OpenSDS mini summit

COPENHAGEN

LINSTOR

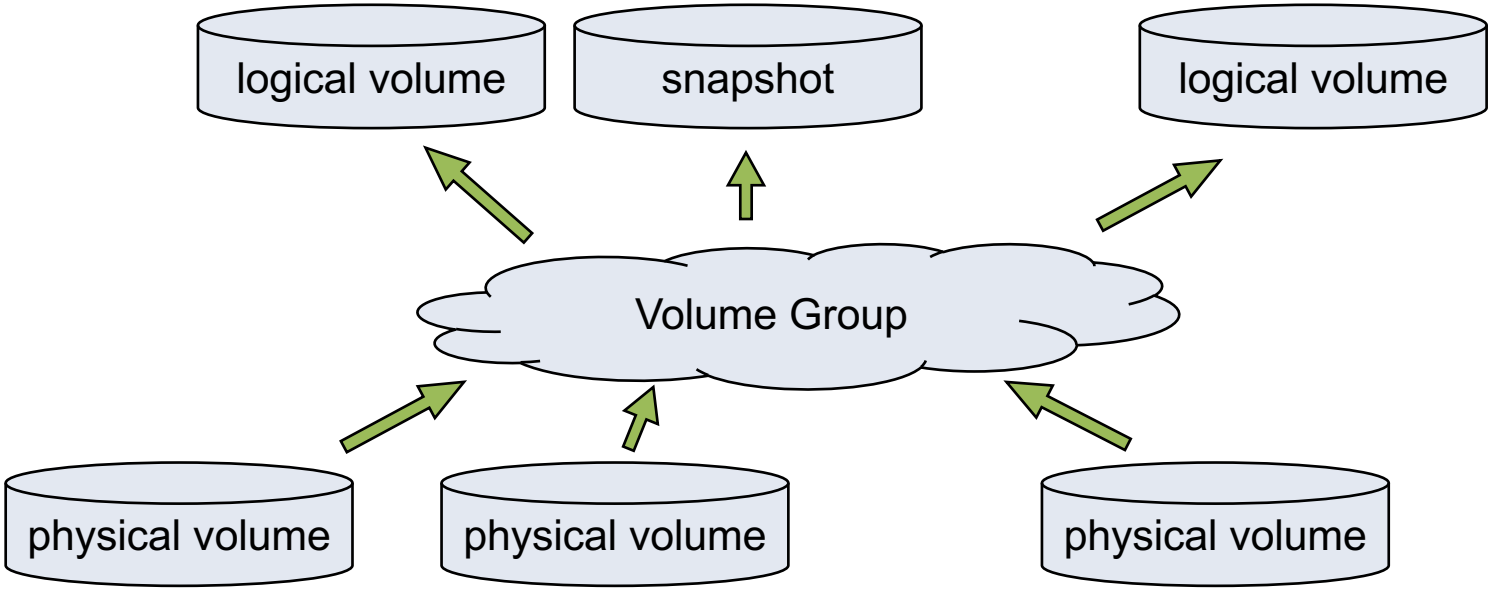
Reliable Storage for HA, DR, Clouds and Containers

Philipp Reisner, CEO LINBIT

Linux Storage Gems

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

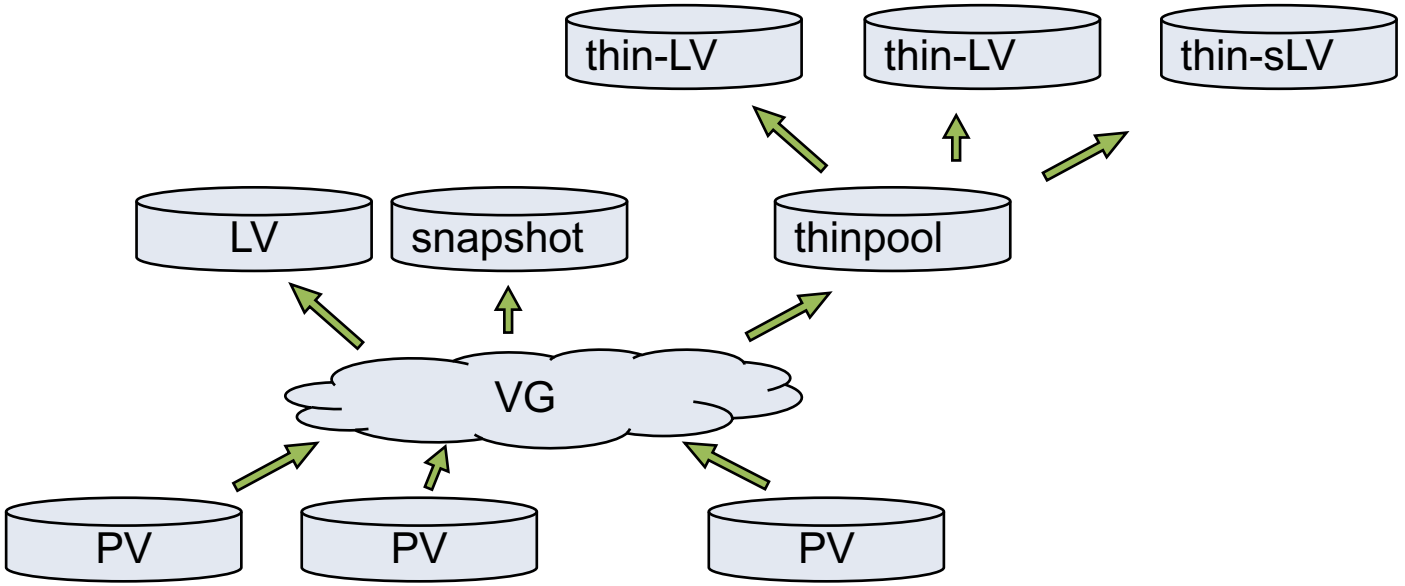
Linux's LVM


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

Linux's LVM

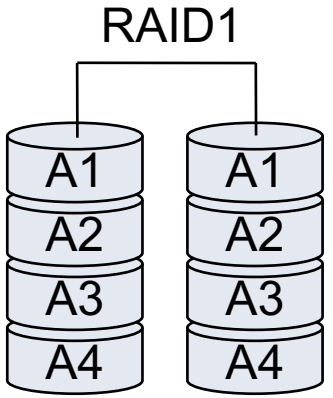
- based on device mapper
- original objects
  - PVs, VGs, LVs, snapshots
  - LVs can scatter over PVs in multiple segments
- thinlv
  - thinpools = LVs
  - thin LVs live in thinpools
  - multiple snapshots became efficient!

Linux's LVM



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

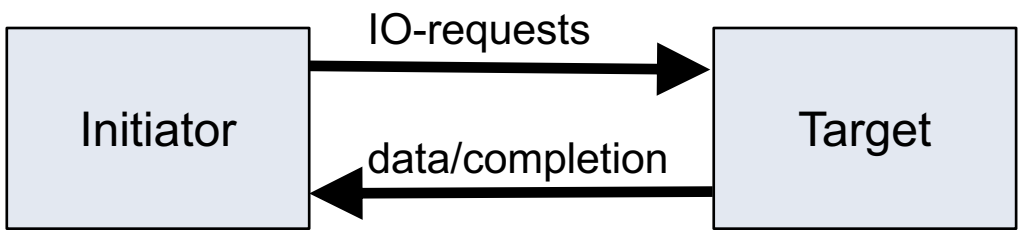
- dm-cache
  - device mapper module
  - accessible via LVM tools
- bcache
  - generic Linux block device
  - slightly ahead in the performance game

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



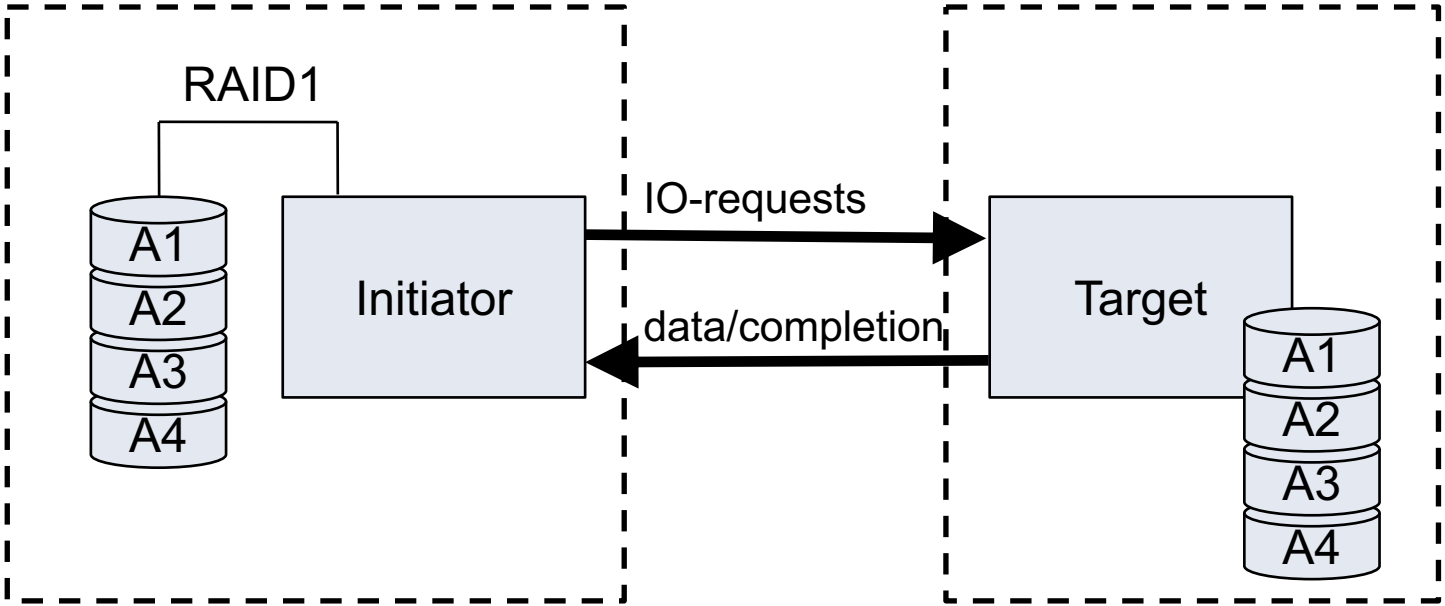
ZFS on Linux

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

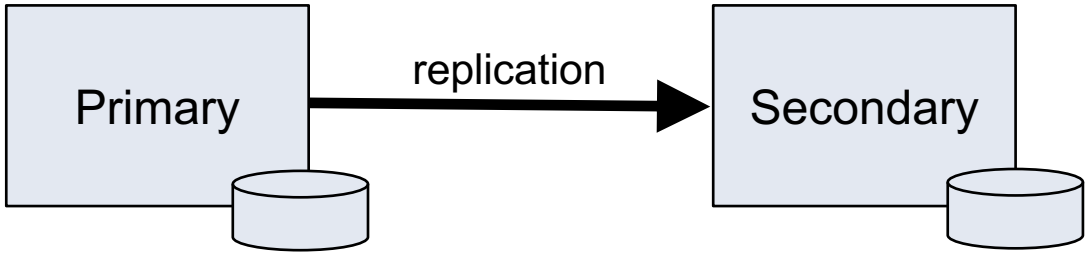
DRBD

Put in simplest form

DRBD – think of it as...

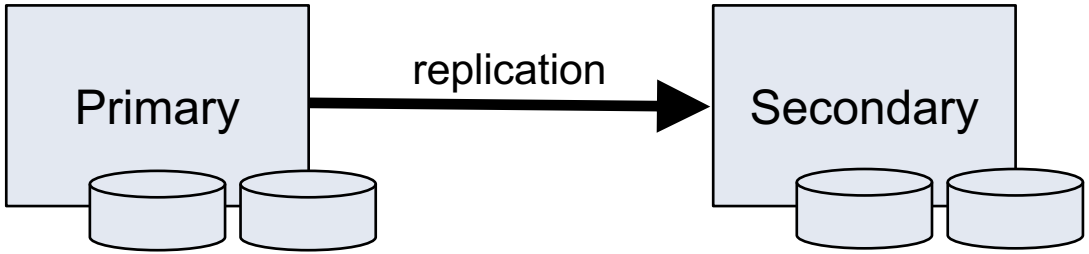


DRBD Roles: Primary & Secondary



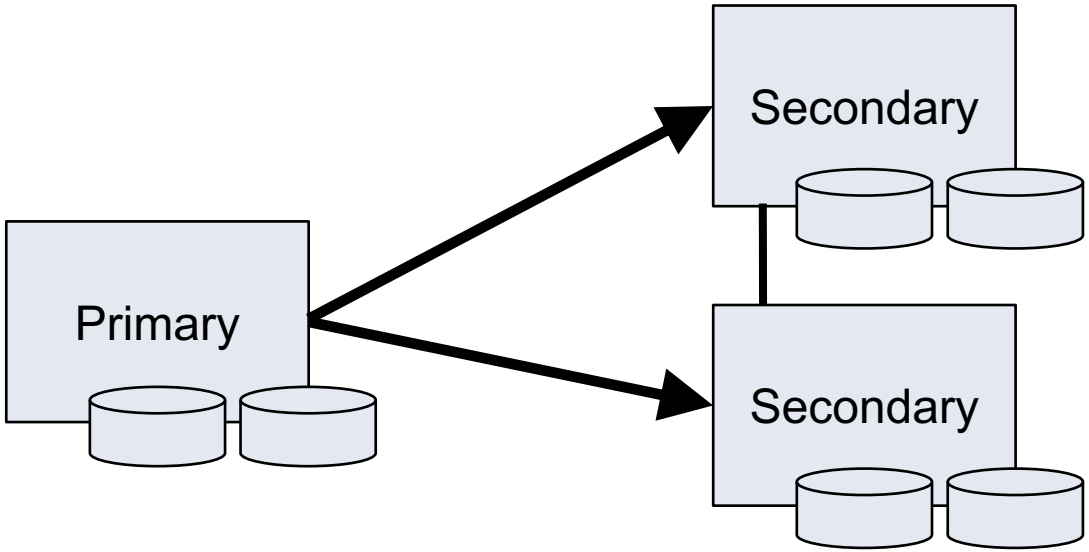
DRBD – multiple Volumes

- consistency group



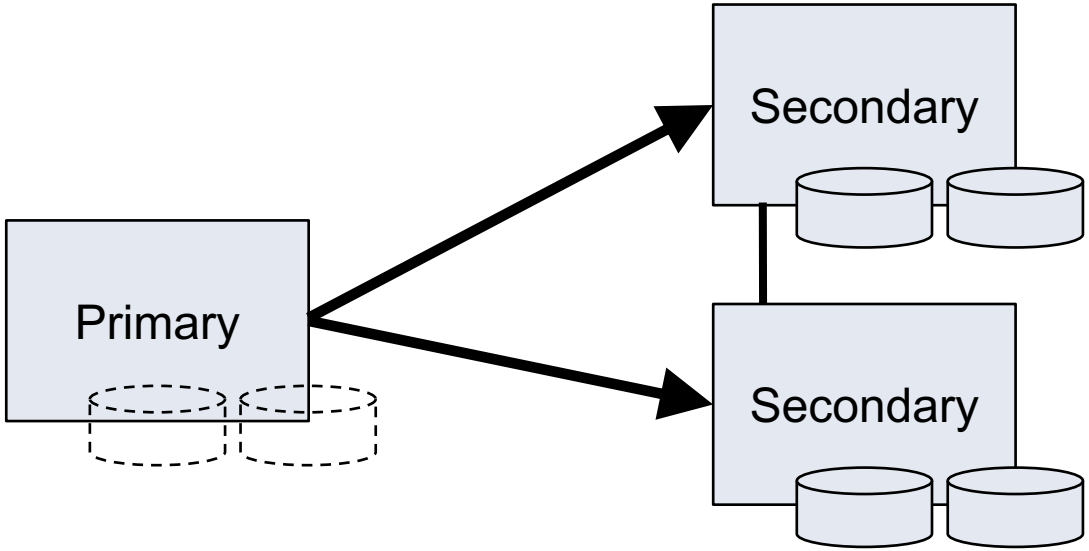
DRBD – up to 32 replicas

- 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 Roadmap

- performance optimizations (2018)
  - meta-data on PMEM/NVDIMMS
  - zero copy receive on diskless (RDMA-transport)
- Eurostars grant: DRBD4Cloud
  - erasure coding (2019)

LINSTOR

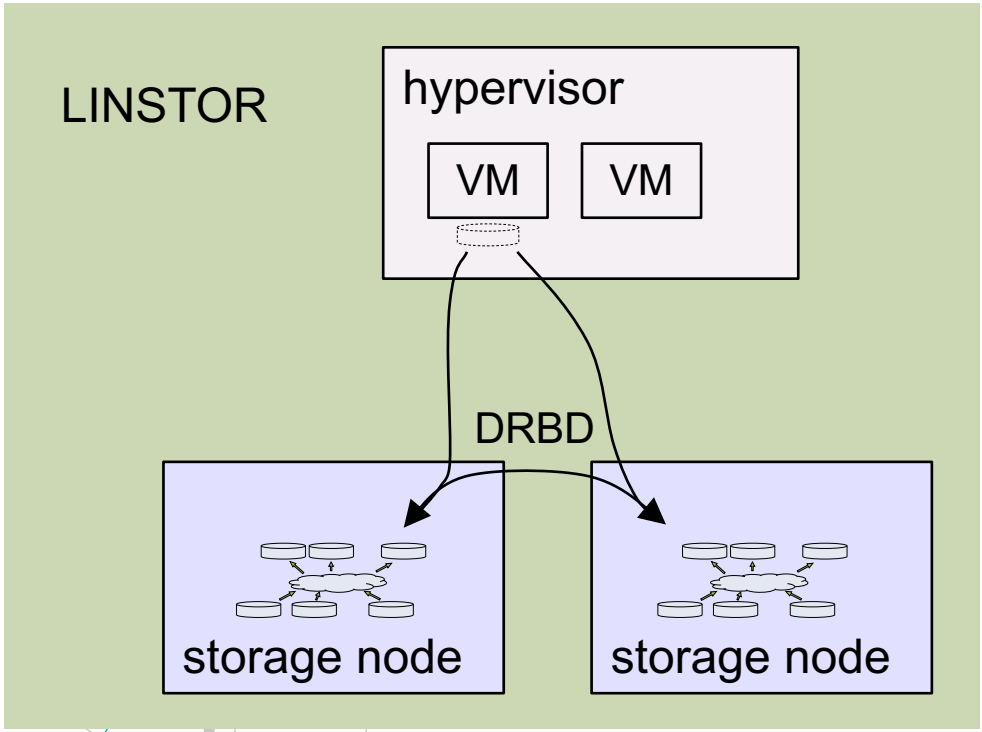
The combination is more than the sum of its parts

LINSTOR - goals

- storage build from generic (x86) nodes
- for SDS consumers (OpenStack Cinder, Kubernetes)
- building on existing Linux storage components
- multiple tenants possible
- deployment architectures
  - distinct storage nodes
  - hyperconverged with hypervisors / container hosts

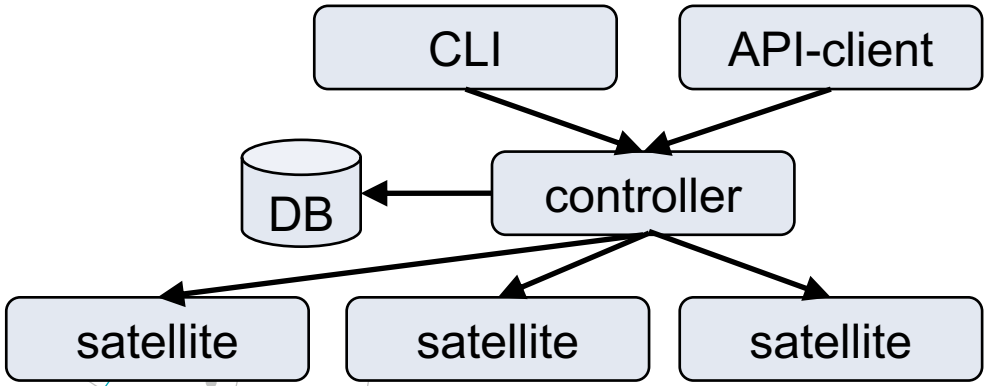
LINSTOR

- controls LVM/ZFS
  - snapshots
  - thin
- multiple VGs
  - for caching SSDs
  - different pools
- controls DRBD



LINSTOR Architecture

- embedded or external SQL data base
  - replicated by DRBD
- one controller process per cluster
  - HA by pacemaker
- one satellite process per node
  - satellite is state less
- API-clients
  - Kubernetes, ...
- CLI



LINSTOR Roadmap

- finish snapshot support (May 2018)
- Swordfish API (August 2018)
  - volume & snapshot management
  - access via NVMe-oF
  - inventory sync from Redfish/Swordfish
- support for multiple sites & DRBD-Proxy (Dec 2018)
- north bound drivers
  - Kubernetes, OpenStack, OpenNebula, Proxmox, XenServer

LINSTOR / DRBD & OpenSDS

- DRBD driver in OpenSDS for host base replication
  - coming soon, contribution of LINBIT
- OpenSDS south bound driver for LINSTOR
  - in planning by LINBIT
  - allows LINSTOR to benefit from OpenSDS' north bound drivers

LINSTOR vs ceph/GlusterFS

- block only
- backend allocation upon volume create
- each replica is a full and consistent copy
- in kernel data path
- control plane completely independent
  - can be restarted, upgraded while IO on existing volumes

Thank you

http://www.linbit.com

OpenSDS COPENHAGEN mini summit