LINBIT

×

HashiCorp Nomad

STORAGE DAY

LINBIT Software Introduction

LINBIT SDS × HashiCorp Nomad

STORAGE DAY

LINBIT

Philipp Reisner

CEO

LINBIT

Rene Peinthor

Solution Architect

LINBIT SDS

When? Why? What?

When is LINBIT SDS a fit?

Persistent Volumes

...for Containers

- Kubernetes
- Nomad
- Docker

Virtualization

- OpenStack
- CloudStack
- OpenNebula
- XCP-ng
- Proxmox

Transaction Processing

- Oracle DB
- PostgreSQL
- MariaDB
- Message queuing systems

Analytic Processing

- DB2 Warehouse
- And similar read intensive workloads

LINBIT SDS × HashiCorp Nomad
STORAGE DAY

Why is LINBIT SDS so fast?

In Kernel data-path

- Reduce number of context switches
- Saving on CPU/memory resources
- Minimal latency for block-IO operations
- Optional load-balancing for READs

Layout at volume allocation

- All participating machines have full replicas, which machines participate determined when creating a volume.
- Be faster at IO submission time
- Saving on CPU/memory

Build on existing components

- DRBD, LVM, ZFS, LUKS, VDO, ...
- Help day2 operations by leveraging on the opera tion teams prior knowledge
- Build on the shoulders of giants

Hyper-Converged

Very well suitable for hyper-converged deployment

- Reduced network load for reads
- Reduces latency
- LINBIT SDS' Low resource consumption leaves most of CPU and memory for workload. About 0.5% of a single core are consumed by DRBD under heavier IO load (measured with an analytics DB)

LINBIT SDS × HashiCorp Nomad

STORAGE DAY

What is LINBIT SDS doing?

Storage Allocation

- 3 to 1000s of nodes
- Multiple tiers
- Multi tenancy
- Complex policies Chassis - rack - room

Data Replication

- Persistence & availability
- Sync / async
- 2,3 or more replicas
- Consistency groups
- Quorum

Network

- Multiple NICs per server
- Multiple networks
- RDMA
- TCP

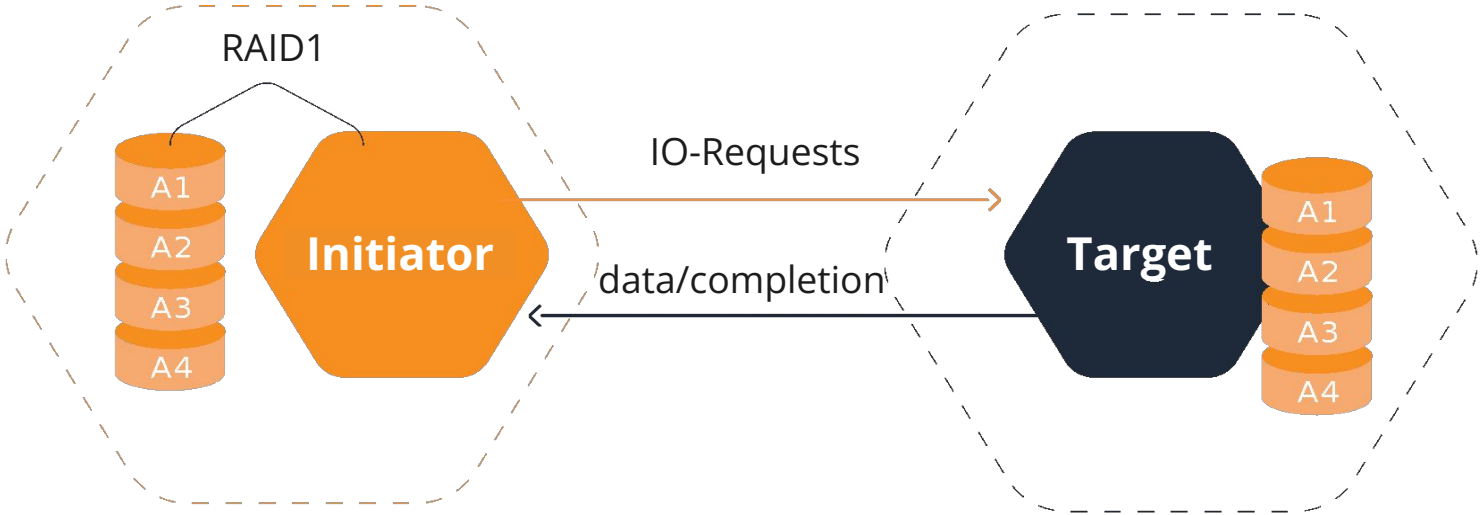
Business continuity

- Continuous data protection
- Multiple sites
- Backups
  SSD - disk - cloud

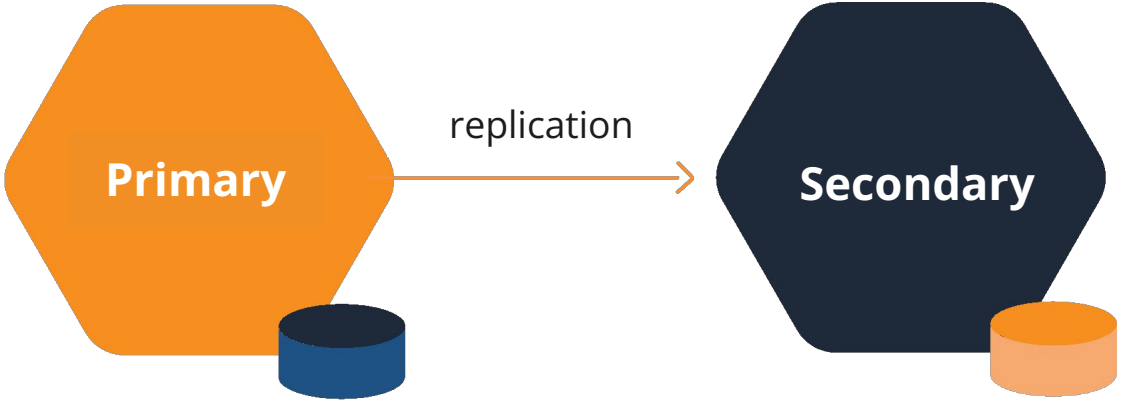
DRBD

Put in simplest form

DRBD – think of it as ...

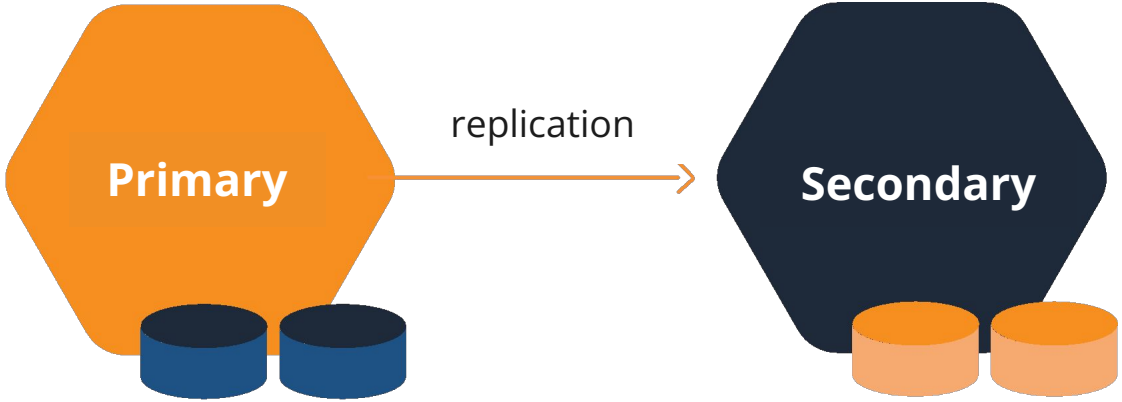


DRBD Roles: Primary & Secondary



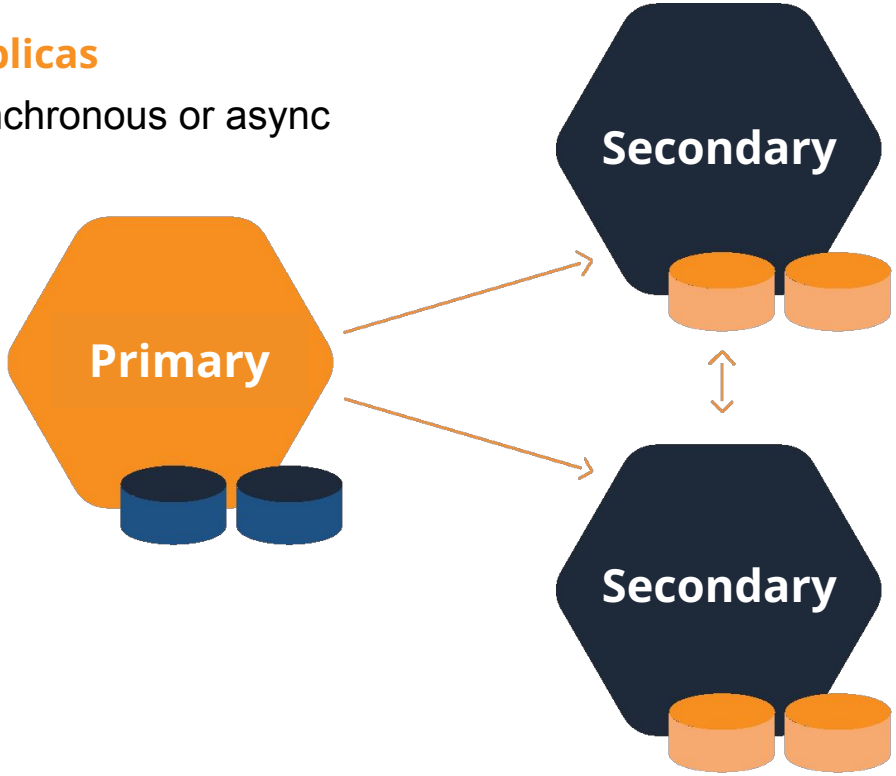
DRBD – multiple Volumes

- consistency group



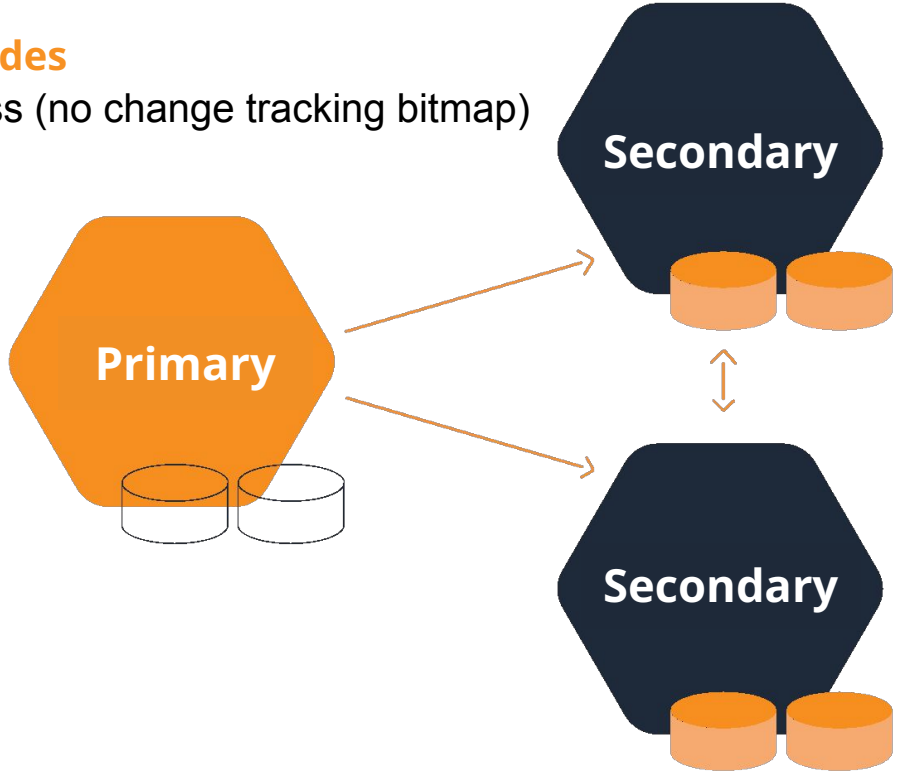
DRBD – up to 32 replicas

- 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 resources per node possible (1000s)
- dual Primary for live migration of VMs only!

DRBD - recent features & ROADMAP

- Recent
  - meta-data on PMEM/NVDIMMS
  - improved, fine-grained locking for parallel workloads
  - Eurostars grant: DRBD4Cloud
  - started DRBD-9.1
- ROADMAP
  - performance optimizations
  - replace „stacking”
  - production release of WinDRBD

LINSTOR

The combination is more than the sum of its parts

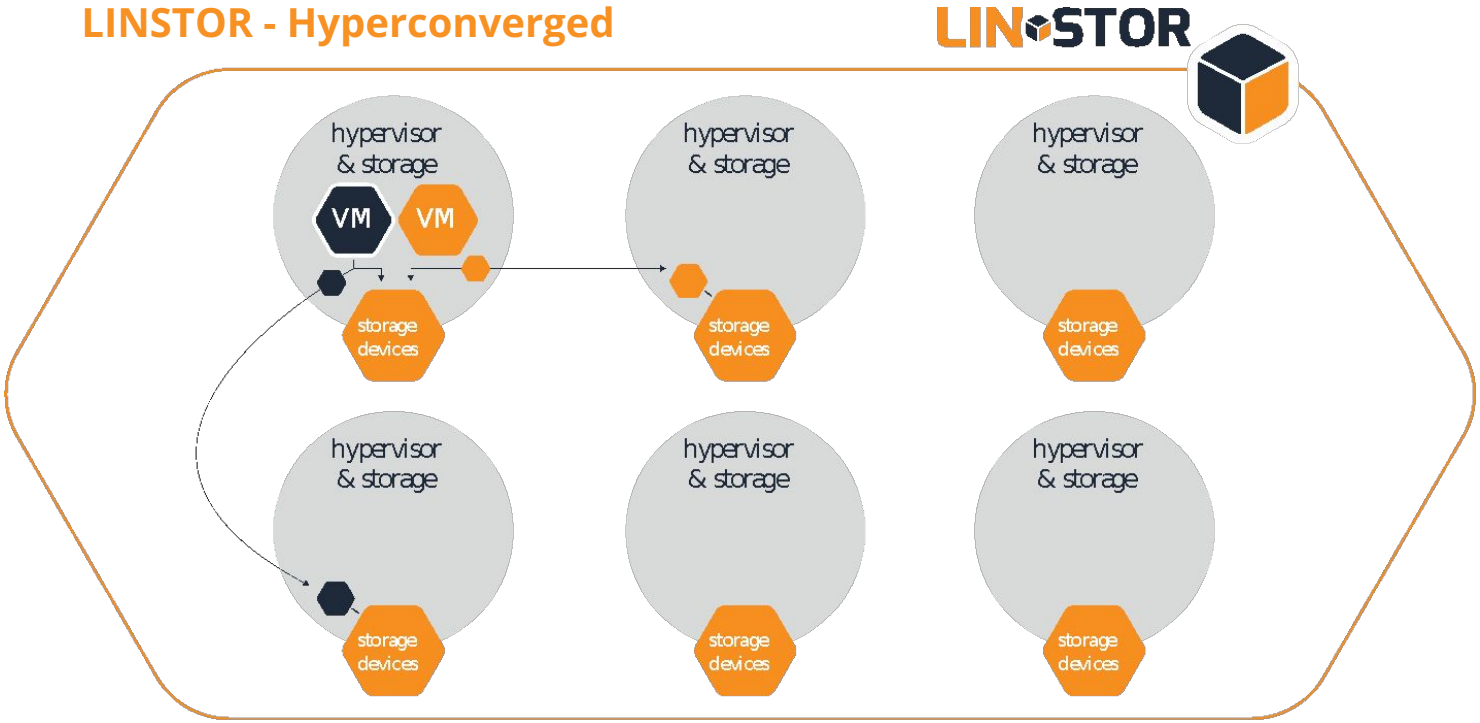
LINSTOR - goals

- storage build from generic Linux 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**

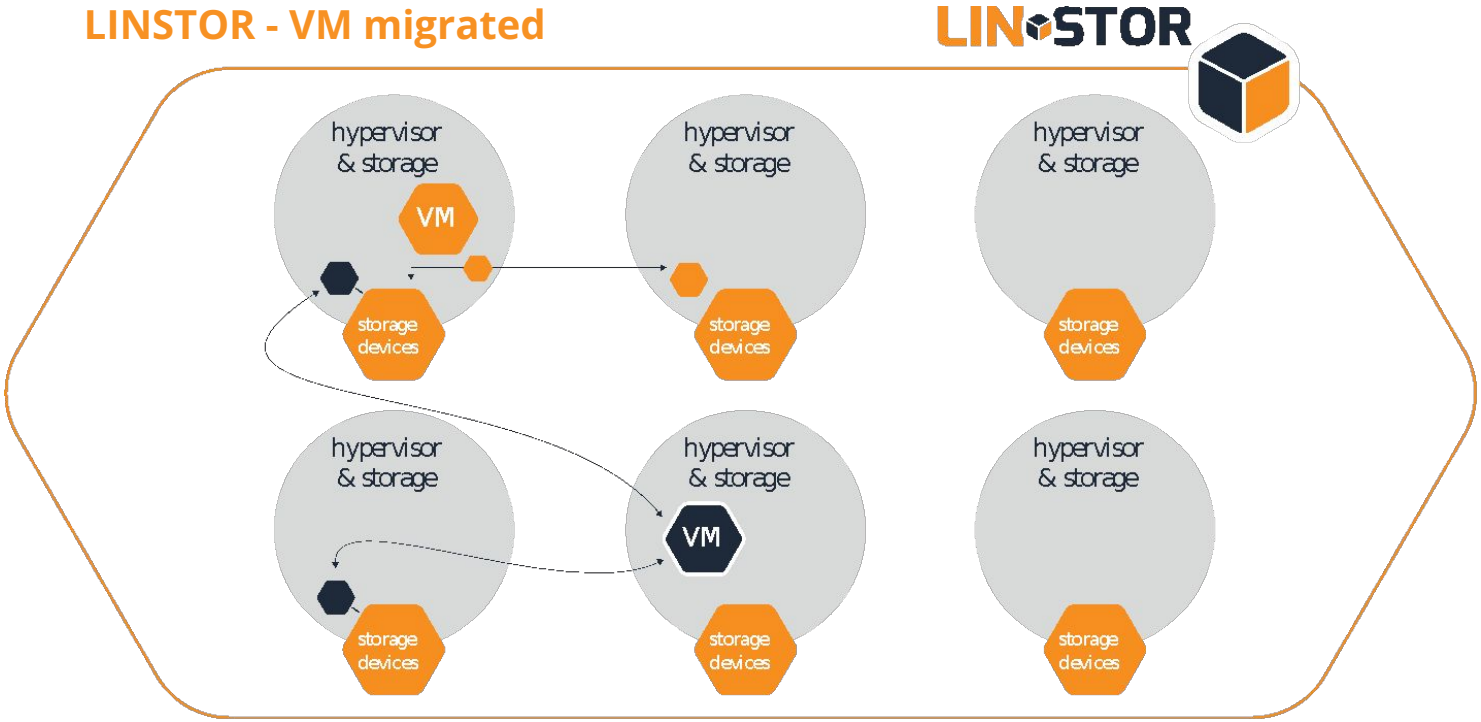
LINSTOR

Example

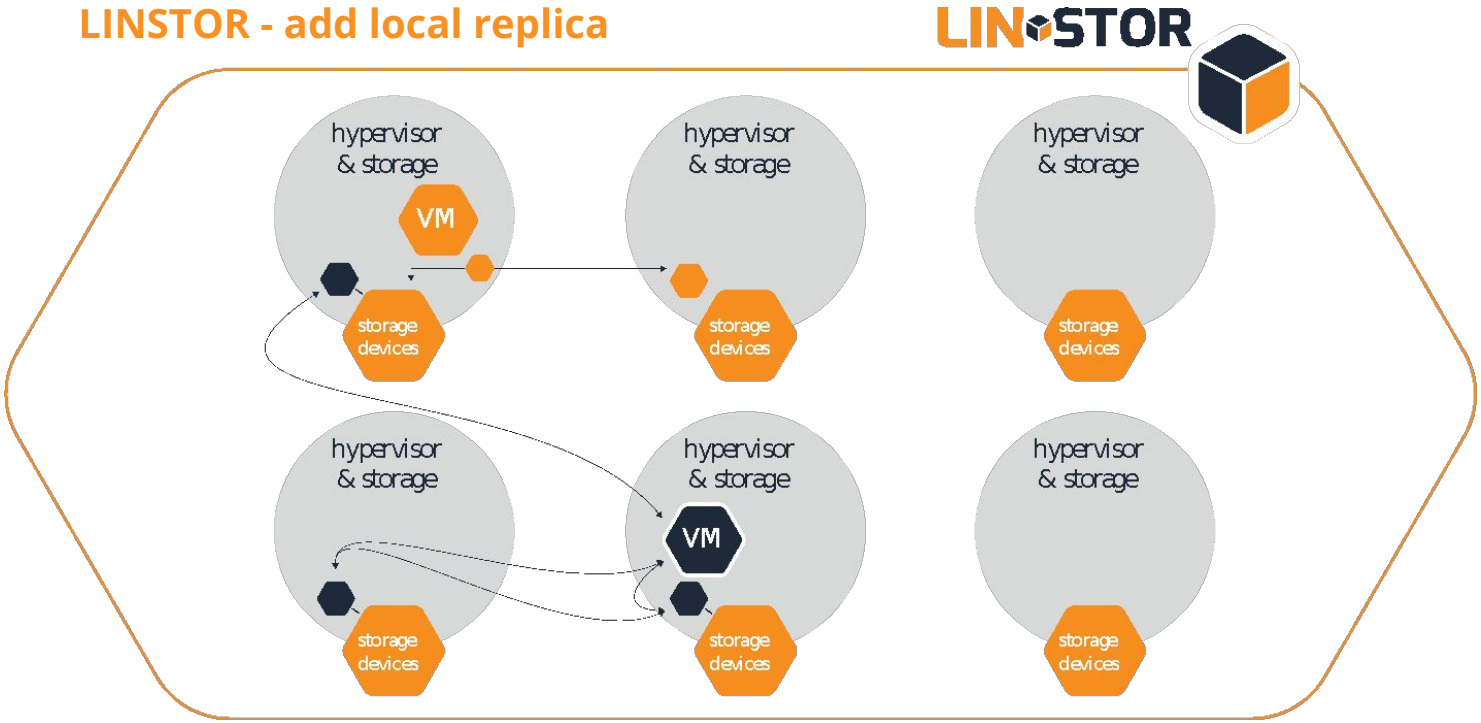
LINSTOR - Hyperconverged



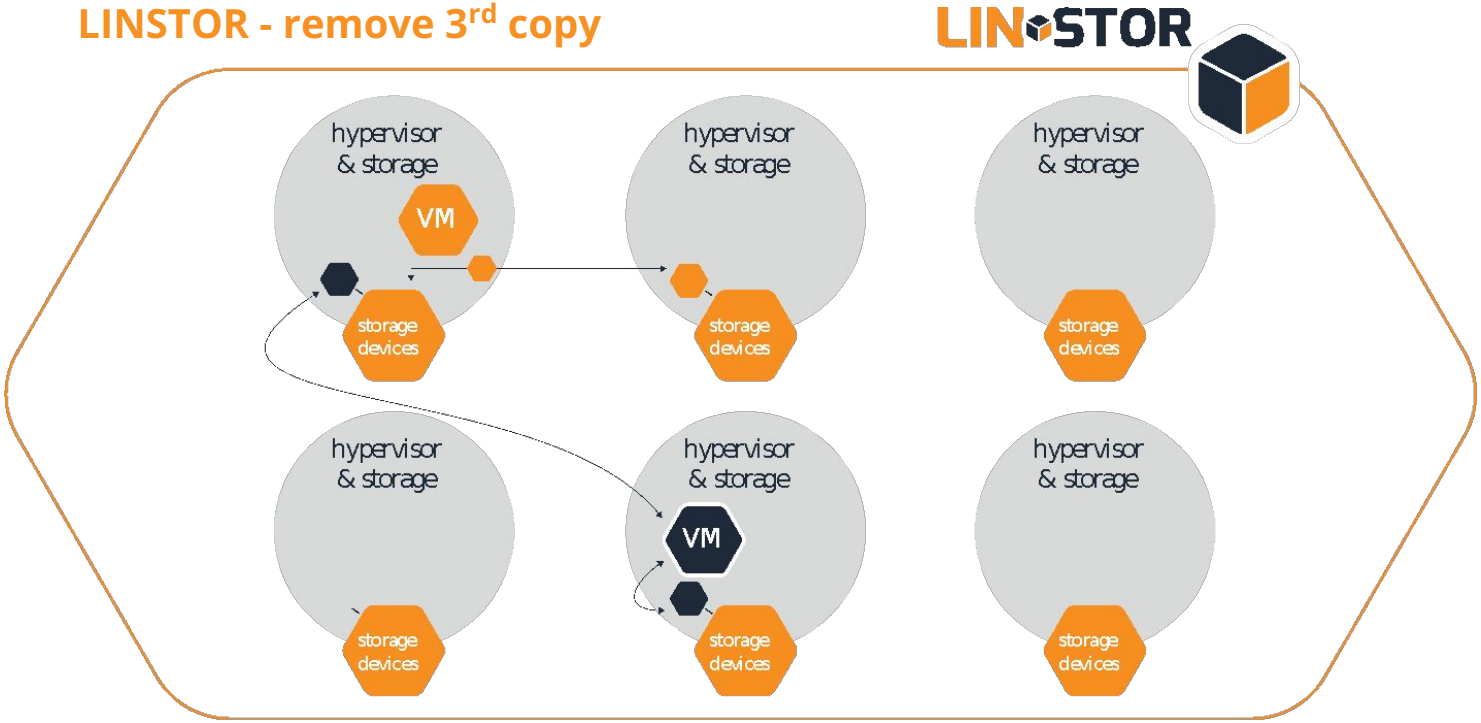
LINSTOR - VM migrated



LINSTOR - add local replica



LINSTOR - remove 3rd copy



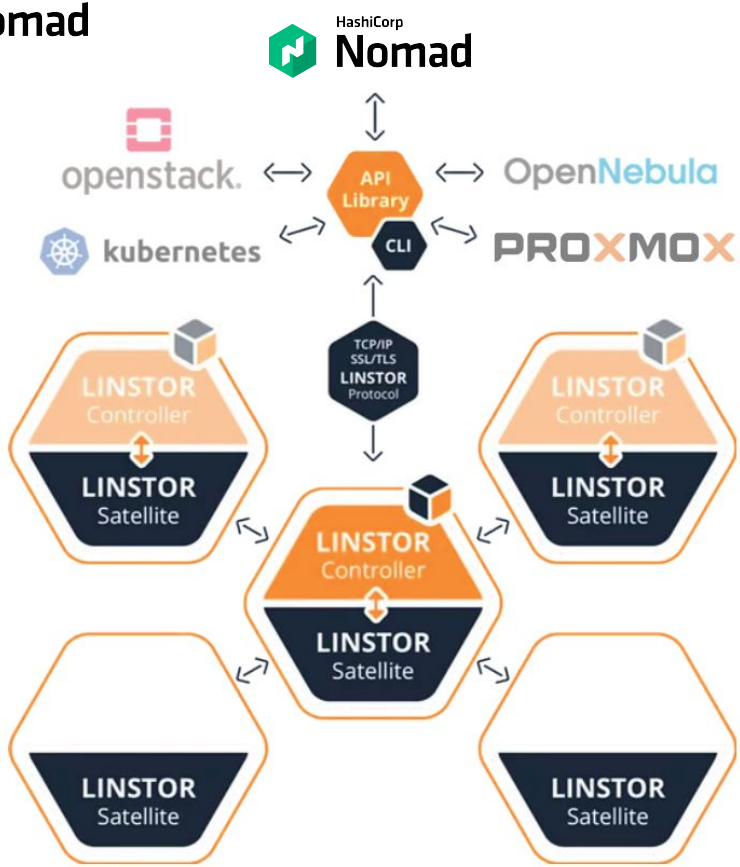
LINBIT SDS × HashiCorp Nomad
STORAGE DAY

LINSTOR

Architecture, objects and functions

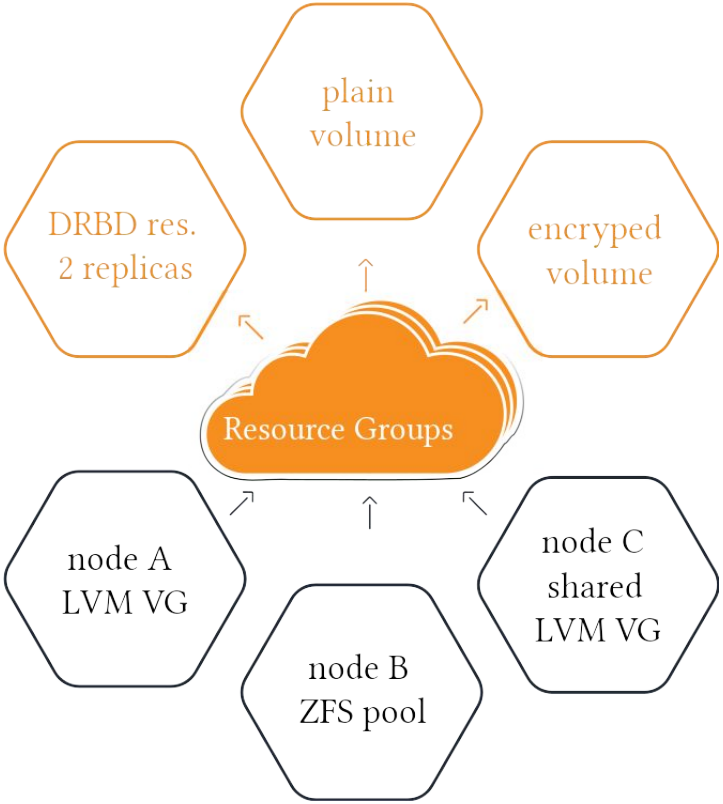
LINBIT SDS × HashiCorp Nomad

STORAGE DAY



LINSTOR Objects

- Nodes
- Resources
  - Volumes
- Snapshots
- Storage Pools
  - Shared
- Resource groups
- Properties
  - Aux properties



LINSTOR Storage Layers

Top (opt)	DRBD
Mid (opt & multiple)	LUKS (encryption), caches, NVMe target & initiator
Bottom (required)	LVM VG, ZFS zPool, Exos, OpenFlex, SPDK, shared LVM VG
Below (opt)	VDO (deduplication), software RAID

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

Q&A

CONTACT

LINBIT HA-Solutions Gmbh
Vivenotgasse 48
1120 Vienna, Austria
sales@linbit.com

HashiCorp Germany GmbH
team-emea-dach@hashicorp.com