IT-DES-DIS

CERN 1954-2004

Oracle database overview

OpenLab Student lecture

13 July 2006

Eric Grancher

Outline

- Who am I?
- What is a database server?
- Key characteristics of Oracle database server
  - Instrumentation
  - Clustering
  - Optimiser
- Usage at CERN

Who am I?

IT-DES-DIS CERN 1954-2004

- Graduated from “École Nationale Supérieure des Télécommunications” and “École Normale Supérieure de Lyon” in 1996 (master parallel computing) , France.
- Long term interest in (rather theoretical) computing.
- At CERN since 1996 in CN/IT department.
- Section leader for the Database Infrastructure Services section in IT / Database and Engineering Services group.
- OakTable network member (64 members worldwide “Oracle scientist, who believes in better ways of administering and developing Oracle based systems“).

Database

- Organized collection of data.
- Schema: structural description of the objects and relationship among them.
- Several models: “flat” (see spreadsheet), “hierarchical” (see file system), network, **relational**, object (ODMG)

Relational model

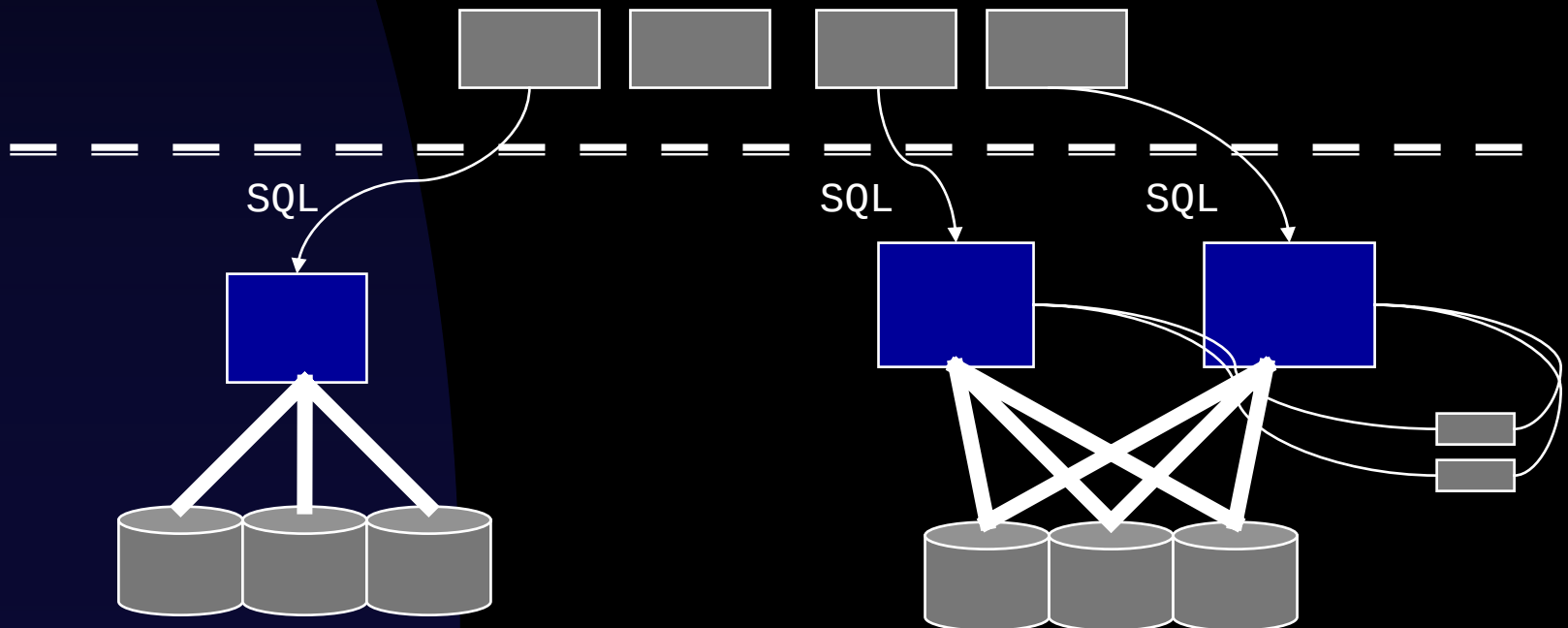
- Model defined with predicate logic and set theory (see E.F.Codd paper 1970).
- Relation database implementations are an approximation of the relational model.
- Multiple tables, keys used to match up rows in different tables (“foreign key”). Unique key, primary key.
- Dimensional: specialized adaptation of the relational model used to represent data in data warehouses.
- Language = SQL (Structured Query Language).

Transactions and integrity

- ACID rules: Atomicity, Consistency, Isolation, Durability.
- Single logical operation is called a transaction (can span a full night!).
- Important implications! Locking (on reading?), several copies of the data, network: two phase commit…

A database server

- Oracle instance = a set of processes that can receive (in SQL) statements, execute them to perform DDL, DML and queries (and administrative tasks).
- Oracle database = a set of files/devices that contain data, indexes, configuration data, metadata.
- Eventually several servers can run Oracle instances attached to the same Oracle database (=cluster, RAC).



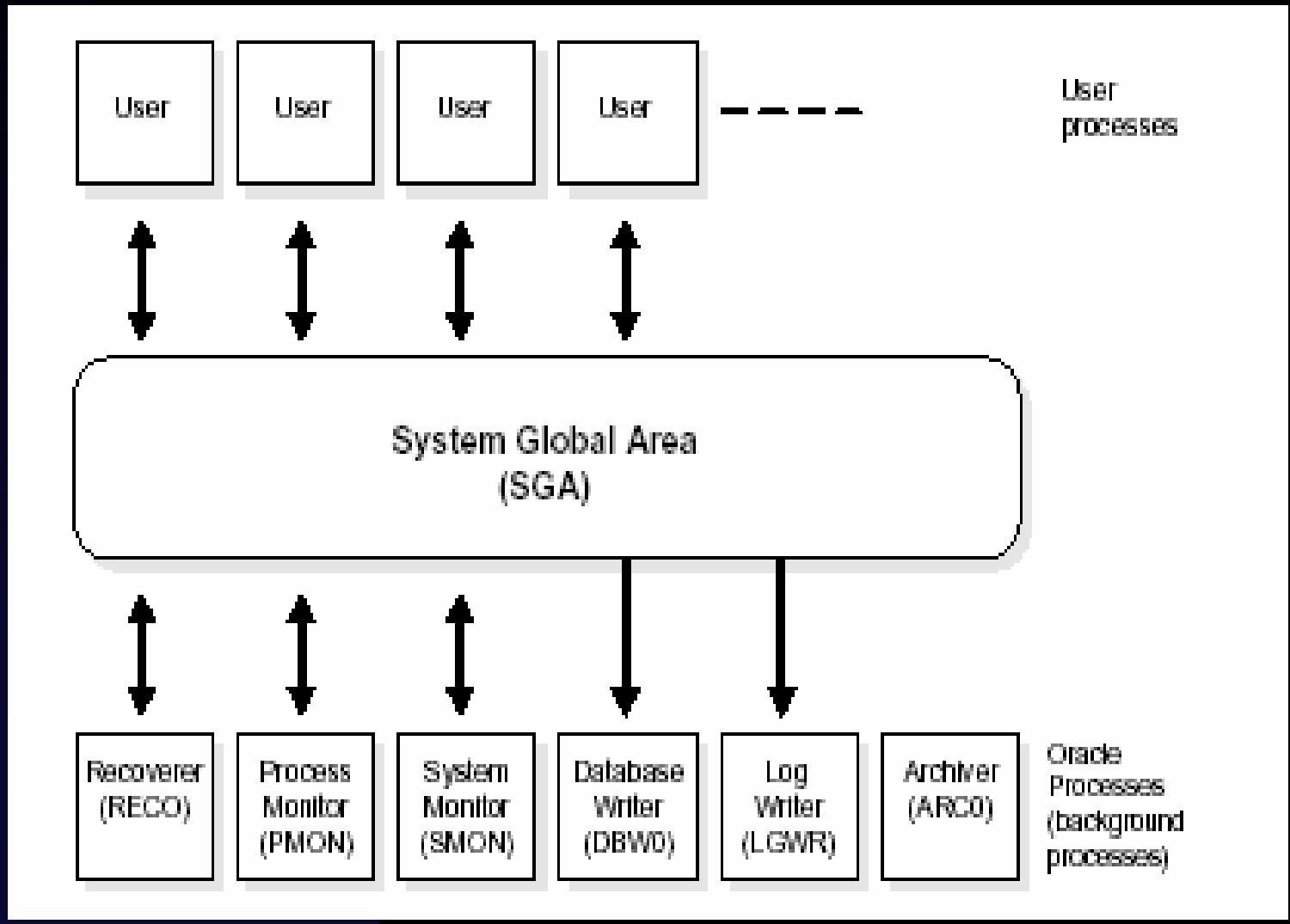
The Logical World

- Few slides thanks to Harvard cscie256
- A Tablespace is a logical division of a database.
- Each tablespace is made up of one or more files, called datafiles. A datafile belongs to one tablespace
- A table is a logical structure, inside a tablespace

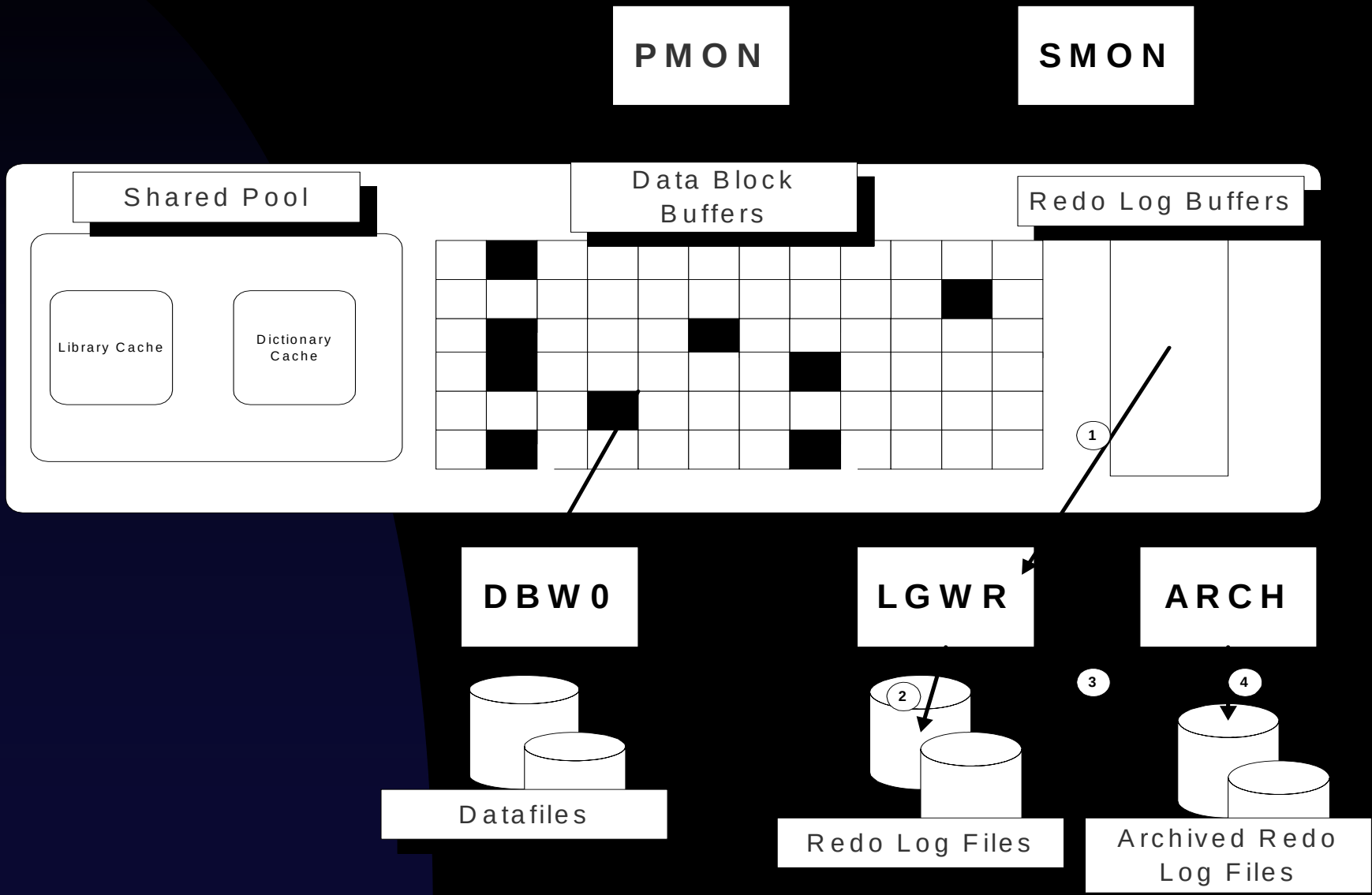
Internal Database Structures

- Table = The way to store data in a database
- Indexes = It is a partial copy of a table, used to help speed up accessing the data in the table
- Views = A method of looking at “some” of the data in a table or group of tables
- Procedures = Blocks of PL/SQL statements, called by applications. They **do not** return a value to the calling program
- Functions = Like a procedure, but **can** return a value to the calling program.
- Packages
  - These hold procedures and functions in logical groups
  - A package can have PUBLIC and PRIVATE elements
  - Public, would be procedures for a USERS usage
  - Private may be called by other procedures in the package
- Triggers
  - Procedures that execute when a specific event occurs
  - Used to augment referential integrity, enforce additional integrity
  - Statement triggers
  - Row triggers
- Sequences
  - Sequential lists of unique numbers, important for concurrency in application
- Users
- Schemas
- Database Links
- Undo Segments (rollback segments)

SGA and processes



SGA structure



System Global Area

- The **Data Buffer Cache** is the cache area where data blocks are read into from the data segments, such as tables, indexes, etc.
- Its size is controlled by the db_cache_size parameter in the init.ora file.
- This space is managed by a least recently used (LRU) algorithm
- If data is not in this area, it must be read from the datafile, so we have disk I/O.

System Global Area

- The **Shared Pool** stores the data dictionary cache and the library cache
- The **Library Cache** holds information about statements that have run against the database
- It allows the sharing of commonly used SQL statements
- It is also managed by an LRU algorithm
- Their sizes are set by the shared_pool_size parameter

System Global Area

- The **Redo log buffer** holds redo data on a transaction, before it gets written to the redo log
- Its size(in bytes) is controlled by the log_buffer parameter

Process architecture

- A process is a mechanism in an operating system that can run a series of steps.
- A process has its own private memory area
- An Oracle database server has
  - User processes
  - Oracle processes

Background Processes

- **PMON** cleans up failed user processes. It wake up periodically to check if it is needed.
- **SMON** checks to see if a database needs recovery, on startup. It also coalesces free space in tablespaces.

Background Processes

- **DBWR** manages the data block buffer cache and the dictionary cache. It handles the batch writes of changed blocks from the SGA to the datafiles. There can be multiple DBWR processes.

Background Processes

- **LGWR** manages the writing of the contents of the redo log buffer, to the online redo log files. It writes the log entries in batches. If the redo logs are mirrored sets, then both are written to simultaneously. There can be multiple LGWR processes

Background Processes

- **ARCH** performs the archiving of the redo log files. LGWR writes to the redo log files in a round robin fashion. When all are full, it over-writes the first one. However, if the database is in archive mode, Oracle takes a copy of this file and stores it on disk.

Interacting with a database server

- SQL is the typical language for interaction. (might be hidden).
- Reference interface is called Oracle Call Interface (C binding, all features exposed).
- Bindings exists for “all” third level languages (from C++ to Python). Specific generic interfaces like ODBC and JDBC. Oracle does not support all of these bindings.

The NULL logic

- Attributes in tables in database management systems can optionally be designated as NULL.
- This indicates that the actual value of the column is unknown or not applicable.
- In Oracle: A null can be assigned but it can not be equated with anything: Even itself!
- “three-value logic”.

“Fun” with NULL…

- See null.gwf

STUDENT_ID	STUDENT_NAME	STUDENT_FIRST
1	SMITH	John
2	SMITH	Bob
3	SMITH	
4	SMITH	
5	LUNDAHN	Kalle Ubbe
6	RUDSHAUG	Atle

Instrumentation

- Oracle database is fully instrumented. Timed wait interface is the key for all tuning exercises.
- select count(*) from v$event_name; -> 875
- Classes of events can be used as first classification.
- Can be seen locally, globally or globally-diff.
- Example: live system, no debug mode.

tkprof

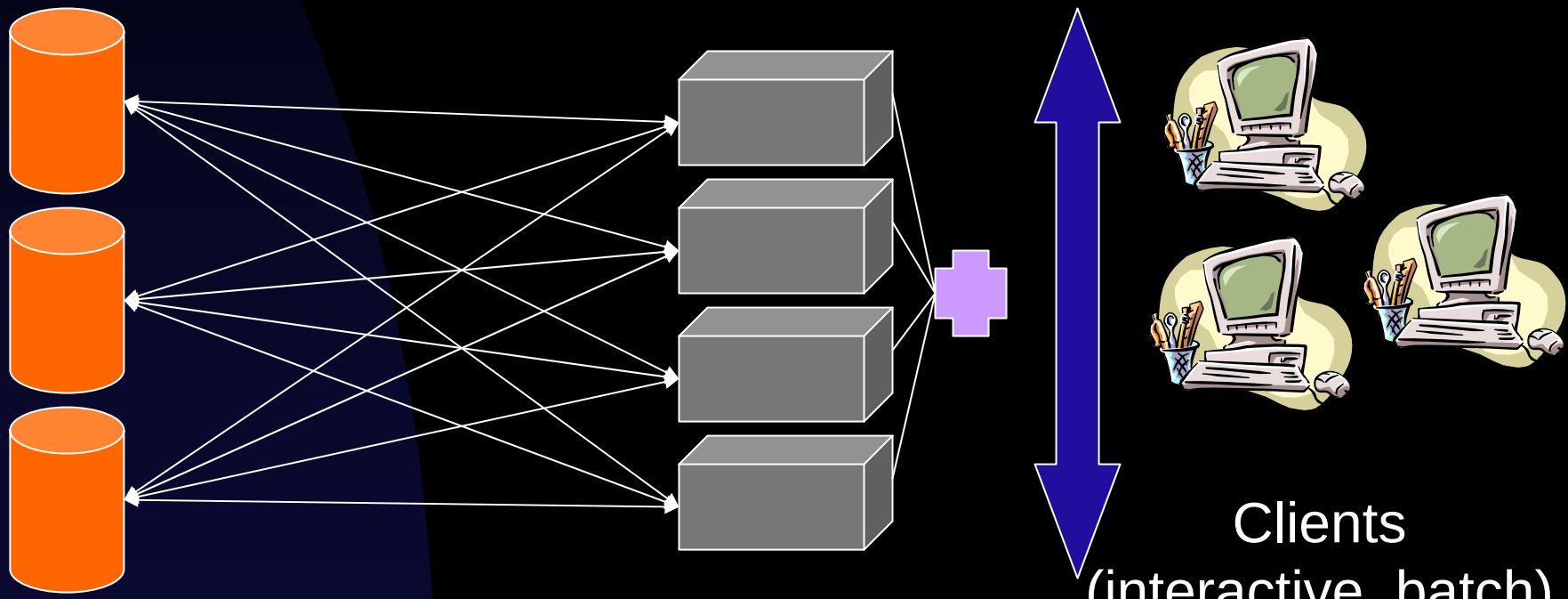

```
UPDATE STUDENTS SET STUDENT_FIRST=UPPER(STUDENT_FIRST)


call     count       cpu    elapsed       disk      query    current        rows
------- ------  -------- ---------- ---------- ---------- ----------  ----------
Parse        0      0.00       0.00          0          0          0           0
Execute     10      0.01       8.81          0         39        149          60
Fetch        0      0.00       0.00          0          0          0           0
------- ------  -------- ---------- ---------- ---------- ----------  ----------
total       10      0.01       8.81          0         39        149          60

Misses in library cache during parse: 0
Optimizer mode: ALL_ROWS
Parsing user id: 23  (GRANCHER)   (recursive depth: 1)

Rows     Execution Plan
-------  ---------------------------------------------------
      0  UPDATE STATEMENT   MODE: ALL_ROWS
      0   UPDATE OF 'STUDENTS'
      0    TABLE ACCESS   MODE: ANALYZED (FULL) OF 'STUDENTS' (TABLE)


Elapsed times include waiting on following events:
  Event waited on                             Times   Max. Wait  Total Waited
  ----------------------------------------   Waited  ----------  ------------
  enq: TX - row lock contention                  12        0.98          8.68
  latch: In memory undo latch                     1        0.00          0.00
  buffer busy waits                               5        0.01          0.01
```

Trace file


```
WAIT #8: nam='enq: TX - row lock contention' ela= 818431
   name|mode=1415053318 usn<<16 | slot=196641
   sequence=286494 obj#=2950721 tim=2541715479178
WAIT #8: nam='enq: TX - row lock contention' ela= 978719
   name|mode=1415053318 usn<<16 | slot=786457 sequence=77545
   obj#=2950721 tim=2541716467066
WAIT #8: nam='enq: TX - row lock contention' ela= 984360
   name|mode=1415053318 usn<<16 | slot=393255
   sequence=211919 obj#=2950721 tim=2541717451912
WAIT #8: nam='latch: In memory undo latch' ela= 9213
   address=15830840048 number=192 tries=1 obj#=-1
   tim=2541717468088
WAIT #8: nam='buffer busy waits' ela= 4 file#=21
   block#=22962 class#=1 obj#=2950721 tim=2541717468387
```

Clustering

- Shared storage (FC, iSCSI, NFS…), interconnects between the nodes.



Scaling with clustering

- Oracle clustering option is called Real Application Cluster but…
- Making application scale with clustering is not trivial:
  - Unless you are working on a “trivially scaling application” (read mostly)
  - Schema design
  - Interaction
- One of the key is to minimize block transfer between the RAC nodes.

RAC scalability example

- Important control application Oracle archiving
- Initially ~100 changes per second
- -- top event is : “SQL*Net message from client“
- -- change of insertion mechanism
- Then ~2000 changes per second
- -- no RAC scalability
- -- partitioning and unique insertion
- -- complete redesign
- Then ~150000 changes per second

Introduction to Cost Based Optimiser

- (simple version) All SQL statements are parsed and if no execution plan is already in the shared pool, the optimiser has to devise an execution plan.
- RBO used to “rule” the (Oracle) world
- CBO introduced with IO view only

Cost of a statement

- Cost (9i/10g)= (
  #SRds*sreadtim +
  #MRds*mreadtim +
  #CPUCycles/cpuspeed
) / sreadtim
- Before 9i, mreadtim=sreadtim and CPUCycles=0 so the formula is number of reads.
- Cost (before 9i)=
  #SRds

System stats

- System stats are mandatory with CBO
- Otherwise the system can not figure difference between single block read and multi block read and has no idea about the CPU
- (automatic gathering as of 10g).

10g SQL optimisation

- Several phases in the optimiser process
  - Logical optimisation
    - View merging
    - Sub-query un-nesting
    - Join predicate transitivity
  - Physical optimisation
    - Access method to every table (full scan, index lookup)
    - Join method for every join (HJ, SM, NL)
    - Join order for the query tables (join(join(A,B),C)

On your own...

- Oracle has released a free version “Oracle Database 10*g* Express Edition“ with restrictions.
- XE will store up to 4GB of user data, use up to 1GB of memory, and use one CPU on the host machine.
- For Linux x86 (Debian, Mandriva, Novell, Red Hat and Ubuntu) and MS Windows.
- http://www.oracle.com/technology/products/database/xe/index.html
- Also available free tools: JDeveloper, SQL Developer.
- Suggestion: install Oracle XE on a PC, redo the examples, use Application Express (automatically installed to create some database application: simple library).

References

- http://en.wikipedia.org/wiki/Database
- http://www.acm.org/classics/nov95/toc.html E.F. Codd
- http://www.psoug.org/reference/null.html NULL
- http://courses.dce.harvard.edu/~cscie256/docs/
- Technet http://www.oracle.com/technology/