



White Paper

Introduction to Global Load Balancing & Cloud Load Balancing

This White Paper aims to provide an introductory overview of the Global and Cloud Load Balancing Services provided by Nebula 6.

About Nebula 6

Nebula 6 is a provider of cutting edge Application Acceleration and Optimisation technology direct from the Cloud. Our aim is to take away the technical complexities, Cap Ex costs and management headaches of deploying state-of-the-art acceleration technologies in-house, and instead leverage our economies of scale and years of web performance engineering experience to deploy services from our own state-of-the-art Cloud infrastructure.

Introduction

For organisations providing vital applications and services over the Internet, the cost of downtime is extremely high, whether that's measured in terms of lost employee productivity, degraded customer experience or a number of other ways in which downtime can adversely affect a business.

“Enterprises should incorporate application acceleration as an integral requirement for any cloud-computing service”

Bjarne Munch, Gartner, 'Optimizing Applications From Cloud Services: The Old Issues, but More Challenging to Solve'



Our Services

- Application Acceleration
- Global Load Balancing
- Cloud Load Balancing
- Global Content Delivery
- Website Resilience & Acceleration
- Bespoke Requirements

Our Platform

- World Class Cloud Infrastructure
- Global PoPs
- Dedicated Architectural Focus on Application Acceleration & Resilience

“

Gartner believes that, through 2013, at least 60% of enterprises will experience slow or inconsistent application performance issues from externally placed applications.

”

Bjarne Munch, Gartner.
Is Your Network Design the Weak Link in Cloud Computing?

Currently there are two widely utilised techniques to minimise the chance of a failure causing downtime in network-based applications; Load Balancing and Global Load Balancing.

Server Load Balancing Within A Single Datacentre

Techniques like server load balancing and clustering are commonly used within a single datacentre to build cluster of fault-tolerant, scalable applications and / or web servers. These clusters are resilient to isolated failures – for example, a server machine developing a hardware fault – and allow the administrator to add more capacity to his application when required.

However, a clustered, fault-tolerant application running in a single datacentre 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 datacentre may become unavailable because of a denial-of-service attack mounted against a different service running in that datacentre, or because of a failure in its local internet connectivity.
- The datacentre may be disrupted due to a catastrophic natural or man-made disaster-power failure because of rolling blackouts, maintenance errors or even terrorist attack.

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

Global Load Balancing between Multiple Datacentres

Global Load Balancing manages how end users are connected to an application or website when that service is hosted in multiple disparate datacentres. As such, there are two main benefits to implementing Global Load Balancing;

- Business Continuity – to ensure that services are always available, even when one or more datacentres becomes unavailable.
- Improve Customer Experience – to load-balance each user to the best datacentre from a choice of several. The choice can be based on datacentre performance and proximity, so that clients are directed to the closest datacentre and / or the best performing datacentre

Global Load Balancing can help achieve these goals through two key types of configuration;

- In an Active-Passive configuration, one datacentre is nominated as Active. The other datacentres are idle for that service. If the active datacentre becomes unavailable, one of the passive datacentres becomes active and all users are directed to it.

Key Features

- Enhance End User Experience & Adoption
- Reduce Infrastructure & Bandwidth Requirements
- Enable Rapid Scaling and Growth
- Achieve Inter-Site Resilience
- Use Geo-Location to Reduce Latency
- Work With Us To Build Customised Solutions To Meet Complex Requirements

“

As enterprises move more applications into external data centers, it has become more difficult to manage application traffic flow and, subsequently, application performance for users placed in branch offices and remote users.

Bjarne Munch, Gartner.
Is Your Network Design the Weak Link in Cloud Computing?

”



- In an Active-Active configuration, all datacentres are used and clients are load-balanced between them based on datacentre performance and proximity.

Potential Active-Active Global Load Balancing Configurations

If datacentres are running in active-active mode, Nebula 6 can Globally Load Balance between them based on a number of different criteria.

Datacentre Availability: If a datacentre has failed, users are directed elsewhere.

Datacentre Performance: Datacentres with better response times are preferred over slower, more overloaded datacentres.

Geographic Proximity: Using a comprehensive database to map IP addresses to geographic distance between the end user and each datacentre, the user is directed to the datacentre that they are physically closest too.

The decision of where to send a user can be based purely on load, purely on geographic location, or on a mixture of the two. The benefits of an active-active load balancing mode are:

- Better Datacentre Utilisation.
- Users Get The Best Possible Level Of Service From The Closest, Best Performing Datacentre.
- The Configuration Still Provides Full Failover In The Event Of A Datacentre Failure.

However, an active-active configuration will not always be appropriate. For example, if the application being balanced cannot be run in multiple datacentres simultaneously – because, say, it depends on a single database or SAN that cannot be continuously replicated over multiple sites – then an active-passive configuration will be more appropriate.

Additionally, one side-effect of an active-active load balancing mode is that an end user may spontaneously be redirected from one datacentre to another when his client software makes a fresh DNS request. For example, the datacentre he is accessing may become overloaded and the load-balancing algorithm may assign him to a different datacentre.

However, if such behaviour is undesirable, it can be overcome by several methods, such as using fully deterministic 'Geo' load-balancing, or using Application-level redirection to detect a user's session and forcibly direct to a particular datacentre when required.

"Efficiency and cost reductions can only be achieved through integration and strategic deployment of an enterprise-wide Application Delivery Network"

Forrester Consulting Thought Leadership Paper, 'Integrating Application Acceleration Into The Network Fabric Is The Future'

Key Features

- Serve Users Around the World Without Deploying Local Infrastructure
- A Fraction of Traditional CDN Costs
- Multiple Interconnected Datacentres
- Worldwide Points-of-Presence
- Unique Cloud Architecture Optimised for Web Performance



Moving business applications into external cloud services will increase the risk of poor application performance and security vulnerabilities due to network connectivity issues. But network services are evolving into cloud network services that can mitigate these risks and provide sufficient on-demand flexibility to support dynamic cloud computing services.



Bjarne Munch, Gartner.
Cloud Network Services Are Essential Enablers of Cloud Computing

Active-Passive Load Balancing Configurations

When the datacentres are running in active-passive mode, the load balancing decision is much simpler. An order in which the datacentres should be used is specified (ie. Primary, Secondary, Tertiary etc) and all users are directed to the primary datacentre so long as it's available.

If the first datacentre fails, all users are directed to the secondary datacentre and so on. It's also possible to configure how the service should fail back when the primary datacentre comes online again. If automatic failback is enabled, users will immediately be directed to the first datacentre again. If it is disabled, users continue to use the second datacentre until a manual decision is taken to revert traffic back to the primary datacentre.

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

Availability and Performance Checking

Nebula 6 Global Load Balancing checks the performance and correct operation of the services in the local datacentre 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.

In addition, performance data can be deducted from the response times from selected monitors, and then used to weight how much each datacentre is used when the Load or Adaptive load balancing algorithm is selected.

Who can benefit from Global Load Balancing?

Global Load Balancing can benefit any organisation;

- Who provides or depends on an internet-based service, such as a public-facing website, or a network-based application for internal use.
- Who cannot countenance service failure, whether this results in lost productivity, lost revenue or lost customers.
- Who wishes to establish an advantageous SLA (Service Level Agreement) with its users or customers, providing them with a superior and competitive level of service.

How Does Nebula 6 Global Load Balancing work?

DNS-based Global Server Load Balancing

Nebula 6 Global Load Balancing functions by manipulating the DNS (Domain Name System) 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.nebula6.com.

Global Load Balancing Methods

- Datacentre Load Balancing - based on observed performance
- Geographic Proximity - routes each user to the closest datacentre
- Adaptive Load Balancing - routes users based on proximity and datacentre load

“

Poor application performance can generally be solved via network redesign or via dedicated application acceleration needed to overcome poor application design.....Changes in application traffic flow between end users and data centers can increase network latency, which can cause application performance issues. Network redesign can solve this but, in many cases, this is not sufficient and there is thus a need to deploy application acceleration.

”

Bjarne Munch, Gartner.
Optimizing Applications From Cloud Services: The Old Issues, but More Challenging to Solve



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.

Different servers in different locations will have different IP addresses. Global Load Balancing controls how domain names are resolved to IP addresses, and thus controls which datacentre clients are directed to.

Conclusion

Nebula 6 are able to provide a comprehensive DNS-based Global Load Balancing solution that provides:

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

Because Nebula 6 deliver Global Load Balancing as a service, the complexity and Capital Expense of deploying a similar solution in house, with a device at each datacentre, is removed, meaning it can be implemented with minimal infrastructure changes and very little operational or financial risk.

Cloud Load Balancing

Cloud Load Balancing can be seen as a natural evolution from Global Load Balancing, just as Global Load Balancing has itself evolved from load balancing multiple servers in a single datacentre.

Cloud Load Balancing can typically be performed across traditional in-house datacentre deployments and/or public, private or hybrid clouds alike.

By combining elements of local server Load Balancing and Global Load Balancing, Cloud Load Balancing enables organisations to distribute user requests across any number of application deployments located in multiple datacentres and / or through different cloud-computing providers.

By extending the existing architectural models for local and global load balancing into the Cloud, Cloud Load Balancing increases the number of intelligence-based options for an organisation to determine where a given application should be delivered from.

As internet delivered services continue to evolve, the amount of factors that potentially need to be considered when delivering web services to end users is increasing at a rapid rate; the ability to meet specific SLAs, the user's device or location, the cost of running the application, compliance with regional regulations and many more.

The ultimate goal of Cloud Load Balancing remains to deliver an application (or website) to a user as quickly as possible with the least amount of resources and for the lowest cost. For organisations focused on this goal, factors that need to be considered include;

- Cost To Execute The Request At A Given Location
- Total Cost To Deliver The Request To A User/Customer.
- Regulatory Compliance And / Or Legal Restrictions.
- Delivery / Response Time and SLA Requirements

When deployed from the Cloud, determining the costs of delivering an application means taking account both the costs incurred by the application instance, as well as the cost of bandwidth used by the request and response. Furthermore, because the costs might be dependent on the total resources used by the application during a specific period of time (such as on a monthly or weekly basis, or even on a time of day basis), these calculations can become extremely complex.



Compliance with regulations and contractual obligations, including SLAs, are often even more complex, as anyone who's dealt with them can testify. One of the ways in which Cloud Load Balancing can assist is that it can be leveraged to help achieve compliance by minimizing the investment necessary to deploy and implement specific geographically-determined services.

For example, an organisation may initially offer customers SLAs or services at a premium that include additional application delivery options, and then subsequently choose to offer these options from a cloud environment to minimise the delivery costs.

In order to adequately meet such contractual obligations after having changed deployment methods, the application delivery infrastructure must be able to identify users using the context of the request data such as IP address, pre-existing cookies, and credentials for which the obligation must be met; secondly, it must be able to correctly determine from which environment the obligation can best be met.

Cloud Load Balancing can be utilised to help frame and enable a global application delivery solution that's able to determine, on a per user/customer basis, the best location from which to deliver an application.

This decision-making process includes traditional Global Load Balancing factors such as:

- Application Response Time.
- Location Of The User.
- Availability Of The Application At A Given Implementation Location.
- Time Of Day.
- Current And Total Capacity Of The Datacentre / Computing Environment In Which An Application Is Deployed.

By taking in all these factors Cloud Load Balancing helps balance business goals, such as cost reduction, with technical goals, such as response time and availability metrics.

Conclusion

Cloud Load Balancing from Nebula 6 facilitates this decision making process with an ease unparalleled by any other service offering, or indeed by traditional local hardware / virtual appliance based solutions. Able to coordinate across application deployments in multiple datacentres, whether in the cloud or traditional DC deployments, and able to perform the tasks of both local server Load Balancing and Global Load Balancing, Nebula 6's Cloud Load Balancing is capable of remarkable reach. When combined with leading edge application acceleration, Cloud Load Balancing can deliver performance and ensure availability, whilst allowing for complex configurations and intelligent, automated decision-making based on multiple factors.

To learn more, contact us on by phone on +44 (0) 870 382 5050 or via email to info@nebula6.com. Alternately, please visit our website at www.nebula6.com

Key Features

- Enhance End User Experience & Adoption
- Reduce Infrastructure & Bandwidth Requirements
- Enable Rapid Scaling and Growth
- Use Geo-Location to Reduce Latency
- Work With Us To Build Customised Solutions
- Serve Users Around the World Without Deploying Local Infrastructure
- A Fraction of Traditional CDN Costs
- Multiple Interconnected Datacentres
- Worldwide Points-of-Presence
- Unique Cloud Architecture Optimised for Web Performance