HITACHI
Inspire the Next

Modern Linux Clients Can Overwhelm HNAS Due to High Parallel Client Writing

Symptom

- HNAS NFS exports can experience very poor performance when a large amount (tens of thousands) of parallel requests occur from modern Linux clients. The HNAS will try to handle the requests, but it can eventually result in 100% of the network receive fibers being used (bossock and/or pi-tcp-socket) which will deny additional network connections and could cause stuck threads on the HNAS and either a hung state for some time or possibly a node reset.

Environment

- Hitachi NAS Platform
  - Hitachi NAS Platform 4040 (HNAS 4040)
  - Hitachi NAS Platform 4060 (HNAS 4060)
  - Hitachi NAS Platform 4080 (HNAS 4080)
  - Hitachi NAS Platform 4100 (HNAS 4100)
  - Hitachi NAS Platform 3080 (HNAS 3080)
  - Hitachi NAS Platform 3090 (HNAS 3090)
  - Hitachi NAS Platform 3100 (HNAS 3100)
  - Hitachi NAS Platform 3200 (HNAS 3200)

Resolution

- Change the Linux clients **tcp_max_slot_table_entries** (has been tested with **RHEL** and and **Ubuntu**) by running the following command (this does not require a reboot, however the Linux client will have to re-mount the HNAS NFS export):
  - `sudo sysctl sunrpc.tcp_max_slot_table_entries=128`
- For Linux servers where the *sysctl* service runs before the *sunrpc* kernel module is loaded, create a **sunrpc.conf** file in the **/etc/modprobe.d** directory with the following parameter within the file (this does not require a reboot, however the Linux client will have to re-mount the HNAS NFS export):
  - `options sunrpc tcp_max_slot_table_entries=128`

Cause

Around 2012, the Linux kernel community decided to change the default size for TCP slots from 16 to 65536, which opens up any particular client to send up to 65536 parallel requests to the HNAS (or most any NAS appliance). It has been determined that 128 is a more ideal number and prevents an overload of client requests going to the HNAS from a single Linux client.

Additional Notes

- Microcode Affected: 11.2.3313.00 and above
- For more information on this issue, click here.

Steps to check what the current setting is and how it is affecting the HNAS:

Linux Ubuntu Client

```
manager@ubuntu1604:/mnt/evs4$ cat /proc/sys/sunrpc/
sunrpc.tcp_max_slot_table_entries   <-- check the setting

65536

manager@ubuntu1604:/mnt/evs4$ sudo time dd if=/dev/zero of=bigfile8 bs=1M
count=100000  <-- start some I/O

100000+0 records in

100000+0 records out

104857600000 bytes (105 GB, 98 GiB) copied, 168.18 s, 623 MB/s

0.02user 49.34system 2:48.20elapsed 29%CPU (0avgtext+0avgdata 3016maxresident)k

24inputs+204800000outputs (2major+337minor)pagefaults 0swaps
```

HNAS

```
$ show-stats running

         1              Running Bossock Fibers (#)

       512  0x00000200 Running pi-tcp-sockets receive Fibers (#)        <-- it goes to 512
right away
```

```
        0            Running pi-tcp-sockets control Fibers (#)
```

Steps to change the current setting and verify the change on the HNAS:

Linux Ubuntu Client

```
manager@ubuntu1604:~$ sudo sysctl sunrpc.tcp_max_slot_table_entries=128   <-- change the setting

manager@ubuntu1604:~$ cat /proc/sys/sunrpc/ sunrpc.tcp_max_slot_table_entries

128

manager@ubuntu1604:~$ sudo umount /mnt/evs4                               <-- un-mount the export

manager@ubuntu1604:~$ sudo mount 172.20.252.24:/evs4 /mnt/evs4/          <--  mount the export

manager@ubuntu1604:~$ sudo time dd if=/dev/zero of=/mnt/evs4/bigfile4 bs=1M
count=100000   <-- start some I/O

100000+0 records in

100000+0 records out

104857600000 bytes (105 GB, 98 GiB) copied, 121.895 s, 860 MB/s

0.05user 48.68system 2:01.90elapsed 39%CPU (0avgtext+0avgdata 3012maxresident)k

16inputs+204800000outputs (1major+338minor)pagefaults 0swaps

manager@ubuntu1604:~$
```

HNAS

$ show-stats running

```
        1            Running Bossock Fibers (#)

      119 0x00000077 Running pi-tcp-sockets receive Fibers (#)    <-- the setting goes no higher than 128 (currently at 119)

        0            Running pi-tcp-sockets control Fibers (#)
```

```
$ show-stats running

       1            Running Bossock Fibers (#)

     111 0x0000006f Running pi-tcp-sockets receive Fibers (#)

       0            Running pi-tcp-sockets control Fibers (#)
```

```
$ show-stats running

       1            Running Bossock Fibers (#)

      76 0x0000004c Running pi-tcp-sockets receive Fibers (#)

       0            Running pi-tcp-sockets control Fibers (#)
```

The *pi-tcp-sockets* never went above **128** after the setting change in “**sunrpc.tcp_max_slot_table_entries**” from 65536 to 128 .

Internal Notes

Related defects are:

D91194: even a small number of modern (from the era of RHEL6.3, Debian Jessie or later) Linux clients can overwhelm HNAS due to highly parallelized client writing

D129935: Problems due to high parallel load tx buffer pending request fix in 13.0 (short term fix)

D130196: Problems due to high parallel load tx buffer pending request fix in 13.0 (longer term fix)

D130908: The high parallel load fix may not actually fix the high parallel load problem when used with a very aggressive internal test app