How SHIR-HA(multiple nodes) handles tasks

Last updated by | Veena Pachauri | Mar 8, 2023 at 11:10 PM PST

If customer is asking why the loading balance of HA cannot ensure the even distribution of traffic:

Let me give a brief explanation for how SHIR-HA(multiple nodes) handle tasks.

```
  1. Each SHIR node has several worker process which is responsible for executing tasks, the number of worker process is called max capacity or max concurrent j
  2. There is only one SHIR node called primary node, which is responsible for pulling tasks from services and has a worker pool that manages workers across all
  3. When a new task pulled off, the primary node will pick a worker from worker pool, and below is the brief logic for picking worker.
          1. Calculate node usage statistics based on all worker status, and pick candidate node which has minimum busy worker number and has at least one avail
          2. Randomly pick one available worker from candidate node.
```

As you can see that our load balance logic is based on task number instead of throughput for now. So it cannot ensure the even distribution of traffic or the balanced CPU/Memory utilization within the nodes.

How good have you found this content?