Internet Networking 236341 — HW4

Ismael Badir , Yasmin Mitkal

1 Load Balancer Policy Overview

The load balancer keeps track of each server’s current workload as a “finish time” (the sum of its queued jobs’ estimated durations), and for every new request it predicts how long each server would take by adding the request’s size, scaled by a server and type multiplier to that server’s finish time. It then selects the server whose predicted completion is smallest, updates its finish time with that prediction, and forwards the request there. By combining both queue length and processing speed, the load balancer ensures that each job goes to the server that will actually finish it first, minimizing overall response latency across servers.

```
        pseudo code for scheduling policy
        Input: finish_times[0..3], MULTIPLIERFACTOR[n][m], request(req_type, run_time)

        req_index = map(req_type)    // e.g., ’P’:0, ’V’:1, ’M’:2
        for i in 0..3:
        predicted[i] = finish_times[i] + run_time * MULTIPLIERFACTOR[i][req_index]
        best = argmin(predicted)
        finish_times[best] = predicted[best]
        return best    // index of server to forward request
```

Listing 1: Scheduling Policy Pseudocode