# Ceph-effectanalyse voor StarOS VNF

## Inhoud



## Inleiding

In dit document wordt beschreven hoe StarOS VNF wordt uitgevoerd, die op Cisco Virtualization Infrastructuur Manager (VIM) draait, wanneer Ceph Storage Service wordt aangetast en wat er kan worden gedaan om de impact te beperken. Het wordt verklaard op de veronderstelling dat de VIM van Cisco als infrastructuur wordt gebruikt maar de zelfde theorie kan op om het even welke OpenStack omgeving worden toegepast.

## Voorwaarden

### Vereisten

Cisco raadt kennis van de volgende onderwerpen aan:

- Cisco Star II9000-OS
- Cisco VIM
- Openstack
- Ceph

### Gebruikte componenten

De informatie in dit document is gebaseerd op de volgende software- en hardware-versies:

- StarOS: 21.16.c9

- Cisco VIM: 3.2.2 (OpenStack Queens)

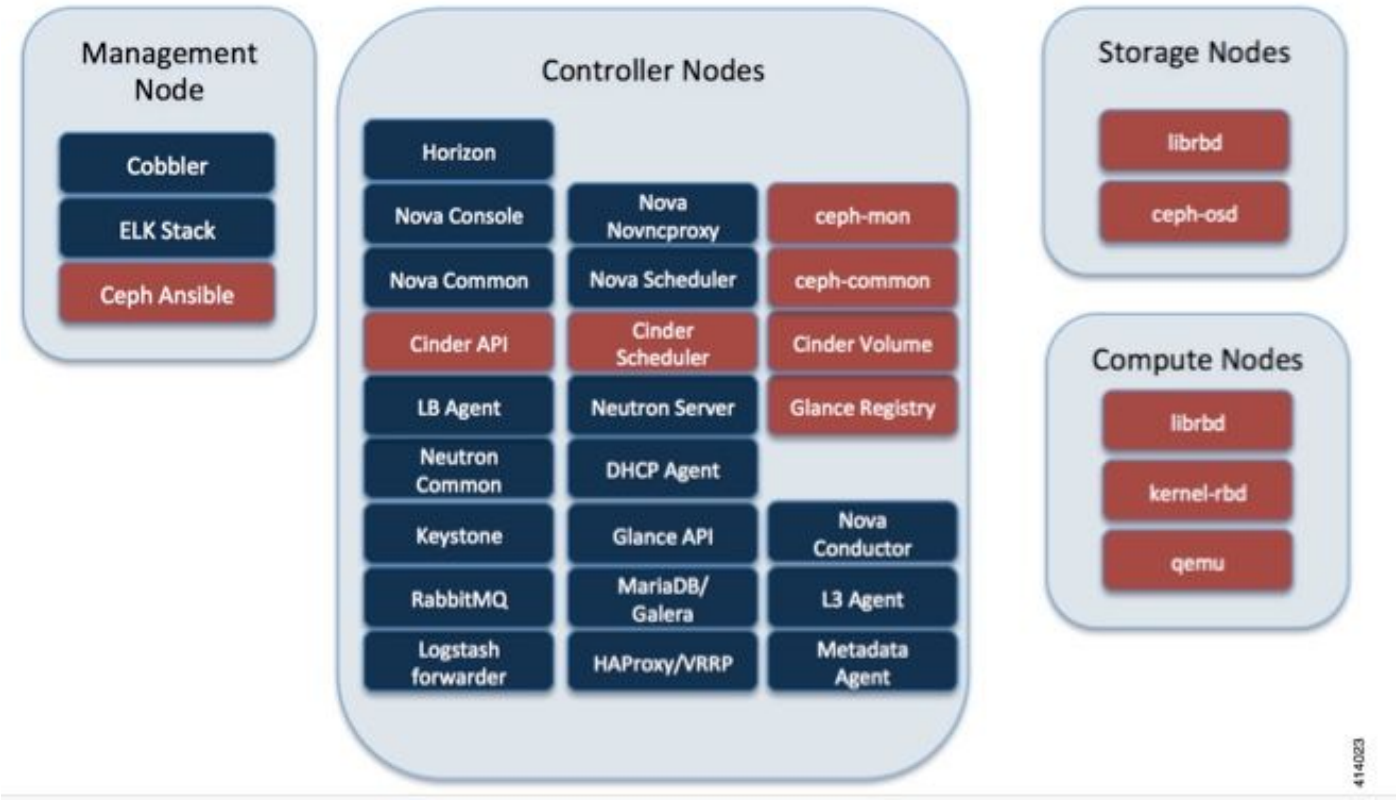
De informatie in dit document is gebaseerd op de apparaten in een specifieke laboratoriumomgeving. Alle apparaten die in dit document worden beschreven, hadden een opgeschoonde (standaard)configuratie. Als uw netwerk levend is, zorg er dan voor dat u de mogelijke impact van om het even welke opdracht begrijpt.

## Afkortingen

| | |
|---|---|
| Cisco VIM | Cisco gevirtualiseerde infrastructuurbeheerder |
| VNF | Virtuele netwerkfunctie |
| Ceph OSD | Opslagdatum voor CEDEX |
| StarOS | Besturingssysteem voor Cisco Mobile Packet Cor oplossing |

## Afdrukken in Cisco VIM

Deze afbeelding is afkomstig uit de Cisco VIM Administrator Guide. Cisco VIM gebruikt Ceph als opslagback-end.




Ceph ondersteunt zowel blokopslag als objectopslag en wordt daarom gebruikt om VM-beelden en -volumes op te slaan die op VM's kunnen worden aangesloten. Meervoudige OpenStack services die afhankelijk zijn van de opslagbackend zijn onder meer:

- Glance (OpenStack Image Service) — Gebruikt Ceph om afbeeldingen op te slaan.
- Cinder (OpenStack Storage Service) — Gebruik Ceph om volumes te maken die aan VM's kunnen worden gekoppeld.

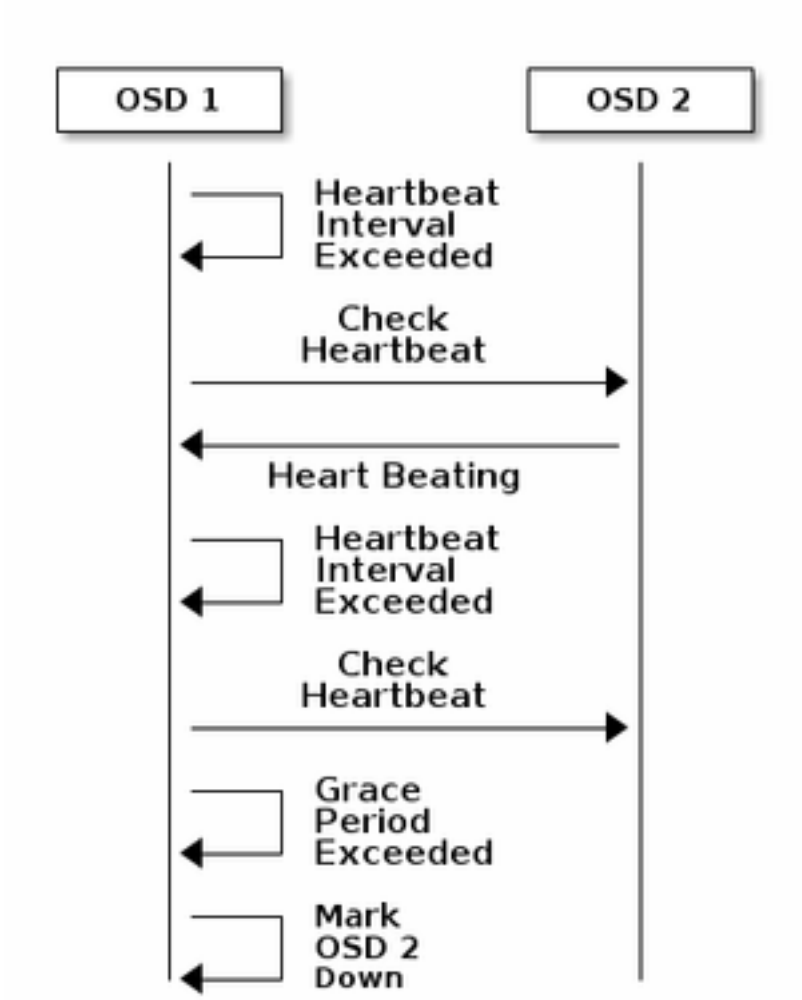
- Nova (OpenStack Computing Service) — Gebruikt Ceph om verbinding te maken met de door Cinder gemaakte volumes.

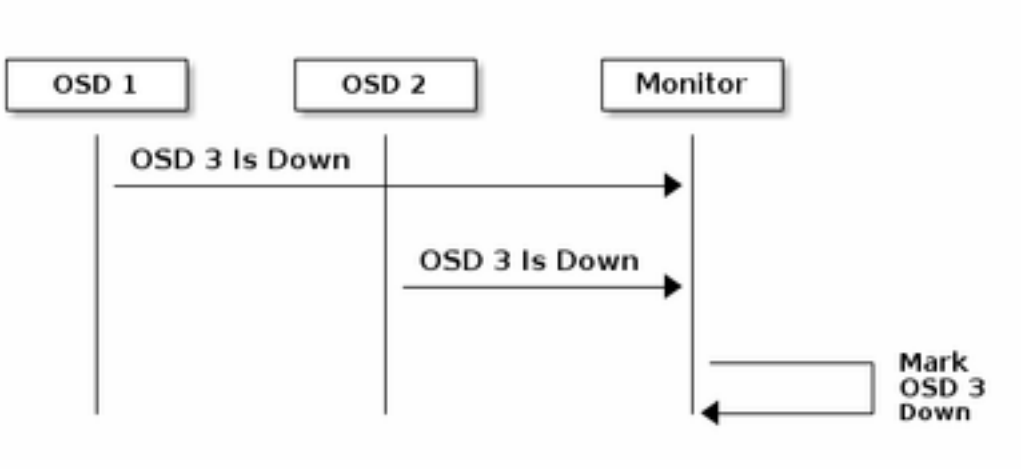
In veel gevallen wordt in Ceph een volume gemaakt voor **/flitser en /hd-aanval** voor StarOS VNF, zoals hier het voorbeeld.

```
openstack volume create --image `glance image-list | grep up-image | awk '{print $2}'` --size 16
--type LUKS up1-flash-boot
openstack volume create --size 20 --type LUKS up1-hd-raid
```

# Basisbeginselen van het bewakingssysteem in Ceph

Hier volgt de verklaring uit het Ceph-document over de controle:
Elke Ceph OSD Daemon controleert de hartslag van andere Ceph OSD Daemons met willekeurige intervallen van minder dan elke 6 seconden. Als een naburige Ceph OSD Daemon geen hartslag vertoont binnen een 20 seconden durende periode, kan de Ceph OSD Daemon de aangrenzende Ceph OSD Daemon overwegen en het naar een Ceph Monitor rapporteren, die de Ceph Cluster Map bijwerkt. Standaard moeten twee Ceph OSD Daemons van verschillende hosts melden aan de Ceph Monitors dat een andere Ceph OSD Daemon is afgevallen voordat de Ceph Monitoring ors erkennen dat de gerapporteerde Ceph OSD Daemon is afgenomen.






Over het algemeen duurt het ongeveer 20 seconden om OSD te detecteren en de Ceph cluster Map wordt bijgewerkt, alleen nadat deze VNF een nieuwe OSD kan gebruiken. Tijdens deze periode wordt de I/O geblokkeerd.

# Invloed van blokkering I/O op StarOS VNF

Als de schijf I/O gedurende meer dan 120 seconden is geblokkeerd, wordt StarOS VPN-herstart. Er is een specifieke controle op xfssyncd/md0 en xfs_db processen die gerelateerd zijn aan disk I/O en StarOS opzettelijke herstart wanneer het een vastgehouden proces meer dan 120 seconden detecteert.

StarOS-debug van console:

```
[ 1080.859817] INFO: task xfssyncd/md0:25787 blocked for more than 120 seconds.
[ 1080.862844] "echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
[ 1080.866184] xfssyncd/md0 D ffff880c036a8290 0 25787 2 0x00000000
[ 1080.869321] ffff880aacf87d30 0000000000000046 0000000100000a9a ffff880a00000000
[ 1080.872665] ffff880aacf87fd8 ffff880c036a8000 ffff880aacf87fd8 ffff880aacf87fd8
[ 1080.876100] ffff880c036a8298 ffff880aacf87fd8 ffff880c0f2f3980 ffff880c036a8000
[ 1080.879443] Call Trace:
[ 1080.880526] [<ffffffff8123d62e>] ? xfs_trans_commit_iclog+0x28e/0x380
[ 1080.883288] [<ffffffff810297c9>] ? default_spin_lock_flags+0x9/0x10
[ 1080.886050] [<ffffffff8157fd7d>] ? _raw_spin_lock_irqsave+0x4d/0x60
[ 1080.888748] [<ffffffff812301b3>] _xfs_log_force_lsn+0x173/0x2f0
[ 1080.891375] [<ffffffff8104bae0>] ? default_wake_function+0x0/0x20
[ 1080.894010] [<ffffffff8123dc15>] _xfs_trans_commit+0x2a5/0x2b0
[ 1080.896588] [<ffffffff8121ff64>] xfs_fs_log_dummy+0x64/0x90
[ 1080.899079] [<ffffffff81253cf1>] xfs_sync_worker+0x81/0x90
[ 1080.901446] [<ffffffff81252871>] xfssyncd+0x141/0x1e0
[ 1080.903670] [<ffffffff81252730>] ? xfssyncd+0x0/0x1e0
[ 1080.905871] [<ffffffff81071d5c>] kthread+0x8c/0xa0
[ 1080.908815] [<ffffffff81003364>] kernel_thread_helper+0x4/0x10
[ 1080.911343] [<ffffffff81580805>] ? restore_args+0x0/0x30
[ 1080.913668] [<ffffffff81071cd0>] ? kthread+0x0/0xa0
[ 1080.915808] [<ffffffff81003360>] ? kernel_thread_helper+0x0/0x10
[ 1080.918411] **** xfssyncd/md0 stuck, resetting card
```

Maar deze is niet beperkt tot de timer 120 seconden. Als de schijf I/O een tijdje is geblokkeerd, zelfs minder dan 120 seconden, kan VNF om verschillende redenen opnieuw worden opgestart. De uitvoer hier is een voorbeeld van een herstart vanwege het disk I/O-probleem, soms een continue StarOS-taakcrash, enzovoort. Het hangt af van de timing van het probleem met actieve disk I/O versus opslag.

```
[ 2153.370758] Hangcheck: hangcheck value past margin!
[ 2153.396850] ata1.01: exception Emask 0x0 SAct 0x0 SErr 0x0 action 0x6 frozen
[ 2153.396853] ata1.01: failed command: WRITE DMA EXT
--- skip ---
SYSLINUX 3.53 0x5d037742 EBIOS Copyright (C) 1994-2007 H. Peter Anvin
```

Een lang blokkerende I/O kan worden beschouwd als een cruciaal probleem voor StarOS VNF en moet zoveel mogelijk worden geminimaliseerd.

# I/O-scenario's met lange blokkering

Gebaseerd op het onderzoek van meerdere klantimplementaties en laboratoriumtesten, zijn er 2 hoofdscenario's geïdentificeerd die een lange blokkerende I/O in Ceph kunnen veroorzaken.

## Mechanisme voor bagagetimer

Er is een hartslagmechanisme tussen OSD's om OSD te detecteren. Gebaseerd op **osd_heartbeat**_value (20 seconden standaard) wordt OSD gedetecteerd als mislukt.
En er is een laggy-timer mechanisme, wanneer er een fluctuatie of flap in OSD-status is, wordt de uitbarsting-timer automatisch aangepast (wordt langer). Dit kan de waarde **van osd_heartbeat**_respe verhogen.

In de normale situatie is de hartslag 20 seconden.

```
2019-01-09 16:58:01.715155 mon.ceph-XXXXX [INF] osd.2 failed (root=default,host=XXXXX) (2
reporters from different host after 20.000047 >= grace 20.000000)
```

Maar na meerdere netwerkflaps van een opslagknooppunt wordt het een grotere waarde.

```
2019-01-10 16:44:15.140433 mon.ceph-XXXXX [INF] osd.2 failed (root=default,host=XXXXX) (2
reporters from different host after 256.588099 >= grace 255.682576)
```

In het bovenstaande voorbeeld duurt het 256 seconden om OSD te detecteren.

## Hardware bij invullen kaart

Ceph kan mogelijk niet tijdig in staat zijn om hardware-uitval van de BANK-kaart op te sporen. Een storing van de inval-kaart leidt tot een soort OSD-ophangsituatie. In dit geval wordt OSD na een paar minuten gedetecteerd, wat voldoende is om StarOS VPN-herstart te maken.

Bij het vasthouden van een DVD-kaart nemen sommige CPU-kernen 100% in status.

```
%Cpu20 : 2.6 us, 7.9 sy, 0.0 ni, 0.0 id, 89.4 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu21 : 0.0 us, 0.3 sy, 0.0 ni, 99.7 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu22 : 31.3 us, 5.1 sy, 0.0 ni, 63.6 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu23 : 0.0 us, 0.0 sy, 0.0 ni, 28.1 id, 71.9 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu24 : 0.0 us, 0.0 sy, 0.0 ni, 0.0 id,100.0 wa, 0.0 hi, 0.0 si, 0.0 st
%Cpu25 : 0.0 us, 0.0 sy, 0.0 ni, 0.0 id,100.0 wa, 0.0 hi, 0.0 si, 0.0 st
```

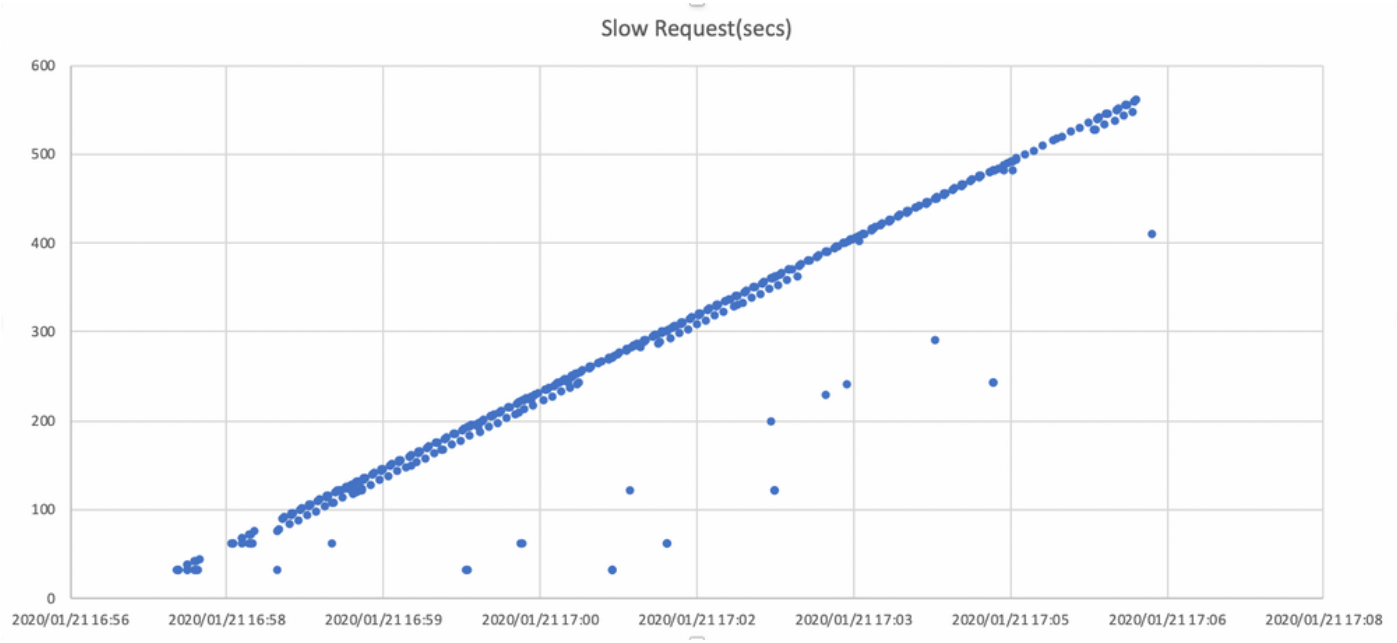
En het eet alle CPU-kernen geleidelijk op en OSD wordt ook geleidelijk minder, met enige tijdspanne.

```
2019-01-01 17:08:05.267629 mon.ceph-XXXXX [INF] Marking osd.2 out (has been down for 602
seconds)
2019-01-01 17:09:25.296955 mon.ceph-XXXXX [INF] Marking osd.4 out (has been down for 603
seconds)
2019-01-01 17:11:10.351131 mon.ceph-XXXXX [INF] Marking osd.7 out (has been down for 604
seconds)
2019-01-01 17:16:40.426927 mon.ceph-XXXXX [INF] Marking osd.10 out (has been down for 603
seconds)
```

Tegelijkertijd worden langzame verzoeken gedetecteerd in **ceph.log**.

```
2019-01-01 16:57:26.743372 mon.XXXXX [WRN] Health check failed: 1 slow requests are blocked > 32
sec. Implicated osds 2 (REQUEST_SLOW)
2019-01-01 16:57:35.129229 mon.XXXXX [WRN] Health check update: 3 slow requests are blocked > 32
sec. Implicated osds 2,7,10 (REQUEST_SLOW)
2019-01-01 16:57:38.055976 osd.7 osd.7 [WRN] 1 slow requests, 1 included below; oldest blocked
for > 30.216236 secs
2019-01-01 16:57:39.048591 osd.2 osd.2 [WRN] 1 slow requests, 1 included below; oldest blocked
for > 30.635122 secs
-----skip-----
2019-01-01 17:06:22.124978 osd.7 osd.7 [WRN] 78 slow requests, 1 included below; oldest blocked
for > 554.285311 secs
2019-01-01 17:06:25.114453 osd.4 osd.4 [WRN] 19 slow requests, 1 included below; oldest blocked
for > 546.221508 secs
2019-01-01 17:06:26.125459 osd.7 osd.7 [WRN] 78 slow requests, 1 included below; oldest blocked
for > 558.285789 secs
2019-01-01 17:06:27.125582 osd.7 osd.7 [WRN] 78 slow requests, 1 included below; oldest blocked
for > 559.285915 secs
```

De grafiek hier toont hoe lang I/O verzoeken met een tijdlijn worden geblokkeerd. De grafiek wordt gecreëerd door de langzame aanvraagdossiers in ceph.log uit te zetten. Het laat zien dat de blokkeringstijd in de loop der tijd langer wordt.




# Hoe de impact te verzachten?

## Naar lokale schijf verplaatsen tijdens Ceph-opslag

De eenvoudigste manier om deze gevolgen te verzachten is door van Ceph-opslag naar een

lokale schijf te verplaatsen. StarOS maakt gebruik van 2 disks, /flitser en /hd-aanval, het is mogelijk om alleen /flitser naar lokale schijf te verplaatsen waardoor StarOS VNF robuuster wordt voor de Ceph-problemen. De negatieve kant van het gebruik van gedeelde opslag zoals Ceph is dat alle VNF die het gebruikt, op hetzelfde moment wordt beïnvloed wanneer een probleem zich voordoet. Door gebruik van een lokale schijf kan de impact van de opslagkwestie worden geminimaliseerd tot VPN dat alleen op het getroffen knooppunt actief is. En de scenario's die in de voorgaande sectie worden genoemd zijn alleen van toepassing op Ceph, dus niet van toepassing op lokale schijf. Maar de keerzijde van de lokale schijf is dat de inhoud van de schijf, zoals StarOS-beeld, configuratie, kernbestand, facturering, niet kan worden bewaard wanneer VM wordt hergebruikt. Dit kan ook van invloed zijn op het VPN-automatische genezingsmechanisme.

## Ceph-configuratie - instelbaarheid

Vanuit het oogpunt van StarOS VNF worden de volgende nieuwe Ceph-parameters aanbevolen om de bovengenoemde blokkerende I/O-tijd te minimaliseren.

<standaardinstellingen>

```
"mon_osd_adjust_heartbeat_grace": "true",
"osd_client_watch_timeout": "30",
"osd_max_markdown_count": "5",
"osd_heartbeat_grace": "20",
```

<nieuwe instellingen>

```
"mon_osd_adjust_heartbeat_grace": "false",
"osd_client_watch_timeout": "10",
"osd_max_markdown_count": "1",
"osd_heartbeat_grace": "10",
```

Het bestaat uit:

- Het lagertimer mechanisme is uitgeschakeld, geen automatische aanpassing
- De vaardige tijd is verkort
- OSD wordt direct als beneden gemarkeerd (standaard 5 keer in de laatste 600 seconden)

De nieuwe parameters worden getest in een lab, de detectietijd voor OSD is gereduceerd tot ongeveer 10 seconden, het was oorspronkelijk ongeveer 30 seconden met de standaardconfiguratie van Ceph.

## Hardware-uitgifte voor monitor-kaart

Voor het hardware-scenario van de BANK-kaart is het wellicht nog steeds moeilijk tijdig te onderkennen als de aard van de kwestie, omdat hierdoor een situatie ontstaat waarin OSD met tussenpozen werkt terwijl I/O wordt geblokkeerd. Er is geen enkele oplossing voor dit, maar het wordt aanbevolen om het hardware-logbestand van de server te controleren op een storing van de inval van de inval van de kaart, of het loggen van de langzame aanvraag in ceph.log door een of ander script en maatregelen te nemen om de aangetaste OSD proactief te verminderen.

## CEPH_OSD_RESEREVED_PCORES-tuning

Dit is niet gerelateerd aan de genoemde scenario's, maar als er een probleem is met de prestaties

van Ceph door zware I/O-activiteiten, kan het verhogen van de waarde van CEPH_OSD_RESEREVED_PCORES de prestaties van Ceph I/O verbeteren. Door standaard is CEPH_OSD_RESEREVED_CORES op Cisco VIM ingesteld als 2 en kan worden verhoogd.