Stephan Seitz <s.seitz@heinlein-support.de>

Hochverfügbares iSCSI Storage

mit ceph

Stephan Seitz <s.seitz@heinlein-support.de>

Wer ist die Heinlein Support GmbH?

- Wir bieten seit 20 Jahren Wissen und Erfahrung rund um Linux-Server und E-Mails
- IT-Consulting und 24/7 Linux-Support mit 20 Mitarbeitern
- Eigener Betrieb eines ISPs seit 1992
- Täglich tiefe Einblicke in die Herzen der IT aller Unternehmensgrößen

Stephan Seitz <s.seitz@heinlein-support.de>

Intro:

Ein kurzer Blick auf die Techniken

Stephan Seitz <s.seitz@heinlein-support.de>

Ein Blick auf ceph

- Flexibler, skalierbarer Object-Store
- Zugriff auf
  - Objekte (librados, S3 via radosgw)
  - software-defined Blockdevices (RBD via Linux Kernel-RBD, FUSE)
  - Filesystem (via ceph MDS)
- „THE FUTURE OF STORAGE™“

Ein Blick auf iSCSI

- Einfach zu etablieren
- Software-defined Blockstorage
- Etablierte Mechanismen für Failover
  - ALUA
  - auch proprietäre Pfadpriorisierungen (EMC, NetApp, …)

Warum dann ceph + iSCSI ?

- Verbindung von modernen und klassischen Storagekonzepten
  - Nicht jede Plattform unterstützt Rados Block Devices (RBD)
  - iSCSI dagegen fast immer
  - CephFS benötigt ebenfalls client-seitige Unterstützung und ist als Dateisystem nicht unbedingt ein Ersatz für Block-Storage
  - iSCSI Targets sind (in Software) üblicherweise nicht über Portale mehrerer Hosts zu betreiben
    - Hochverfügbarkeit nur über Hardware, d.h. mehrere Controller-Units an einer gemeinsamen Backplane
  - Linux TGT besitzt native RBD-Unterstützung (Userland, benötigt bs_rbd, in Mainline seit 02/2013)

Stephan Seitz <s.seitz@heinlein-support.de>

Hands On!

Implementierung und Test

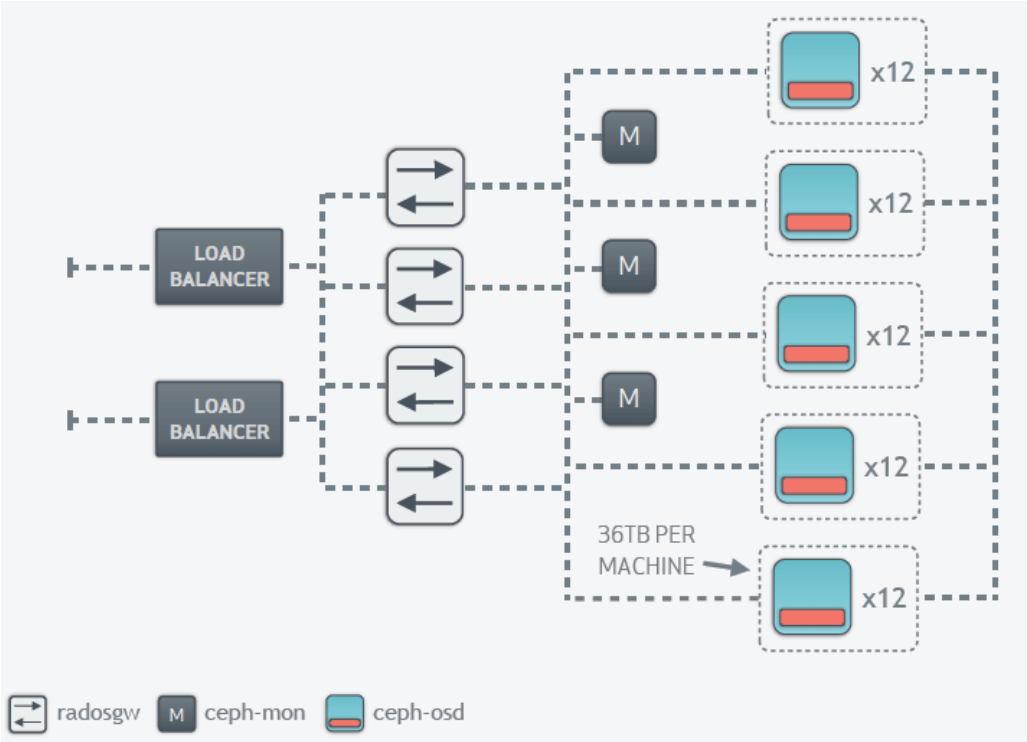
Stephan Seitz <s.seitz@heinlein-support.de>

Prerequisites

- Laufender ceph-cluster
  - Getestet mit Ceph Firefly 0.80.5
- zwei oder mehr Hosts als ceph ↔ iSCSI „Bridge“
  - Idealerweise getrennte NIC für Backing Store (ceph) und iSCSI Portale
  - Getestet mit Ubuntu 14.04 LTS / ceph 0.80.5 und tgt 1.0.43
- Hosts / Cluster / Applikationen / etc. mit dem Bedarf an skalierbarem Storage und „legacy“ Möglichkeiten

Konzept

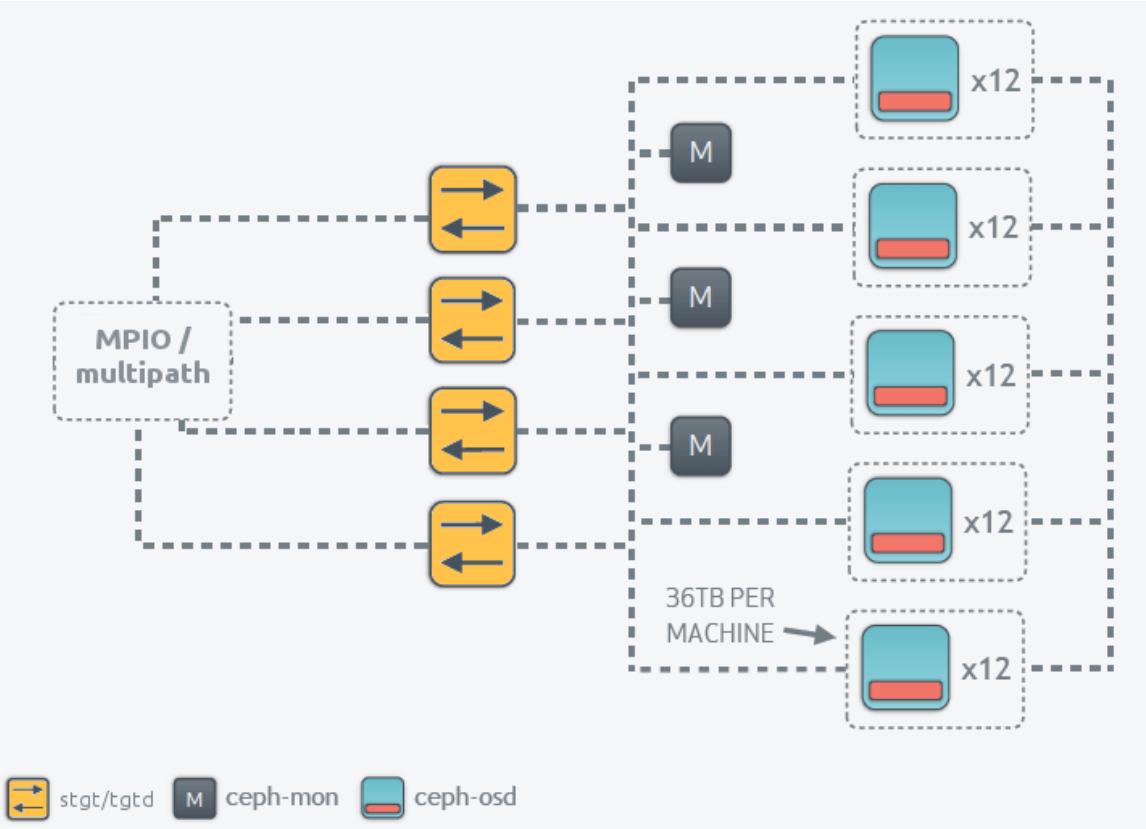
- Ceph kann parallel auf Objekte in OSD zugreifen



- Hier mit multiplen RadosGW Instanzen und S3/HTTP Loadbalancer

Konzept

- Paralleler Zugriff funktioniert auch mit TGT iSCSI Target



Vorbereitung der ceph ↔ iSCSI „Bridge“

- Paketinstallation am Beispiel Ubuntu 14.04 LTS

```
tgt1 # aptitude install ceph tgt
```

- Anbindung an den ceph-Cluster

```
tgt1 # cat /etc/ceph/ceph.conf
[global]
mon_host = 192.168.100.247,192.168.100.249
auth_cluster_required = cephx
auth_service_required = cephx
auth_client_required = cephx
```

```
tgt1 # cat /etc/ceph/ceph.client.admin.keyring
[client.admin]
    key = QXVZBiNWUF2uEhAAnixKgvkpQkWvxQHWSBmr7g==
```

Vorbereitung der ceph ↔ iSCSI „Bridge“

- Erzeugen eines Ceph-RBD, die Basis für eine spätere LUN

```
tgt1 # rbd create --size 3072 ceph-tgt-shared
tgt1 # rbd info ceph-tgt-shared
rbd image 'ceph-tgt-shared':
    size 3072 MB in 768 objects
    order 22 (4096 kB objects)
    block_name_prefix: rb.0.56c7.238e1f29
    format: 1
```

  - Im Beispiel wird ein 3 GB grosses RBD „ceph-tgt-shared“ erzeugt.
  - Ceph-typisch ist der Platz nicht belegt, sondern reserviert.
  - *Das RBD wird nicht über Kernel-RBD verwendet und sollte daher auch nicht gemapped werden!*

Vorbereitung der ceph ↔ iSCSI „Bridge“

- Neues iSCSI Portal erzeugen

```
tgt1 # tgtadm --lld iscsi --op new --mode target \
 --tid 1 \
 -T iqn.2005-10.de.heinlein-hosting:storage.s0.ceph.shared.tgt
```

  - Aufbau des *IQN* : iqn.YYYY-MM.NAMING_AUTHORITY:UNIQUE_NAME

- LUN 1 mit neuem RBD als Storage Backend erzeugen

```
tgt1 # tgtadm --lld iscsi --mode logicalunit --op new \
 --tid 1 --lun 1 --bstype rbd --backing-store ceph-tgt-shared \
 --bsopts "conf=/etc/ceph/ceph.conf;id=admin"
```

- Portal / Target binden

```
tgt1 # tgtadm --lld iscsi --op bind --mode target --tid 1 \
 -I ALL
```

Vorbereitung der ceph ↔ iSCSI „Bridge“

- Wiederholen des ceph- und tgt-Setups auf weiteren Hosts
  - ! Die tgt Konfiguration kann (Stand Version 1.1x) nicht vollständig persistent gespeichert werden.
    Der übliche Weg `tgt-admin –dump > /etc/tgt/conf.d/targets.conf`
    enthält nicht die Parameter bstype und bsopts.
    Je nach Umgebung: Eigene INIT-Skripte, evtl. `/etc/rc.local`
- Ab hier stehen zwei+ iSCSI Targets mit einem Ceph-RBD als LUN bereit
  - Der Storage-Teil ist vollständig :)

Stephan Seitz <s.seitz@heinlein-support.de>

Einbinden der iSCSI Targets

- Paketinstallation am Beispiel Ubuntu 14.04 LTS

```
client # apt-get install open-iscsi multipath-tools
client # service open-iscsi start
client # service multipath-tools start
```

- iSCSI Targets finden

```
client # iscsiadm -m discoverydb -t sendtargets \
 --portal 192.168.100.88 --discover
```

  - Für jeden iSCSI Host wiederholen

```
client # iscsiadm -m node
192.168.100.88:3260,1 iqn.2005-10.de.heinlein-hosting:st..
192.168.100.91:3260,1 iqn.2005-10.de.heinlein-hosting:st..
```

  - Kurzer Check.
    Jeder iSCSI-Host sollte hier mit seiner IP und identischer IQN stehen.

Multipath-Konfiguration

- MP-Konfiguration ***vor*** dem iSCSI Login festlegen

```
client # cat /etc/multipath.conf
defaults {
        user_friendly_names     yes
        polling_interval        2
        path_grouping_policy    group_by_serial
        features                "1 queue_if_no_path"
        path_checker            directio
        rr_min_io               100
        failback                immediate
        no_path_retry           queue
}
[...]
```

  - Alternativ auch jede WWID einzeln in multipaths { multipath { ...

Multipath-Konfiguration und iSCSI Login

- MP-Konfiguration übernehmen …

```
client # echo reconfigure | multipathd -k
```

  - Alternativ auch multipath-tools neu starten

- … und in den iSCSI Portalen anmelden

```
client # iscsiadm -m node --login
Logging in to [iface: default, target: iqn.2005-10.de.he..
Logging in to [iface: default, target: iqn.2005-10.de.he..
Login to [iface: default, target: iqn.2005-10.de.heinlei..
Login to [iface: default, target: iqn.2005-10.de.heinlei..
```

  - für jedes Portal, d.h. jeden iSCSI Host sollte der Login erfolgreich sein

- Done!

Abschliessender Test

- MP-Konfiguration der iSCSI-LUN prüfen

```
client # multipath -ll
ceph-tgt-shared (3300000100000001) dm-3 IET,VIRTUAL-DISK
size=3.0G features='1 queue_if_no_path' hwhandler='0'
wp=rw
`-+- policy='round-robin 0' prio=1 status=active
  |- 28:0:0:1 sdf 8:80 active ready running
  `- 29:0:0:1 sdg 8:96 active ready running
```

  - je nach `multipath.conf` existiert `/dev/mapper/mpath0` oder der konfigurierte Alias. In diesem Beispiel `/dev/mapper/ceph-tgt-shared`

- Fertig!

Weitere sinnvolle Initiator: Hypervisors...

- Erfolgreicher iSCSI-Initiator Test mit XenServer 6.4 (beta)
  - [x] Multipath aktivieren (Am einfachsten via XenCenter)
  - identisch dem vorgestellten Ubuntu 14.04 LTS:
    - alle Portale via iscsiadm discover finden
    - Die /etc/multipath.conf um Device Vendor „`IET*`“ und Product „`VIRTUAL*`“ erweitern. Auch hier `group_by_serial` eintragen.
    - multipathd neustarten bzw. „reconfigure“
    - Ein neues SR über **ein** Portal hinzufügen (Die automatisch generierte Beschreibung des SR erwähnt nur das eine Portal, ein Blick auf die CLI zeigt, dass multipath auf allen Portalen erfolgreich verwendet wird)

Stephan Seitz <s.seitz@heinlein-support.de>

Abschliessende Überlegungen

- Linux TGT muss nicht zwingend iSCSI over IP anbieten
  - RBD sollten auch problemlos via FCoE oder iSER (via RDMA) exportiert werden
  - Export via FC-HBA muss definitiv getestet werden
- Resize?!?
  - Ist nicht trivial, konkret:
    - RBD vergrössern
    - Alle Portale beenden und **ein** Portal starten, danach die restlichen Portale
    - iscsiadm Portal LUN neu einlesen, bzw. „discover“ als harter Weg
    - iscsiadm „logout“ und „login“
    - Multipathd neustarten oder „reconfigure“
    - Eigentlichen Inhalt (pv, partition, Filesystem, etc...) vergrössern
  - Eigentlich kein Unterschied zum Aufwand mit Hardware-SAN ...

Abschliessende Überlegungen

- Kernel RBD ist performanter
  - Kann potentiell Split-Brain und Inkonsistenz verursachen,
    Tests stehen noch aus
  - I/O Caches müssten deaktiviert werden
    - möglicherweise bricht damit ein Grossteil des Performancevorteils weg
  - es könnte der (performantere) SCST verwendet werden
    - mit Direct- bzw. Block-I/O Layer, d.h. auch nur bedingt schneller
  - … to be evaluated

- Natürlich und gerne stehe ich Ihnen jederzeit mit Rat und Tat zur Verfügung und freue mich auf neue Kontakte.
  - Stephan Seitz
  - Mail: s.seitz@heinlein-support.de
  - Telefon: 030/40 50 51 – 44
- Wenn's brennt:
  - Heinlein Support 24/7 Notfall-Hotline: 030/40 505 - 110

Stephan Seitz <s.seitz@heinlein-support.de>

Soweit, so gut.

Gleich sind Sie am Zug:
Fragen und Diskussionen!

Stephan Seitz <s.seitz@heinlein-support.de>

Referenz

- ceph – Storage Cluster Quick Start (ceph-deploy)
  - http://ceph.com/docs/master/start/quick-ceph-deploy/
- Linux SCSI target framework
  - http://stgt.sourceforge.net/
- Generic SCSI Target Subsystem for Linux
  - http://scst.sourceforge.net/
- Multipath I/O
  - http://en.wikipedia.org/wiki/Multipath_I/O
  - http://en.wikipedia.org/wiki/Linux_DM_Multipath
  - /usr/share/doc/multipath-tools/examples/multipath.conf.annotated.gz

Stephan Seitz <s.seitz@heinlein-support.de>

Heinlein Support hilft bei allen Fragen rund um Linux-Server

HEINLEIN AKADEMIE

Von Profis für Profis: Wir vermitteln die oberen 10% Wissen: geballtes Wissen und umfang-reiche Praxiserfahrung.

HEINLEIN HOSTING

Individuelles Business-Hosting mit perfekter Maintenance durch unsere Profis. Sicherheit und Verfügbarkeit stehen an erster Stelle.

HEINLEIN CONSULTING

Das Backup für Ihre Linux-Administration: LPIC-2-Profis lösen im CompetenceCall Notfälle, auch in SLAs mit 24/7-Verfügbarkeit.

HEINLEIN ELEMENTS

Hard- und Software-Appliances und speziell für den Serverbetrieb konzipierte Software rund ums Thema eMail.