

Efficient and Reliable Systems for Building Foundation Models at Scale

Jingren Zhou

Alibaba Group

jingren.zhou@alibaba-inc.com

ABSTRACT

The rapid evolution of AI has led to increasingly large and complex foundation models with trillions of parameters and heterogeneous architectures, making efficient and reliable training and inference systems essential. Training such models requires large-scale distributed computation, effective overlap of computation and communication, sophisticated parallelization strategies, and robust fault-tolerant mechanisms. In addition, the rise of multi-turn and autonomous agents drives new demands for asynchronous reinforcement learning and post-training systems. Inference systems, on the other hand, must support diverse workloads with varying service-level agreements (SLAs), rapidly integrate engineering optimizations, and carefully balance trade-offs among throughput, latency, cost, and availability, particularly in distributed environments. In this talk, I will discuss the major systems challenges involved in building large-scale foundation models, drawing on our experiences developing Qwen large language models and Wan video generation models. I will also present ongoing research and system co-designs for improving training, post-training, and inference at scale, with the goal of making complex AI workloads in cloud environments more efficient, agile, and reliable.

PVLDB Reference Format:

Jingren Zhou. Efficient and Reliable Systems for Building Foundation Models at Scale. PVLDB, 19(12): 4982 - 4982, 2026.
doi:10.14778/3827998.3838714

1 BIOGRAPHY

Jingren Zhou is Chief Scientist and Senior Vice President at Alibaba, where he founded and led the development of Alibaba's flagship AI foundation models, including the Qwen and Wan model series. Today, Qwen is recognized as one of the world's leading large language models and among the most widely adopted open-source AI models globally. Previously, Jingren served as Chief Technology Officer at Alibaba Cloud, where he drove technology innovation and product development across a broad portfolio of cloud computing services. He also played a key role in building advanced AI technologies and infrastructure for personalized search, product recommendation, and advertising on Alibaba's e-commerce platforms. Before joining Alibaba, he was a veteran at Microsoft, focusing on big data and database research and development. His research interests include cloud computing, distributed systems, databases, and large-scale

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 19, No. 12 ISSN 2150-8097.
doi:10.14778/3827998.3838714

machine learning. He has served as PC co-chair and core committee member for many academic conferences and technical forums. He received his PhD in Computer Science from Columbia University and is a Fellow of ACM and IEEE.