The Life of a Token: from Words to Bits on the Wire

This talk will demystify the way we go from Generative Ai interactive chat to actual traffic across Processing Units clusters, for both training and inference workloads, possibly involving multiple agents. It will overview challenges and existing approaches in transformer chain performance modeling and control, request routing to distributed agents, hyperparamenter sizing, showcasing some preliminary experimental results from an academic distributed LLM testbed.