OVM on Exadata

Living in A virtual world

Experiences from the field:
Pieter Van Puymbroek
Philippe Fierens
About Pieter

- Sailor
- Musician
- Engineered systems Oracle DBA
- dad
- Recently Joined the ACE program
- working for Exitas Oracle Platinum Partner in BELGIUM

And ....
likes to be in control
Dreams at night ;-}
About Philippe

Philippe Fierens
Independent
Oracle DBA Architect
Oracle SuperCluster fan
Oracle ACE
and most of all
DAD & husband ;-) 
AND

... @pfierens

pfierens.blogspot.com
3 Membership Tiers
• Oracle ACE Director
• Oracle ACE
• Oracle ACE Associate

Connect:
• oracle-ace_ww@oracle.com
• Facebook.com/oracleaces
• @oracleace

Nominate yourself or someone you know: acenomination.oracle.com

500+ Technical Experts Helping Peers Globally
About Exitas

Quality. Passion. Personality

- Platinum partner
- Support full red stack
  - Consultancy
  - Managed services
  - Reselling
- Team of 25 dba’s

www.exitas.be
Agenda

- What
- Why
- How / Implementation
- Issues
Unboxing the exadata OVM

high availability

VM templates

vm manager Web gui

OVF

openstack

Live Migration

Rapid VM provision cloning

Windows Solaris guest
VMWare / real OVM on Exadata ?
What is it not?

You will not find any of these:

- OVM manager
- Load balancing
- High Availability : in RAC we trust ;-)
- Live Migration
What is it then?
WARNING

Manage Expectations
Why?

- Better name is Virtualised exadata
- See domU’s bonded in 1 cluster as a bare metal exadata
- HA is foreseen by Oracle Grid infrastructure
- Cell’s are NOT virtualised
- Limit Licenses -> Trusted domains
Flexibility

Just because you can, doesn’t mean you should

most operations on VM level not online*:

- No Memory ballooning (mem = maxmem)
  - -> due to IB sr-iov
- CPU adding removal
  - -> as long as vcpus < maxvcpus (License!)
- ASM / Celldisks can be modified online as traditional
- Decent planning required

*not supported in 12.1.2.2.1
Licensing

- Only use options in the environments needed
- EE and storage cell licenses as you’re used to
- Limit number of CPU licensed
- Capacity on demand is possible
- you pay for hwm of max_vcpu!
- Checked with OEM
Security

Designed to isolate environments from each other:

- vlanning on compute nodes supported
- X6-2 allows to add NIC
- partition keys can be used on IB
- cellwall is still present
- ssh can be shutdown on the cells managed through ExaCLI
- Enterprise linux / compute nodes best practices still apply!
Caging

- Resource manager = mandatory
- Cgroups not supported for DB processes
  - Example: Oracle DB + Oracle R => put in own DomU
- Cage your os processes (the ones which aren’t allowed to cap with groups)
Mos Support Addict ;-) 

You love to open MOS tickets ;-)
High level overview

exa123db01

- exa123adm01vm01
- exa123adm01vm02
- exa123adm01vm03
- exa123adm01vm04

vexa123-clu1
exa123-scan1

exa123-adm02vm01

vexa123-clu2
exa123-scan2

exa123-adm02vm02

vexa123-clu3
exa123-scan3

exa123-adm02vm03

vexa123-clu4
exa123-scan4

exa123-adm02vm04

exa123db02
SR-IOV Virtualization with Xen

- SR-IOV exposes VF enabling InfiniBand access for VM
- PF is hosted in Dom0 and is responsible for dynamic allocation of IB resources (PKeys, OP, CQ, memory region, ...) to the VFs
- PF owns QP0 and virtualize QP1 to make it available to all VMs
- Each VM has the Infiniband stack instantiated and can leverage RDMA-enabled protocol for enhanced performance.
Disk layout example

- 12 disks per cell
- 4 data disk groups
- 4 reco disk groups
- data1 + data3 were extended
- reco4 was extended
- Dbfs_dg diskgroup
Move 2 OVM
Upgrade to 12.1
Application upgrade

April QFSDP 2015 because bare metals were running that version
Issue 1

First install done by ACS

• OEDA
  • dos2unix on xml files created by OEDA
  • usernames / groups cannot have _ eg rac_dba

→ fixed in later versions

• GMT issue
  • symlink in /usr/share/zoneinfo to GMT (GMT) due to GMT ilom settings
  • adapt cell.conf

→ fixed in later versions
Issue 1

• ssh key equivalence failing
  • primary groupname and oracle user had the same name, we needed to adapt /etc/profile and /etc/bashrc to make sure the umask is always 022 in the vm’s

—> still valid in image 12.1.2.3.1 / OEDA
  august2016

—> Finally fixed in OEDA april 2017
Issue 2

Oracle 12c
quite some issues
  • expdp impdp
  • truncate generate ORA-0600
  • DBMS_STATS
  • wrong results
  • Cursor S wait for X

bottom line quite some one-offs to put on top of QFSDP April 2015
Issue 3

performance

• insert append slower on OVM
  cluster_database = false
  performance ok
  cluster_database = true
  1 instance or 2 instances up
  performance not ok

weeks and weeks and lots of escalation meeting further
Issue 3 cont

Cpu scheduling issue :
RAC uses other code path then single instance
CPU flags not passed correctly to DomU’s (X5)

result : Two month delay GO Live

details here :
Performance slower in OVM when compared with Baremetal (Doc ID 2080843.1)

--> Also Valid in Exadata X6-2
Issue 4

OEDA doesn’t assign groups for:
- syskm
- sysdg
- bckdba

Solution:
edit $ORACLE_HOME/lib/config.c
+ relink

Recently an enhancement request was opened for this
Issue 5

Stopping a domU takes up to 1 hour !!!!!
-> fixed in QFSDP OCT 2015
Issue 6

unplugging public network did not trigger VIP failover

-> GI 12.1.0.2 image 12.1.2.1.1
   -> 12.1.2.1.1 not supported anymore
-> no fix but must set ping targets

```
srvctl modify network -k 1 -pingtarget "<UsefulTargetIP(s)>"
```

Check certification matrix it changes over time mos 888828.1
Things to consider

- Avoid swapping from DomU’s
- Check 888828.1 for critical issues
  - If necessary PATCH!
- Kernel panic’s seen whilst coupled to ZFSSA
  - new Firmware on the HCA’s loaded
- Always use latest OEDA
  - using OEDA upgrades rpms
    (eg. exadata-ovmutils,...)
- having a lot of OVM’s means:
  each OVM needs GI +DB + OS upgrade
Always use the latest version
—> get it from mos-note: 888828.1
—> Nov 2017 v171130 - Patch 27179375
GIMR not default installed
—> ENABLEMGMGMTDBCONFIG: false -> true
upgrades rpms
- ovmwatch
- ovmwatch
- exadata-ovmutils
- ...
Patching (1/6)

Standard Exadata Patching order:

- Cells
- OS
- Infiniband
- GI + DB
Patching (2/6)

- Use `patchmgr` or `dbnodeupdate.sh` for each DomU and on Dom0:

  ```
eg: ./patchmgr -dbnode /root/dbs_dom0_group_db02
           -dbnode_precheck -dbnode_loc <FULL_PATH>/p22277124_121221_Linux-x86-64.zip
           -dbnode_version 12.1.2.2.1.160119
  ```

- Be careful, 12.2 FIRST domU
- Firmware/Bios is upgraded when Dom0 is upgraded
Collisions on several one-off patches
12.1.2.1.1 image contained a patch on GI to make it work on OVM
Patching (4/6)

- oplan to generate steps
- Rollback the conflicting GI and DB patches
- install the new patches
- don’t run post patch until GI is patched
Patching(5/6)

Patching to 12.2.1.1.1.1 APR 2017

Issues
- domU with snapshots still mount by installer
- libserver.a corrupt RCA still ongoing
- patch mgr bricks IB switch
  - going from 2.1.8 -> 2.2.5-3
    - major upgrade Centos -> OEL
  - issue file permission
  - hwclock issue ...
Patching (6/6)

- dom0 doesn’t boot
- udev rules gone eth0 -> eth7…8
- no IB coming up
- SR still open and ongoing
- rebooting one switch dismounts diskgroup on one node of RAC

All this still ongoing as we speak
to summarise

- Projectplan!
  - platform / application /version upgrade together?
- young product!
  - bugs but maturing / evolving quickly
- have 10vm’s on one exadata means lot of work

- Sizing
- Check certification matrix !!
- Manage Expectations
@pfierens
@vanpupi
philippe@fierensconsulting.eu
pieter.vanpuymbroeck@exitas.be