Glossary - GKE

Subnet:

- Detailed Explanation: A subnet is a way to divide a larger computer network into smaller, more manageable parts. It's like having a big neighborhood, and you split it into different streets or blocks. Each subnet can have its own set of devices, and they can communicate with each other, but it's more challenging for devices in different subnets to talk directly.

Primary IP:

- Detailed Explanation: The primary IP address is the main address assigned to a device on a network. Think of it as your home address. When you want to send or receive data over the internet, this address is used to find your device. It's unique and essential for communication.

Secondary IP:

- Detailed Explanation: A secondary IP address is like having multiple phone numbers for the same person. You might have your primary phone number, but you can also have additional numbers for different purposes. Secondary IPs provide more flexibility and allow your device to handle various types of traffic separately.

RFC 1918:

- Detailed Explanation: RFC 1918 is a set of guidelines that define specific ranges of IP addresses reserved for private use within a network. It's like having a secret code for your house address. These reserved addresses are not used on the public internet, so they are ideal for setting up private networks, like your home Wi-Fi network or a company's internal network.

Kubernetes API object:

- Detailed Explanation: In Kubernetes, an API object is like a blueprint or configuration file that describes how a particular aspect of an application or service should work. Imagine you're building a LEGO castle, and each piece of the castle is represented by a specific instruction manual. These instruction manuals (API objects) tell Kubernetes how to create, manage, and monitor different parts of your applications, like web servers or databases.

256-bit encryption:

- Detailed Explanation: Encryption is a way to protect information by turning it into a secret code. The "256-bit" part refers to the length and complexity of the code. It's like having a very long and complex password that is extremely difficult for anyone to guess. This level of encryption is considered highly secure and is used to protect sensitive data, such as your passwords or credit card information when it's sent over the internet.

TLS (Transport Layer Security):

- Detailed Explanation: TLS is a protocol that ensures secure communication over the internet. It works like a secret handshake between your web browser and a website's server. When you visit a secure website (usually with "https://" in the URL), TLS encrypts the data exchanged between your browser and the server, making it unreadable to anyone who might intercept it. It provides confidentiality and verifies the authenticity of the website.

X.509 certificate:

- Detailed Explanation: An X.509 certificate is a digital document that acts as an identification card on the internet. It contains information about a website or an entity, including their public key. When you visit a secure website, your browser checks the website's X.509 certificate to ensure that it's legitimate. If the certificate is valid and matches the website's identity, it establishes a secure connection using TLS, ensuring that your data is safe from eavesdroppers.

Workload object in GKE:

- Detailed Explanation: In Google Kubernetes Engine (GKE), a workload object is a configuration that specifies how a group of containers should run in a cluster. It defines things like how many containers should be running, what resources they need (like CPU and memory), how they should scale, and how to update them. Workload objects help manage and maintain applications in a Kubernetes cluster, making it easier to run and scale your software.

Borg:

- Borg is a large-scale cluster management system developed by Google. It is used to manage the infrastructure for Google's internal services. Borg is a hierarchical system that consists of a master, workers, and jobs. The master is responsible for scheduling jobs and monitoring the health of the workers. The workers are responsible for running the jobs. The jobs are the units of work that are executed by Borg.

Kubelet:

- Kubelet is a process that runs on each node in a Kubernetes cluster. It is responsible for running containers, pulling images from a registry, and reporting the health of the node to the Kubernetes API server. The Kubelet is the main agent that ensures that containers are running on the nodes in the cluster. It communicates with the Kubernetes API server to get instructions about which containers to run and where to run them. The Kubelet also monitors the health of the containers and reports any problems to the Kubernetes API server.

Kube Proxy:

- Kube proxy is a network proxy that runs on each node in a Kubernetes cluster. It is responsible for routing traffic to containers running on the node. The Kube proxy is a layer 4 (TCP/IP) proxy that sits between the nodes in the cluster and the outside world. It is responsible for routing traffic to the correct containers based on the service definitions in the Kubernetes cluster.

Google Cloud Artifact Registry:

- Google Cloud Artifact Registry is a managed artifact repository service that allows you to store and manage Docker container images, Helm charts, and other artifacts. The Google Cloud Artifact Registry is a fully-managed service that makes it easy to store and manage your artifacts. It provides a secure and scalable repository for your artifacts, and it integrates with other Google Cloud Platform services, such as Kubernetes and Cloud Build.

SLI (Service Level Indicator):

- SLI stands for Service Level Indicator. It is a measure of the performance of a service. An SLI can be anything that can be used to measure the performance of a service, such as the response time, the availability, or the throughput.

SLO (Service Level Objective):

- SLO stands for Service Level Objective. It is a target value for an SLI. An SLO specifies the desired level of performance for a service. For example, an SLO might specify that the response time for a service should be less than 100 milliseconds 99.9% of the time.

SLA (Service Level Agreement):

- SLA stands for Service Level Agreement. It is a contract between a service provider and a customer that defines the SLOs and penalties for not meeting the SLOs. An SLA specifies the level of service that a service provider guarantees to its customers. For example, an SLA might specify that a service provider will refund customers for any downtime that exceeds 1% of the time.

Availability:

- Availability is the percentage of time that a service is available to users. Availability is calculated as the uptime divided by the total time. For example, a service that is up for 99.9% of the time has an availability of 99.9%.

Managed Instance Group:

- Managed Instance group is a collection of Google Compute Engine instances that are managed together. A managed instance group can be used to ensure that there are always a certain number of instances running. For example, you could create a managed instance group with 3 instances. This would ensure that there are always at least 3 instances running, even if one or two instances fail.

Cordon Node:

- Cordon Node is a node that has been marked as unhealthy and should not be used to run workloads. A cordon node is used to prevent workloads from being scheduled on a node that is unhealthy. This can help to prevent the spread of an infection or to isolate a node that is experiencing problems.

Uncordon Node:

- Uncordon Node is a node that has been marked as healthy and can be used to run workloads. An uncordon node is used to allow workloads to be scheduled on a node that has been previously marked as unhealthy. This can be done once the node has been repaired or the infection has been contained.

GCP Cloud Operations:

- GCP Cloud Operations is a suite of tools that help you monitor, troubleshoot, and optimize your Google Cloud Platform (GCP) resources. GCP Cloud Operations provides a variety of tools, such as dashboards, alerts, and logs, that can be used to monitor your GCP resources. It also provides tools for troubleshooting problems and optimizing your resources.

HTTP/HTTPS Traffic:

- HTTP/HTTPS traffic is the type of traffic that is used to transfer hypertext documents and other data over the Internet. HTTP is the unencrypted version of the protocol, while HTTPS is the encrypted version. HTTP is used for unencrypted traffic, while HTTPS is used for encrypted traffic.

Private Service Connect:

- Private Service Connect is a feature of GCP that allows you to connect your Google Cloud services to private networks. Private Service Connect can be used to connect your Google Cloud services to your on-premises network

Virtual Private Cloud (VPC):

- Virtual Private Cloud (VPC) is a logical network that is isolated from other networks in the cloud. A VPC can be used to create a private network for your applications and workloads. VPCs are supported by most major cloud providers, including Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP).

Googleapis:

- Googleapis is a set of APIs that allow you to interact with Google Cloud Platform services. The Googleapis APIs are used to create, manage, and use Google Cloud Platform resources. For example, the Googleapis Compute Engine API can be used to create and manage Google Compute Engine instances.

GCR.IO:

- GCR.IO is the domain name for the Google Cloud Artifact Registry. The Google Cloud Artifact Registry is a managed artifact repository service that allows you to store and manage Docker container images, Helm charts, and other artifacts.

NAT (Network Address Translation):

- NAT stands for Network Address Translation. NAT is a technique that allows you to use one public IP address to connect to multiple private IP addresses. NAT is used to conserve public IP addresses and to improve security.

DNAT (Destination Network Address Translation):

- DNAT stands for Destination NAT. DNAT is a type of NAT that is used to redirect traffic to a specific destination. DNAT can be used to forward traffic to a server in a private network.

Endpoint:

- Endpoint is a logical construct that represents a network resource. Endpoints are used to identify and manage network resources, such as virtual machines, containers, and applications.

GCE_VM_IP:

- GCE_VM_IP is the IP address of a Google Compute Engine instance. The GCE_VM_IP is used to identify and communicate with a Google Compute Engine instance.

Network Endpoint Group (NEG):

- network endpoint group (NEG) is a collection of network endpoints that are treated as a single logical unit. NEGs can be used to simplify the management of network resources.

Untrusted Workloads:

- Untrusted Workloads are workloads that are not trusted to run with the same privileges as other workloads. Untrusted workloads are often isolated from other workloads to reduce the risk of security breaches.

Certificate Authority (CA):

- Certificate Authority (CA) is an entity that issues digital certificates. Digital certificates are used to verify the identity of a website or other entity. CAs are used to secure communications between computers.

Google Front End (GFE) Servers:

- Google Front End (GFE) servers are the servers that are responsible for serving web pages to users. GFE servers are located in data centers around the world.

N2D Instances:

- N2D instances are Google Compute Engine instances that are optimized for performance. N2D instances use the latest Intel® Xeon® Scalable processors and have more memory than other types of Google Compute Engine instances.

C2D Instances:

- C2D instances are Google Compute Engine instances that are optimized for cost. C2D instances use older Intel® Xeon® processors and have less memory than other types of Google Compute Engine instances.