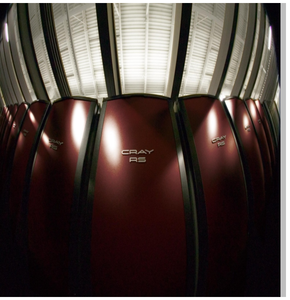


Sandia National Laboratories

Exceptional service in the national interest

Sirocco – An Overview

Matthew L. Curry, Ph.D.

Scalable System Software
Sandia National Laboratories
Albuquerque, NM, USA
mlcurry@sandia.gov

IODC 2016

June 23, 2016

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

What is Sirocco?

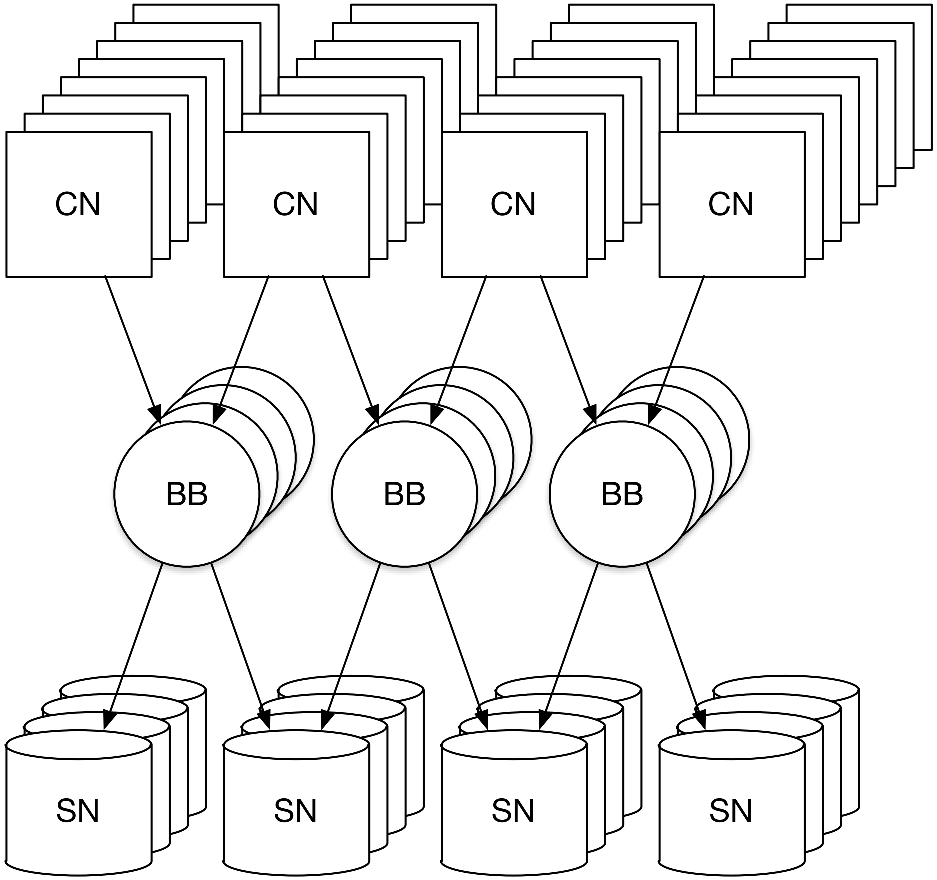
- A low-level distributed object store for large HPC installations
- Not a file system! Think RADOS, not files
  - Lightweight philosophy – Bring your own services
  - Naming, consistency management
- File system/storage API crafted against Sirocco
  - POSIX, HDF5, S3…
- A new entry in the field of non-deterministic storage systems
  - E.g., Zest

How is Sirocco unique?

- Targeted for massive write workloads
  - Checkpoints – All write entire memory ASAP
  - Some can be allowed to fail
  - Often, nothing can happen until checkpoint completes
- Clients choose best locations to write data
  - Local optimization of write performance
  - Write to closest, least burdened servers known
  - Cost: No way to “look up” location of data
  - Benefit: Unreasonably fast checkpoints
- System manages safety/space by moving objects
  - No client notification

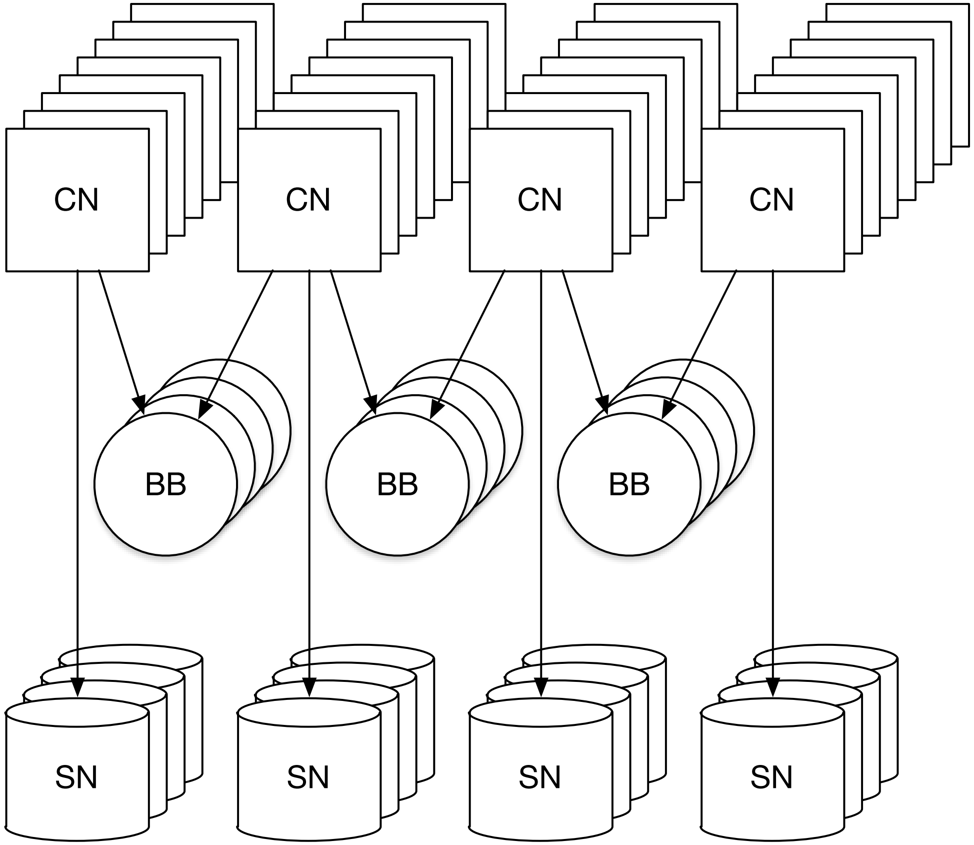
Current state of the art

- Burst buffer is a layer of fast cache between compute nodes and file system
- Burst buffer is sized to accommodate *n* checkpoints, which are asynchronously bled to slow store
- In some models, all I/O is performed through burst buffers



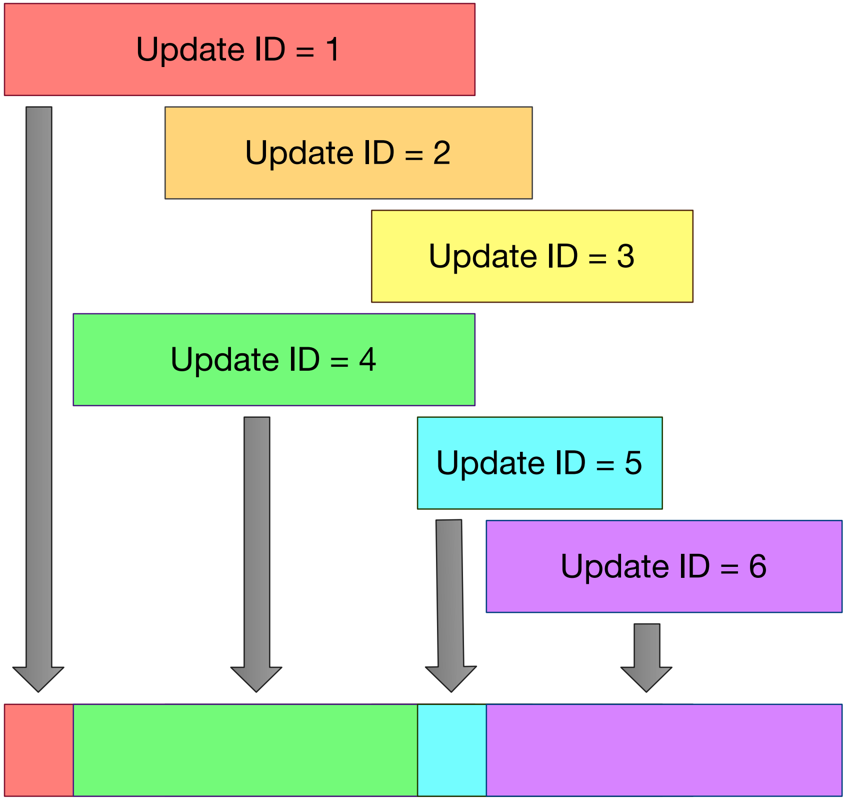
Sirocco model

- Burst buffers are a file-system-level storage resource
  - May be used if beneficial
- Write wherever there is capacity/bandwidth
- System reshuffles data as required
- Because data can move, possible to temporarily introduce resources
  - Compute nodes w/ RAM for storage



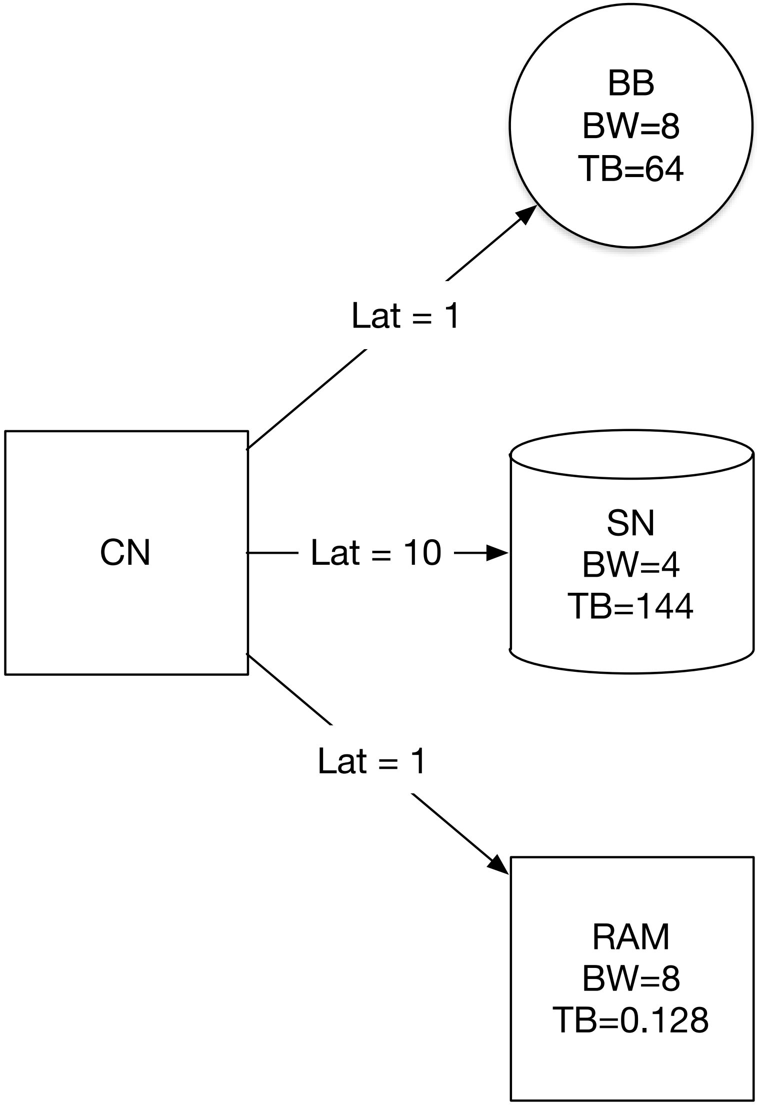
How do clients write?

- Step 1: Determine an *update ID*
  - Update IDs are per-extent logical clocks specified at write time
  - Used to order conflicting writes
  - Trivial for file-per-process or non-overlapping m-to-n checkpointing, both common in NNSA applications
  - More complex write patterns will want to employ a service to manage



How do clients write?

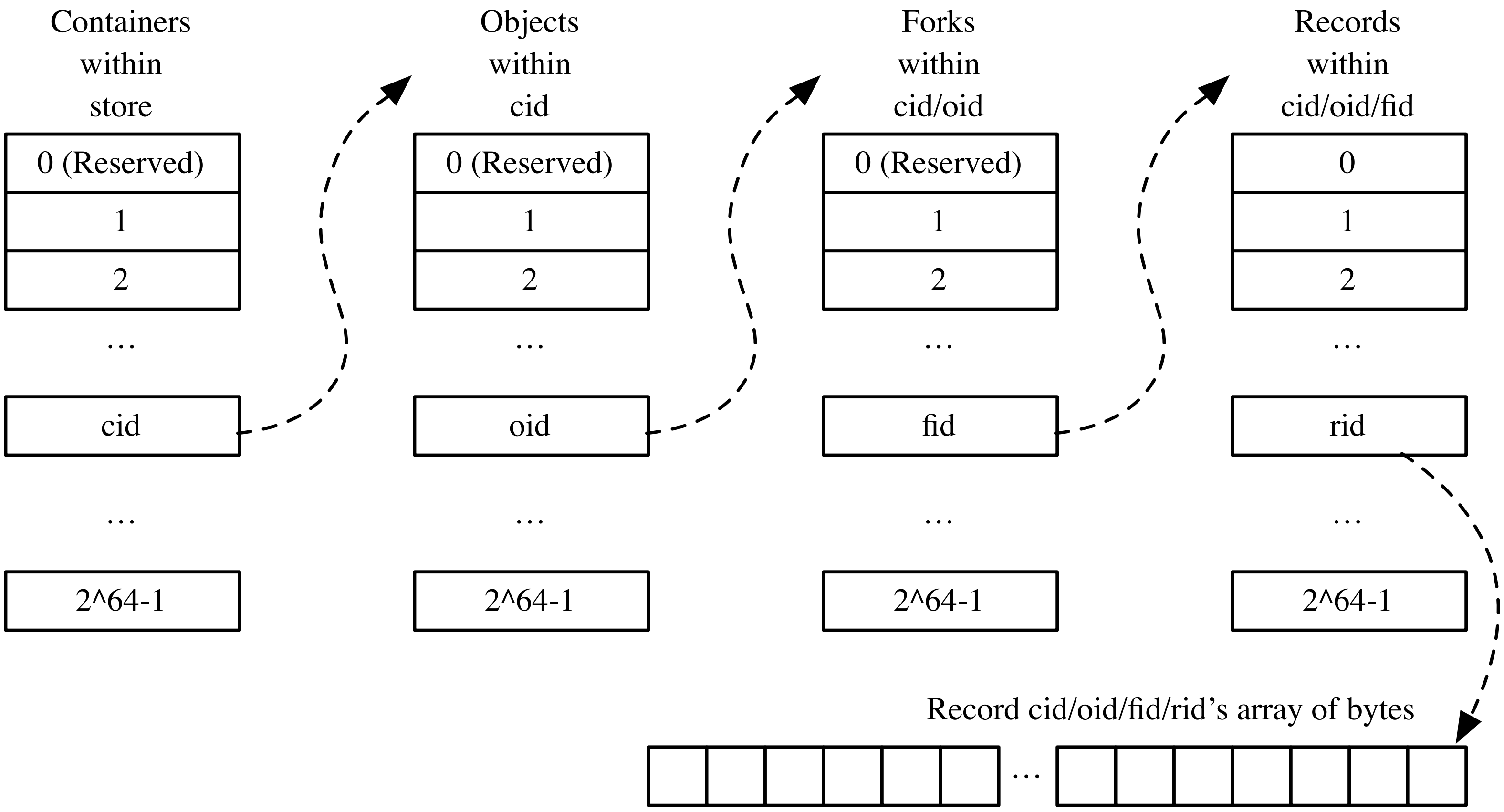
- Step 2: Determine a target
- Select a “fit” target from a local cache
  - Use a membership protocol (e.g., SWIM, SAP2P, etc.) to learn
  - Servers piggyback health/weather back to clients
- Client select target(s) fit for purpose
  - Latency-sensitive workloads use IOPS
  - Bandwidth-sensitive workloads can use bandwidth
  - Both may prefer low latency



How do clients write?

- Steps 3-5: (Don’t worry, they’re easy)
- Set resilience attribute at target
  - Currently number of copies on stable storage
  - Nothing preventing future abstract value
- Write
- Send sync as appropriate to ensure chosen resilience
  - Sirocco takes responsibility

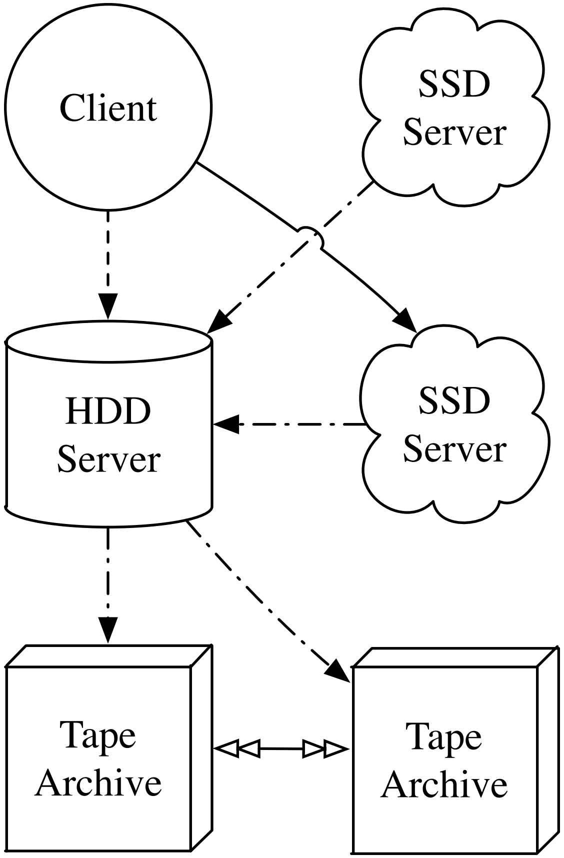
Sirocco Data Model



Karakoyunlu et al., "Toward a Unified Object Storage Foundation for Scalable Storage Systems." Proceedings of the 5th Workshop on Interfaces and Architectures for Scientific Data Storage (IASDS 2013).

How does Sirocco manage data?

- Data moves according to system need, without central management
  - Replication – For resilience
  - Migration – For load balancing, hierarchical storage management
  - Eviction – For cold data with replicas or data with limited lifetime



How do clients read?

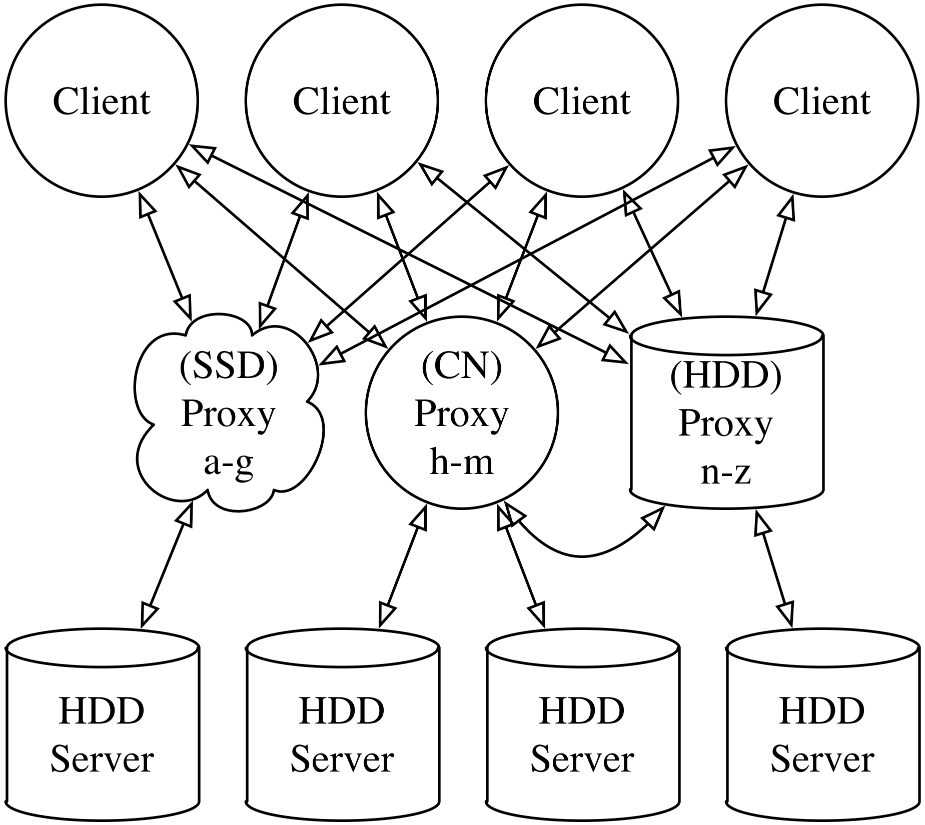
There is no coherent, central index for data.

- Generally: Search.
  - Sun et al. "A Lightweight Data Location Service for Nondeterministic Exascale Storage Systems." ACM TOS, July 2014
  - Sun et al. "A Hierarchical, Triangle-Based Data Location Service for Nondeterministic Exascale Storage Systems." ACM TOS, submitted.
- Not a big concern for checkpoints
  - Checkpoint-restart data is not often read immediately
  - Possible to stage data asynchronously

How do clients read?

There is no coherent, central index for data.

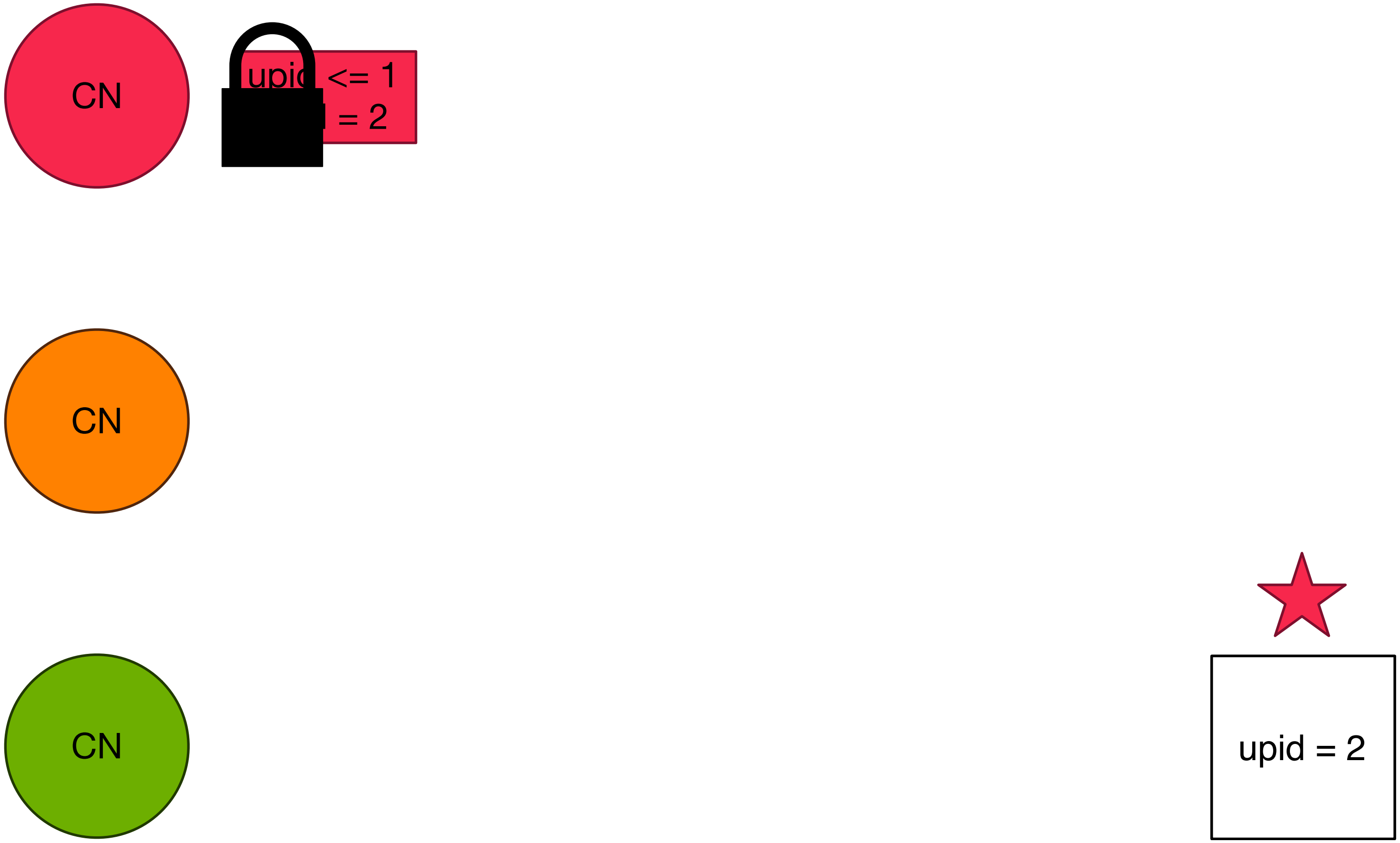
- Clients can cooperatively choose proxies
- Designated storage servers used to manage objects
  - Not necessarily store them
- May be good for metadata
- Future: pNFS-style maps?



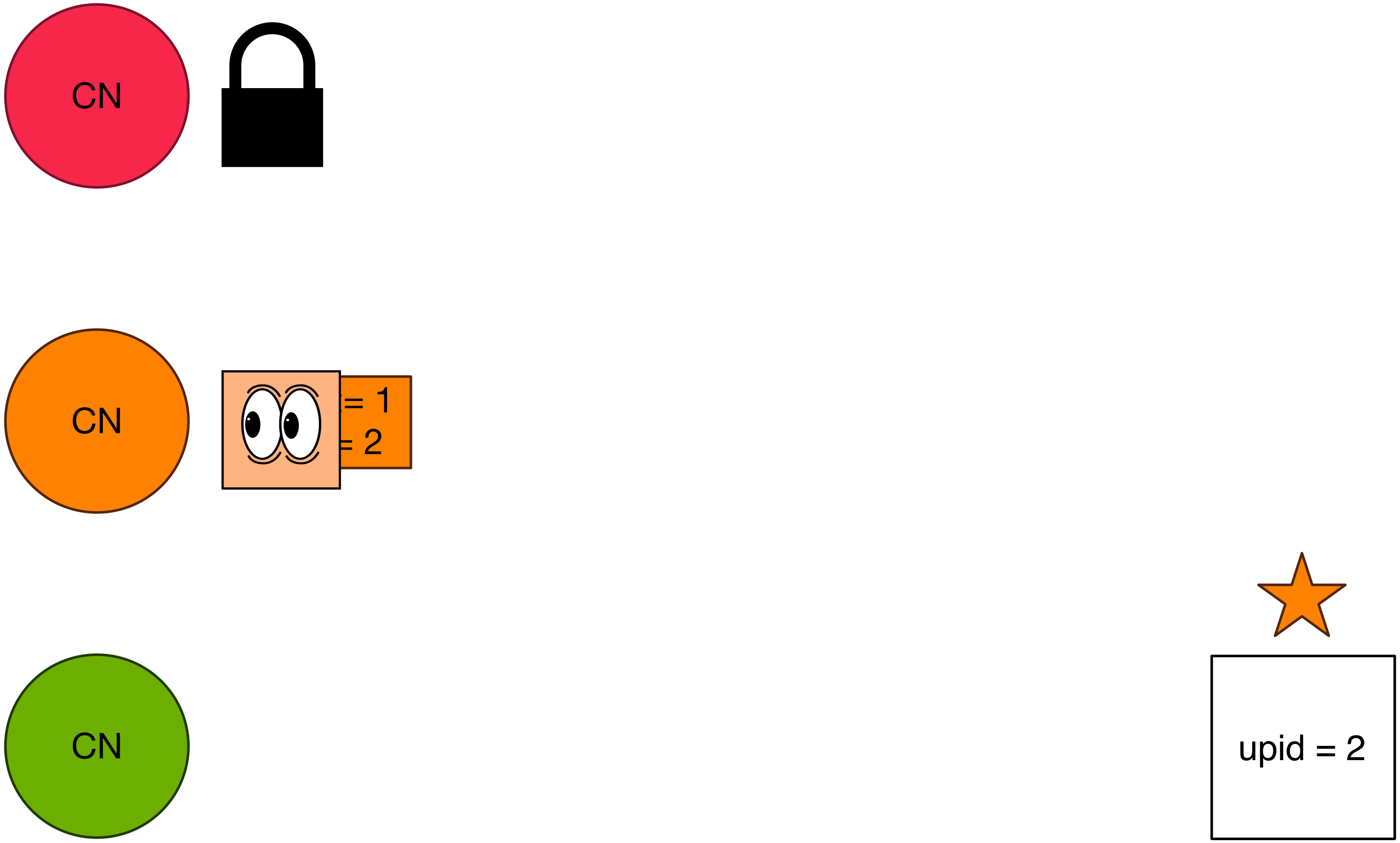
Concurrency Control

- Two widely used methods
  - Pessimistic (i.e., locking)
  - Optimistic (i.e., rollback on conflict)
- Optimistic is provided based on knowledge of update ID.
  - See Karakoyunlu et al. for more information.
- Pessimistic can be provided by external locking library, or by leveraging triggered batches.
  - General; interesting experimentation/optimization for different locking use cases
- Single records on particular storage servers are used to manage synchronization
  - No migration of these records
  - Tendency is to use RAM nodes for these

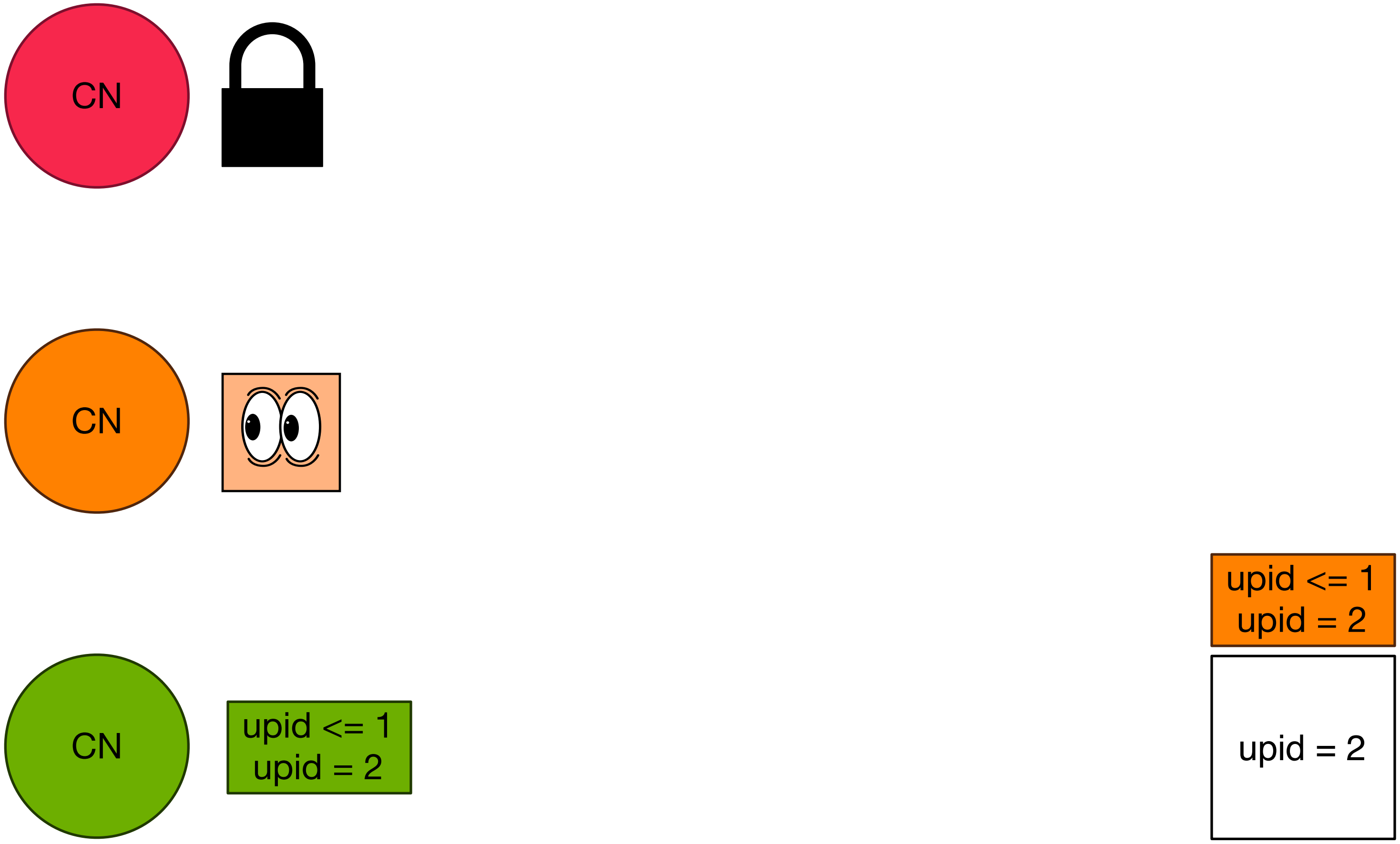
An Example of Triggered Locking



An Example of Triggered Locking



An Example of Triggered Locking



CN

CN

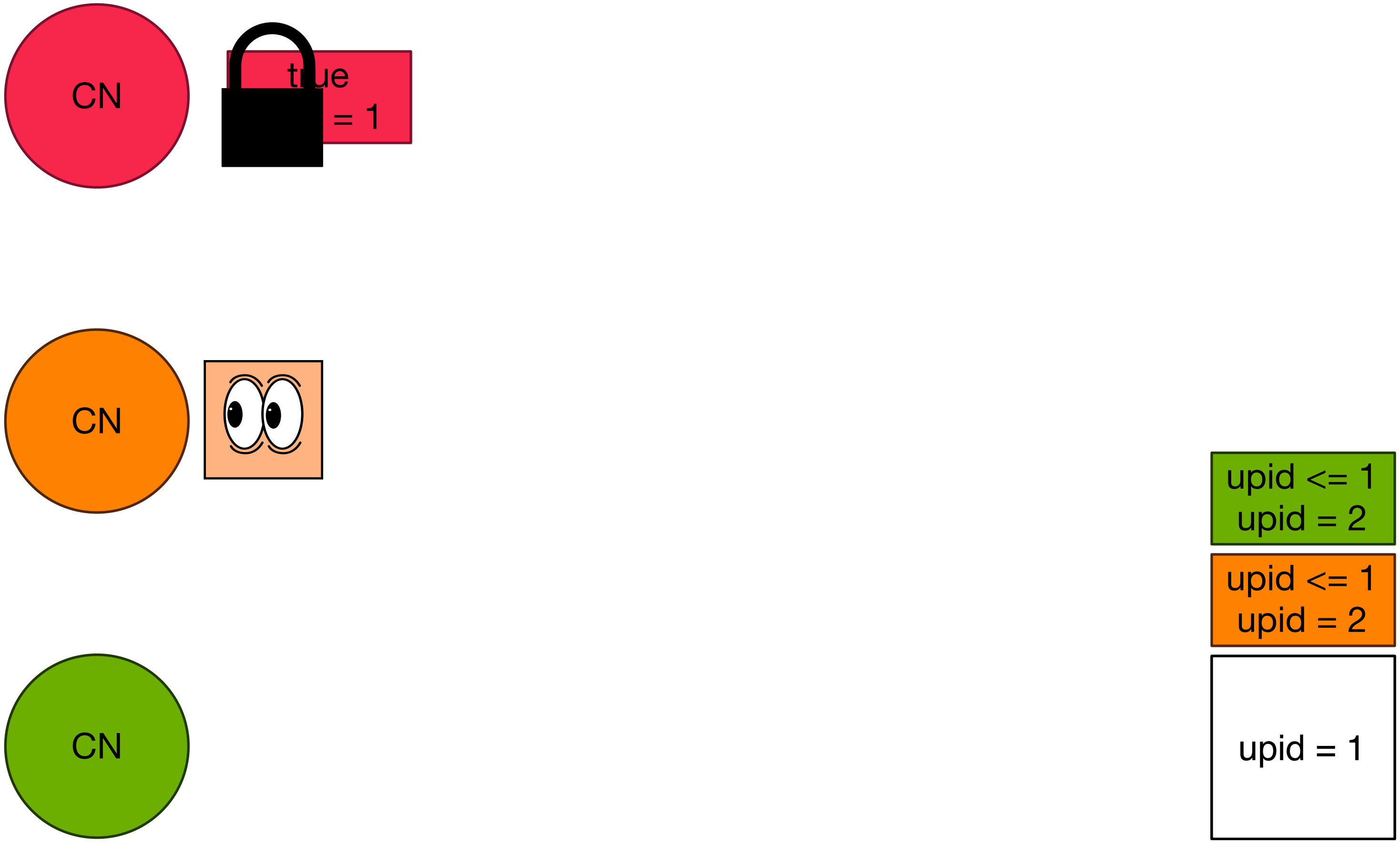
CN

upid <= 1
upid = 2

upid <= 1
upid = 2

upid = 2

An Example of Triggered Locking



CN

true
= 1

CN

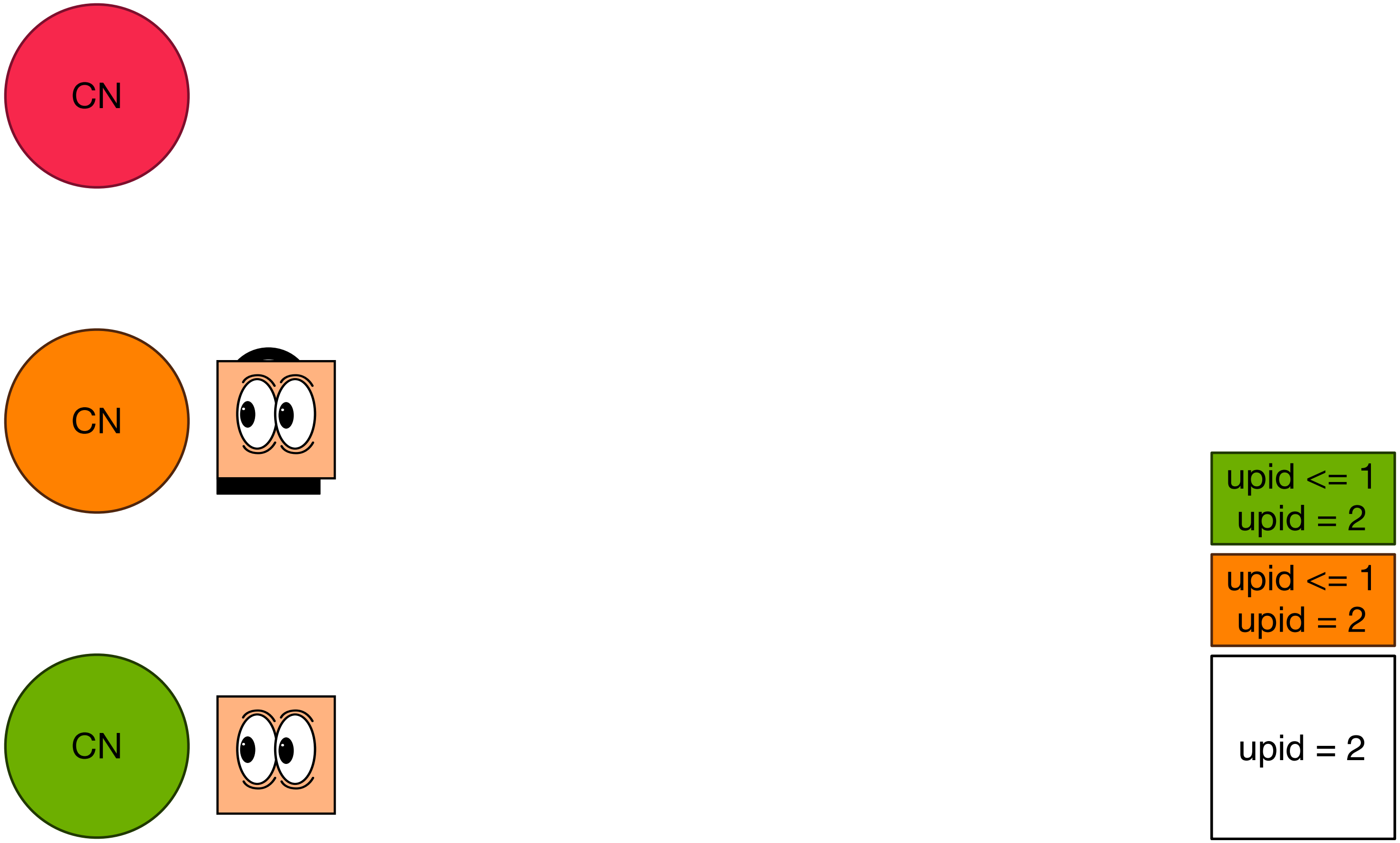
CN

upid <= 1
upid = 2

upid <= 1
upid = 2

upid = 1

An Example of Triggered Locking



CN

CN

CN

upid <= 1
upid = 2

upid <= 1
upid = 2

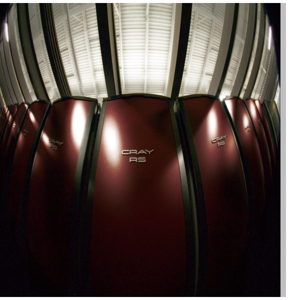
upid = 2

Conclusion

- Sirocco represents a brand new strategy
  - Make writes go as fast as possible under changing system conditions
  - Down with oppressive central control!
- A new object store to host a range of file systems
  - Server- or Client-funded services
- An opportunity to try new things
  - Proxy-based optimizations
  - Variable-length records for rich metadata
  - Storage-based locking primitives

Acknowledgements

- Lee Ward
- Geoff Danielson
- Jay Lofstead



Sirocco – An Overview

Matthew L. Curry, Ph.D.

Scalable System Software
Sandia National Laboratories
Albuquerque, NM, USA
mlcurry@sandia.gov

IODC 2016

June 23, 2016

Sandia National Laboratories

Exceptional service in the national interest

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.