Four routes to multi-site resilience

Maximizing uptime for your critical systems

loadbalancer.org

When downtime isn't an option, you need to use every tool at your disposal.

"You've probably heard, and possibly even used terms like multi-site, multi-cloud, and multi-data center. The concept of having infrastructure, data and services shared across various locations may not be new, but has become much more prevalent in recent years. Back when I was starting out in IT, people took these things out of their own server rooms, and moved them into data centers. Well-supported, well-connected data centers – proper hosted platforms. And they did this to optimize availability.

But things move on, and recent, less predictable times, have invited new trends, like moving up to specific clouds, such as Amazon. But at the moment, it seems to me that organizations are favoring the whole hybrid approach – multi-site, multi-provider and multi-cloud. Why? Well, they don't want to put all their eggs into one basket. You don't have to be an expert in downtime to know that even the larger cloud and SaaS platforms still have outages – just take a look at downdetector.com once in a while!

Unfortunately, these outages still affect business customers in all sorts of industries, as well as their end users. With the growth of multi-cloud and multi-data center environments, we're seeing that people no longer want to have their critical systems in just one place, they want the freedom of doing those things everywhere – with the consistency, resilience and reassurance that provides for business continuity. And while disaster recovery often brings to mind extreme scenarios like earthquakes or fires, it could apply to something as simple and unexplainable as catastrophic data loss. To this end, covering yourself with some form of multi-site resilience into your disaster recovery strategy is something I can't recommend highly enough."

Aaron West, Technical Architect, Loadbalancer.org

THE NEED FOR MULTIPLE SITES

An additional dimension in your route to resilience

In every sector, from retail to healthcare, a significant number of medium-sized organizations – and the vast majority of enterprises – operate across more than one location or site. And despite the current acceleration towards remote, work-anywhere flexibility, whether offering critical services or simply driving business growth, the chances are that many single-site organizations are also considering mitigating disasters with the wide range of off-premise products, services and systems offered by solutions providers.

So what does a typical multi-site organization look like? Although there are similarities in certain sectors – a national 'bricks and mortar' retail chain for example – the structure of organizational sites and services can be as diverse as the organization itself. Even companies that operate chiefly from one location and don't strictly consider themselves 'multi-site' probably still access some off-site services, whether through the cloud, or via a third-party service provider.

The makeup of a multi-site organization could comprise:

- multiple customer-facing locations with separate distribution and storage facilities
- five operationally-identical sites serving communities in different locales
- a global network of regional HQs, operational sites and support functions
- two office buildings, right next to each other

MULTI-SITE BENEFITS - AND CHALLENGES

Operating and accessing services between multiple physical locations can offer plenty of logistical benefits. It can often make more sense for user experience, or perhaps the cost impact on your IT strategy to deploy, serve and maintain applications from a particular location, especially if demand is based at, or near to that location – such as serving virtual desktops to users at a specific office location. But with the necessity of a multi-site approach benefitting organizations of all types, so too it can present challenges.

However, if your services are located in multiple locations, you may also:

- struggle to scale out over more than one office location, data center or cloud
- have difficulty implementing simple solutions to common problems
- be considering complex overlay networks, or worse, something like Anycast.

IN THIS EBOOK, WE'LL EXPLORE:

- techniques to deliver high availability across multiple locations
- active-backup configurations for high availability in any environment
- using Global Server Load Balancing (GSLB) to offer alternatives that don't require common subnets across locations or Border Gateway Protocol (BGP) to achieve scalability.

1. LAYER 2 MULTI-SITE

Our first route, ideal for a campus or dual sites

Most organizations are reliant on their data in some form, whether email, voice or any other critical application for the business – and most of them are looking at providing resilient solutions for that. This could be driven from the perspective of disaster recovery (DR), business continuity, ensuring appropriate capacity, or even just to keep things running day to day. So, traditionally, with a single Data Center, you might have a pair of load balancers, which will load balance across multiple servers, potentially for a wide variety of services. And that's all well and good – if the server fails, the load balancer recognises a failed health check, and then redirects traffic to available nodes.

It has become extremely common for even small and medium enterprises to split their data between two locations.

However, with a single site, there are other risks at play – the chance of a power outage, a specific site issue (for example, flooding), or some other unforseen problem that ultimately results in that site becoming unavailable. For this reason, it's become extremely common for even small and medium enterprises to split their data between two locations. The easiest, probably most common way to achieve that, is to link a campus or pair of buildings with some kind of layer 2 connection. Layer 2 connectivity means that the virtual local area networks (VLANs) and all of the setup that was in the single site, can be replicated across both sites. It requires very little change within overall infrastructure or configuration to manage that.

In these types of scenarios, instead of having a pair of load balancers on each site, it's common to have an active load balancer on one site, and a passive node on the other – stretching the pair that was in the single data center, and putting one in each of the data centers. With this approach, only one load balancer is active at a time, so traffic will have to traverse the links depending on where users are. For example:

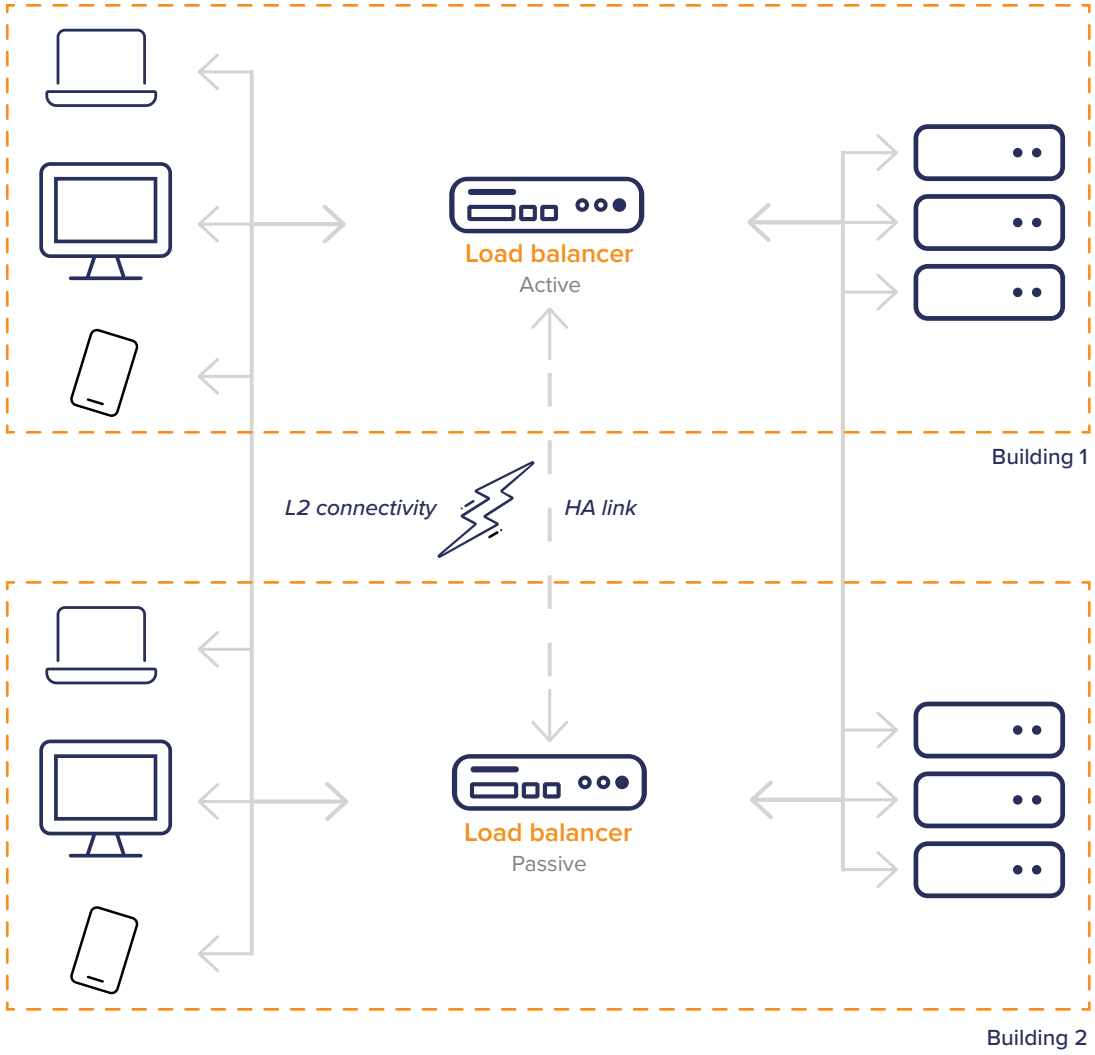
- users in *building two* may have to traverse a link to reach the active load balancer in *building one*
- then, the load balancer will direct them to a service – which could be in data center one *or* data center two, depending on where it's replicated.

This is a fairly simple way of increasing the resilience of a network, which provides some level of business continuity or disaster recovery. And in the event of a load balancing failure, the other one will still take over automatically in the same way as a pair on a single site. Also, normally there's a reliance for the link between the sites to be highly available (HA). Not only HA between the load balancers, but also consider disparately routed, dark fibre, or an SD-WAN-type overlay, to protect your local networks between the two locations.

LAYER 2 MULTI-SITE EXAMPLE

Suitable for:

- *campus-based data centers*
- *overlay (extended L2) data centers*
- *lower cost*
- *simple active/passive cluster*



2. LAYER 3 MULTI-SITE

Route two: optimized for your local users

Another common multi-site deployment method utilizes layer 3. Often, this applies when there's a requirement to have, say, one on-site data center, and one hosted data center – through a standard data center hosting provider, for example. Or, there could be multiple layer 3 locations geographically dispersed around the world, to cater for local users via a local data center, for a higher level of performance.

Utilizing layer 3 applies when there's a requirement for an on-site data center, and a hosted data center

In order to then add the level of resilience that you need, you can add fallback services. Users within a region or data center, on a private network, will usually have a DNS which points them at their local service – either a split horizon DNS scenario or just regional DNS. So, in most cases:

- users will access their local data center (or local building services)
- the load balancers carry out normal health checks on the local service only, before providing them to users
- in the event of a failure, the traffic would be routed through the private wide area network (WAN) and sent to the remote site – a fallback position.

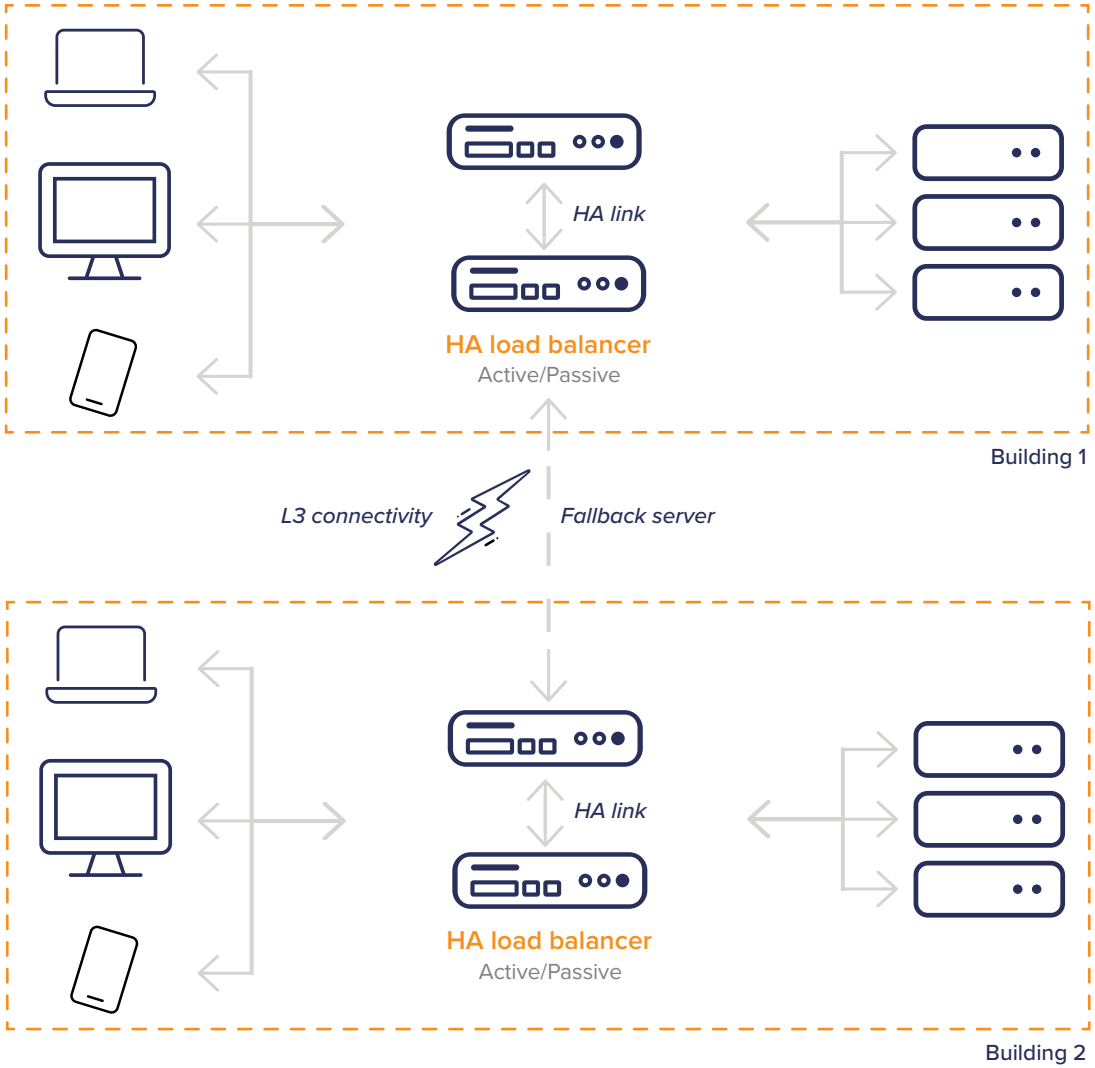
In the scenario we've described, the entire service has to be down on a single site before the load balancer would reroute traffic to one of the remote sites.

While this certainly helps with service availability, it may not help with performance, depending where the other service is geographically. None of this is reliant on any global server load balancing (Global SLB) – this is very much a private network setup, so does come with limitations, ie that your private network needs to be up and running. The benefits of this approach are that you're going to achieve higher performance from the local availability of your service – it's only in those failure scenarios, that you need to fallback on, and access, the remote site.

LAYER 3 MULTI-SITE EXAMPLE

Suitable for:

- *corporate geographic private data centers*
- *hosted private data centers*
- *high availability (HA) within, and between locations*
- *multiple active/passive clusters*
- *limited inter-data center failover*



The rise of Global Server Load Balancing (Global SLB)

Global Server Load Balancing, or Global SLB, is an upgraded version of the traditional round-robin DNS load balancing technique. It's used to distribute internet and corporate network traffic across servers located in different locations, anywhere in the world.

Just as a load balancer distributes traffic between connected servers in a single data center, Global SLB distributes traffic between connected servers in multiple locations, whether these servers are in an organization's own data centers or hosted in the public or private cloud - and able to provide failover between these sites based on the availability and performance profile of the data centers. Should one server in any location fail, Global SLB reroutes traffic to another available server somewhere else in the world.

Global SLB distributes traffic between connected servers in multiple locations, and is able to failover between these sites

By enabling all user traffic to be switched instantly and seamlessly to an alternative data center in the event of an unexpected outage, Global SLB improves the resilience and availability of key applications, and enables organizations to constantly monitor application performance at geographically separate locations. It also ensures the best possible application availability across multiple sites.

During routine maintenance, Global SLB enables organizations to temporarily direct user traffic to an alternative site in order to avoid disruptive downtime.

Note: please don't confuse Global SLB with geographic load balancing, which redistributes traffic dependent on internet-based geo-location.

KEY BENEFITS OF GLOBAL SLB:

SCALING OUT WITH WEIGHTED ROUND-ROBIN

- Scale out across multiple locations and data centers
- Advanced health checks supporting custom scripts and dynamic feedback of real server status
- Adjustable weights allow skewing of traffic loads

TOPOLOGY FOR USER LOCATION AFFINITY OPTIMIZES TRAFFIC

- Topology support, per site location affinity with failover
- Ability to health check a different IP from the member
- Ability to set a default location
- Considers the source of the requesting DNS server

FAILOVER GROUPS PROVIDE STAGED FAILOVER FOR SIMPLE HIGH AVAILABILITY

- Active-backup configuration for high availability
- Supports multiple backups
- Can 'reject' in the event of all members being down

DYNAMIC WEIGHTS AND CUSTOM HEALTH CHECKS

- Custom health checks allow confirming actual application availability and status
- Dynamic weights allow skewing of traffic loads based on server feedback

3. GLOBAL SERVER LOAD BALANCING - MULTIPLE LOCATIONS

Go global with our third route to resilience

Global server load balancing can be introduced, either on its own, or alongside local server load balancing. When designing your deployment scenario, it can offer more flexibility than traditional load balancing alone. Methods such as layer 4 or layer 7 load balancing sit in line and see all traffic from client to server.

This gives a real time view of the application, and its availability – which can be extremely useful. However, sometimes this means the load balancer becomes the bottleneck itself. Methods such as layer 4 DR, or Direct Server Return, can be used

Global server load balancing can can offer more flexibility than traditional load balancing alone

to address some scalability problems, but they only really help improve egress (outbound) traffic – as ingress (inbound) traffic will still flow via the load balancer. This helps make the case for Global SLB with direct-to-node balancing, because its key for some technologies to achieve maximum throughput (for example, storage) – making Global SLB a perfect fit.

OTHER PRIMARY USE

The other way Global SLB is most widely used is to support multiple locations or sites. You may even need an extra layer of load balancing in front of your East Coast and West Coast DCs – favoring one over the other For DR purposes – or just load balancing between the two of them. Standard distribution techniques are supported, such as round robin, failover groups, and also topologies. Topology mapping can be introduced on top of normal round robin load balancing to offer location affinity, to direct users from specific client subnets to a preferred location – shortening the length of the trip to the application and reducing the need to use costly cross-site links.

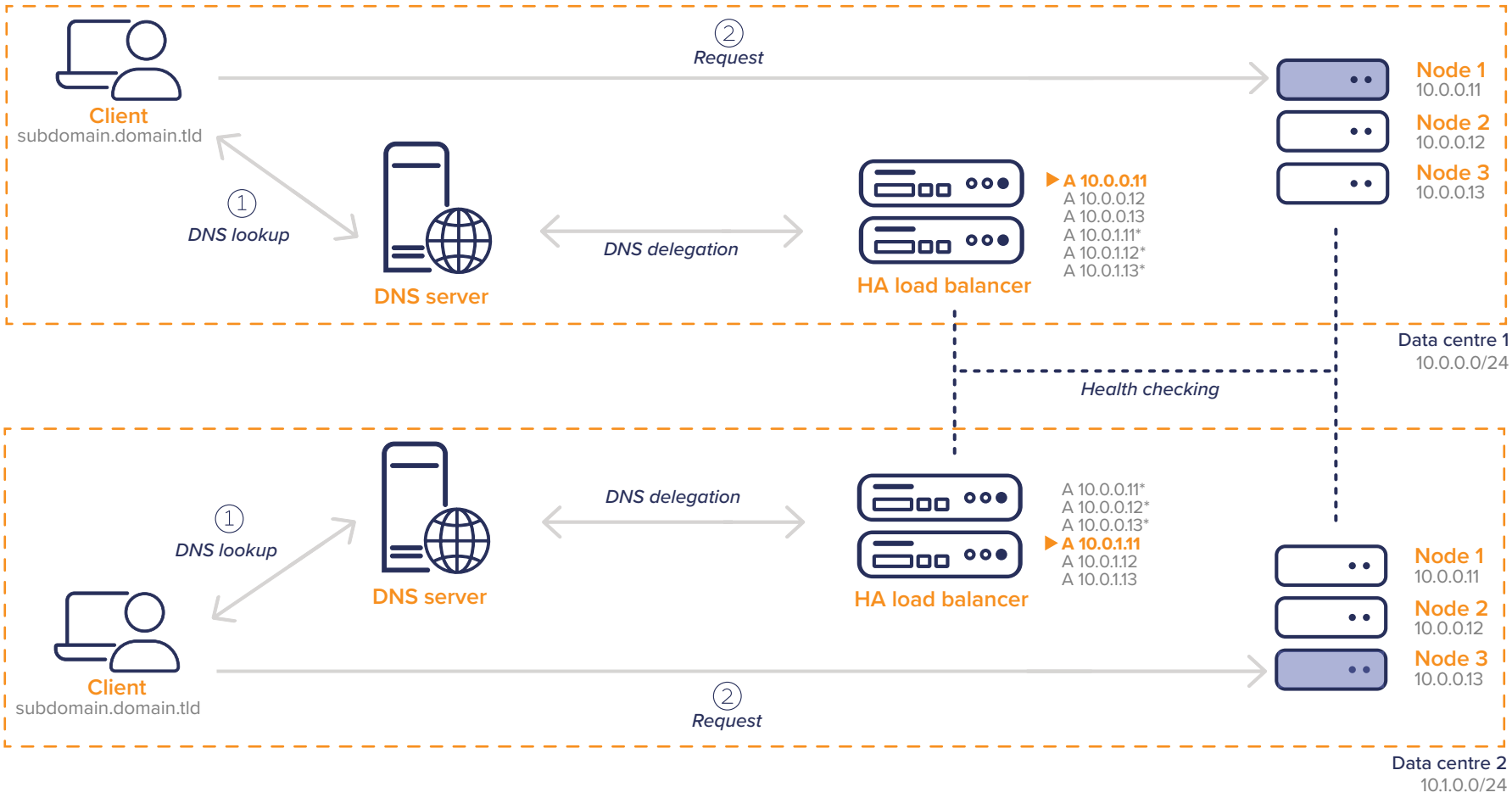
SO IS IT JUST FANCY DNS?

Well, yes *and* no. Typically, what makes a Global SLB is the site awareness and health awareness – knowledge about possible locations and health checks, then the ability to make decisions based on those health checks. Round Robin DNS is a mature, simple solution, but it falls short – without proper health checks in place, it may serve a failed result to customers.

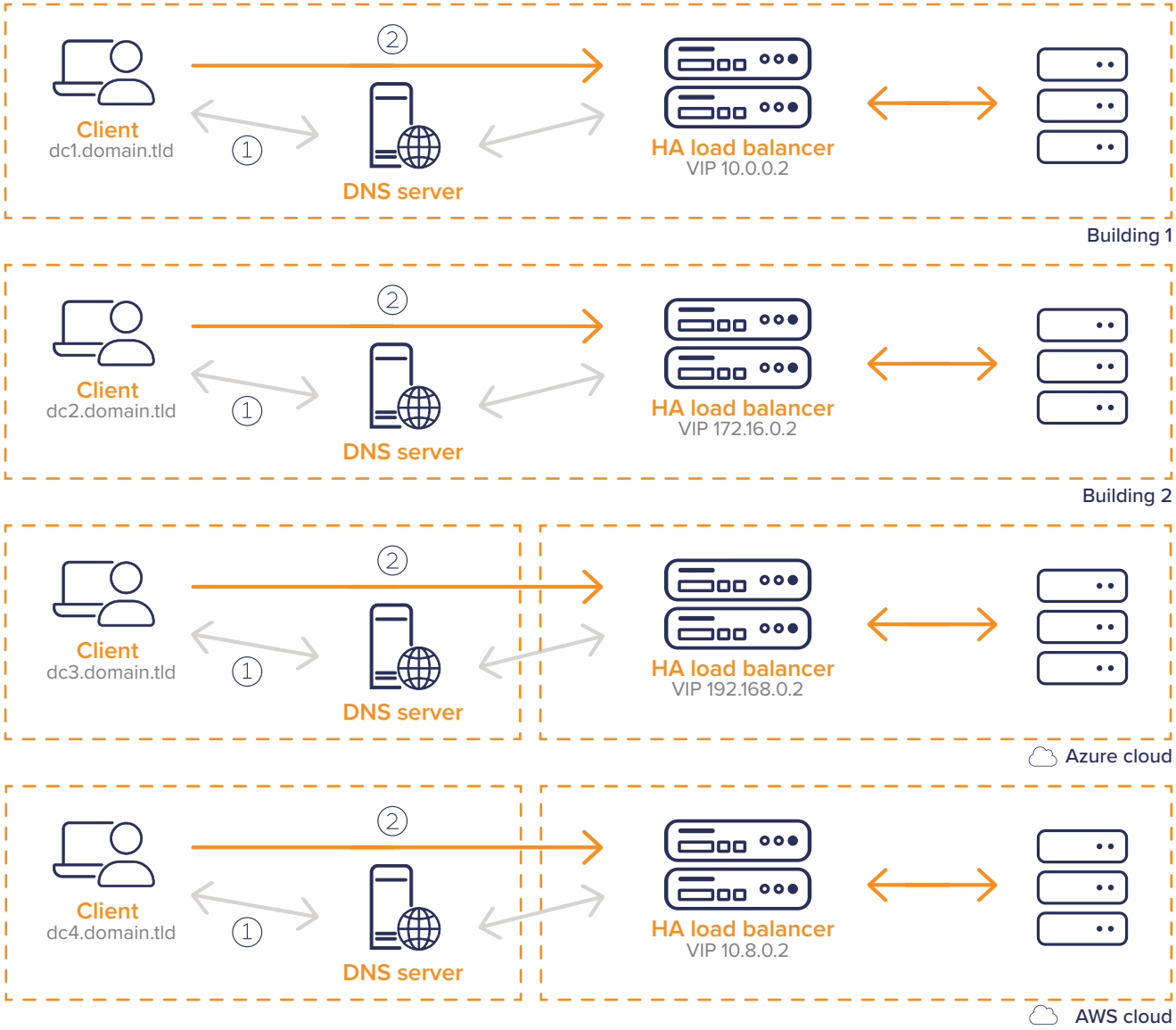
It also lacks awareness of where clients are coming from, and where we might want to send them. Typically it'll just return multiple addresses (hopefully in a random order) which can lead to a bad customer experience when something goes wrong.

Over the next couple of pages we've illustrated two ways to use Global SLB in your multi-site approach, followed by further detail for each on page 12.

GLOBAL SLB – TWO LOCATIONS EXAMPLE



GLOBAL SLB – MULTIPLE LOCATIONS EXAMPLE



GLOBAL SLB – EXAMPLES EXPLAINED

GLOBAL SLB – TWO LOCATIONS

On page 10, we've illustrated an example of two sites. Rather than include extra load balancers at each site, we've shown a direct-to-node approach for simplicity. The client will perform a DNS lookup to their local DNS server. Next, because we've delegated the subdomain that's used by our application, the DNS server knows to forward the request to the load balancer to retrieve an answer (shown as *step one*). Answers will be returned to the DNS server, followed by the client – enabling the DNS server to remember the result, which will typically adhere to the time to live (TTL) that we have configured.

Once the client receives an answer, it can connect to the node returned (shown as *step two*). In this example we're using topology weighted round-robin (TWRR), or our topology algorithm, so clients stay within their local site – unless all nodes fail in that location. The Global SLB will run health checks on both sides, for full knowledge of server availability to direct users to an available, online node.

GLOBAL SLB – MULTIPLE LOCATIONS

Steps one and two detailed above also apply in this example. Site expansion is fully supported – you can deploy this solution across several data centers, sites or even cloud environments, returning up to a maximum of 1,024 results. The solution is hugely scalable, so it can grow alongside your network requirements. Users in a head office can be accessing their on-premise provisions, while any external users can be sent to any of the healthy available sites. In the diagram, we've shown two office locations with local users, and two popular cloud environments, with external users coming from branch office locations – illustrating how diverse the configuration can be. To achieve this we need to delegate our DNS subdomain. Delegating, or configuring what's known as *conditional forwarding,* is straightforward – it's supported with all DNS servers, making this universally easy to configure. While Global SLB only supports A or any records, CNAMES can be configured locally in your DNS infrastructure, and then pointed to the A record that's been delegated to your Global SLB. Generally, the steps to achieve this are very simple, and involve configuring a record for your load balancers, and then delegating (or using conditional forwarding) to send traffic destined for the subdomain, off to the load balancers to get a response.

MORE ON SITE AFFINITY

Site affinity is provided by topology-based load balancing, which distributes the traffic matching your local resources based on the topology, and therefore your location that you have configured within your Global SLB. This offers location affinity, helping users stay within their local office or site, and avoids the use of expensive or contended links, when accessing resources.

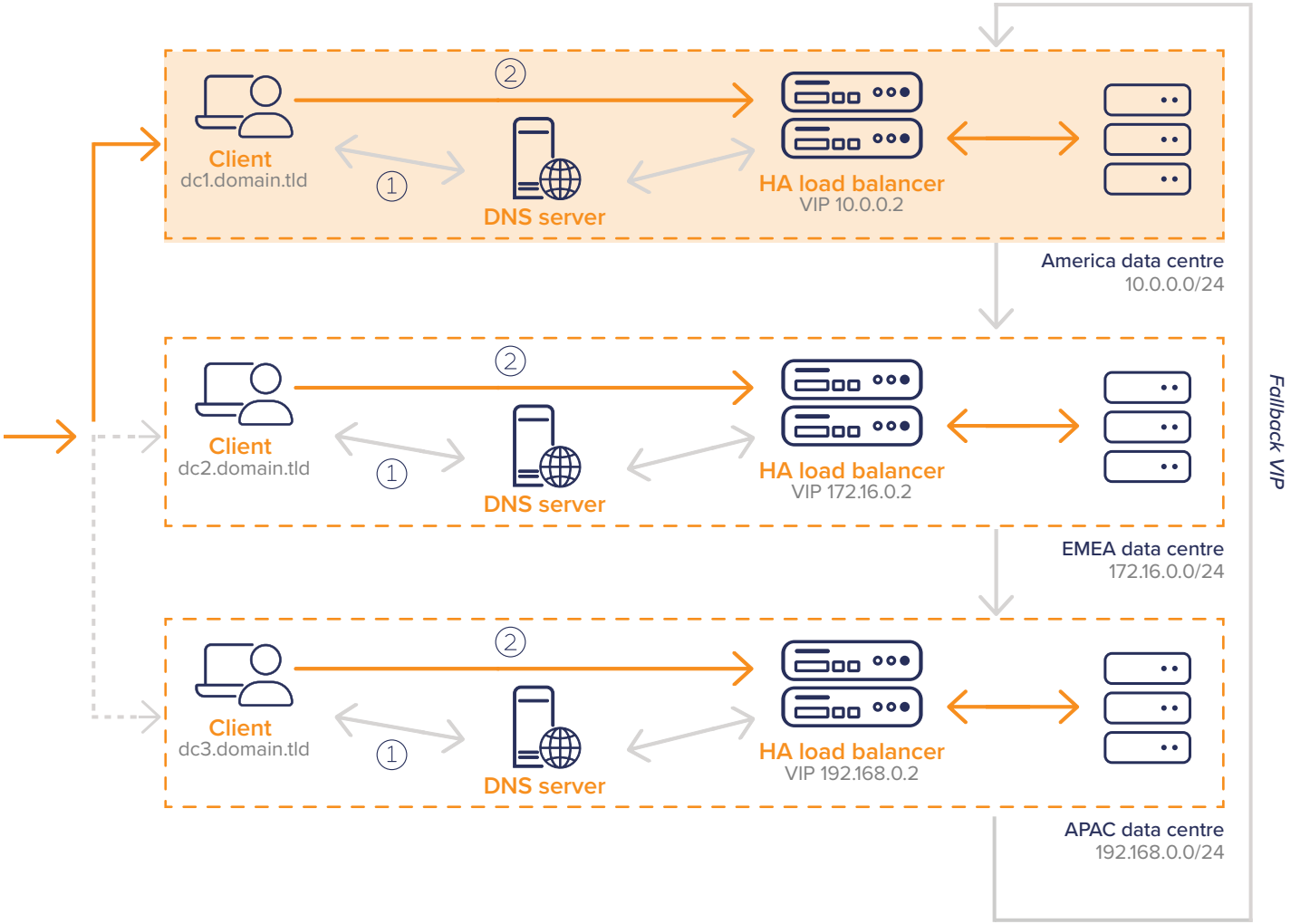
Using the topology-based algorithm TWRR, the Global SLB service can distribute traffic, while matching the source address of the requesting client's DNS server to the destination address in their local site – a simple, very effective solution.

Topologies consist of lists of IPs and subnets grouping sources to destinations, and still offer failover should all local resources be down – failover is up to you. It can failover to any remaining server, or reject. Finally, you can assign 0.0.0.0/0, and use it as a wildcard in the topology to create a default site for any external traffic.

4. FAILOVER GROUPS + EXAMPLE

Enhancing your disaster recovery response

Use failover groups to solve your simple disaster recovery problems, or when to failover between active and passive sites. You can specify multiple failover members, and the Global SLB will prefer the first healthy site or server in that list, only failing over after a negative health check. In our example, we also use a local load balancer for an extra failover layer – we can use a fallback server, with our Global SLB for a 'belt and braces' approach. This method is ideal when you have a DR site that isn't fully functional all the time, or when you have a staple, active/passive application that you need to offer a stage-by-stage failover.



About the company

Loadbalancer.org's mission is to ensure that its clients' businesses are never interrupted. The load balancer experts ask the right questions to get to the heart of what matters, bringing a depth of understanding to each deployment. Experience enables Loadbalancer.org engineers to design less complex, unbreakable solutions - and to provide exceptional personalised support.

To discuss your load balancer requirements with load balancer experts, contact Loadbalancer.org on:

loadbalancer.org

Email us: info@loadbalancer.org

Visit us: www.loadbalancer.org

Call us: +44(0)330 380 1064

Call us: +1 833 273 2566

Follow us: @loadbalancer.org

0362-EB-G-01