Contents

[Load balancer -CPU usage exceeds is calculated per server or for every server in case multiple server ? 1](#_Toc180656206)

[AWS load balancer based on CPU usage threshold can be set for all server ? 3](#_Toc180656207)

[AWS load balancer based on CPU usage can be set per server ? 3](#_Toc180656208)

### 

### Load balancer -CPU usage exceeds is calculated per server or for every server in case multiple server ?

**ChatGPT said:**

When monitoring **CPU usage with a load balancer** in a **multi-server setup**, the way CPU usage is calculated depends on the monitoring and scaling strategy you're using. Let's explore the typical scenarios:

**How CPU Usage is Calculated in Multi-Server Setups with Load Balancers**

1. **Per Server CPU Usage** (Individual Server Monitoring)
   * **CPU usage is monitored independently for each server** behind the load balancer.
   * If one server’s CPU usage exceeds a threshold (e.g., 80%), **only that server** is flagged or scaled (depending on the setup).
   * This method is useful for detecting specific server bottlenecks, like:
     + Uneven load distribution.
     + Resource-intensive processes on certain servers.

**Example:**

* + **Server A**: 85% CPU usage (threshold exceeded)
  + **Server B**: 45% CPU usage  
    **Action**: You might trigger an **alert** or **scale up Server A** only.

1. **Aggregate CPU Usage** (Cluster-wide Monitoring)
   * CPU usage can also be **aggregated across all servers** in the load balancer pool to give an **average CPU usage** for the entire infrastructure.
   * This is useful when making scaling decisions based on the **overall system load** rather than individual servers.
   * For example, in **horizontal scaling**, you might add new servers when the **average CPU usage across all servers** exceeds a threshold.

**Example:**

* + **Server A**: 70% CPU usage
  + **Server B**: 60% CPU usage  
    **Aggregate CPU Usage** = (70% + 60%) / 2 = 65%  
    **Action**: If the aggregate usage exceeds 70%, a new server is added to the pool.

**When to Use Per-Server vs Aggregate Monitoring?**

* **Per-Server Monitoring**:
  + Use this when you need **precise control** over individual servers (e.g., in hybrid environments or for troubleshooting).
  + Good for detecting **load imbalance** issues, such as sticky sessions or uneven request distribution.
* **Aggregate Monitoring**:
  + Ideal for **autoscaling** decisions in **cloud environments** (like AWS Auto Scaling or Kubernetes).
  + Useful when your system distributes traffic evenly and you care about **overall resource utilization** rather than single-server spikes.

**Scaling and Load Balancer Considerations**

* **Horizontal Scaling**: If average CPU usage across the cluster exceeds a certain threshold, the system can add more servers to distribute the load.
* **Vertical Scaling**: If an individual server’s CPU usage spikes, the system might either upgrade the server or redistribute traffic to other servers.

**Example of Aggregate Monitoring in AWS ALB + Auto Scaling Setup**

* AWS **Auto Scaling Group** monitors **average CPU utilization** of all servers behind the **Application Load Balancer (ALB)**.
* If the **average CPU usage exceeds 70%**, a new instance is automatically added to the pool.

**Summary**

* **Per-server CPU usage** is calculated individually for each server and helps in identifying specific bottlenecks or imbalances.
* **Aggregate CPU usage** gives a holistic view across the entire infrastructure, useful for **autoscaling** and **load distribution** decisions.

Both methods are valid and often used together in cloud-native environments to monitor and optimize system performance efficiently.

### AWS load balancer based on CPU usage threshold can be set for all server?

Yes, in **AWS**, you can configure **autoscaling based on CPU usage thresholds** for all servers in an **Auto Scaling Group (ASG)** behind a **Load Balancer** (such as an **Application Load Balancer (ALB)** or **Network Load Balancer (NLB)**).

AWS allows you to set scaling policies based on **average CPU utilization** across the instances or **per-instance metrics**.

### AWS load balancer based on CPU usage can be set per server?

AWS **Elastic Load Balancer (ELB)** itself does not support **CPU-based load balancing** directly out of the box. However, you can achieve **CPU-based load distribution per server** using other AWS services and custom configurations. Below are the ways to set this up using **Auto Scaling Groups (ASG)**, **Target Tracking Policies**, and **Lambda for Dynamic Traffic Shifting**.