Automatic positioning of AI microservices on NextG networks to support interactive holograms[⋆]

Martina Di Bratto[1,*,†], Antonio Origlia[2,†], Jaime Llorca[3,†], Andrea Detti[4,†], Alessandro Mauro[2,†], Marco Grazioso[1,†], Vincenzo Norman Vitale[2,†], Valentina Russo[1,†], Azzurra Mancini[1,*,†], Alessandro Perrino[5,†], Nicholas Napolitano[5,†], Giovanni Della Corte[5,†], Antonia Maria Tulino[2,†], Sergio Piane[6,†] and Margherita Tennirelli[6,†]

[1]*Logogramma s.r.l., Naples, Italy*
[2]*University of Naples Federico II, Naples, Italy*
[3]*University of Trento, Trento, Italy*
[4]*University of Rome Tor Vergata, Rome, Italy*
[5]*Fastweb S.p.A., Italy*
[6]*Alkedo Produzioni S.r.l., Pisa, Italy*

Abstract

Among applications emerging from the advent of new ways to implement Artificial Intelligence, Virtual Humans are a popular example to showcase the capabilities of language models. Such artifacts are complex to manage both in terms of interaction management, with LLMs starting to show their weaknesses in practical use cases but also from the point of view of network management. When Virtual Humans are implemented as remote services, the possibilities offered by NextG networks are stressed since they should provide fast, reliable and self-adaptive configurations to support streamed transmission of dynamically generated contents. Here we present the system architecture that is currently being developed in the framework of the RESTART project to show how intelligent management of microservices over the network can support explainable AI approaches powering holographic Virtual Humans interfaces.

Keywords

NextG networks, XR experiences, RESTART project, virtual humans, conversational AI

1. Introduction

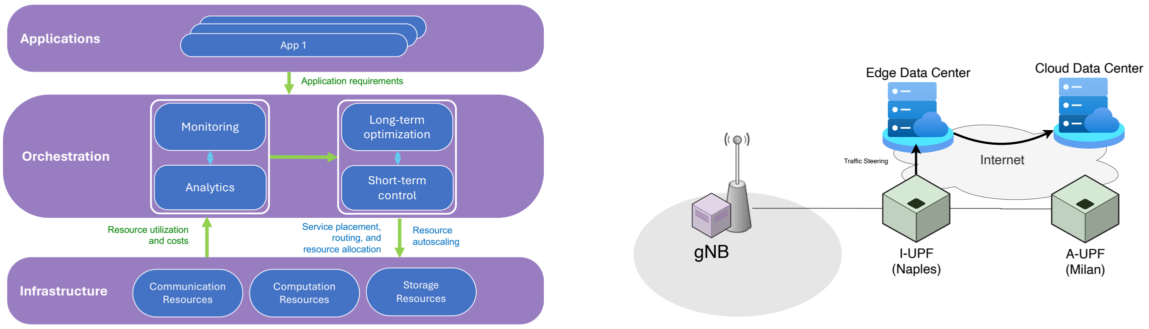
The evolution of Next Generation (NextG) networks introduces transformative challenges and opportunities, particularly regarding service orchestration. Emerging services are increasingly characterized by dynamic, real-time data streams and complex microservice architectures requiring flexible and efficient management strategies. This paper proposes a dual-timescale orchestration strategy that blends a long-term centralized optimization framework, embodied in the IDAGO algorithm, with distributed short-term control policies leveraging Kubernetes-based autoscaling. Furthermore, the integration of advanced Conversational AI and immersive holographic platforms underscores the complexity and interactivity in forthcoming application scenarios. By combining rigorous optimization and dynamic responsiveness, the proposed architecture seeks to optimize resource allocation, maintain service quality, and adapt seamlessly to real-time fluctuations, providing robust and reliable orchestration for sophisticated NextG applications. Interactive applications based on Natural Language Processing techniques, now powered by Generative AI models, have become popular as a way to showcase the

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

[*]Corresponding author.

[†]These authors contributed equally.

mdibratto@logogramma.com (M. Di Bratto); antonio.origlia@unina.it (A. Origlia); jaime.llorca@unitn.it (J. Llorca); andrea.detti@uniroma2.it (A. Detti); alessandro.mauro3@unina.it (A. Mauro); mgrazioso@logogramma.com (M. Grazioso); vincenzonorman.vitale@unina.it (V. N. Vitale); vrusso@logogramma.com (V. Russo); amancini@logogramma.com (A. Mancini); alessandro.perrino@fastweb.it (A. Perrino); nicholas.napolitano@fastweb.it (N. Napolitano); giovanni.dellacorte@fastweb.it (G. Della Corte); antoniamaria.tulino@unina.it (A. M. Tulino); s.piane@alkedoproduzioni.com (S. Piane); m.tennirelli@alkedoproduzioni.com (M. Tennirelli)



(a) (b)

Figure 1: (a) Analytics-driven Multi-scale Orchestration System Architecture. (b) Underlay network and cloud system.

potential of these new technologies. A particularly interesting case is represented by Holographic Virtual Humans providing different kinds of services[1]. In explainable architectures, supporting communication between remote microservices and the client platforms is of critical importance because of real-time communication requirements. Service placement plays a fundamental role in reducing latency, especially in an Edge Computing scenarios.

2. Intelligent service placement

The orchestration of emerging NextG services—comprising disaggregated microservice chains and real-time data streams—requires not only globally optimal end-to-end service deployments but also adaptive mechanisms to respond to real-time changes in network conditions and service demands. This section presents a dual timescale orchestration strategy, built around the cloud network flow optimization framework and associated IDAGO algorithm[1]. The strategy combines a centralized, information-aware end-to-end optimization algorithm for long-term placement, with a distributed short-term control policy for dynamic resource autoscaling in response to changes in traffic conditions. In line with multi-scale orchestration solutions such as [2, 3], we envision end-to-end service optimization algorithms like IDAGO running at centralized controllers (see Fig 1a) that operate at a longer timescale and can leverage a global network view to optimize end-to-end service distribution [4, 5, 6, 7, 8]. These are complemented by distributed control policies operating at a shorter timescale to provide fast reactions based on local, real-time observations [9, 3, 10]

In this context, the goal of the long-term global optimization algorithm is to minimize overall resource cost subject to end-to-end service latency constraints. These constraints are computed based on the average service demands and resource capacities/costs estimates. Complementary, the role of the distributed short-term control policies is to accommodate real-time traffic variations caused by the stochastic nature of service demands via dynamic resource autoscaling. A commonly used mechanism for implementing such policies is the Kubernetes Horizontal Pod Autoscaler (HPA). However, more sophisticated approaches, such as those based on reinforcement learning (RL) [11], can also be employed to achieve finer-grained and adaptive control.

2.1. Dual timescale orchestration

The end-to-end optimized deployment of service functions is performed using IDAGO (Information-Aware DAG Orchestration), a polynomial-time approximation algorithm designed for the orchestration of services modeled as information-aware directed acyclic graphs (DAGs) [1].

Traditional Virtual Network Embedding (VNE)-based approaches are insufficient for orchestrating information-aware DAGs because they do not support stream replication and fail to exploit multicast

[1] The holographic platform is developed by Logogramma with the support of Alkedo Produzioni.

opportunities. IDAGO overcomes these limitations by: a) Exploiting the multicast and replication capabilities of real-time service flows; b) leveraging the recently proposed Cloud Network Flow (CNFlow) optimization framework, which generalizes previous formulations by modeling the joint placement, routing, and resource allocation problem as an *information flow* problem over a cloud-augmented network graph [4, 12]; c) transforming DAGs into functionally equivalent forests to enable the use of tree-based approximation techniques. IDAGO follows a six-step procedure:

1. DAG-to-Forest Transformation: The original information-aware service DAG $\mathscr{R}$ is transformed into a functionally equivalent forest $\mathscr{R}_T$, where each tree corresponds to a destination function. This removes branching by replicating functions with multiple outputs, preserving information flow semantics.
2. LP Relaxation: The MILP for the IA-DAG-DTR problem is relaxed to a linear program on $\mathscr{R}_T$, yielding fractional flows over the cloud-augmented network.
3. Flow Decomposition: Each fractional LP solution is decomposed into a convex combination of valid tree embeddings. For each service tree in the forest, the algorithm identifies a set of mappings from service functions to compute nodes and from commodities to network paths, each with an associated selection probability.
4. Randomized Rounding: For each service tree, a valid embedding is randomly selected based on the decomposition probabilities.
5. Information Flow Computation: Using the selected embeddings, actual flows are computed by taking, for each link, the maximum rate across all commodities carrying the same object; total flow is then summed over all objects.
6. Validation and Iteration: Leveraging its polynomial-time nature and constant-factor, multi-criteria approximation guarantees, IDAGO repeats the rounding process until a solution satisfies the approximation thresholds or the maximum number of iterations is reached.

As services operate, traffic patterns and user demands naturally evolve. To maintain performance, a short-term Kubernetes-based autoscaler dynamically adjusts resource allocations (e.g., CPU and memory) in response to real-time load fluctuations.

Operating at a finer timescale, the Kubernetes autoscaler monitors key runtime metrics—primarily the output data rates of microservices—and adjusts the CPU and memory limits assigned to each microservice pod accordingly. Its primary goals are to: a) maintain service performance by preventing bottlenecks due to under-provisioning; b) optimize resource usage by scaling down when resource demand is low; c) ensure responsiveness between optimization cycles without altering placement decisions made by the long-term algorithm.

The autoscaler is integrated with the Kubernetes orchestration layer, leveraging native Kubernetes APIs. It continuously queries monitoring agents (e.g., Prometheus) for real-time metrics associated with each microservice instance. Based on predefined thresholds and the observed output rate dynamics (e.g., for services like Synthesis and Personalization), the autoscaler updates the resource allocation specifications (e.g., CPU requests/limits) of the pods. These updates are enforced through the Kubernetes control loop, which triggers pod rescheduling if necessary. The autoscaler operates under the constraint of fixed microservice placements—i.e., it does not migrate services across nodes. It is stateless with respect to service dependencies, focusing only on individual service metrics rather than global system objectives.

3. Use case

The application will consist of an interactive Virtual Human presenting Cultural Heritage contents, assuming the case of temporary exhibits to stress the dynamic nature of the solutions studied in RESTART. The system will showcase the situation in which temporary museum exhibits need to be introduced to museum visitors. The application, in its current design, will support a conversational experience aimed at collecting interests and preferences from the users (profiling) to help them select

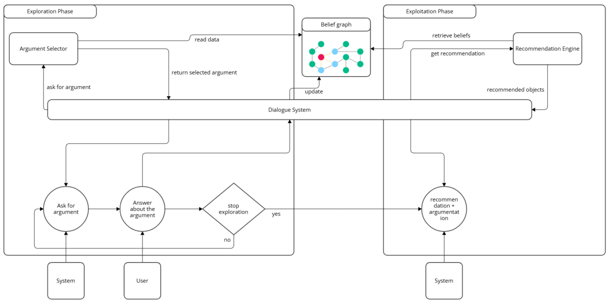


Figure 2: Interaction scheme showing the modules involved in the two phases of the recommendation: 1) exploration phase, for retrieving user preferences and storing them as beliefs; 2) exploitation phase leveraging the collected beliefs and generating recommendations with supporting arguments.

items of interest within the exhibit (recommendation) using personalised messages to support the suggestions (argumentation).

3.1. Interaction design

Argumentative Conversational Recommender Systems (A-CoRS) represent an evolution of Conversational Recommender Systems (CoRS), which recommend relevant information based on user interactions through spoken or textual dialogue. Situated within the field of Argumentation-Based Dialogue (ABD) [13], A-CoRS aim not only to match user preferences but to persuade by aligning recommendations with the user's beliefs and intentions [14]. Grounded in cognitive pragmatics, the proposed system integrates principles such as credibility, importance, relevance, and (un)likeability [15], mapping them onto a graph-based model. Through network analysis with the HITS algorithm [16], authority and hub scores quantify the credibility and importance of each node, while dialogically built belief networks inform relevance via entropy measures, and hard evidence supports the (un)likeability dimension [14]. A Retrieval-Augmented Generation (RAG) approach [17] feeds a Large Language Model (LLM) with structured graph data to produce recommendations followed by context-aware, argument-based explanations [18], enhancing transparency and mitigating hallucinations. The system's development follows two main goals: a hybrid architecture that merges computational dialogue modeling with a focus on beliefs, goals, and rational behavior [19], and a user-centered analysis of trust in machines. Fig. 2 represents the dialogue flow and how the developed modules are used in the recommendation dialogue.

4. Proposed architecture

The proposed setting involves the use of multiple technologies and services distributed over a network organised in a client layer, the holographic platform, an edge layer, running the application and its support services, and a cloud layer, containing aggregated data for analysis purposes. The edge layer represents the most dynamic part of the network, and it is managed using the service placement algorithms, dynamically adjusting location and resource management to optimise traffic and reduce latency. The network infrastructure will support the deployment of the application as described in Section 3. The different components involved in the development of the demonstrator have different technological requirements that need to be considered to design the configuration of the infrastructure. The main components categories can be summarised as: **Compute-oriented services:** services that need access to compute-oriented GPUs to provide their services. These include machine learning models for speech recognition, synthesis, animation, Natural Language Understanding etc...; **Graphics-oriented services:** services that need access to graphics-oriented GPUs to render the front-end of the application, consisting of a virtual human to be projected on the holographic display; **Management-oriented services:** services that need access to CPU, RAM and data storage capabilities to support fast data retrieval and processing for interaction management.

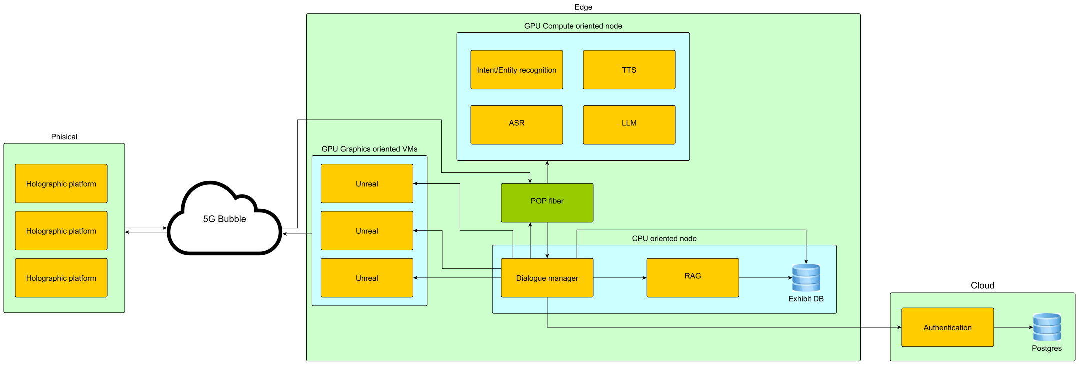


Figure 3: The proposed network architecture demonstrating the organisation of the foreseen application. The represented service positioning is the one expected by expert system designers and should be automatically identified and managed by the proposed AI approaches.

In the current hypothesis, a dedicated edge node will be deployed to host compute-oriented services, another edge node will be dedicated to hosting management-oriented services. The virtual machines running the front-end 3D application will also run on the edge on dedicated hardware. Authentication logic will, instead, be deployed on the cloud as, in this case, centralisation supports security. Also, authentication processes have less strict requirements concerning data transfer speed. The infrastructure will make use of the AI strategies for initial services placement and dynamic positioning, as described in 2. These services will support the back-end controlling the interactive application using natural language processing techniques, as described in 3. A schema of the candidate architecture organisation that is currently being developed is shown in Figure 3.

Figure 1b illustrates the network and cloud infrastructure deployed to support the holographic application. The physical holographic platform connects to a standalone 5G cell in Naples. The 5G PDU session carrying the application traffic is anchored at an intermediate user plane function (I-UPF) in Naples and steered to a co-located edge data center. This topology yields very low latency and high throughput between the holographic platform and the application microservices. Both the edge and core data centers host a Kubernetes (K8s) environment interconnected via an Istio service mesh configured for a multi-cluster deployment. The edge K8s cluster is augmented with GPU-enabled nodes to run Pods dedicated to AI processing, while the edge data center also operates virtual machines (VMs) equipped with graphics-oriented GPUs for Unreal Engine rendering and streaming.

5. Conclusions

This paper presented a comprehensive dual-timescale orchestration strategy aimed at addressing the dynamic demands of NextG service deployments. The innovative approach integrates the IDAGO algorithm for long-term global optimization of service placements, routing, and resource allocation, with responsive, distributed short-term Kubernetes-based autoscaling mechanisms. The described implementation within immersive and interactive applications, such as virtual humans in cultural heritage scenarios, demonstrates the practical value and robustness of the architecture. By effectively balancing global resource efficiency with local adaptability and responsiveness, the proposed solution represents a significant advancement in orchestrating complex, real-time interactive services, paving the way for future developments in NextG technologies.

Acknowledgments

This work was partially supported by the EU under Italy's NRRP (NextGenerationEU), project ”Telecommunications of the Future” (RESTART, PE00000001, E83C22004640001), the PRIN project ”Resilient delivery of real-time interactive services” (E53D23000550006), and the US NSF RINGS program.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. Mauro, A. M. Tulino, J. Llorca, End-to-end orchestration of nextg media services over the distributed compute continuum, arXiv preprint arXiv:2407.08710 (2024). URL: https://arxiv.org/abs/2407.08710, under review at IEEE Transactions on Mobile Computing.

[2] Q. Pagliuca, L. J. Chaves, P. Imputato, A. Tulino, J. Llorca, Dual timescale orchestration system for elastic control of nextg cloud-integrated networks, in: 2024 27th Conference on Innovation in Clouds, Internet and Networks (ICIN), IEEE, 2024, pp. 234–241.

[3] S. Chen, J. Li, Q. Yuan, H. He, S. Li, J. Yang, Two-timescale joint optimization of task scheduling and resource scaling in multi-data center system based on multi-agent deep reinforcement learning, IEEE Transactions on Parallel and Distributed Systems (2024).

[4] M. Barcelo, J. Llorca, A. M. Tulino, N. Raman, The cloud service distribution problem in distributed cloud networks, in: IEEE International Conference on Communication (ICC), IEEE, 2015.

[5] Y. Cai, J. Llorca, A. M. Tulino, A. F. Molisch, Joint compute-caching-communication control for online data-intensive service delivery, IEEE Transactions on Mobile Computing (2023).

[6] K. Poularakis, J. Llorca, A. M. Tulino, L. Tassiulas, Approximation algorithms for data-intensive service chain embedding, in: ACM Mobihoc, ACM, 2020, pp. 131–140.

[7] M. Michael, J. Llorca, A. Tulino, Approximation algorithms for the optimal distribution of real-time stream-processing services, in: ICC 2019-2019 IEEE International Conference on Communications (ICC), IEEE, 2019, pp. 1–7.

[8] M. Rost, E. Döhne, S. Schmid, Parametrized complexity of virtual network embeddings: Dynamic & linear programming approximations, ACM SIGCOMM Computer Communication Review (2019).

[9] H. Feng, J. Llorca, A. M. Tulino, A. Molisch, Optimal dynamic cloud network control, IEEE/ACM Transactions on Networking 26 (2018) 2118–2131.

[10] F. Mason, G. Nencioni, A. Zanella, Using distributed reinforcement learning for resource orchestration in a network slicing scenario, IEEE/ACM Transactions on Networking (2022).

[11] V. N. Vitale, A. M. Tulino, A. F. Molisch, J. Llorca, A flexible multi-agent deep reinforcement learning framework for dynamic routing and scheduling of latency-critical services, in: Proceedings of the IEEE International Conference on Communications (ICC), IEEE, Montreal, Canada, 2025. To appear.

[12] J. Llorca, A. M. Tulino, Cloud network flow: Understanding information flow in nextg cloud-integrated networks, arXiv preprint (2024).

[13] H. Prakken, Historical overview of formal argumentation, in: Handbook of formal argumentation, College Publications, 2018, pp. 73–141.

[14] M. Di Bratto, A. Origlia, M. Di Maro, S. Mennella, Linguistics-based dialogue simulations to evaluate argumentative conversational recommender systems, User Modeling and User-Adapted Interaction (2024) 1–31.

[15] F. Paglieri, C. Castelfranchi, Arguments as belief structures: Towards a toulmin layout of doxastic dynamics? (2005).

[16] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM (JACM) 46 (1999) 604–632.

[17] K. Meduri, G. S. Nadella, H. Gonaygunta, M. H. Maturi, F. Fatima, Efficient rag framework for large-scale knowledge bases, Efficient RAG framework for large-scale knowledge bases (2024).

[18] M. Grazioso, M. Di Bratto, A. Mancini, V. Russo, Towards an explainable argumentation-based dialogue pipeline for conversational recommender systems, in: The 1st Workshop on Risks, Op-

portunities, and Evaluation of Generative Models in Recommender Systems, ROEGEN-RECSYS'24, 2024.

[19] C. Castelfranchi, Reasons: Belief support and goal dynamics, Mathware & soft computing. 1996 Vol. 3 Núm. 1 [-2] p. 233-247 (1996).