REDCENTRIC

SERVER LOAD BALANCING SERVICES SERVICE DEFINITION

SD069 V5.0
Issue Date 31st May 2018

redcentric
business technology. managed.

1) SERVICE OVERVIEW

1.1) SERVICE OVERVIEW

Redcentric offers both local and site based server load balancing services.

Local load balancing is used within a data centre to distribute incoming traffic across a pool of application servers. Local load balancing is used in designs where 1) the group of application servers is is likely to expand over time to accommodate increased traffic volumes, 2) application performance is vitally important and 3) where high availability is required.

Global or Site load balancing is used across data centres to either steer incoming application traffic to an alternative site if the primary fails or to distribute incoming application traffic across multiple data centres to either optimise performance or to deliver specific content to a user based on their location.

2) SERVICE DESCRIPTION

2.1) ASSOCIATED SERVICES

The load balancing services complement Redcentric's extensive portfolio of connectivity, co-location, hosting and other managed services. The services are optimised to deliver applications hosted in Redcentric data centres. The load balancing infrastructure is integrated into the data centre's high capacity Local Area Networks (LAN) and is particularly suitable for enhancing Customer environments built on Redcentric's Infrastructure as a Service (IaaS). Redcentric has extensive experience designing high availability application environments and consultants are available to assist Customers to meet their design objectives.

2.2) LOCAL LOAD BALANCING

2.2.1) Introduction

The Local Load Balancing Service (LLBS) distributes user requests for web site pages and other protected applications across multiple servers that essentially host the same content. Local load balancing is primarily used to manage user requests to heavily used applications, improving performance, minimising outages and generally ensuring that users can access these protected applications. Local load balancing provides fault tolerance by distributing user requests only to alternative servers when a server that hosts a protected application becomes unavailable.

Load balancing can be configured to:

- Distribute all requests for a specific protected web site, application, or resource between two or more identically-configured servers.
- Use any of several different algorithms to determine which server should receive each incoming user request basing the decision on different factors such as which server has the fewest current user connections or which server provides the fastest response.

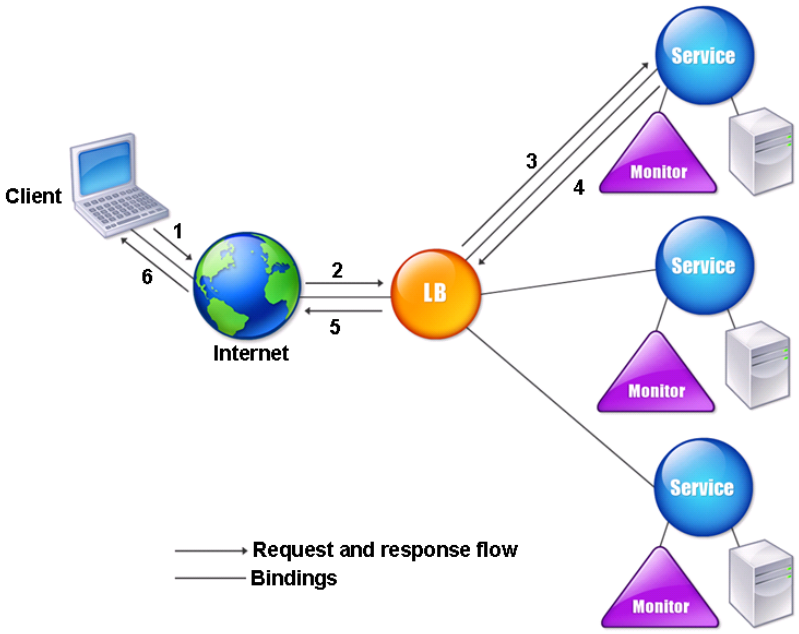
2.2.2) Local Load Balancing Architecture

A load balancing environment includes a load-balancing virtual server and multiple load-balanced application servers. The virtual server receives incoming client requests, uses the load balancing algorithm to select an application server, and forwards the requests to the selected application server.

The following conceptual drawing illustrates a typical local load balancing deployment.

'LB' represents the load-balancing virtual server

Figure 1. Local Load Balancing Architecture



2.2.3) Supported Load Balancing Algorithms

The load balancing virtual server can use one of many algorithms (or methods) to determine how to distribute load among the load-balanced local servers or services that it manages.

Different load balancing algorithms use different criteria. For example, the least connection algorithm selects the service with the fewest active connections, while the round robin algorithm maintains a running queue of active services and distributes each connection to the next service in the queue.

Some load balancing algorithms are best suited to handling traffic on websites, others to managing traffic to DNS servers, and others to handle complex web applications used in e-commerce or on company LANs. The following sub-sections list some of the more common supported load balancing algorithms with a brief description of how each operates. Other algorithms may be available - any such requirements should be identified prior to ordering the Service.

2.2.3.1) The Least Connection Method

When a virtual server is configured to use the least connection load balancing algorithm, it selects the service with the fewest active application sessions (connections). This is the default method because in most circumstances, it provides the best performance.

2.2.3.2) The Round Robin Method

When a load balancing virtual server is configured to use the round robin method, it continuously rotates a list of the services that are bound to it. When the virtual server receives a request, it assigns the connection to the first service in the list, and then moves that service to the bottom of the list. It is possible to assign a different weight to each service; the virtual server performs weighted round robin distribution of incoming connections. It does this by skipping the lower-weighted services at appropriate intervals.

2.2.3.3) The Least Response Time Method

When the load balancing virtual server is configured to use the least response time method, it selects the service with the fewest active connections and the lowest average response time. This method is suitable for HTTP and Secure Sockets Layer (SSL) services only. The response time (also called Time to First Byte, or TTFB) is the time interval between sending a request packet to a service and receiving the first response packet from the service.

2.2.4) Secure Socket Layer Offload

The LLBS can be ordered with Secure Socket Layer (SSL) acceleration. SSL transactions are handled by dedicated hardware in the load balancing appliances, removing load from the application servers. With SSL offloading, the appliance intercepts and processes SSL transactions, and sends the decrypted traffic to the server (unless end-to-end encryption is desirable, in which case the traffic is re-encrypted). Upon receiving the response from the server, the appliance completes the secure transaction with the client. From the client's perspective, the transaction seems to be directly with the server.

2.3) SITE LOAD BALANCING

2.3.1) Introduction

Global Server Load Balancing (GSLB) uses a range of technologies to steer application traffic to geographically distributed resources to achieve one or more of the following:

1. Disaster recovery: Providing an alternate location for hosting resources in the event that the primary location fails
2. Provide a means of minimising user impact when performing maintenance tasks by steering traffic to resources in an alternative location.
3. Load sharing: Distributing traffic between multiple locations to:
   - Minimize bandwidth and other resource costs
   - Accommodate capacity constraints at given locations
   - Limit exposure to various issues, including outages, geographic disruption etc.
4. Performance: Positioning content closer to users can enhance the user's experience
5. Legal Obligations: Occasionally it is necessary to present users with different versions of resources.

Redcentric's Site Load Balancing Service (SLBS) is optimised to steer user traffic to Customer environments deployed within Redcentric data centres and is therefore not a global service per se. Redcentric's SLBS uses precisely the same technologies as GSLB, namely Domain Name Service (DNS) based redirection.

2.3.2) DNS Redirection

The Redcentric infrastructure used to deliver SLBS forms part of the DNS resolution process for the service or sub-domain representing the Customer application(s). Various factors are used to determine what address the system will return to the end user's DNS resolver. The logic is most commonly based on one or more of the following:

1. The load and capacity of resources in the various environments.
2. The originating IP address of the DNS request.
3. Previous requests made from the same Internet Protocol (IP) host or subnet.
4. The health state of the various environments.

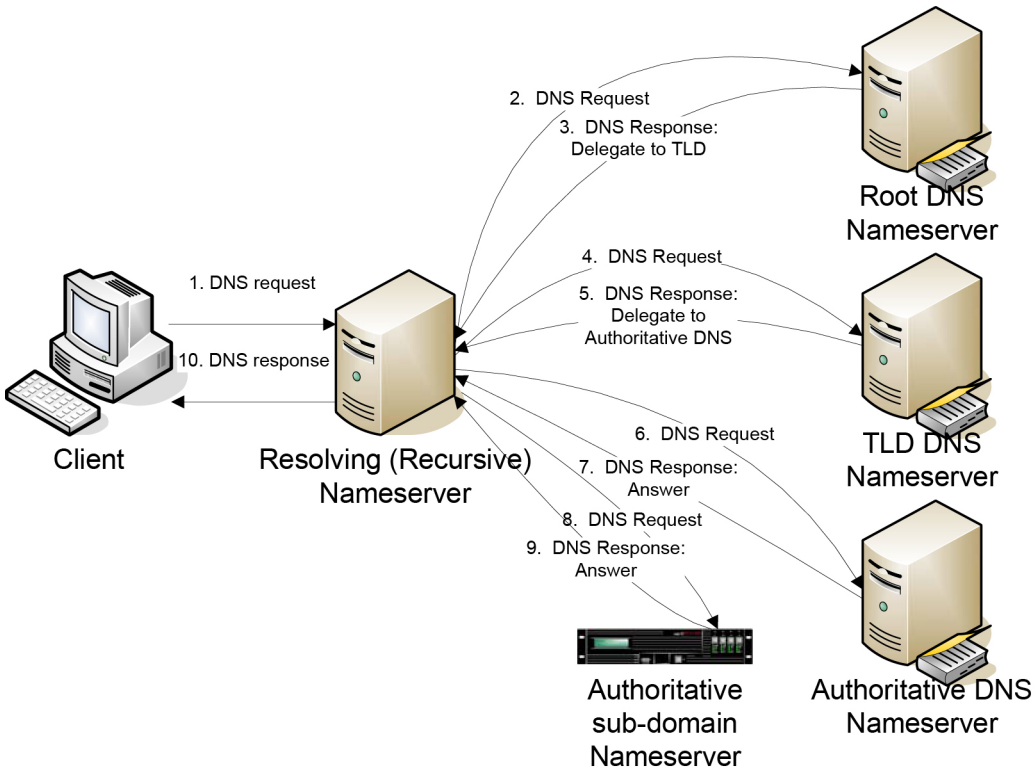
To ensure the various pieces of information are in place, the Redcentric infrastructure uses the following methods to determine the state for proper decision making:

1. Explicit monitors that check for availability of remote resources by accessing the resource itself.
2. Metric Exchange Protocol (MEP) is a communication channel between load balancing devices at different data centres that shares state information
3. SNMP based load monitors, which poll a remote resource for statistics such as CPU load, network load, and so on.

Multiple methods can be used to provide for more system redundancy. MEP and monitors can both determine the availability of a resource.

The conceptual drawing below illustrates a typical global or site load balancing environment.

Figure 2. DNS redirection Global and Site Load Balancing Architecture



2.4) DEPENDENCIES AND INTEROPERATION

When taken together, Redcentric's LLBS and SLBS offer the highest possible levels of functionality. Subject to a comprehensive design and assuming interoperability exists between platforms, Redcentric can also provide either LLBS or SLBS where the Customer or a 3rd-party is responsible for delivery of the other load balancing element. Any reduction in functionality where local and site load balancing is performed on separate hardware, managed by separate parties, will need to be established at the design stage.

2.5) RESOURCE ALLOCATION

The Redcentric platform consists of high performance load balancing hardware appliances supplied by one of the market leaders. The appliances have software that is capable of delivering multiple virtual, completely isolated load balancing instances - these are referred to as virtual servers in the previous sections. Each instance will have finite but dedicated resource allocated depending on the service(s) it delivers. Resources that can be allocated to load balancing instances are:

- Throughput bandwidth
- Memory (RAM)
- Inclusion of an SSL offload Application Specific Integrated Circuit (ASIC)

The specific resources allocated to Customer's instance(s) will form part of the Statement of Work. If it is determined at some point that the resources allocated are insufficient to meet demands of the Customer's environment because the load balancing function is not performing adequately well, the Customer will need to order additional resources.

2.6) SHARED RESOURCE INSTANCES

As well as providing instances which are solely dedicated to one Customer, Redcentric also offers shared instances. In this case, several Customers share the resource allocated to the instance. As resource cannot be allocated to individual Customers within the shared instance, this service is more suited to less demanding, smaller systems where load balancing functionality is required but budget does not extend to a dedicated resource service.

Redcentric monitors the use of resources in shared instances. If Redcentric determines that a Customer is consuming a disproportionate amount of any of the shared resources, Redcentric will give the Customer opportunity to upgrade to a dedicated resource service or may otherwise terminate the Statement of Work.

2.7) SERVICES OFFERED

The following service combinations are available to meet Customer needs.

Dedicated / Shared Instance	Data Centre(s)	Site Load Balancing	Local Load Balancing	SSL Off-load
Dedicated	Reading & Harrogate	Yes	Yes	Yes
Dedicated	Reading & Harrogate	Yes	Yes	No
Dedicated	Reading & Harrogate	Yes	No	N/A
Dedicated	Reading OR Harrogate	No	Yes	Yes
Dedicated	Reading OR Harrogate	No	Yes	No
Shared	Reading & Harrogate	Yes	Yes	Yes
Shared	Reading & Harrogate	Yes	Yes	No
Shared	Reading & Harrogate	Yes	No	N/A
Shared	Reading & Harrogate	No	Yes	Yes
Shared	Reading & Harrogate	No	Yes	No

2.8) IP ADDRESSING

IP addressing of servers and services etc. will be dependent on design. In most cases, Customers will require Internet facing applications to be protected by a firewall and the firewall will perform Network Address Translation (NAT). Consequently a mix of public and private IP address space will be implemented in such a design.

Services operating in shared load balanced instances will need to be addressed using public address space to avoid clashes with the IP addresses of other Customer's services.

2.9) SYSTEM ACCESS FOR CUSTOMERS

Currently no system access is available which allows Customers to view or make changes to their load balancing configuration. Configuration change requests should follow the standard Redcentric process which is detailed in the Customer Service Plan (CSP) document.

2.10) IMPLEMENTATION OF CHANGE REQUESTS

In accordance with the Redcentric change request procedure, all change requests must be submitted by a designated and authorised Customer technical contact. If the Redcentric' engineer cannot validate the change requester against the authorised list, then Redcentric will place the change request on hold and attempt to contact one of the alternative authorised contacts. Redcentric will wait for the request to be ratified by a known authorised contact before proceeding with any change. It is therefore essential that Customers provide accurate and current contact information for their designated and authorised staff.

2.11) REDCENTRIC RESPONSIBILITIES

In the context of providing load balancing services only, Redcentric is responsible for:

- Configuration of the load balancing platform to deliver the service
- General upkeep of the platform including hardware and software upgrades
- Provision of space and power etc.
- Ongoing support of Customer specific load balancing configuration including advice and implementation of minor changes

2.12) CUSTOMER RESPONSIBILITIES

In the context of providing load balancing services only, the Customer is responsible for all other aspects including but not limited to:

- Designing their overall solution
- Undertaking any major design changes required after initial deployment
- Identifying load balance resources required to meet the needs of their solution
- Providing all technical details to Redcentric required to configure the platform including details of services, servers, protocol type, IP addressing scheme, SSL certificate credentials etc.

2.13) MONITORING

Specific requirements for monitoring of servers, services, traffic levels, site and local load balancing activity etc. should be discussed at the design stage. Details of any included monitoring will form part of the Statement of Work. Customers should assume that in the absence of such details in a Statement of Work, no monitoring is provided.

2.14) REPORTING

Specific requirements for reporting should be discussed at the design stage. Details of any included reporting will form part of the Statement of Work. Customers should assume that in the absence of such details in a Statement of Work, no reporting is provided.

2.15) FAULT NOTIFICATION

Redcentric monitors the load balancing platform and notifies Customers of service-wide failure using automated Email/SMS methods as detailed in the CSP.

3) IMPLEMENTATION AND ACCEPTANCE

3.1) ACCEPTANCE CRITERIA

3.1.1) Local Load Balancing

- Incoming traffic is delivered to services running on Servers according to the chosen algorithm
- No traffic is delivered to a service which is unavailable
- When chosen and configured, SSL sessions are decrypted

3.1.2) Site Load Balancing

- DNS responses are returned as part of the resolution process according to the chosen distribution method (either active/back-up sites or active/active sites)
- DNS responses reflect changes to the normal operating environment. I.e. IP addresses of the back-up site are returned when the primary site fails.

4) SERVICE LEVELS AND SERVICE CREDITS

4.1) SERVICE LEVELS

The Service Level applicable to Load Balancing Services is as follows:

Service Level: Availability Measurement Period: Month	
Service Level	Not less than 99.99%

4.2) FLOOR SERVICE LEVEL

The Floor Service Level applicable to Load Balancing Services in respect of Availability shall be 85% in any given Month.

4.3) SERVICE CREDITS

The Service Credits applicable to Load Balancing Services shall be calculated as follows:

$$Service\ Credit = \frac{C\ x\ S}{MS}$$

Where:

- S = the number of seconds by which Redcentric fails to meet the Service Level for Availability in the relevant Month
- C = total Charges payable in respect of Load Balancing Services for the same Month
- MS = the total number of seconds in the same month

5) DATA PROCESSING

5.1) DATA PROCESSING SCOPE

- The Server Load-balancing Services delivers the routing and distribution of IP packets to end devices, typically servers.
- The Server Load-balancing Services does not involve any storage or backing up of data.

5.2) DATA STORAGE AND ENCRYPTION

- The option exists to encrypt/decrypt traffic within the Load-balancing platform. This is commonly referred to as SSL off-load.
- Redcentric does not capture, inspect, analyse, store or share the Customer's traffic/data under normal circumstances.
- Under certain circumstances, when managing a support ticket, Redcentric may capture, inspect, analyse and/or store a small sample of the Customer's traffic in order to investigate and diagnose a very specific problem, e.g. to help resolve a problem relating to IP packet misdirection. Such diagnosis would involve the examination of a small sample of IP packets, but such actions will only be undertaken at the request of and in conjunction with the Customer.

5.3) DATA PROCESSING DECISIONS

- Redcentric does not make any data processing decisions in relation to the Server Load-balancing Services. Any processing of data over Customer systems when using the Server Load-balancing Services is instigated, configured and managed by the Customer, including any decision to use encryption/decryption, and where to implement that encryption/decryption.
- Redcentric Support can be asked by the Customer to intervene in the event of an issue with the Server Load-balancing Services. In such a case Redcentric may make decisions that affect data processing, but such actions will only be undertaken at the request of and in conjunction with the Customer.

5.4) SUB-PROCESSORS

- No other parties are involved in delivering the Server Load-balancing Services, and there are no sub-processors appointed by Redcentric.

5.5) CUSTOMER ACCESS TO DATA

- The Customer controls its own platforms which use the Server Load-balancing Services to carry data, and the Customer therefore has full access to its own data.

5.6) SECURITY ARRANGEMENTS AND OPTIONS

- The Infrastructure delivering the Server Load-balancing Services is hosted at both Redcentric and third party locations. All locations meet physical security standard ISO27002 section 11.1 or equivalent.

HARROGATE (HEAD OFFICE)

Central House
Beckwith Knowle
Harrogate
HG3 1UG

THEALE

2 Commerce Park
Brunel Road
Theale
Reading
RG7 4AB

CAMBRIDGE

Newton House
Cambridge Business Park
Cowley Road
Cambridge
CB4 0WZ

READING

3-5 Worton Drive
Reading
RG2 0TG

LONDON

Lifeline House
80 Clifton Street
London
EC2A 4HB

HYDE

Unit B
SK14 Industrial Park
Broadway
Hyde
SK14 4QF

INDIA

606-611, 6th Floor
Manjeera Trinity Corporate
JNTU – Hitech City Road
Kukatpally, Hyderabad – 72

0800 983 2522
sayhello@redcentricplc.com
www.redcentricplc.com

redcentric
business technology. managed.

bsi.
ISO 9001 Quality Management FS 603185
ISO 22301 Business Continuity Management BCMS 603194
ISO/IEC 27001 Information Security Management IS 603187
ISO/IEC 20000-1 Information Technology Service Management ITMS 668453