Hidden Secrets: I/O-Durchsatz-messung mit Datenbank-Werkzeugen

Frank Schneede, ORACLE Deutschland B.V. & Co. KG

Die Implementierung eines I/O-Subsystems mit einem hohen Durchsatz ist integraler Bestandteil der Infrastruktur für eine zeitgemäße Applikation, sowohl im Data-Warehouse- als auch im OLTP-Umfeld. Defizite im Design des Gesamtsystems wie zum Beispiel eine zu geringe Anzahl eingesetzter Festplatten oder eine nicht ausreichende Netzwerk-Bandbreite machen sich dann im laufenden Betrieb in Form schlechter Antwortzeiten unangenehm bemerkbar.

Um den Kunden zur Vermeidung von Performance-Engpässen eine Hilfe zu geben, hat Oracle einen Architektur-Blueprint entwickelt, der eine auf allen Ebenen ausgewogene System-Landschaft beschreibt. Nach den Prinzipien dieser sogenannten „Well-balanced Architecture“ ist unter anderem die Exadata Database Machine aufgebaut. Den ausgewogenen Idealzustand findet der DBA jedoch in historisch gewachsenen Umgebungen häufig nicht vor; vielmehr hat er für Anwenderklagen über mangelhafte Antwortzeiten Ursachen und Lösungen zu finden.

Der Prozess der Problemlösung beginnt üblicherweise mit der Wait-Event-Analyse eines AWR-Reports. Hierbei deutet sich durch erhöhte Werte für I/O-bezogene Wait-Events (Wait Class „User I/O“ oder „System I/O“) ein Engpass im I/O-Durchsatz an. Um diesen Ansatz weiterzuverfolgen, ist es notwendig, die maximale Rate der I/O-Operationen der Datenbank zu messen, die diese zuverlässig bereitstellen kann. Dieser Vor-

```
[oracle@sccloud034 bin]$ more mytest.lun
/opt/oracle/oradata/CONT_FS/datafile/o1_mf_sysaux_93jpxn9t_.dbf
/opt/oracle/oradata/CONT_FS/datafile/o1_mf_system_93jq0yl9_.dbf
/opt/oracle/oradata/CONT_FS/datafile/o1_mf_undotbs1_93jq6wd7_.dbf
/opt/oracle/oradata/CONT_FS/datafile/o1_mf_users_93jq6rmc_.dbf
/opt/oracle/oradata/CONT_FS/datafile/o1_mf_sysaux_93jqbkf2_.dbf
/opt/oracle/oradata/CONT_FS/datafile/o1_mf_system_93jqbkl7_.dbf

[oracle@sccloud034 bin]$ ./orion -run simple -testname mytest -num_disks 6 -hugenotneeded
ORION: ORacle IO Numbers -- Version 12.1.0.1.0
mytest_20131002_1222
Calibration will take approximately 44 minutes.
.....
[oracle@sccloud034 bin]$
```

Listing 1: Einfache I/O-Durchsatz-Messung mit ORION

```
[oracle@sccloud034 bin]$ more mytest_20131002_1222_summary.txt
ORION VERSION 12.1.0.1.0

Command line:
-run simple -testname mytest -num_disks 6 -hugenotneeded

These options enable these settings:
Test: mytest
Small IO size: 8 KB
Large IO size: 1024 KB
IO types: small random IOs, large random IOs
Sequential stream pattern: one LUN per stream
Writes: 0%
Cache size: not specified
Duration for each data point: 60 seconds
Small Columns:,      0
Large Columns:,      0,      1,      2,      3,      4,      5,      6,      7,      8,
9,     10,     11,     12
Total Data Points: 43

Name: /opt/oracle/oradata/CONT_FS/datafile/o1_mf_sysaux_93jpxn9t_.dbf   Size: 1184899072
Name: /opt/oracle/oradata/CONT_FS/datafile/o1_mf_system_93jq0yl9_.dbf   Size: 828383232
Name: /opt/oracle/oradata/CONT_FS/datafile/o1_mf_undotbs1_93jq6wd7_.dbf Size: 94380032
Name: /opt/oracle/oradata/CONT_FS/datafile/o1_mf_users_93jq6rmc_.dbf    Size: 5251072
Name: /opt/oracle/oradata/CONT_FS/datafile/o1_mf_sysaux_93jqbkf2_.dbf   Size: 639639552
Name: /opt/oracle/oradata/CONT_FS/datafile/o1_mf_system_93jqbkl7_.dbf   Size: 262152192
6 files found.

Maximum Large MBPS=97.70 @ Small=0 and Large=3

Maximum Small IOPS=7100 @ Small=16 and Large=0
Small Read Latency: avg=2249 us, min=342 us, max=85196 us, std dev=1336 us @ Small=16 and Lar-
ge=0

Minimum Small Latency=1182.64 usecs @ Small=7 and Large=0
Small Read Latency: avg=1183 us, min=291 us, max=181059 us, std dev=2216 us @ Small=7 and Lar-
ge=0
......
[oracle@sccloud034 bin]$
```

Listing 2: Zusammenfassung ORION-I/O-Durchsatzmessung

gang heißt „calibration“. Dessen Ziel hängt natürlich auch vom Lastprofil ab, mit der die Datenbank betrieben wird:

- OLTP-Last mit Fokus auf IOPS und Latenz
- DWH-Last mit Fokus auf I/O-Durchsatz

Es gibt seit der Datenbank 11*g*R2 zwei verschiedene Möglichkeiten der „calibration“, die voll unterstützt sind und auch in der aktuellen Version 12c nutzbar sind. Mit vorgestellten Verfahren können I/O-Engpässe in bestehenden Umgebungen aufgezeigt oder auch die I/O-Spezifikation einer neuen Systemlandschaft verifiziert werden. Am Ende steht ein Vergleich beider Ansätze:

- „calibration“ mithilfe eines unabhängigen Utilitys (ORION)
- „calibration“ mithilfe der Oracle-Datenbank

I/O-Durchsatzmessung mit ORION

Bis zur Version 10*g* war der DBA aus schließlich auf die Nutzung eines unabhängigen Utilitys angewiesen, das für die eingesetzte Plattform und die Datenbank-Version vorliegen und installiert werden muss. Seit 11*g*R2 gehört zu diesem Zweck das Tool ORION zum Standardumfang der Datenbank und wird somit auch voll unterstützt. Es erzeugt unabhängig von der Datenbank eine synthetische Last auf dem Speichersystem. Diese Last entspricht von der Charakteristik, also der Verteilung und Art der I/O-Operationen sowie den genutzten Betriebssystem-Aufrufen, einer Last, die eine laufende Datenbank erzeugt.

```
SQL> COLUMN name         FORMAT a40
SQL> COLUMN value        FORMAT a15
SQL> COLUMN ts_name      FORMAT a10
SQL> COLUMN container    FORMAT a10
SQL> COLUMN asynch       FORMAT a10
SQL> SELECT name
  2  ,      value
  3  FROM   v$parameter
  4  WHERE name IN (,timed_statistics‘
  5                ,‘filesystemio_options‘
  6                ,‘disk_asynch_io‘);

NAME                                 VALUE
------------------------------------ ---------------
timed_statistics                     TRUE
filesystemio_options                 SETALL
disk_asynch_io                       TRUE

SQL> SELECT c.con_id     CON_ID
  2  ,      c.name       CONTAINER
  3  ,      t.name       TS_NAME
  4  ,      d.name       NAME
  5  ,      i.asynch_io  ASYNCH
  6  FROM  v$containers  c
  7  ,     v$datafile    d
  8  ,     v$tablespace  t
  9  ,     v$iostat_file i
 10  WHERE TO_NUMBER(c.con_id) = TO_NUMBER(d.con_id)
 11  AND   TO_NUMBER(c.con_id) = TO_NUMBER(t.con_id)
 12  AND   d.ts#               = t.ts#
 13  AND   d.file#             = i.file_no
 14  AND   i.filetype_name     = ,Data File‘
 15  ORDER BY c.con_id
 16  ,        t.name;

CON_ID CONTAINER  TS_NAME  NAME                                 ASYNCH
------ ---------- -------- ------------------------------------ --------
     1 CDB$ROOT   SYSAUX   /opt/.../o1_mf_sysaux_93jpxn9t_.dbf   ASYNC_ON
     1 CDB$ROOT   SYSTEM   /opt/.../o1_mf_system_93jq0yl9_.dbf   ASYNC_ON
     1 CDB$ROOT   UNDOTBS1 /opt/.../o1_mf_undotbs1_93jq6wd7_.dbf ASYNC_ON
     1 CDB$ROOT   USERS    /opt/.../o1_mf_users_93jq6rmc_.dbf    ASYNC_ON
     2 PDB$SEED   SYSAUX   /opt/.../o1_mf_sysaux_93jqbkf2_.dbf   ASYNC_ON
     2 PDB$SEED   SYSTEM   /opt/.../o1_mf_system_93jqbkl7_.dbf   ASYNC_ON
......
     5 SAMPLEPDB2 EXAMPLE  /opt/.../o1_mf_example_93oyhw77_.dbf  ASYNC_ON
     5 SAMPLEPDB2 SYSAUX   /opt/.../o1_mf_sysaux_93oyhw0w_.dbf   ASYNC_ON
     5 SAMPLEPDB2 SYSTEM   /opt/.../o1_mf_system_93oyhwbj_.dbf   ASYNC_ON
     5 SAMPLEPDB2 USERS    /opt/.../o1_mf_users_93oyhwg5_.dbf    ASYNC_ON

17 rows selected.

SQL>
```

Listing 3: Prüfung der Voraussetzungen für „dbms_resource_manager.calibrate“

```
SQL> SET SERVEROUTPUT ON
SQL> DECLARE
  2   lat  INTEGER;
  3   iops INTEGER;
  4   mbps INTEGER;
  5  BEGIN
  6   --DBMS_RESOURCE_MANAGER.CALIBRATE_IO(, , iops, mbps, lat);
  7   DBMS_RESOURCE_MANAGER.CALIBRATE_IO (6, 10, iops, mbps, lat);
  8   DBMS_OUTPUT.PUT_LINE (,max_iops = , || iops);
  9   DBMS_OUTPUT.PUT_LINE (,latency  = , || lat);
 10   DBMS_OUTPUT.PUT_LINE (,max_mbps = , || mbps);
 11  END;
 12  /
max_iops = 6699
latency  = 10
max_mbps = 98

PL/SQL procedure successfully completed.

SQL>
```

Listing 4: I/O-Durchsatz-Messung mit „dbms_resource_manager.calibrate"

Nach Durchführung der Messung hat der DBA einen Richtwert für den Durchsatz der Hardware. Die Mess-Ergebnisse können benutzt werden, um Fehlerquellen abseits der Datenbank auszuschließen oder eine neue Hardware auf deren Eignung hin zu überprüfen. ORION lässt sich für zahlreiche Einsatzmöglichkeiten parametrisieren, an dieser Stelle soll nur ein kurzes Beispiel gezeigt werden.

Die zur Verfügung stehenden LUNs müssen in einem File angegeben werden, das den Namen des durchzuführenden Tests tragen muss. Anschließend wird in dem Beispiel in Listing 1 ein einfacher Test mit Default-Parametern aufgerufen, wobei der Parameter „–hugenotneeded" anzeigt, dass auf der Testumgebung keine „huge pages" zur Verfügung stehen. Der Test fand auf einer virtualisierten Oracle-Linux-Plattform mit sehr schmalen Ressourcen statt. Das Ergebnis ist dann in einer Datei zusammengefasst, von der Listing 2 einen Ausschnitt zeigt. Zusätzlich werden noch weitere Ergebnisdateien (in diesem Beispiel „mytest_<date_time>_mbps.csv", „mytest_<date_time>_iops.csv" und „mytest_<date_time>_lat.csv") erzeugt, die im „csv"-Format vorliegen und nach erfolgter Bearbeitung in MS Excel die Messungen visualisieren.

I/O-Durchsatzmessung mit dem Resource Manager

Ab Oracle Database 11.1 lässt sich die I/O-Durchsatzmessung auf Ebene des Root Containers durchführen. Das Werkzeug zur „calibration" des I/O-Systems steht als Erweiterung des bewährten Database Resource Manager bereit. Das API „DBMS_RESOURCE_MANAGER.CALIBRATE_IO()" erzeugt eine gemischte „Read-only"-Last, bestehend aus folgenden Aktionen:

- Zufällige I/Os in der Größe der parametrisierten „db_block_size"
- Sequenzielle I/Os mit 1 MByte Blockgröße

Da die Prozedur „CALIBRATE_IO()" den Oracle Call-Stack nutzt und die I/O-Operationen gegen Blöcke laufen, die in der Datenbank gespeichert sind, kann man die erzielten Mess-Ergebnisse als sehr realistisch für die tatsächlich erreichbare Performance ansehen. Da „calibration" möglicherweise eine starke Belastung der Datenbank darstellt, ist es sinnvoll, diesen Vorgang zu Zeiten durchzuführen, an denen die Datenbank sehr gering ausgelastet ist. Zu beachten ist auch, dass Datenbanken, die auf den gleichen Speicher zugreifen und zum Zeitpunkt der „calibration" aktiv sind, das Mess-Ergebnis verfälschen können. Um „calibration" mit dem Database Resource Manager nutzen zu können, müssen verschiedene Voraussetzungen eingehalten werden:

- Eingesetzte Datenbank-Version ab 11.1
- Mit Datenbank-Version 12*c* auf „cdb$root"-Container
- Der aufrufende Benutzer hat „SYSDBA"-Privileg
- Asynchrones I/O ist aktiviert („DISC_ASYNCH_IO=TRUE" und „FILESYSTEMIO_OPTIONS=SETALL")
- Zum Zeitpunkt der „calibration" geringe Last auf der Datenbank

Der Initialisierungsparameter „DISC_ASYNCH_IO" ist bereits standardmäßig auf den Wert „TRUE" gesetzt, während der Standardwert für den Parameter „FILESYSTEMIO_OPTIONS" abhängig vom eingesetzten Betriebssystem automatisch auf den optimalen

```
SQL> SELECT * FROM v$io_calibration_status;

STATUS        CALIBRATION_TIME                  CON_ID
------------- --------------------------------- ------
READY         01-OCT-13 04.20.35.397 PM              0

SQL> SELECT start_time
  2  ,       end_time
  3  ,       max_iops           IOPS
  4  ,       max_mbps           MBPS
  5  ,       latency            LAT
  6  ,       num_physical_disks DISKS
  7  FROM dba_rsrc_io_calibrate;

START_TIME                END_TIME                    IOPS MBPS
LAT DISKS
------------------------- ------------------------- ---- ----
--- -----
01-OCT-13 04.10.04.485273 01-OCT-13 04.20.35.397176 6699   98
10     6


SQL>
```

Listing 5: Ergebnis I/O-Durchsatz-Messung im Data Dictionary

	dbms_resource_manager	ORION
PRO	Unterstützung von RAC Verfügbar mit der Datenbank	Stand-alone-Betrieb möglich Grafische Aufbereitungsmöglichkeit
CONTRA	Benötigt laufende Datenbank	Calibration eines RAC nur manuell

Tabelle 1

Wert für diese Plattform gesetzt wird. Im Zweifelsfall ist an dieser Stelle die plattformspezifische Dokumentation zu Rate zu ziehen. „FILESYSTEMIO_OPTIONS“ kann folgende Werte annehmen:

- *asynch*
  Asynchronous I/O wird verwendet, sofern das Betriebssystem das unterstützt.
- *directIO*
  Direct I/O wird verwendet, sofern das Betriebssystem das unterstützt. Direct I/O umgeht den Unix-Buffer-Cache.
- *setall*
  „Asynchronous“ und „Direct I/O“ wird aktiviert.
- *none*
  Deaktiviert „Asynchronous“ und „Direct I/O“. Oracle benutzt normale „synchronous writes” ohne „Direct I/O“-Options.

Die Voraussetzungen lassen sich durch Abfrage des Data Dictionary verifizieren (siehe Listing 3). Anschließend wird „calibration“ durch Aufrufen des API in einem kleinen PL/SQL-Block gestartet. Das API benötigt dazu zwei Parameter als Eingabe:

- *NUM_DISKS*
  Die Anzahl der Platten, die der Datenbank zur Verfügung stehen. Hier ist zu beachten, dass bei ASM-Nutzung lediglich die physikalischen Platten anzugeben sind, die für Daten genutzt werden. Platten für eine „Flash Recovery Area“ sind also auszusparen.
- *MAX_LATENCY*
  Die maximale Latenz in Millisekunden für einen Plattenzugriff.

Mit diesen Informationen aufgerufen, bekommt man nach kurzer Zeit das Ergebnis der „calibration“ angezeigt (siehe Listing 4).

Der Status der Messung ist in der View „V$IO_CALIBRATION_STATUS“ festgehalten, was dem DBA auch während einer laufenden Messung einen Überblick verschafft. Nach Ende der „calibration“ ist das Ergebnis in der Data-Dictionary-View „DBA_RSRC_IO_CALIBRATE“ festgehalten und kann jederzeit abgefragt werden (siehe Listing 5).

Fazit

Mit den beiden vorgestellten Methoden besitzt der DBA ein Portfolio, aus dem er das geeignete Utility auswählen kann. Tabelle 1 gibt eine Auswahlhilfe zur Abgrenzung beider „calibration“-Tools.

Die Erweiterung des Database Resource Manager stellt ein sehr wertvolles Tool dar, um die Einschränkungen der vorliegenden I/O-Architektur zu verstehen. Nach Beendigung einer „calibration“ verfügt der DBA über die Informationen, die notwendig sind, um die Größe und das Design des I/O-Systems anforderungsgerecht zu gestalten.

Weiterführende Informationen

- http://www.oracle.com/webfolder/technetwork/de/community/dbadmin/tipps/io_calibration/index.html
- http://www.oracle.com/webfolder/technetwork/de/community/dbadmin/tipps/orion/index.html
- https://support.oracle.com/CSP/main/article?cmd=show&type=NOT&id=1297112.1

Frank Schneede
frank.schneede@oracle.com