Distributed key-value (KV) systems are a critical part of the infrastructure at many large sites such as Amazon, Facebook, Google, and Twitter. Unfortunately, the ecosystem of these KV systems is a mess---no one existing system meets the needs of all applications. Systems designers worry about running multiple stores from different codebases, vendors, and so on, each optimized for certain application requirements and hardware configuration. We argue that having systems designers worry about running multiple stores from different codebases, vendors, and so on, each optimized for certain application requirements and hardware configuration, is unreasonable and unnecessary.

This dissertation proposes a key-value architecture using a generalization of chain-based replication which can be easily configured to support many points along the KV design continuum. First, we present a new replication protocol, Ouroboros, which extends chain-based replication to allow node additions to any part of the replica chain, minimize blocking during node additions and deletions, and guarantee provably strong data consistency. We use Ouroboros in the implementation of a distributed key-value storage system, FAWN-KV, designed with the goal of supporting the three key properties of fault tolerance, high performance, and generality. Second, we present a generalization of chain-based replication to effectively support a wide range of application requirements using four simple knobs: (a) replica type; (b) replication factor; (c) update mechanism between replicas; and (d) query node selection. We describe Flex-KV, that extends Ouroboros with this generalization. Flex-KV can support DRAM, Flash, and disk-based storage; can act as an unreliable cache or a durable store; and can offer strong or weak data consistency. The value of such a system goes beyond ease-of-use: While exploring these dimensions of durability, consistency, and availability, we find new choices for system designs, such as a cache-consistent memcached, that offer some applications a better balance of performance and cost than was previously available.

Thesis Committee:
David G. Andersen, Chair
Garth A. Gibson
Srinivasan Seshan
Timothy Roscoe, ETH Zurich