DHRUBA BORTHAKUR, ROCKSET

PRESENTED AT PERCONA-LIVE, APRIL 2017

ROCKSDB CLOUD

ROCKSET

OUTLINE

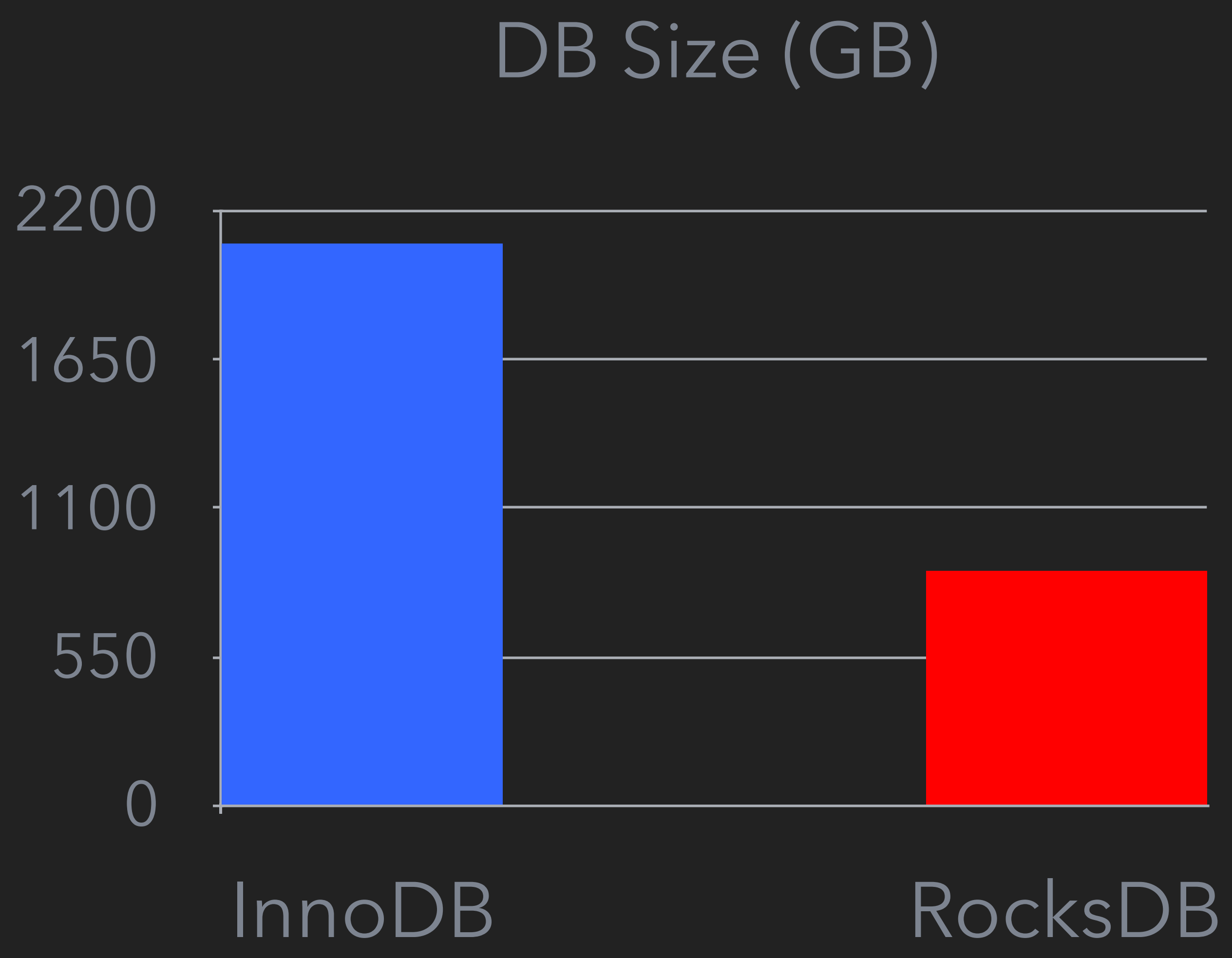
- Why RocksDB-Cloud?
- Differences from RocksDB
- Goals, Design, Architecture
- Next Steps

ROCKSDB STORAGE ENGINE

- Open Sourced from Facebook Engineering
- Log Structured MergeTree
- Embedded c++/java/go library
- Available as MyRocks and MongoRocks
- Used at Microsoft, Yahoo, Netflix,…

ROCKSDB FOCUS ON EFFICIENCY

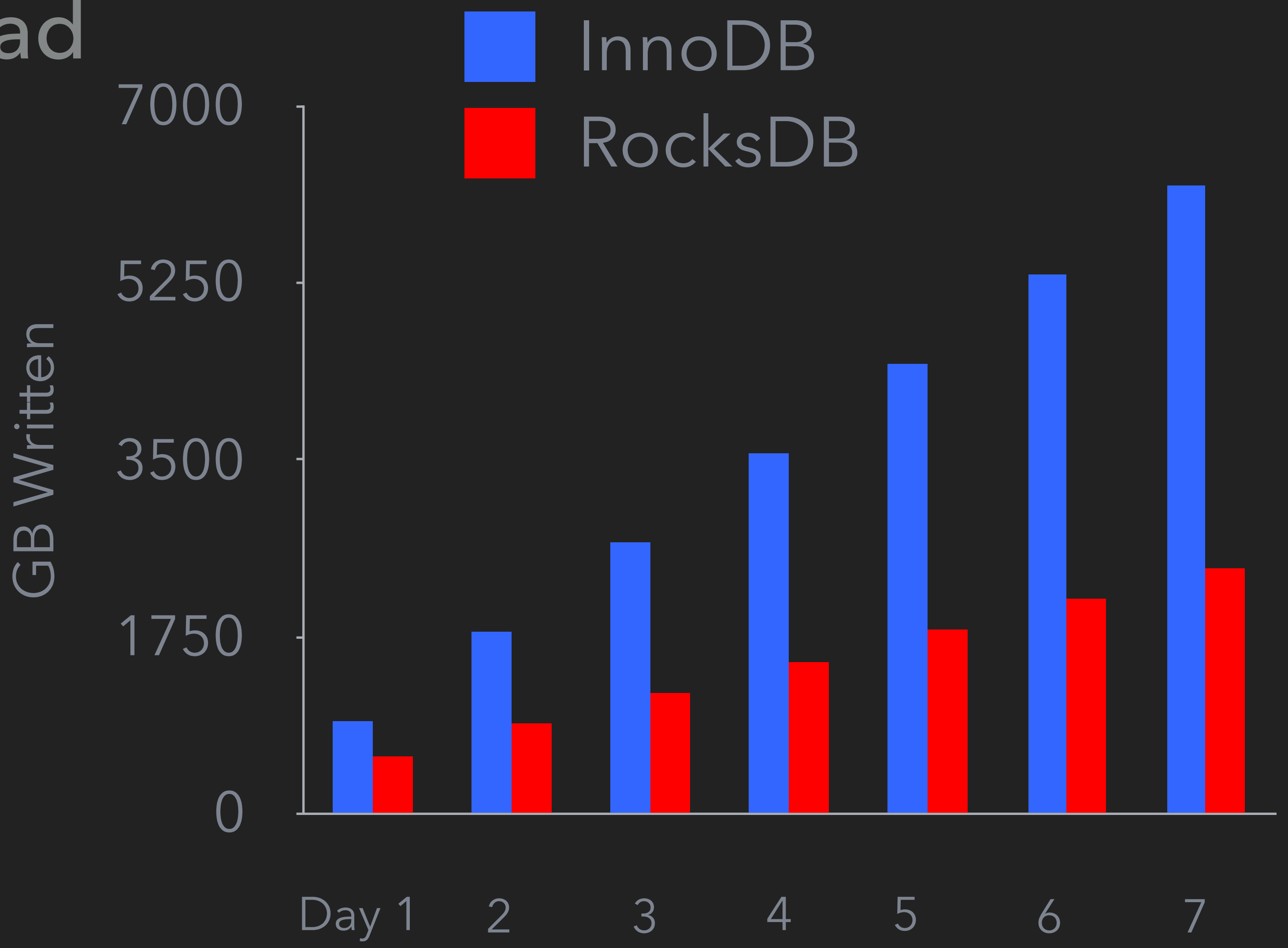
- MyRocks on FB workload
    - 50% smaller



ROCKSET

ROCKSDB FOCUS ON EFFICIENCY

- MyRocks on FB workload
  - 50% smaller
  - 50% lesser IOs



MyRocks is supported by Percona

HAVE YOU USED ROCKSDB FOR CLOUD APPS?

- Rockset started to use RocksDB on AWS
  - build our own replication engine
  - build our own backup system
  - custom code for hot/cold placement
    - RAM, NVMe, SSD , disk

- "Shared Nothing is dead". –– Dewitt @MIT 2017, http://mitdbg.github.io/nedbday/2017/talks/dewitt.pptx

ROCKSDB-CLOUD IS CHEAPER

- RocksDB-Cloud uses locally attached SSD and S3
    - 3x cheaper than 3 way replication
    - If I redesigned HDFS today
        - It won't use 3 way replication
- ***n*** times cheaper than EBS, ***n*** > 1

VISION FOR ROCKSDB-CLOUD

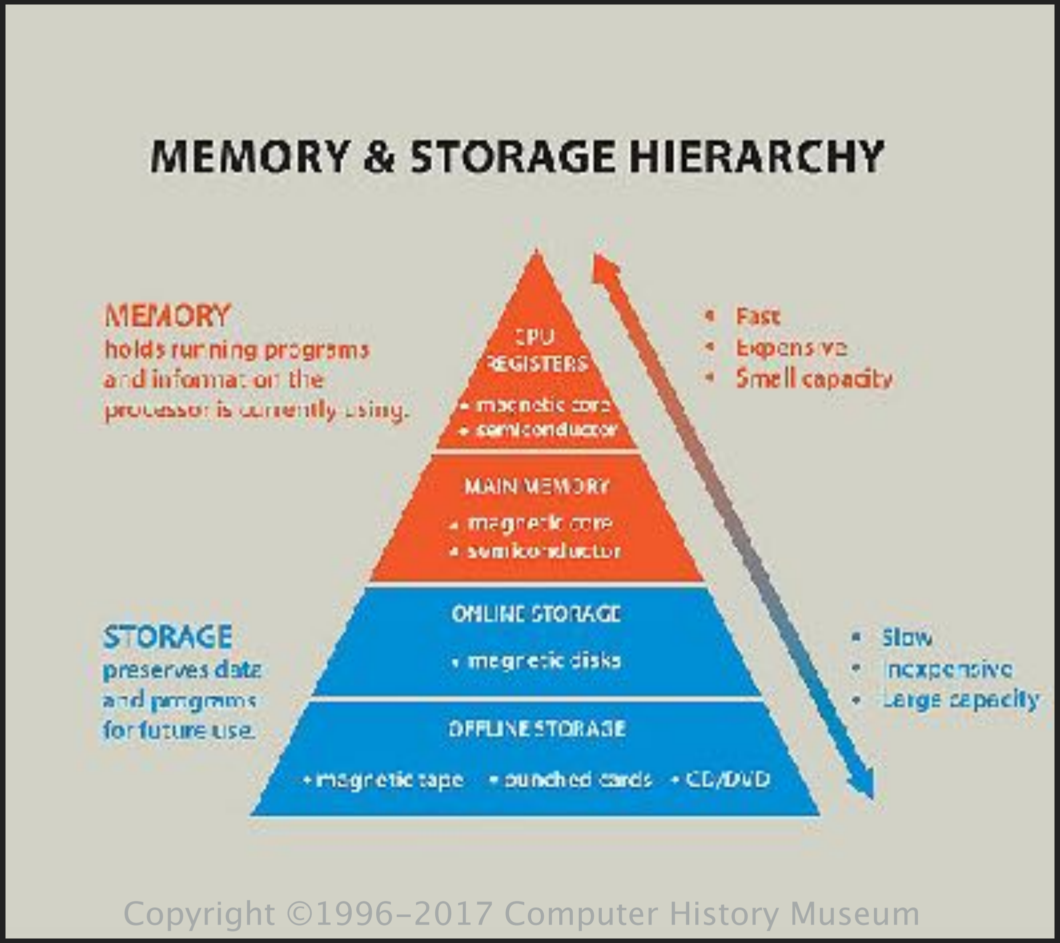
- Optimized for Cloud Applications
- Support AWS, Google Storage, Azure
- Pluggability for other cloud vendors

GOALS

- Durability of data inspite of machine failures
- Replication of data across machines

GOALS

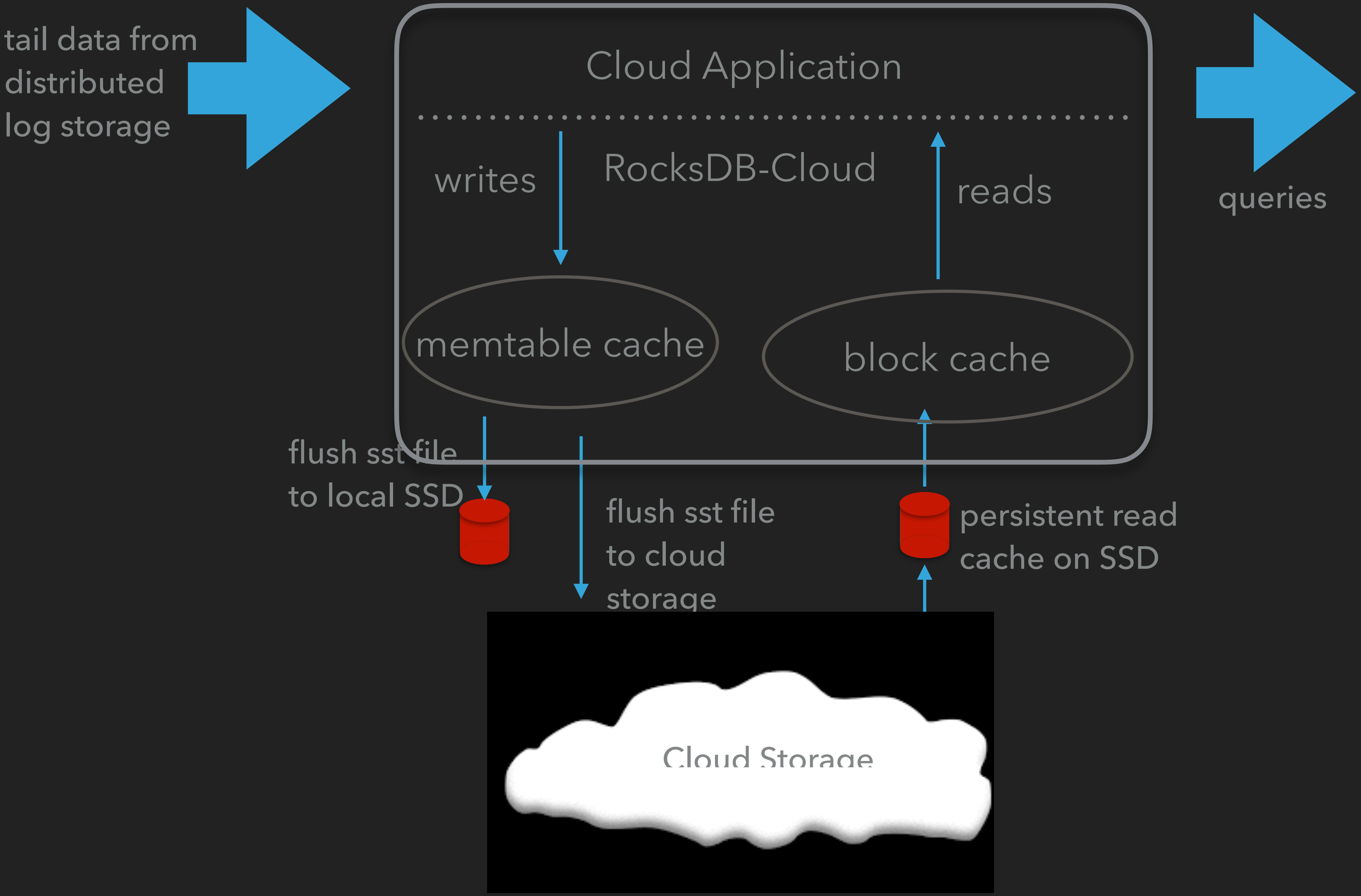
- Durability of data inspite of machine failures
- Replication of data across machines
- Auto placement of hot/cold data on cloud storage hierarchy



GOALS

- Durability of data inspite of machine failures
- Replication of data across machines
- Auto placement of hot/cold data on cloud storage hierarchy
- Portability across cloud vendors

ARCHITECTURE 1: APPLICATION LOG IN FRONT OF DATABASE


tail data from distributed log storage
Cloud Application
RocksDB-Cloud
writes
reads
queries
memtable cache
block cache
flush sst file to local SSD
flush sst file to cloud storage
persistent read cache on SSD
Cloud Storage

ROCKSET

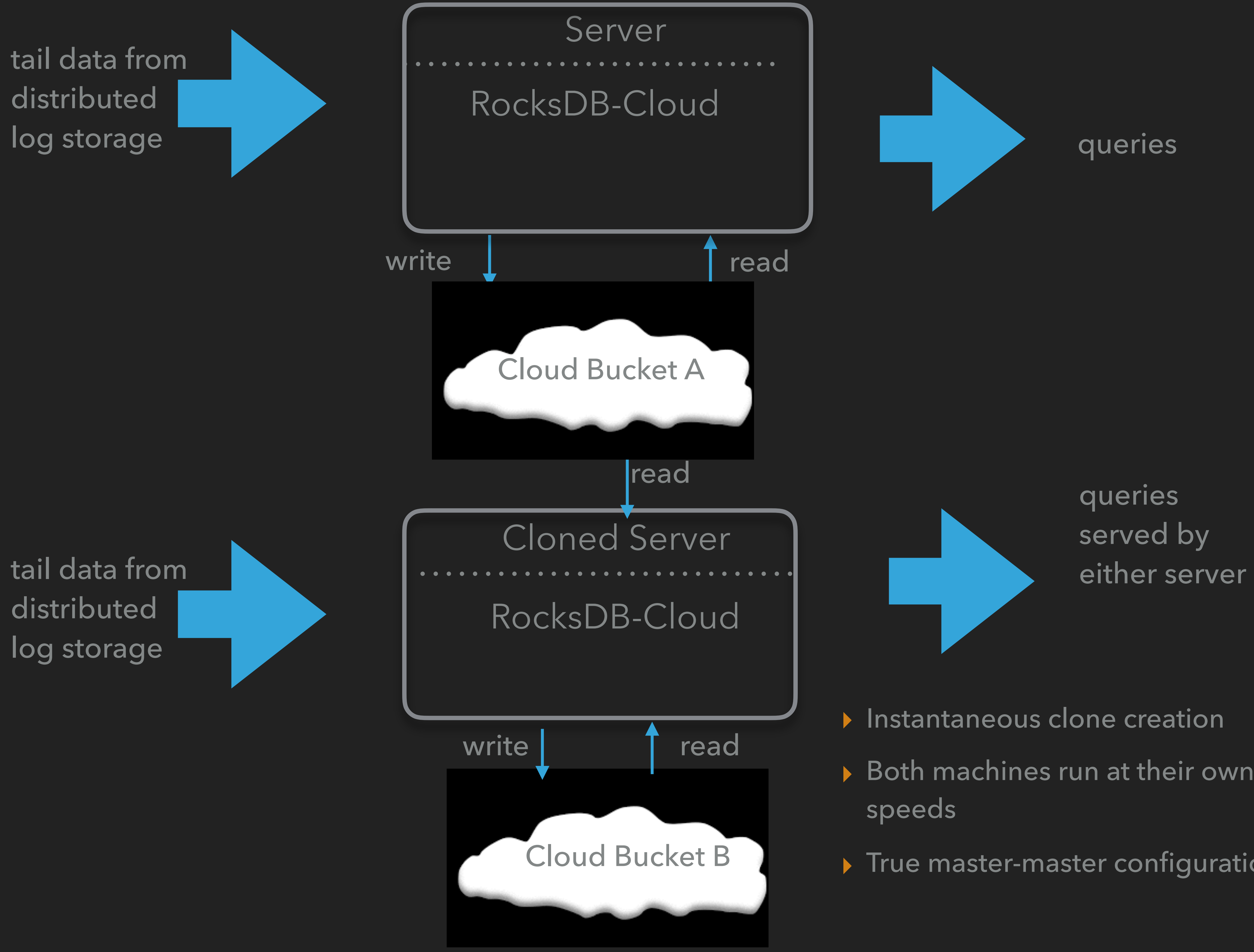
TAILING A LOG

- Tail data from Kafka topic into RocksDB-Cloud
  - Open (local_directory, S3 bucket name)
  - SST file copied to S3 at the time of file close
    - keep or delete local set file
  - Every change to MANIFEST is copied to S3
  - Kafka state stored in RocksDB-Cloud

TAILING A LOG

- Recover data when machine fails
  - Open (local_directory, S3 bucket name)
  - MANIFEST downloaded from S3
  - Download data from sst files on demand
    - Local SSD/disk as persistent cache
  - Restart tailing from Kafka

ARCHITECTURE 1: ZERO COPY CLONES


tail data from distributed log storage
Server
RocksDB-Cloud
queries
write
read
Cloud Bucket A
read
Cloned Server
RocksDB-Cloud
queries served by either server
tail data from distributed log storage
write
read
Cloud Bucket B

- Instantaneous clone creation
- Both machines run at their own speeds
- True master-master configuration

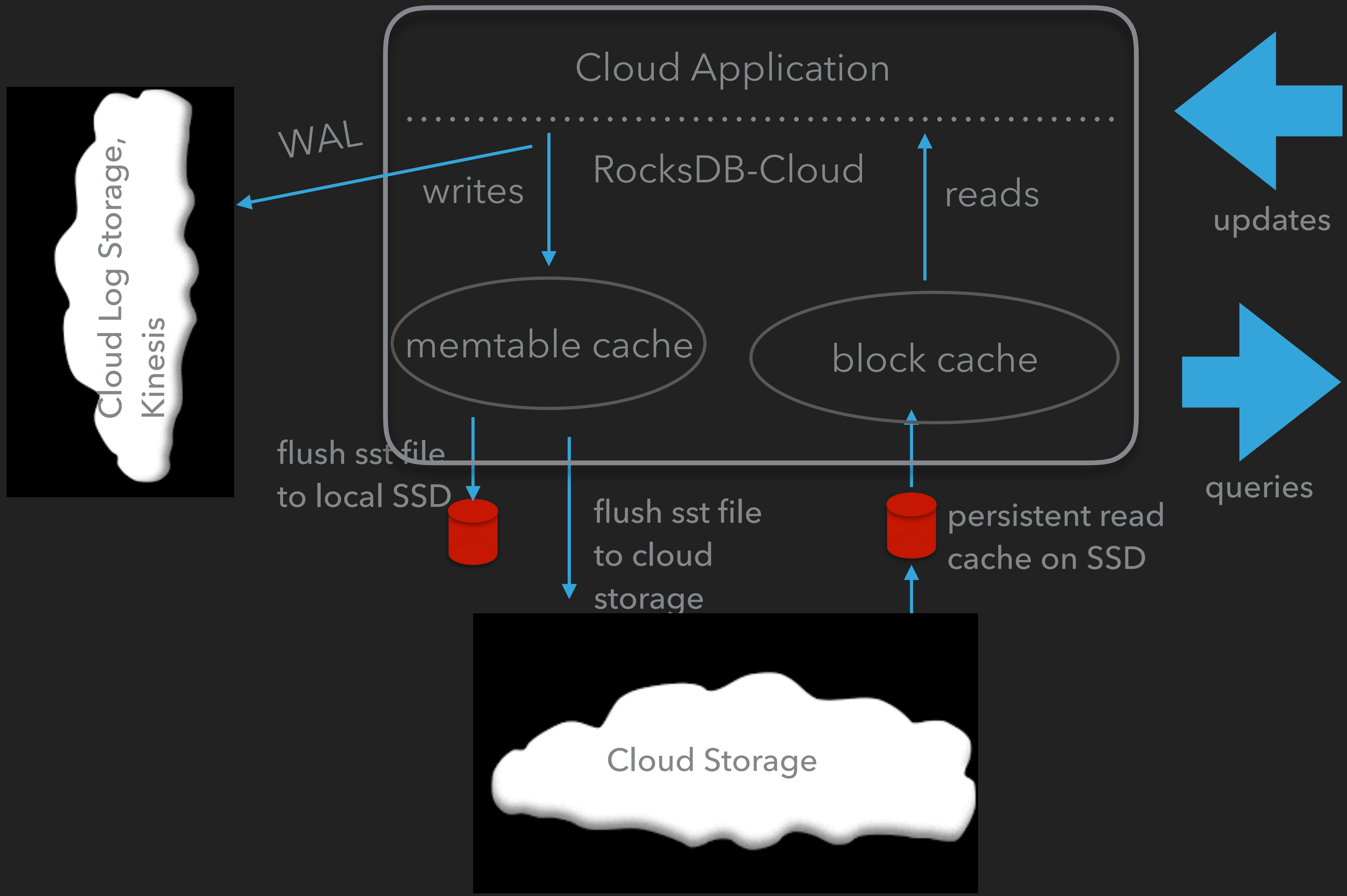
ROCKSET

ZERO COPY CLONES

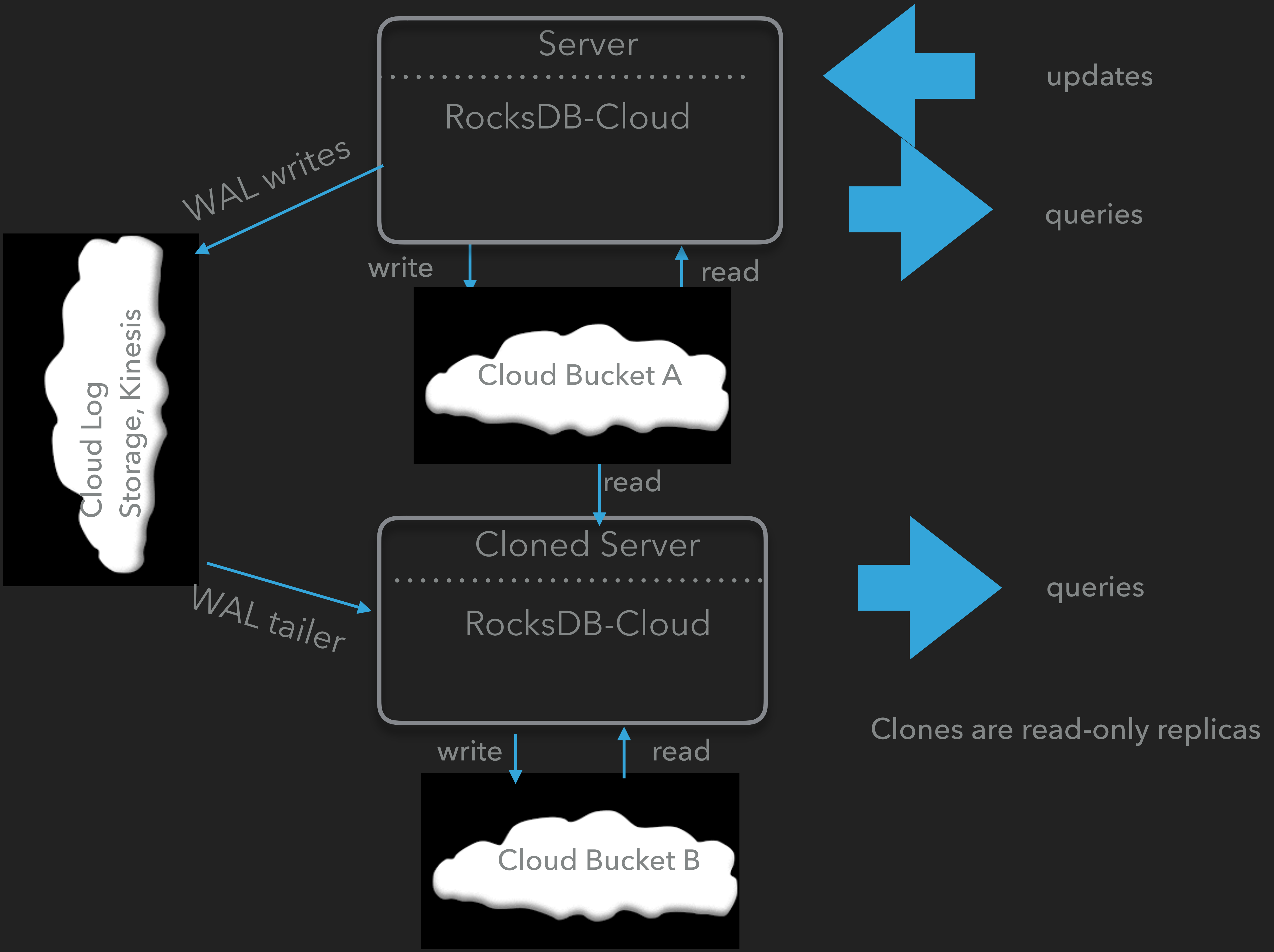
- Purger runs on every Machine
    - Deletes sst files that are not part of any clones
- Both machines run at their own speeds
    - True master-master configuration

ARCHITECTURE 2: DATABASE WAL ENABLED

ROCKSET



ARCHITECTURE 2: ZERO COPY CLONES



Clones are read-only replicas

READ WRITE DB

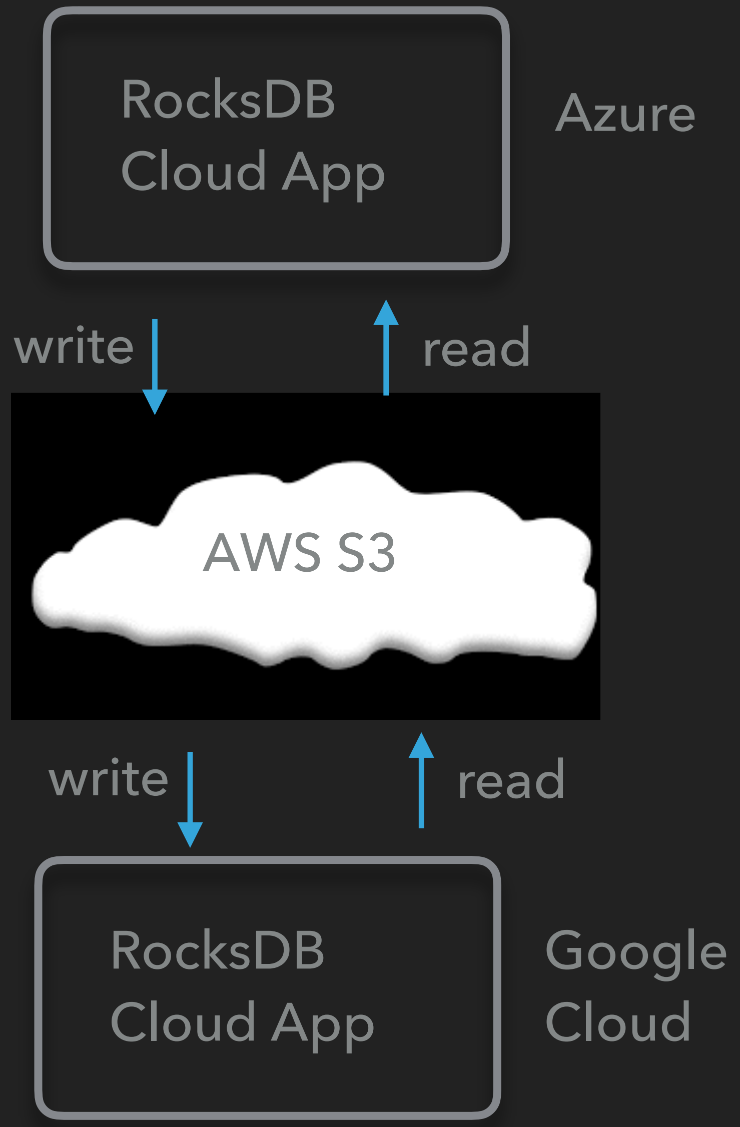
- Master machine:
  - read-write
  - write WAL on AWS-Kinesis
- Slave machine:
  - read-only
  - tail Kinesis and apply

AUTO PLACEMENT OF HOT/COLD DATA

- All Levels L0 - Ln reside in S3
- Levels L0 - L2 typically reside in local SSD and S3
- Cache data from S3 for reads:
  - persistent cache on locally attached SSD
- Support for Intel NVMe

SEAMLESS COPY AMONG S3, AZURE, GOOGLE

- App on Azure can access AWS S3 Storage
- App on Google Cloud can access Azure Storage
- same API on all cloud platforms



COMPATIBILITY

- Pure Open Source
- API compatible with stock RocksDB
- Data format compatible with stock RocksDB
- License compatible with stock RocksDB
  - BSD License

NEXT STEPS

- Support for large size objects
- Support encryption-at-rest

COLLABORATORS

ROCKSET

Microsoft

Pinterest

f

applatix

REFERENCES

- Source code:
  https://github.com/rockset/rocksdb-cloud
- Dev discussions:
  rocksdb-cloud@googlegroups.com
  https://groups.google.com/d/forum/rocksdb-cloud
- Slack Channel:
  #rocksdb-cloud @ https://rockset-io.slack.com