Our approach leverages a dynamic weighted load-balancer informed by each replica's available resources and geographical distance from the replica to the client. These system resources are periodically collected via requests initiated by the authoritative server. Specifically, we monitor CPU utilization, memory utilization, and the number of client connections. CPU and memory usage serve as indicators of a replica's residual capacity to handle additional requests, while the number of client connections provides a complementary measure of system load, under the assumption that a high number of active connections correlates with increased resource contention. We use geographical distance as a heuristic proxy for the RTT between the client and a replica.
The weights are calculated in the following manner:

$$\text{weight} = \alpha \cdot (\beta \cdot m_1 + \gamma \cdot m_2) + \delta \cdot m_3 + \epsilon \cdot m_4$$

Where

$$\begin{aligned} m_1 &= \text{CPU utilization} \\ m_2 &= \text{Memory utilization} \\ m_3 &= \text{Normalized number of client connections} \\ m_4 &= \text{Normalized geographical distance from replica to client} \\ \alpha &= 0.4 \\ \beta &= 0.8 \\ \gamma &= 0.2 \\ \delta &= 0.2 \\ \epsilon &= 0.4 \end{aligned}$$

We forward the request to the replica with the smallest weight.
The pseudocode for the algorithm is the following:

Algorithm 1 Pseudocode for the weighted load-balancing algorithm

```
Do the following upon receiving a client request:
for each replica node of CDN do
    retrieve node's metrics
    $m_1 \leftarrow$ CPU Utilization metric
    $m_2 \leftarrow$ Memory Utilization metric
    $m_3 \leftarrow \frac{\text{min(Number of Connections metric,1000)}}{1000}$

▷ 1000 represents maximum number of reasonable client connections

    Get latitude-longitude pairs corresponding to client and replica
    metric_distance $\leftarrow$ haversine distance between coordinate pairs
    $m_4 \leftarrow \frac{\text{min(distance,20000}km)}{20000km}$

▷ 20000km represents largest distance between any two geographic points; it
can be substituted with maximum expected distance from client to any of the
replicas

    replica's weight $\leftarrow \alpha \cdot (\beta \cdot m_1 + \gamma \cdot m_2) + \delta \cdot m_3 + \epsilon \cdot m_4$

▷ $\alpha$, $\beta$, $\gamma$, $\delta$, and $\epsilon$ are the same as discussed previously

end for
Forward request to replica node with smallest weight
```