SUMMARY

John MacCormick, Nick Murphy, Marc Najork, Chandramohan A. Thekkath, and Lidong Zhou
Boxwood: abstractions as the foundation for storage infrastructure.
In *Proc. of the Symp. on Operating System Design and Implementation (OSDI'04)*, 2004.

DATE: 01-FEB-2010

The publication describes a scalable fault-tolerant system called Boxwood that is intended to be a lower level 'basement' for a storage subsystem. The key point of this implementation is exporting higher-level interfaces from the storage subsystem as compared to traditional raw block interfaces. The system was used as a base for the distributed fault-tolerant NFS filesystem. The choice of exported interfaces is therefore determined by the filesystem requirements, in the presented system they are B-tree and *chunk store*. B-tree service is a concurrent distributed implementation using the *chunk store* as underlying storage, the distributed lock manager to coordinate the nodes' access to shared pages and maintains a write-ahead log to ensure recovery after a failure. Fault-tolerant properties are achieved using a replicated storage*,* node *failure detectors* and the Paxos algorithm to maintain consistent knowledge about which nodes are active. Storage replication is done using *replicated logical devices* which are implemented with two servers replicated synchronously. Two replicas agree on their primary-secondary roles using Paxos, the primary is in charge of receiving client requests and sending updates to the secondary. In the case of a node failure, a global state change is consistently passed using Paxos and the alive node continues in degraded mode, keeping a log of changes pending for the other node. Authors claim that implementing fault-tolerance in lower-level block layer made the implementation of upper-level chunk storage much simpler.

Given the distributed B-tree and chunk storage services, a prototype of NFS file server is described along with some performance results. Performance of different components of Boxwood was measured, including the raw replicated logical device, chunk storage, the B-tree service. Almost linear scalability was shown for raw-level block access and non-contended B-tree access, whereas contended B-tree access scalability was worse than linear.

SUMMARIZED BY: Alexey Karyakin