F5 Friday: Cookie Cutter vApps Realized

Lori MacVittie, 2011-21-10

An architectural solution to the challenge of IP-address dependency.

A rarely mentioned obstacle when attempting to duplicate or migrate enterprise-class applications is IP-dependency. Not just topological dependencies that are easily addressed with dynamic routing and switching protocols in conjunction with a boot script, but internal dependencies – the ones so deeply embedded in the application's "identity" that to change the IP address is to break the installation and render it useless.

These are the applications that, upon asking for an exported image for testing purposes, virtualization experts will tell you is far more efficient to start from scratch, because the IP dependency issue will cause more trouble in the long term than simply starting over. Moving such an application to a public cloud is nearly impossible due to this restriction, and any bursting or data center extension model is out of the question. This is also a problem locally, when attempting to build out a private cloud and IT services. particularly in production environments in which a multi-tenant model is employed by launching multiple instances of the same application with each designated for use by a specific logical group, i.e. a department, project, or business unit.

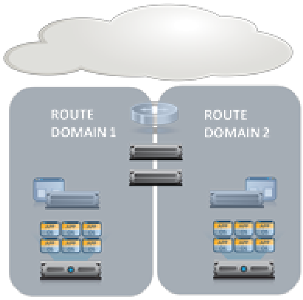
Ultimately what we want is the ability to create cookie cutter applications as a foundational element for IT as a Service. This requires network, security, and application policies – as well as the application – be encapsulated as templates, associated with the application, and applied on a per instance. This ultimately enables application instance sizing and chargeback per logical group, and lays the foundation for push-button IT services in which a department can be one click away from an automated deployment of an application.

What's standing in the way in many cases is the IP address dependency. Applications can't be packaged up neatly into a holistic service along with its requisite network, security, and delivery policies because all of these services are tightly bound to the IP address of the application – and vice-versa. When an application is deployed if it is reassigned a new IP address, every policy will also need to be updated, making the process not only lengthy but fraught with potential for misconfiguration due to stalls or human error.

The dependency on IP addresses within these applications is not going away. To achieve the goal of a more mobile and service-focused data center then, we need is a way to work around the problem. Many see VMware vApps as the solution. But while vApps were designed with mobility and portability in mind, it does not address the IP address dependency obstacle.

A solution to this seemingly unsolvable problem can be found in a collaborative architecture incorporating both global and local application delivery services.

A COLLABORATIVE F5–VMWARE ARCHITECTURAL SOLUTION



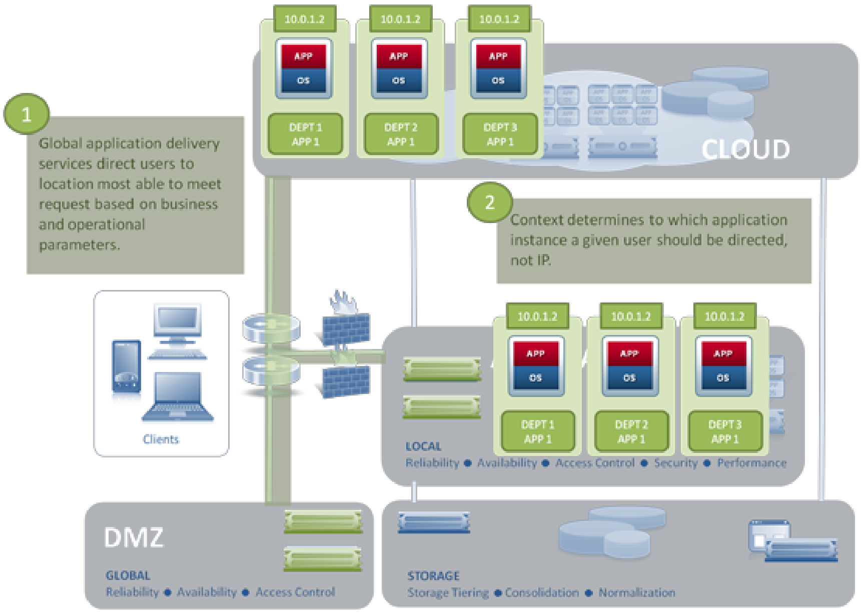
To avoid complexity in multi-DC topologies (and ultimately inter-cloud deployments), it is necessary to reduce the need for coordination between different teams by abstracting network addressing, rules and service names. Bridging networks is not enough – an application and protocol specific approach is needed. VLAN stretching approaches do not differentiate traffic ingress and egress for each datacenter. This means that application traffic can enter one datacenter, traverse the bridged network to the application in the other datacenter, and then return following the same path. As the distance between data centers increases (as is desired for disaster recovery purposes), this "trombone routing" incurs heavy performance penalties due to latency. What we want is not single valid addresses for applications across datacenters, but rather, portable addresses which can then be selected by a global abstraction based on best-path and best-performance for a given client in the context of their locality and the available resources in each datacenter.

Such an architecture is made possible by a rarely mentioned but very powerful feature of BIG-IP systems: route domains.

Route domains give you the ability to segment (isolate) network traffic for different applications on the network. The BIG-IP system can process traffic for each application within its own route domain. Because route domains segment network traffic they can also be used to assign the same IP address or subnet to more than one node on a network. Two nodes on the network can have the same IP address as long as each instance of the IP address resides in a separate routing domain.

The ability to essentially duplicate IP address space in the same environment opens up the ability to create cookie cutter vApps complete with the appropriate network, security, and delivery policies required – an isolated operationally consistent deployment. The problem then becomes ensuring that the right users are routed to the right application instance at the right time.

Using a phased implementation, IT organizations can resolve the issues that prevent the repeatable deployment of enterprise applications locally and globally.


10.0.1.2
APP
OS
DEPT 1 APP 1
DEPT 2 APP 1
DEPT 3 APP 1
CLOUD
1
Global application delivery services direct users to location most able to meet request based on business and operational parameters.
2
Context determines to which application instance a given user should be directed, not IP.
Clients
LOCAL
Reliability ● Availability ● Access Control ● Security ● Performance
DMZ
GLOBAL
Reliability ● Availability ● Access Control
STORAGE
Storage Tiering ● Consolidation ● Normalization

COURTESY F5 NETWORKS

PHASE 1

The focus of phase 1 is the elimination of re-addressing applications at the IP layer in multi-site deployments. This phase relies on BIG-IP Local Traffic Manager (LTM) and in particular route domains to allow the co-existence of architectures utilizing the same IP address space, and BIG-IP Global Traffic Manager (GTM) to determine which site is currently in use as the primary data center.

In an active-standby deployment, this provides site-resilience by ensuring a secondary site is available to assume responsibility for delivering applications in the event of an outage at the primary site. In an active-active deployment, BIG-GTM leverages context shared by the local application delivery controller, BIG-IP LTM, to ensure better performance and availability without sacrificing fault tolerance.

This deployment pattern is based on existing, proven global architectures providing site-resilience and location-based global load balancing.

PHASE 2

This phase also relies on BIG-IP Local Traffic Manager (LTM) and route domains to allow the co-existence of architectures utilizing the same IP address space. Context-awareness is leveraged as a means to properly route users to their designated application deployment. The context can be extracted from the URI or from other variables associated with the user, such as credentials or cookies.

Multiple instances of the application architecture can be launched and co-exist within the data center, each serving a particular logical group. Each group can size applications based on usage needs, and chargeback per department becomes a less complex accounting process as it is based on the instance and its supporting architectural components. Application architectures can be successfully repeated at the logical group level, enabling a smoother transition to IT as a Service and preserving the IP-address dependencies on which many applications rely.

ARCHITECTURE is KEY

As is increasingly the case, the solution to many of the challenges arising from multi-site, cloud computing , and highly virtualized data centers is architectural. Because the challenges often span data center domains – security, networking, storage, compute, and applications – the solution requires cross-domain collaboration, not just of teams but of infrastructure.

Cloud computing really is an exercise in infrastructure integration. By leveraging the strengths and capabilities of various data center components across various domains, solutions can be architected to address even the seemingly unsolvable problems that will continue to frustrate IT as it moves toward a more distributed and highly dynamic data center.

F5 Networks, Inc. | 401 Elliot Avenue West, Seattle, WA 98119 | 888-882-4447 | f5.com

F5 Networks, Inc.
Corporate Headquarters
info@f5.com

F5 Networks
Asia-Pacific
apacinfo@f5.com

F5 Networks Ltd.
Europe/Middle-East/Africa
emeainfo@f5.com

F5 Networks
Japan K.K.
f5j-info@f5.com