FOSDEM21

Software defined storage devroom
Storage management automation

Provides storage for various consumers on a cluster of nodes

Manages backend storage
   (LVM, ZFS, fat- & thin-provisioning, NVMe-oF, etc.)

Manages various optional layers
   (encryption, deduplication, caching, ...)

Manages replication with DRBD

Manages cluster-consistent snapshots

Automates allocation of resources
   (Minor numbers, TCP/IP port numbers, etc.)
Connectors

- Kubernetes
- Docker Swarm, Mesos
- OpenStack Cinder
- OpenNebula
- Proxmox VE
- XenServer, XCP-ng
Hyperconverged

Node-local storage
LINSTOR - VM migrated
LINSTOR - add local replica
LINSTOR - remove 3rd copy
Separate storage nodes

Remote storage
LINSTOR – disaggregated stack

Diagram showing a distributed architectural model with hypervisors, VMs, and storage nodes connected via DRBD.
LINSTOR / failed storage node
Architecture and external components
Controller with transaction-safe persistence
- Stateless Satellites
- Platform independent (Java VM)
- Control-plane / data-plane separation
- Optional Controller high availability with Pacemaker
- HA Persistence: Replicated storage, shared storage, DB replication
Linux's LVM

- Snapshot
- Logical Volume
- Volume Group
- Physical Volume
Linux's LVM

Volume Group

logical volume

physical volume

physical volume

physical volume

snapshot

thin-LV

thin-LV

thin-sLV
Data replication with DRBD
DRBD - Main features

- Linux kernel module, GPL
- Synchronous or asynchronous replication
- Replicates any (random access) block device
- Consistency groups: Multiple volumes per resource
- Each resource available on up to 32 nodes
- Simple TCP/IP replication link
- Multi-pathing: Replication link fail-over
- Can be integrated with cluster resource managers
- Fencing & Quorum
DRBD Roles: Primary & Secondary

Primary ➔ replication ➔ 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 Quorum

- Quorum of the majority, all or a specified number of nodes
- Minimum redundancy
- Upon loss of quorum: suspends or fails I/O
- Minimum of three nodes
DRBD Optimizations

- Infiniband, Dolphin Express
- RDMA (Remote Direct Memory Access):
  - Infiniband/RDMA or RoCE (RDMA over converged Ethernet)
- Multi-path load-balancing with RDMA
- Meta data on PMEM / NVDIMM
Dedicated storage systems
linstor-gateway

- Provides iSCSI and/or NFS resources
- Storage provided by LINSTOR
- High availability provided by Pacemaker/Corosync
- Integrates LINSTOR with Pacemaker
More information

LINSTOR at LINBIT
https://www.linbit.com/linstor

LINSTOR on GitHub
https://github.com/LINBIT/linstor-server

DRBD on GitHub
https://github.com/LINBIT/drbd-9.0