Blog Export: c0t0d0s0.org, http://www.c0t0d0s0.org/

Friday, March 9. 2012

How to show off live migration with a SPARC system?

Yesterday i had the opportunity to show Oracle VM for SPARC in front of customers in action. Not a single slide was used ... Everything was live . The following entry shows what i essentially did in this demo. Perhaps long time users of LDOMS or Oracle VM for SPARC (as they are called today) have already seen for this, however that wasn't the planned audience of this walkthrough. In this example i've configured the control domain, one guest domain, installed it with Solaris 11 and migrated it live (without service interruption) from one system to another)Okay, i started with two unconfigured (okay, to be exact ... deconfigured systems) system of the type SPARC T3-4. So i had plenty resources to play. The first system was node1 listening on 10.128.0.72, the second system was node2 listening on 10.128.0.73.

Just to be sure, i checked the configuration.

```
node1:/# ldm ls
NAME          STATE     FLAGS   CONS    VCPU  MEMORY   UTIL  UPTIME
primary       active    -n-c--  UART    512   261632M  0.4%  5m
```

There was just a single logical domain with all resources (512 virtual CPUs and 256 GB memory were assigned to it. The situation on the second node was the same. No wonder. Same HW config, same SW config.

```
node2:/# ldm ls
NAME          STATE     FLAGS   CONS    VCPU  MEMORY   UTIL  UPTIME
primary       active    -n-c--  UART    512   261632M  0.3%  8m
```

Ensure that you have enabled the vntsd daemon on both systems

```
node1:/# svcadm enable vntsd
node2:/# svcadm enable vntsd
```

Okay, basics were the same, now i had to start the basic config. Important is those single large domains will act as so called control domains, however they will significantly smaller for that task. The already running Solaris 10 was kept unharmed and got the OS of the control domain.

First step was to configure the virtual console server:

```
node1:/# ldm add-vcc port-range=5000-5100 primary-vcc0 primary
```

With this command you configure a console server listening on ports 5000 to 5100 named primary-vcc0 in the domain primary. Okay, the next step was to configure the so called virtual disk server. As long as you don't configure any hardware directly into a domain like a networking card for for iSCSI or a HBA for storage access, the virtual disk server is the daemon that provides the service of storage devices to all guest domains.

```
node1:/# ldm add-vds primary-vds0 primary
```

With this command we have configured a virtual disk server called primary-vds0 in the domain vds. The next step is the configuration of the networking. For this task we configure a virtual switch.

```
node1:/# ldm add-vsw net-dev=igb0 primary-vsw0 primary
```

The virtual switch called primary-vsw0 is running in the domain primary and it's connecting into the real world via the device igb0. When you want to check all the services you have just configured, you can do this with a single command.

```
node1:/# ldm list-services primary
VCC
    NAME          LDOM          PORT-RANGE
    primary-vcc0    primary       5000-5100

VSW
    NAME          LDOM          MAC             NET-DEV   ID   DEVICE     LINKPROP   DEFAULT-VLAN-ID PVID VID
          MTU   MODE   INTER-VNET-LINK
    primary-vsw0    primary       00:14:4f:fa:df:5c igb0     0    switch@0            1           1                   1500
on

VDS
    NAME          LDOM          VOLUME        OPTIONS        MPGROUP       DEVICE
    primary-vds0    primary
```

```
node1:/#
```

At the moment this primary domain is using all the resources. In order to be able to configure some guests, we have to free some room. So at first we reduce the number of assigned crypto units. I just want to give them one.

```
node1:/# ldm set-mau 1 primary
```

In the next step we assign 8 processor to the domain.

```
node1:/# ldm set-vcpu 8 primary
```

Okay, let's check the current configuration.

```
node1:/# ldm ls
NAME             STATE      FLAGS   CONS    VCPU  MEMORY   UTIL  UPTIME
primary          active     -n-cv-  UART    8     261632M  16%   20m
```

Okay, just 8 virtual cpus, hpowever the domain still occupied all the memory in the system. We have to reduce that. Technically it's possible to do this on the running system but getting a running logical domain from 256 GB to 8 GB is quite some work, so most often it is just faster to put the domain in the deferred configuration mode, do the configuration and reboot the system as in this moment nothing runs on the system. When doing deferred reconfiguration, the config change is accepted but it will be executed with the next reboot:

```
node1:/# ldm start-reconf primary
Initiating a delayed reconfiguration operation on the primary domain.
All configuration changes for other domains are disabled until the primary
domain reboots, at which time the new configuration for the primary domain
will also take effect.
```

Now we set the memory of the domain primary to 8 GB

```
node1:/# ldm set-memory 8G primary
------------------------------------------------------------------------------
Notice: The primary domain is in the process of a delayed reconfiguration.
Any changes made to the primary domain will only take effect after it reboots.
------------------------------------------------------------------------------
```

Saving the config to the ILOM, and rebooting the system.

```
node1:/# ldm add-config initial
node1:/# shutdown -y -g0 -i6
```

Okay, while the first system is rebooting, we just repeat the same configuration steps on the second system:

```
node2:/# ldm add-vcc port-range=5000-5100 primary-vcc0 primary
node2:/# ldm add-vds primary-vds0 primary
node2:/# ldm add-vsw net-dev=igb0 primary-vsw0 primary
node2:/# ldm list-services primary
VCC
    NAME             LDOM             PORT-RANGE
    primary-vcc0     primary          5000-5100

VSW
    NAME             LDOM             MAC               NET-DEV   ID   DEVICE     LINKPROP   DEFAULT-VLAN-ID PVID VID
             MTU   MODE   INTER-VNET-LINK
    primary-vsw0     primary          00:14:4f:f9:b9:9c igb0      0    switch@0              1               1                        1500
on

VDS
    NAME             LDOM             VOLUME         OPTIONS          MPGROUP        DEVICE
    primary-vds0     primary

node2:/# ldm set-mau 1 primary
node2:/# ldm set-vcpu 8 primary
Removal of cpu 118 failed, error: Removal of cpu 48 failed, error: Removal of cpu 16 failed, error: Domain primary
permitted removal of only 501 VCPUs
Retry removed 3 more VCPUs from domain primary
node2:/# ldm ls
```

```
NAME            STATE      FLAGS   CONS    VCPU  MEMORY   UTIL  UPTIME
primary         active     -n-cv-  UART    8     261632M  16%   15m
node2:/# ldm start-reconf primary
Initiating a delayed reconfiguration operation on the primary domain.
All configuration changes for other domains are disabled until the primary
domain reboots, at which time the new configuration for the primary domain
will also take effect.
node2:/# ldm set-memory 8G primary
------------------------------------------------------------------------------
Notice: The primary domain is in the process of a delayed reconfiguration.
Any changes made to the primary domain will only take effect after it reboots.
------------------------------------------------------------------------------
node2:/# ldm add-config initial
node2:/# shutdown -y -g0 -i6
```

We now check the config on both systems. On the first system:

```
node1:/# ldm ls
NAME            STATE      FLAGS   CONS    VCPU  MEMORY   UTIL  UPTIME
primary         active     -n-cv-  UART    8     8G       0.6%  8m
```

On the second system.

```
node2:/# ldm ls
NAME            STATE      FLAGS   CONS    VCPU  MEMORY   UTIL  UPTIME
primary         active     -n-cv-  UART    8     8G       8.1%  7m
```

Okay ... i have to explain a little bit ... 10.10.1.37 is a S7000 filer i've used for central storage. In the directory /ldoms/isos i've put a iso of the solaris 11 11/11 text install image.

```
node1:/# mount 10.10.1.37:/export/ldoms /ldoms
```

As i want to install the ldom i will create later on, i add this iso to the virtual disk server as a device:

```
node1:/# ldm add-vdsdev /ldoms/isos/sol-11-1111-text-sparc.iso sol11iso@primary-vds0
```

Okay, i want to demo live migration in this walkthrough, so i need some shared storage. It's obvious why i need shared storage: It' makes no sense to migrate a logical domain to a system that hasn't access to the same disk devices. So i configured my S7000 filer to offer some LUNs via iSCSI. However i have to configure the primary domain in order to actually use this LUNs. However this is pretty easy. At first we tell the iSCSI initator of Solaris 10 that there are disk to find behind 10.10.1.37

```
node1:/# iscsiadm add discovery-address 10.10.1.37
```

Now we tell Solaris to discover the LUNs behing this IP.

```
node1:/# iscsiadm modify discovery -t enable
```

And now we poplulate the /dev tree with the nescessary nodes.

```
node1:/# devfsadm -i iscsi
```

Okay, repeat on the second system.

```
node2:/# mount 10.10.1.37:/export/ldoms /ldoms
node2:/# ldm add-vdsdev /ldoms/isos/sol-11-1111-text-sparc.iso sol11iso@primary-vds0
node2:/# iscsiadm add discovery-address 10.10.1.37
node2:/# iscsiadm modify discovery -t enable
node2:/# devfsadm -i iscsi
```

Okay, let's have a look what the system has found. From the configuration in the filer i knew that there must be something like 600144F0C56DC0FB00004F586FD60004 in the disk id. As the disk is unlabeled at that time the format command will offer to you do this labeling. Do it ... you need a labeled disk.

```
node2:/# format
Searching for disks...done
```

```
c0t600144F0C56DC0FB00004F586FD60004d0: configured with capacity of 9.94GB


AVAILABLE DISK SELECTIONS:
       0. c0t5000C5003A02EE93d0
          /scsi_vhci/disk@g5000c5003a02ee93
[...]
      13. c0t600144F0C56DC0FB00004F586FD60004d0
          /scsi_vhci/ssd@g600144f0c56dc0fb00004f586fd60004
[...]
Specify disk (enter its number):  13
selecting c0t600144F0C56DC0FB00004F586FD60004d0
[disk formatted]

Disk not labeled.  Label it now? Disk not labeled.  Label it now? y


FORMAT MENU:
        disk       - select a disk
[...]
        !     - execute , then return
        quit
format> quit
```

Okay, now check the availability of disk disk on the other server.

```
node1:/# format
Searching for disks...done

c0t600144F0C56DC0FB00004F586FD60004d0: configured with capacity of 9.94GB


AVAILABLE DISK SELECTIONS:
       0. c0t5000C5003A02FD67d0
          /scsi_vhci/disk@g5000c5003a02fd67
[..]
      13. c0t600144F0C56DC0FB00004F586FD60004d0
          /scsi_vhci/ssd@g600144f0c56dc0fb00004f586fd60004
[..]
Specify disk (enter its number):
Specify disk (enter its number):
```

Check ... is there. The next step is the last one in this tour we have to execute on both systems. With this command we assign the disk /dev/dsk/c0t600144F0C56DC0FB00004F586FD60004d0s2 on both nodes as lmtest001iscsibootdisk to the virtual disk server called primary-vds0

```
node1:/# ldm add-vdsdev /dev/dsk/c0t600144F0C56DC0FB00004F586FD60004d0s2
lmtest001iscsibootdisk@primary-vds0
node2:/# ldm add-vdsdev /dev/dsk/c0t600144F0C56DC0FB00004F586FD60004d0s2
lmtest001iscsibootdisk@primary-vds0
```

Okay, now we configure our first guest domain.

At first we just create the domain.

```
node1:/# ldm add-domain lmtest001
```

Now we add 8 virtual CPUs to the domain.

```
node1:/# ldm add-vcpu 8 lmtest001
```

Of course a domain needs memory, so i give it 16 GB.

```
node1:/# ldm add-mem 16G lmtest001
```

Now i'm creating a networking interface for the domain lmtest001 connected to the virtual switch primary-vsw0 and naming it vnet1.

```
node1:/# ldm add-vnet vnet1 primary-vsw0 lmtest001
```

Okay, as my iscsi disk is totally empty, i have to provide an installation image (i could do this via jumpstart or AI, however that would be out of scope of this short article). So i assign the virtual disk sol11iso on the virtual disk server primary-vds0 (remember, we configured them earlier) to lmtest001. To the domain it's named vdisk_iso.

```
node1:/# ldm add-vdisk vdisk_iso sol11iso@primary-vds0 lmtest001
```

Now i have to assign the iscsi boot disk to the domain. The command is quite similar.

```
node1:/# ldm add-vdisk bootdisk lmtest001iscsibootdisk@primary-vds0 lmtest001
```

The next stop is to declare the bootdevice and to tell the system to boot automatically from it.

```
node1:/# ldm set-var auto-boot\?=true lmtest001
node1:/# ldm set-var boot-device=bootdisk lmtest001
```

However, the domain is still inactive and no resources have been binded to the domain

```
node1:/# ldm ls
NAME          STATE     FLAGS   CONS    VCPU  MEMORY   UTIL  UPTIME
primary       active    -n-cv-  UART    8     8G       13%   35m
lmtest001     inactive  ------          8     16G
```

So we bind the resources with a single command:

```
node1:/# ldm bind-domain lmtest001
```

When you look up the status again, you see a state transition. The domain isn't "inactive" any longer, it's now bound.

```
node1:/# ldm ls
NAME          STATE     FLAGS   CONS    VCPU  MEMORY   UTIL  UPTIME
primary       active    -n-cv-  UART    8     8G       0.5%  36m
lmtest001     bound     ------  5000    8     16G
```

Now it's time to start up the domain.

```
node1:/# ldm start-domain lmtest001
LDom lmtest001 started
```

When you look back into the last output of ldm ls you will see the 5000 in the column "CONS" (short for console) for lmtest001. This 5000 is now important to get access to the console of the lmtest001 domain.

```
node1:/# telnet localhost 5000
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.

Connecting to console "lmtest001" in group "lmtest001" ....
Press ~? for control options ..

{0} ok
```

As you see, there is a boot prompt like with a native SPARC machine. As there is no operating system on the device we've called bootdisk, the system doesn't come up but stays in that prompt.

```
{0} ok devalias
bootdisk                 /virtual-devices@100/channel-devices@200/disk@1
vdisk_iso                /virtual-devices@100/channel-devices@200/disk@0
vnet1                    /virtual-devices@100/channel-devices@200/network@0
net                      /virtual-devices@100/channel-devices@200/network@0
disk                     /virtual-devices@100/channel-devices@200/disk@0
virtual-console          /virtual-devices/console@1
name                     aliases
```

Now let's boot from the iso image:

```
{0} ok boot vdisk_iso:f
```

Okay, at first the system comes up from the ISO as you see

```
SPARC T3-4, No Keyboard
Copyright (c) 1998, 2011, Oracle and/or its affiliates. All rights reserved.
OpenBoot 4.33.0.b, 16384 MB memory available, Serial #83470411.
Ethernet address 0:14:4f:f9:a8:4b, Host ID: 84f9a84b.

Boot device: /virtual-devices@100/channel-devices@200/disk@0:f  File and args:
SunOS Release 5.11 Version 11.0 64-bit
```

```
Copyright (c) 1983, 2011, Oracle and/or its affiliates. All rights reserved.
```

Okay, this will now take a while. I won't write about it. It's a standard Solaris 11 install. You know the drill.

After the reboot initiated in the installation procedure, the systems comes up with an installed OS. As you will recognize by the string of the boot device, that you now have booted from the iscsi boot disk.

```
SPARC T3-4, No Keyboard
Copyright (c) 1998, 2011, Oracle and/or its affiliates. All rights reserved.
OpenBoot 4.33.0.b, 16384 MB memory available, Serial #83470411.
Ethernet address 0:14:4f:f9:a8:4b, Host ID: 84f9a84b.



Boot device: /virtual-devices@100/channel-devices@200/disk@1:a  File and args:
SunOS Release 5.11 Version 11.0 64-bit
Copyright (c) 1983, 2011, Oracle and/or its affiliates. All rights reserved.

Loading smf(5) service descriptions: 198/198

Configuring devices.
Loading smf(5) service descriptions: 1/1
Hostname: solaris

solaris console login:
```

Okay, let's play a little bit with the domain. Login into the shell of the system. When you execute an prtdiag, you will see 16 GB of memory and 8 virtual CPUs.

```
jmoekamp@solaris:~$ prtdiag
System Configuration:  Oracle Corporation  sun4v SPARC T3-4
Memory size: 16384 Megabytes

================================ Virtual CPUs ================================


CPU ID Frequency Implementation        Status
------ --------- ---------------------- -------
0      1649 MHz  SPARC-T3             on-line
1      1649 MHz  SPARC-T3             on-line
2      1649 MHz  SPARC-T3             on-line
3      1649 MHz  SPARC-T3             on-line
4      1649 MHz  SPARC-T3             on-line
5      1649 MHz  SPARC-T3             on-line
6      1649 MHz  SPARC-T3             on-line
7      1649 MHz  SPARC-T3             on-line

================================ IO Devices ================================
Slot +            Bus   Name +                          Model
Status            Type  Path
----------------------------------------------------------------------------
jmoekamp@solaris:~$
```

Okay, let's assume we've changed our mind and want a domain with 8 additional virtual CPUs. You can do this while the domain is running:

```
node1:/# ldm add-vcpu 8 lmtest001
node1:/#
```

When you do another prtdiag in the still running domain, you will see 16 virtual CPUs.

```
jmoekamp@solaris:~$ prtdiag
[...]
```

```
================================ Virtual CPUs ================================


CPU ID Frequency Implementation         Status
------ --------- ---------------------- -------
0      1649 MHz  SPARC-T3               on-line
1      1649 MHz  SPARC-T3               on-line
2      1649 MHz  SPARC-T3               on-line
3      1649 MHz  SPARC-T3               on-line
4      1649 MHz  SPARC-T3               on-line
5      1649 MHz  SPARC-T3               on-line
6      1649 MHz  SPARC-T3               on-line
7      1649 MHz  SPARC-T3               on-line
8      1649 MHz  SPARC-T3               on-line
9      1649 MHz  SPARC-T3               on-line
10     1649 MHz  SPARC-T3               on-line
11     1649 MHz  SPARC-T3               on-line
12     1649 MHz  SPARC-T3               on-line
13     1649 MHz  SPARC-T3               on-line
14     1649 MHz  SPARC-T3               on-line
15     1649 MHz  SPARC-T3               on-line
[..]
```

Okay, 8 additional 8G may be a nice idea. So let's add them to the running domain.

```
node1:/# ldm add-mem 8G lmtest001
node1:/#
```

Do another prtdiag, an we see 24 gigs of memory.

```
jmoekamp@solaris:~$ prtdiag
System Configuration:  Oracle Corporation  sun4v SPARC T3-4
Memory size: 24576 Megabytes
[...]
```

However, we aren't really in decision mood today and think that our first config was nice and revert to the old values. So we remove 8 gigs of memory again from the domain lmtest001

```
node1:/# ldm rm-mem 8G lmtest001
node1:/#
```

And again our domain has just 16 GB.

```
jmoekamp@solaris:~$ prtdiag
System Configuration:  Oracle Corporation  sun4v SPARC T3-4
Memory size: 16384 Megabytes
[...]
```

Now we have just to remove the 8 additional vcpus.

```
node1:/# ldm rm-vcpu 8 lmtest001
node1:/#
```

Okay, a last time we will execute prtdiag.

```
jmoekamp@solaris:~$ prtdiag
System Configuration:  Oracle Corporation  sun4v SPARC T3-4
Memory size: 16384 Megabytes

================================ Virtual CPUs ================================
CPU ID Frequency Implementation        Status
------ --------- --------------------- -------
0      1649 MHz  SPARC-T3           on-line
1      1649 MHz  SPARC-T3           on-line
2      1649 MHz  SPARC-T3           on-line
3      1649 MHz  SPARC-T3           on-line
4      1649 MHz  SPARC-T3           on-line
5      1649 MHz  SPARC-T3           on-line
6      1649 MHz  SPARC-T3           on-line
7      1649 MHz  SPARC-T3           on-line

================================ IO Devices ================================
Slot +          Bus   Name +                        Model
Status           Type  Path
------------------------------------------------------------------------
```

Again, back to 8 virtual CPUs in the domain.

Okay, but now back to our demonstration of live migation. I would like to demonstrate the live migration with some network traffic, so i need an ip address. So i configure one one the vnet0 interface card i've created when i was configuring the domain earlier. Log into the domain as root or assume a role that allows you to configure networking.

```
root@solaris:/home/jmoekamp# dladm show-phys
LINK              MEDIA                STATE      SPEED  DUPLEX    DEVICE
net0              Ethernet             unknown    0      unknown   vnet0
root@solaris:/home/jmoekamp# ipadm create-ip net0
root@solaris:/home/jmoekamp# ipadm create-addr -T static -a 10.10.0.74/16 net0ipv4
root@solaris:/home/jmoekamp# route -p add default 10.10.0.1
add net default: gateway 10.10.0.1
add persistent net default: gateway 10.10.0.1
root@solaris:/home/jmoekamp#
```

Just a short test from my local workstation. Just a short remark ... the times are that bad, because my servers were in Scotland and i was in the Düsseldorf FTL lounge connected via VPN over an UTMS line ...

```
PS C:\Windows\System32\WindowsPowerShell\v1.0> ping -t 10.10.0.74

Ping wird ausgeführt für 10.10.0.74 mit 32 Bytes Daten:
Antwort von 10.10.0.74: Bytes=32 Zeit=183ms TTL=253
Antwort von 10.10.0.74: Bytes=32 Zeit=94ms TTL=253
Antwort von 10.10.0.74: Bytes=32 Zeit=142ms TTL=253
```

Okay, let's kick of the live migation. With this command i order the logical domain manager to migrate the domain lmtest001 to the server running a control domain on 10.128.0.73. The password was in my case the root password of that control domain.

```
node1:/# ldm migrate lmtest001 10.128.0.73
Target Password:
node1:/#
```

This will take a while, however you will just get back your prompt in a very unspectacular way. You will get back the prompt as soon as the migration has completeted

What has in the meantime happened?

```
PS C:\Windows\System32\WindowsPowerShell\v1.0> ping -t 10.10.0.74

Ping wird ausgeführt für 10.10.0.74 mit 32 Bytes Daten:
Antwort von 10.10.0.74: Bytes=32 Zeit=497ms TTL=253
Antwort von 10.10.0.74: Bytes=32 Zeit=1371ms TTL=253
Antwort von 10.10.0.74: Bytes=32 Zeit=323ms TTL=253
Antwort von 10.10.0.74: Bytes=32 Zeit=118ms TTL=253
Antwort von 10.10.0.74: Bytes=32 Zeit=80ms TTL=253
Zeitüberschreitung der Anforderung.
Antwort von 10.10.0.74: Bytes=32 Zeit=67ms TTL=253
Antwort von 10.10.0.74: Bytes=32 Zeit=209ms TTL=253
Antwort von 10.10.0.74: Bytes=32 Zeit=86ms TTL=253
Antwort von 10.10.0.74: Bytes=32 Zeit=83ms TTL=253
Antwort von 10.10.0.74: Bytes=32 Zeit=70ms TTL=253
Antwort von 10.10.0.74: Bytes=32 Zeit=105ms TTL=253

Ping-Statistik für 10.10.0.74:
    Pakete: Gesendet = 286, Empfangen = 285, Verloren = 1
    (0% Verlust),
Ca. Zeitangaben in Millisek.:
    Minimum = 66ms, Maximum = 1473ms, Mittelwert = 165ms
STRG-C
PS C:\Windows\System32\WindowsPowerShell\v1.0>
```

Well a ping has been lost. However what's more interesting, there isn't a domain lmtest001 on my server.

```
node1:/# ldm ls
NAME             STATE      FLAGS   CONS    VCPU  MEMORY   UTIL  UPTIME
primary          active     -n-cv-  UART    8     8G       0.5%  2h 6m
node1:/#
```

Because it's on the other one.

```
node2:/ldoms/isos# ldm ls
NAME             STATE      FLAGS   CONS    VCPU  MEMORY   UTIL  UPTIME
primary          active     -n-cv-  UART    8     8G       0.4%  2h 5m
lmtest001        active     -n----  5000    8     16G      0.1%  1h 8m
```

Of course i could migrate back to my old system.

```
node2:/ldoms/isos# ldm migrate lmtest001 10.128.0.72
Target Password:
```

It has disappered from this server

```
node2:/ldoms/isos# ldm ls
NAME             STATE      FLAGS   CONS    VCPU  MEMORY   UTIL  UPTIME
primary          active     -n-cv-  UART    8     8G       0.3%  2h 13m
```

And is now back on the original one.

```
node1:/# ldm ls
NAME             STATE      FLAGS   CONS    VCPU  MEMORY   UTIL  UPTIME
primary          active     -n-cv-  UART    8     8G       0.5%  2h 15m
lmtest001        active     -n----  5000    8     16G      0.2%  1h 18m
node1:/#
```

Neat.

Do you want to learn more?
docs.oracle.com: Oracle VM for SPARC Documentation

Posted by Joerg Moellenkamp in General at 17:03

Very nice write-up. No more downtime for planned hardware maintenance Amazing feature.
Anonymous on Mar 9 2012, 17:32

Hi!

How would you compare two technologies: LDOMs and solaris zones? Which one is going to be developed and used in future? Could you recommend any nice white paper to read about the pros and cons of both those technologies in comparison? Don't you feel that oracle would end development of LDOMs in order to make zoning the only virtualisation technology?
Anonymous on Mar 9 2012, 18:30

LDOM's are hardware partitioning where zones is a virtualized Solaris instance.
http://www.unix.com/solaris/55530-ldom-vrs-container.html

Oracle likes the hardware partitioning it makes purchasers of their high end servers able to utilise them quickly without requiring more/Cheaper hardware. Eg: Oracle DB licensed per Proc.

Hope this helps.
Anonymous on Mar 9 2012, 20:33

The difference of hardware and software virtualisation is clear. What is unclear for me is the following development of both technologies and real benefits of LDOMs in comparison to zones from practical POV. It seems to me that LDOMs are to be killed in future as they provide no real advantages. Am I right or not?

BTW:
"Both Oracle Database 9i R2 and 10g R2 databases have been certified to run in an Oracle Solaris Container. A licensing agreement recognizes capped Oracle Solaris 10 Containers as hard partitions. The ability to license only the CPUs or cores configured in an Oracle Solaris Container provides flexibility, consolidation opportunities, and possible cost savings." From here:
http://developers.sun.com/solaris/docs/oracle_containers.pdf
Anonymous on Mar 11 2012, 07:34

Well, you are wrong. Totally.

1. LDOMs and Zones are mechanisms with different targets.
2. Zones are designed as a lightweight mechanism. LDOMs have a higher level of isolation, but therefore have a higher overhead. Zones are based on a single kernel, each LDOM has it's own kernel. Both situations have advantages and disadvantages. You can have different OS kernels in each LDOM. As you may have noticed, the demo above runs a Solaris 10 control domain with a Solaris 11 guest domain. On Zones you have exactly the same kernel in each zones, which can be problematic if one vendors says "i want this OS patch level", and another says "i want this one" and both are not identical.
3. There is no "XOR" in this question. Most customers use both because in the kind of having LDOMs with a large number of Zones in each of it.
4. The article above should show you one of the read advantage. You can't live migrate with zones, however with LDOMs you could do that. I know that there are vendors that offer some kind of live migration for a kind of zone, however even friends of that architectures admit that those technologies are a mess. And you have to work around with more and more conversion/lookup tables thus increasing the overhead for normal operation.
Anonymous on Mar 11 2012, 08:39

ok. Thanks.
Anonymous on Mar 11 2012, 10:24

Joerg,

thank you for an excellent detailed post. I have a few questions...

so the lesson is, an iSCSI bootdisk coming from a S7000 filer was logically migrated from node1 (a T4 system) to another physical chassis called node2 (a T4 system).

This gives node1 and node2 the ability to "host" or pass back and forth a Solaris OS instance called "lmtest001" that is wholly contained on OBP device "/virtual-devices@100/channel-devices@200/disk@1:a"

Q. What the approximate time in seconds or minutes to accomplish "ldm migrate lmtest001 10.128.0.73"? Is your prompt held for the duration - does it report explicit success or failure of the migration command?

Q. What is the "state" of the Solaris OS instance "lmtest001" during this timeframe? is it running or must it be "halted" with init 0?

I only have 1 T5540 series (need another!), just migrating away from the Sun 6800 (Dynamic Reconfiguration within the same chassis), shared FC storage (Sun 2540) and Veritas (VCS) world.

- svrocket
Anonymous on Mar 12 2012, 04:48

No, the lesson is "How to move a running OS with it's running application from one server to another"!

1. The time for the migration can take some time as the content of the memory of the LDOM is transfered to the other system. LDM will return as soon as migraiton has completed or was ended with an error.

2. lmtest0001 is running during the migration. You can migrate a domain under full production load. Depending on the rate of memory

page changes there may be a short moment, where the domain is freezed at the very end of the migraiton, however as i showed in the example a single ping is "timeouting" here.
Anonymous on Mar 12 2012, 07:03

Jörg, one more question (which I didn't ask on Thursday as were already short on time).

In your example you assign a fixed number of processors to the LDOM. Is it possible to assign "dynamic" resource like you could do while creating resource pools (e.g. "poolcfg -c 'create pset zone_pset (uint pset.min=2; uint pset.max=4)' ") for zones?
Anonymous on Mar 12 2012, 11:20

Yes, you can do that, it's called Dynamic Resource Management.

http://docs.oracle.com/cd/E23120_01/html/821-2854/usingdynamicresourcemanagementpolicies.html#scrolltoc
Anonymous on Mar 12 2012, 11:39

Great. So this would be a facility to "overbook" the hardware?
If so, what gets checked when migrating to another box? I guess at least "vcpu-min" must be matched?

Btw, nice to see that this blog is getting back to life lately.
Anonymous on Mar 12 2012, 11:56

Very nice Joerg. Where can I find the docs/requirements for ldm? Can I do Live Migration with the T1-series T1000 boxes?

And does a live migrated system drag it's NFS mounted file systems with it to the new location?

And can I migrate to unlike T-series machines, like from the T1000 to a T5440 and back? Thanks for your time.
Anonymous on Mar 13 2012, 07:18

Joerg
Please to see some great articles being posted on your site again.....Looks like your site could be my most visited once again.
Hope all is going well
All the best
Stu
Anonymous on Mar 13 2012, 14:16

Good article and questions, here is another one

If you have zones running inside the LDOM guest, can you still Live Migrate the the LDOM with the zones inside it?
Anonymous on Jul 10 2012, 16:33

Of course ... as the Zones are part of the migrated OS, the Zones are migrated with the LDOM.
Anonymous on Jul 12 2012, 07:26