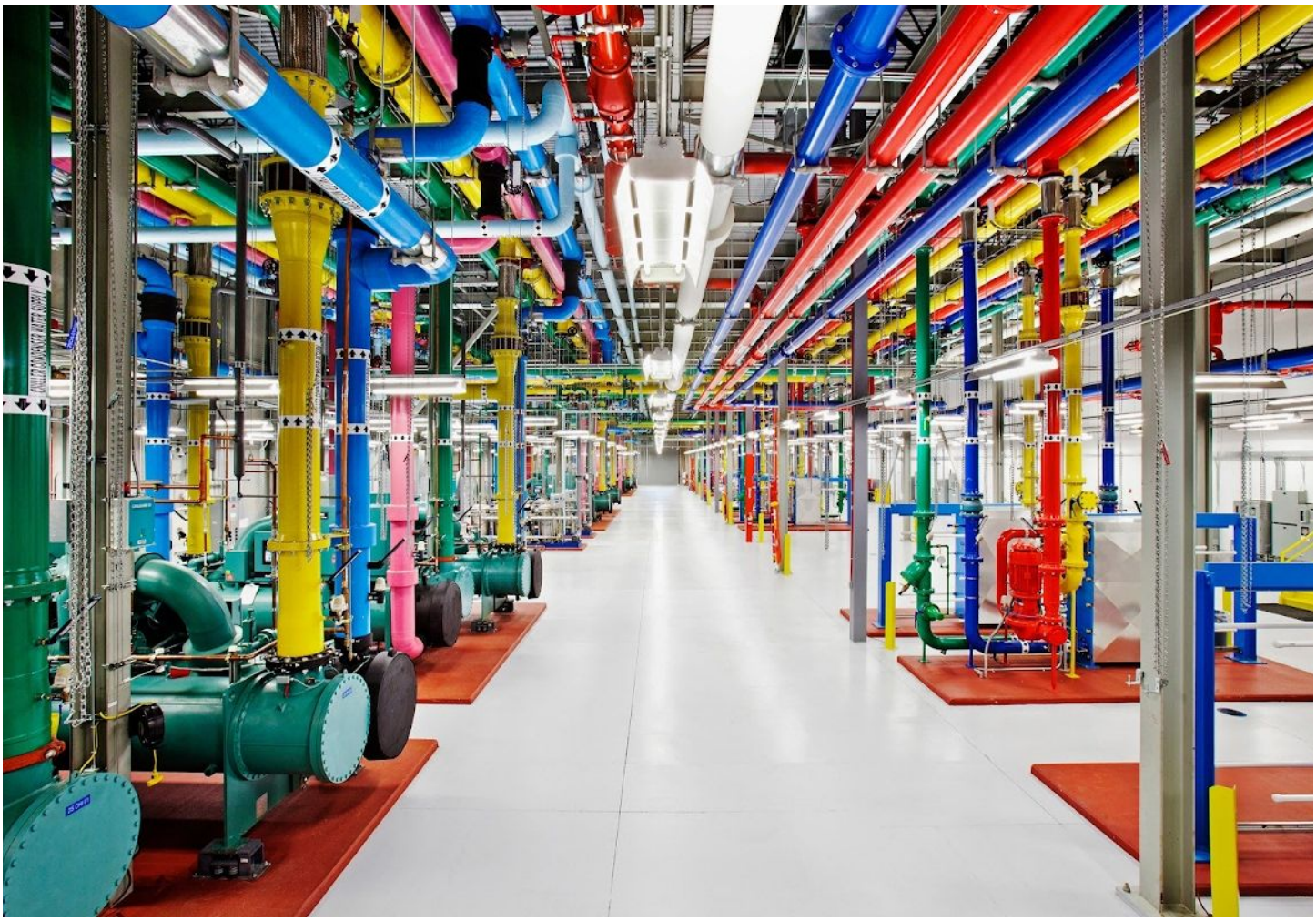
Google Cloud

Application awareness on Cloud Interconnect: overview

Prioritize your business-critical traffic on your Cloud Interconnect



Overview

Google Cloud's Cross-Cloud Network enables scalable and secure network architectures to deploy distributed applications spanning multiple cloud and on-premises networks. High speed Cloud Interconnect offerings with enterprise grade BGP routing and SLA-backed availability are a key component of such architectures that provide a unified cloud experience through appropriate connectivity options for distributed applications.

Application awareness on Cloud Interconnect provides granular controls to prioritize your mission-critical traffic during congestion events or traffic bursts on your Cloud Interconnect.
This solution brief describes how you can leverage application awareness on Cloud Interconnect offerings to enable traffic differentiation and boost the resilience of your distributed applications hosted across multiple cloud or on-premises networks.

Intended audience

Cloud Architects, Infrastructure Architects and Cloud Administrators responsible for provisioning and managing networking for hybrid and multicloud deployments of distributed applications.

Customer challenges

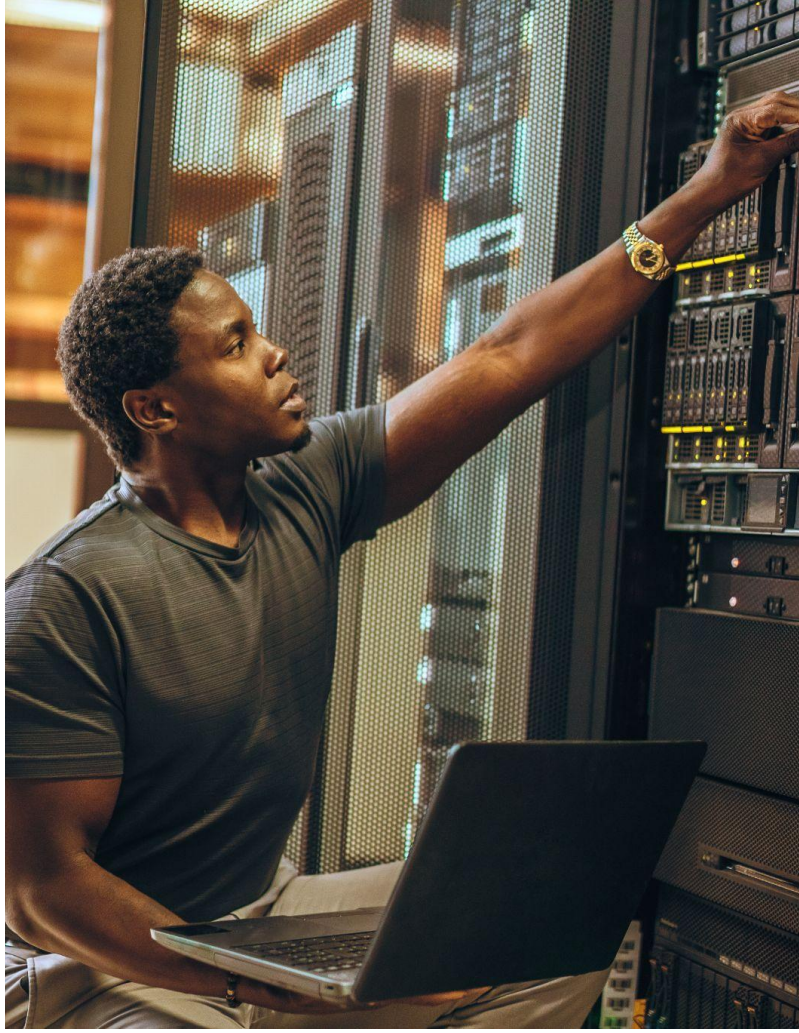
As an increasing number of business-critical workloads move to the cloud, the network traffic patterns across customers' on-prem and cloud deployments are not only getting more diverse, but also have a strong and direct impact on customers' business. Composite applications built across on-prem and multicloud, and accelerated consumption of best-of-breed SaaS offerings add to the complexity. This complexity is further exacerbated for AI/ML workloads where the traffic patterns are different from traditional web applications – traffic bursts, elephant flows, low latency delivery and high bandwidth needs are inherent traits of AI/ML data transfers, model training and inferencing.

Different workloads have different demands from the network. For example:

- Workloads like backups and CI/CD pipeline updates can have sustained bandwidth demands but have higher fault tolerance.
- Real-time transactional workloads leveraging, e.g., Cloud Spanner or BigQuery might have a lower tolerance for latency excursions.

Protecting your business-critical applications becomes increasingly important in an environment with a mix of high priority and low priority traffic sharing your network infrastructure. This is especially true on Cloud Interconnect, which is inherently a fixed bandwidth physical resource that cannot accommodate dynamic bandwidth scaling - hence a traffic spike or surge, e.g., from a backup application, can end up congesting the Cloud Interconnect and cause high latency and even packet loss for other applications sharing the Cloud Interconnect bandwidth.

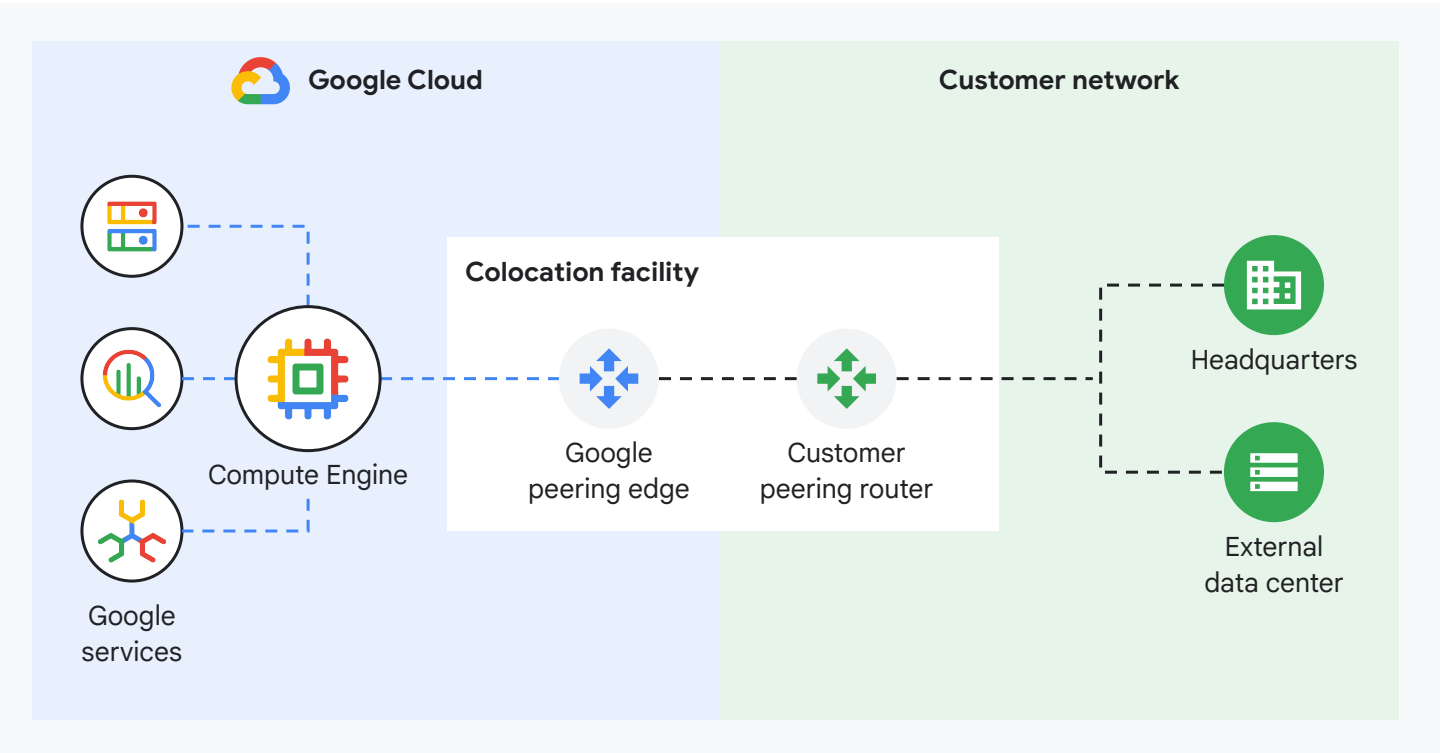
Application awareness on Cloud Interconnect



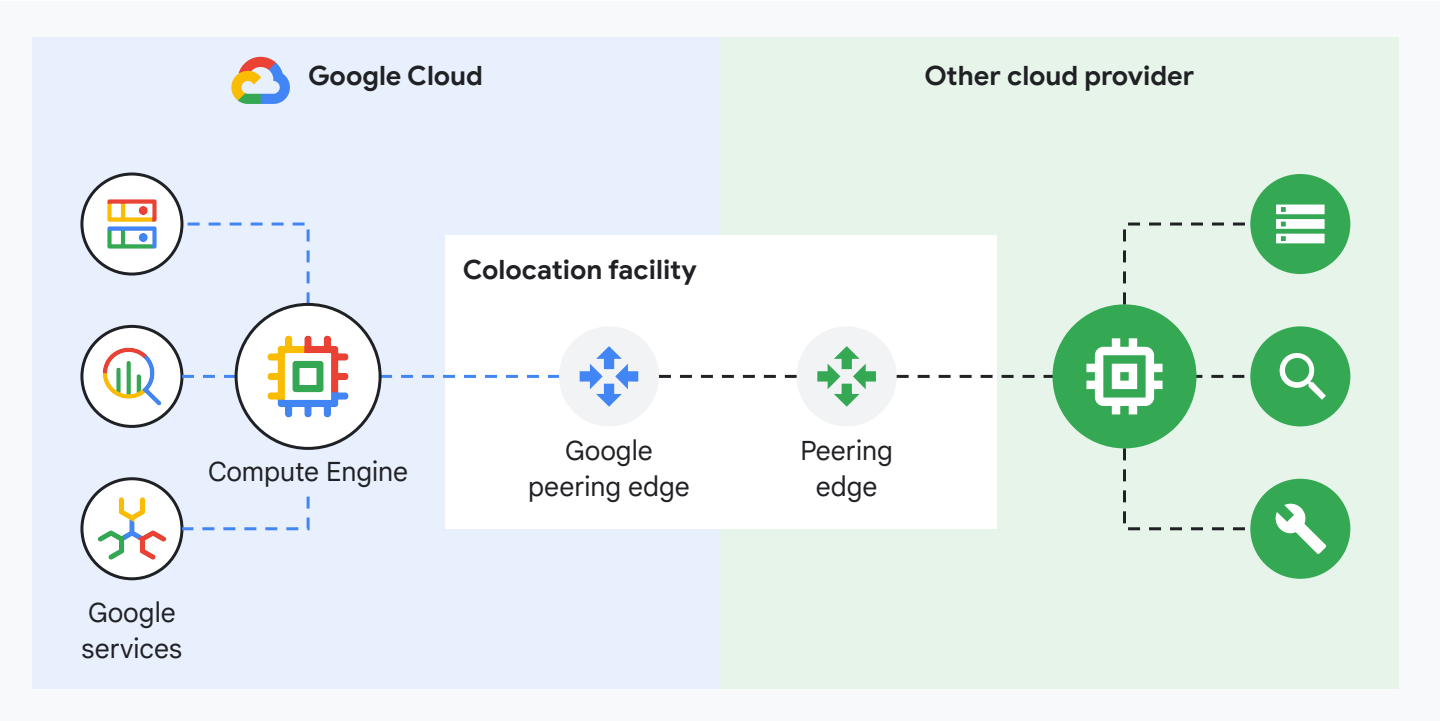
To better understand the value that application awareness can provide, let's first take a quick look at two of Cloud Interconnect offerings for direct physical connections and their value to your distributed workload connectivity.

Dedicated Interconnect provides a direct physical connection between your on-premises environment and Google Cloud. This connection exists in a colocation facility where your on-premises routing equipment meets Google's peering edge. Dedicated Interconnect connectivity is delivered over one or more 10 Gbps or 100 Gbps Ethernet connections.

Network connectivity for Dedicated Interconnect



Network connectivity for Cross-Cloud Interconnect

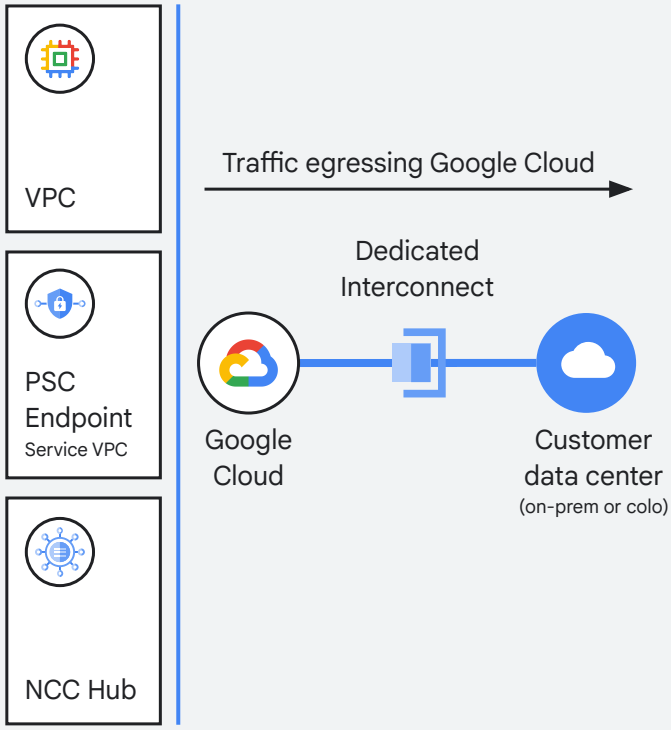


Cross-Cloud Interconnect helps you establish secure and direct high-bandwidth connectivity between Google Cloud and another cloud service provider for your multicloud workloads. Cross-Cloud Interconnect is available in two sizes: 10 Gbps and 100 Gbps.

Application awareness on Cloud Interconnect enables you to prioritize your business-critical applications and ensure their high availability, even during congestion events on your Cloud Interconnect. It helps you reduce your operational overhead and costs via efficient utilization of your Cloud Interconnect along with enhanced visibility. The configurable application awareness policies applied on your traffic egressing Google Cloud enable granular control of how your Cloud Interconnect bandwidth is allocated to your applications.

It is important to follow all the recommended best practices when configuring Cloud Interconnect, in particular, for creating redundant Cloud Interconnect connections with sufficient capacity.

Traffic from Google Cloud resources egressing Google Cloud through Cloud Interconnect



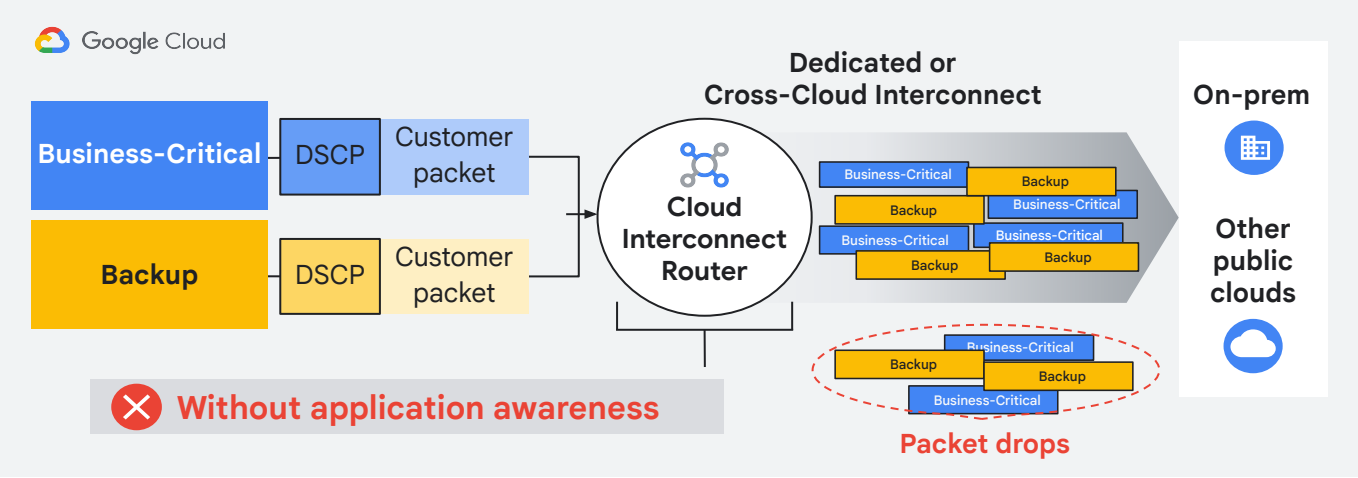
Need for application awareness on Cloud Interconnect

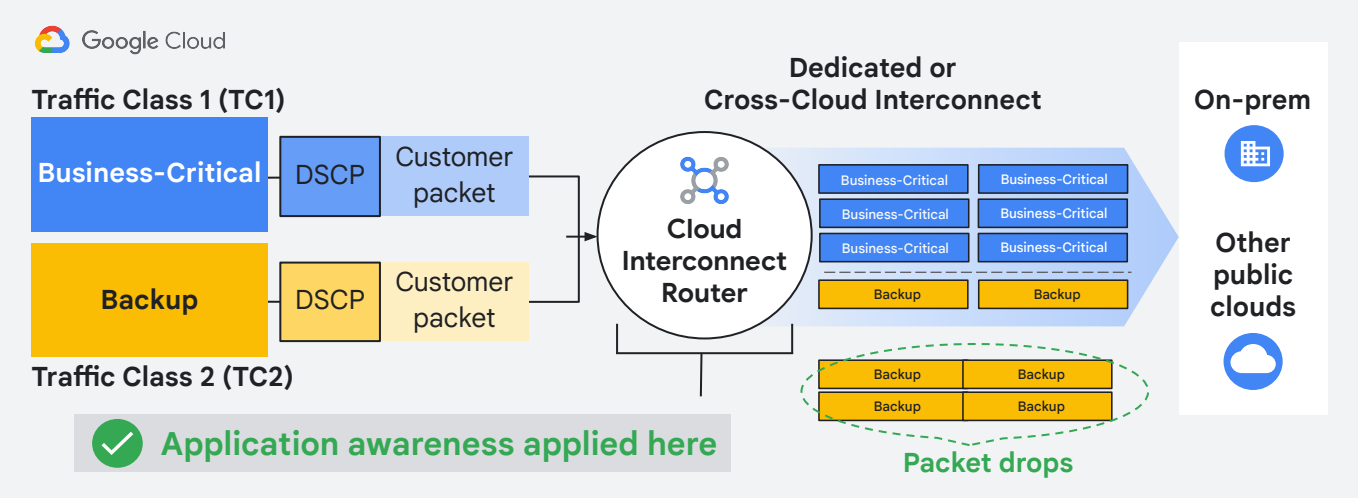
Traditional interconnect solutions lack the capability to prioritize traffic based on application needs. This results in increased costs, inefficient resource utilization, and potential disruption to business operations due to traffic spikes and network congestion events causing packet loss for mission-critical applications.

If you have multiple distributed applications in Google Cloud that communicate to on-prem or to another cloud service provider over Cloud Interconnect, application awareness enables you to control how your Cloud Interconnect capacity is shared across these workloads.

This allows you to prioritize traffic from your business-critical applications with low tolerance for packet loss. In case of a congestion event on your Cloud Interconnect, traffic from your business-critical applications is now safeguarded against indiscriminate drops ensuring their consistent performance and availability.

Traffic egressing Google Cloud via Cloud Interconnect with and without application awareness

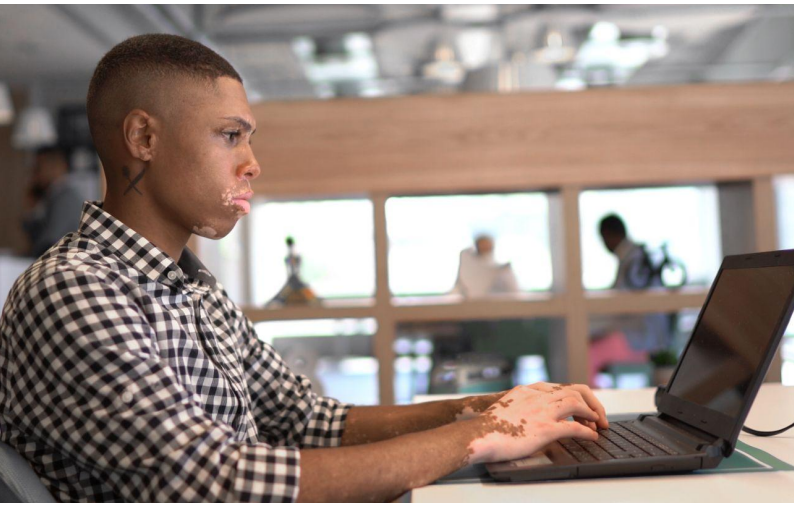




Benefits of application awareness on Cloud Interconnect

Application awareness offers the following benefits:

- **Differentiated service:** Ensure that your lower priority traffic egressing Google Cloud does not adversely impact your higher priority traffic during congestion events on your Cloud Interconnect.
- **Higher reliability and availability:** Ensure that your business-critical and/or real-time applications have high availability even during traffic spikes on your Cloud Interconnect.
- **Optimize costs:** Use your Cloud Interconnect bandwidth more efficiently, avoid excessive over-provisioning of links and save costs. Please note that application awareness on Cloud Interconnect is not a substitute for proper capacity planning.
- **Predictability of spend:** Enables predictable spend with granular control for efficient Cloud Interconnect bandwidth utilization.
- **Scalabile deployment:** Avoid complex workarounds (e.g. mapping different traffic types to different Cloud Interconnect links) with limited scalability.
- **Consistent traffic handling:** Uniform treatment of traffic across on-prem and Cloud Interconnect for easier "lift and shift" of workloads to Google Cloud.
- **SLA-backed solution:** Compared to expensive, self-managed network appliances that add complexity, manual toil, and high operational overhead.
- **Enhanced visibility:** Enhanced monitoring capabilities on your Cloud Interconnect provide per traffic class-level visibility into packets sent, packets dropped, bytes sent, and bytes dropped.



Leverage VPC Flow Logs to obtain detailed insights into your Cloud Interconnect traffic at flow-level granularity (i.e., 5-tuple: src/dst IP, src/dst port, protocol), including bandwidth utilization and traffic class information.

Implementation overview

Application awareness is deployed on your Cloud Interconnect via a three stage process.

1 Traffic marking

Explicitly marking packets

Action: user

2 Traffic classification

User traffic marking mapped to six traffic classes

Action: none

3 Policies via Traffic Map

Options:

- Strict priority
- Bandwidth sharing
- Optional shaper profile

Action: user

1 Traffic marking / marking application intent

Marking is achieved using the differentiated services field codepoint (DSCP) values in the IP header of your packets.

Explicit marking: Marking the DSCP bits in the IP packet header yourself. For example, you could leverage the one of the following options to implement explicit marking.

- On-host agents that leverage the following utilities:
  - iptables
  - setsockopt() (IPv4, IPv6)
  - XDP / eBPF-based filters
- A gateway agent responsible for marking DSCP bits: A customer-provisioned, VM-hosted Gateway agent that aggregates all the traffic and marks the DSCP bits for the traffic in a centralized manner.

  Note that this choice has some tradeoffs in terms of an additional network hop leading to additional latency and a potential network choke-point – it should only be considered if there are no means to leverage the on-host agents on VMs running the workloads, and/or the markings from the distributed sources across the organization cannot be trusted.

2 Network traffic classes

The DSCP marking in the IP packet header is used by Google Cloud to classify your application traffic into six traffic classes (TC1 – TC6). The table below provides the DSCP ranges mapping to each traffic class. Note that for the strict priority policy, TC1 has the lowest priority and TC6 has the highest priority.

DSCP ranges marked by the customer	Traffic class	Description
000xxx	TC1	Typically corresponds to the lowest priority “best_effort” class of traffic.
001xxx	TC2	Typically corresponds to the “low” priority traffic being carried in the network, e.g., bulk copy traffic.
010xxx	TC3	Typically corresponds to the “medium” priority traffic being carried in the network.
011xxx	TC4	Typically corresponds to “high” priority traffic being carried in the network, e.g., streaming.
10xxxx	TC5	Typically corresponds to “critical” traffic carried by the applications, e.g., interactive/user-facing traffic.
11xxxx	TC6	Typically corresponds to the “network_control” or the NC traffic class that includes protocol packets (e.g., BGP, BFD).

Note that six classes are for providing granularity for traffic differentiation.
It is not mandatory to use (i.e., map your traffic to) all six classes.

3 Implement application awareness policies

You assign an application awareness policy via a 'Traffic Map' construct and invoke it on a specific Cloud Interconnect. 'Traffic Map' captures the treatment of the traffic classes. Application awareness offers two policies to implement differentiated treatment of your application traffic.

Shaper profile (optional)

In addition to the queuing policies, you can deploy a shaper profile that controls the maximum bandwidth a traffic class can consume.

The shaper profile can be used to smoothen out traffic bursts for traffic classes with a bursty traffic profile. We recommend using shaper controls for your high priority traffic while using strict priority policy so as to safeguard against starvation scenarios for your low priority traffic.

Strict priority

Higher priority traffic can preempt lower priority traffic when there's congestion on your Cloud Interconnect.

Recommended for:
Traffic profiles where the high priority applications sending traffic over Cloud Interconnect are clearly identified and have predictable and consistent (non-bursty) bandwidth utilization.

Bandwidth percentage

You can assign each traffic class a minimum bandwidth share which is enforced during congestion. Any unused bandwidth can be used by all traffic classes that have packets to send.

With the bandwidth percentage policy, the configured bandwidth reservation is only enforced during a congestion event, i.e., any traffic class can burst as long as there's bandwidth available (no congestion). Any unutilized bandwidth is shared equally among the traffic classes with nonempty queues.

Recommended for:

- Traffic patterns where the relative priority amongst applications is not clear. In such cases, it might be prudent to allocate a bandwidth share to each traffic class that's only enforced during congestion events.
- Scenarios where high priority applications are known and some (or all) of the high priority applications are bursty – the bandwidth percentage policy ensures that those traffic bursts don't starve lower priority traffic while still ensuring efficient bandwidth utilization of your Cloud Interconnect.

Deployment scenarios and architectures

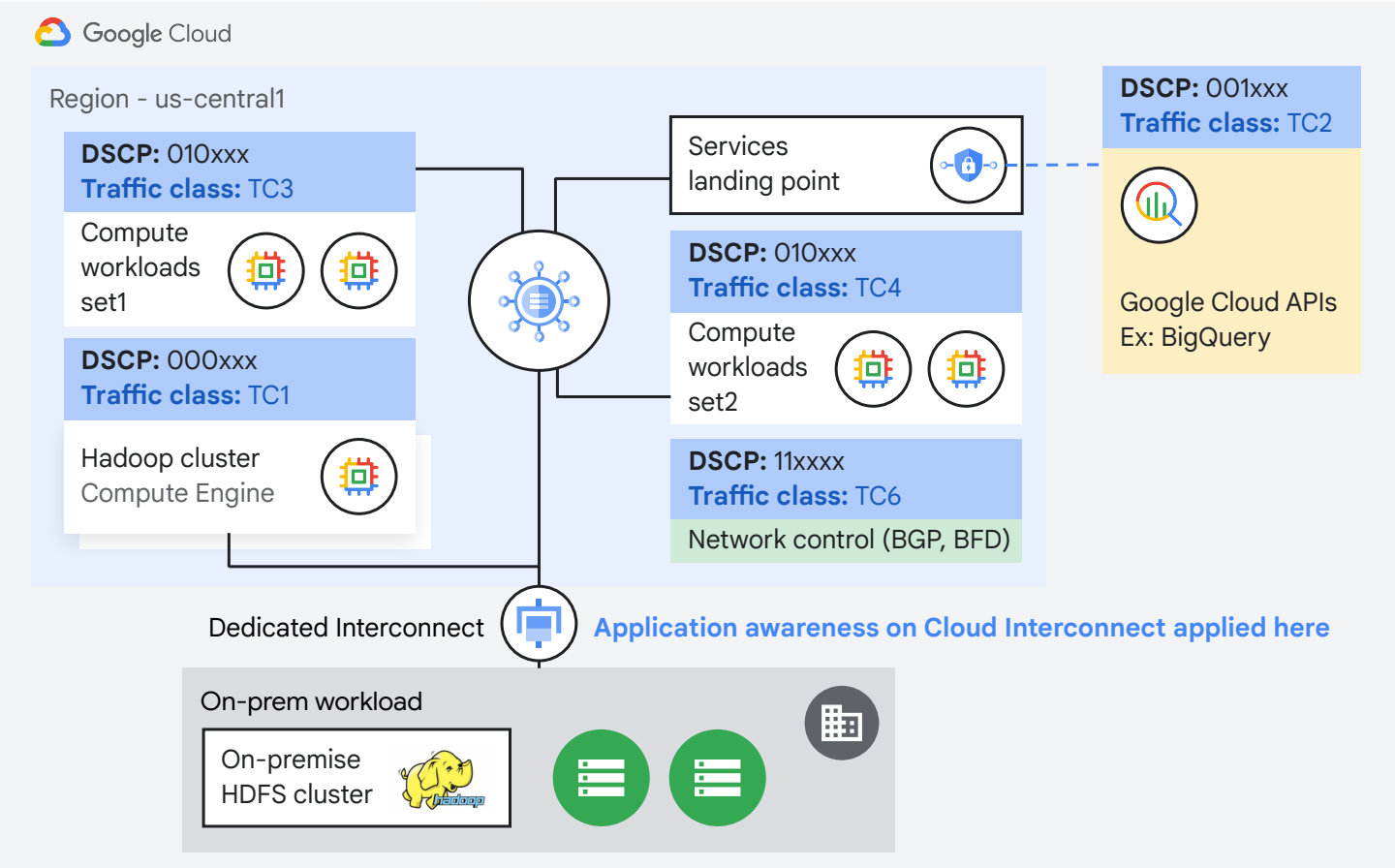
1 Example scenario 1: Google Cloud to on-prem traffic from predictable, non-bursty high priority applications

Consider a scenario where your organization hosts a few distributed applications on Compute Engine - Compute Workloads Set1 and Set2. These workloads are important for your business, require predictable bandwidth (not bursty), and need to communicate with your dataset residing on-premise due to data residency constraints.

Your organization also has a large HDFS workload (heavy bulk traffic and chatty processes) deployed across Google Cloud and your on-premise data center.

Additionally, you have an on-prem workload that consumes a Google Cloud API, e.g., BigQuery via Private Service Connect.

You have limited visibility into the bandwidth usage on your Dedicated Interconnect and no control on segmenting and prioritizing your traffic egressing Google Cloud. You might be experiencing periods of bandwidth utilization spikes and packet loss on your Dedicated Interconnect impacting the availability and performance of your business-critical workloads.



Here's how you can leverage application awareness on your Dedicated Interconnect in this scenario.

You'll first need to mark traffic with appropriate DSCP bits (explicit marking) to create the following classes of traffic.

- **TC6:** Network control (e.g. BGP, BFD)
- **TC4:** High priority traffic: compute workloads set2 (e.g., billing, transactions workloads).
- **TC3:** Medium priority traffic: compute workloads set1 (e.g., user-facing, latency-sensitive apps).
- **TC2:** Low priority traffic: BigQuery API.

  Note that for Google Services traffic egressing Cloud Interconnect, you mark your service request packets with the appropriate DSCP. Google Services will echo back the same DSCP in the service response.

- **TC1:** Unmarked traffic (DSCP 0): HDFS traffic - bursty, bulk data transfer, chatty processes like MapReduce. This is the lowest priority traffic.

You can then deploy a strict priority policy which will ensure that bursty traffic from your Hadoop deployment does not impact your higher priority traffic, helping avoid any packet loss or latency increase for your performance-sensitive workloads during high bandwidth utilization scenarios.

In case of traffic congestion on your Dedicated Interconnect, your network control traffic gets the highest priority, followed by the traffic from your compute workloads set2 (TC4), followed by compute workloads set1 (TC3), followed by your BigQuery traffic (TC2) while the Hadoop traffic (TC1) gets the lowest priority on your Cloud Interconnect. You can optionally configure a shaper profile to safeguard against starvation scenarios for your lower priority traffic.

Application awareness on Cloud Interconnect provides you per traffic class level visibility (per traffic class packets sent, packets dropped, bytes sent, bytes dropped) into the traffic egressing Google Cloud via your Dedicated Interconnect to help you monitor your deployment, track per traffic class bandwidth utilization, and diagnose any issues. You can also leverage VPC Flow Logs for a more granular, flow-level visibility into your Dedicated Interconnect traffic, e.g., to identify your top talkers or isolate flows by traffic class or DSCP value. Use this enhanced visibility to monitor, audit, analyze and optimize your deployment for performance and cost.

2 Example scenario 2: Cross-cloud traffic from bursty applications with a mix of priorities

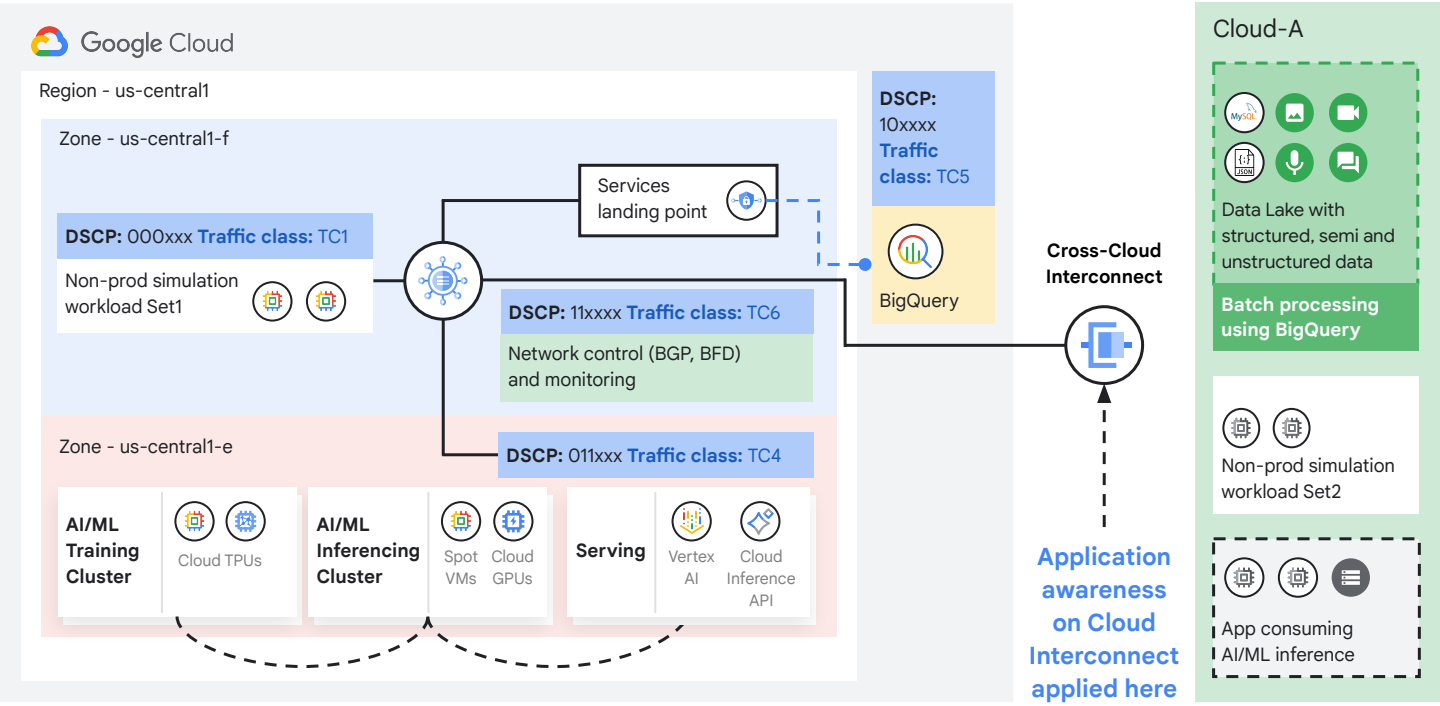
Consider a scenario where your organization has a multicloud data lake estate workload that integrates a batch processing platform hosted in Cloud-A which sends queries to BigQuery. The daily traffic flow per day between these clouds is much higher for traffic egressing Google Cloud to Cloud-A versus the reverse direction traffic on your Cross-Cloud Interconnect.

Your organization prioritizes cross-cloud network control (including BGP, BFD), application control, and the critical monitoring and diagnostics traffic. While the bandwidth requirement for this traffic is very low, it needs to be prioritized and protected over all other traffic per your organization's best practices.

Additionally, your organization uses Google Cloud for its advanced AI/ML infrastructure (pre-training and training on TPUs, custom ML workflows on Vertex AI, inferencing and serving) where the AI/ML model inference is consumed by an enterprise application deployed in Cloud-A. While this cross-cloud AI/ML traffic egressing Google Cloud is not as bandwidth intensive as the BigQuery workload, it is latency sensitive, and can be bursty. This creates significant challenges in protecting other traffic sharing the same Cross-Cloud Interconnect bandwidth.

Your organization also has a non-prod simulation workload being run by your research team on Google Compute Engine, deployed across Google Cloud and Cloud-A. This non-prod traffic is bursty and is to be served only on a best effort basis.

You have little insight into how your Cross-Cloud Interconnect bandwidth is being used, and you lack the ability to segment or prioritize your outbound Google Cloud traffic. This could result in periods of high bandwidth usage and packet loss on your Cross-Cloud Interconnect.



Here's how you can leverage application awareness on Cloud Interconnect in this scenario.

You could mark your network control and your application traffic with appropriate DSCP bits (explicit marking) to map to four classes of traffic: TC6, TC5, TC4, TC1.

- **TC6:** Network/app control and monitoring traffic.
- **TC5:** Critical priority traffic: BigQuery workload which uses the highest bandwidth.
- **TC4:** High priority traffic: AI/ML workload which is latency sensitive.
- **TC1:** Low priority traffic: non-prod compute workload.

Next, implement a bandwidth percentage policy which will ensure the configured bandwidth share for your AI/ML (TC4) and BigQuery (TC5) workloads safeguarding them from packet drops and delays in the event of traffic congestion on your Cross-Cloud Interconnect. You will also need to allocate appropriate bandwidth to the control and monitoring traffic (TC6) across your Cross-Cloud Interconnect.

- **TC1 - 5%**
- **TC2 - 1%**
- **TC3 - 1%**
- **TC4 - 33%**
- **TC5 - 55%**
- **TC6 - 5%**

You will need to ensure that all six traffic classes are allocated a nonzero percentage of bandwidth, and that the total bandwidth allocation adds up to 100. For traffic classes that you don't intend to map any of your application traffic to (e.g., TC2 and TC3 in this example), configure the bandwidth percent as 1% (minimum allowed value). Note that configuring a nonzero bandwidth percent (1%) for the unused traffic classes doesn't imply that the bandwidth will not be available for your traffic from other traffic classes – as noted earlier, the unutilized bandwidth for any traffic class is shared equally among other traffic classes with packets to send.

With this deployment, during a congestion event, bandwidth sharing policy gets enforced on your Cross-Cloud Interconnect enabling you to prioritize and protect your control traffic, and reserve the required bandwidth for your BigQuery workload and the AI/ML workload. Your non-prod best effort traffic is not starved given the configured (5%) bandwidth, and all your traffic classes get to share any unutilized bandwidth.

Application awareness also provides you per traffic class level visibility into the traffic egressing Google Cloud allowing you to monitor per traffic class metrics and diagnose any issues. This visibility, along with the detailed flow-level, traffic-class information from the VPC Flow Logs, can also help you finetune the policies as needed.

Deployment best practices and recommendations

- It is important to follow all the recommended best practices for Cloud Interconnect while leveraging application awareness on Cloud Interconnect. In particular, the best practices for Cloud Interconnect capacity provisioning.
- Since the predefined mapping covers the full DSCP range, all traffic maps to these traffic classes – any unmarked traffic (DSCP bits 000) will map to TC1.
- For application traffic profiles where the high priority application traffic is clearly identifiable, and these applications have a consistent and non-bursty bandwidth utilization profile, it is suggested to use the strict priority policy (along with shaper controls to avoid starvation).

- For application traffic profiles where the priority amongst applications is not clearly identifiable or predictable, it is suggested to use the bandwidth percentage policy.
- For application traffic profiles where the high priority application traffic is identifiable, and some (or all) of these applications are bursty, it is suggested to use the bandwidth percentage policy.
- When using the bandwidth percentage policy, all six traffic classes must be allocated a percentage of bandwidth and the configured bandwidth percentages for the six traffic classes must add up to 100.
- Ensure that your network control traffic, such as BGP and BFD, is either marked with the highest priority class (if using the strict priority policy) or has appropriate bandwidth percent configuration (if using the bandwidth percentage policy).
- Leverage VPC Flow Logs to analyze Cloud Interconnect traffic and fine tune your application awareness policies: VPC Flow Logs provide detailed, flow-level insights into your Cloud Interconnect traffic, including 5-tuple (src/dst IP, src/dst port, protocol) and traffic-class information. These granular insights, including bandwidth utilization, can help you finetune your application awareness policy configurations to optimize performance, availability and costs. You can also monitor and verify that your applications are marking DSCP correctly. Flow Analyzer (available at no additional charge) can help quickly analyze the logs, identify top talkers and drill down into traffic patterns by traffic class or specific DSCP values.

Summary

Application awareness on Cloud Interconnect empowers customers to address the critical challenge of traffic prioritization when deploying business-critical workloads in hybrid or multicloud environments. It enables organizations to ensure consistent performance and reliability for their critical applications, even during peak traffic periods and congestion events on their Cloud Interconnect - allowing businesses to build distributed applications while consuming best-of-breed SaaS and AI/ML services. Application awareness on Cloud Interconnect is a managed solution that can ensure efficient utilization of your Cloud Interconnect bandwidth and reduce your operational overhead and costs with six traffic classes, a choice of strict priority and bandwidth percentage queuing policies and optional shaper control.