CERN

Running Relational Databases on a C-mode Storage Cluster

Ruben.Gaspar.Aparicio_@_cern.ch
CERN, IT Department

UKOUG, Birmingham 28th May 2014

Proton Antiproton collision leading to discovery of W and Z particles. 1984 Nobel Prize: Carlo Rubbia & Simon van der Meer.

About me

- Joined CERN in 2000 to design and implement a J2EE application for accelerator controls
- Joined CERN IT Databases group on 2007
  - From Oracle 9i on
- Project leader of the backup and recovery service till January 2013
- Project leader of the storage infrastructure
- Project leader of the DBaaS service

Agenda

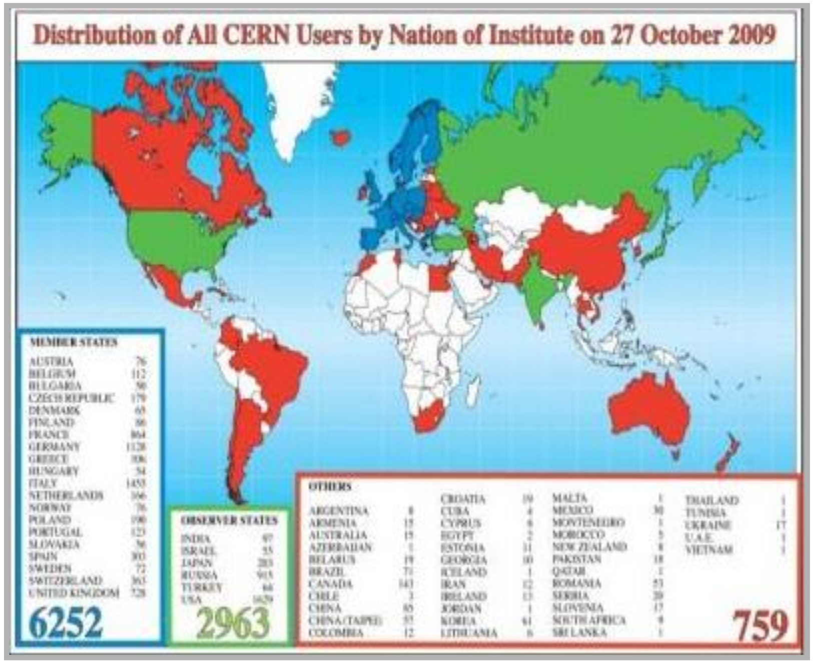
- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
  - Flash cache
  - Flash pool
- Data motion
- Snapshots
- Clonning in Oracle12c
- Backup to disk
- directNFS
- Monitoring
  - In-house tools
  - Netapp tools
- Conclusions

Agenda

- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
  - Flash cache
  - Flash pool
- Data motion
- Snapshots
- Clonning in Oracle12c
- Backup to disk
- directNFS
- Monitoring
  - In-house tools
  - Netapp tools
- Conclusions

CERN

- European Organization for Nuclear Research founded in 1954
- Membership: 21 Member States + 7 Observers
- 60 Non-member States collaborate with CERN
- 2400 staff members work at CERN as personnel + 10000 researchers from institutes world-wide



Distribution of All CERN Users by Nation of Institute on 27 October 2009

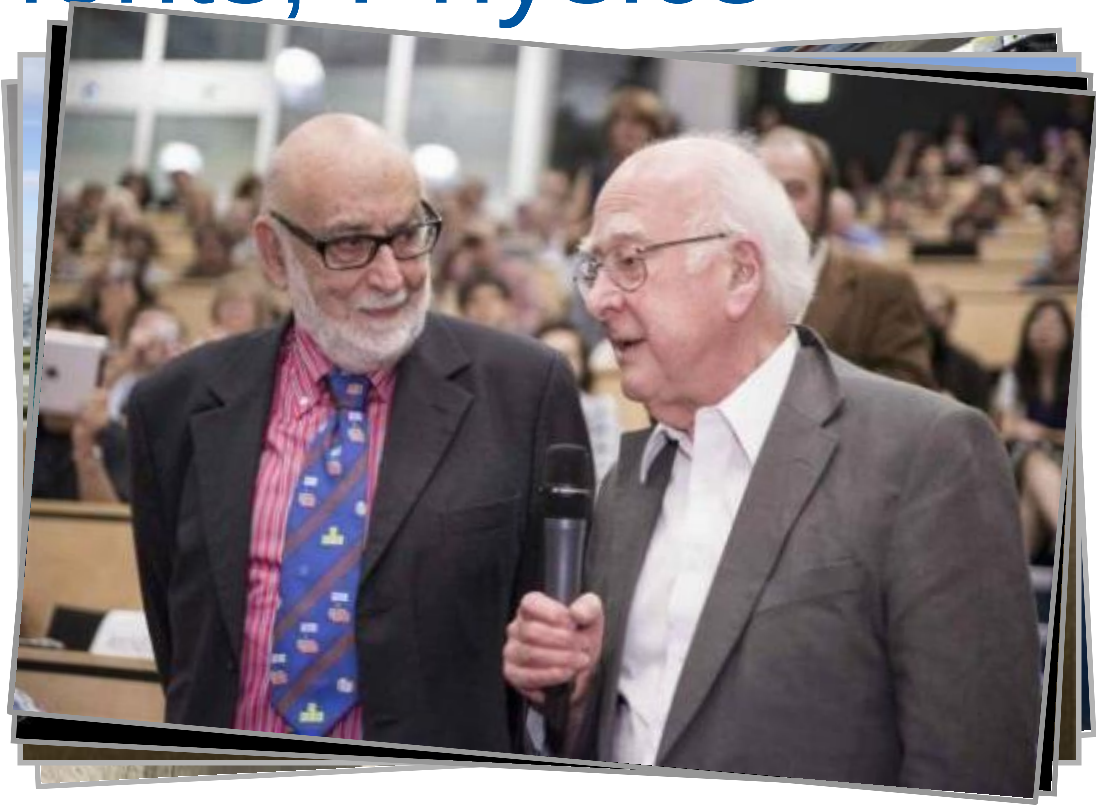
MEMBER STATES	
AUSTRIA	76
BELGIUM	112
BULGARIA	50
CZECH REPUBLIC	179
DENMARK	65
FINLAND	88
FRANCE	864
GERMANY	1128
GREECE	106
HUNGARY	54
ITALY	1453
NETHERLANDS	166
NORWAY	76
POLAND	190
PORTUGAL	123
SLOVAKIA	56
SPAIN	303
SWEDEN	72
SWITZERLAND	363
UNITED KINGDOM	728
6252	

OBSERVER STATES	
INDIA	97
ISRAEL	55
JAPAN	203
RUSSIA	915
TURKEY	64
USA	1629
2963	

OTHERS							
		CROATIA	19	MALTA	1	THAILAND	1
ARGENTINA	8	CUBA	4	MEXICO	30	TUNISIA	1
ARMENIA	15	CYPRUS	6	MONTENEGRO	1	UKRAINE	17
AUSTRALIA	15	EGYPT	2	MOROCCO	5	U.A.E.	1
AZERBAIJAN	1	ESTONIA	11	NEW ZEALAND	8	VIETNAM	1
BELARUS	19	GEORGIA	10	PAKISTAN	18		
BRAZIL	71	ICELAND	1	QATAR	1		
CANADA	143	IRAN	12	ROMANIA	53		
CHILE	3	IRELAND	13	SERBIA	20		
CHINA	85	JORDAN	1	SLOVENIA	17		
CHINA (TAIPEI)	57	KOREA	61	SOUTH AFRICA	9		
COLOMBIA	12	LITHUANIA	6	SRI LANKA	1	759	

LHC, Experiments, Physics

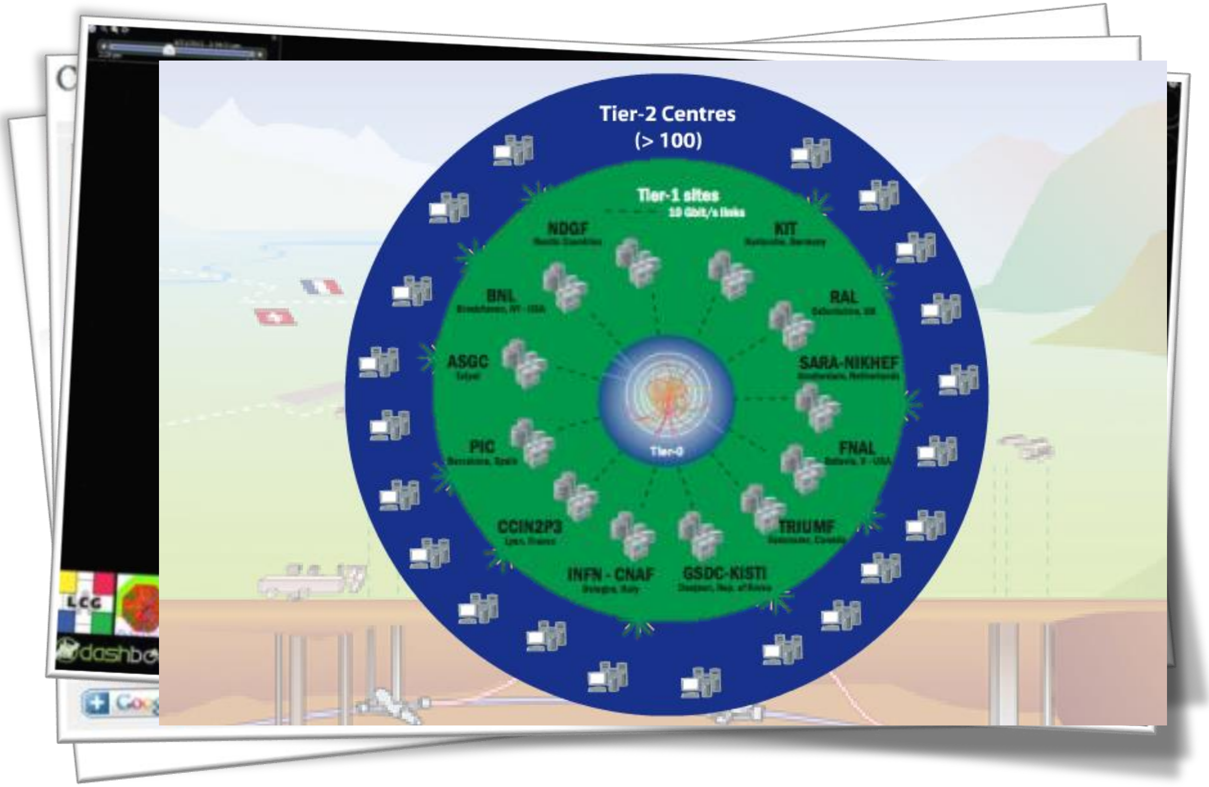
- Large Hadron Collider (LHC)
  - World’s largest and most powerful particle accelerator
  - 27km ring of superconducting magnets
  - Currently undergoing upgrades, restart in 2015
- The products of particle collisions are captured by complex detectors and analyzed by software in the experiments dedicated to LHC
- **Higgs boson discovered!**



- The Nobel Prize in Physics 2013 was awarded jointly to François Englert and Peter W. Higgs *"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"*

WLCG

- The world’s largest scientific computing grid



More than 100 Petabytes
of data stored and analysed.
Increasing: 20+ Petabytes/year

CPU: over 250K cores
Jobs: 2M per day

160 computer centres in 35 countries

More than 8000 physicists with real-time access to LHC data

Agenda

- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
  - Flash cache
  - Flash pool
- Data motion
- Snapshots
- Clonning in Oracle12c
- Backup to disk
- directNFS
- Monitoring
  - In-house tools
  - Netapp tools
- Conclusions

CERN’s Databases

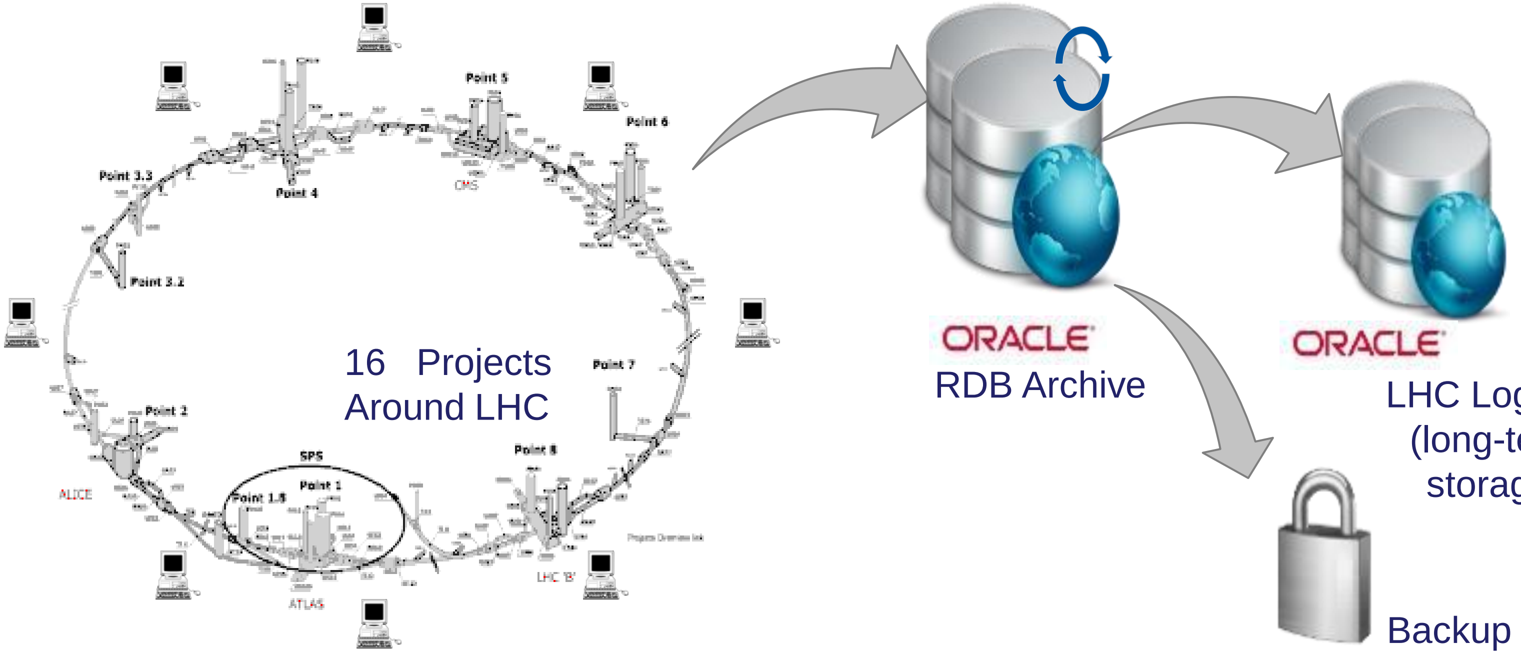
- ~100 Oracle databases, most of them RAC
  - Mostly NAS storage plus some SAN with ASM
  - ~500 TB of data files for production DBs in total

- Examples of critical production DBs:
  - LHC logging database ~170 TB, expected growth up to ~70 TB / year
  - 13 production experiments’ databases ~10-20 TB in each
  - Read-only copies (Active Data Guard)

- But also as DBaaS, as single instances
  - 120 MySQL Open community databases (migrating to 5.6)
  - 11 Postgresql databases (version 9.2, since September 2013)
  - 10 Oracle11g → migrating towards Oracle12c multi-tenancy

Use case: Quench Protection System

- Critical system for LHC operation
    - Major upgrade for LHC Run 2 (2015-2018)
- High throughput for data storage requirement
    - Constant load of 150k changes/s from 100k signals
- Whole data set is transfered to long-term storage DB
    - Query + Filter + Insertion
- Analysis performed on both DBs



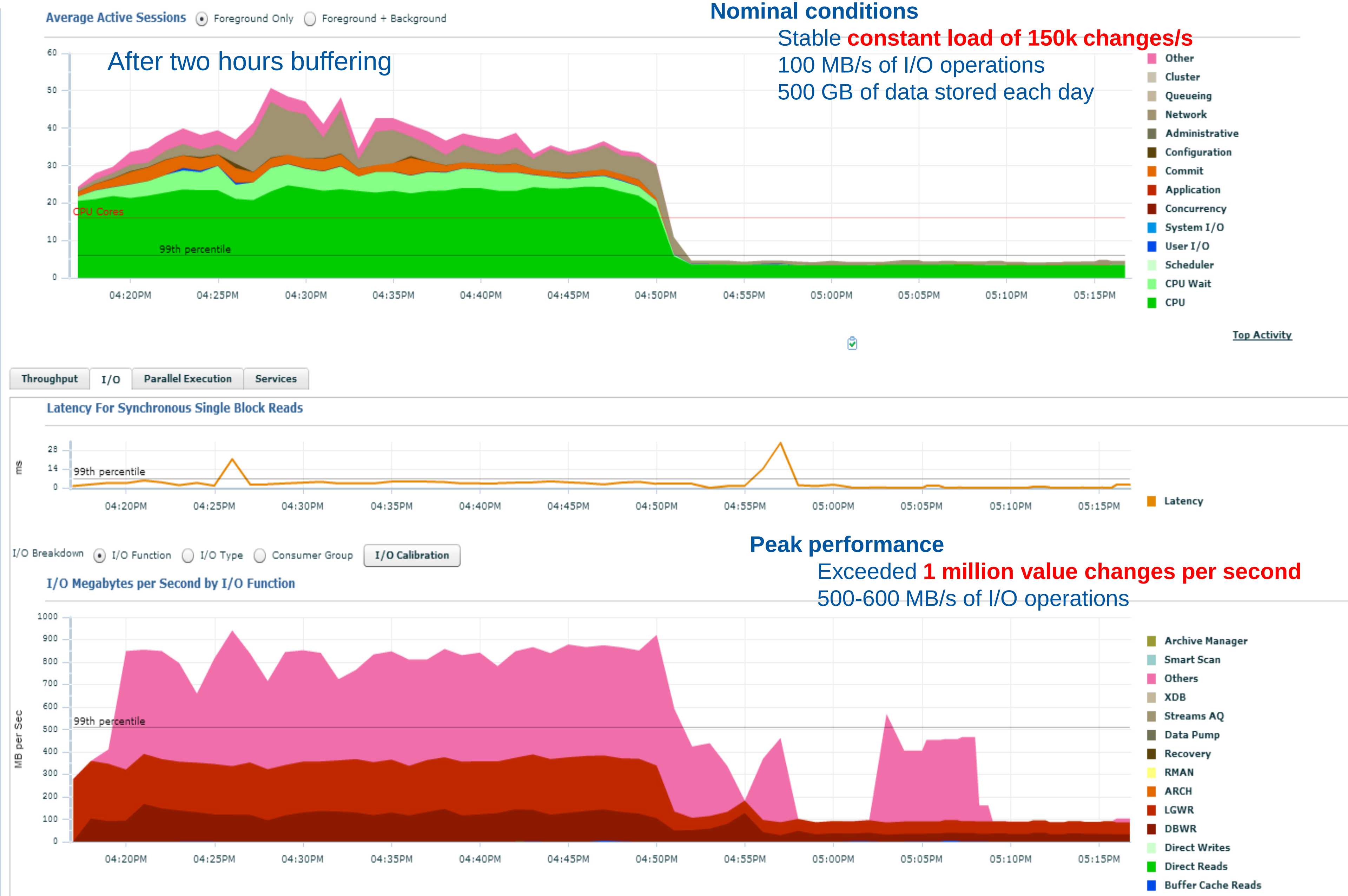
16 Projects Around LHC

RDB Archive

LHC Logging (long-term storage)

Backup

Quench Protection system: tests



After two hours buffering

Nominal conditions

Stable **constant load of 150k changes/s**

100 MB/s of I/O operations

500 GB of data stored each day

Peak performance

Exceeded **1 million value changes per second**

500-600 MB/s of I/O operations

Average Active Sessions ● Foreground Only ○ Foreground + Background

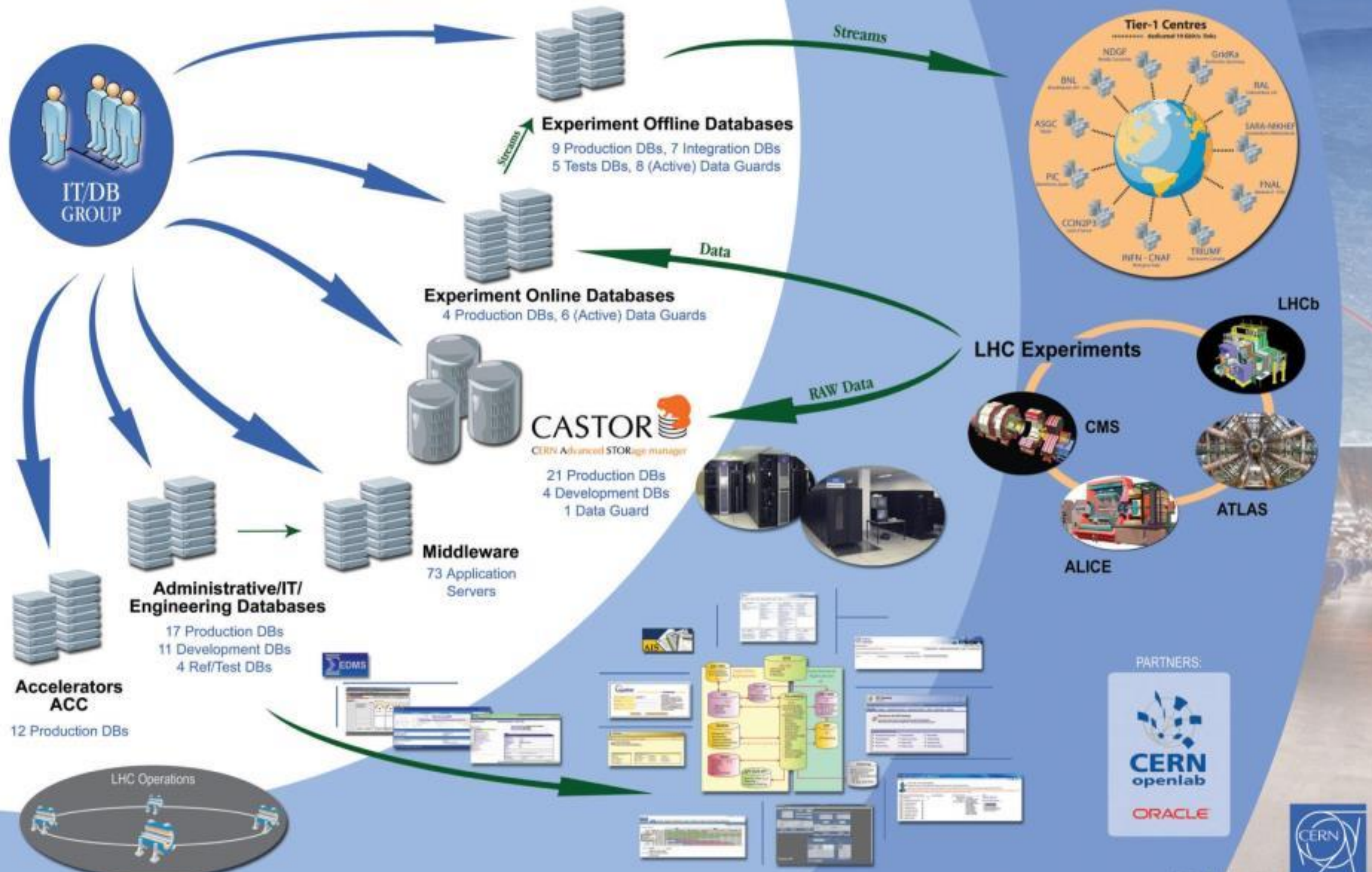
I/O Breakdown ● I/O Function ○ I/O Type ○ Consumer Group

Database SERVICES

At the heart of CERN, LHC and Experiment Operations

CERN IT Department

http://cern.ch/it-dep/db/



IT/DB GROUP

Experiment Offline Databases

9 Production DBs, 7 Integration DBs
5 Tests DBs, 8 (Active) Data Guards

Experiment Online Databases

4 Production DBs, 6 (Active) Data Guards

CASTOR
CERN Advanced STORage manager

21 Production DBs
4 Development DBs
1 Data Guard

Middleware

73 Application Servers

Administrative/IT/Engineering Databases

17 Production DBs
11 Development DBs
4 Ref/Test DBs

Accelerators ACC

12 Production DBs

PARTNERS:

CERN openlab

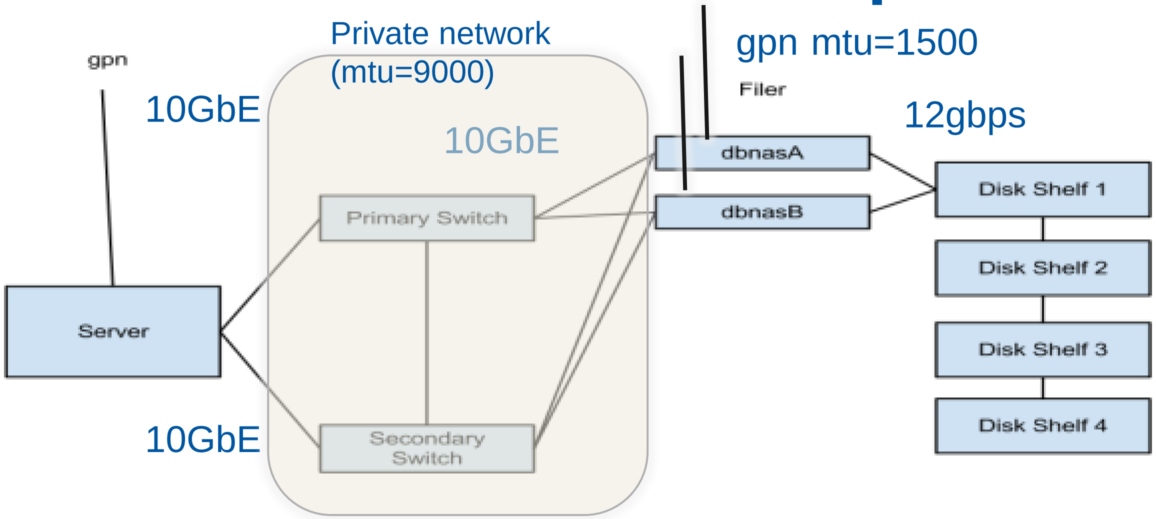
ORACLE

Current status at November [illegible]

Oracle and NetApp at CERN

- 1982: Oracle at CERN, PDP-11, mainframe, VAX VMS, Solaris SPARC 32 and 64 bits
- 1996: Solaris SPARC with OPS
- 2000: Linux x86, local storage
- 2005: Linux x86_64 / RAC / EMC and ASM
- >=2006: Linux x86_64 / RAC / NFS / NetApp
- (96 databases)
- 2011-2012: migration of all (*) databases to Oracle on NetApp

Oracle basic setup



Private network (mtu=9000)

10GbE

10GbE

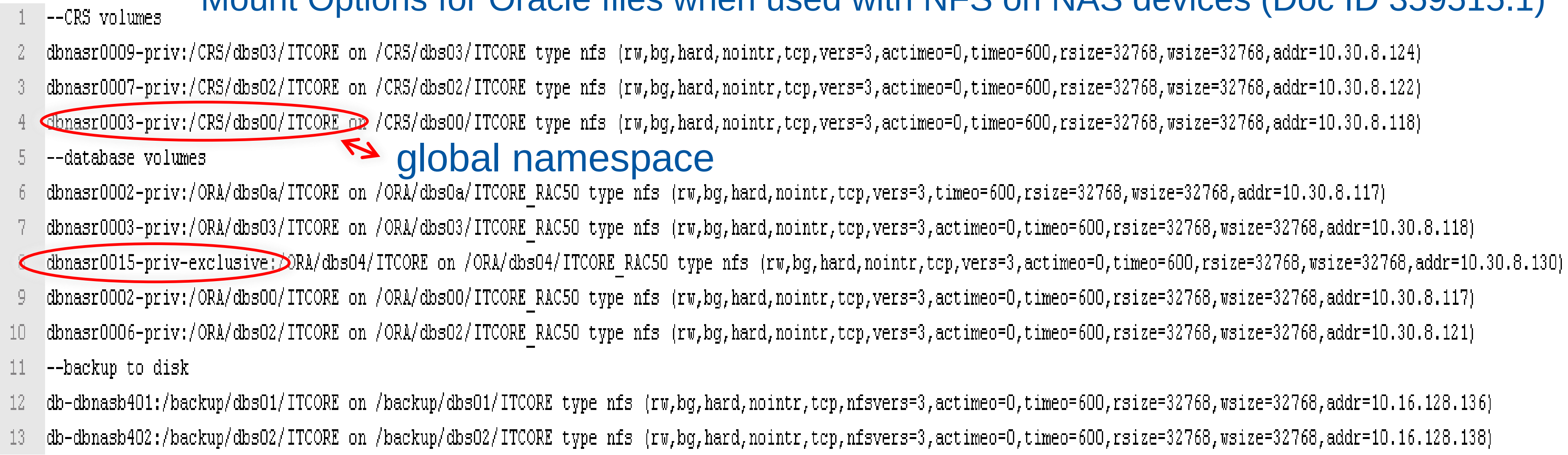
10GbE

gpn mtu=1500

12gbps

Oracle RAC database at least 10 file systems

Mount Options for Oracle files when used with NFS on NAS devices (Doc ID 359515.1)



```
--CRS volumes
dbnasr0009-priv:/CRS/dbs03/ITCORE on /CRS/dbs03/ITCORE type nfs (rw,bg,hard,nointr,tcp,vers=3,actimeo=0,timeo=600,rsize=32768,wsize=32768,addr=10.30.8.124)
dbnasr0007-priv:/CRS/dbs02/ITCORE on /CRS/dbs02/ITCORE type nfs (rw,bg,hard,nointr,tcp,vers=3,actimeo=0,timeo=600,rsize=32768,wsize=32768,addr=10.30.8.122)
dbnasr0003-priv:/CRS/dbs00/ITCORE on /CRS/dbs00/ITCORE type nfs (rw,bg,hard,nointr,tcp,vers=3,actimeo=0,timeo=600,rsize=32768,wsize=32768,addr=10.30.8.118)
--database volumes
dbnasr0002-priv:/ORA/dbs0a/ITCORE on /ORA/dbs0a/ITCORE_RAC50 type nfs (rw,bg,hard,nointr,tcp,vers=3,timeo=600,rsize=32768,wsize=32768,addr=10.30.8.117)
dbnasr0003-priv:/ORA/dbs03/ITCORE on /ORA/dbs03/ITCORE_RAC50 type nfs (rw,bg,hard,nointr,tcp,vers=3,actimeo=0,timeo=600,rsize=32768,wsize=32768,addr=10.30.8.118)
dbnasr0015-priv-exclusive:/ORA/dbs04/ITCORE on /ORA/dbs04/ITCORE_RAC50 type nfs (rw,bg,hard,nointr,tcp,vers=3,actimeo=0,timeo=600,rsize=32768,wsize=32768,addr=10.30.8.130)
dbnasr0002-priv:/ORA/dbs00/ITCORE on /ORA/dbs00/ITCORE_RAC50 type nfs (rw,bg,hard,nointr,tcp,vers=3,actimeo=0,timeo=600,rsize=32768,wsize=32768,addr=10.30.8.117)
dbnasr0006-priv:/ORA/dbs02/ITCORE on /ORA/dbs02/ITCORE_RAC50 type nfs (rw,bg,hard,nointr,tcp,vers=3,actimeo=0,timeo=600,rsize=32768,wsize=32768,addr=10.30.8.121)
--backup to disk
db-dbnasb401:/backup/dbs01/ITCORE on /backup/dbs01/ITCORE type nfs (rw,bg,hard,nointr,tcp,nfsvers=3,actimeo=0,timeo=600,rsize=32768,wsize=32768,addr=10.16.128.136)
db-dbnasb402:/backup/dbs02/ITCORE on /backup/dbs02/ITCORE type nfs (rw,bg,hard,nointr,tcp,nfsvers=3,actimeo=0,timeo=600,rsize=32768,wsize=32768,addr=10.16.128.138)
```

global namespace

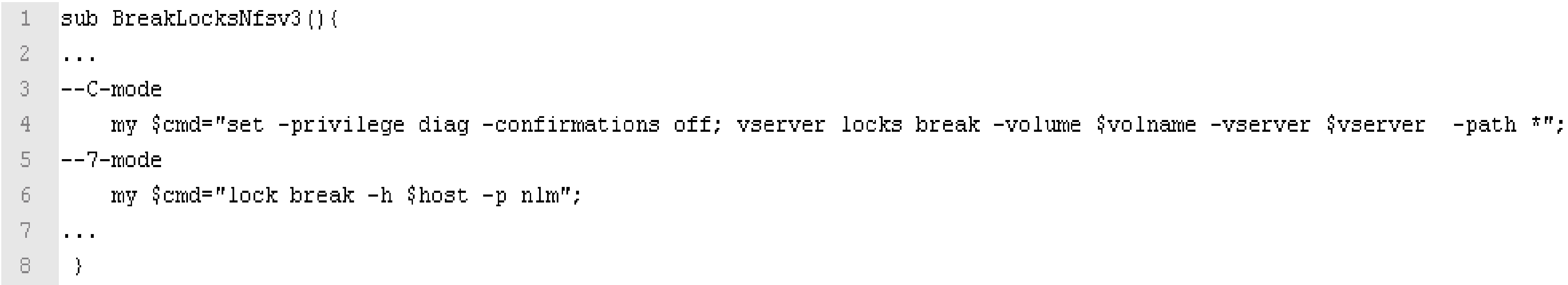
Oracle file systems

Mount point	Content
/ORA/dbs0a/${DB_UNIQUE_NAME}	ADR (including listener) /adump log files
/ORA/dbs00/${DB_UNIQUE_NAME}	Control File + copy of online redo logs
/ORA/dbs02/${DB_UNIQUE_NAME}	Control File + archive logs (FRA)
/ORA/dbs03/${DB_UNIQUE_NAME}*	Datafiles
/ORA/dbs04/${DB_UNIQUE_NAME}	Control File + copy of online redo logs + block change tracking file + spfile
/ORA/dbs0X/${DB_UNIQUE_NAME}*	More datafiles volumes if needed
/CRS/dbs00/${DB_UNIQUE_NAME}	Voting disk
/CRS/dbs02/${DB_UNIQUE_NAME}	Voting disk + OCR
/CRS/dbs00/${DB_UNIQUE_NAME}	Voting disk + OCR

* They are mounted using their own lif to ease volume movements within the cluster

MySQL/PostgreSQL

- Just two file systems on both cases:
  - data
  - binlogs (MySQL) or WALs (PostgreSQL)
- For instances running on an Oracle cluster ware, care must be taken in case of server crash for MySQL instances.
  - "InnoDB: Unable to lock ./ibdata1, error: 11" Error Sometimes Seen With MySQL on NFS (Doc ID 1522745.1)



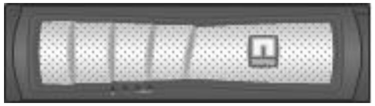
```
sub BreakLocksNfsv3(){
...
--C-mode
    my $cmd="set -privilege diag -confirmations off; vserver locks break -volume $volname -vserver $vserver  -path *";
--7-mode
    my $cmd="lock break -h $host -p nlm";
...
 }
```

Agenda

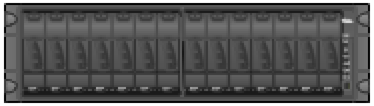
- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
    - Flash cache
    - Flash pool
- Data motion
- Snapshots
- Clonning in Oracle12c
- Backup to disk
- directNFS
- Monitoring
    - In-house tools
    - Netapp tools
- Conclusions

Netapp evolution at CERN (last 8 years)

FAS3000



100% FC disks



2gbps

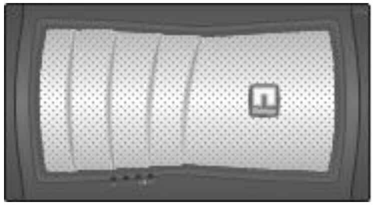
DS14 mk4 FC

Data ONTAP®
7-mode

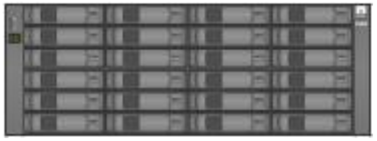
scaling up

scaling out

FAS6200 & FAS8000



Flash pool/cache = 100% SATA disk + SSD

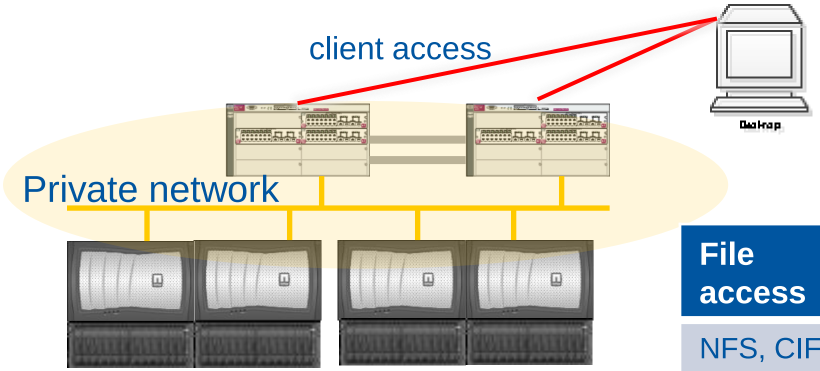


6gbps

DS4246

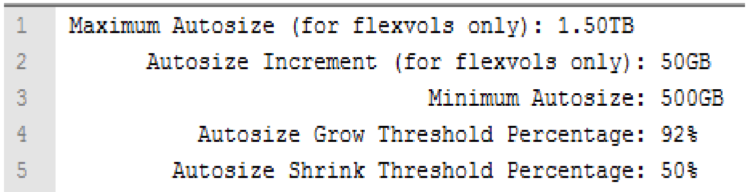
Data ONTAP®
Cluster-Mode

A few 7-mode concepts



autosupport

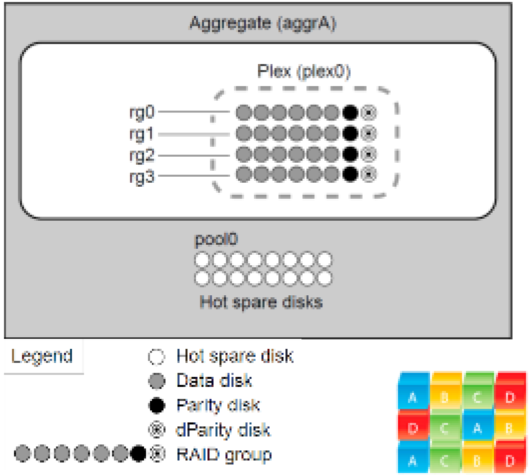
Thin provisioning



```
Maximum Autosize (for flexvols only): 1.50TB
      Autosize Increment (for flexvols only): 50GB
                           Minimum Autosize: 500GB
         Autosize Grow Threshold Percentage: 92%
       Autosize Shrink Threshold Percentage: 50%
```

File access	**Block access**
NFS, CIFS	FC,FCoE, iSCSI

raid_dp or raid4



FlexVolume

Remote **L**an **M**anager

Service **P**rocessor

raid.scrub.schedule

once weekly

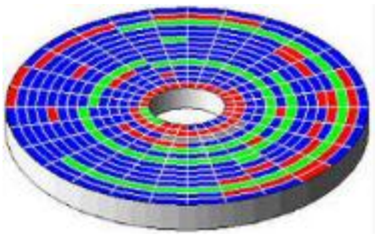
raid.media_scrub.rate

constantly

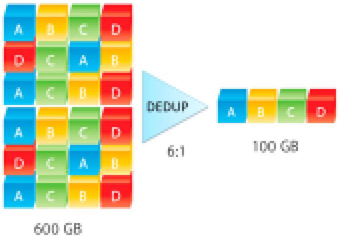
Rapid **RAID R**ecovery

Maintenance center
(at least 2 spares)

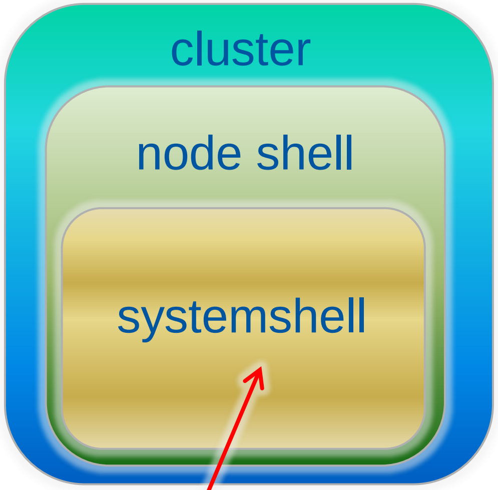
reallocate



ded



A few C-mode concepts

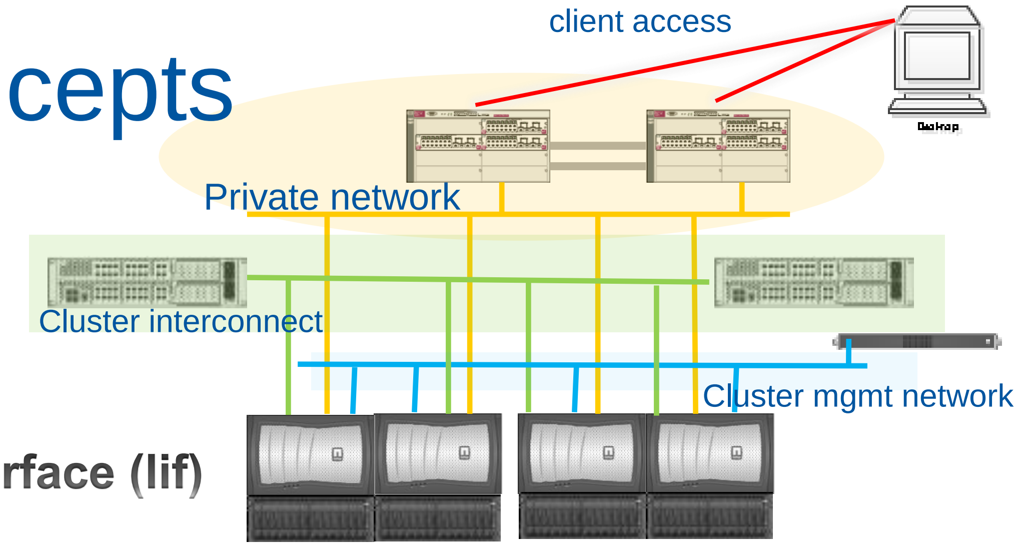

cluster

node shell

systemshell

C-mode

Logging files from the controller no longer accessible by simple NFS export


client access

Private network

Cluster interconnect

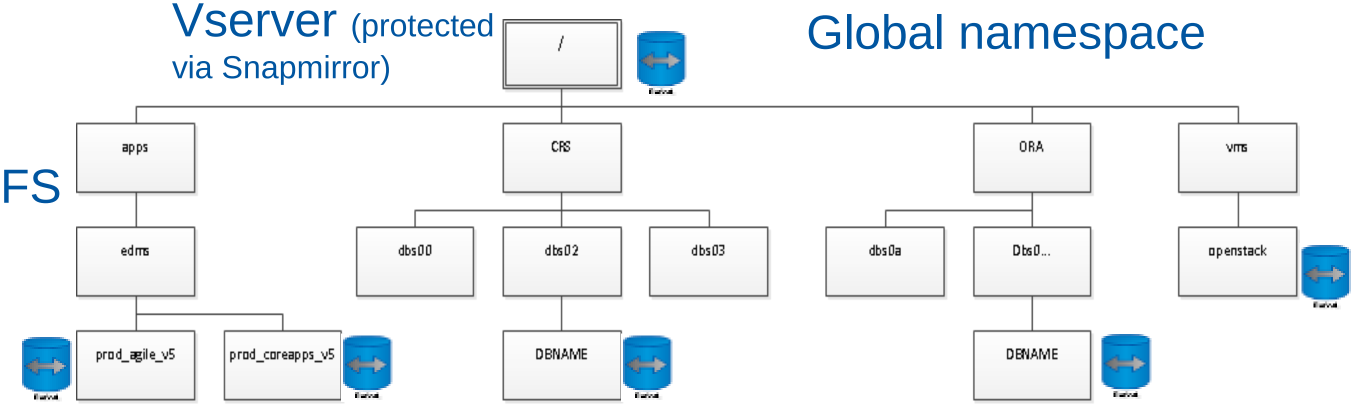
Cluster mgmt network

Logical Interface (lif)

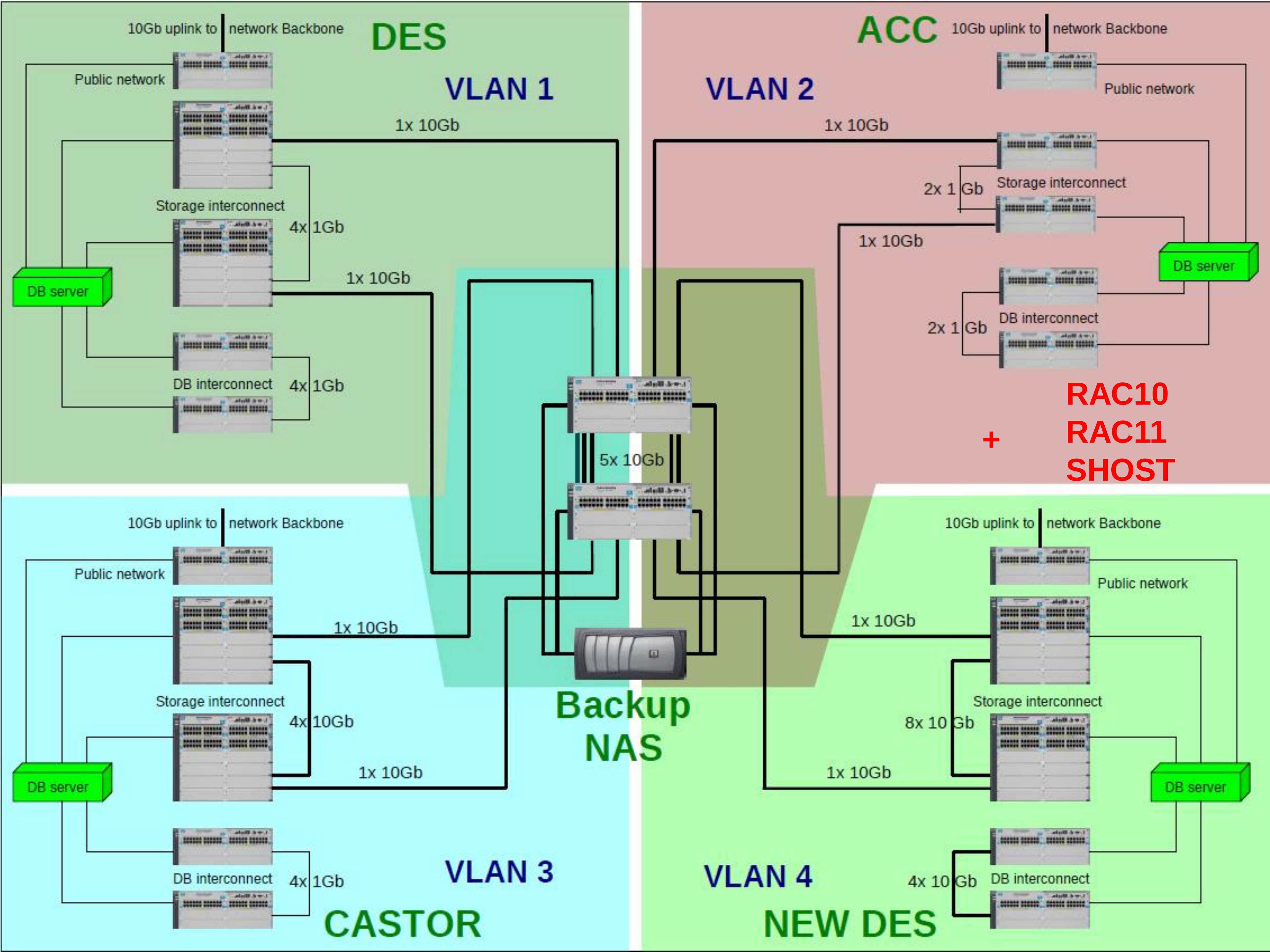
C-mode

`cluster ring show`

RDB: vifmgr + bcomd + vldb + mgmt

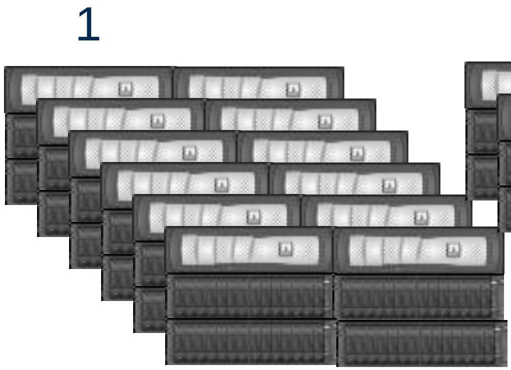

Vserver (protected via Snapmirror)

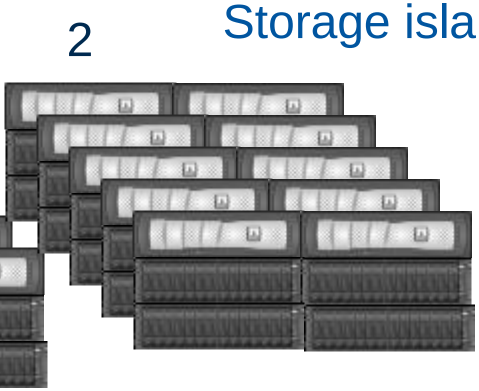
Global namespace


DES
10Gb uplink to network Backbone
Public network
VLAN 1
1x 10Gb
Storage interconnect
4x 1Gb
DB server
1x 10Gb
DB interconnect
4x 1Gb
ACC
10Gb uplink to network Backbone
Public network
VLAN 2
1x 10Gb
2x 1 Gb
Storage interconnect
1x 10Gb
DB server
DB interconnect
2x 1 Gb
RAC10
+
RAC11
SHOST
5x 10Gb
10Gb uplink to network Backbone
Public network
1x 10Gb
Storage interconnect
4x 10Gb
DB server
1x 10Gb
DB interconnect
4x 1Gb
VLAN 3
CASTOR
Backup
NAS
10Gb uplink to network Backbone
Public network
1x 10Gb
Storage interconnect
8x 10 Gb
1x 10Gb
DB server
4x 10 Gb
DB interconnect
VLAN 4
NEW DES

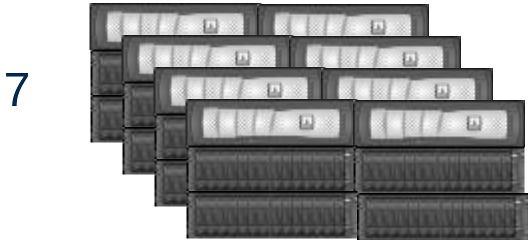
Consolidation

Storage islands, accessible via private network





. . .

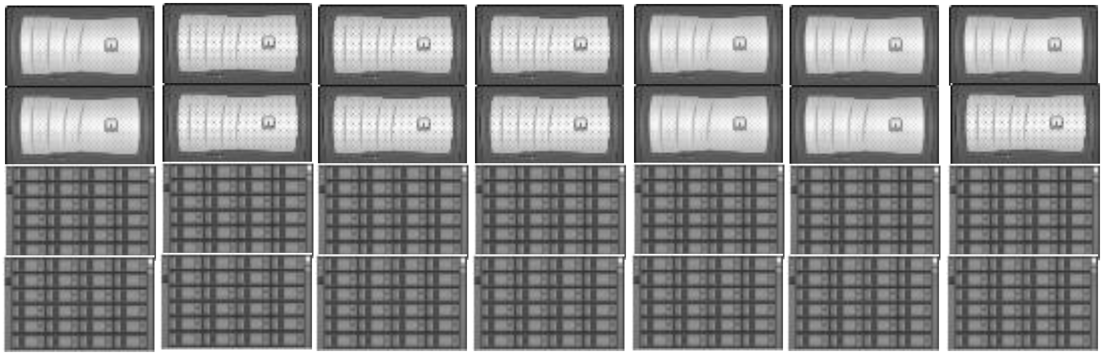


56 controllers (FAS3000) & 2300 disks (1400TB storage)

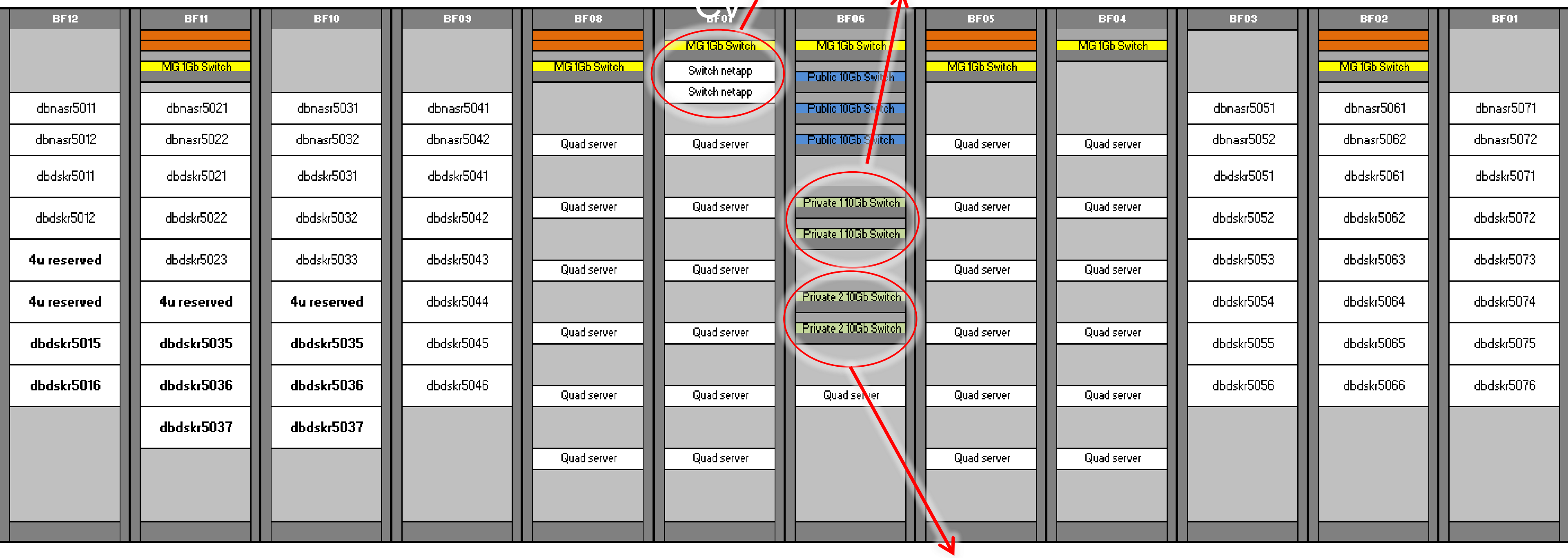
☺ Easy management

☹ Difficulties finding slots for interventions

14 controllers (FAS6220) &
960 disks (1660 TB storage)



RAC50 setup

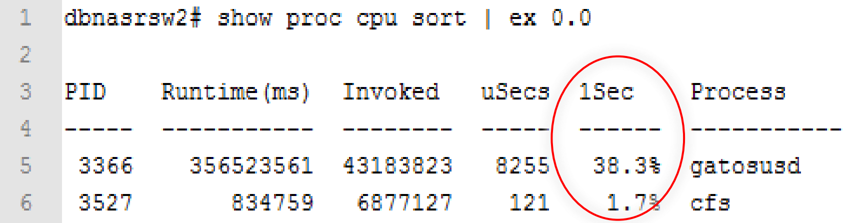


cluster interconnect

primary switch (private network)

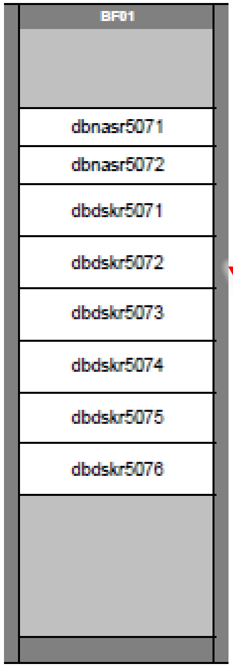
secondary switch (private network)

- Cluster interconnect, using FC gbic’s for distance longer than 5m.
- SFP must be from CISCO

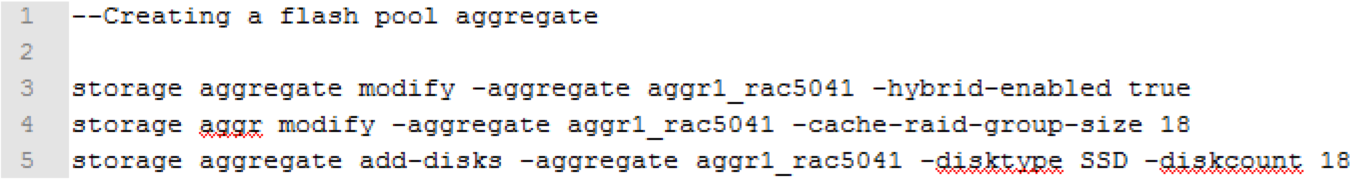


```
dbnasrsw2# show proc cpu sort | ex 0.0

PID     Runtime(ms)  Invoked   uSecs  1Sec    Process
-----   -----------  --------  -----  ------  -----------
 3366     356523561  43183823   8255   38.3%  gatosusd
 3527        834759   6877127    121    1.7%  cfs
```

Configuration details: disk shelves

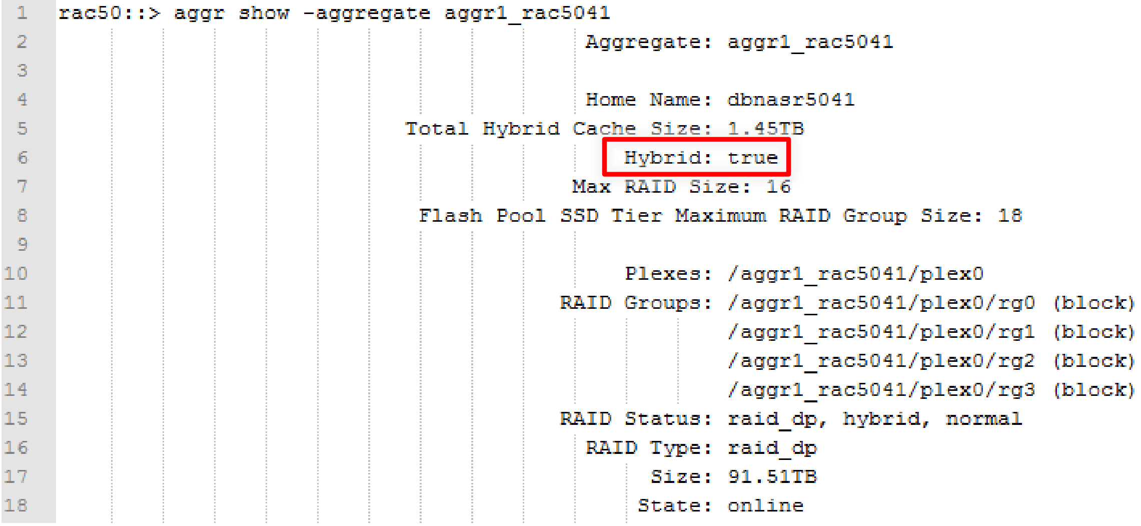


SSD enabled



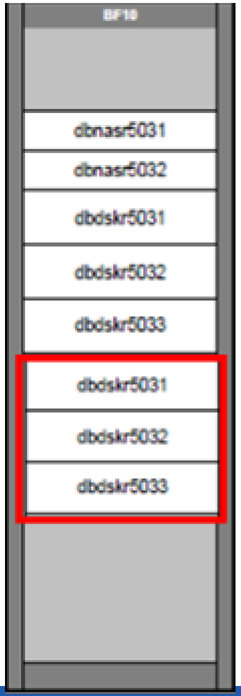
```
--Creating a flash pool aggregate

storage aggregate modify -aggregate aggr1_rac5041 -hybrid-enabled true
storage aggr modify -aggregate aggr1_rac5041 -cache-raid-group-size 18
storage aggregate add-disks -aggregate aggr1_rac5041 -disktype SSD -diskcount 18
```

3 raid groups of 16 disks + 1 SSD raid group of 18 disks



```
rac50::> aggr show -aggregate aggr1_rac5041
                                         Aggregate: aggr1_rac5041

                                         Home Name: dbnasr5041
                           Total Hybrid Cache Size: 1.45TB
                                            Hybrid: true
                                        Max RAID Size: 16
                            Flash Pool SSD Tier Maximum RAID Group Size: 18

                                            Plexes: /aggr1_rac5041/plex0
                                       RAID Groups: /aggr1_rac5041/plex0/rg0 (block)
                                                    /aggr1_rac5041/plex0/rg1 (block)
                                                    /aggr1_rac5041/plex0/rg2 (block)
                                                    /aggr1_rac5041/plex0/rg3 (block)
                                       RAID Status: raid_dp, hybrid, normal
                                         RAID Type: raid_dp
                                              Size: 91.51TB
                                             State: online
```

no SSD enabled



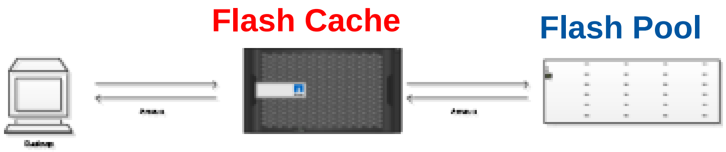
4 raid groups 16 disks, Total usable size: ~135TB

# of diskshelves	Size (TB)	Total
3xDS4246	91.51	732.08
3xDS4243	135.573	542.29

Agenda

- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
  - Flash cache
  - Flash pool
- Data motion
- Snapshots
- Clonning in Oracle12c
- Backup to disk
- directNFS
- Monitoring
  - In-house tools
  - Netapp tools
- Conclusions

Flash cache



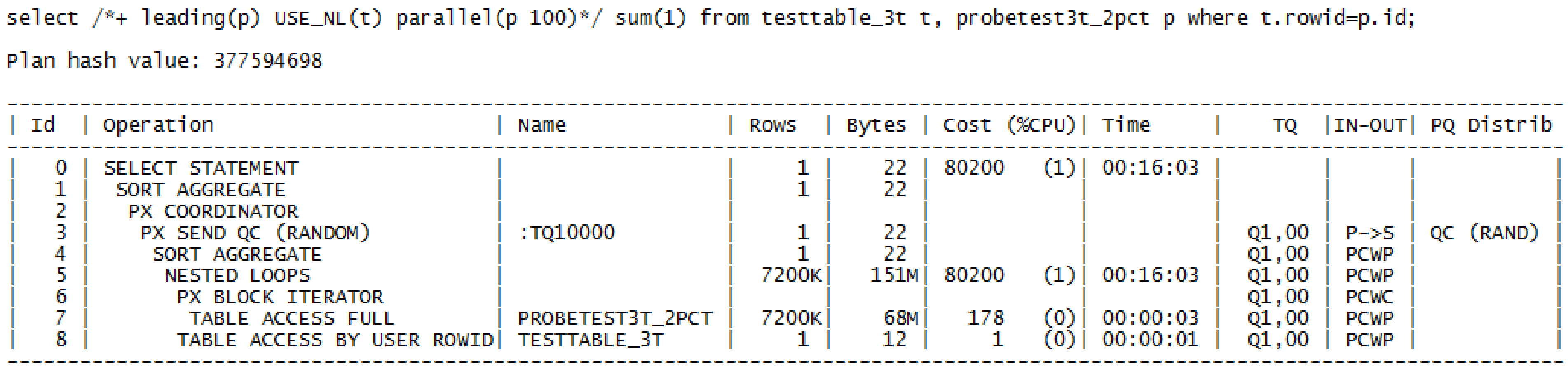
- Helps increase random IOPS on disks
- Warm-up effect (*options flexscale.rewarm*)
  - cf operations (takeover/giveback) invalidate the cache, user initiated ones do not since ONTAP 8.1
- TR-3832 :Flash Cache Best Practice Guide
- For databases
  - Decide what volumes to cache:

```
fas3240>priority on
fas3240>priority set volume volname cache=[reuse|keep]
```

  - `options flexscale.lopri_blocks off`

Flash cache: database benchmark

- Inner table (3TB) where a row = a block (8k). Outer table (2% of Inner table) each row contains rowid of inner table
- v$sysstat ‘physical reads’
  - Starts with db file sequential read but after a little while changes to db file parallel read



```
select /*+ leading(p) USE_NL(t) parallel(p 100)*/ sum(1) from testtable_3t t, probetest3t_2pct p where t.rowid=p.id;

Plan hash value: 377594698

-----------------------------------------------------------------------------------------------------------------------------
| Id  | Operation                    | Name             | Rows  | Bytes | Cost (%CPU)| Time     |    TQ  |IN-OUT| PQ Distrib |
-----------------------------------------------------------------------------------------------------------------------------
|   0 | SELECT STATEMENT             |                  |     1 |    22 | 80200   (1)| 00:16:03 |        |      |            |
|   1 |  SORT AGGREGATE              |                  |     1 |    22 |            |          |        |      |            |
|   2 |   PX COORDINATOR             |                  |       |       |            |          |        |      |            |
|   3 |    PX SEND QC (RANDOM)       | :TQ10000         |     1 |    22 |            |          |  Q1,00 | P->S | QC (RAND)  |
|   4 |     SORT AGGREGATE           |                  |     1 |    22 |            |          |  Q1,00 | PCWP |            |
|   5 |      NESTED LOOPS            |                  |  7200K|   151M| 80200   (1)| 00:16:03 |  Q1,00 | PCWP |            |
|   6 |       PX BLOCK ITERATOR      |                  |       |       |            |          |  Q1,00 | PCWC |            |
|   7 |        TABLE ACCESS FULL     | PROBETEST3T_2PCT |  7200K|    68M|   178   (0)| 00:00:03 |  Q1,00 | PCWP |            |
|   8 |       TABLE ACCESS BY USER ROWID| TESTTABLE_3T  |     1 |    12 |     1   (0)| 00:00:01 |  Q1,00 | PCWP |            |
-----------------------------------------------------------------------------------------------------------------------------
```

*fas3240, 32 disks SATA 2TB, Data Ontap 8.0.1, Oracle 11gR2

Random Read IOPS*	No PAM	PAM + Kernel NFS (RHE5)	PAM + dNFS
First run	2903	795	3827
Second run	2900	16397	37811

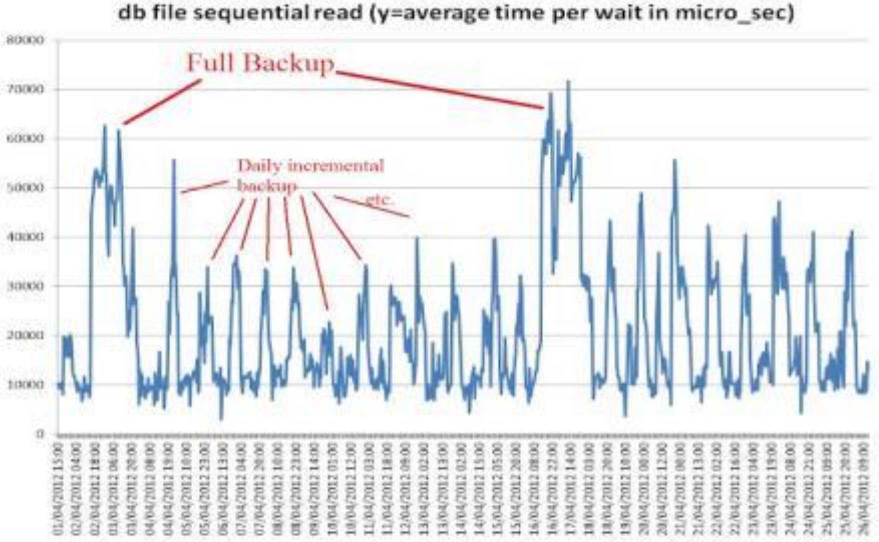
~160 data disks

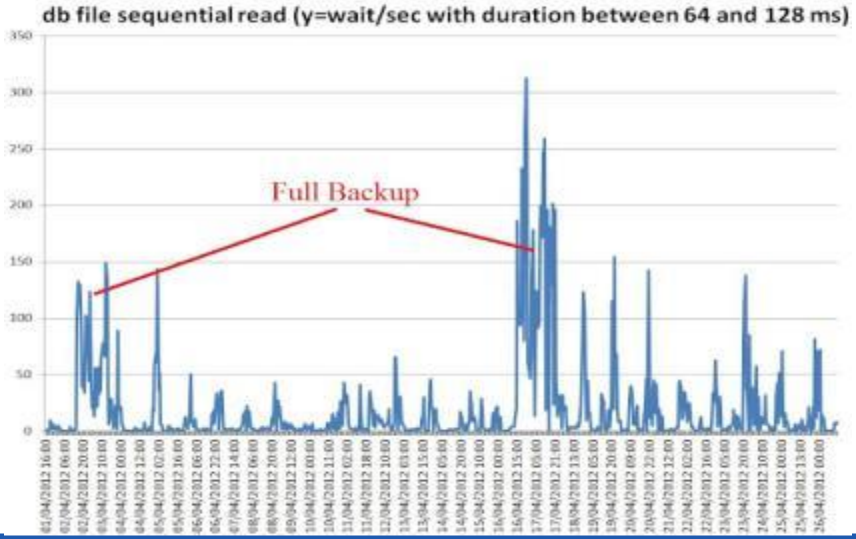
~365 data disks

Flash cache: long running backups…

- During backups SSD cache is flushed
- IO latency increases – hit% on PAM goes down ~ 1%
- Possible solutions:
  - Data Guard
  - **`priority set enabled_components=cache`**
  - Large IO windows to improve sequential IO detection, possible in C-mode:

```
vserver nfs modify -vserver vs1 -v3-tcp-max-read-size 1048576
```

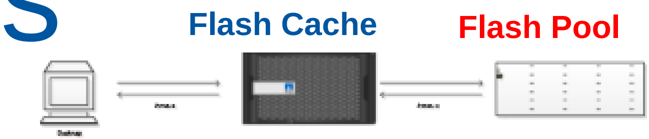

db file sequential read (y=average time per wait in micro_sec)


db file sequential read (y=wait/sec with duration between 64 and 128 ms)

Agenda

- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
  - Flash cache
  - Flash pool
- Data motion
- Snapshots
- Clonning in Oracle12c
- Backup to disk
- directNFS
- Monitoring
  - In-house tools
  - Netapp tools
- Conclusions

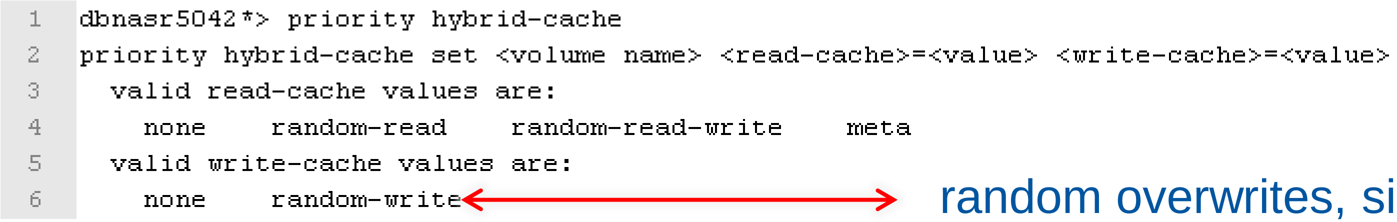
Flash pool aggregates



- 64 bits aggregates
- Aggregate with snapshots, they must be deleted before converting into hybrid aggregate
- SSD rules: minimum number and extensions depending on the model e.g. FAS6000 9+2, 6 (with 100GB SSD)
- No mixed type of disks in a hybrid aggregate: just SAS + SSD, FC + SSD, SATA + SSD. No mixed type of disks in a raid_gp.
- You can combine different protection levels among SSD RAID and HDD RAID, e.g. raid_dp or raid4
- Hybrid aggregate can not be rollbacked
- If SSD raid_gps are not available the whole aggregate is down
- SSD raid_gps doesn’t count in total aggregate space
- Maximum SSD size depending on model & ONTAP release (https://hwu.netapp.com/Controller/Index).
- TR-4070: Flash Pool Design and Implementation Guide

Flash pool behaviour

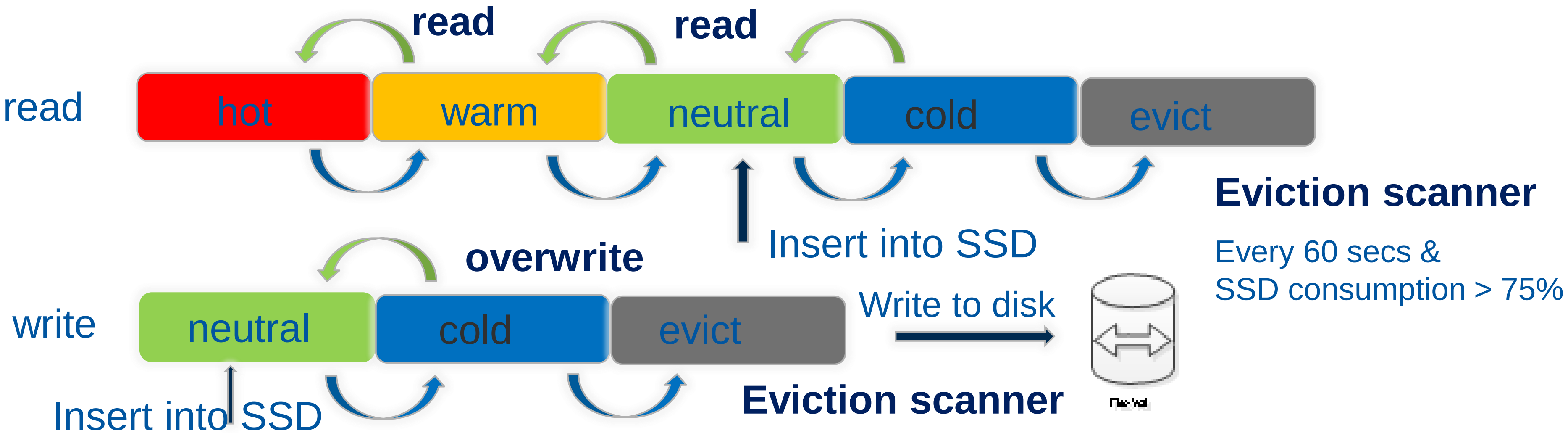
- Blocks going into SSD determined by Write and Read policies. They apply to volumes or globally on whole aggregate.



```
dbnasr5042*> priority hybrid-cache
priority hybrid-cache set <volume name> <read-cache>=<value> <write-cache>=<value>
  valid read-cache values are:
    none    random-read    random-read-write    meta
  valid write-cache values are:
    none    random-write
```

random overwrites, size < 16Kb

- Sequential data is not cached. Data cannot be pinned
- Heat map in order to decide what stays and for how long in SSD cache



read

read

read

hot

warm

neutral

cold

evict

Insert into SSD

Eviction scanner

Every 60 secs &
SSD consumption > 75%

overwrite

write

neutral

cold

evict

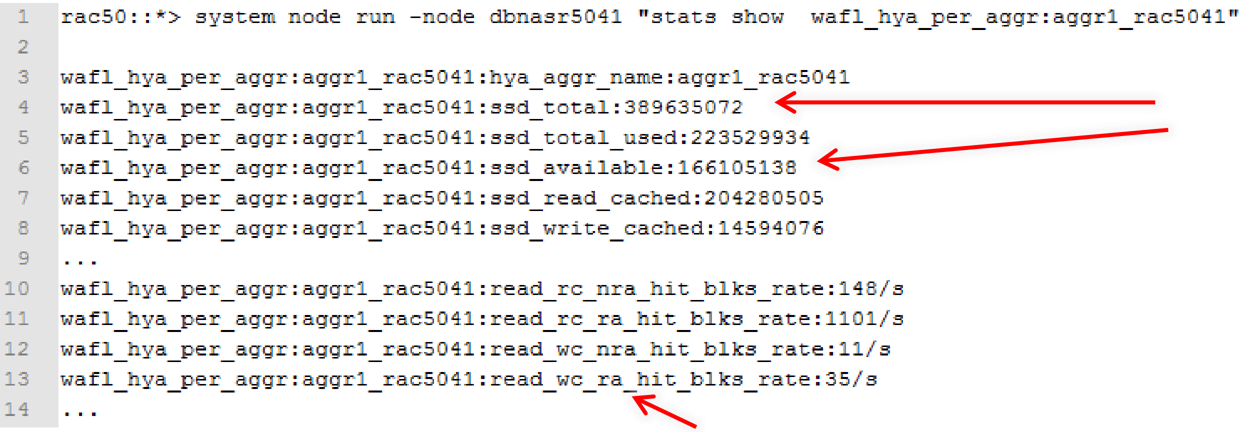
Write to disk

Insert into SSD

Eviction scanner

Flash pool: performance counters

- Performance counters: `wafl_hya_per_aggr (299)` & `wafl_hya_per_vvol (16)`

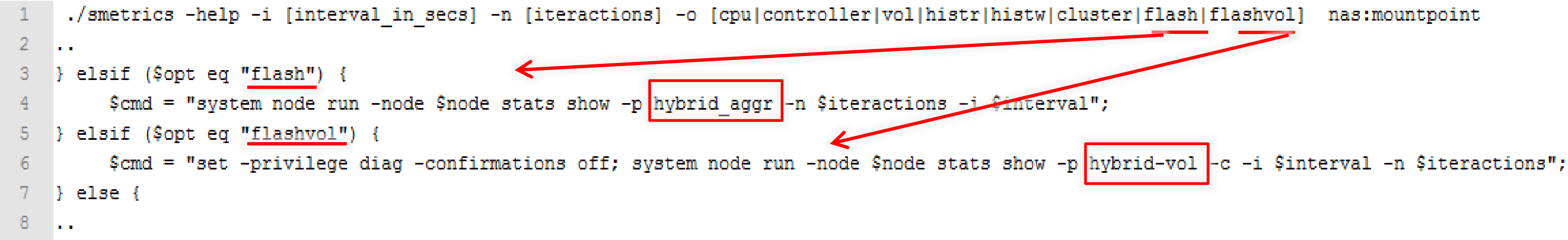


```
rac50::*> system node run -node dbnasr5041 "stats show  wafl_hya_per_aggr:aggr1_rac5041"

wafl_hya_per_aggr:aggr1_rac5041:hya_aggr_name:aggr1_rac5041
wafl_hya_per_aggr:aggr1_rac5041:ssd_total:389635072
wafl_hya_per_aggr:aggr1_rac5041:ssd_total_used:223529934
wafl_hya_per_aggr:aggr1_rac5041:ssd_available:166105138
wafl_hya_per_aggr:aggr1_rac5041:ssd_read_cached:204280505
wafl_hya_per_aggr:aggr1_rac5041:ssd_write_cached:14594076
...
wafl_hya_per_aggr:aggr1_rac5041:read_rc_nra_hit_blks_rate:148/s
wafl_hya_per_aggr:aggr1_rac5041:read_rc_ra_hit_blks_rate:1101/s
wafl_hya_per_aggr:aggr1_rac5041:read_wc_nra_hit_blks_rate:11/s
wafl_hya_per_aggr:aggr1_rac5041:read_wc_ra_hit_blks_rate:35/s
...
```

Around 25% difference in an empty system:
Ensures enough pre-erased blocks to write new data

Read-ahead caching algorithms

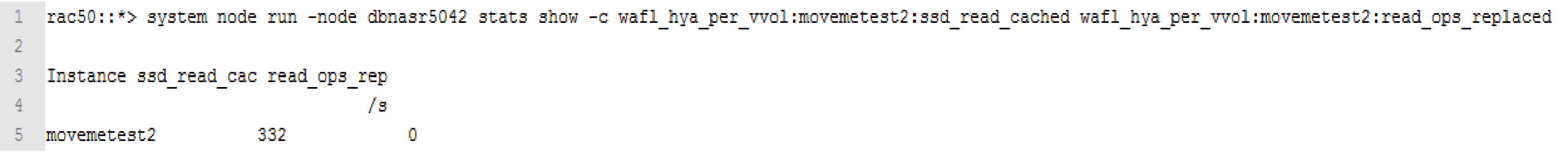
- We have automated the way to query those:



```
 ./smetrics -help -i [interval_in_secs] -n [iteractions] -o [cpu|controller|vol|histr|histw|cluster|flash|flashvol]  nas:mountpoint
..
} elsif ($opt eq "flash") {
     $cmd = "system node run -node $node stats show -p hybrid_aggr -n $iteractions -i $interval";
} elsif ($opt eq "flashvol") {
     $cmd = "set -privilege diag -confirmations off; system node run -node $node stats show -p hybrid-vol -c -i $interval -n $iteractions";
} else {
..
```

Monitoring: selecting counters

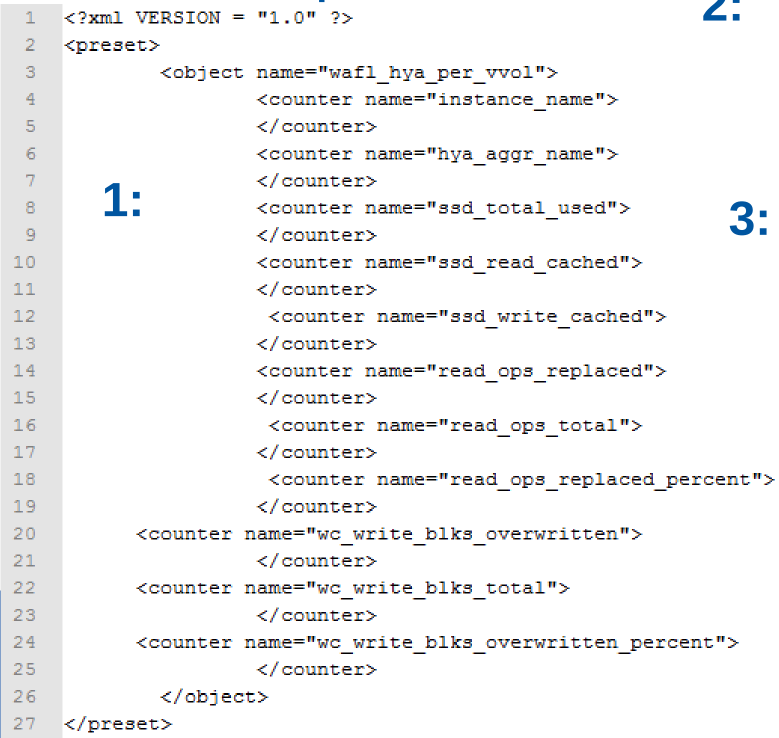
- Ontap 8.2: 37 objects, ~1230 counters
- Viewing the ones you are interested in from CLI can be cumbersome



```
rac50::*> system node run -node dbnasr5042 stats show -c wafl_hya_per_vvol:movemetest2:ssd_read_cached wafl_hya_per_vvol:movemetest2:read_ops_replaced

Instance ssd_read_cac read_ops_rep
                                 /s
movemetest2          332           0
```

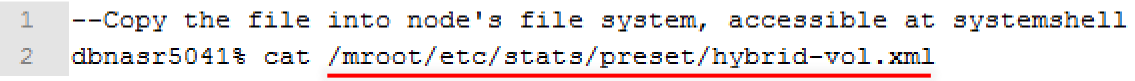
- Use a “preset”:

1:



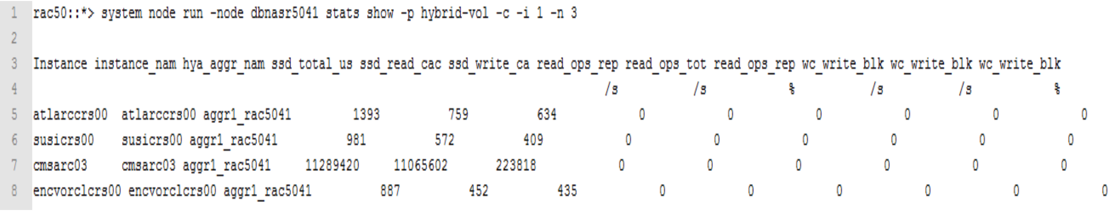
```
<?xml VERSION = "1.0" ?>
<preset>
        <object name="wafl_hya_per_vvol">
                <counter name="instance_name">
                </counter>
                <counter name="hya_aggr_name">
                </counter>
                <counter name="ssd_total_used">
                </counter>
                <counter name="ssd_read_cached">
                </counter>
                 <counter name="ssd_write_cached">
                </counter>
                <counter name="read_ops_replaced">
                </counter>
                 <counter name="read_ops_total">
                </counter>
                 <counter name="read_ops_replaced_percent">
                </counter>
      <counter name="wc_write_blks_overwritten">
                </counter>
      <counter name="wc_write_blks_total">
                </counter>
      <counter name="wc_write_blks_overwritten_percent">
                </counter>
        </object>
</preset>
```

2:



```
--Copy the file into node's file system, accessible at systemshell
dbnasr5041% cat /mroot/etc/stats/preset/hybrid-vol.xml
```

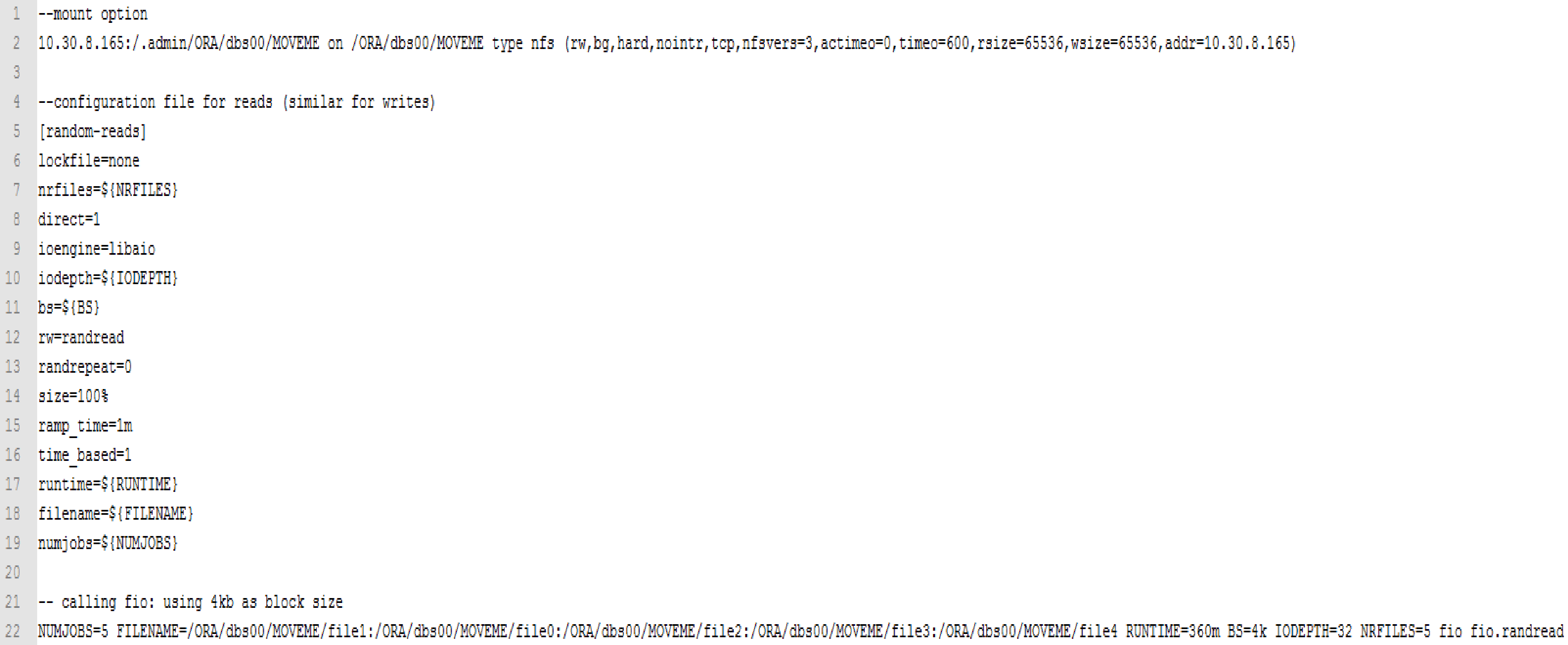
3: rac50::*> system node run -node dbnasr5041 stats show -p hybrid-vol -c -i 1 -n 3



```
rac50::*> system node run -node dbnasr5041 stats show -p hybrid-vol -c -i 1 -n 3

Instance instance_nam hya_aggr_nam ssd_total_us ssd_read_cac ssd_write_ca read_ops_rep read_ops_tot read_ops_rep wc_write_blk wc_write_blk wc_write_blk
                                                                                    /s           /s            %           /s           /s            %
atlarccrs00  atlarccrs00 aggr1_rac5041         1393          759          634            0            0            0            0            0            0
susicrs00    susicrs00 aggr1_rac5041          981          572          409            0            0            0            0            0            0
cmsarc03     cmsarc03 aggr1_rac5041     11289420     11065602       223818            0            0            0            0            0            0
encvorclcrs00 encvorclcrs00 aggr1_rac5041          887          452          435            0            0            0            0            0            0
```

Flash pool behaviour (II)

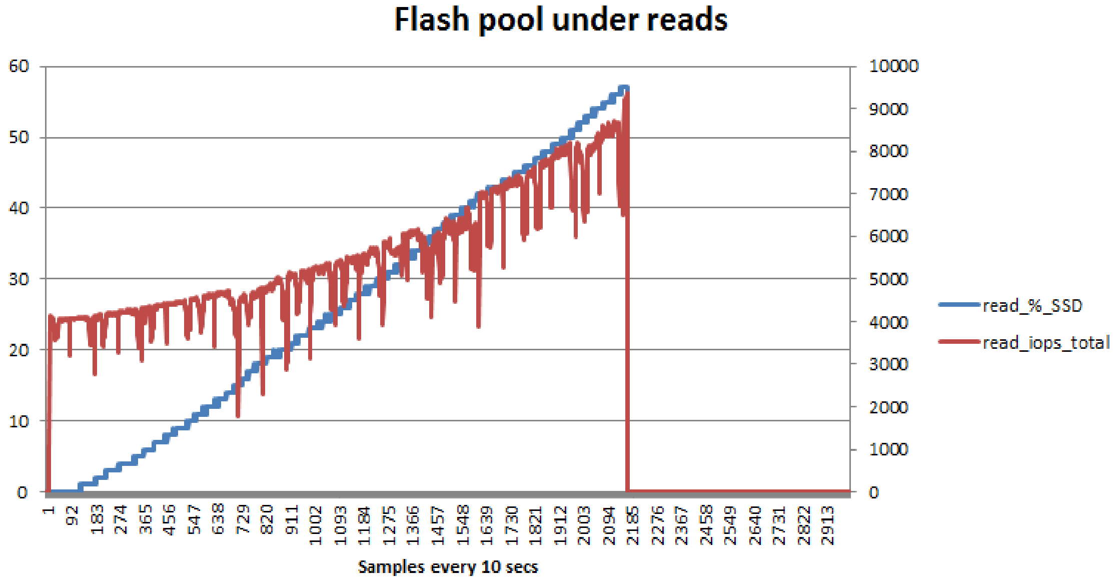
- fio (http://freecode.com/projects/fio) on rhe6.5
- 5x100GB files
- Example of a random read job. Jobs run for 6h.

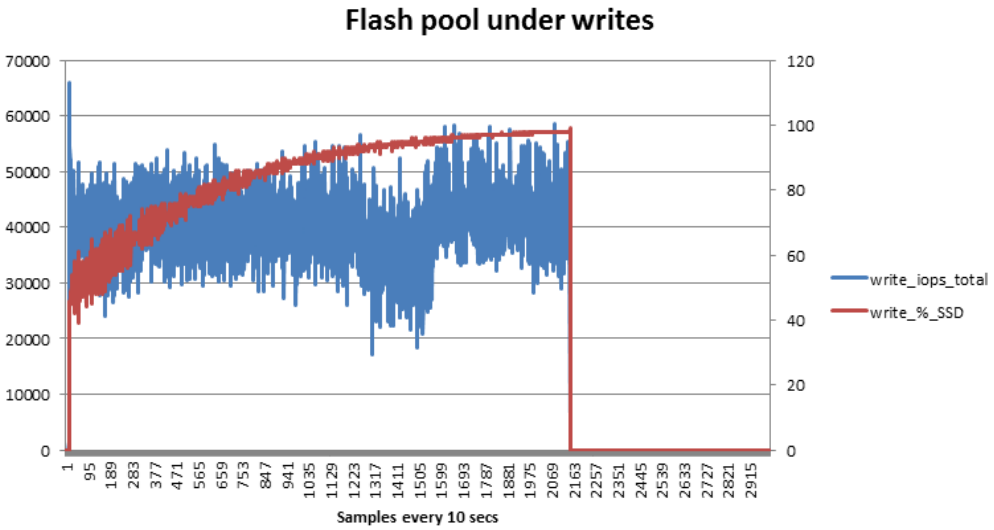


```
--mount option
10.30.8.165:/.admin/ORA/dbs00/MOVEME on /ORA/dbs00/MOVEME type nfs (rw,bg,hard,nointr,tcp,nfsvers=3,actimeo=0,timeo=600,rsize=65536,wsize=65536,addr=10.30.8.165)

--configuration file for reads (similar for writes)
[random-reads]
lockfile=none
nrfiles=${NRFILES}
direct=1
ioengine=libaio
iodepth=${IODEPTH}
bs=${BS}
rw=randread
randrepeat=0
size=100%
ramp_time=1m
time_based=1
runtime=${RUNTIME}
filename=${FILENAME}
numjobs=${NUMJOBS}

-- calling fio: using 4kb as block size
NUMJOBS=5 FILENAME=/ORA/dbs00/MOVEME/file1:/ORA/dbs00/MOVEME/file0:/ORA/dbs00/MOVEME/file2:/ORA/dbs00/MOVEME/file3:/ORA/dbs00/MOVEME/file4 RUNTIME=360m BS=4k IODEPTH=32 NRFILES=5 fio fio.randread
```

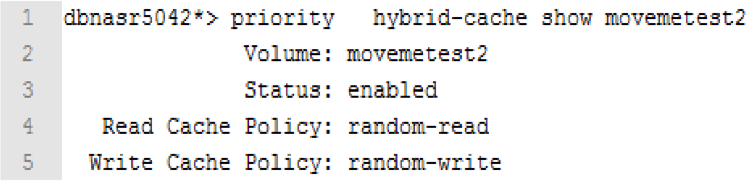
Flash pool behaviour (III)

- Read cache warms slower than write cache
  - reads costs more than writes, ~10 factor.
- After 6hours:
  - 300GB read cache
  - 500GB write cache



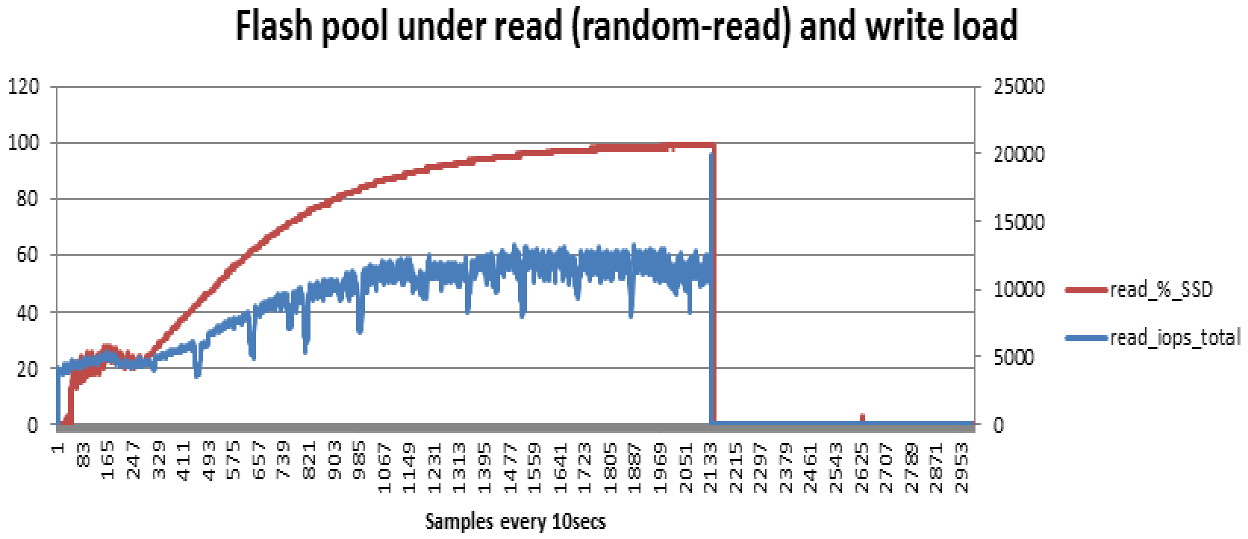


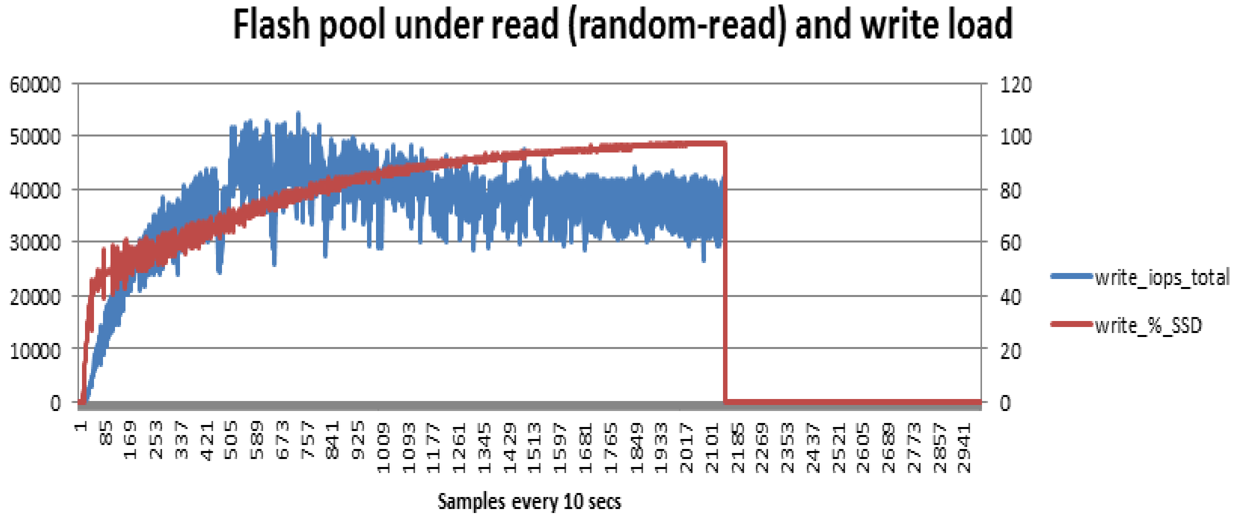
- Stats of SSD consumption can be retrieved using: `wafl_hya_per_vvol` object, at nodeshell in diagnostic level.



```
dbnasr5042*> priority   hybrid-cache show movemetest2
                        Volume: movemetest2
                        Status: enabled
             Read Cache Policy: random-read
            Write Cache Policy: random-write
```

Flash pool behaviour (IV)


Flash pool under read (random-read) and write load


Flash pool under read (random-read) and write load

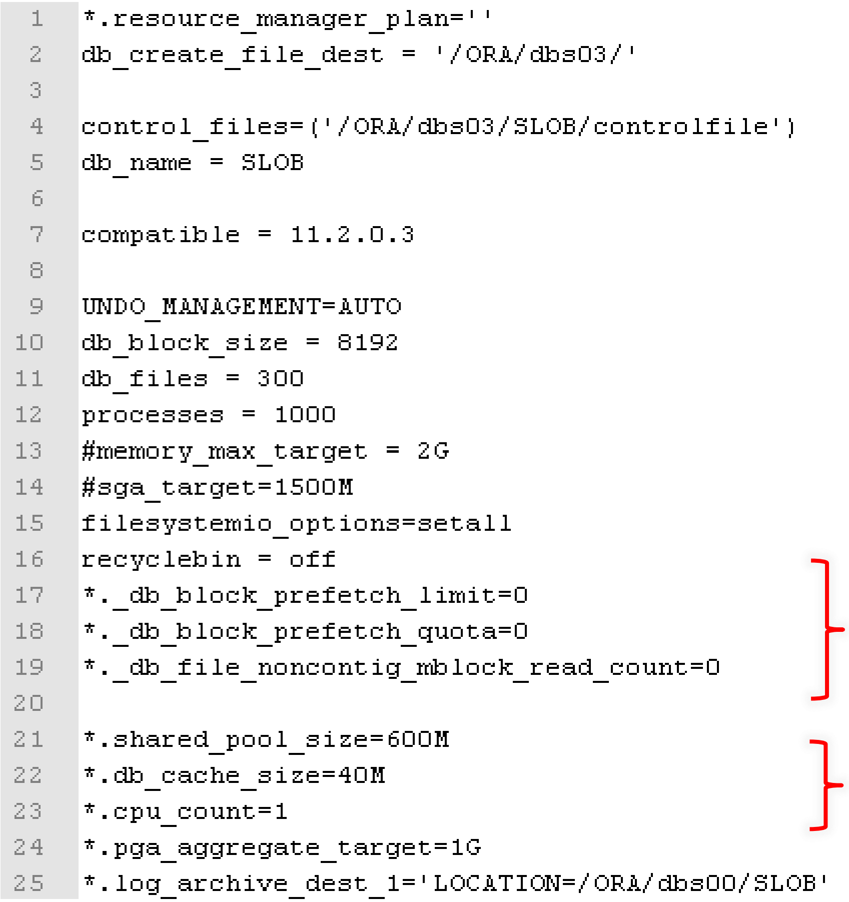
- SSD consumption:
  - 85GB read SSD
  - 493GB write SSD

- Write cache also used for reading → read_%_SSD ~100%
- Not much difference with this workload between: `random-read` & `random-read-write` policies

Test environment

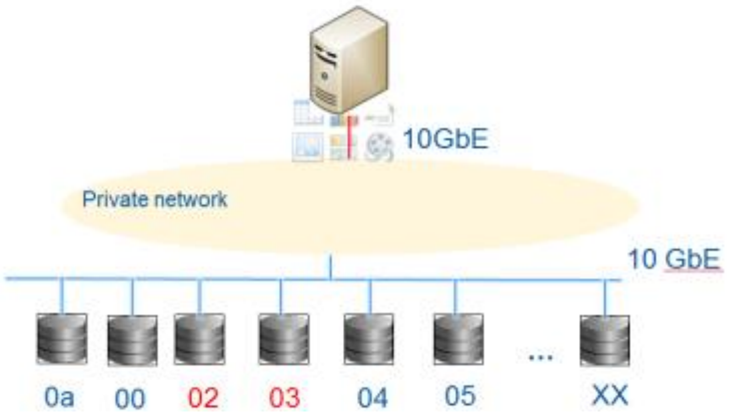
- Testing on a private network
  - Red Hat Enterprise Linux Server release 6.4
  - 16 cores - Intel(R) Xeon(R) CPU E5-2650 0 @ 2.00GHz
  - 128 GB RAM
- Oracle server single instance: 11.2.0.3
- Using SLOB2
- The following graphs were done with a dataset of 1TB

init.ora for testing with SLOB2



```
*.resource_manager_plan=''
db_create_file_dest = '/ORA/dbs03/'

control_files=('/ORA/dbs03/SLOB/controlfile')
db_name = SLOB

compatible = 11.2.0.3

UNDO_MANAGEMENT=AUTO
db_block_size = 8192
db_files = 300
processes = 1000
#memory_max_target = 2G
#sga_target=1500M
filesystemio_options=setall
recyclebin = off
*._db_block_prefetch_limit=0
*._db_block_prefetch_quota=0
*._db_file_noncontig_mblock_read_count=0

*.shared_pool_size=600M
*.db_cache_size=40M
*.cpu_count=1
*.pga_aggregate_target=1G
*.log_archive_dest_1='LOCATION=/ORA/dbs00/SLOB'
```

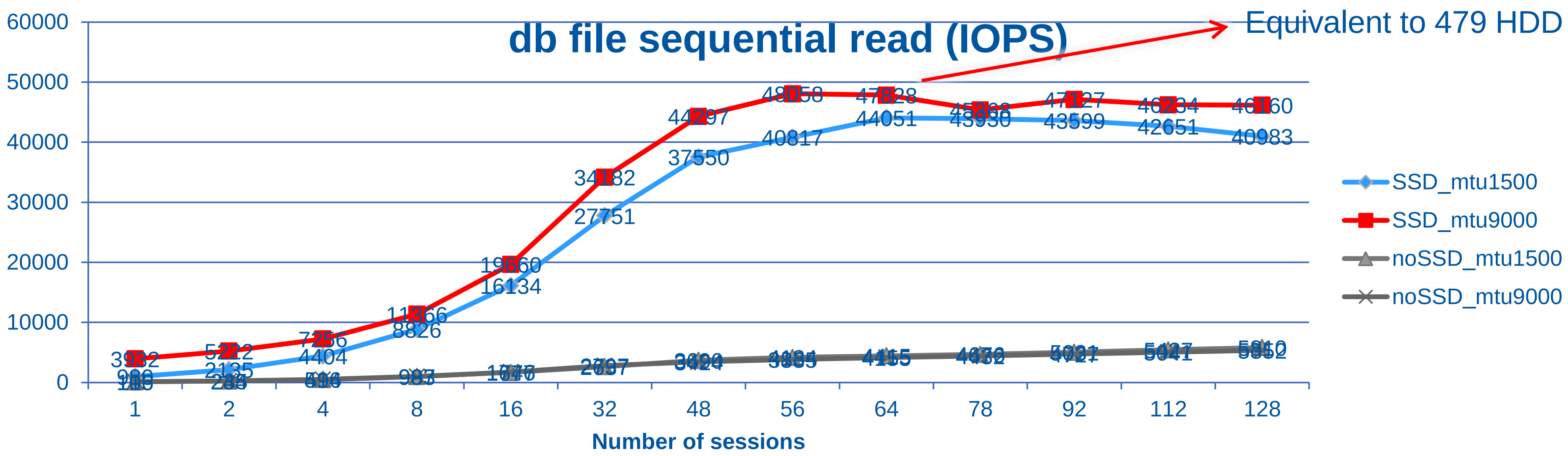
Disable scheduler and resource manager. MOS 786346.1



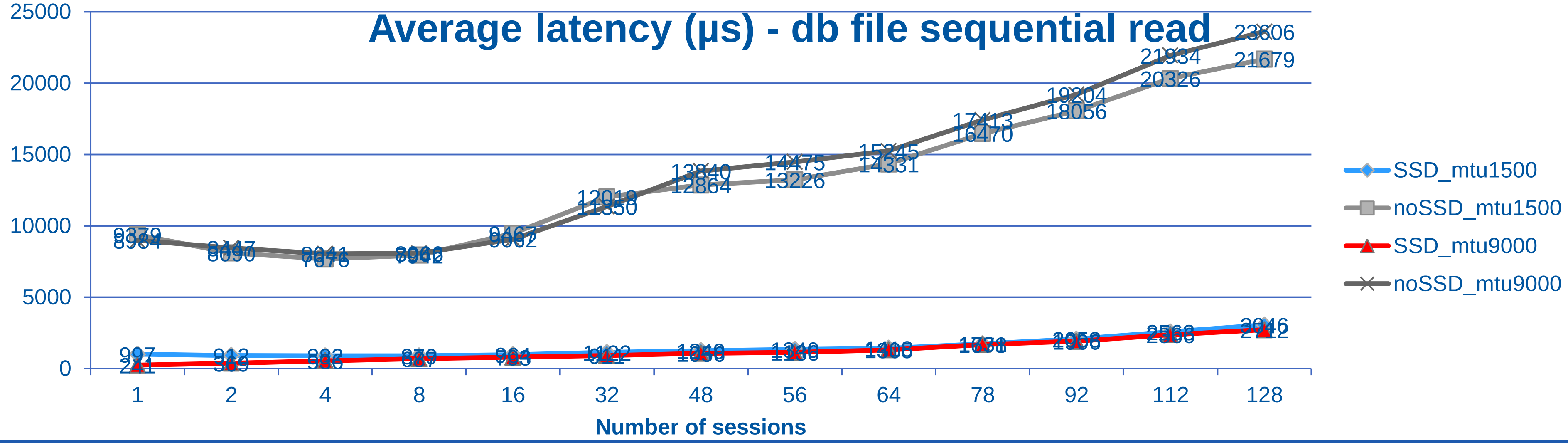
Avoid “db file parallel read” optimization

Small db_cache_size to force IO on storage

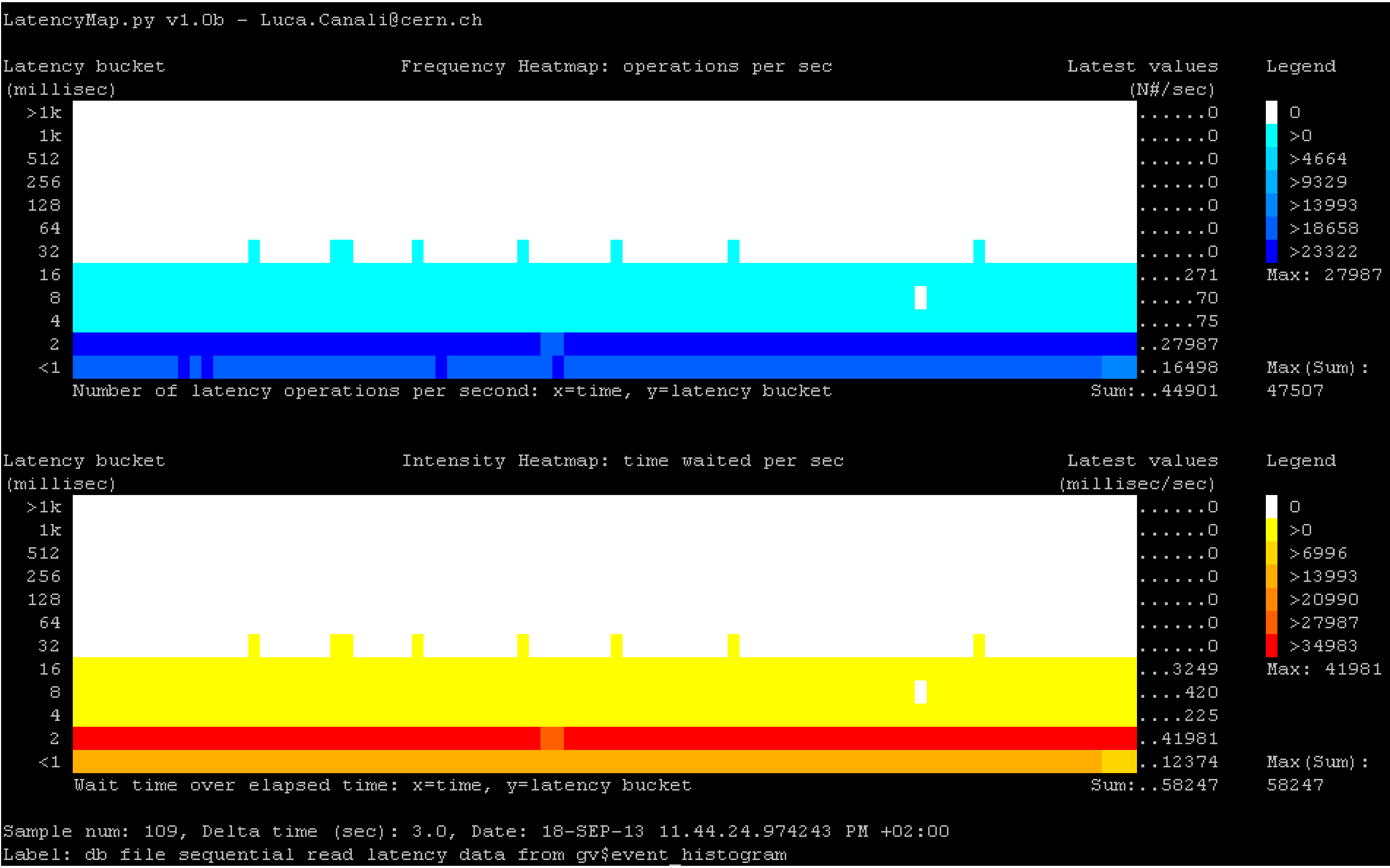
Random Reads



Equivalent to 479 HDD

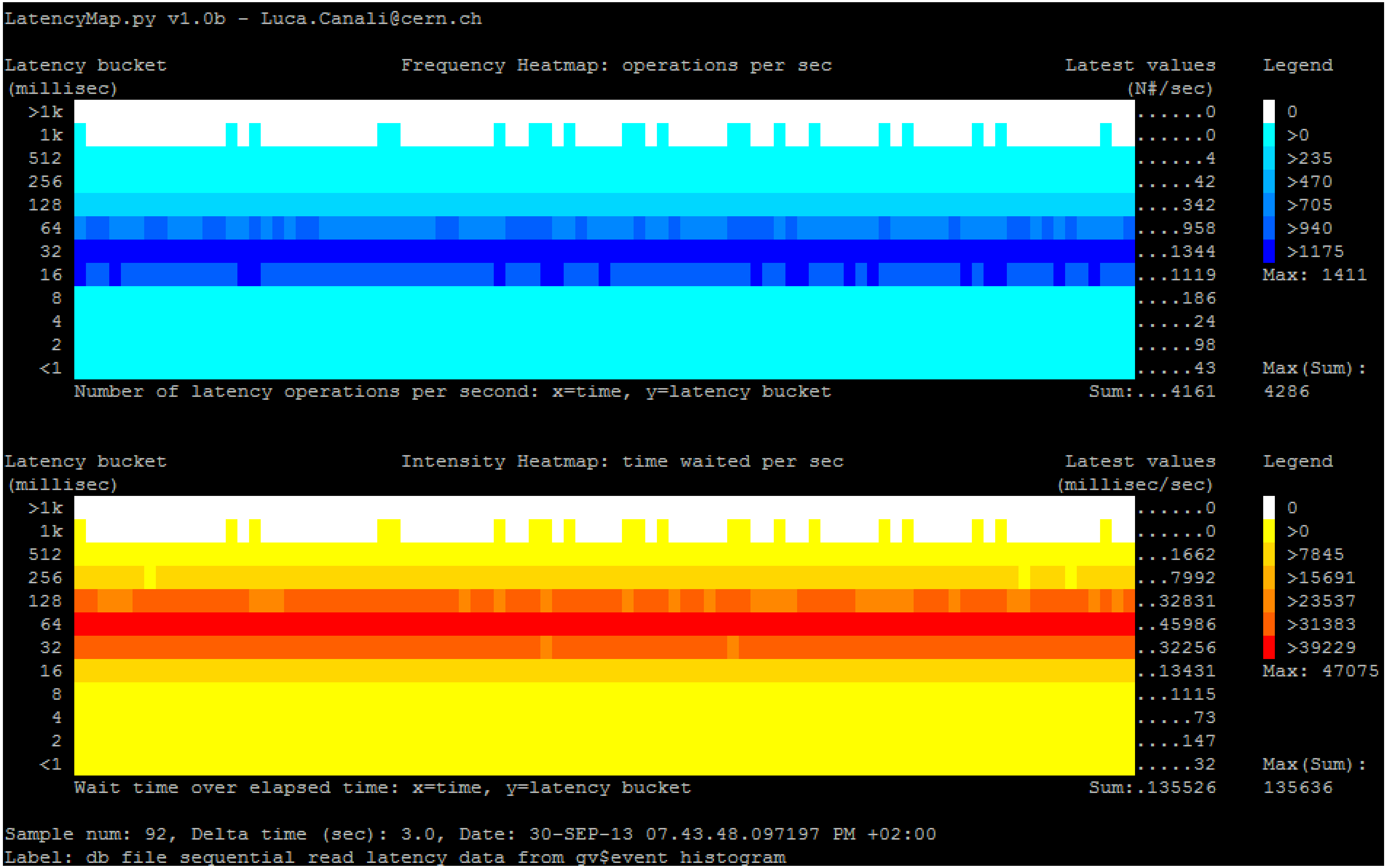


1TB dataset, 100% in SSD, 56 sessions, random reads



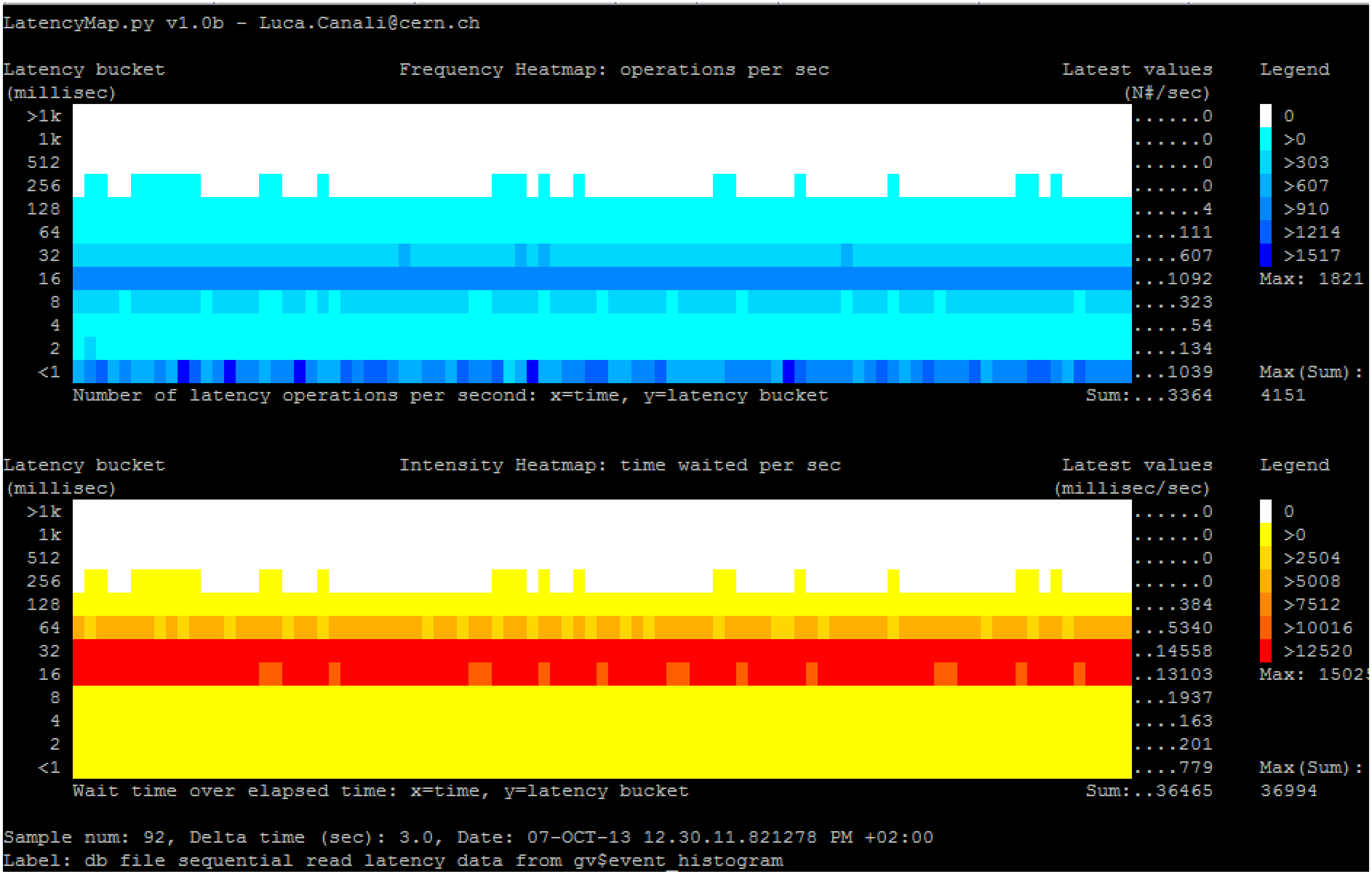
```
LatencyMap.py v1.0b - Luca.Canali@cern.ch

Latency bucket                 Frequency Heatmap: operations per sec                    Latest values    Legend
(millisec)                                                                                   (N#/sec)
  >1k                                                                                  ......0          0
   1k                                                                                  ......0          >0
  512                                                                                  ......0          >4664
  256                                                                                  ......0          >9329
  128                                                                                  ......0          >13993
   64                                                                                  ......0          >18658
   32                                                                                  ......0          >23322
   16                                                                                  ....271          Max: 27987
    8                                                                                  .....70
    4                                                                                  .....75
    2                                                                                  ..27987
   <1                                                                                  ..16498          Max(Sum):
     Number of latency operations per second: x=time, y=latency bucket                 Sum:..44901      47507


Latency bucket                 Intensity Heatmap: time waited per sec                   Latest values    Legend
(millisec)                                                                             (millisec/sec)
  >1k                                                                                  ......0          0
   1k                                                                                  ......0          >0
  512                                                                                  ......0          >6996
  256                                                                                  ......0          >13993
  128                                                                                  ......0          >20990
   64                                                                                  ......0          >27987
   32                                                                                  ......0          >34983
   16                                                                                  ...3249          Max: 41981
    8                                                                                  ....420
    4                                                                                  ....225
    2                                                                                  ..41981
   <1                                                                                  ..12374          Max(Sum):
     Wait time over elapsed time: x=time, y=latency bucket                             Sum:..58247      58247

Sample num: 109, Delta time (sec): 3.0, Date: 18-SEP-13 11.44.24.974243 PM +02:00
Label: db file sequential read latency data from gv$event_histogram
```

10TB dataset, 128 sessions, random reads, disk saturation



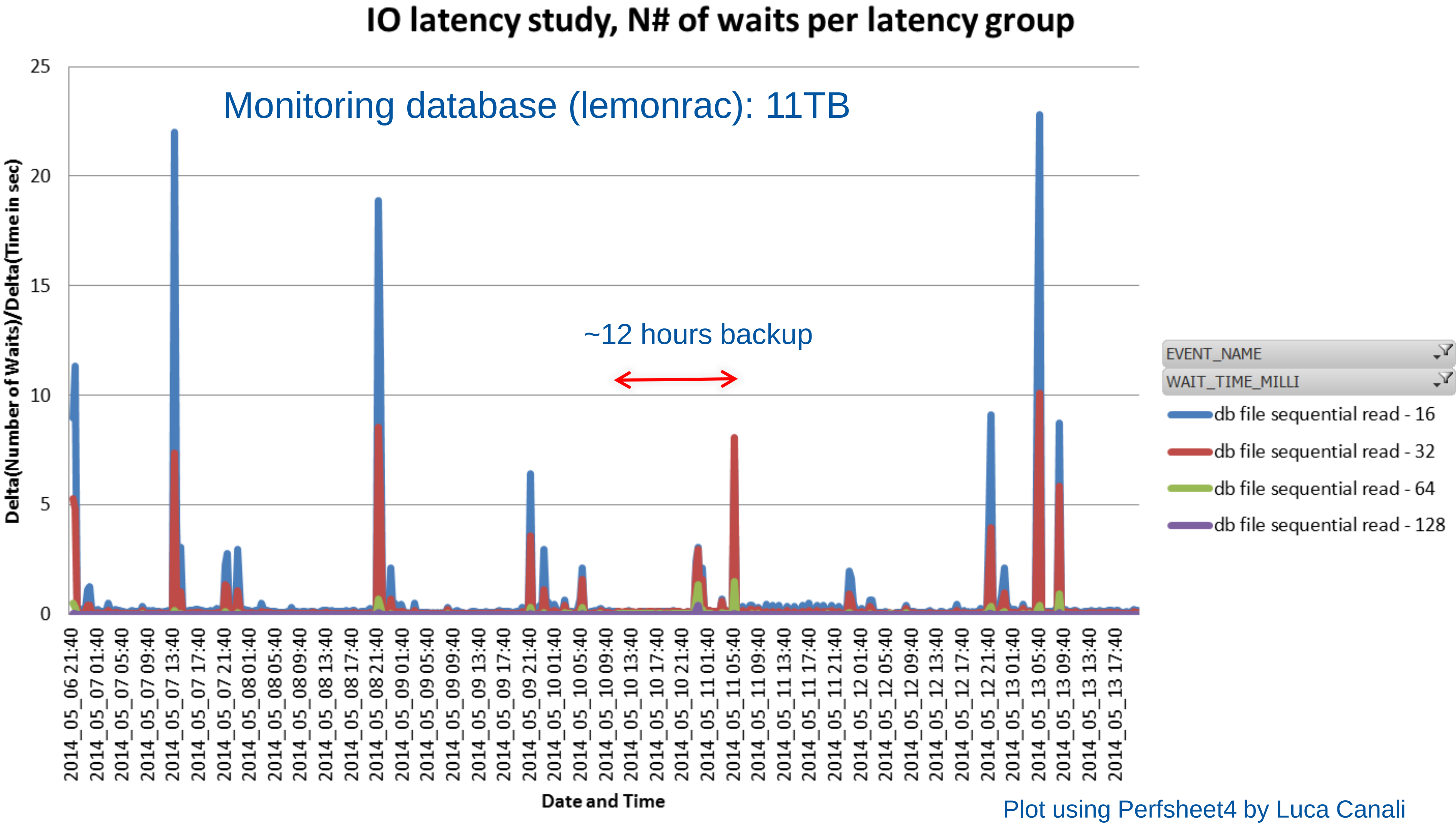
```
LatencyMap.py v1.0b - Luca.Canali@cern.ch

Latency bucket                  Frequency Heatmap: operations per sec                 Latest values    Legend
(millisec)                                                                                 (N#/sec)
  >1k                                                                                  ......0         0
   1k                                                                                  ......0         >0
  512                                                                                  ......4         >235
  256                                                                                  .....42         >470
  128                                                                                  ....342         >705
   64                                                                                  ....958         >940
   32                                                                                  ...1344         >1175
   16                                                                                  ...1119        Max: 1411
    8                                                                                  ....186
    4                                                                                  .....24
    2                                                                                  .....98
   <1                                                                                  .....43        Max(Sum):
      Number of latency operations per second: x=time, y=latency bucket               Sum:...4161     4286


Latency bucket                  Intensity Heatmap: time waited per sec                Latest values    Legend
(millisec)                                                                           (millisec/sec)
  >1k                                                                                  ......0         0
   1k                                                                                  ......0         >0
  512                                                                                  ...1662         >7845
  256                                                                                  ...7992         >15691
  128                                                                                  ..32831         >23537
   64                                                                                  ..45986         >31383
   32                                                                                  ..32256         >39229
   16                                                                                  ..13431        Max: 47075
    8                                                                                  ...1115
    4                                                                                  .....73
    2                                                                                  ....147
   <1                                                                                  .....32        Max(Sum):
      Wait time over elapsed time: x=time, y=latency bucket                           Sum:.135526     135636

Sample num: 92, Delta time (sec): 3.0, Date: 30-SEP-13 07.43.48.097197 PM +02:00
Label: db file sequential read latency data from gv$event histogram
```

10TB dataset, 36% in SSD, 32 sessions, random reads



```
LatencyMap.py v1.0b - Luca.Canali@cern.ch

Latency bucket               Frequency Heatmap: operations per sec             Latest values   Legend
(millisec)                                                                         (N#/sec)
  >1k                                                                         ......0        0
   1k                                                                         ......0        >0
  512                                                                         ......0        >303
  256                                                                         ......0        >607
  128                                                                         ......4        >910
   64                                                                         ....111        >1214
   32                                                                         ....607        >1517
   16                                                                         ...1092        Max: 1821
    8                                                                         ....323
    4                                                                         .....54
    2                                                                         ....134
   <1                                                                         ...1039        Max(Sum):
     Number of latency operations per second: x=time, y=latency bucket        Sum:...3364    4151

Latency bucket               Intensity Heatmap: time waited per sec            Latest values   Legend
(millisec)                                                                    (millisec/sec)
  >1k                                                                         ......0        0
   1k                                                                         ......0        >0
  512                                                                         ......0        >2504
  256                                                                         ......0        >5008
  128                                                                         ....384        >7512
   64                                                                         ...5340        >10016
   32                                                                         ..14558        >12520
   16                                                                         ..13103        Max: 1502
    8                                                                         ...1937
    4                                                                         ....163
    2                                                                         ....201
   <1                                                                         ....779        Max(Sum):
     Wait time over elapsed time: x=time, y=latency bucket                    Sum:..36465    36994

Sample num: 92, Delta time (sec): 3.0, Date: 07-OCT-13 12.30.11.821278 PM +02:00
Label: db file sequential read latency data from gv$event_histogram
```

Flash pool: long running backups



Monitoring database (lemonrac): 11TB

~12 hours backup

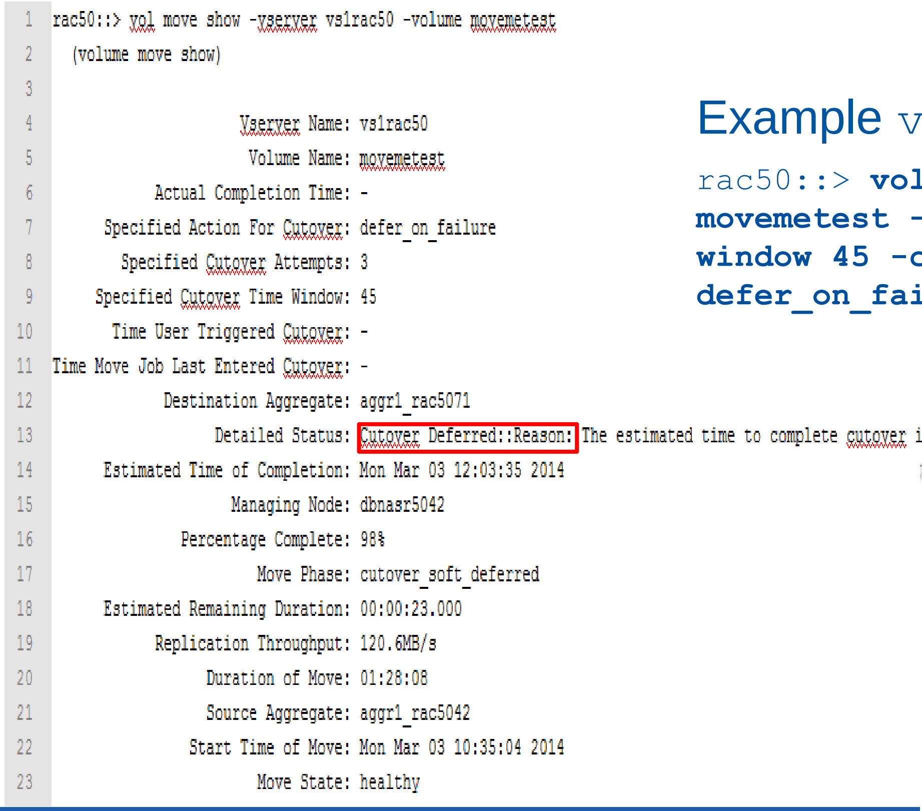
Plot using Perfsheet4 by Luca Canali

Agenda

- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
  - Flash cache
  - Flash pool
- Data motion
- Snapshots
- Clonning in Oracle12c
- Backup to disk
- directNFS
- Monitoring
  - In-house tools
  - Netapp tools
- Conclusions

Vol move

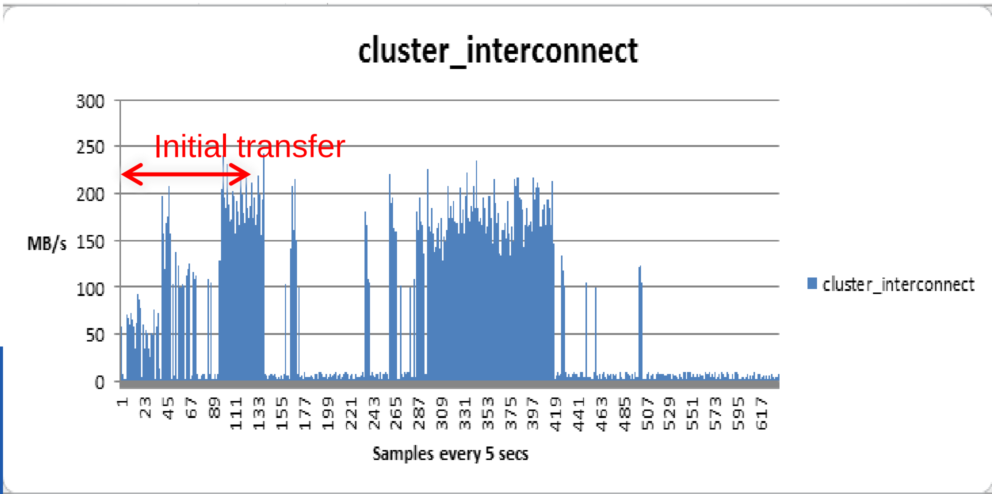
- Powerful feature: rebalancing, interventions,… whole volume granularity
- Transparent but watch-out on high IO (writes) volumes
- Based on SnapMirror technology



```
rac50::> vol move show -vserver vs1rac50 -volume movemetest
  (volume move show)

                    Vserver Name: vs1rac50
                     Volume Name: movemetest
          Actual Completion Time: -
      Specified Action For Cutover: defer_on_failure
        Specified Cutover Attempts: 3
     Specified Cutover Time Window: 45
       Time User Triggered Cutover: -
Time Move Job Last Entered Cutover: -
           Destination Aggregate: aggr1_rac5071
                 Detailed Status: Cutover Deferred::Reason: The estimated time to complete cutover is greater than the cutover window that can be tolerated by the user. Transferring data: 501.0GB sent.
      Estimated Time of Completion: Mon Mar 03 12:03:35 2014
                   Managing Node: dbnasr5042
             Percentage Complete: 98%
                      Move Phase: cutover_soft_deferred
      Estimated Remaining Duration: 00:00:23.000
          Replication Throughput: 120.6MB/s
                Duration of Move: 01:28:08
                Source Aggregate: aggr1_rac5042
              Start Time of Move: Mon Mar 03 10:35:04 2014
                      Move State: healthy
```

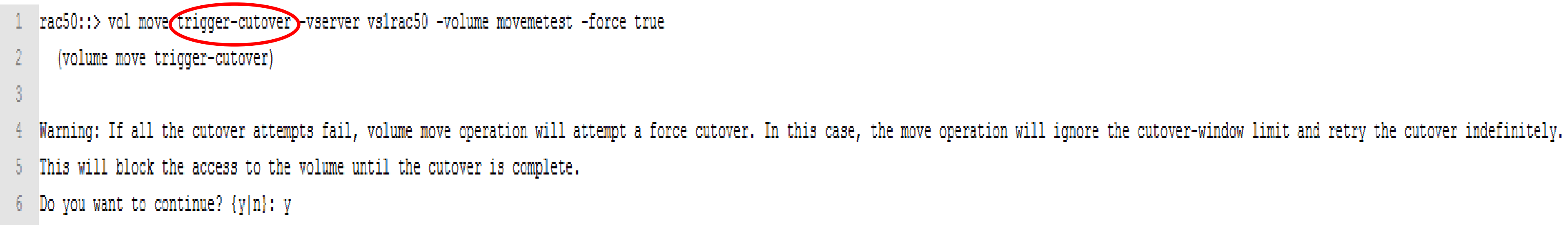
Example `vol move` command:

```
rac50::> vol move start -vserver vs1rac50 -volume movemetest -destination-aggregate aggr1_rac5071 -cutover-window 45 -cutover-attempts 3 -cutover-action defer_on_failure
```



Vol move (II)

- Force cutover: cutover-window will be ignored → client access frozen during cutover duration:



```
rac50::> vol move trigger-cutover -vserver vs1rac50 -volume movemetest -force true
  (volume move trigger-cutover)

Warning: If all the cutover attempts fail, volume move operation will attempt a force cutover. In this case, the move operation will ignore the cutover-window limit and retry the cutover indefinitely.
This will block the access to the volume until the cutover is complete.
Do you want to continue? {y|n}: y
```

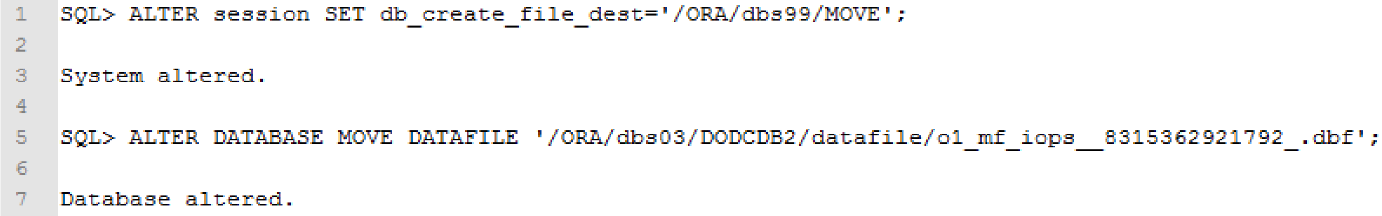
- Flash_pool volumes will need to warm up the SSDs again
  - Probably solved in a future Ontap release
- To avoid interconnect traffic, logical interface (lif) should be moved (NFSv3) to the same controller where new volume is located
  - pnfs (NFSv4.1) netapp implementation redirects IO load to new location without need of remounting.

Vol move (III)

- One lif per data volume
  - To be able to use Ontap move volume feature with no impact on cluster interconnect switch.
  - No need to remount on the new controller hosting the volume.
  - Lif can be moved, once the volume has been migrated.
  - Interconnect just 10 gbps bandwidth (20gbps in next generation)
- Just targeting data volumes
  - Bug ID 540038: Failover groups do not allow specifying a port order
    - Workaround: `network interface failover create`
  - 128 lifs maximum (all types) in Ontap 8.2

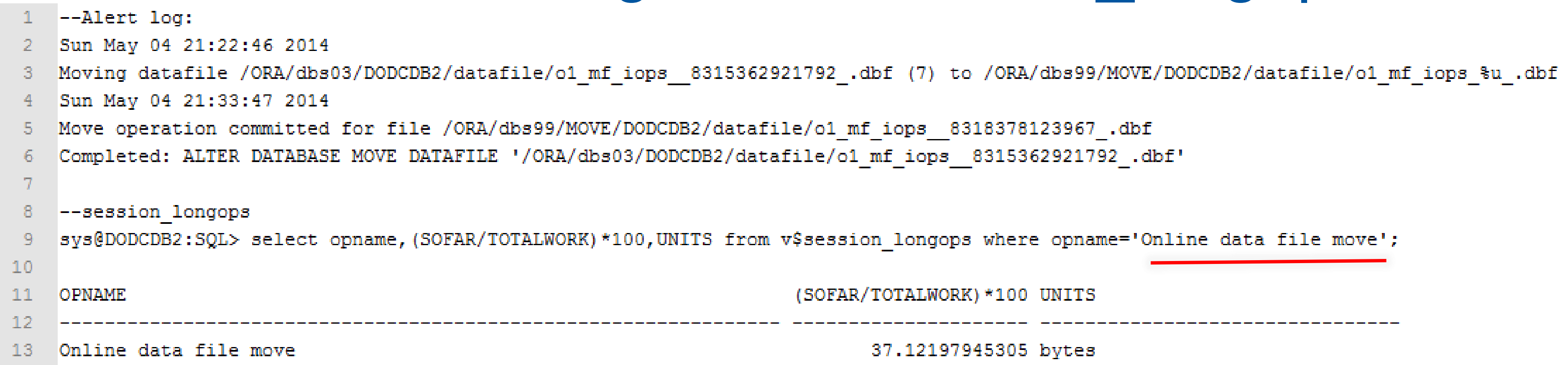
Oracle12c: online datafile move

- Very robust, even with high IO load
  - It takes advantage of database memory buffers
- Works with OMF



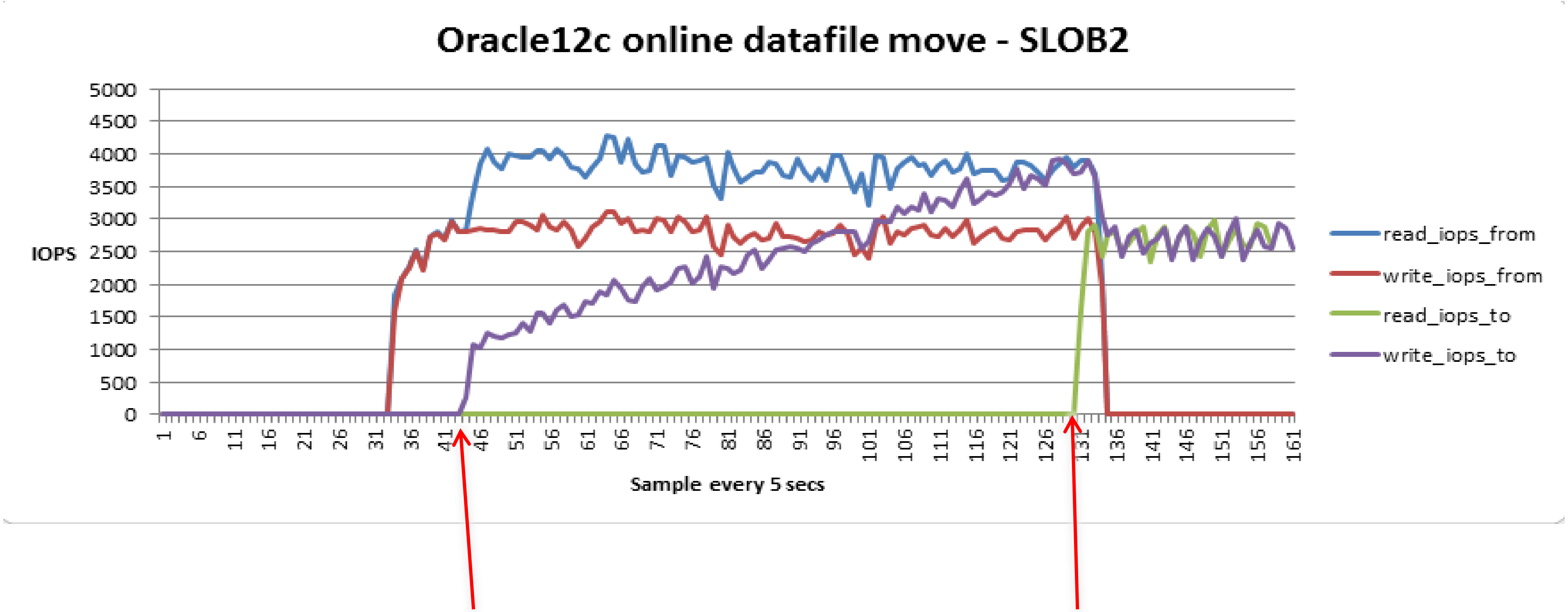
```
SQL> ALTER session SET db_create_file_dest='/ORA/dbs99/MOVE';

System altered.

SQL> ALTER DATABASE MOVE DATAFILE '/ORA/dbs03/DODCDB2/datafile/o1_mf_iops__8315362921792_.dbf';

Database altered.
```

- Track it at alert.log and v$session_longops



```
--Alert log:
Sun May 04 21:22:46 2014
Moving datafile /ORA/dbs03/DODCDB2/datafile/o1_mf_iops__8315362921792_.dbf (7) to /ORA/dbs99/MOVE/DODCDB2/datafile/o1_mf_iops_%u_.dbf
Sun May 04 21:33:47 2014
Move operation committed for file /ORA/dbs99/MOVE/DODCDB2/datafile/o1_mf_iops__8318378123967_.dbf
Completed: ALTER DATABASE MOVE DATAFILE '/ORA/dbs03/DODCDB2/datafile/o1_mf_iops__8315362921792_.dbf'

--session_longops
sys@DODCDB2:SQL> select opname,(SOFAR/TOTALWORK)*100,UNITS from v$session_longops where opname='Online data file move';

OPNAME                                                           (SOFAR/TOTALWORK)*100 UNITS
---------------------------------------------------------------- --------------------- --------------------------------
Online data file move                                                   37.12197945305 bytes
```

Oracle12c: online datafile move (II)



`alter database move datafile`

Move was completed.

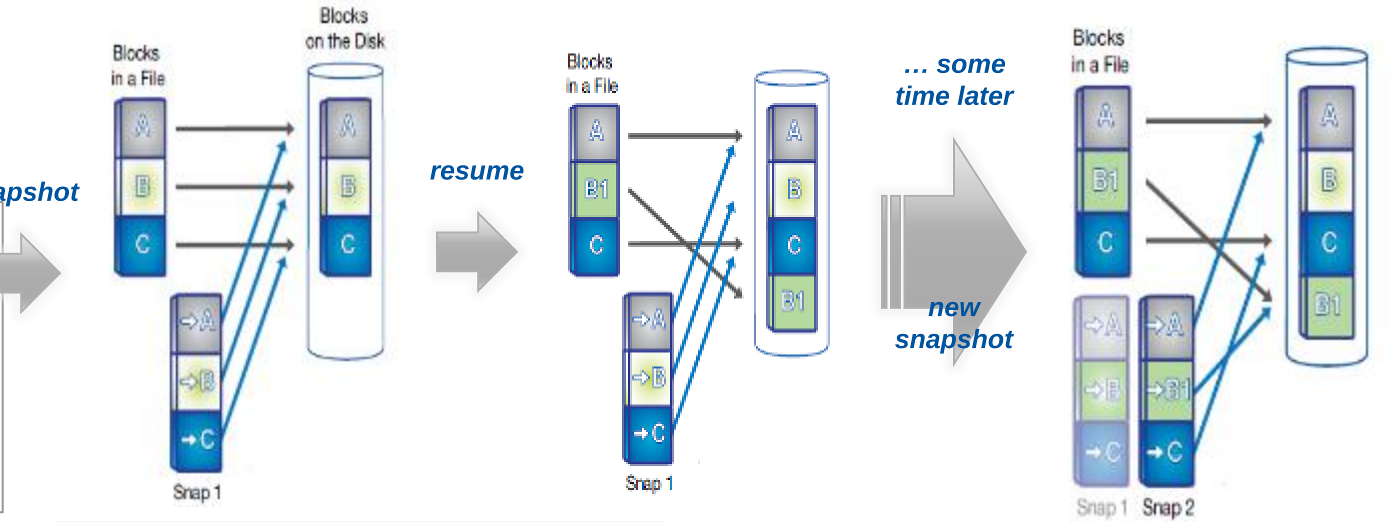
Agenda

- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
  - Flash cache
  - Flash pool
- Data motion
- Backup and Recovery: snapshots
- Clonning in Oracle12c
- Backup to disk
- directNFS
- Monitoring
  - In-house tools
  - Netapp tools
- Conclusions

DBaaS:Backup management

- Same backup procedure for all RDBMS
- Backup workflow:

```
mysql> FLUSH TABLES
       WITH READ LOCK;
mysql> FLUSH LOGS;
            or
Oracle>alter database
begin backup;
            Or
Postgresql> SELECT
pg_start_backup('$SNAP');
```



```
          mysql> UNLOCK TABLES;
                 Or
Oracle>alter database end backup;
                 or
Postgresql> SELECT
pg_stop_backup(),
pg_create_restore_point('$SNAP');
```

Snapshots in Oracle

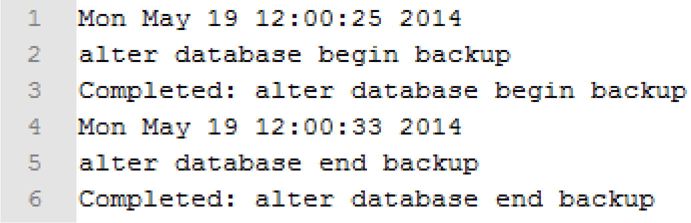
- Storage-based technology
- Speed-up backups/restores: from hours/days to seconds
- Handled by a plug-in on our backup and recovery solution:

```
/etc/init.d/syscontrol --very_silent -i rman_backup start -maxRetries 1 -exec takesnap_zapi.pl -debug -snap
dbnasr0009-priv:/ORA/dbs03/PUBSTG level_EXEC_SNAP -i pubstg_rac50
```

lif Global namespace

- Example:

pubstg: 280GB size, ~ 1 TB archivelogs/day

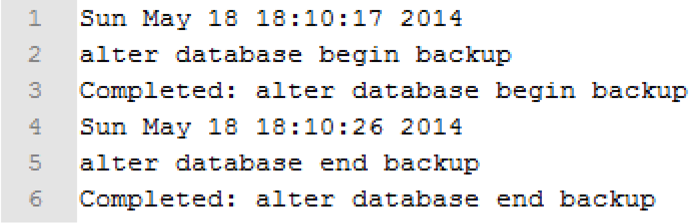


```
Mon May 19 12:00:25 2014
alter database begin backup
Completed: alter database begin backup
Mon May 19 12:00:33 2014
alter database end backup
Completed: alter database end backup
```

8secs

adcr: 24TB size, ~ 2,5 TB archivelogs/day

9secs

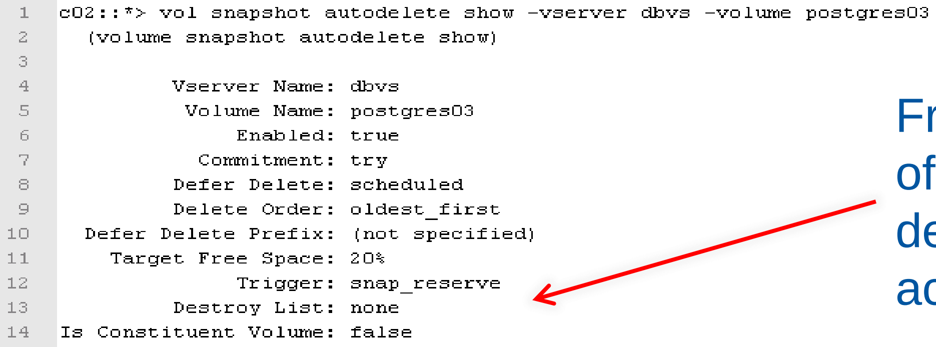


```
Sun May 18 18:10:17 2014
alter database begin backup
Completed: alter database begin backup
Sun May 18 18:10:26 2014
alter database end backup
Completed: alter database end backup
```

- Drawback: lack of integration with RMAN
  - Ontap commands: **`snap create/restore`**
    - Snaprestore requires license
  - snapshots not available via RMAN API
  - But some solutions exist: Netapp MML Proxy api, Oracle snapmanager

Snapshots management / setup

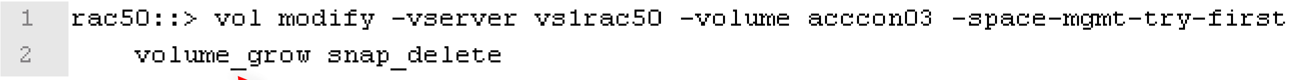
- Managed by the system, with autodeletion policies



```
c02::*> vol snapshot autodelete show -vserver dbvs -volume postgres03
  (volume snapshot autodelete show)

             Vserver Name: dbvs
              Volume Name: postgres03
                  Enabled: true
               Commitment: try
             Defer Delete: scheduled
             Delete Order: oldest_first
      Defer Delete Prefix: (not specified)
        Target Free Space: 20%
                  Trigger: snap_reserve
             Destroy List: none
Is Constituent Volume: false
```

From 20% to 40% of volume size, depending on db activity

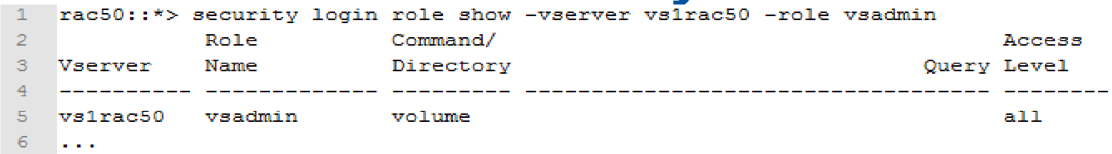
- Primary Space Management Strategy



```
rac50::> vol modify -vserver vs1rac50 -volume acccon03 -space-mgmt-try-first
    volume_grow snap_delete
```

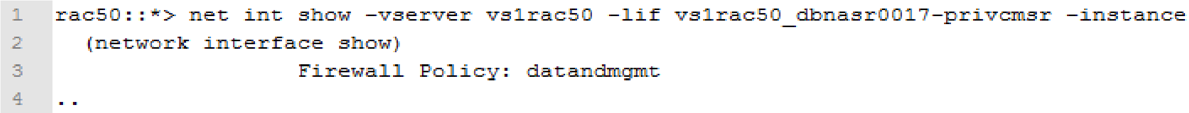
default

- Connect to the lif used to mount file system
  - As user vsadmin:



```
rac50::*> security login role show -vserver vs1rac50 -role vsadmin
           Role          Command/                                         Access
Vserver    Name          Directory                                  Query Level
---------- ------------- --------- ----------------------------------- --------
vs1rac50   vsadmin       volume                                           all
...
```

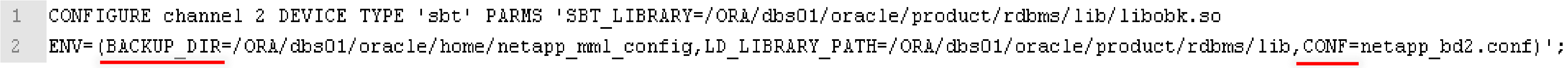
  - Open lif's ssh port:



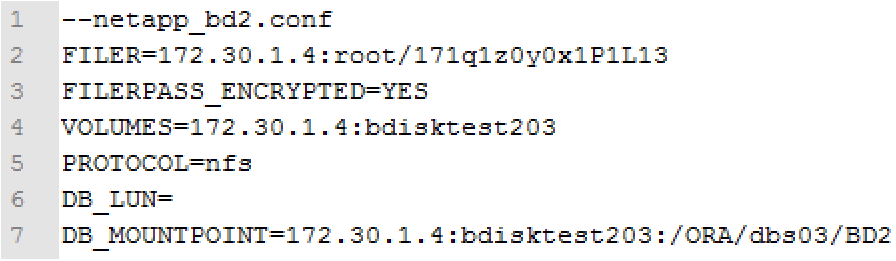
```
rac50::*> net int show -vserver vs1rac50 -lif vs1rac50_dbnasr0017-privcmsr -instance
  (network interface show)
                    Firewall Policy: datandmgmt
..
```

Netapp MML Proxy backup v1

- Implementation of SBT API
- Simple configuration

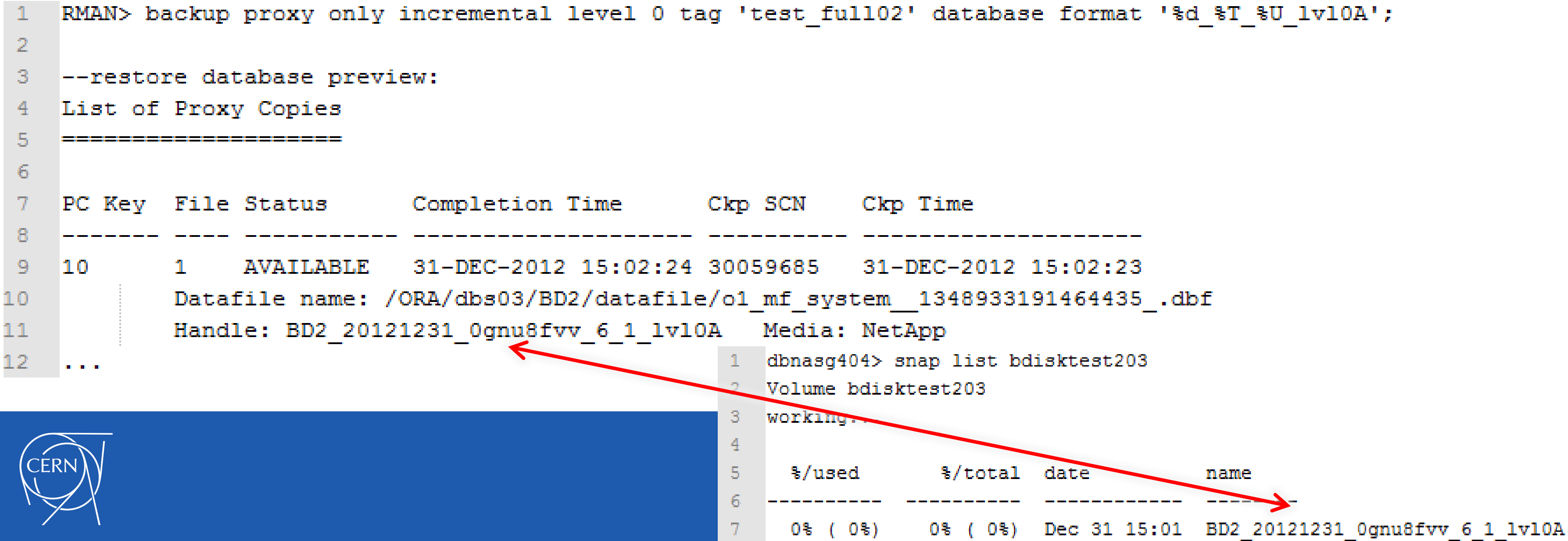


```
CONFIGURE channel 2 DEVICE TYPE 'sbt' PARMS 'SBT_LIBRARY=/ORA/dbs01/oracle/product/rdbms/lib/libobk.so
ENV=(BACKUP_DIR=/ORA/dbs01/oracle/home/netapp_mml_config,LD_LIBRARY_PATH=/ORA/dbs01/oracle/product/rdbms/lib,CONF=netapp_bd2.conf)';
```



```
--netapp_bd2.conf
FILER=172.30.1.4:root/171q1z0y0x1P1L13
FILERPASS_ENCRYPTED=YES
VOLUMES=172.30.1.4:bdisktest203
PROTOCOL=nfs
DB_LUN=
DB_MOUNTPOINT=172.30.1.4:bdisktest203:/ORA/dbs03/BD2
```

- Backups will generate an underlying snapshot



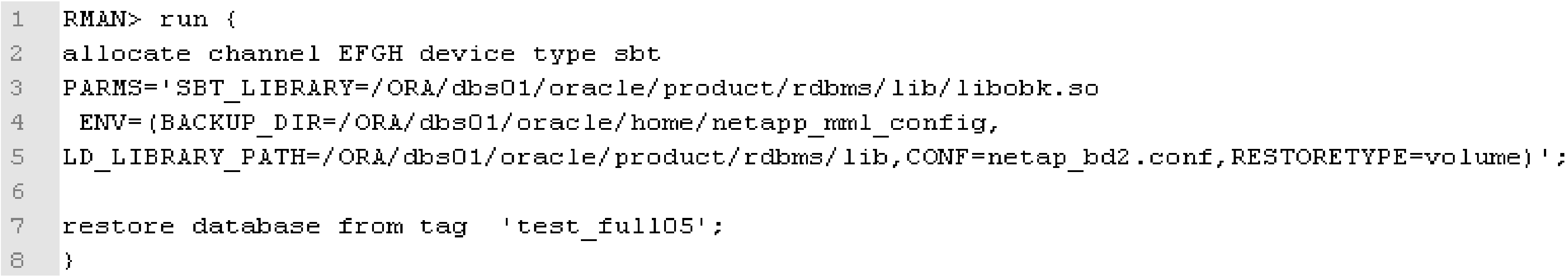
```
RMAN> backup proxy only incremental level 0 tag 'test_full02' database format '%d_%T_%U_lvl0A';

--restore database preview:
List of Proxy Copies
====================

PC Key  File Status      Completion Time      Ckp SCN    Ckp Time
------- ---- ----------- -------------------- ---------- --------------------
10      1    AVAILABLE   31-DEC-2012 15:02:24 30059685   31-DEC-2012 15:02:23
        Datafile name: /ORA/dbs03/BD2/datafile/o1_mf_system__1348933191464435_.dbf
        Handle: BD2_20121231_0gnu8fvv_6_1_lvl0A   Media: NetApp
...
```

```
dbnasg404> snap list bdisktest203
Volume bdisktest203
working...

  %/used       %/total  date          name
----------  ----------  ------------  --------
  0% ( 0%)    0% ( 0%)  Dec 31 15:01  BD2_20121231_0gnu8fvv_6_1_lvl0A
```

Netapp MML Proxy backup v1

- v$proxy views
  - v$proxy_datafile → BACKUP_FUZZY=YES

    (`alter database begin/end backup` being used)
- Restore and delete operations are commanded by environment variables
  - **RESTORETYPE={volume|file|controlvolume}**
  - **DELETETYPE =snap**

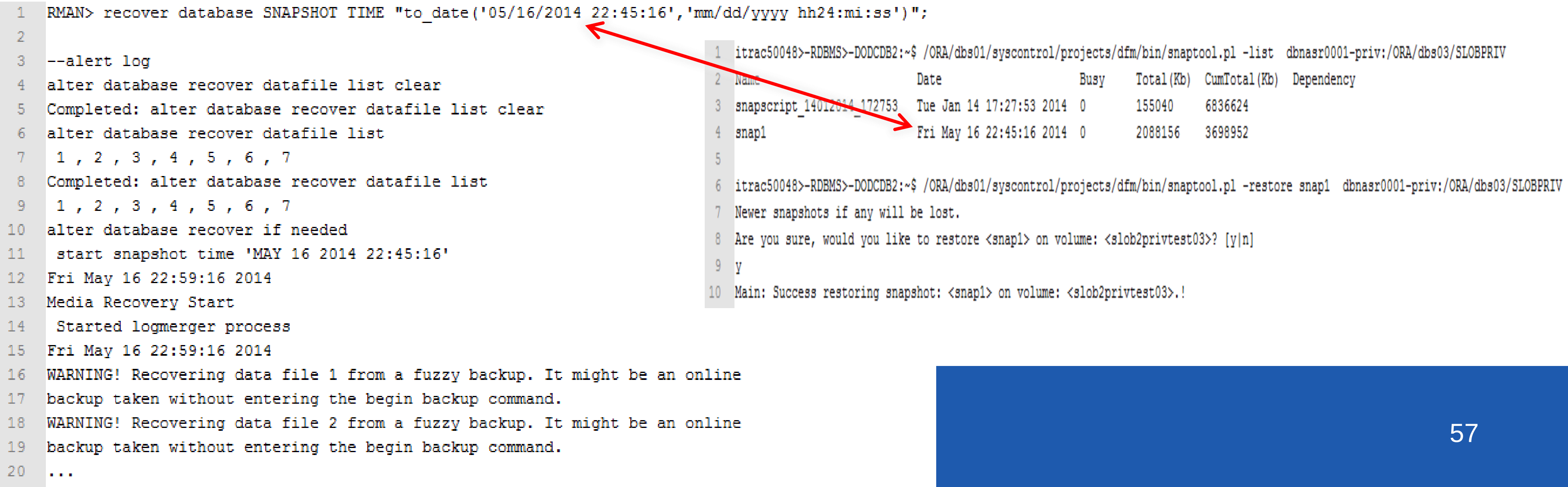


```
RMAN> run {
allocate channel EFGH device type sbt
PARMS='SBT_LIBRARY=/ORA/dbs01/oracle/product/rdbms/lib/libobk.so
 ENV=(BACKUP_DIR=/ORA/dbs01/oracle/home/netapp_mml_config,
LD_LIBRARY_PATH=/ORA/dbs01/oracle/product/rdbms/lib,CONF=netap_bd2.conf,RESTORETYPE=volume)';

restore database from tag  'test_full05';
}
```

- Integration with RMAN API
  - Though disk catalogue is in a file (should be accessible on all instances, RAC), it is not integrated with catalogue/controlfile
- Version 2, it supports Ontap C-mode.
- It is a freely available tool, open community support

Oracle12c: recover snapshot

- RMAN Enhancements in Oracle 12c (Doc ID 1534487.1)
- Under certain conditions no need to set db in backup mode:
  - Database crash consistent at the point of the snapshot AND
  - Write ordering is preserved for each file within a snapshot AND
  - Snapshot stores the time at which a snapshot is completed



```
RMAN> recover database SNAPSHOT TIME "to_date('05/16/2014 22:45:16','mm/dd/yyyy hh24:mi:ss')";

--alert log
alter database recover datafile list clear
Completed: alter database recover datafile list clear
alter database recover datafile list
 1 , 2 , 3 , 4 , 5 , 6 , 7
Completed: alter database recover datafile list
 1 , 2 , 3 , 4 , 5 , 6 , 7
alter database recover if needed
 start snapshot time 'MAY 16 2014 22:45:16'
Fri May 16 22:59:16 2014
Media Recovery Start
 Started logmerger process
Fri May 16 22:59:16 2014
WARNING! Recovering data file 1 from a fuzzy backup. It might be an online
backup taken without entering the begin backup command.
WARNING! Recovering data file 2 from a fuzzy backup. It might be an online
backup taken without entering the begin backup command.
...
```

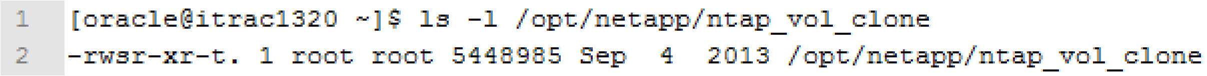
```
itrac50048>-RDBMS>-DODCDB2:~$ /ORA/dbs01/syscontrol/projects/dfm/bin/snaptool.pl -list  dbnasr0001-priv:/ORA/dbs03/SLOBPRIV
Name                       Date                      Busy    Total(Kb)  CumTotal(Kb)  Dependency
snapscript_14012014_172753 Tue Jan 14 17:27:53 2014  0       155040     6836624
snap1                      Fri May 16 22:45:16 2014  0       2088156    3698952

itrac50048>-RDBMS>-DODCDB2:~$ /ORA/dbs01/syscontrol/projects/dfm/bin/snaptool.pl -restore snap1  dbnasr0001-priv:/ORA/dbs03/SLOBPRIV
Newer snapshots if any will be lost.
Are you sure, would you like to restore <snap1> on volume: <slob2privtest03>? [y|n]
y
Main: Success restoring snapshot: <snap1> on volume: <slob2privtest03>.!
```

Agenda

- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
  - Flash cache
  - Flash pool
- Data motion
- Snapshots
- Clonning in Oracle12c
- Backup to disk
- direcNFS
- Monitoring
  - In-house tools
  - Netapp tools
- Conclusions

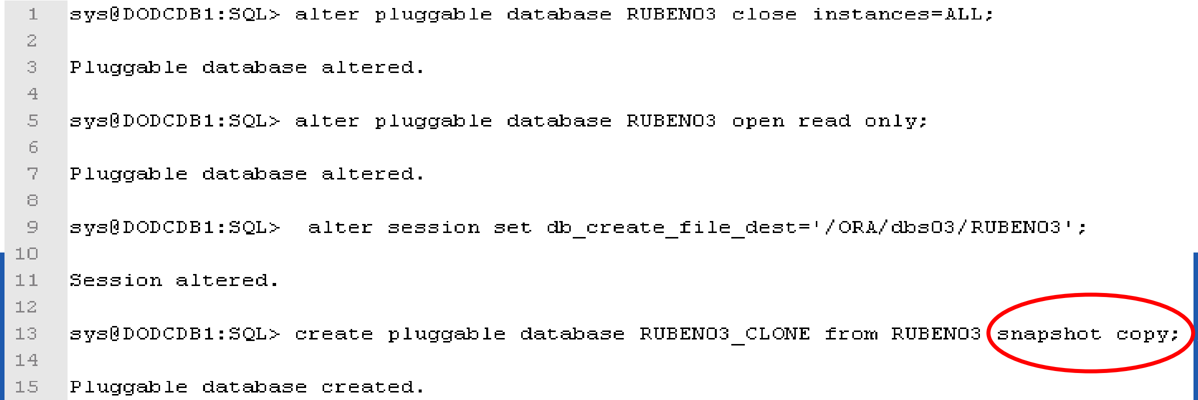
Oracle12c: multi-tenancy cloning

- TR-4266: NetApp Cloning Plug-in for Oracle Multitenant Database 12c
- Patch required on 12.1.0.1 (MOS 16221044)
- Storage credentials stored in an Oracle wallet
- Check dnfs is in use and exports defined at `$ORACLE_HOME/dbs/oranfstab`
- Check plug-in has proper permissions



```
[oracle@itrac1320 ~]$ ls -l /opt/netapp/ntap_vol_clone
-rwsr-xr-t. 1 root root 5448985 Sep  4  2013 /opt/netapp/ntap_vol_clone
```

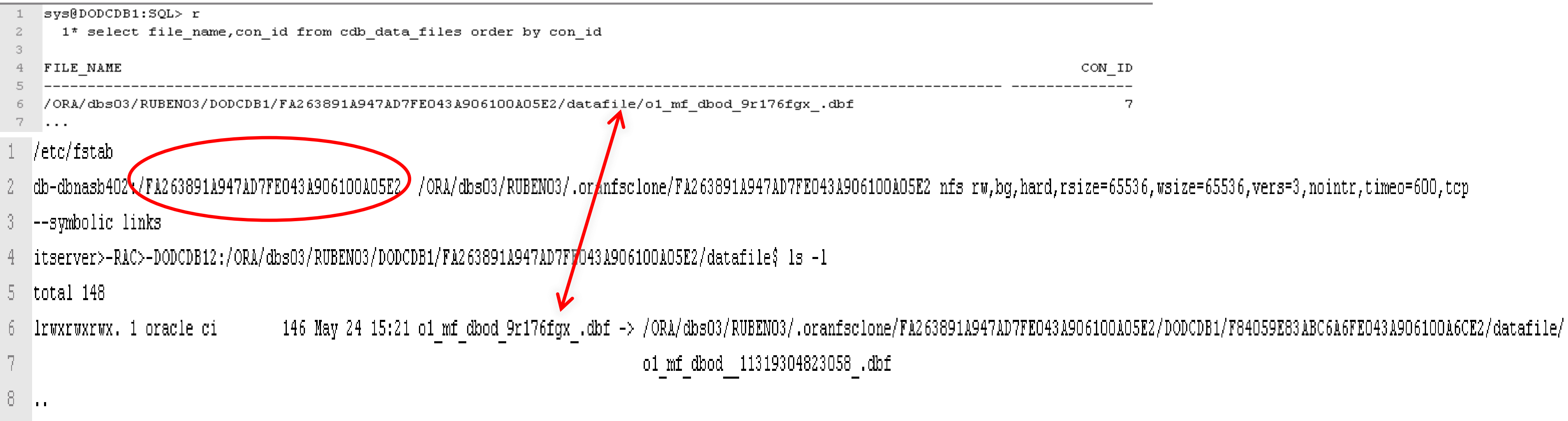
- Using OMF:



```
sys@DODCDB1:SQL> alter pluggable database RUBEN03 close instances=ALL;

Pluggable database altered.

sys@DODCDB1:SQL> alter pluggable database RUBEN03 open read only;

Pluggable database altered.

sys@DODCDB1:SQL>  alter session set db_create_file_dest='/ORA/dbs03/RUBEN03';

Session altered.

sys@DODCDB1:SQL> create pluggable database RUBEN03_CLONE from RUBEN03 snapshot copy;

Pluggable database created.
```

Oracle12c: multi-tenancy cloning

- Mount and file system reference



```
sys@DODCDB1:SQL> r
  1* select file_name,con_id from cdb_data_files order by con_id

FILE_NAME                                                                                                  CON_ID
---------------------------------------------------------------------------------------------------------- -------------
/ORA/dbs03/RUBEN03/DODCDB1/FA263891A947AD7FE043A906100A05E2/datafile/o1_mf_dbod_9r176fgx_.dbf                    7
...
```

```
/etc/fstab
db-dbnasb402:/FA263891A947AD7FE043A906100A05E2 /ORA/dbs03/RUBEN03/.oranfsclone/FA263891A947AD7FE043A906100A05E2 nfs rw,bg,hard,rsize=65536,wsize=65536,vers=3,nointr,timeo=600,tcp
--symbolic links
itserver>-RAC>-DODCDB12:/ORA/dbs03/RUBEN03/DODCDB1/FA263891A947AD7FE043A906100A05E2/datafile$ ls -l
total 148
lrwxrwxrwx. 1 oracle ci       146 May 24 15:21 o1_mf_dbod_9r176fgx_.dbf -> /ORA/dbs03/RUBEN03/.oranfsclone/FA263891A947AD7FE043A906100A05E2/DODCDB1/F84059E83ABC6A6FE043A906100A6CE2/datafile/
                                                                            o1_mf_dbod__11319304823058_.dbf
..
```

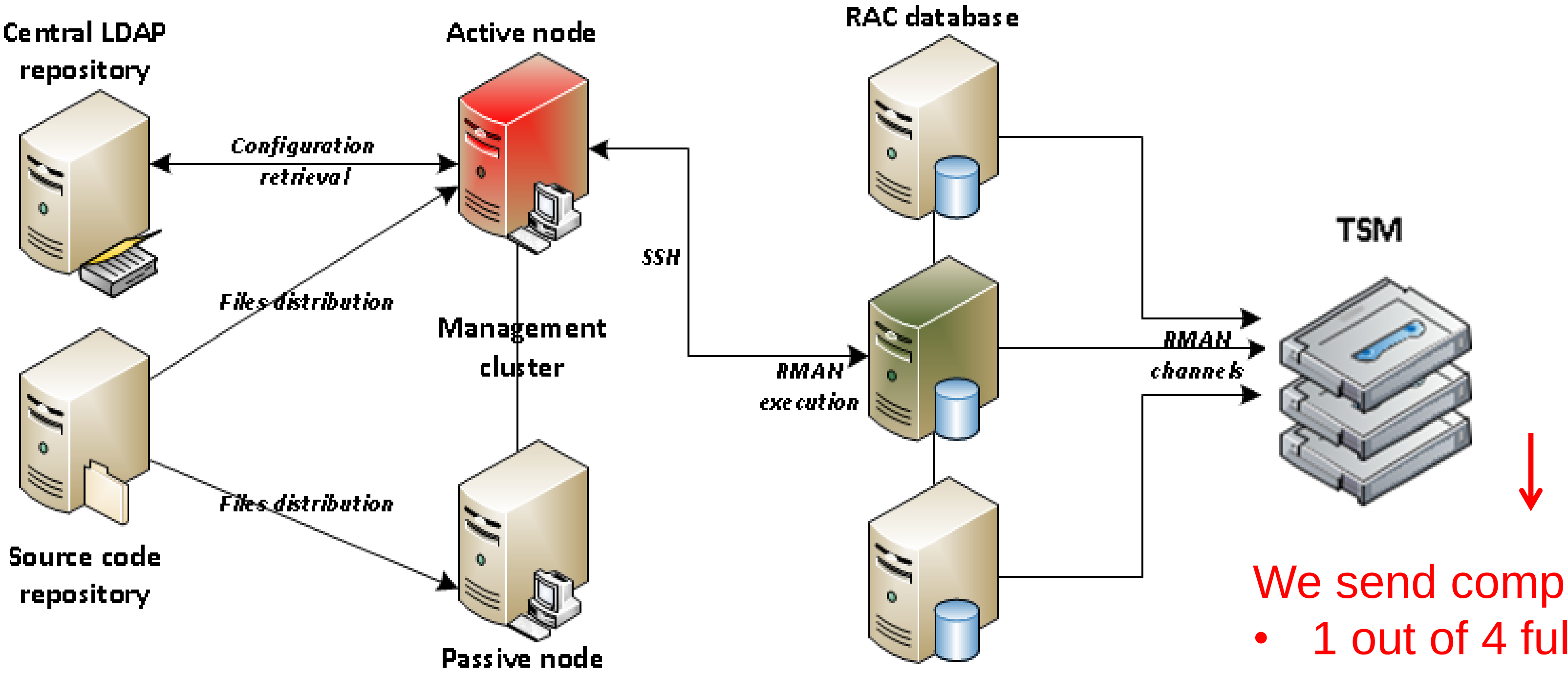
- Single instance is all done
- For RAC:
  - Replicate file system changes to open on other instances
  - CRS service registration/creation
  - Undo changes when clone is destroyed

Agenda

- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
  - Flash cache
  - Flash pool
- Data motion
- Snapshots
- Clonning in Oracle12c
- Backup to disk
- direcNFS
- Monitoring
  - In-house tools
  - Netapp tools
- Conclusions

Backup architecture

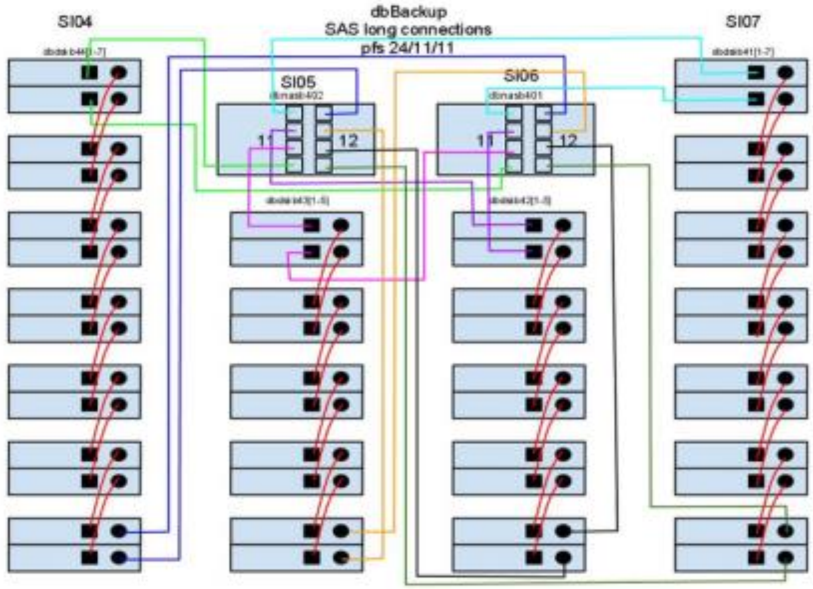
- Custom solution: about 15k lines of code, Perl + Bash
- Flexible: easy to adapt to new Oracle release, backup media
  - Based on Oracle Recovery Manager (RMAN) templates
- Central logging
- Easy to extend via Perl plug-ins: snapshot, exports, RO tablespaces,…



We send compressed:

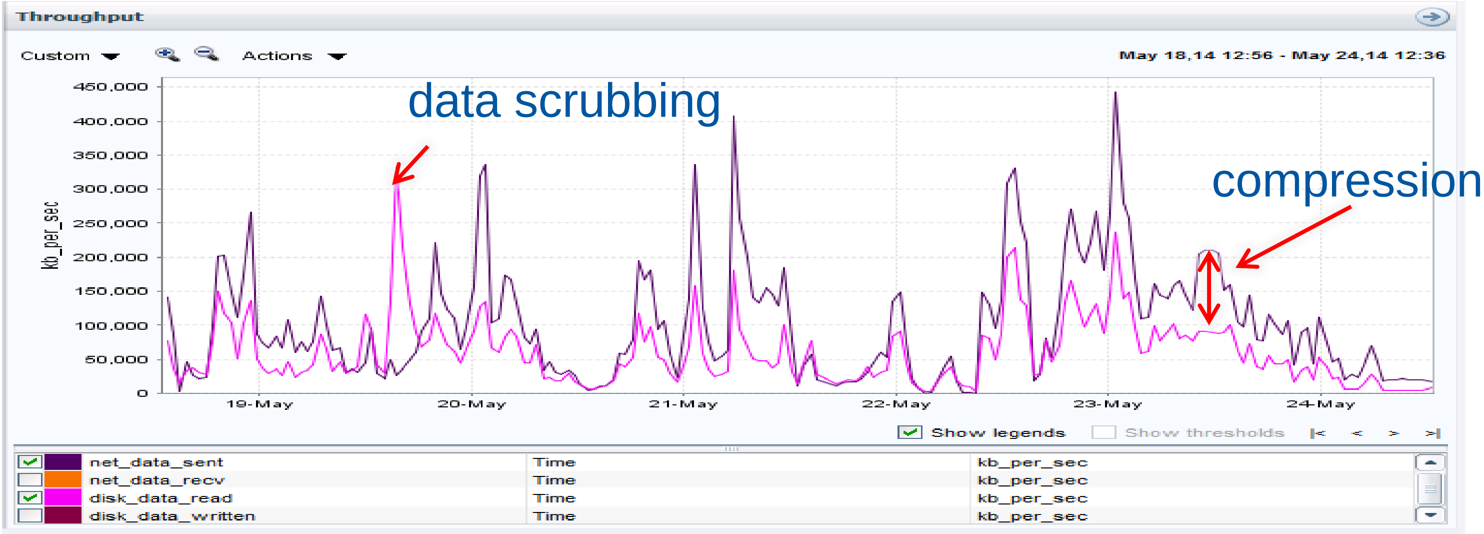
- 1 out of 4 full backups
- All archivelogs

Backup to disk: storage



- 2xFAS6240 Netapp controllers, running ONTAP 8.2.1 C-mode
- 24xdiskshelf DS4243
  - 24x3TB SATA disks each (576 disks)
  - raid_dp (raid6) → 1.1 PB usable space split into 8 aggregates
- 2xquad core 64bit Intel(R) Xeon(R) CPU E5540 @ 2.53GHz
- 10gbps connectivity
- Multipath SAS loops 3 gbps → 6 gpbs maximum throughput (dual path)
- Flash cache 512GB per node (meta data caching)

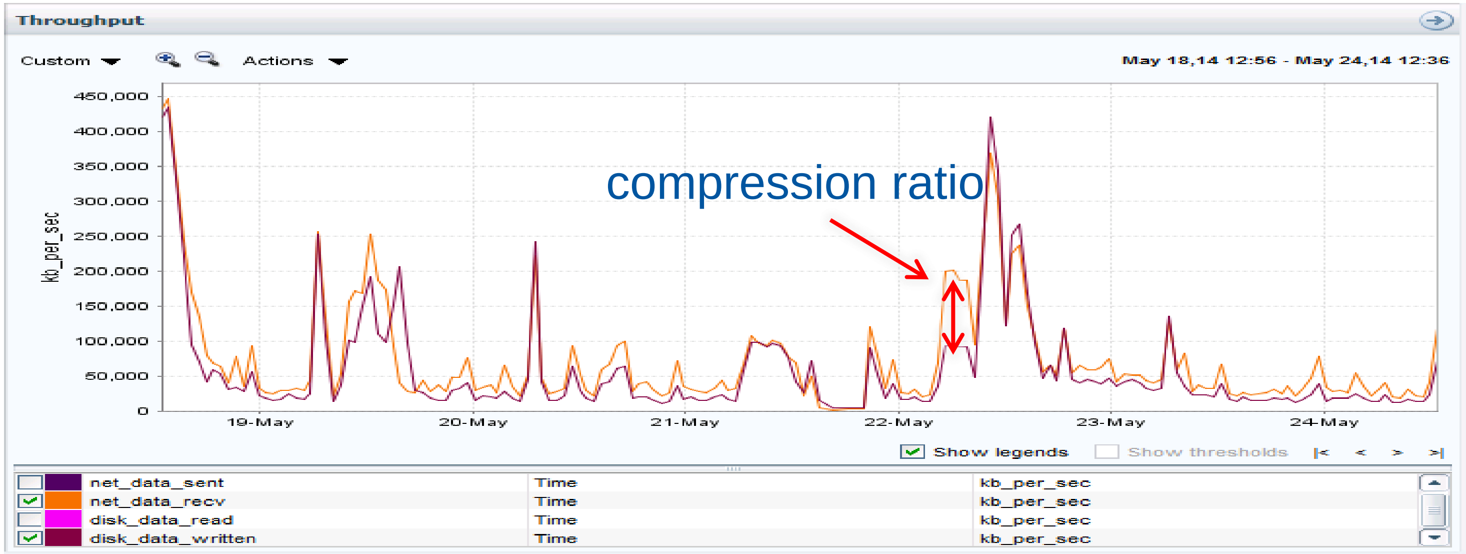
Backup to disk: throughput (one head)



data scrubbing

compression ratio

☑	■	net_data_sent	Time	kb_per_sec
□	■	net_data_recv	Time	kb_per_sec
☑	■	disk_data_read	Time	kb_per_sec
□	■	disk_data_written	Time	kb_per_sec



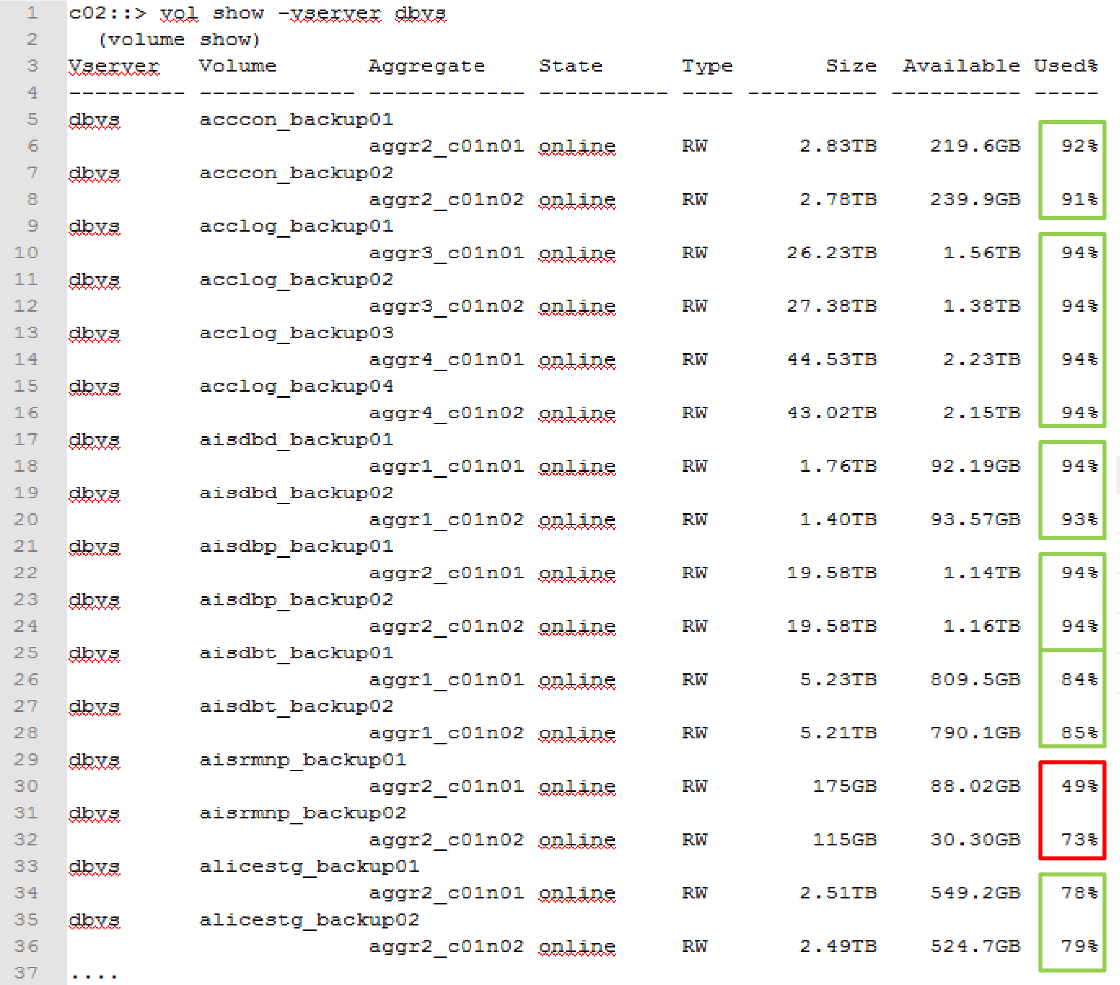
compression ratio

□	■	net_data_sent	Time	kb_per_sec
☑	■	net_data_recv	Time	kb_per_sec
□	■	disk_data_read	Time	kb_per_sec
☑	■	disk_data_written	Time	kb_per_sec

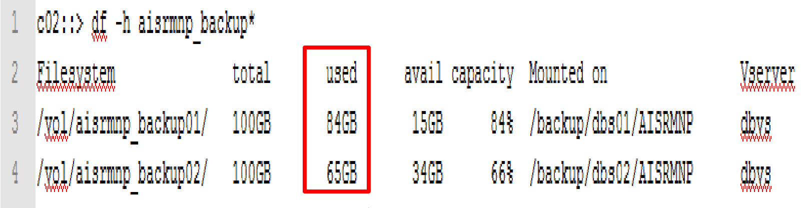
555TB used
538TB saved
due to
compression
mainly but also
deduplication

Backup to disk: space consumption

- The aim is to be as balanced as possible among the volumes assigned to the database

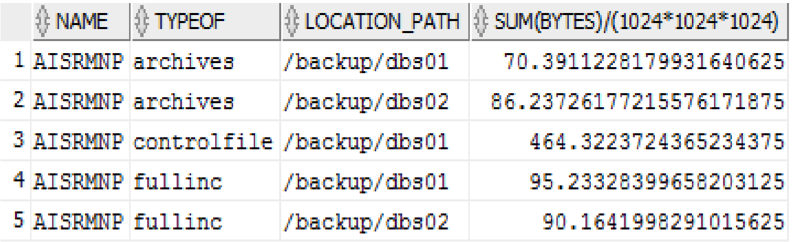


```
c02::> vol show -vserver dbvs
  (volume show)
Vserver   Volume       Aggregate    State      Type       Size  Available Used%
--------- ------------ ------------ ---------- ---- ---------- ---------- -----
dbvs      acccon_backup01
                       aggr2_c01n01 online     RW       2.83TB    219.6GB   92%
dbvs      acccon_backup02
                       aggr2_c01n02 online     RW       2.78TB    239.9GB   91%
dbvs      acclog_backup01
                       aggr3_c01n01 online     RW      26.23TB     1.56TB   94%
dbvs      acclog_backup02
                       aggr3_c01n02 online     RW      27.38TB     1.38TB   94%
dbvs      acclog_backup03
                       aggr4_c01n01 online     RW      44.53TB     2.23TB   94%
dbvs      acclog_backup04
                       aggr4_c01n02 online     RW      43.02TB     2.15TB   94%
dbvs      aisdbd_backup01
                       aggr1_c01n01 online     RW       1.76TB    92.19GB   94%
dbvs      aisdbd_backup02
                       aggr1_c01n02 online     RW       1.40TB    93.57GB   93%
dbvs      aisdbp_backup01
                       aggr2_c01n01 online     RW      19.58TB     1.14TB   94%
dbvs      aisdbp_backup02
                       aggr2_c01n02 online     RW      19.58TB     1.16TB   94%
dbvs      aisdbt_backup01
                       aggr1_c01n01 online     RW       5.23TB    809.5GB   84%
dbvs      aisdbt_backup02
                       aggr1_c01n02 online     RW       5.21TB    790.1GB   85%
dbvs      aisrmnp_backup01
                       aggr2_c01n01 online     RW        175GB    88.02GB   49%
dbvs      aisrmnp_backup02
                       aggr2_c01n02 online     RW        115GB    30.30GB   73%
dbvs      alicestg_backup01
                       aggr2_c01n01 online     RW       2.51TB    549.2GB   78%
dbvs      alicestg_backup02
                       aggr2_c01n02 online     RW       2.49TB    524.7GB   79%
 ....
```



```
c02::> df -h aisrmnp_backup*
Filesystem           total      used     avail capacity  Mounted on            Vserver
/vol/aisrmnp_backup01/  100GB      84GB      15GB      84%  /backup/dbs01/AISRMNP  dbvs
/vol/aisrmnp_backup02/  100GB      65GB      34GB      66%  /backup/dbs02/AISRMNP  dbvs
```

Deduplication applied



	NAME	TYPEOF	LOCATION_PATH	SUM(BYTES)/(1024*1024*1024)
1	AISRMNP	archives	/backup/dbs01	70.3911228179931640625
2	AISRMNP	archives	/backup/dbs02	86.23726177215576171875
3	AISRMNP	controlfile	/backup/dbs01	464.3223724365234375
4	AISRMNP	fullinc	/backup/dbs01	95.23328399658203125
5	AISRMNP	fullinc	/backup/dbs02	90.1641998291015625

- Especial verbs while backing up, e.g. `duration`
- Big files → use `section`

Oracle12c compression

- Oracle 11.2.0.4, new servers (32 cores,129GB RAM)

Intel(R) Xeon(R) CPU E5-2650* 0 @ 2.00GHz

no-compressed (t)	basic	low	medium	high	No-compressed-fs	Inline-compression Netapp 8.2P3
392GB (**devdb11**)	62.24GB(1h54’)	89.17GB (27’30’’)	73.84GB (1h01’)	50.71GB (7h17’)	349GB(22’35’’)	137GB(22’35’’)
Percentage saved (%)	82%	74.4%	78.8%	85.4%	0%	62%

- Oracle 12.1.0.1 new servers

no-compressed (t)	basic	low	medium	high	No-compressed-fs	Inline-compression Netapp 8.2P3
376GB (**devdb11 upgraded to 12c**)	45.2GB (1h29’)	64.13GB (22’)	52.95GB (48’)	34.17GB (5h17’)	252.8GB(22’)	93GB(20’)
Percentage saved (%)	82.1%	74.6%	79%	86.4%	0%	64.5%
229.2GB (**tablespace using Oracle Crypto**)	57.4GB (2h45’)	57.8GB (10’)	58.3GB (44’’)	56.7GB (4h13’)	230GB(12’30’’)	177GB(15’45’’)
Percentage saved (%)	74.95%	74.7%	74.5%	75.2%	0%	22.7%

Agenda

- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
  - Flash cache
  - Flash pool
- Data motion
- Snapshots
- Clonning in Oracle12c
- Backup to disk
- directNFS
- Monitoring
  - In-house tools
  - Netapp tools
- Conclusions

Oracle directNFS

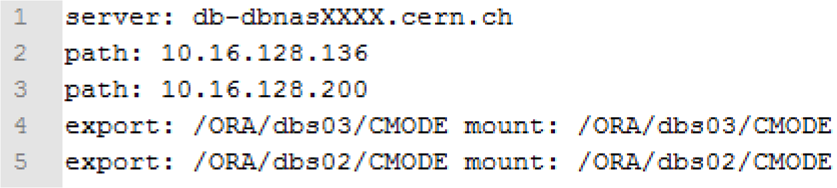
- Set-up: Oracle support note [ID 762374.1]

```
ln -s libnfsodm11.so libodm11.so
```

- dnfs enabled by default in Oracle 12c

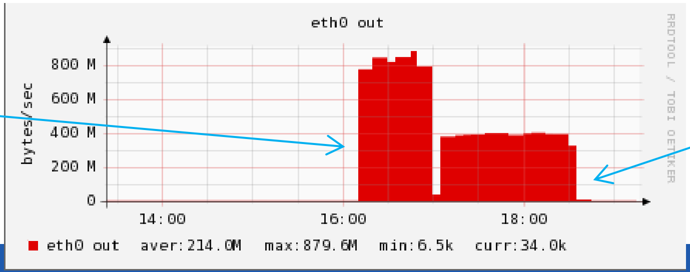
`ln -s libnfsodm12.so libodm12.so` (v$dnfs_servers)

- Multipath. Check note [ID 822481.1]
  - To take advantage of load balancing, failover features: configure **`oranfstab`**:



```
server: db-dbnasXXXX.cern.ch
path: 10.16.128.136
path: 10.16.128.200
export: /ORA/dbs03/CMODE mount: /ORA/dbs03/CMODE
export: /ORA/dbs02/CMODE mount: /ORA/dbs02/CMODE
```

- NFS v4, v4.1 still not supported [ID 1087430.1]
  - automount also not supported
  - Above applies to 11g, Oracle12c supports nfsv4



dnfs

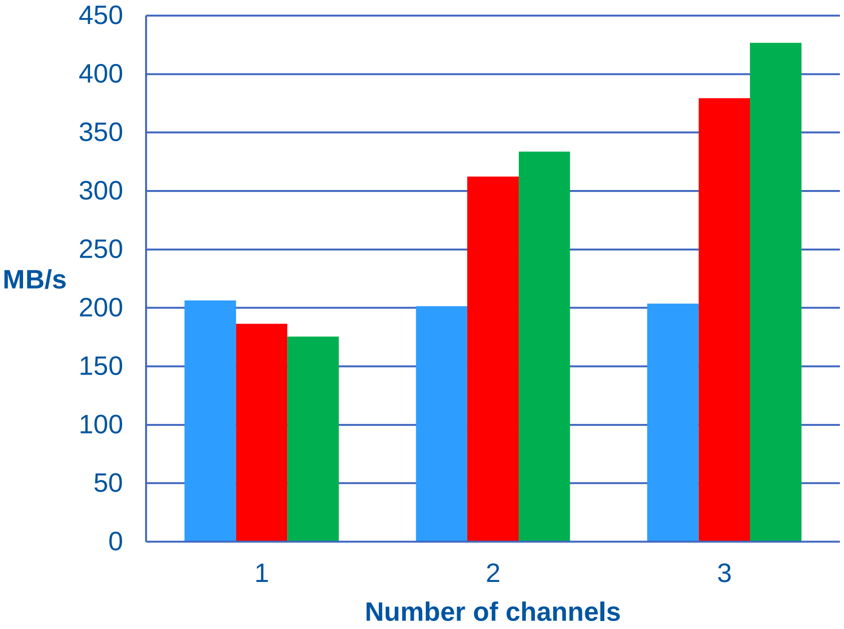
Same operation done with knfs and dnfs

knfs

Oracle directNFS (II)

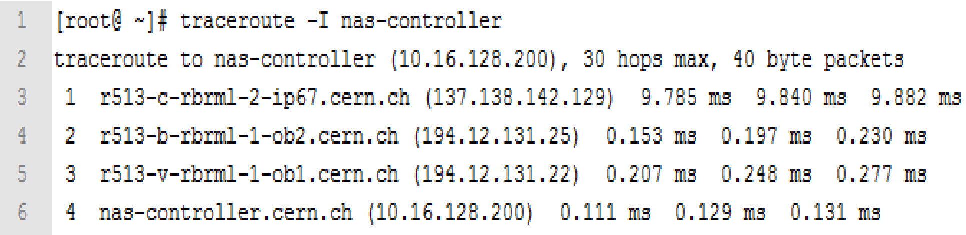
- Mount Options for Oracle files when used with NFS on NAS devices [ID 359515.1]
  - RMAN backups for disk backups kernel NFS [ID 1117597.1]
  - Linux/NetApp: RHEL/SUSE Setup Recommendations for NetApp Filer Storage (Doc ID 279393.1)

RMAN backup to disk*



*Ontap 8.1.1. Fas6240, 72x 3TB SATA disks.

Backup to disk repository in public network (mtu=1500)



```
[root@ ~]# traceroute -I nas-controller
traceroute to nas-controller (10.16.128.200), 30 hops max, 40 byte packets
 1  r513-c-rbrml-2-ip67.cern.ch (137.138.142.129)  9.785 ms  9.840 ms  9.882 ms
 2  r513-b-rbrml-1-ob2.cern.ch (194.12.131.25)  0.153 ms  0.197 ms  0.230 ms
 3  r513-v-rbrml-1-ob1.cern.ch (194.12.131.22)  0.207 ms  0.248 ms  0.277 ms
 4  nas-controller.cern.ch (10.16.128.200)  0.111 ms  0.129 ms  0.131 ms
```

Agenda

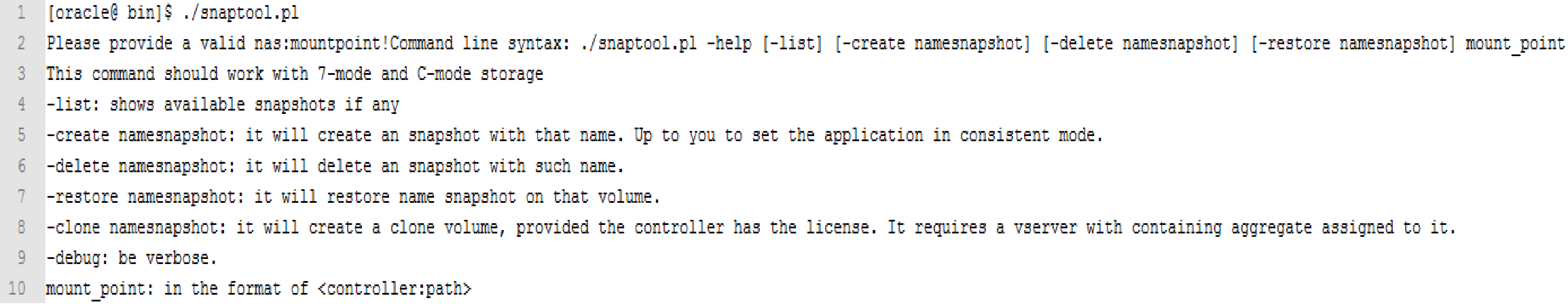
- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
  - Flash cache
  - Flash pool
- Data motion
- Snapshots
- Clonning in Oracle12c
- Backup to disk
- direcNFS
- Monitoring
  - In-house tools
  - Netapp tools
- Conclusions

In-house tools

- Main aim is to allow access to the storage for our DBAs and system admins.
- Based on ZAPI (download NMSDK from NOW), programmed in Perl and Bash about 5000 lines of code
- All work on C-mode or 7-mode, no need to know how to connect to the controllers or ONTAP commands

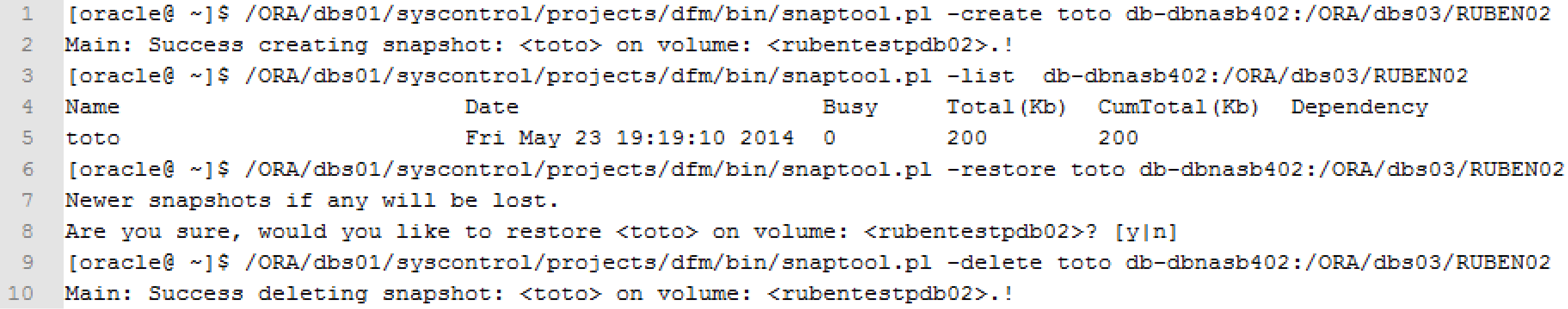
In-house tool: snaptool.pl

- create, list, delete, clone, restore…



```
[oracle@ bin]$ ./snaptool.pl
Please provide a valid nas:mountpoint!Command line syntax: ./snaptool.pl -help [-list] [-create namesnapshot] [-delete namesnapshot] [-restore namesnapshot] mount_point
This command should work with 7-mode and C-mode storage
-list: shows available snapshots if any
-create namesnapshot: it will create an snapshot with that name. Up to you to set the application in consistent mode.
-delete namesnapshot: it will delete an snapshot with such name.
-restore namesnapshot: it will restore name snapshot on that volume.
-clone namesnapshot: it will create a clone volume, provided the controller has the license. It requires a vserver with containing aggregate assigned to it.
-debug: be verbose.
mount_point: in the format of <controller:path>
```

- e.g.

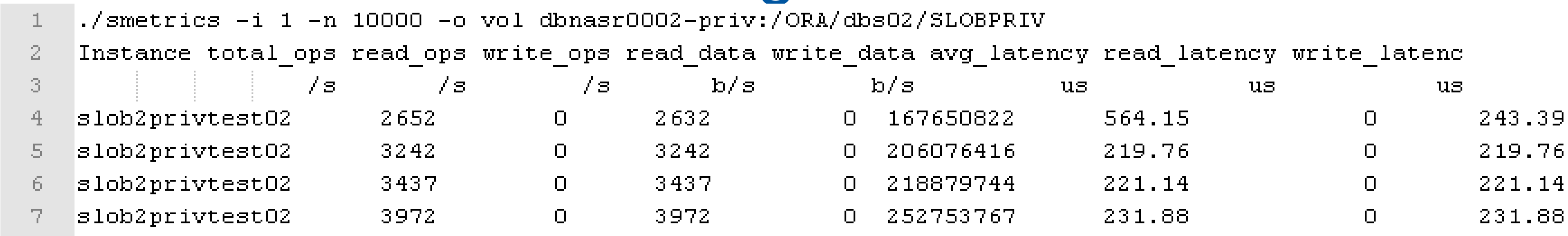


```
[oracle@ ~]$ /ORA/dbs01/syscontrol/projects/dfm/bin/snaptool.pl -create toto db-dbnasb402:/ORA/dbs03/RUBEN02
Main: Success creating snapshot: <toto> on volume: <rubentestpdb02>.!
[oracle@ ~]$ /ORA/dbs01/syscontrol/projects/dfm/bin/snaptool.pl -list  db-dbnasb402:/ORA/dbs03/RUBEN02
Name                             Date                        Busy     Total(Kb)   CumTotal(Kb)   Dependency
toto                             Fri May 23 19:19:10 2014    0        200         200
[oracle@ ~]$ /ORA/dbs01/syscontrol/projects/dfm/bin/snaptool.pl -restore toto db-dbnasb402:/ORA/dbs03/RUBEN02
Newer snapshots if any will be lost.
Are you sure, would you like to restore <toto> on volume: <rubentestpdb02>? [y|n]
[oracle@ ~]$ /ORA/dbs01/syscontrol/projects/dfm/bin/snaptool.pl -delete toto db-dbnasb402:/ORA/dbs03/RUBEN02
Main: Success deleting snapshot: <toto> on volume: <rubentestpdb02>.!
```

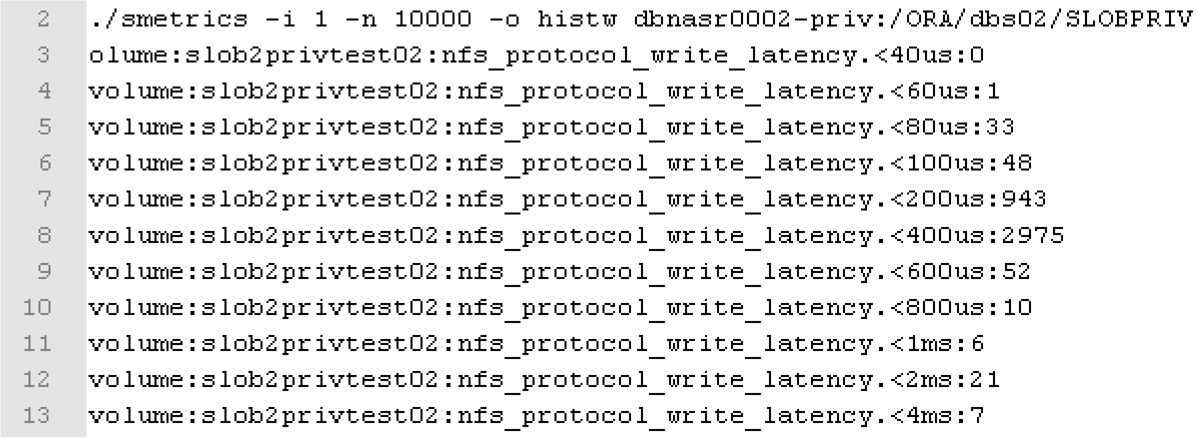
- API available programmatically

In-house tool: smetrics

- Check online statistics of a particular file system or controller serving it
- Volume stats & histograms:



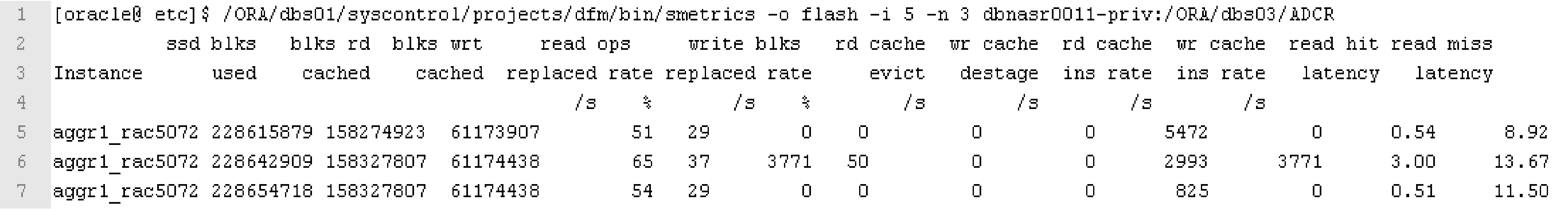
```
./smetrics -i 1 -n 10000 -o vol dbnasr0002-priv:/ORA/dbs02/SLOBPRIV
Instance total_ops read_ops write_ops read_data write_data avg_latency read_latency write_latenc
                /s       /s        /s       b/s         b/s          us           us           us
slob2privtest02      2652        0      2632          0  167650822      564.15            0       243.39
slob2privtest02      3242        0      3242          0  206076416      219.76            0       219.76
slob2privtest02      3437        0      3437          0  218879744      221.14            0       221.14
slob2privtest02      3972        0      3972          0  252753767      231.88            0       231.88
```



```
./smetrics -i 1 -n 10000 -o histw dbnasr0002-priv:/ORA/dbs02/SLOBPRIV
olume:slob2privtest02:nfs_protocol_write_latency.<40us:0
volume:slob2privtest02:nfs_protocol_write_latency.<60us:1
volume:slob2privtest02:nfs_protocol_write_latency.<80us:33
volume:slob2privtest02:nfs_protocol_write_latency.<100us:48
volume:slob2privtest02:nfs_protocol_write_latency.<200us:943
volume:slob2privtest02:nfs_protocol_write_latency.<400us:2975
volume:slob2privtest02:nfs_protocol_write_latency.<600us:52
volume:slob2privtest02:nfs_protocol_write_latency.<800us:10
volume:slob2privtest02:nfs_protocol_write_latency.<1ms:6
volume:slob2privtest02:nfs_protocol_write_latency.<2ms:21
volume:slob2privtest02:nfs_protocol_write_latency.<4ms:7
```

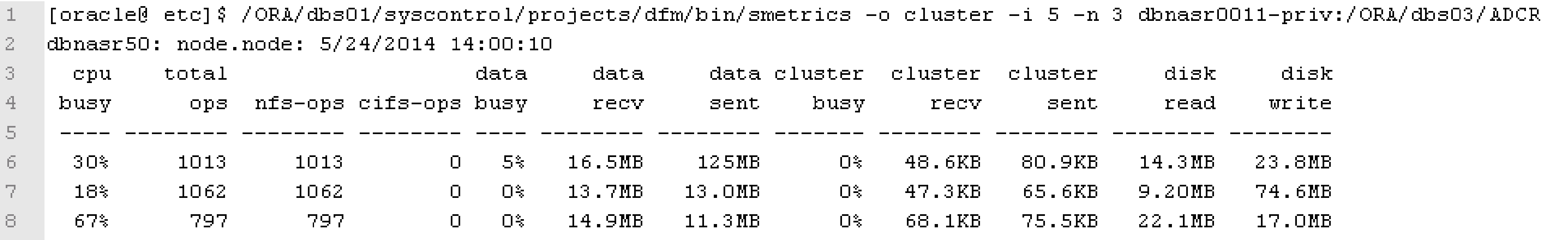
In-house tool: smetrics (II)

- But also SSD consumption per aggregate or vol



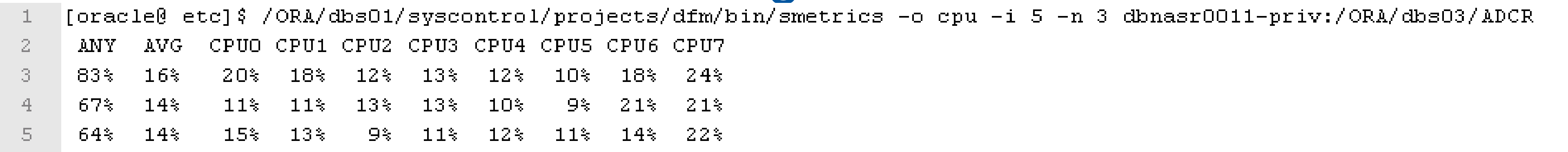
```
[oracle@ etc]$ /ORA/dbs01/syscontrol/projects/dfm/bin/smetrics -o flash -i 5 -n 3 dbnasr0011-priv:/ORA/dbs03/ADCR
          ssd blks   blks rd  blks wrt     read ops     write blks   rd cache  wr cache  rd cache  wr cache  read hit read miss
Instance      used    cached    cached  replaced rate replaced rate     evict   destage  ins rate  ins rate   latency   latency
                                              /s    %       /s    %        /s        /s        /s        /s
aggr1_rac5072 228615879 158274923  61173907        51   29         0    0         0         0      5472         0      0.54      8.92
aggr1_rac5072 228642909 158327807  61174438        65   37     3771   50         0         0      2993      3771      3.00     13.67
aggr1_rac5072 228654718 158327807  61174438        54   29         0    0         0         0       825         0      0.51     11.50
```

- Cluster view:



```
[oracle@ etc]$ /ORA/dbs01/syscontrol/projects/dfm/bin/smetrics -o cluster -i 5 -n 3 dbnasr0011-priv:/ORA/dbs03/ADCR
dbnasr50: node.node: 5/24/2014 14:00:10
  cpu     total                     data      data      data  cluster   cluster   cluster      disk      disk
 busy       ops   nfs-ops  cifs-ops busy      recv      sent     busy      recv      sent      read     write
 ---- --------- --------- --------- ---- --------- --------- -------- --------- --------- --------- ---------
  30%      1013      1013         0   5%    16.5MB     125MB       0%    48.6KB    80.9KB    14.3MB    23.8MB
  18%      1062      1062         0   0%    13.7MB    13.0MB       0%    47.3KB    65.6KB    9.20MB    74.6MB
  67%       797       797         0   0%    14.9MB    11.3MB       0%    68.1KB    75.5KB    22.1MB    17.0MB
```

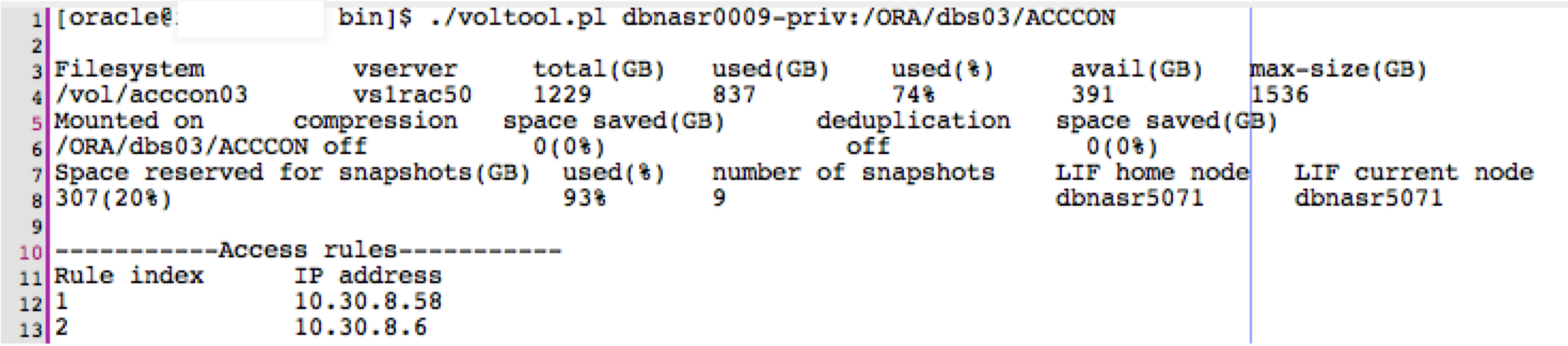
- CPU of controller serving data:



```
[oracle@ etc]$ /ORA/dbs01/syscontrol/projects/dfm/bin/smetrics -o cpu -i 5 -n 3 dbnasr0011-priv:/ORA/dbs03/ADCR
 ANY  AVG  CPU0 CPU1 CPU2 CPU3 CPU4 CPU5 CPU6 CPU7
 83%  16%   20%  18%  12%  13%  12%  10%  18%  24%
 67%  14%   11%  11%  13%  13%  10%   9%  21%  21%
 64%  14%   15%  13%   9%  11%  12%  11%  14%  22%
```

In-house tool: voltool.pl

- Provides information about the volume:



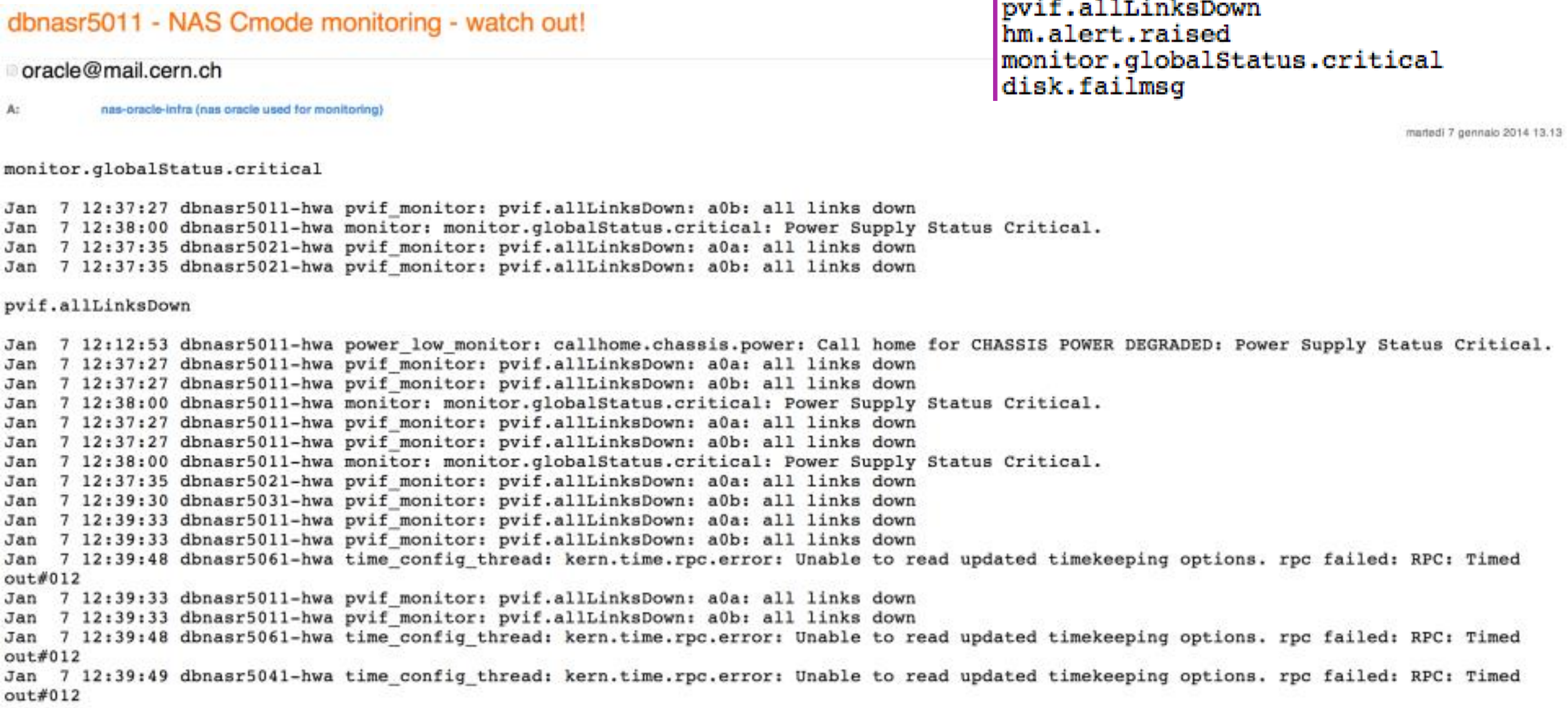
```
[oracle@[redacted] bin]$ ./voltool.pl dbnasr0009-priv:/ORA/dbs03/ACCCON

Filesystem          vserver     total(GB)   used(GB)    used(%)     avail(GB)   max-size(GB)
/vol/acccon03       vs1rac50    1229        837         74%         391         1536
Mounted on      compression    space saved(GB)      deduplication    space saved(GB)
/ORA/dbs03/ACCCON off           0(0%)                 off              0(0%)
Space reserved for snapshots(GB)  used(%)   number of snapshots   LIF home node    LIF current node
307(20%)                          93%       9                     dbnasr5071       dbnasr5071

-----------Access rules-----------
Rule index      IP address
1               10.30.8.58
2               10.30.8.6
```

In-house tool: centralised logging

- `rsyslog` configured for clusters and switches
- Tool allows to regex by type of alert,

It sends emails when a condition is detected:

```
wafl.vol.full
wafl.vol.autoSize.fail
wafl.vol.outOfInodes
wafl.volmove.destination.amd.corrupt
wafl.vvol.exceeded.maxvolsize
pvif.allLinksDown
hm.alert.raised
monitor.globalStatus.critical
disk.failmsg
```



dbnasr5011 - NAS Cmode monitoring - watch out!

oracle@mail.cern.ch

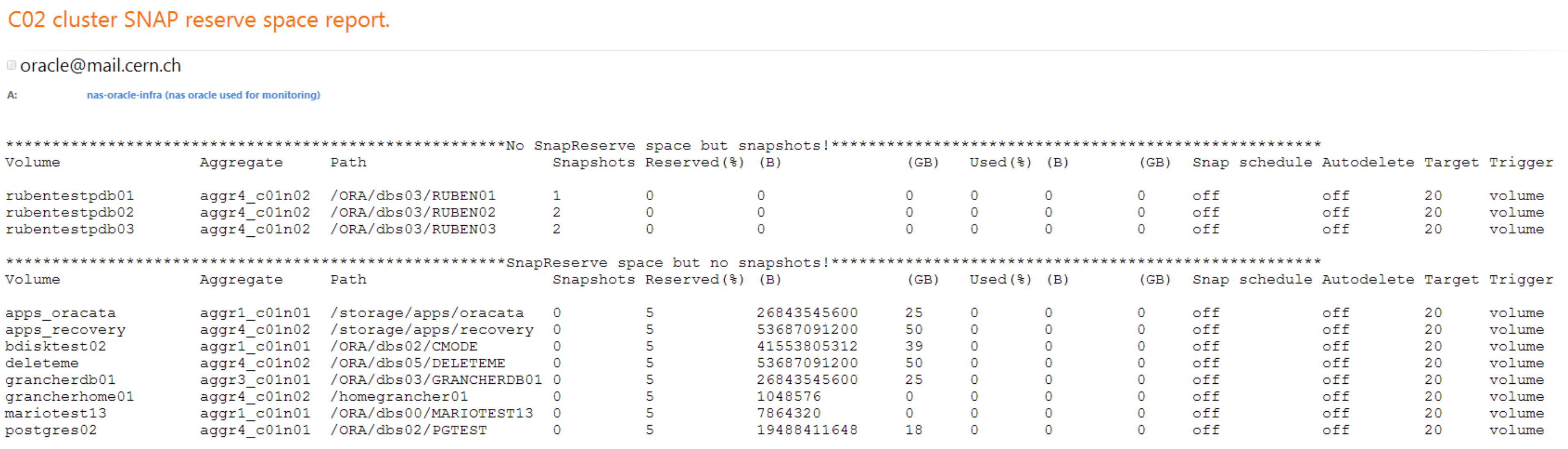
A: nas-oracle-infra (nas oracle used for monitoring)

martedi 7 gennaio 2014 13.13

```
monitor.globalStatus.critical

Jan  7 12:37:27 dbnasr5011-hwa pvif_monitor: pvif.allLinksDown: a0b: all links down
Jan  7 12:38:00 dbnasr5011-hwa monitor: monitor.globalStatus.critical: Power Supply Status Critical.
Jan  7 12:37:35 dbnasr5021-hwa pvif_monitor: pvif.allLinksDown: a0a: all links down
Jan  7 12:37:35 dbnasr5021-hwa pvif_monitor: pvif.allLinksDown: a0b: all links down

pvif.allLinksDown

Jan  7 12:12:53 dbnasr5011-hwa power_low_monitor: callhome.chassis.power: Call home for CHASSIS POWER DEGRADED: Power Supply Status Critical.
Jan  7 12:37:27 dbnasr5011-hwa pvif_monitor: pvif.allLinksDown: a0a: all links down
Jan  7 12:37:27 dbnasr5011-hwa pvif_monitor: pvif.allLinksDown: a0b: all links down
Jan  7 12:38:00 dbnasr5011-hwa monitor: monitor.globalStatus.critical: Power Supply Status Critical.
Jan  7 12:37:27 dbnasr5011-hwa pvif_monitor: pvif.allLinksDown: a0a: all links down
Jan  7 12:37:27 dbnasr5011-hwa pvif_monitor: pvif.allLinksDown: a0b: all links down
Jan  7 12:38:00 dbnasr5011-hwa monitor: monitor.globalStatus.critical: Power Supply Status Critical.
Jan  7 12:37:35 dbnasr5021-hwa pvif_monitor: pvif.allLinksDown: a0a: all links down
Jan  7 12:39:30 dbnasr5031-hwa pvif_monitor: pvif.allLinksDown: a0b: all links down
Jan  7 12:39:33 dbnasr5011-hwa pvif_monitor: pvif.allLinksDown: a0a: all links down
Jan  7 12:39:33 dbnasr5011-hwa pvif_monitor: pvif.allLinksDown: a0b: all links down
Jan  7 12:39:48 dbnasr5061-hwa time_config_thread: kern.time.rpc.error: Unable to read updated timekeeping options. rpc failed: RPC: Timed
out#012
Jan  7 12:39:33 dbnasr5011-hwa pvif_monitor: pvif.allLinksDown: a0a: all links down
Jan  7 12:39:33 dbnasr5011-hwa pvif_monitor: pvif.allLinksDown: a0b: all links down
Jan  7 12:39:48 dbnasr5061-hwa time_config_thread: kern.time.rpc.error: Unable to read updated timekeeping options. rpc failed: RPC: Timed
out#012
Jan  7 12:39:49 dbnasr5041-hwa time_config_thread: kern.time.rpc.error: Unable to read updated timekeeping options. rpc failed: RPC: Timed
out#012
```

In-house logging: reporting

- Reports are not available on OUM 6.1
- It reports anomalies in the usage of snap reserved space



C02 cluster SNAP reserve space report.

□ oracle@mail.cern.ch

A: nas-oracle-infra (nas oracle used for monitoring)

```
******************************************************No SnapReserve space but snapshots!*****************************************************
Volume               Aggregate     Path                    Snapshots Reserved(%) (B)             (GB)   Used(%)  (B)       (GB)  Snap schedule Autodelete Target Trigger

rubentestpdb01       aggr4_c01n02  /ORA/dbs03/RUBEN01      1         0           0               0      0       0         0     off           off        20     volume
rubentestpdb02       aggr4_c01n02  /ORA/dbs03/RUBEN02      2         0           0               0      0       0         0     off           off        20     volume
rubentestpdb03       aggr4_c01n02  /ORA/dbs03/RUBEN03      2         0           0               0      0       0         0     off           off        20     volume

******************************************************SnapReserve space but no snapshots!*****************************************************
Volume               Aggregate     Path                    Snapshots Reserved(%) (B)             (GB)   Used(%)  (B)       (GB)  Snap schedule Autodelete Target Trigger

apps_oracata         aggr1_c01n01  /storage/apps/oracata   0         5           26843545600     25     0       0         0     off           off        20     volume
apps_recovery        aggr4_c01n02  /storage/apps/recovery  0         5           53687091200     50     0       0         0     off           off        20     volume
bdisktest02          aggr1_c01n01  /ORA/dbs02/CMODE        0         5           41553805312     39     0       0         0     off           off        20     volume
deleteme             aggr4_c01n02  /ORA/dbs05/DELETEME     0         5           53687091200     50     0       0         0     off           off        20     volume
grancherdb01         aggr3_c01n01  /ORA/dbs03/GRANCHERDB01 0         5           26843545600     25     0       0         0     off           off        20     volume
grancherhome01       aggr4_c01n02  /homegrancher01         0         5           1048576         0      0       0         0     off           off        20     volume
mariotest13          aggr1_c01n01  /ORA/dbs00/MARIOTEST13  0         5           7864320         0      0       0         0     off           off        20     volume
postgres02           aggr4_c01n01  /ORA/dbs02/PGTEST       0         5           19488411648     18     0       0         0     off           off        20     volume
```

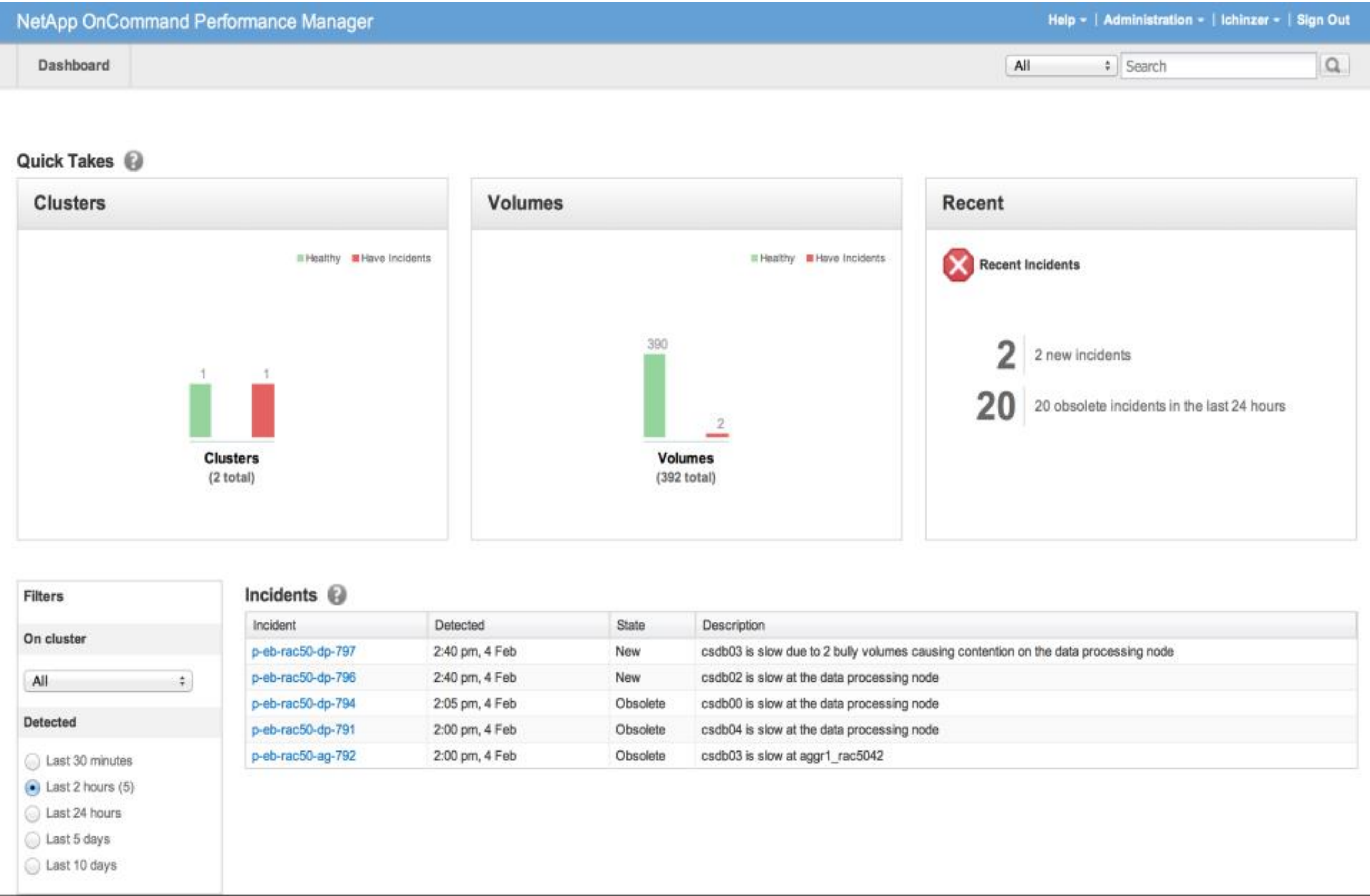
Agenda

- CERN intro
- CERN databases basic description
- Storage evolution using Netapp
- Caching technologies
  - Flash cache
  - Flash pool
- Data motion
- Snapshots
- Clonning in Oracle12c
- Backup to disk
- directNFS
- Monitoring
  - In-house tools
  - Netapp tools
- Conclusions

Netapp monitoring/mgmt tools

- Unified OnCommand Manager 5.2 (linux)
  - Authentication using PAM
  - Extensive use of reporting (in 7-mode)
  - Work for both 7-mode and C-mode
  - Performance management console (performance counters display)
  - Alarms
- OnCommand Performance Manager (OPM) & OnCommand Unified Manager (OUM)
  - Used for C-mode
  - Virtual machine (VM) that runs on a VMware ESX or ESXi Server
- System Manager
  - We use it mainly to check setups
- My Autosupport at NOW website

Netapp OPM 1.0



NetApp OnCommand Performance Manager

Help ▾ | Administration ▾ | lchinzer ▾ | Sign Out
Dashboard
All
Search

Quick Takes

Clusters

🟩 Healthy 🟥 Have Incidents
1
1

Clusters (2 total)

Volumes

🟩 Healthy 🟥 Have Incidents
390
2

Volumes (392 total)

Recent

Recent Incidents

2 | 2 new incidents

20 | 20 obsolete incidents in the last 24 hours

Filters

On cluster

All

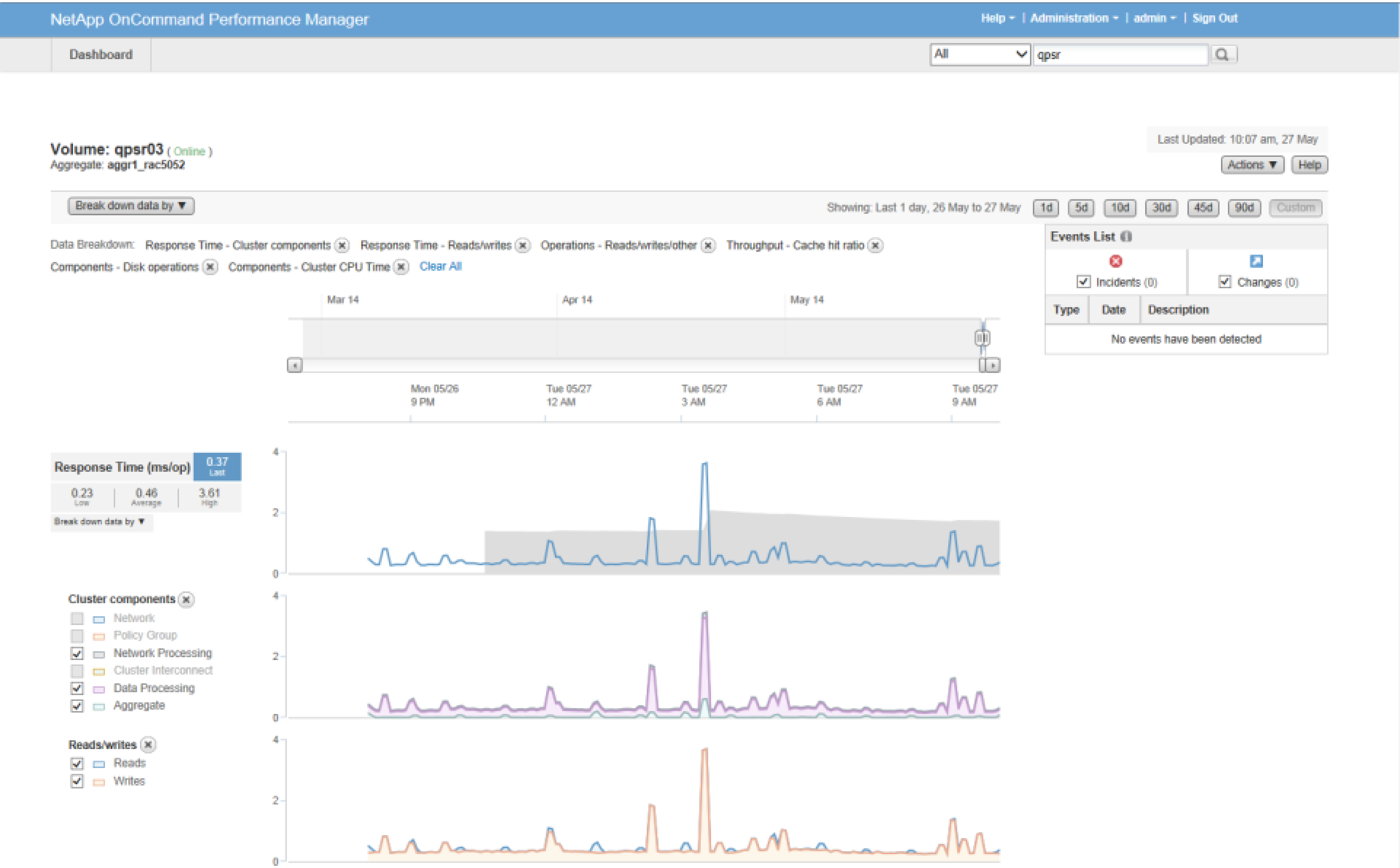
Detected

○ Last 30 minutes
● Last 2 hours (5)
○ Last 24 hours
○ Last 5 days
○ Last 10 days

Incidents

Incident	Detected	State	Description
p-eb-rac50-dp-797	2:40 pm, 4 Feb	New	csdb03 is slow due to 2 bully volumes causing contention on the data processing node
p-eb-rac50-dp-796	2:40 pm, 4 Feb	New	csdb02 is slow at the data processing node
p-eb-rac50-dp-794	2:05 pm, 4 Feb	Obsolete	csdb00 is slow at the data processing node
p-eb-rac50-dp-791	2:00 pm, 4 Feb	Obsolete	csdb04 is slow at the data processing node
p-eb-rac50-ag-792	2:00 pm, 4 Feb	Obsolete	csdb03 is slow at aggr1_rac5042

Netapp OPM 1.0



Volume: qpsr03 (Online)
Aggregate: aggr1_rac5052

Last Updated: 10:07 am, 27 May

Showing: Last 1 day, 26 May to 27 May

Data Breakdown: Response Time - Cluster components ☒ Response Time - Reads/writes ☒ Operations - Reads/writes/other ☒ Throughput - Cache hit ratio ☒ Components - Disk operations ☒ Components - Cluster CPU Time ☒ Clear All

Events List

☑ Incidents (0) ☑ Changes (0)

Type	Date	Description
		No events have been detected

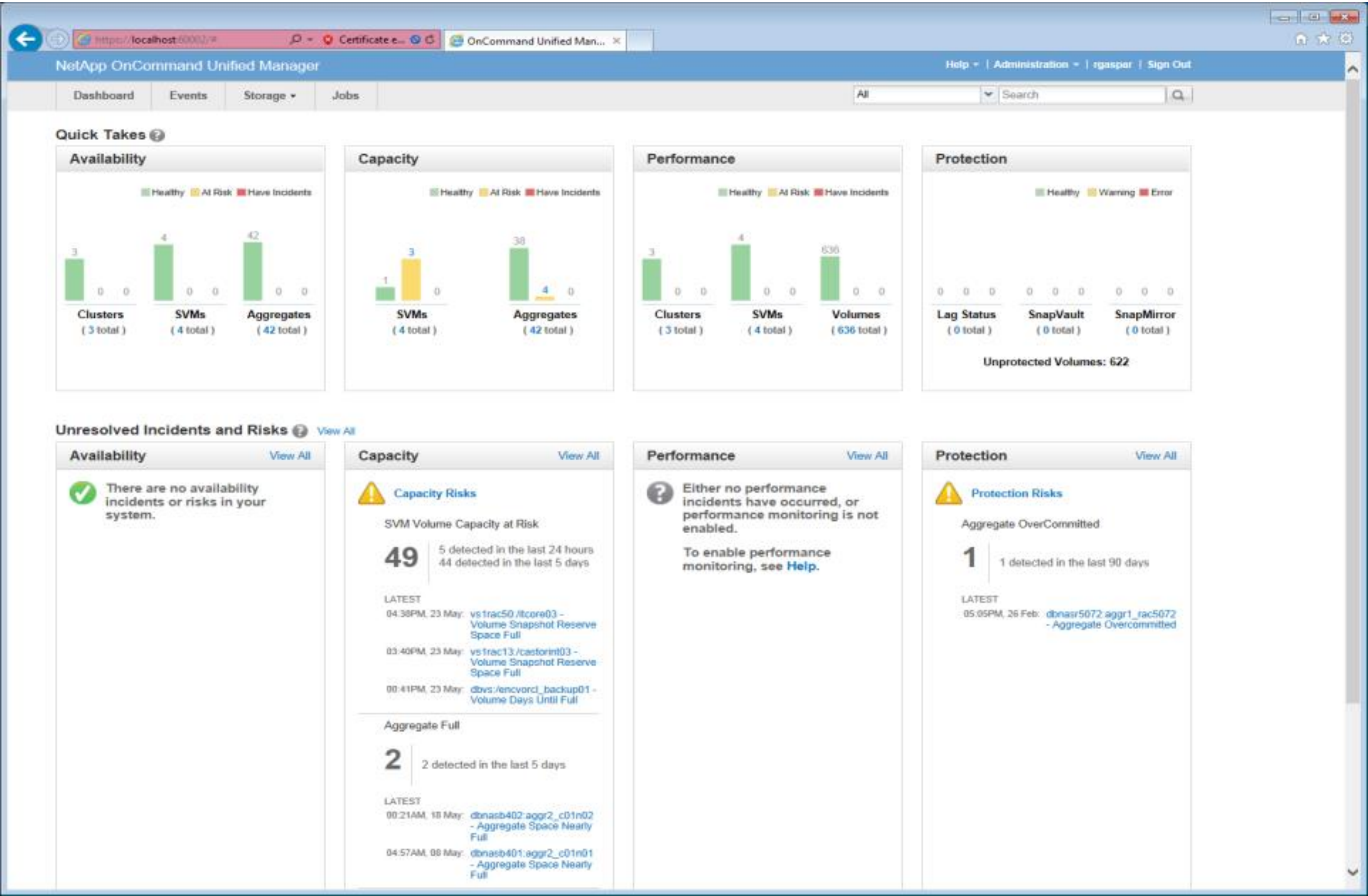
Response Time (ms/op) 0.37 Last

0.23 Low | 0.46 Average | 3.61 High

Cluster components ☒
- □ Network
- □ Policy Group
- ☑ Network Processing
- □ Cluster Interconnect
- ☑ Data Processing
- ☑ Aggregate

Reads/writes ☒
- ☑ Reads
- ☑ Writes

Netapp OnCommand UM 6.1


https://localhost:60002/#
Certificate e...
OnCommand Unified Man...
NetApp OnCommand Unified Manager
Help ▾ | Administration ▾ | rgaspar | Sign Out
Dashboard
Events
Storage ▾
Jobs
All
Search

Quick Takes

Availability

🟩 Healthy 🟨 At Risk 🟥 Have Incidents
3 0 0
4 0 0
42 0 0

Clusters (3 total) SVMs (4 total) Aggregates (42 total)

Capacity

🟩 Healthy 🟨 At Risk 🟥 Have Incidents
1 3 0
38 4 0

SVMs (4 total) Aggregates (42 total)

Performance

🟩 Healthy 🟨 At Risk 🟥 Have Incidents
3 0 0
4 0 0
636 0 0

Clusters (3 total) SVMs (4 total) Volumes (636 total)

Protection

🟩 Healthy 🟨 Warning 🟥 Error
0 0 0
0 0 0
0 0 0

Lag Status (0 total) SnapVault (0 total) SnapMirror (0 total)

Unprotected Volumes: 622

Unresolved Incidents and Risks View All

Availability View All

There are no availability incidents or risks in your system.

Capacity View All

Capacity Risks

SVM Volume Capacity at Risk

49 5 detected in the last 24 hours
44 detected in the last 5 days

LATEST

04:38PM, 23 May: vs1rac50:/tcore03 - Volume Snapshot Reserve Space Full

03:40PM, 23 May: vs1rac13:/castorint03 - Volume Snapshot Reserve Space Full

00:41PM, 23 May: dbvs:/encvorcl_backup01 - Volume Days Until Full

Aggregate Full

2 2 detected in the last 5 days

LATEST

00:21AM, 18 May: dbnasb402:aggr2_c01n02 - Aggregate Space Nearly Full

04:57AM, 08 May: dbnasb401:aggr2_c01n01 - Aggregate Space Nearly Full

Performance View All

Either no performance incidents have occurred, or performance monitoring is not enabled.

To enable performance monitoring, see Help.

Protection View All

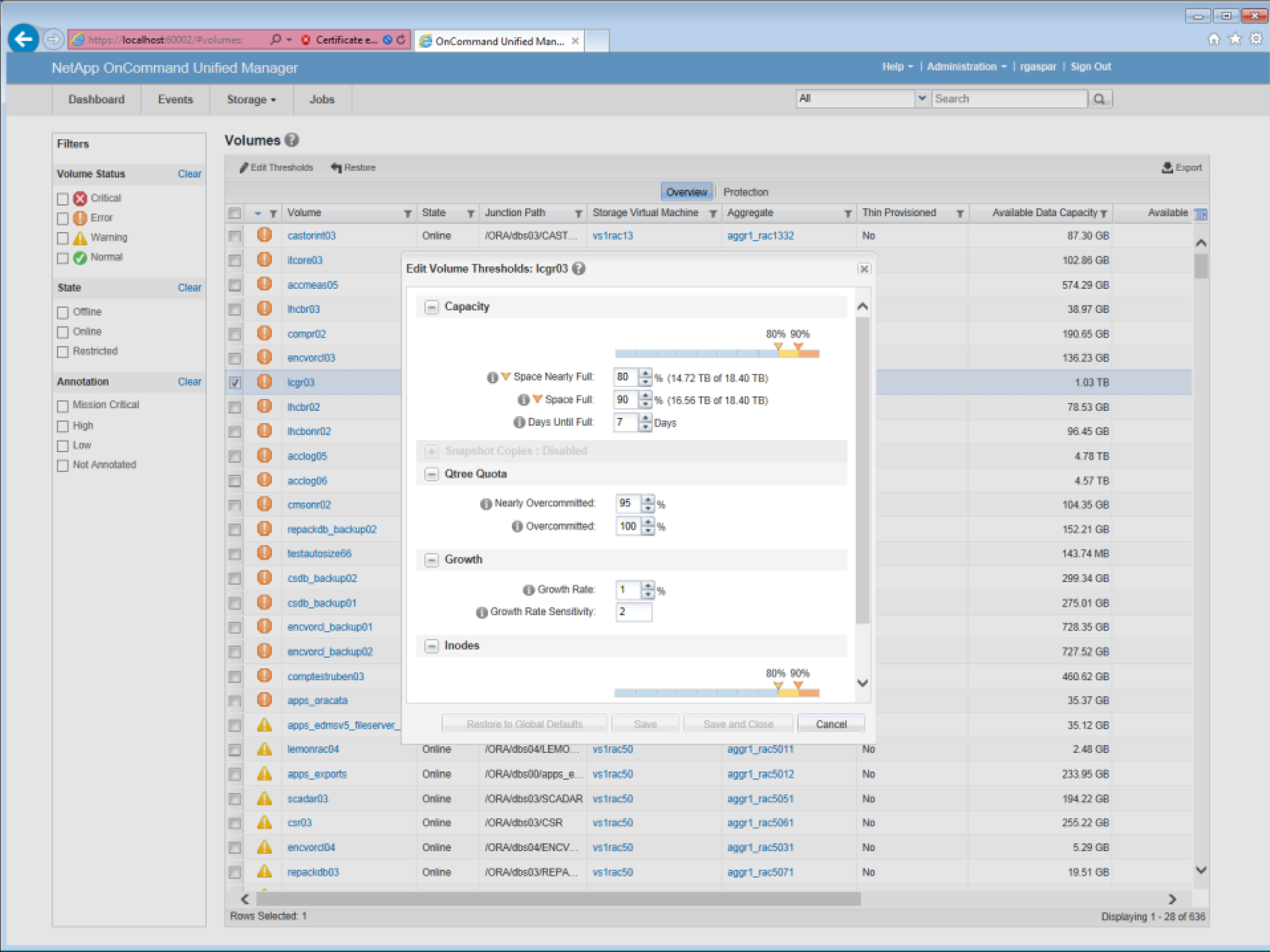
Protection Risks

Aggregate OverCommitted

1 1 detected in the last 90 days

LATEST

05:05PM, 26 Feb: dbnasr5072:aggr1_rac5072 - Aggregate Overcommitted



https://localhost:60002/#volumes: Certificate e... OnCommand Unified Man...

NetApp OnCommand Unified Manager

Help ▾ | Administration ▾ | rgaspar | Sign Out

Dashboard | Events | Storage ▾ | Jobs

All ▾ Search

Filters

Volume Status Clear

- □ Critical
- □ Error
- □ Warning
- □ Normal

State Clear

- □ Offline
- □ Online
- □ Restricted

Annotation Clear

- □ Mission Critical
- □ High
- □ Low
- □ Not Annotated

Volumes

Edit Thresholds | Restore | Export

Overview | Protection

□		Volume	State	Junction Path	Storage Virtual Machine	Aggregate	Thin Provisioned	Available Data Capacity	Available
□	!	castorint03	Online	/ORA/dbs03/CAST...	vs1rac13	aggr1_rac1332	No	87.30 GB	
□	!	itcore03						102.86 GB	
□	!	accmeas05						574.29 GB	
□	!	lhcbr03						38.97 GB	
□	!	compr02						190.65 GB	
□	!	encvorcl03						136.23 GB	
☑	!	lcgr03						1.03 TB	
□	!	lhcbr02						78.53 GB	
□	!	lhcbonr02						96.45 GB	
□	!	acclog05						4.78 TB	
□	!	acclog06						4.57 TB	
□	!	cmsonr02						104.35 GB	
□	!	repackdb_backup02						152.21 GB	
□	!	testautosize66						143.74 MB	
□	!	csdb_backup02						299.34 GB	
□	!	csdb_backup01						275.01 GB	
□	!	encvorcl_backup01						728.35 GB	
□	!	encvorcl_backup02						727.52 GB	
□	!	comptestruben03						460.62 GB	
□	!	apps_oracata						35.37 GB	
□	⚠	apps_edmsv5_fileserver_						35.12 GB	
□	⚠	lemonrac04	Online	/ORA/dbs04/LEMO...	vs1rac50	aggr1_rac5011	No	2.48 GB	
□	⚠	apps_exports	Online	/ORA/dbs00/apps_e...	vs1rac50	aggr1_rac5012	No	233.95 GB	
□	⚠	scadar03	Online	/ORA/dbs03/SCADAR	vs1rac50	aggr1_rac5051	No	194.22 GB	
□	⚠	csr03	Online	/ORA/dbs03/CSR	vs1rac50	aggr1_rac5061	No	255.22 GB	
□	⚠	encvorcl04	Online	/ORA/dbs04/ENCV...	vs1rac50	aggr1_rac5031	No	5.29 GB	
□	⚠	repackdb03	Online	/ORA/dbs03/REPA...	vs1rac50	aggr1_rac5071	No	19.51 GB	

Rows Selected: 1

Displaying 1 - 28 of 636

Edit Volume Thresholds: lcgr03

Capacity

Space Nearly Full: 80 % (14.72 TB of 18.40 TB)

Space Full: 90 % (16.56 TB of 18.40 TB)

Days Until Full: 7 Days

Snapshot Copies : Disabled

Qtree Quota

Nearly Overcommitted: 95 %

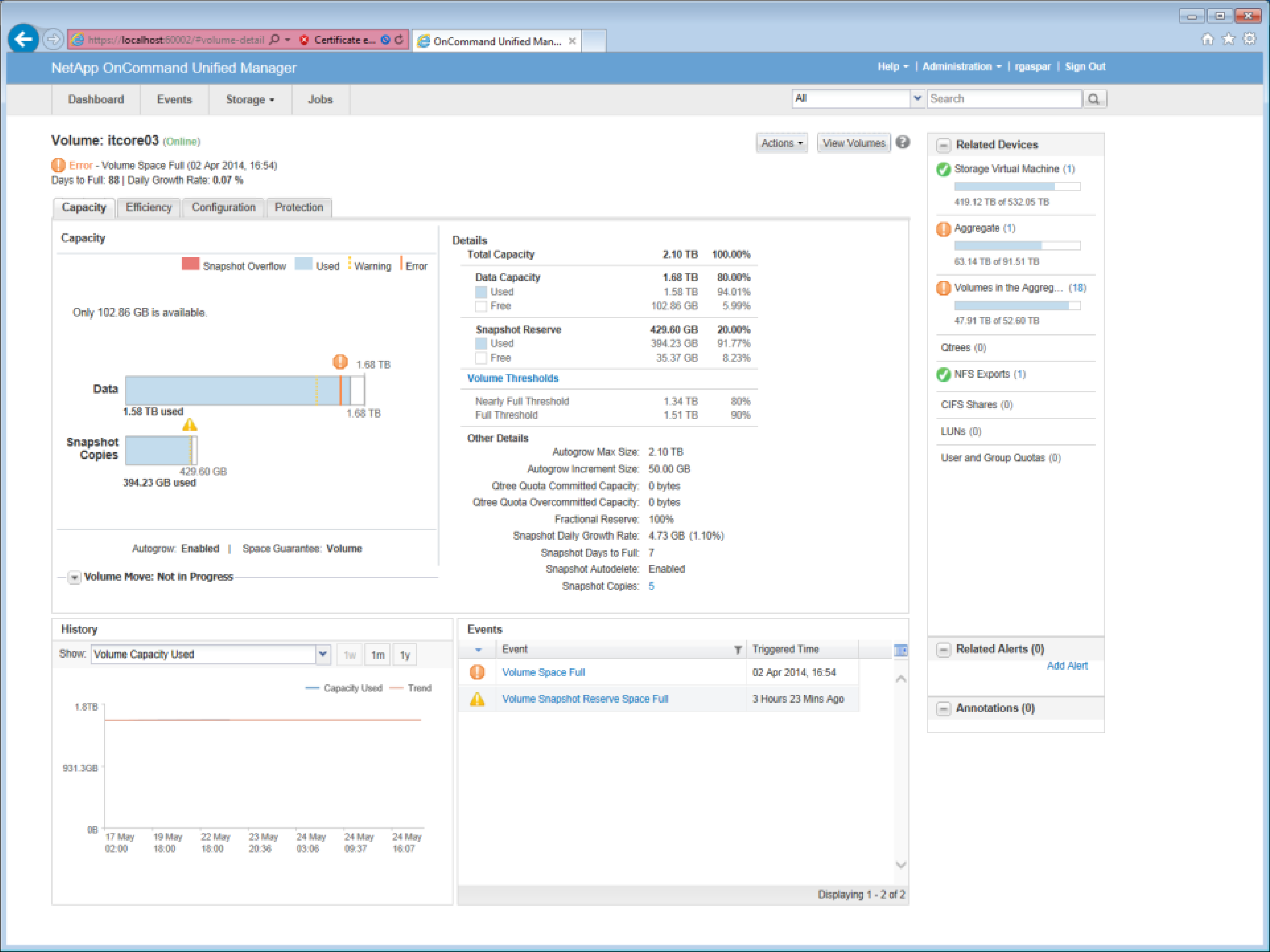
Overcommitted: 100 %

Growth

Growth Rate: 1 %

Growth Rate Sensitivity: 2

Inodes



Volume: itcore03 (Online)

Error - Volume Space Full (02 Apr 2014, 16:54)
Days to Full: 88 | Daily Growth Rate: 0.07 %

Capacity

Only 102.86 GB is available.

Autogrow: Enabled | Space Guarantee: Volume

Volume Move: Not in Progress

Details

Total Capacity	2.10 TB	100.00%
Data Capacity	1.68 TB	80.00%
🟦 Used	1.58 TB	94.01%
□ Free	102.86 GB	5.99%
Snapshot Reserve	429.60 GB	20.00%
🟦 Used	394.23 GB	91.77%
□ Free	35.37 GB	8.23%

Volume Thresholds

Nearly Full Threshold	1.34 TB	80%
Full Threshold	1.51 TB	90%

Other Details

Autogrow Max Size:	2.10 TB
Autogrow Increment Size:	50.00 GB
Qtree Quota Committed Capacity:	0 bytes
Qtree Quota Overcommitted Capacity:	0 bytes
Fractional Reserve:	100%
Snapshot Daily Growth Rate:	4.73 GB (1.10%)
Snapshot Days to Full:	7
Snapshot Autodelete:	Enabled
Snapshot Copies:	5

Related Devices

- Storage Virtual Machine (1) — 419.12 TB of 532.05 TB
- Aggregate (1) — 63.14 TB of 91.51 TB
- Volumes in the Aggreg... (18) — 47.91 TB of 52.60 TB
- Qtrees (0)
- NFS Exports (1)
- CIFS Shares (0)
- LUNs (0)
- User and Group Quotas (0)

Related Alerts (0)

Add Alert

Annotations (0)

History

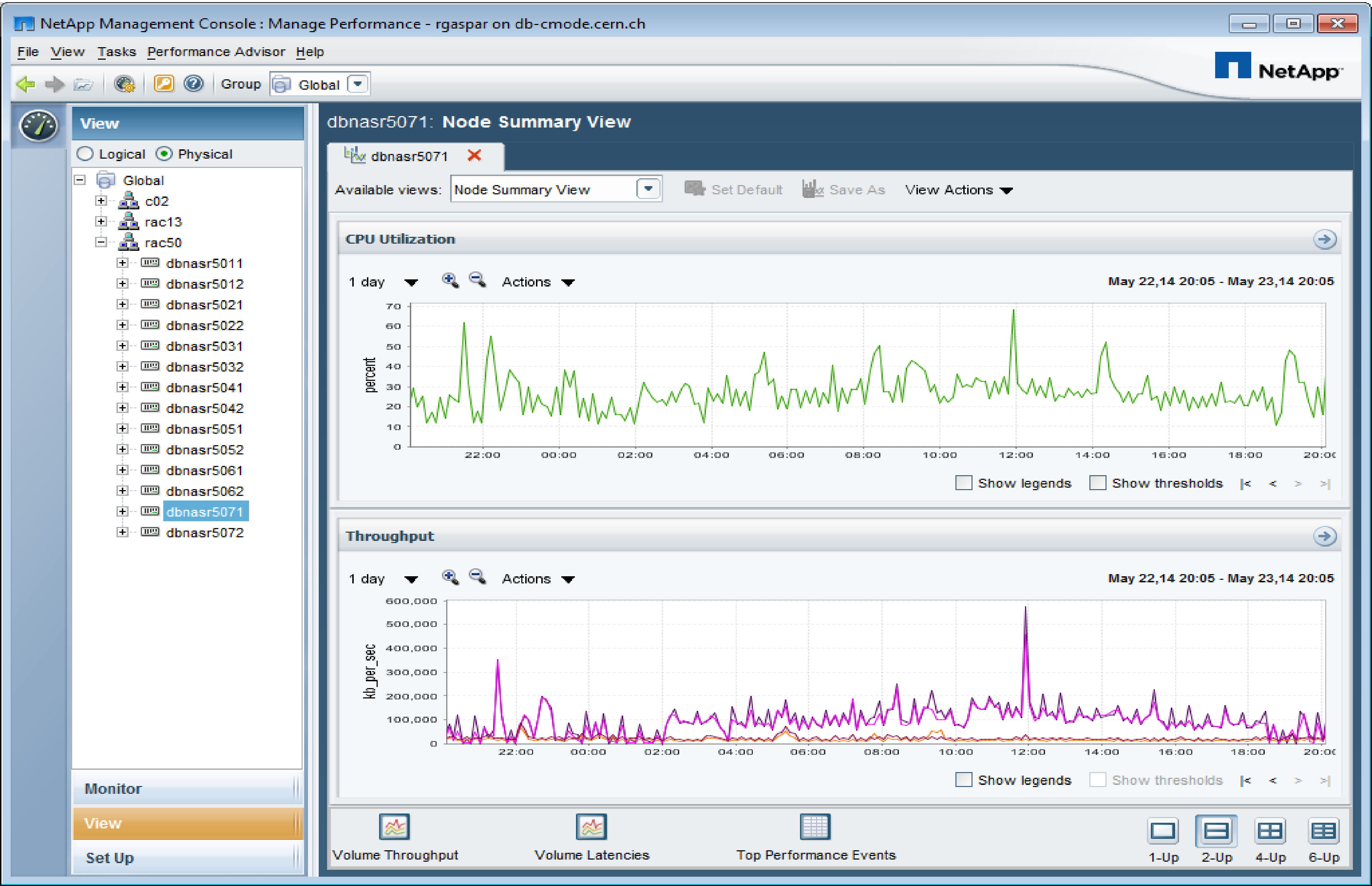
Show: Volume Capacity Used

Events

Event	Triggered Time
Volume Space Full	02 Apr 2014, 16:54
Volume Snapshot Reserve Space Full	3 Hours 23 Mins Ago

Displaying 1 - 2 of 2

Netapp Management console 3.3

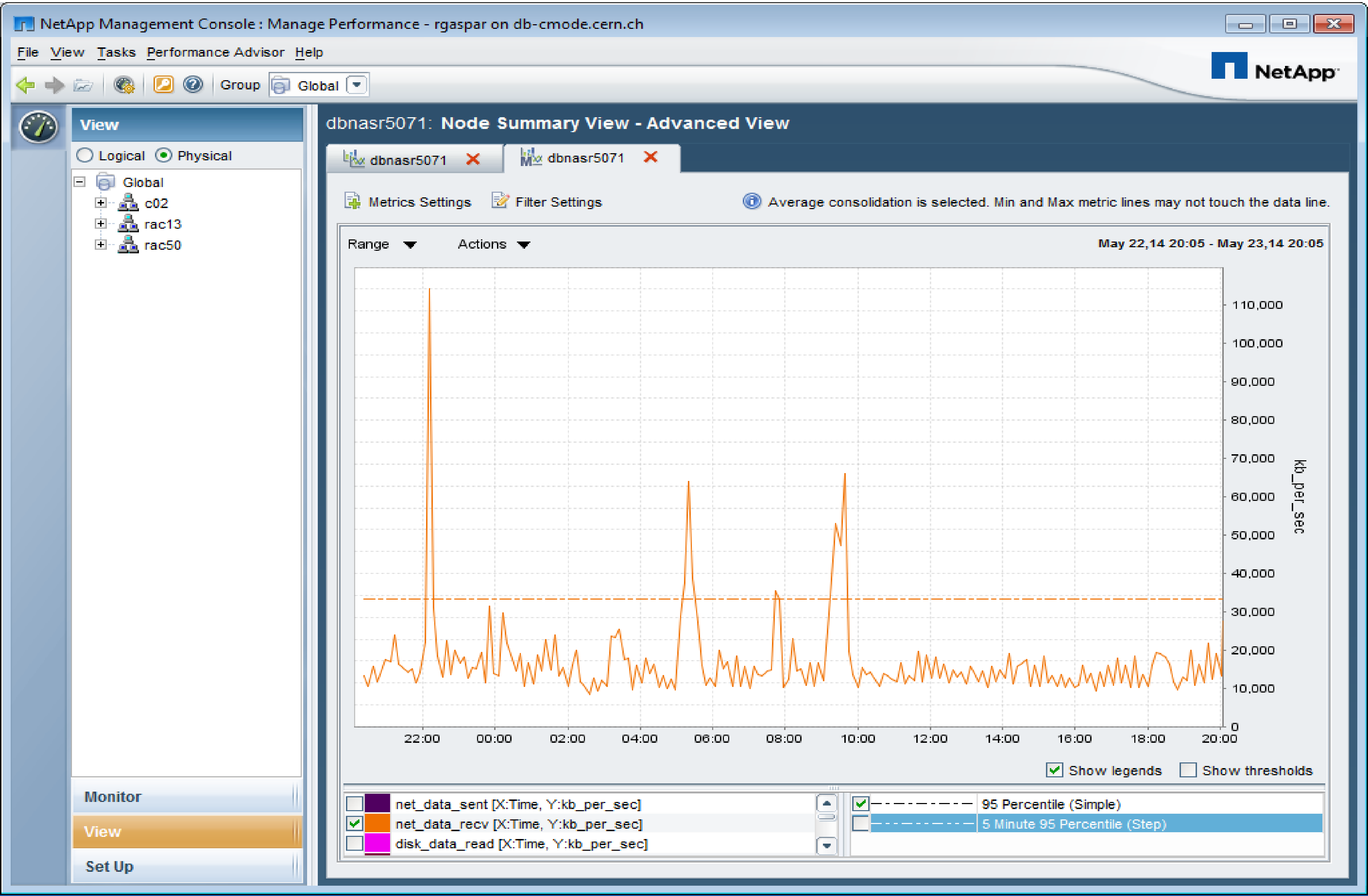


NetApp Management Console : Manage Performance - rgaspar on db-cmode.cern.ch

dbnasr5071: **Node Summary View**

Available views: Node Summary View

Netapp Management console 3.3



NetApp Management Console : Manage Performance - rgaspar on db-cmode.cern.ch

dbnasr5071: **Node Summary View - Advanced View**

Average consolidation is selected. Min and Max metric lines may not touch the data line.

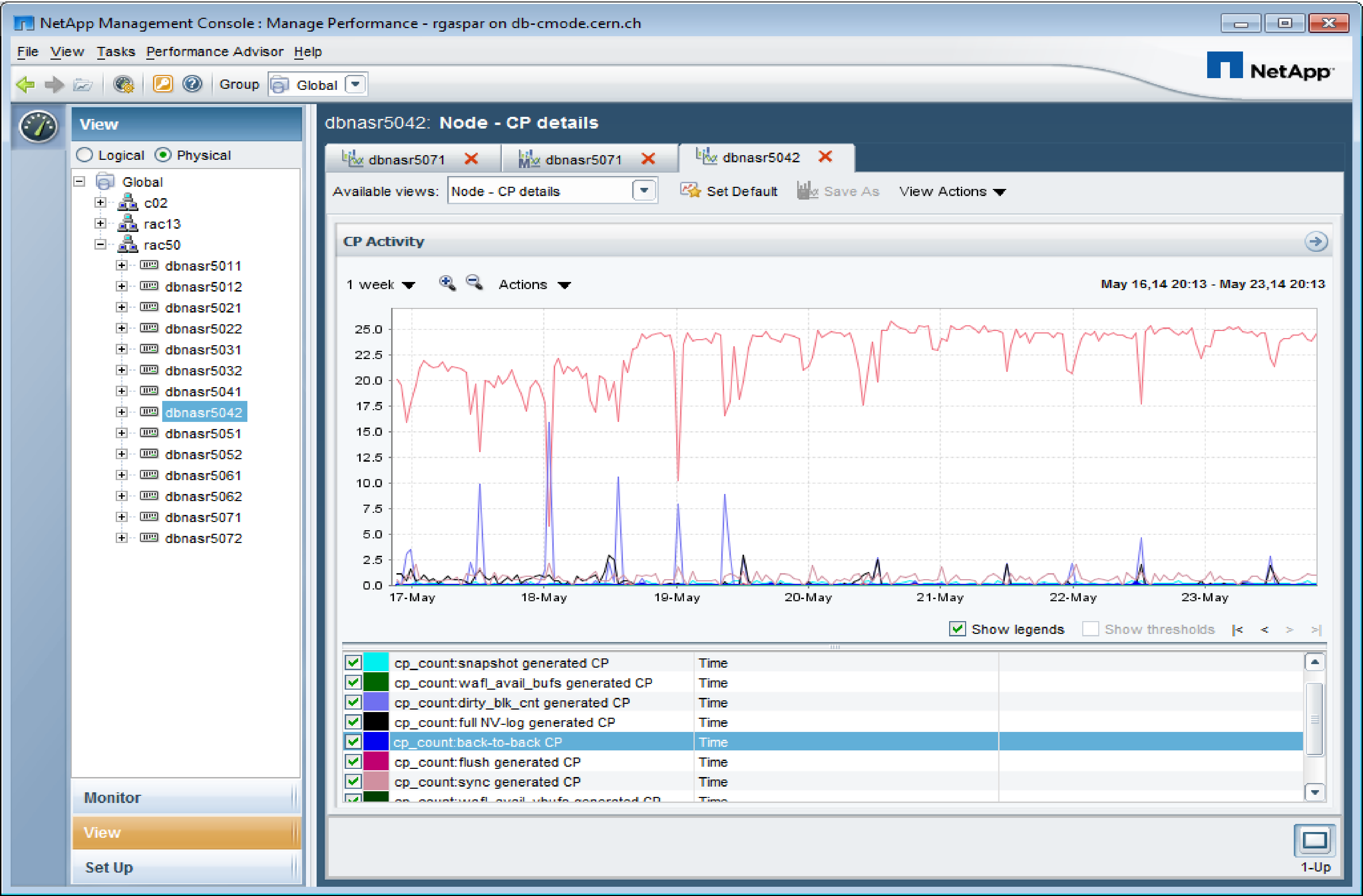
May 22,14 20:05 - May 23,14 20:05

☑ Show legends ☐ Show thresholds

☐ net_data_sent [X:Time, Y:kb_per_sec]
☑ net_data_recv [X:Time, Y:kb_per_sec]
☐ disk_data_read [X:Time, Y:kb_per_sec]

☑ 95 Percentile (Simple)
☐ 5 Minute 95 Percentile (Step)

Netapp Management console 3.3



NetApp Management Console : Manage Performance - rgaspar on db-cmode.cern.ch

dbnasr5042: **Node - CP details**

Available views: Node - CP details

CP Activity

1 week

May 16,14 20:13 - May 23,14 20:13

☑ Show legends □ Show thresholds

☑	🟦	cp_count:snapshot generated CP	Time
☑	🟩	cp_count:wafl_avail_bufs generated CP	Time
☑	🟪	cp_count:dirty_blk_cnt generated CP	Time
☑	⬛	cp_count:full NV-log generated CP	Time
☑	🟦	cp_count:back-to-back CP	Time
☑	🟥	cp_count:flush generated CP	Time
☑	🟥	cp_count:sync generated CP	Time
☑	🟩	cp_count:wafl_avail_vbufs generated CP	Time

Conclusions

- Positive experience so far running on C-mode
- Mid to high end NetApp NAS provide good performance using the FlashPool SSD caching solution
- Flexibility with cluster ONTAP, helps to reduce the investment
- Design of stacks and network access require careful planning

Acknowledgement

- IT-DB colleagues, especially Lisa Azzurra and Miroslav Potocky
- Netapp engineers: Jeffrey Steiner, Nagalingam Karthikeyan, Nicolas Jacquot

?