Isotope: A Case for Block I/O Level Tx Isolation

- Traditionally I/O stacks are too heavy
  - Filesystem and above have rich functionalities
  - Block I/O and below are simple/non-intelligent
  - Time for new abstractions!
- Tx isolation is general
  - Multicore and cloud make concurrency a norm
- Tx isolation is difficult
  - Having one high quality implementation helps
  - Tx isolation in block layer enables easy and clean software design
  - Lowest common software layer that can directly support higher layers
  - New abstractions enable policy/mechanism separation
- Tx isolation can be implemented with negligible overhead
  - Optimistic concurrency control using advanced CPUs and abundant memory

Isotope Design

1) Multi-version Index

2) Timestamp Counter

3) Tx Context

4) Write Buffer

5) Tx Decision Engine

Implementation

- Block I/O layer kernel module (device mapper)
  - Similar to LVM and software RAID
  - Can run on any block devices (Disk, SSD, etc.)
- Based on Gecko, a chain logging storage
- Log-structured design with chaining block devices
- Garbage collection is isolated from writes
- SSD and memory cache included

Applications

- IsoBT and IsoHT
  - Persistent B-tree and hashtable based key-value stores
  - Uses LevelDB APIs
- IsoFS
  - Transactional file system on FUSE
  - PleaseCache() for metadata
- ImgStore using IsoBT and IsoHT
  - IsoBT for metadata and IsoHT for data
- ReleaseTx/TakeoverTx to continue transactions
- Three compositions to handle transactions across, libraries, threads, and processes
- Easy to build transactional applications with Isotope APIs
  - 1K LoC for IsoFS and 150 LoC for ImgStore

Isotope APIs

- BeginTx()
  - Creates a Tx context
  - Treats I/Os before EndTx as a Tx
  - Every write handled in memory
- AbortTx()
  - Terminates a Tx
  - Marks subblock accesses
  - Enables in-memory read
  - E.g. for filesystem metadata
- EndTx()
  - Checks transaction conflicts
  - Persists updates on success
  - Aborts on failure
  - Returns success/failure

Isotope: Transactional Isolation for Block Storage

Ji-Yong Shin1, Mahesh Balakrishnan2, Tudor Marian3, and Hakim Weatherspoon1

1Cornell University, 2Yale University, 3Google

Work funded by NSF and DARPA