June 2023

TELEMETRY-BASED RESOURCE SCALING AND TRAFFIC OPTIMIZATION IN CLOUD INTERCONNECTS

Giorgio Valentini

Pradeep Kanavihalli Subramanyasetty

Venkat Venkatapathy

Madhavi Cherukuri

TELEMETRY-BASED RESOURCE SCALING AND TRAFFIC OPTIMIZATION IN CLOUD INTERCONNECTS

AUTHORS:
Giorgio Valentini
Pradeep Kanavihalli Subramanyasetty
Venkat Venkatapathy
Madhavi Cherukuri

ABSTRACT

Software-defined cloud interconnect (SDCI) connections are established through different providers with specific or proprietary characteristics. However, no existing scaling technology guarantees the automatic provisioning and teardown of an entire network of circuits, from a software-defined wide area network (SD-WAN) router to the cross connect that is required to link to a cloud provider. Presented herein are techniques that not only abstract different implementations (including hypervisors, software images, releases etc.) but also provide for the ability to detect congestion and implement end-to-end remediation from branch devices to cloud workloads. The presented techniques optimize the costs of resources (such as SD-WAN routers and middle-mile connections) without compromising the level of service that is offered by different middle-mile providers; allow an SDCI's automatic scaling of VMs and connections to be tied to the specific network SLA requirements of a user-application combination; and support networking solutions that enable a customer to build automated, scalable, and reliable interconnections that deliver richer cloud-based applications and services to enterprise customers while increasing operational efficiency.

DETAILED DESCRIPTION

Currently, software-defined cloud interconnect (SDCI) connections are established through different providers with specific or proprietary characteristics that may be abstracted through a vendor-supplied customizable facility that simplifies and automates the deployment, configuration, management, and operation of a software-defined wide area network (SD-WAN). For simplicity of exposition, such a facility may be referred to herein as a network management platform. Enterprise entities seeking to implement SD-WAN solutions are often willing to pay a premium for high-speed circuits that adhere to a guaranteed service-level agreement (SLA); however, such entities often have expectations regarding operation of such solutions. For example, it is often expected that an application's availability is guaranteed 24 hours a day, 7 days a week and, further, that costly over-subscription to circuits is not needed when traffic is off of a peak.

Even though some scaling solutions exist, such solutions typically focus on a specific technology and guarantee the scaling of just a specific resource (e.g., virtual network functions (VNFs), containers, pods, etc.). Current solutions typically do not guarantee the automatic provisioning and teardown of an entire network of circuits, from an SD-WAN router to the cross connect that is required to link to a cloud provider.

Techniques are presented herein that not only abstract different implementations (including hypervisors, software images, releases, etc.) but also uniquely detect congestion and implement end-to-end remediation, all the way from branch devices to cloud workloads. Such techniques may be effective regardless of the cloud and middle-mile provider technology of SD-WAN underlays or the application network needs (such as low latency, or high throughput or bandwidth, or low jitter and packet loss, etc.). The presented techniques optimize the costs of resources (e.g., SD-WAN routers, middle-mile connections, etc.) without compromising the level of service that is offered by different middle-mile providers.

The presented techniques tie an SDCI's automatic scaling of virtual machines (VMs) and connections to specific network SLA requirements of a user-application combination. For example, an entity may assign an SLA to an application based on knowledge of the application's bandwidth, latency, and jitter requirements, as well as knowledge regarding how the application can be scaled across multiple concurrent users without affecting the expected SLA. Scaling based on concurrent users is critical to meet SLA requirements and often involves a network management platform retrieving user information directly from a cloud provider.

Given an application's requirements for bandwidth, latency, and jitter (i.e., the application's SLA needs), and the management platform's knowledge of user scaling, the management platform may automatically tune an auto-scaling algorithm to an application's needs. The management platform may retrieve cloud load information regarding concurrent users (as well as an interconnect provider's telemetry regarding bandwidth, latency, and jitter) and calculate the optimal slicing criteria that can be utilized to guarantee the SLA based on all of the above parameters that influence an application.

During times of underutilization, the management platform may scale down to a minimum level of computing power and connections to the cloud. At any point in time, with minimal delay, information regarding the number of users and the telemetry from the interconnect and the cloud can influence a quick response by the management platform to the variable load.

Consider an illustrative example in which an enterprise operating a cloud environment seeks to host a ticket purchasing service for an e-commerce website.

Consider various steps to facilitate techniques of this proposal, as follows. During a first step, the cloud's traffic manager may route a user's request to the e-commerce website that is hosted in the cloud's application service. Next, the cloud's content delivery network (CDN) may serve up static images and content to the user. Next, the user may sign into the application through the cloud's active directory business-to-consumer (B2C) tenant. Then, the user may search for concerts using the cloud's search facility. In response to such searches, concert details can be retrieved from the cloud's structured query language (SQL) database and refer to purchased ticket images that are in located in the cloud's blob storage facility. In some instances, a database query results may be cached in the cloud's cache for redistribution to improve performance.

Next, consider that the user may submit ticket orders and concert reviews that may then be placed in a queue and the cloud's functions may process an order payment and concert reviews.

In some instances, cognitive services may provide an analysis of a concert review to determine a sentiment (such as positive or negative) and the cloud's application insights facilities may provide performance metrics for monitoring the health of the web-based application.

In a backend, datacenters may connect to an SDCI router that has private peering arrangements to the cloud. Application insights and the processing of payments may be downloaded to local servers such that payment applications in the cloud may be scaled up and down while, at the same time, the interconnect provider virtual cross connects (vXCs) may be scaled accordingly.

Consider another example use case that may involve a branch connecting to a cloud with unpredictable and cyclical traffic patterns. Figure 1, below, presents elements of such an example.

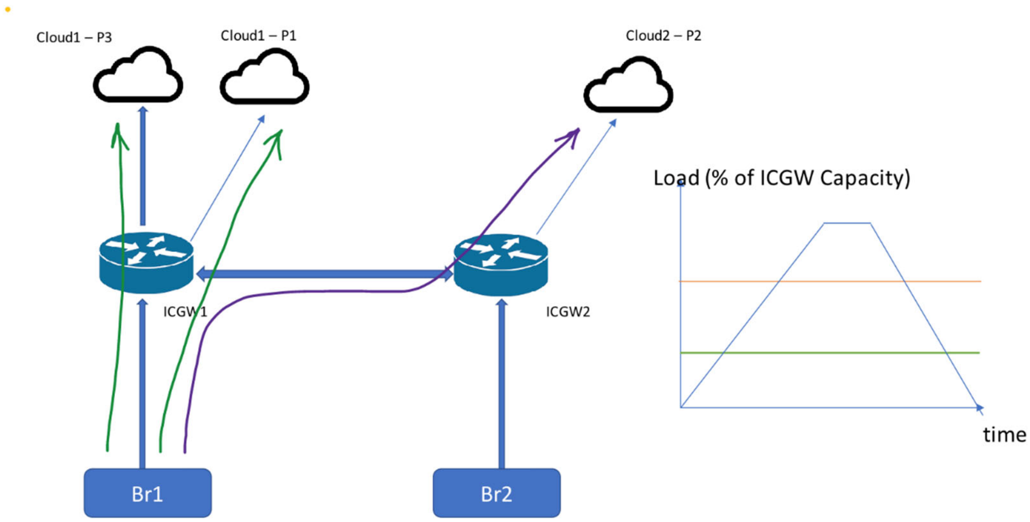

Figure 1: Illustrative Environment

As shown in Figure 1, above, a user's traffic may flow through a middle-mile provider backbone to reach a cloud destination in which each router (such as ICGW1, ICGW2, etc.) and connection from each router into the cloud may involve costly middle-mile infrastructure. It can be difficult for an entity to establish such connectivity in a manner that will automatically scale to the entity's growth, new applications, or traffic volatility. The techniques presented herein offer a solution that automat addresses such challenges in an automated manner. For example, the presented techniques may utilize telemetry from a middle-mile provider to detect application utilization peaks and lows, over time, and then dynamically adjust resources (including both SD-WAN routers and cross connects) based on the same. As a result, an entity may pay only for what capacity is needed at a given point in time, quickly scaling up when (when needed) and then reducing the network footprint when the traffic bandwidth stabilizes to lower levels.

The techniques presented herein encompass a number of elements, including the introduction of auto-scale groups (as opposed to today's interconnect gateway) and the constant monitoring of a traffic load on all such auto-scale groups. A branch may perform equal-cost multi-path (ECMP) routing over all of the inter-converged gateways (ICGWs) of a group to scale up to a maximum available bandwidth and VNF auto-scaling may be extended to include connectivity auto-scaling.

Further elements include an ability to poll a provider's telemetry at fixed intervals (yielding standardized, provider agnostic telemetry data that is gathered from third-party providers for underlay and device utilization), feed such telemetry data to an auto-scale algorithm (which may be finetuned based on an application), and use different telemetry data thresholds depending upon an application's requirements (such as latency, bandwidth, jitter, concurrent users, etc.).

Still further elements include an ability to dynamically deploy and destroy a set of devices on a site in connection with an automatic scaling of instances when load increases above a user's defined threshold and a scaling down when traffic lowers below a minimum threshold. Additionally, cloud provider information may be retrieved (from, for example, a cloud's application insights facilities) concerning concurrent users that are running an application.

Using the branch-to-cloud topology that was described above, the techniques presented herein allow a user to scale bandwidth by adding SDCI gateways based on a key parameter that is selected by the user. Figure 2, below, presents elements of such an activity.

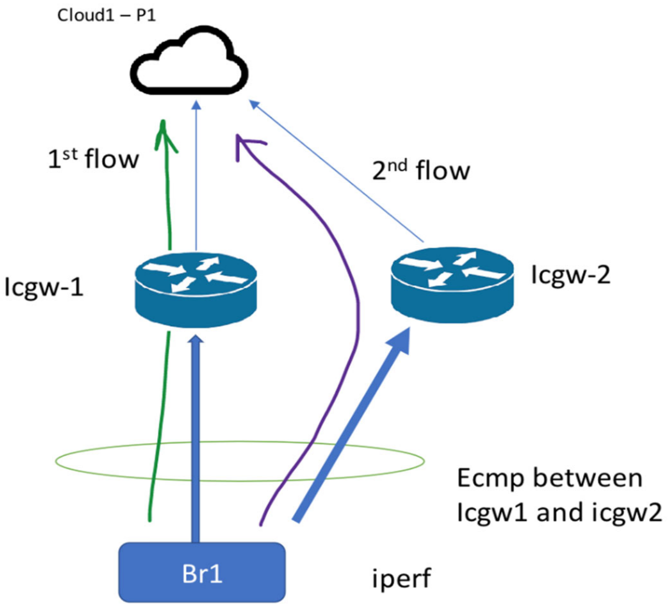

Figure 2: Exemplary Scaling

As described previously, and as illustrated in Figure 2, above, the techniques presented herein not only scale computing power (such as SD-WAN routers) up and down but also scale up and down the network connections from those routers to cloud end points having a greater bandwidth.

Figure 3, below, presents a process flow that captures elements of various operations that utilized through the techniques presented herein.

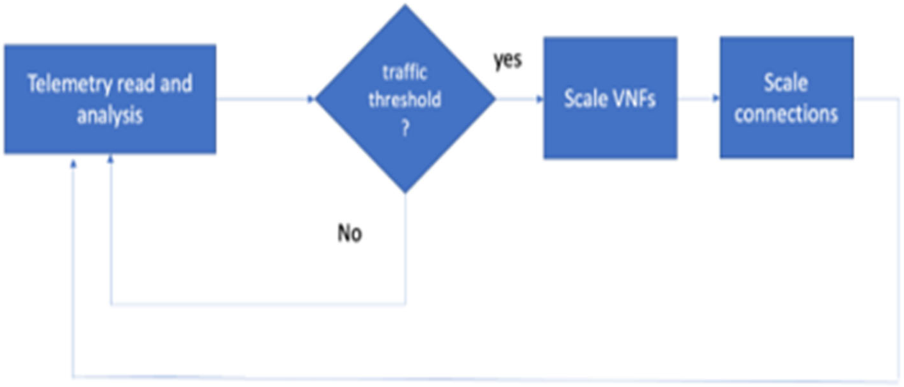

Figure 3: Exemplary Process Flow

As described and illustrated in the above narrative, a novelty of the techniques presented herein encompasses the ability to match middle-mile provider telemetry on a cross connection to a cloud provider's tracking of application usage and user activity. By retrieving different usage statistics, at different points on a network and produced by different providers, the presented techniques may adapt the end-to-end connectivity scaling to latency, bandwidth, and jitter (i.e., network statistics); application requirements (such as real-time, round-trip time (RTT) sensitivity); and concurrent users (to scale resources up and down to ensure that an SLA is met regarding the users that are expected to consume an application).

In summary, techniques (which leverage third-party telemetry, and which are network provider agnostic, virtual instance agnostic, cloud agnostic) have been presented herein that not only abstract different implementations (including hypervisors, software images, releases etc.) but also uniquely detect congestion and implement end-to-end remediation, all the way from branch devices to cloud workloads. Such techniques are effective regardless of the cloud and middle-mile provider technology of SD-WAN underlays or the application network needs (such as low latency, or high throughput or bandwidth, or low jitter and packet loss, etc.). The presented techniques optimize the costs of resources (such as SD-WAN routers and middle-mile connections) without compromising the level of service that is offered by different middle-mile providers; allow an SDCI's automatic scaling of VMs and connections to be tied to the specific network SLA requirements of a user-application combination; and support networking solutions that enable a customer to build automated, scalable, and reliable interconnections that deliver richer cloud-based applications and services to enterprise customers while increasing operational efficiency.