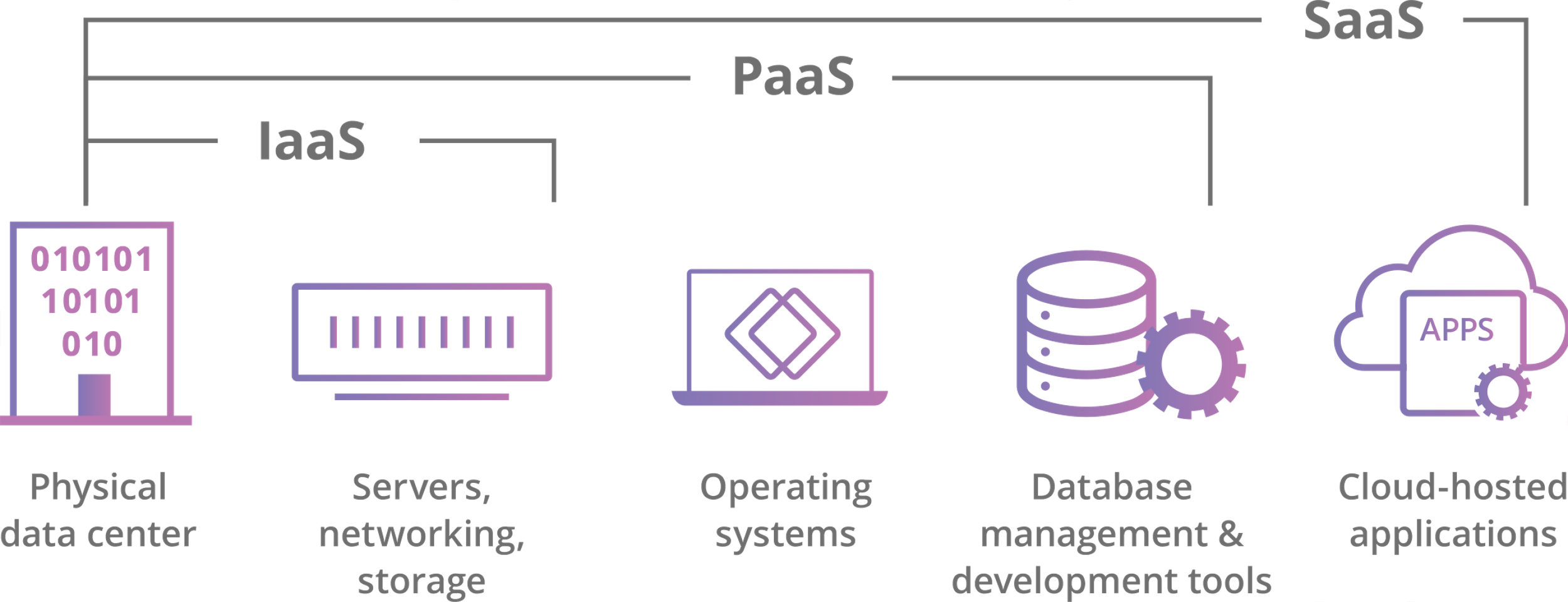
**What is cloud?**

"The cloud" refers to servers that are accessed over the Internet, and the software and databases that run on those servers. Cloud servers are located in data centers all over the world.

**How does cloud computing work?**

Cloud computing is possible because of a technology called virtualization. Virtualization allows for the creation of a simulated, digital-only "virtual" computer that behaves as if it were a physical computer with its own hardware. The technical term for such a computer is virtual machine.

**What are the main service models of cloud computing?**



**Software-as-a-Service (SaaS)**: Instead of users installing an application on their device, SaaS applications are hosted on cloud servers, and users access them over the Internet. SaaS is like renting a house: the landlord maintains the house, but the tenant mostly gets to use it as if they owned it. Examples of SaaS applications include Salesforce, MailChimp, and Slack.

**Platform-as-a-Service (PaaS)**: In this model, companies don't pay for hosted applications; instead they pay for the things they need to build their own applications. PaaS vendors offer everything necessary for building an application, including development tools, infrastructure, and operating systems, over the Internet. PaaS can be compared to renting all the tools and equipment necessary for building a house, instead of renting the house itself. PaaS examples include Heroku and Microsoft Azure.

**Infrastructure-as-a-Service (IaaS)**: In this model, a company rents the servers and storage they need from a cloud provider. They then use that cloud infrastructure to build their applications. IaaS is like a company leasing a plot of land on which they can build whatever they want — but they need to provide their own building equipment and materials. IaaS providers include DigitalOcean, Google Compute Engine, and OpenStack.

Formerly, SaaS, PaaS, and IaaS were the three main models of cloud computing, and essentially all cloud services fit into one of these categories. However, in recent years a fourth model has emerged:

**Function-as-a-Service (FaaS)**: Also known as serverless computing, breaks cloud applications down into even smaller components that only run when they are needed. Imagine if it were possible to rent a house one little bit at a time: for instance, the tenant only pays for the dining room at dinner time, the bedroom while they are sleeping, the living room while they are watching TV, and when they are not using those rooms, they don't have to pay rent on them.

FaaS or serverless applications still run on servers, as do all these models of cloud computing. But they are called "serverless" because they do not run on dedicated machines, and because the companies building the applications do not have to manage any servers.

**What are the different types of cloud deployments?**

In contrast to the models discussed above, which define how services are offered via the cloud, these different cloud deployment types have to do with where the cloud servers are and who manages them.

The most common cloud deployments are:

* **Private cloud**: A private cloud is a server, data center, or distributed network wholly dedicated to one organization.
* **Public cloud**: A public cloud is a service run by an external vendor that may include servers in one or multiple data centers. Unlike a private cloud, public clouds are shared by multiple organizations. Using virtual machines, individual servers may be shared by different companies, a situation that is called "multitenency" because multiple tenants are renting server space within the same server.
* **Hybrid cloud**: hybrid cloud deployments combine public and private clouds, and may even include on-premises legacy servers. An organization may use their private cloud for some services and their public cloud for others, or they may use the public cloud as backup for their private cloud.
* **Multi-cloud**: multi cloud is a type of cloud deployment that involves using multiple public clouds. In other words, an organization with a multi-cloud deployment rents virtual servers and services from several external vendors — to continue the analogy used above, this is like leasing several adjacent plots of land from different landlords. Multi-cloud deployments can also be hybrid cloud, and vice versa.

**Regions, Availability Zones and Edge locations**

**Regions -** are geographic locations where the data centers owned by public cloud service providers’ reside. A region could refer to places, e.g., Africa, Europe, the Middle East, Asia Pacific, the United States, China, etc. Regions could also be named by the Cloud Service Provider. Regions are important when using cloud services. It is important to use regions close to ones proximity. This reduces latency and cost. Choosing a region farther away from you would lead to high latency of service delivery and high cost.

**Availability zone** or **AZ -**  is an isolated location within a geographic region that serves as a physical data center. These locations are where data centers reside and services from Cloud Service Providers originate and operate.

Availability zones serve various purposes such as:

* Compliance
* Customer proximity
* Decrease of latency
* Resource protection, etc.

You can decide to replicate your services across multiple availability zones so as to derive the benefits listed above.

**A single region should have at least 2 availability zones.**

**Edge locations -** involve placing computers, storage, database, and other select cloud services closer to end-users.

They are located in major cities around the world and they are specifically used by CDNs to distribute content to end-users to reduce latency.

**IP Addresses**

**External IP address** - An external IP address is an address assigned to you by your ISP (Internet service provider). It is the address through which the Internet and other computers outside your local network use to identify you. You can also often type "IP address" into search engines such as Google to see your IP address.

**Internal IP address** - The internal IP address is the address that is assigned by your local network router. It generally begins with 192.168.x.x.It is also known as the local IP address. These IP addresses are only seen by other computers in your local network and not by computers connected to an external network, such as the Internet. To know about your Internal IP address, open the terminal/command prompt, search ipconfig, and press enter.

**Static Ip Address** - Static means "Permanent". Static External/Internal IP in GCP means the IP that is fixed for a particular resource. While the creation of a VM instance, a particular IP is assigned(ephemeral IP by default) to it, for static IP you have to reserve an IP from VPC networks and then assign it to the VM instance during the creation. There is no effect of stopping/suspending the VM on the static IP. The advantages of using static IP are that IP never gets changed unlike the Ephemeral, so there is no effect on infrastructural dependencies and the disadvantage of using it is that it is not expensive, you have to pay for the IP separately, even if it's not being used.

**Different ways to create instances**

Create Compute instance - <https://youtu.be/jVPPQ8jCFrE>

Create instance template - <https://cloud.google.com/compute/docs/instance-templates/create-instance-templates>

Create Custom images - <https://cloud.google.com/compute/docs/images/create-custom>