DM

CERN IT Department

Oracle 11g R2 New Features

RAC and ASM

Dawid Wojcik

27 November 2009

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

Outline

- The new Grid Infrastructure
  - Introduction
  - Oracle connectivity redesigned (SCAN, GNS, …)
  - Cluster resources and new ways of cluster database management
  - OCR and voting disks redesigned
- ACFS and ADVM
  - Features
  - Architecture
  - Usage scenarios
- Fun stuff

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

Grid Infrastructure

- Clusterware is now called *Oracle Grid Infrastructure*
  - Installation
    - ASM is now part of Grid Infrastructure
    - OCR and voting disk no longer supported on raw or block devices, only on ASM
    - If installation goes wrong Oracle now provides a deinstallation utility
  - Upgrade
    - Rolling upgrade supported only from 11.1 (for 10g DB is required to be down on all nodes)
    - Possible to install 11.2 Grid with 10.2 DBs (you need to "pin" the nodes)

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

Grid Infrastructure

- Grid Infrastructure has been redesigned
  - New process tree
    - *OPROCD*, *OCLSOMON*, *OCLSVMON* no longer exist
    - Init spawns only one process (*ohasd*), which spawns
      - oraagent
      - orarootagent
      - cssdagent
  - Hangcheck timer no longer required on linux
  - ASM instance is started during Grid Infrastructure startup to allow accessing OCR and voting disks (bootstrap procedure)

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

Single Client Access Name

- SCAN (Single Client Access Name)
  - Allows any client to connect to any database or service in the cluster with just single hostname even if the cluster size changes
  - Load balances connections across all instances providing a service (server side load balancing)
  - Provides failover in case an instance fails and services relocate
  - Allows clients to use EZConect or simple JDBC connection in 'RAC-aware' mode
  - Each cluster (even two node one) has
    - 3 SCAN VIPs
    - 3 SCAN listeners

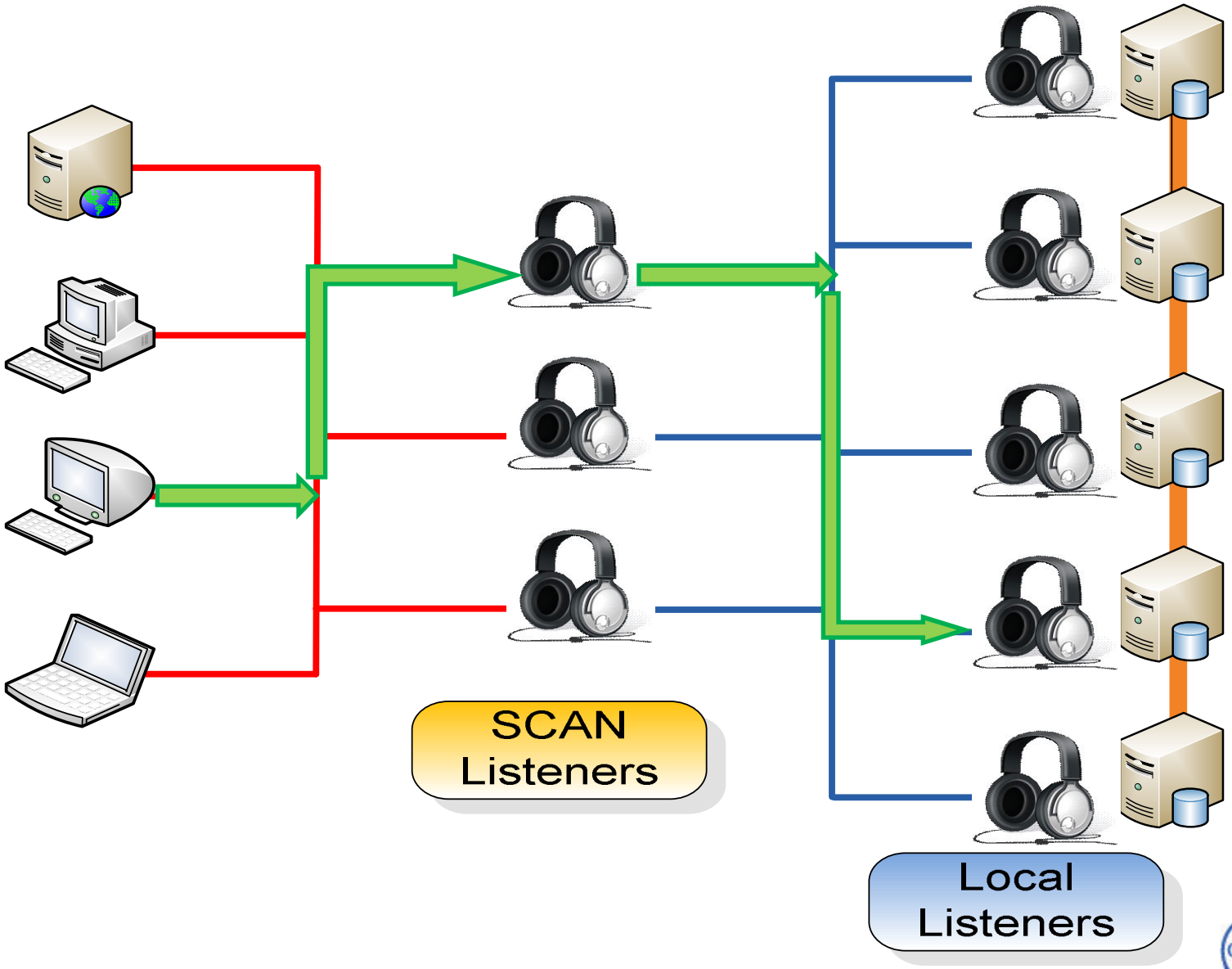
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Single Client Access Name

- SCAN implementation
  - DNS (Domain Name Service)
    - NetOps define DNS alias which resolves one of three SCAN VIPs in round-robin way
  - GNS (Grid Naming Service)
    - Subdomain DNS resolution for SCAN VIPs
    - Can be used for DHCP VIP registration
- SCAN registration
  - Instance registers with local listener on its node
  - Instance registers with SCAN (all SCAN listeners)

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

Single Client Access Name



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

Managing Oracle RAC

- Administrator Managed – “the old way”
  - DBA defines where the database should run and manages list of nodes in RAC
  - DBA defines where services should run within the cluster
- Policy Managed – “the new way”
  - DBA defines resource requirements for each DB running in the cluster
  - DB instances are started “when needed”
  - Goal - remove hard coding services and instance nodes

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

Policy Managed RAC

- Allows grouping more DBs into one cluster for better resource utilization
  - Cluster divided into pools of servers
  - Applications and DBs run in one or more server pools
    - Instances can be started on different nodes each time the cluster starts (!)
  - Resource allocation is defined by 3 attributes:
    - Min – minimum number of servers (default 0)
    - Max – maximum number of servers (default 0 or -1)
    - Importance – 0 (least important) to 1000

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

Policy Managed Services

- Services in RAC managed with server pools can run only in one server pool
- Depending on how many nodes they run they can be
  - Uniform – run in all instances in a pool
  - Singleton – run only in one instance in a pool
- This gives less flexibility than Administrator Managed Services

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

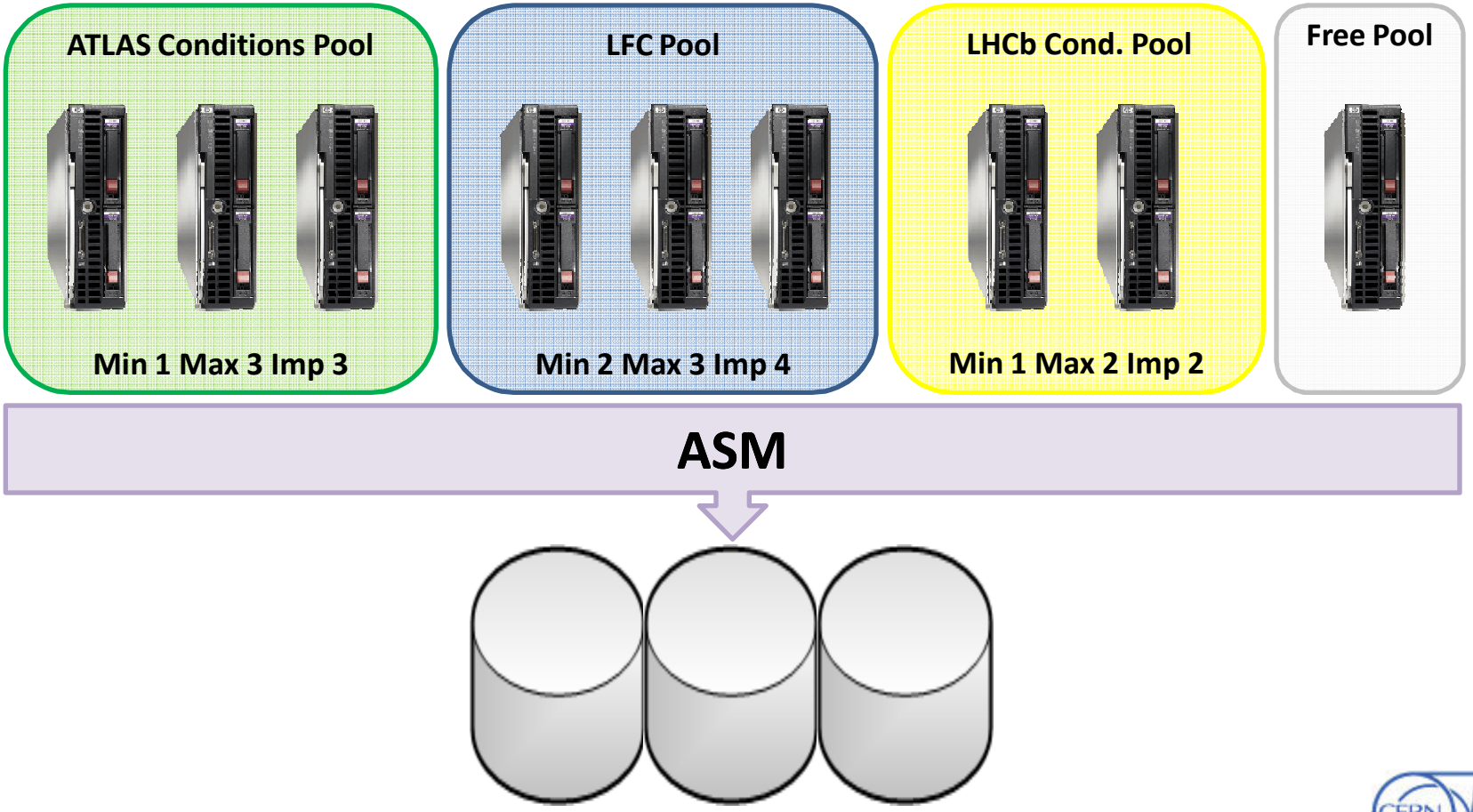
Server assignment in Grid

- Servers are assigned in the following order:
  - Generic server pool (if defined, during upgrade all servers go to generic pool)
  - User defined server pools
  - Free pool
- Grid Infrastructure uses importance of server pool to determine server assignment order:
  - Fill all server pools in order of importance until they meet their minimum
  - Fill all server pools in order of importance until they meet their maximum
  - By Default any left over go into free pool

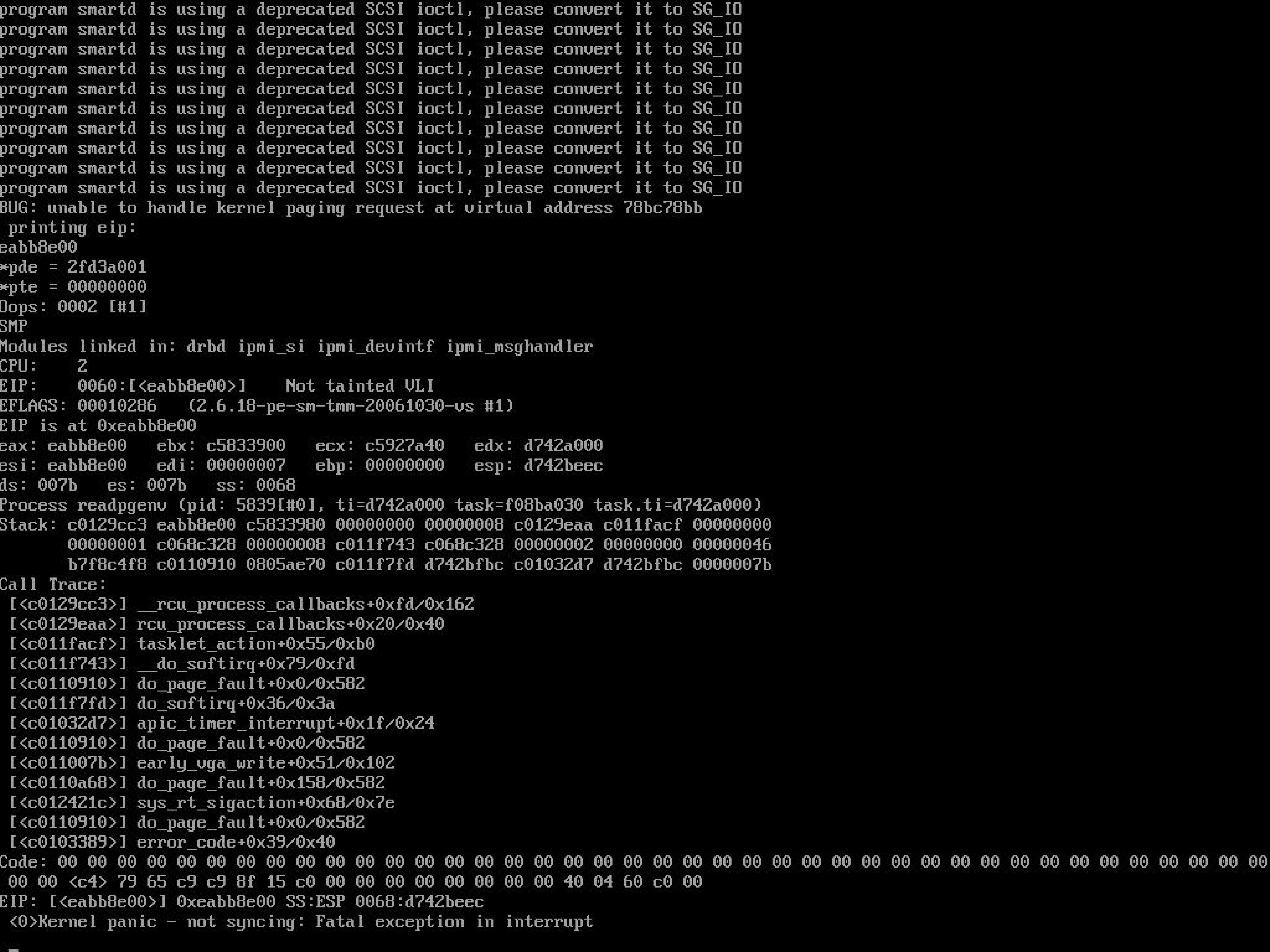
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Server assignment in Grid

- Cluster of 9 nodes is staring up …



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



```
program smartd is using a deprecated SCSI ioctl, please convert it to SG_IO
program smartd is using a deprecated SCSI ioctl, please convert it to SG_IO
program smartd is using a deprecated SCSI ioctl, please convert it to SG_IO
program smartd is using a deprecated SCSI ioctl, please convert it to SG_IO
program smartd is using a deprecated SCSI ioctl, please convert it to SG_IO
program smartd is using a deprecated SCSI ioctl, please convert it to SG_IO
program smartd is using a deprecated SCSI ioctl, please convert it to SG_IO
program smartd is using a deprecated SCSI ioctl, please convert it to SG_IO
program smartd is using a deprecated SCSI ioctl, please convert it to SG_IO
program smartd is using a deprecated SCSI ioctl, please convert it to SG_IO
BUG: unable to handle kernel paging request at virtual address 78bc78bb
 printing eip:
eabb8e00
*pde = 2fd3a001
*pte = 00000000
Oops: 0002 [#1]
SMP
Modules linked in: drbd ipmi_si ipmi_devintf ipmi_msghandler
CPU:    2
EIP:    0060:[<eabb8e00>]    Not tainted VLI
EFLAGS: 00010286   (2.6.18-pe-sm-tmm-20061030-vs #1)
EIP is at 0xeabb8e00
eax: eabb8e00   ebx: c5833900   ecx: c5927a40   edx: d742a000
esi: eabb8e00   edi: 00000007   ebp: 00000000   esp: d742beec
ds: 007b   es: 007b   ss: 0068
Process readpgenv (pid: 5839[#0], ti=d742a000 task=f08ba030 task.ti=d742a000)
Stack: c0129cc3 eabb8e00 c5833980 00000000 00000008 c0129eaa c011facf 00000000
       00000001 c068c328 00000008 c011f743 c068c328 00000002 00000000 00000046
       b7f8c4f8 c0110910 0805ae70 c011f7fd d742bfbc c01032d7 d742bfbc 0000007b
Call Trace:
 [<c0129cc3>] __rcu_process_callbacks+0xfd/0x162
 [<c0129eaa>] rcu_process_callbacks+0x20/0x40
 [<c011facf>] tasklet_action+0x55/0xb0
 [<c011f743>] __do_softirq+0x79/0xfd
 [<c0110910>] do_page_fault+0x0/0x582
 [<c011f7fd>] do_softirq+0x36/0x3a
 [<c01032d7>] apic_timer_interrupt+0x1f/0x24
 [<c0110910>] do_page_fault+0x0/0x582
 [<c011007b>] early_vga_write+0x51/0x102
 [<c0110a68>] do_page_fault+0x158/0x582
 [<c012421c>] sys_rt_sigaction+0x68/0x7e
 [<c0110910>] do_page_fault+0x0/0x582
 [<c0103389>] error_code+0x39/0x40
Code: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
 00 00 <c4> 79 65 c9 c9 8f 15 c0 00 00 00 00 00 00 00 00 40 04 60 c0 00
EIP: [<eabb8e00>] 0xeabb8e00 SS:ESP 0068:d742beec
 <0>Kernel panic - not syncing: Fatal exception in interrupt
```

Handling failures

- If a server leaves the cluster … and the free pool is empty
  - Grid Infrastructure may move servers from one server pool to another only if
    - you have non-default values for min, importance
    - a server pool falls below its minimum
  - Servers to be moved are taken from
    - A server pool that is less important
    - A server pool of the same importance which has more servers than its minimum

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

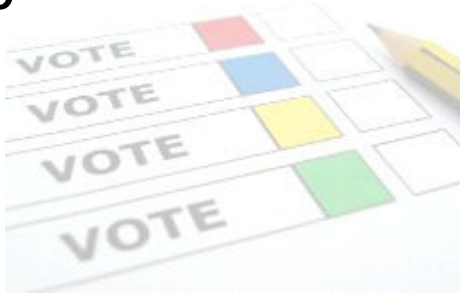
OCR in ASM

- OCR is now just like any file in ASM
  - New Oracle managed file type
  - 1 OCR per diskgroup
  - Good practice to have more than 1 OCR
  - OCR file redundancy follows diskgroup redundancy
  - OCR now holds automatic backups of voting disks! (taken automatically on any cluster configuration change)

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

Voting disk in ASM

- Voting Disks are created on specific disks and Grid Infrastructure knows their location
  - Number of voting disks depends on the redundancy chosen for the diskgroup
    - External – 1 Voting Disk
    - Normal – 3 Voting Disks
    - High – 5 Voting Disks



  - Diskgroup must contain enough failure groups to create each voting disk in a separate failure group!

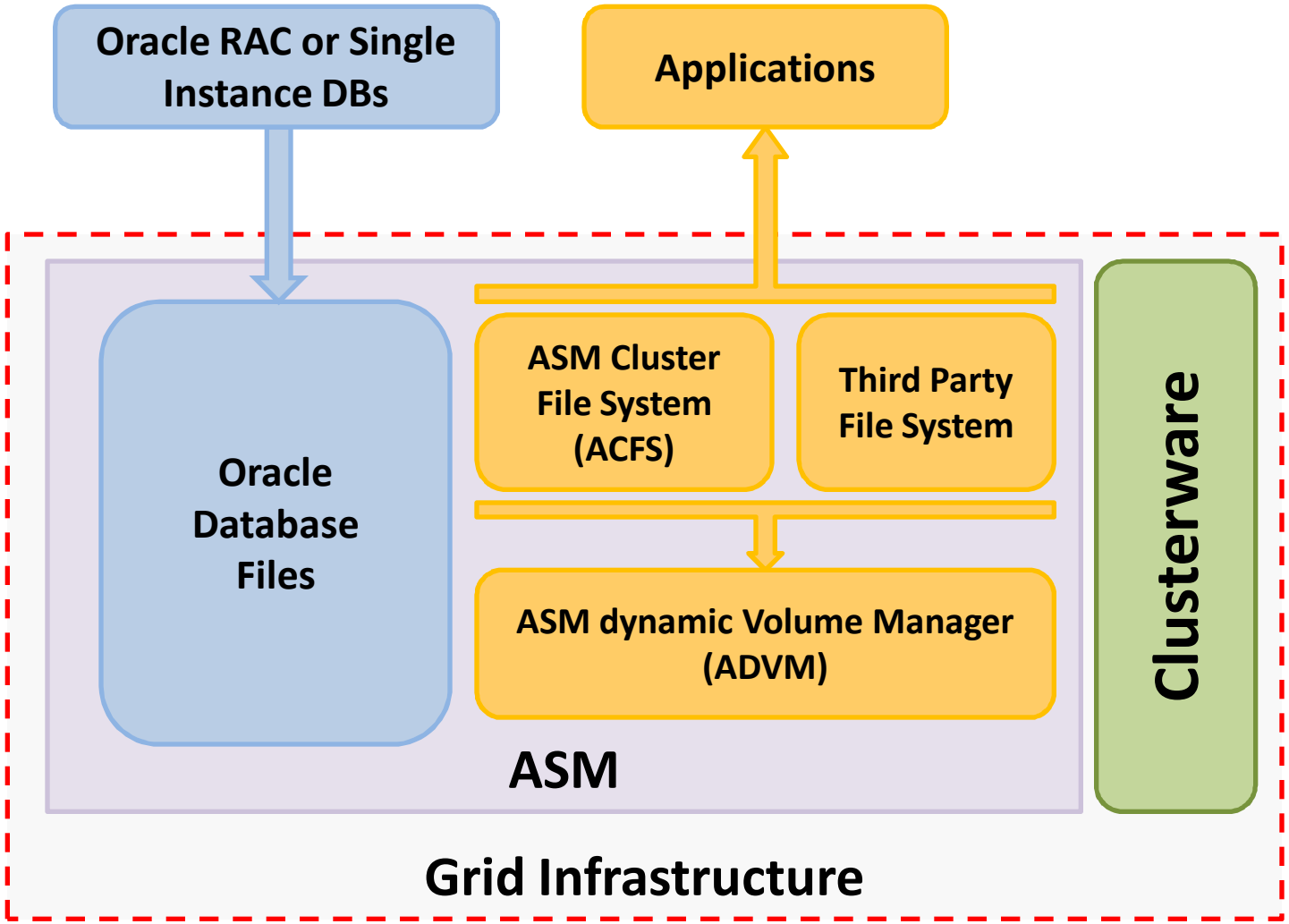
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

ACFS – the revolution

- What is ACFS?
  - ASM Cluster File System – POSIX/X-OPEN compliant File System
    - Can be used cluster-wide or single node only
  - Currently Linux only (RHEL5 +), new platforms coming soon
  - Can be shared using NFS, CIFS, …
  - Online filesystem expansion / shrink
  - mirror protection when using NORMAL redundancy diskgroups
  - read-only snapshots built-in

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

ACFS architecture



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

ACFS – under hood

- ASM Dynamic Volume Manager (ADVM)
  - Carves out logical volumes from an ASM diskgroup that will be exposed to OS as block devices
  - Device name format: `/dev/asm/volume_name-xyz`
  - ADVM volume can be partitioned
  - On top of ADVM volume one can create any file system (ext3, ACFS, ...)
  - Volumes can be resized (make sure the filesystem supports it)
- Kernel modules: oracleacfs, oracleadvm, oracleoks, oracleasm

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

ADVM features

- ADVM Volumes can be exposed (enabled) on any nodes
- Volume redundancy can be
  - Unprotected – for external redundancy diskgroup
  - High, mirror or unprotected – for normal redundancy diskgroup
- Stripe width and IDP (Intelligent Data Placement) parameter can be specified

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

ASM – some new features

- IDP (Intelligent Data Placement)
  - Requires JBOD storage connected (one disk – one LUN)
  - ASM files can be specified to be on external (hot, faster) part of the disks or internal (cold, slower)
- New tools to manage ASM and ACFS
  - Extended asmcmd tool
  - New tool for ACFS – acfsutil
- …

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

ACFS use cases

- BFILEs, external tables, external files storage
- Local file systems, exports, dumps, etc.
- Batch jobs data
- Shared RDBMS home (not recommended if using rolling patches)
- Local RDBMS home
- …

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

ACFS Advantages

- No need for 3rd party cluster file systems
- Uses ASM - known tools
- it will become platform independent CFS
- Simple administration (acfsutil, acfschkdsk, mkfs, ...)
- Cost effective (if you have at least one node licensed for SE or EE - you can use it)
- Snapshots for free
- Very good performance

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

ACFS snapshots

- Provide Point-In-Time images (ACFS only)
- File System functionality
- Can be performed online
- Used for consistent backups
- Limited to 63 snapshots per File System
- Copy on write mechanism (before-images shared between snapshots)
- Snapshots within the same file system - monitor free space!

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

ACFS performance

- Compared ext3 on local disk vs. ext3 on ADVM
  - No difference in write speed
- Typical read/write performance is much better comparing to ext3 (both running on ADVM)

Test type	EXT3 (on ADVM)	ACFS (on ADVM)
Big file creation	37MB/s	**225MB/s**
Extracting 2,5GB tar	180 s	**100 s**
Deleting big dir. tree	**3 s**	11 s
Touching 50k files	**63 s**	75 s

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

Important SQL features

- Recursive Subquery Factoring

```
with x( s, ind ) as
( select sud, instr( s
  from ( select '53  7
     sud from dual )
  union all
  select substr( s, 1,
       , instr( s, ' '
  from x
     , ( select to_cha
         from dual
         connect by ro
       ) z
  where ind > 0
  and not exists ( sel
                   fro
                   whe
                   or
                   or
                 )
)
select s
from x
where ind = 0
/
```

5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

```
28    419  5    8  79'
1 )
* 3
```

http://technology.amis.nl/blog/6404/oracle-rdbms-11gr2-solving-a-sudoku-using-recursive-subquery-factoring

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

Conclusions

- 11gR2 comes with many new features
  - We will test some of the most interesting and publish tests on our TWiki
  - We are eager to try ACFS in production
  - Not ready for Server Pool managed clusters (isolation preferred over consolidation)
    - Except for downstream capture boxes
  - … more useful features to be discovered

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