

Dynamic Request Redirection and Elastic Service Scaling in Cloud-Centric Media Networks

Abstract:

We consider the problem of optimally redirecting user requests in a cloud-centric media network (CCMN) to multiple destination Virtual Machines (VMs), which elastically scale their service capacities in order to minimize a cost function that includes service response times, computing costs, and routing costs. We also allow the request arrival process to switch between normal and flash crowd modes to model user requests to a CCMN. We quantify the trade-offs in flash crowd detection delay and false alarm frequency, request allocation rates, and service capacities at the VMs. We show that under each request arrival mode (normal or flash crowd), the optimal redirection policy can be found in terms of a price for each VM, which is a function of the VM's service cost, with requests redirected to VMs in order of nondecreasing prices, and no redirection to VMs with prices above a threshold price. Applying our proposed strategy to a YouTube request trace data set shows that our strategy outperforms various benchmark strategies. We also present simulation results when various arrival traffic characteristics are varied, which again suggest that our proposed strategy performs well under these conditions.