Technical Notes
IBM Oracle International Competency Center (ICC)

October 15, 2019 ***Email address: ibmoracle@us.ibm.com***

IBM Spectrum Scale and Oracle Database 12cR2 RAC on IBM Power Systems

Introduction

IBM Spectrum Scale, previously known as General Parallel File System or GPFS, is a high-performance clustered file system solution that can be used in various Oracle Database Real Application Cluster (RAC) configurations.

This article provides basic installation requirements and tasks for Oracle customers intending to install the Oracle Database 12cR2 RAC with IBM Spectrum Scale file system on IBM Power® servers with the AIX® operating system. Please see the performance tuning and availability recommendations after the installation section.

Installation Environment

This article is applicable to:

Oracle Database Real Application Cluster (RAC) - Enterprise Edition - Version 11.2.0.4 to 19c
IBM Spectrum Scale - Version 4.1, 4.2 and 5.0
IBM AIX Operating System -- Version 7.1 and 7.2

Acceptable uses of the IBM Spectrum Scale (GPFS) file system include:

1. Shared ORACLE_HOME directory for Oracle Database installation
2. Shared database files for tablespaces and other general database object containers
3. Oracle Clusterware registry and membership files (i.e. Oracle Cluster Registry (OCR) and Vote Disks)*

* For Oracle Database 12cR2, Oracle only supports ASM/NFS storage formats for Oracle Grid Infrastructure OCR and Vote files. The 12.2 upgrade requires that these files are migrated to ASM. IBM Spectrum Scale will still be used for database files in this release. In Oracle Database release 19.3 of 19c, this restriction is removed and IBM Spectrum Scale may be used again for OCR and Vote files

Version Support:

Certified combinations of IBM Spectrum Scale and Oracle Real Application Cluster databases are found using the “Certifications” tab in My Oracle Support. Confirm the versions to be deployed are certified prior to installation.

Solution:

The following installation example uses a 4-node server cluster with shared disks and the required public and private networks as defined in the relevant Oracle RAC installation manuals and MOS notes. For our documentation purposes, a server environment is provided with the following attributes also having all required tuning and prerequisite configuration tasks completed.

1. Four IBM AIX 7.1 TL5 SP1 (with latest fixes) partitions/nodes with host names of NODE1, NODE2, NODE3 and NODE4.

2. Network configuration as follows:

NIC	Host Name	Network	Purpose	Defined in Oracle GI?
en0	admin-srv.domain.com	127.46.70.0/24	Admin net	NO
en1	node1,node2,node3,node4	192.168.67.0/22	Oracle RAC public	YES
en2	node1-p1,node2-p1,node3-p1,node4-p1	10.30.10.0/24	Oracle RAC private #1	YES
en3	node1-p2,node2-p2,node3-p2,node4-p2	10.30.20.0/24	Oracle RAC private #2	YES
en4	node1-gpfs,node2-gpfs,node3-gpfs,node4-gpfs	172.30.10.0/24	IBM Spectrum Scale	NO

Notes: The IBM Spectrum Scale network transmits management and configuration data (no data) which doesn’t require dedicated high-bandwidth connections. Availability on the IBM Spectrum Scale network is provided with standard link aggregation techniques such as 802.3ad and Etherchannel.

3. IBM Spectrum Scale (gpfs*) LPPs installed including latest maintenance level:

```
  gpfs.adv            4.2.3.8    APPLIED   GPFS Advanced Features
  gpfs.base           4.2.3.8    APPLIED   GPFS File Manager
  gpfs.crypto         4.2.3.8    APPLIED   GPFS Cryptographic Subsystem
  gpfs.ext            4.2.3.8    APPLIED   GPFS Extended Features
  gpfs.gskit          8.0.50.75  COMMITTED GPFS GSKit Cryptography Runtime
  gpfs.license.adv    4.2.3.0    COMMITTED IBM Spectrum Scale Advanced Edition License
  gpfs.msg.en_US      4.2.3.7    APPLIED   GPFS Server Messages - U.S. English
```

4. Secure Shell (SSH) and Copy (SCP) installed with user-equivalence for the 'root' user -- also known as password-less configuration.

```
root@node1:/ => ssh node1-gpfs 'oslevel -s; hostname'
7100-05-01-1731
node1

root@node1:/ => ssh node2-gpfs 'oslevel -s; hostname'
7100-05-01-1731
node2

root@node1:/ => ssh node3-gpfs 'oslevel -s; hostname'
7100-05-01-1731
node3

root@node1:/ => ssh node4-gpfs 'oslevel -s; hostname'
7100-05-01-1731
node4

Note:  There should be no prompts for password.
```

5. Create cluster using the specified host name dedicated for IBM Spectrum Scale.

```
Create a node configuration file with definitions as follows:

File name:  /tmp/gpfs.nodes
node1-gpfs:manager-quorum
node2-gpfs:manager-quorum
node3-gpfs:manager-quorum
node4-gpfs:manager-quorum

Create cluster:

root@node1:/ =>  mmcrcluster -N  /tmp/gpfs.nodes --ccr-enable -r /usr/bin/ssh -R
/usr/bin/scp -C rac_cluster

Verify cluster creation with the following command:

root@node1:/ => mmlscluster

GPFS cluster information
========================
  GPFS cluster name:         rac-cluster
  GPFS cluster id:           3091760955773297849
  GPFS UID domain:           rac-cluster
  Remote shell command:      /usr/bin/ssh
  Remote file copy command:  /usr/bin/scp
  Repository type:           CCR

 Node  Daemon node name  IP address   Admin node name  Designation
-----------------------------------------------------------------
   1   node1-gpfs        172.30.10.1  node1-gpfs       quorum-manager
   2   node2-gpfs        172.30.10.2  node2-gpfs       quorum-manager
   3   node3-gpfs        172.30.10.3  node3-gpfs       quorum-manager
   4   node4-gpfs        172.30.10.4  node4-gpfs       quorum-manager

Note:  The cluster was created using a Cluster Configuration Repository (--ccr-enable)
which stores configuration data on all quorum nodes.  This is done in preference of
the legacy option (--ccr-disable) where the configuration is stored on a primary (-p)
and backup or secondary node (-s).  Oracle RAC hub nodes should be given both the
quorum and manager roles.
```

6. License the IBM Spectrum Scale nodes for server functionality.

```
root@node1:/ => mmchlicense server --accept -N node1-gpfs,node2-gpfs,node3-gpfs,node4-
gpfs
```

7. Start the IBM Spectrum Scale Cluster.

```
root@node1:/ => mmstartup -a

Check that all nodes are started:

root@node1:/ => mmgetstate -a

 Node number  Node name    GPFS state
------------------------------------------
       1      node1-gpfs       active
       2      node2-gpfs       active
       3      node3-gpfs       active
       4      node4-gpfs       active

Note:  The nodes will be in an “arbitrating” state before they become “active”.
```

8. Create shared disk devices known as Network Shared Disks (NSDs) used by IBM Spectrum Scale file systems.

```
Create multiple NSD configuration files for each planned file system as follows:

File name: /tmp/gpfs.orahome.nsd
%nsd:
       device=/dev/hdisk2
       nsd=orahome
       usage=dataAndMetadata

File name: /tmp/gpfs.data1.nsd
%nsd:
       device=/dev/hdisk3
       nsd=oradata1
       usage=dataAndMetadata
%nsd:
       device=/dev/hdisk4
       nsd=oradata2
       usage=dataAndMetadata

File name: /tmp/gpfs.data2.nsd
%nsd:
       device=/dev/hdisk5
       nsd=oradata3
       usage=dataAndMetadata
%nsd:
       device=/dev/hdisk6
       nsd=oradata4
       usage=dataAndMetadata


File name: /tmp/gpfs.redo1.nsd
%nsd:
       device=/dev/hdisk7
       nsd=oraredo1
       usage=dataAndMetadata
%nsd:
       device=/dev/hdisk8
```

```
       nsd=oraredo2
       usage=dataAndMetadata


File name: /tmp/gpfs.redo2.nsd
%nsd:
       device=/dev/hdisk9
       nsd=oraredo3
       usage=dataAndMetadata
%nsd:
       device=/dev/hdisk10
       nsd=oraredo4
       usage=dataAndMetadata


Create NSDs:

root@node1:/ => mmcrnsd -F /tmp/gpfs.orahome.nsd            # for shared Oracle db
home
root@node1:/ => mmcrnsd -F /tmp/gpfs.data1.nsd              # for DATA
root@node1:/ => mmcrnsd -F /tmp/gpfs.data2.nsd              # also for DATA
root@node1:/ => mmcrnsd -F /tmp/gpfs.redo1.nsd              # for REDO group A
root@node1:/ => mmcrnsd -F /tmp/gpfs.redo2.nsd              # for REDO group B

Verify NSDs:

root@node1:/ => mmlsnsd

 File system   Disk name    NSD servers
---------------------------------------------------------------------------
 (none)       oradata1     (directly attached)
 (none)       oradata2     (directly attached)
 (none)       oradata3     (directly attached)
 (none)       oradata4     (directly attached)
 (none)       orahome      (directly attached)
 (none)       oraredo1     (directly attached)
 (none)       oraredo2     (directly attached)
 (none)       oraredo3     (directly attached)
 (none)       oraredo4     (directly attached)


root@node1:/ => lspv

hdisk0          00f617d612d351cc       rootvg          active
hdisk1          00f617d631c56ab3       gridvg          active
hdisk2          none                                 orahome
hdisk3          none                                 oradata1
hdisk4          none                                 oradata2
hdisk5          none                                 oradata3
hdisk6          none                                 oradata4
hdisk7          none                                 oraredo1
hdisk8          none                                 oraredo2
hdisk9          none                                 oraredo3
hdisk10         none                                 oraredo4
```

9. Create IBM Spectrum Scale file systems using same NSD configuration files as previous step.

```
root@node1:/ => mmcrfs gfsorahome -F /tmp/gpfs.orahome.nsd -B 512k -A yes --mount-
priority 1 -T /u01/12c
root@node1:/ => mmmount gfsorahome -a
```

```
root@node1:/ => mmcrfs gfsdata1 -F /tmp/gpfs.data1.nsd -B 512k -A yes  --mount-
priority 2 -T /u01/12c/data1
root@node1:/ => mmcrfs gfsdata2 -F /tmp/gpfs.data2.nsd -B 512k -A yes  --mount-
priority 2 -T /u01/12c/data2
root@node1:/ => mmcrfs gfsredo1 -F /tmp/gpfs.redo1.nsd -B 256k -A yes  --mount-
priority 2 -T /u01/12c/redo1
root@node1:/ => mmcrfs gfsredo2 -F /tmp/gpfs.redo2.nsd -B 256k -A yes  --mount-
priority 2 -T /u01/12c/redo2
```

Note: The base file system /u01/12c has a lower mount priority of '1' so it will be mounted before the dependent file systems. The administrator should initially create and mount the base file system first so that dependent file systems will have their mount point directory created and contained in the base file system.

Verify file system creation:

```
root@node1:/ => mmlsnsd

 File system   Disk name    NSD servers
---------------------------------------------------------------------------
 gfsdata1      oradata1     (directly attached)
 gfsdata1      oradata2     (directly attached)
 gfsdata2      oradata3     (directly attached)
 gfsdata2      oradata4     (directly attached)
 gfsorahome    orahome      (directly attached)
 gfsredo1      oraredo1     (directly attached)
 gfsredo1      oraredo2     (directly attached)
 gfsredo2      oraredo3     (directly attached)
 gfsredo2      oraredo4     (directly attached)
```

Mount all file systems:

```
root@node1:/ => mmmount all -a
```

```
root@node1:/ => df -g
Filesystem    GB blocks      Free %Used    Iused %Iused Mounted on
/dev/hd4           3.00      0.99   68%    18050     8% /
/dev/hd2           5.00      2.48   51%    43130     7% /usr
/dev/hd9var        8.00      7.70    4%     4179     1% /var
/dev/hd3          64.00     63.58    1%      381     1% /tmp
/dev/hd1           8.00      7.39    8%       48     1% /home
/dev/hd11admin      0.25      0.25    1%        5     1% /admin
/proc                 -         -    -         -      - /proc
/dev/hd10opt       0.50      0.10   80%     4909    17% /opt
/dev/gfsorahome    120.00    105.82   12%    41280    34% /u01/12c
/dev/gfsdata1    640.00    637.44    1%     4038     1% /u01/12c/data1
/dev/gfsdata2    640.00    637.44    1%     4038     1% /u01/12c/data2
/dev/gfsredo1    300.00    298.51    1%     4038     2% /u01/12c/redo1
/dev/gfsredo2    300.00    298.51    1%     4038     2% /u01/12c/redo2
```

Initial Configuration Considerations and Tuning Parameters:

IBM Spectrum Scale Performance Tuning

Oracle Databases open and access IBM Spectrum Scale files in the correct manner by default. Do not use any special mount options (eg. DIO) for the file systems. The Oracle parameter "filesystemio_options" should remain at the default value of SETALL.

When configuring Network Shared Disk (NSD) devices, there should be only one NSD for each storage LUN -- one-to-one relationship. Also, the storage LUNs should be provisioned from different storage arrays of the same RAID type (i.e. RAID-5 or RAID-10 etc.). When the file systems are created multiple NSDs will be used producing the desirable effect of spreading I/O across the various controllers and cache regions in storage subsystem. This method achieves the general objective of the commonly-used Stripe and Mirror Everything (SAME) strategy.

IBM Spectrum Scale provides the option to set block size for each file system individually and the Oracle-specific recommendations are:

- 512KB is generally suggested.
- 256KB is suggested if the file system is mixed with non-Oracle database files that may be smaller in size
- 1MB is suggested for file systems that are 100TB or larger.

The mount options to suppress atime (-S) and mtime (-E) on data file systems may be helpful in reducing overhead for the file system management and increasing performance. One should understand whether any operating system utilities like backup software uses file modification time and therefore should not be suppressed.

To suppress atime and mtime (either or both), the parameters are set as follows:

mmchfs <device> -E no Disable exact mtime tracking
mmchfs <device> -S yes Suppress atime tracking

For IBM Spectrum Scale versions earlier than 4.2.3, I/O thread tuning parameters are recommended to be initially set as follows:

prefetchThreads=150
worker1Threads=450

For IBM Spectrum Scale versions 4.2.3, 5.0 and later, the I/O thread tuning is controlled by a single parameter, workerThreads. The recommended initial value should be set as follows:

workerThreads=512

IBM Spectrum Scale Availability and Resilience

Availability of the IBM Spectrum Scale cluster is paramount for production or mission-critical databases. As such, the parameter 'minQuorumNodes' may be set to decrease the possibility of losing cluster quorum and incurring unplanned downtime. Quorum loss or loss of connectivity occurs if a node goes down or becomes isolated from its peers by a network failure. Quorum is typically defined as one + half of the explicitly defined quorum nodes in the IBM Spectrum Scale cluster. In small clusters it may be desirable to have the IBM Spectrum Scale cluster remain online with only one surviving node; in that

case, tiebreaker disks must be used. The following parameter values are an example of this configuration option (the names of the tiebreaker disks will be different of course):

```
minQuorumNodes=1
tiebreakerDisks="gfsorahome;gfsdata2;gfsredo1"
```

Although previously stated in this note, it's necessary to reiterate that availability of the IBM Spectrum Scale cluster network is an important consideration. The cluster network should be protected using link aggregation configurations such as IEEE 802.3ad or Etherchannel.

IBM Spectrum Scale makes use of storage subsystems that employ SCSI-3 persistent reservations to control multi-node access to the shared storage. Failover times can be significantly reduced when this parameter is enabled in the file system cluster. IBM tests and certifies vendor storage subsystems for use of this feature. Please see the IBM Spectrum Scale FAQ for supported storage subsystems and further considerations for implementation.

To enable this feature, the following parameters should be set:

```
usePersistentReserve=yes
failureDetectionTime=10
```