RAC parameter tuning for remote access

Carlos Fernando Gamboa, Brookhaven National Lab, US
Frederick Luehring, Indiana University, US

Distributed Database Operations Workhop
CERN Geneva, November 2008

Outline

Introduction

Overview to key OS/Database network parameters

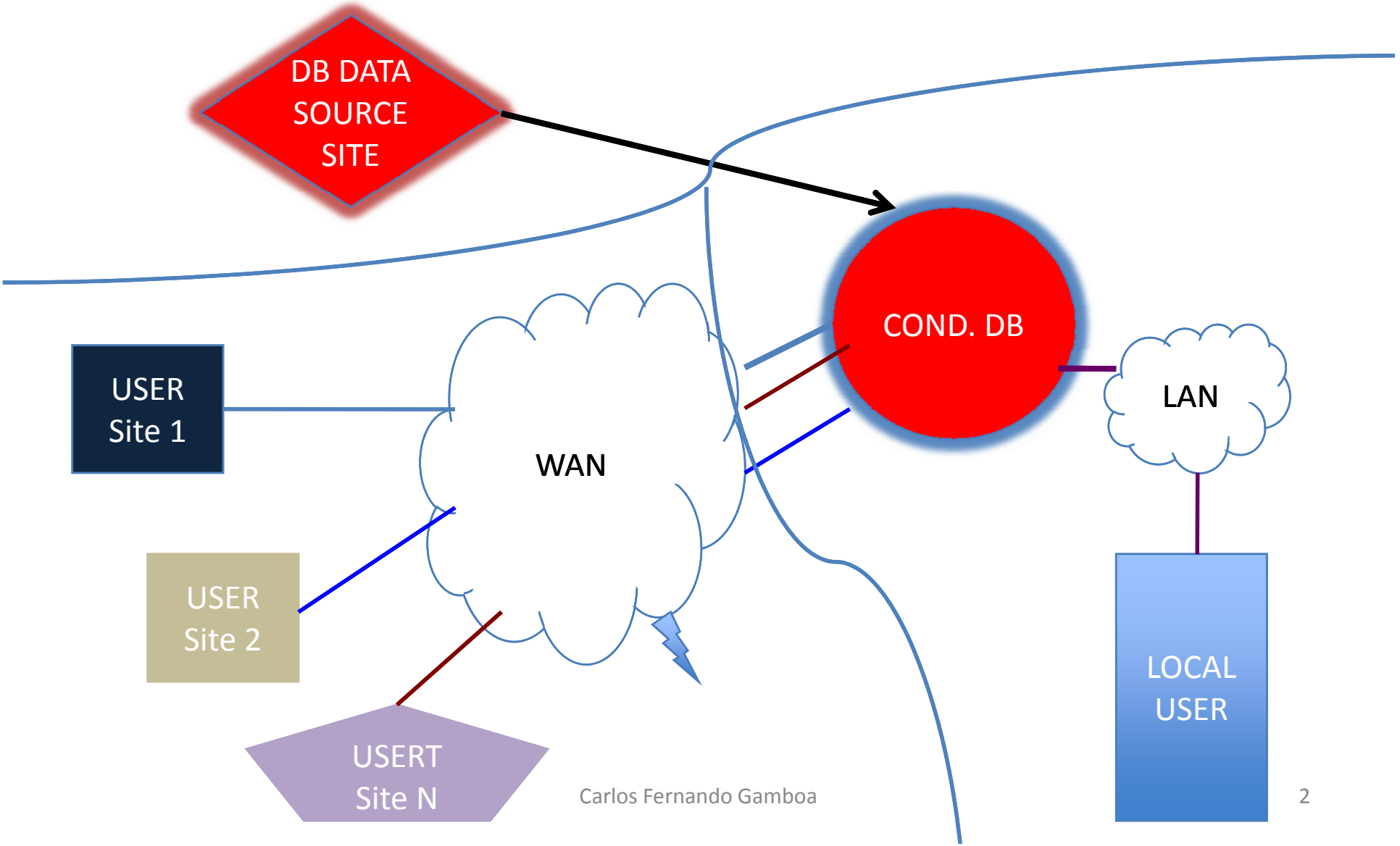
Case overview

Tests

Results

Conclusions

Introduction
-Database Service-



Introduction
- tuning strategy -

-Tuning SQLnet:

Main function of SQLnet:

- to read and write to the TCP send/receive buffers
- to distribute and retrieve data to and from an application such as SQL*Plus and the Oracle Database.

-Can be done at application level:

Setting the ARRAY Fetch Size in the Application

(Metalink doc id. Note:67983.1)

Overview network parameters
- Operative system -

Kernel parameters:

Over 521 different parameter around 365 are tcp parameters

Focus on TCP kernel parameters for a kernel level

2.6, specially on:

- **tcp_moderate_rcvbuf** : If setup as 1 activates autotuning. Thus the receiver buffer size and TCP window size is dynamically updated per connection basis.
- **Memory per connection allocation parameters:**
  - **tcp_rmem:** Memory allocated for TCP rcv buffers
  - **tcp_wmem**: Memory allocated for TCP snd buffers

Overview network parameters
- Operative system-

Memory per connection allocation parameters:

- **tcp_rmem:** Memory allocated for TCP rcv buffers.
  net.ipv4.tcp_rmem = minimum default maximum

- **tcp_wmem**: Memory allocated for TCP snd buffers
  net.ipv4.tcp_wmem = minimum default maximum

Used to set constrains to autotune and controls memory usage under memory stress.

Maximum buffer size for socket buffer declared via the SO_SNDBUF and SO_RECVBF that application can request can be constrained with:

net.core.rmem_max
net.core.wmem_max

Overview network parameters
- Operative System-

Table 1 . Summary default values RHEL 4 kernel level 2.6

Parameter	Min	default	Max
tcp_moderate_rcvbuf		1	
net.ipv4.tcp_rmem	4k	87380	87380*2
net.ipv4.tcp_wmem	4k	16K	128K
net.core.rmem_max		110592	131071
net.core.wmem_max		110592	131071

Overview network parameters
-Oracle database-

Key Network files on oracle

listener.ora (server): Contains information related to listening protocol addresses, about supported services, and parameters that control its Listener process runtime behavior.

sqlnet.ora (client, server): contains the parameters that specify preferences for how a client or server uses Oracle's Network protocol features.

tnsnames.ora (client,server): Maps net services names to connection descriptions. This parameters can be defined as well on this file;

Example configuration on 3D twiki

https://twiki.cern.ch/twiki/bin/view/PSSGroup/RaC

Network File	Parameter		
sqlnet.ora	DEFAULT_SDU_SIZE	RECV_BUF_SIZE	SEND_BUF_SIZE
tnsnames.ora Listener.ora	SDU_SIZE	RECV_BUF_SIZE	SEND_BUF_SIZE

Overview network parameters
- Oracle database -

Session Data Unit (SDU).

- Allows limited control over the packet sizes sent to the NT layer.
- Possible values: 512 to 32767 bytes, **2048 default.**
- Minimizing overhead adjust it to the Maximum Segment Size (MSS) of the network protocol being used. Thus,

MSS=Maximum Transmission Unit – (TCP and IP) header size

=1500 (Ethernet) -20 bytes (TCP) – 20 IP =1460 bytes

- Negotiated by client and server for data retrieval. Minimum value when client and server differ

Overview network parameters
- Oracle database-

- **RECV_BUF_SIZE SEND_BUF_SIZE**
  - Alters the TCP send and receive windows.
  - Not setting this parameters OS buffers sizes will be used

  - This parameters depends on:
    - Network latency among client and database
    - SDU size negotiated
    - OS TCP kernel parameters (previously mentioned)

_ SEND/RCV_BUFFER proportional to the Band Product Delay (BDP)
BDP is proportional to the network latency and the bandwidth

-From Metalink doc 260984.1. “Oracle does not recommend, suggest or dictate any values for these Parameters”

Neither do I…

Hardware used on this test
-general specification -

INDIANA CLIENT

CPU Intel(R) Xeon(TM)

4 CORES 2.80GHz

MEMORY: 5GB

OS: RHEL 4 Kernel level 2.6

NIC: 1000Gb/s

BNL ORACLE CLUSTER DATABSE server

2 Nodes

CPU= 2 dual core 3GHz, 64 bits Architecture

Memory: 9GB SGA, 16GB

RHEL 4 ES kernel level 2.6

NICs= 1000Gb/s

Dual controller 512MB cache per controller, 4Gbps FCP

Physical disks available:2 arrays of 10 disks each.

Disk type: IBM SAS 300GBDrive type: Serial Attached

SCSI (SAS) Capacity: 300GBSpeed: 15Krpm

Sysctl.Conf
-window r/w parameters-

Indiana Client

Parameters	min	Default	Max
net.ipv4.tcp_rmem	4k	87380	87380*2
net.ipv4.tcp_wmem	4k	16K	128K
net.core.rmem_default		110592	
net.core.rmem_max			131071
net.core.wmem_default		11[illegible]	
net.core.w[illegible]ax			131071

System Defaults

BNL Cond DB.

Parameters	min	Default	Max
net.ipv4.tcp_rmem	4096	87380	16777216
net.ipv4.tcp_wmem	4096	65536	16777216
net.core.rmem_default		135168	
net.core.rmem_max			16777216
net.core.wmem_default		135168	
net.core.wmem_max			16777216

Parameter (Tuned for Streams Data Replication)		
SDU_SIZE	RECV_BUF_SIZE	SEND_BUF_SIZE
32767	54750000	54750000

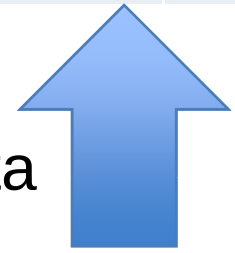
Sysctl.Conf
-window r/w parameters-

CERN Cond DB.

Parameters	min	Default	Max
net.ipv4.tcp_rmem	4k	87380	87380*2
net.ipv4.tcp_wmem	4k	16K	128K
net.core.rmem_default		262144	
net.core.rmem_max			4194304
net.core.wmem_default		262144	
net.core.wmem_max			4194304

BNL Cond DB.

Parameters	min	Default	Max
net.ipv4.tcp_rmem	4096	87380	16777216
net.ipv4.tcp_wmem	4096	65536	16777216
net.core.rmem_default		135168	
net.core.rmem_max			16777216
net.core.wmem_default		135168	
net.core.wmem_max			16777216



Tuned for Streams data replication from CERN

Problem overview
-Reconstruction Job-

Observation from client side

- Long execution time observed at client side when running Reco job.
- RECO jobs last for 30-40 minutes.
- Different thread database activity.
- Not high load on client side observed
- Reco Job resolultion

CLIENT	BNL COND DB	CERN COND DB
BNL Worker Node	~3 minutes	
CERN Worker Node		~3 minutes

Database Server observation

- 7 Threads connected to the database.
- 1 Thread on constant activity (sending/receiving data, about 30MB).
- Longest idle time observe was 37 minutes.
- Not database locks observed
- Not database load at job execution
- Data retrieved from DB cache
- RTT from BNL DB to client 30ms

Observation from Database side

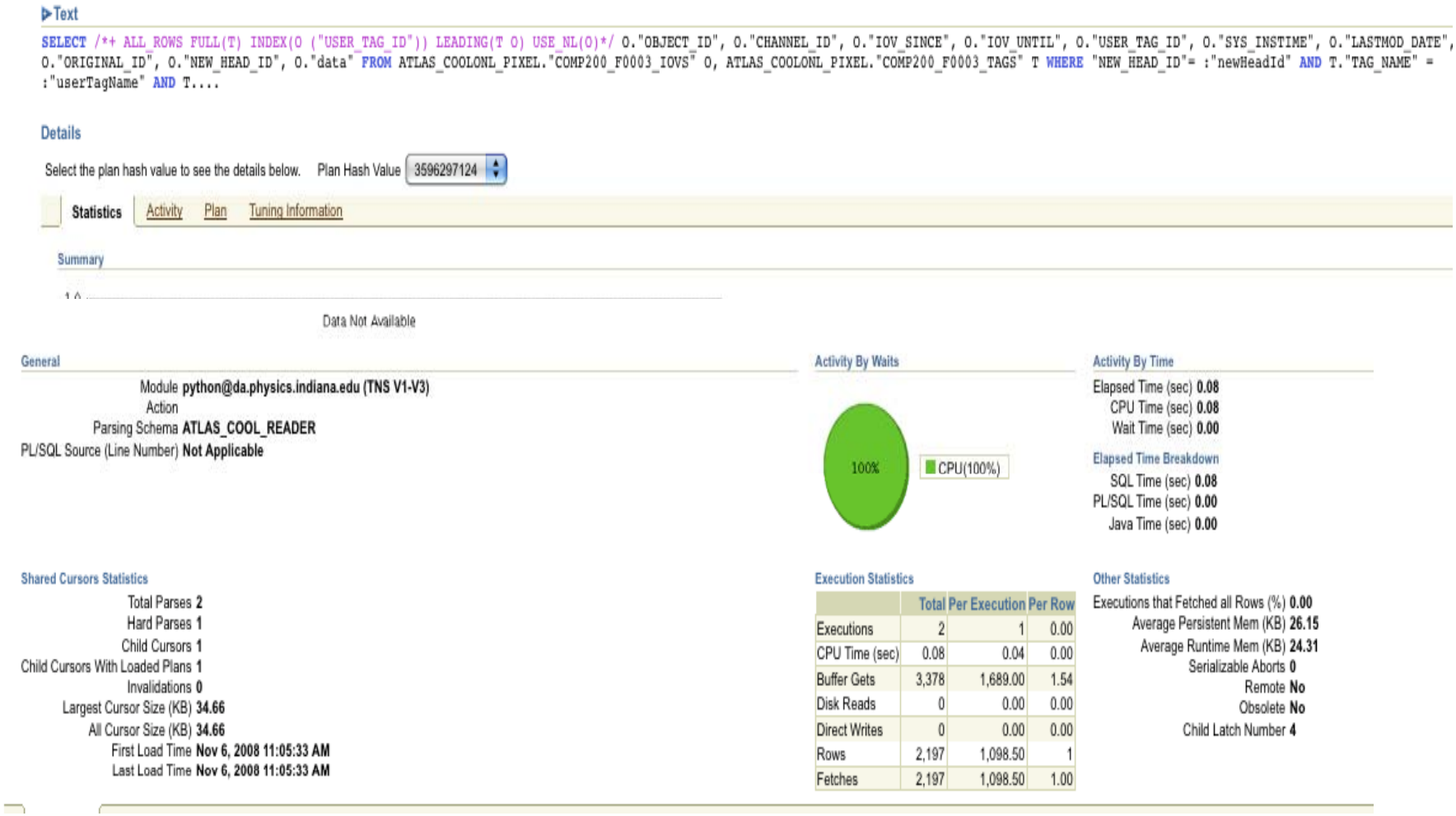
- Trace files show significant wait times events

  SQL*Net message from client

```
SELECT /*+ ALL_ROWS FULL(T) INDEX(O ("USER_TAG_ID")) LEADING(T O) USE_NL(O)*/ O."OBJECT_ID", O."CHANNEL_ID", O."IOV_SINCE", O."IOV_UNTIL", O."USER_TA
, O."SYS_INSTIME", O."LASTMOD_DATE", O."ORIGINAL_ID", O."NEW_HEAD_ID", O."moduleID", O."ModuleSpecialPixelMap_Clob" FROM ATLAS_COOLONL_PIXEL."COMP200
```

- Different “select count(*) ……” queries were found as well.

Monitoring
-OEM-



▶Text

```
SELECT /*+ ALL_ROWS FULL(T) INDEX(O ("USER_TAG_ID")) LEADING(T O) USE_NL(O)*/ O."OBJECT_ID", O."CHANNEL_ID", O."IOV_SINCE", O."IOV_UNTIL", O."USER_TAG_ID", O."SYS_INSTIME", O."LASTMOD_DATE",
O."ORIGINAL_ID", O."NEW_HEAD_ID", O."data" FROM ATLAS_COOLONL_PIXEL."COMP200_F0003_IOVS" O, ATLAS_COOLONL_PIXEL."COMP200_F0003_TAGS" T WHERE "NEW_HEAD_ID"= :"newHeadId" AND T."TAG_NAME" =
:"userTagName" AND T....
```

Details

Select the plan hash value to see the details below. Plan Hash Value 3596297124

Statistics | Activity | Plan | Tuning Information

Summary

Data Not Available

General

Module **python@da.physics.indiana.edu (TNS V1-V3)**
Action
Parsing Schema **ATLAS_COOL_READER**
PL/SQL Source (Line Number) **Not Applicable**

Activity By Waits

100% CPU(100%)

Activity By Time

Elapsed Time (sec) **0.08**
CPU Time (sec) **0.08**
Wait Time (sec) **0.00**

Elapsed Time Breakdown

SQL Time (sec) **0.08**
PL/SQL Time (sec) **0.00**
Java Time (sec) **0.00**

Shared Cursors Statistics

Total Parses **2**
Hard Parses **1**
Child Cursors **1**
Child Cursors With Loaded Plans **1**
Invalidations **0**
Largest Cursor Size (KB) **34.66**
All Cursor Size (KB) **34.66**
First Load Time **Nov 6, 2008 11:05:33 AM**
Last Load Time **Nov 6, 2008 11:05:33 AM**

Execution Statistics

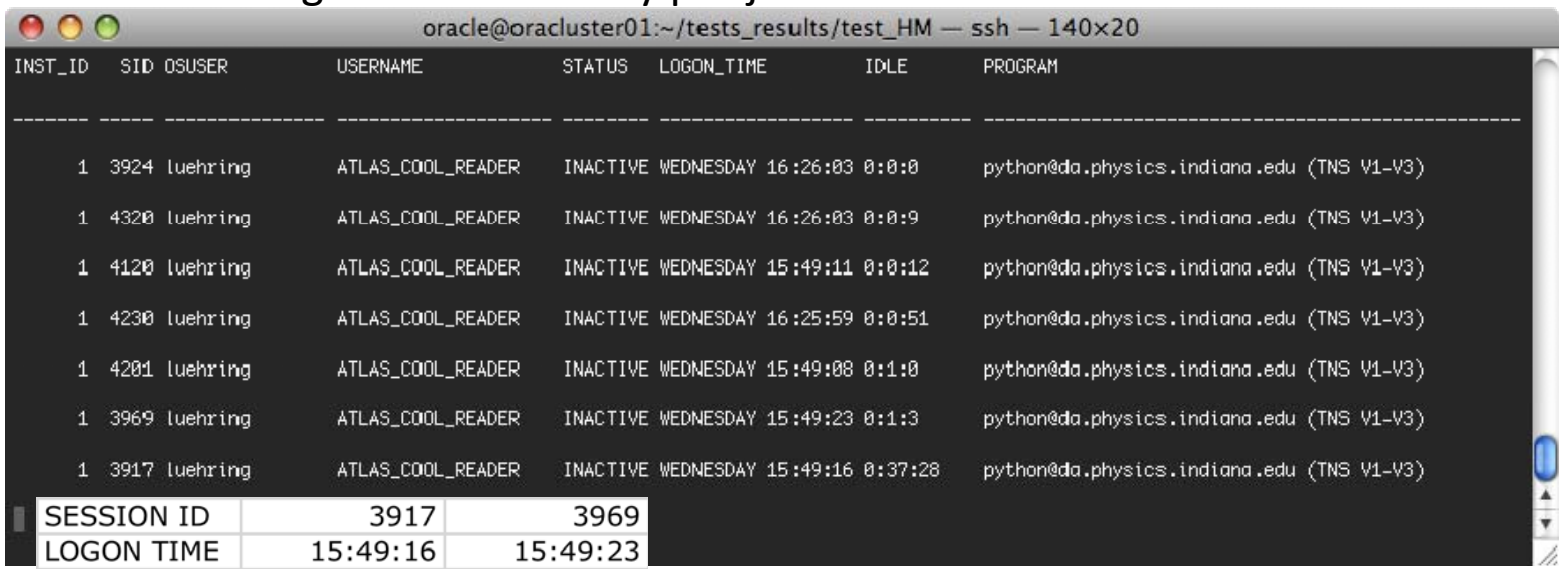
	Total	Per Execution	Per Row
Executions	2	1	0.00
CPU Time (sec)	0.08	0.04	0.00
Buffer Gets	3,378	1,689.00	1.54
Disk Reads	0	0.00	0.00
Direct Writes	0	0.00	0.00
Rows	2,197	1,098.50	1
Fetches	2,197	1,098.50	1.00

Other Statistics

Executions that Fetched all Rows (%) **0.00**
Average Persistent Mem (KB) **26.15**
Average Runtime Mem (KB) **24.31**
Serializable Aborts **0**
Remote **No**
Obsolete **No**
Child Latch Number **4**

Idle times during job execution

Monitoring idle time activity per job run



SESSION ID	3917	3969
LOGON TIME	15:49:16	15:49:23
Time sample	IDLE TIME	
155300	0:03:36	0:00:00
155513	0:05:51	0:00:00
155707	0:07:44	0:00:00
160734	0:18:10	0:00:00
161047	0:21:24	0:00:00
161747	0:28:25	0:00:00
161753	0:28:31	0:00:00
161806	0:28:43	0:00:00
161814	0:28:52	0:00:00
162017	0:30:55	0:00:00
162028	0:31:04	0:00:00
162113	0:31:49	0:00:00
162652	0:37:28	0:01:03

Idle state

IPERF tests

-First set of IPERF tests -

IPERF client at IDIANA to IPERF server at BNL database cluster (node 1)

Test	TCP window (MB)	Time (seconds)	Results (MBytes)	Troughput Mbits/s
1	1	30	165	46.1
2	1	100	566	47.5
3	2	10	50.8	42.5
4	16	10	54.6	45.8

IPERF warning message on every test result
TCP window size: 256 KByte (WARNING: requested 2.00 MByte)

Indiana parameters OS kernel parameters

```
# sysctl -p
net.ipv4.ip_forward = 1
net.ipv4.conf.default.rp_filter = 1
net.ipv4.conf.default.accept_source_route = 0
kernel.sysrq = 0
kernel.core_uses_pid = 1
vm.oom-kill = 0
```

IPERF tests
- Second set of IPERF tests -

IPERF client at IDIANA to IPERF server at BNL database cluster after tuning

Test	TCP window (MB)	Time (seconds)	Result (MBytes)	Troughput (Mbits/s)
1	1	30	395	110
2	16	10	227	186

IPERF warning message on last test result

TCP window size: 32.0 MByte (WARNING: requested 16.0 MByte)

Indiana new set of OS kernel parameters
tcp parameters

```
# sysctl -p
net.ipv4.ip_forward = 1
net.ipv4.conf.default.rp_filter = 1
net.ipv4.conf.default.accept_source_route = 0
kernel.sysrq = 0
kernel.core_uses_pid = 1
vm.oom-kill = 0
net.ipv4.tcp_rmem = 4096 87380 16777216
net.ipv4.tcp_wmem = 4096 65536 16777216
net.ipv4.tcp_timestamps = 1
net.ipv4.tcp_window_scaling = 1
net.ipv4.tcp_sack = 1
net.ipv4.tcp_dsack = 0
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216
```

IPERF tests
-Third set of IPERF tests -

IPERF client at IDIANA to IPERF server at BNL database cluster after tuning

Test	TCP window (MB)	Time sec	Result (MBytes)	Troughput Mbits/s
1	1	30	450	126
2	16	10	210	175

IPERF warning message on last test result

TCP window size: 32.0 MByte (WARNING: requested 16.0 MByte)

Indiana new set of OS kernel parameters
tcp parameters

```
# sysctl -p
net.ipv4.ip_forward = 1
net.ipv4.conf.default.rp_filter = 1
net.ipv4.conf.default.accept_source_route =
kernel.sysrq = 0
kernel.core_uses_pid = 1
vm.oom-kill = 0
net.ipv4.tcp_rmem = 4096 87380 33554432
net.ipv4.tcp_wmem = 4096 65536 33554432
net.ipv4.tcp_timestamps = 1
net.ipv4.tcp_window_scaling = 1
net.ipv4.tcp_sack = 1
net.ipv4.tcp_dsack = 0
net.core.rmem_max = 33554432
net.core.wmem_max = 33554432
```

TEST RESULTS

Date Tests	Test / tnsnames.ora	SDU (Bytes)	SND_BUFFER (Bytes)	RCV_BUFFER (Bytes)	Finish Time Reco Job (mm:ss:ms)	Max Idle observed minutes
10/28/08	Prior tuning	2048	OS	OS	35:00-40:00	~30-37
10/28/08	Job default (DBrelease)	2048	OS	OS	19:49.25	~18
10/28/08	2	8352	14250000	14250000	18:09:75	~13-14
10/28/08	3	32767	14250000	14250000	18:14:21	~15
11/04/08	4	2048	14250000	54750000	19:29.98	~18
11/04/08	5	8352	14250000	54750000	18:20:83	~16
11/04/08	6	31744	14250000	54750000	18:08:95	~17
11/06/08	7	2048	OS	OS	21:05.31	~17:15
11/06/08	8	8760	29200000	14600000	20:30:22	~13-14
11/06/08	9	31120	29200000	14600000	20:45:22	~15

Conclusions

- The Indiana Client TCP kernel configuration affected the resolution time when retrieving data from BNL database.
- Network latency affected Job time resolution.
- Although the initial idle time was considerably reduced there still a significant overhead of idle time which for my simple TRT histogram job seems to be ~15 minutes after tuning of the client side.
- Contact your local Network admin.

Acknowledgements

- Many thanks to:
  - THOM SALUKE Indiana University

References

Metalink Doc.

- 44694.1
- 1005123.6
- 260984.1
- 125021.1

Linux Man page:

- TCP
- sysctl.conf

Oracle Database 10g Real Application Clusters Handbook, McGraw Hill Osborne Media; 1 edition (November 22, 2006)

Online documentation

Oracle database concepts 10.2

http://download.oracle.com/docs/cd/B19306_01/server.102/b14220/toc.htm