Towards an intelligent 6G architecture: the case of jointly Optimised handover and Orchestration

Xenofon Vasilakos*, Sharday Olowu‡, Reza Nejabati*, Dimitra Simeonidou*
Smart Internet Lab, Department of Electrical and Electronic Engineering
University of Bristol, Bristol, United Kingdom
*name.surname@bristol.ac.uk, ‡xz19483@bristol.ac.u

***Abstract*—Modern communication networks are intrinsically softwarised and programmable, hence essentially agile. However, critical functionality comes from legacy generations, limiting agility and posing performance and other considerations for demanding 5G services and - in the future - 6G services. This paper discusses issues with legacy functionality and envisions high-level intrinsically intelligent design toward a 6G architecture. We focus on the case of machine learning model-based network functionality that can jointly optimise both user handover and service resource orchestration, rather than engaging in costly and uncertain prediction-based actions to accommodate the stringent requirements of 6G services. Moreover, we discuss this paradigm as a design pilot for other intrinsically intelligent 6G network functions.**

***Index Terms*—6G, Machine learning, SDN, Mobility**

I. INTRODUCTION

Following the rising 5G era, future Sixth Generation (6G) communication networks will be characterised by an unprecedented need for intrinsic function and service-hosting agility. The first steps towards this have already taken place in 5G by adapting the Software Defined Networking (SDN) and Network Function Virtualisation (NFV) paradigms for implementing network and service functions. Nevertheless, much of the supported functionality or design philosophy in 5G comes as a legacy from previous generations, posing considerations towards (even radically) redesigning Network Functions (NFs) in a future 6G architecture.

One notable example of such a legacy NF is User Equipment (UE) HandOver (HO) with its design roots stemming from the 4G era, and with implications on contemporary 5G MANagement and Orchestration (MANO) of network resources, running services and even control decisions. The HO and MANO functions are clearly decoupled; in essence, they consider each other as a *"mutual black box"*. To deal with this architecture design reality, the community has produced a significant body of research on solutions aiming to optimise 5G resource-related decisions from lower-network to application/service layers. Notable examples include Machine Learning (ML) model-based HO predictions [1][2][3][4] and/or purpose-built mobility patterns such for Virtual Network Function (VNF) [5], cache placement [6], favourable edge-node selection [7], multimedia content adaptation [8] or layer-3 wireless multicast [9].

Despite the significant exhibited benefits of such and other works[1], the above "mutual black box" between HO and resource handling poses a *gap* as state of the art solutions come with certain costs in part due to prediction and other uncertainties. Notable examples include long-term forecasts (e.g., at the level of minutes [1]) necessary for spawning new or synching large states between existing service replica nodes like containers or virtual machines [1] [10] and so forth.

Besides the time dimension uncertainties, other vital factors refer to space and other scalability aspects such as the number of cells or users or addressing important cost/performance dilemmas. The most evident dilemmas refer to adopting (deploying and maintaining) *individualised* ML models per user versus one model for a group of users, or a service versus a service type (such as a slice), or a cell versus a cellular region composed by many cells, and so forth.

Finally, 6G services and technology leaps [11][12][13] will only aggravate the above issues, and pose new ones as well as new ways and opportunities to address them. 6G services should be expected to be largely interactive, include immersive environments, high and internment user mobility (physical or virtual), and in general combine the most stringent requirements [10][11] across all known 5G use case categories (eMBB, URLLC and mMTC). Moreover, 6G extreme edge devices (IoT, smartphones, smartwatches, smartglasses, etc.) will be capable of much more. That includes not only hosting and running more resource-hungry models or leveraging multiple network access technologies at the same time (6G cellular/cell-less [11], device P2P, Wi-Fi and other parallel path-link connections); but also improving functionalities in terms of QoS and QoE, and in ways assessed not only by Key Performance Indicators (KPIs) (e.g., resource capacity, access-latency or reliability) but also Key Value Indicators (KVIs) like privacy/security and social fairness (e.g., by fighting network resource starvation against privileged users).

A. *Contributions: a design vision for 6G architecture*

The current paper focuses on the HO-MANO "mutual black box" gap and the challenges discussed in Section II.A, and tries to *propose appropriate design pillars towards a 6G architecture.* At the heart of this lies a novel 6G NF

[1] Section II.A provides a more detailed discussion on critical problems and challenges motivating our interest.

approach to *jointly and intelligently optimising resource MANO and* UE *HO*:

- **Analysing and tearing down the HO-MANO "mutual black box" gap:** move beyond standard legacy algorithmic HO processes such as the timer-based A3-RSRP algorithm or algorithms A2-A4-RSRQ, described in 3GPP reference [13] to jointly decided and optimise HO and MANO actions. The goal is to avoid prediction uncertainties and related costs with reactive and/or proactive approaches based on real-time monitoring and "open box" access between HO and resource-handling policies.
- **Highlight the potential of ORAN in an envisioned 6G architecture:** Actively involve the Multi-access Edge Computing (MEC) platforms and the large prospects of Open Radio Access Network (ORAN) RAN Intelligent Controller (RIC) on HO and resource handling decisions to improve resource usage efficiency and service KPI metrics.
- **Preliminary ML model-based results for HO predictions:** The results support our argument for facilitating intelligent NFs (in this case, the joint HO-MANO) within the 6G architecture.
- **Other aspects of the envisioned 6G architecture:** We discuss the joint HO-MANO as a pilot 6G function for design other NFs in the envisioned 6G architecture. We also discuss the role and engagement of the extreme edge (mobile devices) in future 6G NFs.

In what follows, Section II discusses the issues posed by legacy NFs in 5G and future 6G, as well as the related work. Section III outlines a high-level design approach towards a 6G architecture, emphasising the roles of ORAN RICs and MLOps. Section IV presents preliminary evaluation results for HO perdition supporting our argument that HO prediction is error-prone and should be abolished in a 6G architecture. Finally, Section V concludes this paper by wrapping up its main points and discussing how the joint HO-MANO can serve as a pilot for designing other intelligent 6G NFs or extending NFs and processes from ORAN to the extreme edge.

II. Background and State of the art

Legacy NFs have been designed for a different landscape than what is gradually forming in 5G and in the future towards converging to a new 6G era [11] [12]. Nevertheless, NFs' algorithmic processes such as for UE HO, have been largely successful and able to handle user mobility as of very recently. As a result, they are still in use nowadays, gradually maturing 5G deployments to decide when a user will be handed over from one cell, S (namely, the source), to the next, D (namely, the destination).

The most important aspects of these algorithms are time and signal quality [13][14][15]. The most notable examples include the A3-RSRP algorithm or algorithms A2-A4-RSRQ, described in 3GPP reference [13], which decide handovers after parametrised monitoring of UE (i) Received Signal Strength Indicators (RSSI), specifically Reference Signal Received Power (RSRP) or Reference Signal Received Quality (RSRQ), and a Time To Trigger (TTT) duration along with the former measurement. In the case of A3-RSRP, for instance, if the RSRP from an adjacent cell rises above the one of the currently UE serving cell by a value equal or greater than a HO hysteresis threshold, then a HO event is triggered if the RSRP difference remains this for a TTT period.

In this paper we use the HO-MANO relation as a paradigm problem representing others from legacy NF designs. Though largely successful in practice, still HO NFs ignore the MANO function or higher-level service aspects simply because these dimensions were *not* considered for designing former network generations. Back at the time, services where not demanding today's agility, where much more static, where less resource-hangry and stringent, and not designed or expecting to integrate into (or have privileged access to) the network itself such as in the case of service slices.

A. Challenges from past generation architectures

The following subsections use the HO-MANO paradigm to outline the most critical challenges of today's and future NFs that a 6G architecture must address.

1) The HO-MANO "mutual black box" gap: The "mutually black box" is a critical gap identified even in today's 5G between service MANO and HO operations. It can be immediately identified by the lack of HO internal-parameter knowledge or APIs exposed to service MANO and vice versa. Without such knowledge or any way to access any information via APIs or notification processes, 5G/6G services can *merely try to predict* imminent HO events with significant uncertainty and other costs (see Section II.A.4).

2) Legacy HO ignores 5G/6G landscape: The fundamentals of the handover procedure have remained the same from legacy networks (LTE) even in the 5G era: a UE reports monitoring measurements to its source cell S about neighbouring - hence, candidate destination D – cells, i.e. the physical layer Cell Identifiers – (PCIs) and RSSI measurements. Then, S decides to start the HO to "best" cell D, and UE and D complete the HO. More details can be found in the relevant 3GPP specification (Sec 4.9) [14] and [15] with variations including direct or not preparatory commutation between S and D.

a) Ignoring technology leaps: Legacy HO ignores technology leaps and the opportunities of (i) more distributed HO decisions in the context ORAN [25] (by central or distributed units); and (ii) by extreme edge devices themselves that could have a role part (even individually decide) about their connectivity and HO status (e.g., leveraging multiple interfaces and/or P2P inter-connectivity).

b) Lack of Intelligence: Lack of intelligence in A3-RSRP or A2-A4-RSRQ is evidently a problem for 6G and even 5G. Stativity of rules, parameters and even parameter update processes cannot capture a programmable network's dynamicity and volatility. Examples include, and are not limited to, user and service mobility (physical or virtual), network composition with continuously added newly-spawned nodes, altered nodes and link network or VNF composition regarding resource allocation.

3) Increased complexity of everything: Programmable networks pose a far more complex and dynamic landscape where there is *no* "one static algorithm matches all" solution. NFs such as HO must be intelligent (i.e., ML-based) to capture and address a *non*-static landscape of (i) cells and other network resource/components; (ii) UE and service expectations (QoS/QoE) or dynamic behaviour (arbitrary mobility or failure/churn) and (iii) their reflection on training data[2]; (iv) resource congestion and starvation risks (v) especially against non-privileged users and services; (vi) more adaptive to dynamics SLAs.

4) Significant costs of Intelligent HO predictions: Given the existing HO processes, there is a need to predict UE HO for MANO and even Radio Access Network (RAN) control operations. But that comes at a series of costs, spanning from (i) prediction uncertainty and its impact on uncertain resource allocation decisions; to (ii) significant use of resources, including human-involved effort and compute resources for full ML life cycles (from gathering training data to retraining/deploying/maintaining a model). And all that, of course, at the cost of (iii) energy consumption which is high for ML model training and runtime [16].

B. Impact of challenges

Without intelligent control or any means to access or influence HO, the state of the art (see Section II.C) tries to predict handover in order to take *proactive* MANO actions that improve/guarantee service quality, such as in the case of [1]. Such proactive actions include aggressive resource allocation, such as in [10] setting up Kubernetes nodes at the handover destination network edge and proactively syncing the service state there to avoid delay-critical service downtime after handover. Other examples include service migration [7].

Nonetheless, HO prediction potential comes with challenges: handover accuracy, time-to-handover accuracy, exposure/availability of necessary monitoring data [17], scalability of ML models regarding the number of users and cells, etc. Moreover, there is consideration regarding the time it takes for MANO and/or container orchestration actions to happen and guarantee optimised seamless service delivery.

For delay-critical services, actions must occur as soon as possible after predictions for imminent handover events [10]. And even if the model decision and/or action time after a prediction is considered solved, there may be implications regarding input data staleness such as for Radio Resource Control (RRC), e.g., for assigning Resource Blocks (RBs) to users for which the combined time for model access to current monitoring and decision is at sub-millisecond scales [18].

On the very contrary, there are other service types for which predictions must refer to farther in the future (e.g., in the scale of minutes [10] instead of 10s of milliseconds). For the latter, as well as for necessary MANO/container actions that take long, such as spawning a new container (1-2 minutes), handover predictions tend to be less accurate.

[2] This is known in the literature as the “dataset shift” problem, analysed extensively here: https://mitpress.mit.edu/books/dataset-shift-machine-learning.

C. State of the Art

1) HO prediction: State of the art in UE mobility and HO prediction aims predominantly at two goals, often combined together: (i) seamless mobility, i.e. no/zero service interruption upon HOs, and (ii) node selection for application/service/function placement. The most popular techniques include UE profiling [19][20], exploiting handover history patterns and user trajectory prediction [6][9][20], radio link characteristics and cross-layer optimization [7][21][22], and finally, ML model-based [1][2][3][4] solutions. Out of all the above, ML model-based approaches and particularly those *integrated into MEC* platform solutions, fall within the scope of this paper and are analysed further below.

The authors of [1] use intelligent handover prediction models between radio 5G Base Stations. Specifically, they apply a Transfer Learning (TL) technique by conducting aligned simulation and actual testbed training using a combination of (i) Long Short-Term Memory (LSTM) or gradient boost regression with classification models (N/XGBoost) for filtering out any received signal power and compute resource prediction input outliers to (ii) predict the destination serving Multi-access Edge Computing (MEC) point and cell (hence, cellular handovers) that takes over UE's service after handover. Posing similar research traits, the work of [3] uses prediction assisted handover based on multilayer perception neural networks to reduce the handover time delays.

Next, the work of [10] covers that of [1] from a broader MEC perspective in high-mobility scenarios. The work leverages predictions from models like in [1] as the means to identify favourable edge nodes for hosting service replicas for handover UEs. As a result, the experienced service downtime after handovers should be non-perceivable.

Another work that focuses on the MEC side is [2], specifically on *decentralised 5G deployments* with services in distributed resources in the MEC architecture to locate services topologically close to UEs. The authors explore Recurrent Neural Networks (RNN) using LSTM for UE mobility prediction for automotive scenarios. Handover prediction integrated with service migration in 5G systems is studied in [7] by using two prediction mechanisms to forecast mobile UE's handover events by exploiting user-network association patterns. Just like [1], [10] and [2], this work refers to MEC scenarios for identifying favourable edge nodes. Last, likewise to [7], the work of [23] aims at efficient handovers by using mobility pattern history and user trajectory prediction.

2) MEC and RAN adaptation: Regarding MEC and radio access platforms with ML model-hosting capabilities, Low-Latency MEC [24] is a 3GPP fully compliant open-source platform offering multiple APIs that align to ETSI MEC specifications. The platform is SDN programmable using the OpenFlow protocol while fully integrated with the FlexRAN [18] SD-RAN controller. As a result, LL-MEC addresses three types of latency: (i) user transport latency, (ii) underlying network (control) latency between MEC-hosted apps and MEC-performed actions; and (iii) application latency. All these types of latencies are relevant for hosting ML-models at the edge, which according to [10]

need (ultra) low-latency access to monitoring data as well as to transmitting decided model actions.

ORAN [25][26] is widely considered as the most viable solution for next generation RAN. It offers disaggregated RAN functionality composed of Open Centralised Units (O-CUs), Open Distributed Units (O-DU), and Open Radio Units (O-RUs) units, using open interface specifications between elements implemented over vendor-neutral hardware using open, programmable interfaces and standards. The most important element of ORAN is its RAN Intelligent Controller (RIC), i.e. an SDN architecture component responsible for controlling and optimising RAN functions. Section III discusses ORAN in more detail.

III. Architectural Vision

The diagram and discussion below present a high-level adaptation of our architectural vision, which largely involves ORAN aspects and capabilities. The vision tries to capture our background discussion, and particularly the multi-facet role(s) of integrated ML models into a 6G architecture. Moreover, it has certain design features that try can intrinsically address the critical challenges identified in Section II.A., particularly the *"mutually black box" gap* that we discuss as a representative problem and pilot solution within the architecture.

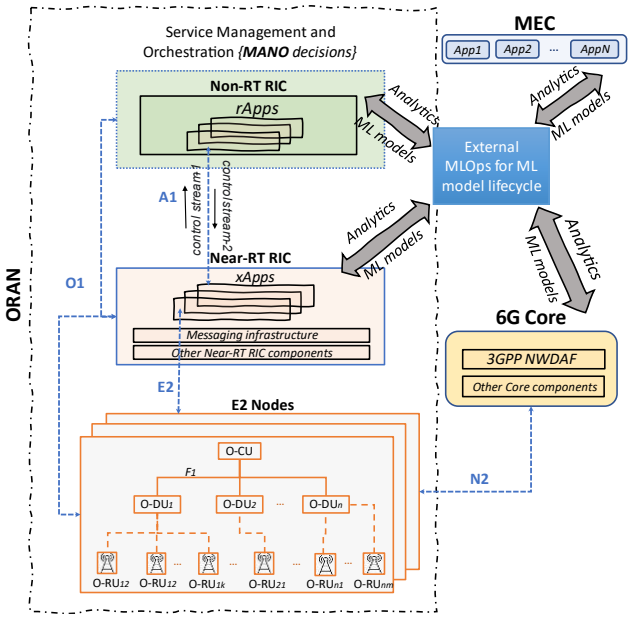

Fig. 1. Leveraging ORAN RIC for ML model-based joint or cooperative MANO and control decisions.

In more detail, Fig. 1 shows the relationship between ORAN and its internal components, emphasising Near-RT (near real-time) and non-RT Radio Interface Controls (RICs), MEC, the Core, and an external ML model lifecycle framework. Services can be running at the MEC (especially delay-critical ones), or the Core or ORAN as Network Apps. Notice the interfaces. A1 allows communication between management and orchestration rApps and control xApps, while E2 connects xApps to E2 node elements for controlling purposes such as HO. An E2 node has a one-to-one relationship with a near real-time RIC, but one RIC can connect to multiple E2 nodes. The protocols that go over E2 allow controlling and optimising the E2 node elements and resource usage. O1 is the interface between management entities in the MANO/O-RAN parts. Last, N2 connects the E2 nodes to the Core.

A joint MANO-HO solution can be native to this architecture via ML models running as *xApps* over the Near-RT RIC by being responsible for monitoring HO critical data and for taking optimised HO decisions. Near-RT RIC offers a platform hosting microservice-based applications called xApps. Such ML models implemented as xApps can have near *real-time access* to data for taking near real-time control decisions such as HO. Besides HO control decisions, xApp ML models can take other types of control decisions like wireless resource block scheduling.

On the one hand, xApp ML models can work closely with rApps (discussed below) in more than ways, such as for detecting monitoring anomalies (e.g., with LSTM-like models); or by updating rApps regarding imminent or longer-term control decisions likewise to HO (namely, "control stream-1" Fig. 1).

On the other hand, MANO optimisation ML models can be fed with xApp input over interface A1 to take optimised orchestration decisions. One or more of such models, each baring single or multi-objective resource optimisation goals, can be running as Non-RT rApps as part of a centralised ORAN Service Management and Orchestration (SMO) Framework. Examples of such models include Reinforcement Learning (RL) model-based solutions like [27][28][29] or other supervised ML models like those analysed in [30] or for performance profiling in [31]. As defined by the ORAN specification, this is non-real-time, i.e. takes more than a second. Therefore, any MANO decisions will be taken based on constantly fed HO xApp input, including updates on imminent or long-term HO. This implies either a particular element of HO prediction or HO/MANO coordination to jointly decide and optimise both HO and resource handling.

A major difference compared to traditional HO predictions lies in continuous real-time access to HO data, which is currently not available but instead guessed or extracted implicitly, leading to prediction uncertainty. Also, xApps can explicitly schedule HOs based on feedback input from SMO rAPPs, hence following a reverse approach ("control stream-2" in Fig. 1) compared to "control stream-1" above.

Finally, note that a major advantage of xApps and rApps lies in being third-party software, enabling 6G services to control or influence MANO and HO, assuming this is allowed by a corresponding SLA.

A. The distinct MLOps component

The Machine Learning Operations (MLOps) component is an external component of the poposed architecture, using external interfaces with MEC, both ORAN RICs and the Core. MLOps is a core ML function that takes continuously fed data analytics and raw monitoring data for executing the complete ML models life cycle. The latter includes developing, (re-)training/maintaining and deploying models, and then continuously monitoring/reviewing their performance in order to replace them with others or retrain/improve them. In our previous works, a complete approach to MLOps is provided in [17], while preliminary

data pipelining with offline or online model (re-)training is provided in works [28] [30][31] and [1][27], respectively.

B. ML models and data for the intelligent NFs

We aim to explore and combine the following data analytics (stats and raw input) with training, deploying and maintaining appropriate ML models. Again, we use the joint HO-MANO intelligent NF as a paradigm:

- *State of Network resources:* nodes for hosting the service and local resources like GPU, CPU, memory, NIC buffers, number of UEs per node/cell, etc.
- *UE monitoring*: utilise everything from the physical wireless to higher network layer(s): Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), link/path statistics at all layers, including capacity, latency, availability, jitter, etc.
- TTT and hysteresis: these parameters stem from the events considered by A1-A4 3GPPP HO specs. Get such input from live monitoring and associate it with a need for proactive control [9] or MANO actions like path steering for switching to alternative service points [10].
- Leverage *service/app data* regarding service-*level* performance extending to user QoE metrics and other sources of information (possibly with user privacy approval like GPS location).
- *Energy-awareness*: monitor and consider the consumption of UE battery, O-RU energy, and all network components utilised, such as MEC resources.

IV. Preliminary simulations on ML HO predictions against A3

Next, we present our preliminary simulation results for assessing the uncertainty of HO predictions in support of our position for coordinating HO and MANO optimisation, rather than predicting the behaviour of the HO NF for enhancing MANO.

A. Setup

Our simulations were done with the well-known C++ ns-3 simulation environment and the ns3-gym framework[3] that enables ML model integration to 5G simulations.

We trained a Long Short-Term Memory (LSTM) model and an XGBoost (decision tree) model. These models were used to live forecast a UEs' RSRP values and finally to predict the serving cell physical identity (S-PCI) using the XGBoost model with the former forecasted values as its input. The latter forecasted RSRP, as well as S-PCI predictions, were fed back to the simulation environment as actions.

All simulations were carried out to assess the performance of the system under different circumstances, including a scalability study and corresponding assessment using different arrangements of either 2 or 4 microcells to provide high-speed connectivity, even in built-up urban areas.

To emulate realistic mobility, we utilised real taxi mobility traces from a publicly available San Francisco taxis mobility traceset [33].

B. Results and conclusions

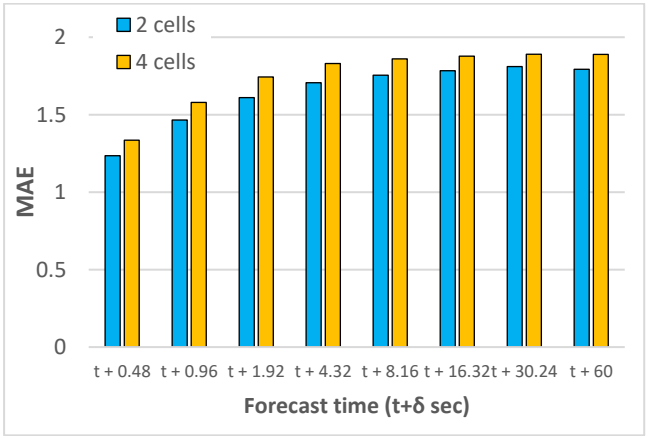

Fig. 2. Mean Absolute Error results (y axis) of LSTM models trained for different forecast lengths in 2-cell and 4-cell network scenarios. Time periods on the x-axis refer to now (t) plus some delta (δ) for the prediction period (~0.5 sec., ~1 sec., … ~0.5 mins, 1 min).

Fig. 2 shows Mean Absolute Error (MAE) results of LSTM models trained for different forecast t+δ periods in a 2-cell and 4-cell network scenarios. The metric expresses the magnitude of the difference between real and predicted RSRP values in dB.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

As seen in the graph, MAE is lowest when predicting imminent RSRP in just one δ (~0.5 sec.) step ahead. However, it increases as the forecast δ increases, for both 2-cell and 4-cell scenarios, without the scale of cell number affecting this behaviour.

- **Conclusion 1: *There is uncertainty* in RSRP predictions.** The uncertainty gets aggravated with longer future period predictions despite the exhibited robustness of LSTM. For instance, MAE is 1.79 dB for a 60-sec forecast, which is 2.3% higher than that of a t+ 8.16s forecast (1.75 dB) in the 2-cell scenario.

Moreover, an increased number of cells (i.e., four) seems to have a negative impact on MAE, yet with a small difference compared to the case of two cells. Therefore, we cannot safely conclude on the exact implications or their extent of cell scalability on RSRP predictions.

[3] https://apps.nsnam.org/app/ns3-gym/.

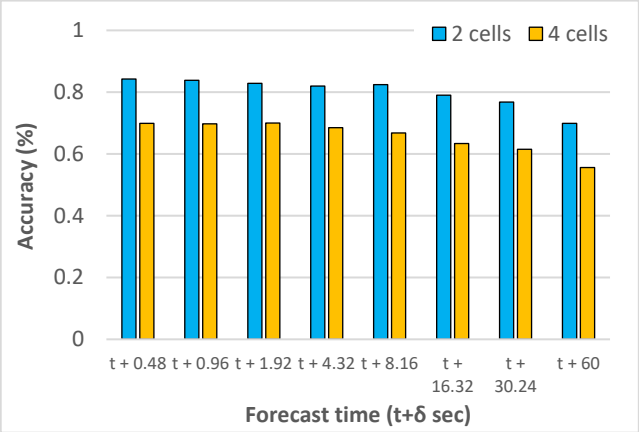

Fig. 3. Accuracy results of XGBoost S-PCI classifiers trained for different forecast lengths in 2-cell and 4-cell network scenarios.

Fig. 3 shows the accuracy of service cell predictions as a percentage of correct predictions overall predictions. Remember that the predictions are based on corresponding LSTM RSRP predictions passed as input to the XGBoost model.

The XGBoost model shows fair or good accuracy scores for S-PCI prediction, particularly for the 2-cell scenario. The absence of confidence intervals denotes the use of realistic traces, rather than conducting randomised mobility simulation repeats. In both scenarios, the accuracy reduces as the forecast length increases. This is because LSTM-based input is more error-prone due to greater uncertainty, which then propagates to the classification which is based on these values.

- **Conclusion 2: *HO prediction suffers from a significant uncertainty, particularly for longer-term predictions*.** Despite the LSTM robustness, the final HO classification-prediction is error-prone due to the "amplified" propagation of otherwise seemingly small errors in RSRP predictions by the LSTM first layer.

It is worth noting that these results are in accordance with the results of [1], which were conducted over a custom simulation plus a real testbed environment. The current results stem from a standardised simulation environment, including a further scalability study and realistic vehicle mobility traces over a real wide-area realistic use case, as opposed to human walking mobility over an urban square used in [1].

- **Overall conclusion: HO prediction is error-prone. A 6G should facilitate a different HO NF to avoid erroneous predictions.** This is important for both imminent or longer-term forecasts, i.e. (according to [1][9] and other works from the literature) for cases where service state must be rapidly synced prior to HO or for cases of transferring an entire service/creating a consistent replica, respectively.

V. HANDOVER AND BEYOND FUTURE WORK PLAN

The joint HO-MANO 6G function and its place in the high-level architecture of Fig. 1 are discussed in this paper as a feasibility proof of concept for integrating intelligence in a future 6G architecture. Intrinsic ML model-based intelligence in a future 6G design stems from the currently developing state of the art technologies like ORAN.

Moving beyond the discussed HO-MANO 6G paradigm case and the corresponding pilot NF proposition for addressing the "mutual black box" gap (see Section II), integrating ML model-based intelligence for other NFs is feasible and can address *the increased complexity* in 6G in more than one ways:

- First, by jointly optimising multiple objectives under dynamic and possibly unforeseen conditions.
- Secondly, by minimising – if not entirely removing – static rules and any remaining elements of traditional human administration. Both of the latter make approaching optimisation goals and conducting timely actions infeasible under dynamic conditions, especially for delay-critical use cases.
- On the very contrary, a design along the lines of the high-level architecture of Fig. 1 complies with the 3GPP-defined Zero-touch network and Service Management (ZSM) vision [32].
- As discussed and verified by our preliminary simulation-based evaluation results, predictions come at the cost of uncertainty and other implications, particularly when forecasting network behaviour for longer times in the future. Unlike that, the architecture of Fig. 1 *eliminates* the need for HO *predictions* to enhance MANO; and by induction, the need for predictions regarding other NF actions (e.g., new UE or service registration) that currently also suffer from a "mutual black-box" gap with MANO.
- The previous point highlights the need to allow NF actions to be co-decided or coordinated with MANO such as via control streams 1 and 2 in Fig. 1. Given the nature or an ORAN RIC environment, near-RT control xApps and non-RT rApps such as for MANO operations can adapt a pub/sub model of communication such as proposed in Information Centric Architectures [34][35]. The latter use dynamic naming and name resolutions schemes that allow scalable, secure and scoped-based communication among x/rApps via unicast, multicast/broadcast and concast via pub/sub messages. This can benefit the purposes of an extendable 6G service-based architecture over an infrastructure that leverages ORAN and NWDAF (e.g., for monitoring [36] and ML models execution).

Regarding our future work, this includes exploring the join HO-MANO solution along the lines of the posed architecture, including the MLOps component discussion.

In addition, we plan to focus on possible extensions that can strengthen the above design and specifically engage the so-called extreme edge. We may replace the sole global view of intelligent NFs deployed in ORAN RIC with a more UE-peer approach. Engaging the mobile devices (aka the "extreme edge") for NF decisions such as autonomously decided HO can come with pros, cons and trade-offs. Depending on the NF and the exact approach taken (model type, protocol architecture such as for control messaging,

etc.), energy consumption needs may be low or high for running autonomous ML models over the (usually) power-restricted UE devices.

On the other hand, engaging the extreme edge comes with a lot of advantages and options. It can be done (fully/semi-) autonomously with peer UE relations or under an ORAN hierarchy. This may allow *individualised* (e.g., user personalised) model training that optimises user-specific KPIs for improving their QoE. Moreover, individualised models have intrinsic privacy and security advantages expressed by KPIs and/or KVIs, thus can improve social welfare by avoiding resource starvation from a centralised authority. One notable example includes the ability of modern devices to exploit multiple wireless interfaces and technologies for P2P connections and corresponding network access. Under such scenarios, the centralised ORAN is alleviated by HO and other NFs, and of course, from consuming RAN and other resources. Alternatively, intelligence may be "shared" between both the extreme and ORAN RIC apps, e.g., via federated learning schemes, which can combine the benefits of both worlds.

ACKNOWLEDGEMENT

This work is supported by the H2020 European Project 5GASP (grant agreement No. 101016448).

REFERENCES

[1] N. Uniyal et al., "Intelligent Mobile Handover Prediction for Zero Downtime Edge Application Mobility," 2021 IEEE Global Communications Conference (GLOBECOM), pp. 1-6, 2021.

[2] U. Fattore et al., "AutoMEC: LSTM-based user mobility prediction for service management in distributed MEC resources," Proceedings of the 23rd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, 2020.

[3] J. F. Abuhasnah and F. K. Muradov, "Direction prediction assisted handover using the multilayer perception neural network to reduce the handover time delays in lte networks", Procedia computer science, vol. 120, pp. 719-727, 2017.

[4] M. Feltrin and S. Tomasin, "A machine-learning-based handover prediction for anticipatory techniques in Wi-Fi networks", 2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN), pp. 341-345, 2018

[5] T. Taleb, M. Bagaa and A. Ksentini, "User mobility-aware Virtual Network Function placement for Virtual 5G Network Infrastructure," 2015 IEEE International Conference on Communications (ICC), 2015, pp. 3879-3884, doi: 10.1109/ICC.2015.7248929.

[6] X. Vasilakos et al., "Proactive selective neighbor caching for enhancing mobility support in information-centric networks," Proceedings of the second edition of the ICN workshop on Information-centric networking, 2012.

[7] H. Abdah, J. P. Barraca and R. L. Aguiar, "Handover Prediction Integrated with Service Migration in 5G Systems," ICC 2020 - 2020 IEEE International Conference on Communications (ICC), 2020, pp. 1-7, doi: 10.1109/ICC40277.2020.9149426.

[8] A. Siris et al., "Exploiting mobility prediction for mobility & popularity caching and DASH adaptation," 2016 IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2016, pp. 1-8, doi: 10.1109/WoWMoM.2016.7523523.

[9] X. Vasilakos et al., "Mobility-based proactive multicast for seamless mobility support in cellular network environments," Proceedings of the Workshop on Mobile Edge Communications, 2017.

[10] X. Vasilakos et al., "Towards Zero Downtime Edge Application Mobility for Ultra-Low Latency 5G Streaming." 2020 IEEE Cloud Summit. IEEE, 2020.

[11] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan and M. Zorzi, "Toward 6G Networks: Use Cases and Technologies," in IEEE Communications Magazine, vol. 58, no. 3, pp. 55-61, March 2020, doi: 10.1109/MCOM.001.1900411

[12] CISCO, "Cisco Annual Internet Report (2018–2023) White Paper," [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[13] ETSI TS 138 331 V15.3.0 (2018-10), "5G; NR; Radio Resource Control (RRC); Protocol specification," 3GPP TS 38.331 version 15.3.0 Release 15, [Online]. Available: https://www.etsi.org/deliver/etsi_ts/138300_138399/138331/15.03.00_60/ts_138331v150300p.pdf

[14] ETSI TS 123 502 V15.2.0 (2018-06), "5G; Procedures for the 5G System (3GPP TS 23.502 version 15.2.0 Release 15)," 3GPP TS 23.502 version 15.2.0 Release 15, [Online]. Available: https://www.etsi.org/deliver/etsi_ts/123500_123599/123502/15.02.00_60/ts_123502v150200p.pdf

[15] Techplayon, "5G SA Inter gNB Handover – Xn Handover", [Online]. Available: https://www.techplayon.com/5g-sa-inter-gnb-hanodver-xn-handover/

[16] E. García-Martín et al., ""Estimation of energy consumption in machine learning, Journal of Parallel and Distributed Computing, Volume 134, pp. 75-88, 2019, https://doi.org/10.1016/j.jpdc.2019.07.007"

[17] X Vasilakos et al. "Integrated Methodology to Cognitive Network Slice Management in Virtualized 5G Networks.", arXiv preprint arXiv:2005.04830, 2020

[18] X. Foukas et al. "FlexRAN: A flexible and programmable platform for software-defined radio access networks." Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies. 2016.

[19] D. Barth, S. Bellahsene and L. Kloul, "Mobility prediction using mobile user profiles", Modeling Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS) 2011 IEEE 19th International Symposium on, pp. 286-294, 2011.

[20] Z. Becvar, "Efficiency of handover prediction based on handover history", Journal of Convergence Information Technology, vol. 4, no. 4, pp. 41-47, 2009.

[21] X. Chen, F. Mériaux and S. Valentin, "Predicting a user's next cell with supervised learning based on channel states", Signal Processing Advances in Wireless Communications (SPAWC) 2013 IEEE 14th Workshop on, pp. 36-40, 2013.

[22] Y.-H. Choi et al., "Cross-layer handover optimisation using linear regression model", Information Networking2008. ICOIN 2008. International Conference on, pp. 1-4, 2008.

[23] R. Ahmad et al. Efficient handover in lte-a by using mobility pattern history and user trajectory prediction. Arabian Journal for Science and Engineering, 43(6):2995–3009, 2018.

[24] N. Nikaein, et. al., "LL-MEC: Enabling Low Latency Edge Applications," 2018 IEEE 7th International Conference on Cloud Networking (CloudNet), 2018, pp. 1-7, doi: 10.1109/CloudNet.2018.8549500.

[25] I. Chin-Lin and K. Sachin, "O-RAN Minimum Viable Plan and Acceleration towards Commercialization", O-RAN alliance, White Paper, 29 June 2021.

[26] S. K. Singh, R. Singh and B. Kumbhani, "The Evolution of Radio Access Network Towards Open-RAN: Challenges and Opportunities," 2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), 2020, pp. 1-6, doi: 10.1109/WCNCW48565.2020.9124820.

[27] M. Bunyakitanon et al. "End-to-end performance-based autonomous VNF placement with adopted reinforcement learning." IEEE Transactions on Cognitive Communications and Networking 6.2 (2020): 534-547.

[28] Y. Bi et al., "Multi-Objective Deep Reinforcement Learning Assisted Service Function Chains Placement." IEEE Transactions on Network and Service Management 18.4 (2021): 4134-4150.

[29] X. Vasilakos et al., "Towards Low-latent & Load-balanced VNF Placement with Hierarchical Reinforcement Learning." 2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom). IEEE, 2021.

[30] M. Bunyakitanon et al., "Auto-3P: An autonomous VNF performance prediction & placement framework based on machine learning." Computer Networks 181 (2020): 107433.

[31] S. Moazzeni et al. "A Novel Autonomous Profiling Method for the Next-Generation NFV Orchestrators." IEEE Transactions on Network and Service Management 18.1 (2020): 642-655.

[32] ETSI GS ZSM 007 V1.1.1 (2019-08), "Zero-touch network and Service Management (ZSM); Terminology for concepts in ZSM", European Telecommunications Standards Institute, Sophia Antipolis Cedex, France, White Paper, 2019. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/ZSM/001_099/002/01.01.01_60/gs_ZSM002v010101p.pdf.

[33] M. Piorkowski et al., CRAWDAD dataset epfl/mobility (v. 2009-02-24), traceset: cab, downloaded from https://crawdad.org/epfl/mobility/20090224/cab, https://doi.org/10.15783/C7J010, Feb 2009.

[34] K. Katsaros et al. "On inter-domain name resolution for information-centric networks." International Conference on Research in Networking. Springer, Berlin, Heidelberg, 2012.

[35] K. Katsaros et al., "On the inter-domain scalability of route-by-name information-centric network architectures." 2015 IFIP Networking Conference (IFIP Networking). IEEE, 2015.

[36] X. Vasilakos et al., "Elasticsdk: A monitoring software development kit for enabling data-driven management and control in 5g." NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium. IEEE, 2020.

AUTHORS

Xenofon Vasilakos (Member, IEEE) is a Lecturer in AI for Digital Infrastructures with the University of Bristol, the U.K, and a member of Bristol Digital Futures Institute and Smart Internet Lab. His current research focuses on SDN/NFV programmable 6G Cloud and Multi-access Edge Computing network architectures leveraging machine learning model-based solutions. He has also conducted research on network slicing, distributed and Information Centric Networking architectures. He has received Greek and French Government grants, awards and an accolade, and has worked for the FIA award-winning project FP7-PURSUIT. CV: http://pages.cs.aueb.gr/~xvas/pdfs/detailedCV.pdf

Sharday Olowu is a final year undergraduate student of Computer Science and Electronics at the University of Bristol. During her time at Bristol, she has explored a wide range of areas within electronic engineering and computer science, specialising in topics involving artificial intelligence. She has held positions of Course Representative and Teaching Assistant at the university, and achieved the Bristol PLUS Award. She holds an offer to study the MPhil in Advanced Computer Science at the University of Cambridge with a DeepMind AI Scholarship.

Reza Nejabati (Senior Member, IEEE)} is the Chair Professor of Intelligent Networks and the Head of the High-Performance Network Group with the Department of Electrical and Electronic Engineering, University of Bristol, U.K. He is also a Visiting Professor and a Cisco Chair with the Cisco Centre for Intent-Based Networking, Curtin University, Australia. He has established successful and internationally recognised research activities in “Autonomous and Intent Based Networks,” as well as “Quantum Networks.” Building on his research, he co-founded a successful start-up company (Zeetta Networks Ltd.). It has currently 25 employees and 6m VC and external funding. His research received the prestigious IEEE Charles Kao Award in 2016 and has done important contributions in 5G, smart city, quantum communication, and future Internet experimentation.

Dimitra Simeonidou (Fellow, IEEE)} is a Full Professor with the University of Bristol, the Co-Director of the Bristol Digital Futures Institute and the Director of Smart Internet Lab. Her research is focusing in high performance networks, programmable networks, wireless-optical convergence, 5G/6G and smart city infrastructures. She is increasingly working with social sciences on topics of digital transformation for society and businesses. She has been the Technical Architect and the CTO of the Smart City Project Bristol Is Open. She is currently leading the Bristol City/Region 5G urban pilots. She has authored and co-authored over 600 publications, numerous patents, and several major contributions to standards. She has been the co-founder of two spin-out companies, the latest being the University of Bristol VC funded spin-out Zeetta Networks, delivering SDN solutions for enterprise and emergency networks. Prof. Simeonidou is a Fellow of Royal Academy of Engineering and a Royal Society Wolfson Scholar.