LeoFS

Scaling and High Performance Storage System

Yosuke Hara - @yosukehara

A Researcher of R.I.T. and **Tech Lead LeoFS**

with Hiroki Matsue, LeoFS Support and Rakuten Software Engineer

楽天 Rakuten

R.I.T. Rakuten Institute of Technology

LeoFS is an **Unstructured Object Storage for the Web** and a highly available, distributed, eventually consistent storage system.

LeoFS

The Lion of Storage Systems

LeoFS was published as OSS on July of 2012

leo-project.net/leofs

楽天 Rakuten

Overview

Brief Benchmark Report

Multi Data Center Replication

LeoFS Administration at Rakuten

Future Plans
“NFS” Support and more

楽天 Rakuten

Overview

楽天 Rakuten

Tokyo, Japan

LeoFS

The Lion of Storage Systems

HIGH Availability

LeoFS Non Stop

3 Vs in 3 HIGHs

Velocity: Low Latency
Minimum Resources

Volume: Petabyte / Exabyte
Variety: Photo, Movie, Unstructured-data

HIGH Cost Performance Ratio

HIGH Scalability

楽天 Rakuten

LeoFS Overview

Request from
Web Applications / Browsers
w/HTTP over REST-API / S3-API

Load Balancer

Keeping High Availability
Keeping High Performance
Easy Administration

Gateway

Manager

(Erlang RPC)

(Erlang RPC)

(TCP/IP,SNMP)

Nagios
ZABBIX

Monitor

GUI Console

Storage

Storage Engine/Router
RAM
Metadata Object Storage

Storage Engine/Router
RAM
Metadata Object Storage

Storage Engine/Router
RAM
Metadata Object Storage

楽天 Rakuten

Gateway

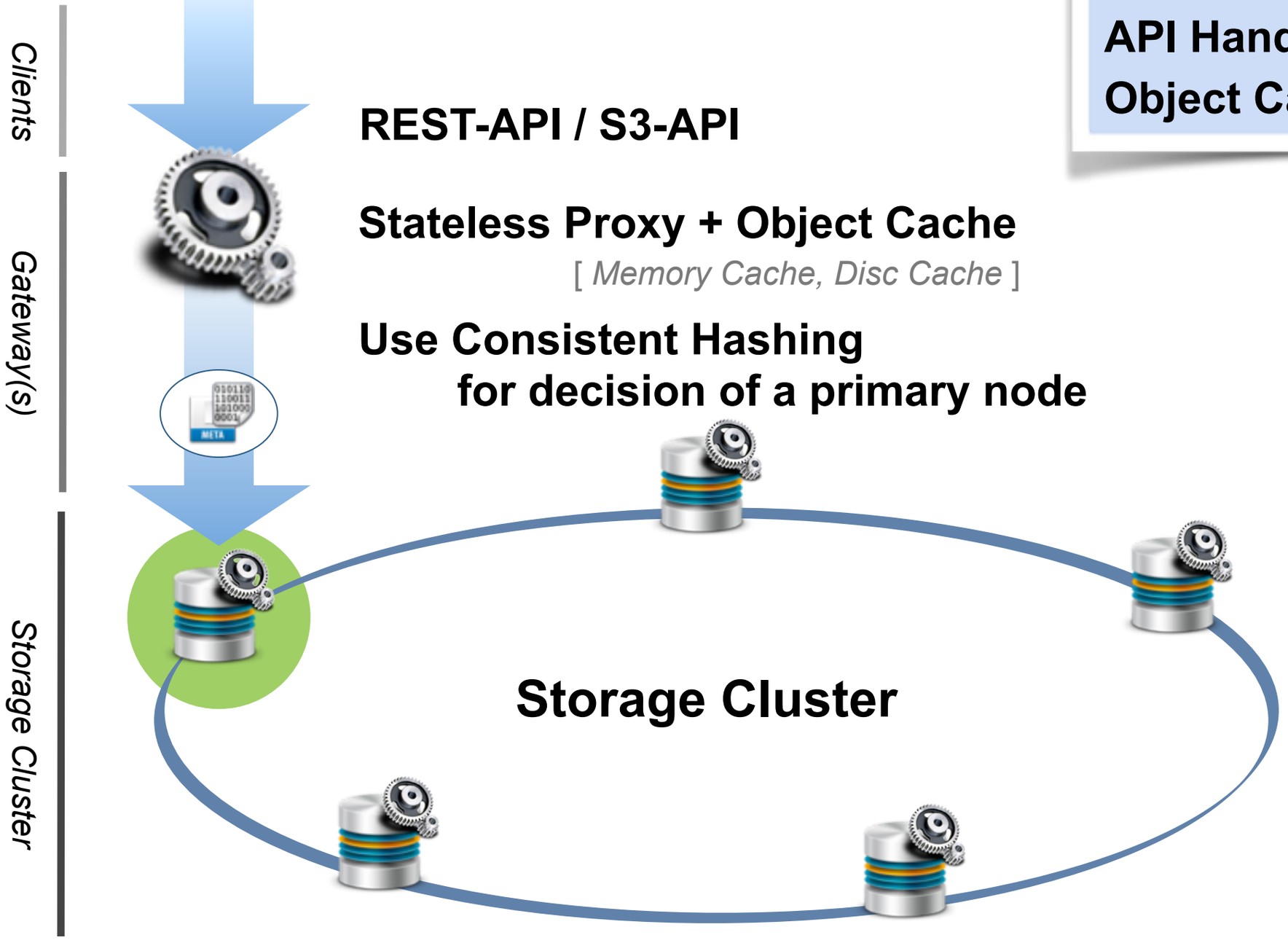
楽R天　RRakuten

LeoFS Overview - Gateway

HTTP Request and Response

Built in *Object Cache Mechanism*

Fast HTTP Server - Cowboy
API Handler
Object Cache Mechanism



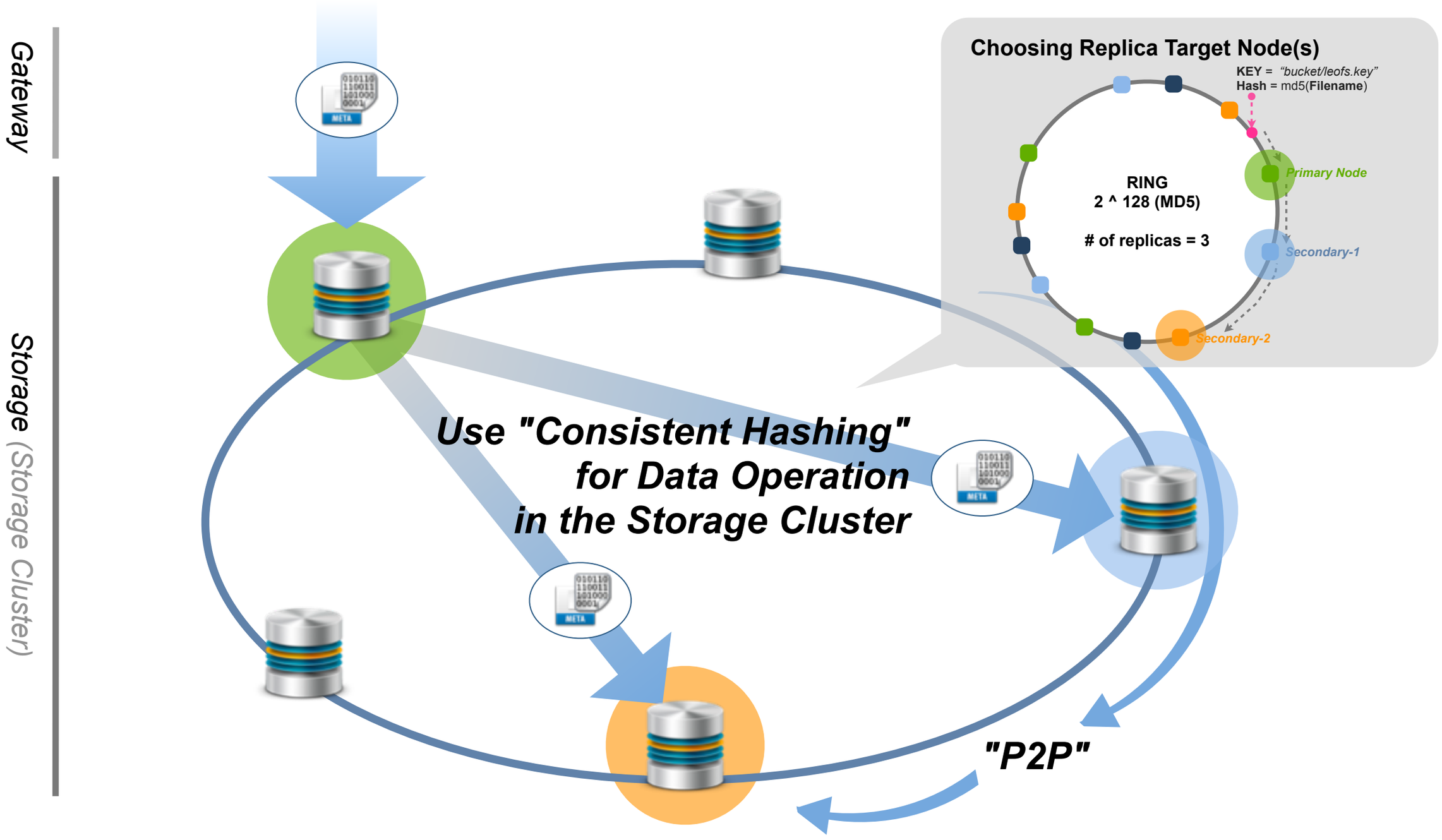
楽天 Rakuten

Storage

楽天 Rakuten

LeoFS Overview - Storage

WRITE: Auto Replication
READ : Auto Repair of an Inconsistent Object with Async

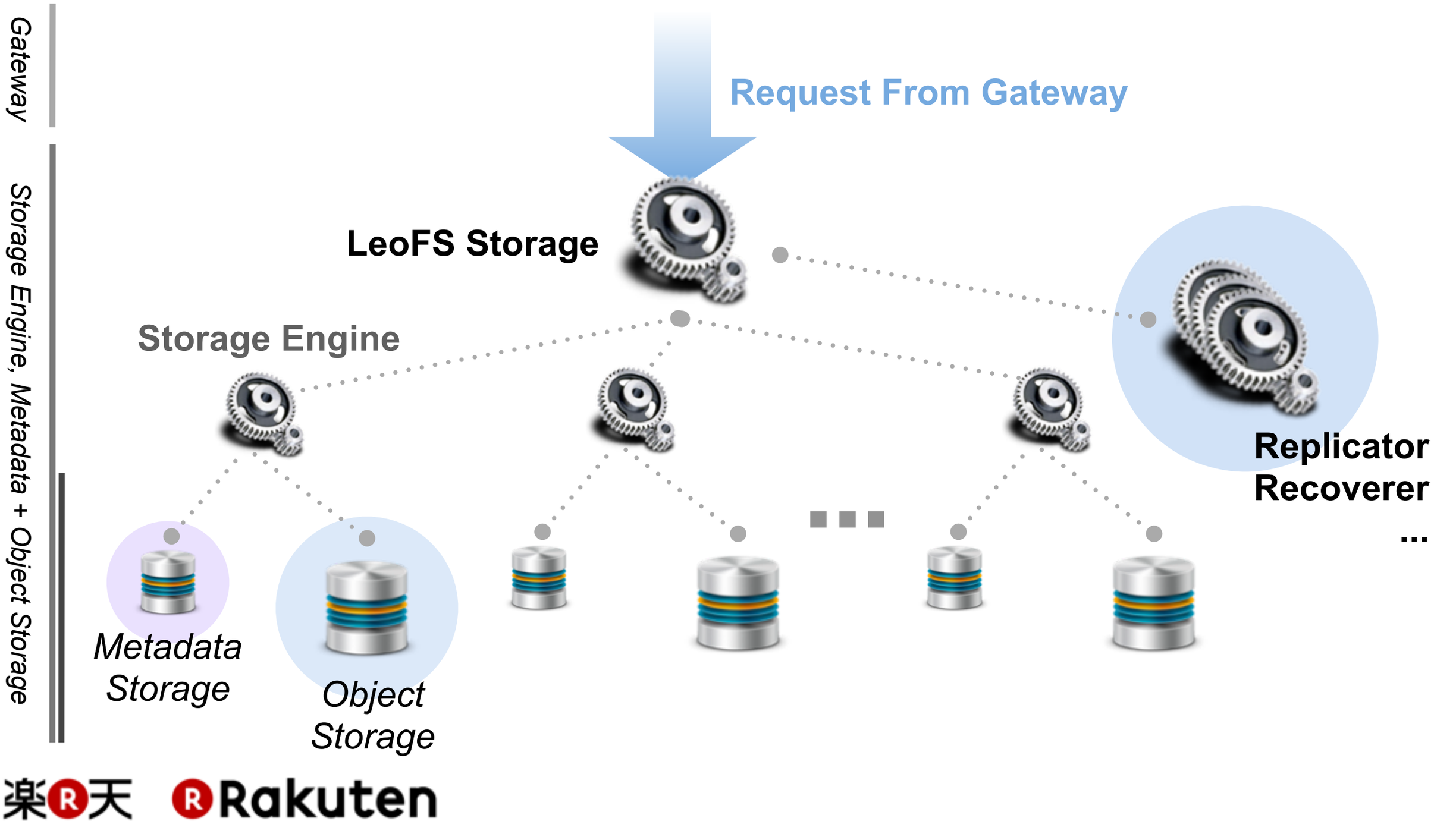


楽天 Rakuten

LeoFS Overview - Storage

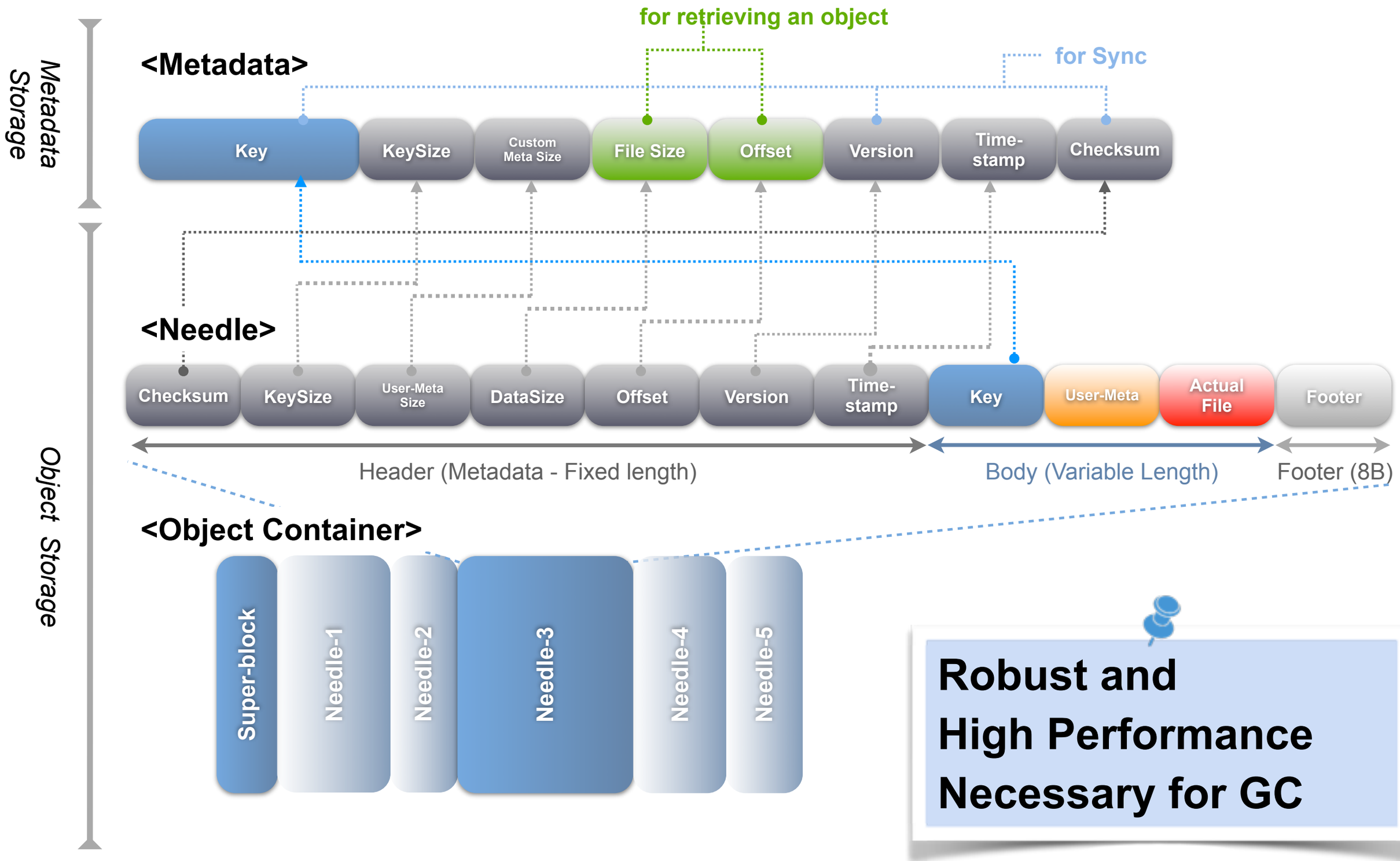
Storage consists of *Object Storage* and *Metadata Storage*

Includes *Replicator* and *Recoverer* for the eventual consistency



楽天 Rakuten

LeoFS Overview - Storage - Data Structure



Robust and
High Performance
Necessary for GC

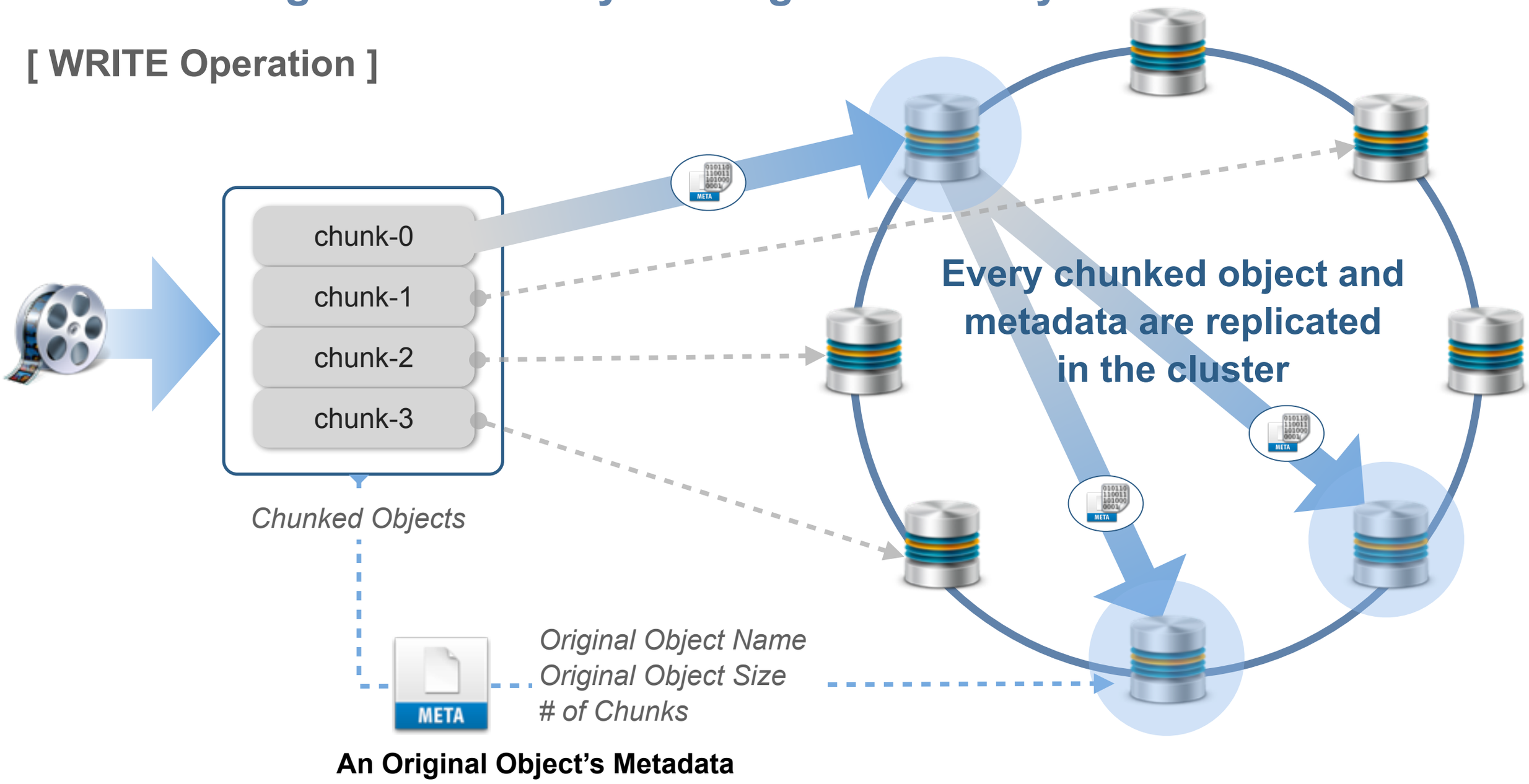
楽天 Rakuten

LeoFS Overview - Storage - Large Object Support

To Equalize Disk Usage in Every Storage Node

To Realize High I/O efficiency and High Availability

[WRITE Operation]



An Original Object's Metadata

Client(s) ***Gateway*** ***Storage Cluster***

楽天 Rakuten

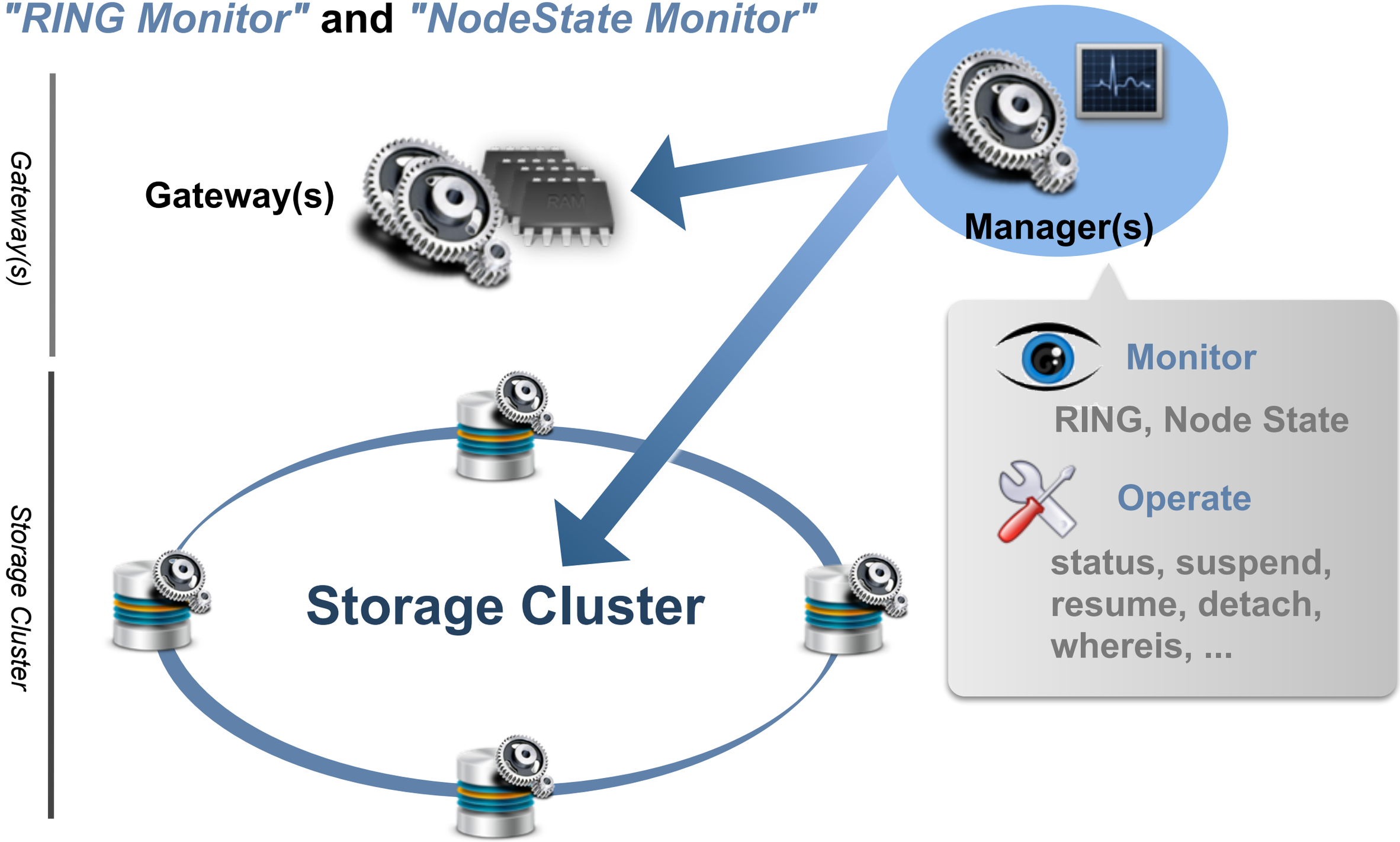
Manager

楽R天　RRakuten

LeoFS Overview - Manager

Operate LeoFS - Gateway and Storage Cluster

"RING Monitor" and *"NodeState Monitor"*



楽天 Rakuten

Brief Benchmark Report

楽天 Rakuten

Hokkaido, Japan

Brief Benchmark Report

Summary of the benchmark results

LeoFS kept in a stable performance through the benchmark

Bottleneck is Disk I/O

The cache mechanism contributed to reduce network traffic between Gateway and Storage

楽天 Rakuten

Brief Benchmark Report

1st Case:

Group of Value Ranges (HDD)

Storage:5, Gateway:1, Manager:2
R:W = 9:1

source: https://github.com/leo-project/notes/tree/master/leofs/benchmark/leofs/20140605/tests/1m_r9w1_240min

2nd Case:

Group of Value Ranges (HDD)

Storage:5, Gateway:1, Manager:2
R:W = 8:2

source: https://github.com/leo-project/notes/tree/master/leofs/benchmark/leofs/20140605/tests/1m_r8w2_120min

楽天 Rakuten

Brief Benchmark Report

Server Spec - Gateway:

CPU	Intel(R) Xeon(R) CPU X5650 @ 2.67GHz * 2 (12 cores / 24 threads)
Memory	96GB
Disk	HDD - 240GB RAID0
Network	10G-Ether

Server Spec - Storage x5:

CPU	Intel(R) Xeon(R) CPU X5650 @ 2.67GHz * 2 (12 cores / 24 threads)
Memory	96GB
Disk	HDD - 240GB RAID0 (System)
	HDD - 2TB RAID0 (Data)
Network	10G-Ether

楽天 Rakuten

Brief Benchmark Report - 1st Case (HDD / R:W=9:1)

Environment:

Network	10Gbps
OS	CentOS release 6.5 (Final)
Erlang	OTP R16B03-1
LeoFS	v1.0.2

System Consistency Level: [N:3, W:2, R:1, D:2]

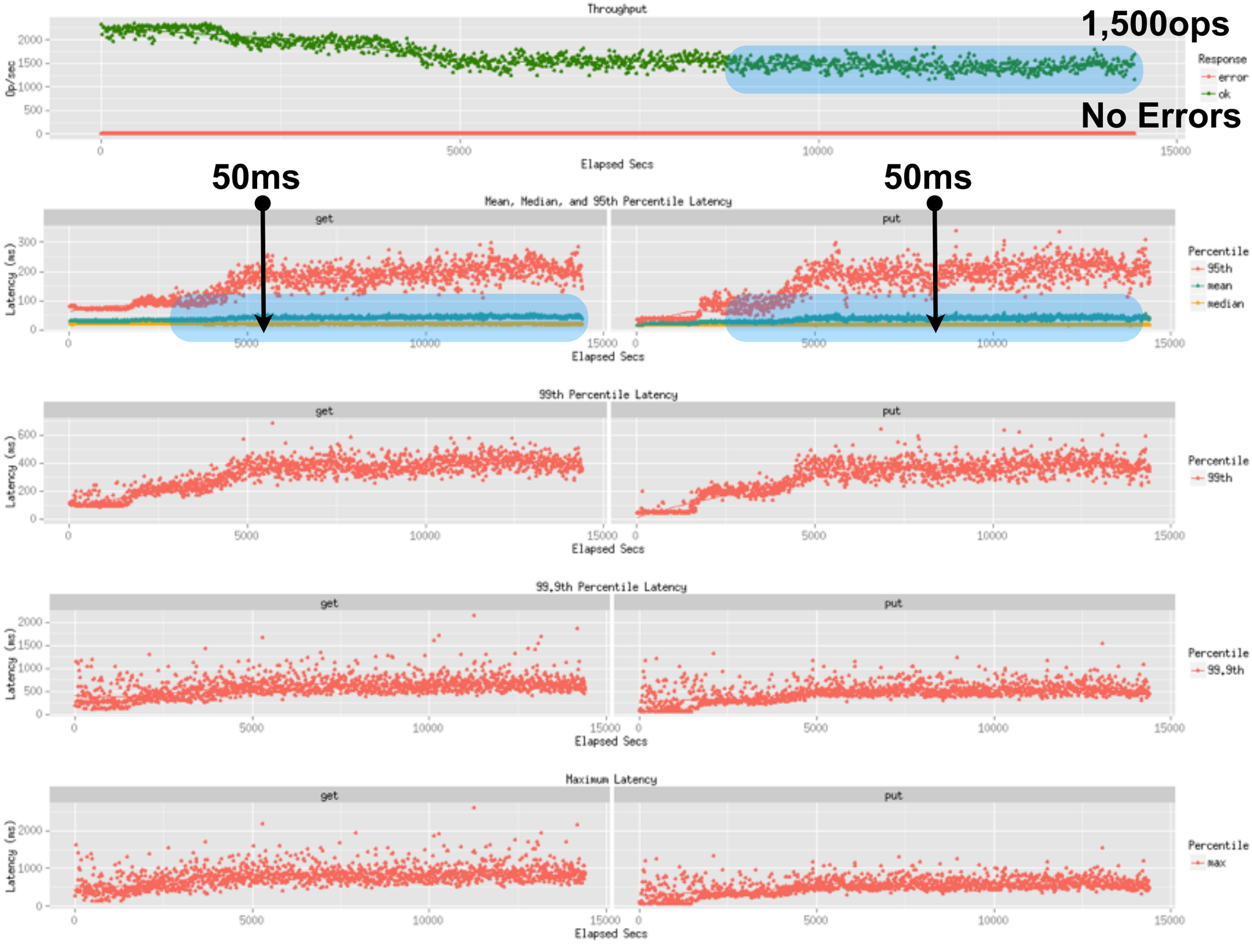
Benchmark Configuration:

Duration	4.0h
R:W	**9:1**
# of Concurrent Processes	64
# of Keys	100,000
Value Size	Range (byte) / Percentage

Range (byte)		Percentage
1024	10240	24%
10241	102400	30%
10241	819200	30%
819201	1572864	16%

楽天 Rakuten

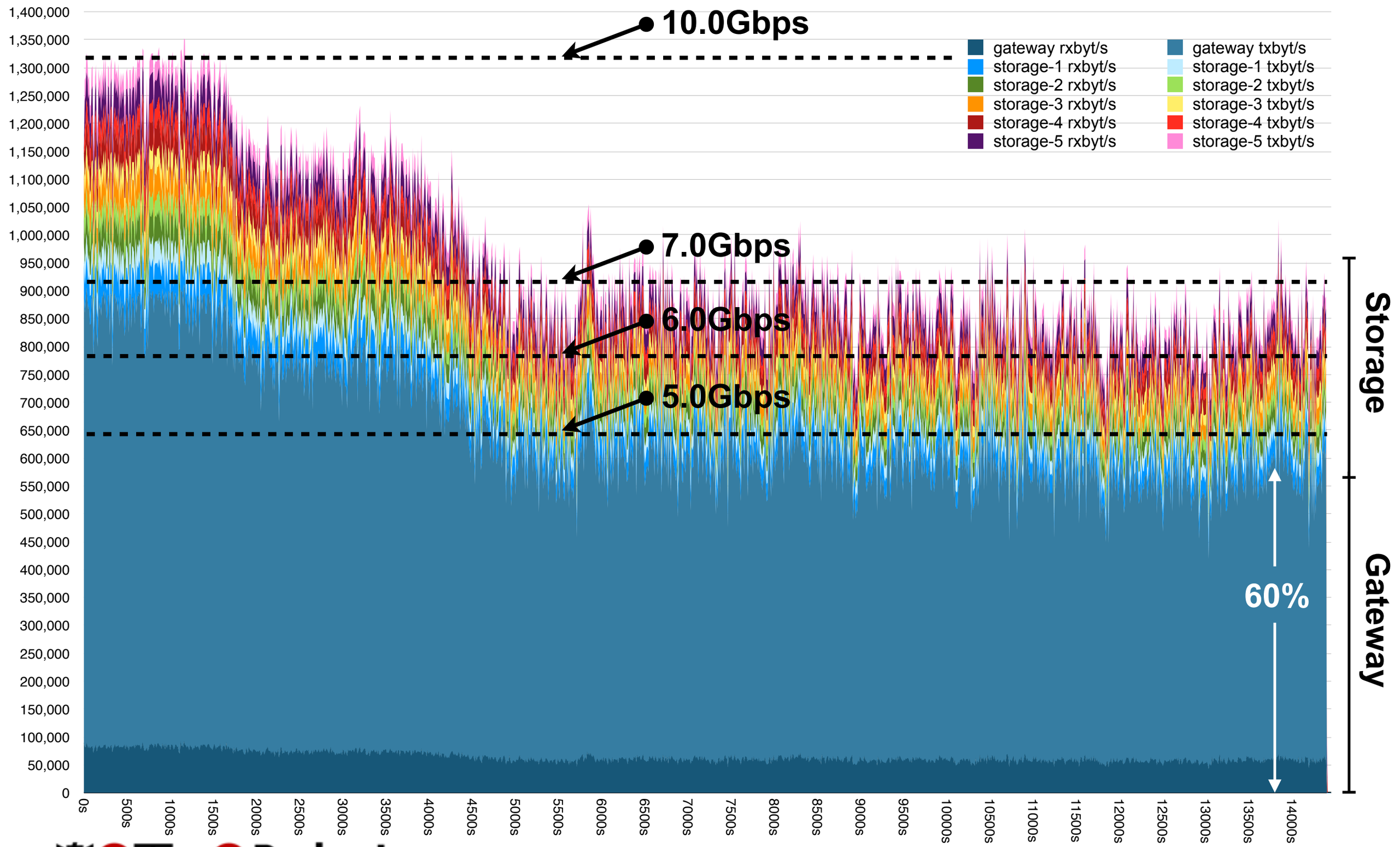
Brief Benchmark Report - 1st Case (HDD / R:W=9:1)



楽天 Rakuten

source: https://github.com/leo-project/notes/tree/master/leofs/benchmark/leofs/20140601/tests/1m_r9w1_240min

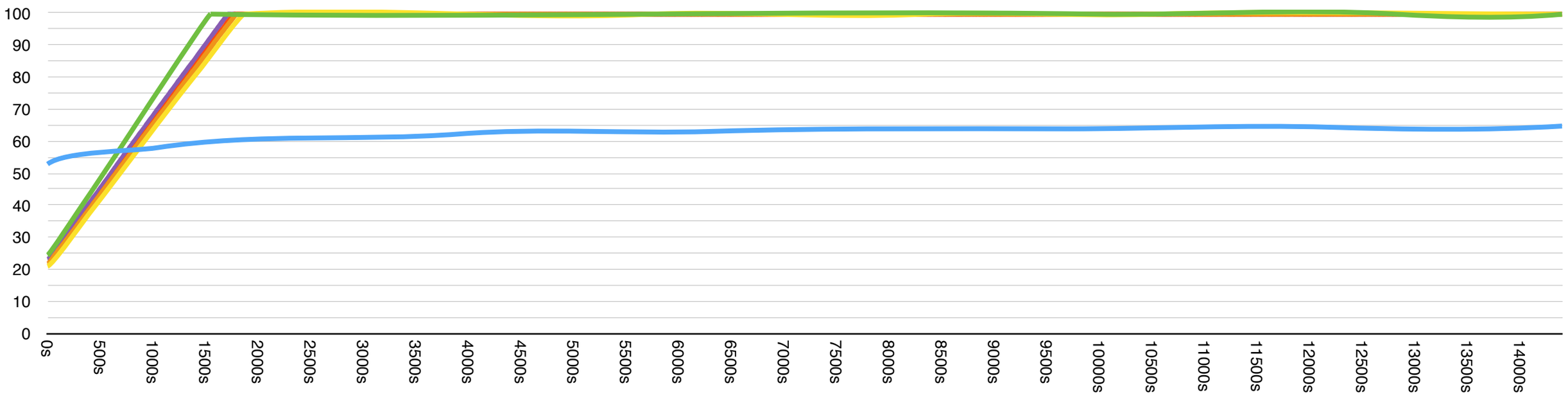
Brief Benchmark Report - 1st Case / Network Traffic



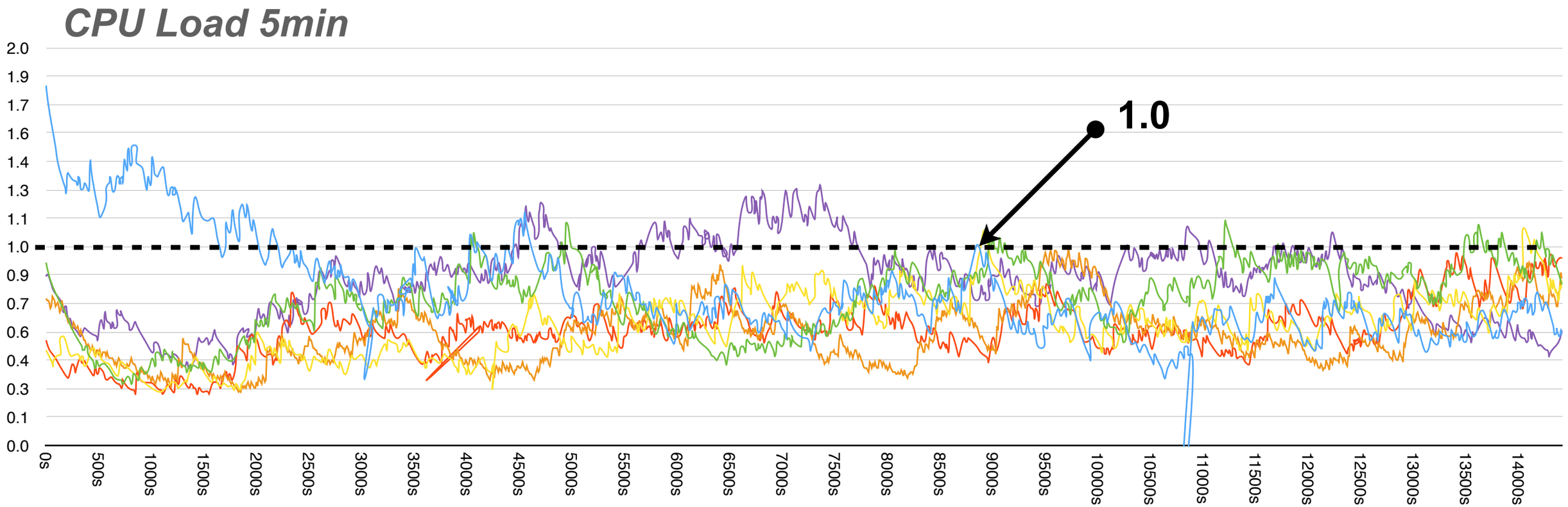
楽天 Rakuten

Brief Benchmark Report - 1st Case / Memory and CPU

Memory Usage



CPU Load 5min



楽天 Rakuten

Brief Benchmark Report - 2nd Case (HDD / R:W=8:2)

Environment:

Network	10Gbps
OS	CentOS release 6.5 (Final)
Erlang	OTP R16B03-1
LeoFS	v1.0.2

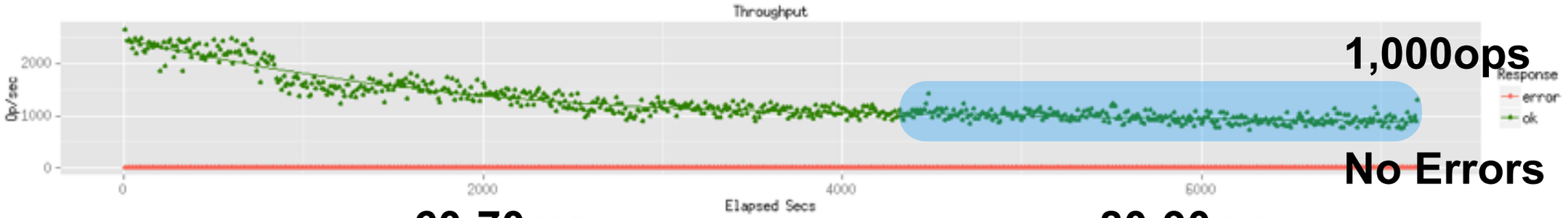
System Consistency Level: [N:3, W:2, R:1, D:2]

Benchmark Configuration:

Duration	2.0h
R:W	**8:2**
# of Concurrent Processes	64
# of Keys	100,000
Value Size	Range (byte) 1024 – 10240: 24% 10241 – 102400: 30% 10241 – 819200: 30% 819201 – 1572864: 16%

楽天 Rakuten

Brief Benchmark Report - 2nd Case (HDD / R:W=8:2)

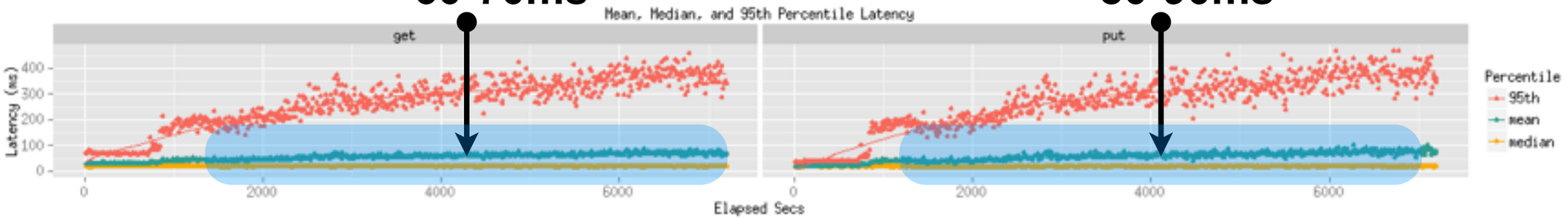


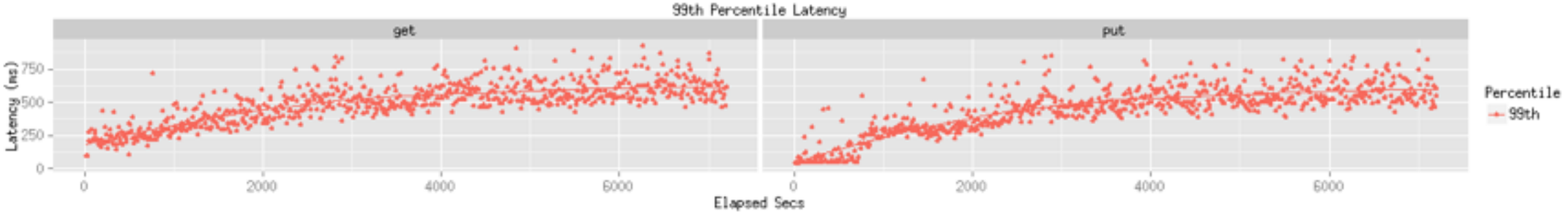
1,000ops

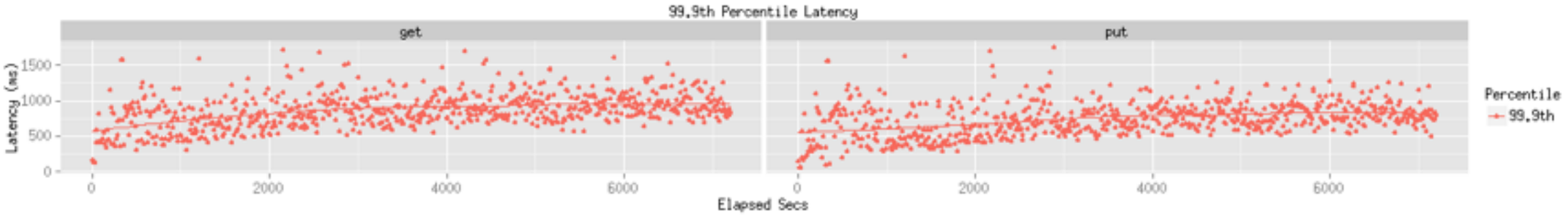
No Errors

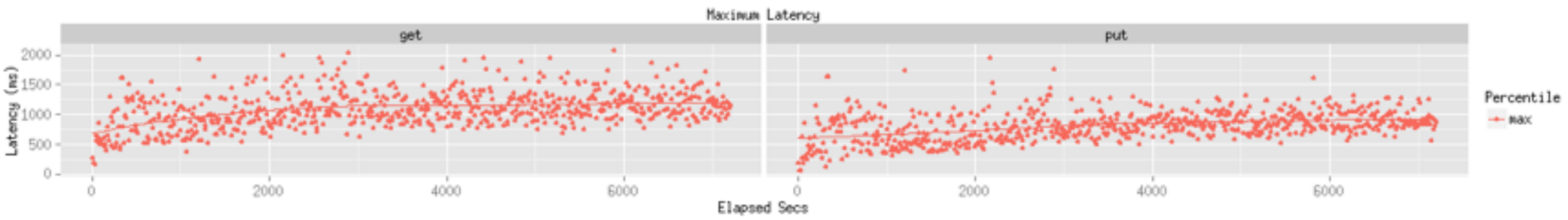
60-70ms

80-90ms









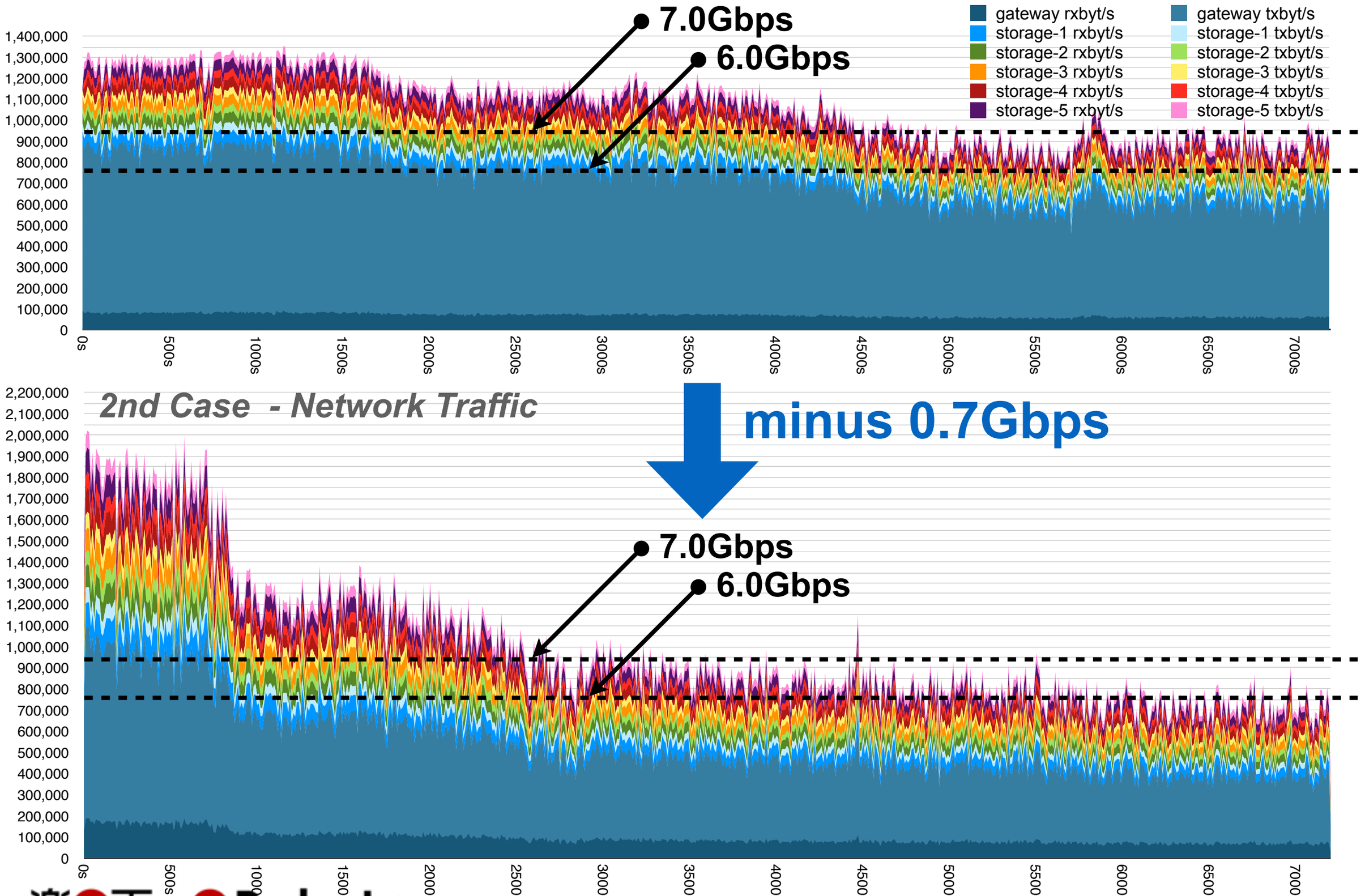
楽天 Rakuten

Compare 1st case with 2nd case

楽R天　RRakuten

Brief Benchmark Report

1st Case - Network Traffic



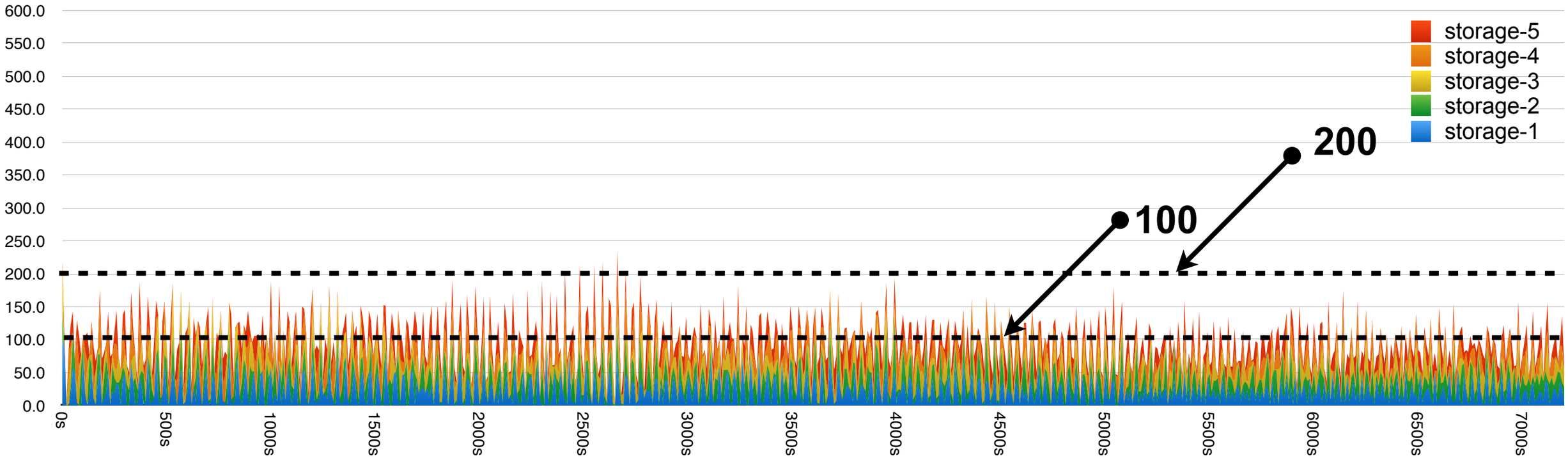
2nd Case - Network Traffic

minus 0.7Gbps

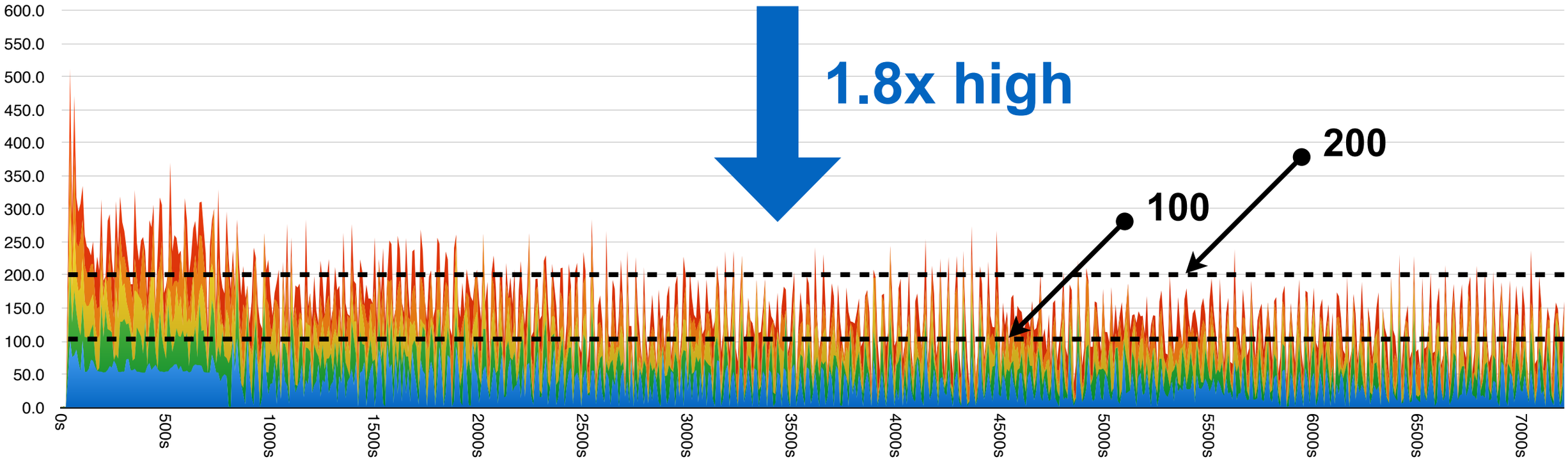
楽天 Rakuten

Brief Benchmark Report

1st Case - Disk util%



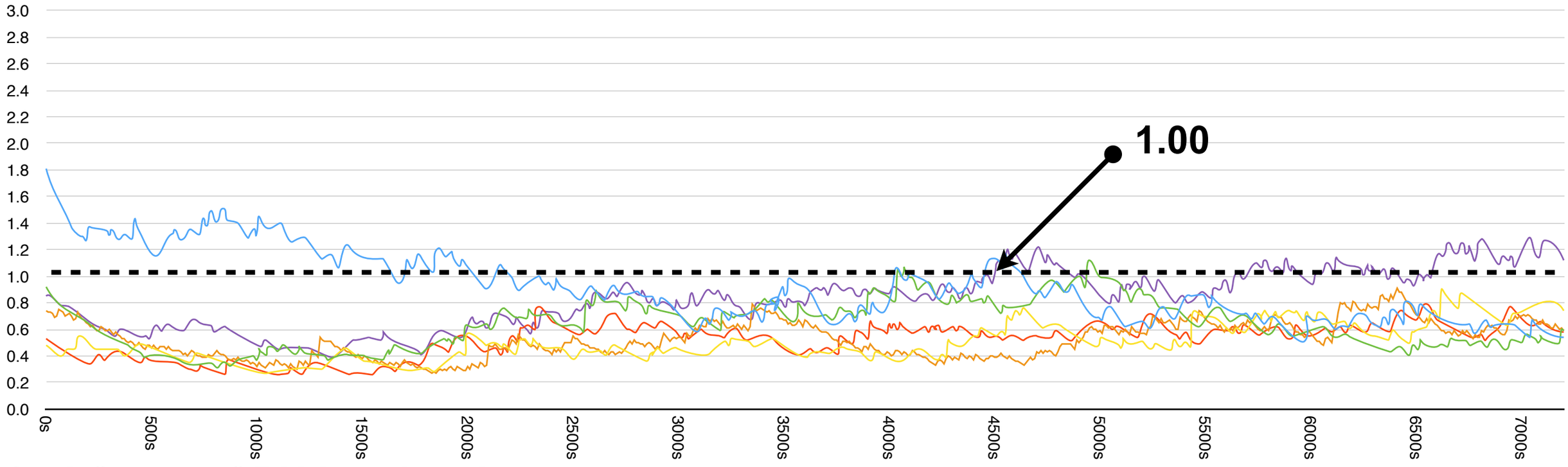
2nd Case - Disk util%



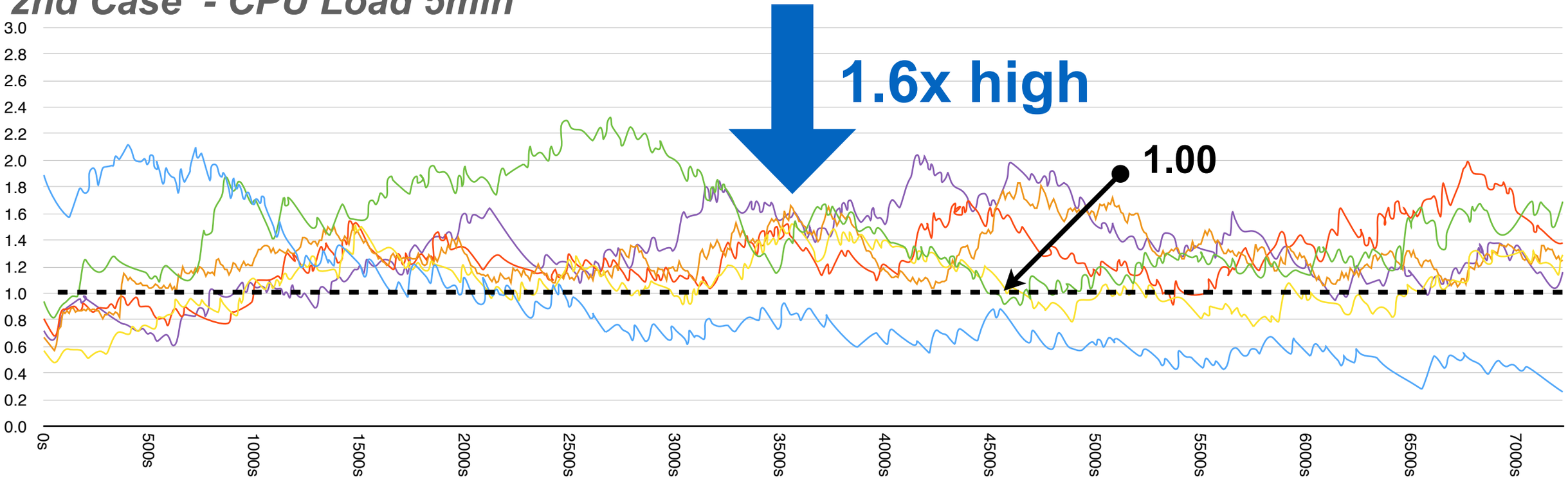
楽天 Rakuten

Brief Benchmark Report

1st Case - CPU Load 5min



2nd Case - CPU Load 5min



楽天 Rakuten

Conclusion:

LeoFS kept in a stable performance through the benchmark

Bottleneck is Disk I/O

The cache mechanism contributed to reduce network traffic between Gateway and Storage

楽R天 RRakuten

Multi Data Center Replication

楽天 Rakuten

Multi Data Center Replication

楽天 Rakuten

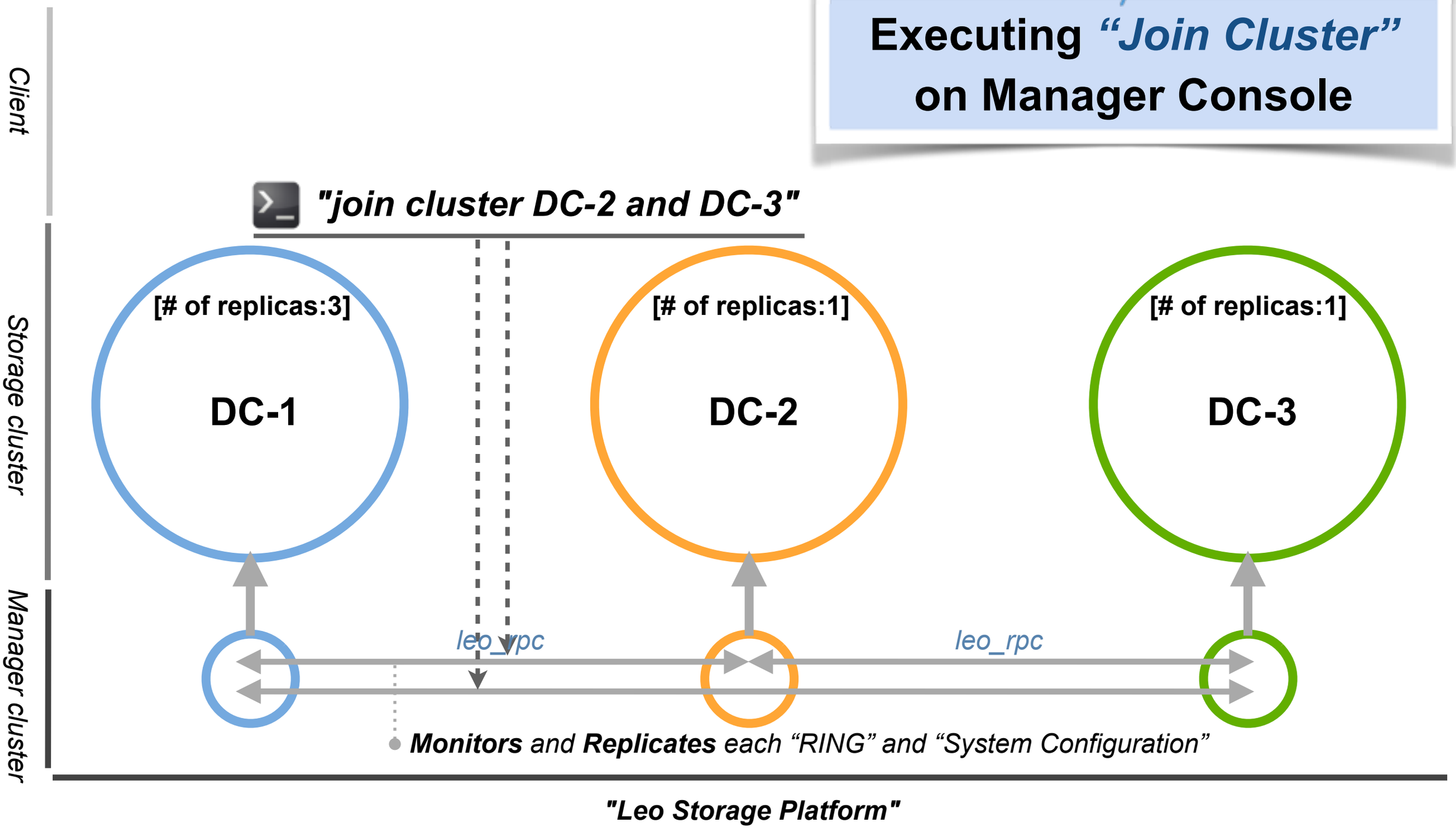
Designed it as simple as possible

1. Easy Operation to build **multi clusters**.
2. **Asynchronous data replication** between clusters

   Stacked data is transferred to remote cluster(s)
3. **Eventual consistency**

Preparing MDC Replication

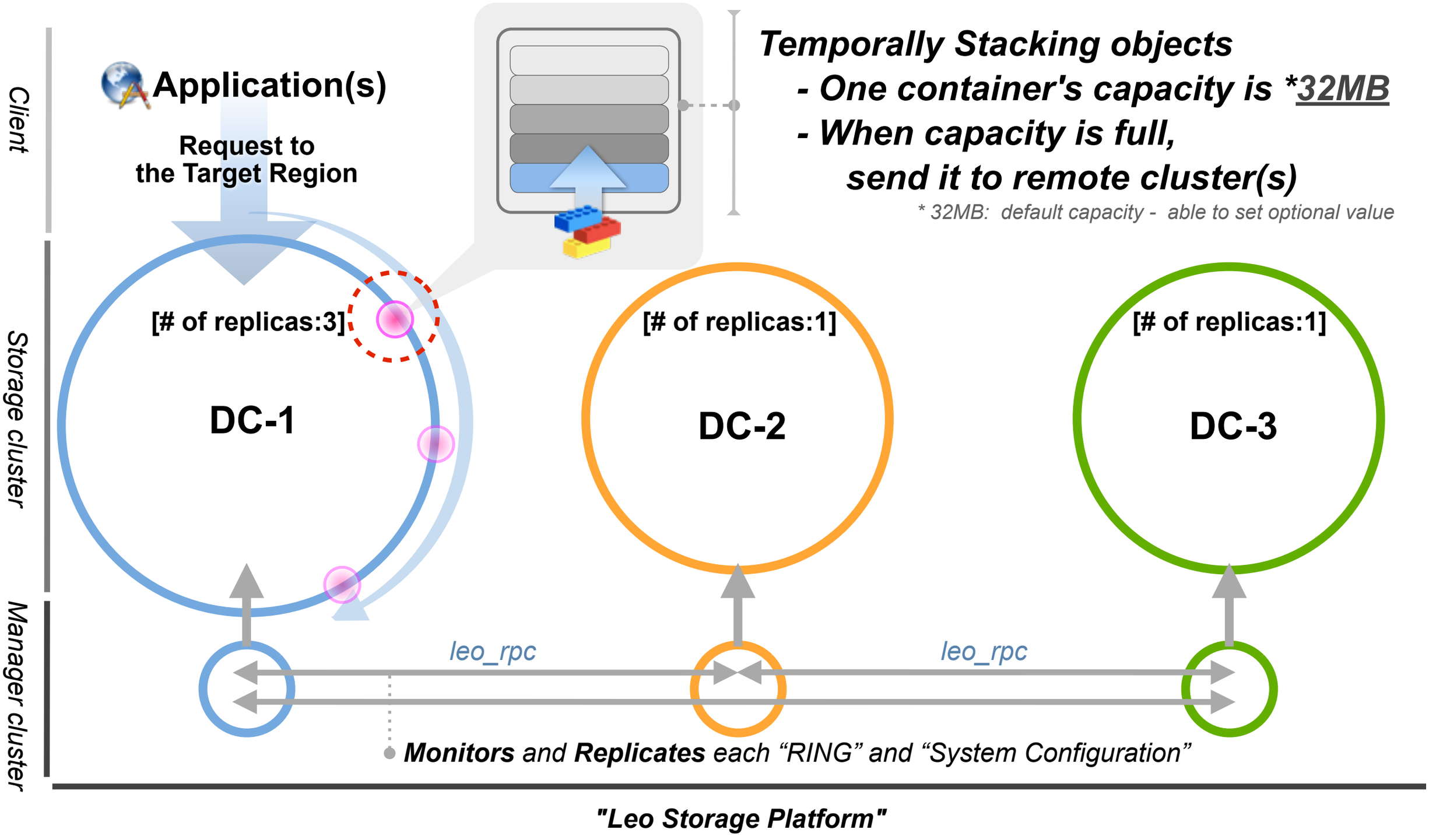
Executing *"Join Cluster"* on Manager Console



楽天 Rakuten

Multi Data Center Replication

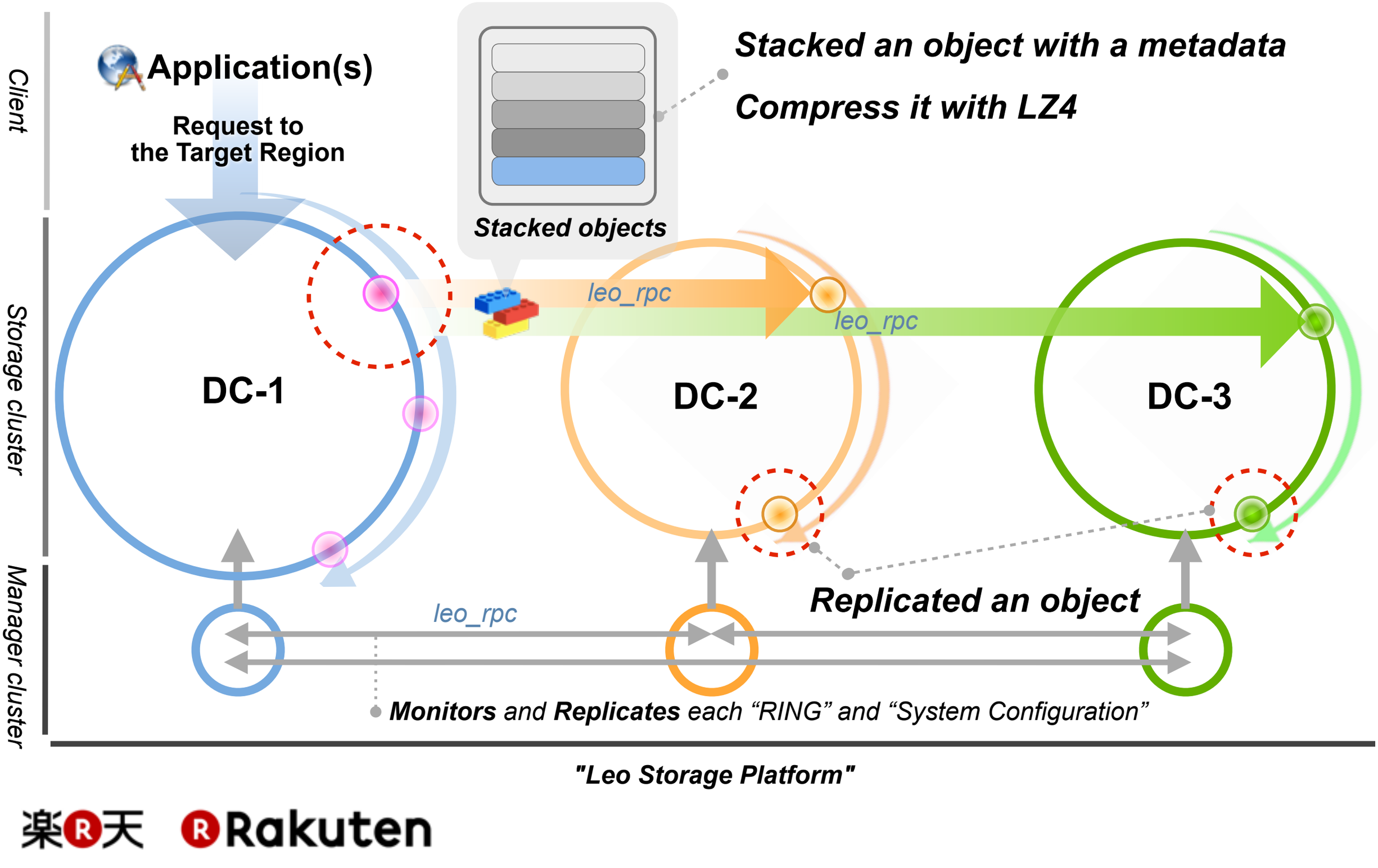
Stacking objects



楽天 Rakuten

Multi Data Center Replication

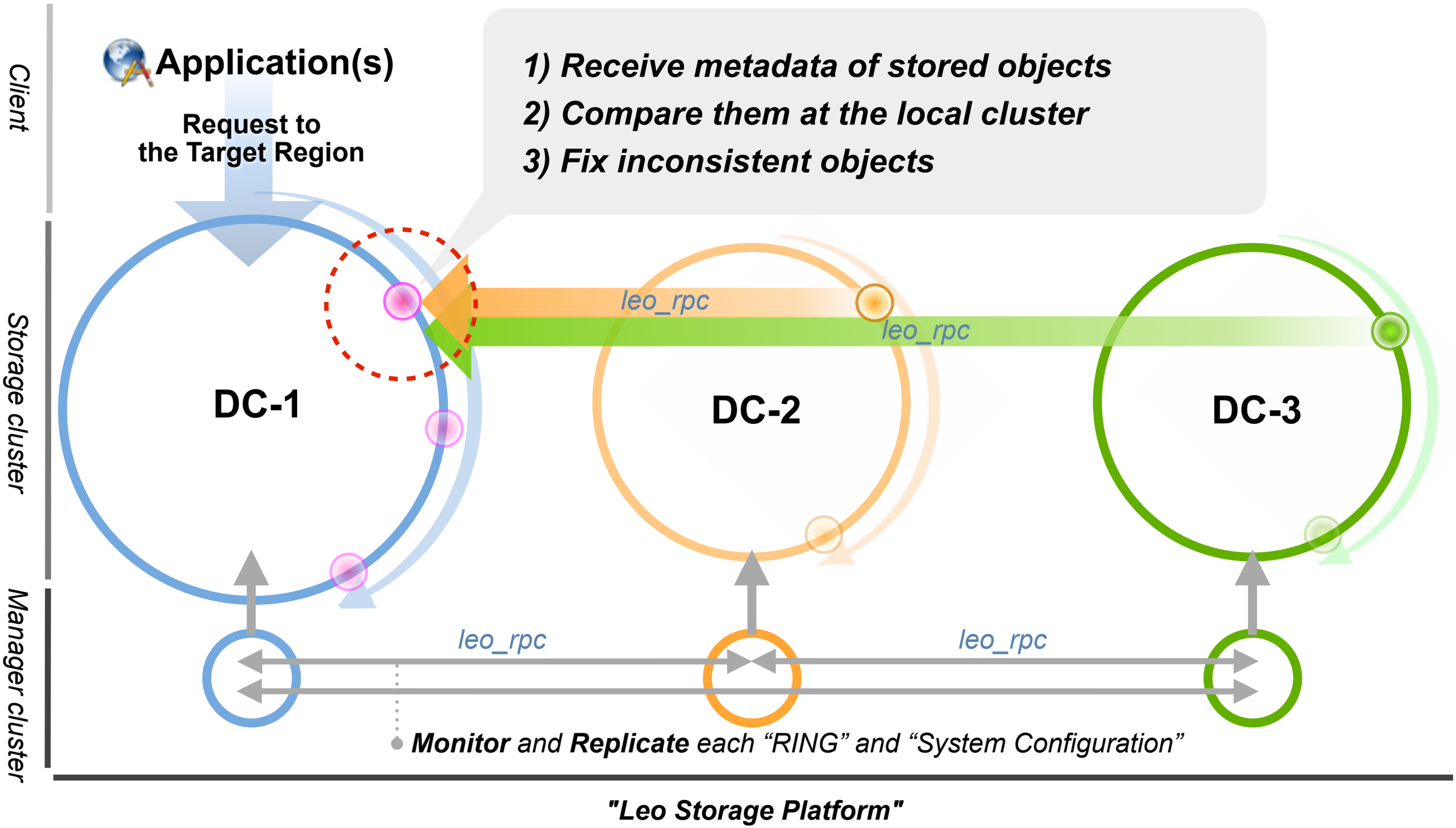
Transferring stacked objects



楽天 Rakuten

Multi Data Center Replication

Investigating stored objects



楽天 Rakuten

LeoFS Administration at Rakuten

Presented by Hiroki Matsue
Rakuten Software Engineer

楽天 Rakuten

LeoFS Administration at Rakuten

Storage Platform

File Sharing Service

Others

Portal Site

Photo Storage

Background Storage of OpenStack

楽天 Rakuten

Storage Platform

楽天 Rakuten

Storage Platform - Scaling the Storage Platform

Reduce Costs

High Reliability

Easy to Scale

S3-API

楽天 Rakuten

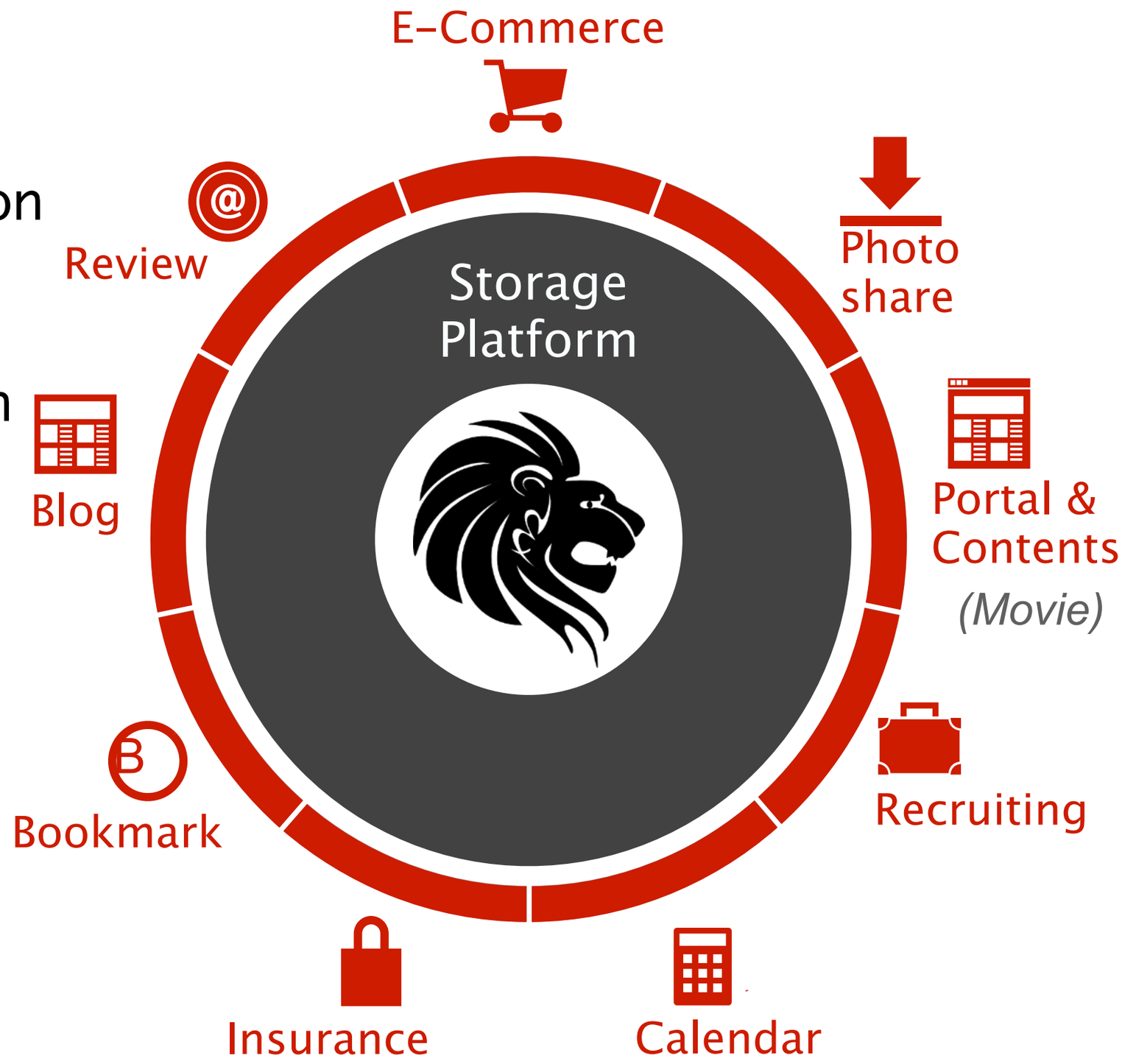
Using Various Services

Total Usage: 450TB

# of Files: 600Million

Daily Growth: 100GB

Daily Reqs: 13Million

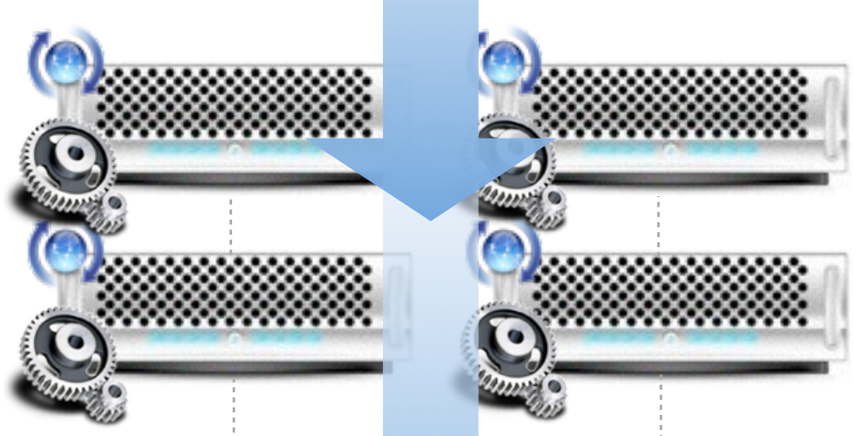


楽天 Rakuten

Storage Platform - System Layout

Requests from
Web Applications / Browsers
w/HTTP over S3-API

Load Balancer / Cache Servers



Gateway x 4

Storage x 14

(Erlang RPC)

(Erlang RPC)

(TCP/IP,SNMP)

Total disk space: 600TB
Number of Files: 600Million
Access Stats:
800Mbps *(MAX)*
400Mbps *(AVG)*

Manager x 2

Nagios
Ganglia

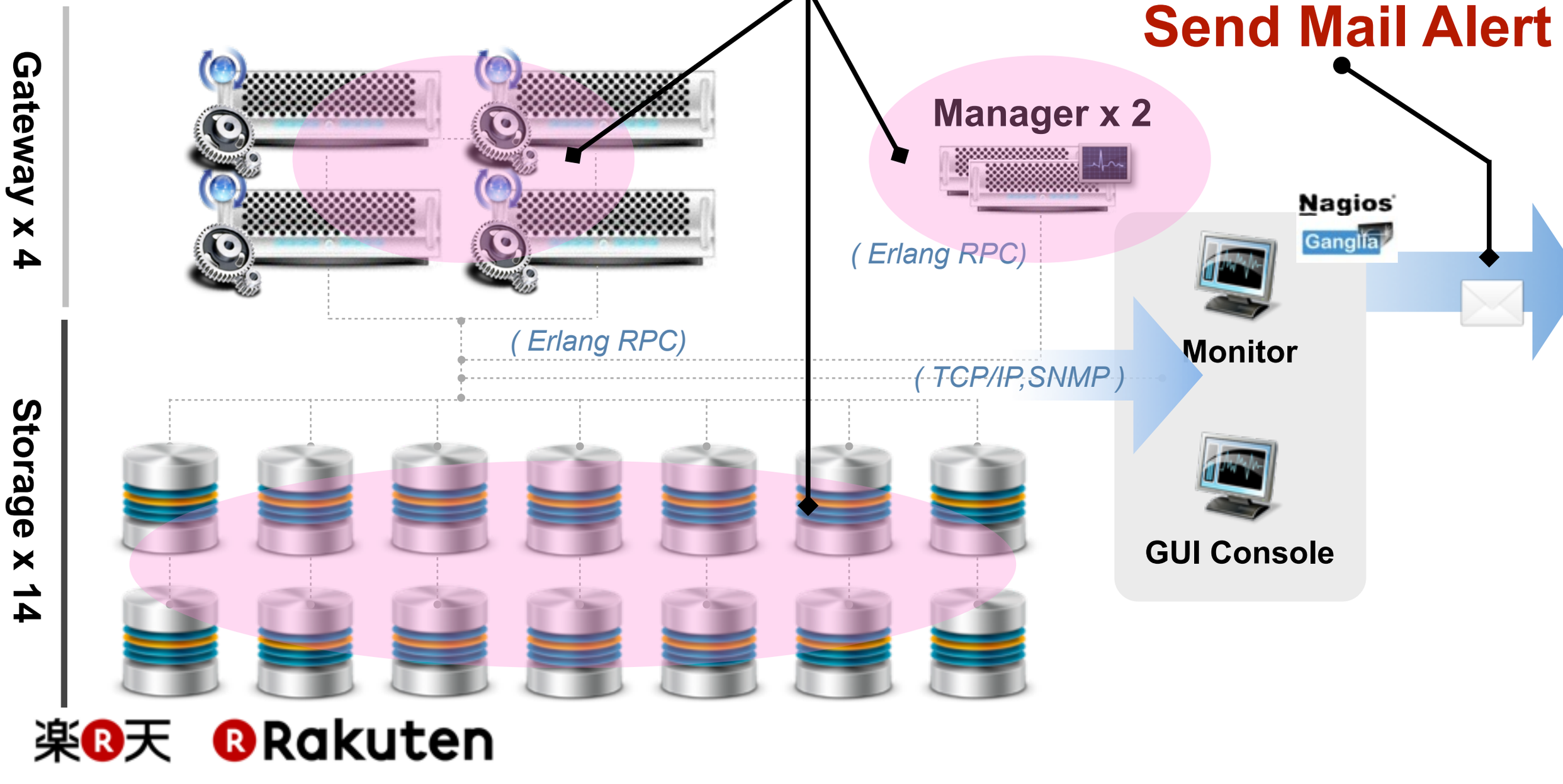
Monitor

GUI Console

楽天 Rakuten

Storage Platform - Monitor

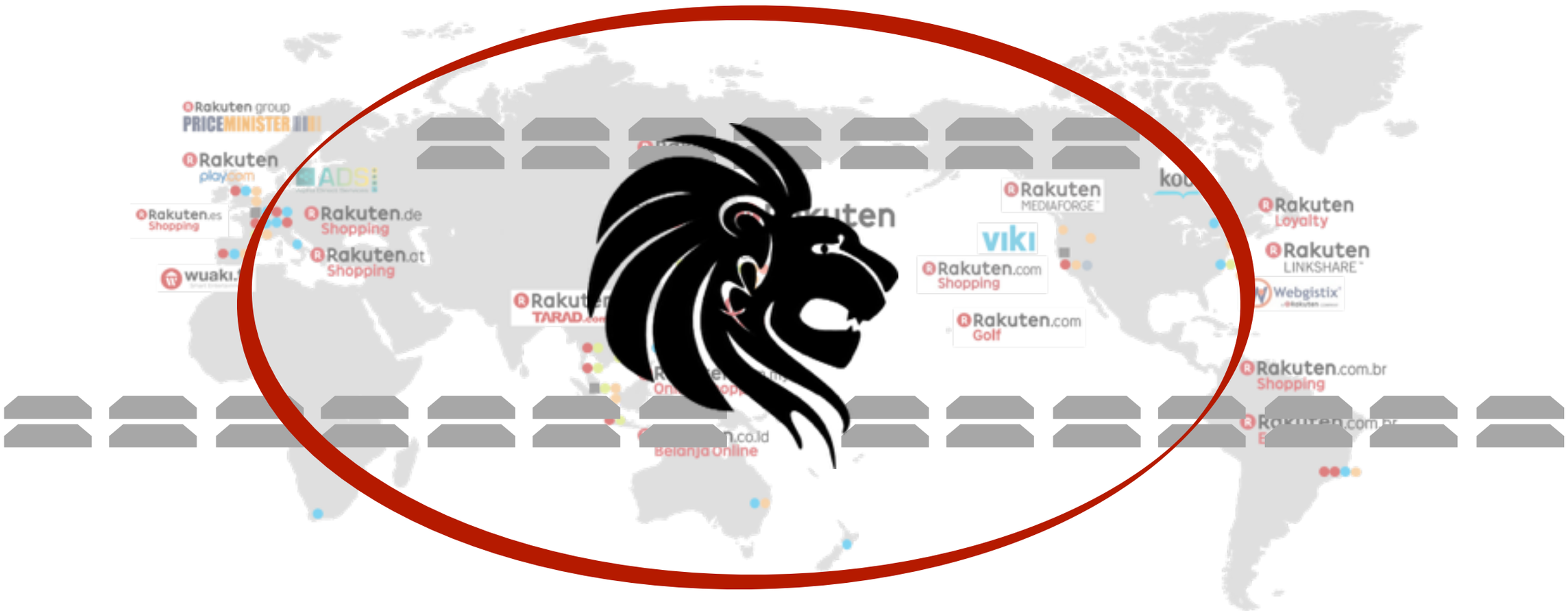
Status Collection *(Ganglia)*
Status Check *(Nagios)*
Port + Threshold Check



楽天 Rakuten

Storage Platform - Spreading Globally

Covering All Services with Multi DC Replication



楽天 Rakuten

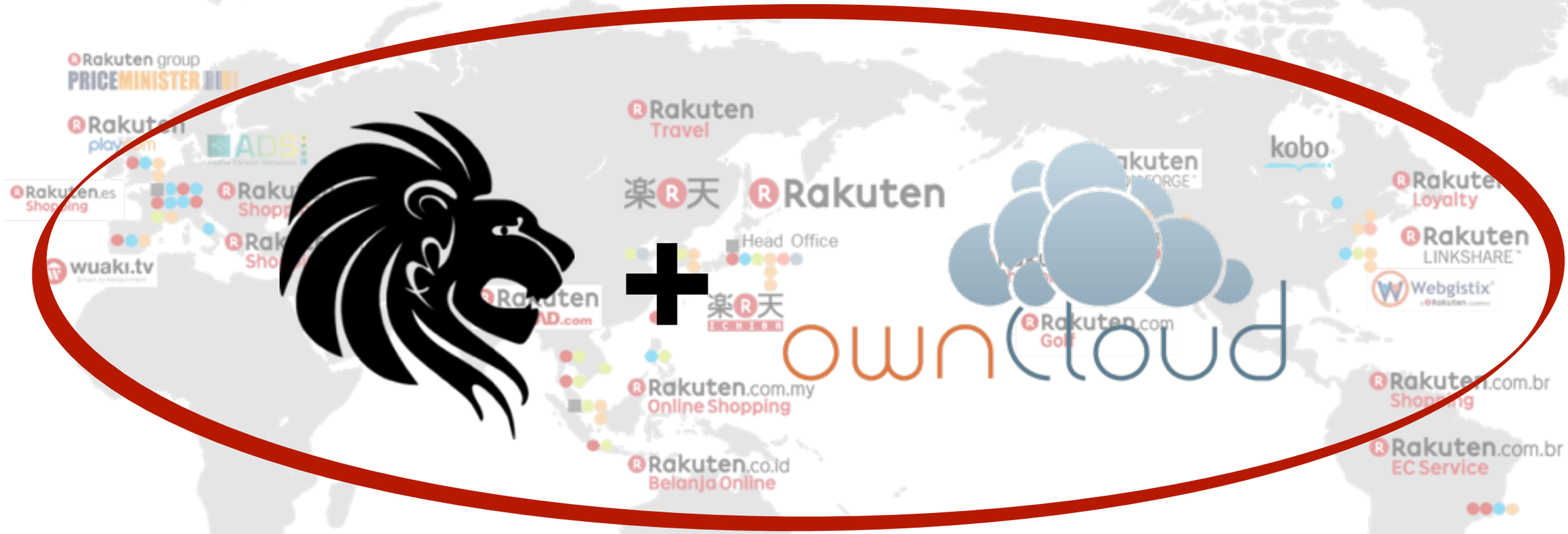
File Sharing Service

+

https://owncloud.com/

楽天 Rakuten

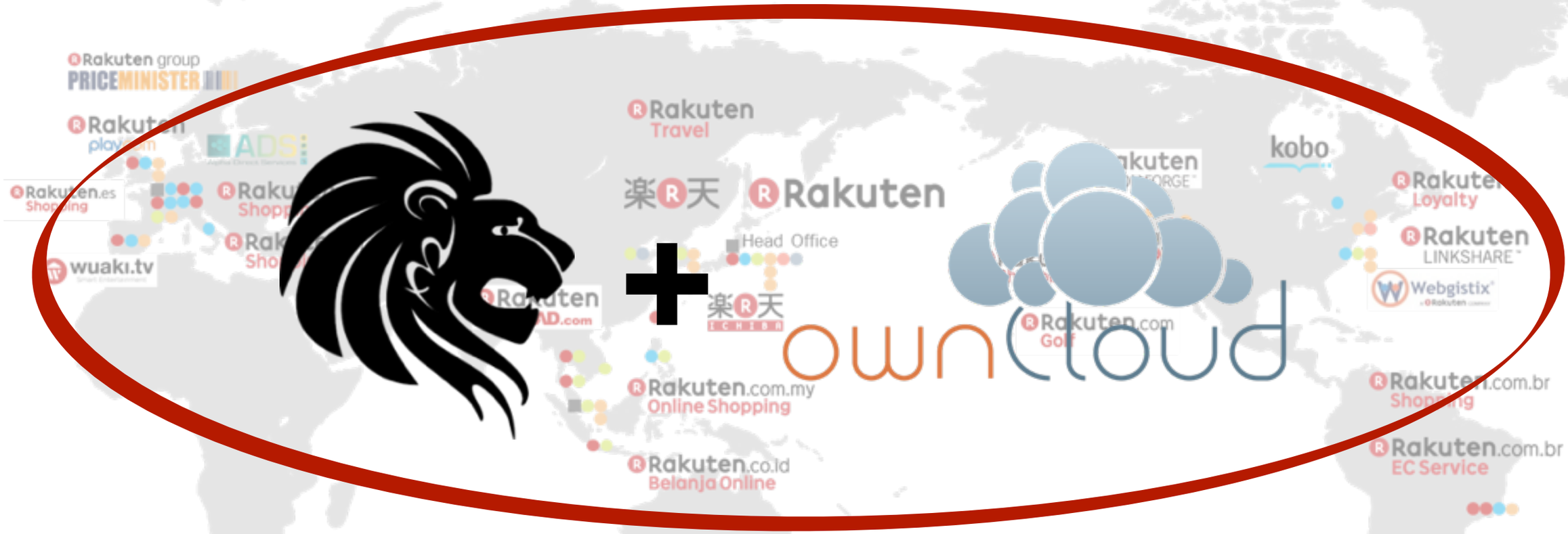
File Sharing Service - Required Targets



Reduce Costs
Handle Confidential Files
Store Large Files
Scale Easily

楽天 Rakuten

File Sharing Service - Usage



Share Docs and Movies with Group Companies

Over 20 Companies, Over 10 Countries

Over 4,000 Users, Over 10,000 Teams

楽天 Rakuten

File Sharing Service - System Layout

Web GUI File Browser

ownCloud

Authenticate Users

LDAP

MariaDB

Manager x 2

(Erlang RPC)

(TCP/IP,SNMP)

Monitor

GUI Console

ROMA

Manage Configurations

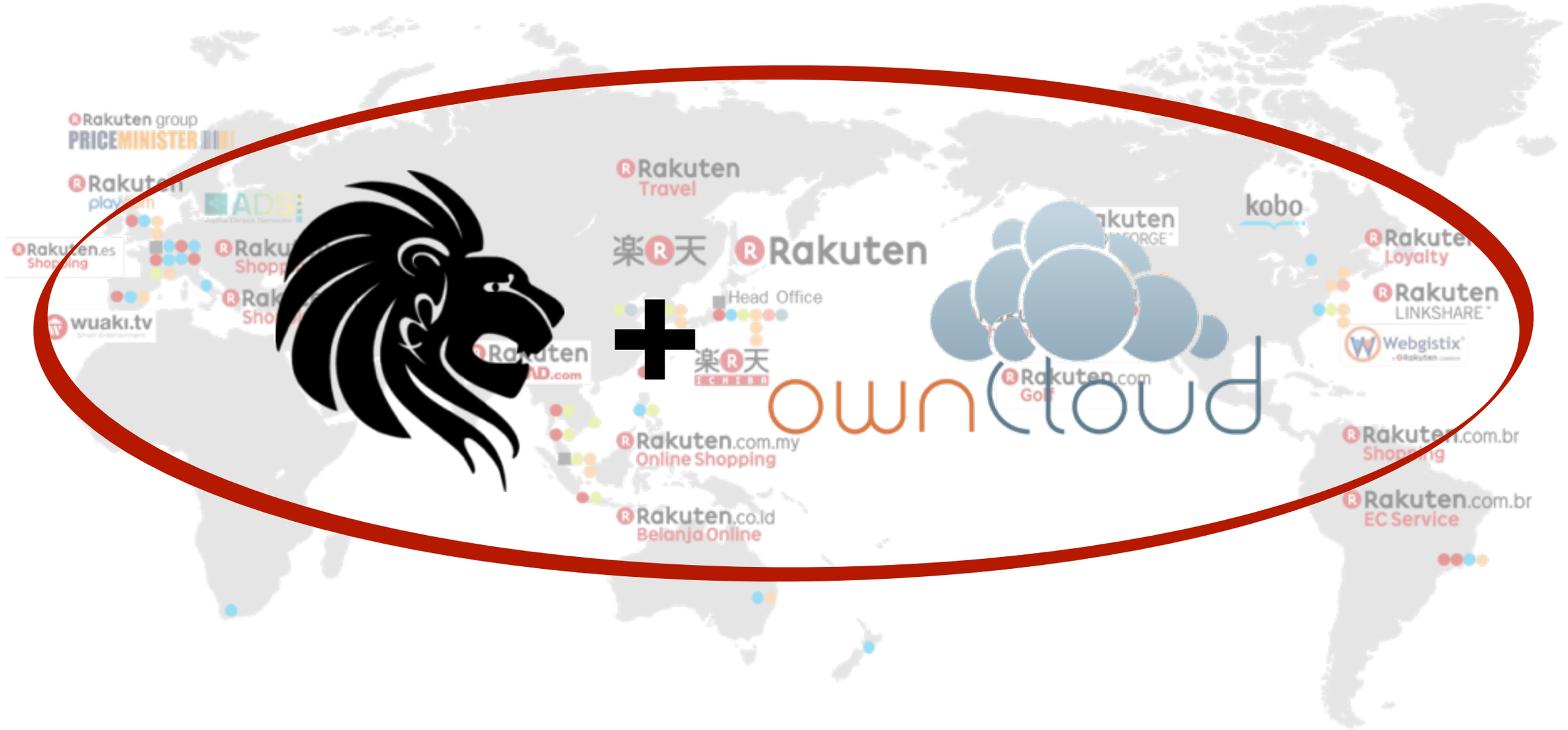
Manage Login Session (KVS)

楽天 Rakuten

File Sharing Service - Future Plans

Cover 25 Countries/Regions

Over 20,000 Users



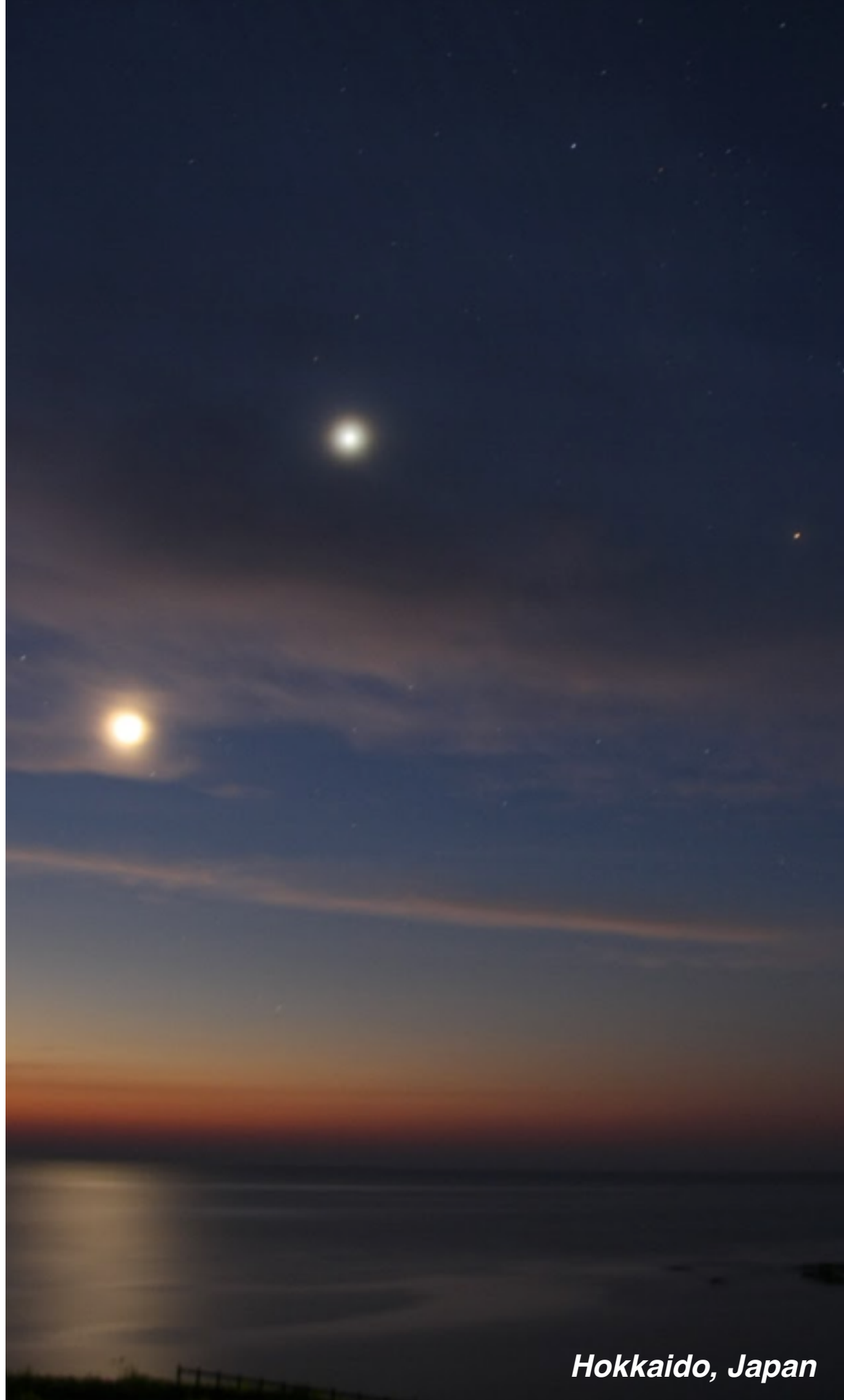
楽天 Rakuten

Empowering the Services and the Users Through the Cloud Storage

楽天　Rakuten

Future Plans

楽天 Rakuten

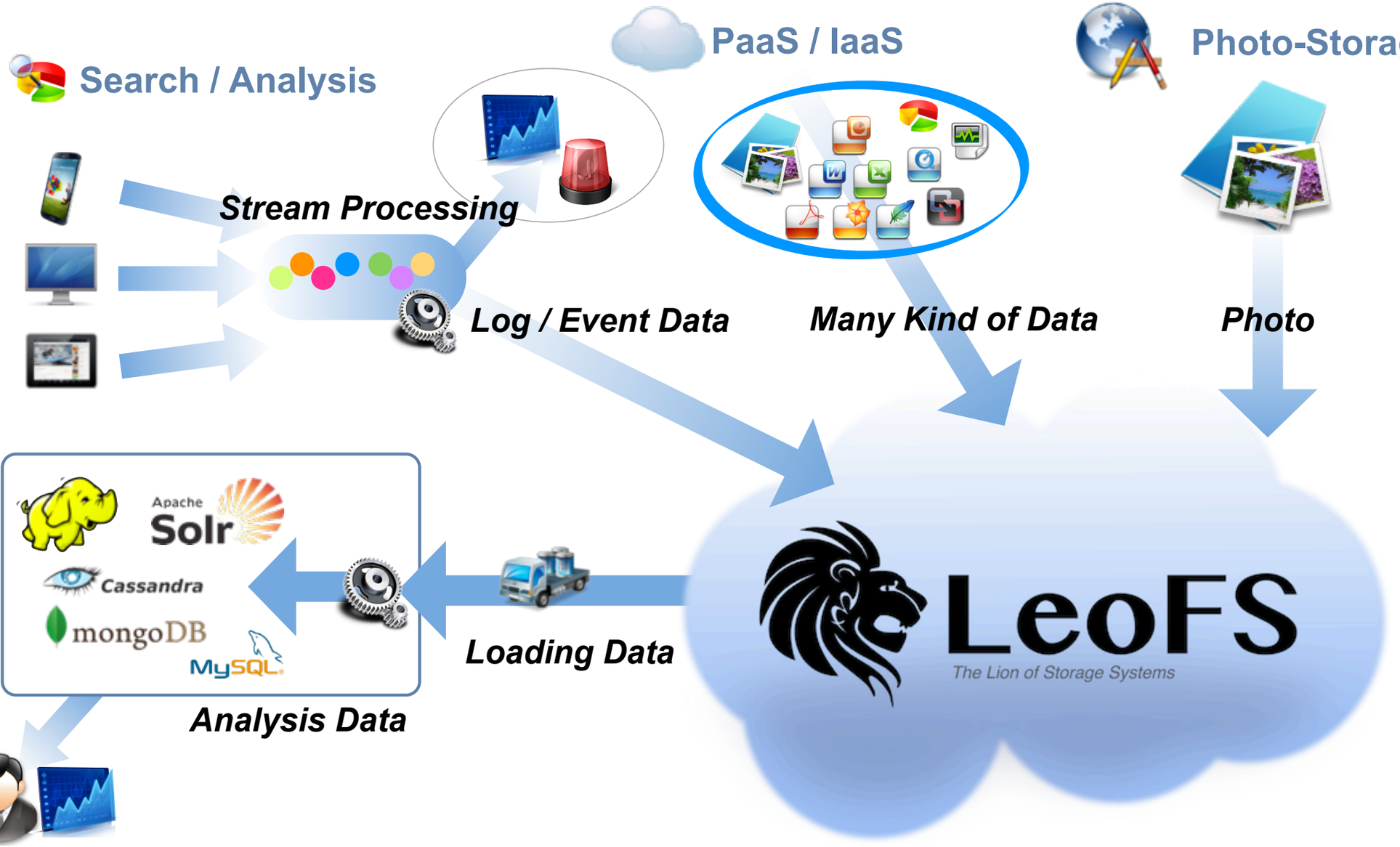


Hokkaido, Japan

Future Plans

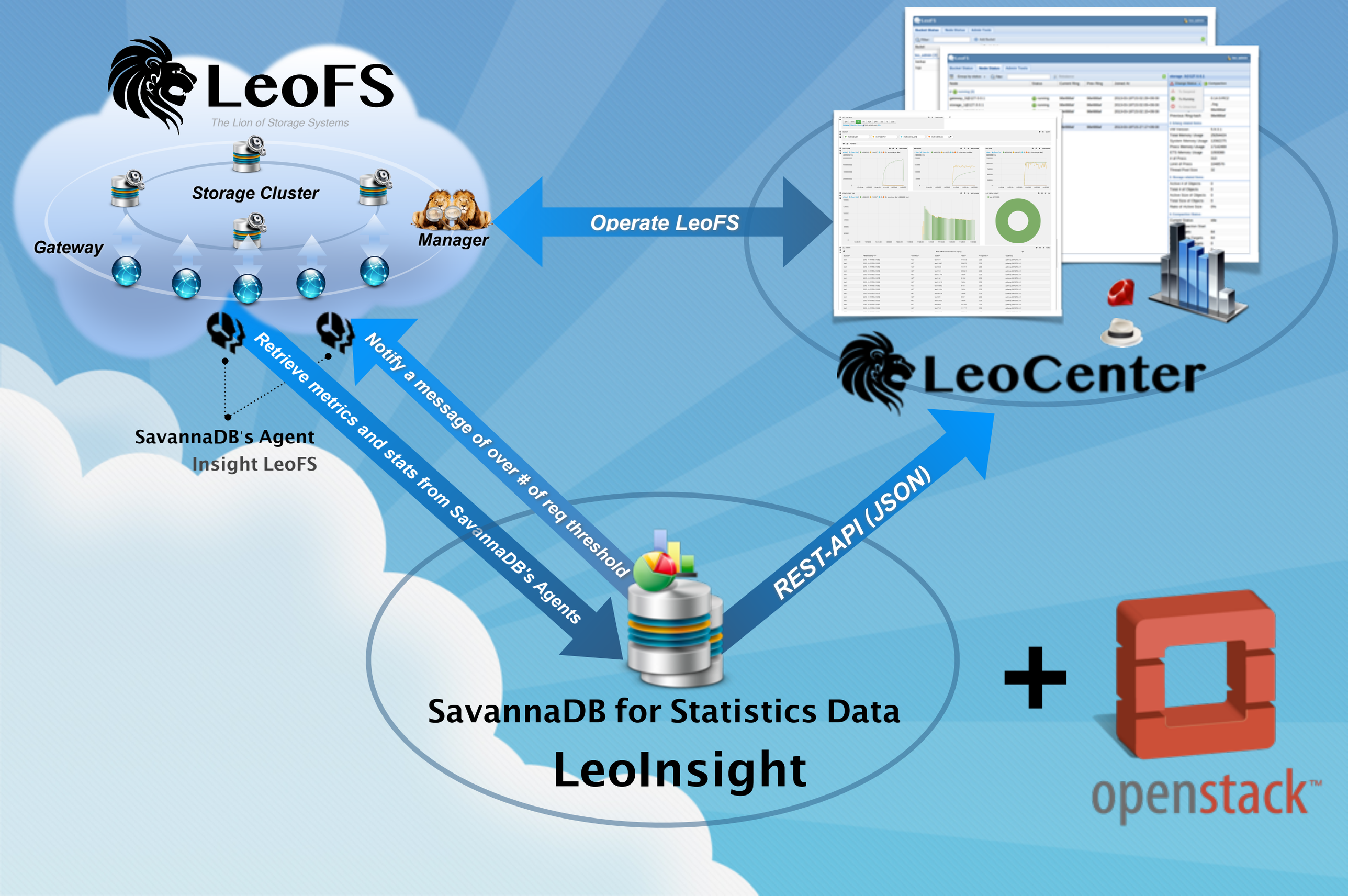
NFS Support

Data-HUB: Centralize unstructured data in LeoFS



楽天 Rakuten

Future Plans



Set Sail for “Cloud Storage”

Website: leo-project.net
Twitter: @LeoFastStorage
Facebook: www.facebook.com/org.leofs