Configuration and Tuning the OLTPbenchmark workload on the IBM Power S824

Updated: 4/22/2016

Benchmark Configuration

The System Under Test has the following configuration:

1 x IBM Power S824 with
  4 x 3.52 Ghz POWER 8 processors (6 cores / chip – 24 cores total)
  512 GB memory
  4 x 300 GB Internal HDD
  1 x Dual port 16Gb Fiber HBA
  1 x 10Gb Network controller
  AIX 7.1 TL3 SP3 (7100-03-03-1415)
  Oracle Database 12c Release 12.1.0.2 Enterprise Edition
  Oracle Grid Infrastructure 12c Release 12.1.0.2

1x Oracle Server X5-2L (for Redo logs) with
  2 x 2.4 GHz Intel Xeon E5-2630 v3
  128 GB memory
  14 x 600 GB SAS2 disks
  1 x Dual port 16Gb Fiber HBA
  1x 12Gb SAS PCIe RAID
  Solaris 11.2 SRU 9

1x Oracle Server X5-2L (for Database files) with
  2 x 2.4 GHz Intel Xeon E5-2630 v3
  128 GB memory
  4 x 600 GB SAS2 disks
  4 x 1.6TB NVMe SSD
  1 x NVMe PCIe 3.0 Switch Ctrl
  1 x Dual port 16Gb Fiber HBA
  Solaris 11.2 SRU 9

1x Brocade 6510 Fiber Channel Switch
**Oracle Database Installation and Configuration**

Oracle Database 12c Release 1 (12.1.0.2) and Oracle Grid Infrastructure 12c (12.1.0.2) for the IBM Power S824 was installed from the following location:


The Installation documentation is located here:


The Database Release Notes for IBM AIX on Power Systems is located here:

[https://docs.oracle.com/database/121/AXDBN/toc.htm](https://docs.oracle.com/database/121/AXDBN/toc.htm)

Pre-installation of Oracle Grid Infrastructure requires enable I/O completion ports, which is documented here:

[https://docs.oracle.com/database/121/CWAIX/preaix.htm#CWAIX168](https://docs.oracle.com/database/121/CWAIX/preaix.htm#CWAIX168)

**System Tuning**

The following system tuning options were used:

2. Use 16 MB pages for the database SGA:
   
   ```
   $ vmo -o lpg_regions=5120 -o lgpgsize=16M -o v_pinshm=1
   ```
3. Enable Simultaneous Multithreading Mode
   
   ```
   $ smtctl -t 8 -w now
   ```
4. Change oracle user capabilities
   
   ```
   $ chuser capabilities=CAP_BYPASS_RAC_VMM,CAP_PROPAGATE,CAP_NUMA_ATTACH oracle
   ```
5. Increase maxuproc
   
   ```
   $ chdev -l sys0 -a maxuproc=1024
   ```
6. Set max fsize = -1 in /etc/security/limits
7. Other tunings sourced from published benchmarks that were added:
   
   ```
   $ dscrctl -n -s 2
   $ ctctrl -P memtraceoff
   $ errctrl -P errcheckoff
   ```
Storage Configuration

The System Under test used two Oracle Server X5-2L systems for the database files and redo logs.

The Oracle Server X5-2L used for the database files contained 4x 1.6TB NMVe SSDs exported as LUNS to the database system using Oracle Solaris Volume Manager and the Oracle Solaris COMSTAR technology. Each server was installed with Oracle Solaris 11.2 SRU 9 with the following options added to the /etc/system file:

```bash
set tune_t_fsflushr = 5
set autoup = 300
set idle_cpu_no_deep_c=1
exclude: ehci
exclude: uhci
```

Installing and configuration of the Oracle SVM and COMSTAR software can be found here:


The four SSDs were initially configured as a single 5.8 TB SVM volume and then into eight separate soft partitions of 700 GB each:

```bash
# metastat -p
d0 1 4 /dev/rdsk/c3t1d0s0 /dev/rdsk/c4t1d0s0 /dev/rdsk/c5t1d0s0 /dev/rdsk/c6t1d0s0 -i 128b
d3008 -p /dev/md/rdsk/d0 -o 10276045824 -b 1468006400
d3007 -p /dev/md/rdsk/d0 -o 8808039296 -b 1468006400
d3006 -p /dev/md/rdsk/d0 -o 7340032768 -b 1468006400
d3005 -p /dev/md/rdsk/d0 -o 5872026240 -b 1468006400
d3004 -p /dev/md/rdsk/d0 -o 4404019712 -b 1468006400
d3003 -p /dev/md/rdsk/d0 -o 2936013184 -b 1468006400
d3002 -p /dev/md/rdsk/d0 -o 1468006656 -b 1468006400
d3001 -p /dev/md/rdsk/d0 -o 128 -b 1468006400
```

Each volume then exported as a separate COMSTAR LUN using the `stmfadm` command:

```bash
# stmfadm create-lu -p guid=666677778888899990520070FE0D3001 /dev/md/rdsk/d3001
# stmfadm create-lu -p guid=666677778888899990520070FE0D3002 /dev/md/rdsk/d3002
# stmfadm create-lu -p guid=666677778888899990520070FE0D3003 /dev/md/rdsk/d3003
# stmfadm create-lu -p guid=666677778888899990520070FE0D3004 /dev/md/rdsk/d3004
# stmfadm create-lu -p guid=666677778888899990520070FE0D3005 /dev/md/rdsk/d3005
# stmfadm create-lu -p guid=666677778888899990520070FE0D3006 /dev/md/rdsk/d3006
# stmfadm create-lu -p guid=666677778888899990520070FE0D3007 /dev/md/rdsk/d3007
# stmfadm create-lu -p guid=666677778888899990520070FE0D3008 /dev/md/rdsk/d3008
```

The Oracle Server X5-2L used for the redo logs contained fourteen 600 GB SAS disks where two disks were used to install the Solaris Operating system. Twelve 600 GB disks were configured as a single 3 TB raid 1 Virtual Drive using the RAID management interface in the Oracle Server X5-2L firmware.
The virtual drive was configured with a 256K segment size and Current Write Cache Policy set to Forced Write Back.

The Solaris Volume Manager utility was used to configure the single 3TB LUN into 3 smaller 1 TB soft partitions:

```
# metastat -p
d0 1 1 /dev/rdsk/c0t0d0s0
d3000 -p /dev/md/rdsk/d0 -o 4294967680 -b 2147483648
d2000 -p /dev/md/rdsk/d0 -o 2147483904 -b 2147483648
d1000 -p /dev/md/rdsk/d0 -o 128 -b 2147483648
```

Each soft partition was then exported as a separate COMSTAR LUN using the `stmfadm` command:

```
# stmfadm create-lu -p guid=666677778888999900520050AD0D1000 /dev/md/rdsk/d1000
# stmfadm create-lu -p guid=666677778888999900520050AD0D2000 /dev/md/rdsk/d2000
# stmfadm create-lu -p guid=666677778888999900520050AD0D3000 /dev/md/rdsk/d3000
```

The two Oracle Server X5-2L storage servers each have two fiber channel cables connected to four ports on a Brocade 6510 Fiber switch. There are two fiber cables from the IBM Power S824 also connected to the fiber switch.

### Database Configuration

On the IBM Power S824, the LUNS exported by the two Oracle Server X5-2L systems are visible using the Oracle 12c Grid Infrastructure software. The Oracle command `asmca` was used to configure the DATA disk group using the eight 700 GB LUNS for a total of 5.6 TB and create a single LOG disk group using 2 of the 1TB LUNS to stripe redo across both LUNS. The third 1TB LUN was not used during testing.

The STRIPE property for the LOG disk group was set to `FINE` before creating the database.

```
# sqlplus / as sysasm <<EOT
ALTER DISKGROUP LOG MODIFY TEMPLATE onlinelog ATTRIBUTES (FINE);
EOT
```

Database creation script.

```
#!/bin/sh
DB=tpcc
DB_DIR="+DATA"
SYS=${DB_DIR}/sys
SYS_AUX=${DB_DIR}/sysaux
LOG1="+LOG/log1"
LOG2="+LOG/log2"
LOGSZ="400G"
TEMP=${DB_DIR}/temp
UNDO=${DB_DIR}/undo
sqlplus <<EOF
CONNECT / as sysdba
shutdown;
```
after creating the database, create the user \textit{tpcc} and the tablespaces for the workload. The tablespace for the orderline table was configured to use 16K blocksize.

\begin{verbatim}
# sqlplus "/as sysdba" << EOT
CONNECT / as sysdba
drop user tpcc;
create user tpcc identified by tpcc;
grant connect, resource to tpcc;
grant all privileges to tpcc identified by tpcc;
alter user tpcc temporary tablespace temp_ts;
exit
EOT

# sqlplus "/as sysdba" << EOT
create bigfile tablespace tpcctab
        datafile '$DB_DIR/tpcctab' size 200G reuse;
create bigfile tablespace tpcc_ol
        datafile '$DB_DIR/tpcc_ol' size 400G reuse blocksize 16k;
alter user tpcc default tablespace tpcctab;
EOT
\end{verbatim}
**Database Listener**

To take advantage of memory affinity, the IBM Power S824 was configured with four Resource Sets based on the processor numbers as follows:

```bash
# mkrset -c 0-23 96-119 test/orapset1
# mkrset -c 24-47 120-143 test/orapset2
# mkrset -c 48-71 144-167 test/orapset3
# mkrset -c 72-95 168-191 test/orapset4
```

The database was then configured with four listener processes, each listening on a the same network IP, but using a separate port number:

```bash
# cat ${ORACLE_HOME}/network/admin/listener.ora

LISTENER_TPCC_1=(DESCRIPTION=(ADDRESS_LIST=(ADDRESS=(PROTOCOL=tcp)(HOST=192.168.222.1)(PORT=1522))))
LISTENER_TPCC_2=(DESCRIPTION=(ADDRESS_LIST=(ADDRESS=(PROTOCOL=tcp)(HOST=192.168.222.1)(PORT=1523))))
LISTENER_TPCC_3=(DESCRIPTION=(ADDRESS_LIST=(ADDRESS=(PROTOCOL=tcp)(HOST=192.168.222.1)(PORT=1524))))
LISTENER_TPCC_4=(DESCRIPTION=(ADDRESS_LIST=(ADDRESS=(PROTOCOL=tcp)(HOST=192.168.222.1)(PORT=1525))))
```

After each listener process was started, it was bound to one of the four resource sets created earlier.

```bash
# export MEMORY_AFFINITY=MCM
# lsnrctl start LISTENER_TPCC_1
# lsnrctl start LISTENER_TPCC_2
# lsnrctl start LISTENER_TPCC_3
# lsnrctl start LISTENER_TPCC_4

# LSNR_PID=`ps -ef | grep LISTENER_TPCC_1 | grep -v grep | awk '{print $2}'`
# sudo attachrset -F test/orapset1 $LSNR_PID
# LSNR_PID=`ps -ef | grep LISTENER_TPCC_2 | grep -v grep | awk '{print $2}'`
# sudo attachrset -F test/orapset2 $LSNR_PID
# LSNR_PID=`ps -ef | grep LISTENER_TPCC_3 | grep -v grep | awk '{print $2}'`
# sudo attachrset -F test/orapset3 $LSNR_PID
# LSNR_PID=`ps -ef | grep LISTENER_TPCC_4 | grep -v grep | awk '{print $2}'`
# sudo attachrset -F test/orapset4 $LSNR_PID
```

**Network Tuning**

```bash
/usr/sbin/no -ro ipqmaxlen=512
/usr/sbin/no -o sb_max=4194304
/usr/sbin/no -o rfc1323=1
/usr/sbin/no -o tcp_recvspace=65536
/usr/sbin/no -o tcp_sendspace=65536
/usr/sbin/no -o udp_recvspace=655360
/usr/sbin/no -o udp_sendspace=65536
```
The Oracle parameters file in `${ORACLE_HOME}/dbs/inittpcc.ora`, contained the following values:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>aq_tm_processes</td>
<td>0</td>
</tr>
<tr>
<td>compatible</td>
<td>12.1.0.1.0.0</td>
</tr>
<tr>
<td>control_files</td>
<td>+DATA/cntrlspec</td>
</tr>
<tr>
<td>cpu_count</td>
<td>48</td>
</tr>
<tr>
<td>db_16k_cache_size</td>
<td>20G</td>
</tr>
<tr>
<td>db_block_checking</td>
<td>FALSE</td>
</tr>
<tr>
<td>db_block_checksum</td>
<td>FALSE</td>
</tr>
<tr>
<td>db_block_size</td>
<td>8192</td>
</tr>
<tr>
<td>db_cache_advice</td>
<td>off</td>
</tr>
<tr>
<td>db_cache_size</td>
<td>40G</td>
</tr>
<tr>
<td>db_file_multiblock_read_count</td>
<td>128</td>
</tr>
<tr>
<td>db_files</td>
<td>256</td>
</tr>
<tr>
<td>db_name</td>
<td>tpcc</td>
</tr>
<tr>
<td>db_writer_processes</td>
<td>16</td>
</tr>
<tr>
<td>ddl_lock_timeout</td>
<td>30</td>
</tr>
<tr>
<td>deferred_segment_creation</td>
<td>FALSE</td>
</tr>
<tr>
<td>disk_asyncio_io</td>
<td>TRUE</td>
</tr>
<tr>
<td>dml_locks</td>
<td>1000</td>
</tr>
<tr>
<td>filesystemio_options</td>
<td>setall</td>
</tr>
<tr>
<td>lock_sga</td>
<td>TRUE</td>
</tr>
<tr>
<td>log_buffer</td>
<td>100663296</td>
</tr>
<tr>
<td>log_checkpoint_interval</td>
<td>0</td>
</tr>
<tr>
<td>log_checkpoints_to_alert</td>
<td>TRUE</td>
</tr>
<tr>
<td>open_cursors</td>
<td>2400</td>
</tr>
<tr>
<td>parallel_max_servers</td>
<td>100</td>
</tr>
<tr>
<td>processes</td>
<td>8000</td>
</tr>
<tr>
<td>query_rewrite_enabled</td>
<td>FALSE</td>
</tr>
<tr>
<td>replication_dependency_tracking</td>
<td>FALSE</td>
</tr>
<tr>
<td>sessions</td>
<td>12000</td>
</tr>
<tr>
<td>shared_pool_size</td>
<td>16G</td>
</tr>
<tr>
<td>statistics_level</td>
<td>BASIC</td>
</tr>
<tr>
<td>timed_statistics</td>
<td>FALSE</td>
</tr>
<tr>
<td>trace_enabled</td>
<td>FALSE</td>
</tr>
<tr>
<td>transactions</td>
<td>12000</td>
</tr>
<tr>
<td>transactions_per_rollback_segment</td>
<td>1</td>
</tr>
<tr>
<td>undo_management</td>
<td>AUTO</td>
</tr>
<tr>
<td>undo_retention</td>
<td>30</td>
</tr>
<tr>
<td>undo_tablespace</td>
<td>undo_ts</td>
</tr>
<tr>
<td>_fast_cursor_reexecute</td>
<td>TRUE</td>
</tr>
<tr>
<td>_fg_sync_sleep_usecs</td>
<td>500</td>
</tr>
</tbody>
</table>
**Driver Configuration**

The workload was driven over a 10Gb network link from an external Oracle Server X5-2 to the IBM Power S824 database server configured as follows:

Oracle Server X5-2 with

- 2 x 2.6 GHz Intel Xeon E5-2660 v3
- 128 GB memory
- 2 x 600 GB SAS2 disks
- 1 x Dual port 10 Gb Network controller
- Oracle Linux 6.5
- Oracle Database 12c Release (12.1.0.2)

The Oracle Database 12c Release 12.1.0.2 was installed on the driver system to provide the necessary Oracle shared libraries to drive the load.

The following tuning was added to the `/etc/sysctl.conf` file:

```
fs.file-max = 6815744
kernel.sem = 250 32000 100 128
kernel.shmmax = 4096
kernel.shmall = 1073741824
kernel.shmmni = 4398046511104
net.core.rmem_default = 262144
net.core.rmem_max = 4194304
net.core.wmem_default = 262144
net.core.wmem_max = 1048576
fs.aio-max-nr = 1048576
net.ipv4.ip_local_port_range = 9000 65500
```

To distribute the workload equally across the four listener processes on the IBM Power S824, load balancing was enabled on the driver system using the `tnsnames.ora` configuration file as follows:

```
# cat $[TNS_ADMIN]/tnsnames.ora
TPCC=(DESCRIPTION =(ADDRESS_LIST =(LOAD_BALANCE=on)
  (ADDRESS = (PROTOCOL = TCP)(HOST = 192.168.222.1)(PORT = 1522))
  (ADDRESS = (PROTOCOL = TCP)(HOST = 192.168.222.1)(PORT = 1523))
  (ADDRESS = (PROTOCOL = TCP)(HOST = 192.168.222.1)(PORT = 1524))
  (ADDRESS = (PROTOCOL = TCP)(HOST = 192.168.222.1)(PORT = 1525)))
(CONNECT_DATA =(SERVICE_NAME = tpcc))
```
**Workload Configuration**

The OLTPbenchmark kit version 2.16 was installed on the driver system to drive the load. Using the workload *Options* tab, configured the following TPC-C Schema options:

- Oracle Service Name: TPCC
- TPC-C User: tpcc
- TPC-C Password: tpcc
- TPC-C Default Tablespace: tpcctab
- Order Line Tablespace: tpcc_ol
- TPC_C Temporary Tablespace: temp
- Partition Order Line Table: enabled
- Number of Warehouses: 800
- TPC-C Driver Script: AWR Snapshot Driver Script
- Total Transactions per User: 2000000
- Rampup: 10 mins
- Test Duration: 10 mins

Using the workload *Virtual Users* tab, configured 340 users, 1000 ms User Delay, and log output to Temp with a Unique log file name.

Based on initial test runs showing contention wait events with the *cache chains buffers* latch, added partitioning to the NEW_ORDER table and 2 indexes for the ORDERS table using the following changes to the file hdb_tpcc.tcl – modifications are in red:

```sql
set sql(7) "CREATE TABLE NEW_ORDER (NO_W_ID NUMBER, NO_D_ID NUMBER, NO_O_ID NUMBER, CONSTRAINT INORD PRIMARY KEY (NO_W_ID, NO_D_ID, NO_O_ID) ENABLE ) ORGANIZATION INDEX NOCOMPRESS INITRANS 4 MAXTRANS 16 PCTFREE 10 PARTITION BY HASH(NO_W_ID) PARTITIONS 10"
set sql(6) "CREATE UNIQUE INDEX ORDERS_I1 ON ORDERS (O_W_ID, O_D_ID, O_ID) INITRANS 4 MAXTRANS 16 PCTFREE 10 GLOBAL PARTITION BY HASH(O_W_ID) PARTITIONS 10"
set sql(7) "CREATE UNIQUE INDEX ORDERS_I2 ON ORDERS (O_W_ID, O_D_ID, O_C_ID, O_ID) INITRANS 4 MAXTRANS 16 PCTFREE 10 GLOBAL PARTITION BY HASH(O_W_ID) PARTITIONS 10"
```

To reduce the latency of *library cache: mutex X* wait events, added the recommended workarounds from the following blog article:

Result Summary

The OLTPbenchmark peak score reported in the workload log file using the configuration described above was 3,609,832 Oracle TPM at 1,219,451 NOPM. The TPM number was calculated by the workload using the Transactions metric from Workload Repository snapshots collected at the start and end of the test period and are not official TPC results.

<table>
<thead>
<tr>
<th>Load Profile</th>
<th>Per Second</th>
<th>Per Transaction</th>
<th>Per Exec</th>
<th>Per Call</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB Time(s):</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>DB CPU(s):</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Background CPU(s):</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Redo size (bytes):</td>
<td>321,576,095.7</td>
<td>5,345.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logical read (blocks):</td>
<td>6,203,812.8</td>
<td>103.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Block changes:</td>
<td>1,884,611.5</td>
<td>31.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical read (blocks):</td>
<td>4,204.1</td>
<td>0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical write (blocks):</td>
<td>28,926.6</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read IO requests:</td>
<td>4,198.3</td>
<td>0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Write IO requests:</td>
<td>8,054.2</td>
<td>0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read IO (MB):</td>
<td>36.4</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Write IO (MB):</td>
<td>238.9</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IM scan rows:</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session Logical Read IM:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User calls:</td>
<td>46,151.2</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parses (SQL):</td>
<td>27,866.1</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hard parses (SQL):</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SQL Work Area (MB):</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logons:</td>
<td>0.1</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Executes (SQL):</td>
<td>1,243,039.6</td>
<td>20.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rollbacks:</td>
<td>101.5</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transactions:</td>
<td>60,163.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Oracle made reasonable efforts to ensure the accuracy and validity of its testing and suggest that knowledgeable database and software engineers run it themselves to validate these results.