

Errors writing large files via NFS

Source: <http://unix.derkeiler.com/Mailing-Lists/Tru64-UNIX-Managers/2003-09/0183.html>

From: Meiklejohn David – dimeikl (david.meiklejohn_at_acxiom.com)

Date: 09/25/03

Date: Thu, 25 Sep 2003 13:40:18 +1000

To: Tru64-UNIX-Managers@ornl.gov

Managers,

We have a strange problem (aren't they all?), where we are unable to copy files larger than a certain size to a NFS server. This server exports a filesystem (/archive) to a number of other Tru64 boxes. The problem of being unable to write large files to this filesystem shows up on all the client systems, although the definition of "too large" varies. Some of the clients are on the same network as the server, and access it through a lag interface. The other clients are in another data centre and connect through a WAN link via a single (non-lag) interface. Since there are quite distinct network paths involved, I'm pretty sure we're not seeing a network problem.

The server is running Tru64 v5.1A. It was at patch kit 4, and running with no problems, then we installed PK5. We then had some problems with applications failing to write out large files (e.g. 2GB) properly. The jobs would appear to complete, but the output file would be smaller than expected. Sorting of large files (we use AST Syncsort) would fail. We raised a call with HP, but they had no reports of any similar problem with PK5 and no suggested solution. So we backed it out by deleting the patches, using dupatch. The jobs that were failing now run ok again, but we now have this NFS problem.

We are pretty sure that it is it an NFS problem, and not the filesystem nor any other network later, since there are no problems with using 'rcp' to copy files onto this server. And yet the same file, copied from the same client over the same network connection, to the same location, will fail to copy successfully if we use 'cp' to copy it to an NFS mount.

To explore this, I tried using 'dd' to create files on the various clients, in the filesystem mounted from the NFS server. That would allow me to see how far it got before falling over.

What is really interesting is that this problem depends on the block size. E.g. from one client, with NFS parameters (v3, rw, tcp, hard, intr) over a WAN link, I get:

```
root@bunyip:/archive # dd if=/dev/zero of=blah bs=65536
```

Tru64–UNIX–Managers: Errors writing large files via NFS

```
dd write error: I/O error
3654+0 records in
3653+0 records out
```

```
root@bunyip:/archive # dd if=/dev/zero of=blah bs=65536
dd write error: I/O error
3527+0 records in
3526+0 records out
```

```
root@bunyip:/archive # dd if=/dev/zero of=blah bs=8192
dd write error: I/O error
26057+0 records in
26056+0 records out
```

```
root@bunyip:/archive # dd if=/dev/zero of=blah bs=8192
dd write error: I/O error
31833+0 records in
31832+0 records out
```

```
root@bunyip:/archive # dd if=/dev/zero of=blah bs=4096
148337+0 records in
148336+0 records out
```

I.e. with either 64k or 8k blocks, an I/O error shows up after about 230MB have been written. But reduce the block size to 4k, and it had written out *>500MB before I killed it. Subsequent testing shows that I can create arbitrarily large files if the block size is 4k or less, but if anything larger than 4k is used, I can't create any file bigger than 250MB.*

So I thought, maybe setting the NFS read and write buffer sizes on the client to 4k would fix the problem. But no, that only changes the error message:

```
root@bunyip:/archive # dd if=/dev/zero of=blah bs=65536
dd: UNABLE TO write complete record
3128+0 records in
3127+1 records out
```

It still falls over in that 200–250MB range.

Each of these failures corresponds to a `/var/adm/messages` log entry of: `"vmunix: NFS3 write error 5 on host brian"`.

And yet there is no corresponding error logged on the NFS server (brian), either in messages or the event logs.

Repeating the same test on another client, with the same filesystem mounted in the same way (v3, rw, tcp, hard, intr), over the same WAN link, also fails, but with a different file size limit:

Tru64-UNIX-Managers: Errors writing large files via NFS

```
root@echidna:/archive # dd if=/dev/zero of=blah bs=65536
dd write error: I/O error
11386+0 records in
11385+0 records out
```

I.e. this client can write out about 730MB before the I/O error hits.
Again, we see "NFS3 write error 5" in the messages file.

Where it gets really interesting is when we look at a client on the same LAN (connected to the same switch) as the server, where the filesystem is mounted over UDP:

```
root@dropbear:/archive # dd if=/dev/zero of=blah bs=65536
dd: UNABLE TO write complete record
18093+0 records in
18092+1 records out
```

I.e. it still falls over, but the limit is >1GB.

Finally, remounting that filesystem onto the same client, using TCP instead of UDP, leads to:

```
root@dropbear:/archive # dd if=/dev/zero of=blah bs=65536
126087+0 records in
126086+0 records out
dd close error: Interrupted system call
```

I killed it after it wrote out more than 8GB. So there is no sign of any problem for a TCP mount over a local network. But there are problems for a UDP mount on a local network, and for TCP mounts over WAN links. I also tried UDP over the WAN, but that has always been unusable, as is well known.

So – does anyone have any ideas?

Thanks,

David Meiklejohn
Acxiom

The information contained in this communication is confidential, is intended only for the use of the recipient named above, and may be legally privileged.

If the reader of this message is not the intended recipient, you are hereby notified that any dissemination, distribution, or copying of this communication is strictly prohibited.

If you have received this communication in error, please re-send this communication to the sender and delete the original message or any copy of it from your computer system. Thank You.