~~Living on the edge~~
Exadata adventures

Yavor Ivanov

Disclaimer

On this presentation I will show many problems we encountered during the setup of a custom Exadata configuration in Mtel, and their solutions.

Please remember, that we have built some **very specific, custom implementation**. Most of the problems we encountered would not exist if we used the **standard preconfigured DB Machine**.

Also, please remember, that we were the first in the world to try such an implementation, at **the first days of Exadata existence**. Many of the problems are resolved by Oracle **by now.**

Agenda

- What is an Exadata1 system
  - What is an Exadata1 storage cell
  - What is Infiniband
  - What is HP Oracle DB machine
- Installation adventures
- Post-installation adventures
- Practical Exadata performance

Oracle Exadata1 storage cell

- HP ProLiant DL180 G5
  - 2 quad-core Intel Xeon Processor E5430 (2.66GHz)
  - 8GB memory
- 12 SAS or SATA HDD
- Oracle Enterprise Linux 5.1 + Oracle CellOS
- 1–HP InfiniBand Dual Port HCA

What is InfiniBand

- Industry Standard
- Open Source software
- Very good performance
  - Latency: ~1.07µs *(Mellanox ConnectX HCAs)*
  - Data transfer: current limit is 120 Gbit/s raw (96 Gbit/s useful) *(quad-rate 12X link)*

- Many features
  - RDMA, Atomic Operations, Shared Receive Queue
  - Hardware multicast, Quality of Service …
- Topology
  - InfiniBand uses a switched fabric topology (as in mainframe computers, fibre channel, etc.), opposed to a hierarchical switched network like Ethernet.

HP Oracle Database Machine

- 14 Exadata1 Storage cells
- 8 HP ProLiant DL360 database servers
  - 2 quad-core Intel Xeon Processor E5430 (2.66GHz)
  - 32GB memory
  - 1–HP InfiniBand Dual Port HCA
  - 4–146GB SAS 10K hard disk drives
- 4 Voltaire ISR9024D InfiniBand switches
- Oracle ASM + Oracle RAC

HP Oracle Database Machine capabilities (for full rack)

- With 450 GB SAS drives:
  - Up to 14 GB/sec of raw, uncompressed I/O throughput
  - Up to 1 TB/hour data loading
  - Up to 21 TB of user data
    (User data capacity is computed after mirroring and after allowing space for database structures such as temp, logs, undo, and indexes. User data capacity is uncompressed)
- With 1000 GB SATA drives:
  - Up to 10.5 GB/sec of raw, uncompressed I/O throughput
  - Up to 1 TB/hour data loading
  - Up to 46 TB of user data

Exadata system: proven to be GRID ready!

- Scalability - you can start with **3 cells** (minimum) and add more capacity as you need it.
  - Currently the biggest implementation in production is **3 full rack** DB machines working together (+ 1 rack for Data Guard)
- Balance - When adding capacity to the storage system, you add a whole cell:
  - HDD space
  - CPUs on the storage layer
  - Infiniband bandwidth

Mtel's implementation differences

- We are the first in the world to run (**successfully**) custom Exadata implementation
  - Exadata cells delivered by Oracle
  - HP DL360 servers with faster CPUs and more RAM, delivered by HP
  - Rack, Infiniband switches, cables, etc. delivered by HP
  - RHEL 5.2 on the computing nodes (instead of OEL 5.1)

Agenda

- What is an Exadata system
  - What is an Exadata storage cell
  - What is Infiniband
  - What is HP Oracle DB machine
- Installation adventures
- Post-installation adventures
- Practical Exadata performance

Hardware install

- Done by HP, at our site
- 2 days
- No problems

Software install

- Massive support from Oracle, since this is the first custom Exadata install in the world
  - On site:
    - Two consultants from Oracle RAC Pack
    - Daniela Milanova, *Oracle Bularia*
- Exadata software preinstalled on storage cells (OEL 5.1 + Oracle CellOS)
- RHEL 5.2 preinstalled on db nodes by Mobiltel engineers

Day 1

- We checked the hardware in the server room. Everything looks fine
- Some minor setup on the Exadata cells
  - IP settings
  - Naming
  - etc.
- The consultants start running scripts on the db nodes. Some minor Linux problems are found and instantly resolved by Mtel's engineers

Challenge 1

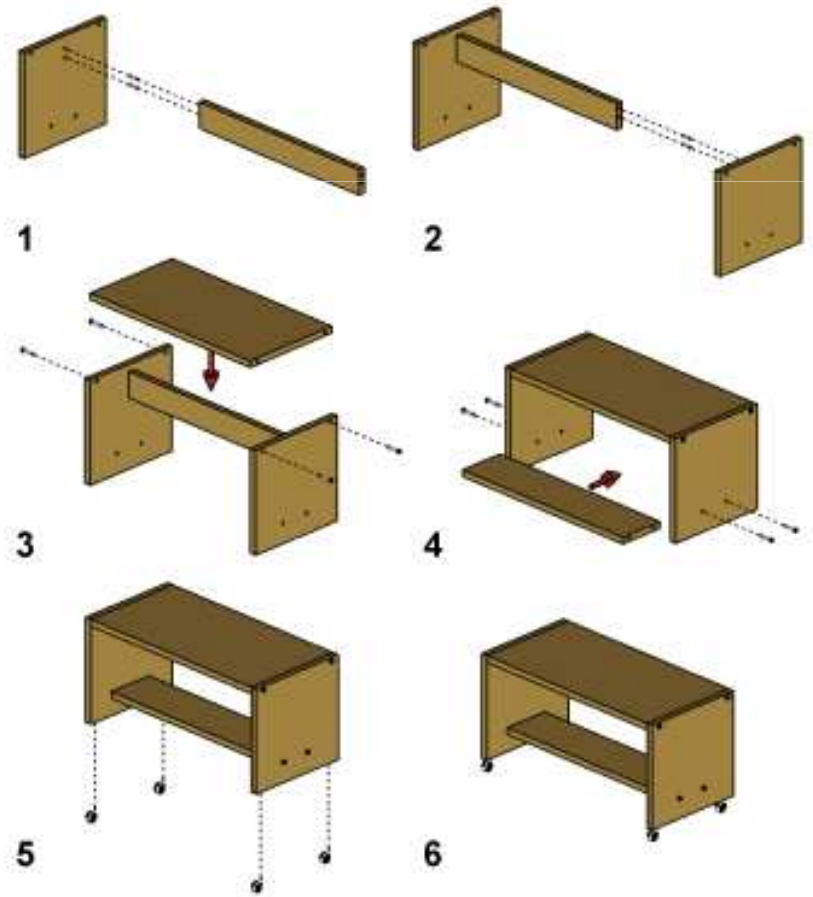
- There were no written installation prerequisites for the db nodes
  - No Linux package requirements
  - No kernel and network configuration requirements
- We used RAC 11.1 preinstallation requirements

We were assured by Oracle that it will work on RHEL 5.2, as long as we use Oracle's RDS (more on that later)

Currently there are some guidelines in Metalink note 757553.1 "Oracle Exadata Setup/Configuration Best Practices"

Challenge 2

- There was no usable installation guide
- Fortunately we had consultants from RAC Pack on-site



Challenge 3: OFED

- OFED stands for **OpenFabrics Enterprise Distribution –** a software package that supports InfiniBand and the RDS protocol
- We need to install **Oracle's version** of OFED
- We need to install the **latest Oracle version** of OFED
- All the nodes (Exadata cells and db nodes) MUST have exactly **the same version** of OFED

Challenge 3, contd.

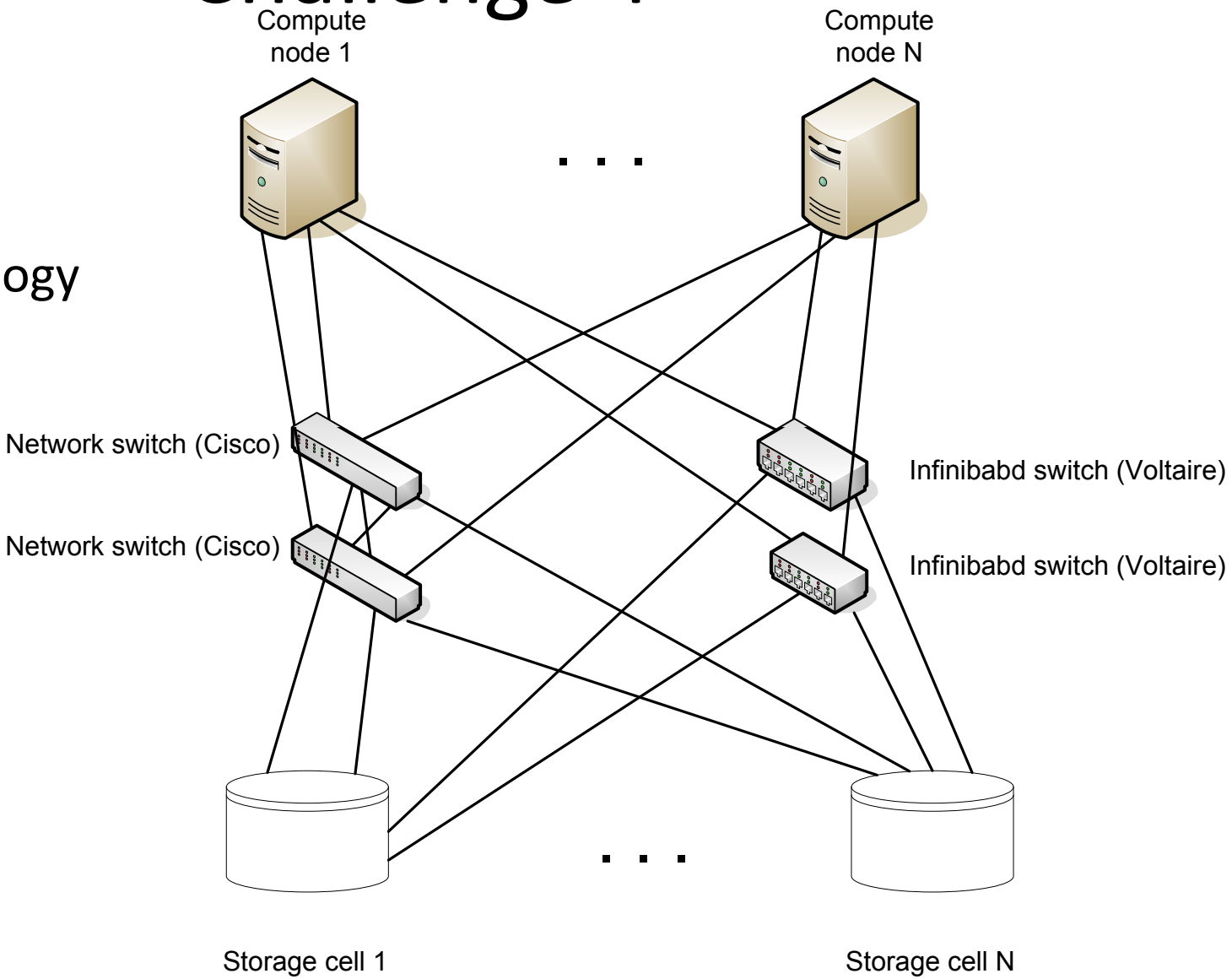
- We find downloadable OFED from Oracle on http://oss.oracle.com/el5/SRPMS-rds/
- We download OFED 1.3.1-10 (*OFED **10***) for both kernels and install in on all nodes
- It seems to work on cells, but not on db nodes
- At 21:00 we give up

Day 2

- The problem: we cannot make OFED work on our home-made db nodes
- The consultants gave us an image of computing node from the DB machine to test with
  - OEL 5.1 with all packages preinstalled and configurations already done
- Because of the minor differences in the hardware, we (hardly) install it, but with some errors
- RDS over Infiniband link seems to be working fine:
  - Between the db nodes
  - Between the storage cells
  - **But not between (any db node) and (any storage cell)**

Challenge 4

Something is missing in the network topology



Day 2, contd.

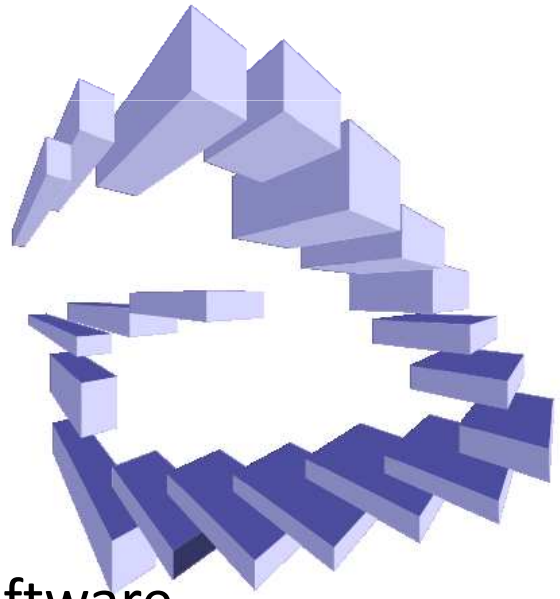
- The system seems to work with DB Machine's image, but
  - We do not have support with OEL
  - Our sysadmins do not have experience with OEL
  - There were some errors during the installation
  - The disk layout is not fine and we cannot change it
  - The system becomes black box

Evening 2

- We have conference call with some Exadata engineers/developers in UK and USA. They assure us that Exadata should work and is supported with any platform that supports Oracle RAC, as long as we use Oracle's OFED
- After a short discussion we decide to reinstall RHEL 5.2.
- This time we download OFED 1.3.1-9's source from Oracle, compile it for kernel 2.6.18-92 and it seems to run fine

Day 3

- We finally start to prepare for installation
  - We copy lots of scripts from Exadata cell to db node
- The only “installation guide” is a script called `/work/steps.sh` which should install everything
- We begin to run it step-by-step
- It has 26 steps
  - Check the hardware
  - Checks the network
  - Checks the installed software
  - Checks kernel parameters
  - Creates OCR and voting disk devices
  - Installs and patches CRS, ASM and DB software



Challenge 5

- At 10:30 we see that we missconfigured the disk layout on one of the db nodes
- We need to reinstall it completely
  - Setup the RAID
  - Install RHEL 5.2
  - Compile, install and configure OFED
  - Configure network – addresses, bounding, etc.
- It took 30 minutes to our sysadmin to do it
- The consultants from Oracle started to say "He is very efficient!" ☺

Challenge 6, 7, 8, 9

- The script `steps.sh` has lots of… side effects, like
  - After step 5 one should check if `dcli` is working
  - Step 6 generates wrong `private_ib_hosts` file
  - Step 9 damages `modprobe.conf` and the db nodes loose network connectivity
  - At step 10 some files are missing; we copy over from some storage cell
  - etc.
- Hopefully, after our experiences some of the bugs are fixed

Challenge 10

- Step 15 configures 6 partitions on the first 3 cells and present them through iSCSI on the db nodes. This are for OCR and Voting disks:

```
lrwxrwxrwx 1 root root 3 May 15 14:43 /dev/ocrvota -> sdb
lrwxrwxrwx 1 root root 4 May 15 14:43 /dev/ocrvota1 -> sdb1
lrwxrwxrwx 1 root root 4 May 15 14:43 /dev/ocrvota2 -> sdb2
lrwxrwxrwx 1 root root 3 May 15 14:43 /dev/ocrvotb -> sdc
lrwxrwxrwx 1 root root 4 May 15 14:43 /dev/ocrvotb1 -> sdc1
lrwxrwxrwx 1 root root 4 May 15 14:43 /dev/ocrvotb2 -> sdc2
lrwxrwxrwx 1 root root 3 May 15 14:43 /dev/ocrvotc -> sda
lrwxrwxrwx 1 root root 4 May 15 14:43 /dev/ocrvotc1 -> sda1
lrwxrwxrwx 1 root root 4 May 15 14:43 /dev/ocrvotc2 -> sda2
```

- A side effect – when we run step 15 we loose network connectivity to the db node we use.
- God save the iLO

Challenge 11

- Step 18 – ASM home installation and ASM setup
- When DBCA tries to setup ASM, it needs listener
- When NetCA tries to start listeners, it says it `cannot find IP configuration in the file CELLIP.ORA`
- After less then 3 hours of investigation we discover that this errors means, "**rds module is not loaded in the linux kernel**"
  (Of course, Google and Metalink cannot say anything about this at that time)

Challenge 12

- DBCA hangs when tries to create ASM diskgroups
- We leave this for the next day
  (it is about 22:00)

Day 4

- The plan:
  - Setup ASM
  - Create a database
  - Load some test data
  - Test the system

Challenge 12 – contd.

- DBCA hangs when we setup ASM instances
- It hangs on “Creating diskgroups”
- Almost every query in the ASM instances hangs

Challenge 12 - solution

- It took us less than 15 hours intensive debugging (with support engineers from UK, Germany and US) to find it is caused by **incorrect OFED installation**

- **BUG 7514146** has an OFED-10 tarball for db node's kernel

Day 4: 2:00-3:00 in the night

- We created the diskgroups
- We created a database
- **This finishes the installation**

Agenda

- What is an Exadata system
  - What is an Exadata storage cell
  - What is Infiniband
  - What is HP Oracle DB machine
- Installation adventures
- Post-installation adventures
- Practical Exadata performance

Post-installation adventures: the easy ones

- HP delivered 4 ib cables for inter-switch connection
- We had one minor cell failure with long root-cause haunting

Post-installation adventures: the hard one

- The speed of the Infiniband connections was not as good as expected:

```
# infinicheck -d –p
. . .
##    [(1) Every DBNODE to its STORAGE CELL    ######
          -------Throughput results using rds-stress --------
        600 MB/s and above is expected for runs on quiet machines
node01-priv to cell01-priv : 498 MB/s...WARNING
. . .
####      [(2) Every DBNODE to its PEER        ######
          --------Throughput results using rds-stress --------
        1100 MB/s and above is expected for runs on quiet machines
node01-priv to node02-priv : 691 MB/s...ALERT
. . .
```

Post-installation adventures: the hard one

- It took really lot of efforts to diagnose the issue
- We have involved Oracle engineers from USA, UK and Germany, 3 members of RAC Pack, and some Oracle Developers
- The main work force was from HP:
  - 3 engineers from Bulgaria
  - 1 from Germany
  - 3 from India
  - 2 from *HP High-Performance Computing Team* in US

Post-installation adventures: the hard one

- We checked lots of things:
  - OS kernel configuration
  - Infiniband Switch configuration
    - Upgraded the firmware
  - Infiniband fabric
  - OFED versions and configurations
  - BIOS and Firmware versions on cells and db nodes
  - If there is some problem with HDD, motherboards, IB HCAs, etc.
- We have made tons of diagnostics

Post-installation adventures: the hard one

- And the problem was…

(please, don’t laugh too loudly)

- The Infiniband adapters in the DB nodes were on PCI x4 slots (half height).

PCI x4 half-height
PCI x8 full-height
CAUTION:
CAUTION:

Post-installation adventures: the hard one

- We moved the IB adapters to PCI x8 slots, the results became like this:

```
# infinicheck -d –p
. . .
##    [(1) Every DBNODE to its STORAGE CELL    ######
          -------Throughput results using rds-stress --------
        600 MB/s and above is expected for runs on quiet machines
node01-priv to cell01 : 665 MB/s...OK
. . .
####      [(2) Every DBNODE to its PEER        ######
          --------Throughput results using rds-stress --------
        1100 MB/s and above is expected for runs on quiet machines
node01-priv to node02-priv : 1188 MB/s...OK
. . .
```

Post-installation: patching

- We had to install a patch to CellOS 1 month after the install. We had NO problems during the (long and complex) patching procedure
- Regular patches appear every ~2 months. They need 2-5 hours downtime
  - We are pressing Oracle to make them rolling-installable

Exadata support policy

- For everything (even hardware failures) the customer have to raise SR with Oracle. Then they may open a ticket with HP
- You cannot change anything on the cells (even firmware upgrade) if it is not approved by Oracle
- The support is very prompt. Some of the best support engineers in Oracle work for Exadata

Agenda

- What is an Exadata system
  - What is an Exadata storage cell
  - What is Infiniband
  - What is HP Oracle DB machine
- Installation adventures
- Post-installation adventures
- Practical Exadata performance

App 1: XXX application

- The business demand was stopped for several years. On the old platforms we could not run it (MS SQL server). So we load the data in Oracle DB on Exadata
- Web-based, read-only application (new data is loaded on batches every 5 minutes)
- **Largest table is ~ 150 million rows**
  - Partitioned
  - Local indexes
- **OLTP-like load**: short, indexed queries, returning ~1-10 rows
  - **Average query time: < 1 sec**

App 2: Exadata as a migration platform

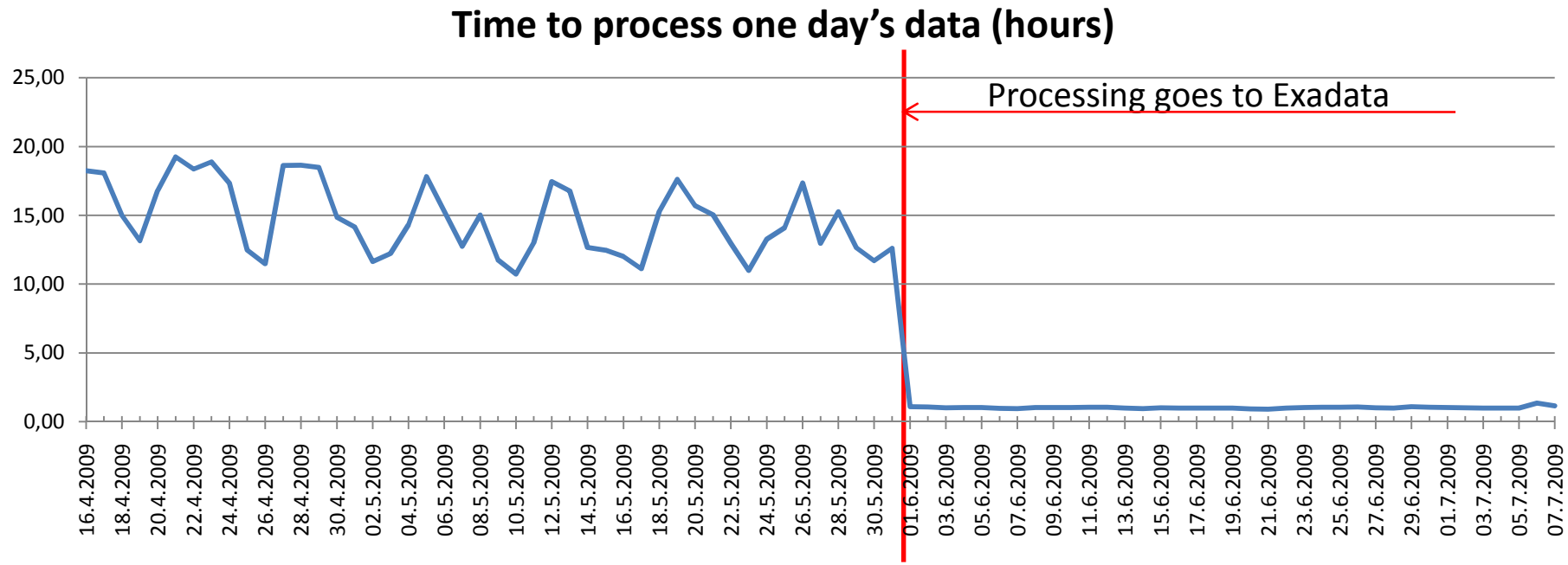
- Great performance

- It is faster to
  - Transfer some tens of millions of rows from another system to Exadata
  - and do the processing on Exadata
- than
  - processing the data on the source system

App 3: NII application

- NII refresh procedure
  - DWH-like load: aggregations over tens of millions of rows
  - Executed daily on one day's data
  - Optimized with all DBA tools and experience we have

Execution statistics:



App 4: YYY application

- Unique automation for an important service
- Possible thanks to the speedup in NII application
- Aggregations on hundreds thousands rows…
- … in seconds, of course ☺

App 5: Online CDR database

- Data in the terabytes range
- Used by really lots of departments, issuing many different (and heavy) ad-hoc queries, reports, checks, etc.

- We are still testing it

App 6: Exadata as a staging area for ETL

- We are still testing it

Q & A