Building Oracle on LXC Linux Containers

Gilbert Standen, Solution Architect
Violin Memory

Presenter Information

- Gilbert Standen
- Solution Architect, Oracle & Virtualization, Violin Memory, Inc.
- Presenter, NYOUG, 2007, 2008, 2014
- Presenter, AUSOUG, “Oracle Communities with 20:20 Foresight”, 2006, 2007
- Presenter, (Violin Memory All Flash Arrays), Oracle Open World, 2014
- Production Oracle DBA (20 Years)
- Violin Memory All-Flash-Array (AFA) Evangelist
- Ubuntu Linux Evangelist (Oracle EE DB on Ubuntu / Oracle RAC on Ubuntu)
- Virtualization Evangelist
- Mountain Biker and Digital Trance Music Bum

Presentation Description

- Today I want to show you the amazing performance and manageability advantages of running Oracle DB workloads in LXC Linux Containers.
- I will share names of the very largest datacenter organizations in the world which have converted part or all of their infrastructure to Linux Containers.
- Specifically, we will look at the implementation of Oracle EE databases in Linux Containers and hopefully include some practical hands-on information.
- I will share with you some performance graphs generated by SLOB running Oracle databases in LXC Linux Containers on unremarkable commodity SuperMicro dual-socket servers attached to a Violin 6616 SLC all-flash array.
- Why am I doing this? My main reason is to hopefully convince you that Linux Containers are the best high-performance way to run any Oracle databases at high-performance, also achieving efficient utilization on Linux.

Credits and Recognition

- A large number of people and organizations made this presentation possible.
- Caryl Lee, Executive Director, NYOUG
- Michael Olin, President, NYOUG
- Violin Memory Lab Support (Justin Pham, Jalal Schenk, Charles Vuong)
- Kurt Schmidt
- The woman to whom I am married 20 years, Dr. Yelena Belyaeva-Standen
- My mother, Mary Elizabeth "Betty" Standen (nee Bell)
- My father, who passed away in 1996, who believed you can be/do anything
- The large and growing number of people blogging on this topic.
- The FlashDBA website http://www.flashdba.com

Oracle Performance Assessment Service (OPAS)

- Most performance graphs used in this presentation for comparing Oracle database running in OEL 6.6 KVM guest vs. Oracle database running in OEL 6.6 LXC Linux Container were generated using the Violin Memory Oracle Performance Assessment Service (OPAS), a FREE online performance tool.
- OPAS is a FREE web tool at http://awr.vmem.com . It takes as input a zip file containing a set of Oracle Automated Workload Repository (AWR) reports in TEXT format ONLY and generates online a performance assessment report complete with graphs on latency, throughput, IOPS and many more metrics.
- OPAS is available on the www and is a FREE TOOL to quickly get a performance assessment of an Oracle database. This is a service for which some companies have charged several thousand dollars or more.
- A couple of the graphs at near end were generated by www.chartgo.com

Resources

My blog: https://sites.google.com/site/nandydandyoracle

The above site has very detailed hands-on, step-by-step, how-to guides on how to deploy Oracle 12c Grid Infrastructure Standalone Oracle Restart enabled databases, and Oracle 12c GNS ASM RAC Clusters in OEL 6 LXC Linux Containers. These guides include copious blockquote command codeblocks.

Guides are there, and I have tested, Oracle 12c Standalone and RAC deployments in Oracle Enterprise Linux 6.5 LXC Linux Containers on both Oracle Enterprise Linux 6.6 and on Ubuntu 14.04.1. Slides here are intended just to give a flavor of the overall simplicity of LXC.

Wim Coaekerts Blogs on Linux Containers…

Violin MEMORY

- "This really shows the power of containers on Linux and Linux itself. We have all these various Linux distributions but inside lxc (or docker) you can run ubuntu, debian, gentoo, yourowncustomcrazything and it will just run, old versions of OL, newer versions of OL, all on the same host kernel."
- "I can run OL6.5 and create OL4, OL5, OL6 containers or docker images but I can also run any old debian or slackware images at the same time."
- "To try out LXC, you need to install the LXC packages. **LXC is capable of using Oracle VM Oracle Linux templates as a base image to create a container.** "
- "The use of btrfs in particular and being able to easily create clones/snapshots of container images. To get started : have an Oracle Linux 6.5 installation with UEK3 and LXC installed and configured."
- "LXC by default uses '/container' as the directory to store container images and metadata" [Note: On Ubuntu it uses '/var/lib/lxc' as the default container storage directory).

The "Big Picture" on Linux Containers

Since this presentation is about "Building Oracle on LXC Linux Containers" it doesn't really cover the big picture. James Bottomley has described that here:

The Future of Containers in Linux and OpenStack

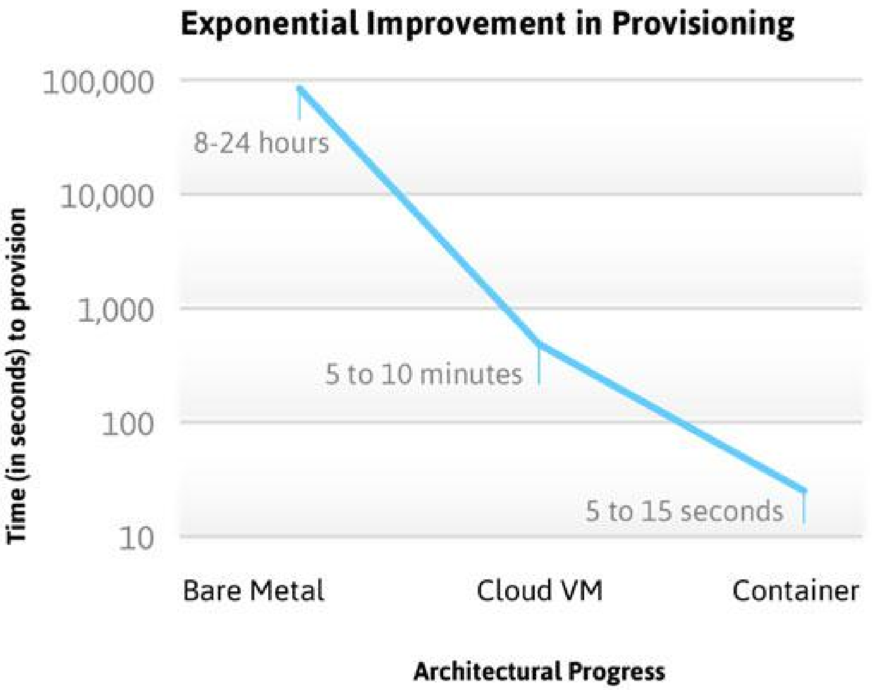
James Bottomley is CTO of Virtualization at Parallels. He is a kernel developer and maintainer of the SCSI subsystem and PA-RISC architecture. He is also the Chair of the Linux Foundation Technical Advisory Board (TAB). Highly recommended viewing.

Also, this link by Joe Beda, Sr. Staff Software Engineer, Google Cloud Platform, from June 2014 should also be viewed. Google spins up **2 billion containers per week**, or about 3,300 containers every second, of every day, of every year.

Everything at Google Runs in A Container

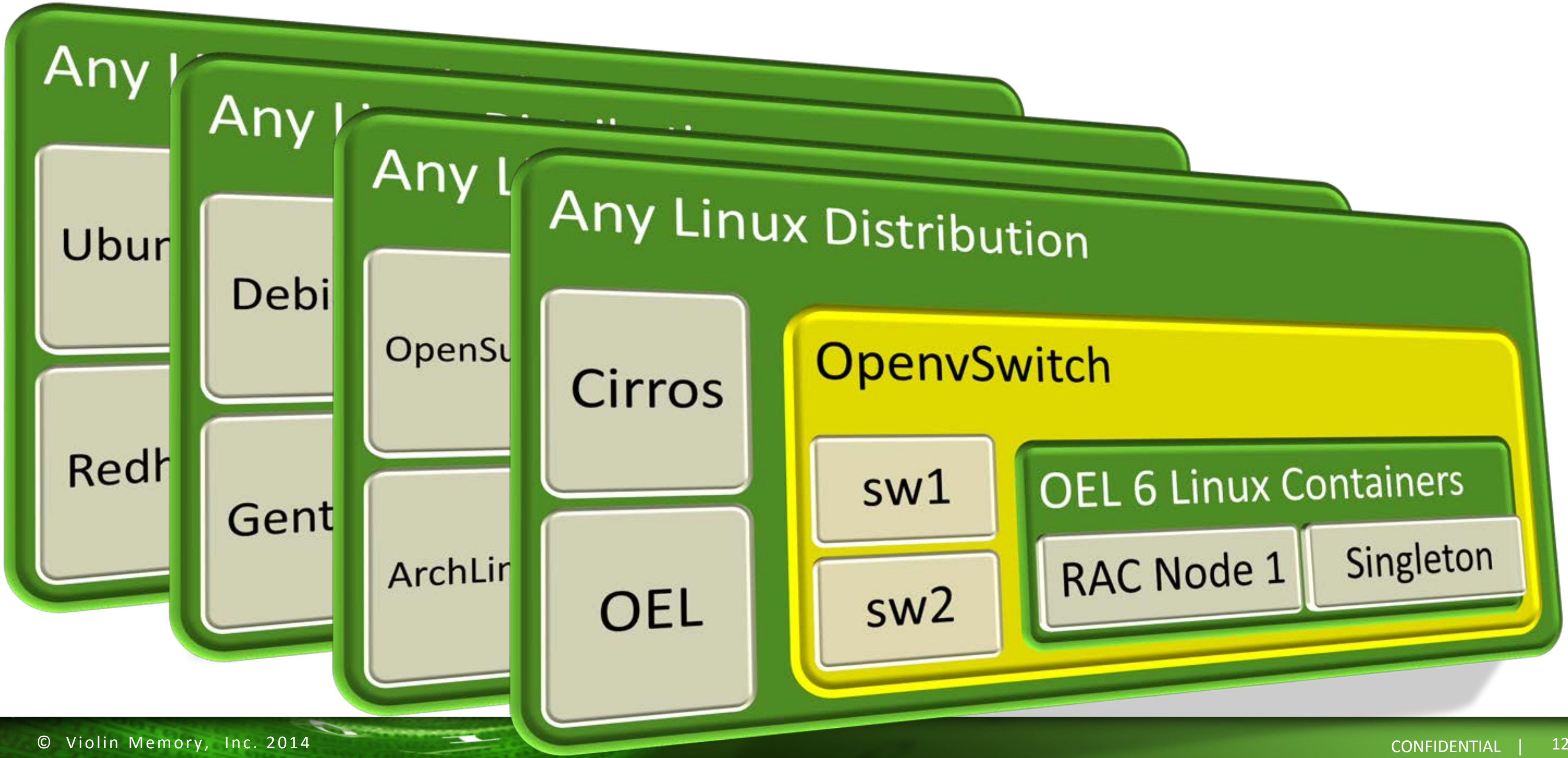
Service Providers Already Using Linux Containers

Large Hosting Providers Using Linux Containers

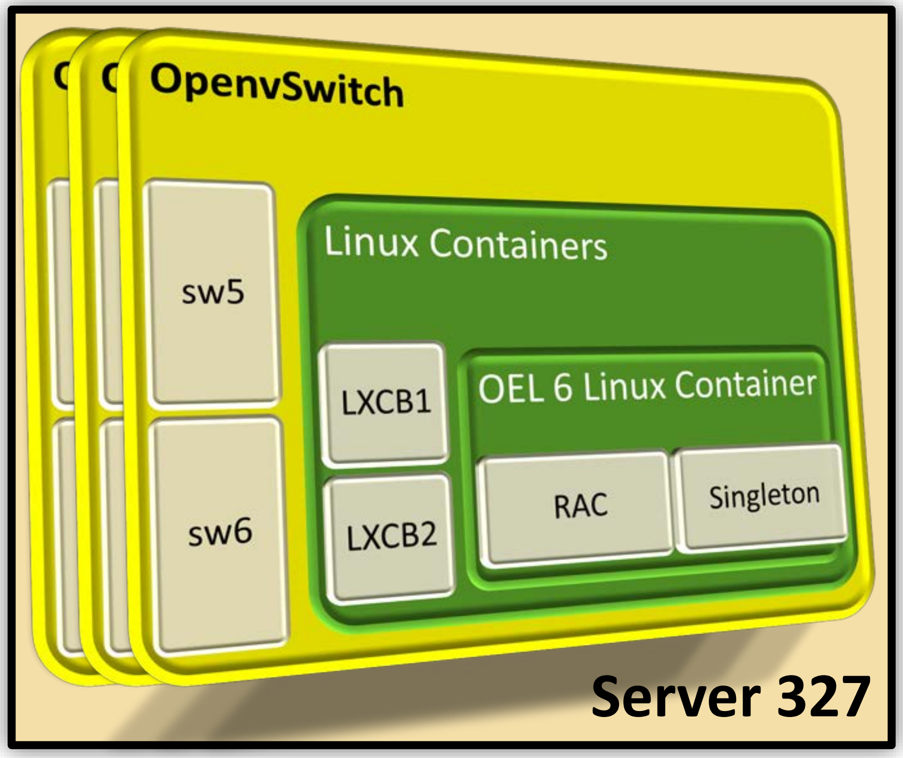


Other Technology Providers – You & Me, and…

A Tale of Two (or more) Linux Architectures

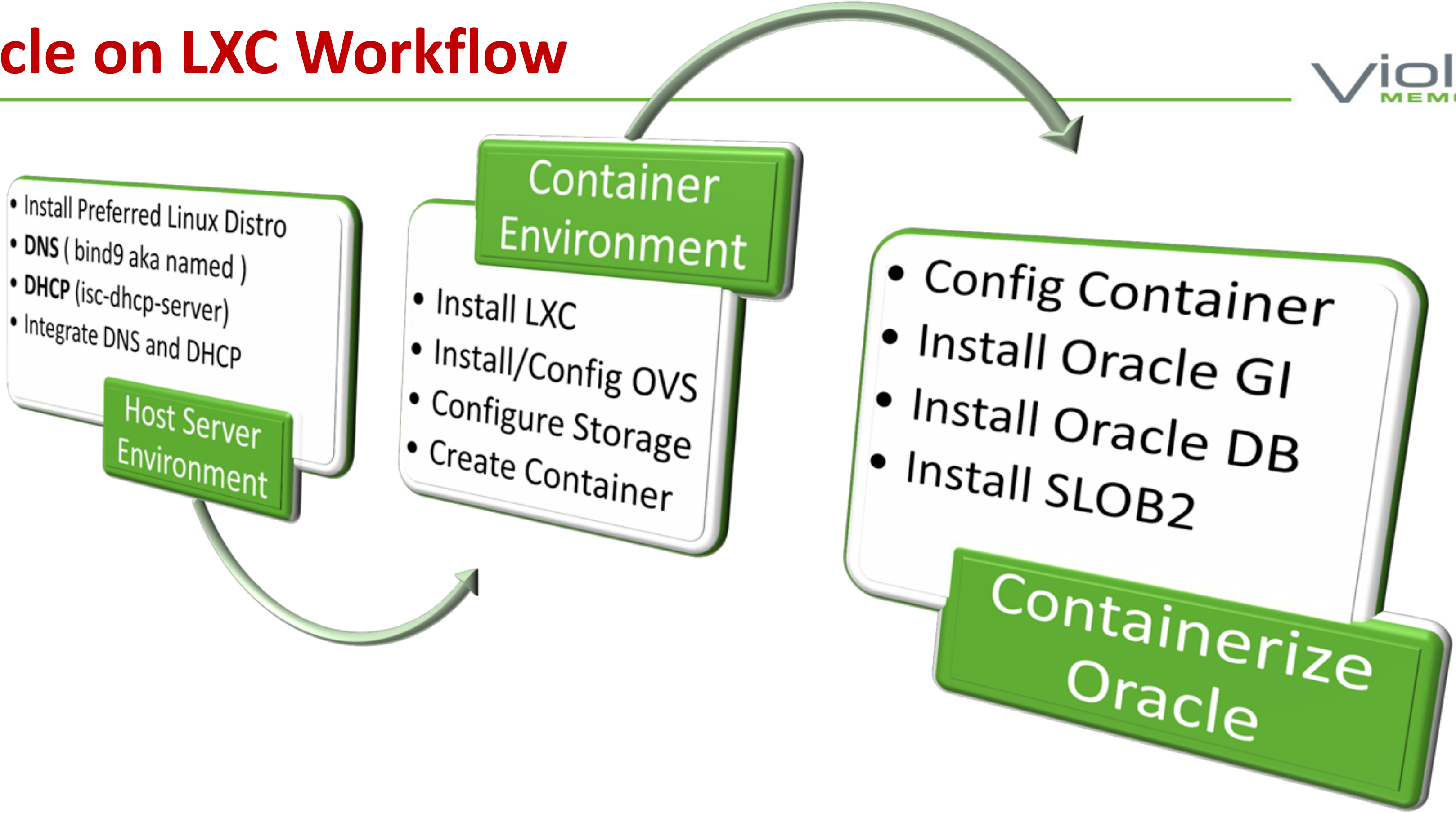


A Tale of Portability: Move Switches with Container



Server 701

Oracle on LXC Workflow



Configure Container Environment (OEL6 / UEK3)

Step 1

- wget http://public-yum.oracle.com/beta/public-yum-ol6-beta.repo
- sed -i s/enabled=0/enabled=1/g public-yum-ol6-beta.repo

Step 2

- yum install openvswitch lxc tunctl bind-* dhcp oracleasm-*
- service openvswitch start; chkconfig openvswitch on; ovs-vsctl show
- configure DHCP DNS autoupdate; create an openvswitch

Step 3

- lxc-create -t oracle -n <container-name> (installs most recent OEL)
- lxc-create -n ol59 -t oracle --R 5.9. (installs specific OEL version)
- configure storage for Oracle ASM (UDEV, multipath, ASMLib)

Create an OpenvSwitch

Step 1
- ovs-vsctl add-br sw1
- ovs-vsctl add-port sw1 s1

Step 2
- ip link set up dev sw1
- ip addr add 10.207.39.1/24 dev sw1

Step 3
- ip route replace 10.207.39.0/24 dev sw1
- ovs-vsctl set port sw1 tag=10

Containerize Oracle

Step 1
- create “/container/fstab” file; add Oracle ASM storage to fstab
- edit “/container/config” for openvswitch networking

Step 2
- cd /container/<container-name>
- edit container configuration (config) file

Step 3
- configure container for openvswitch(es) networking
- verify DHCP assigns address on container boot
- Install Oracle GI and Oracle DB; Install SLOB; benchmark

Connecting LXC Container to OpenvSwitch

```
#!/bin/bash
ovsBr='sw9'
sudo ovs-vsctl add-port ${ovsBr} $5
sudo ovs-vsctl set port lxcora02-asm2 tag=70
sudo ovs-vsctl set port ${ovsBr} tag=70
sudo ovs-vsctl set port ${ovsBr} trunks=60,70
```

```
lxc.network.type = veth
lxc.network.flags = up
lxc.network.script.up = /etc/network/if-up.d/lxcora02-asm2-ifup-sw9
lxc.network.script.down = /etc/network/if-down.d/lxcora02-asm2-ifdown-sw9
lxc.network.veth.pair = lxcora02-asm2
lxc.network.name = eth6
lxc.network.mtu = 9000
lxc.network.ipv4 = 172.221.40.10
lxc.network.hwaddr = 00:16:3e:24:c8:4b
```

```
#!/bin/bash
ovsBr='sw9'
sudo ovs-vsctl del-port ${ovsBr} $5
```

Some Notes on Web Connectivity for Containers

- So that networking from LXC Linux Containers will resolve, ping, and communicate with external (www) IP addresses, several configurations must be made according to the instructions here thanks to Venu Murthy.
- The **'/etc/selinux/config'** file must use the "**permissive**" setting. Do NOT use the "disabled" setting in OEL6.
- Include the below settings in the /etc/sysctl.conf as shown below.

  net.ipv4.conf.default.rp_filter = 0
  net.ipv4.conf.all.rp_filter = 0

- Do NOT disable iptables. Add rules to allow multicast, open up EM Express port. Complete details are at my blog and Venu's link.

  -A INPUT -m pkttype --pkt-type multicast -j ACCEPT

Configuring "fstab" for LXC Container

- Storage for the Oracle database can be presented to the LXC container using the LXC 'fstab' file which lives in the same directory as the 'config' file which on OEL will be '/container/<container-name>'. Below is an example of contents of that file.

  /dev/oracleasm/disks dev/oracleasm/disks none bind,create=dir

- This file tells the LXC container to mount /dev/oracleasm/disks on the host at /dev/oracleasm/disks on the LXC container, and to create the mount point if it does not already exist.
- This is one way to present storage to an LXC container.
- This is a good place to note that you CANNOT load modules inside a container. Any module you need, say 'oracleasm' must be loaded at the host level. There are hooks that can be used to access the module from the host.

Oracle 12c (12.1.0.2.0) ASM Flex Cluster and GNS

- Oracle GNS (Grid Naming Service) has particularly interesting implications for migration of RAC nodes running in LXC Linux Containers
- Because only a single IP address need be defined in corporate DNS (and the rest of the IP required for RAC handled by Oracle GNS) the mechanics of migrating an LXC containerized RAC node from one server host to another server host is thus greatly simplified.
- Even without live migration, one can envision updating local_listener on one of the RAC nodes, and draining off connections to the other RAC nodes, and then shutting down the drained RAC node and bringing it up on another host. The software Openvswitch and even the rootfs of the migrating RAC container can be pre-staged on the destination host.

LXC: Learning from the Past, Getting it Right…

- You can create your own company repository of containers and cut deployment time for Oracle systems to seconds rather than days by prebuilding everything then just deploying the container.
- All these hardware tricks are completely unnecessary for Containers. Things like VT-d, VT-x, etc. were simply for hypervisors.
- Hardware manufacturers will be able to turn away from chip optimizations for hypervisors.
- The Technical Advisory Board of Linux Foundation and other key players learned a great deal from the Xen vs. KVM "disaster". It was decided and is completed that kernel version 3.12 supports in-tree ALL containerization schemes on a standardized containerization set of kernel components.
- The youth of Linux is ending and the golden age of Linux is just beginning.

Linux Container Security…a Brief Mention

- LXD is coming promising security equivalent to a hypervisor.
- Hostile root has been a part of container technology for over 10 years ('root' in the container / 'nobody' in the host).
- Linux containers use POSIX file 'capabilities' to tell LXC what capabilities to drop before starting.
- Cgroups, apparmor, selinux, seccomp, and user namespace are also used.

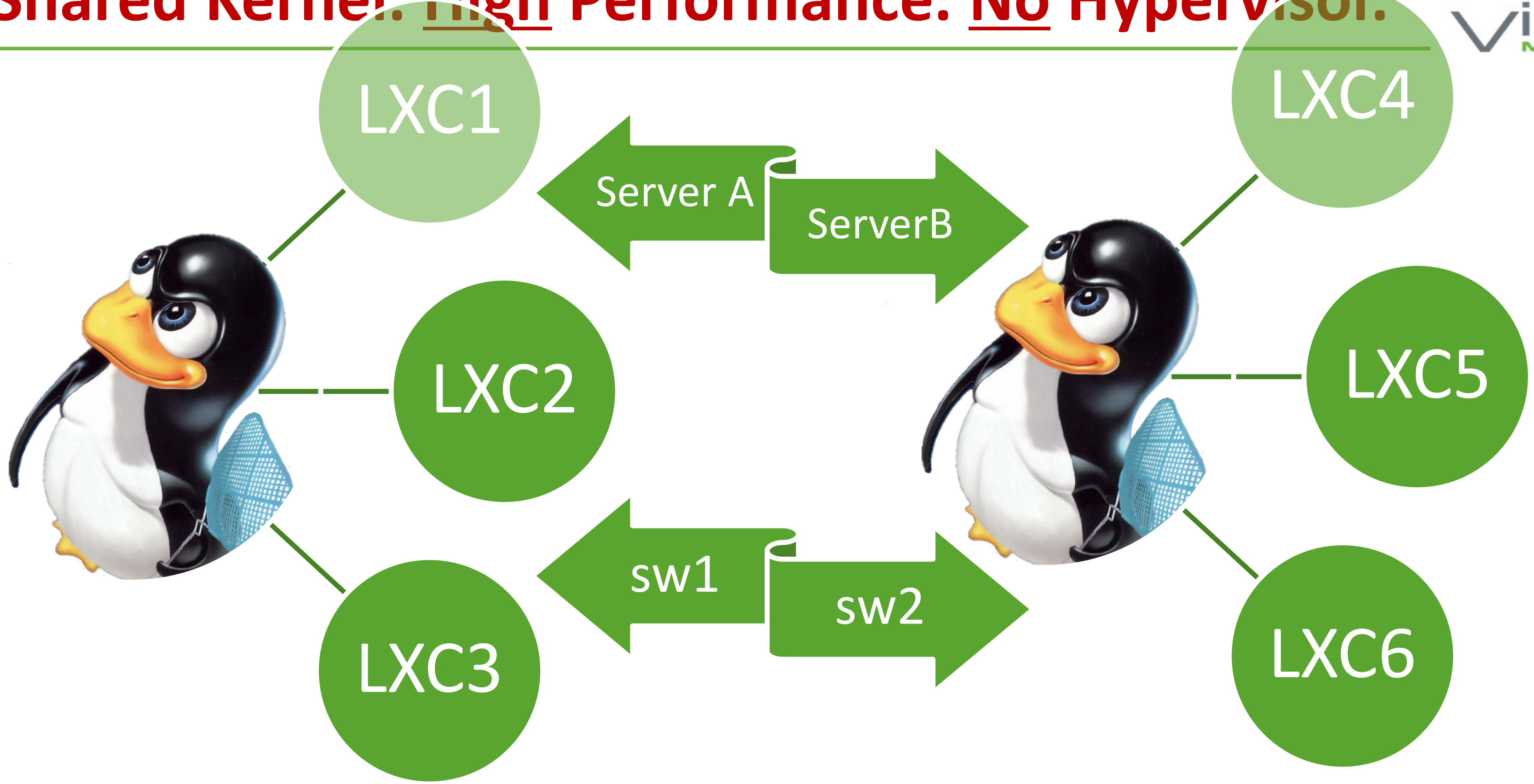
Hypervisors and Containers

- Hypervisors provide virtual emulated hardware. Examples are VMWare, KVM, VirtualBox and Xen.
- The problem with hypervisors is that the emulated hardware introduces additional bootup time, addition disk space requirements, and reduced IO, network, CPU performance.
- Over the years, various hardware and software strategies were used to try to boost performance.
- Linux Containers are based on sharing the one kernel and resources of a single operating system.
- Linux Containers use server bare metal hardware, use minimal disk space, and deliver bare-metal performance. The main problem one might envision is upgrading a kernel on a Container host.
- As noted, Oracle has solved the container kernel upgrade problem using their ksplice technology…

Oracle Ksplice (zero-downtime security updates)

- On July 21, 2011, Oracle acquired Ksplice Inc.
- Ksplice provides zero downtime update technology for Linux kernels.
- Ksplice Uptrack can apply 100% of the important kernel security updates released by Linux vendors without rebooting which is critical for hosts running hundreds of Linux Containers requiring zero-downtime operation.
- Ksplice Uptrack is available for Oracle Linux, free of charge, for Oracle Linux customers with a Premier support subscription.
- Anyone can use Ksplice Uptrack for free on Ubuntu Desktop and Fedora to learn how it works and test it out.
- 100,000+ servers at over 700 companies are protected with Ksplice Uptrack.
- Over 2 million production updates successfully applied by Ksplice Uptrack.

Shared Kernel. High Performance. No Hypervisor.



Violin Memory 6600-04-SLC All-Flash Array "6616"

Characteristic	6600
Model	**Violin 6616**
Flash Type	SLC
Raw Capacity (TiB/TB)	16 TiB/17.5 TB
Maximum Usable Capacity (TiB/TB)*	10 TiB/11 TB
Maximum 4KB IOPS (Mixed)	1,000,000
Maximum Bandwidth 100% Reads)	
Minimum Latency	200 µsec
VIMM Cnt (Data+Hot Spares)	60 + 4

SOFTWARE:

Software Version: G5.6.0

Installed Images:

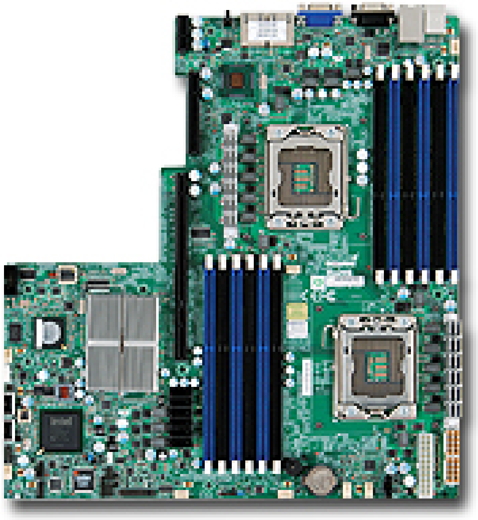
Partition 1

vMOS G5.5.2 #1-GA 2013-04-18 11:02:10 x86_64 common@eng-builds-fir:vmg:023edae

Partition 2 (currently booted) (to boot next)

vMOS G5.6.0 #2 2013-12-09 13:16:04 x86_64 root@ci-fir08:vmg:dc6a52f

SuperMicro X8DTU (Nehalem)



Key Features

1. Intel® Xeon® processor 5600/5500 series, with QPI up to 6.4 GT/s
2. Intel® 5520 (Tylersburg) Chipset
3. Up to 192GB DDR3 1333/1066/800MHz ECC Registered DIMM/48GB Unbuffered DIMM
4. Intel® 82576 Dual-Port Gigabit Ethernet Controller
5. 6x SATA2 (3 Gbps) Ports via ICH10R Controller
6. Left Side 1 (x16) PCI-E 2.0 + 1 UIO **or** 1 (x16) PCI-E 2.0 Right Side
7. Two dual-port 8Gb HBA’s connecting server to Violin 6616 SLC AFA

An unremarkable server, typical commodity server, average performance.

Specific Oracle Initialization Parameter Settings

- Parameters used were those prescribed by Chris Buckel at his "FlashDBA" website for use with SLOB2 benchmarking. The settings are shown on the following slides for reference and convenience.
- These parameter settings are designed to MINIMIZE buffer cache hit ratio in order MAXIMIZE physical reads and to MINIMIZE logical reads so that latency to storage is tested. Therefore, the buffer cache is dialed down to 16Mb.

Special Oracle Init Settings to Drive Disk IO

```
*._db_block_prefetch_limit=0
*._db_block_prefetch_quota=0
*._db_file_noncontig_mblock_read_count=0
*._disk_sector_size_override=true
*.cpu_count=1
*.db_block_size=8192
*.db_cache_size=16M
*.db_files=200
*.db_recovery_file_dest_size=200G
*.diagnostic_dest='/u01/app/oracle'
*.filesystemio_options='setall'
*.log_buffer=134217728
*.parallel_max_servers=0
*.pga_aggregate_target=1G
*.processes=500
*.recyclebin=OFF
*.resource_manager_plan=''
*.remote_login_passwordfile='EXCLUSIVE'
*.shared_pool_size=1G
*.undo_tablespace='UNDO1'
```

- Disabled Oracle Scheduler
- Five Redo Log groups, Multiplex
- Multiplexed across ASM DG's (2)
- Two members in each group
- Size 1 Gb each*, "blocksize 4096" redo logs for flash
- *for KVM optimization run, 10Gb redo logs were used to help reduce log switch waits.
- All Oracle installs were "separation environment" GI owned by "grid" and DB owned by "oracle" using Oracle Restart.
- All SLOB harness tests were run against Oracle 12c (12.1.0.2.0) pluggable databases (PDB) and all databases used for results in the presentation were singleton (non-RAC) databases.
- Note that ALL charts are Oracle AWR data!

SLOB2 Harness Configuration Files

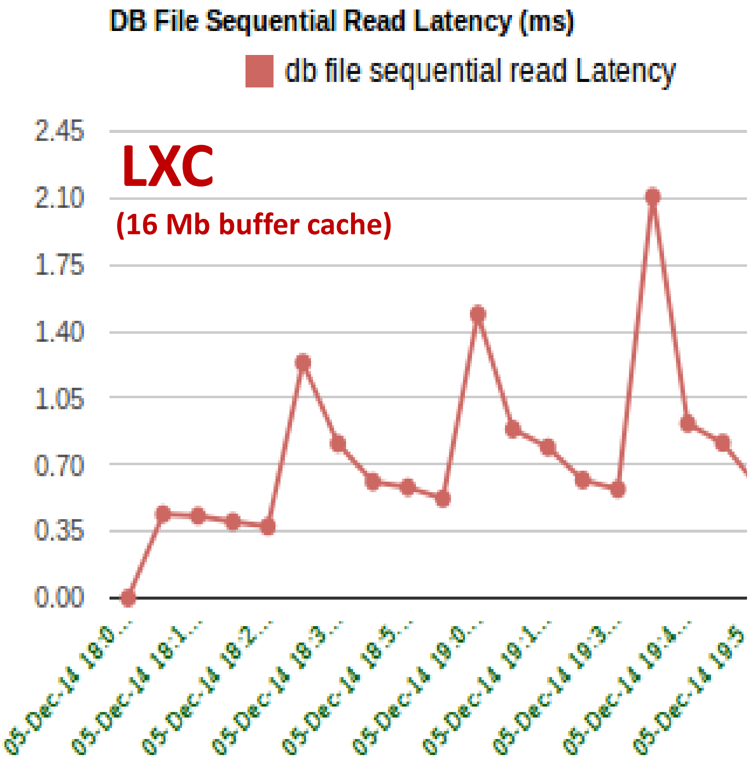
Slob2-harness.inf

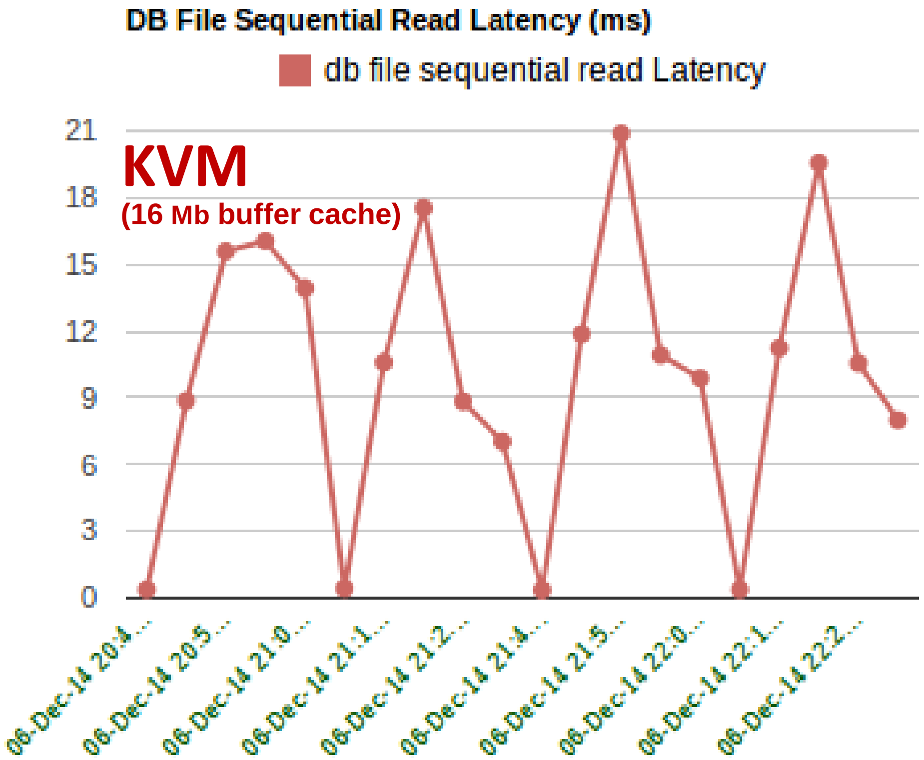
```
</pre>
# slob2-harness.inf
# Metadata file for slob2-harness.sh
# The UPDATE_PCT section defines the percentage of
transactions run by SLOB which are DML
UPDATE_PCT
0
30
60
90
/UPDATE_PCT
# The WORKERS section defines the number of SLOB
processes to test
WORKERS
001
032
064
096
128
/WORKERS
# EOF
<pre>
```

Slob.conf

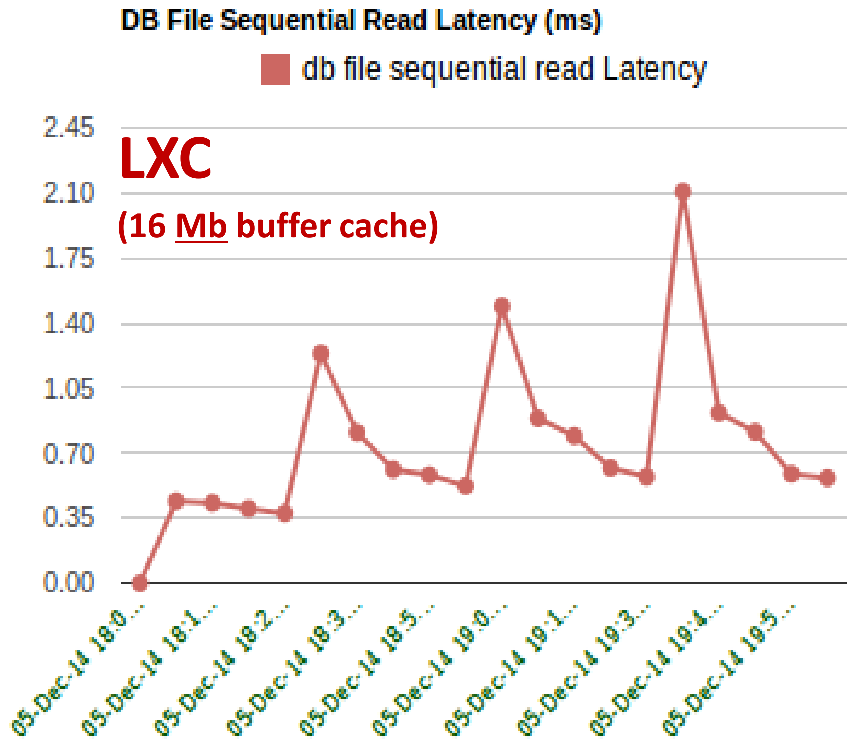
```
RUN_TIME=300
WORK_LOOP=0
SCALE=10000
WORK_UNIT=256
REDO_STRESS=HEAVY
LOAD_PARALLEL_DEGREE=8
SHARED_DATA_MODULUS=0

# Settings for SQL*Net connectivity:
#ADMIN_SQLNET_SERVICE=slob
#SQLNET_SERVICE_BASE=slob
#SQLNET_SERVICE_MAX=2
#SYSDBA_PASSWD="change_on_install"

export UPDATE_PCT RUN_TIME WORK_LOOP SCALE
WORK_UNIT LOAD_PARALLEL_DEGREE REDO_STRESS
SHARED_DATA_MODULUS
```

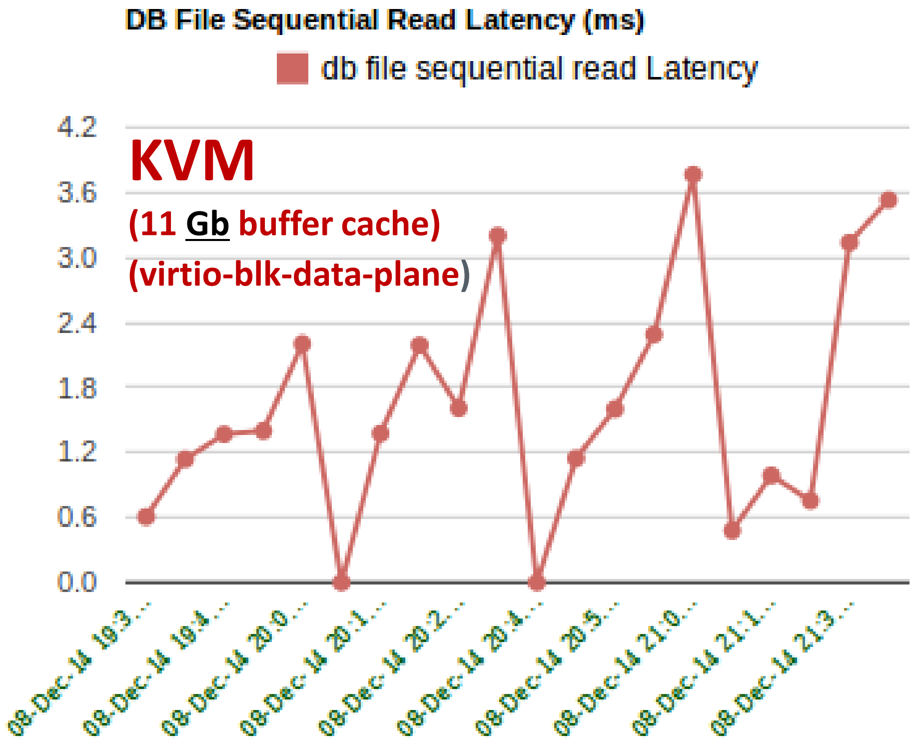
DB Sequential Read Latency (LXC vs KVM)



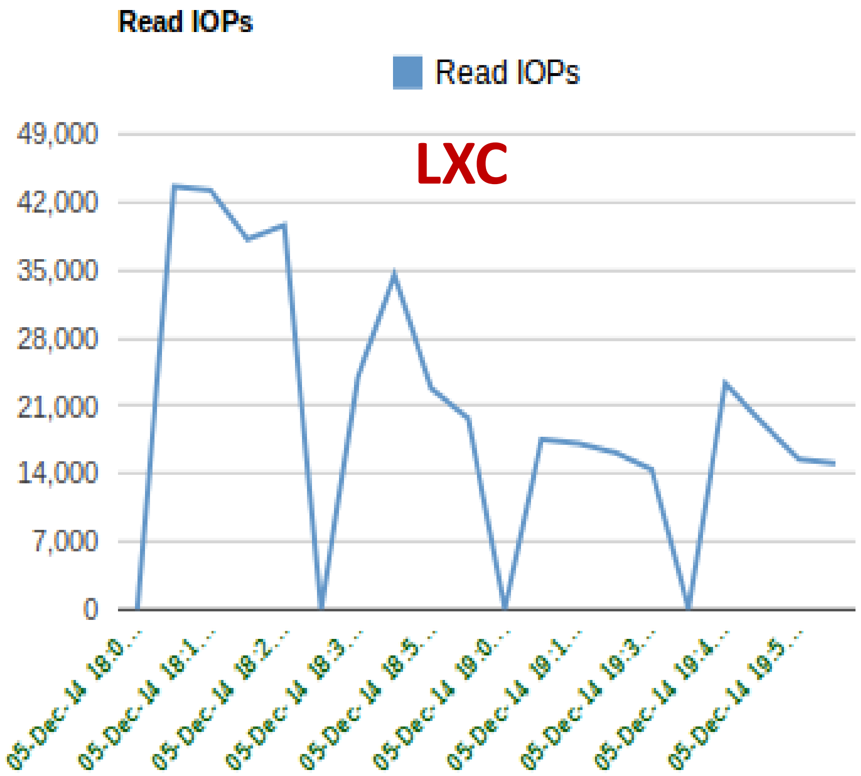


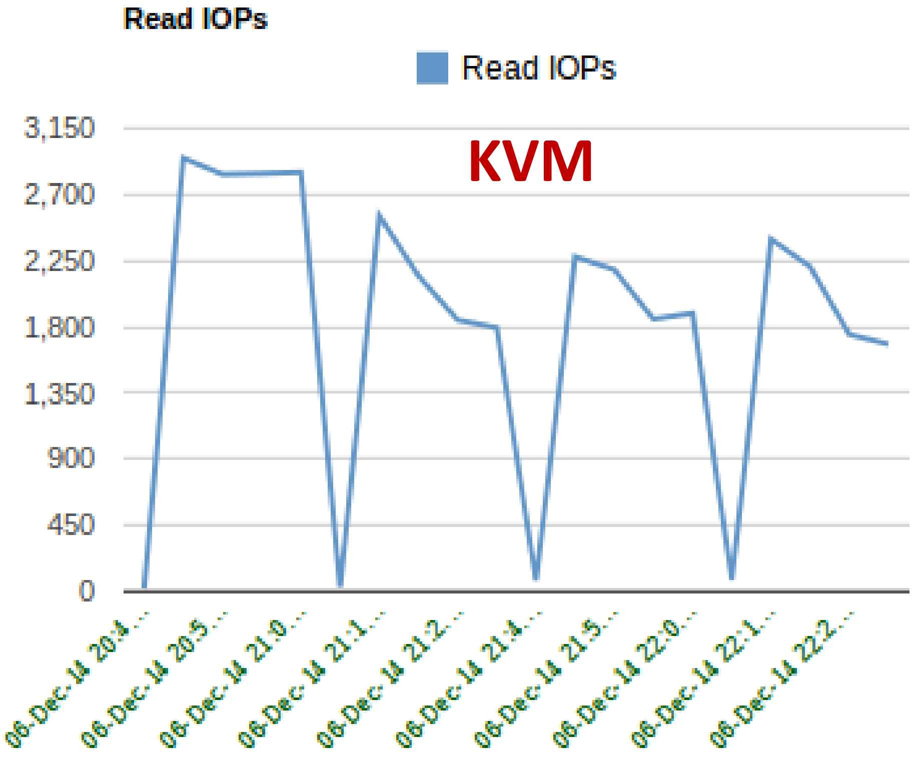
DB Sequential Read Latency (LXC vs KVM)



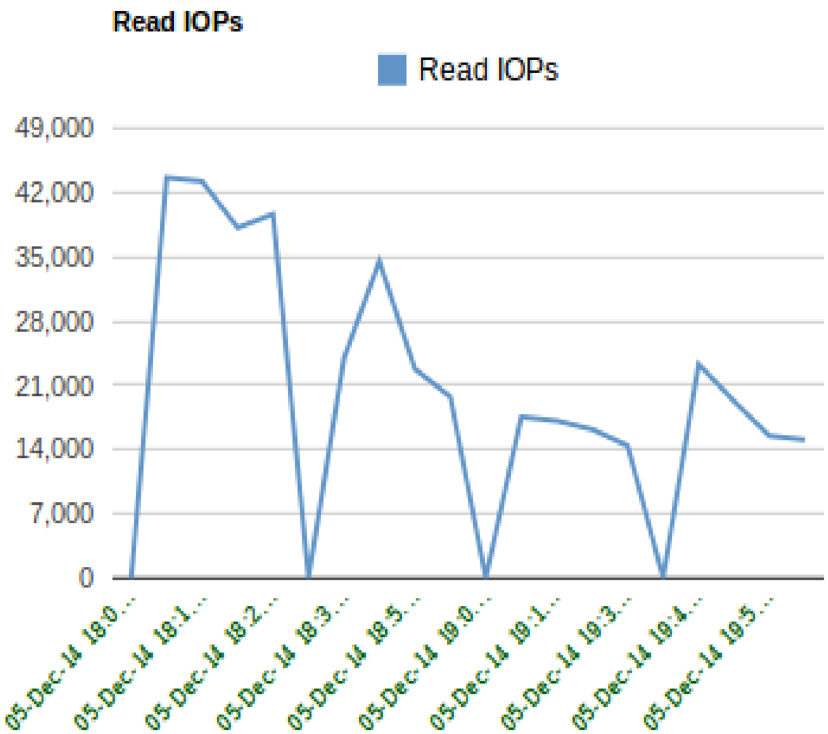


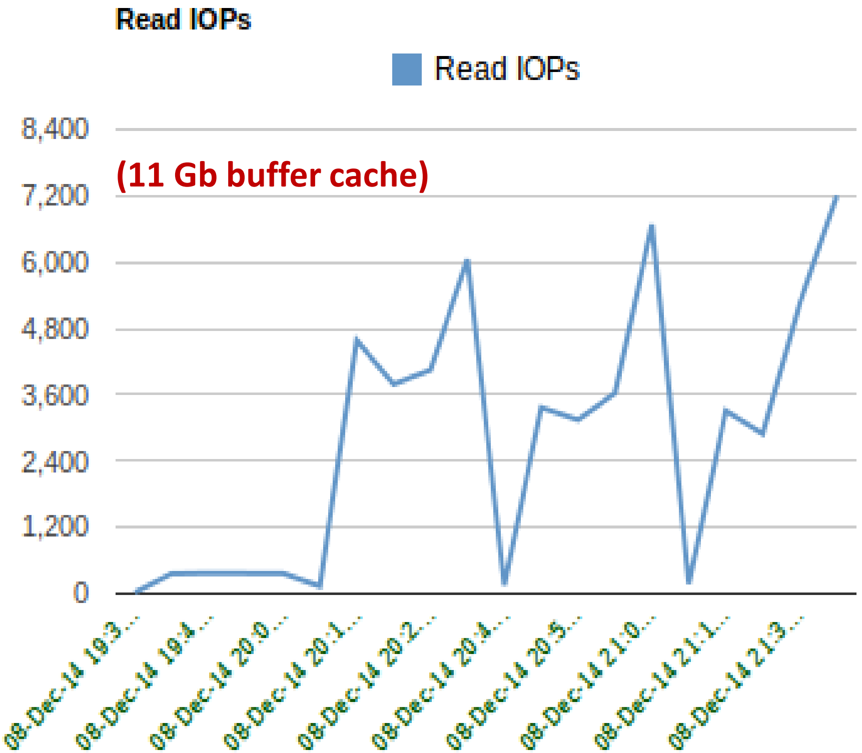
DB Read IOPS Performance (LXC vs KVM)



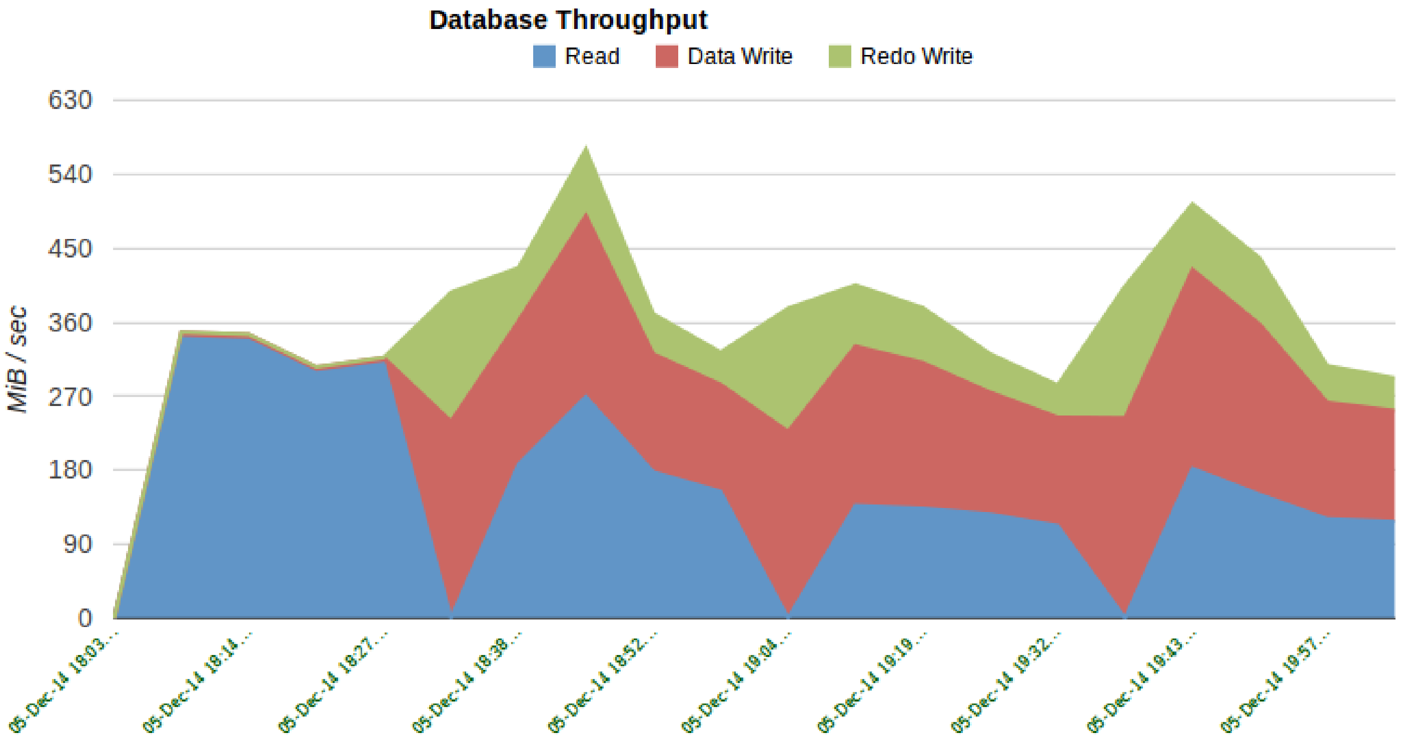


DB Read IOPS Performance (LXC vs KVM)

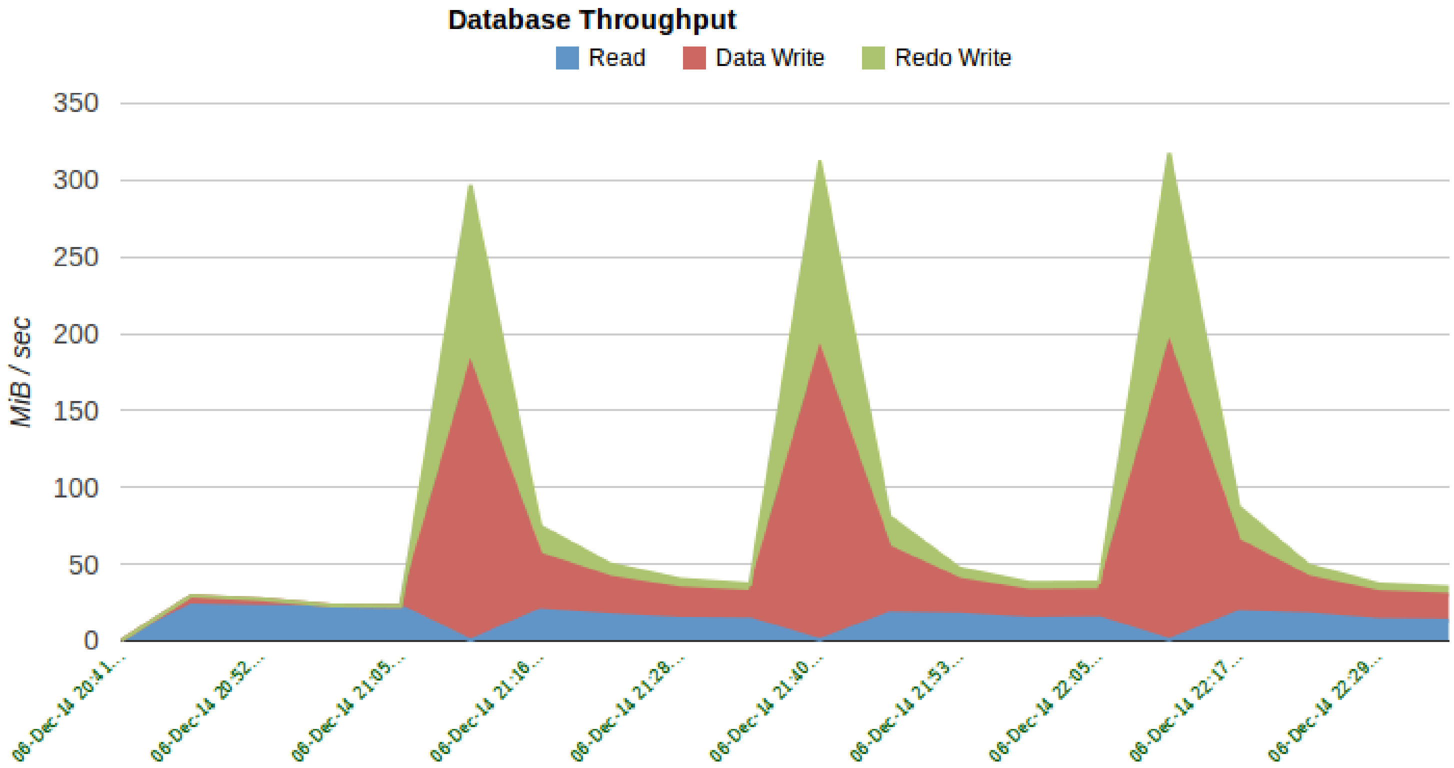




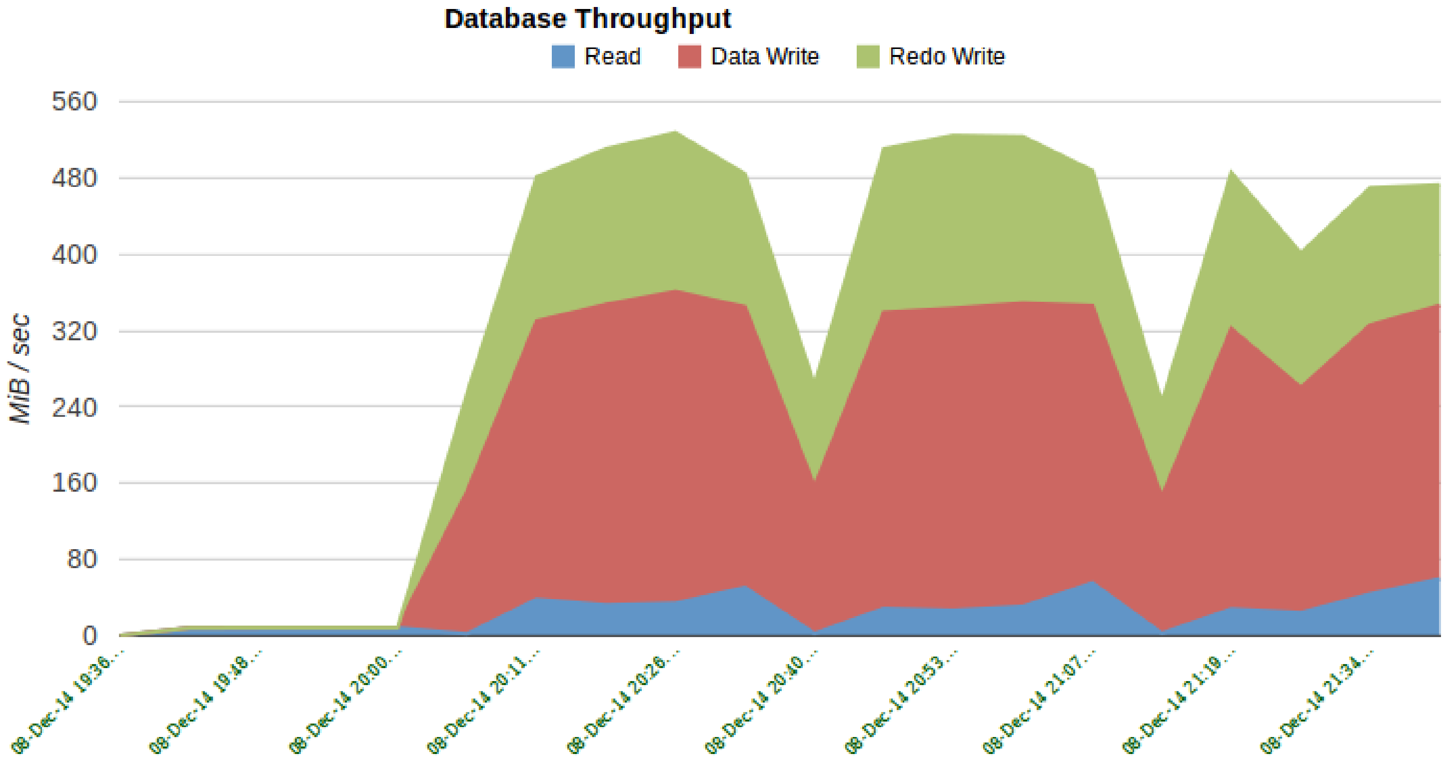
Oracle DB Throughput LXC (16 Mb buffer cache)



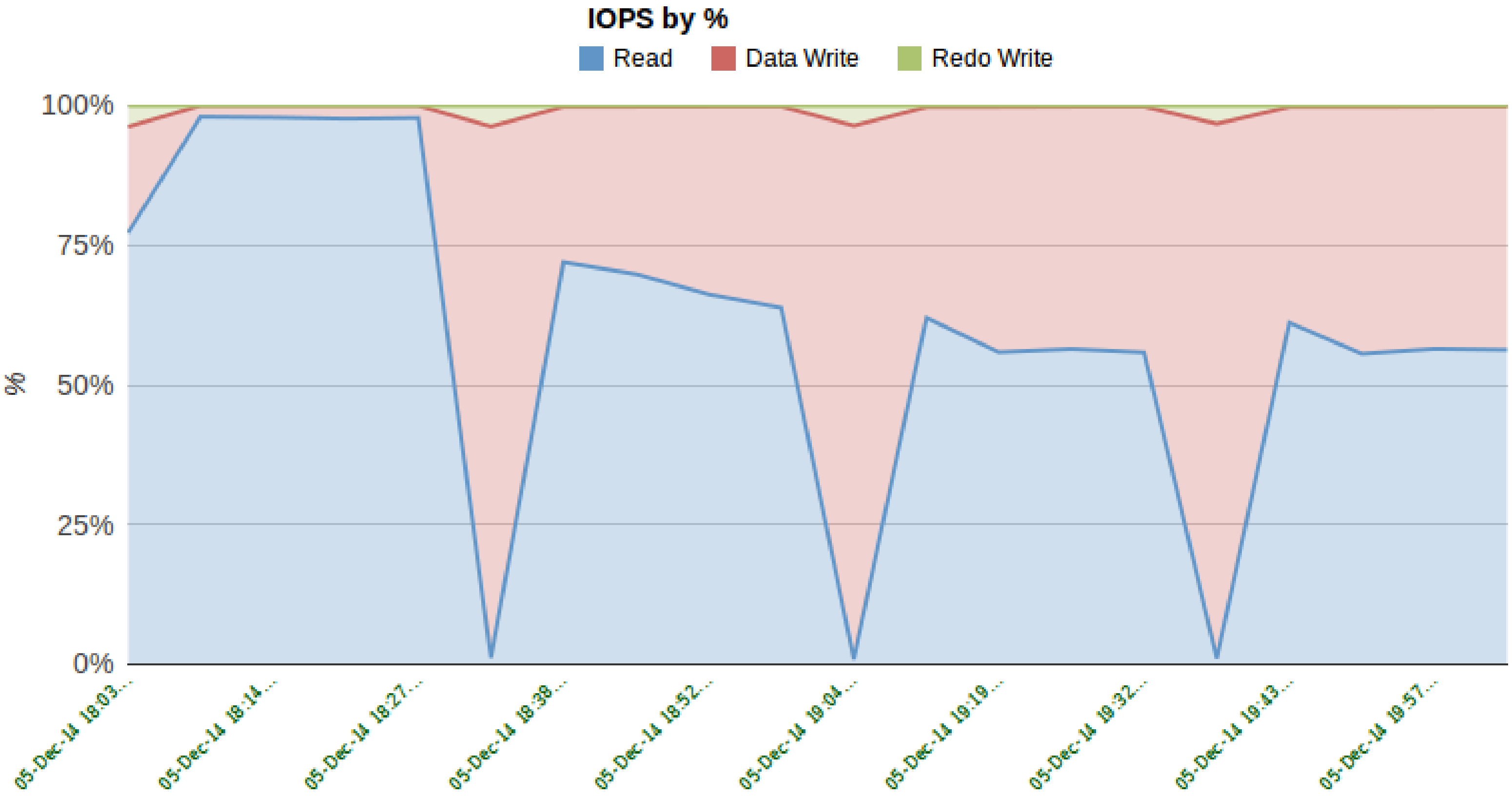
Oracle DB Throughput KVM (16 Mb buffer cache)



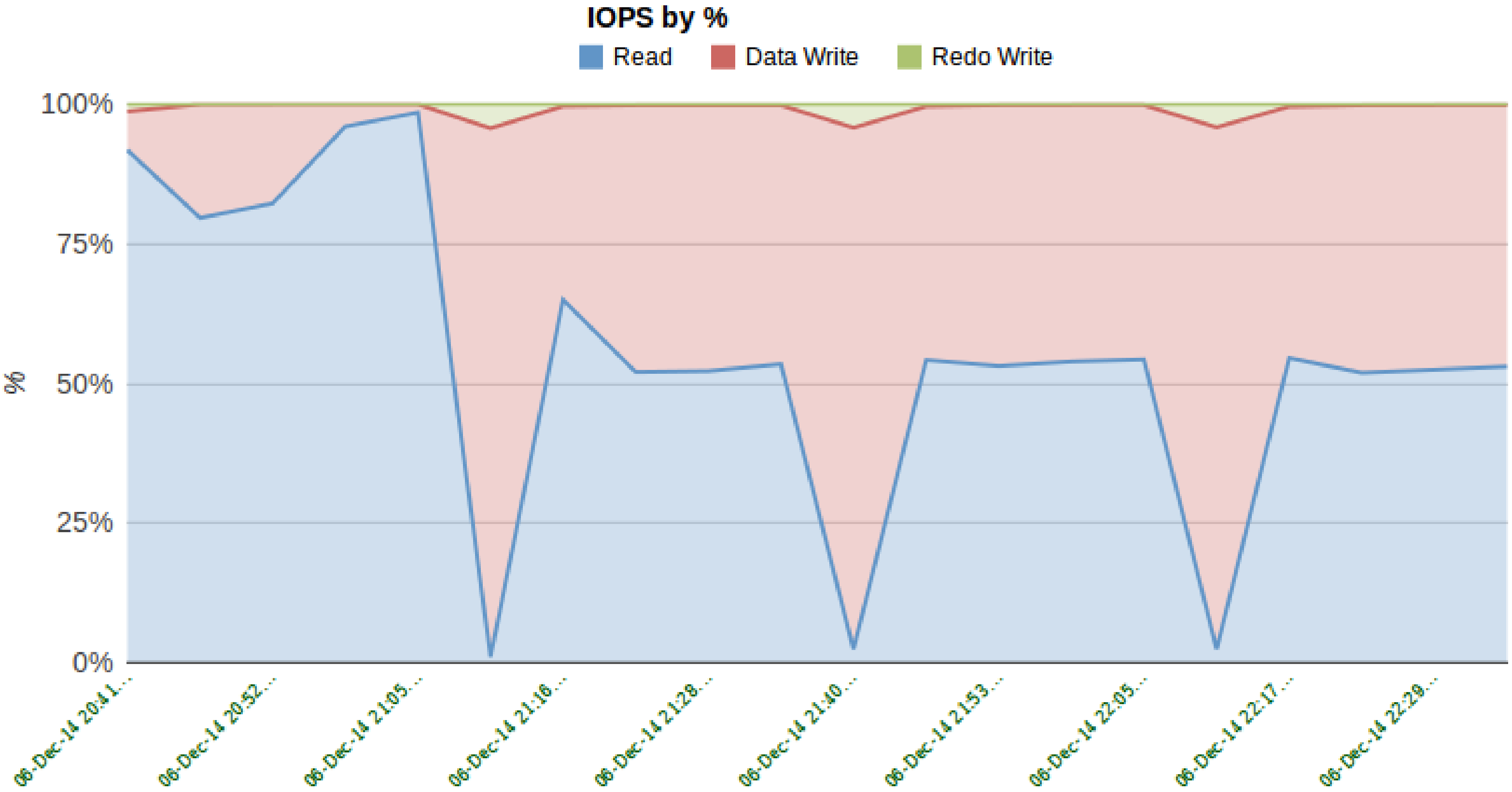
Oracle DB Throughput KVM (11Gb buffer cache)



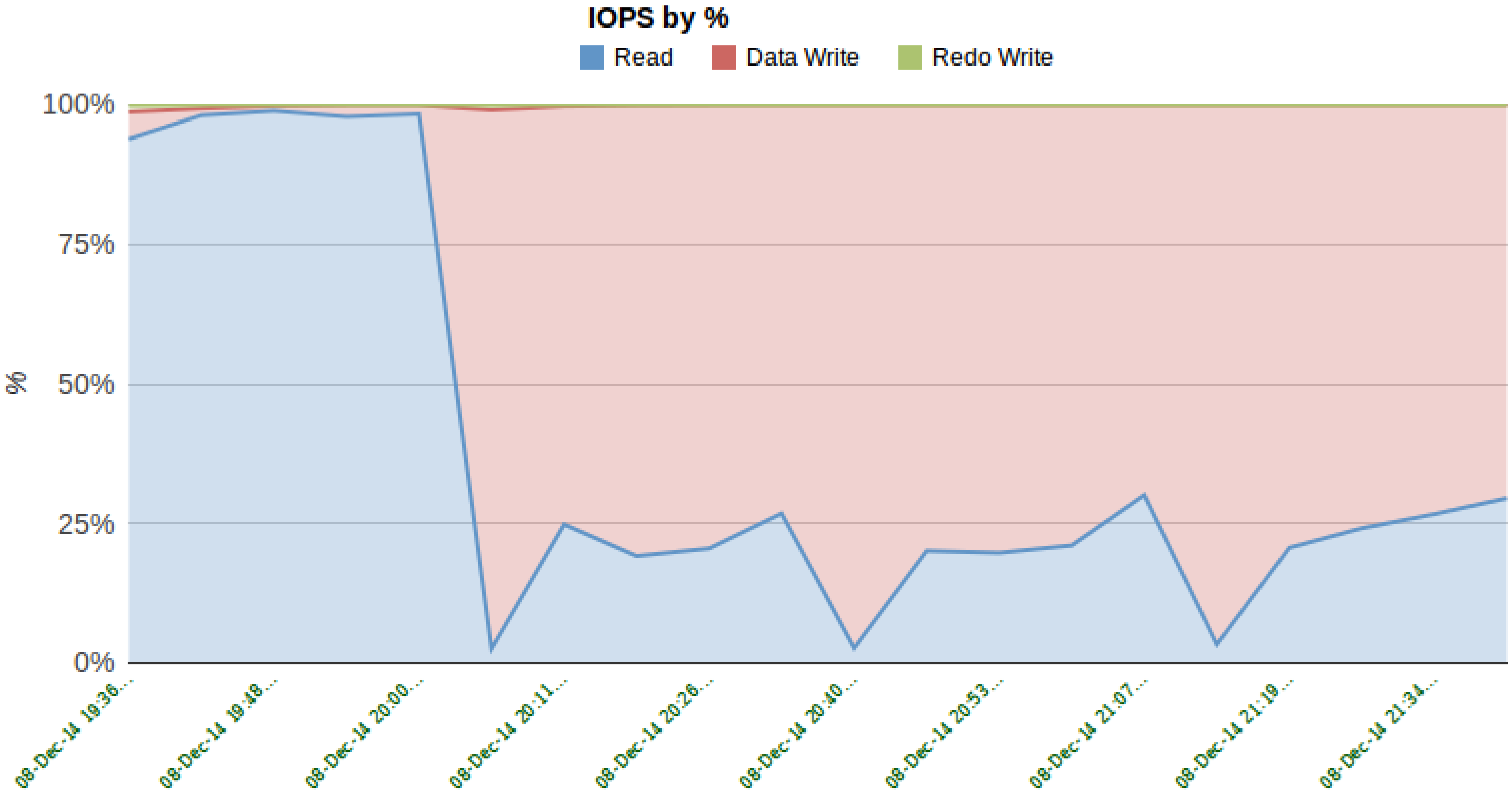
DB IOPS by Percent LXC (16 Mb buffer cache)



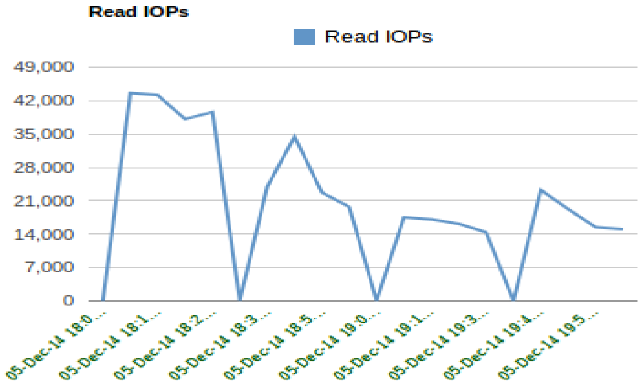
DB IOPS by Percent KVM (16 Mb buffer cache)

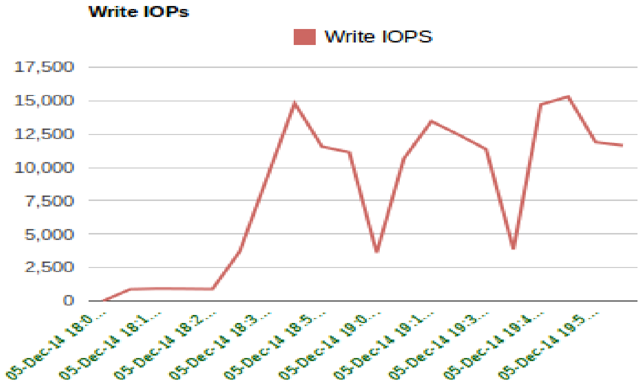


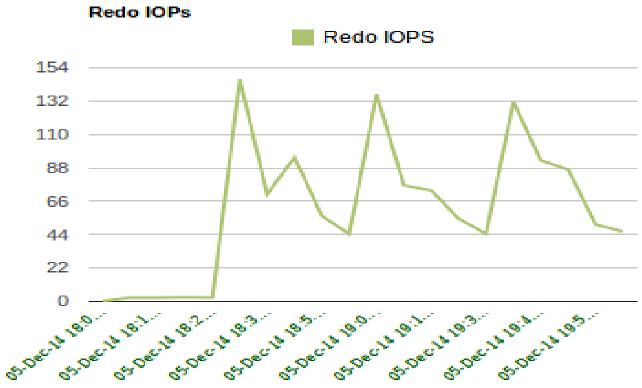
DB IOPS by Percent KVM (11 Gb buffer cache)

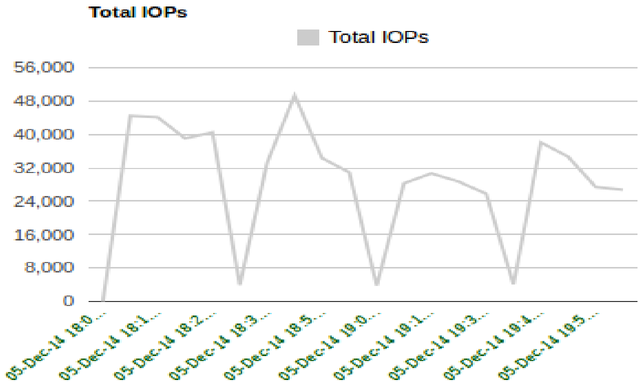


All Database IOPS LXC (16 Mb buffer cache)

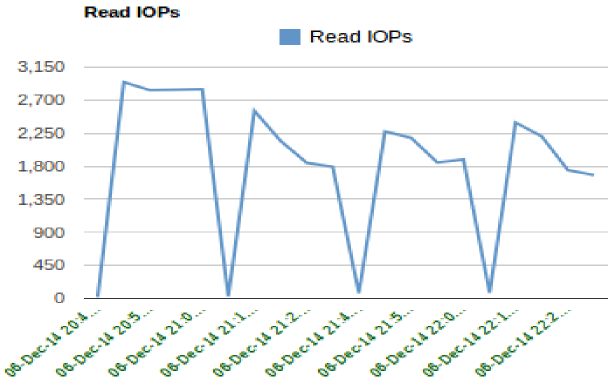


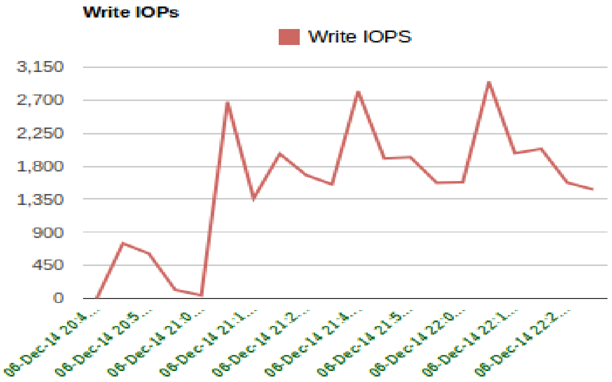


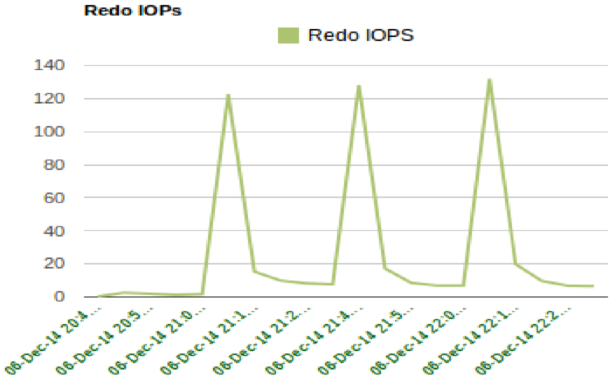


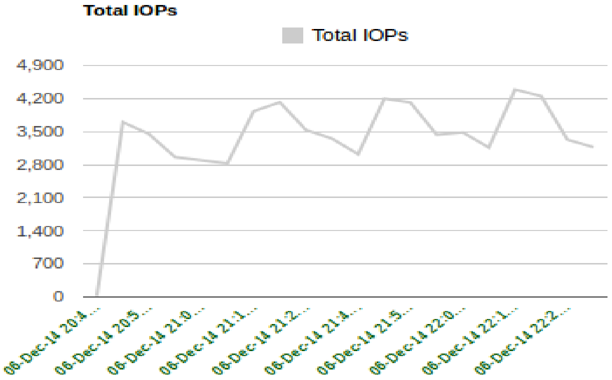


All Database IOPS KVM (16 Mb buffer cache)

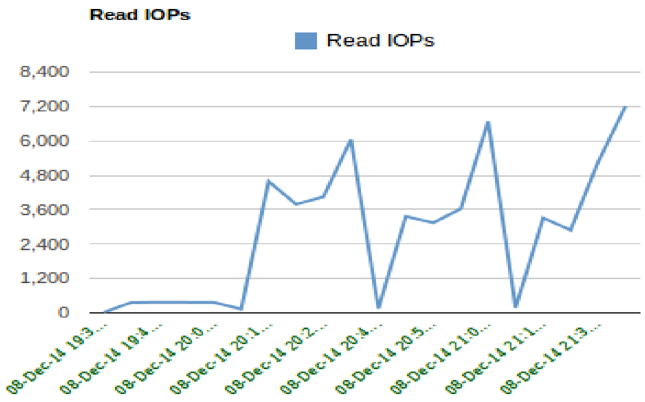


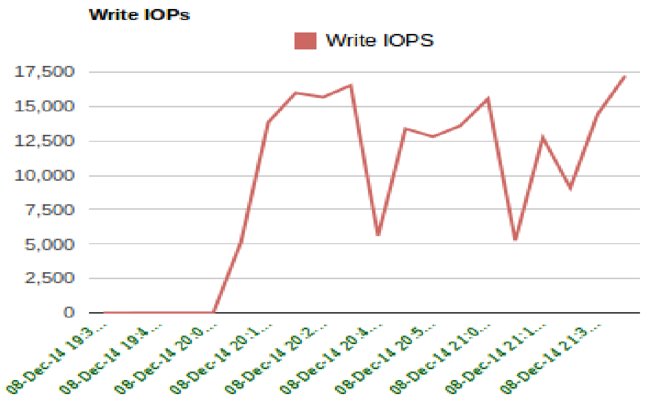


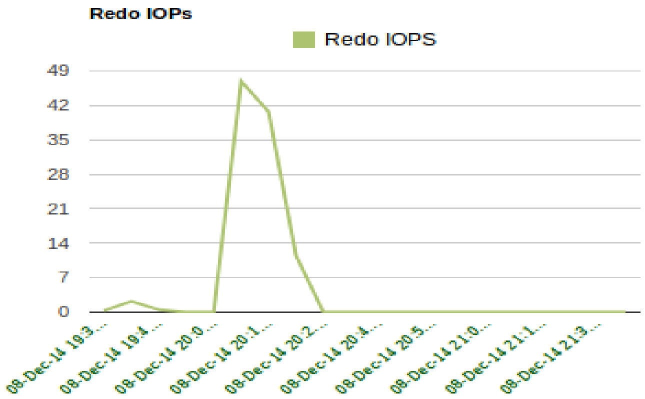


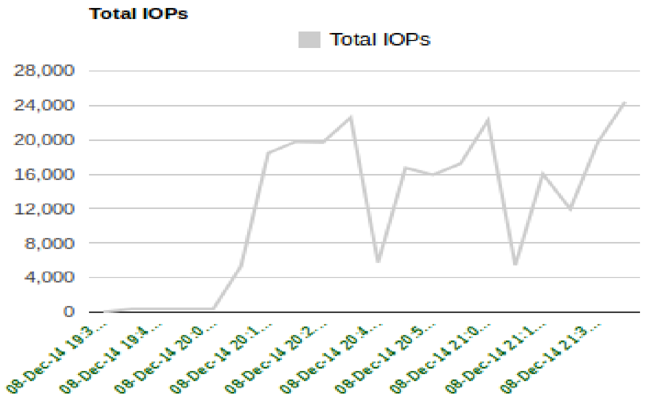


All Database IOPS KVM (11 Gb buffer cache)

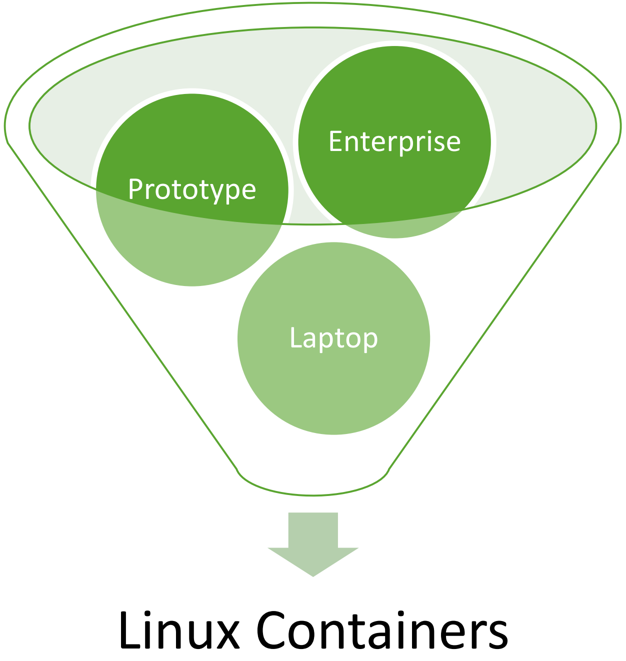








The Linux Container Convergence Phenomenon



- Linux Containers, and the necessary soft infrastructure of DNS, DHCP, and OpenvSwitch can run on your desktop or laptop in exactly the same way they will run in the enterprise. Also, since desktop is a performance-limited environment, Linux Containers are ideal for max at-home performance.
- Linux Containers are ideal for enterprise, because they fully leverage the power/expense of high-speed computing equipment.

Seven Advantages of Linux Containers

- Massive reduction in deployment (creation) time
- Massive reduction in bootup time (boot starts at init)
- Massive virtualization performance improvement (bare metal performance)
- Massive virtualization manageability improvement (files can be edited at host level)
- Consolidation of Enterprise to “Standard” (non-hypervisor) generic Linux skillsets
- Linux Containers deliver all the efficient resource utilization
- No hypervisor, no virtualized hardwar, bare-metal performance

Container Manageability

Ever had a virtual machine image file “go bad”? VM would not start. If you were smart or lucky, maybe you had a backup of the image file. If not, then the next resort is to try and fix the problem. This usually involves tools like “guestfish” or “virsh-rescue” or starting the VM up with a rescue ISO, to fix the misconfiguration or the corruption.

With Linux Containers, this will NEVER happen, because the Container filesystem resides on the same filesystem as the host.

Cgroups have been around in Oracle for awhile…

How I Used CGroups to Manage System Resources In Oracle Linux 6

by Ginny Henningsen (and contributing author Lenz Grimmer), January 2012

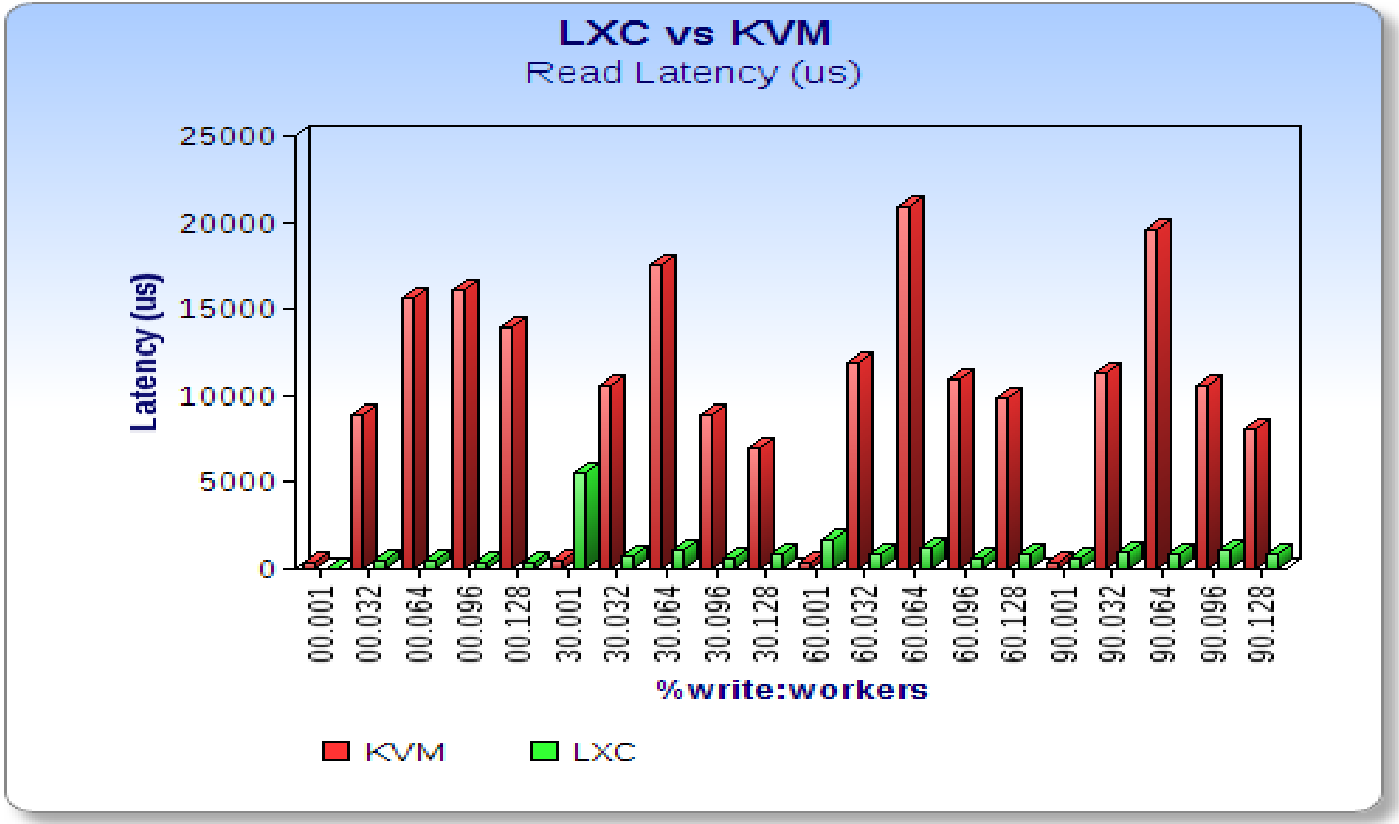
How I used kernel resource controllers in Oracle Linux 6 to manage the allocation of system resources.

http://www.oracle.com/technetwork/articles/servers-storage-admin/resource-controllers-linux-1506602.html

Linux containers takes makes strategies like cgroups part of a complete resource utilization structure called containers.

However, If you are running and OEL6.5 box, for example, and you install Oracle products in OEL 6.5 containers running On that box, is that supported? Why would it not be supported? The entire deployment is OEL 6.5, and Linux Containers are in some ways an extension of the cgroups approach to system resource management, formalized as Linux Containers.

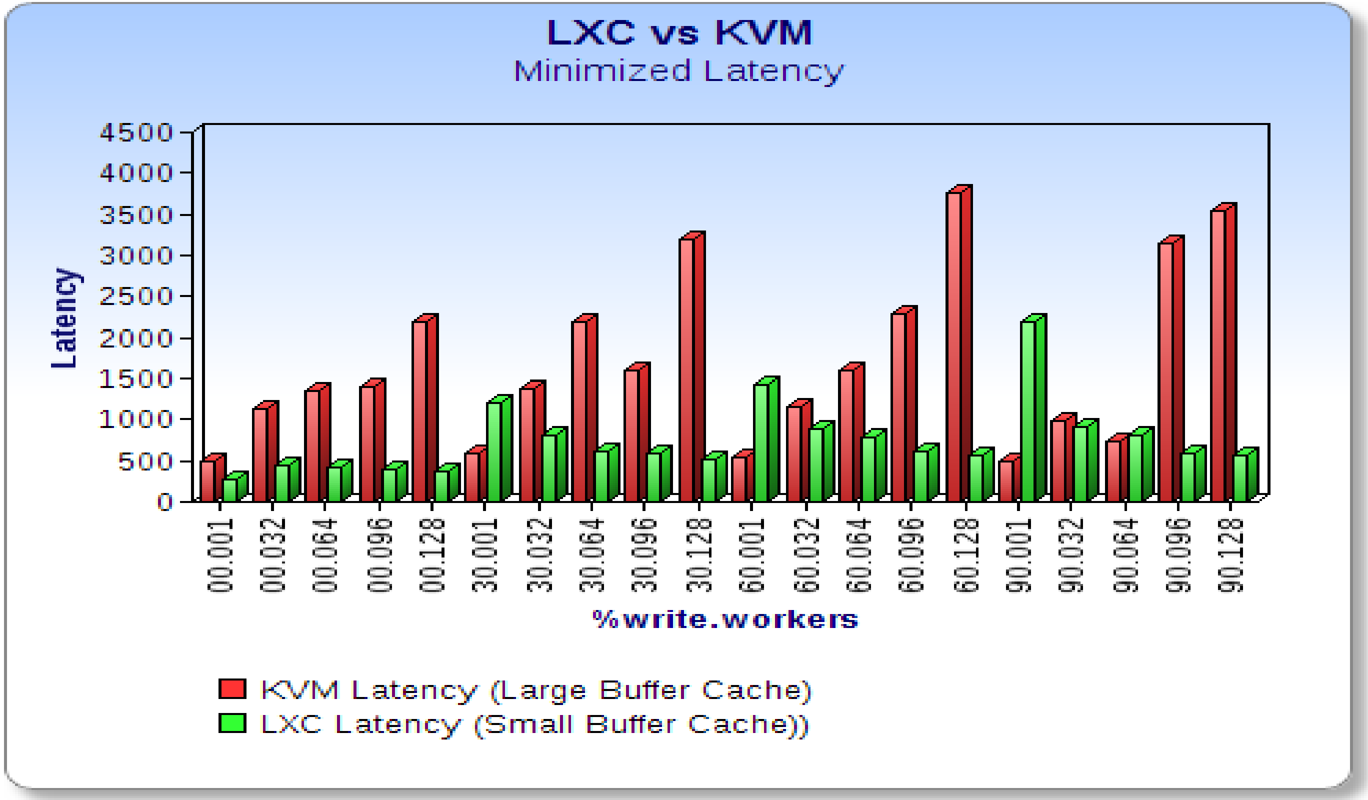
Read Latency (16 Mb Buffer Cache)



When buffer cache size is mimimized, it is observed that on a low-latency AFA such as the Violin 6616 SLC that the Linux Containerized Oracle database continues to have sub-millisecond latency even though physical read percentage is relatively high.

The KVM Oracle database on the other hand, has latency In the very high ms range because the small buffer size reveals poor physical IO latencyof KVM relative to LXC.

Read Latency (KVM 11Gb Buffer Cache), LXC (16 Mb Buffer Cache)



KVM

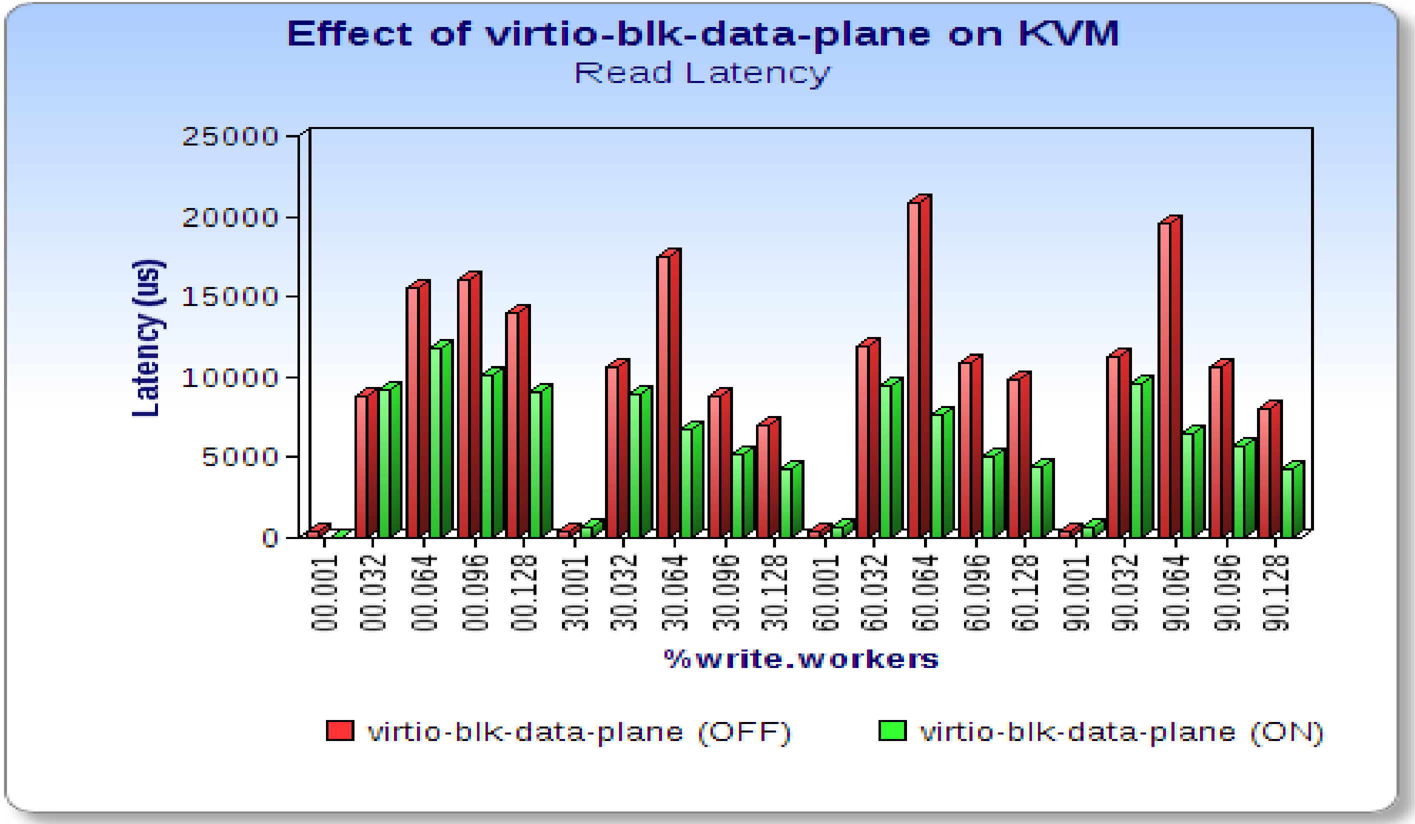
- Virtio-blk-data-plane (ON)
- Hugepages (ON)
- Buffer Cache (11 Gb)

By increasing the size of the Oracle buffer cache to 11 Gb finally better latency is Achieved (4 millisecond max) but this comes at a cost of relatively huge memory requirement compared to LXC.

LXC

Buffer Cache (16 Mb) and so LXC delivers much LOWER Latency at MINIMUM buffer cache on Violin 6616 SLC AFA.

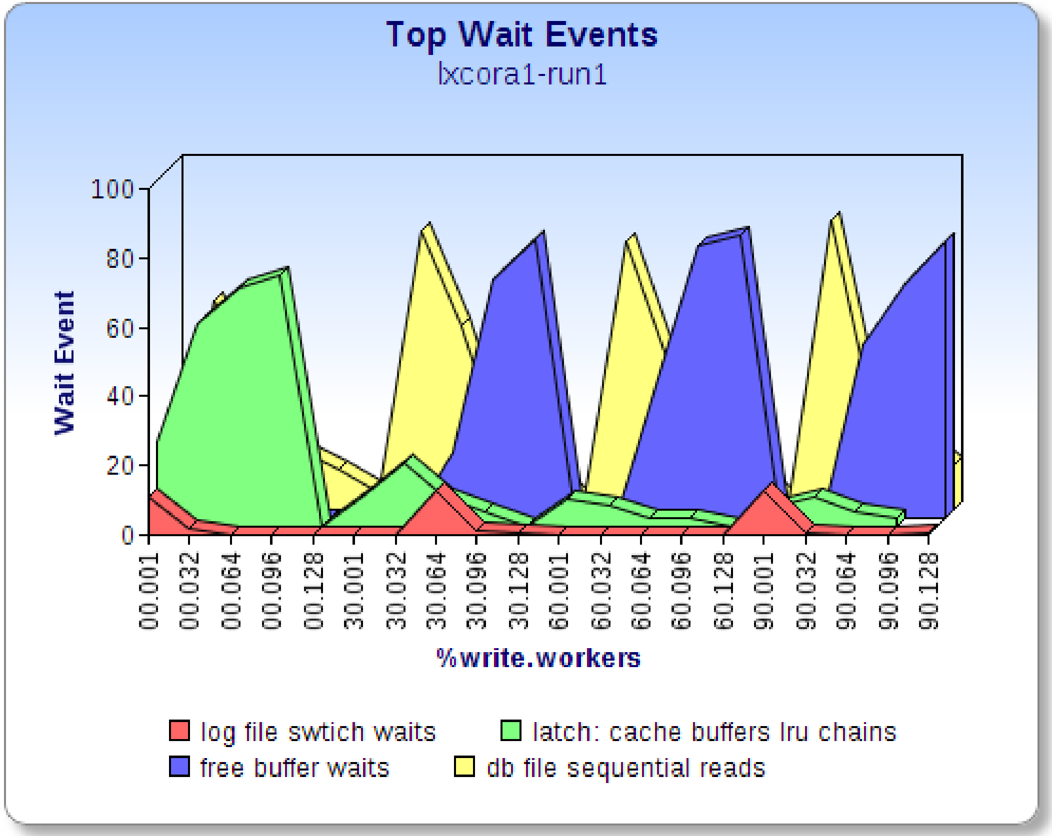
Read Latency KVM (Effect of virtio-blk-data-plane on read latency)

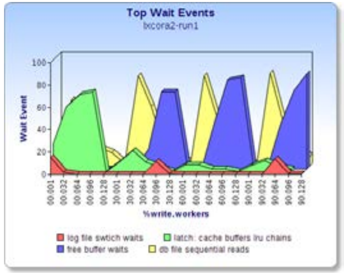


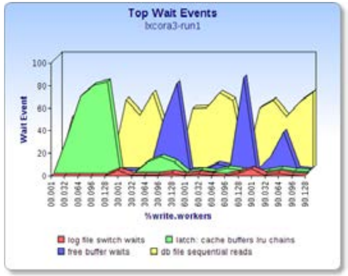
Virtio-blk-data-plane is A storage mechanism for KVM which bypasses QEMU for accessing data.

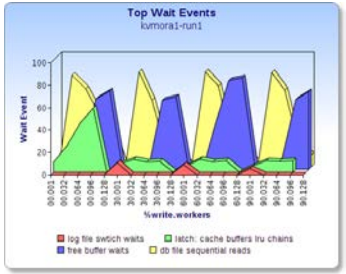
Unofficially, my testing now and in recent past Indicates vbdp is using memory heavily to get better latency to main storage results. Thus virtio-blk-data-plane seems to require relatively large memory provisioning.

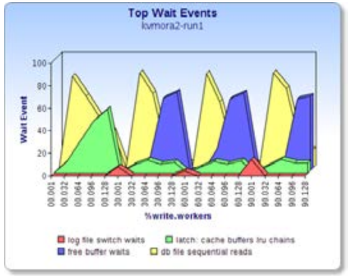
Top Waits LXC vs KVM (density tests)

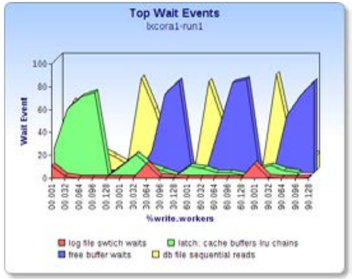


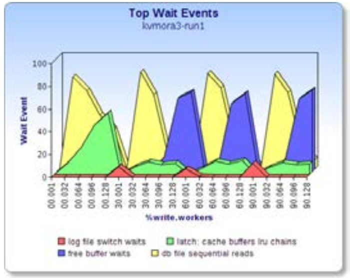












Building Oracle on LXC in OEL UEK[2,3] Kernels (References)

- Wim Coaekerts Blog (October 9, 2011)
  Getting LXC, OpenvSwitch, and BTRFS from Beta Repo.
  Oracle VM templates can be converted to containerization.
- Converting Oracle VM templates to Linux Containers
  Oracle® Linux Administrator's Solutions Guide for Release 6
    Chapter 9 Linux Containers
    Chapter 10 Docker
- However, best to use UEK3 Kernels.
- The 3.12 kernel provides unified Container support This means
- That on 3.12, LXC, OpenVZ, Parallels, .

References

IBM Research Report
An Updated Performance Comparison of Virtual Machines and Linux Containers
Wes Felter, Alexandre Ferreira, Ram Rajamony, Juan Rubio
IBM Research Division
RC25482 (AUS1407-001) **July 21, 2014**

http://domino.research.ibm.com/library/cyberdig.nsf/papers/0929052195DD819C85257D2300681E7B/$File/rc25482.pdf

This paper compares LXC Linux Containers and KVM Guest but does so from the OS point of view (but does include a section on comparison of MySQL DB running in LXC Container vs. MySQL running in KVM guest).

Contacts & Web Presence

- gstanden@vmem.com
- gilstanden@hotmail.com
- http://sites.google.com/site/nandydandyoracle
- http://www.consultingcommandos.us
- http://m.consultingcommandos.us
- http://www.violin-memory.com
- http://www.facebook.com/oraclehsc
- http://www.facebook.com/gilbert_standen
- Twitter: @yoyo_gbiker
- **THANK YOU. QUESTIONS?**