Edge Network

This lesson discusses edge networks in detail.

What is an edge network?

An *edge network* is an integral part of delivering low latency data to the end-users in the cloud infrastructure, while simultaneously reducing the load on the origin servers. By origin servers, I mean the backend servers on which we deploy our service.
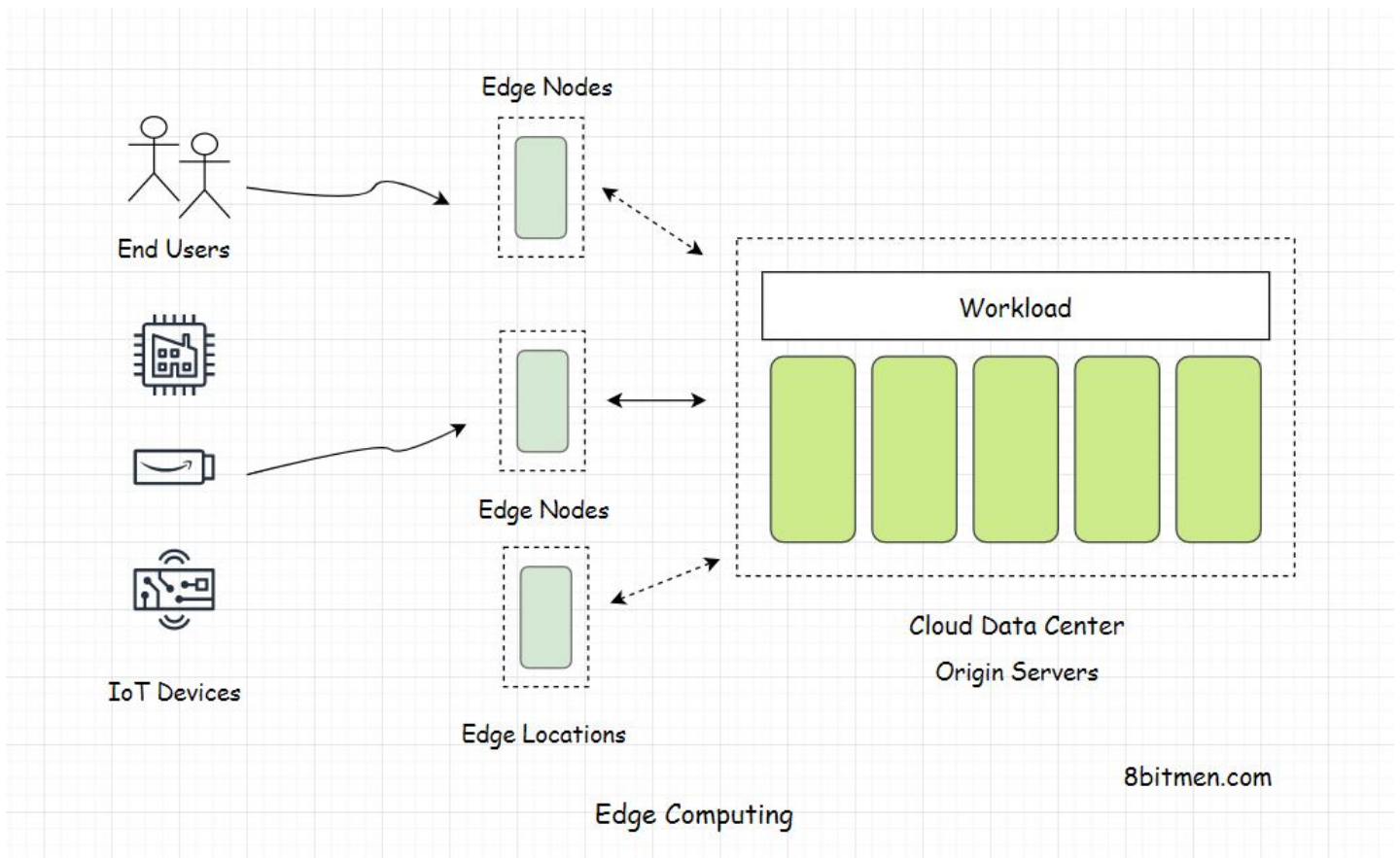
Edge simply means placing the infrastructure nearest to the end-users, that is, right

at the edge of the cloud infrastructure's global footprint. The edge network helps

deliver static content via *CDN* and contains compute nodes to process user requests at the edge locations, as opposed to routing them to the origin server.

Running *FaaS* on edge location is one example of this. Read *IBM Edge Functions* and *AWS Lambda@Edge* for more info.

Cloud platforms have several edge locations around the globe to serve content with minimal latency. _ Content Delivery Networks (CDNs)_ a component that delivers static content to the end-users and cuts down the latency and the load on the origin server, leverages the edge network of the cloud platform to function.

There are three primary upsides of using an edge network. Let's have a look at what they are.

Upsides of using an edge network

- Low latency – Improves user experience
- Geographic distribution of computing power - Facilitates a scalable architecture
- Cuts down the load on the original cloud data centers

Today most of the world is online. *IoT* devices like smartphones, voice assistants, and digicams are household things. This has put an immense load on cloud data centers and network bandwidth. There is a limit to how much load an infrastructure can handle.

To tackle this ever-growing need for computing, cloud providers have moved compute power to the end-users as opposed to processing everything on a centralized data center.

You will understand this with the help of a simple example, but before that, let's look at the edge computing trend first.

Edge computing

We know that the cloud providers have set up their infrastructure closest to the end-users. Also, the *IoT* devices are now developed with additional compute capabilities containing business logic.

This enables them to run computers on the premises where they are installed and provide the response to the end-user there, as opposed to streaming the

information over to the cloud, getting it processed in the data center, and fetching the response from there.
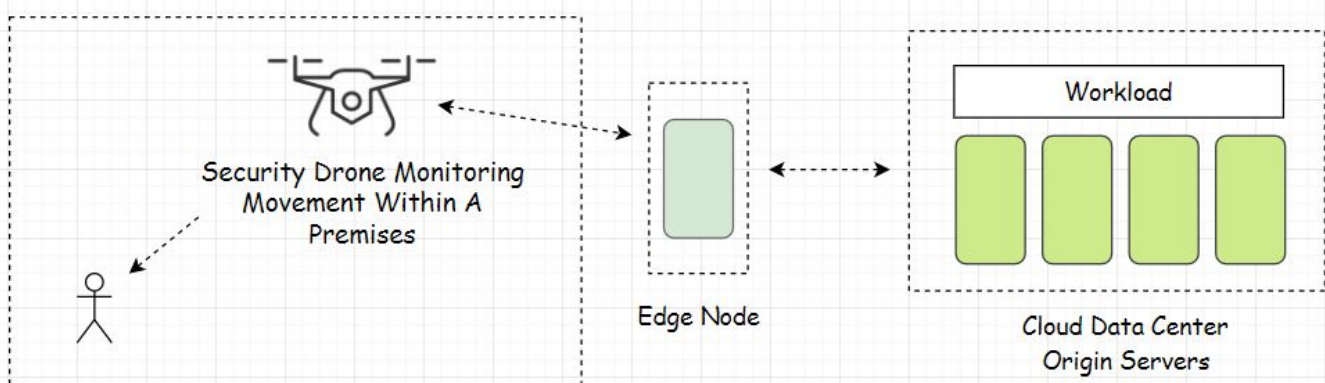
Putting business logic and compute in an IoT device is pretty similar to having a *fat client* in a *client-server architecture*. Let's understand this with the help of an example.

Edge compute example

Imagine an *AI* powered drone monitoring the premises. Its primary task is to detect the movement of objects on the premises, capture the visual, stream the information to the backend server for the algorithms to process the information, and based on the response, trigger an alarm.

Now, what if we move the intruder detection logic from the backend server to the drone? It can process the info and trigger the alarm immediately without the need to contact the backend server. This flow has literally no network latency.

Even if the drone needs to stream some information that it cannot process to the backend, we can process it on the edge node as opposed to streaming it all the way to the origin server.

Security Drone Monitoring Movement Within A Premises

Edge Node

Workload

Cloud Data Center Origin Servers

Leveraging Edge Network To Improve Drone's Response Time

8bitmen.com

By moving computers to the drone and the edge location, we have cut down the latency, improved the response time of the drone, and reduced the load on the

network and our origin data center. Edge locations facilitate scalable infrastructure and local data processing.

We should also keep this in mind that not all the business logic can be moved to the local device. It may spawn new security risks. There is no silver bullet. Trade-offs are always involved when designing systems.

So, folks, this concludes the global deployment chapter. In the next chapter, we will have a look at the cloud computing stack and the technologies involved in application deployment.