

Stonesoft Technical Note

StoneGate Server Load Balancing An Explanation and Typical Scenarios

By Paul Brettle

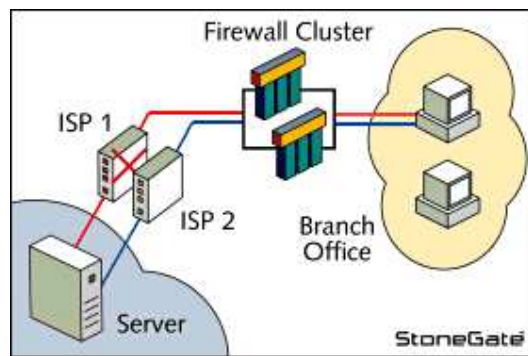
Friday, 09 March 2007

Server Load Balancing

With the demands of modern businesses, scalability of servers is now seen as a critical issue. The move to a centralised data centre solution has obvious cost savings with reduced administration and management. However, these solutions do cause problems when managing the load and availability of the servers and the services they provide. The increased criticality of the servers means that it becomes important to ensure up time and performance at all time. This causes headaches for many organisations as they balance cost and complexity with availability.

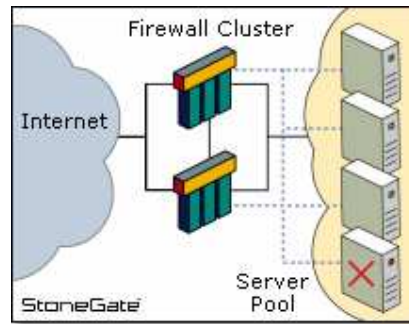
Stonesoft provides an extremely flexible platform for security and availability with the StoneGate solution. With its ability to support multiple network links and multi-homed ISP connections, it is ideally placed to balance traffic to network server.

Multi-link provides support multiple Internet connections simply and easily. This means that it is possible to present a server on two different network connections maximising availability should any one fail. Fortunately, StoneGate manages this for you without having to resort to complex DNS solutions.



But this does not necessarily mean that this load balancing is for just one server. In fact a pool of servers can be defined with the most suitable one selected to spread the load. This can be done on a simple availability check, or can use a much more sophisticated agent to provide more accurate information. This means that it is now simple to present a pool of servers on multiple network links without having to resort to application re-writing or

complex software based solutions. By providing the load balancing at the network layer, compatibility is maintained and flexibility



To further complement the Server Load Balancing provided by StoneGate, a server load agent could be used. Included as part of the StoneGate solution, it is able to test for a large number of conditions and feedback this information to the StoneGate load balancers. It is then possible to make a much more accurate decision based on the CPU load, amount of disk space available and even if a process is running correctly or not. More sophisticated tests can also be used; for example a URL can be tested and its response compared as well as supporting custom scripts for any unusual systems.

This data is then collected by the agent and constantly fed back to the StoneGate load balancing system. These metrics are then weighted and an accurate selection of which server to use next is made. This evens out the load between the servers and ensures that no single one is overloaded, as can be the case with round-robin type load balancing systems.

The server load balancing can be manually maintained easily. Network links can be disabled if necessary and servers can be removed from a pool for maintenance. This can all be carried out from the management GUI remotely. This ease of management makes the StoneGate solution uniquely easy to ensure uptime.

Finally, the StoneGate solution is a fully-fledged firewall with EAL 4+ certification. Additionally StoneGate can be easily clustered to remove single points of failure and maintain availability in all situations. Therefore StoneGate is not only a server load balancing solution, but an integral network security solution.

Server Load Balancing

- Allow for multiple connections, including ISP links to be balanced
- Network based load balancing to reduce complexity and impact
- No third-party server load balancing solutions needed
- Server agent to make accurate load decisions

StoneGate Solution

- Firewall – EAL 4+ with full stateful firewall with application intelligence
- Clustering – ensure maximum uptime with load sharing clustering
- Central administration – enterprise class management and control

Frequently Asked Questions

What is Server Load Balancing?

Server load balancing is where the traffic to a number of servers, typically two or more, is intelligently routed to the most available destination. As far as the client whom is making the network connection, the pool of servers is simply just a single server. The network load is then balanced between the pool of servers and hence ensures an even spread and the highest availability as necessary.

What is the Sever Monitoring Agent?

With StoneGate, an agent is available to provide more realistic information as to what is happening on a server. It is possible to use network availability as a test to use a server in a pool, but this is not always the best guide for the load on a particular server. Therefore an agent is available to install onto the server to provide more meaningful information to StoneGate. Typically this will be data such as memory usage, processor load, server availability and even additional information such as free disk space.

How does StoneGate provide Load Balancing to Servers?

StoneGate makes use of a special configuration option called a “Server Pool” object. This is then used in the rulebase for the firewall and allows one or more server to be in the pool. Traffic to the server is then passed through the rulebase for security and then intelligently routed to the best server to process that traffic. The decision to send the traffic to which server pool member is based on availability of that server as well as other data such as load etc.

Can StoneGate support Multiple ISP's?

Yes, StoneGate can support multiple ISP connections through the Multi-Link technology. This means it is possible to load balance network traffic from two or more Internet links to two or more server pool members. This reduces the impact of a network failure such as an ISP connection, the other Internet links will continue to process the network traffic. This can be coupled with the server pool object so that it is possible to balance traffic to a pool of servers on multiple ISP links with ease.

Is StoneGate now a single point of failure now?

If StoneGate is providing a server pool and is making use of multiple ISP links then yes, it is now the single point of failure in the resilient network configuration. In this circumstance it is highly recommended to make use of the in-built clustering that is available. It is then possible to reduce the impact of a failure of any part of the highly available solution to a bare minimum. For organisations wishing to improve the availability of their externally facing applications, this is a flexible and effective solution to do this.

StoneGate Server Pool Configuration

Provide below is a sample configuration dialog box for a server pool object. In this case it is possible to see that two ISP links have been used to two server members in the Pool.

netilla.example.com - Properties

Name:

Category: OBJECTS [HEADQUARTERS] Categories...

Comment:

External Addresses

NetLink	IP Address	Status	Proxy ARP Entry
hqcluster.netlink.cogent	78.56.120.202	Enabled	YES
hqcluster.netlink.sprint	89.44.230.202	Enabled	YES

Enable Dynamic DNS Updates

DNS Server: Fully Qualified Domain Name:

Server Pool Members

Element	IP Address
ftp-1.example.com.local	192.168.1.211
ftp-2.example.com.local	192.168.1.212

Allocate traffic to servers by:



Monitoring: Ping Agent Port:

OK Cancel Help

Further information is provided in the following pages.

Defining External Addresses



StoneGate makes use of what is known as NetLinks. It is these that are used to provide the multiple ISP support with the Multi-Link functionality. Each NetLink defines the ISP and address range in use.

External Addresses		
NetLink	IP Address	Status
 hqcluster.netlink.cogent	78.56.120.202	Enabled
 hqcluster.netlink.sprint	89.44.230.202	Enabled

In the example above, there are two NetLinks in the configuration. Each NetLink has an IP address to which the Server Pool is to use. Because there are two ISP links each one has a different address range. In this case there is one connection from Cogent and the other from Sprint. The Server Pool is then available on both of these specific addresses.

Defining Server Pool Members

Typically a server pool has two members in it. However, it is perfectly acceptable to have just a single member. In this example there are two servers on two different IP addresses.

Server Pool Members	
Element	IP Address
 ftp-1.example.com.local	192.168.1.211
 ftp-2.example.com.local	192.168.1.212

In this particular example the IP addresses are actually sequential. However, there is no restriction to the addresses used. They can be on different ranges and even different locations. The only requirement is that StoneGate can directly contact the pool members.

Defining Traffic Allocation Scheme

By default a server pool makes use of a simple connectivity check for the members. This is carried out by a simple ping message sent from the firewall to the members to check they are still available. If any member does not respond then it is assumed that it is not available and hence is removed from the pool automatically. Traffic is then routed to the other pool members.



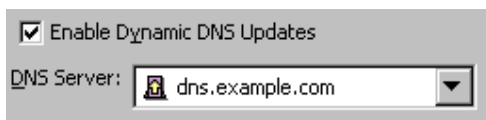
Allocate traffic to servers by:

Monitoring: Ping Agent

It is also possible to use the Server Monitoring Agent which will feedback more accurate information about a member server. A simple network connectivity check is not necessarily the best guide and the agent on the server can monitor other key metrics. Using the Agent it is possible to directly check that the server processes are running as well as the overall load on the server. This information is fed back to StoneGate and a more informed decision is made on the selection of which server to use.

Dynamic DNS Support

When services are being made available to the Internet, they are almost always accessed via their domain name. Since StoneGate provides the ability to offer the same service on different ISP connections, it is possible to use a Dynamic DNS service.



Enable Dynamic DNS Updates

DNS Server:

When configured, the Dynamic DNS configuration will constantly check the ISP links and should any one of the available links be unavailable a DDNS update is sent. This means that the ISP link that is not available will be removed from the DNS resolution for that particular service name. When the ISP link becomes available again, a further update will

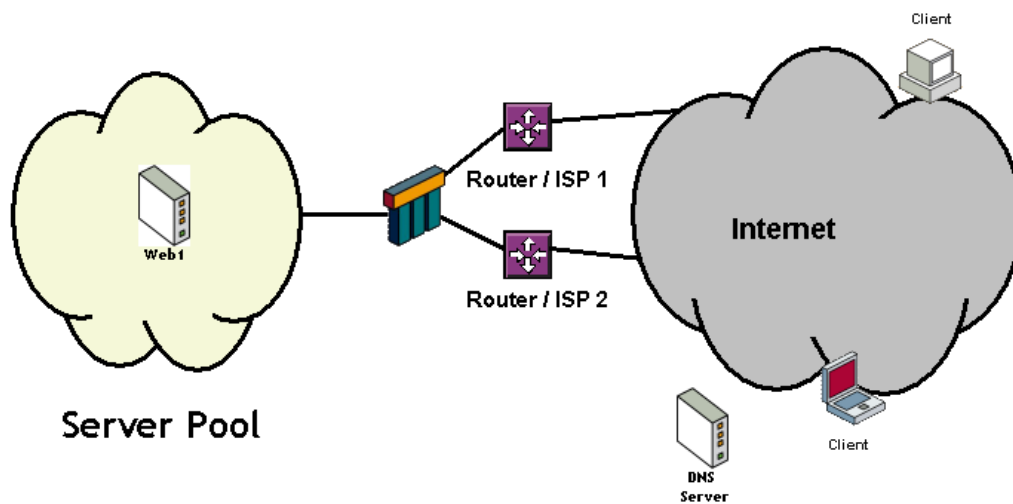
take place to add the IP address back in again. This further minimises the impact of an ISP failure on the availability of the network server.

Please note that it is not essential to use the Dynamic DNS configuration. By default all TCP/IP clients will do a DNS lookup to convert the name to an IP address. If the DNS record for that name responds with two or more IP addresses, the client will randomly select one address to use. If that is unavailable for any reason the next IP address is used. The Dynamic DNS system reduces the chances of this happening by ensuring that the records are updated, but it is not essential to use this feature.

Typical Deployment Scenarios

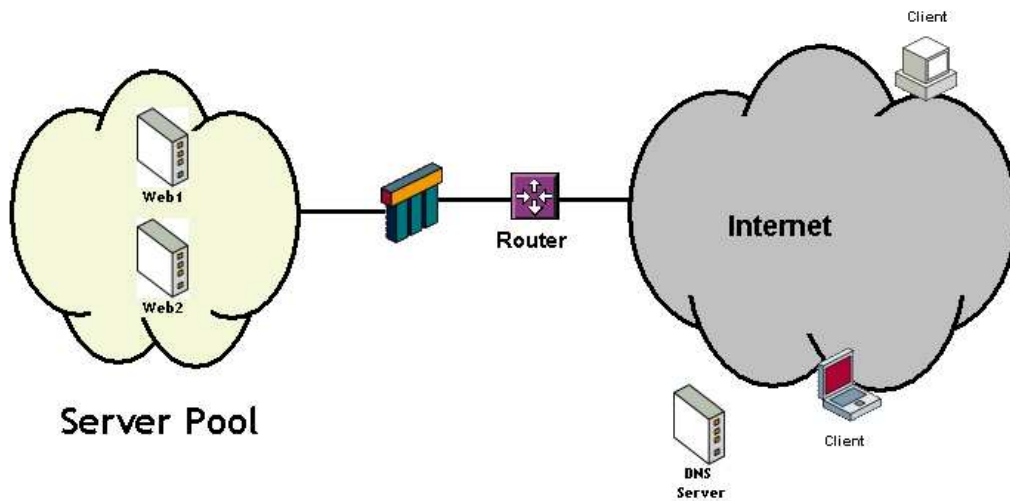
Two ISP connections – 1 Member to Server Pool

In this scenario, a single member server is made available on two separate ISP links to the Internet. This adds availability and resilience to the server in that it is accessible from two different ISP routes. In this case there is still only a single server and it therefore not possible to balance the load between multiple servers.



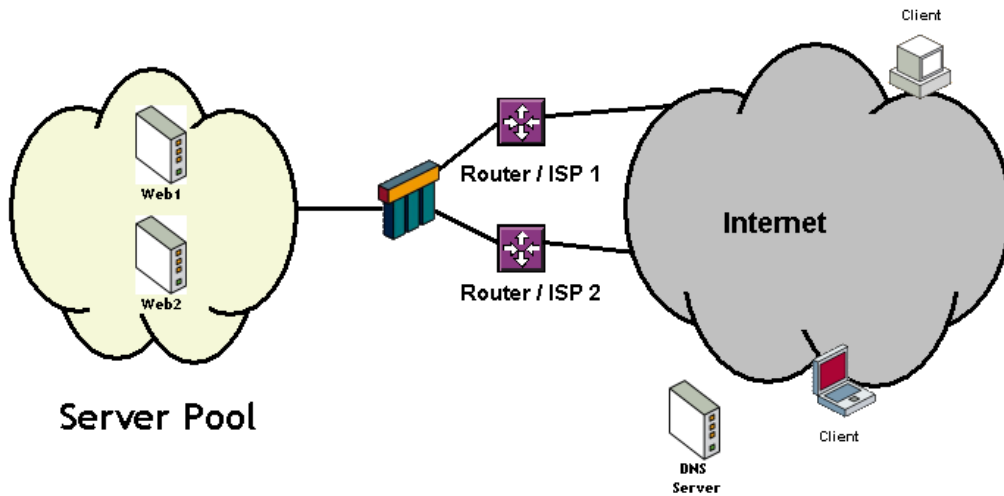
Single ISP connection – 2 Members to Server Pool

In this scenario two separate member servers are presented to the Internet through a single ISP link. This adds availability to the service being offered, as there are two servers to share the load for the application being presented. However, since there is only a single ISP link there is still only one route to the Internet and back.



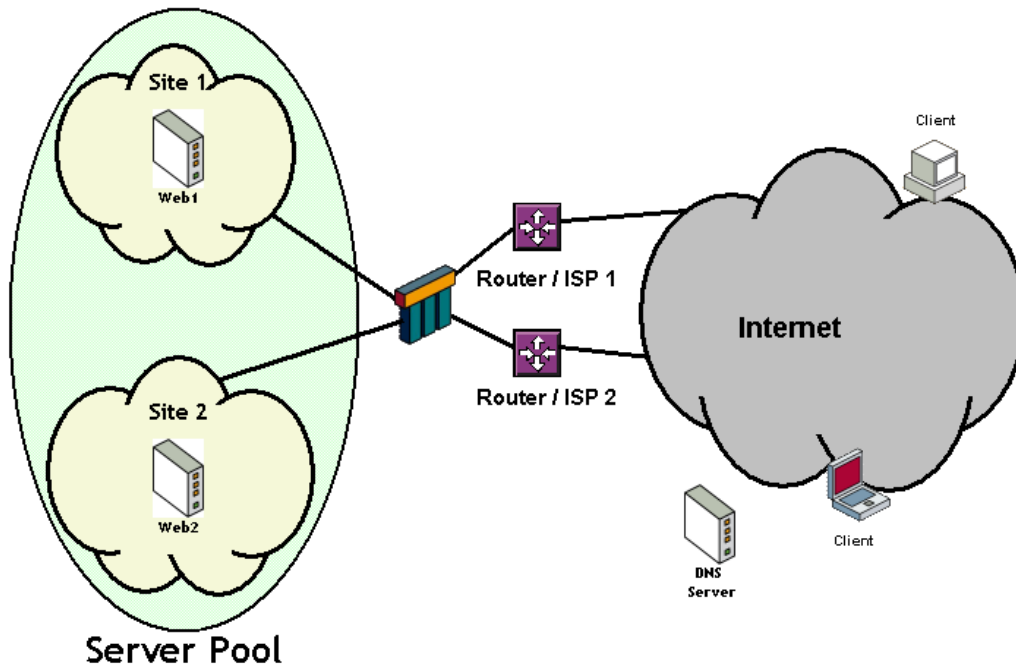
Two ISP connections – 2 Members to Server Pool

In this scenario, the most popular by far, an application is presented to the Internet on two separate ISP connections. Additionally there are two servers that can be used for balancing the load. Therefore in this scenario it is possible to still carry on operating the application even when one server and one ISP link is unavailable.



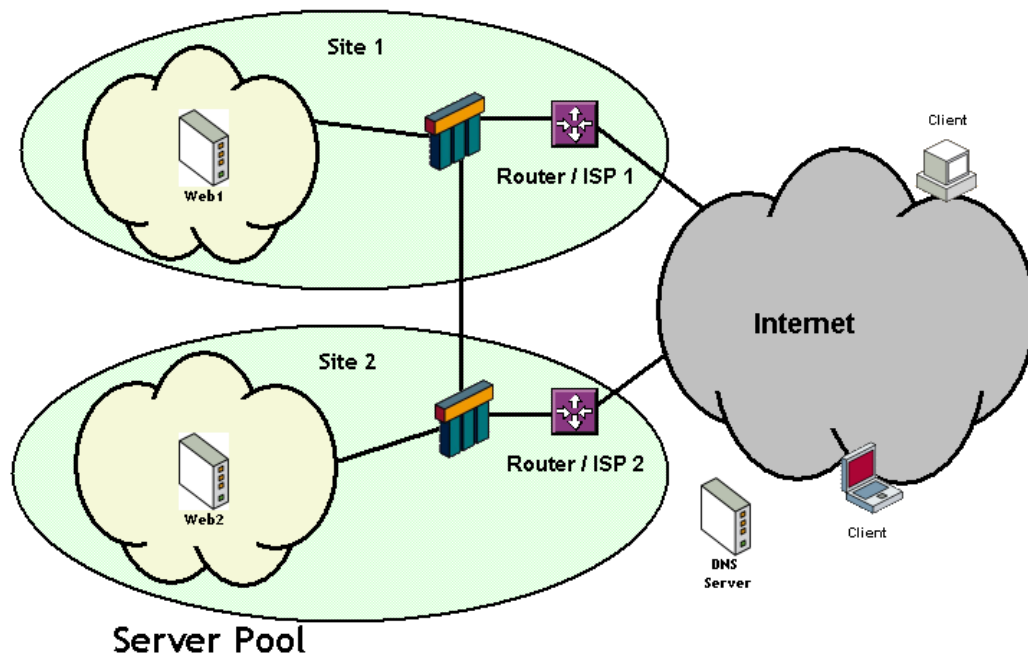
Two ISP connections – 2 Members to Server Pool located on different sites

In this particular scenario there is little difference to the one above. However, the locations of the two servers are now in separate sites. Since StoneGate makes use of IP addressing rather than location, the load is still balanced to the separate servers in the pool.



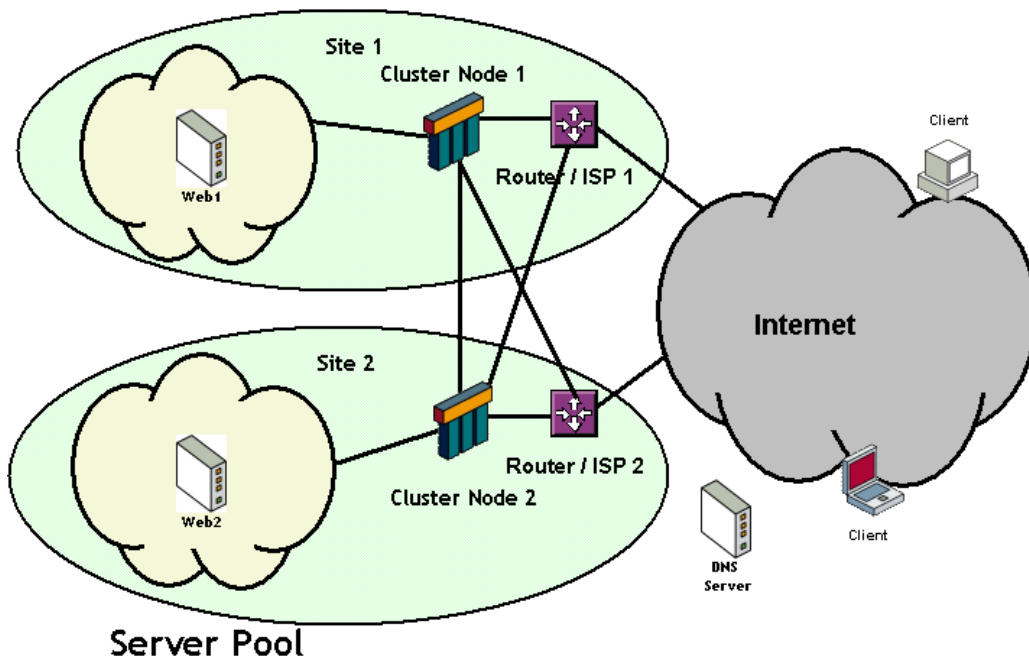
Separate StoneGate on Two sites – 2 x ISP, 2 x Member Servers

In this particular scenario the complexity is much higher but the resiliency is also. Here, not only are the member servers to the pool in separate sites, but the StoneGate appliances and the ISP connections are. Therefore it is possible to present the service to the Internet even if an entire site is not available. Additionally, traffic will be routed between sites if the member server or ISP link is not available. However, there is still a requirement for a direct connection between the two sites and the StoneGate appliances are not clustered across the sites.



Clustered StoneGate on Two sites – 2 x ISP, 2 x Member Servers

This scenario represents the most complex, but equally the highest resiliency solution to the Server Pool members. In this case the StoneGate appliances are clustered, but across the two available sites. This means that the impact to the application being presented to the Internet is minimised to the smallest possible level. Here full resiliency is provided for all parts of the solution. However, it should be noted that a direct link is required between the two sites for the StoneGate communications as well as access to the other sites ISP link. Layer 2/3 switching typically resolves this network connectivity issue.



Even though it is this last scenario that illustrates the clustering of the StoneGate appliances, this facility is available on all scenarios. It is strongly recommended to make use of the StoneGate clustering for all scenarios since it ensures that the appliance is not the single point of failure for the solution. If clustering is not used then a failure of the StoneGate appliance will prevent the solution from operating.