NFS/RDMA Update

- Check OFA Developer Workshop Presentation NFS/RDMA Update
Why not NFS/RDMA?

• Better Performance:
  ◦ High throughput
  ◦ Low latency
  ◦ Less CPU utilization

• Better Price:
  ◦ Utilize existing fabrics: no cost moving from IPoIB to RDMA

• Distros support:
  ◦ RHEL 7.1 support: client
  ◦ Oracle UEK3: client

Large I/O Bandwidth (iozone 16M)

Large I/O BW
(single thread vs multiple threads)

- RDMA write iops
- IPOIB write iops
- RDMA read iops
- IPOIB read iops

16m-1thread  16m-6threads  16m-1thread-vf  16m-6threads-vf
Small I/O IOPS (iozone 4K & 8K)

Small I/O IOPS (single thread vs. multiple threads)

- RDMA write iops
- IPOIB write iops
- RDMA read iops
- IPOIB read iops
NFS Server Set Up

NFS/RDMA server:

1. Exportfs: /etc/exports
2. Start rdma service: service rdma start
3. Load svcrdma module: modprobe svcrdma
4. Start NFS service: service nfs start
5. Configure IPoIB interface
6. Add “rdma 20049” to portlist:
   
   ```
   echo "rdma 20049" > /proc/fs/nfsd/portlist
   ```
7. Check exportfs: exportfs -v
NFS Client Set Up

NFS/RDMA client:

1. Start rdma service: service rdma start
2. Load xprtrdma module: modprobe xprtrdma
3. Start NFS service: service nfs start
4. Configure IPoIB interface:
5. Mount: mount -t nfs -o vers=3,proto=rdma,port=20049,wsize=256k,rsize=256k Server-IpoIB address:/export/dir /mountpoint
6. Check mount: mount
Benchmark tool: iozone

iozone: (http://iozone.org/)

◦ Operations:
  • Read – reading a file that already exists in the filesystem.
  • Write – writing a new file to the filesystem.
  • Re-read – reading a file again.
  • Re-write – writing to an existing file.
  • Random Read – reading random information from the file.
  • Random Write – writing to a file in various random locations.

◦ Single stream measurement:
  • iozone -I -c -r 4k -s 1g -f /mnt/tmp1

◦ Multiple stream measurement:
  • iozone -I -c -l 2 -u 2 -r 16k -s 1g -t -F /mnt/tmp1 /mnt/tmp2
Workload Simulation Tool: fio

fio - flexible I/O tester (http://pkgs.repoforge.org/fio/)

- Throughput (Read+Write IOPS Aggregate)
- Average Latency (Read+Write Latency Averaged Together)
- Max Latency (Peak Read or Write Latency)
- Basic parameters:
  - IO type, depth, size,
  - Block size
  - Num files,
  - Num Threads
configuration file sample:

```
[global]
direct=1
size=1G
bsrange=4k-4k
timeout=300
numjobs=4 ; 4 simultaneous threads for each job
ioengine=libaio
[f1]
rw=write
[f2]
stonewall
rw=randwrite
[f3]
stonewall
rw=read
[f4]
stonewall
rw=randread
```
RPC Latency: mountstat

- mountstats

- per-op statistics

READ:

- 8193 ops (33%) 0 retrans (0%) 0 major timeouts
- avg bytes sent per op: 136 avg bytes received per op: 262224
- backlog wait: 0.005248 RTT: 41.043208 total execute time: 41.069694 (milliseconds)

- rpc operation backlog wait: queued for transmission
- rpc operation response time: RTT
- rpc operation total execute time: RTT + queue time
RPCDEBUG:

rpcdebug -vh
rpcdebug -m module
rpcdebug -m module -s flags...
rpcdebug -m module -c flags...

Setting these flags causes the kernel to emit messages to the system log in response to NFS activity.

Setting -m rpc -s xprt call trans

Specify which module's flags to set or clear. Available modules are:

- nfsd The NFS server.
- nfs The NFS client.
- nlm The Network Lock Manager, in either an NFS client or server.
- rpc The Remote Procedure Call module, in either an NFS client or server.
Deployment

Any plan to move from NFS/IPoIB to NFS/RDMA?

How many clients?

How many servers?

What's your workload?

What's your fabrics?

What's your favorite distribution?
Thank You