4-2017

TANGO: Performance and Fault Management in Cellular Networks through Cooperation between Devices and Edge Computing Nodes

Saurabh Bagchi
Purdue University, sbagchi@purdue.edu

Nawanol Theera-Ampornpunt
Purdue University, ntheeraa@purdue.edu

Mostafa Ammar
Georgia Institute of Technology, ammar@cc.gatech.edu

Ellen Zegura
Georgia Institute of Technology, ewz@cc.gatech.edu

Tarun Mangla
Georgia Institute of Technology, tmangla3@gatech.edu

See next page for additional authors

Authors
Saurabh Bagchi, Nawanol Theera-Ampornpunt, Mostafa Ammar, Ellen Zegura, Tarun Mangla, Rajesh K. Panta, and Kaustubh Joshi

TANGO: Performance and Fault Management in Cellular Networks through Cooperation between Devices and Edge Computing Nodes

Saurabh Bagchi[Ψ], Nawanol Theera-Ampornpunt, Mostafa Ammar, Ellen Zegura, Tarun Mangla, Rajesh K. Panta, Kaustubh Joshi

***Abstract*—Cellular networks have become an essential part of our lives. With increasing demands on its available bandwidth, we are seeing failures and performance degradations for data and voice traffic on the rise. In this paper, we propose the view that fog computing, integrated in the edge components of cellular networks, can partially alleviate this situation. In our vision, some data gathering and data analytics capability will be developed at the edge of the cellular network and client devices and the network using this edge capability will coordinate to reduce failures and performance degradations. We also envisage proactive management of disruptions including prediction of impending events of interest (such as, congestion or call drop) and deployment of appropriate mitigation actions. We show that a simple streaming media pre-caching service built using such device-fog cooperation significantly expands the number of streaming video users that can be supported in a nominal cellular network of today.**

I. INTRODUCTION

Cellular networks have become a part of the infrastructure that we rely on without pausing to think of the enormous complexity that underlies such networks. These networks serve an amazing diversity of mobile devices and in an amazing diversity of radio environments. Failures in such networks are not uncommon and we take their consequences in our stride as a fact of life, grumbling about call drops and data disconnections. Similarly, we also encounter performance degradations, either in voice quality of calls or connection speeds. The amount of data on cellular networks has been increasing rapidly [1], driven in large part by the amount of streaming audio and video media that is being consumed by resource-rich end devices, such as, smartphones and tablets. This fact, together with crunch for available spectrum, will likely cause the incidences of faults and performance degradations to become more frequent and more severe in the near to mid-term future. In this paper, we use the term *disruption* or *disruptive event* to refer to the combination of faults and performance degradations.

There has been a trend to migrate some critical services to the cellular network (or at least partially use the cellular network), such as, the E-911 service for emergencies. With this migration, disruptive events can even be a matter of life-and-death. The overall issue of disruptions is exacerbated in densely crowded public venues, such as, sports and music venues and civic events. Additionally, the volume of media uploaded on cellular network is also growing, putting stress on the uplink capacity, which has tended to be lower than the downlink capacity. Some very public demonstrations of the effects have come from the City of Seattle requesting the patrons at a Seahawks game to limit their social media use (September 2014) and blackouts of cellular networks at disaster sites, such as the Boston Marathon bombing (April 2013).

In this paper, we propose the view that fog computing, integrated in the edge components of cellular networks, can partially alleviate this situation[1]. In this instantation of fog computing, some data gathering and data analytics capability will be developed at the edge of the cellular network. Client devices and the network using this edge capability will coordinate to reduce failures and performance degradations. We also envisage proactive management of disruptions including prediction of impending events of interest (such as, congestion or call drop) and deployment of appropriate mitigation actions.

The primary difficulty for proactive performance or fault management is that in today's networks, there is little collaboration between the end devices and the cellular network. To the end devices, the network is considered essentially as a black box and conditions about the network, such as, congestion or overload of some elements of the cellular infrastructure, are not made visible to the applications on the mobile end devices. Likewise, much of the application eco-system within a device is not made visible to the network. Thus, the network is not aware of the latency requirement of an application or the demand it is going to place on the network in the near future.

We, therefore, argue for a broad solution direction that enables proactive fault and performance management through cooperation between fog computing nodes in the network and the devices, a TANGO *of sorts.* Consider the following example of a user consuming a high-bandwidth media stream while being mobile. It is possible for the user to experience buffering and pauses due to network congestion. In the TANGO system, the cellular network can inform a "smart" streaming media

Saurabh Bagchi and Nawanol Theera-Ampornpunt are with Purdue University, e-mail: {sbagchi, ntheeraa}@purdue.edu. Mostafa Ammar, Ellen Zegura, and Tarun Mangla are with Georgia Tech, e-mail: {ammar, ewz, tmangla3}@gatech.edu. Rajesh K. Panta and Kaustubh Joshi are with AT&T Labs Research, e-mail: {rpanta, kaustubh}@research.att.com

[Ψ] Contact author: Saurabh Bagchi

[1]For ease of exposition, we will sometime shorten "fog/edge computing nodes in the cellular network" simply as "edge network".

application on the device when connectivity is predicted to be poor in the near future, based on the user's mobility pattern (say, she is going to be entering a tunnel) and current load in the network. The application can then initiate pre-caching of the content, based on the predicted usage profile. The fog layer would need to predict the network condition in the cells that the user is predicted to move to, model how the application behaves on the specific kind of user device in response to the specific level of network congestion, and possibly predict the mobility of the user. We show the results for a sample service that provides video content in an adaptive manner, adapting to near-future network conditions in cells the user is likely to move to. We see that for short videos (à la YouTube), TANGO with perfect location prediction can boost the network capacity in terms of the number of users that can be supported by 147%. For long videos (à la NetFlix), the improvement is 168%.

The rest of the paper is structured as follows. In Section II, we provide the relevant background on cellular networks, to understand where the elements of the fog computing layer can be positioned. In Section III, we put forward our broad vision for the solution. In Section IV, we provide the architecture of our proposed solution approach TANGO. In Section V, we describe our sample media streaming service followed by its evaluation in Section VI. In Section VII, we lay out the technical and the commercial challenges for the solution approach to become feasible. Then we conclude the paper.

II. Problem Background and Current Approaches

Here we first give a short background on the cellular network components and then discuss the solution approaches used in practice today to mitigate these problems.

A 4G LTE network consists of three major components—the User Equipment (UE), the Radio Acess Network (RAN), and the Core Network (CN). UEs are mobile devices which connect to base stations, also called Evolved Node Bs or eNodeBs, in RAN. The CN connects the RAN to the Internet. The Mobility Management Entity (MME) in CN is the control component responsible for handling signaling between UEs and CN. The Home Subscriber Server (HSS) contains users' subscription data and roaming information for managing movement of UEs. The Packet Data Network Gateway (P-GW) manages IP address allocation for UEs, and enforces Quality of Service and flow-based charges defined in the Policy Control and Charging Rules Function (PCRF). The Serving Gateway (S-GW) serves as the local mobility anchor when the UE moves between eNodeBs.

Ensuring good quality of service is challenging when millions of mobile devices and thousands of base stations are spread across a large geographic area with a wide range of radio environements. Currently, the UEs and the cellular network appear as black-boxes to each other with no collaboration between them. Taking the example of video streaming services (e.g. Dynamic Adaptive Streaming over HTTP or DASH, Apple HTTP Live Streaming, Microsoft Individualized-Integrated Book Smooth Streaming, and Adobe HTTP Dynamic Streaming), most support streaming videos at variable bit rates (*i.e.*, video qualities) using adaptive bit-rate (ABR) technology. The source video is divided into smaller chunks. The streaming server stores multiple versions of each chunk, pre-encoded at different bit rates. The video player running on the mobile device adaptively selects the appropriate bit rate based on network conditions and device capabilities. Simple prediction algorithms are used by mobile devices to estimate the network bandwidth that will be available to download the next video chunk, such as, using the average of the bandwidth observed over the last X measurement intervals. The ABR mechanism tries to maximize the video quality by choosing the highest bit rate the network can support without causing video pauses. The ABR system uses local information, available at the mobile device, to estimate future bandwith availability. However, the network also possesses some information that can be useful to improve video streaming. For example, each eNodeB has information about the number of concurrent video streaming applications present in a given cell at any given time. In [2], we show that, through collaboration between mobile devices and the network, this information can be conveyed to mobile devices to improve the performance of a video streaming application. Specifically, we show that the information about the total video demand in a cell can be used by mobile devices to coordiante their individual demands at different time instants to reduce the peak traffic in a cell, while still avoiding video pauses.

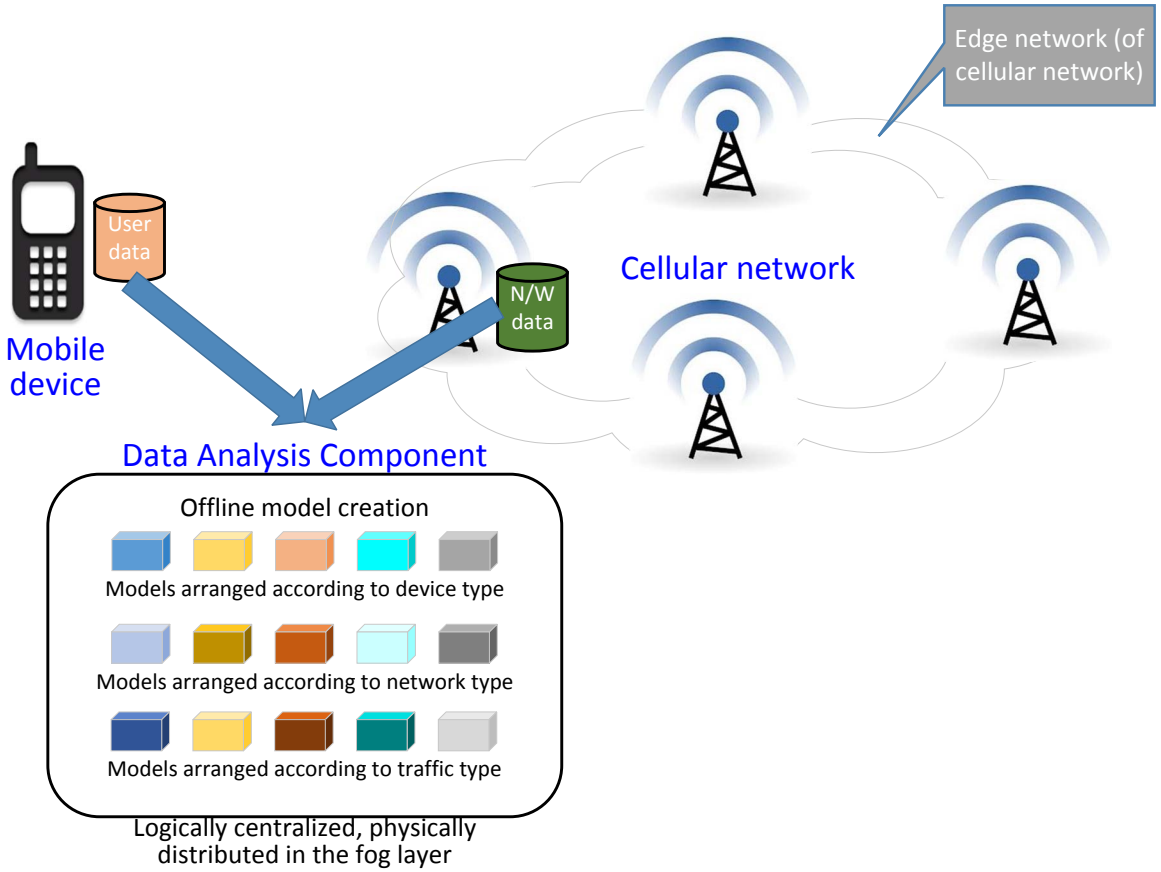
III. Solution Direction

For our broad solution direction of cooperation between the edge network and end devices for managing disruptions, there is a set of functionalities that need to be in place. We show a schematic of the two broad phases in Figure 1 showing the offline model building and the online prediction using the model.

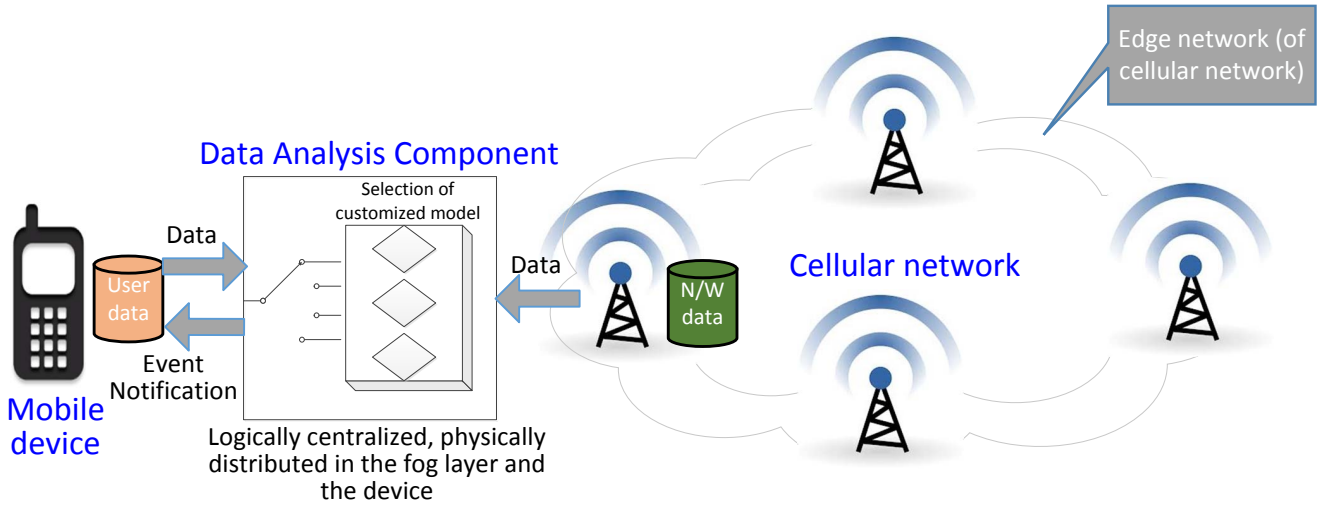
A. Components

First we discuss the different functional components that need to be in place for the solution direction that we are laying out.

1) *Data analysis at the network elements and device.* Cellular networks collect a wealth of data about individual mobile devices and aggregate quality of the network. This data is distributed among its various elements, such as, the Radio Network Controller or base stations. We posit that it is possible to do analysis of this data to determine events of interest that are likely to happen in the near future, such as, network congestion or disconnection of a device. It is likely that the data analysis models have to evolve over time due to the dynamic nature of the environment. In our framework of cooperation between the network and devices, the data analysis will also use information from the individual devices, such as, the nature of traffic being generated by each application, for example, to time shift some delay-tolerant traffic. One key requirement is that the data be available in real-time and the predictions from the data analytic models be also made in real-time. This



(a) **The offline training phase that builds the statistical data models that will be used to predict impending failures at runtime.**



(b) **The online failure prediction using customized models (per device type, per cell, per RNC, etc.) and event notification to devices for possible mitigation.**

Fig. 1: **The offline and the online phases of data analytic prediction of performance degradation or failure events.**

is so that impending disruptive events can be detected and possibly mitigated.

2) *Personalized data models*. In our prior work [3], [4], we have found that the data analytic models have to be built taking into account the spatial, temporal, and personalized conditions, in order to achieve high accuracy. Spatial implies, for example, that a model that is applicable to midtown Atlanta will not carry over to the Purdue campus, while temporal implies that a model that is applicable to the Purdue campus during a regular semester day will have to be recalibrated to develop a model for a home football game. Personalization of the model may be at different granularities—a separate model may be needed for each OS of the device, or at a finer granularity, each make and model of the device, and definitely for each class of cellular network, such as, 4G-LTE, 3GPP-LTE, or Mobile WiMax. The granularity has implication for how many models will need to be computed, stored, and retrieved at the fog computing layer; a finer granularity of models will increase the load for each of these.
3) *Mitigation actions.* The prediction of an impending event of interest should trigger a root cause analysis. The root cause will map to mitigation actions, either initiated by the mobile device or by an edge network element. Examples of such mitigation actions are: the application on the mobile device delaying a part of its traffic that is delay tolerant with the intent of easing the congestion; the network forcing a handoff of the mobile device from one base station to another to pre-empt an impending disconnection.
4) *Flexible mapping of functionality to the network edge*. There is a move toward executing some of the cellular network functions in virtualized environments, running on general purpose computing nodes, a trend that is described by the term *Network Function Virtualization* (NFV). We should map the functionality needed from the network flexibly on to the virtual machines running at the edge network. However, we need not aspire for completely flexible mapping in the sense that any function can go to any computing element. There are some co-locations which are obvious from an efficiency standpoint. For example, the functionality that tries to predict the mobility of an end device may well be co-located with the location registry of the cellular network. We anticipate that there will be hierarchical functionality provided, some at the very network edge which will be local to a geographical area served by a base station, and some aggregation functionality provided further inside the network edge, which will be less localized.

B. Proactive Management

Existing approaches to handle the problems laid out earlier are mainly reactive in nature, *i.e.*, *after* the failure or the performance degradation has happened and has affected the end user. Existing proactive solution approaches involve a combination of one or more of the following means.

- Upgrading the capacity of the cellular network, *e.g.*, adding new cell towers in the vicinity of high traffic areas. A more dramatic example would be upgrading the cellular network infrastructure itself, such as, from LTE to LTE-Advanced.
- Offloading traffic to local Wi-Fi hotspots. This has the potential to divert some traffic, though Wi-Fi coverage tends to be quite localized and this has reduced the potential impact of this measure.
- Adding temporary cellular capacity, such as, through micro or Femto cells near a congested area. This can be done in an agile manner when some source of congestion is expected to last for a specified period of time, such as, during a game day.
- Applying traffic management techniques in software that are meant to utilize the existing cellular capacity more efficiently. For example, a technique could be to prioritize emergency network traffic (such as, an E-911 call) over less critical network traffic.

The fog computing approach of TANGO is complementary to all of these approaches, with differing levels of effort needed for integration. While addition of cellular capacity, either transient or permanent, is easy to integrate, the offloading to Wi-Fi is more subtle in that it adds a mitigation action to the repertoire of TANGO.

IV. Our Proposed Solution Architecture

Figure 2 shows an overview of our proposed framework and how the existing device and network elements interact with it. The main components shown are: 1) applications, 2) Tango framework interface, separated between the device and the network parts, 3) service components, and 4) parts of the cellular network edge, the ones that provide valuable service to TANGO.

Applications on the device.

The applications follow the standard definition of mobile applications. The distinguishing characteristic is that they wish to be made aware of some characteristics of the network so that they can react to them, ultimately leading to a better quality of experience for the user of the applications. Different applications may query the network for different characteristics, e.g., one application may wish to find out the degree of congestion in the cell to which it is connected, while another application may wish to get historical information about the duration of the disconnection, when a call drop occur. The network is also free to not provide the response to some queries, whether due to competitive commercial reasons or because such information is not readily available.

Device and network components of the TANGO framework.

The framework comprises a device component and a network component, the former resides on the device while the latter resides on elements of the network edge, such as, the RAN. The device component stores data locally, processes such data, communicates with the network components in both uplink and downlink capacity, and communicates with the application.

The network component is physically distributed in the edge of the cellular network, co-located with existing elements of

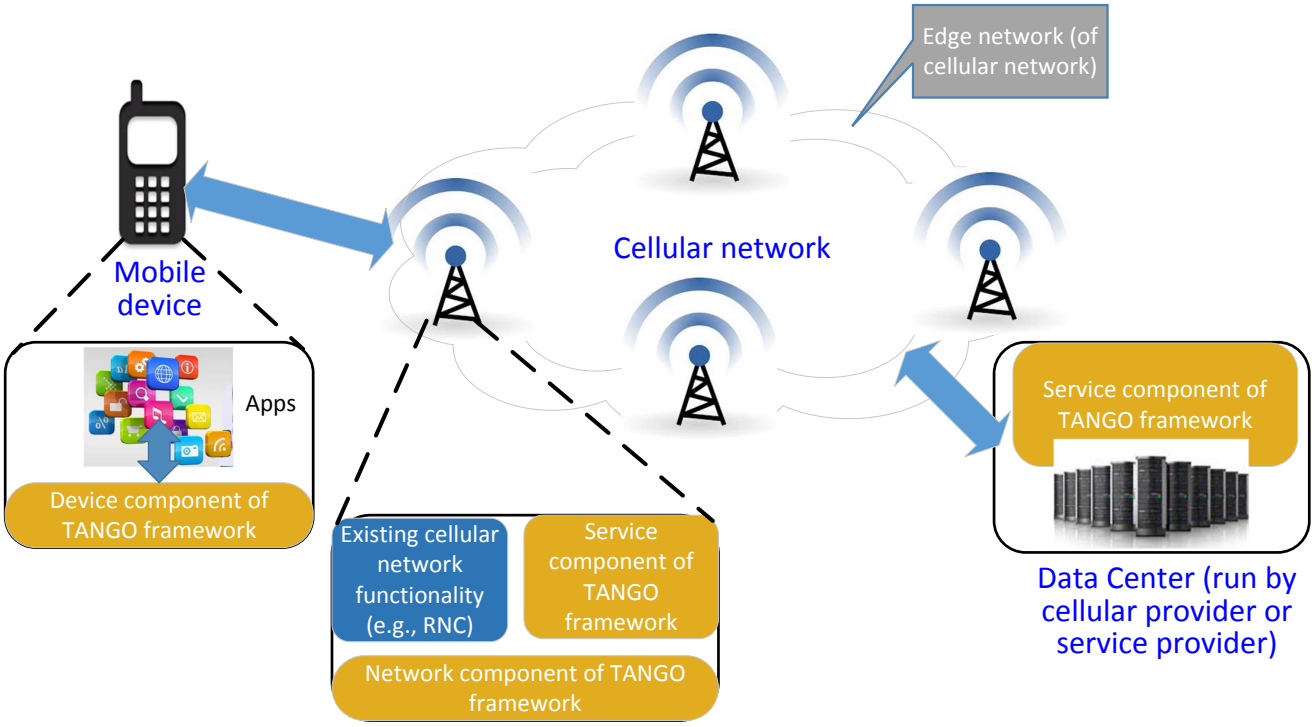


Fig. 2: **An overview of the elements of the TANGO framework and how the device and the network components interact with them.**

the current cellular network services, such as, the eNodeB, RNC, or RAN. For operational reasons, the network component elements may execute on a separate Virtual Machine (VM) on the same physical machine as the existing network services, or on closely located physical machines. Since we would like to support a large number of concurrent users (potentially everyone in the network), we must be able to divide the load among different instances of service components. The framework supports partitioning by location (based on the current primary cell of the device) and by a hash of the device's identifier.

Service component of the TANGO framework.

The service component may exist on the edge of the network or backend servers provided by the cellular network provider, the latter typically for resource intensive services. The service component provides a myriad of services that may be desired by the users of the cellular network. Sample services could come from the *entertainment sector*, such as, high definition video, *industrial sector*, such as, asset tracking for mobile assets (as the number of movable assets proliferate such as tablets and laptops, industries struggle to keep good inventory of these), *mobile commerce sector*, such as, more secure purchasing options that correlate the location of the user with the location of the transaction, and *advanced communications sector*, such as, context-aware telecommunications service (*e.g.*, if a user is in a privacy sensitive zone, then video will be blurred or disabled). The services do not form a fixed set but there will be some churn in them as commercial interest may cause some new services to be unveiled and some services to be discontinued. The service component interacts with the network component to provide the added functionality and in some sense, is not as closely tied in to the cellular network. For example, a service may be to provide estimate of the load on the network. This relies on historical load data from the cellular network, but also exogenous information such as popular events in the locality. An application registers with a particular service and the service component is responsible for figuring out which instance of the network component it needs to interact with to provide the requisite functionality. In terms of ownership, while we initially expect that the service components for various services will be provided by the cellular network service provider, in the future, we anticipate a more distributed ownership model. In this model, different service providers can stand up their own services and have a commercial contract with the cellular network service provider to tap into the resources of the network component.

V. Congestion Avoidance in Media Streaming

In this section, we describe a service that takes advantage of cooperation between mobile device and cellular network in order to improve user experience.

During network congestion, applications that have high bandwidth requirements such as multimedia streaming are severely affected. Currently, streaming applications rely on buffers and bit-rate adaptation to combat against temporary connectivity problems such as congestion. While larger buffers are effective in preventing playback disruption caused by congestion, it incurs higher bandwidth cost in the case where

the user does not watch or listen to the entire video or song. Ideally, the buffer should be large when poor connectivity is expected, and small when connectivity is good. Since the network knows about its current load in the various cells, it is in a position to predict a user's connection quality. On the other side, the application on the device knows important playback states such as current buffer level. Thus, with cooperation between the mobile device and the network, the buffer size can be dynamically adjusted to be appropriate for the current condition. We call this the *Data Pre-caching Service*.

To provide enough lead time for mitigation action, the Data Pre-caching Service continuously predicts the user's future location based on the current trajectory. When a user is predicted to enter a congested area in the near future, the service sends a pre-caching alert to the streaming application. The application can then increase its buffer size, in order to mitigate the effect of impending congestion. The service runs on some infrastructure at the edge of the cellular network, where it can closely monitor the local network status.

In order to avoid the energy overhead of frequent GPS measurements, the mobility prediction model uses current and past cell sectors as input (which are known to both the device and the network), and gives a list of potential future cell sectors the user is likely to visit as output. The interested reader can find further details of this service in [4].

VI. Evaluation of Congestion Avoidance in Media Streaming

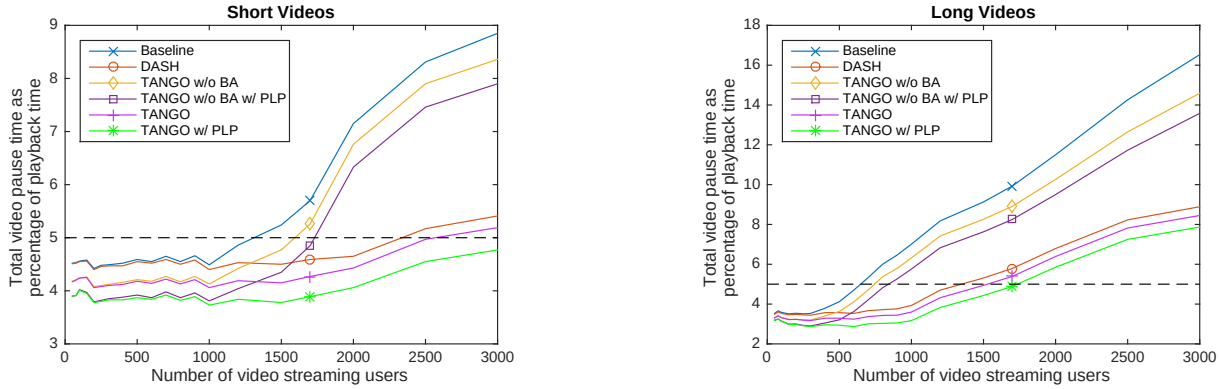
We evaluate the data pre-caching service by quantifying how effective it is at reducing the amount of video pause time due to rebuffering, in the presence of congestion. We rely on simulations which enable us to model situations where thousands of streaming users are present in a small area. Streaming users are modeled using two types of workload, short videos averaging 5 minutes (similar to say YouTube viewing habits) and long videos of length 1 hour (similar to say Netflix viewing habits). Congestion affecting a cell is simulated by setting *the background traffic* for the cell to 95% of its maximum bandwidth throughout the simulation. We simulate congestion due to such background traffic in 5% of the cells. The additional traffic due to video streaming causes additional congestion in a variable fraction of cells.

In the simulation, each video streaming client's playback and buffer states are tracked. Cells are simulated as having fixed bandwidth. We use real traffic traces from the cellular network of a major US-based cellular network provider for the simulated users' mobility pattern as well as background traffic. The area covered is 1.66 miles by 1.91 miles in downtown San Francisco, USA over a 1 hour period during the afternoon rush hour. Users watching short videos follow abandonment rate described in [5], while users watching long videos watch the video until the end. In the short video case, pre-caching content for the next video relies on a hypothetical *next video predictor*, which gives a list of 3 potential videos as the next video the user will watch, with 80% chance that one of the 3 videos will actually be watched next. The predictor's prediction accuracy directly affects the effectiveness of the data pre-caching service since pre-caching an incorrect video will only increase bandwidth usage without helping the pause time.

We compare six different approaches: 1) baseline, 2) DASH, 3) TANGO without bit-rate adaptation, 4) TANGO without bit-rate adaptation but with perfect location prediction, 5) TANGO, and 6) TANGO with perfect location prediction. The buffer size is the same for all approaches: 30 seconds for short videos and 240 seconds for long videos. In baseline, the player simply tries to keep the buffer full at all times. In DASH and TANGO variants with bit-rate adaptation, the video bit-rate is chosen from four possible values, 300, 400, 600, and 1000 Kbps, based on download rate in the past 20 seconds. TANGO variants with perfect location predictor are included in order to show the significance of mobility prediction in TANGO. When a pre-caching alert is sent to the player in TANGO, the player increases the buffer size so that the current video as well as all 3 predicted next videos are buffered. When the alert is no longer active, it switches back to the original buffer size.

The simulation results for short videos are shown in Figure 3a. On average, 18% of simulated users are active at any point in time. For example, for the point on the X-axis corresponding to 1,000, there are on average 180 users concurrently downloading video. With the typical tolerable pause time of 5% (of total playback duration), the current cellular network can only support 1,311 users, while TANGO, even with imperfect location prediction, can support 2,568 users, an increase of 96%. Two distinct patterns are clearly seen for two regions of the number of users. When there are less than 1,500 video streaming users, bit-rate adaptation barely provides any benefit, while TANGO reduces video pause time by roughly 10%. The benefit increases significantly when TANGO is paired with perfect location predictor. When there are more than 1,500 video streaming users, however, bit-rate adaptation becomes critical. It also reduces the rate of growth of pause time with the number of users, thus the cellular network is able to support a larger number of mobile video users. With 1,500 or more video streaming users, a significant fraction of cells are congested, so there is less opportunity for TANGO without bit-rate adaptation to move traffic from congested areas to uncongested areas. For example, at 2,000 video streaming users, on average 29% of the cells with users are congested. On the other hand, bit-rate adaptation is effective as it reduces the video quality and thus the required bandwidth, letting traffic from more users fit into the same cellular network.

The simulation results for long videos are shown in Figure 3b. On average, 50% of simulated users are active at any point in time, compared to 18% for the short video case. The patterns are similar to those of short videos. However, the network's capacity is reached earlier, at roughly 650 video streaming users. This is because the session length of users watching short videos tends to be shorter than those watching long videos and thus long videos present more concurrent users. With the tolerable pause time of 5%, the current cellular network can only support 648 users, while TANGO, even with imperfect location prediction, can support 1,527 users, an increase of 136%.



(a) **Video pause time for short videos**

(b) **Video pause time for long videos**

Fig. 3: **Video pause time (as percentage of total playback duration) as a function of number of video streaming users for long-video workload. BA refers to bit-rate adaptation, while PLP refers to perfect location prediction. Dashed line indicates users' tolerance of pause time.**

VII. The Road Ahead

The vision that we have laid out for the cooperation between mobile devices and the edge of the cellular network poses several technical and commercial challenges. For some of these, we can try to adapt existing solutions to other problems and for some, we need to create new solutions. Below we provide our view of these challenges.

Systems management. The system view that we have laid out makes for a daunting systems management problem. It brings together in one system two big challenges of systems management—a large number of devices that need to be managed (the user equipments), and distributed nature of the elements to be managed (the devices themselves, plus also the network component and the service component). We expect that declarative high-level systems management modules will need to be developed for handling this confluence of challenges. Declarative means the administrators will be able to specify ("declare") desirable properties from the system and a compiler or interpreter will translate this to the procedural detailed instructions to achieve these properties. Further, the properties must be specifiable at a high level using rich semantics. For example, it should be specifiable that some class of network information will *not* be available to a device when it is roaming on a cellular provider's network that is different from the home network of the device. There is inspiration to be drawn from declarative systems management that has been applied for large-scale tasks [6].

Control Architecture. Our envisioned system requires control and coordination functions to enable allocation and orchestration of system resources and entities. In some instances, real-time, adaptive control is required while in others, longer term coordination is sufficient. These functions are necessary to insure providing efficient fog processing capabilities. Control functions can be distributed among the system entities or they can be centralized in a centralized system controller. In the latter case, the question of where such control should be instantiated and how to insure its scalability and fault tolerance become important. Control architecture alternatives need to be designed in a manner that insures low overhead.

Privacy. In our system model, some user-specific information, *i.e.*, what is resident on the device, is being made available to the fog layer of the cellular network. This raises obvious privacy questions. While in the current cellular networks, user information is already available to the cellular network provider, in TANGO that information is available to a distributed set of entities. It is of course possible to encrypt the personalized information but this severely curbs the kinds of queries that can be run against such information and thus, the kinds of services that can be a part of TANGO.

Economic incentives. The economic incentives must be aligned to create the cooperation between the devices and the cellular network. We envisage that this will take two broad forms. First, the third-party service providers that are developing services using the network and the device information will compensate the cellular provider. Second, a better management of potential disruptions (hard faults as well as performance degradations) for the cellular devices will improve the quality of experience for the consumer, while using the cellular provider's available spectrum more efficiently. Under these two broad forms, there still remain to be worked out tiering of the services and the pricing strategies. For example, service level agreements may be made available guaranteeing a certain application quality of service (such as, percent of pause time for streaming media), in which case some shared responsibility is implied between the application provider and the cellular network provider.

VIII. Conclusion

In this paper, we have laid out a vision for proactive management of performance problems and hard failures in cellular networks. Such a vision is realizable through cooperation between users' mobile devices and computing at the edge of the cellular network, through data collection and data analysis built into the fog computing layer. We describe the components of a potential architecture, which we call TANGO. We outline challenges centered around technical questions, privacy, and economic incentives that will need to be solved to make the vision a reality.

ACKNOWLEDGMENT

The authors acknowledge the support from the National Science Foundation through the NeTS program (grant numbers CNS-1409506 and CNS-1409589) as well from AT&T through their Virtual University Research Initiative (VURI), which were used to carry out the activities described in this paper. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Ericsson, "Ericsson Mobility Report: On the Pulse of the Networked Society," https://www.ericsson.com/res/docs/2015/mobility-report/ericsson-mobility-report-nov-2015.pdf.

[2] C. Shi, K. Joshi, R. K. Panta, M. H. Ammar, and E. W. Zegura, "Coast: collaborative application-aware scheduling of last-mile cellular traffic," in *Proceedings of the 12th annual international conference on Mobile systems, applications, and services (Mobisys)*. ACM, 2014, pp. 245–258.

[3] Y.-S. Wu, S. Bagchi, N. Singh, and R. Wita, "Spam detection in voice-over-ip calls through semi-supervised clustering," in *2009 IEEE/IFIP International Conference on Dependable Systems & Networks*. IEEE, 2009, pp. 307–316.

[4] N. Theera-Ampornpunt, T. Mangla, S. Bagchi, R. Panta, K. Joshi, M. Ammar, and E. Zegura, "TANGO: Toward a More Reliable Mobile Streaming through Cooperation between Cellular Network and Mobile Devices," in *IEEE 35th Symposium on Reliable Distributed Systems (SRDS)*. IEEE, 2016, pp. 1–10.

[5] A. Ali-Eldin, M. Kihl, J. Tordsson, and E. Elmroth, "Analysis and characterization of a video-on-demand service workload," in *Proceedings of the 6th ACM Multimedia Systems Conference*. ACM, 2015, pp. 189–200.

[6] T. L. Hinrichs, N. S. Gude, M. Casado, J. C. Mitchell, and S. Shenker, "Practical declarative network management," in *Proceedings of the 1st ACM workshop on Research on enterprise networking*. ACM, 2009, pp. 1–10.