The shift towards Cloud based services is accelerating, whether it be Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) or Software-as-a-Service (SaaS). We are seeing the emergence of Cloud Service Providers (CSPs), who are standing up massive Data Centers, delivering on the Cloud’s promise of ubiquitous access, rapid elasticity and pay-per-use. These providers connect their geographically distributed Data Centers together into a common resource pool, to deliver a variety of multi-tenant, Cloud Services. When a cloud consumer requests the service, one of these Data Centers will be chosen to deliver the requested service, over the network.
Consider a cloud consumer in San Francisco who requests Video storage, organization and retrieval services from a SaaS provider, and her request is fulfilled by a randomly selected Data Center, located in South Africa. Now if that consumer tries to view her uploaded video hosted on the SaaS cloud provider, the video bits have to be transported halfway across the globe. It’s quite possible that the latency experienced maybe too high, resulting in degraded Quality of Experience (QoE) for that Cloud consumer.

When all consumers in a given geography are pointed to the nearest Data Center, that Data Center becomes prone to overload. How can we still maintain the consumer’s experience with the Cloud Service? This article explores various technologies to enhance and maintain Cloud service QoE.

**IP ADDRESS CONVERSION TO GEOGRAPHIC LOCATION**

The Cloud consumer’s IP address can be converted into location information including Country, Region, City, Zip or into a latitude/longitude pair. There are APIs [1] available on the web, which can help translate an IP address into geo-location information, accurate up to City level. For mobile consumers, other attributes such as WiFi access point or Cell Tower can be used to similar effect. Accuracy rates for these geolocation lookups vary, and sometimes the consumer may incorrectly show up as being located in an adjacent or nearby city.

Notwithstanding the accuracy and reliability issues, geolocation does provide an easy technique to locate the cloud consumer. Even a somewhat inaccurate geolocation is better than none at all. In the case of our cloud customer from San Francisco, at the least we would obtain enough information from geolocation to place the Cloud Service in the Northern California region – a vast improvement over selecting the location in South Africa, halfway across the globe!

**NETWORK PATH CHARACTERISTICS**

Application Layer Traffic Optimization (ALTO) is an IETF Working Group initiative aimed at facilitating the sharing of network related information with applications, thus enabling applications to make better choices in selecting their “peer” for P2P communications such as file sharing and media streaming [2]. The ALTO concepts can also be extended to Content Delivery Networks (CDNs) [3] where the network information shared by the Network Service Provider through the ALTO server can help to redirect the content requester to the “best” Content Node.

The CDN scenario could be applicable to Cloud consumers, which can use the information available through the ALTO server to determine the best Data Center to obtain the Cloud Service from. Examples of information that can be made available through ALTO are operator’s policies, geographical location or network topology – information that is typically considered longer lasting, and does not change rapidly.

ALTO can be a win-win for both the Service Provider and the Cloud consumer. Service Providers have an incentive to make network information available through the ALTO service, since it can decrease the overall amount of Cloud Service traffic across different networks, thus reducing network costs. Consumers can expect better performance, and QoE, by leveraging the information available through ALTO servers.
Continuing with our example scenario, let’s say that our consumer enjoyed the quality of the service and started recommending it to her friends or at work, or through positive reviews on local community social networks. If this turns viral, this cloud service could unexpectedly achieve a lot of popularity in San Francisco. This results in a lot of consumers being serviced from the nearby Data Centers, increasing the load on these Data Centers to many folds above average. This could result in resource contention inside the Data Center for compute, storage or network resources, ultimately degrading the QoE for all consumers.

One solution would be to move some of the Virtual Machines running the Cloud service (and their associated storage) to nearby Data Centers with spare capacity, thus helping to balance the load. Each Virtual Machine has an IP address, which is how ultimately consumers are connected to it. Moving a live Virtual Machine, which is currently serving several consumers, from one Data Center to another over a typical IP/MPLS network could be disruptive since the IP address of the Virtual Machine would need to change in accordance with the subnet in the new Data Center. Existing connections (typically TCP sessions) from the consumer to the Virtual Machine would need to be re-established, thus interrupting the user’s experience with the cloud service.

The above problem can be solved by extending the LAN fabric across Data Centers. Essentially, Data Center Interconnect technologies such as Virtual Private LAN Service (VPLS) or Overlay Transport Virtualization (OTV) can be used to connect multiple Data Centers over a Layer 2 mesh, allowing the migrated VMs to maintain their IP address and existing connections. While VPLS is offered by Network Service Providers, OTV could be deployed by Cloud Providers themselves to interconnect their Data Centers via an Layer 2 overlay running on top of the Service Provider’s IP/MPLS network.

Cloud consumers need to be served by an optimal Data Center to enhance their experience with the Cloud service. Geo-location and ALTO based methods could be used to accurately and reliably identify such a Data Center. Cloud providers need a way to avoid degrading performance, in the event of unexpectedly high, localized demand from consumers. Building Layer 2 connectivity between Data Centers allows them to non-disruptively migrate load in such scenarios, and preserve the user experience.

**Additional Resources**


[4] “OTV draft”, Hasmit et al., Click here or visit http://bit.ly/dNhXtC.

Disclaimer: Opinions expressed in this article and in any corresponding comments are the personal opinions of the original authors, not those of Cisco Systems Inc.