NoQueue Real-Time Offloading Framework

Award # CNS-1329755 (UCLA), CNS-1329644 (CMU), CNS-1329644 (UCSD), and CNS-1329650 (UCSB)
Type: Frontier; Start Date: June 2014

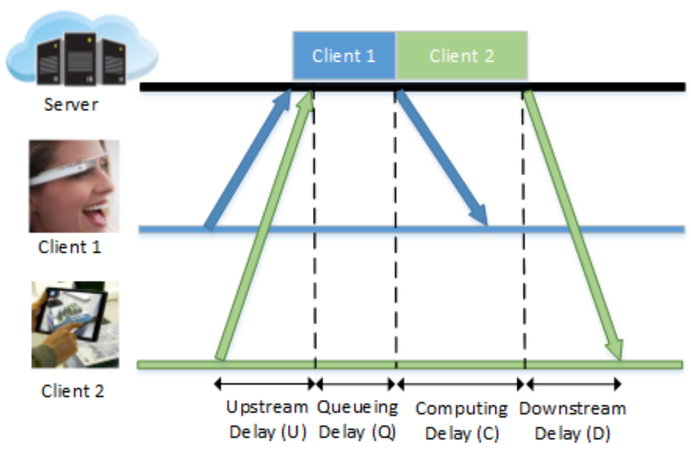
Zhou Fang (UCSD)
Co-Authors: Mulong Luo, Rajesh Gupta (UCSD)

Introduction

NoQueue is an offloading framework for serving real-time workload. Embedded devices benefit from offloading computation intensive tasks to more powerful servers in several aspects:

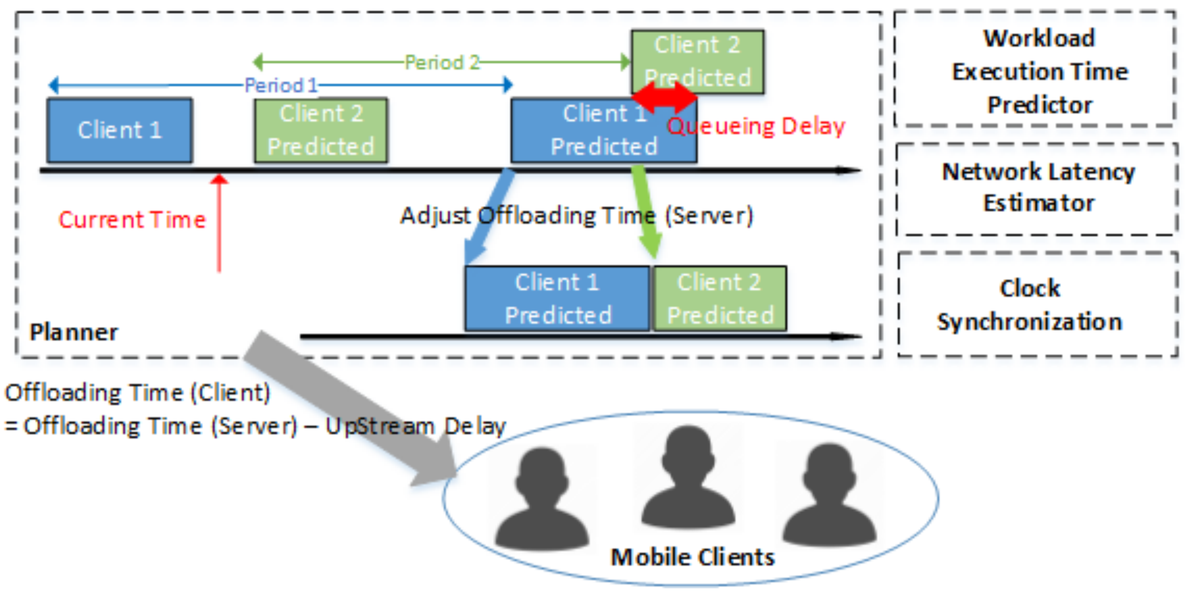
- data processing latency
- power consumption
- simplified embedded system development

Offloading real-time tasks is challenging because of unpredictable network latency (U, D), workload execution time variation (C) and multi-client interference on server resource (Q).



NoQueue adopts the ***predict-plan*** strategy to eliminate offloading task conflict to reduce server side queueing delay. It adopts **workload execution time prediction**, **clock synchronization** and **network latency estimation**. It also provides SLA to clients, manages server resource/admission control.

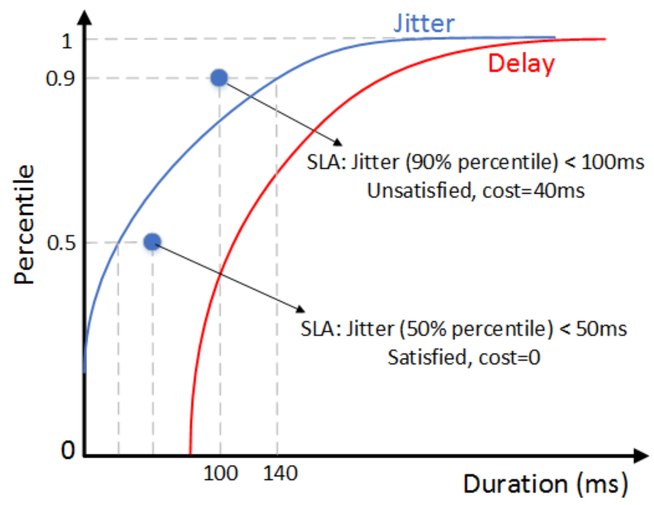
Design



- Mobile clients offload tasks to server with an adjustable period
- All client **clock**s are synchronized to server via NTP. Remaining offset is estimated (by NTP algorithm) and compensated in client timestamps.
- **Upstream delay (U)** is estimated by TCP retransmission timeout timer algorithm (RFC 6298)
- **Predictor** estimates workload execution time using time series linear regression (one model per client, training online or offline)
- **Planner** runs continuous simulation on future task conflict, adjusts future offloading time, and sends new times to clients

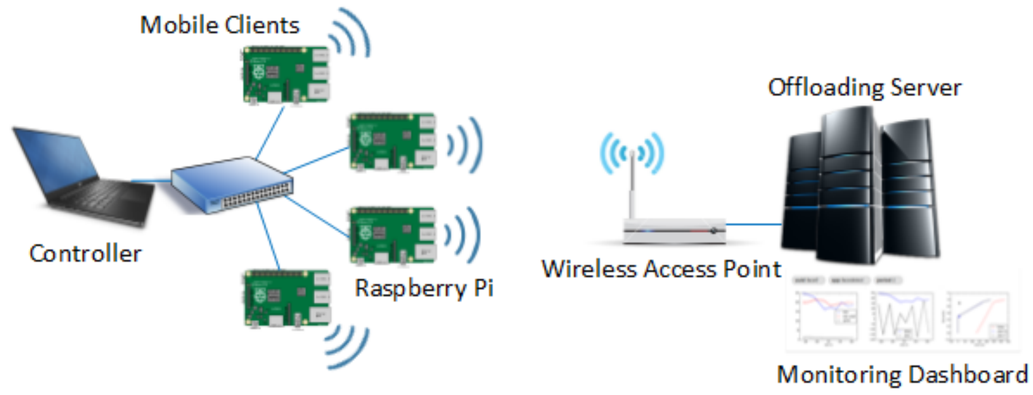
Service Level Agreement (SLA)

SLA is defined as a list of desired values on percentile of jitter of task interval. If measured metric exceeds SLA, the difference results in a cost. The maximal SLA cost becomes the client's cost.



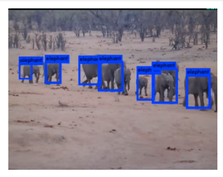
Planner adjusts offloading time based on client **cost**. The queueing delay to be removed is divided proportionally to adjust client1/client2.

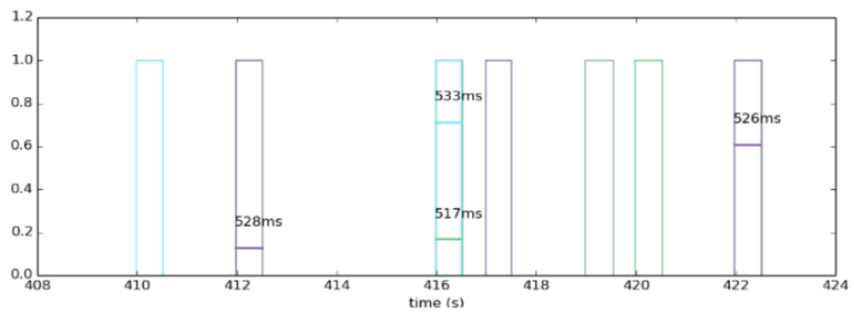
Testbed



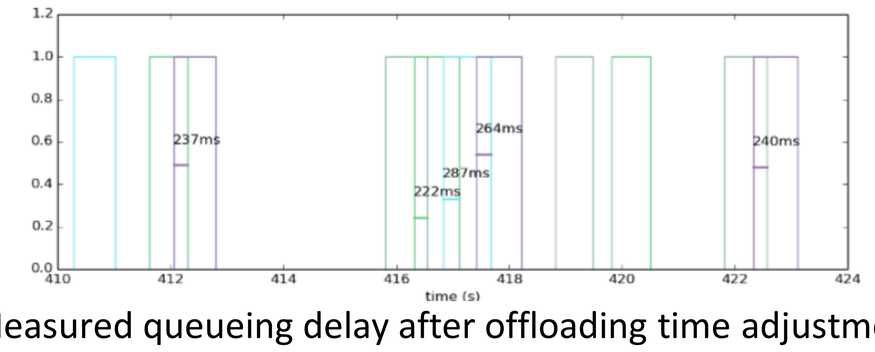
Dashboard

Testcase: Neural Network Object Detection (https://github.com/pjreddie/darknet)





Queueing delay predicted by Planner. Each pulse is the time slot occupied by an offloading task. Clients have different colors.



Measured queueing delay after offloading time adjustment.