

Performance Modeling of Computer Systems and Networks

Prof. V. de Nitto Personè

AA 2018/2019

Project

Mobile devices are still limited in computing capabilities and battery lifetime. The problem of enhancing user experience may find a solution in cloud computing.

Consider a "two-layer" cloud system, consisting of an edge cloud server (cloudlet) and a remote cloud server, where the cloudlet is at "one-hop" distance from a set of mobile devices. Applications running on these mobile devices autonomously select some of their tasks for offloading to an external server (e.g., because of performance or energy saving reasons), and send an **offloading request** to a **controller** located on the cloudlet. Upon receiving a request, the controller takes a decision about whether the task should be sent to the cloudlet or the cloud, with the goal of minimizing the mean response time.

Typically, tasks hosted by the cloud server benefit of a higher execution rate, but suffer for greater network delay. We assume that the remote cloud server has virtually unlimited resources, hence it is always able to guarantee absence of interference among any number of tasks allocated to it. On the other hand, the cloudlet has limited resources, so that it is able to guarantee absence of interferences among tasks allocated to it as long as their number does not exceed a given threshold N .

Let us assume users belong to two classes and denote with n_i the number of class i tasks in execution on the cloudlet. The controller can take the following simple decision upon task arrival events:

Algorithm 1

```
○ arrival:
    if  $n_1 + n_2 = N \rightarrow$  send on the cloud
    else accept the task on the cloudlet
```

Consider the following parameters:

- the mean arrival rates for the two classes are:
 $\lambda_1 = 4$ task/s, $\lambda_2 = 6.25$ task/s
- services on cloud can be assumed exponential, with mean rates:
 $\mu_{1\text{cloud}} = 0.25$ task/s and $\mu_{2\text{cloud}} = 0.22$ task/s
while services on cloudlet are hyperexponential with $p=0.2$ for both classes and
 $\mu_{1\text{clet}} = 0.45$ task/s and $\mu_{2\text{clet}} = 0.27$ task/s.

Note that, as we stated above, the service time includes the transmission time. For this reason, we assume $\mu_{1\text{clet}} > \mu_{1\text{cloud}}$ for both classes. Moreover, execution on the cloudlet of a class 1 task is more convenient than the execution of a class 2 task.

- Design a next-event simulator for the system above with access control *Algorithm 1* and $N=20$.
- Determine if the system is stationary or not and design the experiments accordingly.
- Evaluate:
 - the *system* response time and throughput both global and per-class;
 - the per-class effective¹ cloudlet throughput;
 - the per-class cloud throughput;
 - the *class* response time and mean population, both for the cloudlet and for the cloud;
 - Analyze the transient system statistics and in case of existence the steady-state statistics with the appropriate approach respectively.
- Define a queueing model for the whole system.
- Derive an analytical solution² and validate the simulation results.

¹ For "effective" throughput we mean the completed tasks rate.

² You can consider exponential distributions for cloudlet too.

By considering that execution on the cloudlet of a class 1 task is more convenient than execution of a class 2 task, design a different access control *Algorithm 2* with the aim to improve the global system response time. Motivate the choices in *Algorithm 2*.

- c.1. Implement *Algorithm 2*.
- c.2. Evaluate indices 1., 2., 3., 4., in a.3 and other important indices specific for *Algorithm 2*.
- c.3. Compare all results with those obtained with *Algorithm 1*.
- c.4. Modify queueing model in step b.1. to include *Algorithm 2*.
- c.5. Derive an analytical solution³ and validate the simulation results.

The steady state or transient statistics should be computed with a 95% confidence level. Please, consider as a guideline the paper “MANET Simulation Studies: The Incredibles”, S. Kurkowski, T. Camp, M. Colagrosso, *Mobile Computing and Communications Review*, Volume 9, Number 4.

The project must be delivered at least one week before the oral test.

The students should prepare a presentation of the obtained results. The time should not exceed 10 minutes per student in a group (20 minutes for a student alone).

Documents to be delivered:

- The written report should include: conceptual and analytical models, the methodology used for estimating the statistics, the results, both graphical and numerical form, with the relative comments
- the listing code.

Evaluation grid

- clarity
- pertinence
- completeness.

³ You can consider exponential distributions for cloudlet too.