



White Paper

General Public Use

G/On

Load Balancing and Redundancy
Version 5.0

Connect, secure and grow

1. Scope

This document describes how load balancing and redundancy is normally handled in typical G/On installations to provide high availability configurations.

Introduction

In larger G/On installations supporting either critical applications requiring very high uptime or a high number of simultaneous users, load balancing and redundancy become critical features.

Giritech supports the fundamental features directly in the product, and maintains a set of guidelines for designing and building a strong solution. This White Paper outlines these guidelines.

More detailed information can be required by contacting Giritech directly, especially on how to map these generic guidelines to your specific IT infrastructure.

Load Balancing

Load balancing lets you spread server access across more than one server. This is not only a performance issue but, designed correctly, also becomes part of a high availability solution.

In general Giritech recommends using network level load balancing as this is substantially more efficient, in terms of performance, than using higher level load balancing.

Giritech therefore recommends using standard IP load balancers from, for instance, Cisco to perform the load balancing in front of the G/On servers.

The diagram below indicates how this could be implemented.

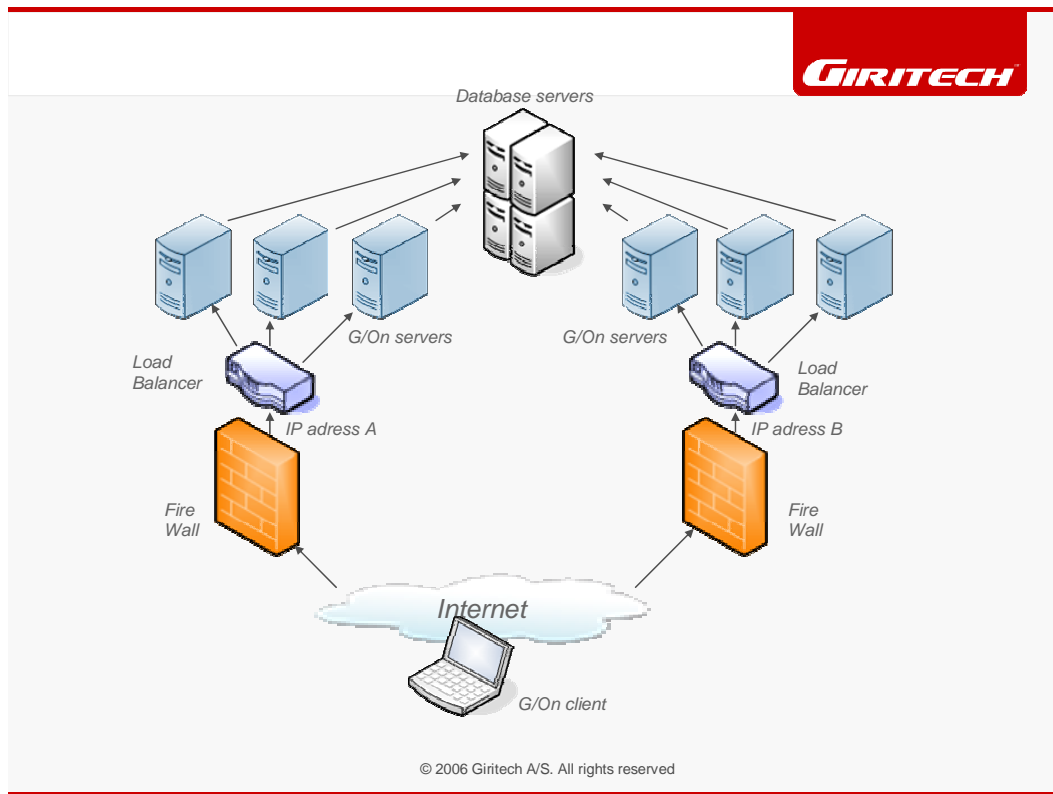
The G/On client (USB or Desktop) will automatically (if configured correctly) switch to an alternative server address (IP address A or B) if one should fail. It first tries address A and then switches to B if that fails.

The two firewalls in the diagram indicate two separate locations. Behind each firewall the load balancers distribute the incoming traffic between the connected G/On servers (three at each location in this example).

The G/On servers use a common EMCADS database (which is also redundant as indicated on the figure) to authenticate the client and the user and grant access according to the common setup in the database.

This configuration ensures that no matter which G/On server the users connects to they will always gain the same access rights to the network behind the G/On servers. It also ensures that any one server or appliance in the setup can fail without any remote user losing service.

The result is no Single Point of Failure (SPOF) exists in the path from the user to the application – provided of course that the applications are also redundantly configured to support full service.



Redundancy and Failover

Spreading the load across multiple servers “automatically” provides a minimum level of redundancy by having the users connect to different servers. It will therefore only be the users on the affected server that lose service in the event of server breakdown; all other users will obviously continue operation.

However, this is typically not a sufficient solution in IT environments where service access is critical. Combining the load balancing solution with a redundancy solution, allowing users to access more servers, is more effective.

G/On supports real redundancy by allowing the same users to automatically switch to alternative servers in the event that they cannot connect to their primary server.

This is implemented partly in the G/On client, which can be configured to automatically fall back to other servers (or “server farms” as indicated on the figure) if the primary server fails. The clients maintain a list of allowed servers (and their URLs or IP addresses – indicated as IP address A or IP address B on the figure). This list is controlled and updated from the central G/On server.

Secondly the G/On server can authenticate users via the common database configuration. This ensures complete transparency for users regarding which access rights they are allocated when logging on. Obviously this requires that the underlying database configuration also supports redundancy as well which will normally be the case in critical infrastructures as indicated on the figure above.

G/On connections are inherently very stable and automatically attempt to reconnect if a connection to a G/On server is lost. This means that if a server goes down, the reallocation to an alternate G/On server can be configured to be almost completely transparent for the enduser. If for instance the company is using Terminal Server or Citrix, the application session is not even lost. All the user experiences is a simple re-login with username and password (stateless failover).

Summary

Combining G/On's failover features with a network level load balancing setup and a properly configured database a fully redundant, a G/On installation can be designed to satisfy even the most demanding applications.

This White Paper confines itself to the general principles and recommendations behind this solution. Please contact Giritech for specific information on how to ensure your infrastructure matches these guidelines and therefore meets the requirements of your users and applications.