LINBIT

LINSTOR

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

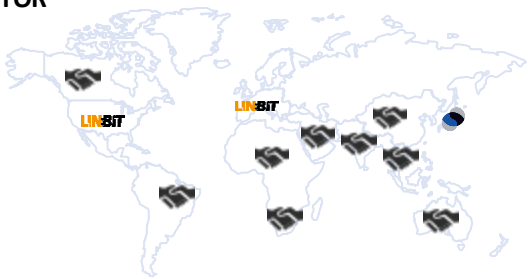
Philipp Reisner, CEO LINBIT

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)

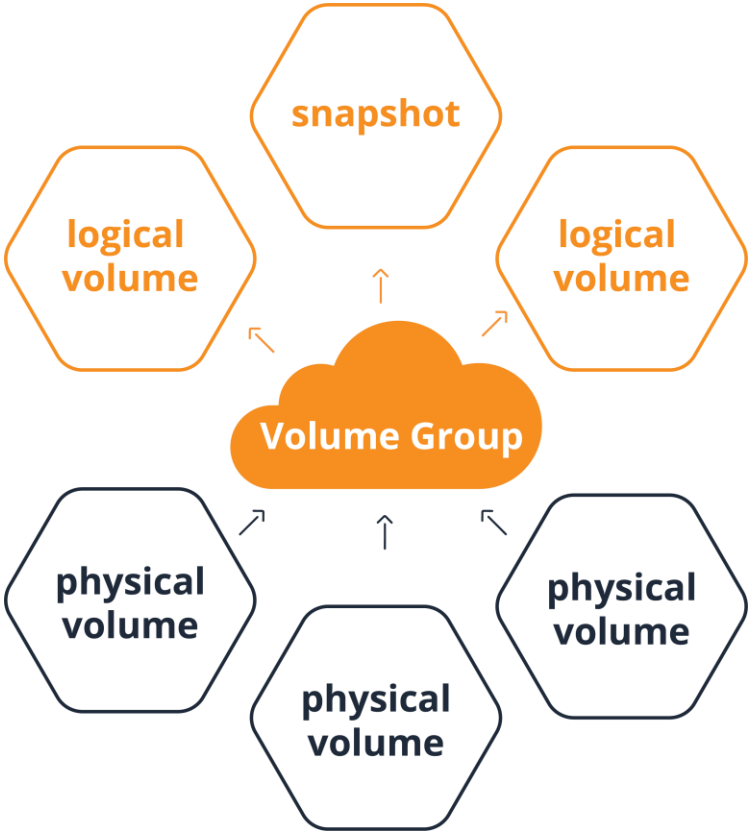
Market leading solutions since 2001, over 600 customers

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

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

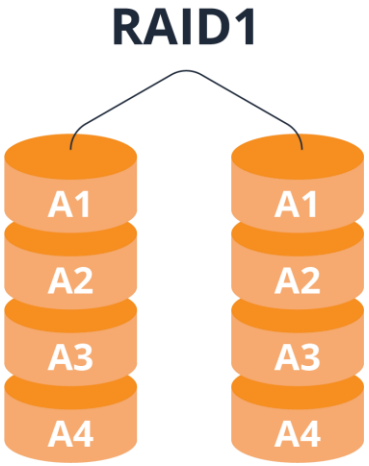
LINBIT

- 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
snapshot
thin-sLV
logical volume
thinpool
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
- letd, STGT, SCST
  - mostly historical
- **LIO**
  - iSCSI, iSER, SRP, FC, FCoE
  - SCSI pass through, block IO, file IO, user-specific-IO
- NVMe-OF
  - target & initiator

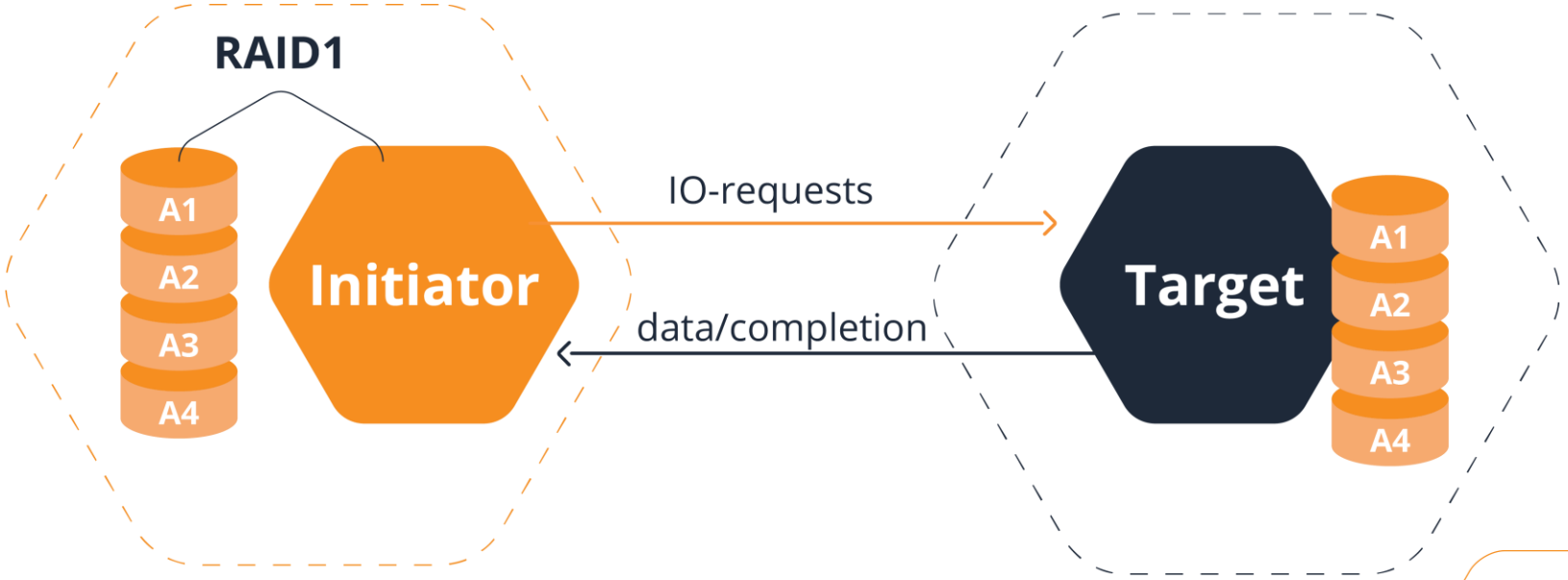
ZFS on Linux

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

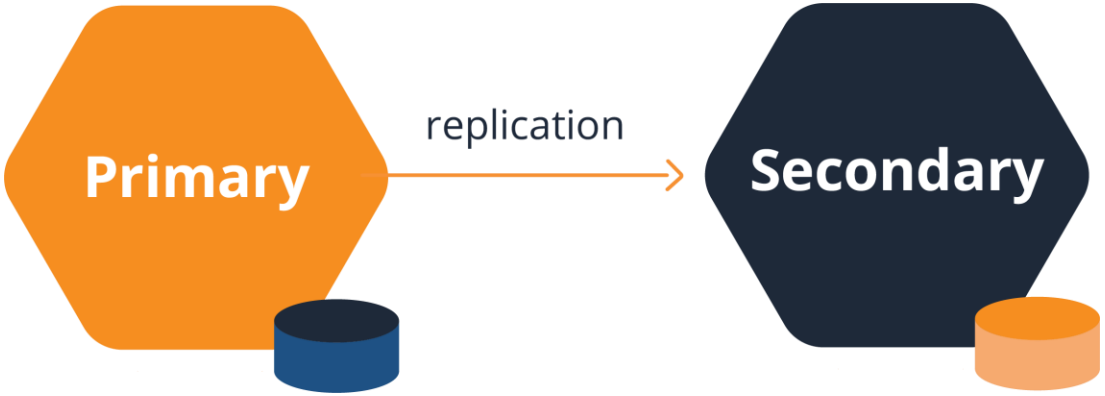
DRBD

Put in simplest form

DRBD – think of it as …

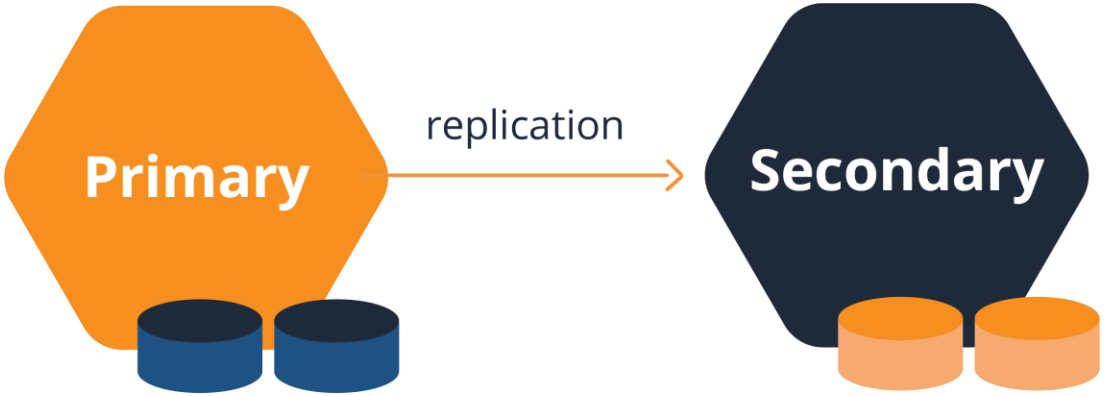


DRBD Roles: Primary & Secondary



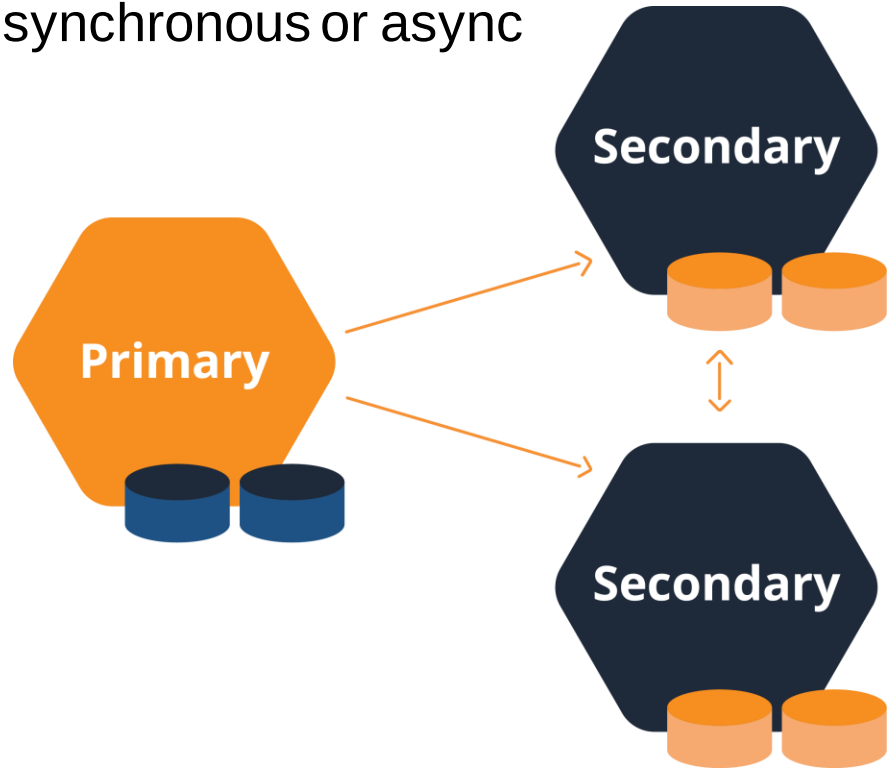
DRBD – multiple Volumes

- consistency group



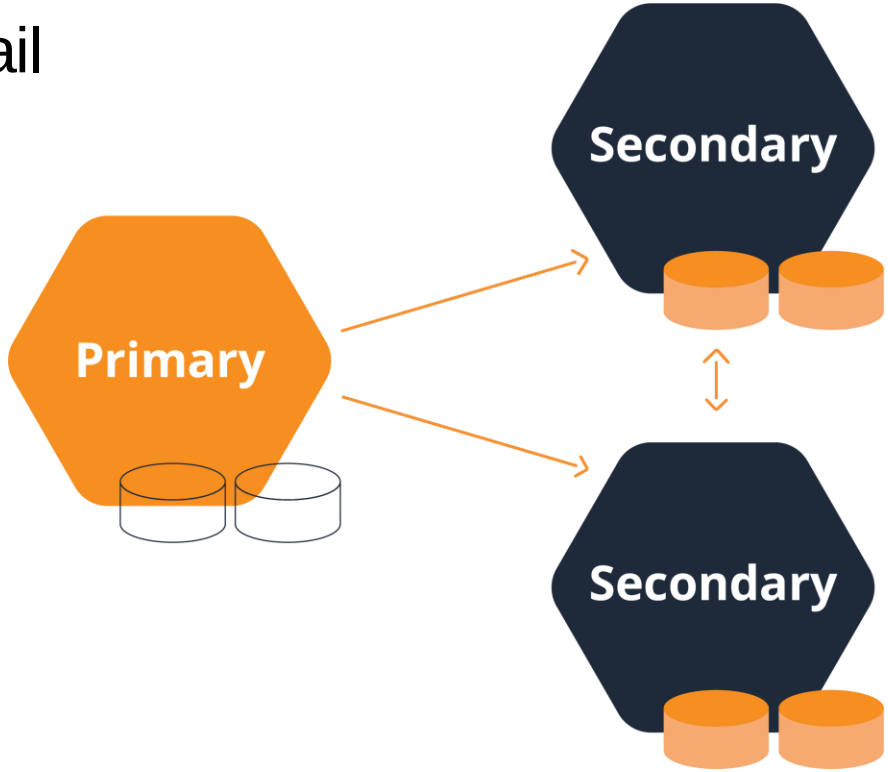
DRBD – up to 32 replicas

- each may be synchronous or async



DRBD – Diskless nodes

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



DRBD - more about

- 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

- 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

LINSTOR

The combination is more than the sum of its parts

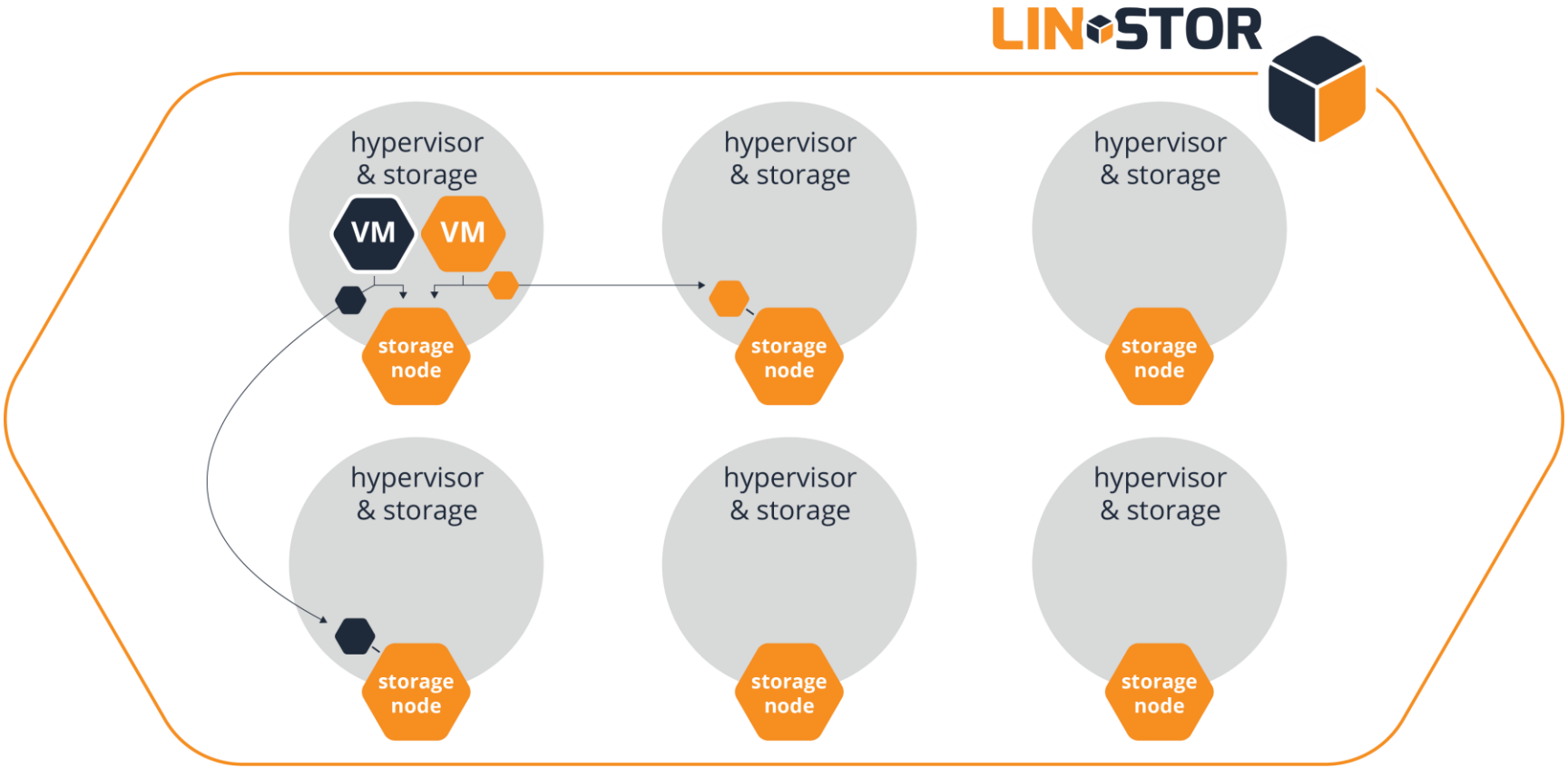
LINSTOR - goals

- 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**

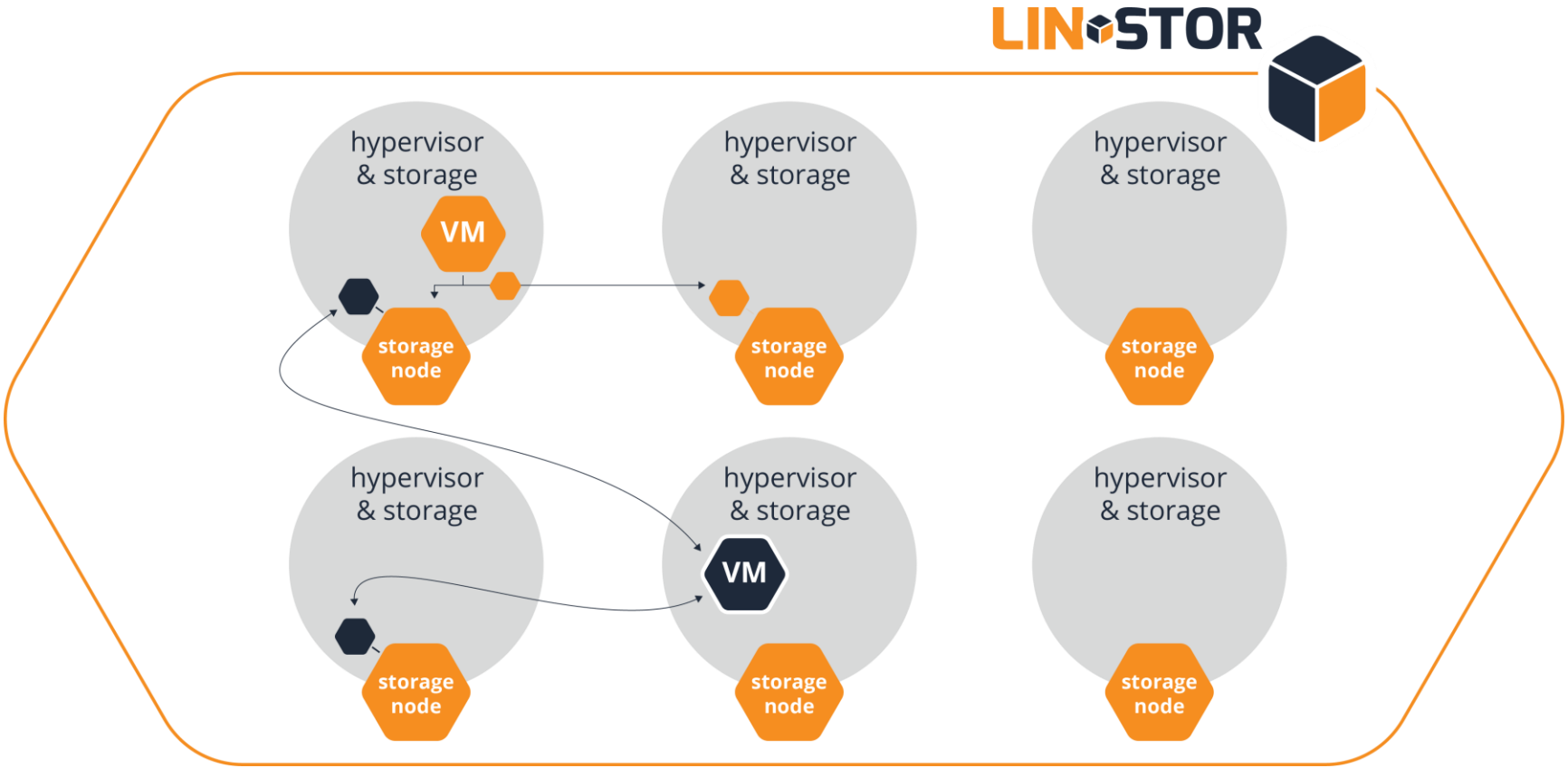
LINSTOR

Examples

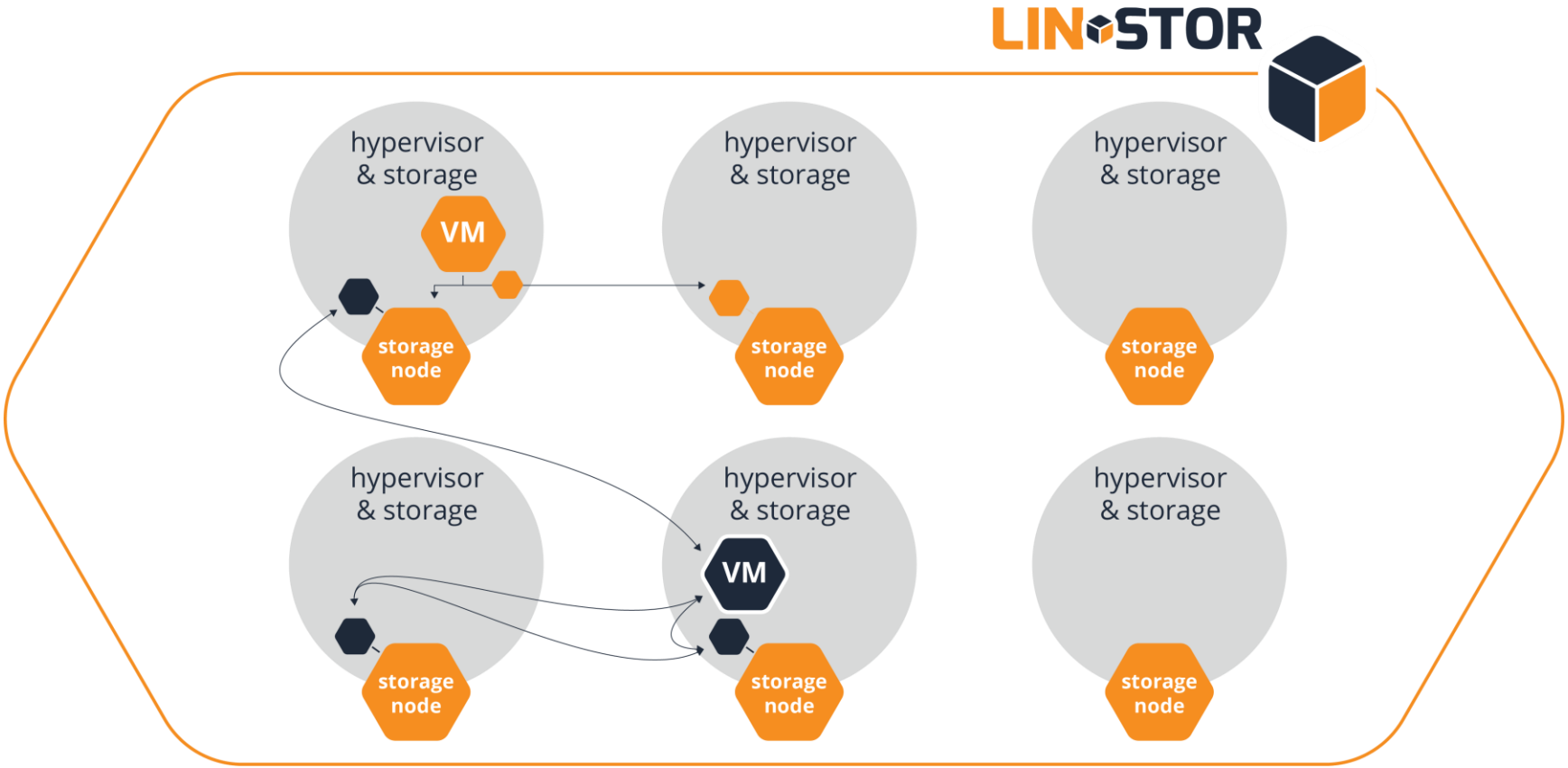
LINSTOR - Hyperconverged



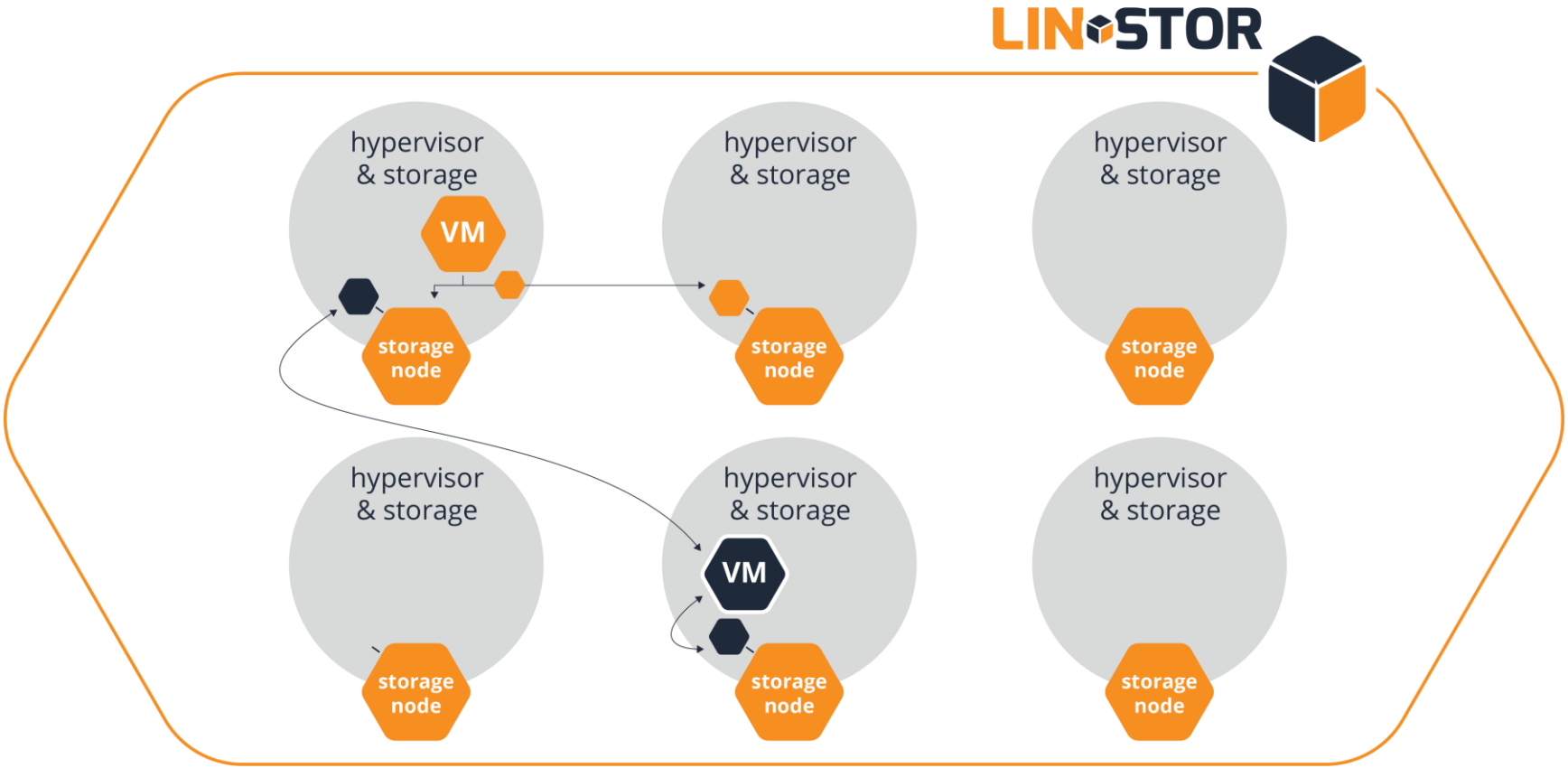
LINSTOR - VM migrated



LINSTOR - add local replica

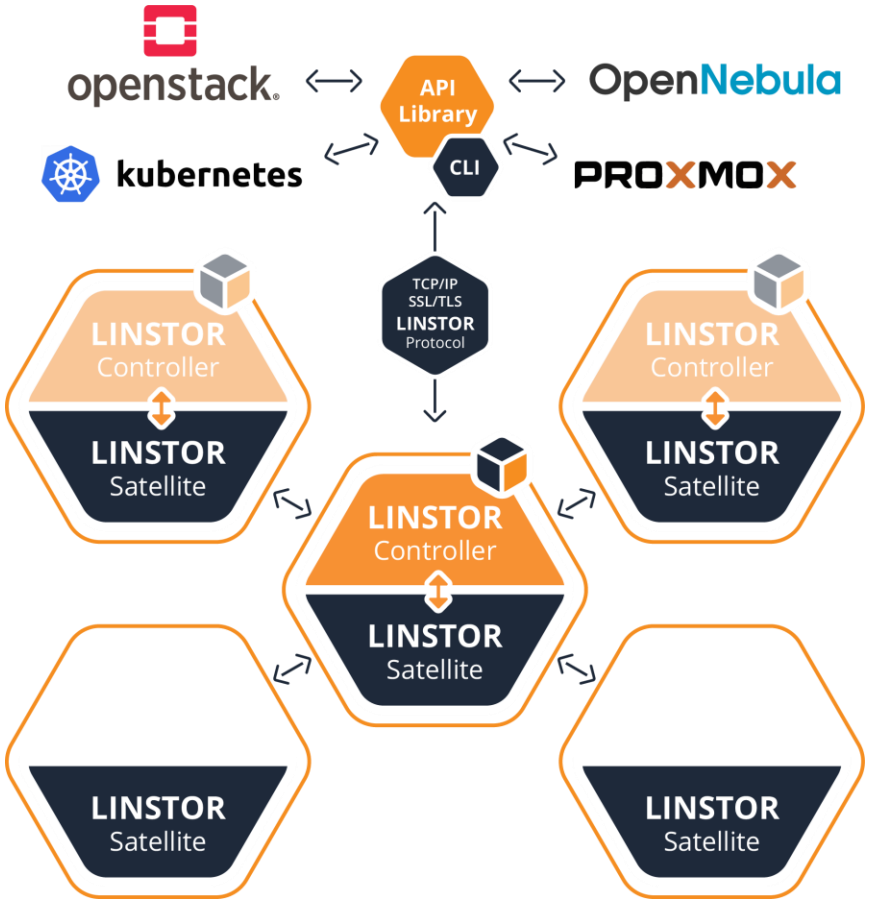


LINSTOR - remove 3rd copy



LINSTOR

Architecture and functions


openstack.
kubernetes
API Library
CLI
OpenNebula
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 diffeent 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 preferred 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

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

LINSTOR SDS & Piraeus Datastore

	LINBIT SDS	Piraeus Datastore
Container base Img	Red hat UBI	Debian
Available	drbd.io LINBIT customers only	dockerhub, quay.io publicly
Support	✓ Enterprise, incl 24/7	Community only
OpenShift/RHCOS	✓	n.a.
DBRD driver	Pre-compiled for RHEL kernels	Compile from source
Contains	LINSTOR, DRBD, operator, YAML-files, Helm chart, CSI-driver	

LINSTOR – Kubernetes

Kubernetes	LINSTOR
Storage Class	Resource Group
Persistent Volume	Resource / Volume

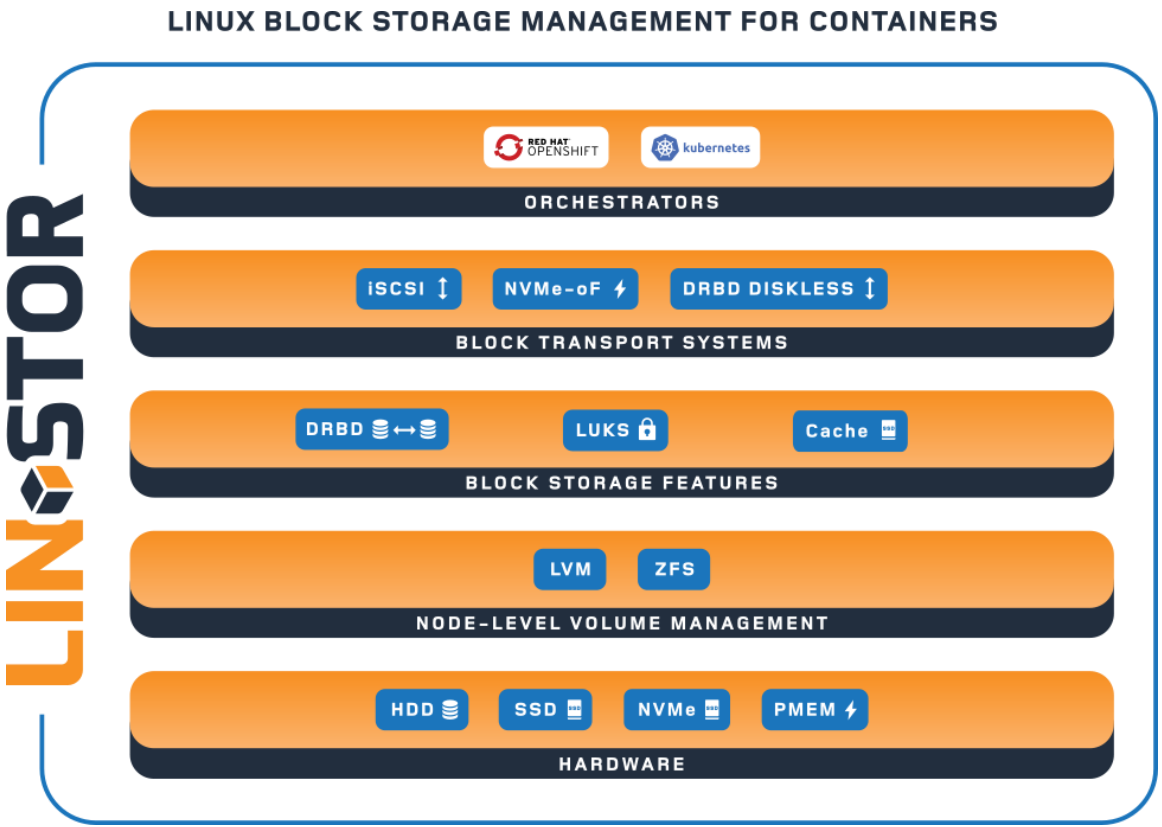
Case study - intel

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.

Summary

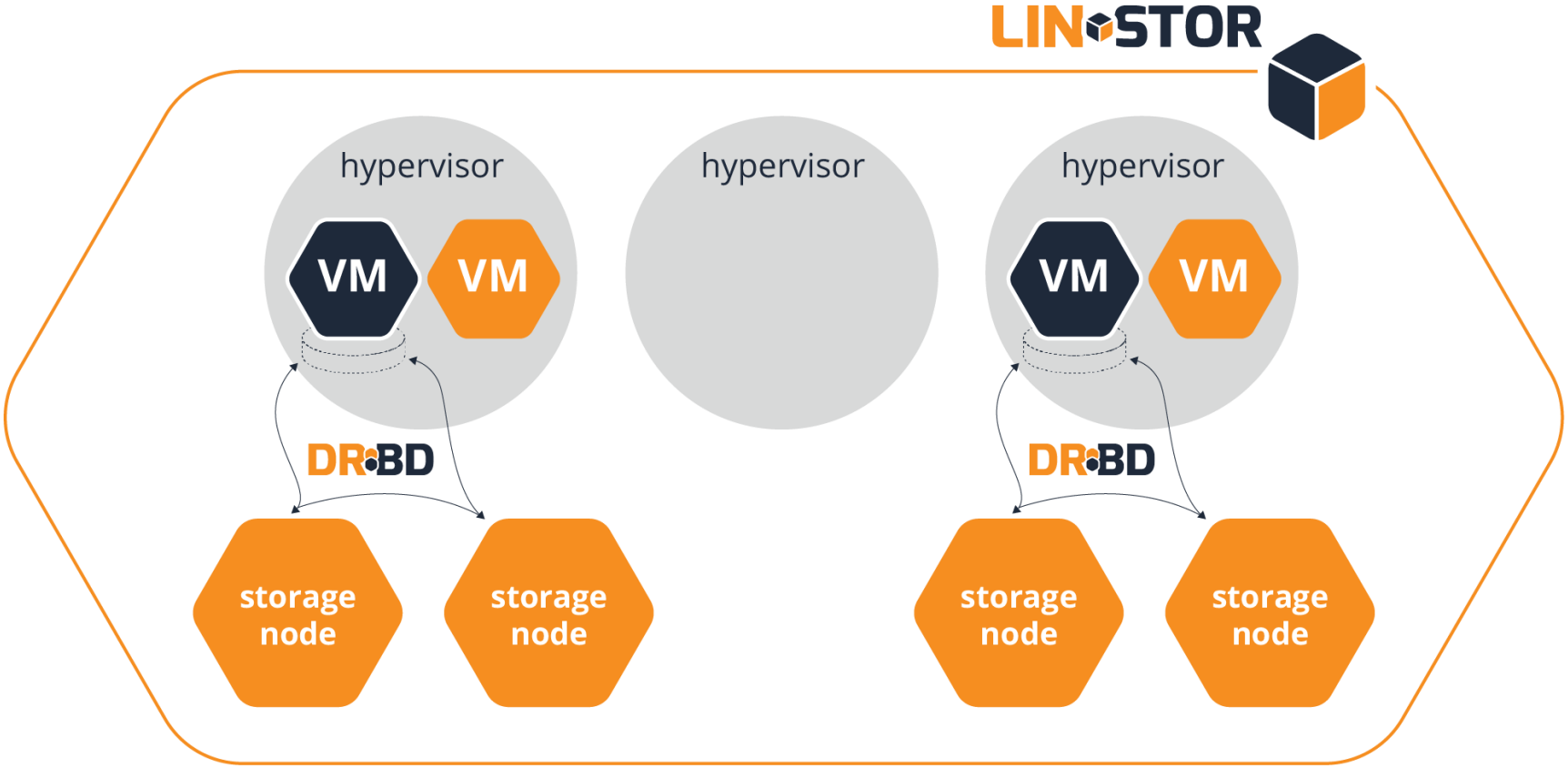

LINUX BLOCK STORAGE MANAGEMENT FOR CONTAINERS
LINSTOR
RED HAT OPENSHIFT
kubernetes
ORCHESTRATORS
iSCSI
NVMe-oF
DRBD DISKLESS
BLOCK TRANSPORT SYSTEMS
DRBD
LUKS
Cache
BLOCK STORAGE FEATURES
LVM
ZFS
NODE-LEVEL VOLUME MANAGEMENT
HDD
SSD
NVMe
PMEM
HARDWARE

Thank you

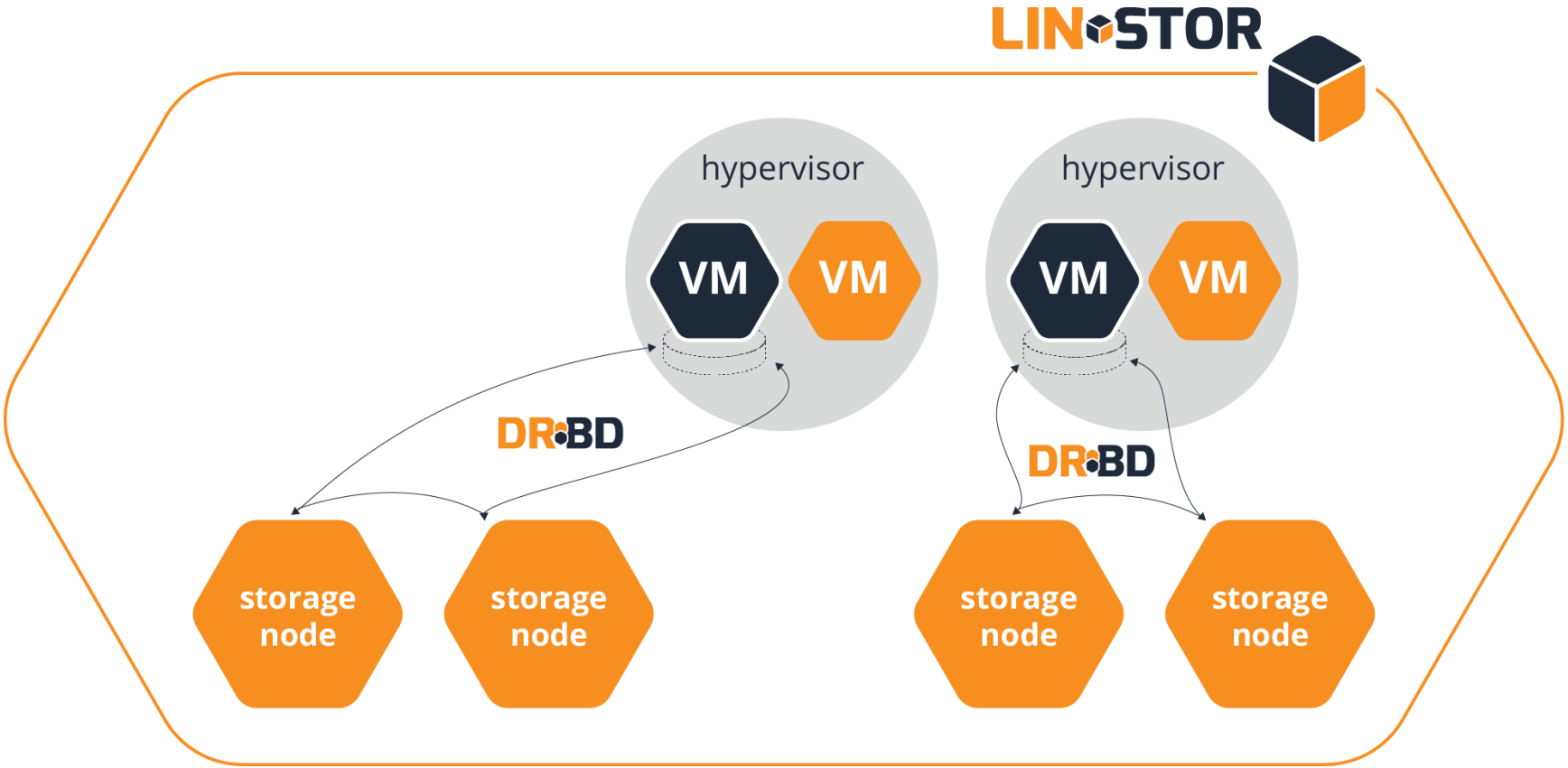
https://www.linbit.com

Appendix Slides: Example Disaggregated Architecture

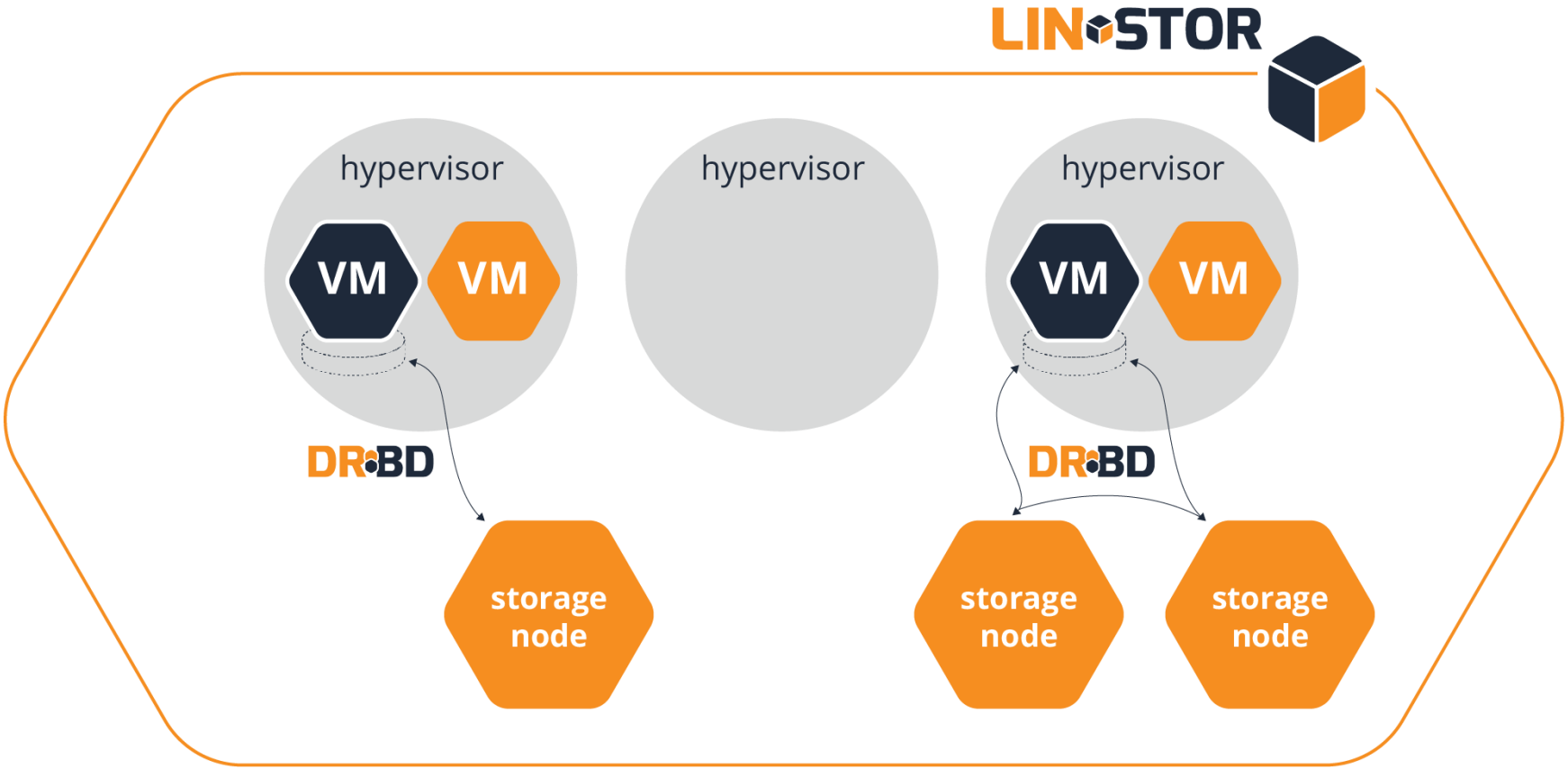
LINSTOR – disaggregated stack



LINSTOR / failed Hypervisor



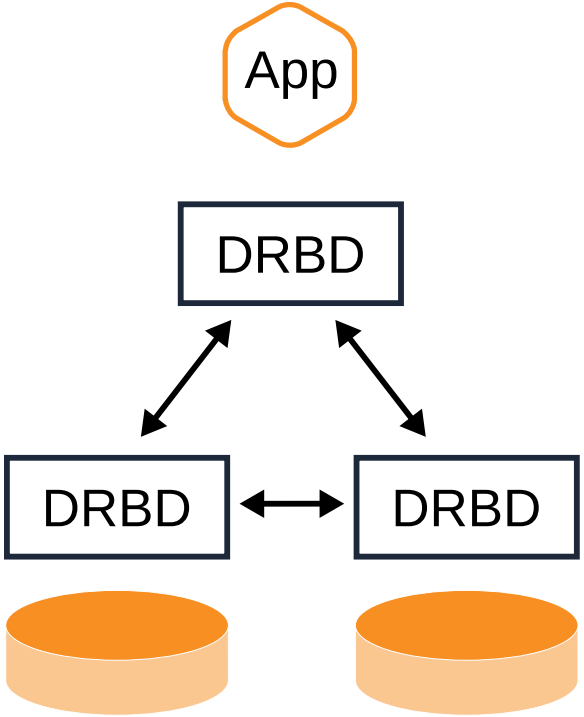
LINSTOR / failed storage node



LINSTOR

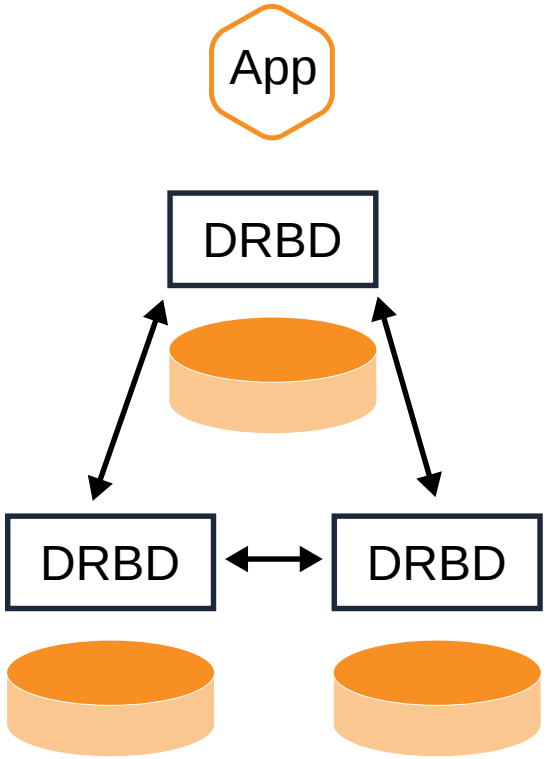
Appendix Slides: Possible Storage Stacks

LINSTOR Storage Stacks



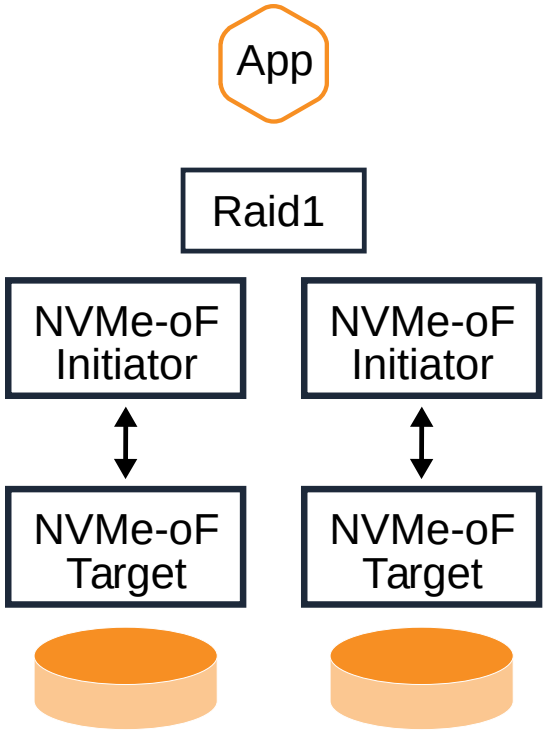
- Disaggregated Storage
- Classic enterprise workloads
  - Data bases
  - Message queues
- Typical Orchestrators
  - OpenStack, OpenNebula
  - Kubernetes
- Flexibly redundancy (1-n)
- HDDs, SSDs, NVMe SSDs

LINSTOR Storage Stacks



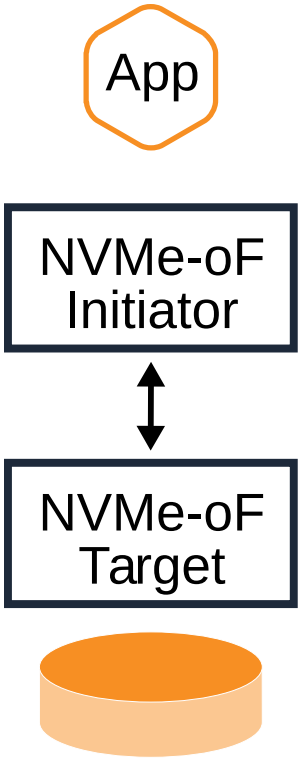
- Hyperconverged
- Classic enterprise workloads
  - Data bases
  - Message queues
- Typical Orchestrators
  - OpenStack, OpenNebula
  - Kubernetes
- Flexibly redundancy (1-n)
- HDDs, SSDs, NVMe SSDs

LINSTOR Storage Stacks



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

LINSTOR Storage Stacks

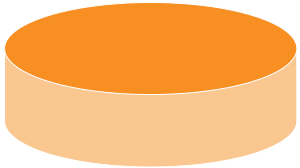


- Disaggregated
- Cloud native workload
  - Ephemeral storage
- Typical Orchestrator
  - Kubernetes
- Application handles redundancy
- Best suited for NVMe SSDs

LINSTOR Storage Stacks

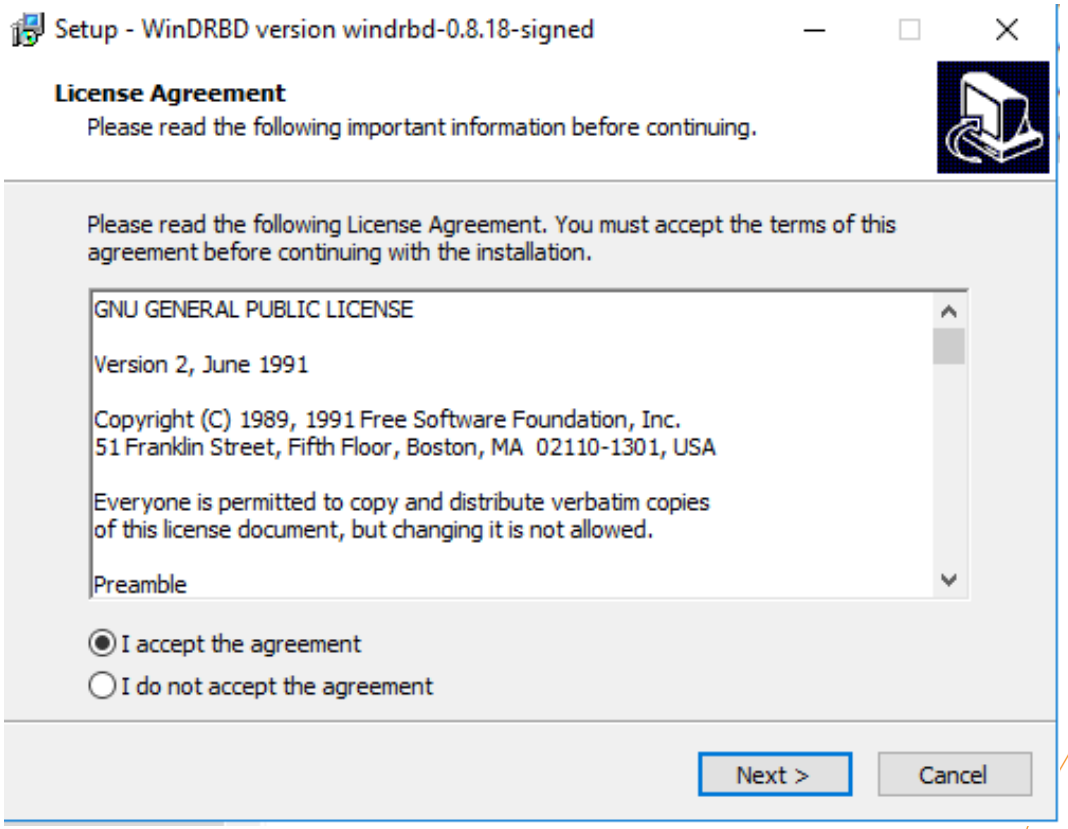
- Hyperconverged
- Cloud native workload
  - Ephemeral storage
  - PMEM optimized data base
- Typical Orchestrator
  - Kubernetes
- Application handles redundancy
- PMEM, NVDIMMs

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

WinDRBD


Setup - WinDRBD version windrbd-0.8.18-signed
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
Next >
Cancel

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