Case studies – Atlas and PVSS Oracle archiver

Luca Canali, IT-DM

Database Developers' Workshop

July 8th, 2008

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

PVSS Oracle Archiver

CERN IT Department

- Commercial SCADA system
  - CERN use-case for the PVSS Oracle archiver has required several optimizations
  - In particular PVSS ‘out of the box’ in 2005 could insert at about ~100Hz
  - Requirement from the experiments: 3 orders of magnitude higher (at peak time)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

PVSS Optimization

CERN IT Department

- A joint effort
  - Input from experiments, IT/CO, ETM (vendor) and DBAs
  - Performances have been improved considerably
  - Thorough stress testing
  - Many lessons learned
  - A few slides follow focusing on techniques that can be of interest to HEP application developers

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

PVSS and high insert performance

CERN IT Department

- Bulk insert
- Fastest insert method
  - Direct path insert, i.e. insert /*+ append*/ ..
  - Bypasses the database cache, very fast, does not see slowdown from cluster cache management
  - It's a bulk operation
    - Implementation detail: data are buffered in temporary tables before 'bulk' insert
- The table is further partitioned
  - The insert process is partition-aware
  - Clients insert into a given partition

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

PVSS and large data volumes

CERN IT Department

- Very large amounts of data are created
  - Data is inserted into an ‘active table’ at a given time
  - When the table reaches a threshold it is archived
  - A new active table is used.
- Space-speed tradeoff
  - When insert rate is slow direct path insert leaves behind blocks with a lot of free space
  - The outcome is space ‘wastage’

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Space Management

CERN IT Department

- Compacting space
  - Archived tables can be recreated ('alter table move') to improve space filling
    - Implementation note: an update on the metadata tables is needed when changing tablespace
  - Extra gain by using oracle compressed table
    - 3x to 10x improvement measured (Atlas)
- Column order
  - Placing 'null columns at the end' saves space
  - ~20% in some tests

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Experimenting with IOTs

CERN IT Department

- Index Organized Tables
  - Data is stored in a index
  - Excellent clustering of data with the primary key
    - i.e. retrieval by PK is optimized
  - Saves space for the extra index compared to a ‘normal (heap) table
    - The problem of space wastage is drastically reduced
  - Can have performance problems if additional indexes are needed
  - Direct path insert is not available
    - max rate is about 80KHz (on quadcore machines)
  - Currently work in progress (not ‘certified’ yet)..

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

High Availability

- PVSS can work using a filesystem based buffer
  - This allows for the DB to be unavailable for short times
  - Fits in the service failover-clustering approach to Oracle availability (RAC)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Resource Usage

CERN IT Department

- Examining PVSS write-only activity (100KHz):
  - HW: 4-node RAC (dual single core Xeon) and ~32 SATA disks (4 arrays of 8 disks in RAID 10 configuration)
  - Data flow: 8 MB/sec (i.e. about 1TB in a WE)
  - 70 MB/sec writes to datafiles (of which 2/3 to undo segments!)
  - 40 MB/sec redo writes
- Notes:
  - PVSS is IO intensive, ultimately IO-bound
  - Oracle needs quite a few CPU cycles for I/O ops
    - With quadcore, 1 node has enough CPU for 100KHz (tested, with CPU at about 50%)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Query Tuning

- Running arbitrary queries on the system will not perform neither scale
  - The queries have been defined and ‘packaged’ for many use cases
  - Queries that retrieve data by PK are the most performing
  - Additional indexes are possible but with a cost on insert
    - Optionally the replicated offline schema can afford more/different indexing
- Agreement with the users community on the ‘allowed’ list of queries is fundamental

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Conclusions

- PVSS Oracle archiver has been tuned to scale and perform
  - at the peak insert rate values expected from LHC experiments
- Main lessons
  - Tuning works best if you put together a team of developers, application owners and DB experts
  - In addition, stress testing, and performance measuring is fundamental
  - Users requirements need to be defined as soon as possible

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Acknowledgments

CERN IT Department

- Many thanks to Gancho, Florbela, Jim (Atlas), Manuel, Wayne, and Piotr (IT-CO), Eric and Anton (IT-DES), Edward (ETM/ Siemens), Dawid, Eva, Jacek, Maria, Miguel (IT-DM).