Nan-nan Zhao, Ji-guang Wan, Jun Wang, Chang-sheng Xie, 2016. A reliable power management scheme for consistent hashing based distributed key value storage systems. *Frontiers of Information Technology & Electronic Engineering*, **17**(10):994-1007. http://dx.doi.org/10.1631/FITEE.1601162

A reliable power management scheme for consistent hashing based distributed key value storage systems

Key words: Consistent hash table (CHT), Replication, Power management, Key value storage system, Reliability

Corresponding author: Ji-guang Wan
E-mail: jgwan@mail.hust.edu.cn

ORCID: http://orcid.org/0000-0002-7219-7856

Motivation

Current energy issues with key value storage systems

- **Key value storage systems**
  - Dynamo at Amazon, Cassandra at Facebook, and Voldemort at LinkedIn …
  - Consistent hash table (CHT)
    - High scalability
    - Load balance
    - Simplifying the lookup operations
- **Server energy conservation has become a priority**
  - With the increase in the sheer volume of digital data, storage and server demands are on a rapid increase.
  - Server energy cost constitutes a significant part of a data center's power bill

Traditional replication under consistent hashing

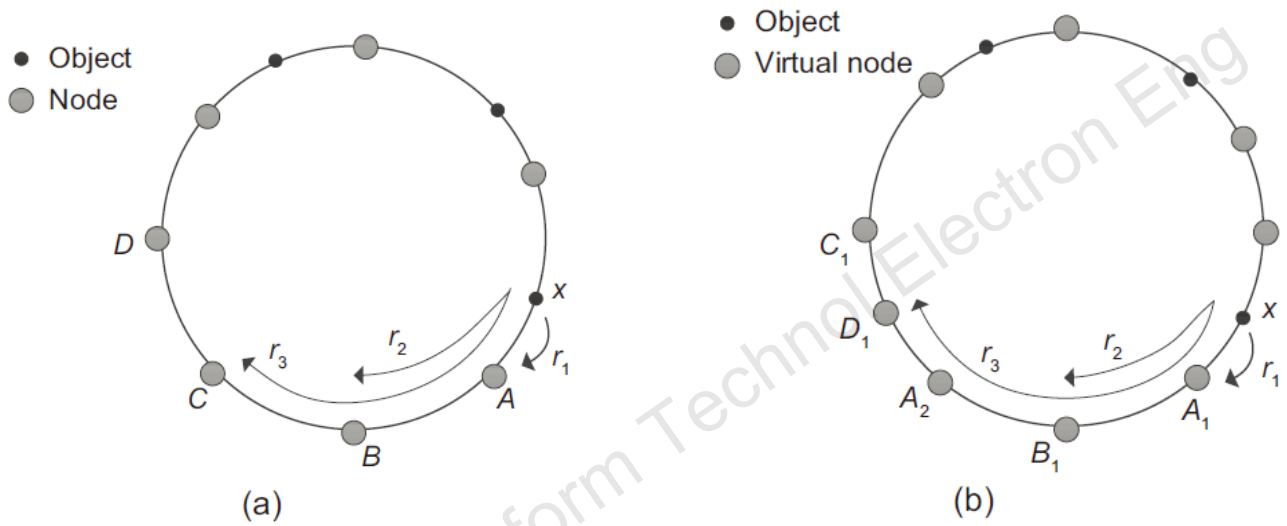

Fig. 1 Traditional replication under consistent hashing without (a) or with (b) virtual nodes

- The traditional replication strategy prevents subsets of nodes from powering down without violating data availability[1].

1 D. Harnik, D. Naor, and I. Segal. Low Power Mode in Cloud Storage Systems.

GreenCHT design

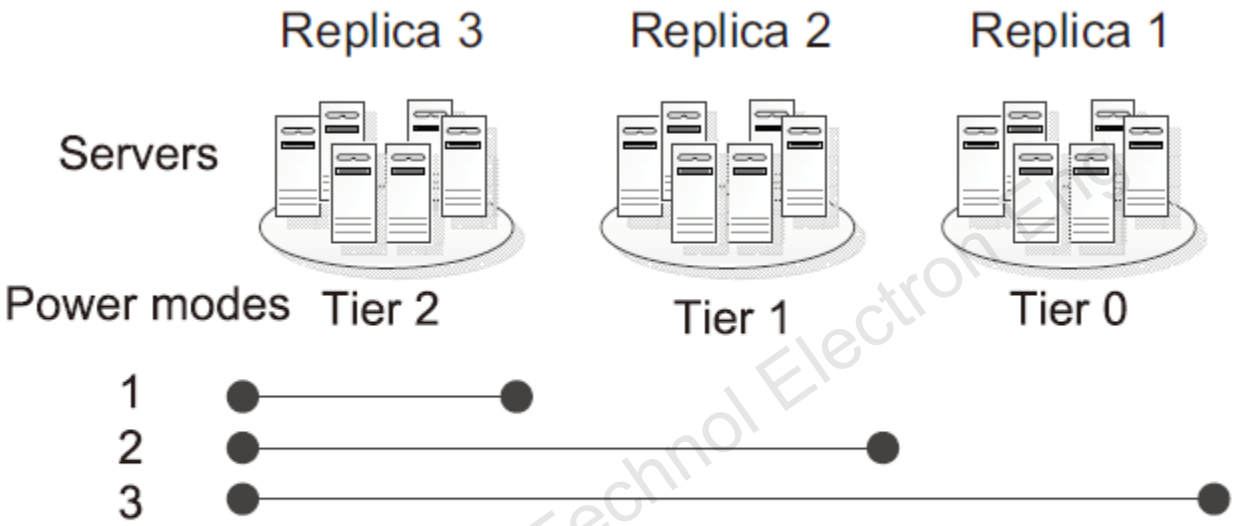

Fig. 2 Tiering and power modes with a replication factor of 3

- **Availability**
  - Allow $\frac{(R-t)N}{R}$ of the nodes to be powered down
- **Different power modes**
  - Sustain different workload levels

Multi-tier replication

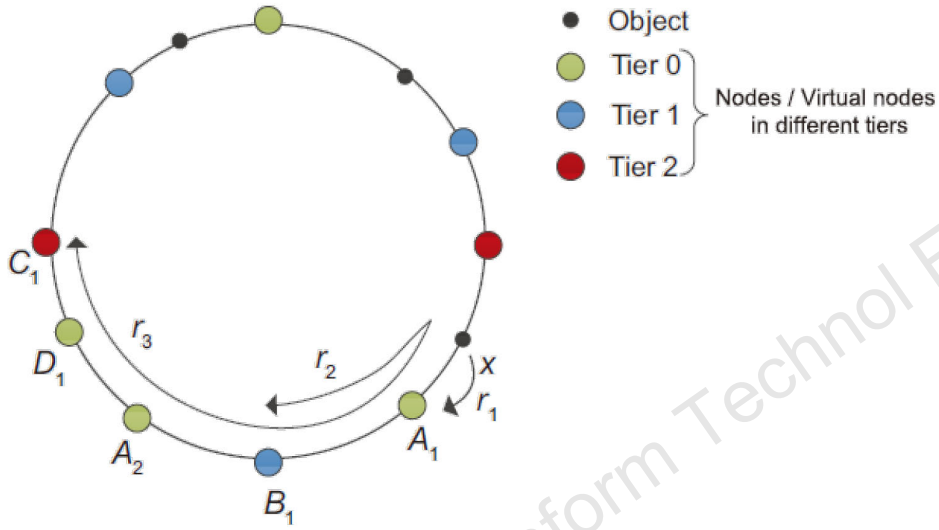

Fig. 3 Multi-tier consistent hash ring and replica layout (M-CHT)

Table 1 Allocation of replicas

Object x	Successor 1
Tier 0	r_1
Tier 1	r_2
Tier 2	r_3

- **Scalability**
  - When server *n* joins or leaves the system, certain objects will be migrated between server *n* and its successor in the same tier.

Log-store

Table 2 Log-replicas allocation

Object x	Successor_1	Successor_2	Successor_3	… …	Successor_R
Tier 0	r_1	-	-	-	-
Tier 1	r_2	Log r_1	-	-	-
Tier 2	r_3	Log r_1	Log r_2	-	-
… …	… …	… …	… …	… …	… …
Tier R-1	r_R	Log r_1	Log r_2	… …	Log r_{R-1}

- **Availability and reliability**
  - All the writes to standby replicas are offloaded to log-store, which exists in active nodes in higher tiers.
- **Parallelism of writes**
  - Replicas and log-replicas are stored in different nodes.
- **Scalability**
  - When a node enters or leaves, certain objects will be migrated between the node and its successor in the same tier. It will not influence other nodes.

Power mode scheduler

- **Track the load**
  - **Hour**
- **Predict the load**
  - **ARMAX model**
- **Choose the power mode**
  - $P = \left\lceil \frac{L_{predict}}{L_{tier}} \right\rceil$

GreenCHT prototype

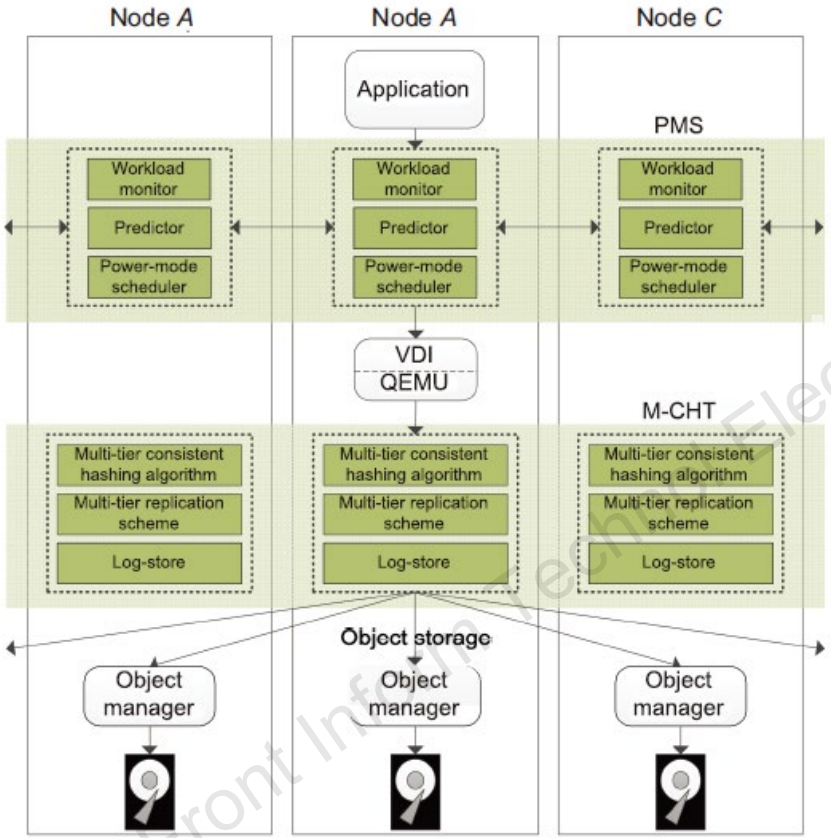

Fig. 6 GreenCHT system architecture

- **We implemented our multi-tier replication scheme on Sheepdog**
  - Modify its original data distribution and replication algorithm.
  - The power mode scheduler runs in the user space to schedule nodes to be powered-down and powered-up.

Evaluation

Power savings

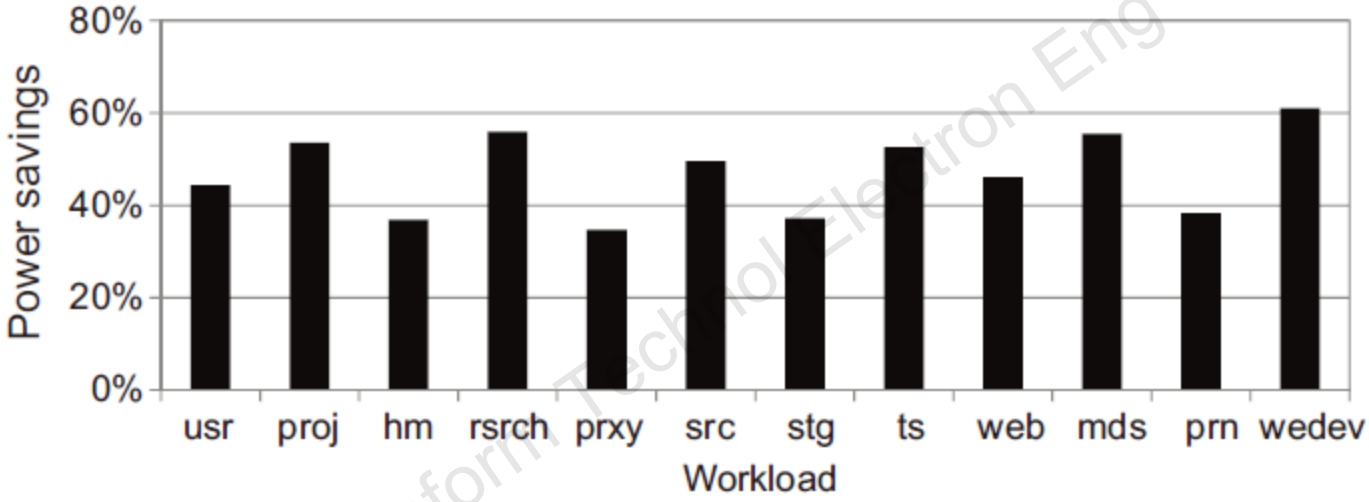

Fig. 12 Power savings for GreenCHT by using power mode scheduler's decisions

- **GreenCHT can reduce power consumption by up to 35%–61%**

Latency

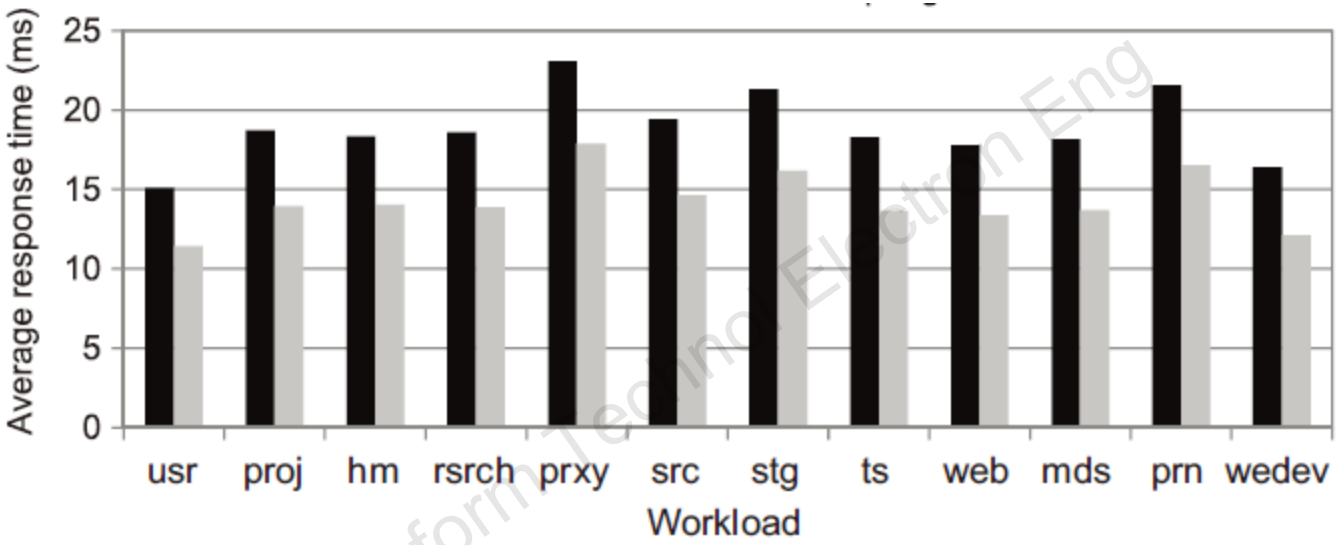

Fig. 14 Average response time for the two schemes under 12 traces

Conclusions

- We proposed a reliable power management scheme, GreenCHT, for consistent hashing based distributed key value storage systems.
- GreenCHT uses multi-tier replication to achieve considerable power savings while ensuring data availability and a reliable distributed log store to ensure reliability as well as fault tolerance of the whole system.
- To provide power proportionality and ensure good performance, GreenCHT dynamically adapts to workload variation under the required performance constraint by using a predictive power mode scheduler (PMS).
- GreenCHT was implemented based on the Sheepdog storage cluster by modifying Sheepdog's original data distribution and replication algorithm. By replaying 12 real-world traces, we observed that GreenCHT can reduce power consumption by up to 35%–61%.