The Promise of Serverless Computing within Peer-to-Peer Architectures for Distributed ML Training

Amine Barrak

Department of Computer Science and Mathematics
University of Quebec at Chicoutimi, Québec, Canada
mabarrak@uqac.ca

Abstract

My thesis focuses on the integration of serverless computing with Peer to Peer (P2P) architectures in distributed Machine Learning (ML). This research aims to harness the decentralized, resilient nature of P2P systems, combined with the scalability and automation of serverless platforms. We explore using databases not just for communication but also for in-database model updates and gradient averaging, addressing the challenges of statelessness in serverless environments.

Introduction

Distributed machine learning emerges as an essential tool to manage the intricate dance between burgeoning data sizes and model complexities. By leveraging the strength of multiple computational nodes for parallel processing, ML accelerates large model training through parallel processing across multiple nodes. Datasets are spread across workers, each handling their model replicas and synchronizing periodically for model convergence. (Yuan et al. 2022).

Distributed training has given rise to a multitude of architectural designs, many of which are fundamentally based on the structures of the Parameter Server (PS) and Peer-to-Peer (P2P) architectures (Verbraeken et al. 2020). Each of these represents a unique methodology for orchestrating the management and distribution of tasks and data across nodes in a distributed system, with their own set of benefits and challenges. In the Parameter Server architecture, the worker nodes perform computations on their respective data partitions and communicate with the parameter server to update the global model (Li et al. 2013). In contrast, Peer-To-Peer architectures distribute the model parameters and computation across all nodes in the network, eliminating the need for a central coordinator (Šajina, Tanković, and Ipšić 2023).

Recognizing these challenges, the convergence of ML and serverless computing platforms has emerged as a compelling solution. Serverless computing offers the dual advantage of automating resource management while dynamically scaling resources, liberating developers from the intricate tasks of infrastructure management (Shafiei, Khonsari, and Mousavi 2022; Barrak, Petrillo, and Jaafar 2022). However, directly porting ML systems to a serverless environment presents limitations, including statelessness, restricted function communication, and limited execution times (Sarroca and Sánchez-Artigas 2024).

Within the vast scope of serverless architectures in distributed ML, a conspicuous gap remains. While the integration of serverless computing with the Parameter Server architecture has been extensively scrutinized, revealing benefits like cost savings (Sarroca and Sánchez-Artigas 2024), enhanced scalability (Grafberger et al. 2021), and performance boosts (Sampé et al. 2018), the Peer-to-Peer (P2P) framework, known for its decentralized nature and resilience, remains less explored in the context of serverless computing. Yet, the idea of melding the resilience and decentralization of P2P architectures with the dynamic scalability and automation of serverless platforms presents a tantalizing prospect for distributed ML training, potentially accelerating ML algorithms and offering transformative benefits to the ML community. It is in this juncture that our research resides.

Problem Statement: While the integration of serverless computing with the Parameter Server (PS) architecture in distributed Machine Learning (ML) has been explored, its combination with Peer-to-Peer (P2P) architectures remains understudied. This gap could hinder the potential benefits of decentralization, resilience inherent in P2P coupled with serverless scalability, and the transformative acceleration of ML algorithms for the broader ML community. Our research seeks to investigate the implications and advantages of this integration for optimized distributed ML training.

Research Questions and Timeline

We commence our exploration into the integration of serverless computing within the realms of the Machine Learning (ML) pipeline. Specifically, we aim to understand how serverless computing has been integrated into the various stages of the ML pipeline.

RQ1: How has serverless computing been integrated into the various stages of the ML pipeline?

Research Background: In June 2022, 53 papers touched upon this subject, with approximately 30% specifically addressing the incorporation of serverless computing for ML training. These findings were subsequently published in the IEEE Access Journal (Barrak, Petrillo, and Jaafar 2022).

Serverless in ML – The PS Architecture: The predominant methodology employed for ML training was the Parameter Server (PS) architecture. However, a critical limitation arises from this approach - a single point of failure at the server side, particularly as the model scales.

Primary Benefit of Serverless: One of the notable advantages of serverless computing is its inherent dependence on cloud providers, eliminating the need for direct infrastructure management.

Towards a Fault-Tolerant Architecture – P2P Integration: To circumvent the vulnerabilities of the PS architecture, we transitioned to a Peer-to-Peer (P2P) system. While this overcomes the centralized failure point, it introduces communication overhead between peers. Our inspiration to integrate serverless within P2P was driven by the desire for heightened fault tolerance, the specifics of which have been documented in an archive (Barrak, Petrillo, and Jaafar 2023).

RQ2: How can serverless computing support P2P architectures for distributed ML training?

To address the challenges posed by integrating serverless platforms into P2P architectures for distributed ML training, our proposed method comprises several stages: (1) Assigning datasets and splitting them into batches; (2) Computing gradients in parallel using serverless technology, after which the results are stored in a dedicated database; (3) Retrieving and averaging the computed gradients; (4) Communicating these averaged gradients with other peers in the network; and (5) Averaging all received gradients from the peers, which then allows for the model to be updated accordingly.

We identified that computing gradients is the most resource-intensive task. Leveraging serverless computing, we optimized parallel batch processing within each peer. The associated results showed a staggering 97.34% improvement in gradient computation time and up to 5.4 times more expensive when juxtaposed against traditional P2P training methods. This work has been published in IC2E conference on June 2023 (Barrak et al. 2023b). However, introducing serverless within peers led to communication overhead, primarily due to their stateless nature. This required constant interaction with databases for result storage.

We found that, within each peer, the training delays primarily stem from model updates and gradient averaging.

RQ3: How can we mitigate the communication constraints?

Several works have been proposed to optimise communication in distributed training (Abdi et al. 2023). Our exploration led us to Redis, which supports scripting within the database, and was adapted to suit our gradient averaging requirements. This significantly reduced the time overhead. We modified RedisAI, a module for Redis that adds AI capabilities to Redis, to support direct ML model updates within the database.

Additionally, to maintain the integrity of our Peer-to-Peer (P2P) model, we have implemented strong aggregation algorithms to mitigate any potential Byzantine behavior among peers. Our latest research findings were accepted at the QRS conference on September 21, 2023 (Barrak et al. 2023a).

RQ4: How do serverless ML training workflows differ between P2P and PS architectures, and what paths of convergence and divergence emerge when comparing their performance, cost-efficiency, and resilience?

Moving forward, our ambition for the next months is to craft a comprehensive mapping study that elucidates the pros and cons of P2P versus PS architectures in serverless ML training, with an emphasis on performance, cost-effectiveness, and resilience.

References

Abdi, A.; Rashidi, S.; Fekri, F.; and Krishna, T. 2023. Efficient distributed inference of deep neural networks via restructuring and pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 6640–6648.

Barrak, A.; Jaziri, M.; Trabelsi, R.; Jaafar, F.; and Petrillo, F. 2023a. SPIRT: A Fault-Tolerant and Reliable Peer-to-Peer Serverless ML Training Architecture. *Proceedings of the 23rd International Conference on Software Quality, Reliability and Security (QRS)*.

Barrak, A.; Petrillo, F.; and Jaafar, F. 2022. Serverless on Machine Learning: A Systematic Mapping Study. *IEEE Access*, 10: 99337–99352.

Barrak, A.; Petrillo, F.; and Jaafar, F. 2023. Architecting Peer-to-Peer Serverless Distributed Machine Learning Training for Improved Fault Tolerance.

Barrak, A.; Trabelssi, R.; Jaafar, F.; and Petrillo, F. 2023b. Exploring the Impact of Serverless Computing on Peer To Peer Training Machine Learning. *2023 IEEE 11th International Conference on Cloud Engineering (IC2E)*.

Grafberger, A.; Chadha, M.; Jindal, A.; Gu, J.; and Gerndt, M. 2021. Fedless: Secure and scalable federated learning using serverless computing. In *2021 IEEE International Conference on Big Data (Big Data)*, 164–173. IEEE.

Li, M.; Zhou, L.; Yang, Z.; Li, A.; Xia, F.; Andersen, D. G.; and Smola, A. 2013. Parameter server for distributed machine learning. In *Big learning NIPS workshop*, volume 6.

Šajina, R.; Tanković, N.; and Ipšić, I. 2023. Peer-to-peer deep learning with non-IID data. *Expert Systems with Applications*, 214: 119159.

Sampé, J.; Vernik, G.; Sánchez-Artigas, M.; and García-López, P. 2018. Serverless data analytics in the ibm cloud. In *Proceedings of the 19th International Middleware Conference Industry*, 1–8.

Sarroca, P. G.; and Sánchez-Artigas, M. 2024. Mlless: Achieving cost efficiency in serverless machine learning training. *Journal of Parallel and Distributed Computing*, 183: 104764.

Shafiei, H.; Khonsari, A.; and Mousavi, P. 2022. Serverless computing: a survey of opportunities, challenges, and applications. *ACM Computing Surveys*, 54(11s): 1–32.

Verbraeken, J.; Wolting, M.; Katzy, J.; Kloppenburg, J.; Verbelen, T.; and Rellermeyer, J. S. 2020. A survey on distributed machine learning. *Acm computing surveys (csur)*, 53(2): 1–33.

Yuan, B.; Wolfe, C. R.; Dun, C.; Tang, Y.; Kyrillidis, A.; and Jermaine, C. 2022. Distributed Learning of Fully Connected Neural Networks Using Independent Subnet Training. *Proc. VLDB Endow.*, 15(8): 1581–1590.