LOAD BALANCER
/ technical datasheet

gigas

LOAD BALANCER

Gigas' Load Balancer is a virtual device that manages servers and assigns or balances client requests. The balancer incorporates a mechanism which regularly checks each of the servers, allowing – in a totally seamless way for the user – any nodes (or servers) that are down to be excluded from the pool.

Gigas' balancer has been developed by us and thus does not have a commercial license. It therefore enables us to offer a technologically advanced product at a very competitive price. A VLAN contract is needed prior to its installation. Clients can easily incorporate both products to their Cloud Datacenter in minutes via the Gigas control panel. As this is a virtual device, the user can enjoy it exclusively, resetting it at any time in addition to increasing any resources as required.

MAIN ADVANTAGES

Increases the system's load volume
The balancer allows the capacity of the systems among which the load is distributed to be increased. Whatever the limit of a server's web connections, when a second server is installed and is balanced, the two servers will work in parallel, allowing the system to support twice the amount of concurrent connections. Furthermore, by adding several front-end loads, connection speed is also increased.

More secure
The balancer is the only element that is exposed to internet and that has a public address. The servers to which the balancer distributes the load are connected to a private address. This means that if someone tries to hack into the system, they would only have access to the load balancer and would not be able to directly access the web servers.

High availability service
All virtual servers deployed by Gigas comprise a failover system. In other words, if the server (the physical computer where the virtual server is installed) has a problem, the system takes the virtual server from the physical server, reroutes it to another one and restarts it. This failover or high availability process can result in a downtime of a few minutes until the transfer of the server has been completed. For clients who cannot afford any downtime, the solution is to install a balancer. This constantly distributes requests between the two servers and checks the health of both. Before sending a request, the balancer checks that the server is working properly. In the case of a problem, instead of directing said request to a server (which reads it an as error) it is rerouted to another one. This grouping of distributed systems ensures that the correction of any server failure is seamless for clients.

High availability clusters
Cluster solutions ensure total availability in certain situations such as the failure of a balancer. When clustered, the balancer is comprised of, and works with, two devices (which monitor each other). When a request reaches the balancer, it functions as an entity or cluster and the request is rerouted to the web servers, but if there is a problem with the balancer, the other clustered balancer takes over the service in a seamless way and continues rerouting the requests. A high performance environment and high availability are thus ensured for clients. High availability of the balancer however, does not make their management more complicated as the system is responsible for maintaining the synchronized configuration between them.

KEY POINTS

- Virtual device for each client
- Increases load capacity and connection speed
- More secure, protects the server
- Seamless failover system for users
- High availability, cluster solution
- Balancing can be configured according to client requirements
- Comprehensive information with advanced statistics
- Virtual IP addresses
- Auto-scaling function

Features

CONFIGURING THE LOAD BALANCER
Gigas's Load Balancer offers clients three types of balancing:

1. **Round Robin or random balancing:** when clients want requests to the load balancer rerouted to the next server in line at random so that they sometimes go to one and sometimes to the other.
2. **Least conexión**: for clients who wish to add instructions to the balancing so that requests are distributed to where there is less load, thus ensuring a consistent load level among the balanced servers.
3. **Source of origin balancing/session maintenance:** Balancing in which a client is mapped to a same server. Specific applications exist where unless a server has failed, all requests originating from the same client will always go to the same server. This can be clearly seen in the case of control panels where the user needs

to log in to access them. If, when the user next visits the page, the balancer reroutes them to another server, the user will have to log in again. To avoid this, clients can choose this balancing option so that all requests from the same client are always rerouted to the same server in order to ensure session maintenance.

AUTO-SCALING

This feature allows the balancer to be responsible for monitoring the status of servers next in line. This means that when some of the instructions that the client has defined in advance are followed, the number of servers among which the service is being distributed is automatically increased. It is useful for automatically covering spikes in demand and traffic, and precludes the need of having a technician to monitor the system.

This advanced feature developed by Gigas can be purchased with balancing and VLAN options.

There are basically 3 drivers or instructions that can set by users when the balancer adjusts the resources:

- **Planned:** clients can define the exact moment when they require an increase in resources.
- **By maximum CPU threshold when using a server for a specific length of time:** the client sets an instruction so that when a server is using more than a certain % of the CPU during a specified length of time it detects a saturation problem. The balancer then automatically creates new servers within the load balanced environment.
- **By maximum number of sessions:** clients usually know the number of sessions that their server can support. An instruction can thus be given to the balancer so that when this maximum is approached, it automatically increases the overall performance of the service by adding new servers.

LOAD BALANCING STATISTICS

Users can check the load balance statistics via their own control panel. They can see how the balancer behaves whenever there is a request, how many servers it is balancing, to which server each request is being rerouted, what the status of each server is, if they are alive, if there is a problem, etc. All of this can be seen from the statistics that are included in the balancer.

ADVANCED CONFIGURATION

Other advanced balancing options are also available, such as:

4. **Consistent hashing:** enables the static assignment of requests to a server, so when servers are added or removed from the balanced group, the existing assignment is not affected
5. **Minimum/maximum number:** the client can set a minimum or maximum number of connections for a specific server. When this server reaches a maximum number of connections, the next request is directly rerouted to the other server.
6. **Priority between servers:** this can be set for servers that are balancing priority levels when rerouting requests.
7. **Cookie Jsession**: a program based on this type of cookie that allows static assignments to be set between a request that reaches the balancer and the server to which it is routed.

VIRTUAL IP ADDRESSES (VIPS)

The balancer has two types of IP addresses associated with it; a physical IP and a private IP, the latter being the one used to connect directly to the servers whose traffic is balanced. In addition to these two IPs, there is a third one called a virtual balancer IP (an IP with which the client configures their domain). This virtual IP is a public IP that floats between two balancers and is assigned to one or the other depending on the settings. It can therefore be associated with the first one, and, in the case of a problem, will automatically associate itself with the second one which will continue to receive all the requests, making the failover seamless for users.

ADDONS

Gigas' Load Balancer is a separate product offering various add-ons:

- Cluster option, to ensure uninterrupted, high availability service.
- Incorporation of virtual IPs, so that a single balancer can balance multiple web instances, multiple services, or web services with different IPs. This is necessary for certain configurations, and in our case the client can add on as and when they need according to their requirements.

TECHNICAL FEATURES

High availability
Throughput: up to 200 Mbits per second per load balancer
Up to 15,000 recurrent connections per load balancer
Allows various VIPs to be configured
Different balancing and health check systems for balanced servers
Real time statistics
Auto-scaling solutions enabled

Ref_BCa_v01