Title	Using context-awareness for storage services in edge computing
Author(s)	Pérez-Torres, Rafael; Torres-Huitzil, César; Truong, Thuy; Buckley, Donagh; Sreenan, Cormac J.
Publication date	2021-03-31
Original citation	Pérez-Torres, R., Torres-Huitzil, C., Truong, T., Buckley, D. and Sreenan, C. J. (2021) 'Using Context-Awareness for Storage Services in Edge Computing', IT Professional, 23(2), pp. 50-57. doi: 10.1109/MITP.2020.3043164
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://ieeexplore.ieee.org/document/9391745 http://dx.doi.org/10.1109/MITP.2020.3043164 Access to the full text of the published version may require a subscription.
Rights	**© 2021, IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works**
Item downloaded from	http://hdl.handle.net/10468/11213

Using Context-Awareness for Storage Services in Edge Computing

Rafael Pérez-Torres, César Torres-Huitzil, Thuy Truong, Donagh Buckley, and Cormac J. Sreenan

Abstract—Modern mobile networks face a dynamic environment with massive devices and heterogeneous service expectations that will need to significantly scale for 5G. Edge Computing approaches aim at enhancing scalability through strategies like computation offloading and local storage services, which will be fundamental to deploying large-scale distributed applications. Unlike the cloud, edge resources are limited, which calls for novel techniques to handle large volumes of up- and down-stream data under a changing environment. Being closer to data consumers and producers, a compelling view is to adopt context-aware techniques for enabling the edge to work with patterns from mobile traffic at different spatio-temporal scales. In this article we overview the challenges and opportunities of edge storage from the perspective of context-awareness. We introduce a conceptual architecture to learn and exploit context information for enhancing uplink and downlink scenarios. Finally, we outline future directions for edge applications.

Index Terms—Edge Computing, Context-awareness, Edge Storage, Internet of Things, Multimedia Delivery

1 Introduction

Mobile networks are shifting from an early focus on connectivity towards content-centric communications with user and machine-type exchanges in a 5G world. A myriad of devices will transmit data and demand services anytime, anywhere, with growing expectations regarding Quality of Service (QoS) and Quality of Experience (QoE). This irruption of connected devices creates a dynamic environment that will burden the capacity of mobile networks [15], [17].

Edge computing aims at overcoming these issues by leveraging the network's communication, computing, and storage resources. Edge computing refers to the enabling technologies for computation at the edge of the network, operating on downstream data from cloud services and upstream data on behalf of mobile devices known as user equipments (UEs) [10]. The edge concept is embodied in the ETSI *Multi-Access Edge Computing* (MEC) standard for latency reduction, location awareness, real time network telemetry, and energy savings across hosts [12]. MEC is a critical technology for 5G [1].

Conducted research has focused on edge computing and communications, with MEC storage assisting many of these solutions and creating innovative Edge Storage Services (ESSs) like mobile personal clouds (e.g. Dropbox) and mobile content delivery networks (*CDNs*). ESSs rely on channels for content distribution, shown in Fig. 1, and its integration with computing functionalities is required, as mere transmission speed improvements will not fulfil the demands of future information-centric networks [12].

The mobile Internet challenges ESSs due to its rich context comprising traffic bursts from Internet of Things (IoT) devices, changing wireless conditions, and users with varying routines and mobility [3]. Context refers to any information that characterises the situation of an entity [7] for decision making. Conventional storage management was not designed for this context-rich and dynamic traffic conditions. Cloud storage operates without the resource constraints and changing conditions at the edge, while edge services can benefit from local context information. Thus, we advocate the inclusion of context-awareness as a core design feature in ESSs to enable new edge applications and business models for Mobile Network Operators (MNOs) [3].

Here, we study representative case studies for the use of context-awareness in edge storage. Furthermore, our main contribution is a conceptual distributed architecture that learns and exploits context-information from mobile traffic to address several issues faced by ESSs. Finally, we outline future directions for context-awareness in ESSs.

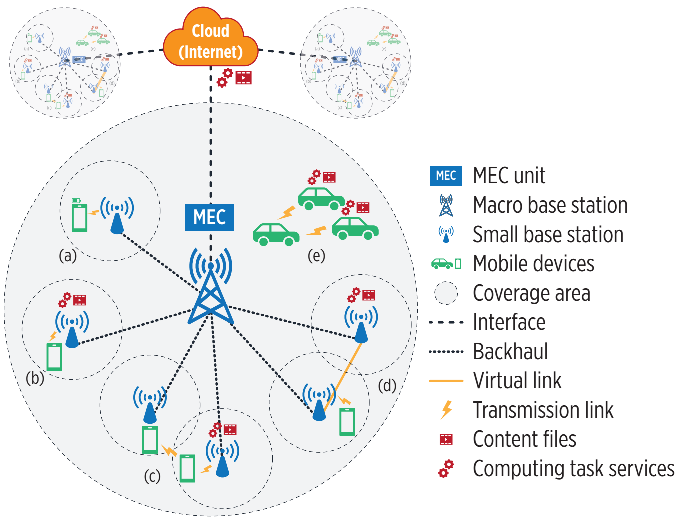
2 Case studies for edge storage

Edge computing enables a wide range of downlink- and uplink-specialized ESSs (shown in Fig. 2), with multimedia delivery and IoT systems as representative cases [8]. Context-awareness, specifically mobility and content popularity, is the key to improve their decision making.

2.1 Multimedia systems

Multimedia systems focus on downlink content delivery to mobile devices. Nevertheless, massive requests and the dynamic conditions of wireless channels make mobile networks struggle. Although emerging techniques address some of these issues (e.g., DASH for video streaming),

- *R. Pérez-Torres (corresponding author) and C. J. Sreenan are with the School of Computer Science & IT, University College Cork, Cork, Ireland. E-mail: {r.perez,cjs}@cs.ucc.ie*
- *C. Torres-Huitzil is with the Instituto Tecnológico y de Estudios Superiores de Monterrey - Campus Puebla, School of Engineering and Sciences, Puebla, Mexico. E-mail: torresc@tec.mx*
- *T. Truong and D. Buckley are with OCTO Research and Strategy Office, Dell EMC Research Europe, Cork, Ireland. E-mail: {thuy.truong,donagh.buckley}@dell.com*



(a) A mobile device requests content or offloads computing tasks to the cloud.
(b) The content is provided by a small base station thanks to a caching strategy.
(c) A multi-hop strategy to deliver content using paths different than the local small base station.
(d) Content is delivered from another small base station by using a macro base station as intermediator, creating a virtual link.
(e) Services are provided using Device to Device (D2D) communication, avoiding Infrastructure-based communication.

Fig. 1. Channels for computation offloading, content sharing and distribution in modern mobile network architectures (Adapted from [9]).

further QoS improvement for individuals and groups is needed. For example, ESSs can employ content popularity over edge servers to reduce latency by caching videos under an *edge CDN* approach. Moreover, users-content attributes like the type, location and time of content requests can help to characterize mobile traffic. Group-tailored content recommendations enable prefetching services at community level to step up ESSs performance. Similarly, mobility-aware prefetching can instruct where and when to push content to maintain content hit rate and reduce traffic.

2.2 IoT systems

IoT systems hold attributes that defy typical data management strategies: a) massive data, b) short and frequent uplink transmissions in a many-to-one fashion, c) strict latency requirements, d) data that quickly expire, and e) devices with varying mobility.

For instance, Connected and Autonomous Vehicles (CAVs) feature sensors to collect data for varying purposes like autonomous driving, insurance evidence, etc. Since CAVs can act as distributed sensor hubs [3], it makes sense to assemble them in a *reverse CDN* with edge caching to reduce the latency and energy consumption in uplink transmissions [7]. Context-awareness can contribute to locally adjust cache techniques by exploring on the transient IoT data streams using the distributed edge computing resources. Yet, CAVs, as both content producers and consumers, require simultaneous uplink and downlink management. Their density and mobility will generate dynamic traffic in time and space, leading to resource congestion and underuse across cells, calling for mobility-aware caching.

3 Context-awareness for improved edge storage

Unlike a centralised cloud, context information from mobile traffic can help ESSs to improve the use of resources across the network [1]. As shown in Fig. 3, the UE and network hosts provide different features that act as the building blocks for a wide range of edge applications at individual or group scales. Similarly, the traffic data available for MNOs has personalized, real-time, multi-sensory and spatio-temporal context features regarding habits like transportation and social interaction [3]. Furthermore, the edge has access to information regarding the quality of link connections to UE. Although MNOs analyse traffic only for billing and basic network management, it has potential for decision making, as recently explored in WSN and IoT services [7].

3.1 The synergy between context-awareness and edge computing

The massive mobile traffic prevents cloud-centralised approaches for context inference. Edge computing can assist on this, reducing backhaul overhead through distributed online data analytics and storage [7], [1]. The edge is a natural fit as it promptly captures local attributes of a) app-level data (users and requests), b) content metadata (popularity, request times), and c) network-level data (changes in wireless links and edge resources). Context resembles pictures taken by panoramic cameras for city-scale sensing [17], which reveal patterns to improve ESSs performance [10]. We argue that to significantly enhance resource use, the edge must become intelligent through local analytics assisted by global insights from the cloud.

3.2 Spatio-temporal attributes in mobile traffic

The intertwined spatio-temporal attributes of users and content requests allow to characterise the strong temporal periodicity and geographic locality of mobile traffic [17]. Although spatio-temporal features enhance mobility and network traffic prediction, they have not been widely studied for edge resources management [15].

The spatio-temporal data can be analysed at different scales: at the edge, to detect immediate events regarding small zones and individuals, and at the cloud, focusing on long-term patterns from crowds over larger zones.

3.3 Context-awareness beyond ESSs

There is an increasing interest on self-decision making in mobile networks to reduce the dependency on human operators (Knowledge-Defined Networking). Self-learning and self-decision making allow enhanced performance not only for ESSs but for other services in 5G networks [6]. While self-learning enables the automatic inference of patterns, self-operation can exploit them for adjusting to predicted changes in mobile traffic and link quality. Such predictions are the base for robust data management strategies like link-aware caching, popularity-based caching, and mobility-aware prefetching.

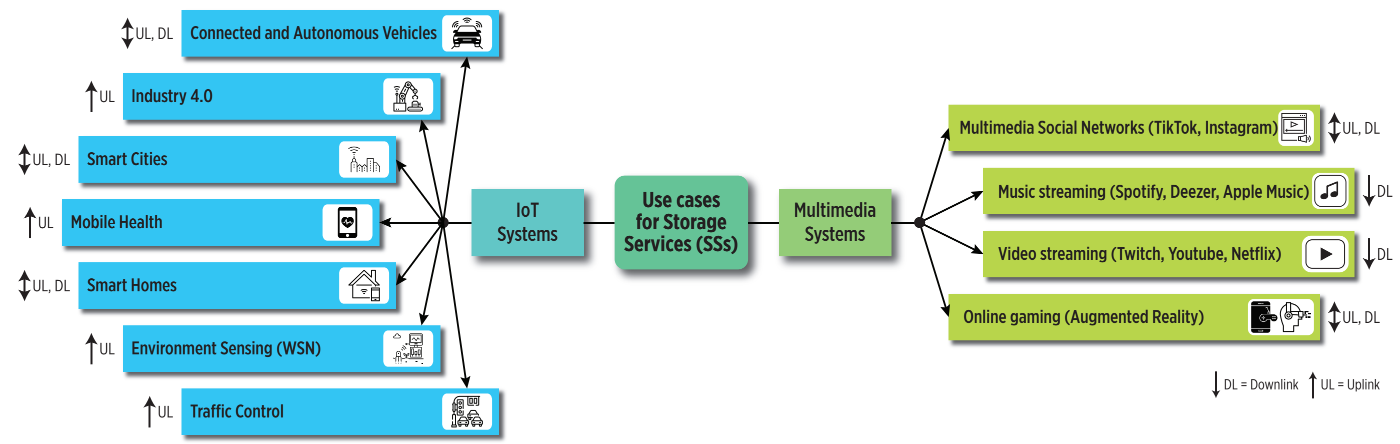


Fig. 2. A taxonomy of use cases for ESSs. ESSs can be categorized depending on their main data traffic direction into downlink- and uplink-specialized.

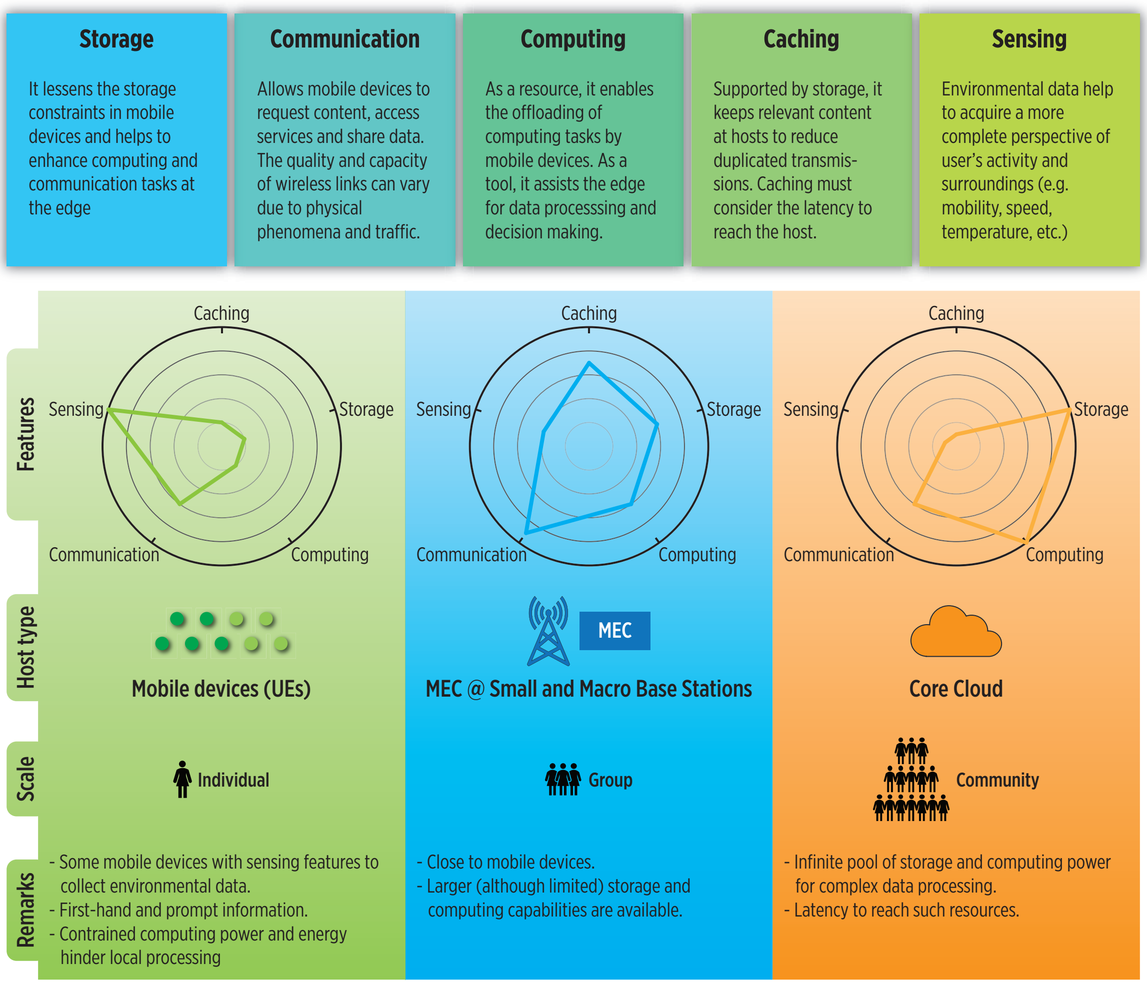


Fig. 3. The features of connected entities across the hierarchy of edge systems. Each level focuses on analytics for individuals or groups according to the size of the covered area.

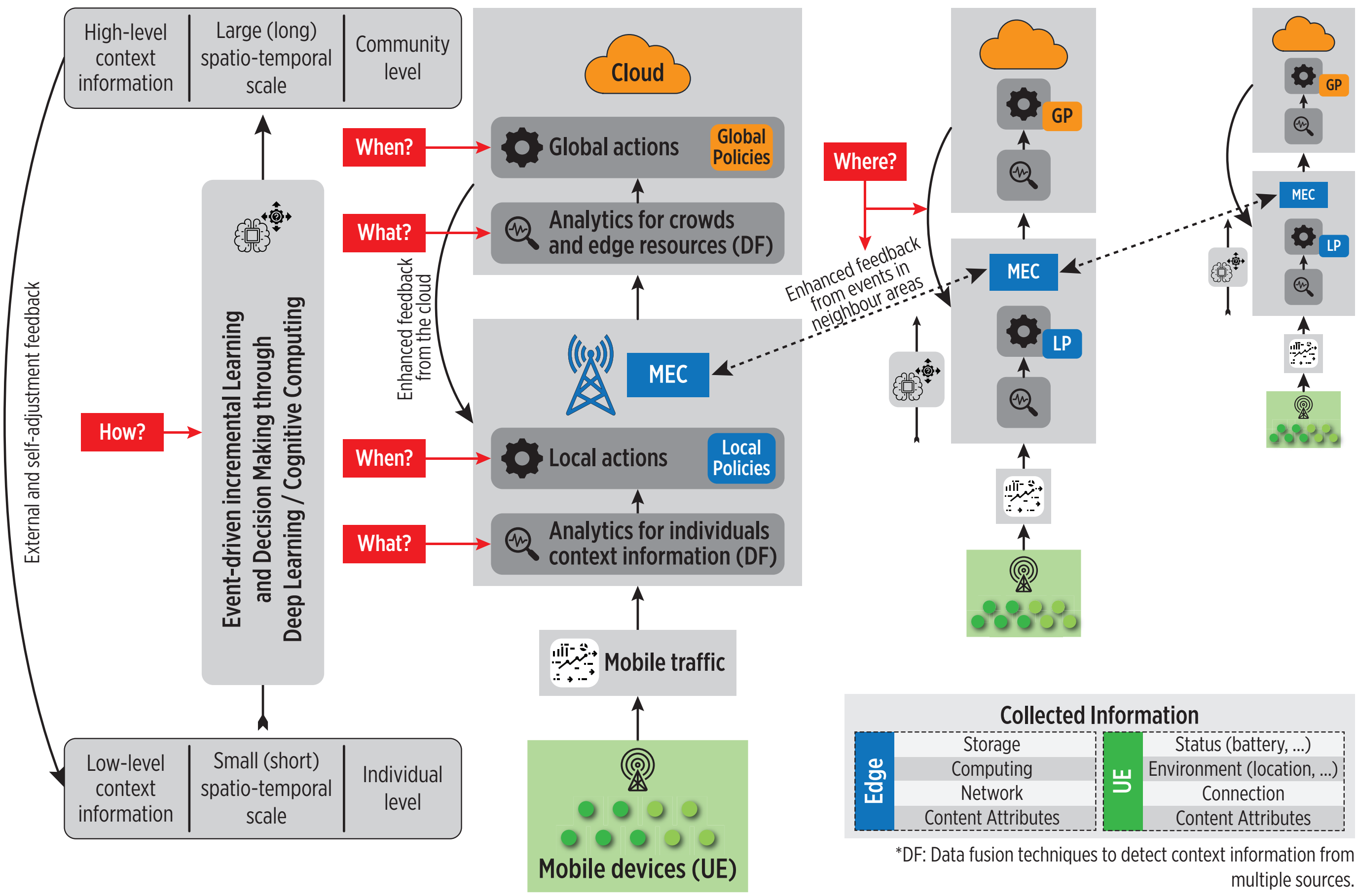


Fig. 4. A conceptual architecture for context-aware ESSs, including the simultaneous and incremental learning and exploitation of context information, and cooperative edge servers for information sharing and distributed control. This platform can be deployed as an edge middleware accessible through an API using Virtualized Network Function (VNF) and AI-enabled devices.

4 The Challenges for Deploying ESSs

Conventional techniques used by ESSs (e.g. buffering, caching) come from memory management, peer-to-peer, and distributed systems, which ignore the dynamics of mobile Internet regarding a) the wireless medium, b) devices' mobility, c) content *virality*, d) uplink transmission of transient data, and e) devices' capabilities. We propose the context-aware and distributed architecture shown in Fig. 4 to address these changes and discover *what*, *when*, *where*, and *how* to allocate content. Our event-driven architecture incrementally learns and exploits context information from mobile traffic to enhance ESSs operation. Figure 4 highlights the actors and issues, and the context information types captured by our architecture for reacting to trends in mobile traffic. Our architecture helps to answer the previous questions as follows:

- **What?**: We rely on local and global analytics for individuals and groups to detect content popularity changes in both uplink and downlink. Storing popular content reduces delivery latency and backhaul overhead. Link-aware caching helps to select the quality of cached content according to the quality of links to UE.
- **Where?**: Our architecture can implement mobility-aware prefetching using mobility information and inter-edge feedback, allocating content at the predicted location of users [15]. Recall that the higher in the network hierarchy the content is stored, the more users it serves at the cost of an increased latency.
- **When?**: In our architecture, local and global actions can trigger reactions after receiving requests (reactive caching), or even before through prefetching (proactive caching). This impacts on how fast ESSs react to traffic trends. Prefetching calls for predictive features [16] focused on popularity and UE's mobility [3][12] (e.g., when and where content will be requested?).

How to orchestrate content allocation is the most complex design feature of an ESS as it must simultaneously coordinate the interaction of its modules and the collection of input data. This enables a) flexible and dynamic content routing, b) detection of relevant events, and c) decisions to make under conflictive scenarios.

Data collection is possible through polling and event driven approaches. In the polling approach, modules frequently ask for detected events, which can lead to unnecessary requests and overhead. Under the event-driven approach, a publish-subscribe strategy notifies modules once events are detected. Although more complex, we follow the event-driven approach as it supports the asynchrony in traffic events, consumes less energy than polling (modules only activated when needed), and favours the incremental

TABLE 1
Relevant framework approaches for resource optimization, context inference, and context exploitation in mobile networks.

Approach family	Pros	Cons	Example technique	Target use	Application
Optimization	* Based on strict mathematical models, producing actual minimized values.	* Make strong assumptions about objective functions. * Disregard the uncertainty and dynamics of mobile traffic [14].	Alternating Direction Method of Multipliers (ADMM)	Resource optimization	Maximizing QoS and QoE of video streaming [5], maximizing MNOs revenue by reducing the use of hired network, storage, and computing [13], [11].
Deep Learning	* Benefit from massive and unlabelled data. * Automatic feature extraction uncovering complex correlations in mobile traffic.	* Low interpretability of decisions (black box model). * Hyperparameters configuration. * Computing demands.	Convolutional Neural Networks (CNN)	Context inference	Mobile traffic classification, spatial mobile analytics for trajectory prediction [17].
			Recurrent Neuronal Networks (RNN)	Context inference	Network-wide spatio-temporal data modelling [17].
			Long Short-Term Memory Networks (LSTM)	Context inference	Mobile traffic forecasting [17].
			Deep Policy Gradient, Deep Q-Networks (DQN)	Context exploitation	Dynamic orchestration of networking, caching, and computing resources [17].
Cognitive Computing	* Allow simultaneous learning and exploitation at a scale. * Explicit features are learned in a short- and long-term memory, which applications can exploit for resource self-decision making. * A customizable perception-action cycle controls the interaction between humans and devices [4].	* The relevant events for the system must be individually defined. * Individual pattern recognition techniques must be selected to control perception of events.	Cognitive Dynamic Systems	Context inference and exploitation	ESSs for IoT data processing and storage [4], health care systems [2].

processing of data. System coordination is possible through:

- Fixed approaches: With optimization techniques that solve resource models and the impact of requests (e.g., storage and computing) using heuristics.
- Policy-based: Using parameterised policies to react to changes in edge resources (e.g., low storage). Multiple policies can handle complex scenarios, although disambiguation measures must exist to solve conflicts.
- Autonomous operation: An objective (e.g., increasing cache hit) is incrementally achieved, self-adjusting according to taken decisions.

Our architecture supports combining these approaches, e.g. in an autonomous ESS that generates input data for policies and self-adjusts according to decisions made.

5 Frameworks for context-awareness in ESSs

Research on context-aware ESSs is still in its early stage. As shown in Table 1, optimization, deep learning, and cognitive approaches have been studied for context learning and exploitation in ESSs, network control, and computation offloading. The approach is selected depending on the target use, the input features for analysis, and the required flexibility to adjust to changes.

Machine Learning (ML) techniques like Deep Learning (DL) offer advantages to classic optimization approaches, including the support for unlabelled and massive mobile traffic. DL has been explored thanks to advances on optimization algorithms and parallel computing. Indeed, edge computing benefits from DL while providing the infrastructure for its deployment. Cognitive computing allows both learning and exploitation of context information, enabling an ESS to adjust its configuration towards a goal. Furthermore, cognitive computing is recognized as a key framework for Knowledge-Defined Networking on which networks self-organize to meet system and users' requirements [6].

6 Future directions

Spatio-temporal attributes from mobile traffic will create more robust strategies to address *where* and *when* to allocate content. The study of short- and long-term mobility of individuals and groups will help to adjust the network to slow and abrupt changes, inferring events like concerts and commuting. Content sharing between edge servers will further alleviate backhaul bottlenecks, while in the uplink flexible processing workflows will control the incremental processing of data. For uplink caching, replacement strategies based on popularity or data freshness are mandatory [16].

As storage is not an isolated resource, developments will focus on joint resource management [16], [12]. The architecture in Fig. 4 contributes on this by overseeing and reacting to changes on all edge resources. Native support for similar architectures will produce mobile networks with out-of-the box features to tackle current and future issues in ESSs and further services. Indeed, these architectures will produce advances on infrastructure planning, energy-aware networking (turning off idle base stations), malware control, and new business models from users profiling (e.g., content recommendation and spatio-temporal ads). Thus, user's privacy must be addressed for data multi-tenancy, and data governance must be ensured as data could reach different jurisdictions [8].

7 Conclusions

ESSs are key to enhance modern networks performance, and context-awareness, as a core design feature, can contribute to address ESSs' challenges. Furthermore, context-awareness contributes to enhanced infrastructure planning and even creates new business models for MNOs based on users profiling. We overviewed issues and opportunities offered by context information and presented a conceptual distributed architecture for its inference from mobile traffic data. This architecture helps to uncover spatio-temporal patterns in user's mobility and requests for uplink and downlink services. We envision future mobile networks with native context-awareness powered by ML and cognitive features for efficient and autonomous ESSs.

8 Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) and is co-funded under the European Regional Development Fund under Grant Number 13/RC/2077.

References

[1] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie. Mobile Edge Computing: A Survey. *IEEE Internet of Things Journal*, 5(1):450–465, feb 2018.

[2] M. Chen, W. Li, Y. Hao, Y. Qian, and I. Humar. Edge cognitive computing based smart healthcare system. *Future Generation Computer Systems*, 86:403–411, sep 2018.

[3] X. Cheng, L. Fang, X. Hong, and L. Yang. Exploiting Mobile Big Data: Sources, Features, and Applications. *IEEE Network*, 31(1):72–79, jan 2017.

[4] S. Feng, P. Setoodeh, and S. Haykin. Smart Home: Cognitive Interactive People-Centric Internet of Things. *IEEE Communications Magazine*, 55(2):34–39, feb 2017.

[5] C. Liang, Y. He, F. R. Yu, and N. Zhao. Enhancing Video Rate Adaptation With Mobile Edge Computing and Caching in Software-Defined Mobile Networks. *IEEE Transactions on Wireless Communications*, 17(10):7013–7026, oct 2018.

[6] A. Mestres, M. J. Hibbett, G. Estrada, K. Ma'ruf, F. Coras, V. Ermagan, H. Latapie, C. Cassar, J. Evans, F. Maino, J. Walrand, A. Rodriguez-Natal, A. Cabellos, J. Carner, P. Barlet-Ros, E. Alarcón, M. Solé, V. Muntés-Mulero, D. Meyer, and S. Barkai. Knowledge-Defined Networking. *ACM SIGCOMM Computer Communication Review*, 47(3):2–10, sep 2017.

[7] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos. Context Aware Computing for The Internet of Things: A Survey. *IEEE Communications Surveys & Tutorials*, 16(1):414–454, 2014.

[8] C. Puliafito, E. Mingozzi, F. Longo, A. Puliafito, and O. Rana. Fog Computing for the Internet of Things. *ACM Transactions on Internet Technology*, 19(2):1–41, apr 2019.

[9] M. Sheng, C. Xu, J. Liu, J. Song, X. Ma, and J. Li. Enhancement for content delivery with proximity communications in caching enabled wireless networks: architecture and challenges. *IEEE Communications Magazine*, 54(8):70–76, aug 2016.

[10] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu. Edge Computing: Vision and Challenges. *IEEE Internet of Things Journal*, 3(5):637–646, oct 2016.

[11] Z. Tan, F. R. Yu, X. Li, H. Ji, and V. C. M. Leung. Virtual Resource Allocation for Heterogeneous Services in Full Duplex-Enabled SCNs With Mobile Edge Computing and Caching. *IEEE Transactions on Vehicular Technology*, 67(2):1794–1808, feb 2018.

[12] C. Wang, Y. He, F. R. Yu, Q. Chen, and L. Tang. Integration of Networking, Caching, and Computing in Wireless Systems: A Survey, Some Research Issues, and Challenges. *IEEE Communications Surveys and Tutorials*, 20(1):7–38, 2018.

[13] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang. Joint computation offloading, resource allocation and content caching in cellular networks with mobile edge computing. In *2017 IEEE International Conference on Communications (ICC)*, volume 16, pages 1–6. IEEE, may 2017.

[14] M. Wang, Y. Cui, X. Wang, S. Xiao, and J. Jiang. Machine Learning for Networking: Workflow, Advances and Opportunities. *IEEE Network*, 32(2):92–99, mar 2018.

[15] R. Wang, X. Peng, J. Zhang, and K. B. Letaief. Mobility-aware caching for content-centric wireless networks: modeling and methodology. *IEEE Communications Magazine*, 54(8):77–83, aug 2016.

[16] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. YANG, and W. Wang. A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications. *IEEE Access*, 5:6757–6779, 2017.

[17] C. Zhang, P. Patras, and H. Haddadi. Deep Learning in Mobile and Wireless Networking: A Survey. *IEEE Communications Surveys & Tutorials*, 21(3):2224–2287, 2019.