En4S: Enabling SLOs in Serverless Storage Systems

Serverless computing promises scalability and cost-efficiency by decomposing monolithic tasks into small, stateless, self-contained functions. As functions only reserve hardware resources during their lifetime, and serverless providers such as Amazon Lambda define strict data size limits, data required for the whole lifetime of a monolithic task needs to be kept in an external ephemeral data store. This approach increases costs and introduces performance variability, causing serverless applications to violate service level objectives (SLOs). Traditional cloud storage solutions, such as AWS S3 and Redis, fail to provide low-cost and the enforcement of SLOs, while prior works on disaggregated data stores do not scale sufficiently due to: (1) increased scheduling costs when supporting many SLOs; (2) performance degradation in the presence of burst allowances and worsened interference with lenient ones; and (3) failed service differentiation with increased number of SLO. These challenges make SLO enforcement in serverless environments difficult, leading to unpredictable performance and costs that undermine the benefits of serverless computing.

We introduce En4S, a high-performance, flash-based storage system designed for data-intensive serverless applications. En4S employs a profile-based scheduling framework with adaptive strategies to efficiently scale to many tenants with different SLOs. Key features include dynamic tenant handling, adaptive burst control, token reclaim control, and various optimizations to minimize scheduling costs while maintaining superior performance. By re-enabling SLO enforcement for disaggregated flash storage in cloud-native environments, En4S is crucial for modern serverless applications. Our implementation on Amazon EC2 and Lambda demonstrates substantial performance and cost improvements while reliably ensuring SLO compliance, enhancing the viability of serverless storage systems.