Serverless computing is becoming an increasingly representative cloud computing model. Despite its benefits in ease of programming, auto-scaling, and fine-grained billing, it suffers from challenges in guaranteeing SLOs, resource utilization, and cold start latency. The increasing function sizes from recent innovations in large ML models further deteriorate these challenges. We present PPFaaS, a serverless framework centered on the idea of using dynamic pipelining to support efficient autoscaling. This framework replaces the replication-based autoscaling with pipelining to optimize SLOs, tail latency, and resource consumption. Through a diverse set of applications and workloads, PPFaaS simultaneously outperforms the current state-of-the-art system in multiple dimensions: on average, it improves SLO attainment by 29%, reduces 98th percentile tail latency by 73%, and decreases memory and GPU consumption by 22% and 17%, respectively.