Extreme Edge Computing Challenges on the Edge-Cloud Continuum

Sherif B. Azmy[1], Rawan F. El-Khatib[1], Nizar Zorba[2], Hossam S. Hassanein[3]
[1]Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada
[2]Electrical Engineering Department, Qatar University, Doha, Qatar
[3]School of Computing, Queen's University, Kingston, ON, Canada
Email: {sherif.azmy,rawan.elkhatib}@queensu.ca, nizarz@qu.edu.qa, hossam@cs.queensu.ca

***Abstract*—The evolution of AI and LLMs is transforming service provision, revealing the limitations of traditional cloud computing in latency and privacy. Edge computing overcomes such limitations by moving resources closer to users, but often overlooks user-owned devices. This work examines the Edge-Cloud continuum, including the role of Extreme Edge Computing (XEC), which extends computational reach to user-owned devices. We discuss how this continuum can lead to reduced latency, enhanced privacy, and improved energy efficiency, while noting the challenges posed by heterogeneous and unpredictable user-owned resources.**

I. Introduction

Artificial Intelligence (AI) and Large Language Models (LLMs) introduced a radical change to the service provision market. Both technologies have the potential to enable a wide range of innovative and disruptive applications.

However, that potential is limited by the traditional structure of service provision. Cloud computing, as a centralized paradigm, is not capable of catering to the demands of these new applications. In particular, aspects regarding latency and privacy are key limitations. Many of these applications require the deployment of a sandboxed AI model on the Edge to guarantee privacy and minimal latency. For example, Magic Circle seeks to utilize an LLM to create a group game in which the model listens to a conversation between friends in which they tell stories about themselves, then the LLM generates a trivia [1]. Such an LLM needs to be highly responsive to ensure the participants' engagement, and to be secure as such stories told are bound to have personal details; both are aspects at which the enterprise-owned cloud lacks the capability to ensure both of these aspects [2], [3].

Edge computing seeks to overcome the limitations of the cloud by dividing and conquering demand. It achieves so by bringing service deployment closer to the end-user [4], [5]. Edge Computing is a promising solution that has proven its reliability, however it is not fully democratized and the rules and regulations regarding the training of models remain vague and unclear. In addition, traditional edge computing stops at the bounds of enterprise-owned infrastructure, leaving out potential to achieve lower latency at the very extreme of the edge, on user-owned devices [3]. Extreme Edge Computing (XEC), that is, computing done as close as possible to the end-user (regardless of the devices being user-owned or enterprise-owned), extends the Edge-Cloud continuum to actively include user-owned devices [6]. In addition to accessing a wealth of resources, it also improves the effective usage of idle resources - in contrast to servers and the like - distributing energy costs, leading to a greener form of edge computing.

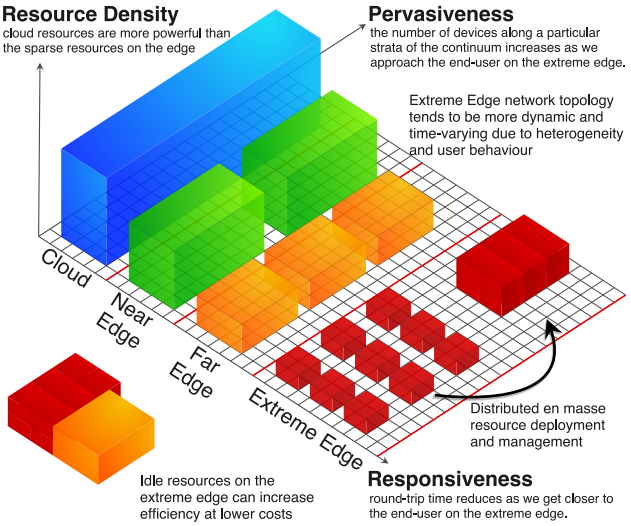

Fig. 1. Traversing the Edge-Cloud Continuum introduces a number of trade-offs. Illustrated in the figure is the trade-off between resourcefulness, responsiveness, and pervasiveness. It is possible, in theory, to leverage idle resources in a distributed fashion, to provide service. However, the nature of the extreme edge and its devices introduce barriers to achieving this.

While still in its infancy, numerous technologies such as 6G and beyond, WebAssembly [7], Unikernels and Microcontainers [5] bring computing on the extreme edge closer to being a reality. However, leveraging user-owned idle resources on the Extreme Edge comes with challenges that stem from the heterogeneous and unpredictable nature of devices on the extreme edge [3], [4]. In this work, we provide an overview of the Edge-Cloud continuum, highlight a few challenges on it, as well as emphasize its fortuitous potential in reducing latency, and enhancing privacy.

II. The Edge-Cloud Continuum

In Figure I, we provide a perception of the Edge-Cloud Continuum that looks at the resource density which are highly

concentrated and rich in the cloud, as opposed to the other extremity on the extreme edge in which resource density is scarce, and resources are sparsely distributed on different devices. We also illustrate the pervasiveness in the count of devices. The cloud is comprised of centralized data centers and sophisticated hardware. However, this gets more distributed and more pervasive as we traverse the continuum through the near and far edges, with the highest pervasiveness and ubiquity being on the extreme edge; user-owned devices are numerous and go everywhere. By being in the user's immediate vicinity, user-owned devices are theoretically capable of achieving the best possible raw responsiveness. Efficient orchestration of such resource-constrained devices can synthesize a "meta edge server", providing service at the very extreme edge of the network. However, such orchestration not possible without taming the heterogeneity and uncertainty inherent to XEC or handling its dynamic nature.

III. Extreme Challenges Along the Continuum

An inherent characteristic of approaching the extreme edge is the intermittent availability of resources and the unique access profile per user (which stems from human behaviour that is believed to be predictable to an extent [8]). This intermittence introduces substantial levels of uncertainty and risk to service providers, which blur their ability to provide a certain Service Level Agreement (SLA) in addition to the burden of losing service due to mobility, lack of control, or unreliable connectivity. As a result, techniques such as device discovery and recruitment, re-attribution, prediction, etc., are vital for the successful orchestration of XEC in a semi-centralized manner, but comes at significant overhead. A completely distributed approach, such as self-organizing compute clusters and having an open-market in which computational resources of such devices become a marketable commodity can be more efficient [3].

As we traverse the continuum from the Near Edge (or Fog) to the Far Edge and the Extreme Edge, the demand for self-organization and decentralization increases, and challenges arise in the following aspects:

a) Privacy: Privacy is a major concern as it seeks to restrict service providers from information about the end-user and their usage of the service. In AI and LLM-driven world, such data enhances models. Thus, the nuances of what can be shared need to be explored. When it comes to the extreme edge, guaranteeing privacy increases in complexity due to the involvement of such third-party user-owned infrastructure. Service providers might be honest-but-curious or even outright malicious. With the novel applications on the Extreme Edge, privacy becomes more sensitive.

b) Security: sandboxed execution protects the user-device, but if the device itself is malicious, it might jeopardize the end-user. Likewise, if it travels a long way to a centralized cloud, although encrypted, it remains exposed.

c) Connectivity: heterogeneous devices employ different communication technologies at different locations, and they are subject to their owners' whims and mobility.

d) Modularity: lightweight virtualization is not sufficiently lightweight. WebAssembly, microcontainers, and unikernels are nascent technologies that are not entirely portable from one device to another, particularly in the case if a node churns and another node replaces it instantly.

e) Complexity: highly dynamic environments incur frequent node churning and limited resources, requiring real-time self-organization to guarantee a service and its provision.

f) Resourcefulness: devices on the extreme edge are weak on their own, but when regarded as clusters they are capable of providing resource-intensive services (e.g., rendering game assets on an intercity train [5]). Moreover, exploiting social patterns to predict spatiotemporal changes can help increase XEC's resourcefulness.

g) Uncertainty: the closer we get to the extreme edge the more dynamic and time-varying the compute-resources topology becomes. A major source of this uncertainty is human behaviour. Thus, predictive techniques or active interventions targeting the user themselves (e.g., incentives, gamification, etc) are recommended.

IV. Conclusion

Navigating the Edge-Cloud continuum requires careful consideration of different aspects along it. While there is a wealth of resources on user-owned devices, XEC comes at a significant cost, requiring highly dynamic, reliable, and analytic orchestration. While the current state-of-the-art seems to hint the rise of the XEC on the horizon, there remains much work to be done to address these challenges and tap into the true potential of XEC.

Acknowledgement

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant ALLRP 549919-20; in part by the Distributive, Ltd. This work was also supported in part by Qatar University under Grant IRCC-2024-494.

References

[1] A. Chen, J. Lu *et al.*, "Next-Gen Gaming: AI Souls, Real-time Culture, Personalized Avatars," A16Z Podcast, Andreesen Horowitz, 4 2024. [Online]. Available: https://a16z.com/podcast/next-gen-gaming-ai-souls-real-time-culture-personalized-avatars/

[2] M. Nieke, L. Almstedt, and R. Kapitza, "Secure mobile webassembly services on the edge," ser. Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking, 2021, p. 13–18.

[3] S. B. Azmy, N. Zorba, and H. S. Hassanein, "Incentive-Vacation Queueing for Edge Crowd Computing," *IEEE Internet of Things Journal*, vol. 11, no. 8, p. 13167–13179, 2024.

[4] D. Milojicic, "The Edge-to-Cloud Continuum," *Computer*, vol. 53, no. 11, p. 16–25, 2020.

[5] A. J. Ferrer, "Beyond Edge Computing, Swarm Computing and Ad-Hoc Edge Clouds," 2023.

[6] S. B. Azmy, N. Zorba, and H. S. Hassanein, "Incentive-vacation queueing in extreme edge computing: An analytical reward-based framework," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 2183–2195, 2024.

[7] P. P. Ray, "An Overview of WebAssembly for IoT: Background, Tools, State-of-the-Art, Challenges, and Future Directions," *Future Internet*, vol. 15, no. 8, p. 275, 2023.

[8] Q. Liang and E. Modiano, "Survivability in Time-Varying Networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 9, p. 2668–2681, 2016.