Lambert: Achieve High Durability, Low Cost & Flexibility at Same Time

Open source storage engine for exabyte data in Alibaba

Coly Li <bosong.ly@alibaba-inc.com>
Robin Dong <haodong@alibaba-inc.com>

About Speakers

Coly Li

Member of technical architecture committee, in charge of storage engineering of AIS (Alibaba Infrastructure Service)

Robin Dong

Chief software engineer, cold data storage of AIS (Alibaba Infrastructure Service).

Content

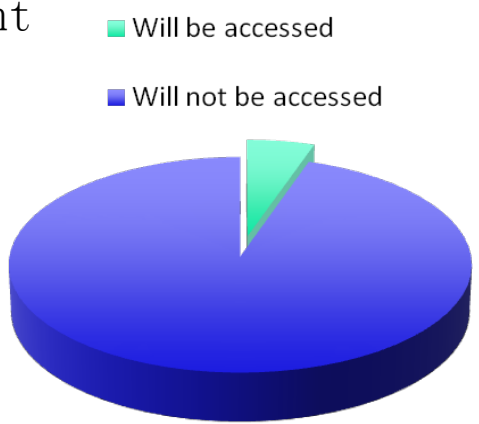
- Cold Data in Real World
- Work Load characteristic and Technical Challenge
- System design and topology
- Motivation of using open source technology
- Improvement & Contribution to sheepdog project
- Credit to Open Source Community

Cold Data in Real World

From an internal storage system, we observe different access pattern unlike existing backup/restore usage.

In last 18 months,

- Less than 5% data is accessed (read/delete)
- Each read access does not touch whole data object
- Only a small range of data is accessed from a single data object



This is typical cold data access pattern, existing backup/restore system is expensive to handle cold data storage work load.

Work Load characteristic

High performance is NOT first priority,

- Internal customer doesn't require high throughput or low latency
- Access latency in hours is acceptable

High Durability is critical,

- Not high availability (failure of network, power supply, mainboard)
- Data is safe on storage media, after failure is recovered

Work Load characteristic (Cont.)

Extremely low cost is required,

- Data is required to be accessible for many years, even whole company life cycle
- Data set increases very fast, especially when online business goes very well (US $9.3 billion in GMV on 11.11 Shopping Festival)
- Extremely low storage cost is critical, in Exabyte cold data storage

Technical Challenge

High Durability & Low Cost in same time,

- Reliable & cheap storage media
- Multiple copies and fast failure recovery

Technical Challenge (Cont.)

Deployment in third-party data centers,

- Power supply, cooling, rack capacity might be variable in different data centers
- Cold storage hardware is required to scale from Petabyte to Exabyte, with different power supplies (e.g. from 4KW to 8KW)

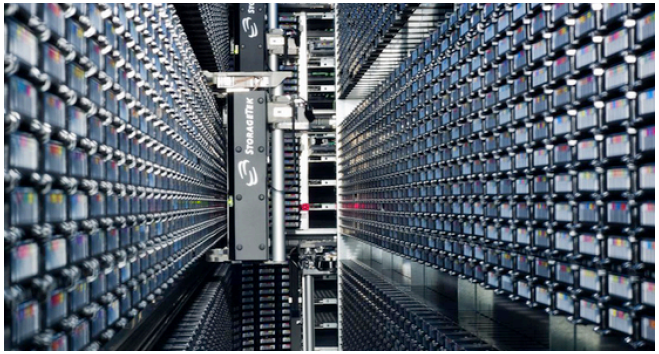
Not every data center (we have) is perfect like this

Technical Challenge (Cont.)

Tape or Blue-Ray disk does not work perfectly in our case **currently**,

- Tape is cheap, but automatic tape library is expensive, if not in large scale deployment
- Is Blue-Ray disk cheaper in long term? We need help from industry to prove it, in our work load.

In our current situation, we prefer mechanical hard disk as storage media for cold data.



tape library with automatic robot
http://www.boston.com/bigpicture/2009/11/large_hadron_collider_ready_to.html

Facebook Blue-Ray storage for cold data
http://www.burnworld.com/blu-ray-the-future-of-data-storage/

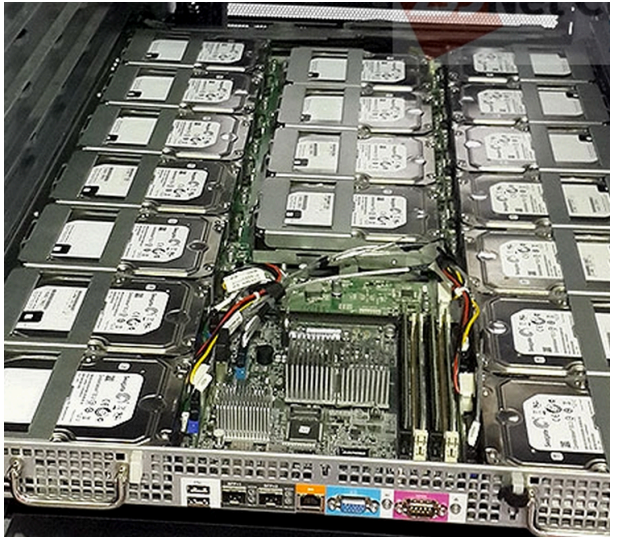
System design and topology

Hardware is designed for low cost & high density storage

- 18 hard disks (3.5 inch) in a single 1U case
- 4T or 8T low performance hard disk
- 32U cases in a single rack
- Low cost & power consume CPU and memory

The hardware design is called Project Scorpio.

Check open data center committee website for more chinese information:
http://www.opendatacenter.cn/data/manual/index.html

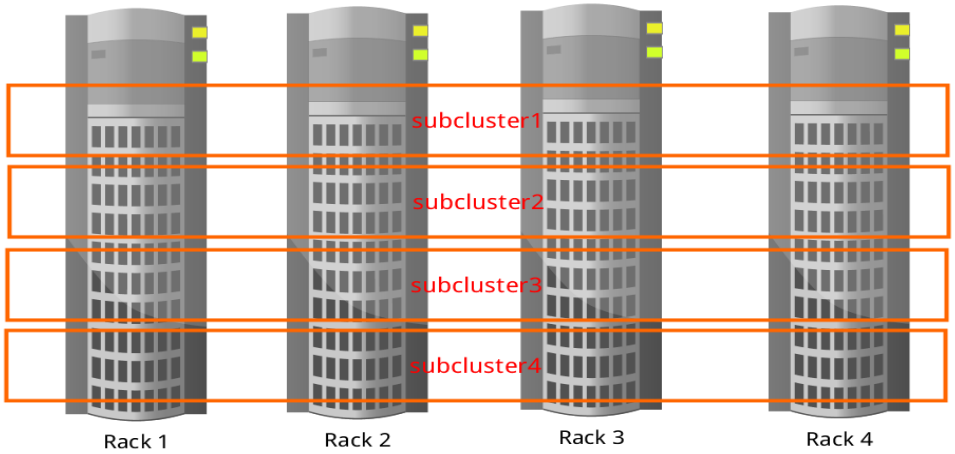


Project Scorpio storage hardware reported by ZDnet (chinese tech news media) ,
http://solution.zdnet.com.cn/2014/0724/3028228.shtml

System design and topology (Cont.)

Deployment unit

- Hardware is extended in deployment unit (e.g. 4 Scorpio rack).
- In each unit, there are several distributed storage clusters.
- The sub-cluster is minimum unit of software defined storage.
- Single data object will be only stored within single specific sub-cluster.
- When a sub-cluster is full, turn it into sealed state.

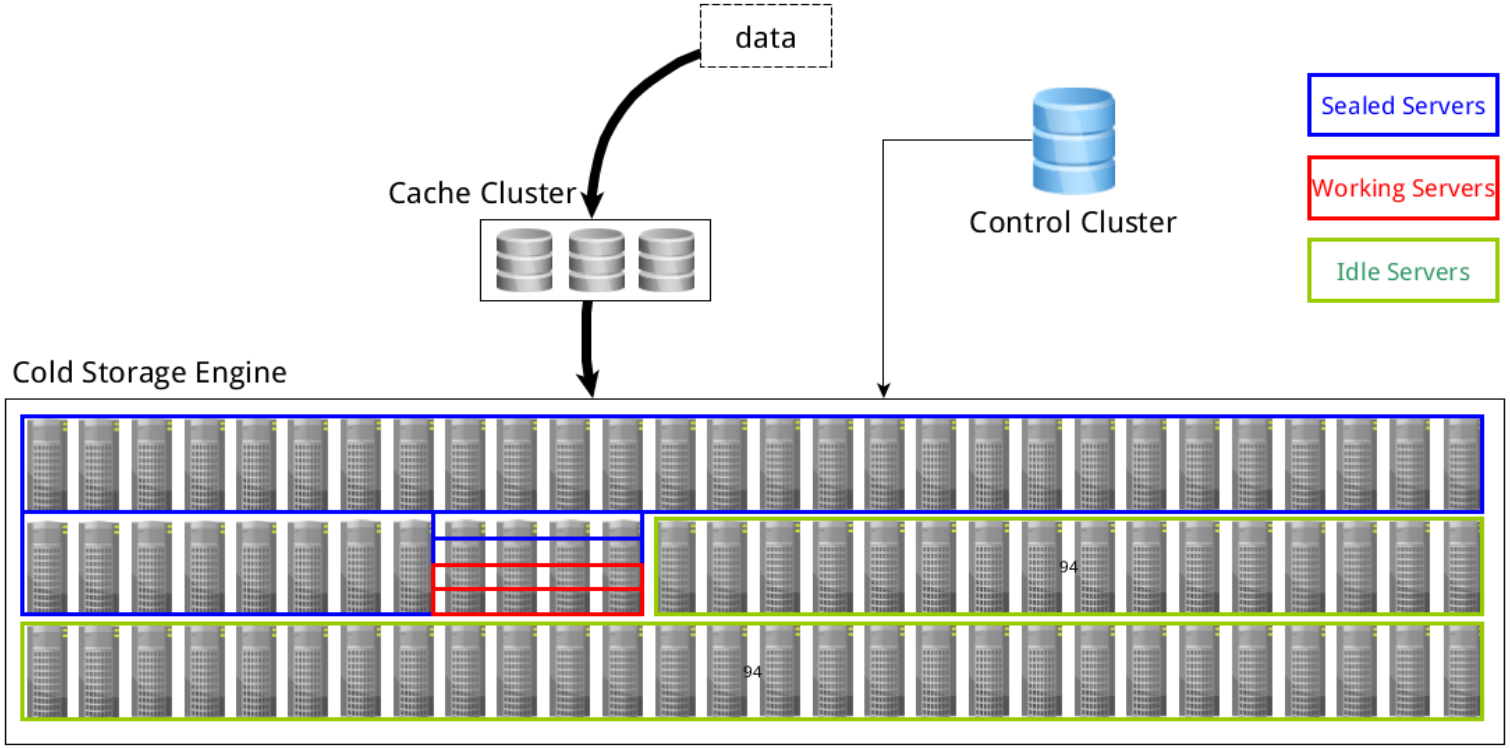


System design and topology (Cont.)

Simple software defined storage

- Sub-cluster is a distributed storage cluster as minimum software storage unit.
- Just adding more sub-clusters (deployment unit) when extend storage capacity.
- No matter how large a cold storage system is, we only encounter scalability issue of a single sub-cluster. It is much easier.
- Small storage cluster means simple, simple means reliable and stable in large scale systems.

System design and topology (Cont.)



Deploy in large scale, only a small group of sub-clusters are in working state

Motivation of using open source technology

For the distributed storage system of sub-cluster, we need,

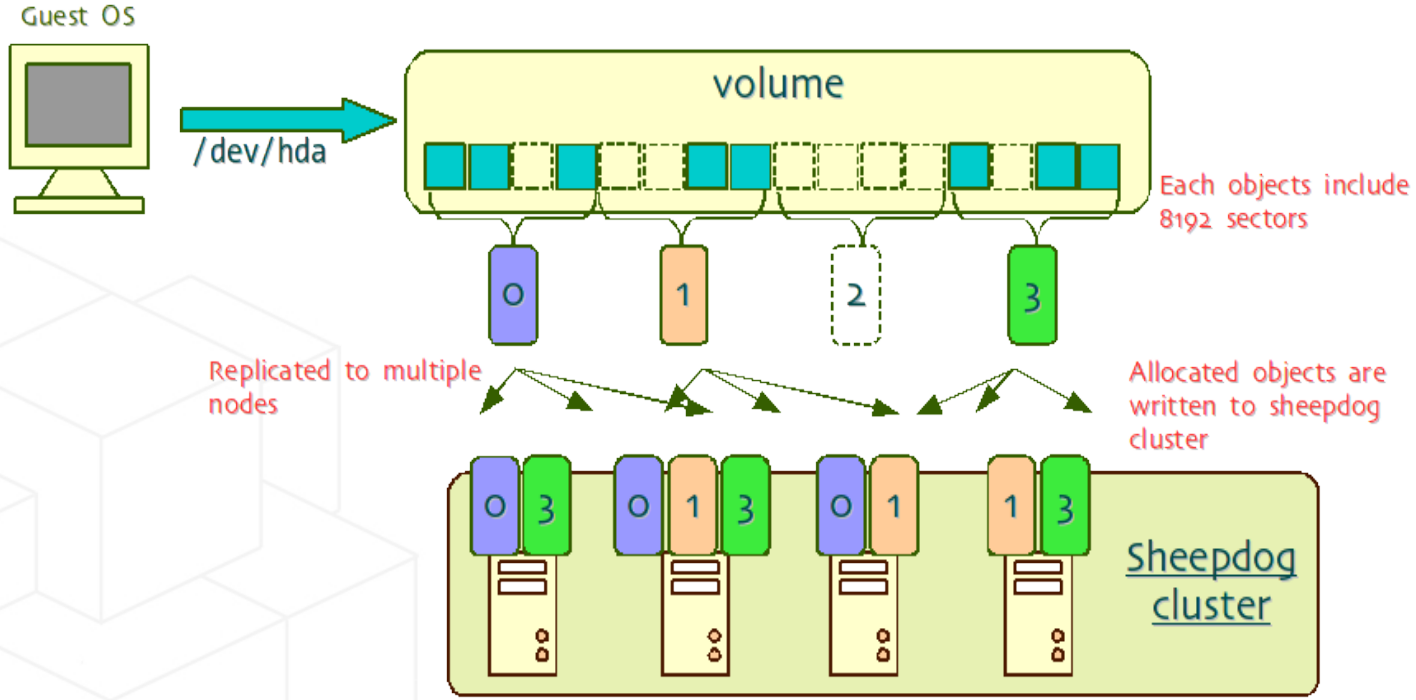
- Simplicity: easy to improve, optimize and maintain
- Consistent Hashing: distributed block storage
- Erasure Code: less data duplication with higher durability
- RESTful API: swift interface for data store, access and control

Other than developing from scratch, it is more efficient to start from a simple open source project as code base for cold data storage engine.

The code name of this cold data storage engine is called: **Lambert**

Motivation of using open source technology (Cont.)

Sheepdog volume, the code base to start.



Improvement & Contribution to sheepdog project

In early 2013, sheepdog project only has consistent hashing storage framework. Since June 2013, Alibaba contributes engineering source to improve sheepdog for cold data storage,

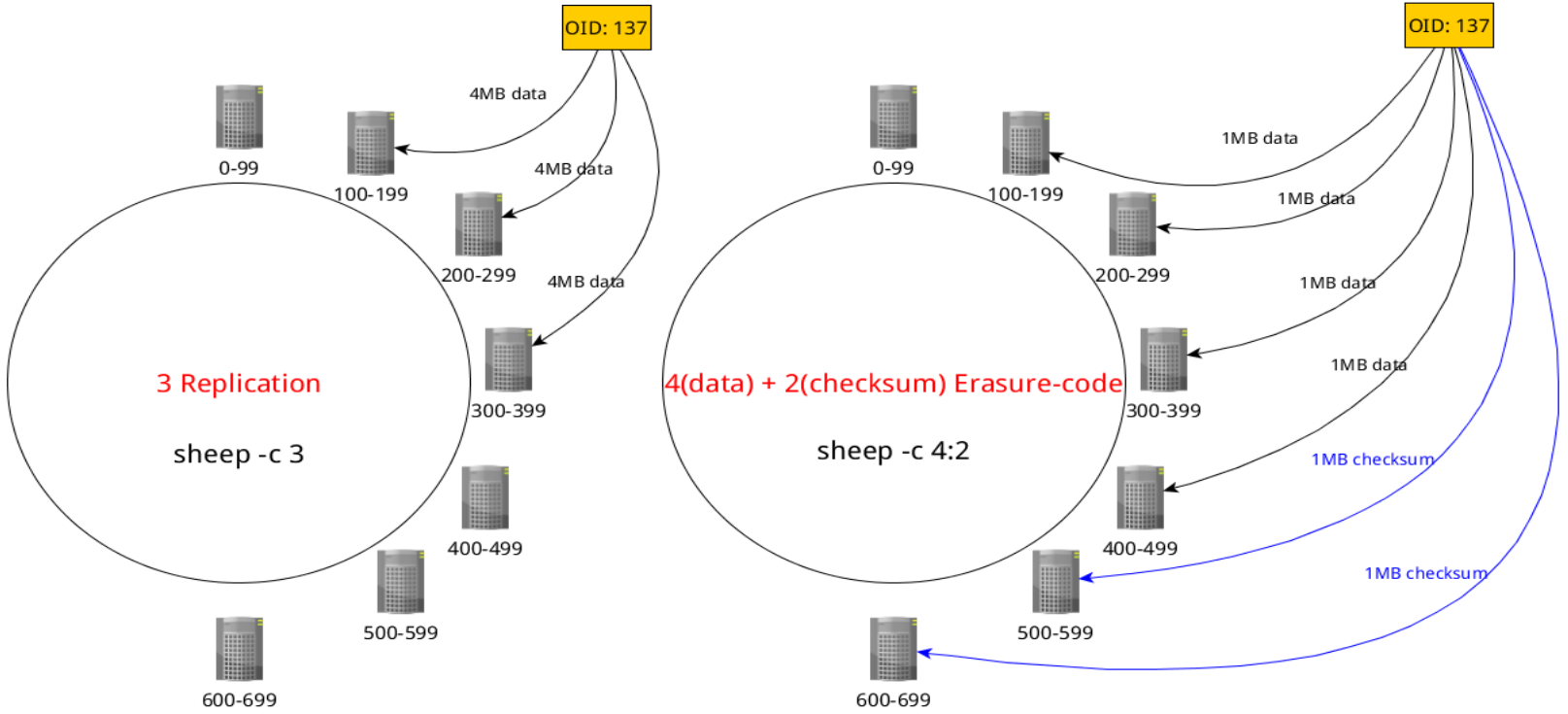
- Yuan Liu, implements erasure code support with zfec, and many other sheepdog improvement.
- Robin Dong, implements (1) RESTful API, complying with Openstack Swift interface spec; (2) hyper volume, for big data object storage; (3) data recovery performance improvement

All general patches are back to sheepdog upstream. Nov 2014, Lambert starts online service in selected data center.

Zfec project: http://freecode.com/projects/zfec , Openstack swift project: https://swiftstack.com/openstack-swift/

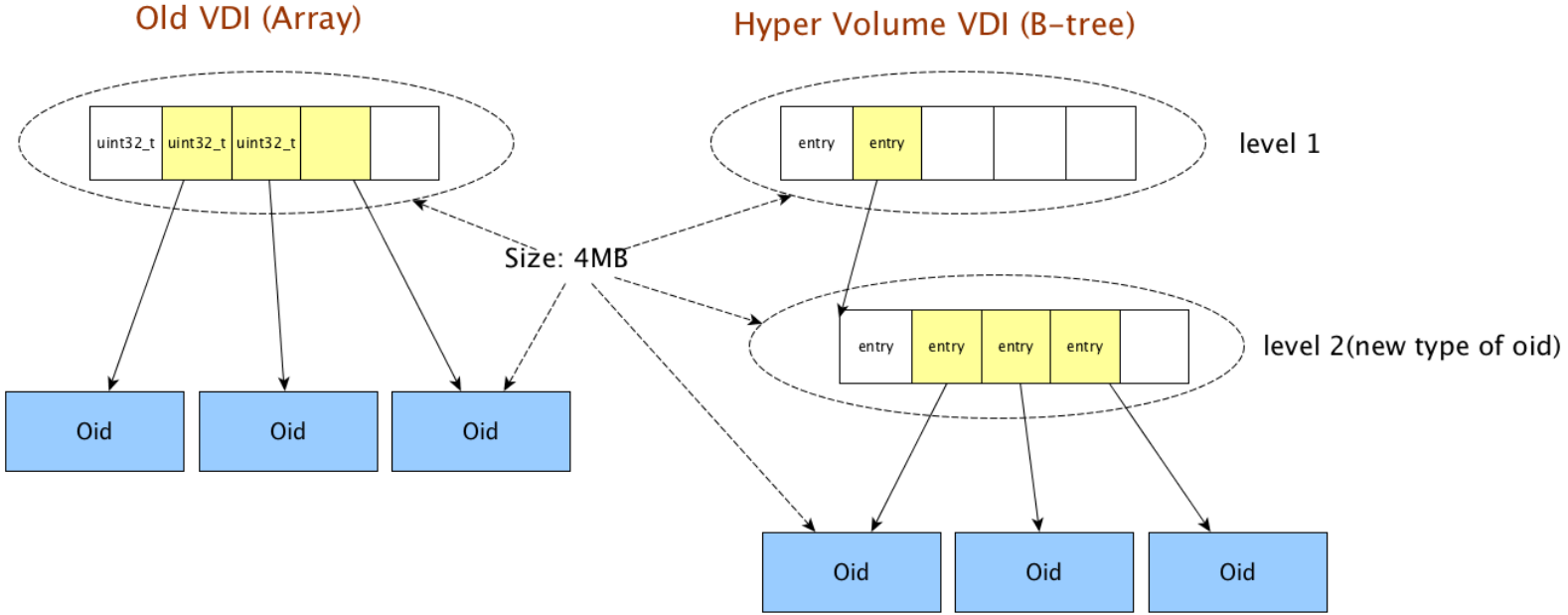
Improvement & Contribution to sheepdog project (Cont.)

Erasure Code



Improvement & Contribution to sheepdog project (Cont.)

Hyper volume implementation

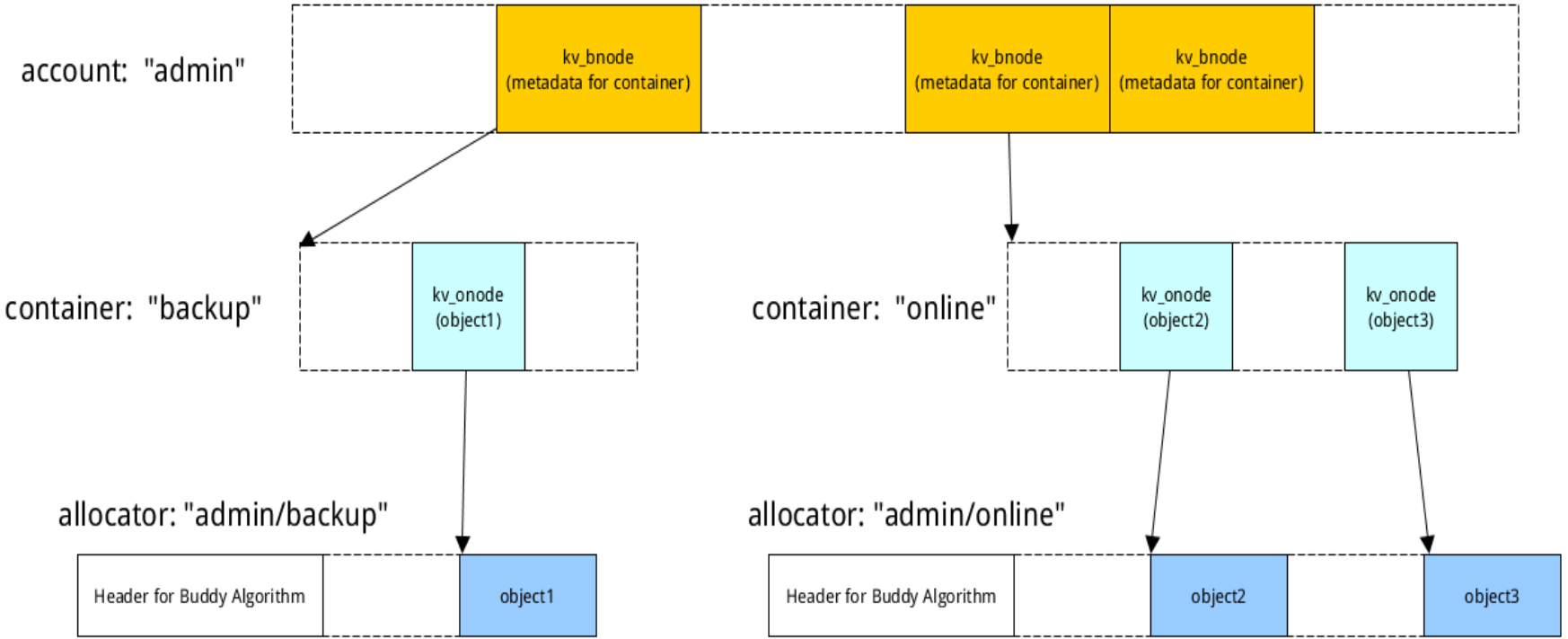


Capacity: (4MB / sizeof(uint32_t)) x 4MB = 4TB

Capacity: (4MB / sizeof(entry)) x (4MB / sizeof(entry)) x 4MB > 600PB

Improvement & Contribution to sheepdog project (Cont.)

Swift interface implementation



Improvement & Contribution to sheepdog project (Cont.)

Data recovery optimization,

- Enable multiple threads for data object recovery.
- Optimize recovery mode:

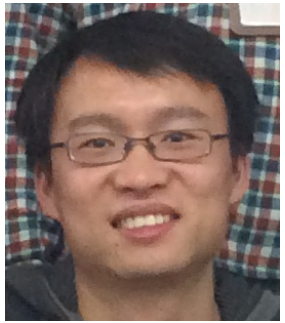
  Before (Node mode): If one disk failed, the server will fetch data from other servers and calculate the EC to recovery its own lost data.

  After (Disk mode): All the servers in subcluster will participate the recovery work.

Recovery performance increases 4 times, which results much better data storage durability.

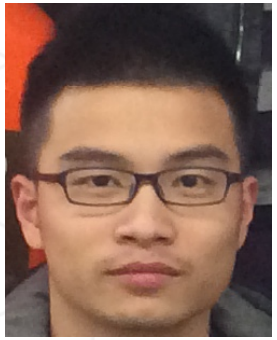
Hats off to Lambert Engineering Team


Coly Li


Robin Dong


Yuan Liu


Guining Li


Kai Zhou

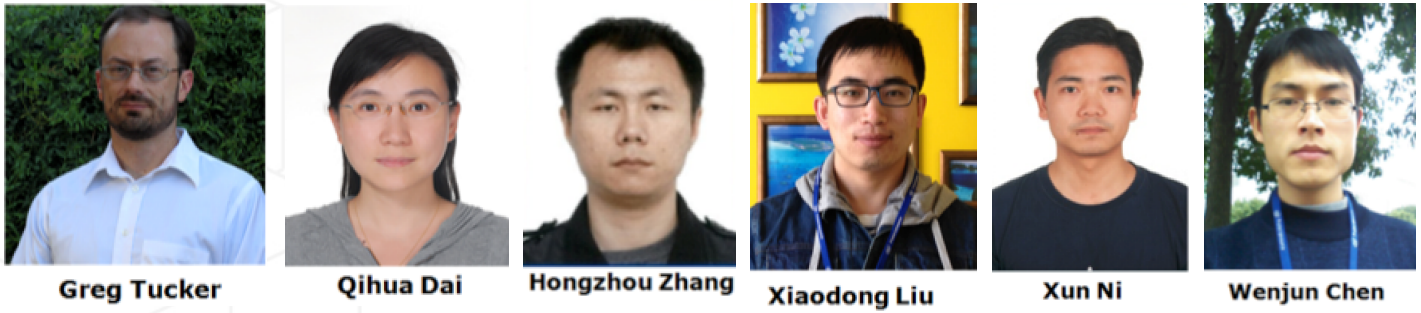

Bingpeng Zhu


Meng An

Credit to open source community

Without cooperation with open source community, we are not able to move forward fast.

- Sheepdog community builds a simple and elegant code base for our cold data storage engine. Top 10 contributors of sheepdog project[1]:
  liuy, kazum, mitake, levin108, RobinDong, fujita, liangry, kylezh, yunkai, arachsys
- Intel open source engineering team help us to accelerate generic storage algorithm[2] performance on Intel ATOM processors.



[1] Some people in top 10 contributors work{s,ed} in Alibaba as well, they are liuy, levin108, RobinDong, yunkai.
[2] Performance acceleration is included but not limited to erasure code, hashing, and other storage related algorithms.

Great Thanks to You All !