253.4.

In order to deploy a GLB solution effectively, you need a good understanding of how the DNS system functions.

## Other GLB Designs

Other techniques are sometimes used to load balance users across several globally-distributed data centers.

## Application Level Redirection

Some application protocols, such as HTTP, allow for “redirection” messages.  A user accesses www.pulsesecure.net, but receives a redirect sending him to us.Pulse Secure.com, which resolves to just one of the data centers.

This method is effective at controlling precisely which data center a user is sent to, but it does not cater for data center failure, and users may bookmark or distribute links to www.pulsesecure.net, bypassing the load-balancing decision.

Generally, this method needs to be implemented by using a DNS-based GLB system to ensure that www.pulsesecure.net always available, and a traffic management device to control how and when users are redirected.

The Pulse Secure community Web site includes more information on how to use Pulse Secure TrafficScript to implement layer-7 application redirection for GLB persistence.

### Anycast Routing Advertisements

When a customer has many locations, and has good control over their routing advertisements, they can use the Anycast methodology for advertising their IP blocks, as described in RFC 4786 as a best practice.

Pulse Secure Virtual Traffic Manager (Pulse Secure vTM) can implement Anycast functionality using RHI (Route Health Injection). However, when used in conjunction with DNS-based GLB service, Anycast can be used as a building block for more efficient DNS resolution, improving GLB performance.

(Implementation of RHI for Anycast GLB is outside the scope of this document)

## Typical Pulse Secure Virtual Traffic Manager Deployment

With Pulse Secure vTM, Global Load Balancing can be deployed with minimal interference or disruption to the existing infrastructure.

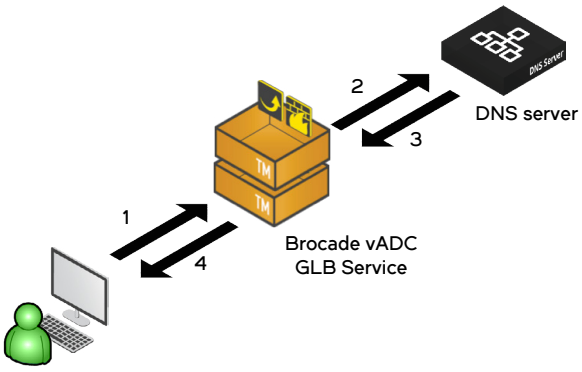
Pulse Secure Virtual Traffic Manager works alongside the existing DNS infrastructure, taking the DNS responses and manipulating them to control where each remote user is directed. In this guide, we will assume that the Pulse Secure vTM is operating in front of an existing DNS server. However, Pulse Secure vTM can also act as the DNS server for this deployment. From a logical point of view however, the internal DNS server can be thought of as an "external" DNS server, and the GLB configuration is performed in the same way.

### Begin with Round-Robin DNS

For example, suppose that the service www.pulsesecure.net is hosted in two different locations, with IP addresses 21.2.12.1 and 45.4.54.5. Without a GLB capability, the DNS server would normally be configured to return both of these IP addresses when queried about www.pulsesecure.net. The IP addresses would be returned in a different order each time using a process called Round-Robin DNS, and clients would connect to one of the data centers. It is important to note that Pulse Secure vTM will only modify A records that match the GLB services hosted by Traffic Manager. If other IP addresses are included in the response, they will NOT be modified.

### Add in Pulse Secure Virtual Traffic Manager

Pulse Secure GLB builds on this standard configuration by manipulating the round-robin DNS responses:




1. The end user makes a DNS request for www.pulsesecure.net.
2. Pulse Secure vTM forwards the request to a DNS server (either external, or built-in DNS).
3. The DNS server responds with all IP addresses in a round-robin fashion.
4. Pulse Secure vTM chooses one IP address and masks out the others from the response.

The key load-balancing decision that Pulse Secure Virtual Traffic Manager performs is to decide which IP address(es) should be returned to each remote user. This decision directly controls which data center each remote user uses.

Just one change needs to be made to the DNS information so that clients make DNS lookups through the GLB device rather than directly to the DNS servers. This change can be made by altering the NS record for the domain, or by adding a CNAME. Please refer to the Pulse Secure vTM documentation for more information.

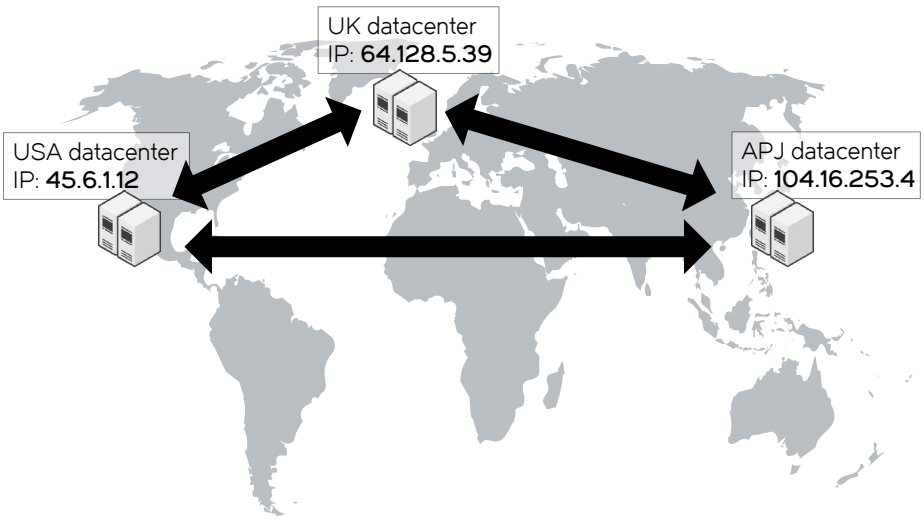
## DNS TTLs

DNS information is usually cached by intermediaries across the network. This caching behavior helps to reduce the amount of DNS traffic, but can interfere with the operation of a DNS-based GLB service.

An important element in a DNS response is the Time To Live (TTL) value. This value informs any intermediaries as to how long the DNS response can be cached for. Pulse Secure Virtual Traffic Manager can rewrite TTL values in the DNS responses, overwriting a long default value with a much shorter one. The effect of the change (increased DNS traffic) can be easily observed using the real-time visualization tools in Pulse Secure Virtual Traffic Manager, so you can chose a suitable value that balances traffic rates with responsive failover.

## How Does Pulse Secure vTM Global Load Balancing Work in Practice?

One or more instances of Pulse Secure Virtual Traffic Manager are deployed in each data center. Pulse Secure vTM monitors the performance and availability of the local data center, and broadcasts that information to the other Pulse Secure vTM instances in the other data centers. This way, every Pulse Secure vTM instance device knows the availability and performance of every data center.

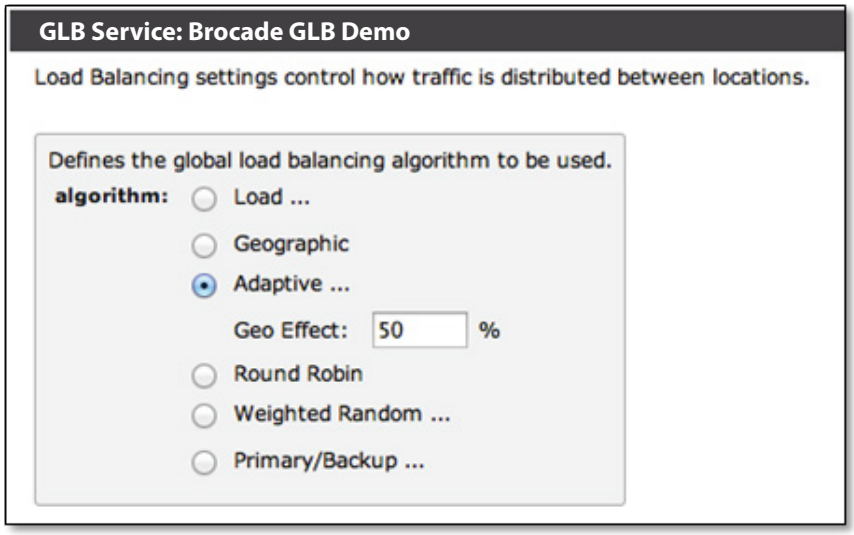


## Active-Active Load Balancing Configurations

Any of the Pulse Secure vTM instances may receive a DNS request for a service running in the data centers. When the data centers are running in active-active mode, Pulse Secure vTM chooses which data center the user should be directed to. This decision is based on three criteria:

- Data Center Availability: If a data center has failed, users are not directed there
- Data Center Performance: Data centers with better response times are preferred over slower, more overloaded data centers
- Geographic Proximity: Pulse Secure Virtual Traffic Manager uses a comprehensive database that maps IP address to geographic location, and calculates the geographic distance between the end user and each data center.

The decision criteria can be tuned so that it is based purely on load, purely on geographic location, or on a mixture of the two:




An active-active load balancing model gives better data center utilization, and users get the best possible level of service from the closest, best performing data center. In addition, the configuration provides full failover in the event of a data center failure.

However, you may prefer not to use an active-active configuration if the applications cannot be run in multiple data centers simultaneously—for example, they depend on a single database or SAN that cannot be continuously replicated over multiple sites. In this case, an active-passive configuration is more appropriate.

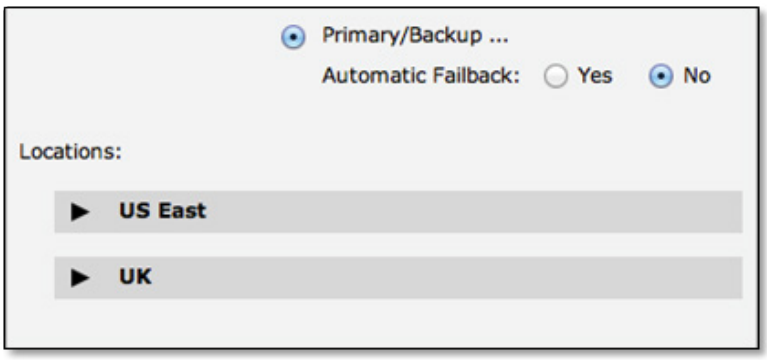
Another side-effect of an active-active load balancing model is that an end user may spontaneously be redirected from one data center to another if the client software makes a fresh DNS request. For example, the first data center may become overloaded and the load-balancing algorithm may assign subsequent requests to an alternate data center.

This behavior can be solved in several ways. You can use the fully deterministic "Geo" load-balancing method, or you can use application-level redirection to detect user's sessions and force requests to a particular data center when required, using Pulse Secure Virtual Traffic Manager to manage session persistence.

The Pulse Secure community Web site includes more information on how to implement multisite session persistence with Pulse Secure Virtual Traffic Manager

## Primary-Backup Load Balancing Configurations

When the data centers are running in primary-backup mode, the load balancing decision is much simpler. You first specify the order in which the data centers should be used:




In this example, all users are directed to the first data center (Hudson in this case) so long as that data center is available. If the first data center fails, all users are directed to the second data center (Cambridge); you can build arbitrarily long chains of data centers for multiple levels of failover.

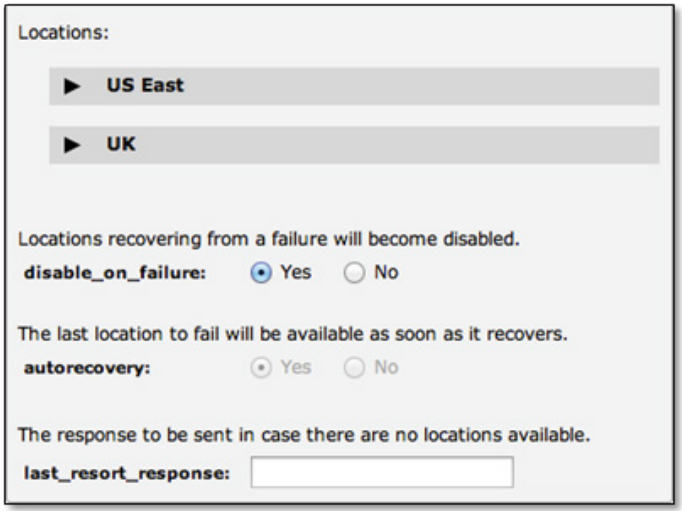
If the first data center recovers, you can specify how the service should fail back. If automatic failback is enabled, users will immediately be directed to the first data center again. If it is disabled, users continue to use the second data center until the administrator manually indicates that the first data center is ready to receive traffic again.

The benefit of this configuration is that it gives a very deterministic, controllable disaster recovery solution, ideally suited for complex, stateful applications.

## Hybrid Active/Passive Load Balancing Configurations

For more complex failover scenarios, Pulse Secure Virtual Traffic Manager also supports the ability to configure behavior when failure is detected. The first option is to allow automatic failback, so the site that is designated as primary will always be active if available. Next is to perform a “disable on failure” to explicitly force a site to go offline and stay offline. This is useful when a manual process is needed to re-sync locations once they go offline. Third, is to allow “last location recovery,”w on the basis that the last location that was online can always be considered the safest to bring back online, due to database consistency.

These options are shown here:




## Availability and Performance Checking

Pulse Secure Virtual Traffic Manager checks the performance and correct operation of the services in the local data center using a range of application monitors. These monitors can run simple tests like network pings, or complex tests like HTTP GETs, to verify that returned pages match particular criteria.

Performance data can optionally be deduced from the response times from selected monitors, or it can be supplied separately using SOAP or REST APIs. This performance data is used to weight how much each data center is used when the Load or Adaptive load balancing algorithm is selected. Traffic Manager can also run an external connectivity monitor to verify that the local data center has connectivity to an upstream location on the Internet.

Pulse Secure Virtual Traffic Manager broadcasts the health and performance data to the other Traffic Manager instances in the other data centers. It determines the availability of other data centers from the health and performance information broadcast from the Traffic Manager instances in those data centers. For this reason, organizations typically operate a pair of instances in each data center, thus removing a potential single-point-of-failure within each data center.


The Global Map view in Pulse Secure Virtual Traffic Manager shows real-time site activity, ideal for displaying in a network operations center or support group.

## Conclusion

Pulse Secure Virtual Traffic Manager can provide a complete DNS-based Global Load Balancing solution that provides:

- Business Continuity in the event of catastrophic data center failure
- Improved Customer Experience by routing users to the closest, best performing data center

Pulse Secure Virtual Traffic Manager is very easy to deploy, with minimal infrastructure changes and very little operational risk. The rich real-time visualization and reporting gives a clear picture of the effectiveness of the Global Load Balancing configuration and the activity of global users at any time.

## Find out More

To find out more about Pulse Secure Virtual Traffic Manager and Global Load Balancer solutions, or to arrange a demonstration or product evaluation, please visit http://www.pulsesecure.net/vadc.

**Corporate and Sales Headquarters**
Pulse Secure LLC
2700 Zanker Rd. Suite 200
San Jose, CA 95134
www.pulsesecure.net



Pulse Secure®