Software Defined Storage
the Linux Way

ALWAYS ON AVAILABLE EFFICIENT
Who is LINBIT?

● 2001 Founded in Vienna Austria
  ● Primarily a company built around DRBD
● 2008 USA offices opened
● 2010 DRBD is merged into the mainline Kernel (2.6.33)
● 2013 LINBIT begin our first SDS type project
● 2016 LINBIT begins our second SDS project (LINSTOR)
● 2019 LINBIT teams up with Daocloud to help develop the Piraeus Datastore
LINBIT Software-Defined Storage

**LINBIT HA**
- NFS / CIFS / iSCSI
- KVM / VMWare / Xen
- Databases
- Fileservers
- Webservers
- NagiosXI
- Messaging (MQ)
- Nearly any other app

**LINBIT SDS (LINSTOR)**
- Container-native
  - OpenShift
  - Kubernetes
  - Docker
- Cloud-native
  - OpenNebula
  - OpenStack
  - Proxmox VE

**Must be**
- Highly reliable
- Cost effective
- Easy to provision
- Easy to scale
OS-Based Storage Technology

- Linux already provides several storage gems:
  - LVM
  - RAID
  - SSD cache tiers
  - De-duplication
  - Targets & initiators

Native Storage Management Capabilities
Linux 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 are efficient!
**RAID**

**Capabilities**

- Original MD code
  - Controlled via mdadm
  - 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
Cache Devices

- dm-cache
  - device mapper module
  - accessible via LVM tools

- bcache
  - generic Linux block device
  - slightly ahead in the performance game
ZFS on Linux

- Popular in the Ubuntu ecosystem
- Has support in other distros as well.
  - Mainly from repos provided by the ZFSonLinux team

- Has its own
  - logic volume manager (zVols)
  - thin provisioning
  - RAID (RAIDz)
  - caching with SSDs (ZIL, SLOG)
  - and a file system!
VDO Inline Deduplication

• Virtual Data Optimizer (VDO) since RHEL 7.5
  • Red hat acquired Permabit and has open sourced it: https://github.com/dm-vdo

• Linux upstreaming is in preparation

• In-line data deduplication

• Kernel part is a device mapper module

• Indexing service runs in user-space

• Asynch or synchronous writeback

• Recommended to be used below LVM
### Targets & Initiators

#### Capabilities

- **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

#### Targets & initiators

![Diagram showing IO-requests and data/completion between Initiator and Target]
DRBD – Mainline Linux Kernel

- 1000’s of Nodes
  - Up to 32 Synchronous or async replicas per volume
  - Automatic partial resync after connection outage
  - Multiple resources per node possible (1000s)

- Diskless nodes
  - Intentional diskless (no change tracking bitmap)
  - Disks can fail

- Reliable
  - A node knows the version of the data is exposes
  - Checksum-based verify & resync
  - Split brain detection & resolution policies
  - Fencing
  - Quorum
  - Dual Primary for live migration of VMs only!
LINSTOR - Goals

Goals

- Don’t recreate the wheel
  - Use existing linux storage components
- Centralized control plane separated from the storage plane
- Allow multi-tenancy
- Multiple Deployment Architecture
  - Dedicated storage nodes Vs. Hyperconverged, etc.
- Serve SDS Consumers (Cinder, Kubernetes, etc)
- Use existing off-the-shelf-hardware

Approach
# LINSTOR - Structure

## Capabilities
- Controls LVM/ZFS
  - Snapshots
  - Thin/Thick provisioned
- Multiple VGs
  - For caching SSDs
  - Different pools
- Controls DRBD
  - Replicates Data

## Structure

![LINSTOR Architecture Diagram](image-url)
Container Storage

LINUX BLOCK STORAGE MANAGEMENT FOR CONTAINERS

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
# New LINSTOR Features

## Complete
- Snapshot support
- Multiple geo-diverse sites with DRBD Proxy
- File-based storage (via “loop”)
- Secure REST API (HTTPS)
- Multi-user REST API (LDAP)
- PMEM backing storage for DRBD metadata
- Plugins for several orchestrators and cloud platforms
- Can make calls to Swordfish APIs
- Manager NVMe-oF targets and initiators

## Roadmap
- iSCSI Targets: creating and attaching (Q4 2019)
- VDO deduplication (2020)
- DRBD 10 (already in alpha)
  - Performance Improvements
  - PMEM caching + journaling
  - Erasure Coding
  - Request Forwarding (“resource chaining”)
Piraeus Datastore

https://piraeus.io/

Piraeus

Coming soon...
## Resources

### High Availability

### Disaster Recovery
- Video Demo (6 minutes) https://www.youtube.com/watch?v=Sf0IPCiIDWk

### Software-Defined Storage
- Private Cloud: https://www.linbit.com/en/linbit-sds-private-cloud/