Title

Computation Offloading for Smart Touristic Sites - COSMOS

Organization

Institute of Communication and Computer Systems – ICCS, https://www.iccs.gr

Experiment Description

The COSMOS project develops a framework which enables the dynamic offloading of processing workloads from mobile devices to edge clouds to facilitate the deployment of smart touristic applications in crowded cultural areas, by capitalizing on the features provided by the 5GINFIRE functionalities.

COSMOS Objectives

- Dynamic computation offloading mechanism for the minimization of the energy consumption of the portable devices.
- Workload profiling for the computation of resource requirements of the VNF chain towards efficient VNF placement and load balancing.
- User mobility estimation by using measurement of motion sensors. The estimated position is taken into account for the offloading decision.
- Performance evaluation (i.e., measurement of provisioning and scaling timescales, micro-benchmarks to identify potential performance bottlenecks) with the VNFs (e.g., object recognition) that are deployed in COSMOS.

The performance of the proposed framework is evaluated with the following scenario. COSMOS project capitalizes the University of Bristol 5G Testbed MEC infrastructure within the Millennium Square in the center of Bristol, and users moving in the vicinity of the Square. Visitors of this crowded place camera-equipped Raspberry Pi's to take snapshots of a Point of Interest (PoI) and get useful sight information via an object recognition service. Google's TensorFlow deep learning framework, as image classification service, was selected for the specific use case and was retrained in order to classify images.
IoT devices are generally limited in terms of energy availability and processing power. Hence, to enhance energy saving according to MEC principles for IoT-enabled applications, mobile users connect via multiple Wireless Access Points located in the Millennium Square and offload their computing intensive requests to a cluster of Edge servers. This placement enables low-latency access to the servers, capable of serving the users' requests in an on-demand fashion as depicted in Figure 1.

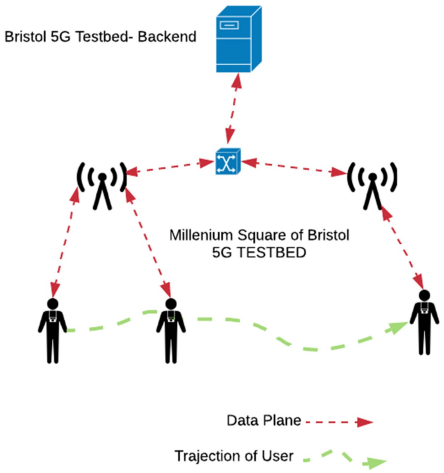

Figure 1 COSMOS use case overview

Cosmos Architecture

An overview on the system architecture is shown in Figure 2. The Bristol University testbed provides OSM Release 4 operating as an NFV management and orchestration tool that is connected with OpenStack, which acts as Virtual Infrastructure Manager (VIM) that controls the Virtual Deployment Units (VDUs). The system architecture of COSMOS follows a top-down design, meaning that there exists a VDU, namely; the Centralized Controller that dictates the decisions needed for the load balancing of the incoming workload, while at the bottom layer 3 VDU's namely; TensorFlow VDUs with different flavors, are deployed by the Bristol's OpenStack node. The proposed architecture is generally applicable in single-site MEC infrastructures and can be easily expanded towards Edge-to-Cloud or Edge- to-Edge collaboration.

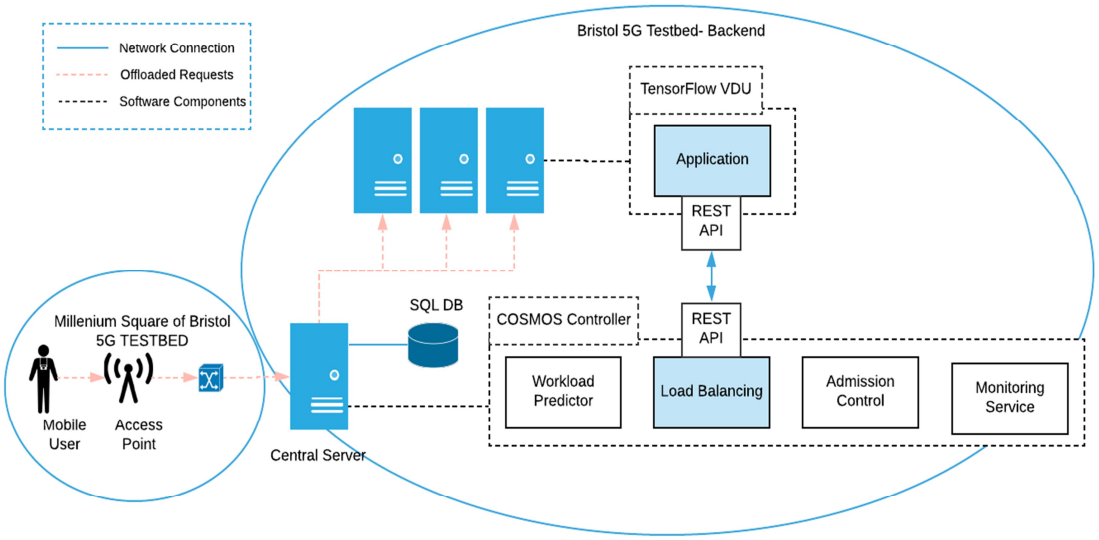

Figure 2 COSMOS architecture

Evaluation - Results

User Mobility Estimation

Assuming a user who is moving in the Millennium Square following the path on Figure 3, the purpose of this experiment is to demonstrate the accuracy of the estimation technique. The user repeatedly walks over the path, starting from O to F point, and holds a Raspberry PI device mounted with an IMU sensor. The results of the 5 most interesting instances are presented here. Overall the deviation of the estimations never exceeded the following results.

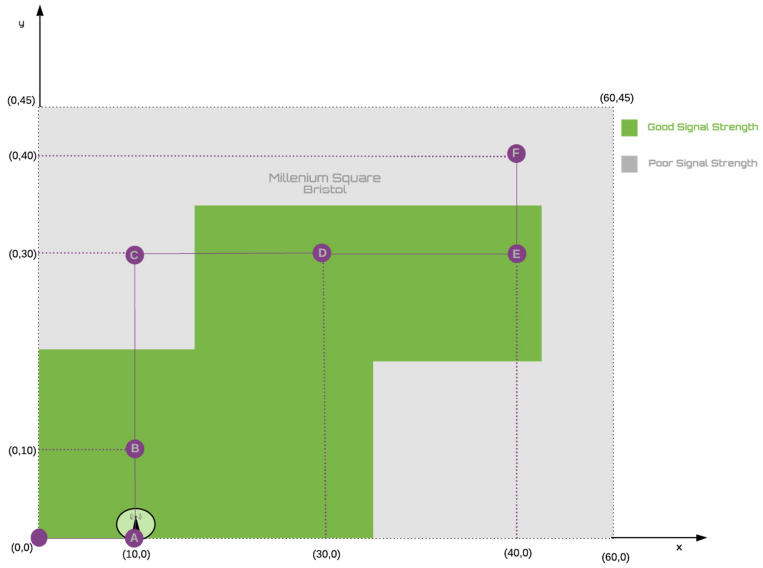

Figure 3 - Millennium Square Sketch

Figure 4 shows the coordinates (star) of the points of the path and the estimated positions (cross mark) that is measured by the IMU sensors.

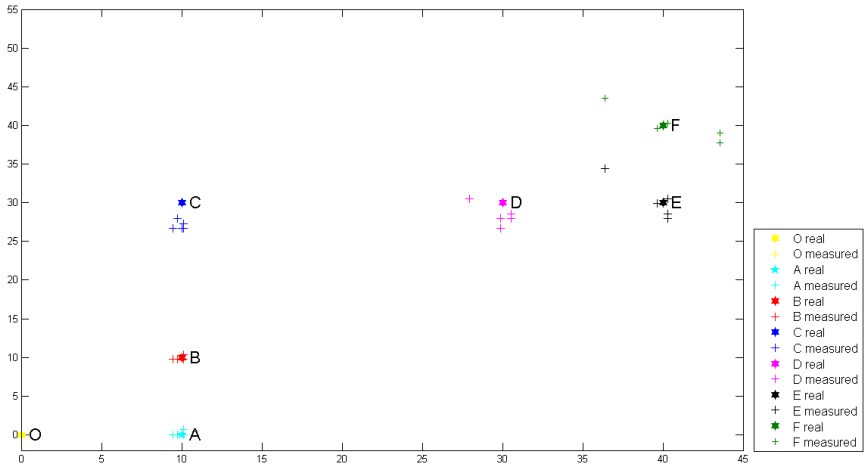

Figure 4 - Evaluation of Mobility Estimation Algorithm

Load Balancing and Resource Allocation

The performance of COSMOS Load Balancing mechanism and the workload prediction method are evaluated. At the beginning of every time interval, the COSMOS Controller predicts the volume of the workload for the next interval and determines the proper topology of operating VDUs and the distribution of the request to them. The distribution of the request to the activated VDUs is performed by solving a MILP problem, which is based on the VDU's flavor in Table 1. Each VDU's flavor corresponds to specific operating conditions of the VDU and they are designed is such way to server specific number of requests and keep the average response time below a specific threshold, as Table 1 presents. We mention that the average response time of each VDU flavor is not the highest acceptable value, but it is set at the half of this value. We underline that the response time of the VDUs is four time less than the execution of the image processing on the Raspberry Pi devices. Thus, the offloading decision is meaningful and beneficial for the users.

Table 1 - Resource Profiling

Flavor	Average Response Time (sec)	Number of Requests
Small - 2 cores	5	3
Medium - 4 cores	5	14
Large - 8 cores	5	27

In this experiment, we simulated a workflow of requests in order to exploit the full capabilities of the COSMOS architecture. A dataset of 5.5 hours is created. The experiment is divided into three phases. As depicted in Figure 5, for the first 5000 sec the average number of incoming requests is increasing gradually from 2 per interval to 20 per interval and finally back to 4 requests per interval. The second phase last 11000 sec and the average number of requests varies from 30 to 48, which is the peak value. During the last phase, the average volume of the requests decreases gradually from 20 to 10 requests until the end of the experiment. Figure 6 depicts the performance of the workload predictor. The red line represents the real requests per interval, while the blue line indicates the predicted requests by Kalman filter. As it is shown by the graph, the accuracy of the workload estimator is high and especially it quickly identifies the sudden changes of the incoming requests.

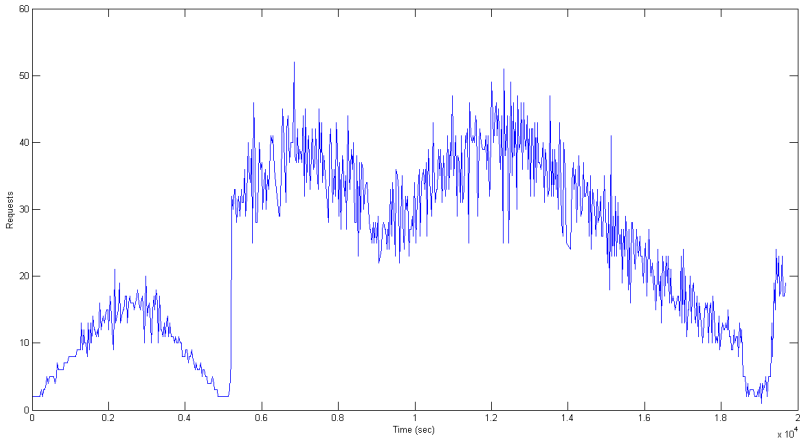

Figure 5 - Average Number of Requests

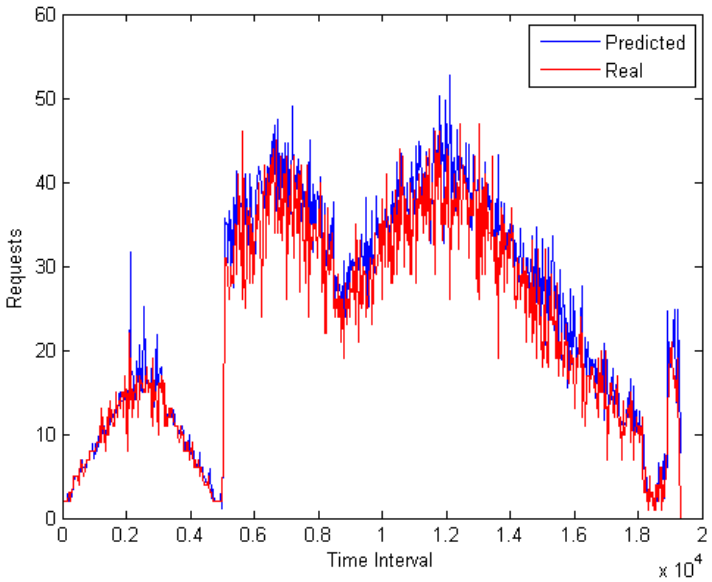

Figure 6 - Workload Prediction

Figure 7 illustrates the decision of the load balancing mechanism. As it shown on Figure 7, the combination of two VDUs is sufficient to serve the total number of incoming requests. Even for the most intensive workload, there is no need to activate all VDUs. This is due to the accurate resource profiling and the extraction of the specific operating points.

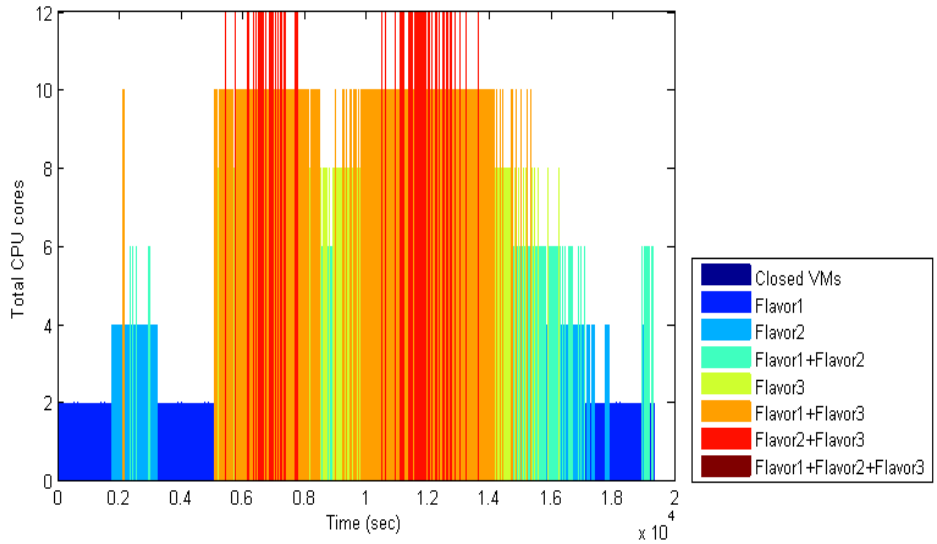

Figure 7 - Flavors Utilization over time

The main objective of COSMOS is to satisfy the QoS requirements and make the offloading process beneficial for the users. Thus, the threshold value of the average response time of the offloaded requests is set four times below the local execution on the mobile device. Regardless the phase of the experiment, the response time remains below the highest acceptance value (10 sec), as it is shown on Figure 8 . Only in few time intervals, this value is actually violated when the incoming requests reach their peak value.

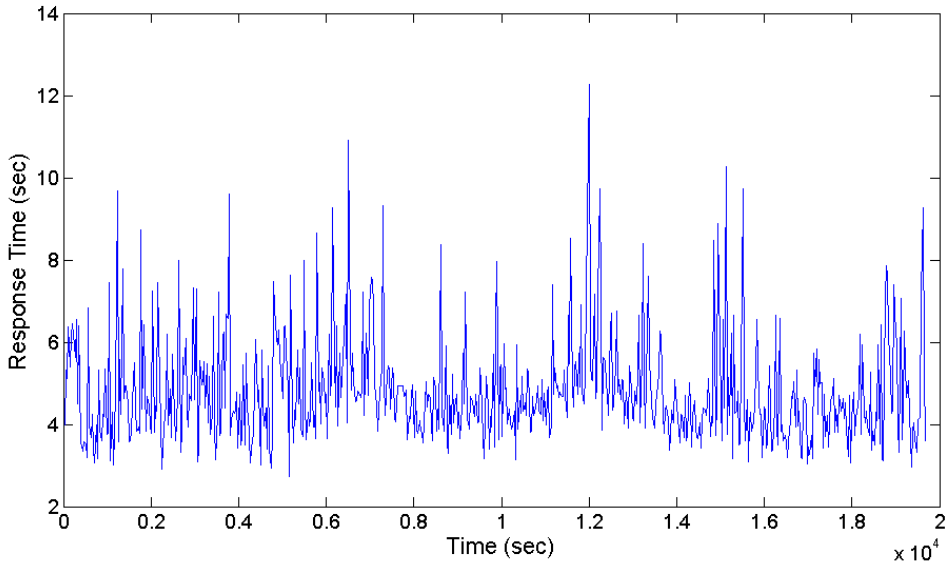

Figure 8 – Average Response Time

Conclusions

The experiments conducted on the 5GinFIRE Bristol Infrastructure allows us to draw some significant conclusions regarding the performance of COSMOS and gain important insights regarding its applicability and scalability in real world environments. Our work focused on two main aspects of the problem, namely a) Load Balancing of incoming requests to the

available MEC resources based on the predicted workload and b) the mobility estimation of a user's trajectory and position, and c) the effect of the wireless signal strength and the user's position on the offloading decision. The main conclusions that were drawn from the experimental results are:

- The impact of the load balancing technique is twofold. First, it is essential for guaranteeing the QoS metrics of time and secondly the over or under- utilization of the deployed VDU's is avoided.
- The resource profiling methodology determines feasible operating points for the image recognition service and facilitate the load balancing decision.
- The workload prediction method based on Kalman Filter has significant accuracy and predicts sudden spikes of workload.
- The estimation of the user's position and trajectory, exploiting the capabilities of an inexpensive IMU sensor, is precise and applicable to mobile phones too.
- All functionalities of the COSMOS controller are generic and applicable on other types of MEC/IoT-enabled applications.