# Zero Trust Security for GCP Workloads with Zscaler Cloud Connector

Reference Architecture — Zscaler for Users

# Contents

# About Zscaler Reference Architectures Guides

The Zscaler™ Reference Architecture series delivers best practices based on real-world deployments. The recommendations in this series were developed by Zscaler's transformation experts from across the company.

Each guide steers you through the architecture process and provides technical deep dives into specific platform functionality and integrations.

The Zscaler Reference Architecture series is designed to be modular. Each guide shows you how to configure a different aspect of the platform. You can use only the guides that you need to meet your specific policy goals.

## Who Is This Guide For?

The Overview portion of this guide is suitable for all audiences. It provides a brief refresher on the platform features and integrations being covered. A summary of the design follows, along with a consolidated summary of recommendations.

The rest of the document is written with a technical reader in mind, covering detailed information on the recommendations and the architecture process. For configuration steps, we provide links to the appropriate Zscaler Help site articles or configuration steps on integration partner sites.

## A Note for Federal Cloud Customers

This series assumes you are a Zscaler public cloud customer. If you are a Federal Cloud user, please check with your Zscaler Account team on feature availability and configuration requirements.

## Conventions Used in This Guide

The product name ZIA Service Edge is used as a reference to the following Zscaler products: ZIA Public Service Edge, ZIA Private Service Edge, and ZIA Virtual Service Edge. Any reference to ZIA Service Edge means that the features and functions being discussed are applicable to all three products. Similarly, ZPA Service Edge is used to represent ZPA Public Service Edge and ZPA Private Service Edge where the discussion applies to both products.

Notes call out important information that you need to complete your design and implementation.

Warnings indicate that a configuration could be risky. Read the warnings carefully and exercise caution before making your configuration changes.

## Finding Out More

You can find our guides on the **Zscaler website** (**https://www.zscaler.com/resources/reference-architectures**).

You can join our user and partner community and get answers to your questions in the **Zenith Community** (**https://community.zscaler.com**).

## Terms and Acronyms Used in This Guide

| Acronym | Definition |
|---|---|
| C2 | Command & Control |
| DC | Data Center |
| DNS | Domain Name System |
| DoH | DNS over HTTPS |
| FQDN | Fully Qualified Domain Name |
| GCP | Google Cloud Platform |
| ICMP | Internet Control Message Protocol |
| IoT | Internet of Things |
| IP | Internet Protocol |
| NAT | Network Address Translation |
| SIPA | Source IP Anchoring |
| SSL | Secure Socket Layer (superseded by TLS) |
| TCP | Transmission Control Protocol |
| TLS | Transport Layer Security |
| UDP | User Datagram Protocol |
| URL | Universal Resource Locator |
| VPC | Virtual Private Cloud |
| ZDX | Zscaler Digital Experience |
| ZIA | Zscaler Internet Access |
| ZPA | Zscaler Private Access |
| ZTE | Zero Trust Exchange |

## Icons Used in This Guide

The following icons are used in the diagrams contained in this guide.

Zscaler Zero Trust Exchange

ZIA or ZPA Service Edge

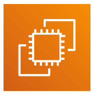
Zscaler Central Authority

Zscaler App Connector

Zscaler Cloud Connector

Google Cloud

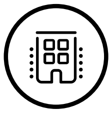
AWS Cloud



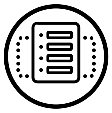
AWS Application or Workload

Azure Cloud

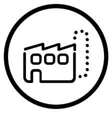
Azure Application or Workload



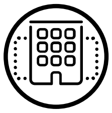
Branch Office Location



Data Center

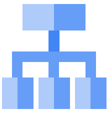


Factory Location

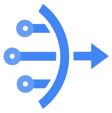


Headquarters Location

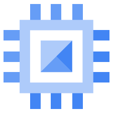
Internet



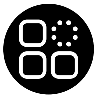
Cloud Load Balancing



Cloud NAT



Compute Engine

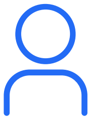


Generic Application or Workload

Data Tunnel

Positive / True Badge

Negative / False Badge



Authorized User

Bad Actor

Threat Actor

# Introduction

The shift to cloud services has rebuilt the enterprise data center off-premises and outside of traditional security boundaries. Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) enable organizations to quickly build out and scale their platforms and services. Securing services across multiple clouds, vendors, and support features requires a different approach than that of the traditional data center.

Securing this communication through the layering of legacy Access Control Lists (ACL), on-premises firewalls, and service-chaining has always been both complicated to build and difficult to maintain. Private applications were accessed via virtual private networks (VPNs) that extended the network to locations in an any-to-any access model. This large, flat network gave users a single location to connect to for access to private applications.

Leveraging the cloud breaks these models. You now have multiple vendors across different clouds, products, and services. Your policy must be interpreted at each cloud and application to determine how best to implement it with the tools available. This risk goes up given a mistake, potentially exposing your organization to a host of network-born attack vectors.

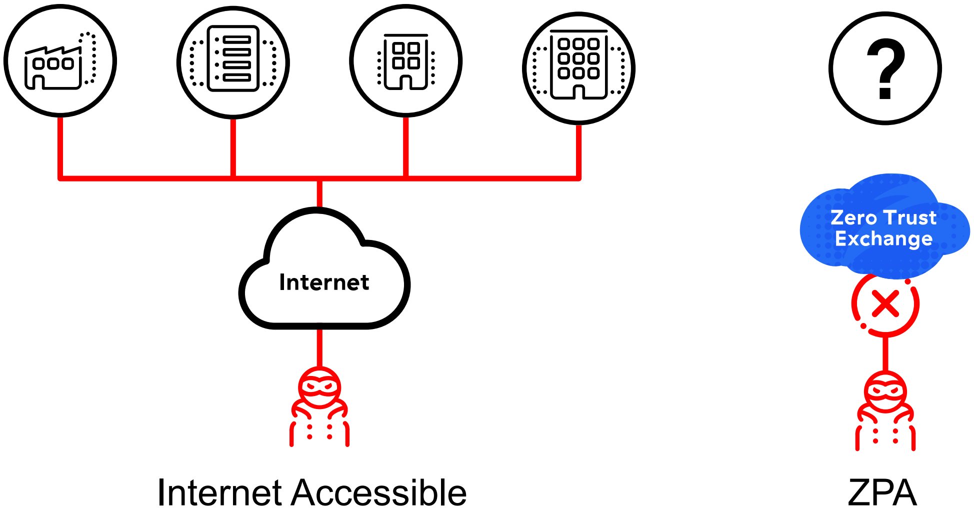


*Figure 1. ZPA makes your applications invisible outside of your organization*

Ideally, an organization's security policy should be at the foundation of its network design. Connectivity to and from devices happens as a product of the security policy and not the other way around. This is the heart of the Zscaler Zero Trust Exchange (ZTE) model. Users must be authorized before they can connect to that service. Even knowing the application's hostname and the services it provides won't give the attacker any information, as that service won't resolve until the user authenticates. Your applications are effectively hidden from the internet and each other until you define policy to allow access.

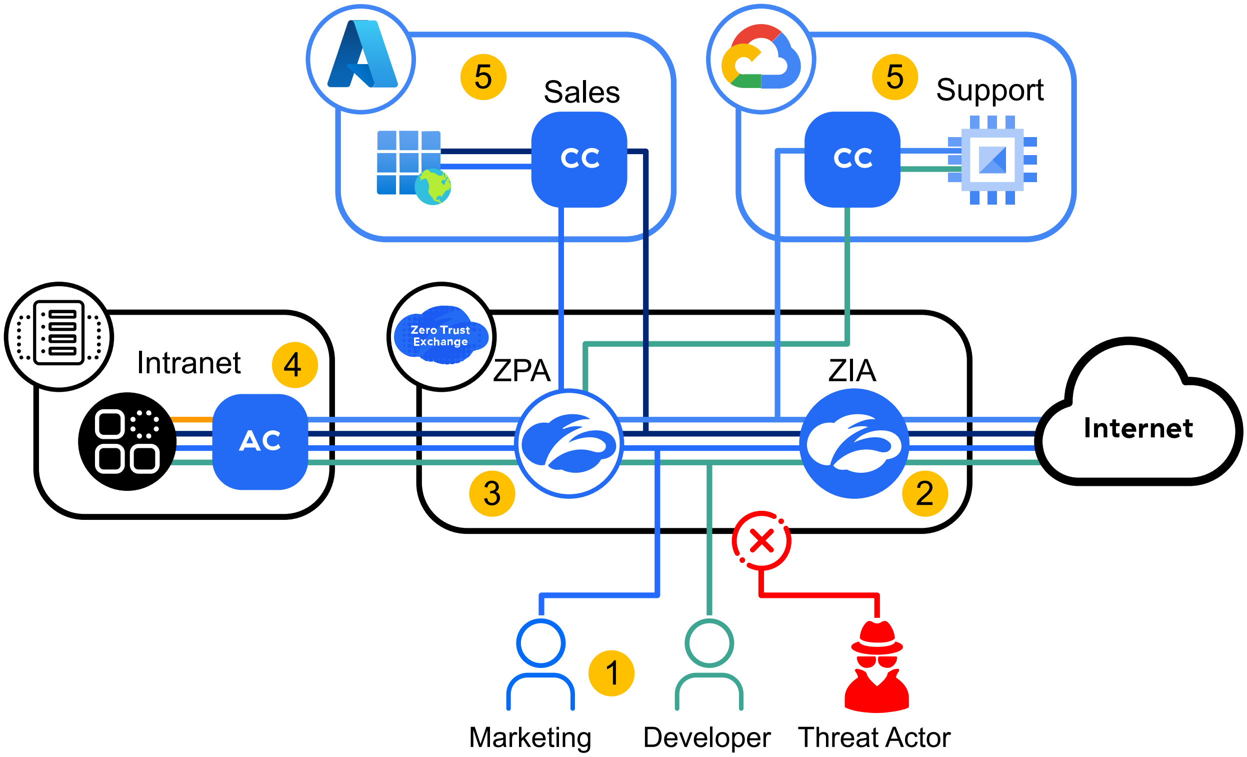


*Figure 2. Zero Trust principles applied to users and workloads*

1. Authentication – All users must first authenticate to Zscaler. Based on multiple criteria such as user group membership, device posture, and location, the user is assigned a set of policies. These include the ability to see internal applications.
2. ZIA Service Edge – When traffic from users or workloads needs to be routed to the internet, a ZIA Service Edge inspects the traffic. If your policy allows the traffic out, the return traffic is also scanned for malicious content on its way back to the user.
3. ZPA Service Edge – Traffic from users or workloads bound for other internal applications is handled by ZPA. Based on the user's authentication and assigned policy, only approved resources are resolved. All other resources are hidden from unauthorized users as if the services do not exist.
4. App Connector – Sitting in front of internal applications, App Connector allows ZPA connections to applications for authorized users.
5. Cloud Connector – Deployed in front of your internal applications, Cloud Connector creates a set of outbound tunnels to ZIA and ZPA. They decide where the tunnel connects based on policy.

In the previous model, your developers have workloads in Google Cloud Platform (GCP), and your users in marketing have workloads running in Microsoft Azure. All users have access to internal applications in the data center, as well as general internet access. In this case, each user and workload are limited to which applications they can resolve and access, based on the policy applied to them.

Your marketing user (blue) can access their workloads in Azure, the data center, and internet based on policy. Your developer (green) can access their workloads in AWS, the data center, and the internet. Finally, your workloads in GCP, Azure, or your data center can all reach one another and the internet via the ZTE, without the need to set up additional VPN links. All these connections are subject to the policy you set.

Cloud Connector ensures that cloud workloads adhere to organizational security policy when accessing both public and private endpoints. This is achieved by intelligently forwarding traffic to the Zscaler Internet Access (ZIA) and Zscaler Private Access (ZPA) platforms. Cloud Connector also enables multi-cloud connectivity and enforces a security policy for cloud-to-cloud traffic.

## Key Features and Benefits

- Reduce complexity by connecting directly to the internet and eliminate the need for complex routing configurations through SNAT, transit gateways, and transit hubs.
- Total visibility, control, and awareness for workload communications. Centralized logging and real-time streaming can also be used with third-party monitoring solutions.
- Flexible scalability with elastic, horizontal scaling made possible through the Zero Trust Exchange architecture, which operates in over 150 global data centers.
- High availability with built-in automatic failover with N+2 redundancy is provided for forwarding and security. Additional redundancy can be built into the Google deployment via Auto Scaling groups and warm pools.
- Lower operational costs by removing expenses associated with complex network configurations, network service replication, and hidden costs for cloud connectivity.

## New to Zscaler Cloud Connector or Google Cloud Platform?

If this is your first time reading about Zscaler Cloud Connector or GCP, we encourage you to explore the following resources:

- If you are new to GCP, Google offers a range of courses for different roles. To learn more about GCP courses, refer to **Google Cloud Training and Certification** (**https://cloud.google.com/learn/training**).
- To watch a demonstration of Zscaler Cloud Connector, refer to **Zero Trust Your Cloud Workloads** (**https://www.youtube.com/watch?v=S-g_qmuxnqU&t=1845s**).
- ZIA provides outbound internet protection for users. Learn more at **Zscaler Internet Access** (**https://www.zscaler.com/products/zscaler-internet-access**).
- ZPA provides private access to applications, not networks. Learn more at **Zscaler Private Access** (**https://www.zscaler.com/products/zscaler-private-access**).
- To learn more about Zero Trust, see **It Starts With Zero** (**https://www.zscaler.com/it-starts-with-zero**).
- To learn more about the Zero Trust architecture, refer to the National Institute of Standards and Technology (NIST) paper **Zero Trust Architecture** (**https://www.nist.gov/publications/zero-trust-architecture**).

# Cloud Infrastructure Protection Using Cloud Connector

As organizations began moving workloads to the cloud, securing those resources was always a challenge. Securing applications against attack both from outside actors and malicious content on legitimate sites led some organizations to provide access to applications over VPN. Attempting to leverage legacy security products added latency and frustration for users. As cloud usage expanded, data now moved between the cloud and the user, between systems in the cloud vendor, and between applications across cloud vendors and your legacy data center.

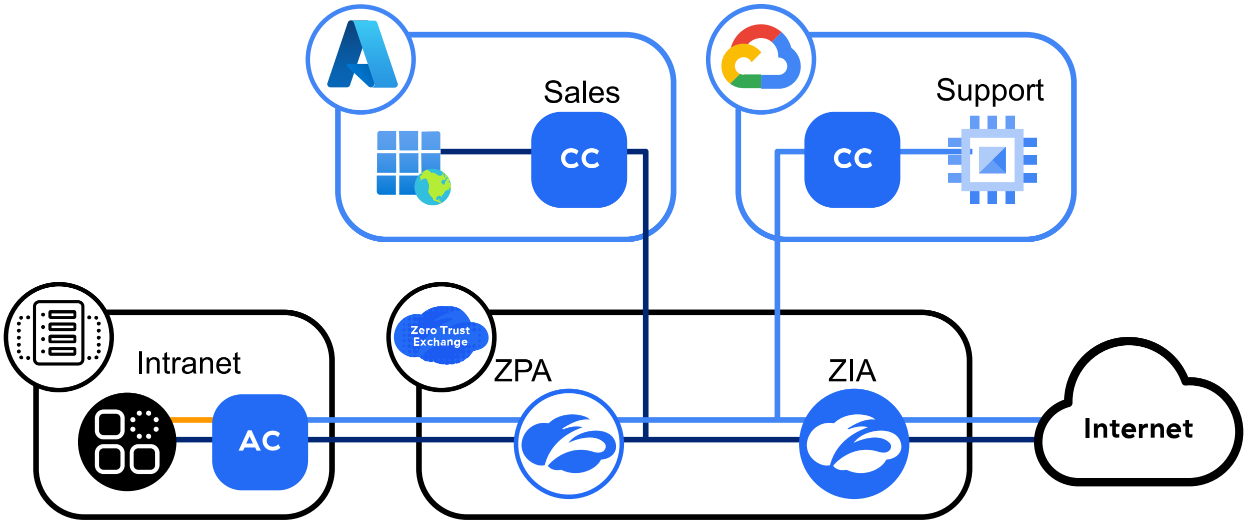


*Figure 3. Workload communication between private and public applications*

The communication can be from private workloads (IaaS or physical DC) to public workloads (SaaS internet application), or between private workloads (IaaS to IaaS, or physical DC to IaaS). Securing these communications channels with physical or virtual appliances is cumbersome and can lead to inconsistent configuration.

In the previous example, our application *sales.azure.internal.safemarch.com* sits behind a Cloud Connector with access to both ZPA and ZIA platforms. In this model, the workload can reach out to the support workloads in AWS, allowing the sales team to file support and product requests without logging into the support portal. The sales portal is accessed by our intranet workload in our data center to pull deals and rankings for the company dashboard. Finally, our sales workload can reach the internet to update our cloud CRM, which in turn only accepts connections from Zscaler IP addresses for our tenant.

Zscaler Cloud Connector virtual machines extend the security of ZIA and ZPA to cloud native workloads. ZIA protects your workload traffic communicating with a public application. ZPA protects your communications between private workloads. This allows organizations to secure all workload communications over any network. The Zscaler Zero Trust Exchange allows workloads to communicate with each other with a granular security policy applied.

- Applications-to-Internet Communications for applications that might need to access any internet or SaaS destination, such as third-party APIs, software updates, etc. A scalable, reliable security solution that inspects all transactions and applies advanced threat prevention and data loss protection controls.
- Application-to-Application Communication to other public clouds and corporate data centers for multi/hybrid cloud connectivity. Delivered with better security and a dramatically simplified operational model, as compared with traditional solutions like proxies, virtual firewalls, and IDS/IPS.
- Application-to-Application Communications within a Virtual Private Cloud by securing process-to-process communications. This achieves microsegmentation of traffic with no changes to the application or the network.

Cloud Connector is delivered in several form factors. It is available as a virtual appliance on Google Cloud Platform, Amazon Web Services, and Microsoft Azure, as well as VMs for on-premises deployment.

If you are deploying on Google GCP:

- Zscaler recommends the n2-standard-2 instance type for deploying Cloud Connector as it offers the best mix of performance and cost. To view available instance types, refer to the **Google Cloud documentation** (**https://cloud.google.com/compute/docs/machine-resource**).

For on-premises deployments, the image requires:

- VMware ESXi and CentOS/Linux (KVM) images
- 2 virtual CPUs
- 4 GB of RAM

## Deploying Cloud Connector VMs via Scripts

Zscaler recommends deploying Cloud Connector instances by leveraging Terraform scripts. The Zscaler Terraform scripts provide complete end-to-end automation for deploying Cloud Connector instances and supporting network components. Terraform's goal is to be as "hands-off" as possible by automatically configuring items without user intervention. For a detailed look at deploying with Terraform scripts, see **Deploying Cloud Connector via Terraform Scripts**.

## High Availability Deployment Design

Cloud Connector leverages GCP's internal passthrough Network Load Balancer functionality to achieve high availability and horizontal scalability. In this model, inbound traffic from your workloads are directed to the IP address of the internal passthrough Network Load Balancer. When traffic returns from the internet, the Cloud Connector appliance strips off Datagram Transport Layer Security (DTLS) encapsulation and forwards the traffic back to the originating workload.

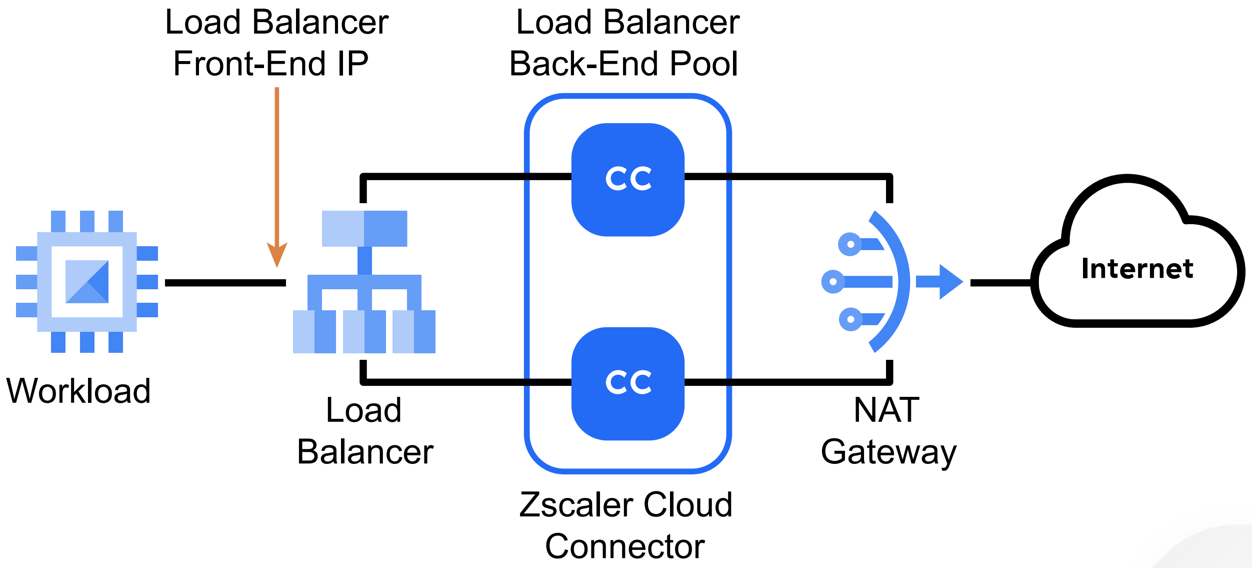


*Figure 4. Cloud Connectors receive outbound traffic from the load balancer*

Zscaler recommends a minimum of two Cloud Connector appliances, each in a different GCP zone. Workloads within those same GCP zones should then leverage their respective Cloud Connector appliances. If a Cloud Connector appliance fails, load balancer functionality automatically redirects traffic to the active appliance in the adjacent GCP zone.

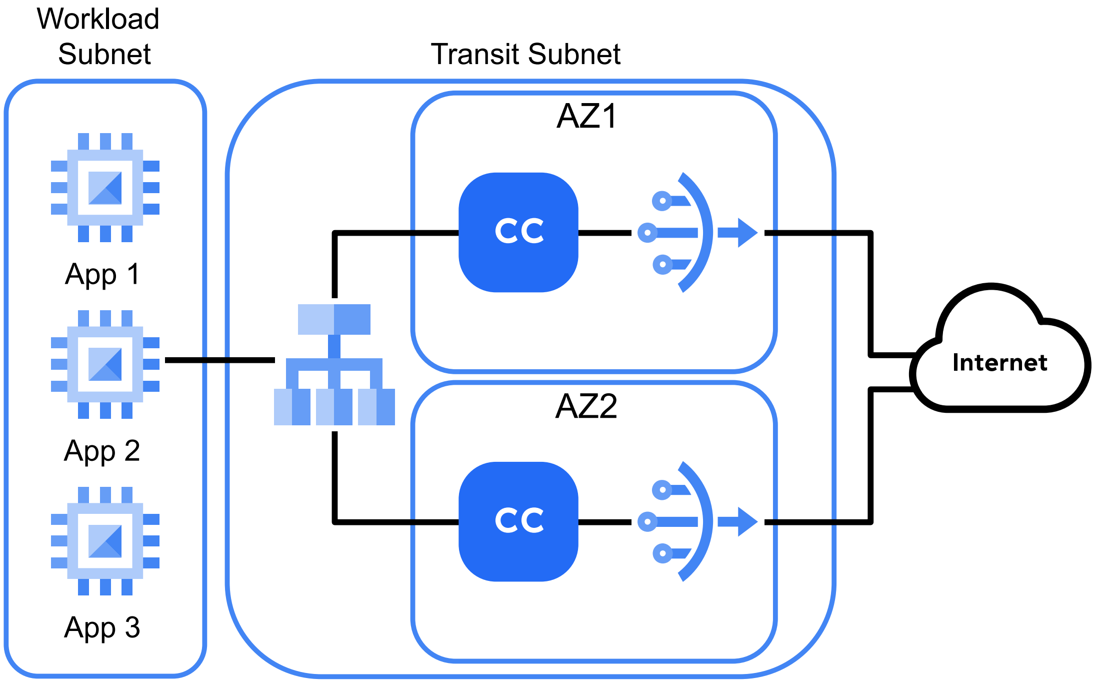


*Figure 5. Cloud Connectors deployed in redundant pairs across GCP zones*

As part of an effective High Availability and horizontal scaling strategy, GCP's internal passthrough network load balancer can be employed to allow organizations to automatically distribute traffic across multiple cloud connector instances.

The load balancer works by accepting traffic using a forwarding rule that contains a workload-facing IP address and distributing received traffic to a pool of resources. In the context of workload communications with Cloud Connector, cloud workloads that wish to send traffic can simply direct their traffic towards a single IP Address in the VPC at the Front-end of the Load Balancer. The Load Balancer then selects a healthy Cloud Connector resource from the back-end pool and forwards the traffic accordingly.

This functionality uses HTTP Ping to monitor the health of the Cloud Connector appliances. The distribution algorithm used by default is 2-tuple using source and destination IP Address, though this is configurable. The way that an internal passthrough Network Load Balancer distributes new connections depends on whether you have configured failover.

With failover enabled, the ILB continues to forward to healthy instances during normal operation. However, an administrator can opt to forward to primary instances if all instances are unhealthy (or simply drop the traffic). If failover is not enabled, the ILB will forward connections to healthy instances as part of normal operation. If no healthy instances are available, the ILB will resort to forwarding blindly to unhealthy instances as a last resort.

The Load Balancer functionality requires the Cloud Connector appliance to spawn an HTTP service for probing. This must happen during the initial build of the appliance. When building an appliance, ensure you specify an HTTP Probing Port in the Terraform script to ensure this service spawns correctly. Similarly, ensure that any Firewall Rules configured will allow access to this port.

Terraform scripts can be used to automate deployments, or a script can be built manually. Zscaler provides a Terraform template for your use at About Cloud Automation Scripts (https://help.zscaler.com/cloud-connector/about-cloud-automation-scripts).

For more information, refer to the Google Cloud documentation (https://cloud.google.com/load-balancing/docs/internal/).

## Scalability of Cloud Connector Instances

Cloud Connector supports two methods of scaling: vertical and horizontal. With vertical scaling, the Cloud Connector can be deployed with a higher footprint of vCPU and RAM. However, throughput and connection capacity scales linearly with additional resources. You will eventually hit the maximum throughput limit for the appliance. Additionally, the failure of a larger appliance requires more connections to fail over to the backup solution, which faces the same throughput restrictions.

The following image is simplified for clarity. Redundant instances of Zscaler Cloud Connector should be deployed in all instances.

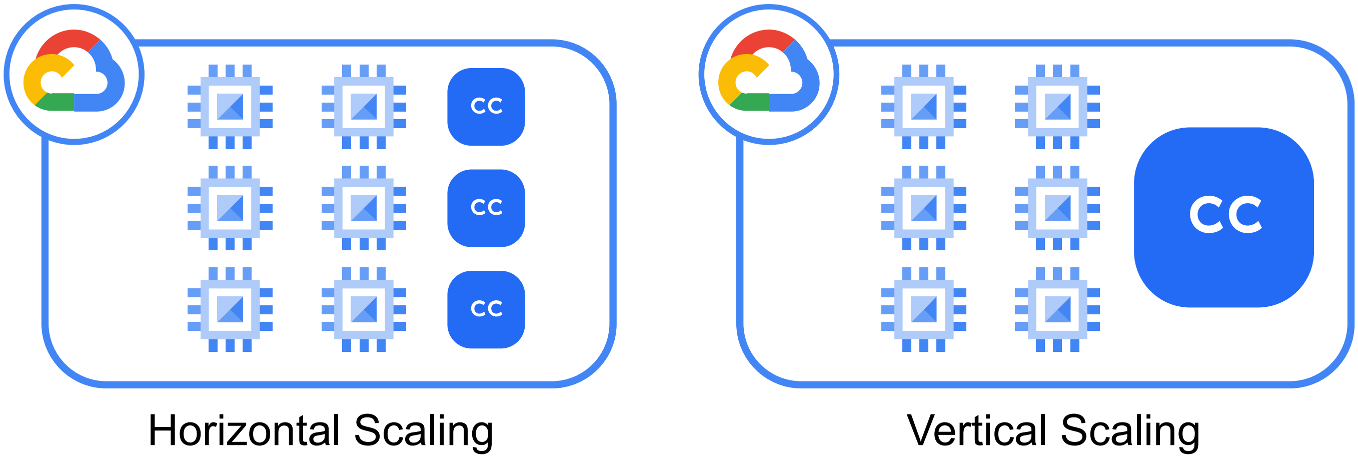


*Figure 6. Horizontal and vertical scaling of Cloud Connectors*

Cloud Connector can also be scaled horizontally, wherein multiple appliances are deployed within multiple GCP zones around a region. Inbound traffic to the Cloud Connector appliance can then be load-balanced across all available paths. Either or both methods are supported when considering current and future throughput requirements. Typically, horizontal scaling is more scalable and fault-tolerant, and avoids any cloud provider platform limits.

## Cloud Connector Logging and Service Dashboards

Cloud Connector can use built-in logging functionality through the Insights page of the portal. Zscaler streams all logs to centralized log locations, allowing you to view logs from across your organization. The dashboard has views for Session Insights, DNS Insights, and ZIA Tunnel Insights. All three facilities allow you to review traffic that passes through the Cloud Connector appliance from a different perspective.

- Learn more about Analyzing Traffic Using Insights (https://help.zscaler.com/cloud-connector/analyzing-traffic-using-insights).
- Learn more about ZIA dashboards at About Dashboards (https://help.zscaler.com/zia/about-dashboards).
- Learn more about ZPA dashboards at Dashboard and Diagnostics (https://help.zscaler.com/zpa/dashboard-diagnostics).

Cloud Connector supports both the Nanolog Streaming Service (NSS) for ZIA use cases and Log Streaming Service (LSS) for ZPA use cases. NSS uses a virtual machine (VM) to stream traffic logs in real time to your Security Information and Event Management (SIEM) system, such as Splunk or ArcSight. LSS operates in a similar way, with the deployment of a ZPA App Connector VM that receives the log stream and then forwards it to the log receiver.

Both services enable real-time alerting and correlation of logs with your other devices. NSS and LSS can be configured from the Cloud & Branch Connector Admin Portal.

NSS and LSS require separate subscriptions for each virtual machine.

- Learn more about Nanolog Streaming Service (https://help.zscaler.com/zia/about-nanolog-streaming-service).
- Learn more about Log Streaming Service (https://help.zscaler.com/zpa/about-log-streaming-service).

## Upgrading Your Cloud Connectors

Cloud Connector runs the Zscaler OS in the virtual machine. Software updates and OS updates are provided by Zscaler via automatic upgrades. When a Cloud Connector is deployed, the software is automatically upgraded to the latest version. Cloud Connector instances check for new software daily. If a new version is available, the Cloud Connector will upgrade itself automatically at midnight local time, based on the deployed cloud region.

This automatic check and update means it is critical that your Cloud Connector locations are accurate. An inaccurate location can lead to upgrades occurring in the middle of the day. Always specify exactly where the Cloud Connector is located when deploying the virtual machine.

As a matter of redundancy during upgrades, Cloud Connector is installed in pairs within a GCP zone. Multiple pairs of Cloud Connectors should be instantiated within different GCP zones, thereby minimizing the impact of service upgrades or infrastructure failures.

Cloud Connector is based on Zscaler OS, and therefore the software updates and OS updates are provided and automatically applied by Zscaler. When a Cloud Connector is deployed, the software is automatically updated to the latest version. A Cloud Connector then checks for new software daily and upgrades itself automatically at midnight (local time, based on the deployed cloud region).

You can configure this upgrade window from the Cloud & Branch Connector Admin Portal. As mentioned throughout this document, Zscaler recommends that Cloud Connector appliances be deployed as redundant, high-availability instances. Specifically, we recommend deploying two appliances per GCP zone with a minimum regional cluster size of four (two in GCP zone 1 and two in GCP zone 2). The Zscaler software upgrade process will upgrade one instance of a pair at a time, providing availability for the GCP zone from the remaining instance.

Zscaler recommends that Cloud Connector appliances be deployed as redundant, high-availability appliances. Specific to software upgrades performed by Zscaler, this ensures that you incur no downtime. When an appliance is rebooted to accept a new update, Google load balancer automatically moves traffic over to the redundant, active appliance.

Although cloud IaaS providers such as GCP are responsible for ensuring the security and availability of their infrastructure, organizations are ultimately still responsible for the security of their workloads, applications, and data. To learn more about the shared responsibility model, refer to the Microsoft documentation (https://docs.microsoft.com/en-us/azure/security/fundamentals/shared-responsibility).

# Deployment and Design Options

The following section outlines your options when deploying Cloud Connector. You can design your network using the tools that best match your cloud deployment. We recommend that you review each use case to familiarize yourself with the various options, which can be combined to meet your organization's deployment needs.

## Pre-Deployment Considerations

The following sections provide some general design recommendations common to all deployment types.

### Cloud NAT vs. Public IP Addressing

Zscaler recommends that the Cloud Connector appliances leverage the Cloud NAT functionality of GCP for outbound internet access, as opposed to assigning public IP addresses to each appliance:

- Cloud NAT can provide outbound internet access for multiple Cloud Connectors using a single public IP address.
- Cloud NAT utilizes a single public IP address as opposed to the expense of purchasing a public IP address for each instance.
- Cloud NAT is stateful in its operation. Traffic initiated from the "inside" zone towards the "outside" public internet is allowed, as well as the corresponding return traffic. Inbound traffic is dropped, preventing attackers from directly targeting the Cloud Connector appliance from the outside. Public IP addresses assigned directly to the Cloud Connector, by contrast, allow bidirectional communication and expose these hosts to the public internet from the outside.

In cases where public IP addresses assigned directly to the Cloud Connector are the only option, Zscaler recommends that GCP firewall rules be adjusted to drop inbound traffic to the Cloud Connector appliances and reduce the attack vector as much as possible.

For an up-to-date listing of Zscaler public IP ranges, ports, and protocols in use, see **Zscaler Config** (**https://config.zscaler.com/zscaler.net/cenr**).

## Cloud Connectors and GCP Zones

GCP regions are clusters of physical data centers strategically positioned within metropolitan areas around the globe. GCP zones are simply smaller subsections of regions and generally refer to the individual facilities hosting the hardware. A zone can consist of one or more data centers that interconnect with each other via high-speed, low-latency links. The network performance is sufficient to accomplish synchronous replication between zones, making high availability easy. These data centers have separate power, cooling, and redundancy mechanisms in place such that a failure of one zone has no impact on other zones within the same region. Hence, it is best practice to distribute cloud workloads and their redundant counterparts across zones within a region.

## Network Connectivity

Cloud Connector appliance has a service interface where workload traffic is brought in and where DTLS tunnels are terminated towards the Zscaler cloud. Both interfaces are associated with a service and management virtual private cloud (VPC), respectively. When created, VPCs are associated with a region and transparently stretch across zones. Because of this, VPCs don't need to be created for each zone as you would in other IaaS providers.

## Virtual Compute

GCP allows the administrator to select the zone for an instance as part of the instance creation process. This ensures that individual Cloud Connector appliances exist on physically separate pieces of underlying hardware from one another. As such, when building high-availability pairs of Cloud Connector appliances, Zscaler recommends that each appliance be instantiated within different zones. The ideal deployment is two appliances per zone.

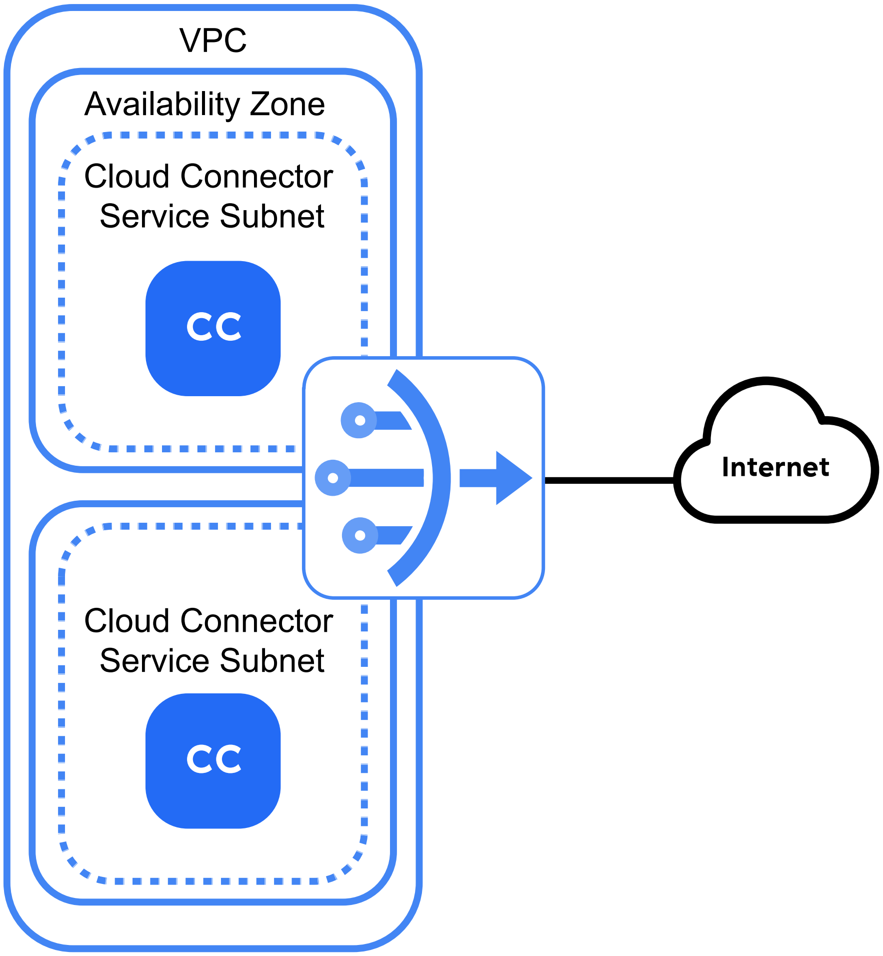

VPC
Availability Zone
Cloud Connector Service Subnet
CC
Cloud Connector Service Subnet
CC
Internet

*Figure 7. Cloud Connector appliances in two different GCP zones*

## Deploying Cloud Connector via Terraform Scripts

Zscaler Terraform scripts provide complete end-to-end automation to not only deploy Cloud Connector appliances, but all the secondary and tertiary components as well in a repeatable and predictable way. The scripts are divided into two categories:

- Greenfield – Blank slate deployments
- Brownfield – A GCP deployment already exists

It is important to note that Terraform does not modify brownfield deployments. When executing Terraform scripts, new VNets, route tables, subnets, and VM instances are spawned to support the current workflow. It is your responsibility to integrate the new deployment into your existing environment. This can mean that the new Cloud Connector VNet is peered with existing VNets, or that new workloads are installed within the Cloud Connector VNet. Bear this in mind when considering whether Terraform is the correct option to use when integrating with your brownfield environment.

To download the Zscaler Terraform scripts, refer to **Zscaler GitHub** (**https://github.com/zscaler**) to download the latest copies and be notified when updates occur. A brief overview of the available scripts is included here. Always check GitHub for the latest updates and documentation.

### Greenfield

- Base 1CC (Greenfield) – This deployment type is intended for greenfield deployments such as new production or lab networks. It deploys a fully functioning sandbox environment in a new management and service VPC with a test workload VM and bastion host.
- Base 1CC with ZPA (Greenfield) – This deployment type is intended for greenfield deployments such as new production or lab networks. It deploys a fully functioning sandbox environment in a new management and service VPC with a test workload VM and bastion host. This deployment type adds support for ZPA by instantiating Google Cloud DNS to add DNS redirection to cloud workloads. In addition to the resources created in the Base 1CC method, this option also creates a Google Cloud DNS forward zone intended for ZPA app segment DNS redirection.

For inbound access into GCP workloads, App Connector must also be installed as part of a separate workflow.

- Base CC with ILB (Greenfield) – This deployment type is intended for greenfield deployments such as new production or lab networks. It deploys a fully functioning sandbox environment in a new management and service VPC with a test workload VM and bastion host. This is a high-availability deployment method using an internal passthrough Network Load Balancer.

### Brownfield

- CC with ILB (Brownfield) – This deployment type is intended for use in existing production deployments. All network infrastructure resources have conditional “byo” variables that can be input if they already exist. This can include VPC, Subnet, Cloud Router, and Cloud NAT.

## Directing Traffic to Cloud Connector

Cloud Connector acts as a gateway to cloud workloads. Directing traffic through the Cloud Connector is as simple as modifying the default gateway route of the workload VPC route table to point to the appliance, or to the internal passthrough Network Load Balancer. In most circumstances, this ensures traffic to ZIA and ZPA are steered correctly. For ZIA, internet-bound traffic is pointed to the nearest ZIA Service Edge. DNS traffic that requires modification for ZPA use cases where redirection to an App Connector is necessary is also forwarded appropriately to the Zscaler DNS service.

For example, with a single instance of Cloud Connector, the workload route table can be updated with a default route using the IP address of the service interface of the Cloud Connector appliance as the default gateway. The Cloud Connector appliance uses the service VPC and route table created during the deployment process. The default route for the Cloud Connector's route table should point towards the NAT Gateway, also created in the deployment process. A public subnet and route table should have also been created during the deployment process and reference the corresponding internet gateway with its default route.

The following image is simplified for clarity. Redundant instances of Zscaler Cloud Connector should be deployed in all instances.

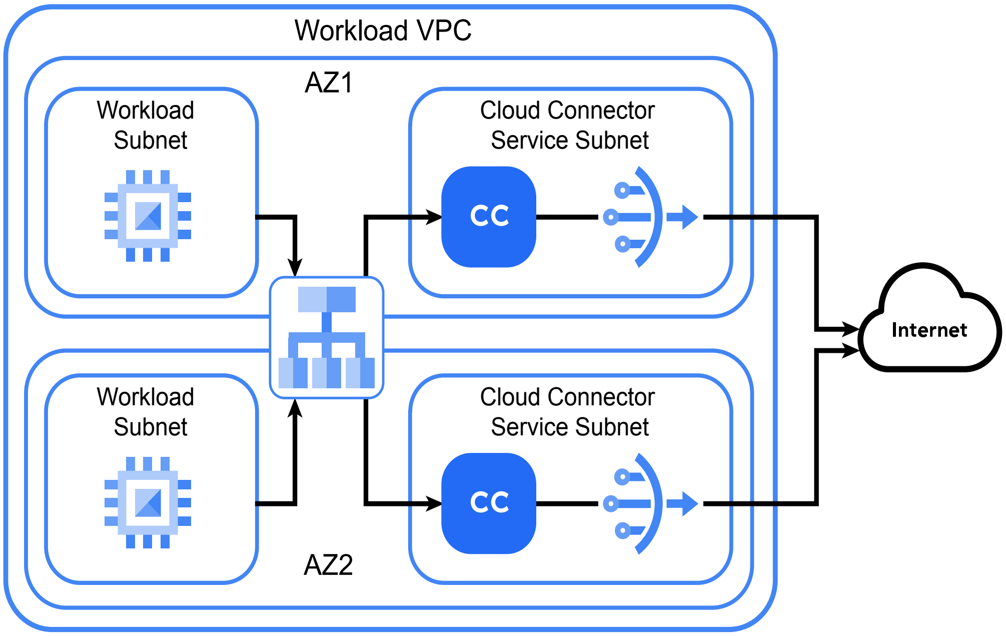

Workload VPC
AZ1
Workload Subnet
Cloud Connector Service Subnet
CC
Internet
Workload Subnet
Cloud Connector Service Subnet
CC
AZ2

*Figure 8. Default routes for workloads go through the Cloud Connector*

In the case of hub and spoke, wherein the Cloud Connector service VPC is the "hub" and workload VPCs are the "spokes," the Cloud Connector service VPC should be peered with all workload VPCs. A default route in the service VPC route table of the Cloud Connector appliance directs traffic towards the NAT Gateway by default. In the workload VPC, a default route is present to direct traffic across the VPC peering towards the Cloud Connector appliance. As with all network traffic, ensure you have routing set up as well so that returning traffic from the internet is correctly directed back towards the initiating host.

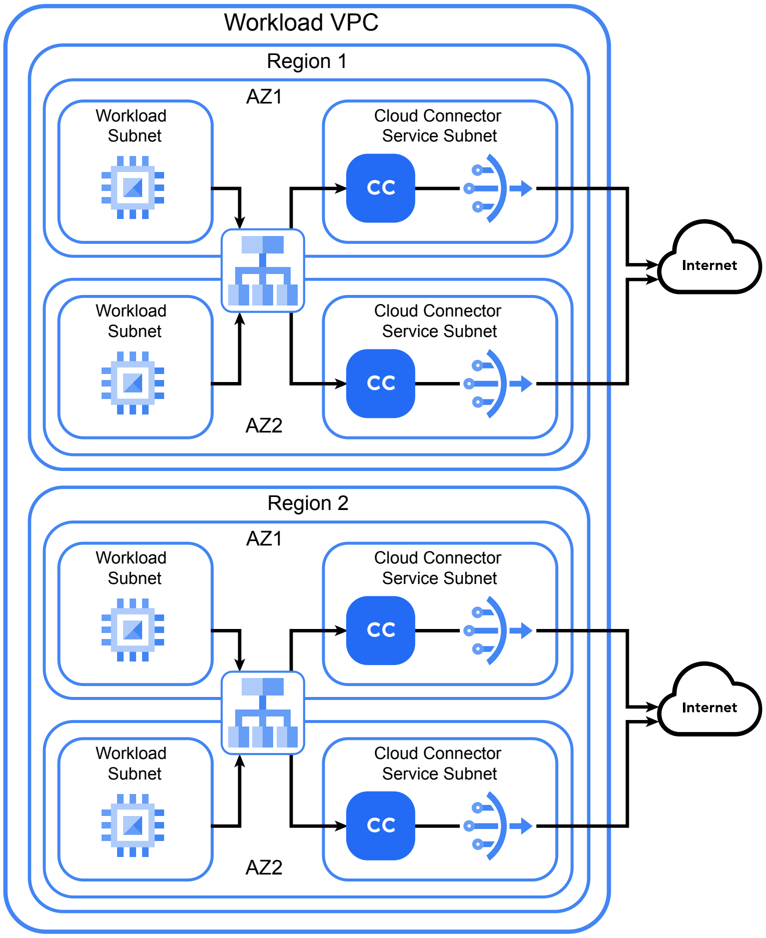


*Figure 9. Transit VNets provide access to one or more private VPCs*

This approach works for both single region deployments and a shared multi-region model wherein the Cloud Connector service VPC is peered with the workload VPC. However in the Shared VPC model, the network tagging is inverted to achieve regional redundancy. This is because tagged routes are not exported over peering connections; only the non-tagged default route is exported. Additionally, it should be noted that regional routing is not possible in a shared multi-VPC model.

## Forwarding Options

When traffic has reached the Cloud Connector, there are four Traffic Forwarding options available to direct traffic out of the Google cloud:

- **Direct** – Traffic matching the criteria defined bypasses the Cloud Connector and is routed out of the service interface, where it follows Google route tables towards the destination.
- **Zscaler Internet Access (ZIA)** – Traffic matching the criteria defined is forwarded to the ZIA cloud for inspection.
- **Zscaler Private Access (ZPA)** – Traffic matching the criteria defined is forwarded to the ZPA cloud for inspection.
- **Drop** – Traffic matching the criteria is dropped by the Cloud Connector.

Each of the four options permits the administrator to define a range of match criteria. In general, macro forwarding logic can be defined within the Cloud & Branch Connector Admin Portal, whereas ZIA or ZPA can perform more granular inspection.

Traffic Forwarding policy is in the Policy Management section of the Cloud & Branch Connector Admin Portal. Rule creation and assessment models ZIA and ZPA workflows. More specific rules should be ordered near the top, while more broad rules ordered towards the bottom. Match criteria is as follows:

## General

- **Location** – Locations identify the various VPCs from which your workloads send traffic. As Cloud Connector appliances are brought online, the VPC they are installed within automatically populates this menu. It should be noted, however, that in a Transit/Egress VPC scenario, downstream VPCs do not automatically populate. In such a case, you must use Source or Destination FQDN (recommended) or IP as match criteria. In ZIA, if the traffic is from a known location, the service processes the traffic based on the location settings. For example, the service checks whether the location has authentication enabled and proceeds accordingly. It also applies any location policies that you configure and logs internet activity by location.
- **Location Group** – If necessary, location groups can be created to organize various cloud VPCs, such as a “Dev VPCs” location group, “Prod VPCs” location group, etc. If there are many locations and associated sub-locations within your organization, consider using location groups.
- **Branch and Cloud Connector Groups** – Branch and Cloud Connector groups allow you to match traffic transiting specific Cloud Connector appliances.

## Source

- **Source IP Groups** – When multiple source IP addresses must be matched across multiple policy rules, it is operationally more efficient to create source IP groups. These groups allow you to organize IP addresses for easier rule creation and visualization.
- **Source IP Addresses** – This match criteria allows you to specify the source IP address of the workload.

## Destination

- **Destination FQDN/IP** – For individual FQDN (recommended) or IP address matching, enter the value you want to be matched in this field.
- **Destination FQDN/IP Group** – You can group together destination FQDNs (recommended) and the IP address that you want to control in a Forwarding Policy rule by specifying FQDN, IP addresses, countries where servers are located, and URL categories.

Wildcard domain identifiers (“*”) are not currently supported.

- **Destination Country** – This match criteria allows you to specify the destination country of the remote machine.

Destination criteria is not supported when ZPA is selected as the Forwarding Method.

After configuring a Forwarding Method and match criteria, you must choose an action. By default for ZIA use cases, the Cloud Connector appliance uses geolocation to locate a ZIA Public Service Edge in geographic proximity to the appliance. Alternatively, you can manually specify which Service Edge to use by configuring a gateway under the Forwarding Methods section of the Administration menu. Zscaler recommends using geolocation where possible.

Gateway selection criteria is not supported when ZPA is selected as the Forwarding Method. Cloud Connector automatically selects a broker.

Lastly, specifically for ZPA use cases, Cloud Connector also allows for the filtering of DNS requests/responses. In the Administration menu within DNS Control, administrators can add additional rules to permit or deny specific DNS requests from workload segments. More importantly, this functionality can be used to determine which traffic gets consumed by ZPA, and therefore which synthetic IP pool is used to address traffic within Microtunnels.

To view configuration instructions, see **Configuring Traffic Forwarding Rules** (**https://help.zscaler.com/cloud-branch-connector/configuring-traffic-forwarding-rule**).

## Choosing the Correct Design Model

Cloud Connector is extremely flexible in the ways in which it can be deployed: directly adjacent to the workloads it services, or in a dedicated island by itself wherein traffic can be directed through it via Google networking constructs like Shared VPC. There is no single design model that fits every environment. Many organizations pull elements from all design models to suit their goals. There are three main questions to ask when determining how best to get started:

### Is ZPA a requirement?

ZPA requires workload DNS queries to transit the Cloud Connector so a synthetic IP address can be assigned to the connection. Consider how DNS is employed within the cloud. If using cloud-hosted DNS servers, it is possible that DNS resolution requests are never directed across the Cloud Connector which would break ZPA. For this reason, automation scripts implement Google Cloud DNS to intercept and redirect DNS traffic. Google Cloud DNS is not required, however, consider how DNS resolution requests inherently transit Cloud Connector, such as if a public DNS server outside of the cloud is used. Additionally, if this cloud implementation also services inbound requests from remote clouds, consider pointing App Connectors towards real DNS servers in this scenario.

### Is high availability a requirement?

Zscaler recommends that high availability be employed in all use cases. However, when deploying directly into the workload VPC, compute costs can quickly spiral out of control. For this reason, you might consider using dedicated Transit/Egress VPCs peered with a Shared VPC. This allows you to maintain high availability, without a large compute footprint. This model likely requires that functions like Google shared VPC are implemented.

### Will Cloud Connector be deployed within the workload VPC, or in a dedicated VPC?

For small environments with only a handful of VPCs, Cloud Connector instances can be deployed directly within the workload VPC. However, the number of VPCs and EC2 instances tend to increase as an organization grows larger and invests further in the cloud. As new VPCs are added, they will require new appliances. As you consider where the Cloud Connector appliances will be installed, ensure you plan for adequate growth in the number of workloads and VPCs that Cloud Connector will protect. If the future state of the environment becomes operationally cumbersome, or if the environment already contains several VPCs, it might be best to consider a Shared VPC approach with a dedicated Transit/Egress VPC for Cloud Connector.

## Use Case: Direct to Internet Using Zscaler Internet Access

Implementing Cloud Connector to provide outbound internet access through ZIA is one of the first steps to cloud workload protection. The following deployment model represents a recommended option that can be leveraged to satisfy this business requirement and offer a foundation to build on when looking to implement services like ZPA.

In this model, Cloud Connectors can be installed directly into the workload VPC adjacent to the individual workloads they service. As with all deployment models, Zscaler highly recommends deploying Cloud Connector in high availability.

The following image is simplified for clarity. Redundant instances of Zscaler Cloud Connector should be deployed in all instances.

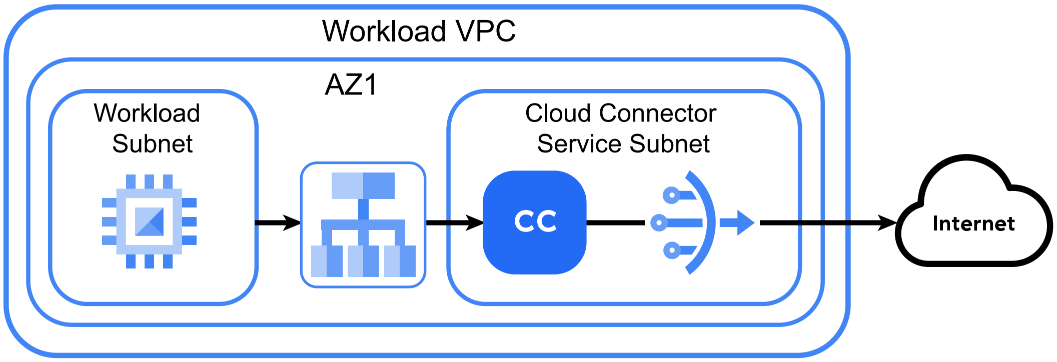


*Figure 10. Redundant workload model*

The primary benefit to this design option is its simplicity and time to implement. Since each Cloud Connector instance is spawned within the workload VPC that it services, routing is made simple. With Terraform, Zscaler automation can implement this model in a matter of minutes. From a cost perspective, you are only paying for egressing data fees one time (as the workload traffic leaves the Cloud Connector).

If you have many workload VPCs, however, this design option can be cumbersome. Any cost savings associated with egress fees can be eliminated by the increased compute footprint, since separate Cloud Connector instances are required per workload VPC. Additionally, this option requires the modification of many route tables to direct traffic accordingly, which is further complicated when high availability enters the picture.

When deploying this option via Terraform, the Cloud Connector instance is placed into a programmatically created VPC with new subnets and route tables. You can install this model using the Terraform scripts *base_cc_ilb* or *cc_ilb* as the deployment type.

## Use Case: Integrating with a Shared VPC

Cloud Connector can also be placed in a dedicated VPC where outbound workload traffic is first directed through a centralized hub. This model uses a Shared VPC (also known as a Transit, Security, or Egress VPC). This design solution is growing in adoption as organizations seek to address scalability concerns and operational deficiencies imposed by legacy inter-VPC networking.

This model closely resembles a traditional hub-and-spoke network since the Transit/Egress VPC, where Cloud Connector operates, receives traffic from many workloads spoke VPCs through a hub Shared VPC. As with all deployment models, Zscaler highly recommends deploying Cloud Connector in high availability. As workload traffic enters the Shared VPC from one availability zone, the transit gateway attempts to direct that traffic to a Cloud Connector appliance that exists in the same availability zone. This lowers cost and latency while also providing a mechanism for leveraging all available Cloud Connector appliances in an Active/Active fashion. Cloud Connector appliances provide high availability for one another, while simultaneously servicing traffic from their own availability zone.

The following image is simplified for clarity. Redundant instances of Zscaler Cloud Connector should be deployed in all instances.

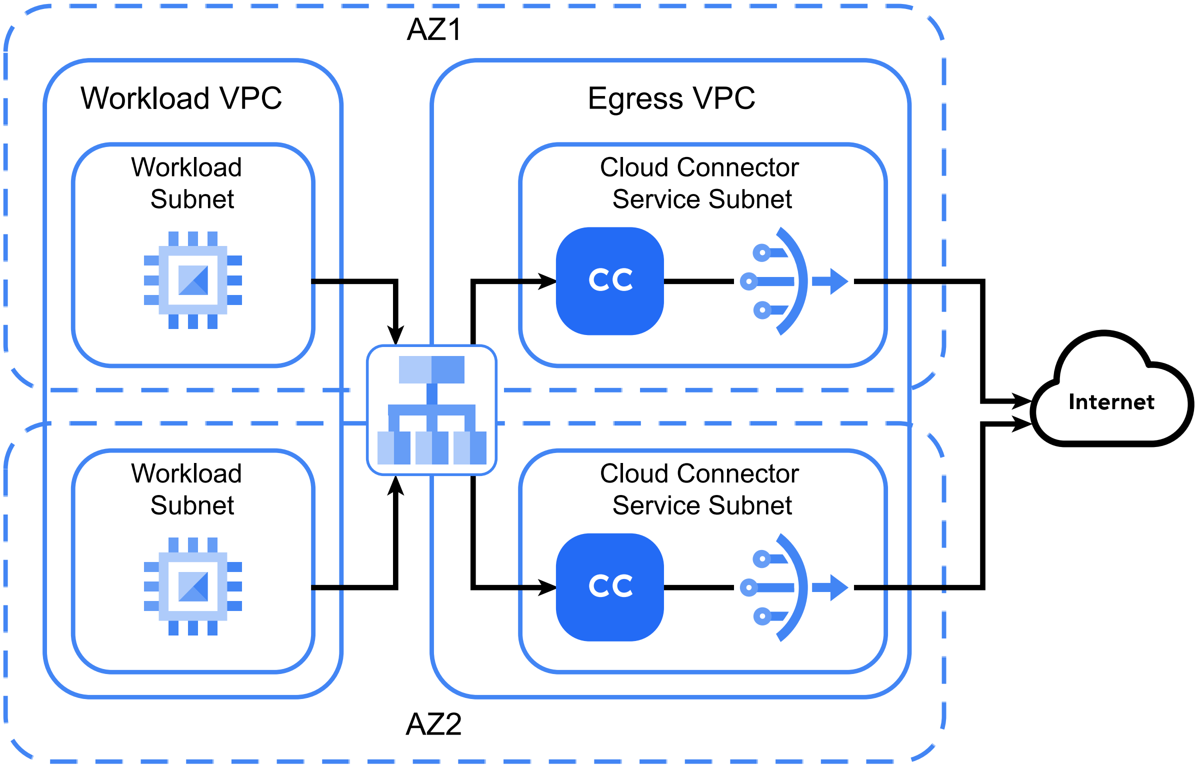


*Figure 11. Deploying Cloud Connector using a Transit/Egress VPC with Shared VPC*

Deploying Cloud Connector using a Shared VPC allows the organization to simplify cloud routing and reduce the compute footprint required when deploying directly to the workload VPCs. In this option, only a single pair of Cloud Connector appliances is necessary for the Shared VPC. Spoke VPC workloads requiring internet or private access are simply directed towards the internal passthrough Network Load Balancer using a simple default route, where they can then be directed towards the Cloud Connector appliances for outbound routing.

Zscaler recommends using a Shared VPC in conjunction for ZIA and ZPA use cases. For ZPA-only use cases, VPC Peering without a Shared VPC is not recommended. Depending on the size of the environment, Shared VPC can incur additional costs that could eliminate the cost savings from a reduced compute footprint.

You can implement this design option using the Terraform *base_cc_ilb* or *cc_ilb* deployment type.

The Shared VPC architecture is limited to 25 VPC peering connections, and lacks flexibility to accommodate regional routing. For these reasons, both Zscaler and Google recommend the single Shared VPC approach.

## Use Case: Integrating Zscaler Private Access

Assuming that Cloud Connector has been deployed and traffic directed through it, we can now add support for ZPA. This use case is growing in popularity as organizations seek to depart from legacy VPN technologies to interconnect cloud and on-premises workloads. An important consideration with Cloud Connector is that it is designed to facilitate outbound workload traffic towards a remote destination. When the destination is in a location you control, we must consider how this traffic ingresses into the remote facility. We do this using the Zscaler App Connector appliance, where App Connector VMs sit adjacent to the workloads they provide access to.

This model builds on the foundation provided in the direct-to-internet and Shared VPC use cases discussed previously. Cloud Connector provides outbound connectivity for cloud workloads to an on-premises data center, which uses App Connector VM appliances sitting in an application server segment to provide inbound connectivity. Both appliances build DTLS tunnels to the ZPA Broker and establish a Microtunnel between the source workload in the cloud and the destination data center workload. The traffic within the Microtunnel targets synthetic proxy IP addresses inside the Cloud Connector and App Connector, respectively.

The following image is simplified for clarity. Redundant instances of Zscaler Cloud Connector should be deployed in all instances.

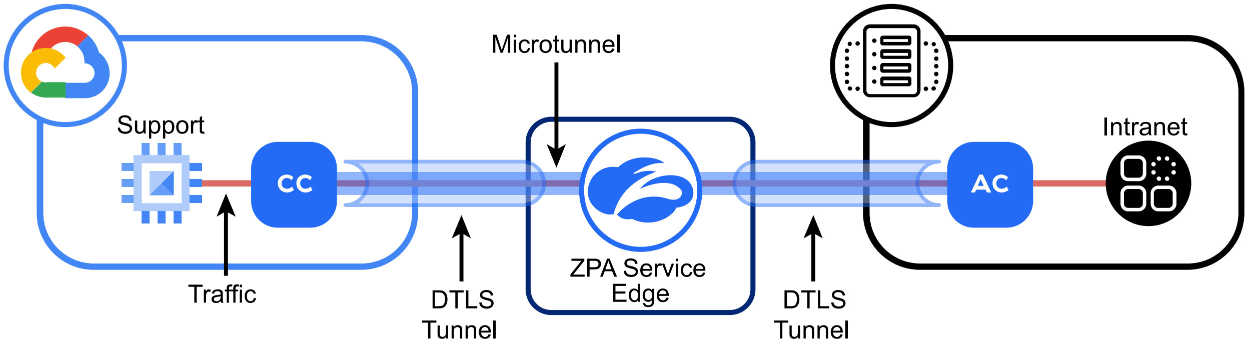


*Figure 12. Cloud and on-premises workloads meet at the ZPA Service Edge*

All communication between ZPA components travel within a client and server certificate-verified TLS connection. Within this TLS-encrypted Zscaler Tunnel, a microtunneling protocol exists. Select components of ZPA run through this encrypted Microtunnel end to end. Because the client and server use certificates issued by Zscaler, it is cryptographically impossible for ZPA to experience a Man-in-the-Middle (MITM) attack. The client certificates are verified against an organization's Certificate Authority (CA) and the server certificates are verified against Zscaler's CA, which cannot be spoofed by any third-party compromised CA.

ZPA only accepts connections from the Zscaler Cloud Connector and the App Connector instances that present a client certificate signed by a CA associated with each tenant. Zscaler Cloud Connector and App Connector only connect to ZPA service components that present a certificate signed by the ZPA infrastructure PKI.

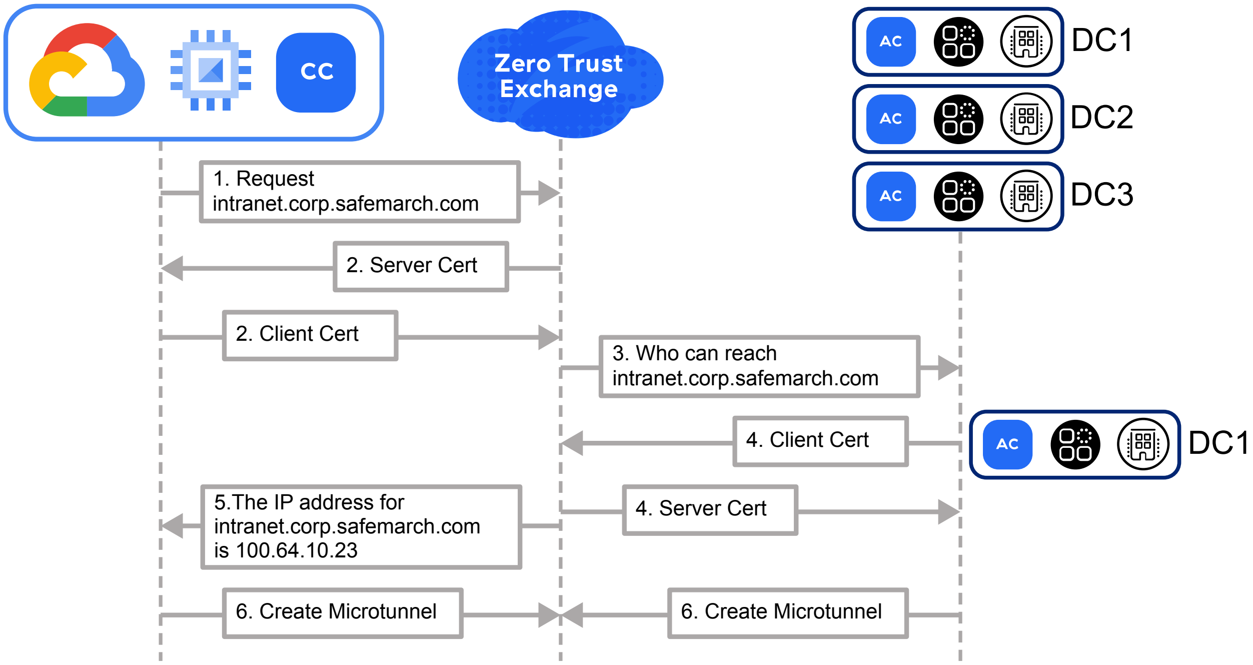

CC
Zero Trust Exchange
AC
DC1
AC
DC2
AC
DC3
1. Request intranet.corp.safemarch.com
2. Server Cert
2. Client Cert
3. Who can reach intranet.corp.safemarch.com
4. Client Cert
AC
DC1
5.The IP address for intranet.corp.safemarch.com is 100.64.10.23
4. Server Cert
6. Create Microtunnel
6. Create Microtunnel

*Figure 13. Authentication and tunnel setup between workloads and internal apps*

1. A workload requests access to an application.
2. The ZPA Service Edge and Zscaler Cloud Connector authenticate via certificate exchange.
3. If the workload is authorized to access the requested application, the ZPA Service Edge determines which App Connector can service the request.
4. The ZPA Service Edge and Zscaler App Connector authenticate via certificate exchange.
5. The workload is presented with the synthetic IP of the application.
6. A Microtunnel is established between the Zscaler Cloud Connector and Zscaler App Connector.

Zscaler Cloud Connector recognizes the internal applications that are available via ZPA. Access to these applications is defined by ZPA based on policies. Using information received from the ZPA Public Service Edge or ZPA Private Service Edge, Cloud Connector intercepts workload requests for applications, and then forwards those requests to the ZPA cloud.

No network information is required to access available applications. To facilitate secure private connections that are abstracted from the physical network, Cloud Connector associates permitted internal applications with a set of synthetic IP addresses. When a workload sends out a DNS request, Cloud Connector can recognize the domain as an internal application being protected by ZPA. Cloud Connector then intercepts the DNS request and delivers a DNS response to the workload that uses the synthetic IP address associated with the internal application.

To intercept and modify DNS requests, Cloud Connector must “see” the initial request from the cloud workload. To facilitate this, Zscaler recommends adding Google Cloud DNS support. By default, cloud workloads leverage Google Cloud DNS. However, this traffic never crosses the Cloud Connector and can break ZPA.

For every domain or DNS A Record configured, Google Cloud DNS ensures that the resolution request is first sent to the DNS resolver endpoint. The Google Cloud DNS endpoint then redirects the resolution request to a public (or internal) DNS server, where the request crosses Cloud Connector.

- Note that wildcard domain identifiers are supported, but it is recommended that you use specific DNS A Records for applications.
- You must revisit Google Cloud DNS configuration to add new or additional domains and DNS A Records.
- Google Cloud DNS is only required when internal Google Cloud DNS servers are used, where the DNS request does not traverse the Cloud Connector appliance. If using public DNS servers, Google Cloud DNS can safely be omitted.

## Use Case: Securing Traffic Between Clouds

Multi-cloud deployments, where workloads are spread across more than one cloud provider, are becoming more common as organizations look to provide hosting across more than one vendor. You can choose to host your cloud workloads in more than one cloud or, for redundancy or geoproximity, in multiple regions of the same cloud service provider. This use case focuses on how to solve for the challenges faced in this scenario and how we can secure this traffic using the ZPA model discussed **previously**.

Using the ZPA model as a basis, this model builds on the fact that remote application destinations secured by ZPA might not be in an on-premises data center. Instead, these applications exist within a different cloud region or in a different cloud service provider altogether. As originating cloud workloads send requests to remote applications, Cloud Connector routes them to the appropriate App Connectors in the destination cloud.

The following image is simplified for clarity. Redundant instances of Zscaler Cloud Connector should be deployed in all instances.

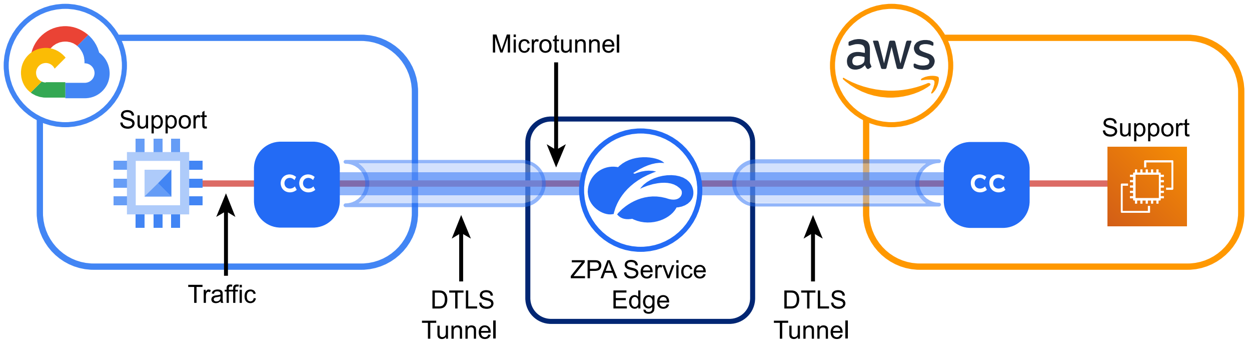


*Figure 14. Workload-to-workload communication across cloud providers*

Communication between two cloud workloads via ZPA mirrors the illustrations described in the **previous section**, so they are not repeated here. However, in the interest of discussing the underlay architecture, the previous figure depicts how App Connectors can be installed to facilitate this use case.

You can install this model using the Terraform scripts *base_cc_ilb* or *cc_ilb* as the deployment type.

## Use Case: Source IP Anchoring

Some cloud applications or web services restrict access based on the source IP address of the traffic. These applications require that traffic originates from a pre-registered static IP address. In the past, this address would typically be the public IP address of the organization's internet gateway. Access is denied to user traffic that originates from other IP addresses within or outside the organization. This includes the IP addresses of Zscaler data centers that are not pre-registered with the service. Source IP Anchoring (SIPA) allows an organization to use a common exit point in the cloud to satisfy this requirement.

This use case leverages Cloud Connector along with an organization's ZPA infrastructure discussed previously to allow traffic to be steered towards a specific egress point within the network. As cloud workloads send requests to internet-based web services requiring a specific source IP address, Cloud Connector routes the traffic to the appropriate App Connectors that reside in the on-premises or in-cloud data centers that have the required IP address allocated to them. Outbound traffic from these App Connectors then follow the associated route tables for their respective location and pass through a NAT Gateway. This allows the registered IP address to be used for all traffic in the organization.

## Optional SIPA Traffic Inspection with ZIA

There are two deployment options to be aware of when deploying SIPA. The first option, depicted below, allows egress traffic to utilize Source IP Anchoring functionality without any sort of traffic inspection.  In this model, the requests to the specific service are not inspected and are forwarded on to their ultimate destination. This is typically used by ZPA-only customers.

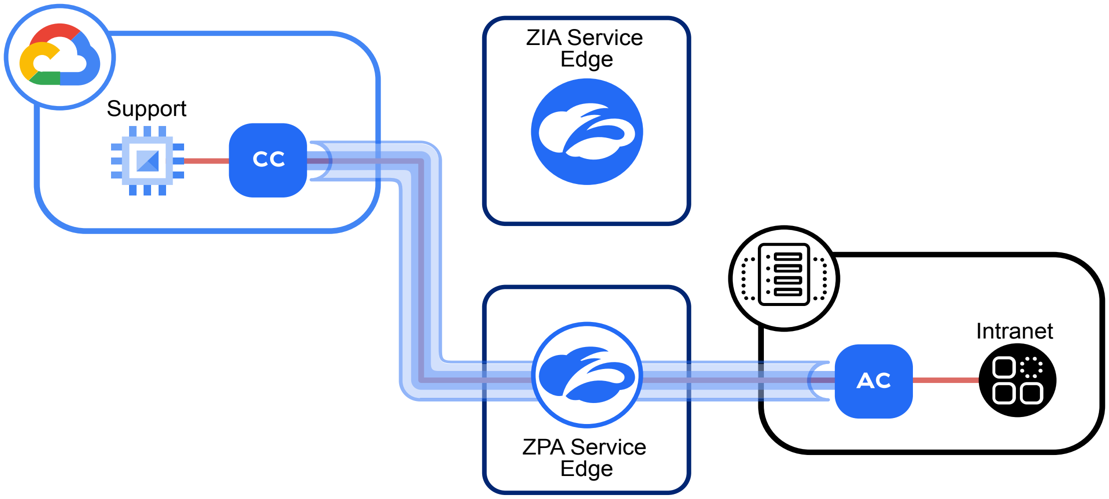


*Figure 15. SIPA traffic is forwarded to a specific ZPA App Connector acting as a broker, and then goes directly to the service*

Alternatively, traffic can be routed through ZIA first, allowing the traffic to be subjected to the organization's security policy prior to being routed to the remote App Connector and, ultimately, out to the internet. If you are a ZIA customer, Zscaler strongly recommends sending your workload traffic through your ZIA tenant, ensuring your security policies are adhered to by all the devices in your network.

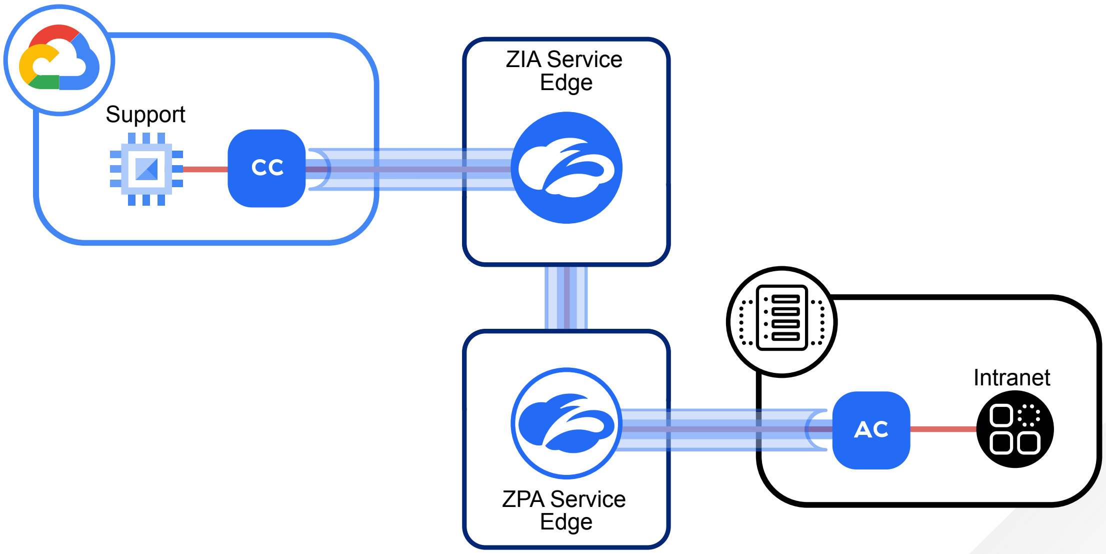


*Figure 16. SIPA traffic is forwarded to a ZIA Service Edge for inspection, then to a specific ZPA App Connector acting as a broker, and then goes directly to the service*

## Summary

Connecting workloads to the internet across different networks is difficult. What makes this harder is the traditional approach used by organizations to solve this challenge, such as using technologies like VPNs and firewalls. While the outcome of connecting these workloads is achieved, the cost to achieve these goals is significant:

- Risk of lateral threats and internet-based attacks by overextending the trusted network across the internet using VPN and WAN technologies.
- Complexity increases because of complicated route filtering, multiple network hops, and fragmented policy management.
- Poor visibility across application connectivity paths and increased network blind spots.
- Costs rise due to overprovisioning network services and the use of virtual appliances such as firewalls, IPs, routers, and other point products in cloud environments.
- Limited scale and performance from the increase in network and security services used in cloud environments.

As a result, there is a need for a better approach. Zscaler Cloud Connector is a cloud-native Zero Trust access service that provides fast and secure app-to-app, app-to-internet connectivity across multi-cloud environments. With integrated, automated connectivity and security, it reduces complexity and cost, and provides a faster, smarter, and more secure alternative to legacy network solutions.

## About Zscaler

Zscaler (NASDAQ: ZS) accelerates digital transformation so customers can be more agile, efficient, resilient, and secure. The Zscaler Zero Trust Exchange protects thousands of customers from cyberattacks and data loss by securely connecting users, devices, and applications in any location. Distributed across more than 150 data centers globally, the SASE-based Zero Trust Exchange is the world's largest inline cloud security platform.