This is a postprint version of the following published document: M. Becattini et al., "Dynamic MEC resource management for URLLC in Industry X.0 scenarios: a quantitative approach based on digital twin networks," 2024 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0 & IoT), Firenze, Italy, 2024, pp. 372-376, doi: 10.1109/MetroInd4.0IoT61288.2024.10584165.

Dynamic MEC resource management to enable ultra-reliable and low latency communications (URLLC) in industry X.0 scenarios: a quantitative, digital twin network based approach

Marco Becattini, Laura Carnevali, Giovanni Fontani, Leonardo Paroli, Leonardo Scommegna
Department of Information Engineering
University of Florence
Florence, Italy
{marco.becattini, laura.carnevali, giovanni.fontani1, leonardo.paroli, leonardo.scommegna}@unifi.it

Maryam Masoumi, Ignacio de Miguel
Department of Signal Theory, Communication and Telematics Engineering
Universidad de Valladolid
Valladolid, Spain
{maryam.masoumi@, ignacio.miguel@tel.}uva.es

Fabrizio Brasca
WindTre
Milan, Italy
fabriziogabrio.brasca@windtre.it

***Abstract*—The use of innovative technologies within industry X.0, including but not limited to AR/VR (Augmented Reality/Virtual Reality), Autonomous Robotics and advanced security systems, requires applicative interconnection between a large number of IoT machines and devices. These connections are required to be on ultra-reliable and low latency communications (URLLC), to enable optimal performance of the new technologies. Moreover, the concepts of reliability and low latency are deeply interconnected. From the perspective of a device, a service that exceeds a certain time threshold in its response is considered unresponsive and therefore unreliable.**

In this paper, we present an innovative approach to quantitatively evaluate reliability in URLLC settings, leveraging the use of digital twin networks (DTN) with a specific focus on mobile edge computing (MEC) and its application to industry X.0 scenarios.

DTNs are created with data collected from MECs within an industry X.0 setting, where MECs act as service-providers for the whole ecosystem.

DTN thus is a representation of the overall network and of the service it provides; services are associated within the DTN with the probability, evaluated a posteriori, for those service to be requested and the related capability of a given MEC to satisfy such request within the time boundaries of the requests themselves, under all the possible configurations, in terms of service readily available on the MEC. This allow evaluating the probability for a given set of services to be managed in time by a given MEC configuration.

To achieve this, we propose a solution to transform DTN first into a simulation-oriented representation, specifically a DAG, and then, in a mathematical modeling language, in our case, Structured Tree, that allows near real time quantitative prediction, performed by a function within the MEC, about the MEC capability to manage the expected service load.

The prediction function first evaluates if the current MEC configuration can sustain the probable incoming requests. In case of a negative response, the set of most probable requests are passed to an optimizer function, which evaluates if there is any potential configuration that can sustain the expected incoming requests. If so, such a configuration is pre-computed in the MEC level, thus enabling a near real time switch of the services configuration as soon as the expected pattern of requests starting to come in.

Results obtained so far show potential for this approach to confer the MEC better request handling capabilities, by providing a near real time re-configuration ability, deployed within the MEC itself.

***Index Terms*—Industry X.0, Ultra-reliable and low latency communications (URLLC), Digital Twin Networks (DTN), Mobile Edge Computing (MEC), Mathematical modeling languages.**

I. Introduction

Industry X.0 is an umbrella term to define the ongoing progressive automation and digitalization of industries, which has started with industry 4.0 [1] and is progressively evolving into industry 5.0.

Industry 5.0 differs and represent an evolution with respect to industry 4.0, as it is more focused , sustainability, and resiliency [2]. These characteristic empathize the need for a telecommunication solutions within Industry X.0 that provide unprecedented level of reliability combined with ultra low latency. This class of solutions are usually named under the umbrella term of ultra-reliable and low latency communications (URLLC).

A key element to provide such solutions is represented by Mobile Edge Computing (MEC): a single unit that combines high speed communication capabilities, usually 5G powered, with a server that is able to provide services. These services are containerized and deployed in form of micro-service [3].

MEC devices are therefore a key foundational element for Industry X.0, as they can be seen as the provider of all those services which requirements falls under the URLLC umbrella. The initial URLLC requirements were outlined by 3GPP in [4].

Specifically, this document stipulates that the desired average value for URLLC user-plane (U-plane) latency should be 0.5 ms in both the uplink and downlink. It is relevant to note that there is no reliability requirement attached, a limitation that will be addressed in this paper.

It's important to note that U-plane latency refers to the time taken to successfully deliver a packet from the starting point of the layer 2 protocol on the transmitting end to the end point of the layer 2 protocol on the receiving end [5]. Such short time requires that MEC units (or MECs) are able to respond to incoming request in the order of milliseconds. This consequently mandates MECs to be in a very specific operational condition:

1) To be able to process an incoming request immediately (i.e. a zero queue condition)
2) To have the services required from incoming request already initialized (i.e. deployed and active, in form of micro-service within an active container instance within the MEC)

Any queuing time, or any request to initialize a specific service, should only occur after a request necessitating it has been received by the MEC would result in service-time for the request that surpasses URLLC service level agreements (SLA) and in general those SLA that are associated with Industry X.0, as assessed in dedicated studies [6].

As a consequence, even offloading (i.e routing the request to a nearby MEC) is not viable and cloud scaling is also clearly precluded.

A simple solution would be a linear scaling of the available MECs, however this solution is not practicable from several points of view: it is economically unsustainable owing to the expense of MEC units, and it is ethically and financially questionable due to the over-scaled set-up's economic and CO2 production costs.

Hence it arises the need for a novel approach that allows MEC to perform at their maximum capacity in a variety of scenario: such an approach requires to be able to predict the incoming request and pre-configure the MEC level accordingly.

II. Contribution

In this paper we propose an approach that, according to preliminary experiments, it is able to both predict the incoming request and reconfigure the MEC level dynamically, in near real time, thus allowing the strict SLA of industry x.0 to be satisfied, leveraging the use of digital twin and quantitative methods.

The approach is named 3Zero, as it is intended to provide a zero touch, zero latency in service set up and zero fault as a consequence of service being non responding within the strict time frame specification for URLLC.

The proposed approach is based on the construction of an architecture, named 3Zero, that is able to cover all the required passages previously outlined, which are here detailed:

1) **Data Capture**: in order to create the digital twin, probes are deployed, in form of micro-services, that are able to register the incoming request to the MEC unit and the time of the completion time of these requests. It is relevant to underline that it is important to sample all possible incoming request and the associate functions that the MEC unit uses to process them, in order to have enough information to create the digital representation of the overall ecosystem, that comprises the MEC unit, its requests and the functions that process them.
2) **Digital Twin Creation**: with the information gathered at step 1 it is possible to create a digital twins that represent the incoming request and the related functions that process them. This digital twin contain the information on the structure of the requests and the processing functions, together with the historical information on the time necessary to fulfill a given request.
3) **Digital Twin Hierarchization**: the digital twins described in the previous step represent the atomic element of the ecosystem. In order to represents composite element of the ecosystem, as a request which is made by several sub-request and therefore creates a set of order functions to be called to address, namely, a workflow of function calls. These composite functions are developed by intentionally imbuing the digital twin with properties that enable composability. A dedicated function is then able to create composites of digital twin, which are a hierarchical superior entity of digital twin, hence digital twins of digital twins. The structure is recursive, to allow n-layer hierarchies, with an n number that is not constrained as for the need to represent any MEC-based ecosystem. It is important to outline that in the hierarchy of digital twins, higher level digital twins contain a composite and processed version of the information that are present in its lower level digital twins, this allows for representation of behaviour statistics of the underlying digital twins' functions.
4) **Digital Twin Networking**: the digital twins, arranged in proper hierarchy as described in the previous step,

represent a static version of the ecosystem that do not comprise the inter-relationship between elements of ecosystem. To depict these connections, digital twins are linked through associations forming a Digital Twin Network, which provide information on how requests are inter-related and how functions are interconnected. This allows the association of the set of functions needed to satisfy any given composite request, in all its sub-requests, thus providing a representation of a workflow.

5) **Workflow conversion to mathematical model and quantitative prediction**: now that we have a representation of the workflow, it is possible to map it into a mathematical model, as Petri Net or Structured Tree [7], that allows quantitative analysis to find relevant insights. The insight we focused on this work is the performance of the MEC level (comprising all the MEC units over the applicative domain), or, more properly, the probability function (cumulative or density distribution function, respectively CDF or PDF) for the MEC level to be able to process a workflow within a given time.
6) **MEC configuration optimization**: now that we have the insights on MEC functions capability to address workflows, it becomes possible to determine which MEC level DTN configuration (the set of all active functions or DT micro-services in the network) has the highest probability to perform withing the SLA, both in terms of probability that a given request is submitted and the time necessary to complete it.
   As MEC units work via containerization, a given DTN configuration is the set of all active containers and the functions they contain: hence, configuration optimization within this approach is a resource management strategy where deployed (through containerization) functions are the ones that have the highest probability to complete the most probable incoming requests.

The above outlined steps are automatically managed by the proposed 3Zero architecture, which is based and fully compliant to the ITU-T-REC-Y.3090-202202 standard [8].

In order to allow the level of autonomy required by the above outlined approach, the architecture leverage the use of **Reflection** architectural pattern to support code introspection and modification at runtime, enhancing system flexibility and adaptability [9].

It includes capabilities such as retrieving class information, dynamic object instantiation, method invocation, property access and modification, interface and inheritance checks, and accessibility modifications.

The Reflection pattern divides the system into the Knowledge Level (KNL) and the Operational Level (OPL), with the former managing the virtualized (i.e. represented by DT and DTN) views, and the latter handling deployed, active services.

III. Experimentation

After designing the architecture as previously described, we implemented it and validated it, covering till now the steps from 1 to 5, through an experimentation conducted on a virtualized environment, by using Docker.

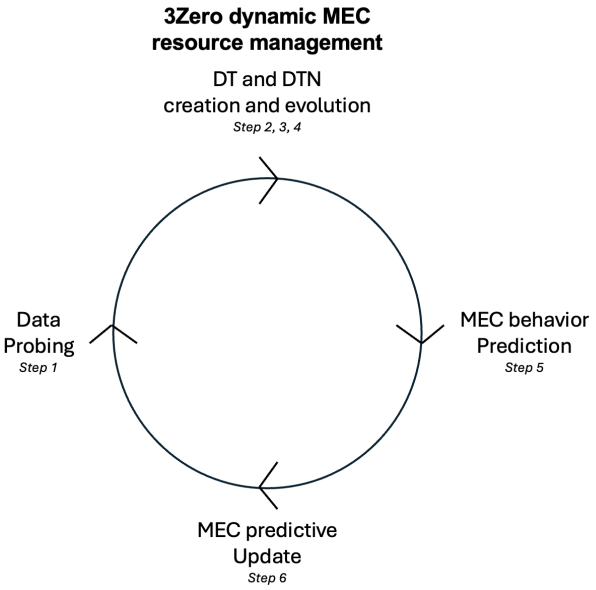

Fig. 1. The dynamic, continuous cycle of resource management allocation within the MEC in 3Zero architecture.

In the architecture, Dockerfiles are used for the definition of images containing all the necessary information to build and execute the services and their dependencies (libraries, modules, etc.).

In addition, these images bring with them other positive factors such as: Automated image creation, service reproducibility, image versioning, and simplified deployment.

We utlized the Docker virtualization environment to deploy the DTN as a network of interconnected containers, with each container housing a node of the Directed Acyclic Graph (DAG) of micro-services, which constitute the representation of DTs.

This process autonomously creates containers and installs the requisite services within them, exposing the top-level endpoints of the DTN to end-users, facilitating the submission of requests to the newly established micro-service-based DTN.

The prediction and optimization functions, as previously cited in the approach, are modelled as extension of the Knowledge Level (modeling and meta-modeling) of the reference architecture, and it is introduced thinking about the possibility of being able to provide the environment for the development of simulations of any type that aim to guarantee the instantiability of DT networks via docker containers and without the need to deploy services on real servers and giving the possibility of carrying out integration testing for more complex services that require the interaction of multiple micro-services.

For the creation of complex services, associated to hierachical functions interconnected via the DTN, we developed within the architecture a Docker controller, specifically aimed at instantiate micro-services using related .JAR/.WAR files,

creating the associated Docker containers and define the virtual network hosting the services.

To conduct further tests, we created different workflows, i.e. composition of requests, and applied uniform distributions probability to the estimated time for a given function to complete its request.

Although the architecture allows the use of various Distribution types such as exponential, exponential, polynomial, deterministic, etc., we used uniform distributions with [0,1] support for a first round of experimentation, due to their simplicity.

IV. Conclusions

The results obtained so far via experimentation show that near real time dynamic MEC resource management that enables ultra-reliable and low latency communications (URLLC) is possible under the proposed 3Zero approach, at least till step 5.

As for the step 6, it is known from literature that it possible to perform configuration optimization, with a similar approach to the one used in step 5. Next activities would integrate step 6 within the existing implementation of the 3Zero, using an already validated approach, outlined in [10].

The approach proposed in this paper differs and innovate with respect to current ones in its relying on quantitative predictive methods to evaluate a given MEC level DTN configuration to withstand the probable incoming request and thus to proactively and dynamically rearrange the configuration to continuously maximize the probability that all requests are managed within the URLLC SLAs.

Moreover, this approach provide benefit from an economic perspective, by dynamically optimizing resource management within a MEC, thus allowing maximum exploitation of a given MEC unit and making unnecessary horizontal or vertical scaling of the hardware.

There are also positive effects in terms of limitation of non renewable resource utilization, as functions are instantiated in the MEC only if are deemed necessary. It is relevant to note that even if this is also offered by out-of-the-box container resource managers, such tools instantiate a service only after a request has arrived, thus making them not employable in a URLLC scenario.

The promising proposed approach needs to be validated on a live scenario: the next steps in this research is to apply the 3Zero on a working MEC provided by WindTre, in order to compare the result so far obtained via computer simulation in a field environment.

Acknowledgment

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE0000001 - program "RESTART").

This work has received funding from the EU H2020 research and innovation programme (MSCA GA No 953442, IoTalentum).

References

[1] H. Lasi, P. Fettke, H. Kemper, T. Feld, and M. Hoffmann, "Industry 4.0," *Business & Information Systems Engineering*, vol. 6, no. 4, pp. 239–242, 2014.

[2] J. Leng, W. Sha, B. Wang, P. Zheng, C. Zhuang, Q. Liu, T. Wuest, D. Mourtzis, and L. Wang, "Industry 5.0: Prospect and retrospect," *Journal of Manufacturing Systems*, vol. 65, pp. 279–295, 2022.

[3] K. Jiang, H. Zhou, X. Chen, and H. Zhang, "Mobile edge computing for ultra-reliable and low-latency communications," *IEEE Communications Standards Magazine*, vol. 5, no. 2, pp. 68–75, 2021.

[4] 3GPP, "3gpp tr 38.913 v15.0.0: Study on scenarios and requirements for next generation access technologies; (release 15)," Tech. Rep., 2018.

[5] Stefanović, "Industry 4.0 from 5g perspective: Use-cases, requirements, challenges and approaches," in *2018 11th CMI International Conference: Prospects and Challenges Towards Developing a Digital Economy within the EU*, 2018, pp. 44–48.

[6] S. Abraham, A. K. Paul, R. I. S. Khan, and A. R. Butt, "On the use of containers in high performance computing environments," in *2020 IEEE 13th International Conference on Cloud Computing (CLOUD)*, 2020, pp. 284–293.

[7] L. Carnevali, R. Reali, and E. Vicario, "Eulero: A tool for quantitative modeling and evaluation of complex workflows," in *International Conference on Quantitative Evaluation of Systems*. Cham: Springer International Publishing, 2022.

[8] International Telecommunication Union, "ITU-T Y.3090: TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU," ITU, Standard Y.3090, February 2022, future networks.

[9] F. Buschmann, R. Meunier, H. Rohnert, P. Sommerlad, and M. Stal, *Pattern-Oriented Software Architecture: A System of Patterns*. Germania: Wiley, 2013.

[10] L. Carnevali, M. Paolieri, B. Picano, R. Reali, L. Scommegna, and E. Vicario, "A quantitative approach to coordinated scaling of resources in complex cloud computing workflows," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 9, no. 1, p. 12, 2024.