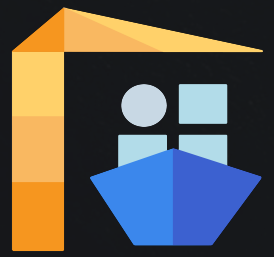
LINBIT DaoCloud



100% Opensource

Piraeus Datastore

Webinar Agenda

1. Core Technologies behind Piraeus
2. About Cloud Native Storage
3. Project Piraeus: Dynamic Provisioning, Resource Management and HA for Local Volumes
4. Demo

DaoCloud

Piraeus Datastore

Core Technologies

Speaker: Philipp Reisner

Philipp Reisner

Founder and CEO of LINBIT in Vienna/Austria.

His professional career has been dominated by developing **DRBD**, a storage replication for Linux. While in the early years (2001) this was literally writing kernel code.

Today he leads a company of about 30 employees with locations in Vienna/Austria and Portland/Oregon with an open source business model offering support subscriptions to customers around the globe.

Email: philipp.reisner@linbit.com

Leading Open Source OS based SDS

LINBIT

COMPANY OVERVIEW

- **Developer** of **DRBD and LINSTOR**
- **100% founder owned**
- Offices in **Europe and US**
- Team of **highly experienced Linux experts**
- Exclusivity **Japan: SIOS**

PRODUCT OVERVIEW

- **Leading Open Source Block Storage** (included in Linux Kernel (v2.6.33)
- **Open Source DRBD** supported by proprietary LINBIT products / services
- **OpenStack** with **DRBD Cinder driver**
- **Kubernetes Driver**
- **Install base of** >2 million

DRBD

LINBIT
DR HA SDS

LINSTOR | PACEMAKER

RDMA MODULE | DRBD PROXY

Support, Consulting, Training

REFERENCES

Google SanDisk T-Mobile IBM MICRO FOCUS

amazon.com intel at&t CISCO NTT redhat

ORACLE ERICSSON YAHOO! SkySQL Mellanox TECHNOLOGIES

SOLUTIONS

DRBD Software Defined Storage (SDS)

New solution (introduced 2016)

Perfectly suited for SSD/NVMe high performance storage

DRBD High Availability (HA), DRBD Disaster Recovery (DR)

Market leading solutions since 2001, over 600 customers

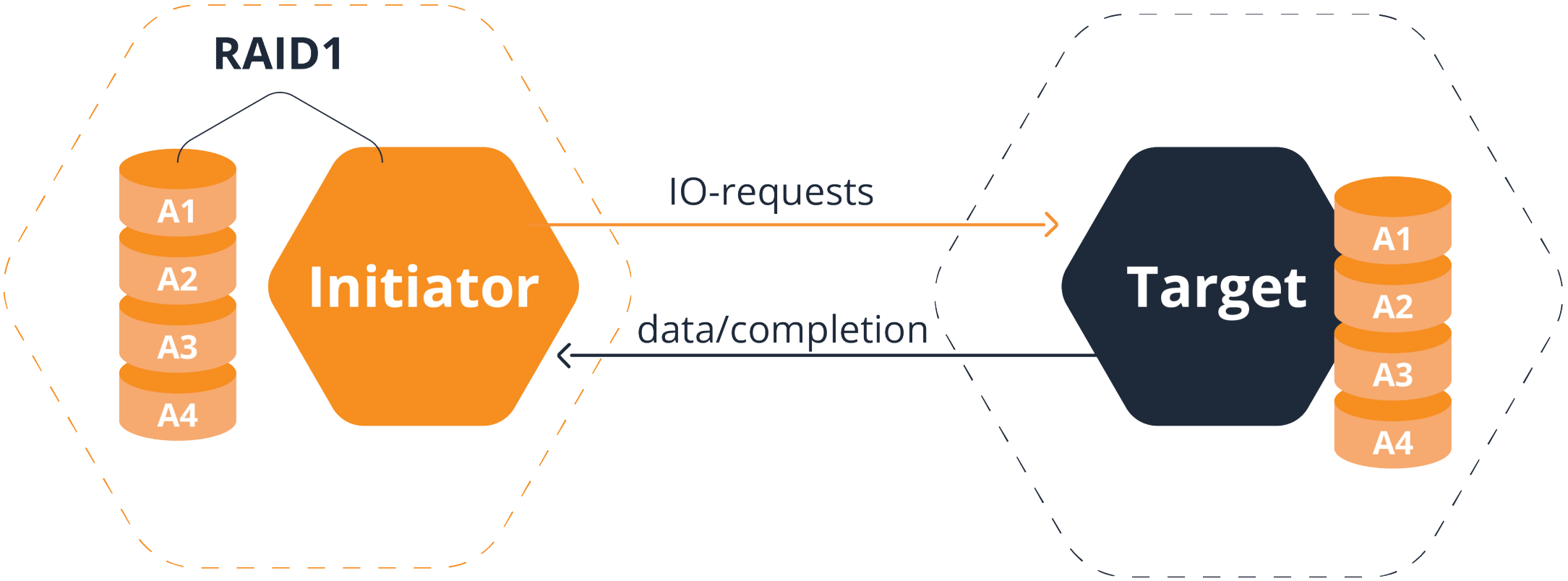
Ideally suited to power HA and DR in OEM appliances (Cisco, IBM, Oracle)

LINBIT

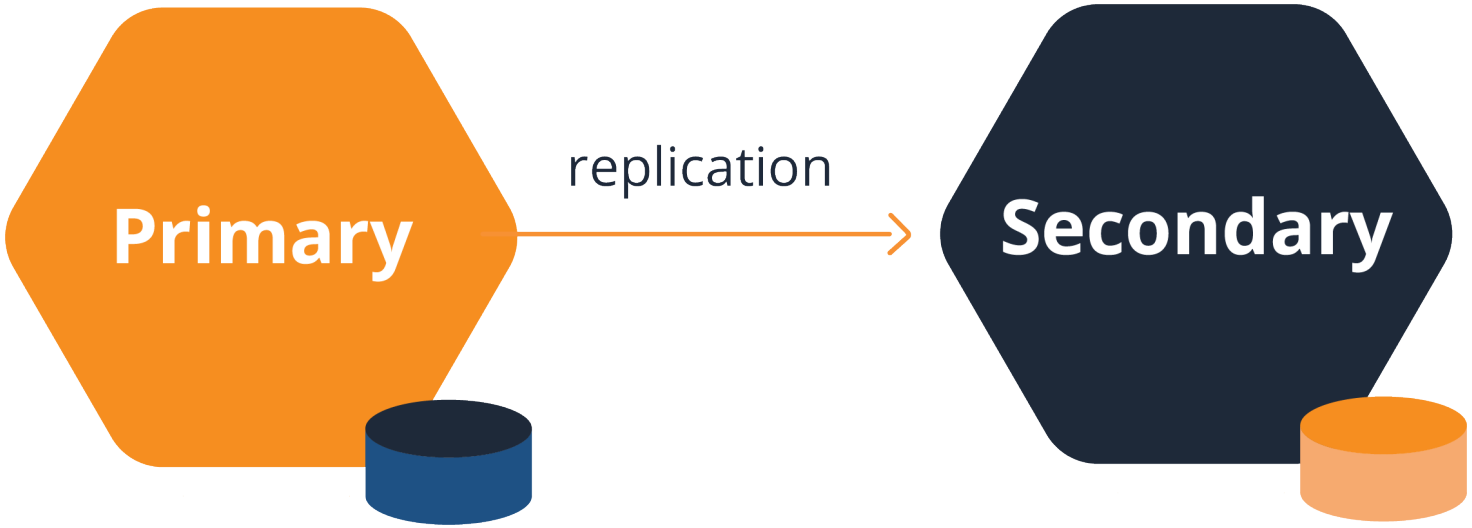
DRBD

Put in simplest form

DRBD – think of it as ...

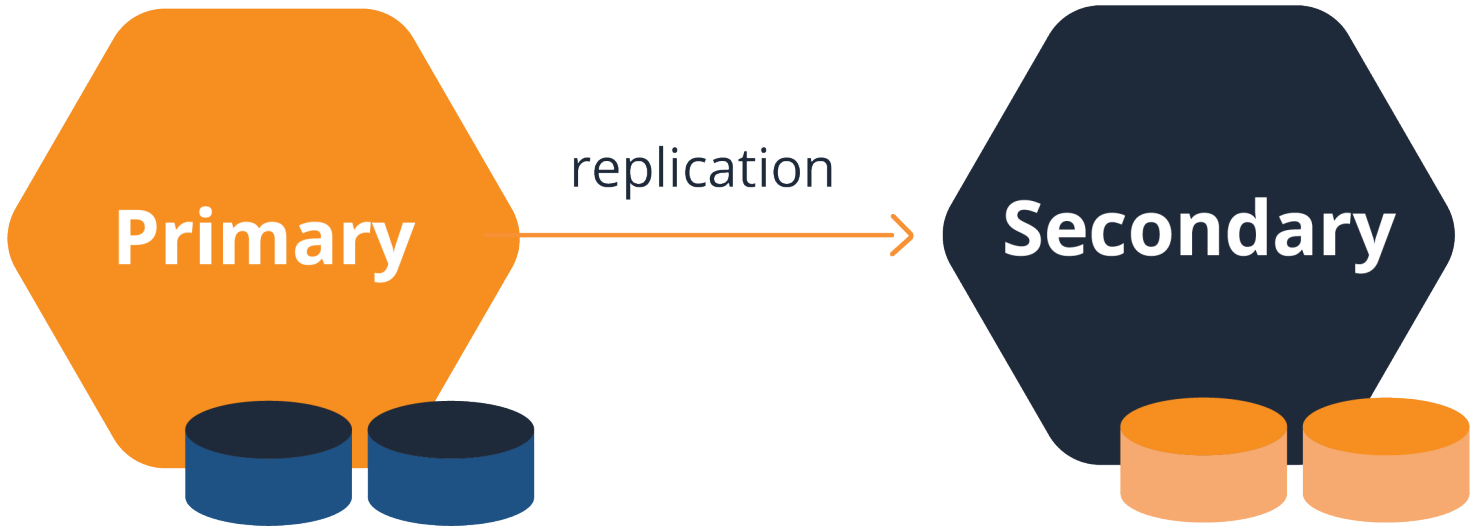


DRBD Roles: Primary & Secondary



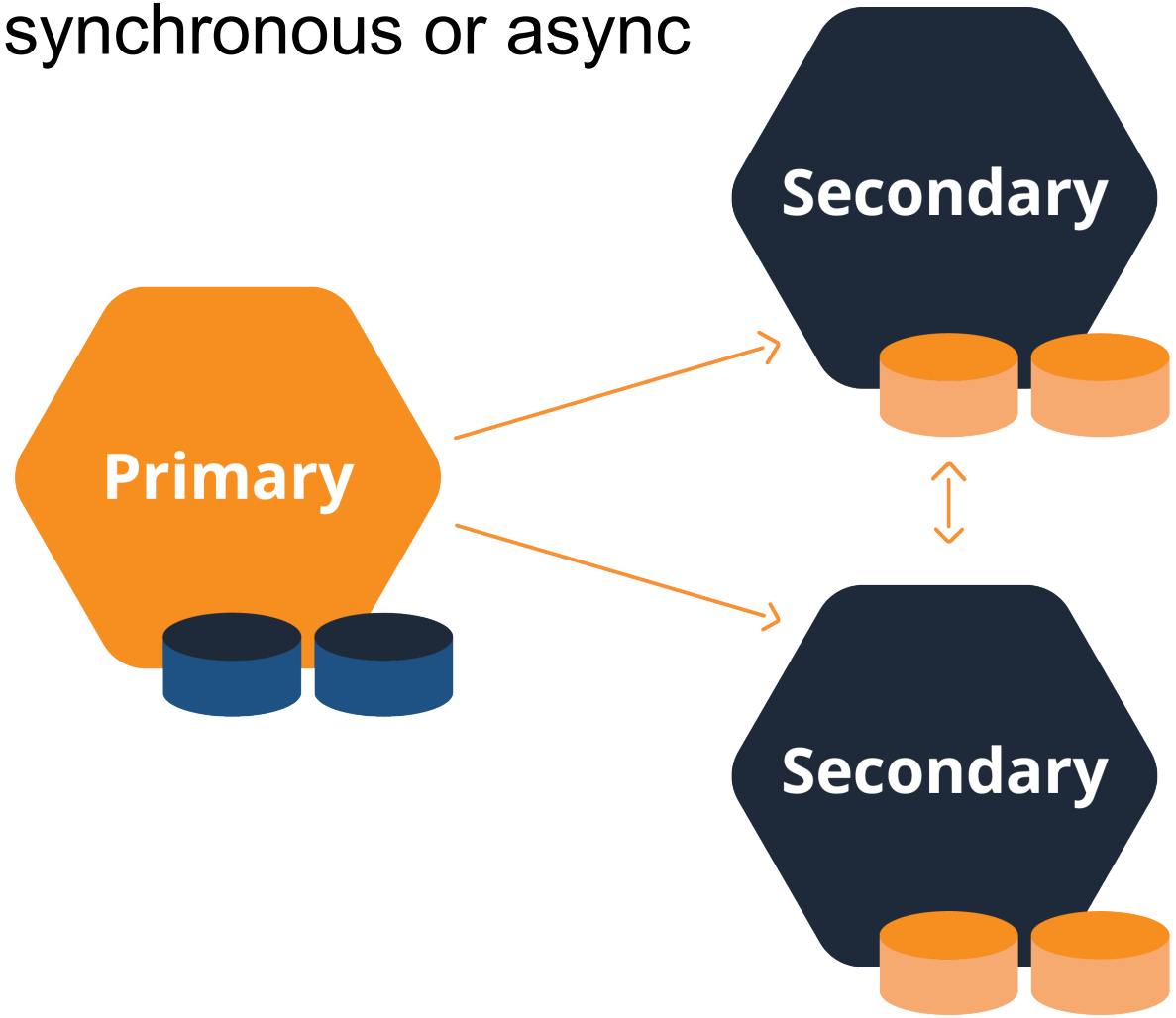
DRBD – multiple Volumes

- consistency group



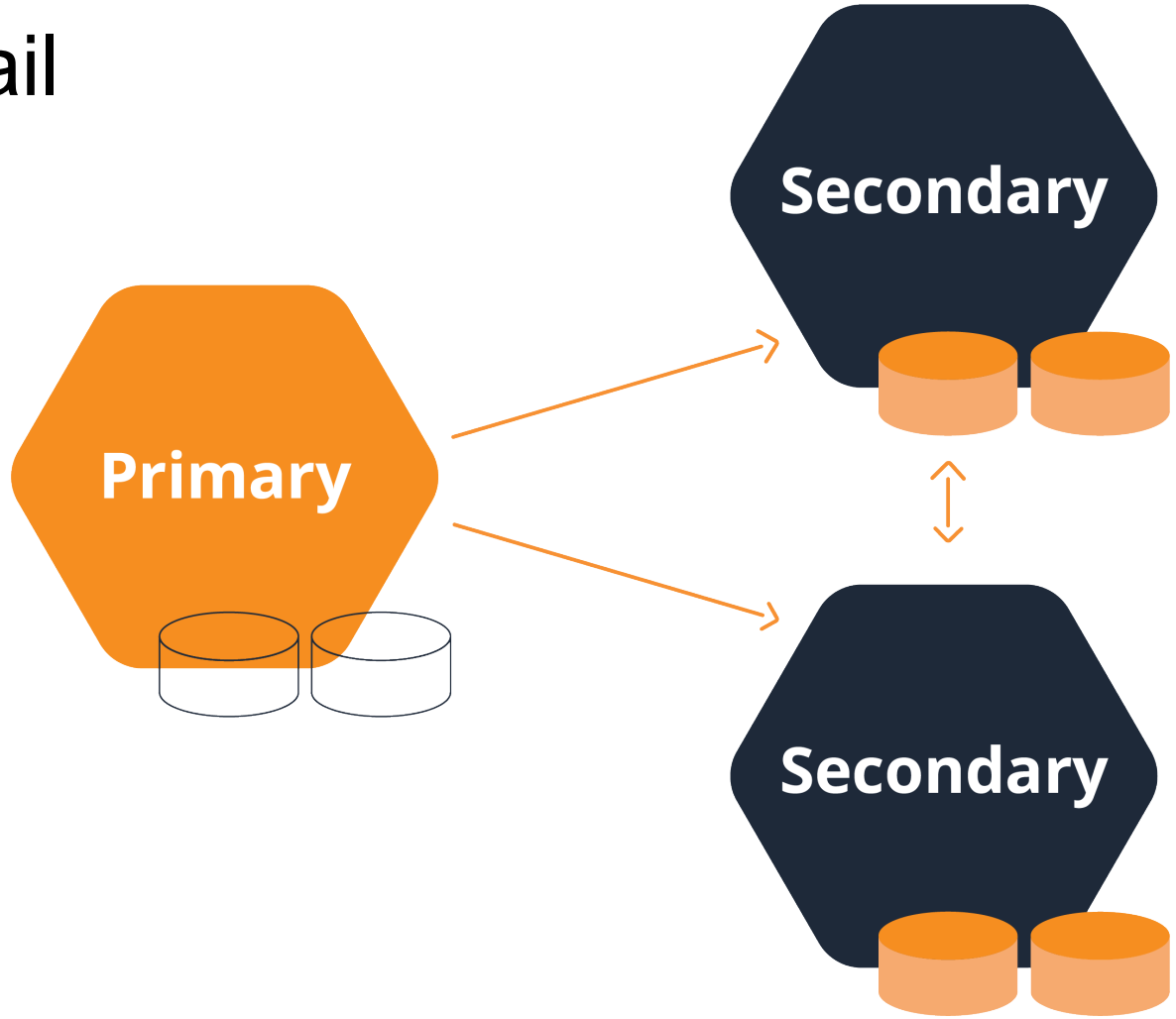
DRBD – up to 32 replicas

- each may be synchronous or async



DRBD – Diskless nodes

- intentional diskless (no change tracking bitmap)
- disks can fail



DRBD - more about

LINBIT

- a node knows the version of the data is exposes
- automatic partial resync after connection outage
- checksum-based verify & resync
- split brain detection & resolution policies
- fencing
- quorum
- multiple resouces per node possible (1000s)
- dual Primary for live migration of VMs only!

DRBD Recent Features & ROADMAP

LINBIT

- Recent optimizations
  - meta-data on PMEM/NVDIMMS
  - Improved, fine-grained locking for parallel workloads
- ROADMAP
  - Eurostars grant: DRBD4Cloud
    - erasure coding (2020)
  - Long distance replication
    - send data once over long distance to multiple replicas

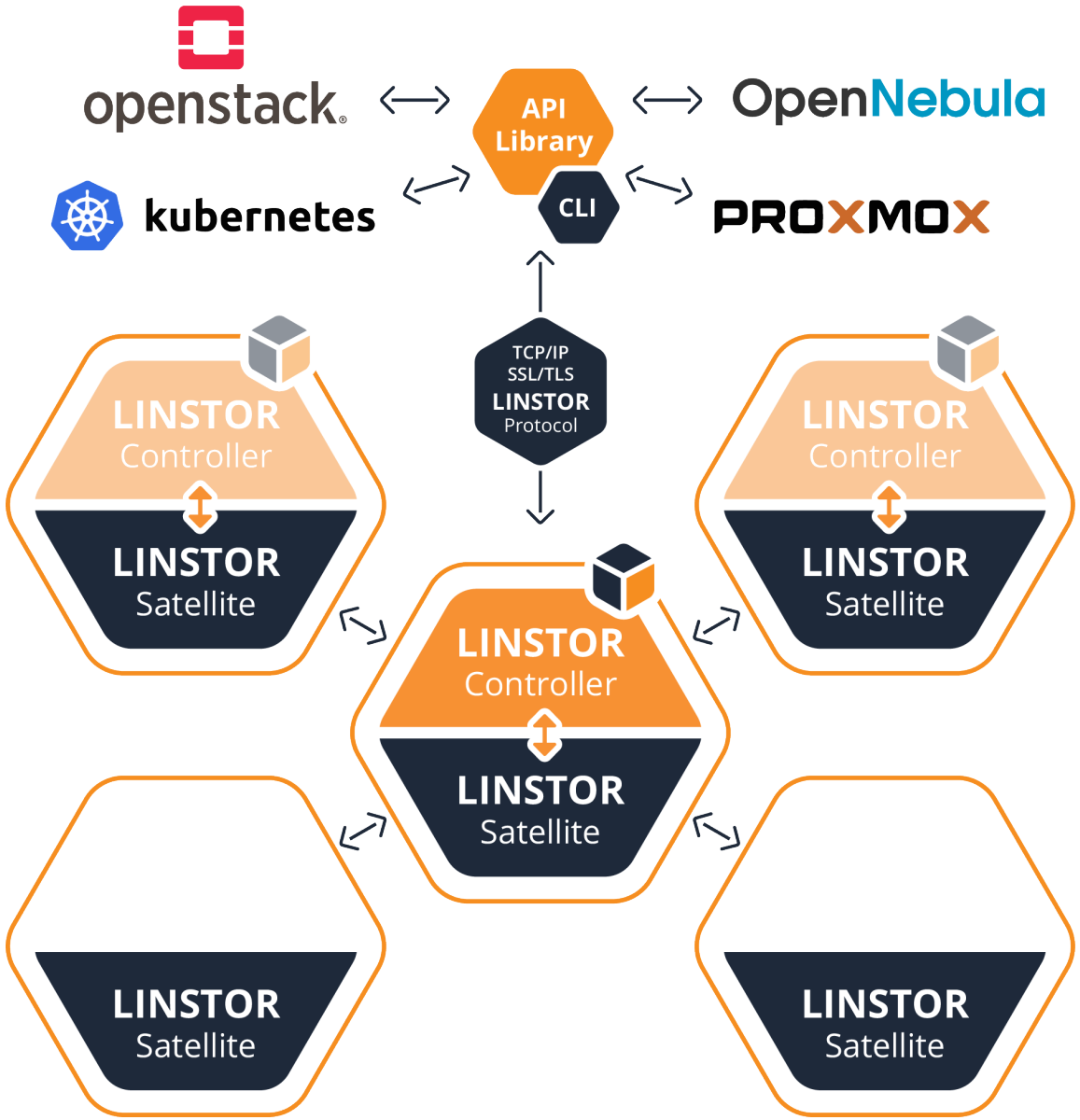
LINBIT

LINSTOR

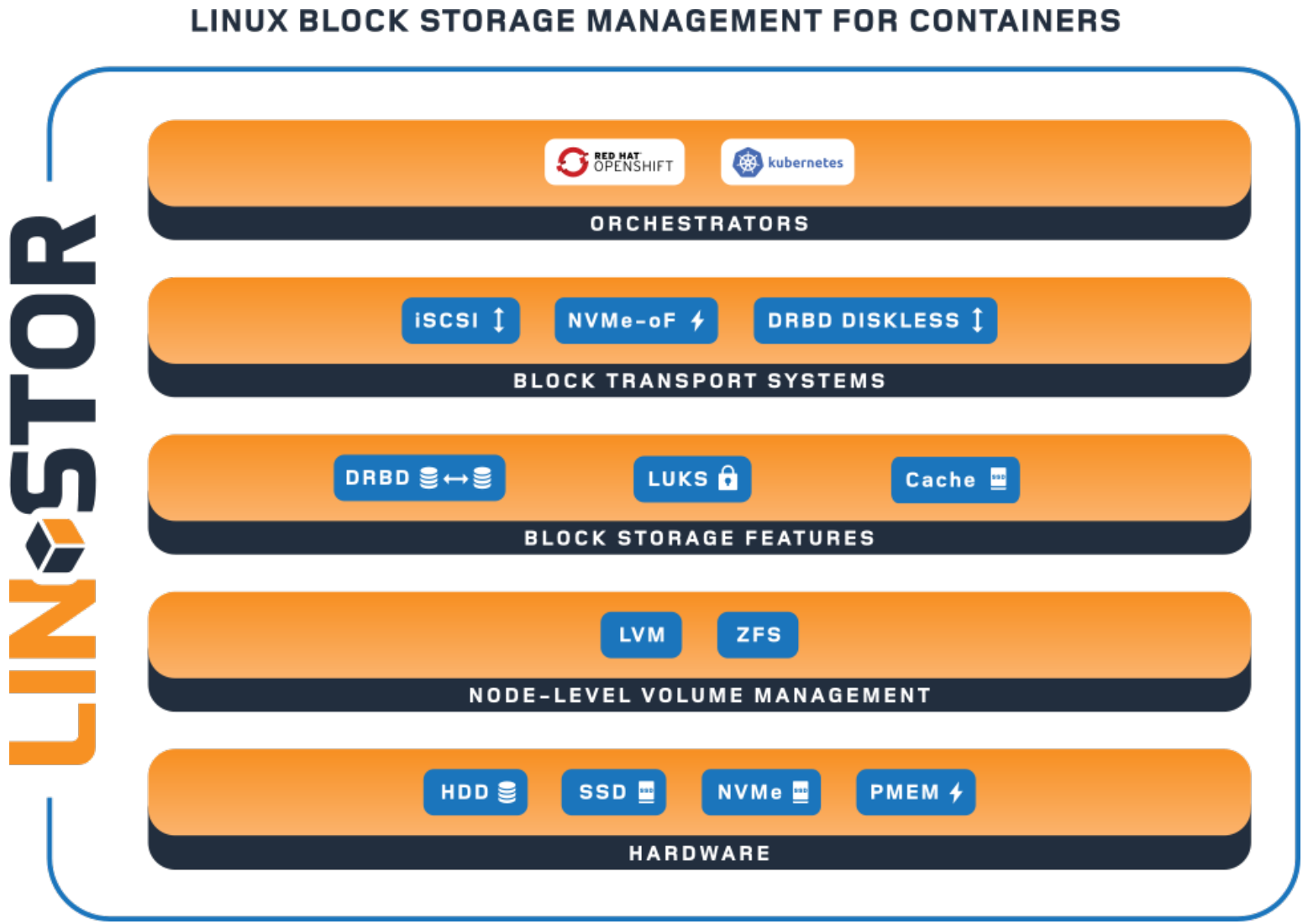
The combination is more than the sum of its parts

LINSTOR - goals

- storage build from generic (x86) nodes
- for SDS consumers (K8s, OpenStack, OpenNebula)
- building on existing Linux storage components
- multiple tenants possible
- deployment architectures
  - distinct storage nodes
  - hyperconverged with hypervisors / container hosts
- LVM, thin LVM or ZFS for volume management (stratis later)
- **Open Source, GPL**


openstack
API
Library
OpenNebula
kubernetes
CLI
PROXMOX
TCP/IP
SSL/TLS
LINSTOR
Protocol
LINSTOR
Controller
LINSTOR
Satellite
LINSTOR
Controller
LINSTOR
Satellite
LINSTOR
Controller
LINSTOR
Satellite
LINSTOR
Satellite
LINSTOR
Satellite

Summary


LINUX BLOCK STORAGE MANAGEMENT FOR CONTAINERS
LINSTOR
RED HAT OPENSHIFT
kubernetes
ORCHESTRATORS
iSCSI
NVMe-oF
DRBD DISKLESS
BLOCK TRANSPORT SYSTEMS
DRBD
LUKS
Cache
BLOCK STORAGE FEATURES
LVM
ZFS
NODE-LEVEL VOLUME MANAGEMENT
HDD
SSD
NVMe
PMEM
HARDWARE

LINSTOR connectors

- Kubernetes
  - CSI Driver
  - YAML, Operator, Helm chart
- OpenStack/Cinder
  - since Stein release (April 2019)
- OpenNebula
- Proxmox VE
- XenServer / XCP-ng

Case study - intel

Intel® Rack Scale Design (Intel® **RSD**)
is an industry-wide architecture for disaggregated, composable infrastructure that fundamentally changes the way a data center is built, managed, and expanded over time.

LINBIT working together with Intel

LINSTOR is a storage orchestration technology that brings storage from generic Linux servers and SNIA Swordfish enabled targets to containerized workloads as persistent storage. LINBIT is working with Intel to develop a Data Management Platform that includes a storage backend based on LINBIT's software. LINBIT adds support for the SNIA Swordfish API and NVMe-oF to LINSTOR.

Piraeus Datastore

- Publicly available containers of all components
- Deployment by single YAML-file
- Joint effort of LINBIT & DaoCloud

https://piraeus.io

https://github.com/piraeusdatastore

DaoCloud

DaoCloud

Piraeus Datastore

Cloud Native Storage

Speaker: Sun Liang

Sun Liang

Dr. Liang Sun is Chief Storage Architect of DaoCloud.

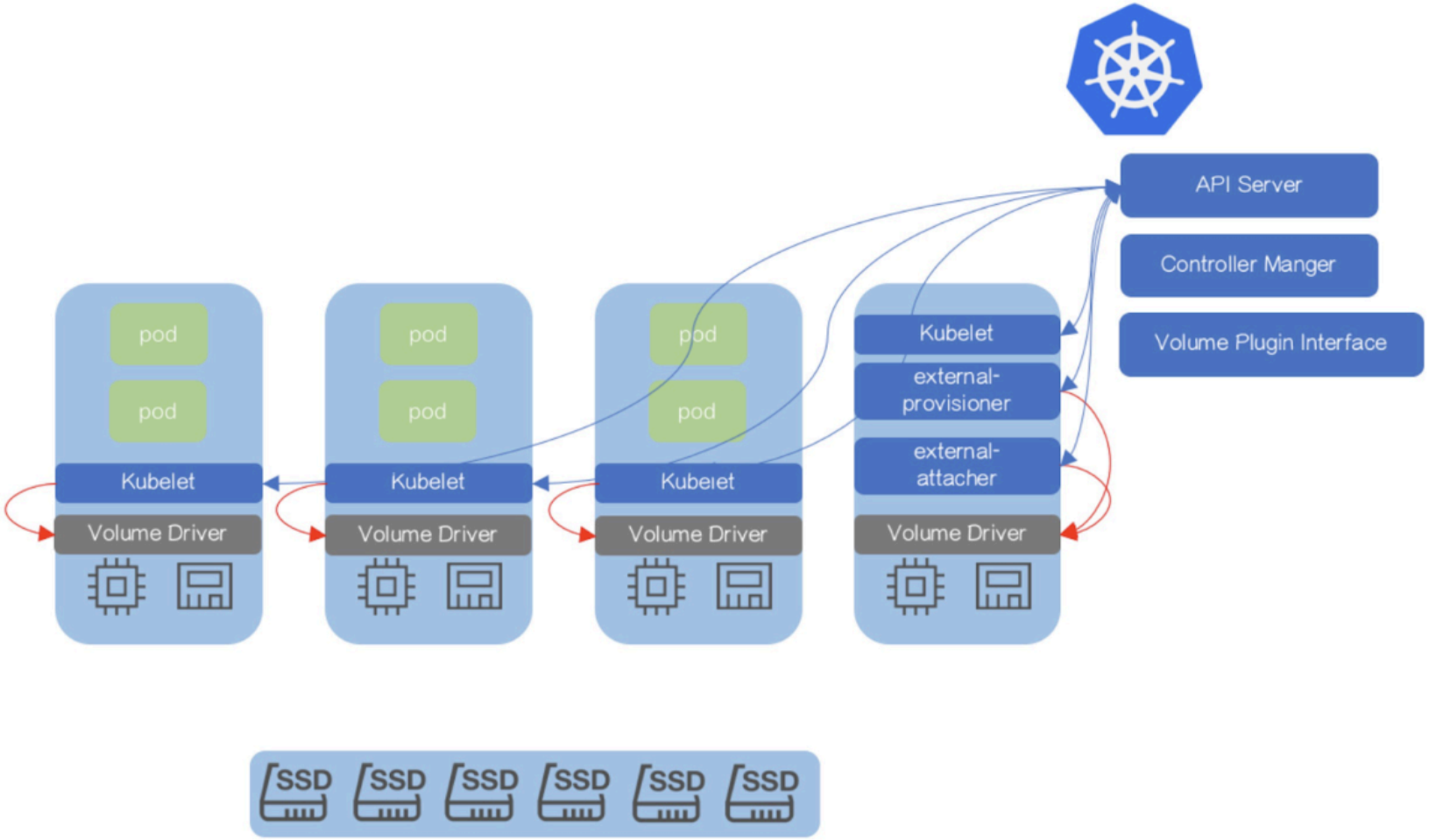
He has more than 13 years of working experience in EMC, Pure Storage, AWS, and worked on many storage products such as NAS, Object, etc. with a focus on storage, container and cloud computing.

Email: liang.sun@daocloud.io

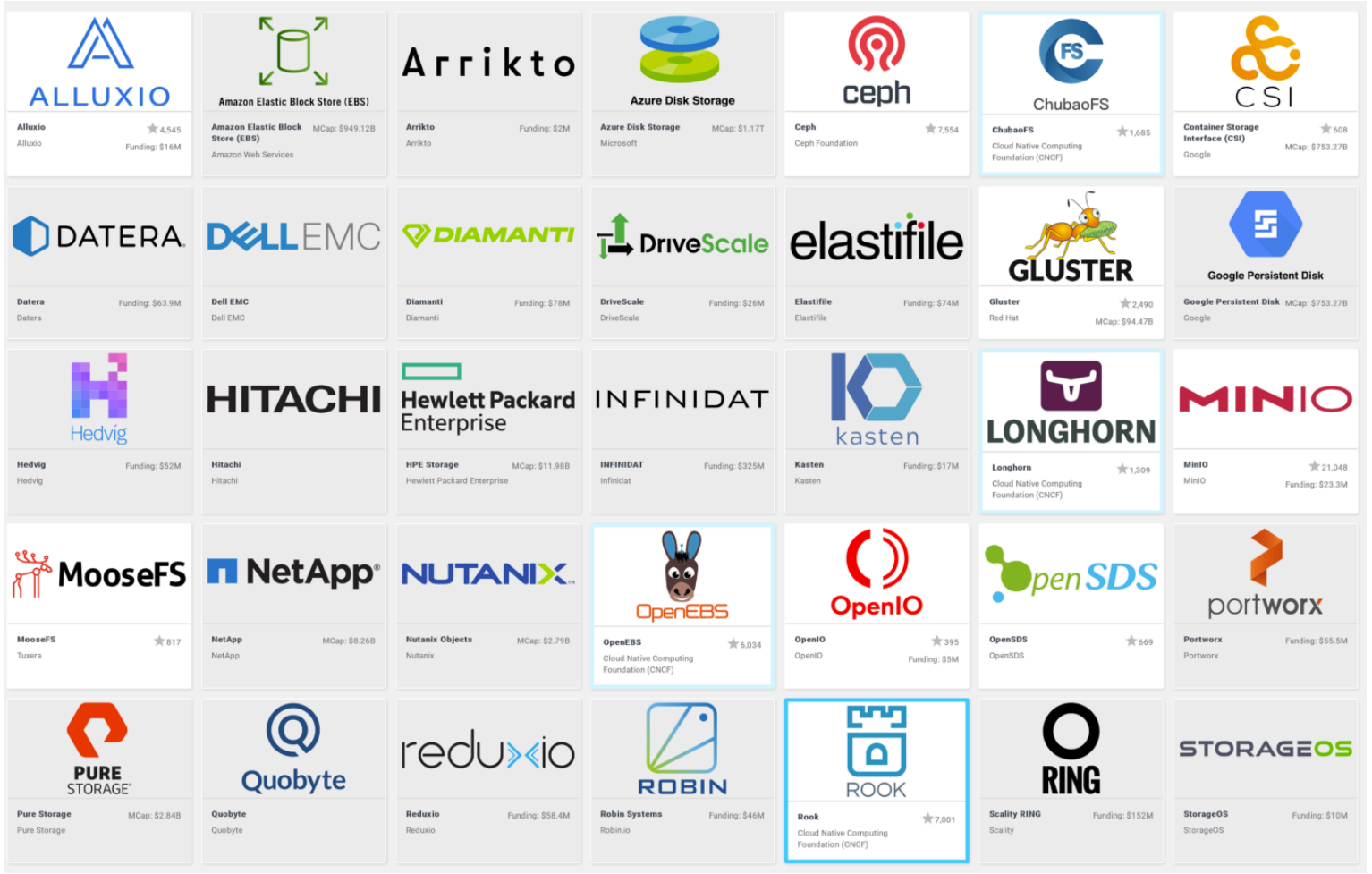
Cloud Native Storage

- Scalability
- High Availability
- Containerized
- Observable
- Multi-tenancy
- Serviceable
- … …

Kubernetes Container Storage Interface (CSI)

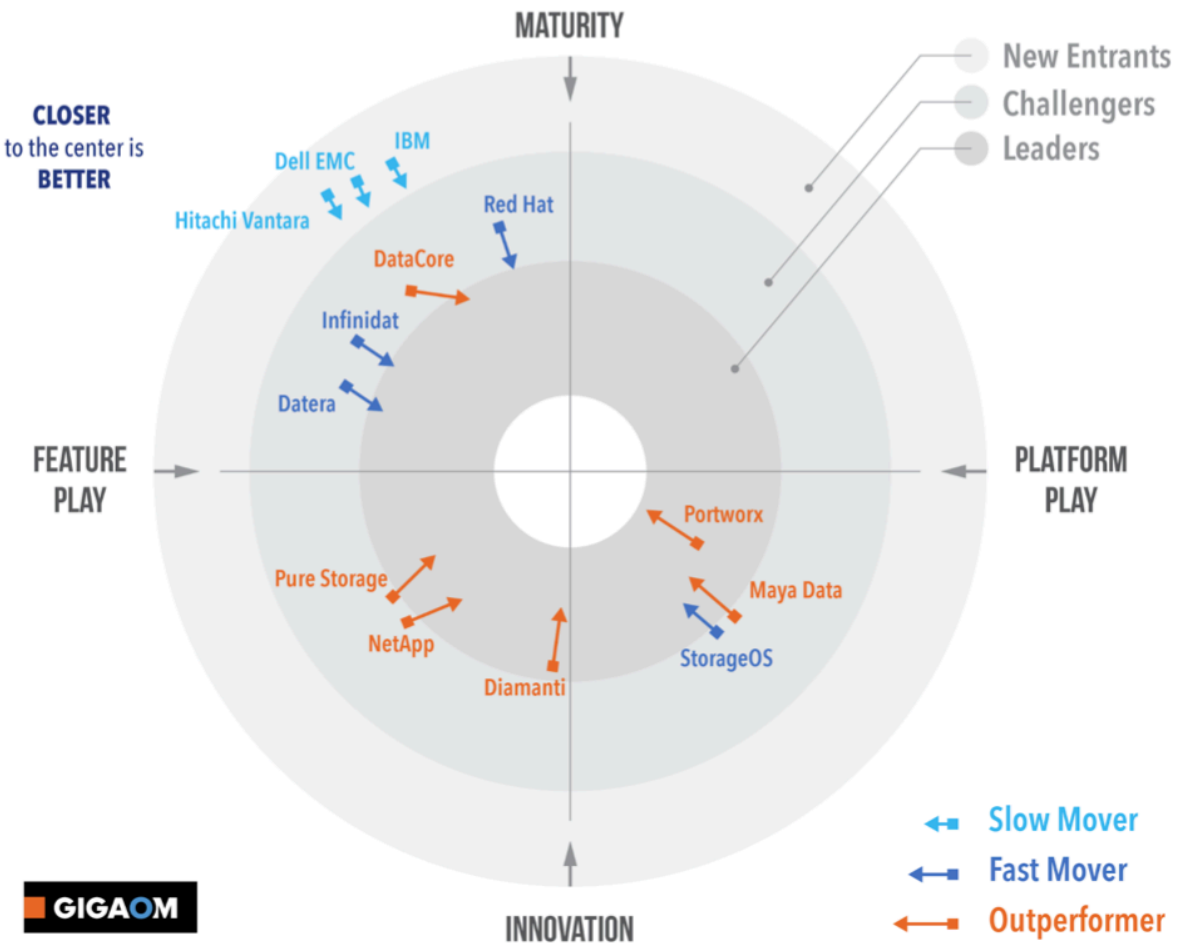


Current storage players in CNCF landscape



Market Assessment

FIGURE 1. **GIGAOM RADAR DATA STORAGE FOR KUBERNETES**



	KEY CRITERIA					EVALUATION METRICS					
	DATA SERVICES	PERFORMANCE	MULTI-TENDACY	SECURITY	MONITORING & ALERTING	ARCHITECTURE	SCALABILITY	FLEXIBILITY	EFFICIENCY	MANAGEABILITY EASE OF USE	PARTNER ECOSYSTEM
DATACORE	+	+++	+++	++	++	++	++	+++	++	+++	+
DATARA	+	+++	++	+++	++	++	+++	++	+++	++	++
DELL EMC	+	++	++	++	++	+	+	+	++	+	+++
DIAMANTI	+++	+++	++	+++	+++	++	+++	++	+++	+++	+
HITACHI VANTARA	+	++	++	++	++	+	+++	+	++	+	++
IBM	+	++	+	++	+	+	+	+	++	+	++
INFINDAT	++	+++	+++	++	++	++	+++	++	+++	++	++
MAYA DATA	+++	++	+	++	+++	++	++	+++	++	+++	++
NETAPP	+	+++	++	+++	+++	++	+++	++	+++	+++	+++
PORTWORX	+++	+++	++	+++	+++	+++	+++	+++	++	+++	+++
PURE STORAGE	+	+++	+++	+++	+	+++	+++	++	+++	++	+++
RED HAT	+	+	+	++	+++	+	+	+	+	+++	+++
STORAGEOS	+	+++	+++	+++	++	+++	+++	++	++	++	+

GIGAOM

Legend:
+++ Strong focus and perfect fit of the solution
++ The solution is good in this area, but there is still room for improvement
+ The solution has limitations and a narrow set of use cases
- Not applicable or absent.

PIRAEUS
Local Storage

- High Performance
- High Availability
- Software Defined Storage
- Cloud Native
- 100% Open Source

… …

1. http://github.com/piraeusdatastore
2. http://piraeus.io

CNCF Communities Contributions（DaoCloud）

上 云

模版编排

HELM 第9名

开发环境

MINIKUBE 第9名

云原生应用

Vitess 第16名

感 知

服务网格

Istio 第11名

全链路追踪

Skywalking 第7名

服务网关

envoy 第18名

配置中心

Apollo 第5名

运 营

监控告警

Prometheus 第10名

安装运维

kubeadm 第2名

kops 第3名

KUBESPRAY 第18名

控 制

调度

kubernetes 第23名

计算

docker 第8名

rkt 第18名

网络

CNI 第14名

CoreDNS 第5名

By StackAlytics 2020-03

References

1. https://github.com/cncf/sig-storage
2. https://www.cncf.io/webinars/introduction-to-cloud-native-storage/
3. https://www.youtube.com/watch?time_continue=561&v=ZRp8G9UUC8U&feature=emb_logo
4. https://docs.google.com/document/d/1Cek8jJ2SPt4xx7Tnx7ih_m4DxzSimj_w26qYHnfrrRQ/edit#heading=h.edh0hqbub2ib
5. http://github.com/piraeusdatastore
6. http://piraeus.io

DaoCloud

Piraeus Datastore

Project Piraeus

Speaker: Alex Zheng

Alex Zheng

Alex is senior storage engineer in DaoCloud and PM of Piraeus project.

He holds a bachelor degree of computer engineering from Virginia Tech. Before joining DaoCloud, he worked in EMC as a tech specialist for ScaleIO software defined storage.

Email: alex.zheng@daocloud.io

Project Piraeus

Name	Piraeus (sea port of Athens Greece)
Founders	DaoCloud and LINBIT
Definition	Dynamic Provisioning, Resource Management and High Availability for **Local Volumes**
Goal	Solve container persistence challenge within Kubernetes nodes
Webpage	https://piraeus.io/ github.com/piraeusdatastore

KubeCon North America, San Diego, Nov 2019

Piraeus Datastore

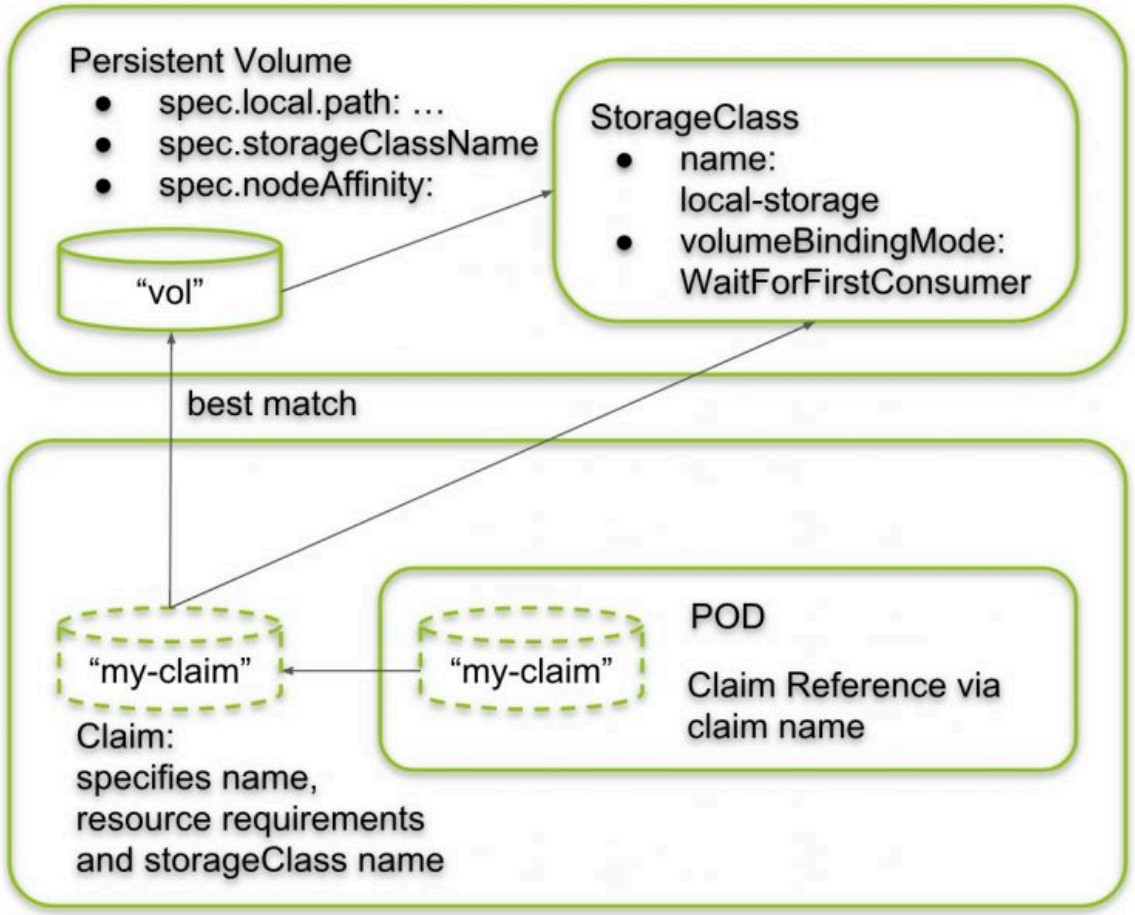
Kubernetes Local Persistent Volumes

Local Persistent Volume feature became GA in Kubernetes version 1.14. It implements Volume.nodeAffinity and WaitForFirstConsumer.

The physical backend of local volumes includes HDD, SSD, RAID and also SAN/EBS.

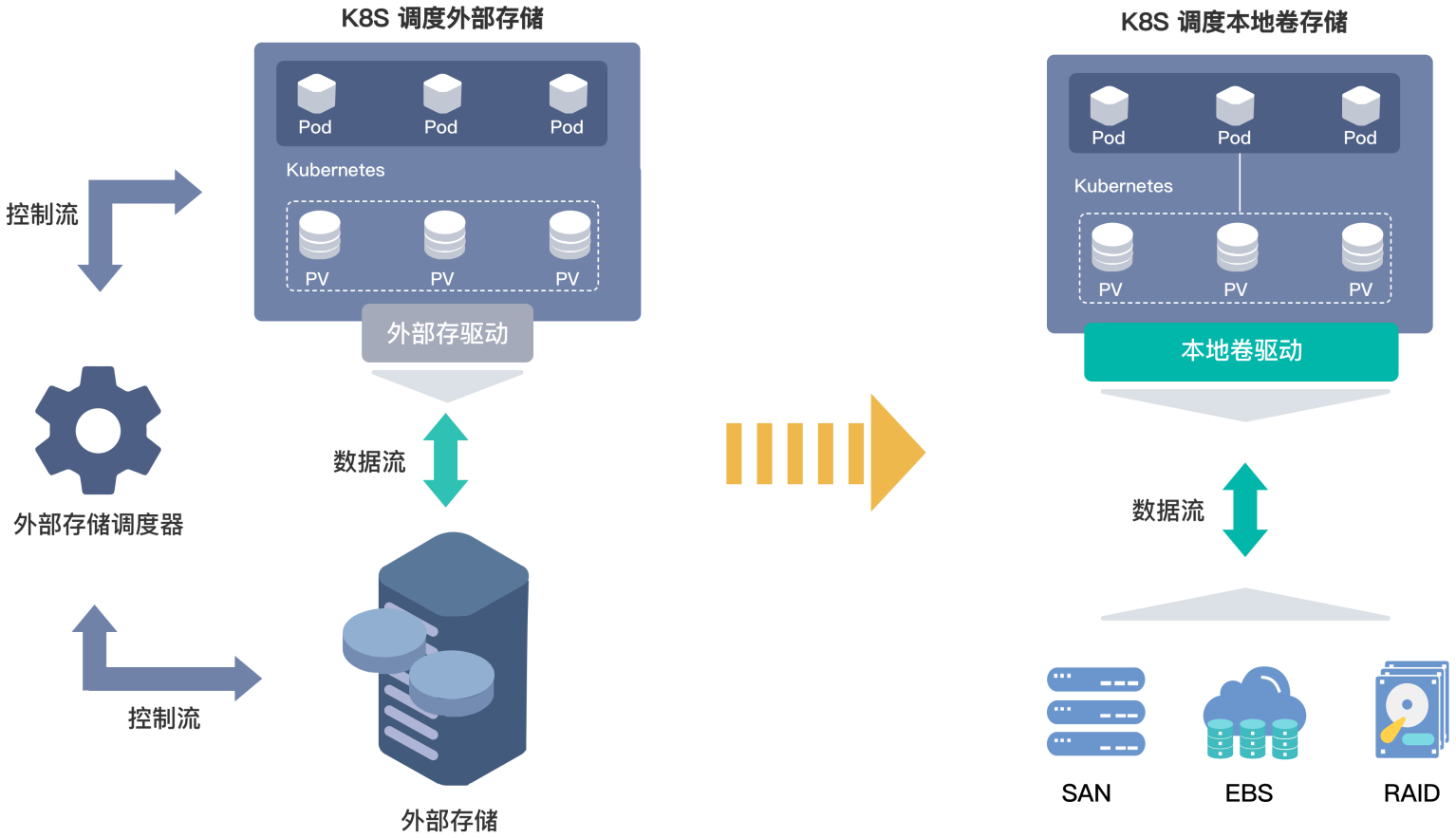
“With the Local Persistent Volume plugin, Kubernetes workloads can now consume high performance local storage using the same volume APIs that app developers have become accustomed to.”

https://kubernetes.io/blog/2019/04/04/kubernetes-1.14-local-persistent-volumes-ga/



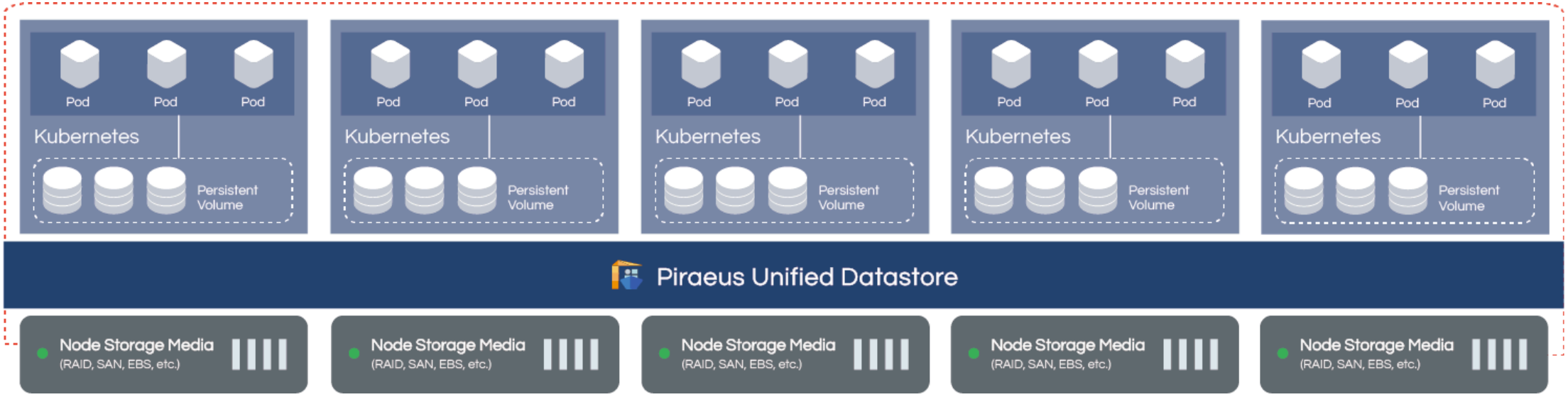
https://vocon-it.com/

Piraeus avoids PaaS <=> IaaS storage control flow



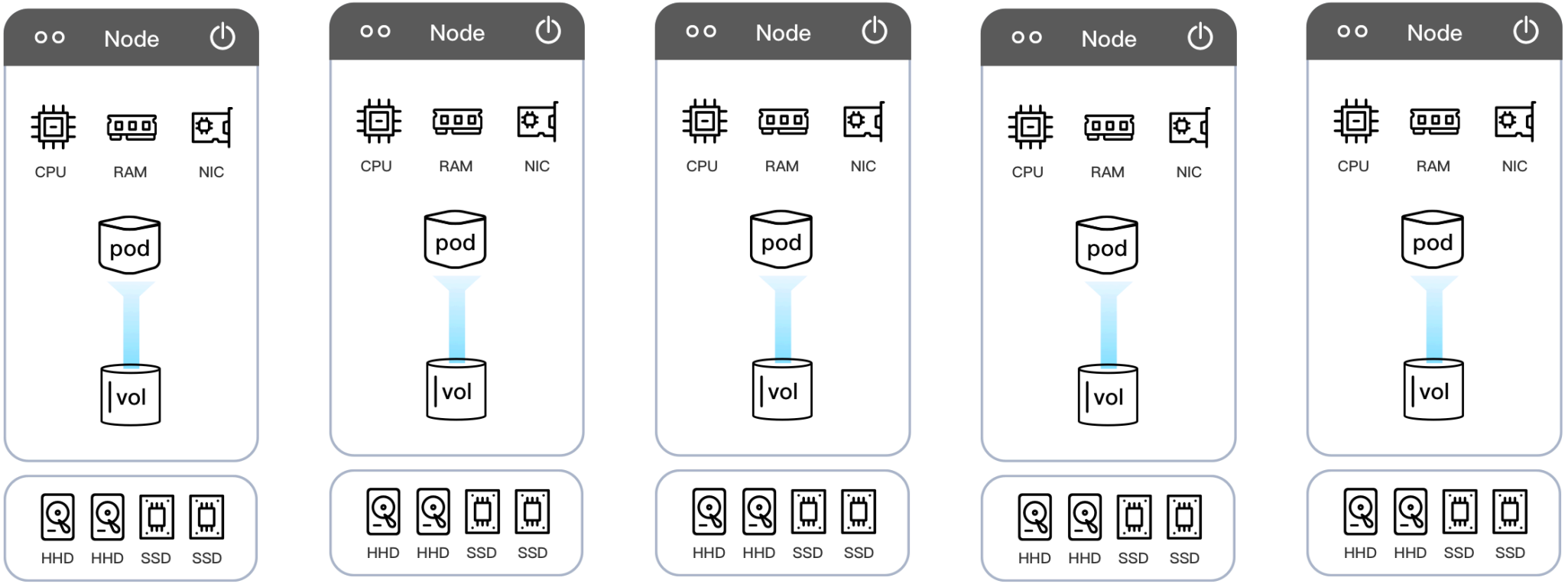
Piraeus is distributed

- Manages local disk space of each node

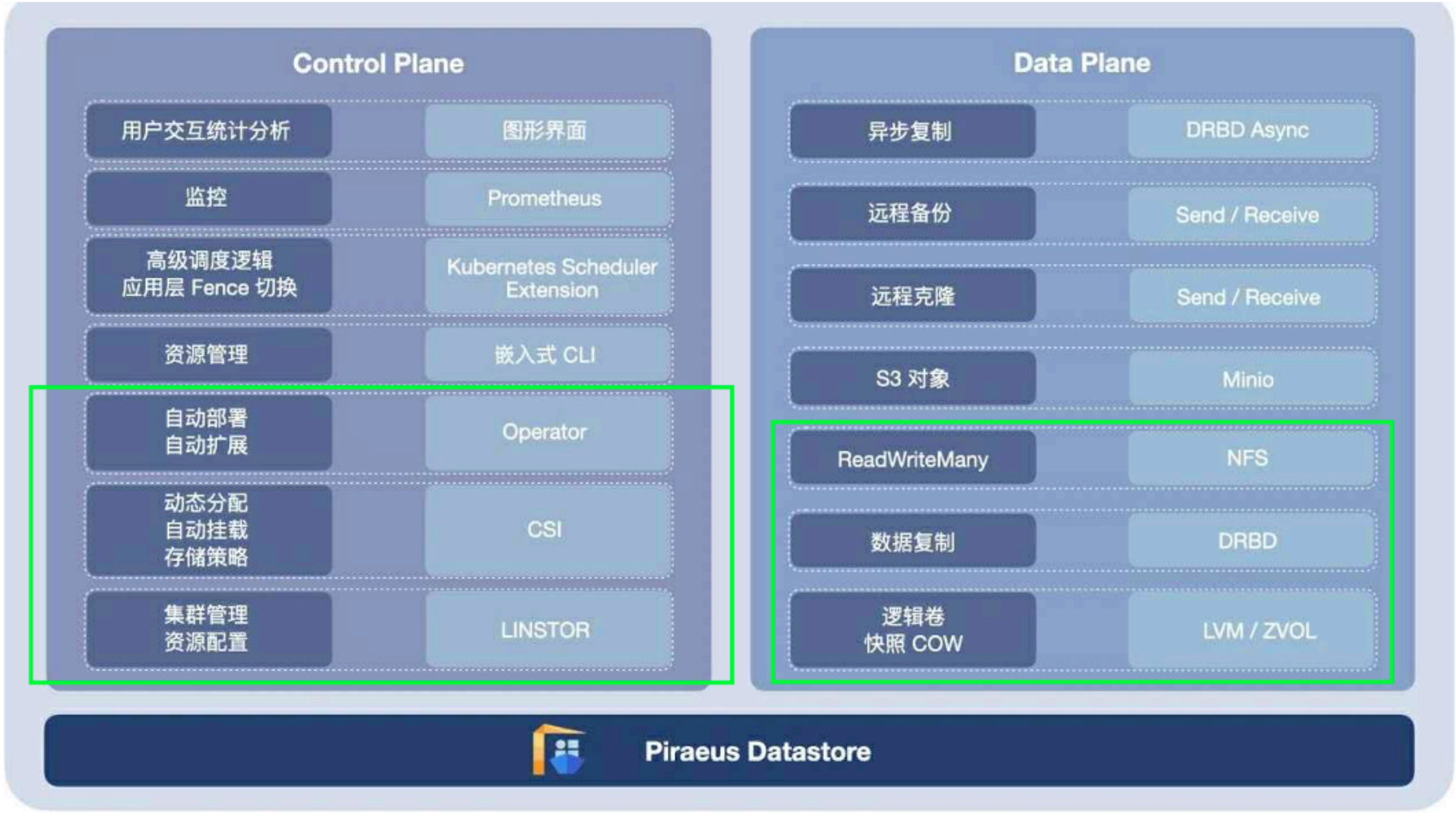


Piraeus is disaggregated

- Scalability matches Kubernetes on 1:1 ratio



Piraeus stack (proposed)



Key part of Piraeus is to provision local volume dynamically with HA option

POD Node Affinity vs. Volume Node Affinity

- The two are equivalent in syntax
- Volume Node Affinity also indicates data locality
- POD will not start if there are POD/Volume node affinity conflict, for example:

POD Node Affinity:

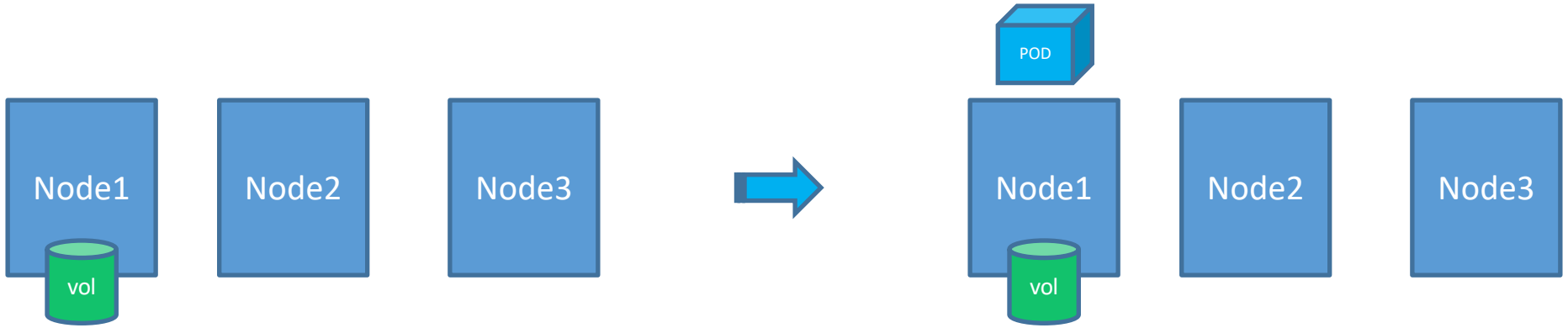
```
affinity:
  nodeAffinity:
    required:
      nodeSelectorTerms:
      - matchExpressions:
        - key: kubernetes.io/hostname
          operator: In
          values:
          - "k8s-worker-1"
```

Volume Node Affinity:

```
affinity:
  nodeAffinity:
    required:
      nodeSelectorTerms:
      - matchExpressions:
        - key: kubernetes.io/hostname
          operator: In
          values:
          - "k8s-worker-2"
```

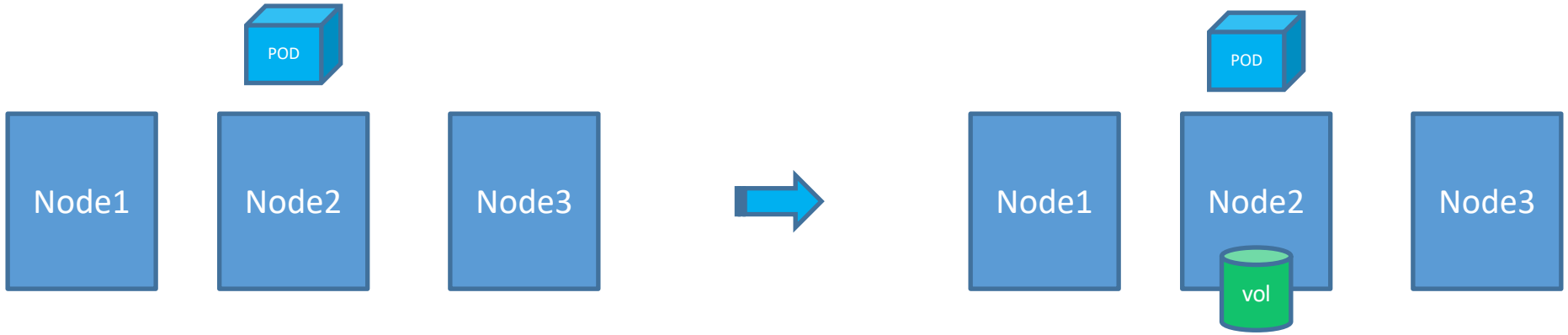
VolumeBindingMode: Immediate

- PV is bound right after PVC creation
- VolumeNodeAffinity is immutable after PV creation
- POD starts on the node where volume is provisioned



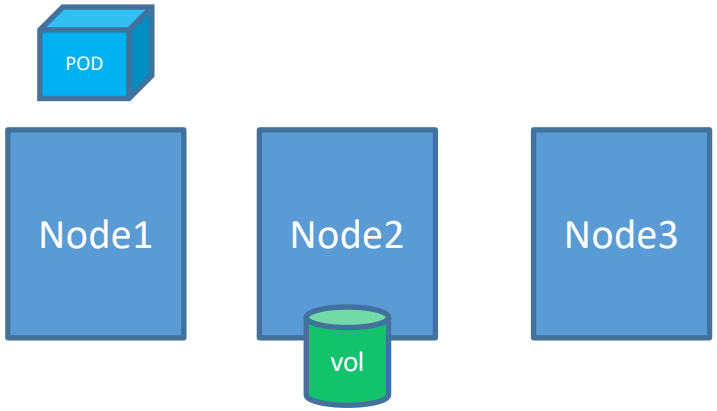
VolumeBindingMode: waitForFirstConsumer

- PV is pending after PVC creation
- Volume is provisioned after POD is scheduled

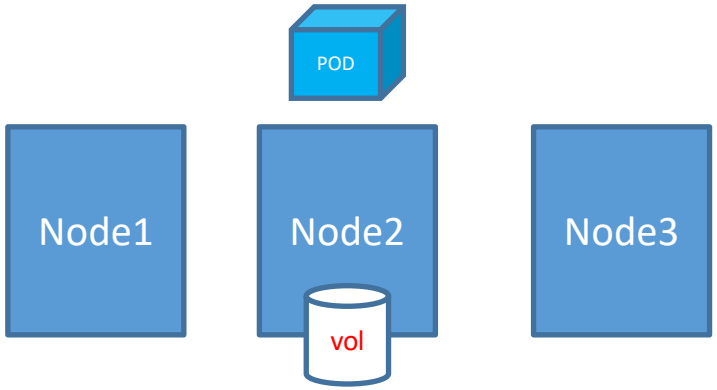


How to deal with Not-So-Ideal Situations ?

- Node is either drained or taint:NoSchedule
- POD has AntiAffinity

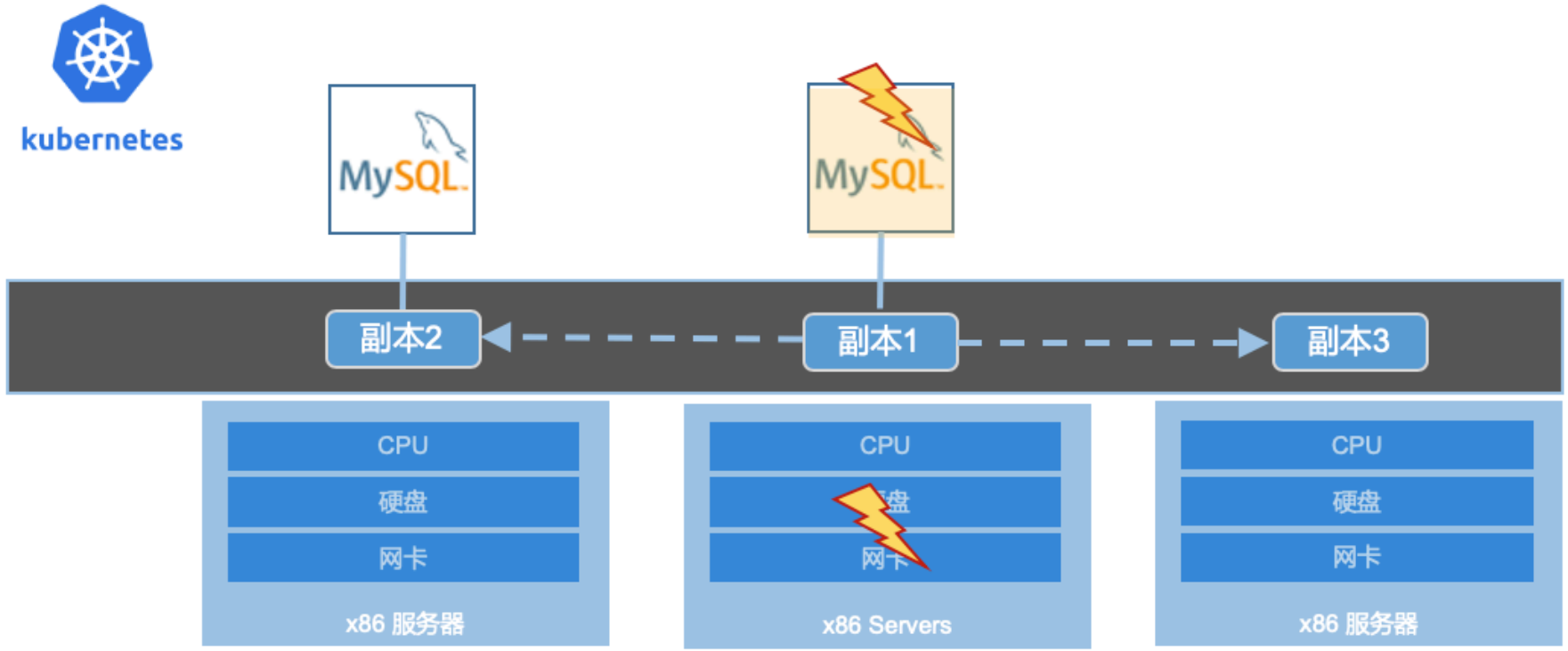


- Node has not enough disk spaces

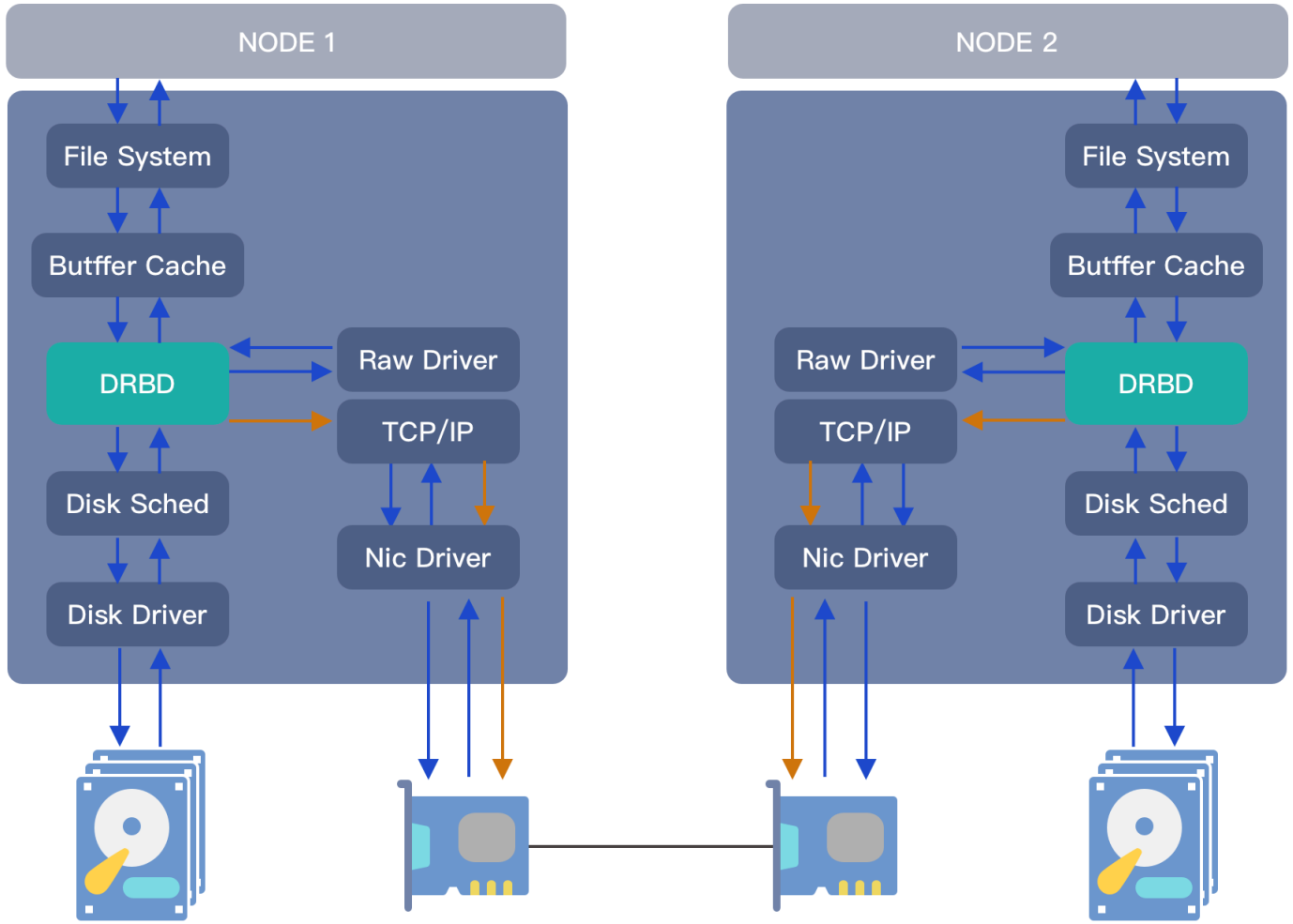


Remote access to help …

What about HA?



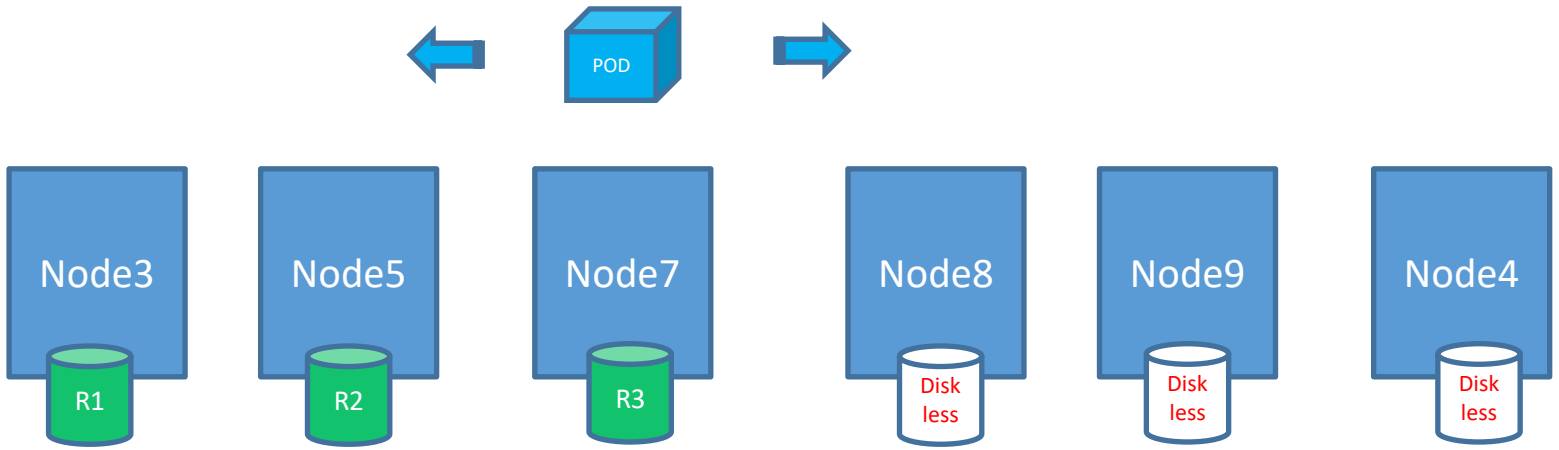
Piraeus uses DRBD for HA



DRBD has been merged into Linux kernel for over 10 years. It is widely battle-tested in enterprise production environment by IBM and Intel.

Ideal provisioning

- Prioritize local volume for the best effort
- Use remote access if above is not possible



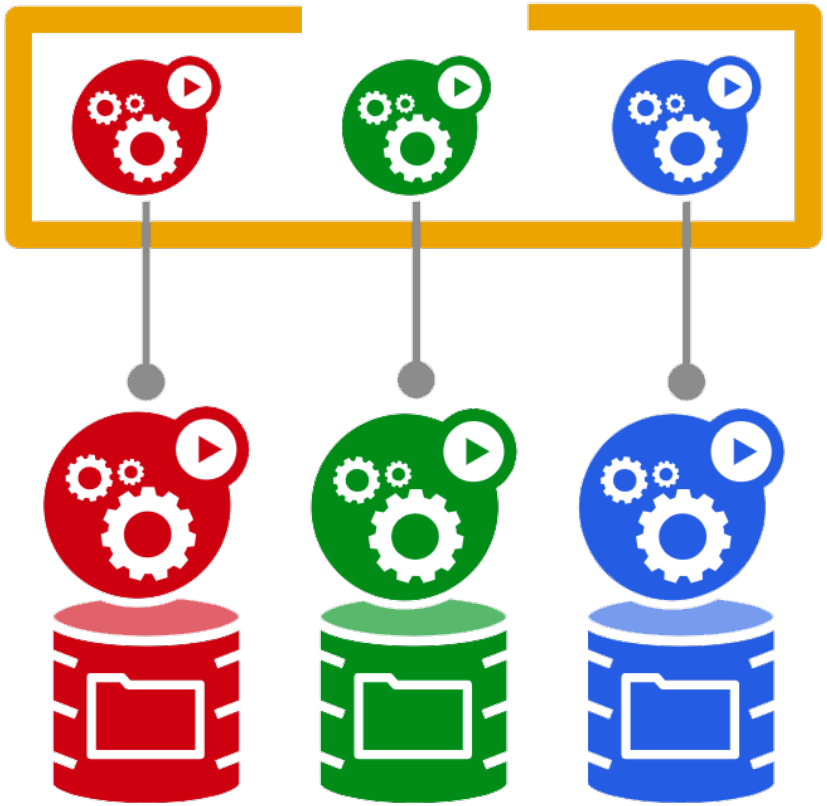

```
nodeAffinity:
    required:
      nodeSelectorTerms:
      - matchExpressions:
        - key: linbit.com/hostname
          operator: Exists
    preferred:
    - weight: 100
      preference:
        matchExpressions:
        - key: linbit.com/hostname
          operator: In
          values:
          - k8s-worker-3
          - k8s-worker-5
          - k8s-worker-7
```

DaoCloud

Piraeus Datastore

Demo

MySQL StatefulSet



In this demo, we will simulate a 3-node MySQL cluster.

Each MySQL node will use a 3-replicate Piraeus volume to store its data.

Such setup provides an **enterprise-level data continuity** for a datastore, as there are replication on both application and storage level.