Scaling the Metaverse: An AI Perspective

Tania Lorido-Botran
Roblox
San Mateo, CA (USA)
tbotran@roblox.com

ABSTRACT

When one hears the word Metaverse, it is automatically associated with millions of users, immersive experiences and its potential to change our lives. But, what enables the Metaverse to function at such a scale? Through an AI lens, let us examine the challenges associated with handling 55 million daily users within the Roblox Metaverse.

Infrastructure, A tale of Edge Computing: The Tale of the Metaverse starts with two key components, which are the end-users and the infrastructure necessary to power it. On one side, there are millions of users connected to multiple experiences (games, concerts, activities…), sharing audio, video and text information. Measuring the Quality of Experience (QoE) is very subjective and user-dependent, but can be modeled based on two main factors: low latency and good quality graphics (frames per second or FPS). In order to ensure low latency, the computation needs to happen geographically close to the end user. The first option is using the client device (whenever is logical to do so) to slash the latency completely. However, the game logic still needs to be executed in a server and some workloads might need to be offloaded when the client device does not have enough resource computation. As opposed to traditional data centers, edge computing places computation nodes all around the globe, ensuring that the end-users are geographically close to them. Although some services might require bare-metal execution, most experiences can be executed on top of some form of virtualized environment such as containers to ease deployment, scaling and migration.

Multi-model Machine Learning Workloads: Most workloads can be represented as a DAG workflow (Directed Acyclic Graph), that is, several tasks with multiple dependencies among them. Some of the tasks might be executed in parallel, while others require the upstream tasks to finish before the downstream tasks can start. Let us look at a very prevalent type of Metaverse workloads. Each user in the Metaverse is represented with an avatar, which can reflect real time features of the human behind the device, such as facial expressions [1].

ICPE'23 Companion, April 15-19, 2023, Coimbra, Portugal.

ACM ISBN 979-8-4007-0072-9/23/04.
https://doi.org/10.1145/3578245.3584920

Such functionality is powered by Machine Learning (ML) or Deep Learning (DL) models, typically arranged in the form of a workflow that combines several models towards one goal (e.g. object detection). Auto-scaling Multi-model ML workloads imposes a new challenge, as it allows for multiple scaling options: different model variants (e.g. quantization, architecture), choice of accelerator, horizonal and vertical scaling for each separate model. The scaling plan has to accommodate for multiple objectives: it requires ensuring a minimum overall model performance, low end-to-end latency to satisfy user QoE, low energy consumption in certain devices (e.g. phones) and optimized resource usage across heterogeneous nodes.

Holistic Resource Optimization at Scale: The multi-model ML workload case is just one small example of what it takes to Scale the Metaverse. Overall, it requires orchestrating multiple layers of workloads across heterogeneous, distributed client and edge nodes. Reinforcement Learning has largely been applied for data center resource management problems, such us container placement optimization, network congestion control, data center cooling control or client/server task offloading. It has the potential to automate resource orchestration for the Metaverse edge infrastructure. ImpalaE [2] was proposed as an example of a distributed and scalable RL-based optimizer for edge environments that maximizes the overall resource utilization. However, ImpalaE requires to be expanded in various ways. A hierarchical, multi-agent approach would be needed to deal with several layers of scheduling: mapping end-users to different edge nodes, assigning containers to different locations, individual workflow autoscaling, etc. Besides that, making RL production-ready still has some pitfalls (i.e. the need for offline training or hyperparameter tuning) that still need to be overcome.

CCS Concepts

• Artificial Intelligence, Machine Learning, Operations Research

Author Keywords

Metaverse; reinforcement learning; data center; resource optimization

ACM Reference format:

Lorido-Botran,Tania. 2023. Scaling the Metaverse: An AI Perspective. In *Companion of the 2023 ACM/SPEC International Conference on Performance Engineering, April 15-19, 2023, Coimbra, Portugal.* ACM, New York, NY, USA, 2 pages. https://doi.org/ https://doi.org/10.1145/3578245.3584920

BIOGRAPHY

Dr. Tania Lorido-Botran is a Research Scientist at Roblox. Prior to that, she worked at Microsoft and the Pacific Northwest National Laboratory.

During her PhD, she had the opportunity to spend one year at Rice University and also did two internships at VMware and HP Labs. Dr. Lorido-Botran received her PhD from the University of Deusto in Spain with a Cum Laude distinction, and her master's degree in Distributed systems from University of the Basque Country with a highest marks distinction. Her current research interests span across ML for systems, data center sustainability and fault-tolerance.

REFERENCES

[1] Fast Facial Animation from Video, Navarro et al., SIGGRAPH Talks 2021

[2] ImpalaE: Towards an optimal policy for efficient resource management at the edge. Lorido-Botran, T., Bhatti, M.K.. doors 2021.