Software Defined Storage (SDS)

ALWAYS ON AVAILABLE EFFICIENT
The obvious approach to enterprise grade SDS
LINBIT Software-Defined Storage

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

**LINBIT SDS**
- 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
Container Storage

LINUX BLOCK STORAGE MANAGEMENT FOR CONTAINERS

ORCHESTRATORS

BLOCK TRANSPORT SYSTEMS

iSCSI ⬅️ ⬆️ NVMe-oF ⬅️ ⬆️ DRBD DISKLESS ⬅️ ⬆️

BLOCK STORAGE FEATURES

DRBD ⬅️ ⬆️ LUKS 🔒 Cache 🔒

NODE-LEVEL VOLUME MANAGEMENT

LVM ⬅️ ZFS ⬆️

HARDWARE

HDD ⬅️ SSD ⬆️ NVMe ⬅️ PMEM ⬆️
Capabilities

- 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
  - `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`
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!
Cache Devices

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

- bcache
  - generic Linux block device
  - slightly ahead in the performance game
Linux’s Inline Deduplication

- 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

- 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
DRBD – Mainline Linux Kernel

**Capabilities**

- 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!

**DRBD**
Capabilities ➔ Products ➔ Solution
# LINSTOR - Goals

<table>
<thead>
<tr>
<th>Goals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Build storage from generic (x86) nodes</strong></td>
</tr>
<tr>
<td>• Serve SDS consumers (OpenStack Cinder, Kubernetes, Custom platforms)</td>
</tr>
<tr>
<td>• Allow multi-tenancy</td>
</tr>
<tr>
<td>• Enable multiple Deployment architectures</td>
</tr>
<tr>
<td>• Distinct storage nodes</td>
</tr>
<tr>
<td>• Hyperconverged with hypervisors / container hosts</td>
</tr>
<tr>
<td>• Don’t recreate the wheel</td>
</tr>
<tr>
<td>• Use existing Linux storage components</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
</tbody>
</table>
Capabilities

- Controls LVM/ZFS
  - Snapshots
  - Thin
- Multiple VGs
  - For caching SSDs
  - Different pools
- Controls DRBD

LINSTOR

[Diagram showing the integration of openstack, CLI, and kubernetes with LINSTOR components like Controller and Satellite]
## New LINSTOR Features

### Complete
- Snapshot support
- Multiple geo-diverse sites with DRBD Proxy
- File-based backing storage (via “loop”)
- Secure REST API (HTTPS)
- Multi-user REST API support (LDAP)
- PMEM backing storage for DRBD metadata
- Support for several orchestrators and cloud platforms
- Swordfish API
  - Can manage 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”)
# Resources

<table>
<thead>
<tr>
<th>High Availability</th>
<th>Disaster Recovery</th>
<th>Software-Defined Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><a href="https://www.linbit.com/en/linstor/">LINSTOR:</a></td>
</tr>
</tbody>
</table>