Sheepdog: Yet Another All-In-One Storage For Openstack

Openstack Hong Kong Summit
Liu Yuan 2013.11.8

Who I Am

- Active contributor to various open source projects such as Sheepdog, QEMU, Linux Kernel, Xen, Openstack, etc.
- Primary core contributor of Sheepdog project and co-maintains it with Kazutaka Morita from NTT Japan
- Technically lead the storage projects based on Sheepdog for internal uses of www.taobao.com
- Contacts
  - Email: namei.unix@gmail.com
  - Micro Blog: @ 淘泰来

Agenda

- Introduction - Sheepdog Overview
- Exploration - Sheepdog Internals
- Openstack - Sheepdog Goal
- Roadmap - Features From The Future
- Industry - How Industry Use Sheepdog

Introduction

Sheepdog Overview

What Is Sheepdog

- Distributed Object Storage System In User Space
  - –Manage Disks and Nodes
    - •Aggregate the capacity and the power (IOPS + throughput)
    - •Hide the failure of hardware
    - •Dynamically grow or shrink the scale
  - –Manage Data
    - •Provide redundancy mechanisms (replication and erasure code) for high-availability
    - •Secure the data with auto-healing and auto-rebalanced mechanisms
  - –Provide Services
    - •Virtual volume for QEMU VM, iSCSI TGT (Perfectly supported by upstream)
    - •RESTful container (Openstack Swift and Amazon S3 Compatible, in progress)
    - •Storage for Openstack Cinder, Glance, Nova (Available for Havana)

Sheepdog Architecture

Zookeeper: membership management and message queue

No meta servers!

Global Consistent Hash Ring and P2P

sheep daemon

Node event: global rebalance

Disk event: local rebalance

dog admin tool

Private Hash Ring

Gateway	Gateway	Gateway
Store	Store	Store
1TB, 1TB, 1TB, 4TB	1TB, 1TB, 2TB	1TB, 2TB, X
	Hot-plugged	Auto unplugged on EIO

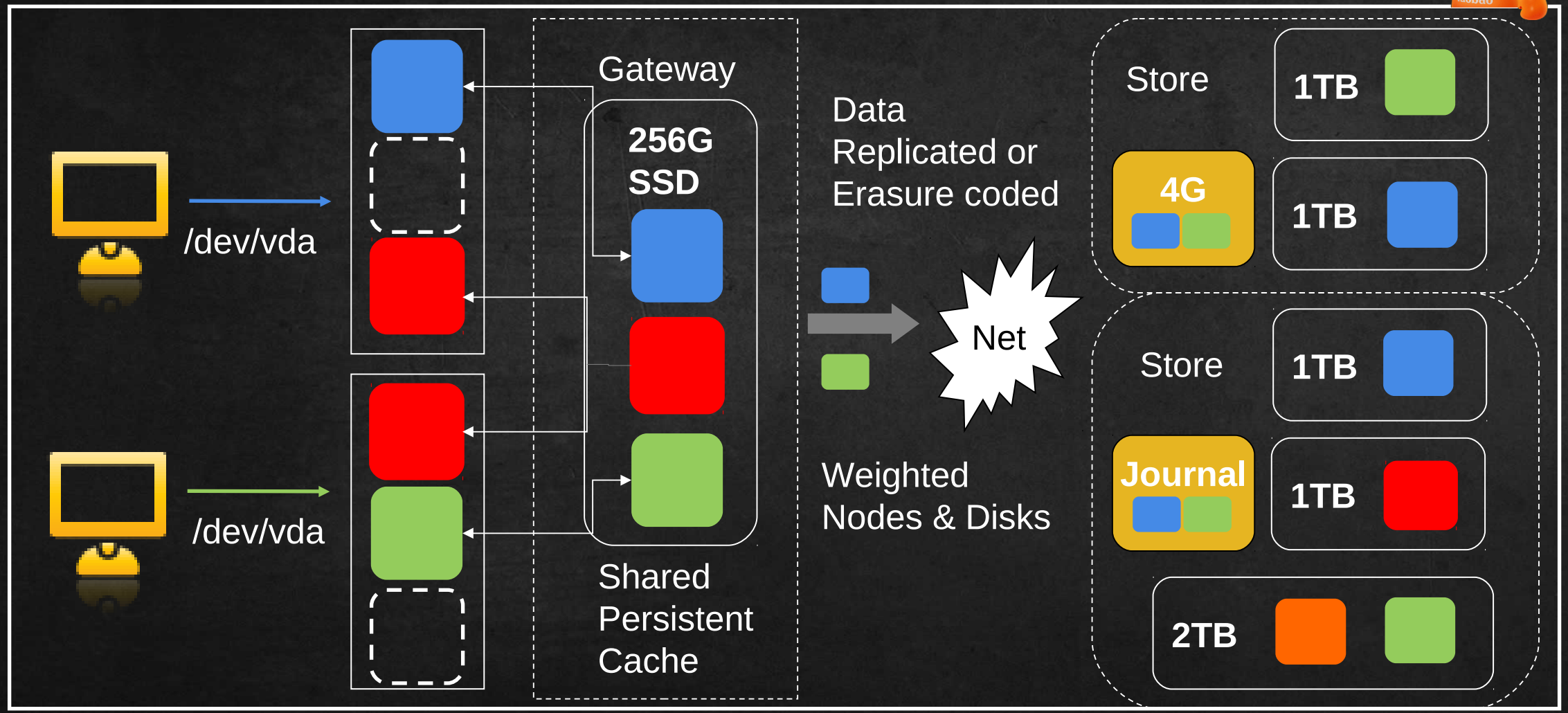
Why Sheepdog

- Minimal assumptions of underlying kernel and file system
  - –Any type of file systems that support extended attribute(xattr)
  - –Only require kernel version >= 2.6.32
- Full of features
  - –snapshot, clone, incremental backup, cluster-wide snapshot, discard, etc.
  - –User-defined replication/erasure code scheme on VDI(Virtual Disk Image) basis
  - –Auto node/disk management
- Easy to set up the cluster with thousands of nodes
  - –Single daemon can manage unlimited number of disks in one node as efficient as RAID0
  - –as many as 6k+ for a single cluster
- Small
  - –Fast and very small memory footprint (less than 50 MB even when busy)
  - –Easy to hack and maintain, 35K lines of code in C as of now

Exploration

Sheepdog Internals

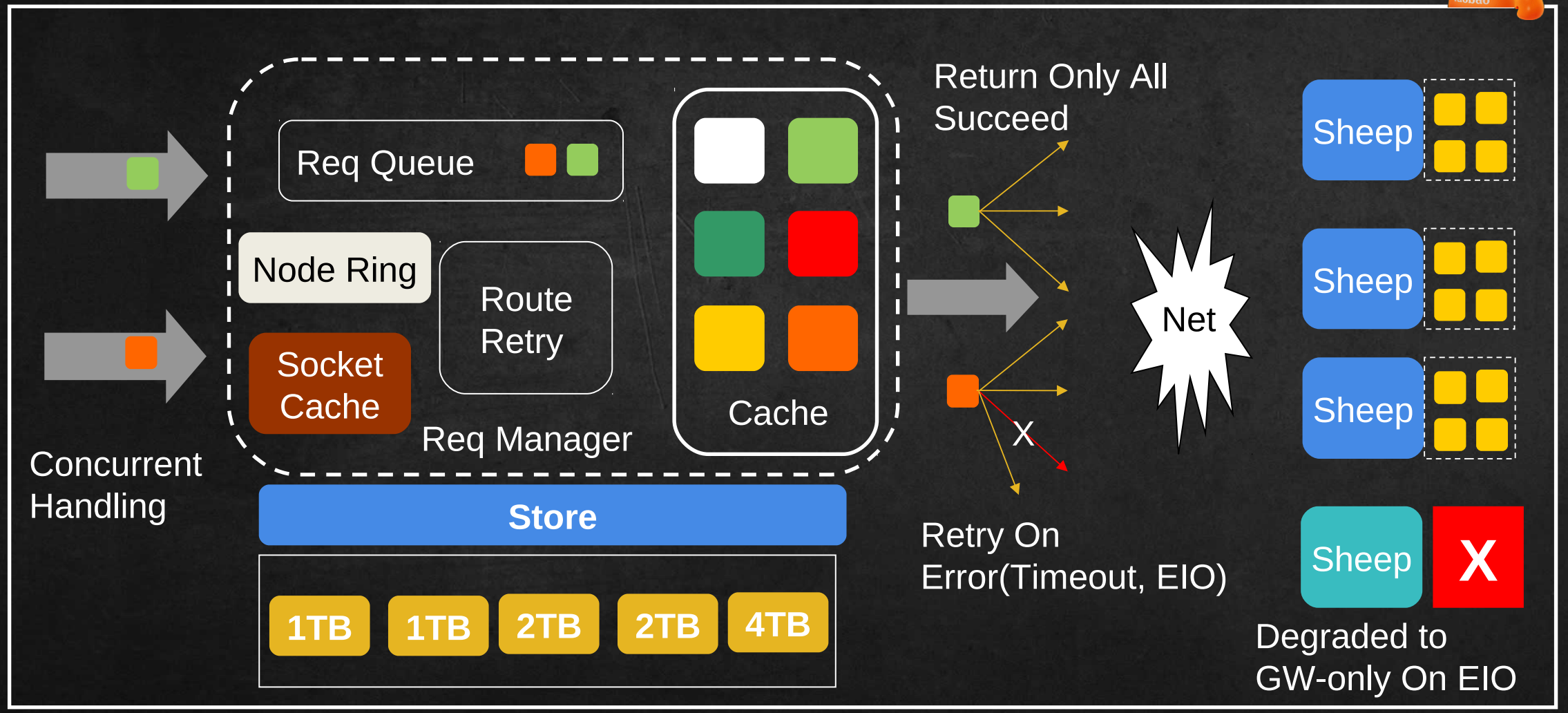
Sheepdog In A Nutshell



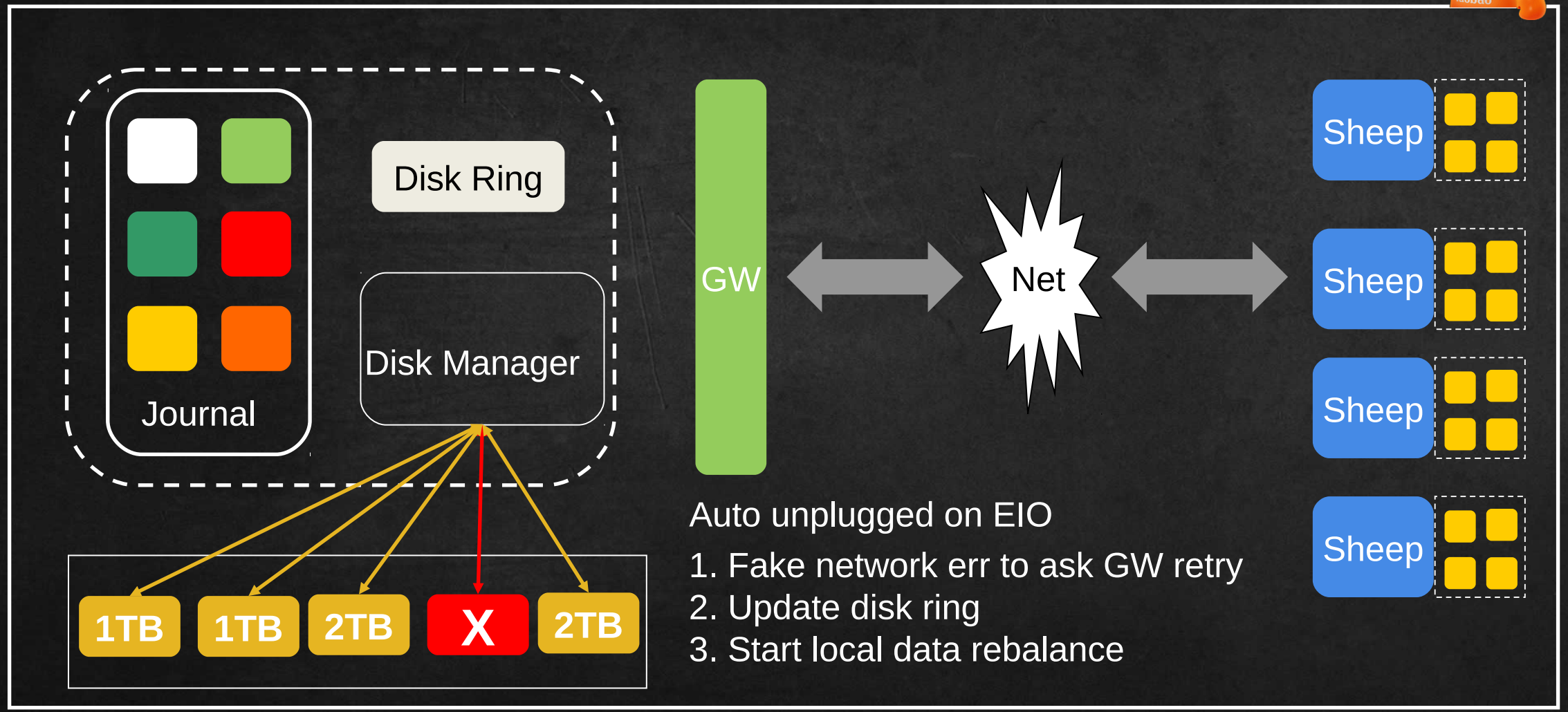
Sheepdog Volume

- Copy-On-Write Snapshot
  - –Disk-only snapshot, disk & memory snapshot
  - –Live snapshot, offline snapshot
  - –Rollback(tree structure), clone
  - –Incremental backup
  - –Instant operation, only create 4M inode object
- Push many logics into client -> simple and fast code !
  - –Only 4 opcodes for store, read/write/create/remove, snapshot is done by QEMU block driver or dog
  - –Requests serialization is not handled by Sheepdog but client
  - –Inode object is treated the same as data object

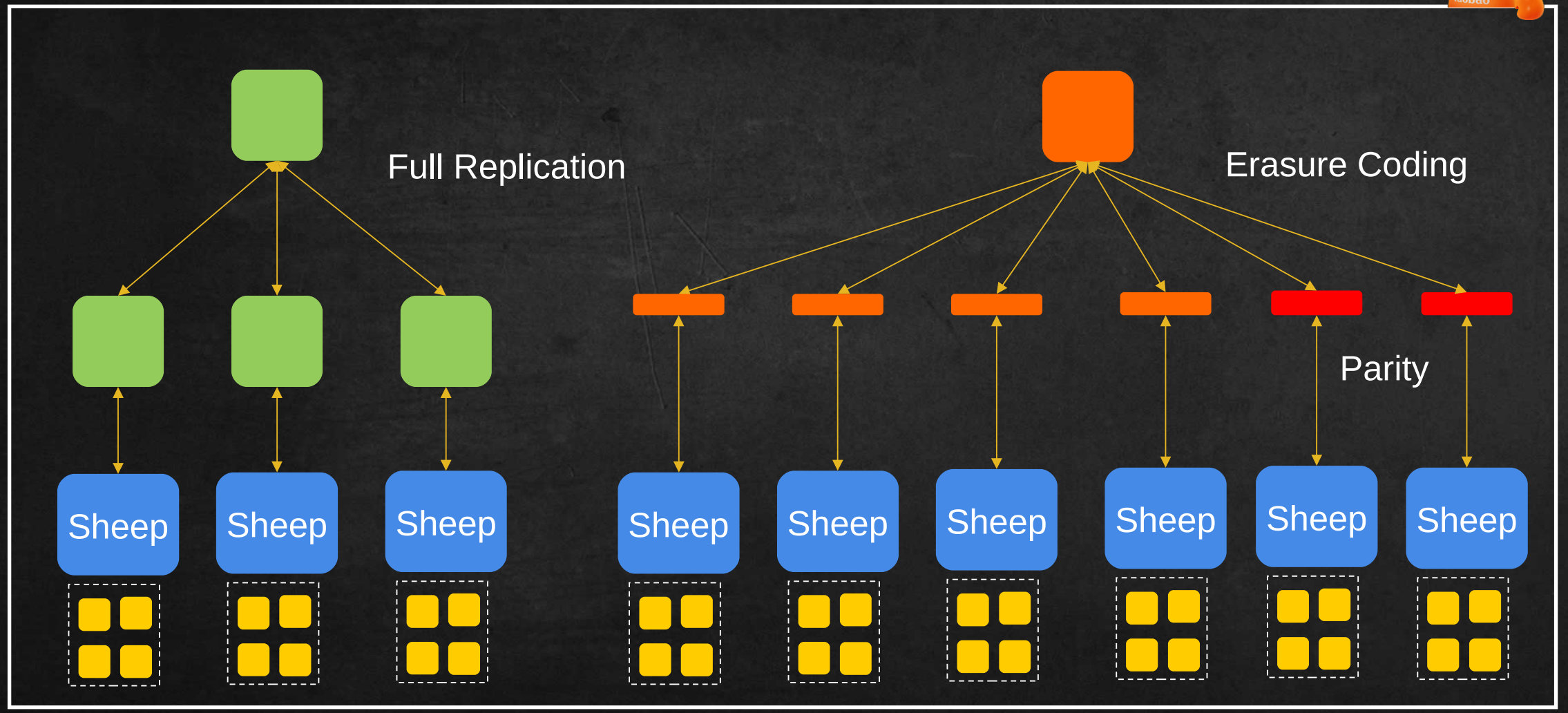
Gateway - Request Engine



Store - Data Engine

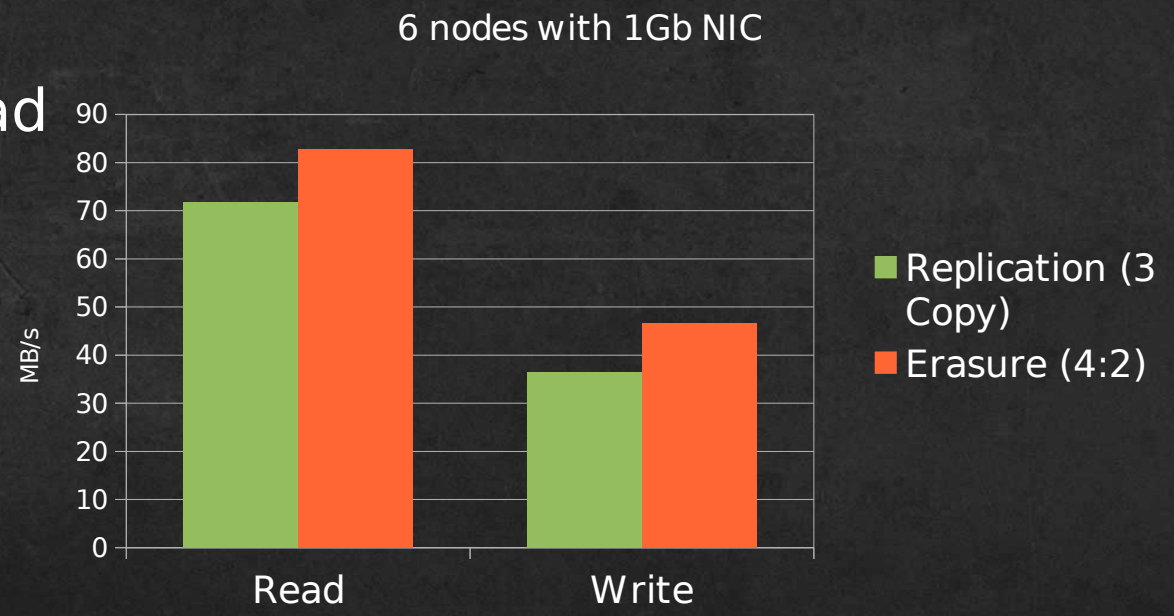


Redundancy Scheme

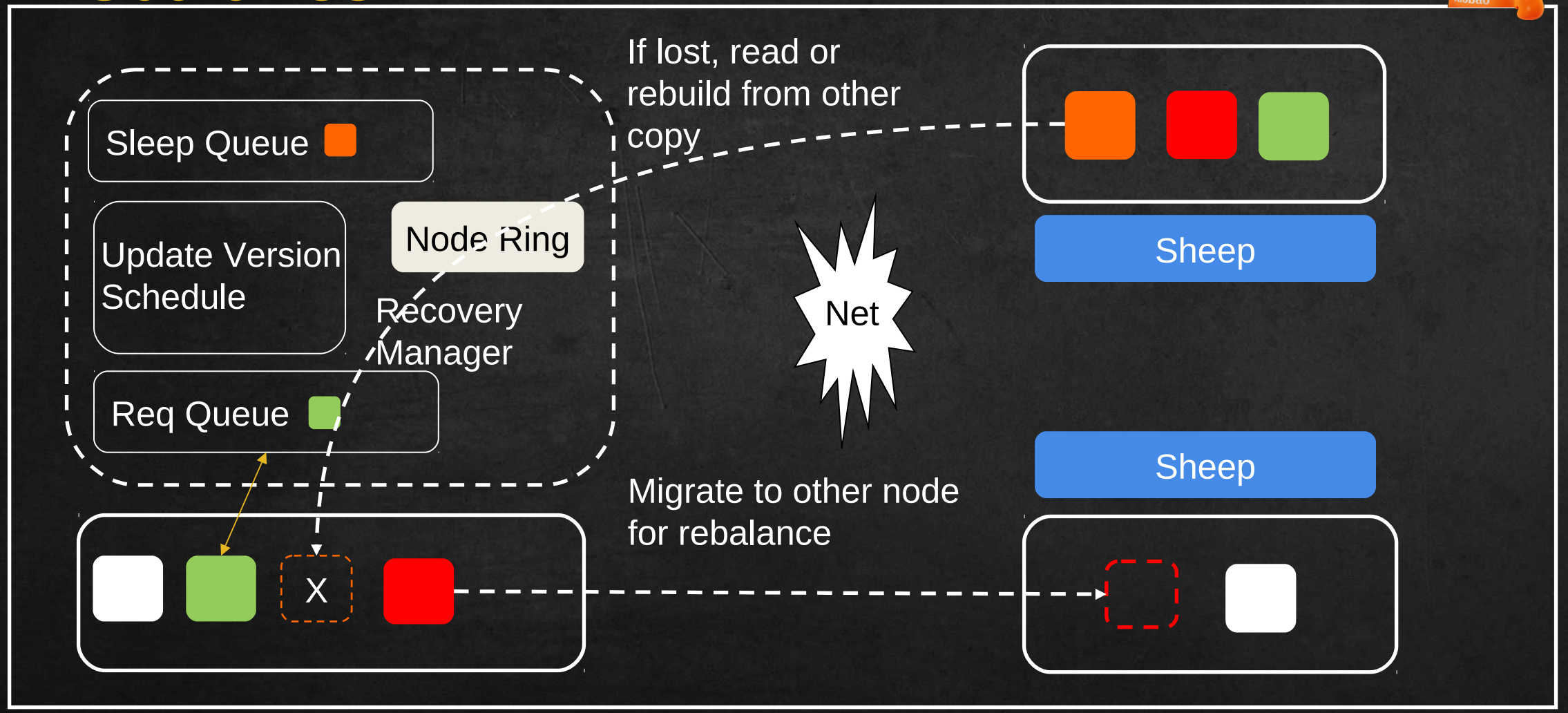


Erasure Coding Over Full Replication

- Advantages
  - –Far less storage overhead
  - –Rumors breaking
    - •Better R/W performance
    - •Support random R/W
    - •Can run VM Images !
- Disadvantages
  - –Generate more traffic for recovery
    - •X/Y times data (Suppose X data, Y parity strips)



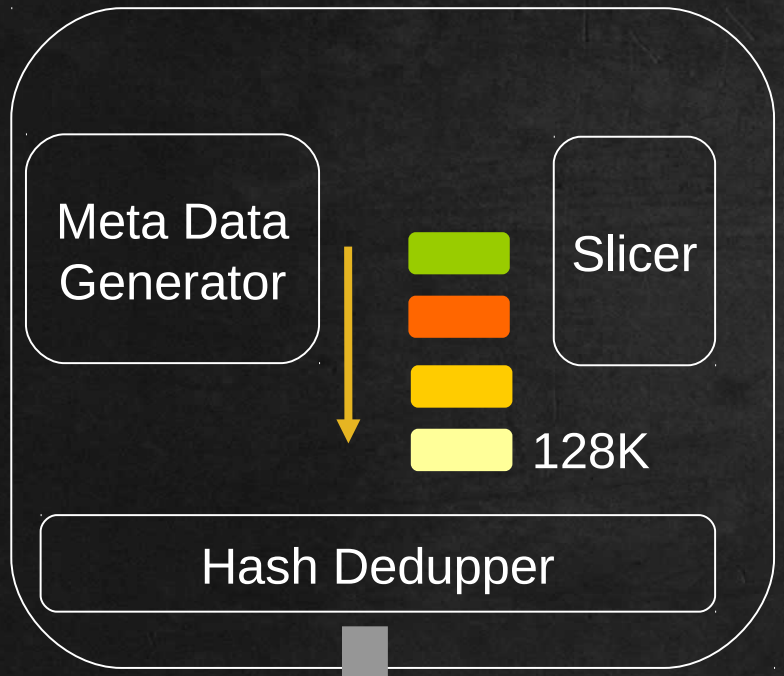
Recovery - Redundancy Repair & Data Rebalance

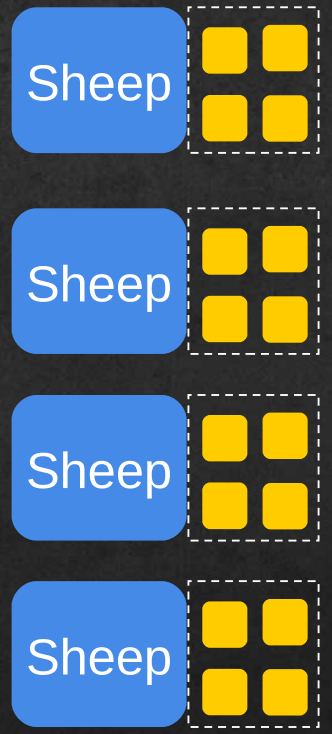


Recovery Cont.

- Eager Recovery As Default
  - –Allow users to stop it and do manual recovery temporarily
- Node/Disk Events Handling
  - –Disk event and node event share the same algorithm
    - •Handle mixed node and disk events nicely
  - –Subsequent event will supersede previous one
    - •Handle group join/leave of disks and nodes gracefully
- Recovery Handling Transparently to the Client
  - –Put requests for objects being recovered on sleep queue and wake it up later
  - –Serve the request directly if object is right there in the store

Farm - Cluster Wide Snapshot





- Incremental backup
- Up to 50% dedup ratio
- Compression doesn't help

Dedicated Backup Storage

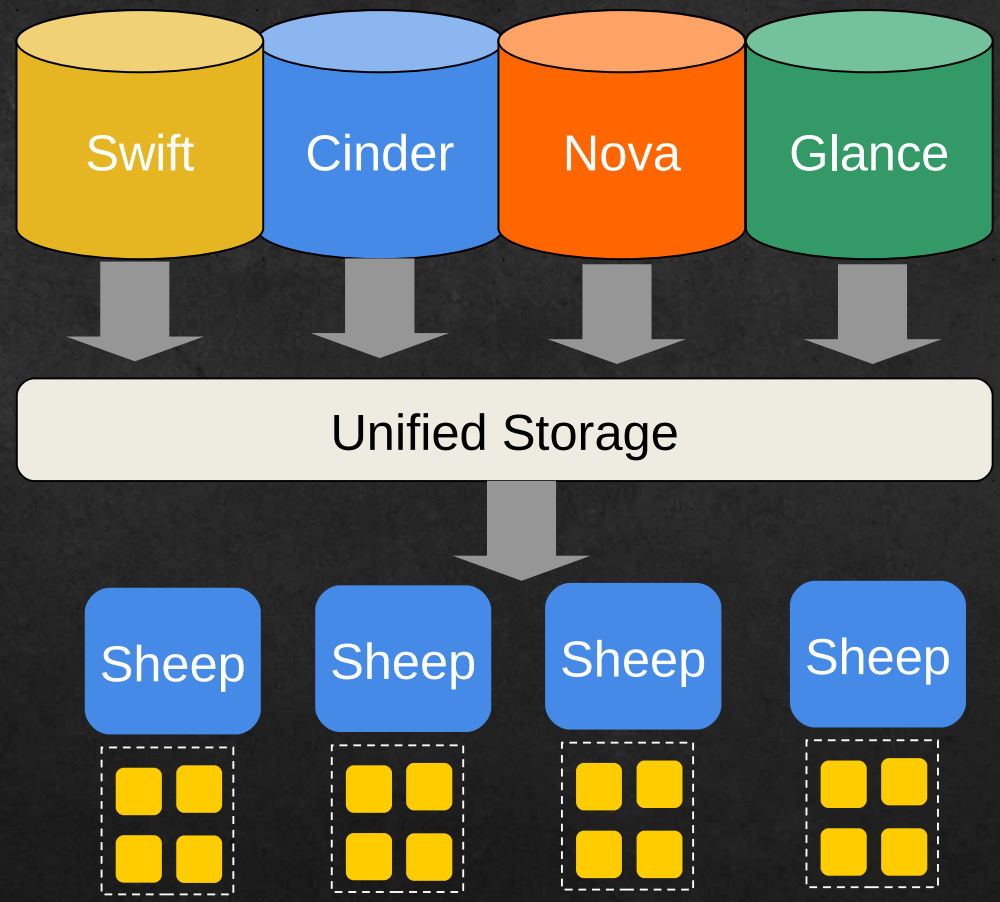
Think of Sheepdog On Sheepdog ? Yeah!

Openstack

Sheepdog Goal

Openstack Storage Components

- Cinder - Block Storage
  - –Support since day 1
- Glance - Image Storage
  - –Support merged at Havana version
- Nova - Ephemeral Storage
  - –Not yet started
- Swift - Object Storage
  - –Swift API compatible In progress
- Final Goal - Unified Storage
  - –Cope-On-Write anywhere ?
  - –Data dedup ?



Roadmap

Features From The Future

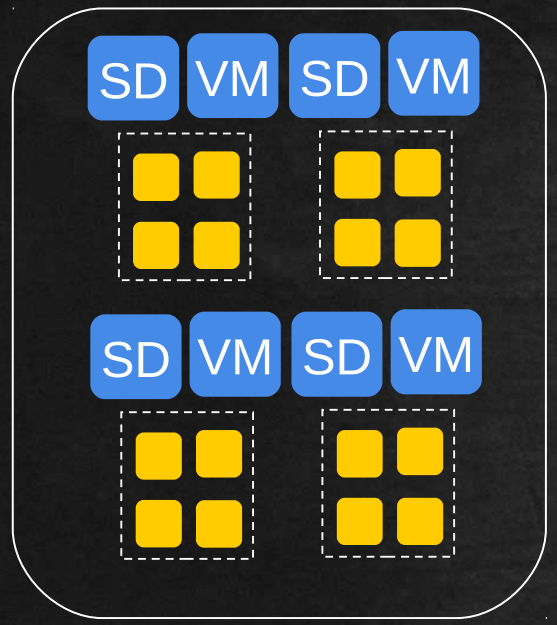
Look Into The Future

- RESTful Container
  - –Plans to be Openstack Swift API compatible first, coming soon
- Hyper Scale Volume
  - –256PB Volume, coming soon
- Geo-Replication
- Sheepdog On Sheepdog
  - –Storage for cluster wide snapshot
- Slow Disk & Broken Disk Detecter
  - –Deal with dead D state process hang because of broken disk in massive deployment

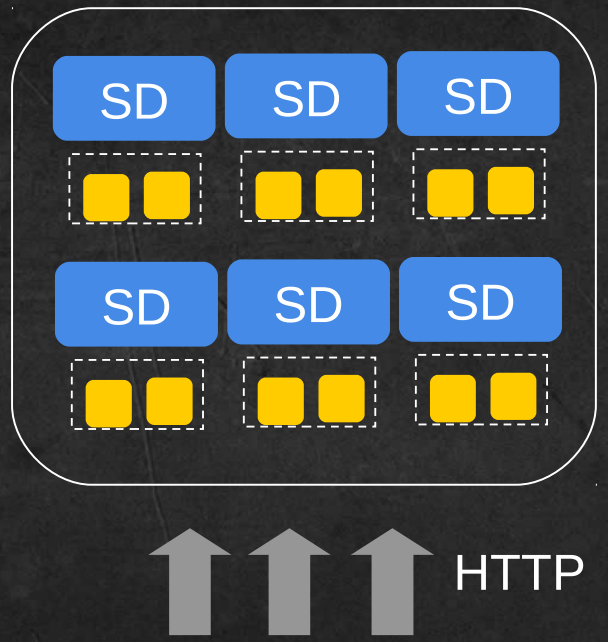
Industry

How Industry Use Sheepdog

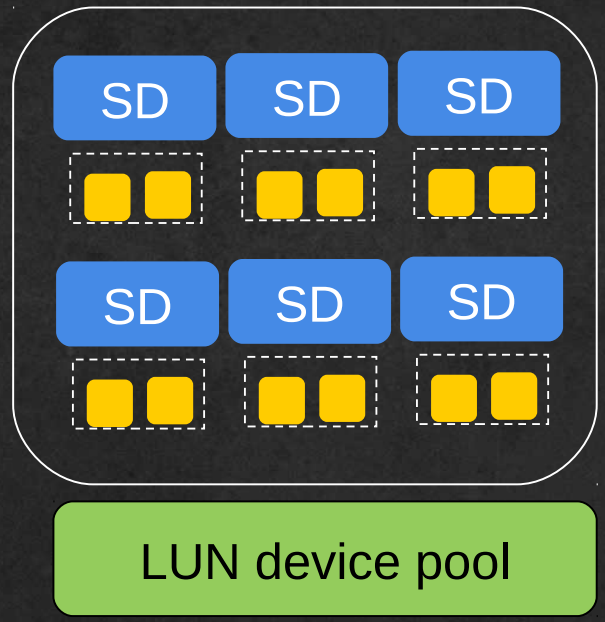
Sheepdog In Taobao & NTT



VM running inside Sheepdog Cluster for test & dev at Taobao

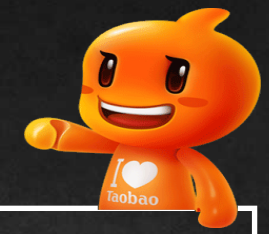


Ongoing project with 10k+ ARM nodes for cold data at Taobao



Sheepdog cluster run as iSCSI TGT backend storage at NTT

Other Users In Production



Extensys

SWIFT The global provider of secure financial messaging services

美橙互联™ CNDNS.com

海云捷迅 www.awcloud.com

Any more users I don't know ?

Q & A

Homepage

http://sheepdog.github.io/sheepdog/

Try me out

git clone git://github.com/sheepdog/sheepdog.git

Go Sheepdog !