

Location-Aware Distributed Virtual Disk Storage for OpenStack

Keiichi SHIMA <keiichi@ijlab.net>
IIJ Innovation Institute Inc.

CloudOpen Europe 2014
2014-10-15

Note Well

- This proposal is **NOT** an autonomously distributed storage solution

Background

- Widely spread virtualization deployment
- Integrated distributed datacenter
- Flexible resource management

Virtualization

- There is no service provider who doesn't use virtualization
- Many Internet services are now built on top of virtualized computers
- Easy to deploy, scale out when necessary, and scale down when shutting down services

Integrated Datacenter

- Datacenters become bigger and bigger
 - Aiming to reduce operation cost
 - Single point vulnerability
- Approaches to distributed datacenters
 - Geographically distributed and widely integrated

Resource Management

- Resources for virtual infrastructures
 - CPU and memory
 - Network
 - Storage
- We need to locate or relocate these resources flexibly

So far

- CPU and memory relocation
 - Most of the virtualization platforms support this
- Network relocation
 - Software defined networking
 - Intensively being researched and developed

Storage Resource

- Storage resource relocation
 - still not in operation status

Current Status

- Integrated
 - NFS or iSCSI storage devices
- Distributed
 - Ceph (RBD)
 - GlusterFS
 - Sheepdog

Current Status

- Integrated solution
 - Single point of failure (typically)
 - Difficult to relocate resources

Current Status

- Distributed approaches
 - Difficult to adapt non-uniform environment
 - Administration issues (in the sense of a resource location control)

Idea is

- To use only good parts of the integrated and distributed solutions

Objectives

- Flexibility
- Redundancy
- Locality

Flexibility

- Flexible management of location
 - Place the data at the specified location
 - not at somewhere in a storage cluster
- Operation unit
 - Per virtual disk
 - Per block of a virtual disk
- Flexibility in relocation timing
 - Try to prevent to disturb other traffic

Redundancy

- Replication of a virtual disk
 - Even more than two replicas
 - Dynamic operation of replication

Locality

- Locate the actual data of a virtual disk as near as possible to the virtual machine using it

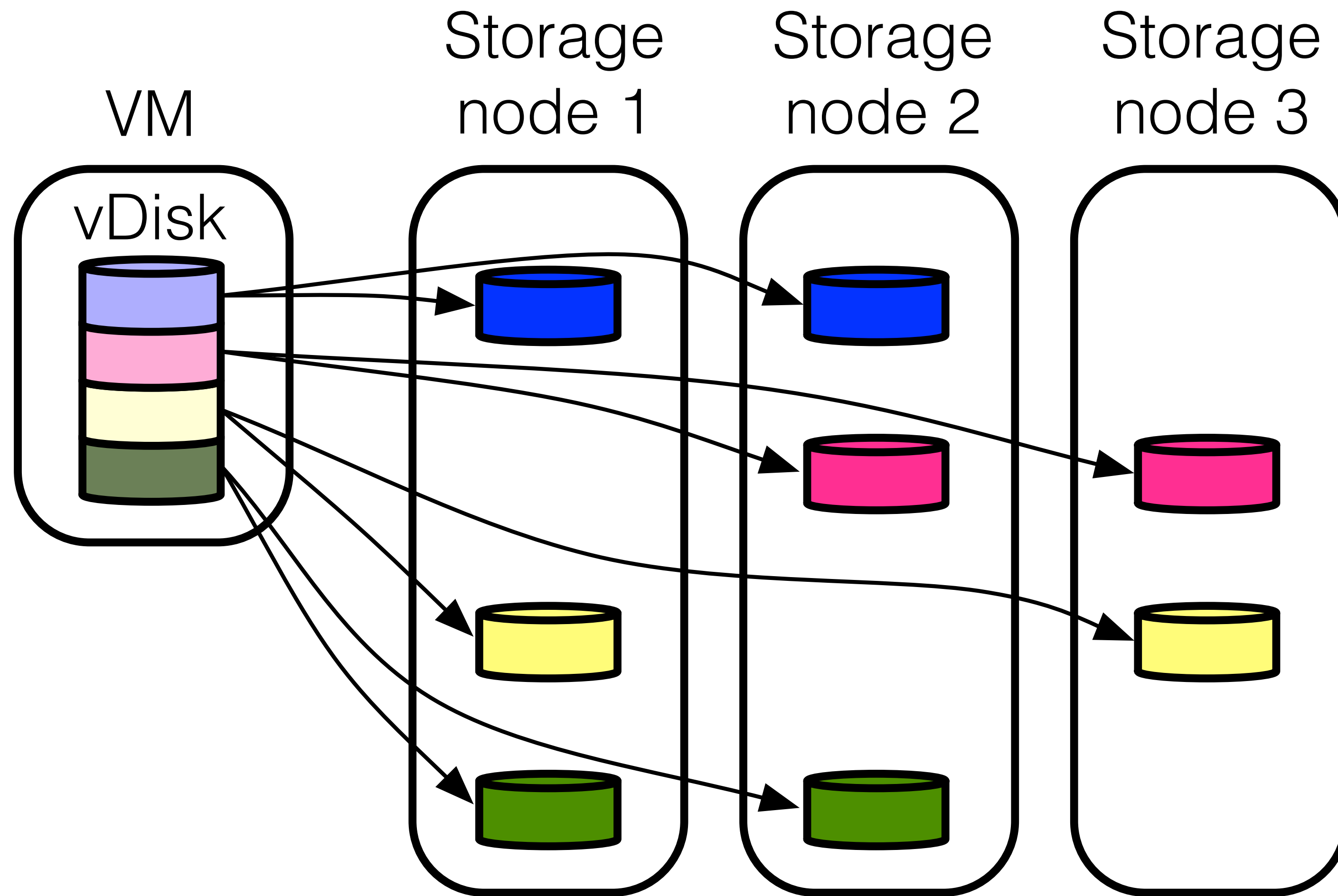
Operation Scenarios

- Consolidation
 - Consolidate virtual machines to minimize the # of running hypervisors
- Move
 - Push virtual machines out from a hypervisor to upgrade the hypervisor
 - Move virtual machines from one DC to another DC, e.g. due to new DC launch, or discontinue
- Efficient resource usage
 - Use best performing pair of hypervisors and storage nodes

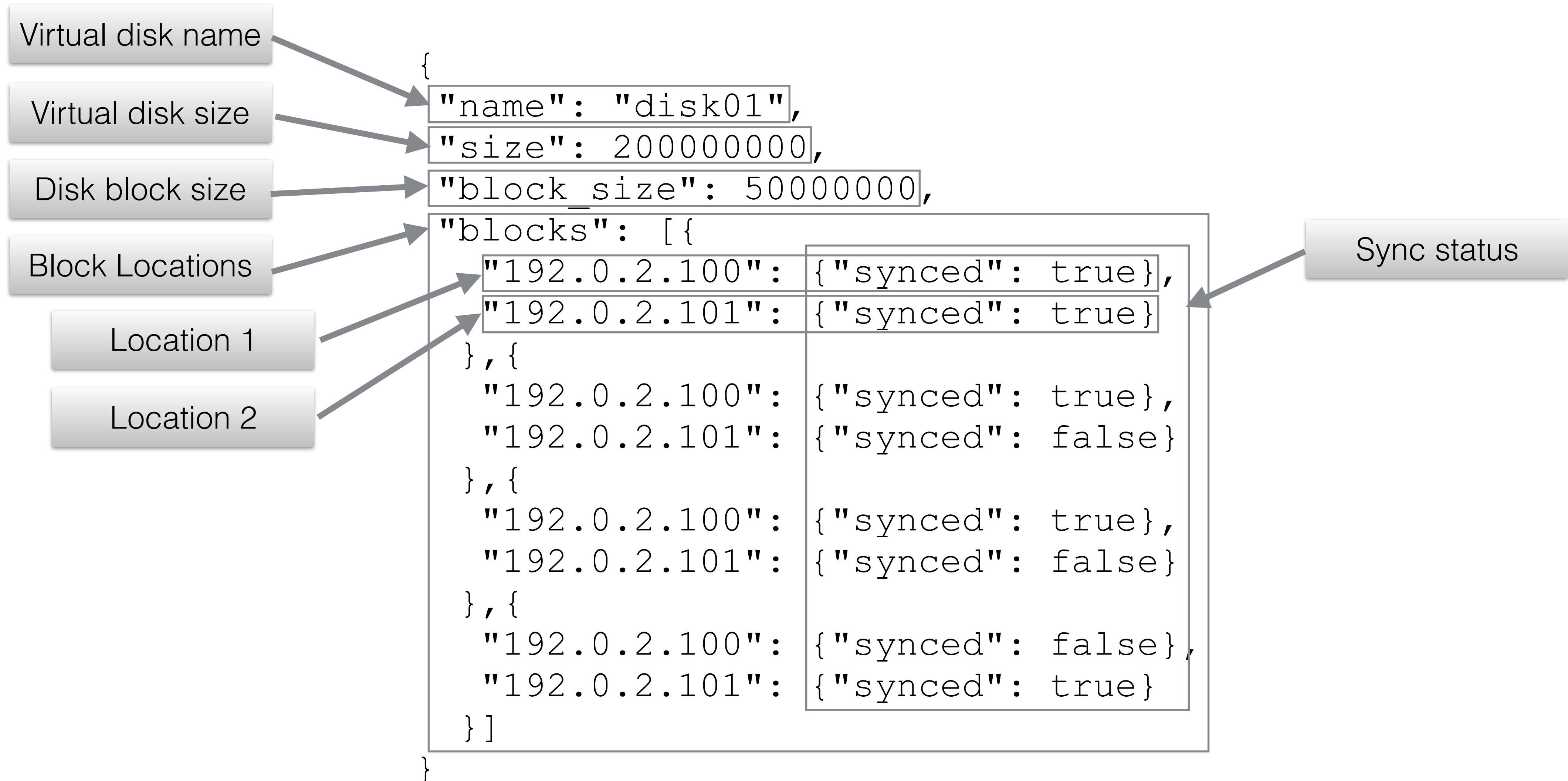
UKAI

- A virtual storage device for virtual machines
 - Location control per block
 - Replication operation per block
 - No global lock required
 - Since a virtual storage device is owned by only one virtual machine at the same time

UKAI Virtual Disk Concept



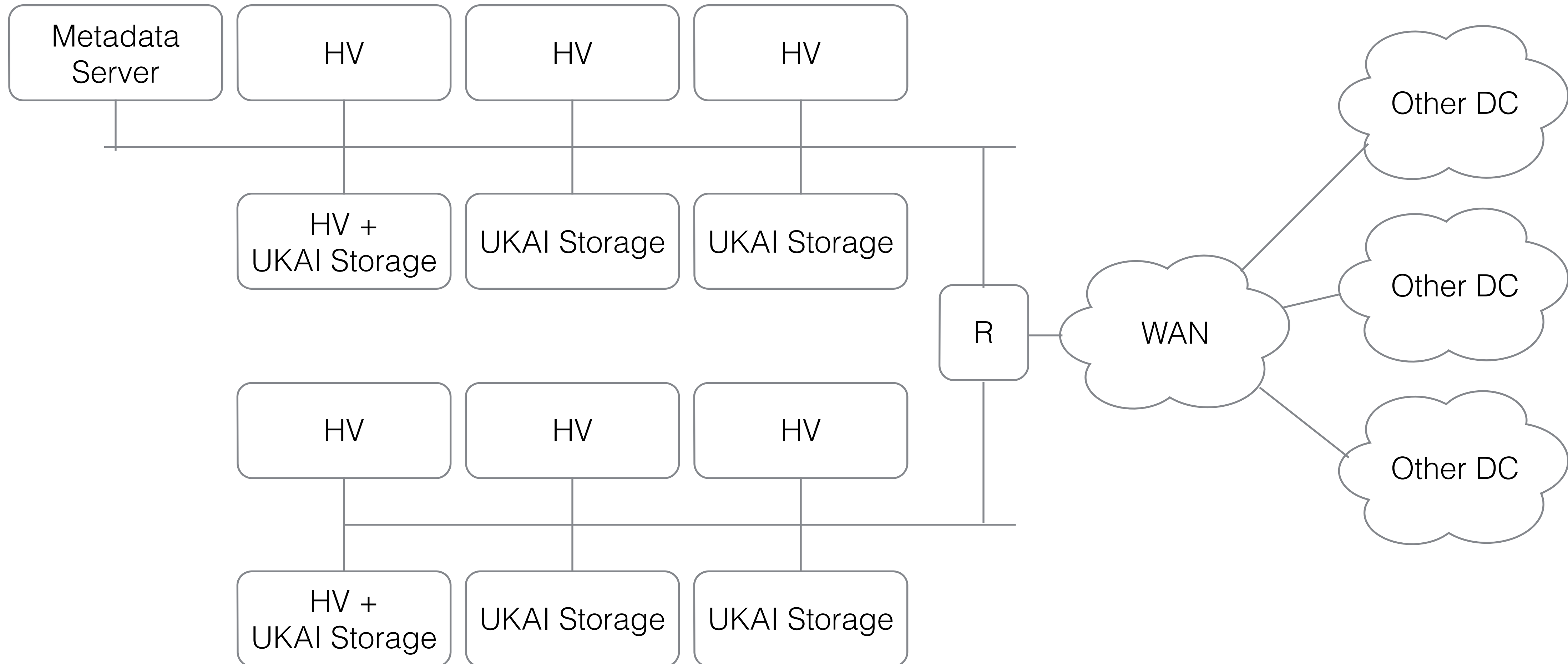
Metadata Structure



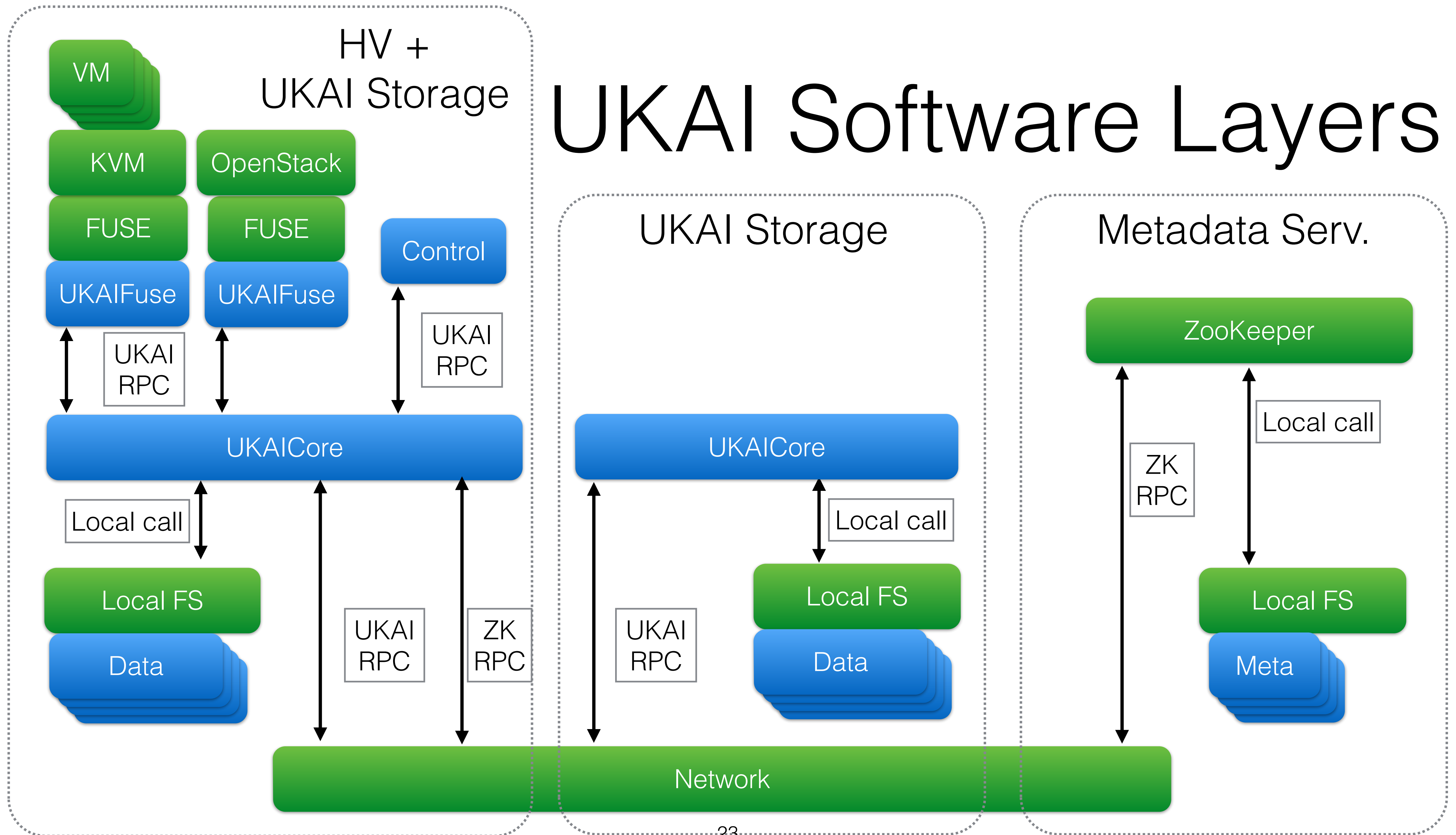
Metadata Server

- The server nodes keeping virtual data disk structure information
- Apache ZooKeeper is used

Typical Node Layout



UKAI Software Layers



Supported Disk Admin Ops

- Create a disk image / Destroy a disk image
 - Size and block size is configurable
- Add location / Delete location
 - Per block granularity
- Synchronize
 - Per block granularity
- Misc (e.g. getting image metadata info)

OpenStack Integration

- Compute (Nova)
 - A new virtual disk mount adapter module for UKAI is required
- Block storage (Cinder)
 - A new storage resource management module for UKAI is required
- Object storage (Swift)
 - Out of scope of this proposal

Demo

Hurdles

- Performance
 - At this moment, worse than NFS
 - Possible reasons: FUSE, RPC, no-caching
- Management facility
 - 100% manual operation is not always required
 - Need some support tools for node selection based on the predefined hypervisor/storage locations

Summary

- We need a new storage backend system to support distributed and integrated virtual infrastructure operation
- UKAI: a location-aware distributed storage system for virtual machines
 - Flexible location management, redundancy, and locality are provided
- OpenStack integration is possible
- Need more development efforts to get better performance

Availability

- UKAI
 - <https://github.com/keiichishima/ukai>
- OpenStack support
 - The UKAI branch in <https://github.com/keiichishima/nova>
 - The UKAI branch in <https://github.com/keiichishima/cinder>

Questions?

- UKAI
 - <https://github.com/keiichishima/ukai>
- OpenStack support
 - The UKAI branch in <https://github.com/keiichishima/nova>
 - The UKAI branch in <https://github.com/keiichishima/cinder>