

CERN Achieves Database Scalability and Performance with Oracle and NetApp

session S319046

Eric Grancher
eric.grancher@cern.ch
CERN IT department

Steve Daniel
steve.daniel@netapp.com
NetApp

<https://edms.cern.ch/document/1093461/1>



Image courtesy of Forschungszentrum Jülich /
Seitenplan, with material from NASA, ESA and
AURA/Caltech

Driving Storage as a Value Center



Reduce Complexity

- Unified infrastructure
 - Combines technology and process seamlessly

Maximize Asset Utilization

- Storage efficiency
 - Protect data while avoiding data duplication
 - Provide multi-use datasets without copying
 - Eliminate duplicate copies of data
 - Reduce power, cooling & space consumption

Control Hidden Costs

- Comprehensive data management
 - Complete data protection
 - Application-level end-to-end provisioning
 - Policy-based automation

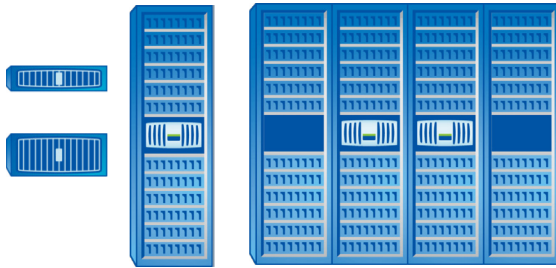
Flexible Storage

A Single, Unified Platform



Single, Unified Storage Platform

Low-to-High Scalability



Multiple Networks



Multiple Protocols

SAN
NAS
iSCSI

Multiple Disks

FC
SATA
SSD

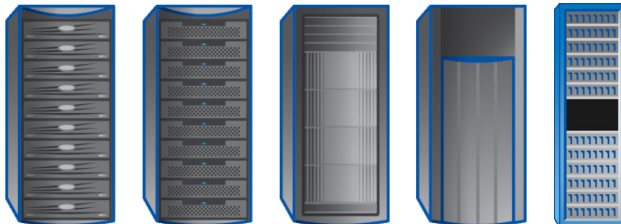
Unified Management

- Same tools and processes: learn once, run everywhere
- Integrated management
- Integrated data protection

Storage Virtualization



Multi-vendor virtualization



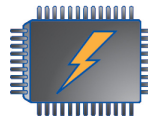
Unified Flash



FlashCache

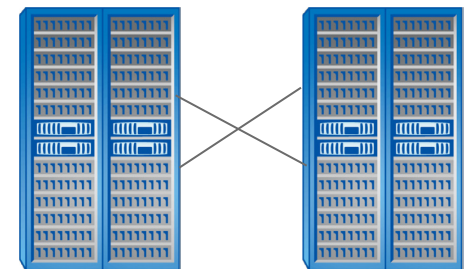


SSD



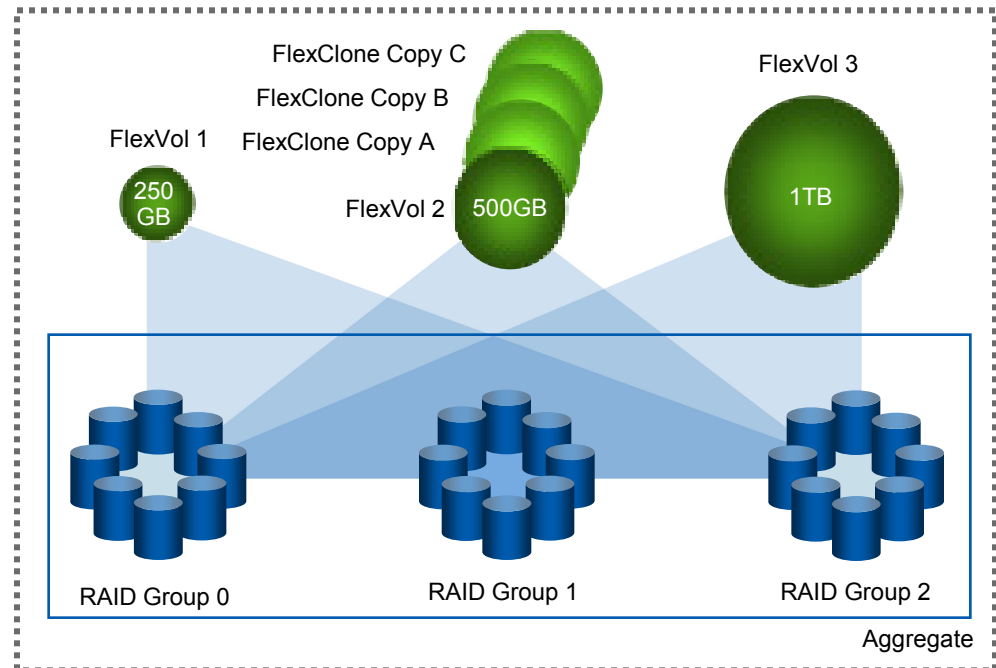
FlexCache

Unified Scale Out



FlexClone Writable Copies

- Application development often requires substantial primary storage space for essential test operations such as platform and upgrade rollouts
- FlexClone® improves storage efficiency for applications that need temporary, writable copies of data volumes
- Creates a virtual “clone” copy of the primary dataset and stores only the data changes between parent volume and clone
- Multiple clones are easily created
- Resulting space savings of 80% or more



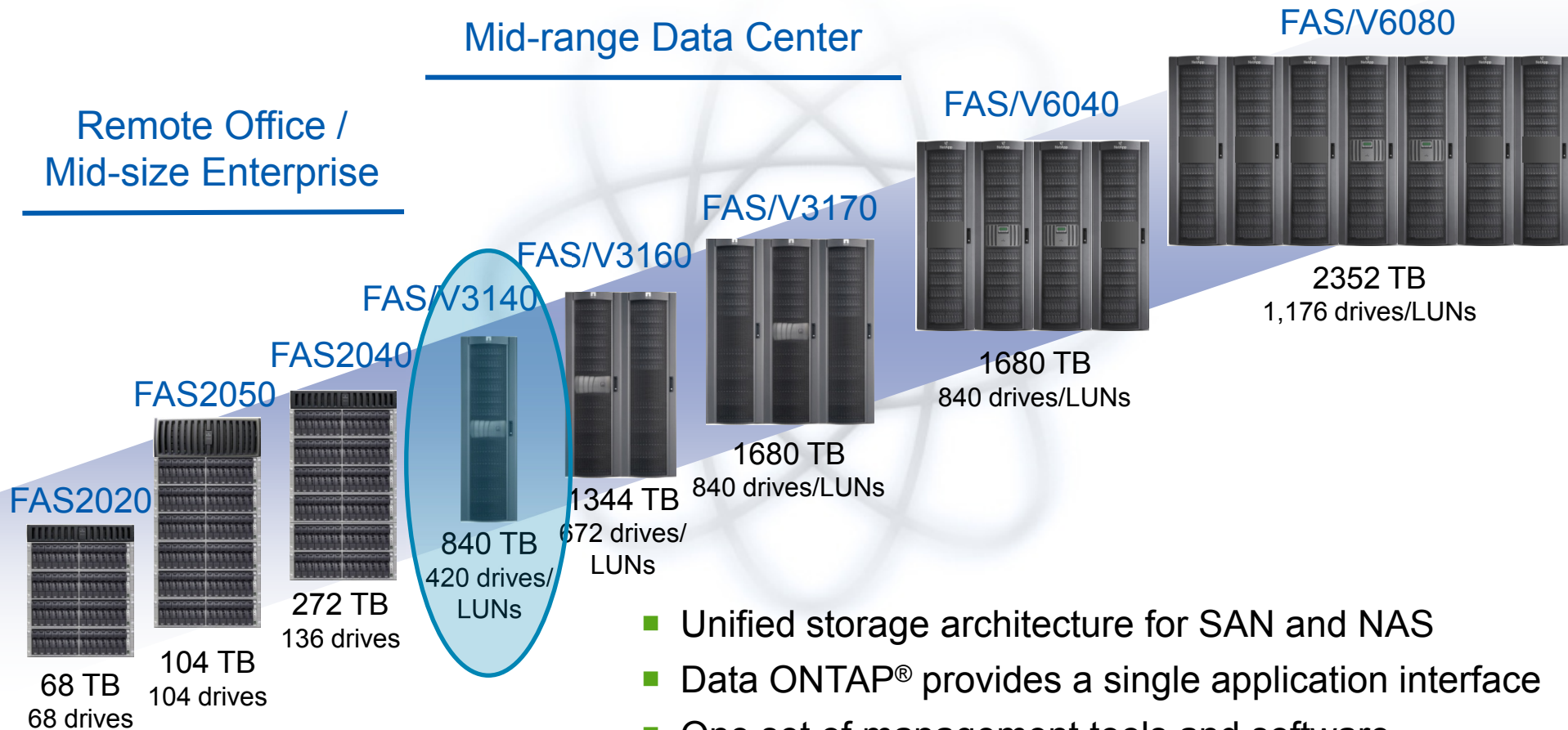


NetApp® Unified Storage Architecture

High-end Data Center

Mid-range Data Center

Remote Office /
Mid-size Enterprise



- Unified storage architecture for SAN and NAS
- Data ONTAP® provides a single application interface
- One set of management tools and software
- V-Series for heterogeneous storage virtualization

- Few words about CERN and computing challenge
- Oracle and RAC at CERN and NetApp for accelerator databases example
- DataOntap 8 scalability
 - PVSS to 150 000 changes/s
 - IO operations per second
 - Flash Cache
 - 10GbE
- Oracle DB on NFS experience
- Oracle VM experience
- Reliability and simplicity
- Conclusions



- Joined CERN in 1996 to work on Oracle database parallelism features
- OakTable member since April 2005
- Team leader for the Database Services section in the CERN IT department
- Specific interest in database application and storage performance

CERN

Annual budget: ~1000 MSFr (~600 M€)

Staff members: 2650

+ 270 Fellows,

Member states: 20

+ 440 Associates

+ 8000 CERN users

Basic research

Fundamental questions

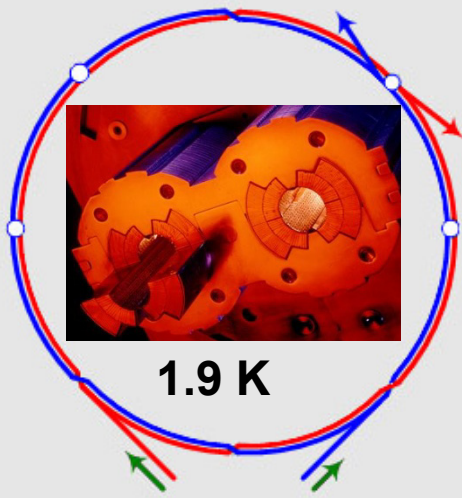
High E accelerator:

Giant microscope ($p=h/\lambda$)

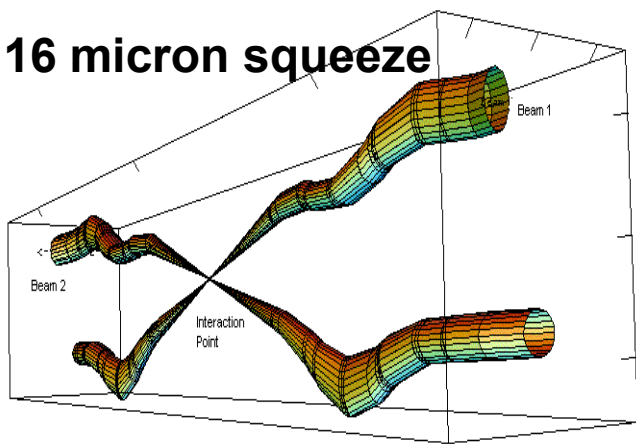
Generate new particles ($E=mc^2$)

Create Big Bang conditions

Large Hadron Collider - LHC



16 micron squeeze

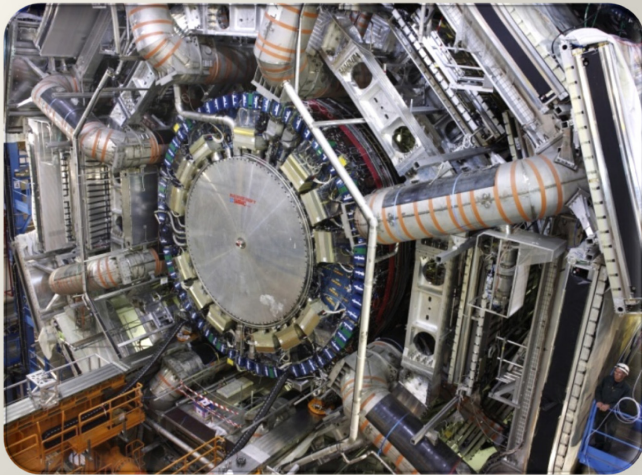
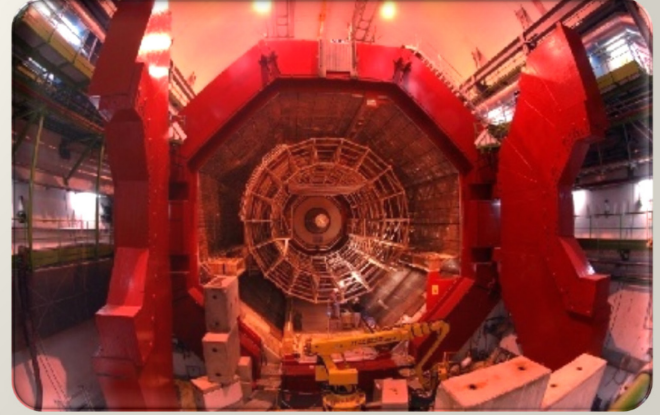
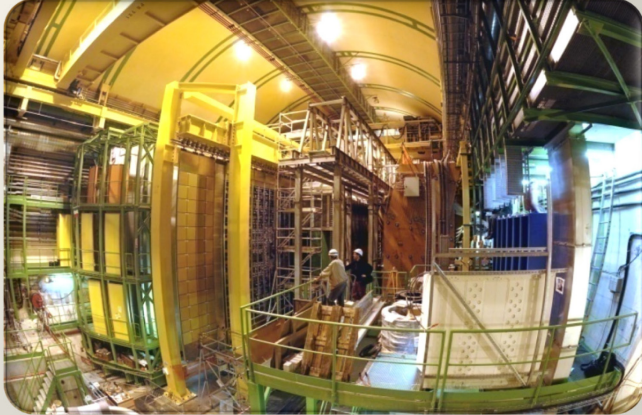


- 27 km circumference
- Cost ~ 3000 M€ (+ detectors)
- Proton-proton (or lead ion) collisions at 7+7 TeV
- Bunches of 10^{11} protons cross every 25 nsec
- 600 million collisions/sec
- Physics questions
 - Origin of mass (Higgs?)
 - Dark matter?
 - Symmetry matter-antimatter
 - Forces – supersymmetry
 - Early universe – quark-gluon plasma
 - ...



LHC accelerator and experiments

CERN
IT
Department



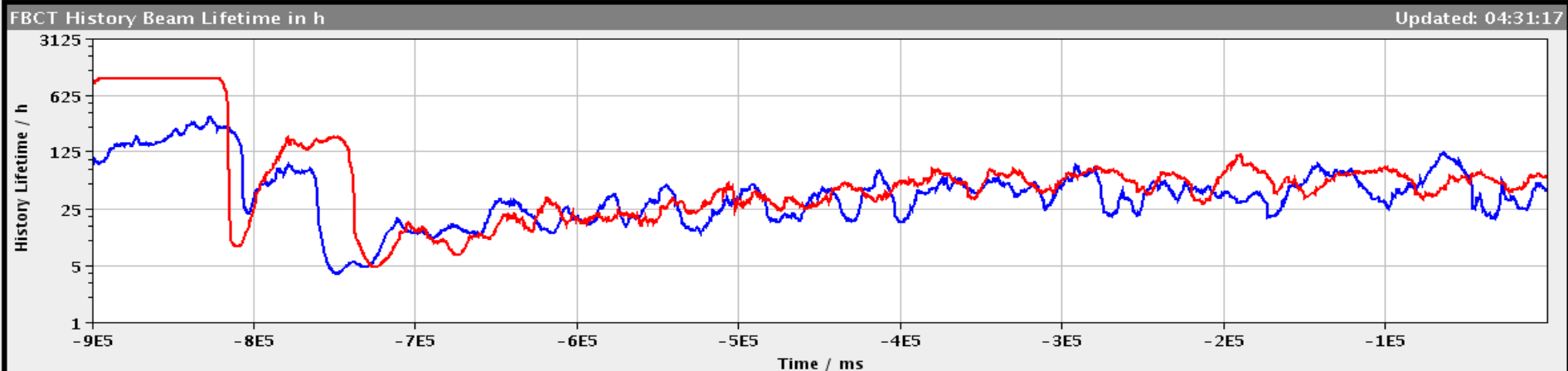
J. Knobloch/CERN: European Grid Initiative – EGI

LHC Instantaneous Luminosity: August Record

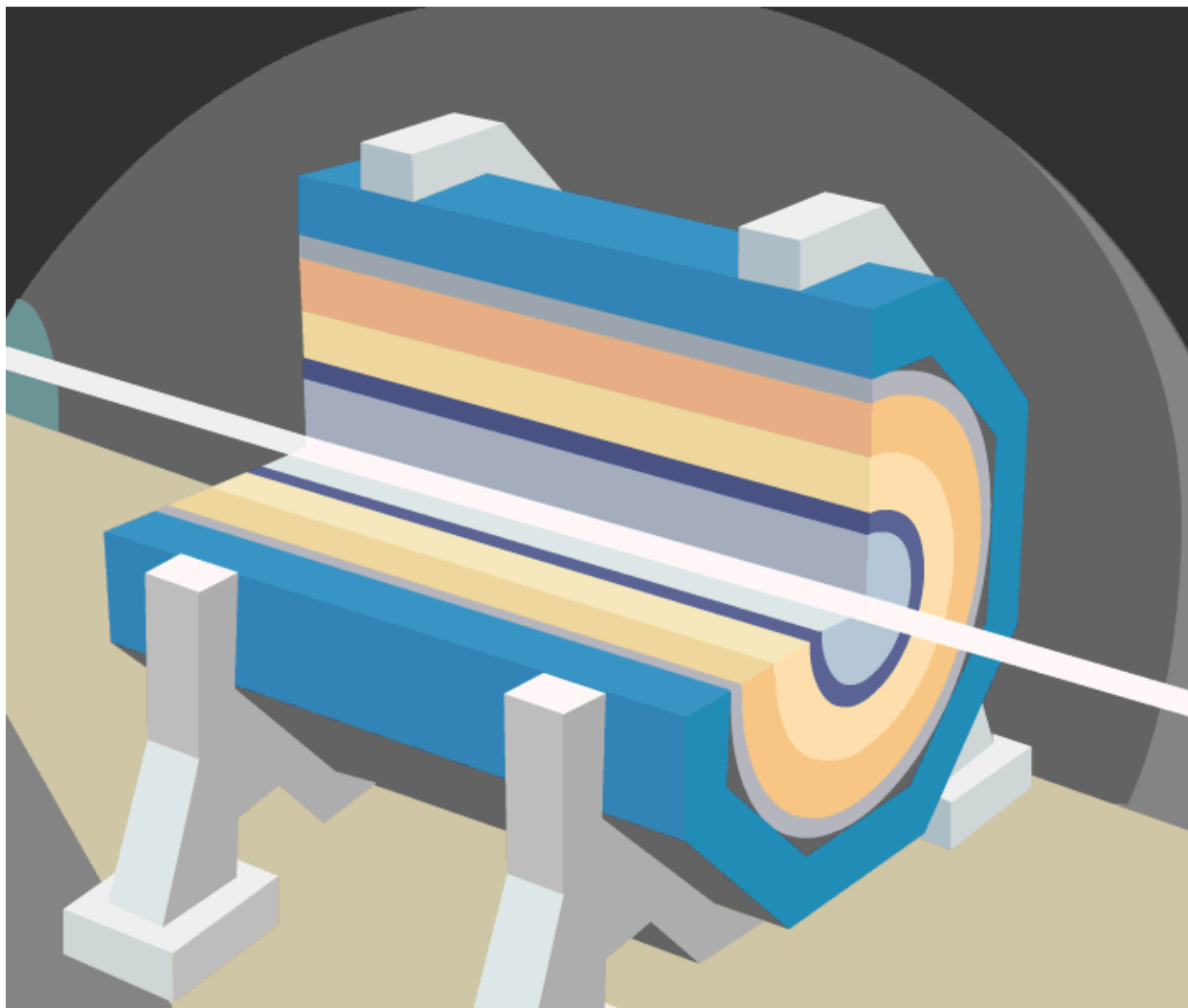
26-Aug-2010 04:24:46 Fill #: 1303 Energy: 3500 GeV I(B1): 5.51e+12 I(B2): 5.23e+12

Experiment Status	ATLAS PHYSICS	ALICE NOT READY	CMS STANDBY	LHCb PHYSICS
Instantaneous Lumi (ub.s) ⁻¹	10.456	0.138	10.719	8.882
BRAN Luminosity (ub.s) ⁻¹	9.573	0.137	7.914	7.327
Fill Lumiosity (nb) ⁻¹	2.0	0.0	2.0	1.7
BKGD 1	0.018	0.019	20.644	0.197
BKGD 2	16.000	0.290	0.002	4.773
BKGD 3	5.000	0.008	0.003	0.106

LHCb VELO Position **OUT** Gap: 58.0 mm STABLE BEAMS TOTEM: **STANDBY**



Slide from Ralph Assmann <http://op-webtools.web.cern.ch/op-webtools/vistar/vistars.php?usr=LHC1>

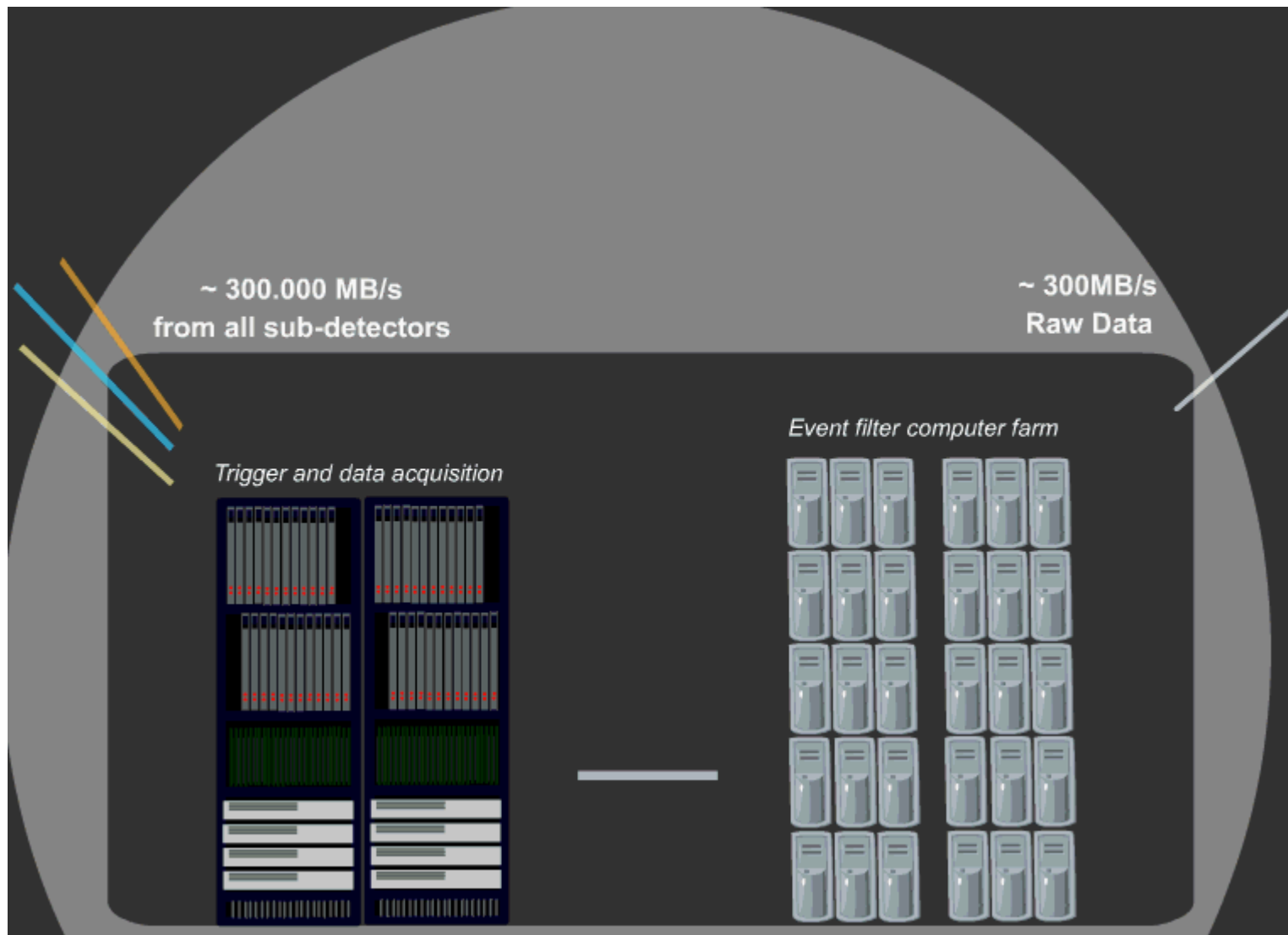


Luminosity :
 $10^{34} \text{cm}^{-2} \text{s}^{-1}$

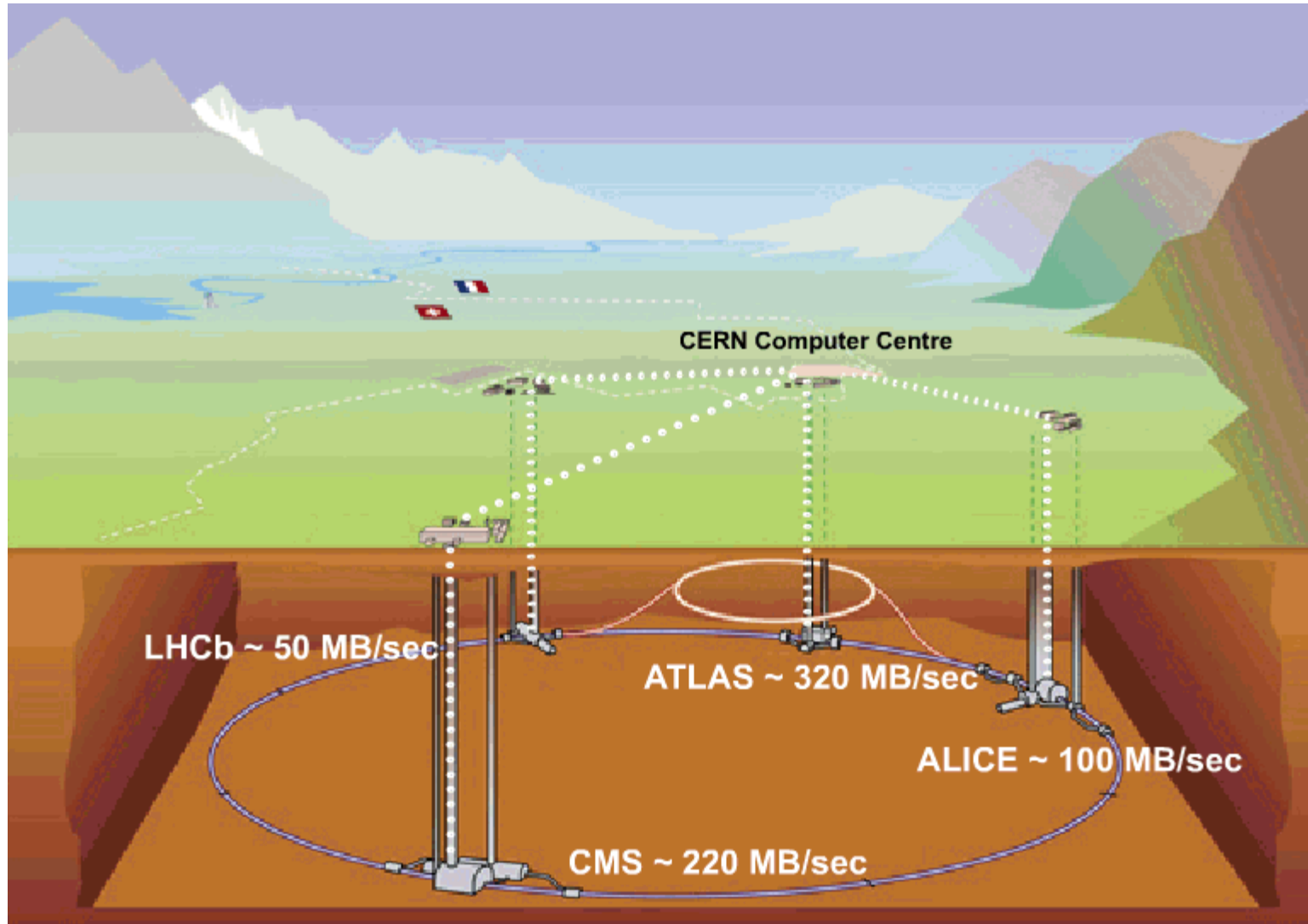
40 MHz – every 25 ns

20 events overlaying

Trigger & Data Acquisition

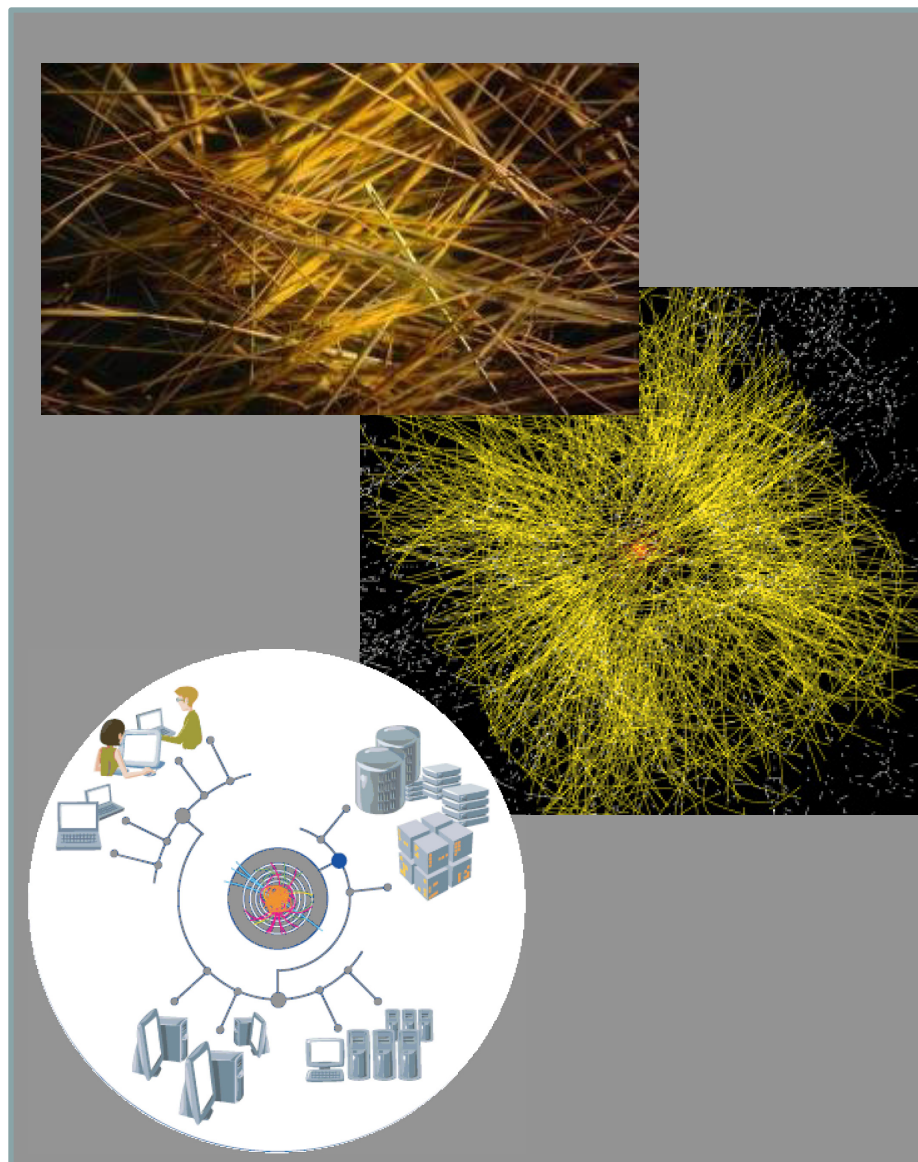


Data Recording



LHC Computing Challenge

- Signal/Noise 10^{-9}
- Data volume
 - High rate * large number of channels * 4 experiments
 - ➔ 15 PetaBytes of new data each year
- Compute power
 - >140 sites
 - ~150k CPU cores
 - >50 PB disk
- Worldwide analysis & funding
 - Computing funding locally in major regions & countries
 - Efficient analysis everywhere
 - ➔ GRID technology



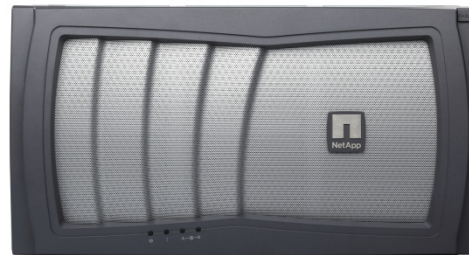
- Few words about CERN and computing challenge
- Oracle and RAC at CERN and NetApp for accelerator databases example
- DataOntap 8 scalability
 - PVSS to 150 000 changes/s
 - IO operations per second
 - Flash Cache
 - 10GbE
- Oracle DB on NFS experience
- Oracle VM experience
- Reliability and simplicity
- Conclusions

- 1982: Oracle at CERN

Conclusion

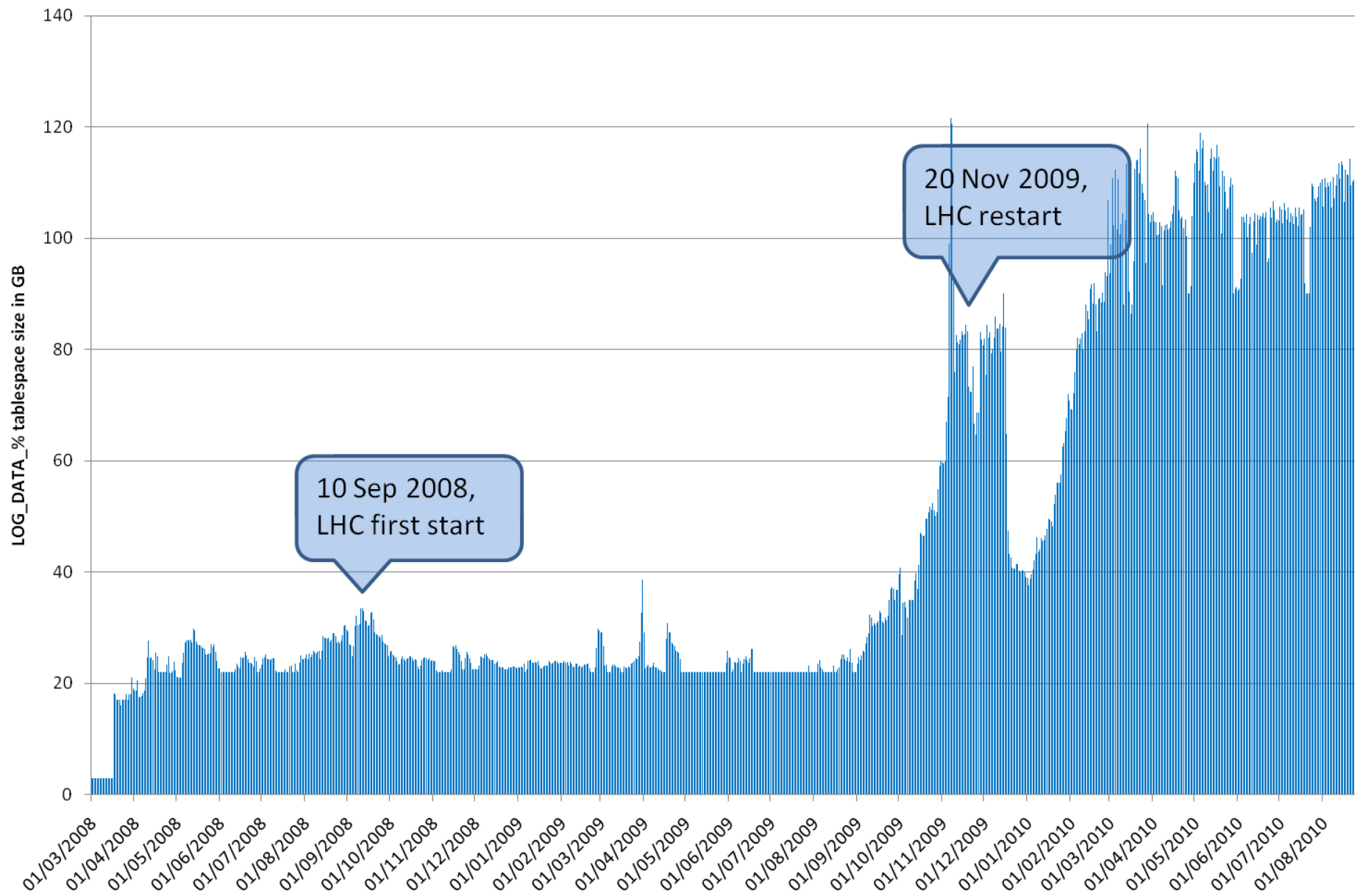
A database management system is a tool for storing, modifying and retrieving data. ORACLE has been chosen for the LEP project because very little training or computer experience is required before a user can effectively use the database.

- Solaris SPARC 32 and 64
- 1996: Solaris SPARC with OPS
- 2000: Linux x86 on single node, DAS
- 2005: Linux x86_64 / RAC / EMC with ASM
- ≥ 2006 : Linux x86_64 / RAC / NFS / NetApp
 - (now, 96 databases)

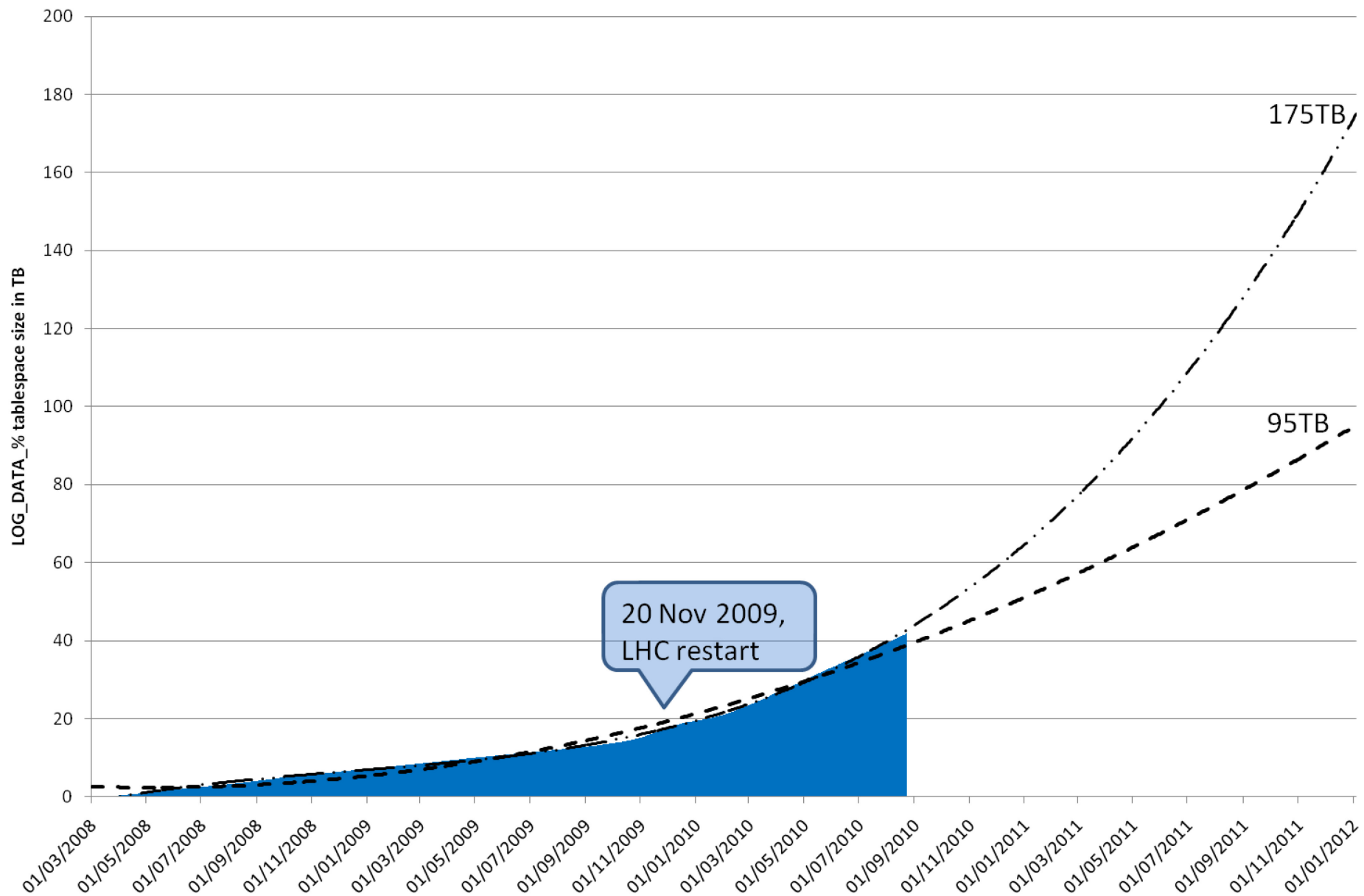


- Use cases
 - ACCCON
 - Accelerator Settings and Controls Configuration **necessary to drive all accelerator installations**, unavailability may require to stop accelerator operation
 - ACCLOG
 - Accelerator long-term Logging database
 - **3.5TB growth per month**

ACCLOG daily growth (GB/day)



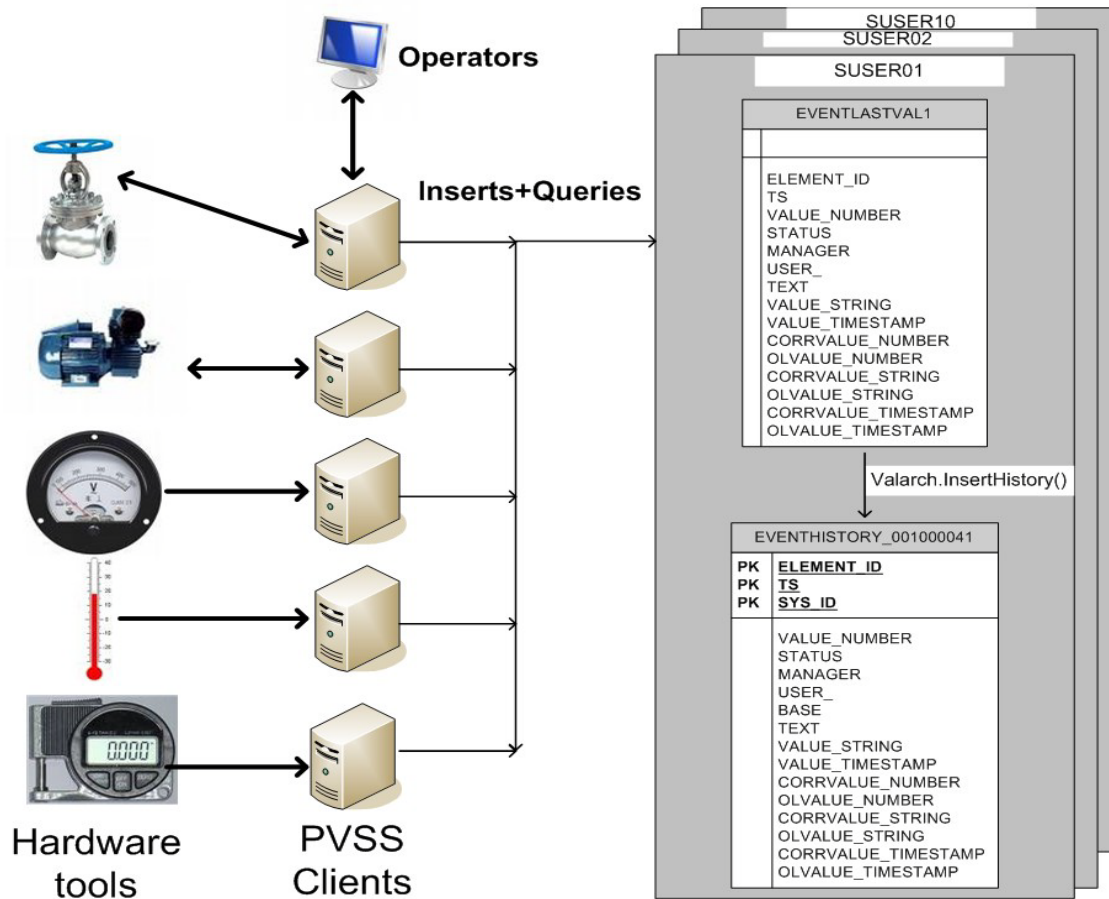
ACCLOG total space



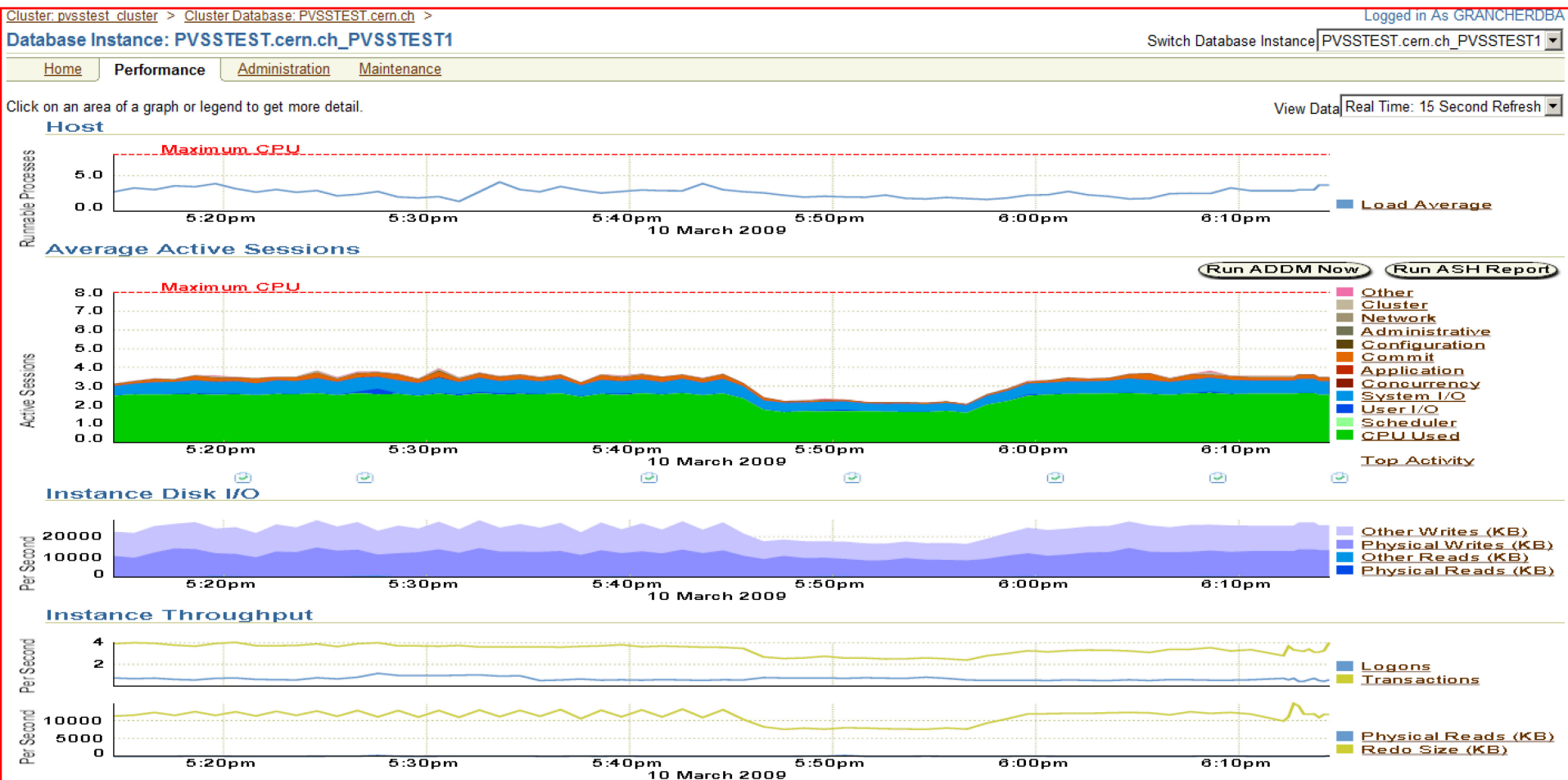
- Implementation
 - Oracle RAC 10.2.0.5 with partitioning
 - Intel x86_64
 - NetApp 3040 and 3140 with Data OnTap8-7 mode
 - Example aggregate dbdsk210
 - Data 12 August 2010 to ~mid July 2011
 - RAID-DP
 - 30 SATA disks, each “2TB”
 - 2 raid groups
 - 38 743GB usable

- Few words about CERN and computing challenge
- Oracle and RAC at CERN and NetApp for accelerator databases example
- **DataOntap 8 scalability**
 - PVSS to 150 000 changes/s
 - IO operations per second
 - Flash Cache
 - 10GbE
- Oracle DB on NFS experience
- Oracle VM experience
- Reliability and simplicity
- Conclusions

- Target = 150 000 changes per second (tested with 160k)
- 3 000 changes per client
- 5 nodes RAC 10.2.0.4
- 2 NAS 3040, each with one aggregate of 13 disks (10k rpm FC)



- Load on one of the instances, stable data loading



- NVRAM plays a critical role in order to have **write operations happen quickly**

Load Profile

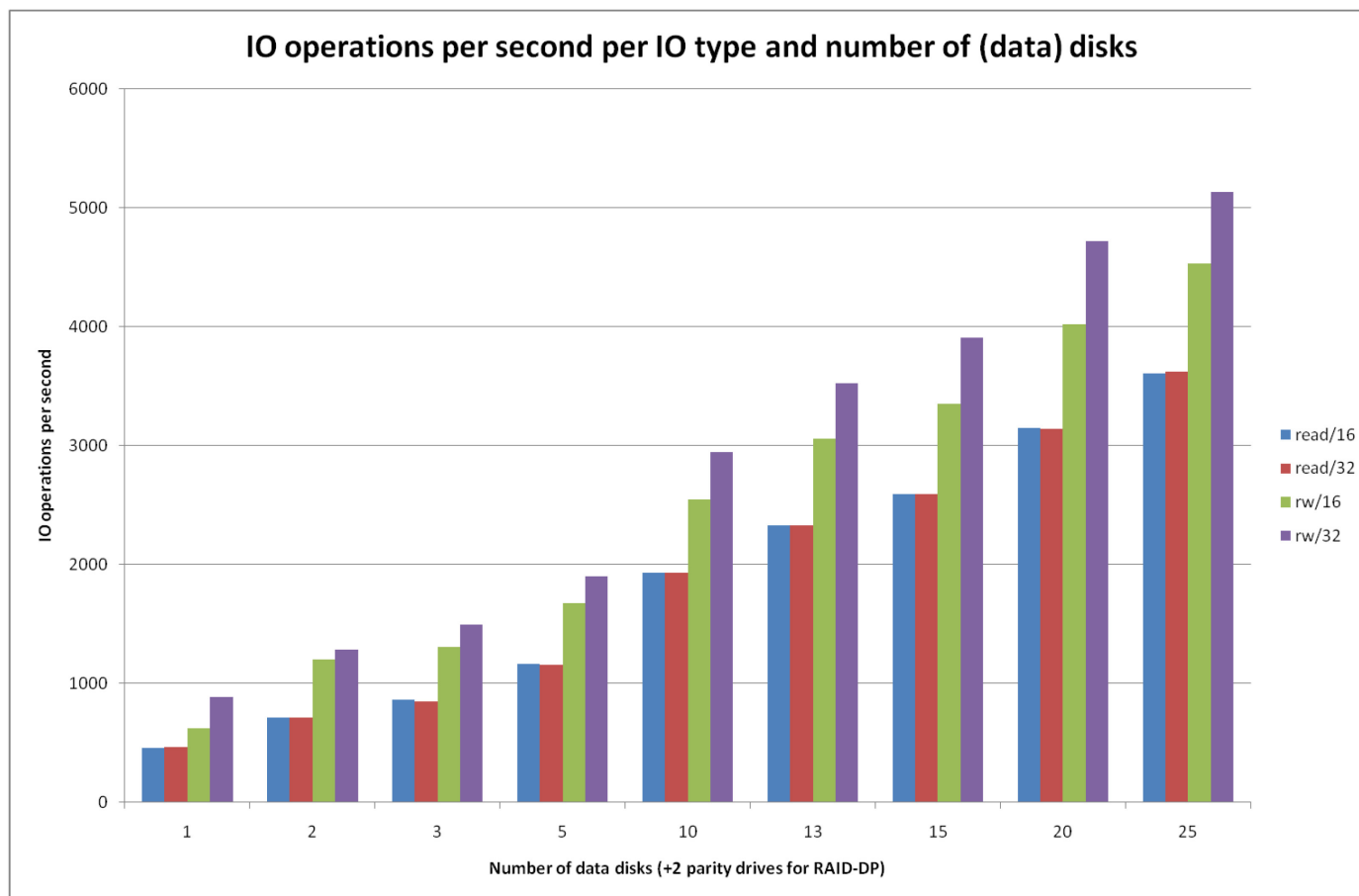
	Per Second	Per Transaction
Redo size:	12,245,111.95	3,701,088.03
Logical reads:	64,352.96	19,450.70
Block changes:	79,638.91	24,070.88
Physical reads:	2.47	0.75
Physical writes:	1,546.05	467.29

```
dbsrvc235>-RAC>-PVSSTEST1:~/work/pvsstest/changestorage$ ssh -2 root@dbnasc210 sysstat -x 1
```

CPU	NFS	CIFS	HTTP	Total	Net kB/s	Disk kB/s	Tape kB/s	Cache	Cache	CP	CP	Disk
					in out	read write	read write	age	hit	time	ty	util
64%	5506	0	0	5506	136147 1692	1568 207148	0 0	>60	100%	82%	Df	79%
58%	5626	0	0	5626	139578 1697	1040 137420	0 0	>60	100%	62%	D	58%
57%	5420	0	0	5420	127307 1618	1080 136384	0 0	>60	100%	79%	D	62%
61%	5142	0	0	5142	130298 1562	1927 149545	0 0	>60	100%	57%	Dn	57%

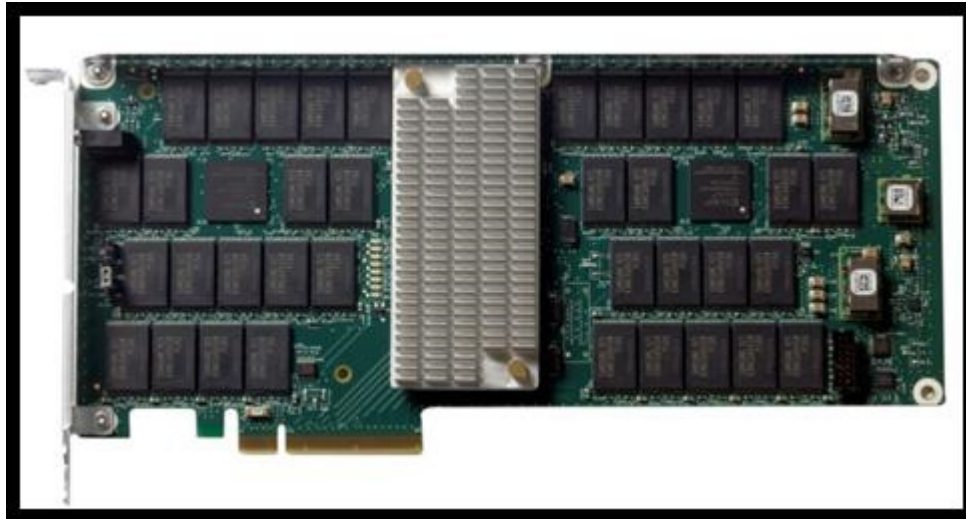
- Few words about CERN and computing challenge
- Oracle and RAC at CERN and NetApp for accelerator databases example
- DataOntap 8 scalability
 - PVSS to 150 000 changes/s
 - **IO operations per second**
 - Flash Cache
 - 10GbE
- Oracle DB on NFS experience
- Oracle VM experience
- Reliability and simplicity
- Conclusions

- DataOntap 8 enables stripping over large number of disks (depends on FAS model and disk size)
- Enables very good scalability



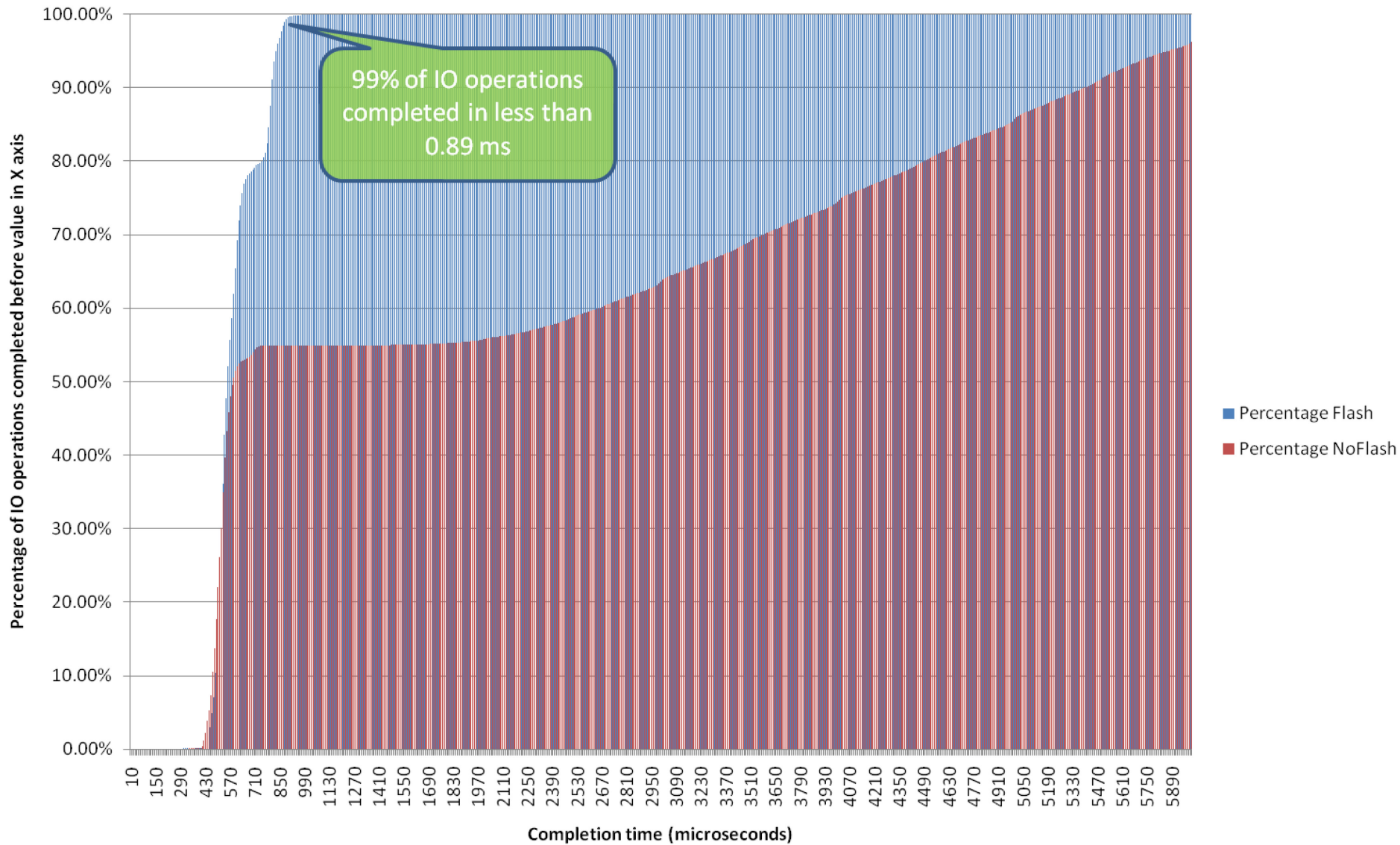
- Few words about CERN and computing challenge
- Oracle and RAC at CERN and NetApp for accelerator databases example
- DataOntap 8 scalability
 - PVSS to 150 000 changes/s
 - IO operations per second
 - **Flash Cache**
 - 10GbE
- Oracle DB on NFS experience
- Oracle VM experience
- Reliability and simplicity
- Conclusions

- Help to increase random IOPS on disks
- Warm-up effect will be an increasingly important issue (2 level of large caches is likely of help)
- For databases
 - select volumes for which caching will benefit (not archive redo logs for example)
 - set “flexscale.lopri_blocks on”



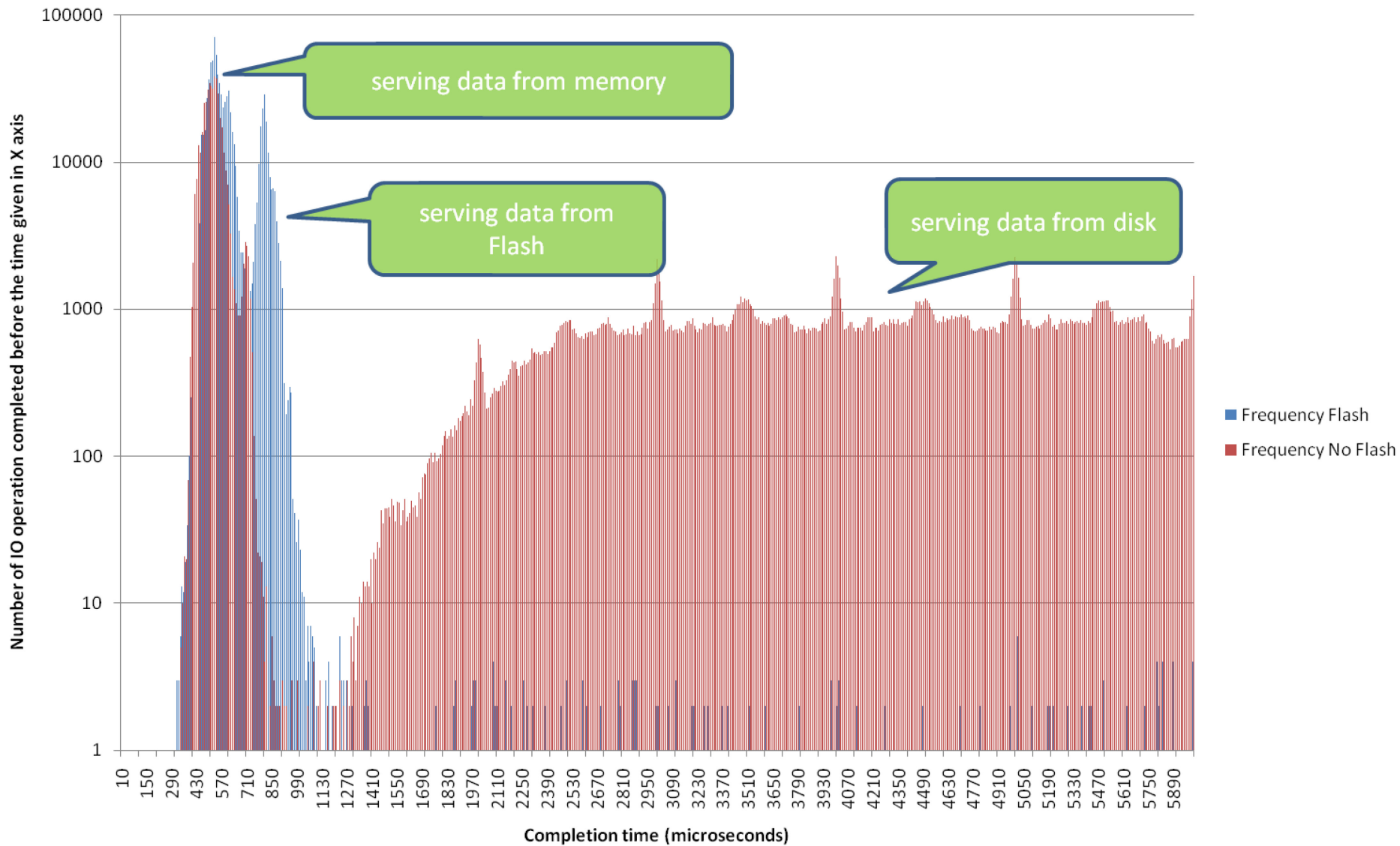
IOPS and flash cache

Histogram of percentage of IO completion time



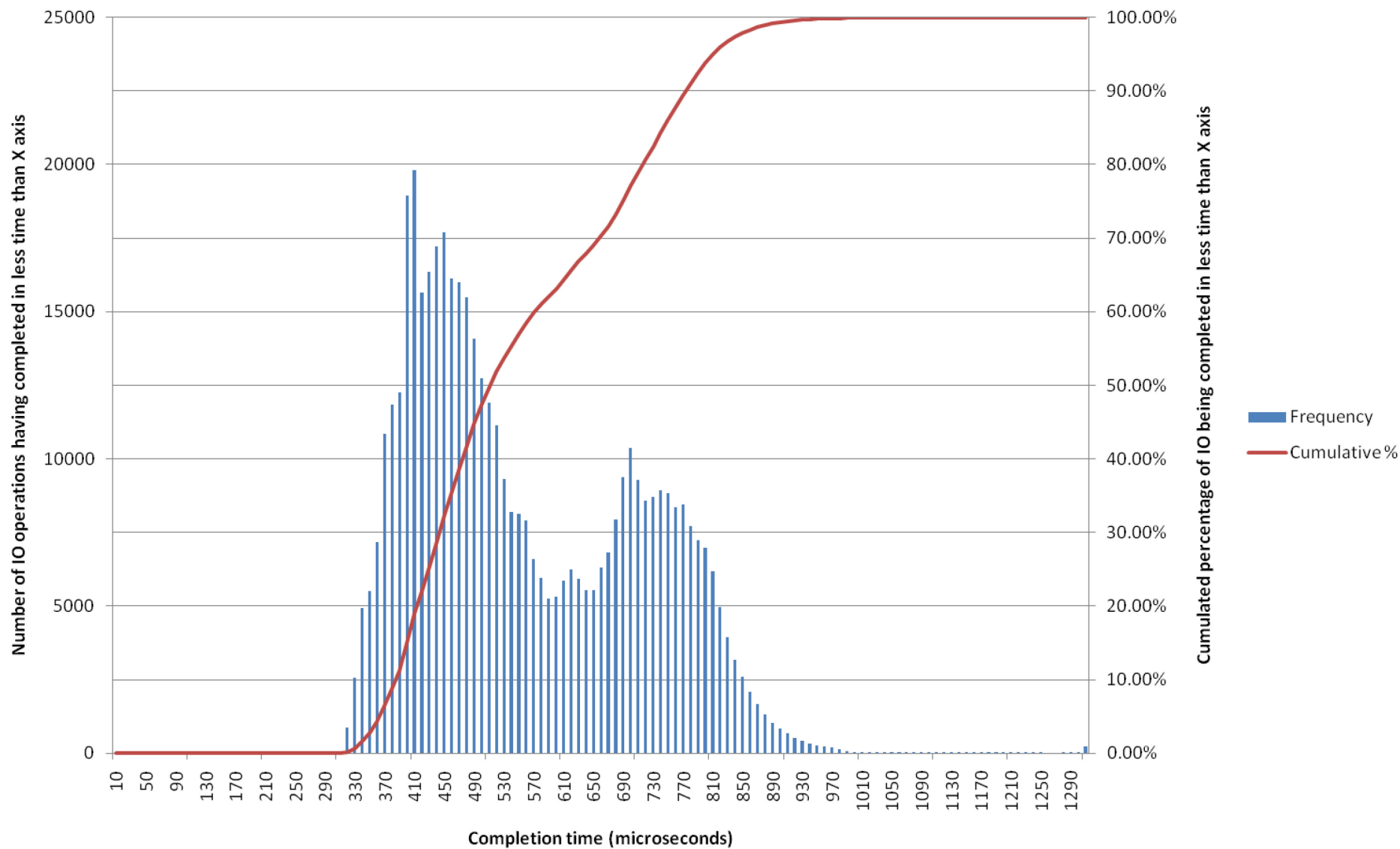
IOPS and flash cache

Distribution of IO operations completion time



IOPS and flash cache

Histogram of IO completion (Oracle trace, 10046 event)



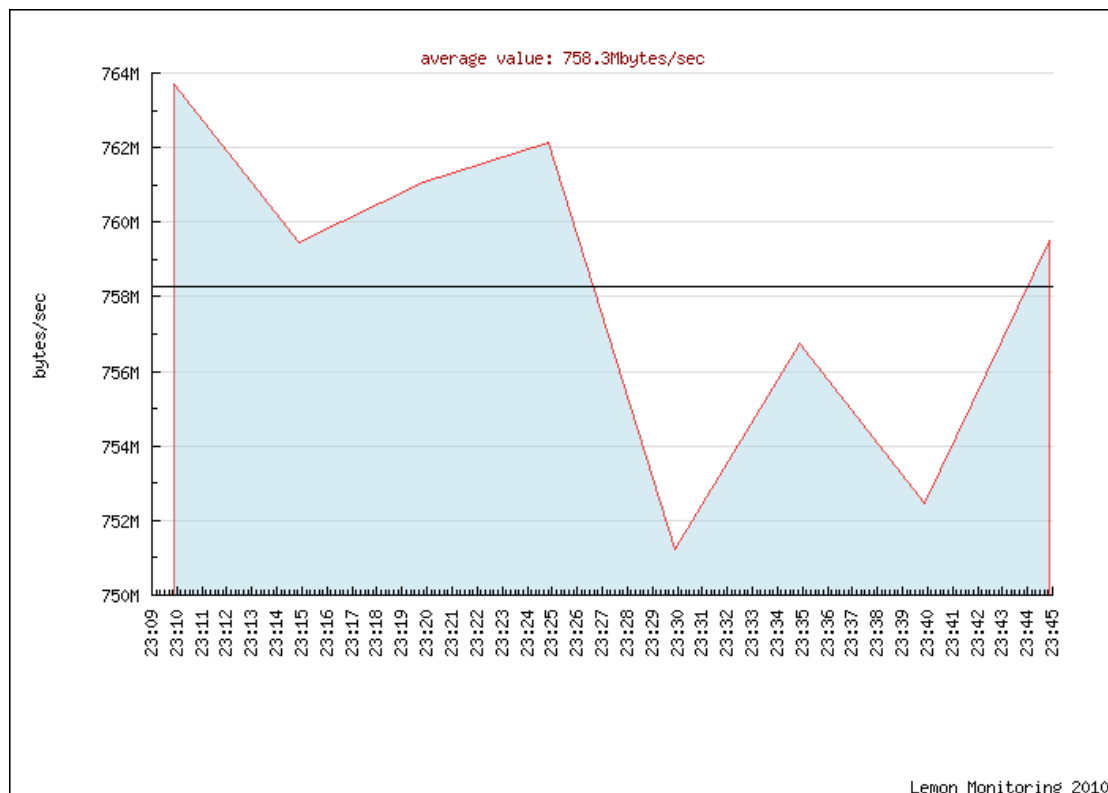
- Few words about CERN and computing challenge
- Oracle and RAC at CERN and NetApp for accelerator databases example
- DataOntap 8 scalability
 - PVSS to 150 000 changes/s
 - IO operations per second
 - Flash Cache
 - **10GbE**
- Oracle DB on NFS experience
- Oracle VM experience
- Reliability and simplicity
- Conclusions

- 10GbE is becoming mainstream (cards, switches)

TX: 289Mb (/s), RX: 6.24Gb (/s)

TOTAL: 6.52Gb (/s) (19% CPU)

- CPU usage
- NAS: 3140 cluster
- Host: dual E5410 with Intel 82598EB 10-Gigabit card



- Few words about CERN and computing challenge
- Oracle and RAC at CERN and NetApp for accelerator databases example
- DataOntap 8 scalability
 - PVSS to 150 000 changes/s
 - IO operations per second
 - Flash Cache
 - 10GbE
- **Oracle DB on NFS experience**
- Oracle VM experience
- Reliability and simplicity
- Conclusions

- One of the nicest features of Oracle11g
 - Enables using **multiple paths to storage**
- Makes Oracle on NFS from simple to extremely simple
 - Just a symlink in \$ORACLE_HOME/lib
 - List of paths to be declared

```
Oracle instance running with ODM: Oracle Direct NFS ODM Library Version 2.0
```

```
...
```

```
Direct NFS: channel id [0] path [dbnasg301] to filer [dbnasg301] via local [] is UP
```

```
Direct NFS: channel id [1] path [dbnasg301] to filer [dbnasg301] via local [] is UP
```

- Promising with NFS 4.1/pNFS
 - Scalability, “on demand”
 - Move of volumes, upgrades

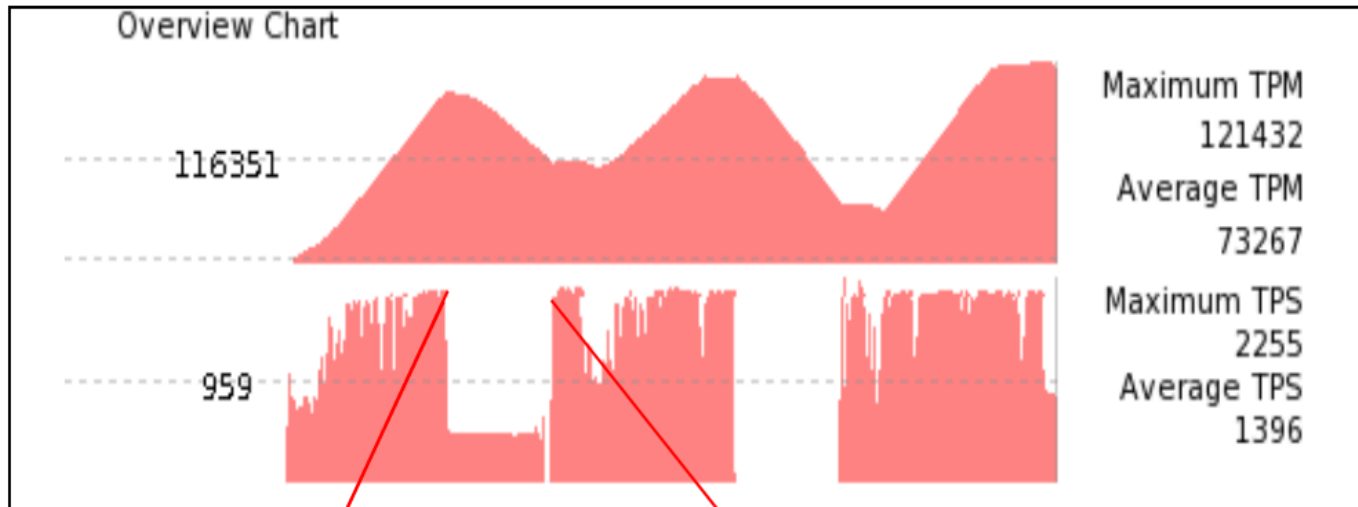
- Have `reallocate` enabled by default (backup!) and `filesystem_options = setall` (async+directIO)
- NetApp NVRAM makes writing fast (see PVSS testcase)
 - Key for OLTP commit time
- DataOntap 8 enables large aggregates (40TB on 3140, up to 100TB on 61xx)
 - Gain in management
 - Gain in performance
- NFS or TCP/IP overhead, CPU usage (large transfer): network roundtrip and disk access
- Scales much better than what many think

- Use NFS/DNFS (11.1 see Note 840059.1 /11.2)
 - Resilient to errors
 - TCP/IP and NFS extremely stable and mature
 - **Extremely simple**, good productivity per DBA
 - Use different volumes for log files, archive redo logs and data files
 - Have several copies of control files and OCR on different aggregate / filer (at least different aggregates)
- Split storage network
 - Cost for the switches is not very high
 - Use MTU = 9000 on the storage network

- Few words about CERN and computing challenge
- Oracle and RAC at CERN and NetApp for accelerator databases example
- DataOntap 8 scalability
 - PVSS to 150 000 changes/s
 - IO operations per second
 - Flash Cache
 - 10GbE
- Oracle DB on NFS experience
- **Oracle VM experience**
- Reliability and simplicity
- Conclusions

- NFS is **extremely well suited for virtualisation**
- Mount database volumes from the guest
 - Separation OS/data
 - Scalability (add mount points if necessary)
 - Same as physical
 - can easily migrate from “physical” to/from “virtual”
- Disk access might more expensive than local
 - Limit swap (do you need any swap?)
 - Check for file inexistence (iAS SSL semaphores)
 - $5.4 \cdot 10^{-6}$ second per “stat” system call on local filesystem
 - $18.1 \cdot 10^{-6}$ second per “stat” system call on NFS mounted filesystem

Oracle VM live migration



Node 1

```
#xm list
Name      ID  Mem VCPUs  State  Time(s)
Domain-0  0  834   8    r----- 1773.7
virt04    8 4096   8    -b----- 517.4
```

xm migrate virt04 node2 --live

```
# xm list
Name      ID  Mem VCPUs  State  Time(s)
Domain-0  0  834   8    r----- 1785.7
migrating-virt04 8 4096   8    r----- 538.3
```

```
# xm list
Name      ID  Mem VCPUs  State  Time(s)
Domain-0  0  834   8    r----- 1851.5
```

Node 2

```
# xm list
Name      ID  Mem VCPUs  State  Time(s)
Domain-0  0  834   8    r----- 2410.8
```

```
# xm list
Name      ID  Mem VCPUs  State  Time(s)
Domain-0  0  834   8    r----- 2444.8
virt04    11 4096   0    -bp---- 0.0
```

```
# xm list
Name      ID  Mem VCPUs  State  Time(s)
Domain-0  0  834   8    r----- 2481.1
virt04    11 4096   8    -b----- 6.4
```

From
Anton
Topurov

- Few words about CERN and computing challenge
- Oracle and RAC at CERN and NetApp for accelerator databases example
- DataOntap 8 scalability
 - PVSS to 150 000 changes/s
 - IO operations per second
 - Flash Cache
 - 10GbE
- Oracle DB on NFS experience
- Oracle VM experience
- **Reliability and simplicity**
- Conclusions

- Simplicity

- Shared log files for the database (`tail alertSID*.log`)
- No need for ASM, day to day simpler operations
 - Operations under stress made easier (copy control file with RMAN)
 - Rename a file in ASM 10.2?
 - Install a 2 nodes RAC with NFS or ASM (multi-pathing, raw on 10.2, FC drivers, ASM ...)



- Reliability

- Do a snapshot before upgrade
- Simplicity is key for reliability (even experienced DBA do basic errors linked with complex storage)
- More robust than ASM “normal” redundancy
- RAID-DP (double parity)

- Disks are larger and larger
 - speed stay ~constant -> issue with speed
 - **bit error rate stay constant** (10^{-14} to 10^{-16}), increasing issue with availability

- With x as the size and α the “bit error rate”

$$P_{failure}(mirror) = 1 - (1 - \alpha)^x$$

$$P_{failure}(raid5, n + 1) = 1 - (1 - \alpha)^{nx}$$

$$P_{failure}(raid6, n + 2) = ((1 - \alpha)^n + n\alpha(1 - \alpha)^{n-1})^x$$

$$P_{failure}(triplemirror) = 1 - (1 - \alpha^2)^x$$

Disks, redundancy comparison (2/2)

	5	14	28
1 TB SATA desktop			Bit error rate 10^{-14}
RAID 1	7.68E-02		
RAID 5 (n+1)	3.29E-01	6.73E-01	8.93E-01
~RAID 6 (n+2)	1.60E-14	1.46E-13	6.05E-13
~triple mirror	8.00E-16	8.00E-16	8.00E-16

	5	14	28
450GB FC			Bit error rate 10^{-16}
RAID 1	4.00E-04		
RAID 5 (n+1)	2.00E-03	5.58E-03	1.11E-02
~RAID 6 (n+2)	7.20E-19	6.55E-18	2.72E-17
~triple mirror	3.60E-20	3.60E-20	3.60E-20

	5	14	28
1TB SATA enterprise			Bit error rate 10^{-15}
RAID 1	7.96E-03		
RAID 5 (n+1)	3.92E-02	1.06E-01	2.01E-01
~RAID 6 (n+2)	1.60E-16	1.46E-15	6.05E-15
~triple mirror	8.00E-18	8.00E-18	8.00E-18

	5	14	28
10TB SATA enterprise			Bit error rate 10^{-15}
RAID 1	7.68E-02		
RAID 5 (n+1)	3.29E-01	6.73E-01	8.93E-01
~RAID 6 (n+2)	1.60E-15	1.46E-14	6.05E-14
~triple mirror	8E-17	8E-17	8E-17

- Few words about CERN and computing challenge
- Oracle and RAC at CERN and NetApp for accelerator databases example
- DataOntap 8 scalability
 - PVSS to 150 000 changes/s
 - IO operations per second
 - Flash Cache
 - 10GbE
- Oracle DB on NFS experience
- Oracle VM experience
- Reliability and simplicity
- **Conclusions**

- Well supported (recommendations at NetApp NOW and Oracle MOS)
- Well managed (AutoSupport, new DOT releases include firmware/...)
- **Very good scalability in performance and size** with Data Ontap 8
- **Impressive stability, cluster failover “just works”**, non-disruptive upgrade (all upgrades since 2006)
- Checksum, scrubbing, multipathing...
- RAID-DP **double parity** (always more important)
- Snapshots and associated feature

- CERN has standardised part of its database infrastructure (all for accelerators, mass storage and administrative applications) on NetApp/NFS
- DataOntap 8 (7 mode) provides scalability, ease of maintenance and management
- Our experience is that Oracle/NFS on NetApp is a rock-solid combination, providing performance and scalability
- Scalability with 64bits aggregate, 10Gb/s Ethernet, Direct NFS, flash caching
- Oracle VM on NFS is simple, extensible and stable

Q&A

session S319046

Steve Daniel, Steve.Daniel@netapp.com

Eric Grancher, Eric.Grancher@cern.ch

- Required Diagnostic for Direct NFS Issues and Recommended Patches for 11.1.0.7 Version
<https://supporthtml.oracle.com/ep/faces/secure/km/DocumentDisplay.jspx?id=840059.1&h=Y>
- Oracle : The Database Management System For LEP
<http://cdsweb.cern.ch/record/443114>
- Oracle 11g Release 1 Performance: Protocol Comparison on Red Hat Enterprise Linux 5 Update 1 <http://media.netapp.com/documents/tr-3700.pdf>