LINBIT

LINSTOR

Resilient and Fast Persistent Container Storage Leveraging Linux’s Storage Functionalities

Philipp Reisner, CEO LINBIT

Presenter: Robert Altnoeder

LINBIT - the company behind it

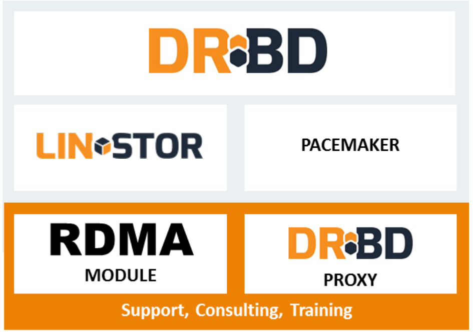
LINBIT

COMPANY OVERVIEW

- **Developer** of **DRBD**
- **100% founder owned**
- Offices in **Europe and US**
- Team of **30 highly experienced Linux experts**
- Partner in **Japan**

LINBIT
LINBIT
LINBIT

TECHNOLOGY OVERVIEW



DRBD

LINSTOR

PACEMAKER

RDMA MODULE

DRBD PROXY

Support, Consulting, Training

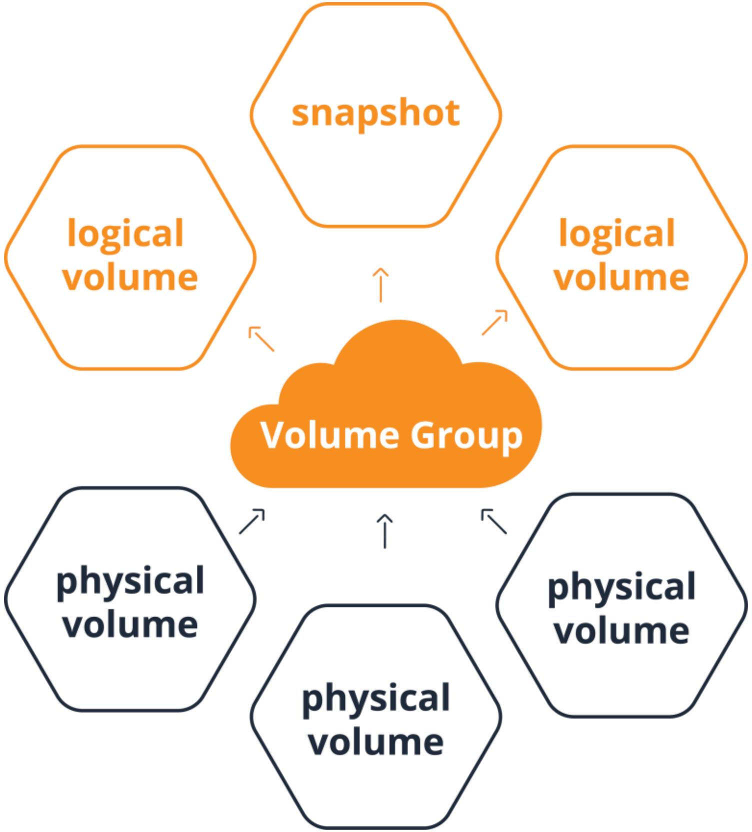
REFERENCES

Google SanDisk intel cisco IBM MICRO FOCUS redhat NTT

Linux Storage Gems

LVM, RAID, SSD cache tiers, deduplication, targets & initiators

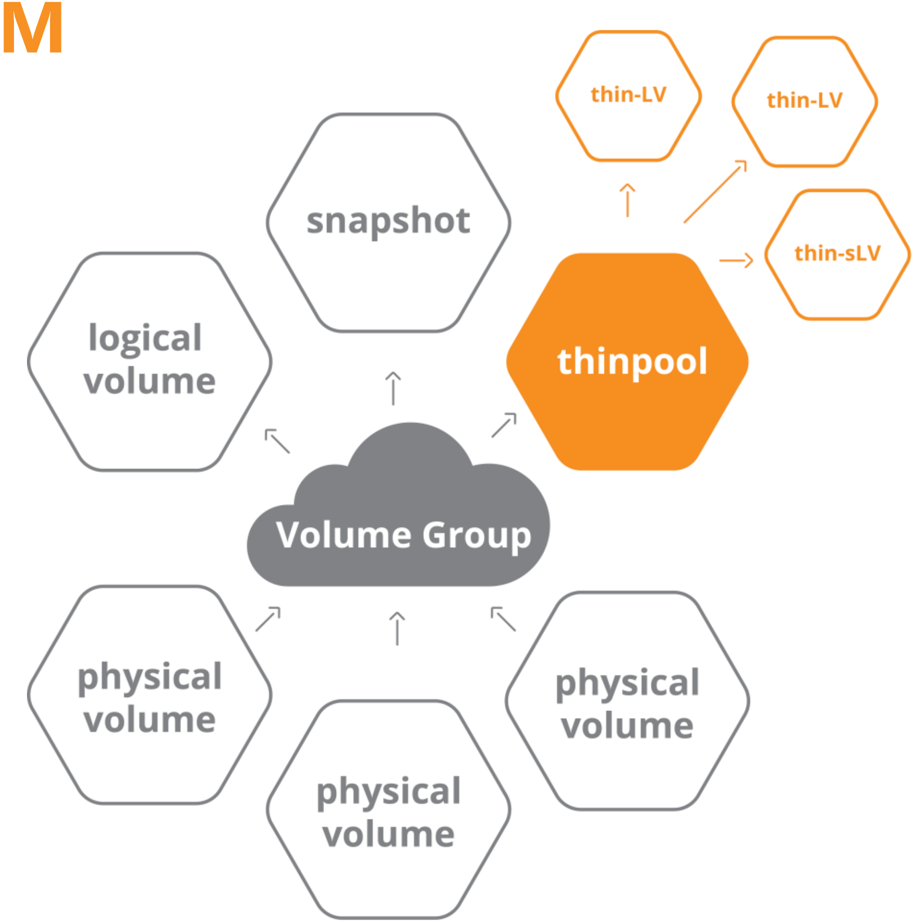
Linux's LVM


snapshot
logical volume
logical volume
Volume Group
physical volume
physical volume
physical volume

Linux's LVM

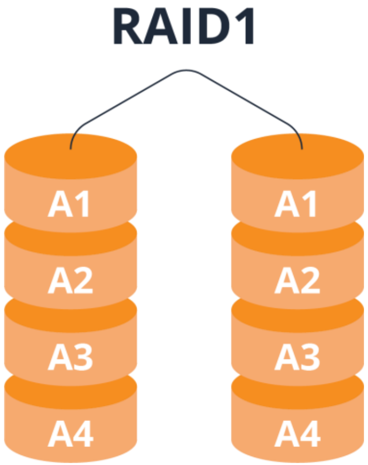
- based on device mapper
- original objects
  - PVs, VGs, LVs, snapshots
  - LVs can scatter over PVs in multiple segments
- thinlv
  - thinpools = LVs
  - thin LVs live in thinpools
  - multiple snapshots became efficient!

Linux's LVM


thin-LV
thin-LV
thin-sLV
snapshot
thinpool
logical volume
Volume Group
physical volume
physical volume
physical volume

Linux's RAID

- original MD code
  - mdadm command
  - Raid Levels: 0,1,4,5,6,10
- Now available in LVM as well
  - device mapper interface for MD code
  - do not call it ‘dmraid’; that is software for hardware fake-raid
  - lvcreate --type raid6 --size 100G VG_name



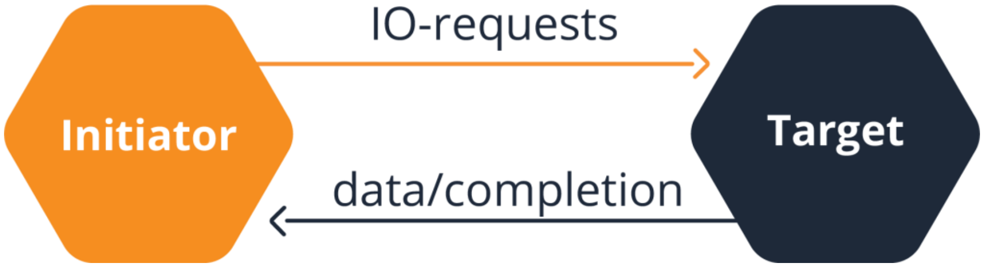
SSD cache for HDD

- dm-cache
  - device mapper module
  - accessible via LVM tools
- bcache
  - generic Linux block device
  - slightly ahead in the performance game

Linux's DeDupe

- Virtual Data Optimizer (VDO) since RHEL 7.5
  - Red hat acquired Permabit and is GPLing VDO
- Linux upstreaming is in preparation
- in-line data deduplication
- kernel part is a device mapper module
- indexing service runs in user-space
- async or synchronous writeback
- Recommended to be used below LVM

Linux's targets & initiators



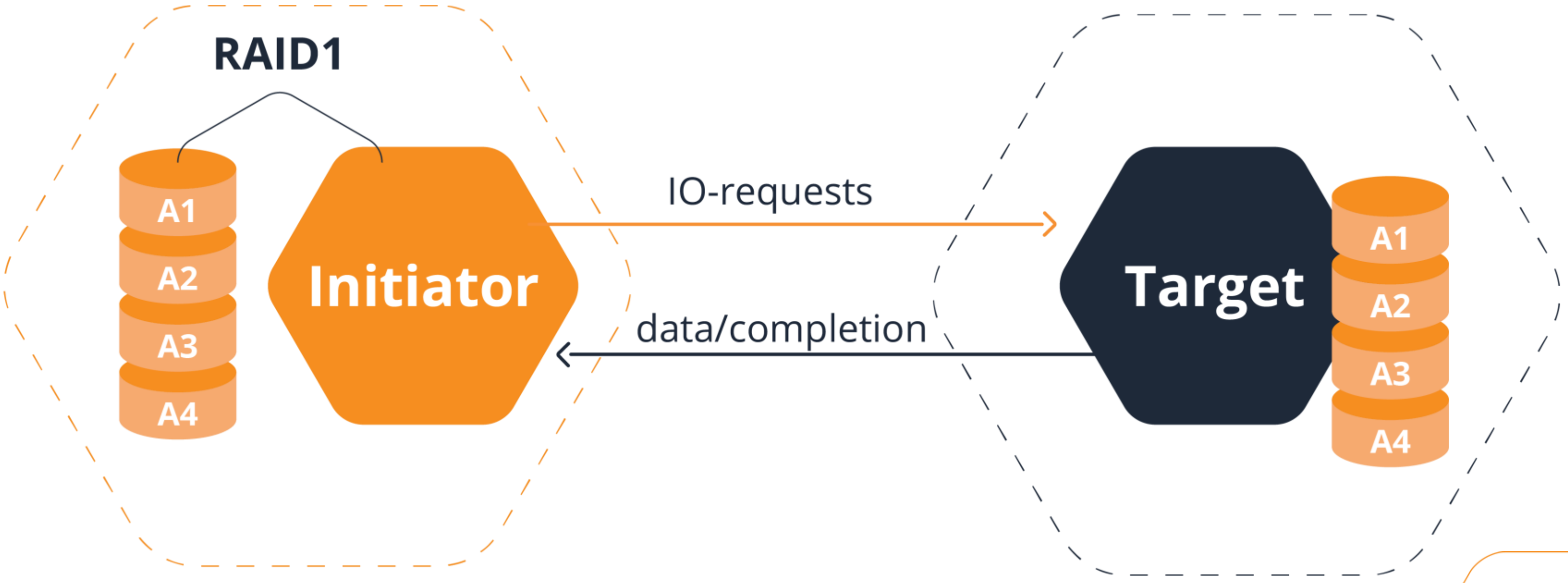
- Open-ISCSI initiator
- Ietd, STGT, SCST
  - mostly historical
- **LIO**
  - iSCSI, iSER, SRP, FC, FCoE
  - SCSI pass through, block IO, file IO, user-specific-IO
- NVMe-OF
  - target & initiator

ZFS on Linux

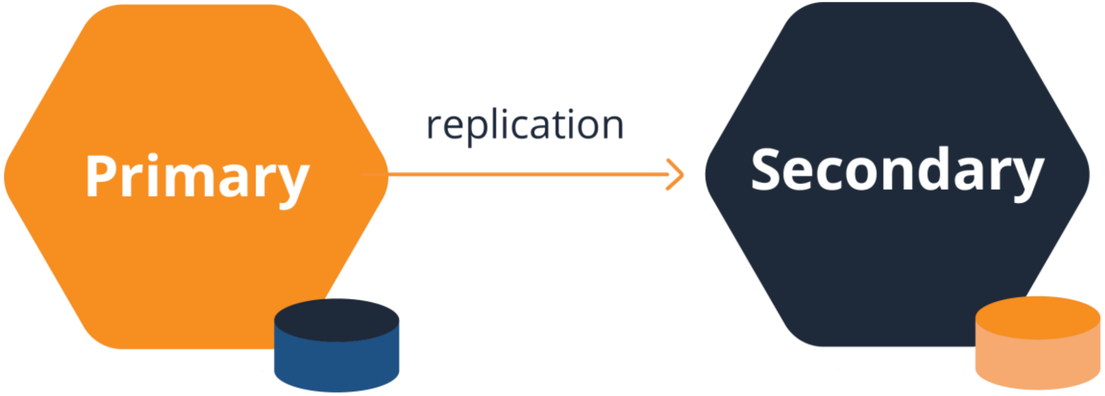
- Ubuntu eco-system only
- has its own
    - logic volume manager (zVols)
    - thin provisioning
    - RAID (RAIDz)
    - caching for SSDs (ZIL, SLOG)
    - and a file system!

Put in simplest form

DRBD – think of it as …

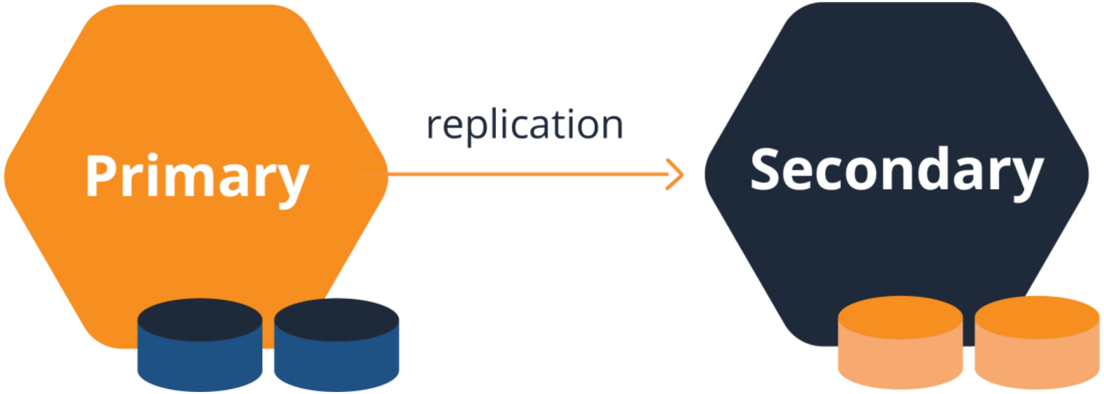


DRBD Roles: Primary & Secondary



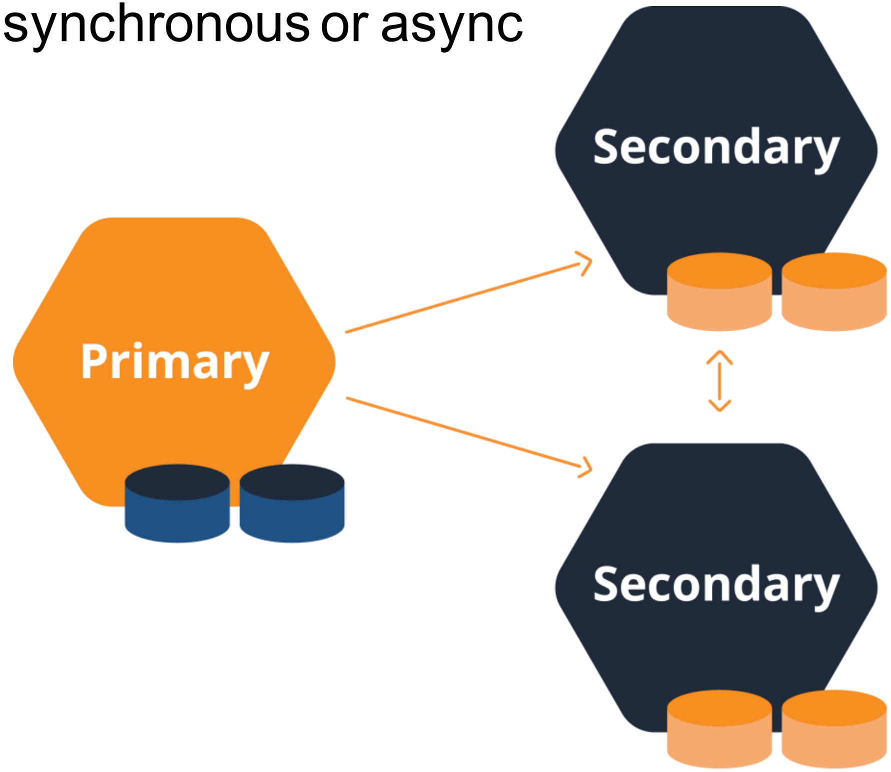
DRBD – multiple Volumes

- consistency group



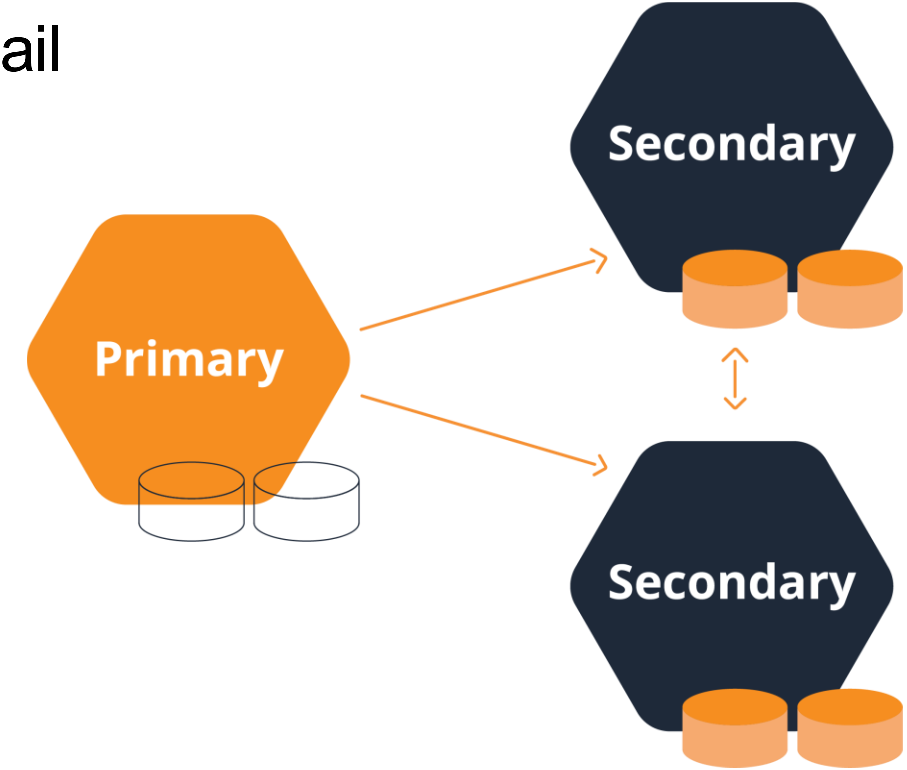
DRBD – up to 32 replicas

- each may be synchronous or async



DRBD – Diskless nodes

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



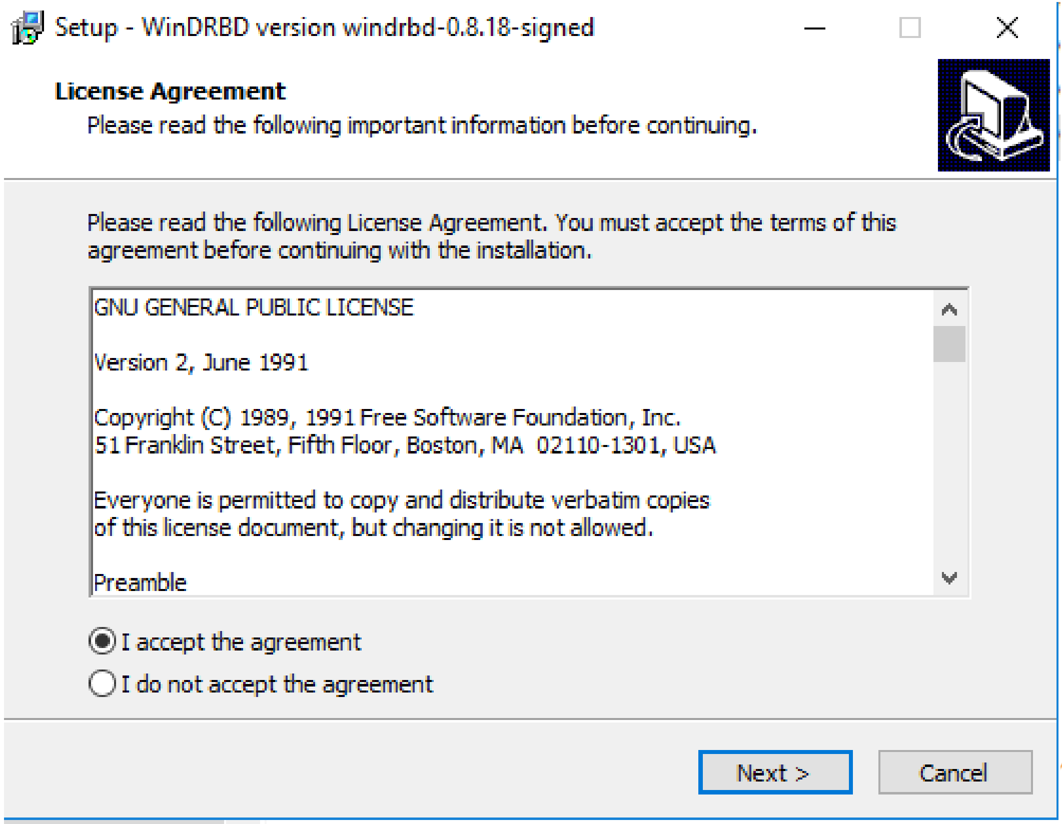
DRBD - more about

- a node knows the version of the data it 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 Roadmap

- performance optimizations
  - meta-data on PMEM/NVDIMMS
  - zero copy receive on diskless (RDMA-transport)
  - no context switch send (RDMA & TCP transport)
  - Improve resync speed
- Eurostars grant: DRBD4Cloud
  - erasure coding (2019)
- Long distance replication
  - send data once over long distance to multple replicas

WinDRBD



License Agreement

Please read the following important information before continuing.

Please read the following License Agreement. You must accept the terms of this agreement before continuing with the installation.

GNU GENERAL PUBLIC LICENSE

Version 2, June 1991

Copyright (C) 1989, 1991 Free Software Foundation, Inc.
51 Franklin Street, Fifth Floor, Boston, MA 02110-1301, USA

Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.

Preamble

● I accept the agreement

○ I do not accept the agreement

WinDRBD

- in public beta
  - https://www.linbit.com/en/drbd-community/drbd-download/
- Windows 7sp1, Windows 10, Windows Server 2016
- wire protocol compatible to Linux version
- driver tracks Linux version with one day release offset
- WinDRBD user level tools are merged into upstream

WinDRBD ROADMAP 2019

- fix multiple connections (Februar)
- add auto-promote (March)
- enable WinDRBD for boot and drive C: (March, April)
- review/rework spinlock & RCU primitives (May)
- POCs with customers (starting in July)

LINSTOR

The combination is more than the sum of its parts

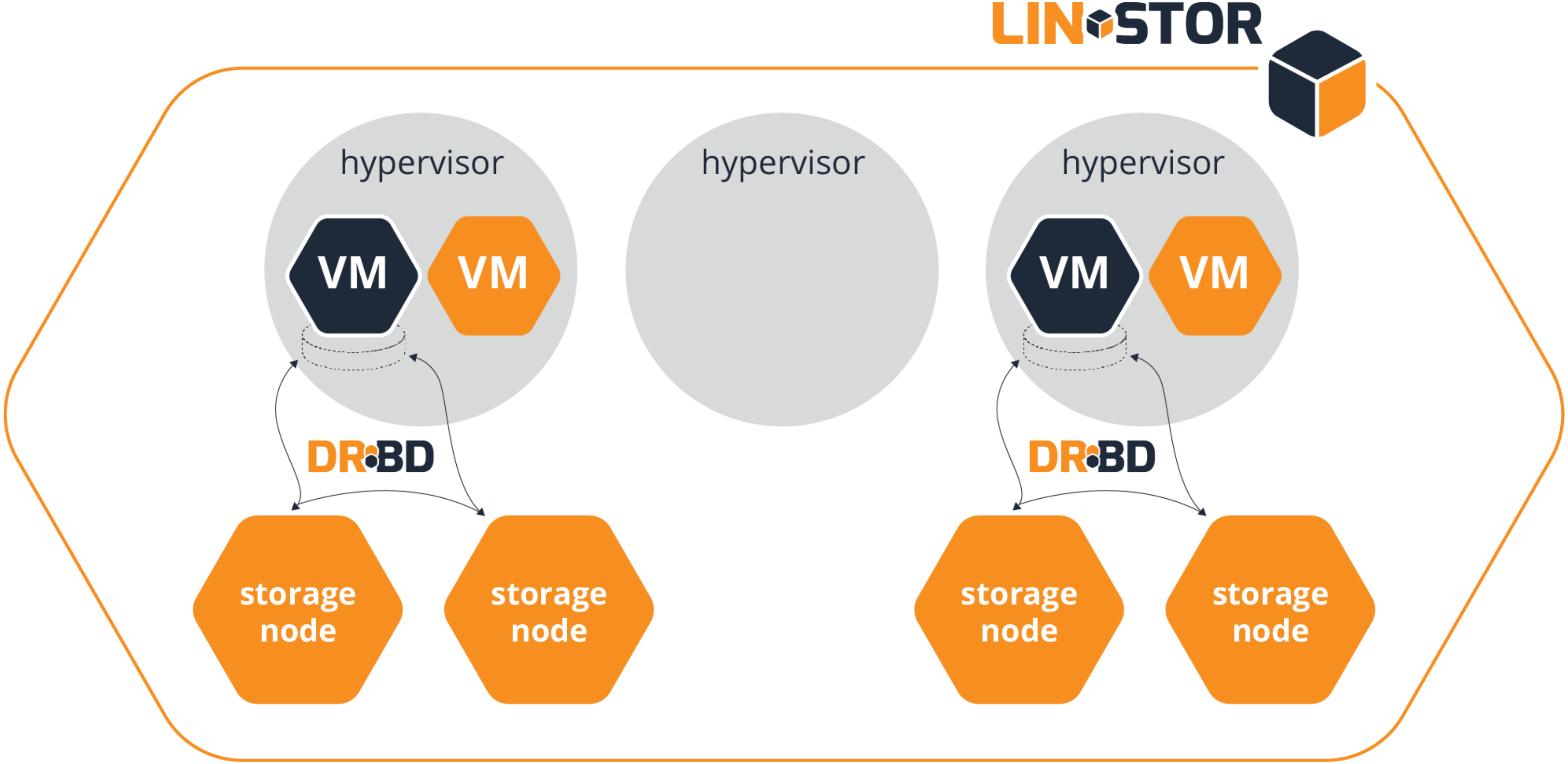
LINSTOR - goals

- storage built 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**

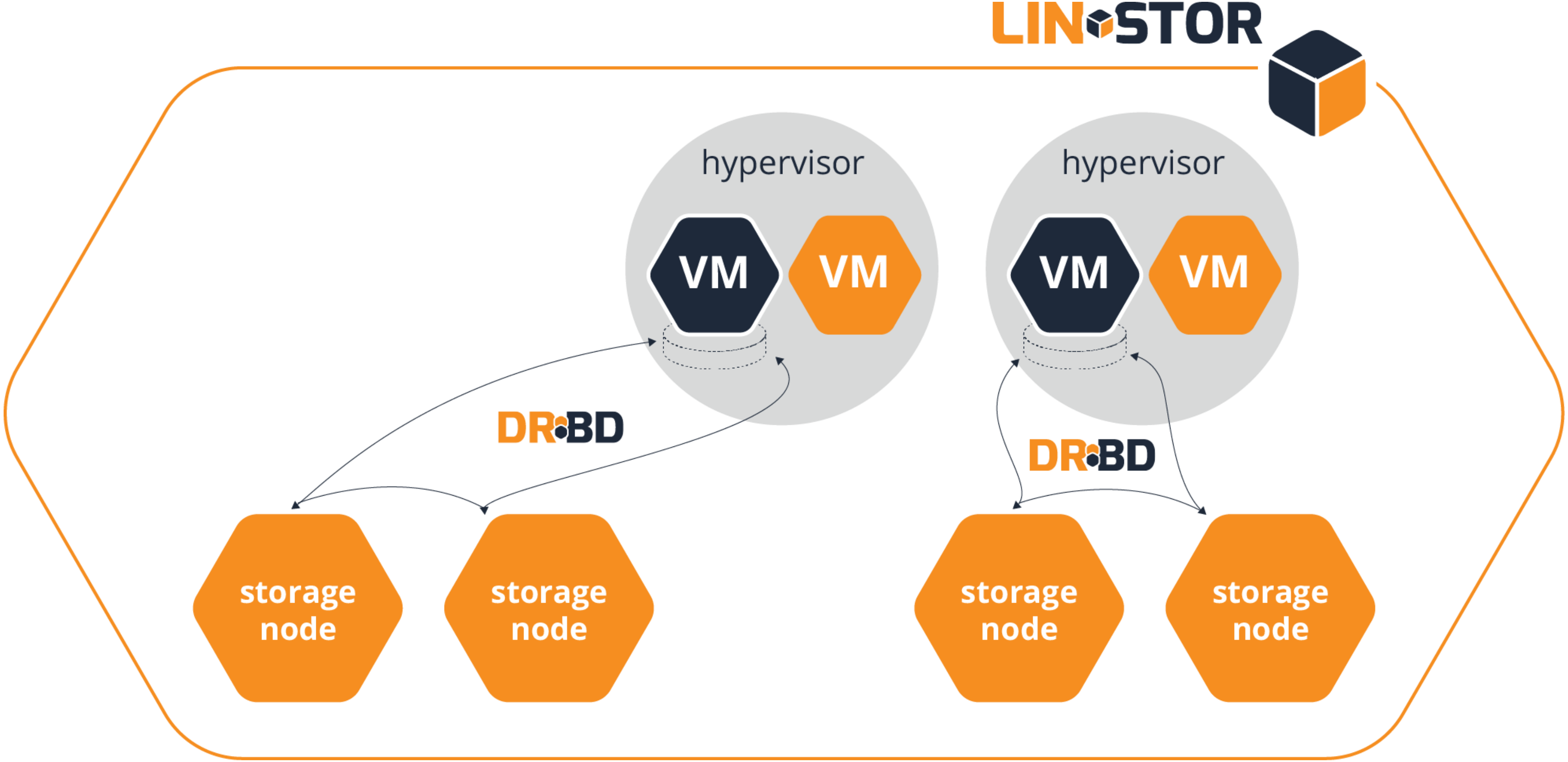
LINSTOR

Examples

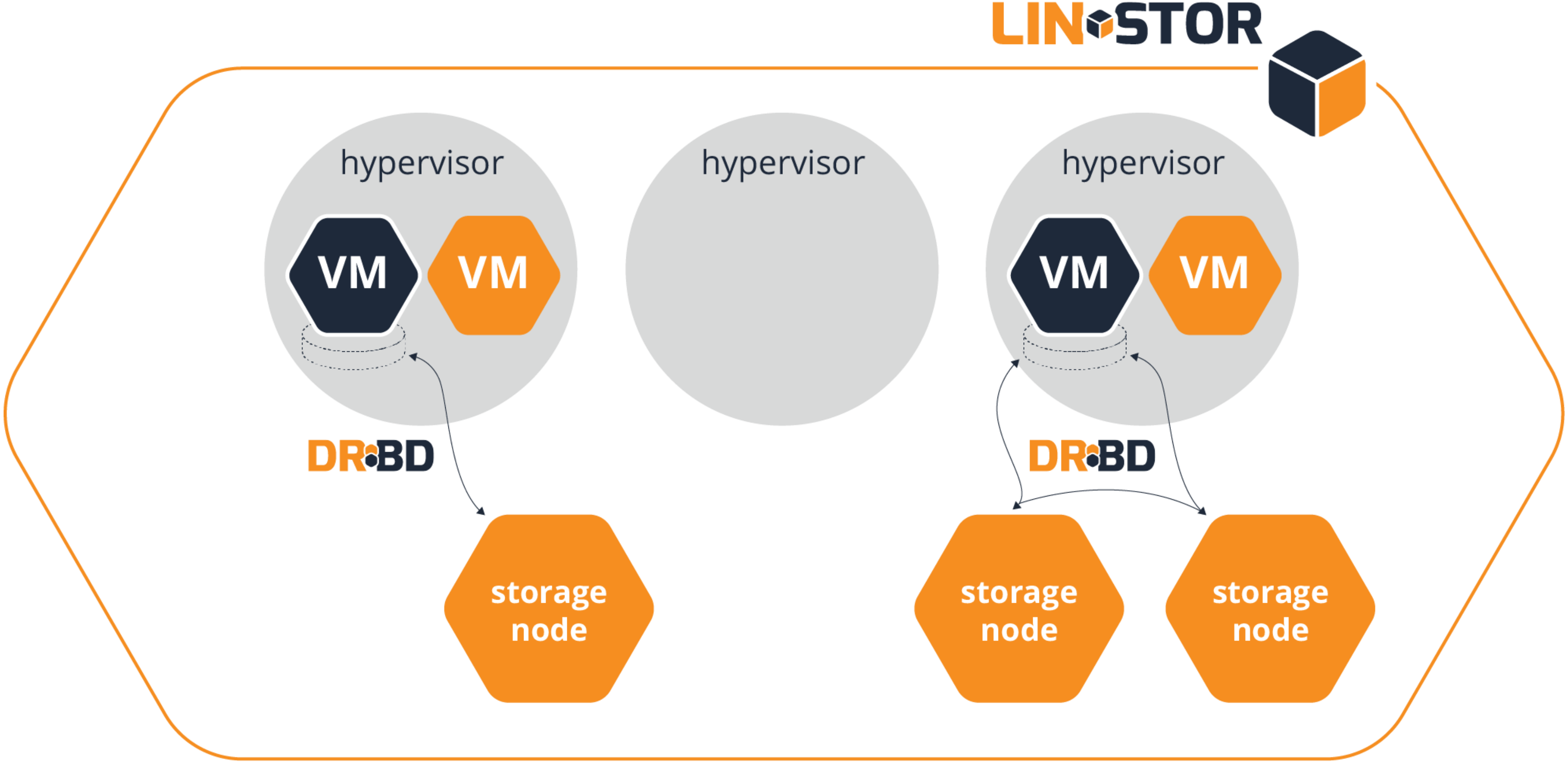
LINSTOR – disaggregated stack



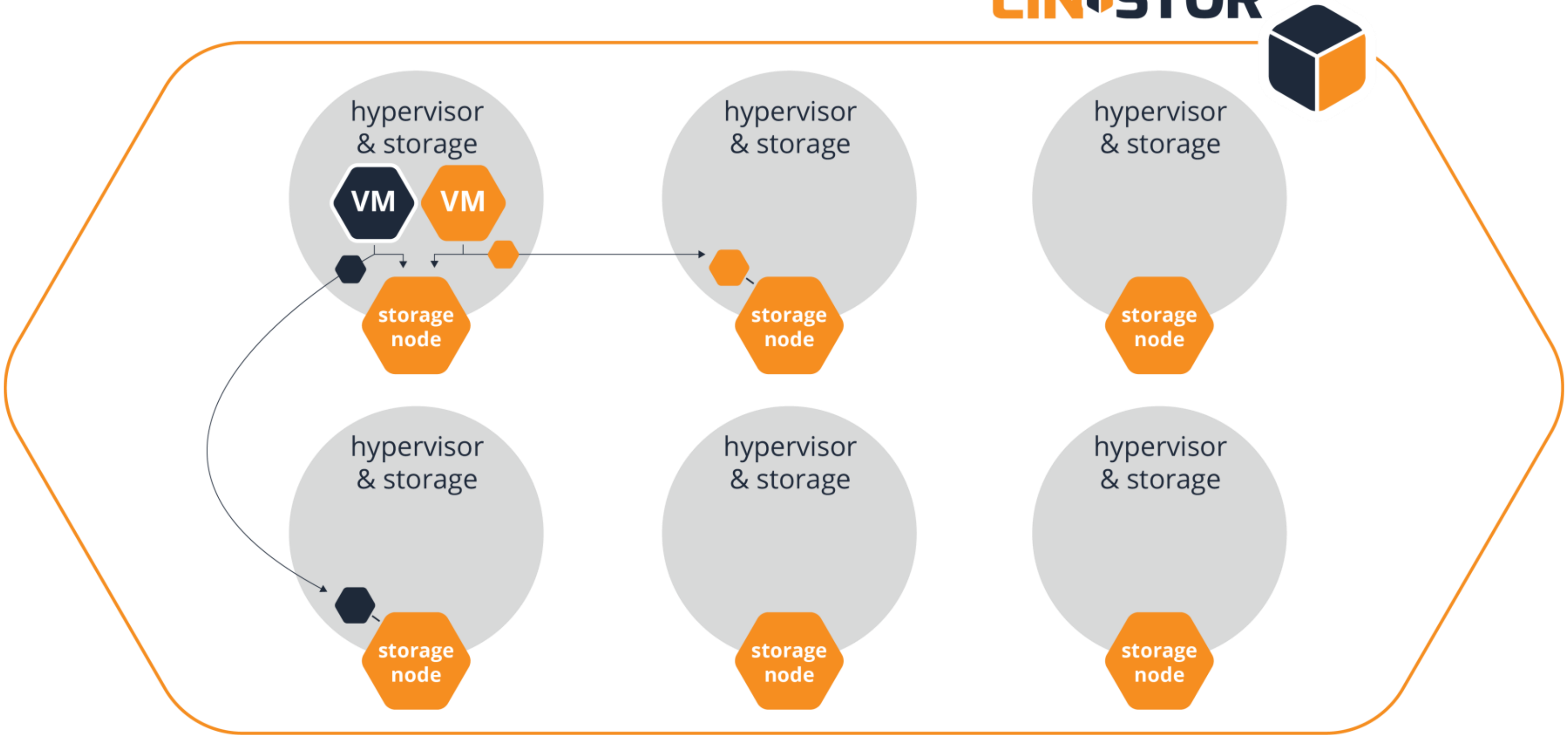
LINSTOR / failed Hypervisor



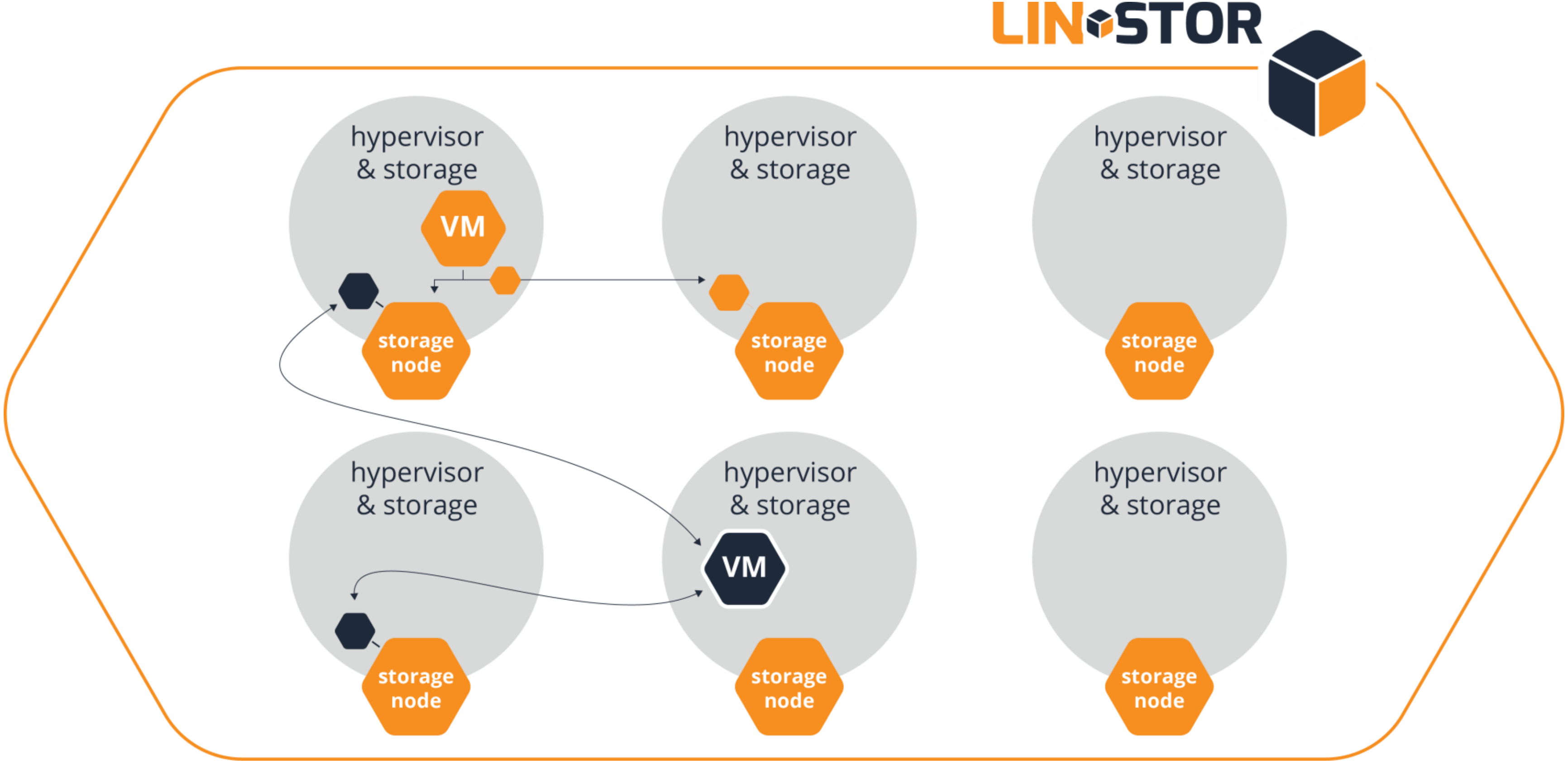
LINSTOR / failed storage node



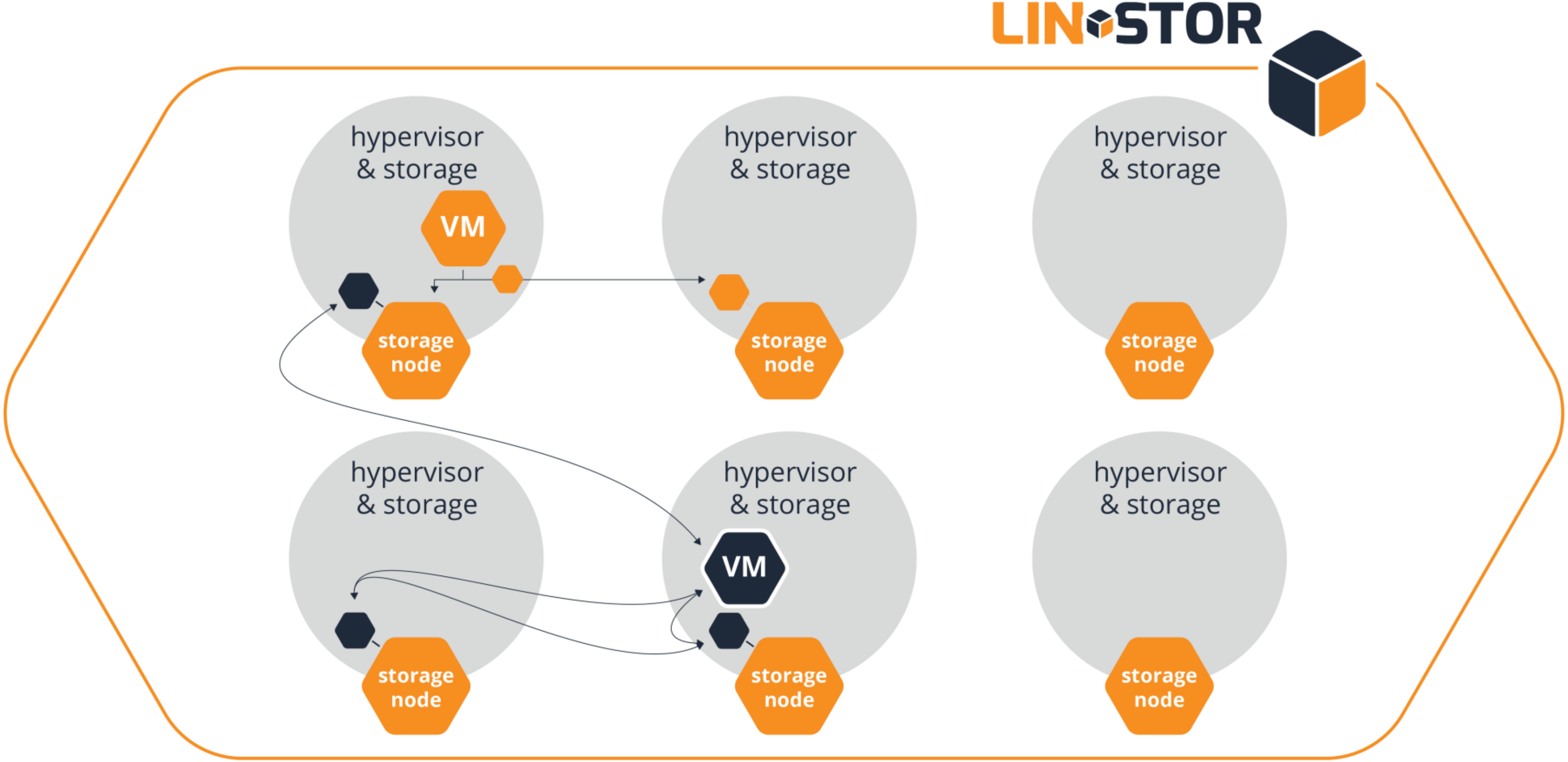
LINSTOR - Hyperconverged



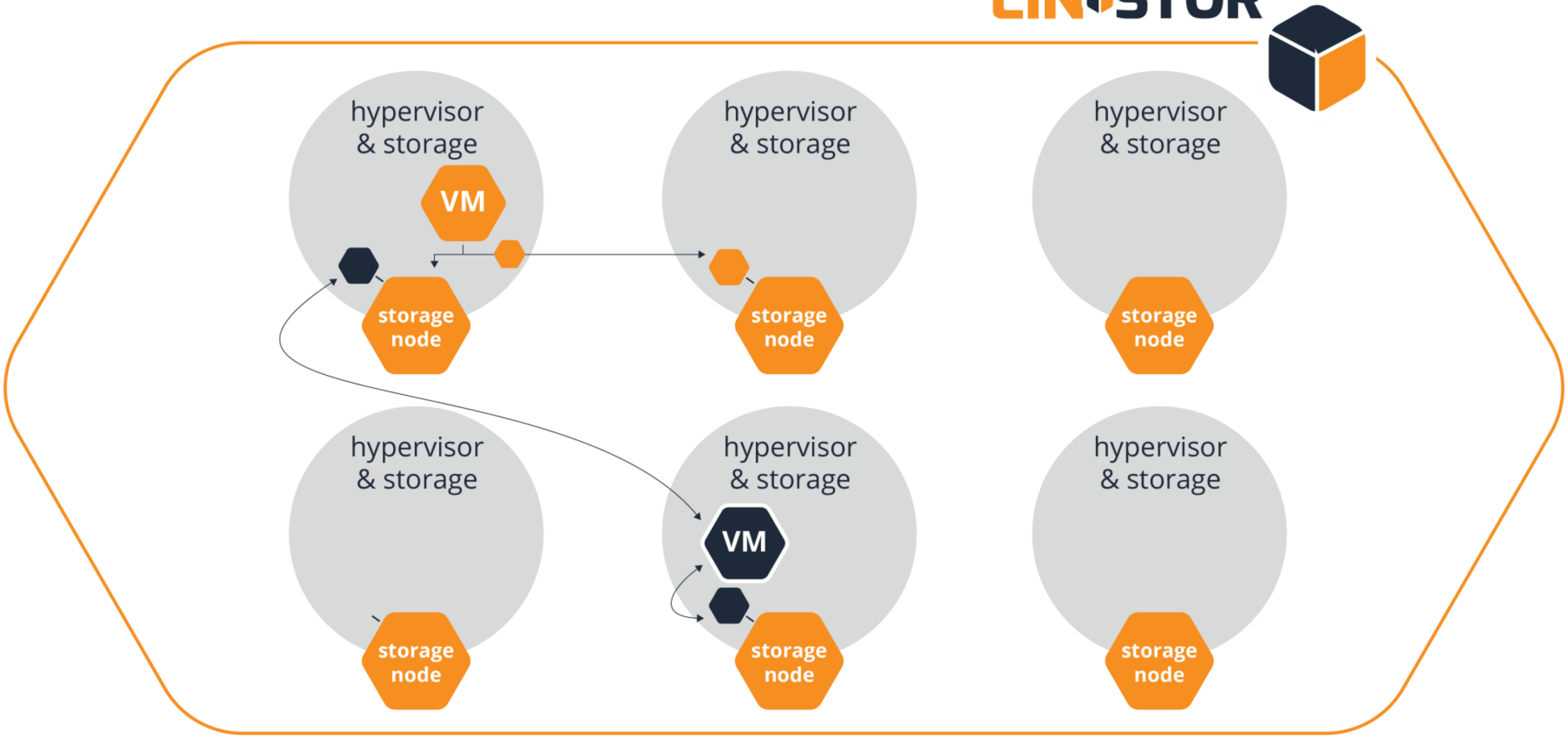
LINSTOR - VM migrated



LINSTOR - add local replica

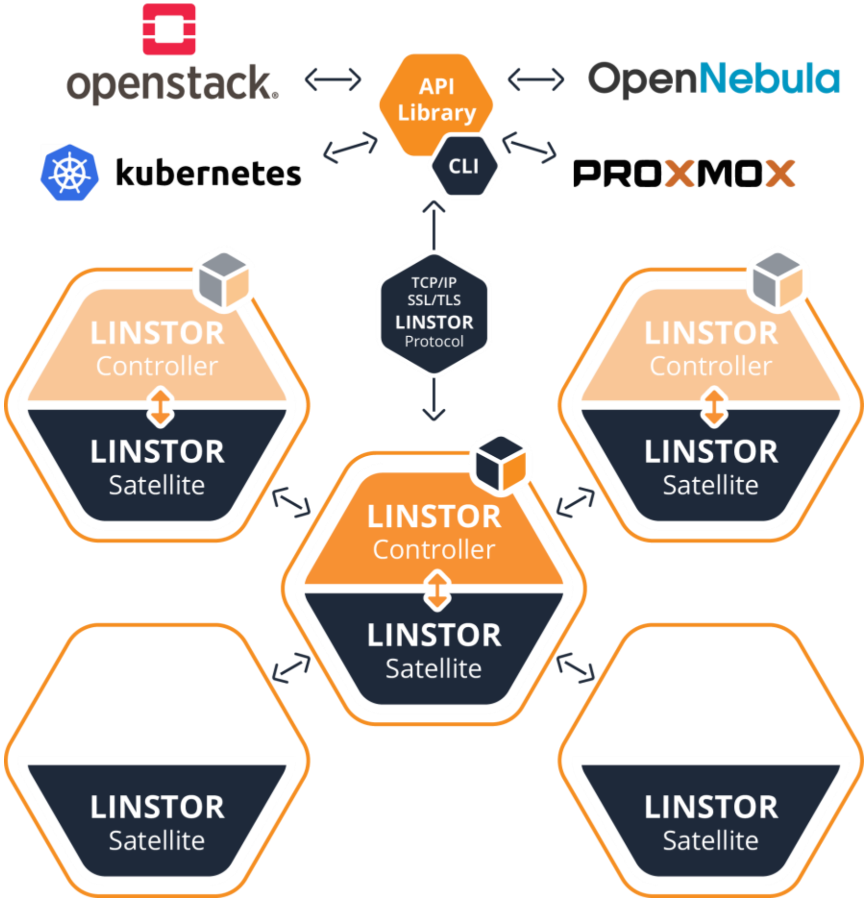


LINSTOR - remove 3rd copy



LINSTOR

Architecture and functions


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

LINSTOR data placement

- arbitrary tags on nodes
  - require placement on equal/different/named tag values
- prohibit placements with named existing volumes
  - different failure domains for related volumes

Example policy
3 way redundant, where two copies are in the same rack but in different fire compartments (synchronous) and a 3rd replica in a different site (asynchronous)

Example tags
rack = number
room = number
site = city

LINSTOR network path selection

- a storage pool may prefer a NIC
  - express NUMA relation of NVMe devices and NICs
- DRBD’s multi-pathing supported
  - load balancing with the RDMA transport
  - fail-over only with the TCP transport

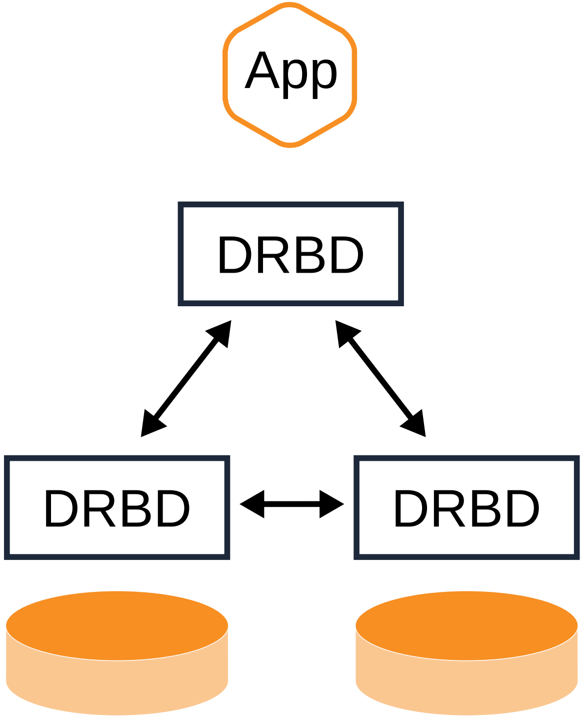
LINSTOR connectors

- Kubernetes
  - FlexVolume & External Provisioner
  - CSI (Docker Swarm, Mesos)
- OpenStack/Cinder
  - since Stein release (April 2019)
- OpenNebula
- Proxmox VE
- XenServer / XCP-ng

LINSTOR Roadmap

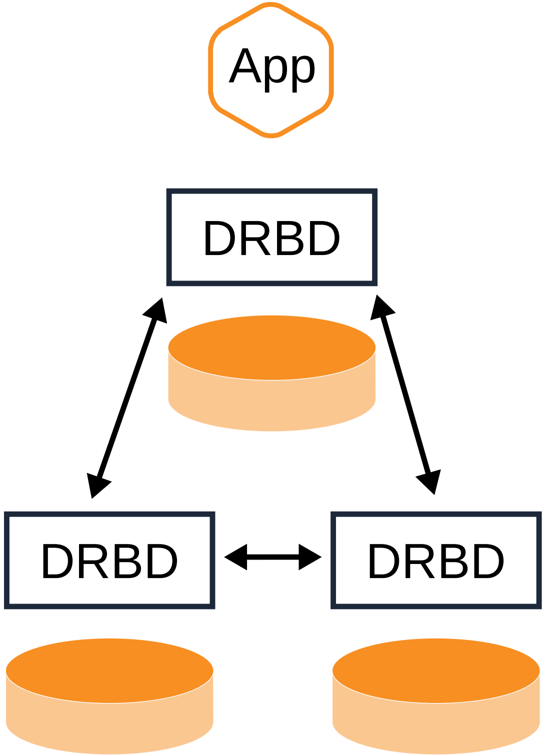
- generalize LINSTOR for other IO stacks
  - MD-Raid & NVMe-oF
  - optional HW discovery & VG automatic VG creation
  - bcache & deduplication
- Linux NVMe-oF initiator & targets
- GUI based on REST-API
- auto-placement policies as LINSTOR objects
- LINSTOR & WinDRBD (?)

LINSTOR Storage Stacks



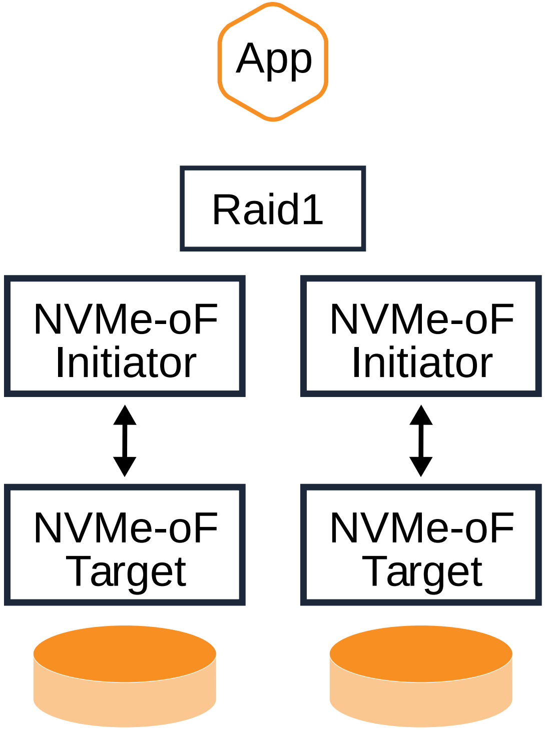
- Disaggregated Storage
- Classic enterprise workloads
  - Data bases
  - Message queues
- Typical Orchestrators
  - OpenStack, OpenNebula
  - Kubernetes
- Flexible redundancy (1-n)
- HDDs, SSDs, NVMe SSDs

LINSTOR Storage Stacks



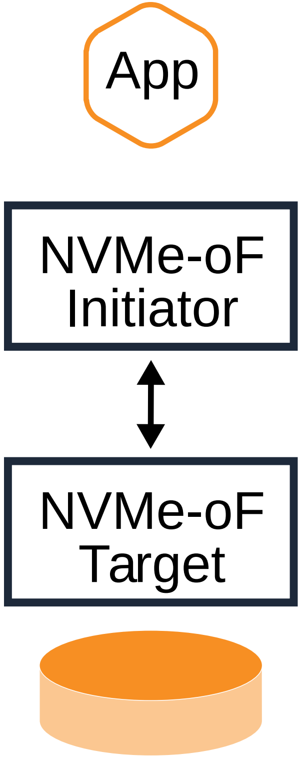
- Hyperconverged
- Classic enterprise workloads
  - Data bases
  - Message queues
- Typical Orchestrators
  - OpenStack, OpenNebula
  - Kubernetes
- Flexible redundancy (1-n)
- HDDs, SSDs, NVMe SSDs

LINSTOR Storage Stacks



- Disaggregated
- Classic enterprise workloads
  - Data bases
  - Message queues
- Typical Orchestrators
  - OpenStack, OpenNebula
  - Kubernetes
- NVMe SSDs, SSDs

LINSTOR Storage Stacks

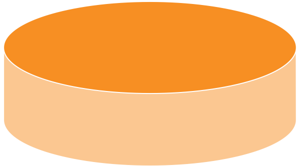


- Disaggregated
- Cloud native workload
  - Ephemeral storage
- Typical Orchestrator
  - Kubernetes
- Application handles redundancy
- Best suited for NVMe SSDs

LINSTOR Storage Stacks

- Hyperconverged
- Cloud native workload
  - Ephemeral storage
  - PMEM optimized data base
- Typical Orchestrator
  - Kubernetes
- Application handles redundancy
- PMEM, NVDIMMs

LINSTOR Slicing Storage



- LVM or ZFS
- Thick – pre allocated
  - Best performance
  - Less features
- Thin – allocated on demand
  - Overprovisioning possible
  - Many snapshots possible
- Optional
  - Encryption on top
  - Deduplication below

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.

Thank you

https://www.linbit.com