

Date: 8/17/2012

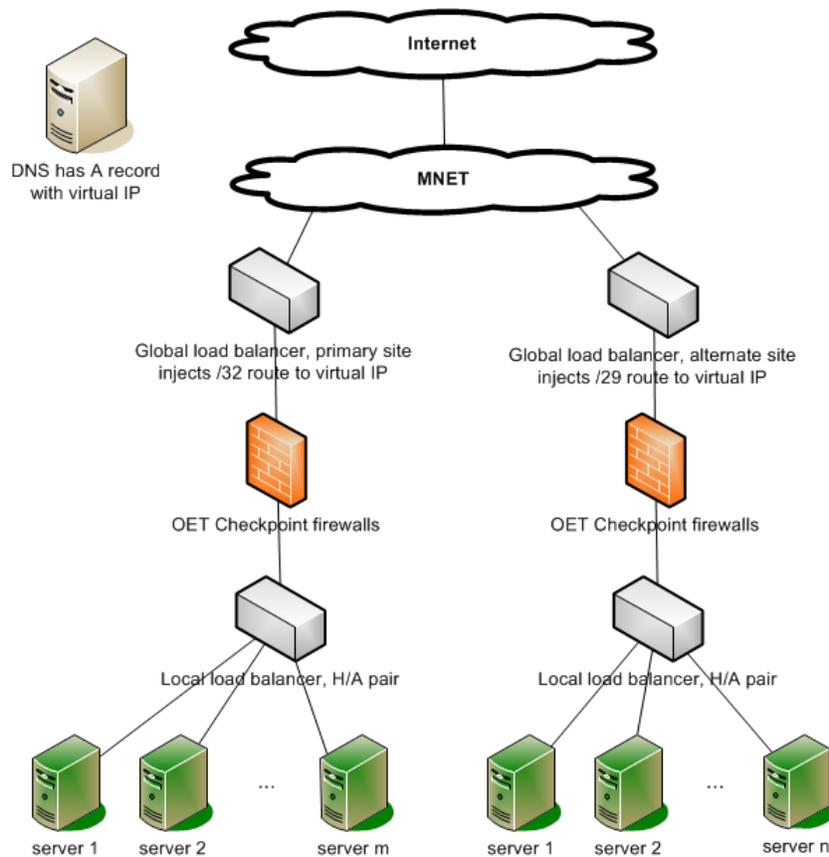
# Enterprise Architecture Office Resource Document Design Note - Load Balancing and High Availability

MN.IT Services offers load balancing services that can be used to spread load among multiple servers, implement high availability, or both.

This service is based on balancing at the IP address level. It does not operate by altering records in the DNS. The reason is that DNS records are (excessively) cached by other ISPs (in violation of the applicable RFCs) and thus the altered records are not seen by customers in a timely fashion.

We offer two levels of load balancing: global and local. They are illustrated in Figure 1.

**Figure 1: Load Balancing**



## Global

In this case, the load is balanced between two locations based on availability of a service. This is a yes/no availability and it does not pro-rate or scale traffic between the locations.<sup>1</sup> The configuration is normally for an active/passive basis, although active/active is available.

Normally, one of the locations is an MN.IT Services' facility. However, the services fronted by those load balancers may be in any location that can be supported by the network.

Each installation must be carefully engineered. Depending on the engineering analysis, it may not be possible for MN.IT Services to offer the service in all cases.

MN.IT Services currently uses Cisco devices for this service. Each global instance will receive its own VIP and is managed in its own virtual context.

## Implementing the Global Service

A virtual IP (VIP) in its own /29 block is defined for each service.<sup>2</sup> The alternate location injects the route for the /29 block and the primary location injects the (more specific) /32 route for the IP itself. If the primary location fails, the more specific route is no longer injected (and hence dropped); the traffic is then routed to the /29 block. If a service is provided with both locations online at all times, we allocate two VIPs and associated blocks, with one site injecting the /32 for one VIP and the /29 for the other and the other site injecting the other set.

The DNS record for the named service contains an A record with the VIP. If the service is a web page, the DNS will normally contain both the www and non-www versions of the name. For example:

```
example.org      A      1.2.3.4
www.example.org  A      1.2.3.4
```

The service has both locations online at all times, both VIPs will be in the DNS:

```
example.org      A      1.2.3.4
                 A      5.6.7.8
www.example.org  A      1.2.3.4
                 A      5.6.7.8
```

If one site has to be taken offline for maintenance, the DNS is not changed. Instead, the virtual instance of the global load balancer for that VIP is turned off (and back on again when maintenance has been completed).

While a /30 would be sufficient address space for both the primary and failover addresses, a /29 is used to provide for the case where we may expand to a third data center.

## Local

Local balancing spreads load among two or more servers. Load spreading can be pro-rated, round robin, or other methods.

MN.IT Services offers local load balancing at selected MN.IT Services facilities. In some cases, we can use VPNs and/or VLANs to extend these servers to other locations.

---

<sup>1</sup> Very crude capacity adjustments can be made by use of multiple IPs per data center.

<sup>2</sup> These will be /126 and /128 blocks for IPv6 addresses. Since the entire /64 is allocated for each network, sufficient address space is available for a third data center.

Each installation must be carefully engineered. Depending on the engineering analysis, it may not be possible for us to offer the service in all cases.

MN.IT Services currently uses Cisco devices for this service. Each local instance is managed in its own virtual context.

The servers behind the local load balancers have addresses that are *different* than the VIP address.