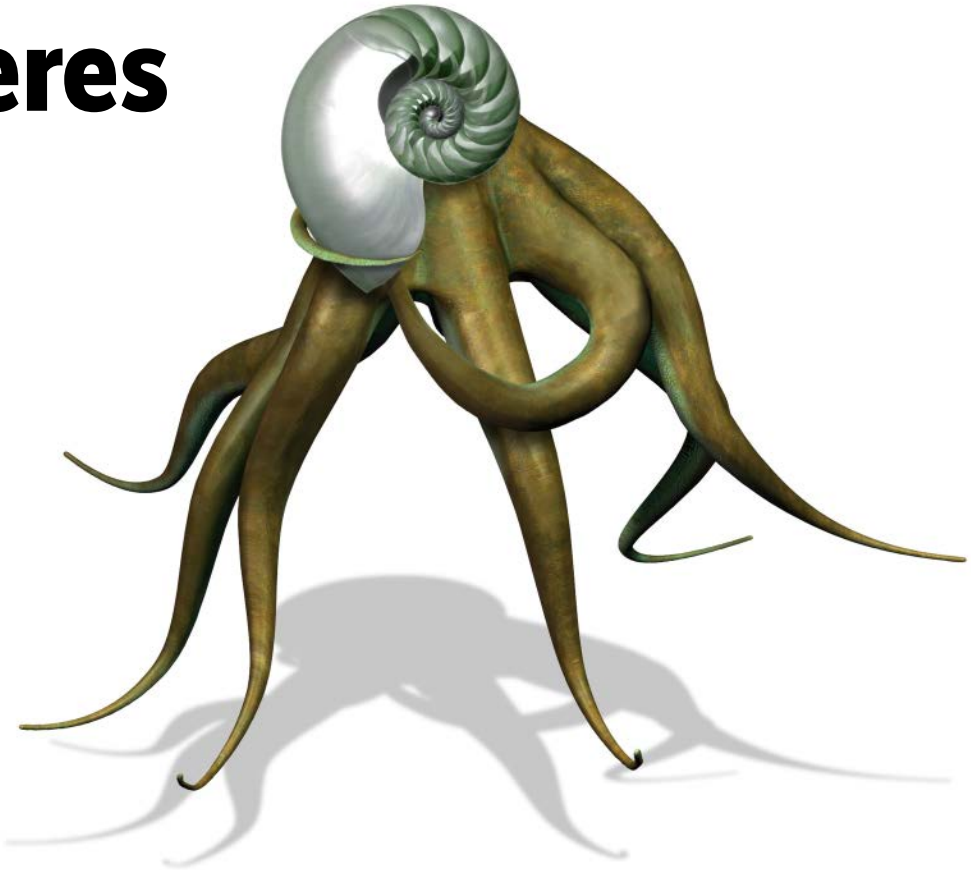
Ceph-Version Nautilus mit neuem Dashboard

Ein anderes Gesicht

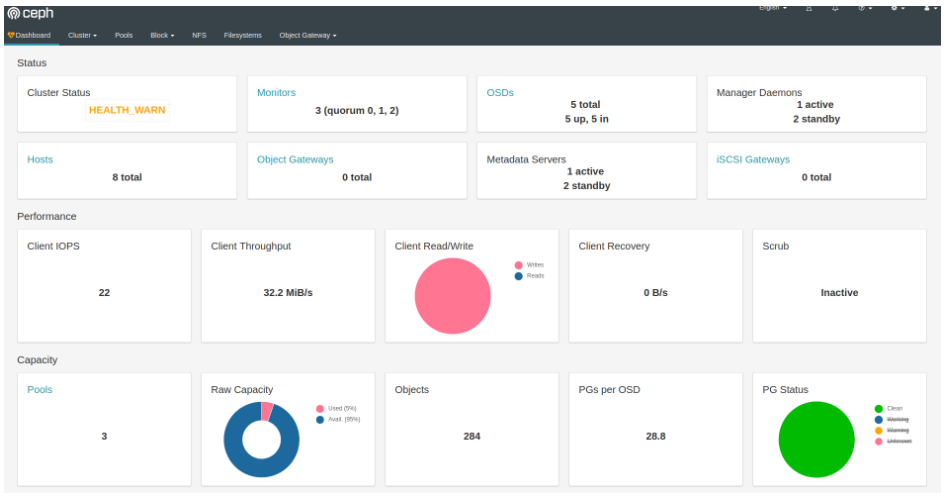
Michel Raabe



Vereinfachungen und mehr Spielraum für den Administrator haben sich die Ceph-Entwickler für die neue Version Nautilus zum Ziel gesetzt.

Nur einen Monat hatte der Release Candidate der neuen Ceph-Version Bestand, schon erschien im März 2019 die erste stabile Ausgabe von Nautilus alias Ceph v14.2.0. Mit der Wahl des Namens haben die Entwickler einen weiteren Kopffüßer oder Cephalopoda zum Paten erhoben.

Wie gewohnt bringt auch diese Release eine ganze Reihe neuer Funktionen mit. Den wohl größten sichtbaren Umbruch gab es aber beim Dashboard, der internen grafischen Schnittstelle von Ceph (siehe Abbildung 1). Die Entwickler haben hierfür das von SUSE gepflegte Projekt openATTIC als Grundlage genommen und mit SUSEs Hilfe ihr eigenes Dashboard nahezu ersetzt.



Im neuen Dashboard von Nautilus hat der Administrator alle wichtigen Informationen zum Status, zur Performance und zur Kapazität seines Ceph-Clusters im Blick (Abb. 1).

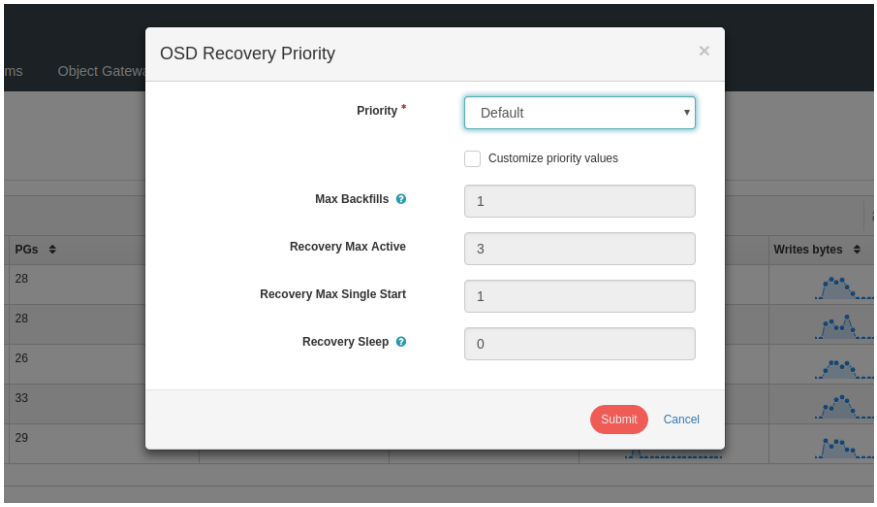
Dadurch lassen sich viele alltägliche Aufgaben wie das Anlegen von RBDs (Rados Block Devices) oder das Anpassen von Konfigurationsparametern einfach und schnell per Dashboard erledigen. Heraus sticht der neue Schalter zum „Tunen“ von Rebalance-Operationen (siehe Abbildung 2). Sie verteilen die Placement Groups (PGs) neu, wenn ein OSD (Object Storage Daemon) hinzukommt oder ausfällt. Zudem sind viele nützliche Kommandos nun nicht mehr dem Ceph-Administrator vorbehalten, sondern stehen jetzt auch den Anwendern zur Verfügung.

Helferlein stärker eingebunden

Integriert hat Red Hat in das Dashboard einen einfachen CRUSH Map Viewer, womit sich zumindest Buckets und Classes anzeigen lassen (siehe Abbildung 3). Des Weiteren lässt sich nun eine bestehende Grafana-Instanz in das Dashboard integrieren – ein Tab weniger im Browser.

Dass das Dashboard die Option bietet, sich per Single Sign-on anzumelden, bringt ein weiteres Stück Komfort. Dazu ist lediglich auf der Kommandozeile ein selbst signiertes Zertifikat zu erstellen und ein neuer Benutzer anzulegen. Die Befehle dazu liefert Ceph gleich mit, wie in Listing 1 zu sehen.

Neu ist auch eine zentrale Schnittstelle für angepflanschte Konfigurationsmanagement-Systeme wie ceph-ansible oder DeepSea. Damit lassen sich nun aus Ceph heraus API-Calls gegen die Konfigurationsmanagmentsysteme starten, die dann zum Beispiel OSDs ersetzen, neue Services ausrollen oder einfach nur eine Festplatten-LED blinken lassen.

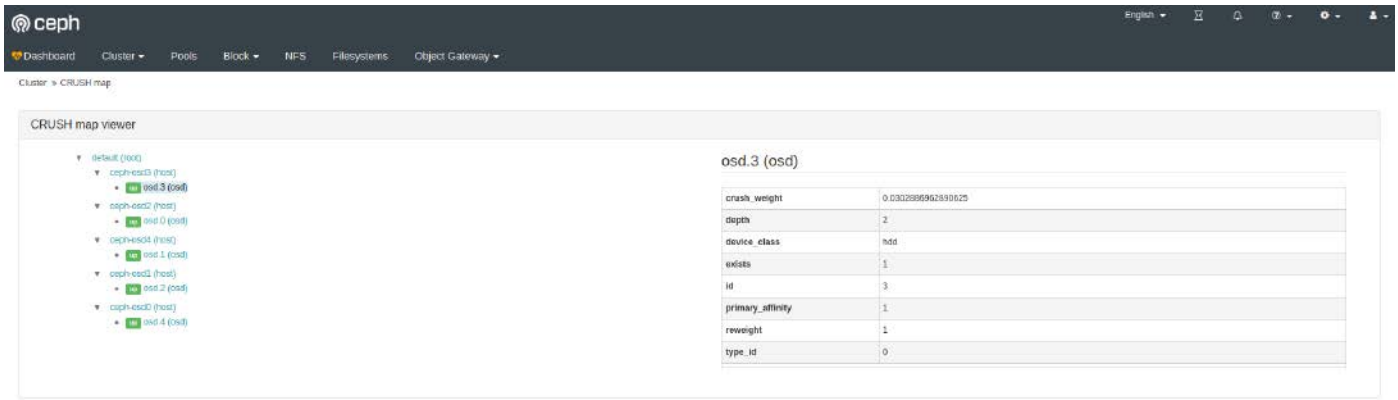


Der Ressourcenbedarf von Recovery- und Rebalancing-Operationen lässt sich nun regulieren (Abb. 2).

Placement-Groups im Griff

Licht bringt die neue Version endlich in das dunkle Geheimnis um die Placement Groups (PGs). Was lange eine Menge Know-how verlangte, kann der Ceph-Cluster nun selbstständig einstellen – nach oben und nach unten, also Richtung OSDs und Richtung Pool. Mit der Funktion Placement Group Autoscale kann der Administrator über die Zielgröße eines Pools die Zahl der PGs ermitteln und entweder automatisch setzen lassen oder anpassen.

Eine weitere offensichtliche Neuerung bildet das Disk Failure Prediction Module im ceph-mgr. Es kann Festplattenausfälle vorhersagen. Das Modul nutzt dazu die SMART-Werte, etwa den `Wear_Leveling_Count` für SSDs, die es im Local-Modus oder über einen Cloud-Dienst wie ProphetStor auswertet. Zu bedenken ist dabei allerdings, dass eine SSD-Firmware beispielsweise einen nicht vorher-



Der neue CRUSH Map Viewer ist zwar noch einfach gestrickt, zeigt aber zumindest die Buckets und Classes an (Abb. 3).

sagbaren Bug haben kann. Die `ceph`-Befehle zum Aktivieren des Moduls und zum Konfigurieren für den lokalen Modus zeigt Listing 2 ebenso wie die Vorhersage des Moduls bei Eingabe des Befehls `ceph device ls`:

Zahlreiche Anpassungen und Verbesserungen gab es zudem unter der Haube von Ceph. Geschlossen ist nun eine Sicherheitslücke: Mit dem MSGR2-Protokoll lassen sich die Verbindungen zwischen den Monitoren, Managern und OSDs nun verschlüsseln. Dazu hat Red Hat den TCP-/UDP-Port 3300 offiziell bei der IANA beantragt. Für die Zukunft plant Red Hat, mit dem neuen Protokoll auch Kerberos zu nutzen.

Verbesserungen gab es auch bei dem noch recht neuen internen Dateisystem Bluestore. Zum Beispiel haben die Entwickler den RAM-Verbrauch verringert.

Top für alle Verbraucher

Den RBDs hat Red Hat eine `top`-Funktion spendiert, die über die Ceph Manager Daemons die Daten sammelt und die Top-Verbraucher auflistet, gefiltert nach Pools, Objects und Clients. Als Aufruf genügt der Befehl:

```
rbd perf image iotop
```

Namespaces sind für den RadosGW im Ceph nichts Neues. Jetzt gibt es diese Namespaces, also die Trennung der Daten über einen Namen, auch für RBDs. Vorteil: Nun lässt sich endlich ein Pool für mehrere Nutzer bereitstellen, aber zugleich die Trennung der Daten der einzelnen Nutzer gewährleisten.

Dazu legt der Administrator die Ceph Keys für einen User an und versieht sie mit einem Namespace:

```
ceph auth get-or-create client.user [...] ⤸
      osd 'profile rbd pool=rbd namespace=tux'
```

Zudem kann er über die Ceph Keys auch das Source-Netz eintragen, aus dem ein Benutzer auf die Ressourcen zugreift.

```
ceph auth get-or-create client.user [...] ⤸
   osd 'profile rbd pool=rbd namespace=tux ⤸
                          network 10.20.30.0/24'
```

Danach muss der Administrator nur noch die Namespaces im RBD anlegen und die Images erstellen.

```
rbd namespace create --namespace tuxrbd ⤸
        create --namespace tux --size 5G tuxrbd
```

Die Images sind dann nur im jeweiligen Namespace sichtbar.

```
# rbd ls --long --namespace tux
NAME    SIZE   PARENT FMT PROT LOCK
tuxrbd 5 GiB   2
```

Ohne die Angabe des Namespace bleibt deshalb die Ausgabe leer:

```
# rbd ls --long
#
```

Viele sind besser als eins

Eine weitere Funktion begegnet den Anwendern im Ceph-Dateisystem. Per Default gibt es in Ceph nur genau ein Filesystem, in der Regel cephfs. Allerdings gibt es Anwendungsfälle, in denen es einfacher und notwendig ist, ein weiteres Filesystem zu haben statt mehrerer Keys gepaart mit Rechten auf bestimmte Verzeichnisse. Bereits Jewel alias Ceph v10 bietet die Option, mehr als ein Filesystem zu aktivieren – das war bisher aber als experimentell eingestuft. Nun ist diese Funktion als stabil gekennzeichnet. Man erstellt es mit dem Befehl

```
ceph fs volume create <name>
```

Fazit

Mit dem neuen Dashboard und den vielen anderen Verbesserungen und Erweiterungen rechtfertigt Ceph den Versionssprung in eine neue Major Release. Sowohl SUSE als auch Red Hat haben die neue Version schon auf ihrer Roadmap. Damit wird auch Ceph Nautilus den Weg in ihre Distribution finden. (sun@ix.de)

Listing 1: Selbst signierte Zertifikate fürs Single Sign-on erstellen

```
# ceph dashboard create-self-signed-cert
Self-signed certificate created
# ceph dashboard ac-user-create admin password administrator
{"username": "admin", "lastUpdate": 1551253931, "name": null, "roles": ["administrator"],
"password": "$2b$12$eQZJgRpRtjWdtYcs5DLVVuvCgUTNIW4mlsC39F/eefVrGYyW.iP7K", "email": null}
```

Listing 2: Disk Failure Prediction Module einrichten

```
# ceph mgr module enable diskprediction_local
# ceph config set global device_failure_prediction_mode local
# ceph device ls
DEVICE                                                HOST:DEV DAEMONS LIFE EXPECTANCY
LVM_PV_1QFuQQ-ekcY-T9n3-3nht-aMGO-tnhP-DMKKwo_on_/dev/sdb_drive-scsi1 ceph-osd3:sdb osd.3 >3M
LVM_PV_QwXbCd-zSGw-Rdyn-IT35-htHC-Uyq6-23OMsr_on_/dev/sdb_drive-scsi1 ceph-osd4:sdb osd.1 >4w
LVM_PV_RlHAgQ-S9AW-5c9c-WY9v-MVYV-MzSw-cyOeXO_on_/dev/sdb_drive-scsi1 ceph-osd1:sdb osd.2 >4w
LVM_PV_Wm3BQn-JHMd-bxM1-MKXU-FB6e-aM8F-Coleb1_on_/dev/sdb_drive-scsi1 ceph-osd2:sdb osd.0 >33h
LVM_PV_cAMvKs-5E8V-cOfv-aRMd-QAtt-CkXo-ejyfrT_on_/dev/sdb_drive-scsi1 ceph-osd0:sdb osd.4 >12M
```

Michel Raabe

arbeitet als Consultant und Trainer mit den Schwerpunkten Cluster, Hochverfügbarkeit und Virtualisierung bei der B1 Systems GmbH.