Spandana: Reconciling Strict SLOs with Low Cost under Fine-Grained Load Fluctuations

Cloud-based online services must contend with a load that exhibits significant fluctuations at sub-second granularity. Despite such load volatility, services must meet strict Service Level Objectives (SLOs), which forces a tradeoff between over-provisioning resources in order to meet SLO and achieving high resource utilization and cost efficiency. None of the existing approaches, which include reactive and proactive autoscalers, serverless (FaaS), and hybrid clusters combining virtual machines (VMs) and serverless workers, are able to reconcile this tension under fine-grained load fluctuation. We introduce Spandana, an architecture that resolves this tension by decoupling SLO enforcement from cost optimization through a two-level control plane. A light-weight controller colocated with each application VM (the local level) steers each arriving request to the VM if the expected service time falls within the SLO target; otherwise, the request is forwarded to a stock FaaS layer (e.g., AWS Lambda). Doing so achieves high VM utilization and cost-efficiency while ensuring strict SLO compliance through the use of elastic serverless instances when needed. At the global level, Spandana's resource allocator determines the most-efficient VM provisioning considering both VM and serverless costs as well as traffic volatility. Our evaluation shows that Spandana exhibits extremely strict SLO adherence, consistently achieves CPU utilization in the range of 76-86%, and reduces costs by 5-44% over three SOTA baselines.