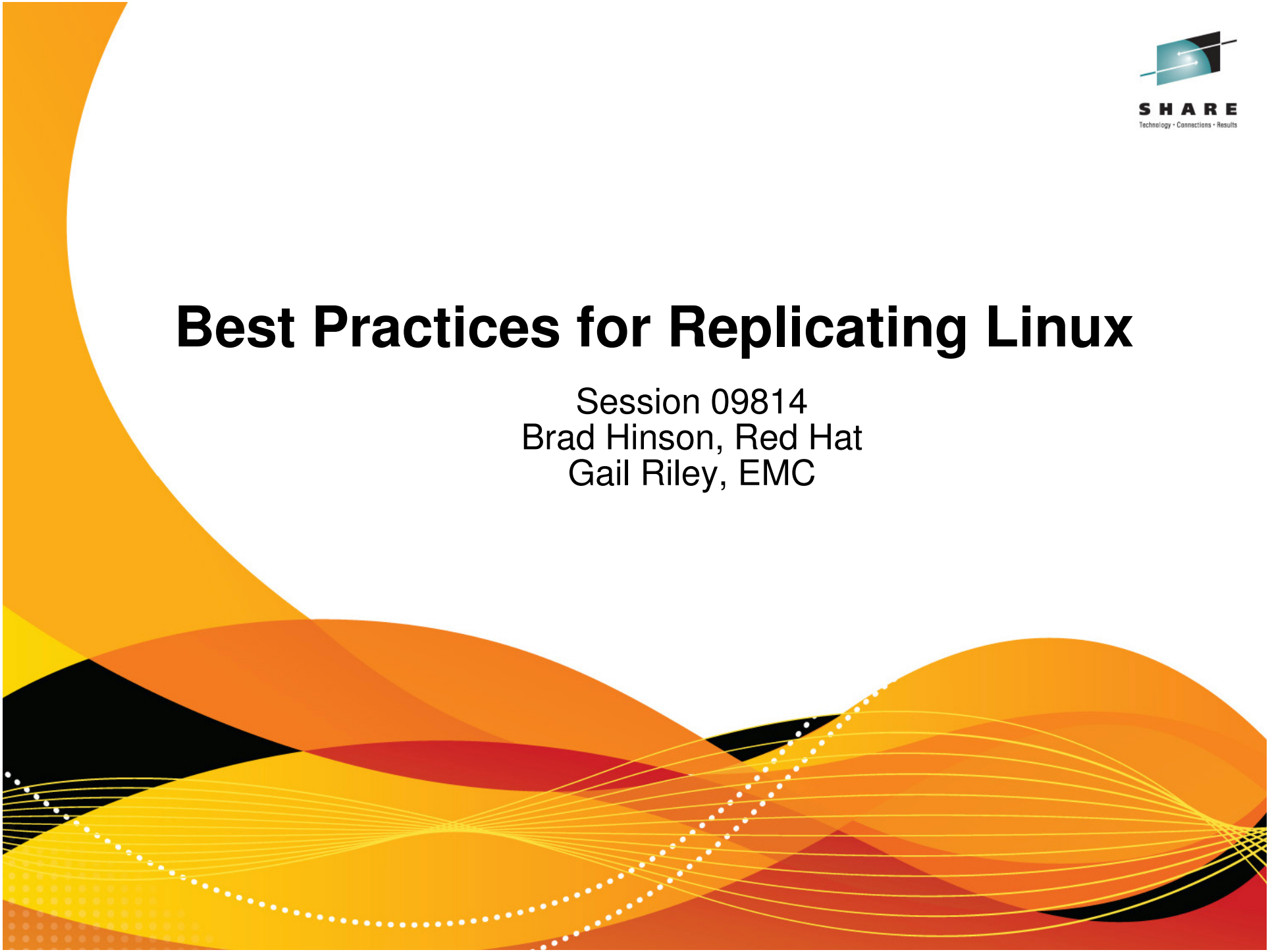


Best Practices for Replicating Linux

Session 09814

Brad Hinson, Red Hat
Gail Riley, EMC

Objectives

After completing this session, you will be able to:

- Discuss the considerations when implementing replication
- Understand the Red Hat clone process
- Describe the tasks for accessing a Local and Remote replica in a Linux on System z environment

Disaster Recovery versus Disaster Restart

- Most business critical applications have some level of data interdependencies
- Disaster recovery
  - Restoring previous copy of data and applying logs to that copy to bring it to a known point of consistency
  - Generally implies the use of backup technology
  - Data copied to tape and then shipped off-site
  - Requires manual intervention during the restore and recovery processes
- Disaster restart
  - Process of restarting mirrored consistent copies of data and applications
  - Allows restart of all participating DBMS to a common point of consistency utilizing automated application of recovery logs during DBMS initialization
  - The restart time is comparable to the length of time required for the application to restart after a power failure

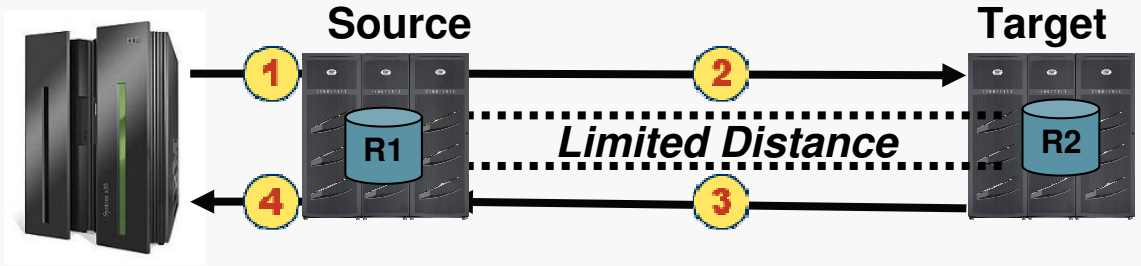
Forms of Remote Replication

- Synchronous Replication
  - Identical copies of data across storage systems where writes are committed across to remote systems/sites first which increases execution time
    - ***Source = Target***
- Asynchronous Replication
  - Data is a point-in-time consistent copy but writes happen locally and are sent across to remote systems/sites at a periodic interval
    - ***Source ≅ Target***
- Data Distribution -
  - Data is copied from one storage system to another without maintaining a consistent recoverable copy
    - ***Source ≠ Target***

Symmetrix Remote Data Facility: Two Site solutions

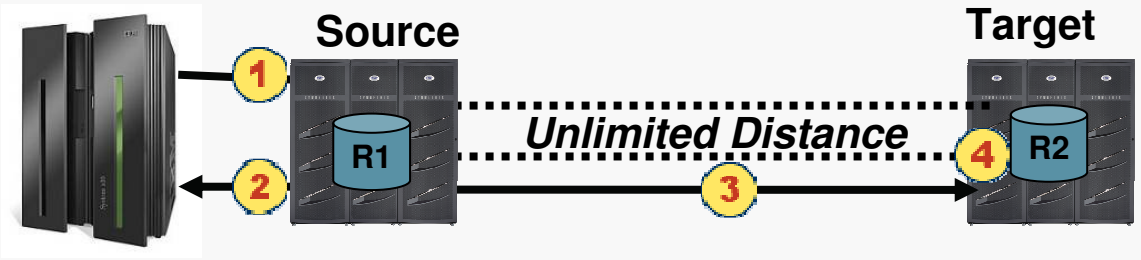
SRDF/Synchronous

- No data exposure
- Some performance impact
- Limited distance



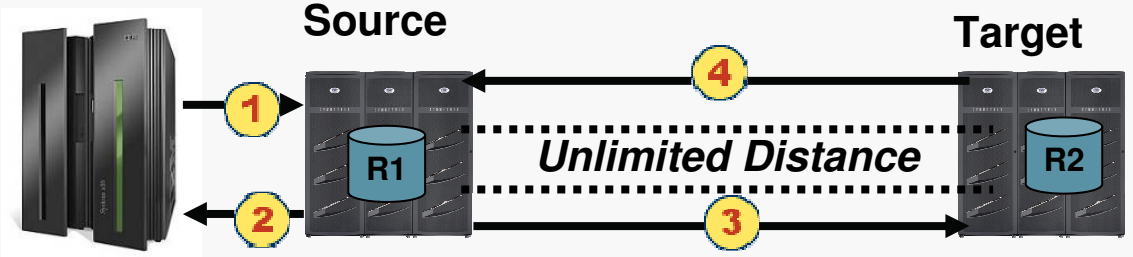
SRDF/Asynchronous

- Predictable RPO
- No performance impact
- Extended distance



SRDF/AR

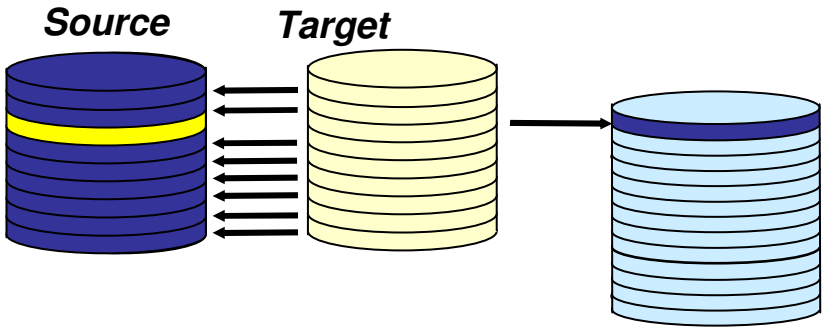
- Data Movement solution
- No performance impact
- Unlimited distance



Forms of Local Replication

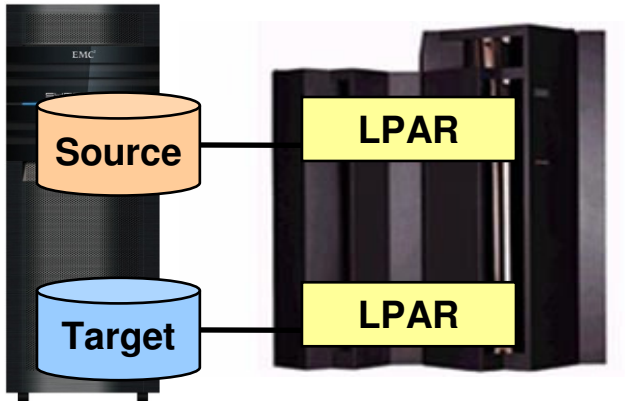
- Full Volume Copy - Clone
  - Data is copied from the Source Device to a Target Device of equal size and emulation

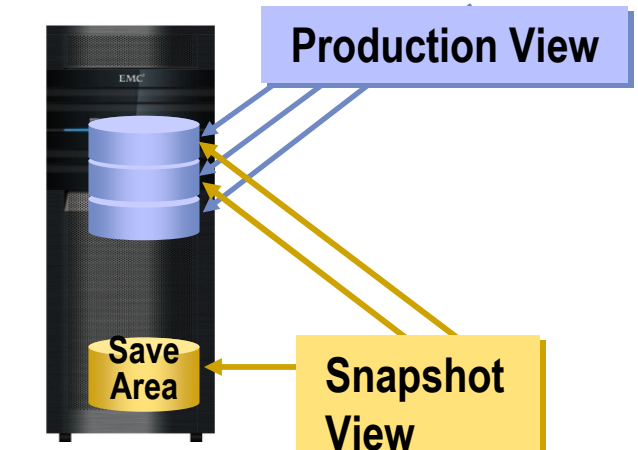
- Pointer Based Replication - Snap
  - The Target Device is a virtual device housing a collection of pointer between the Source and a reserve area for a point-in-time view



TimeFinder – Local Replication

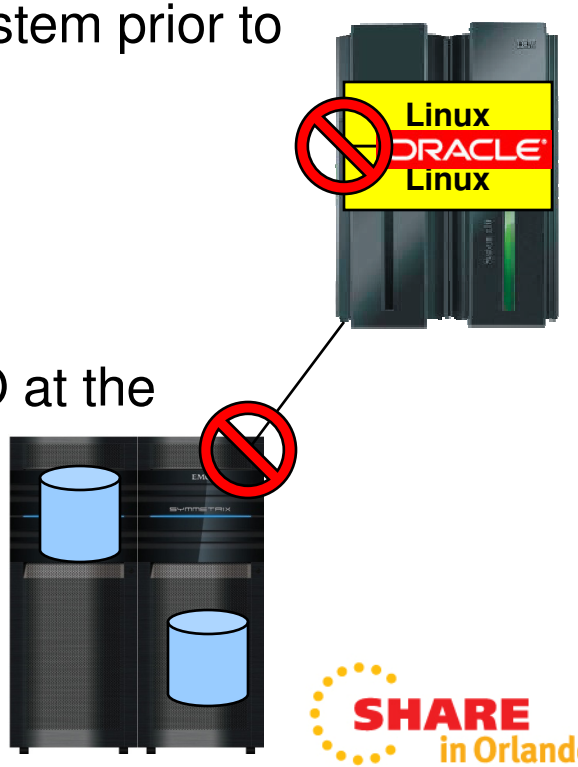
- Clone
  - Provides up to 16 concurrent, instant Point-in-Time:
    - Copies of a Volume
  - Immediately accessible after activation
    - The CLONE is completed in the background in the Symmetrix
  - Target device can be larger than Source
- Snap
  - SNAP'S create logical point-in-time "snapshots" of a source volume
  - Requires only a fraction of the source volume's capacity (based on percentage of writes)
  - Multiple Snapshots can be created from a source volume and are available immediately
  - Snapshots support read / write processing
  - Supports mainframe and open systems host environments





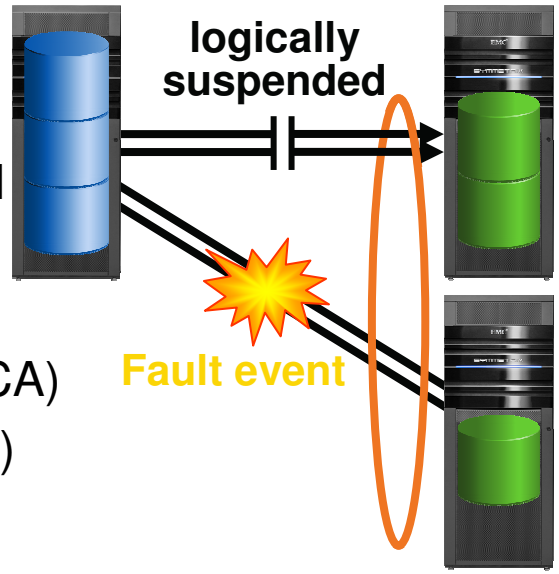
Creating a TimeFinder Consistent Copy

- Different options depending on application and host requirements
- Server
  - Pause I/O at the Server Level to provide a Consistent Point-in-Time Copy
- Application
  - Stop the application and unmount the file system prior to activate or split
  - Database hot backup mode
  - Database freeze/thaw
- Symmetrix based
  - Enginuity Consistency Assist (ECA) holds IO at the Symmetrix until all Splits/Activate complete



SRDF/Consistency Groups Overview

- Preserves dependent-write consistency of devices
  - Ensures application dependent write consistency of the application data remotely mirrored by SRDF operations in the event of a rolling disaster
  - Across multiple Symmetrix systems and/or multiple SRDF groups within a Symmetrix system
- A composite group comprised of SRDF R1 or R2 devices
  - Configured to act in unison to maintain the integrity of a database or application distributed across Symmetrix systems
- Included with SRDF/S and SRDF/A
  - SRDF/S using Enginuity Consistency Assist (ECA)
  - SRDF/A using Multi Session Consistency (MSC)



Ensures dependent-write consistency of the data remotely mirrored by SRDF

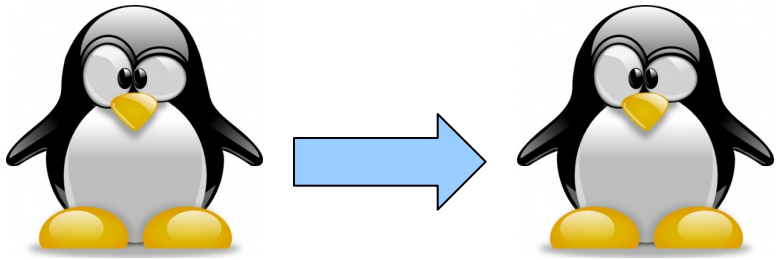
Linux on System z Replication Devices

- The Symmetrix SRDF and TimeFinder replicate disk drives
  - FBA
    - SCSI/FBA devices
    - z/VM edev
  - CKD
- The Symmetrix supports the z/VM FlashCopy command

Replication Options

- Storage array supplied replication process for local and remote replication
- Linux Operating Systems utilities
  - Red Hat clone rpm – local replication
  - rsync for remote directory refresh
- Create your own local replication process

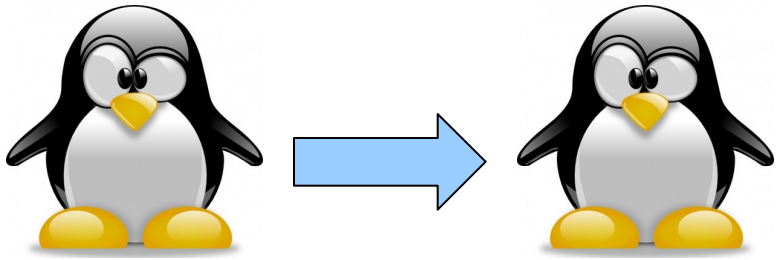
Red Hat Clone rpm



- Provided with RHEL Virtualization Cookbook
  - http://www.vm.ibm.com/devpages/mikemac/SG247932.tgz
  - http://people.redhat.com/bhinson/clone/ (latest copy)
- Requirements
  - Cloner guest, source guest (separate guests, cloner can't clone itself)
  - z/VM user definition for new/target clone must exist
  - Cloner must have privilege class B for FlashCopy and attach*
  - For “dd” options, cloner must LINK disks to copy
    - OPTION LNKNOPAS or
    - LINK password set to “ALL” for read & write
    - MDISK definitions for DASD, not DEDICATE
  - For LVM installs, cloner Volume Group name must be different from source

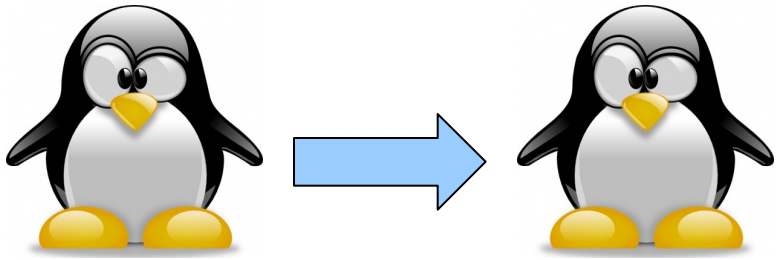
*attach is used for FCP port access

Red Hat Clone rpm



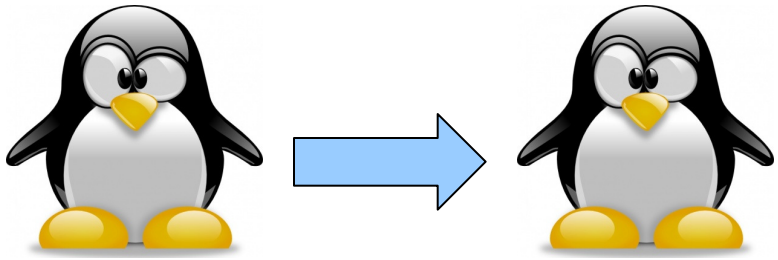
- Configuration file (/etc/sysconfig/clone)
  - AUTOLOG=
    - Boot guest automatically after cloning
  - CLONE_METHOD=
    - FlashCopy “auto” or Linux “dd”
  - CLONE_FCP=
    - symclone or Linux “dd”
- Clone configuration files (/etc/clone)
  - rhel.conf.sample: sample values. Copy to {target ID}.conf
  - Similar values can be copied to shared.conf

Red Hat Clone rpm



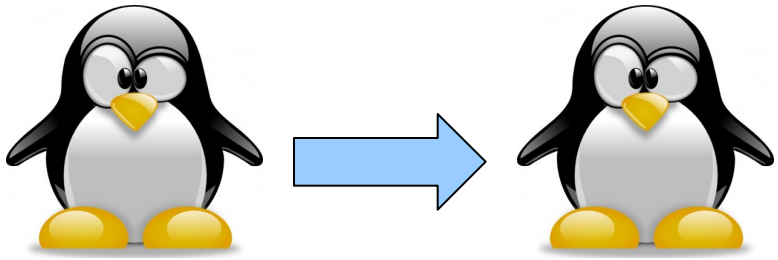
```
# rpm -ivh clone-1.0-12.s390x.rpm
Preparing...                ########################################### [100%]
   1:clone                  ########################################### [100%]

# cp /etc/clone/rhel.conf.sample /etc/clone/newguestID.conf
# vi /etc/clone/newguestID.conf

# clone -v masterguestID newguestID

This will copy disks from masterguestID to newguestID
Host name will be:  newguestID.s390.bos.redhat.com
IP address will be: 10.16.105.65
Do you want to continue? (y/n): y

[…]
Invoking Linux command: dasdfmt -p -b 4096 -y -F -f /dev/dasdd
cyl  3338 of  3338 |###############################################| 100%
Invoking Linux command: dd bs=4096 count=600840 if=/dev/dasdc of=/dev/dasdd
[...]
```

Red Hat Clone rpm



- CLONE_FCP=dd
  - Read zFCP configuration on source system
  - Specify zFCP configuration of target system
    - /etc/clone/zfcp-{target}.conf
  - Attach source and target FCP port to cloner
  - Clone will bring both sets of LUNs online, use Linux “dd” to copy
- CLONE_FCP=symclone
  - Specify device group in configuration (SYMDG=)
  - Clone calls Symmetrix command-line utilities:
    - symclone {create, activate}
    - symclone {verify} gives updates until copy complete
    - symclone {terminate} to break connection

Red Hat Clone rpm



clone -v masterguestID newguestID
[…]

Calling symclone to copy FCP disks ...

Execute 'Create' operation for device group
'*clone-dg*' (y/[n]) ? **y**
[…]
Execute 'Activate' operation for device group
'*clone-dg*' (y/[n]) ? **y**
[…]
waiting for symclone to complete...
None of the devices in the group '*clone-dg*' are in 'Copied' state.
None of the devices in the group '*clone-dg*' are in 'Copied' state.
[…]
All devices in the group '*clone-dg*' are in 'Copied' state.

Execute 'Terminate' operation for device group
'*clone-dg*' (y/[n]) ? **y**

Clone rpm - prereq's for symclone

- On the Linux instance where the clone will be executed
  - Solutions Enabler is required
    - Minimum of 1 gatekeeper required
  - Create a Symmetrix device group containing the Symmetrix device (symdev) source and symdev target devices

CKD Replication Considerations

- Minimal changes may be required for CKD local and/or remote replication, but it depends…..
- Minidisks
  - Full or partial – if replicating z/VM, no directory changes needed at remote site
  - mdisk rdev – same as DEDICATE
  - Avoid duplicate VOLSER at same LPAR, site
- DEDICATE/ATTACH
  - No change if real device address is the same at the primary and backup site
  - Use virtual addresses to mask changes at the Linux layer

SCSI Considerations

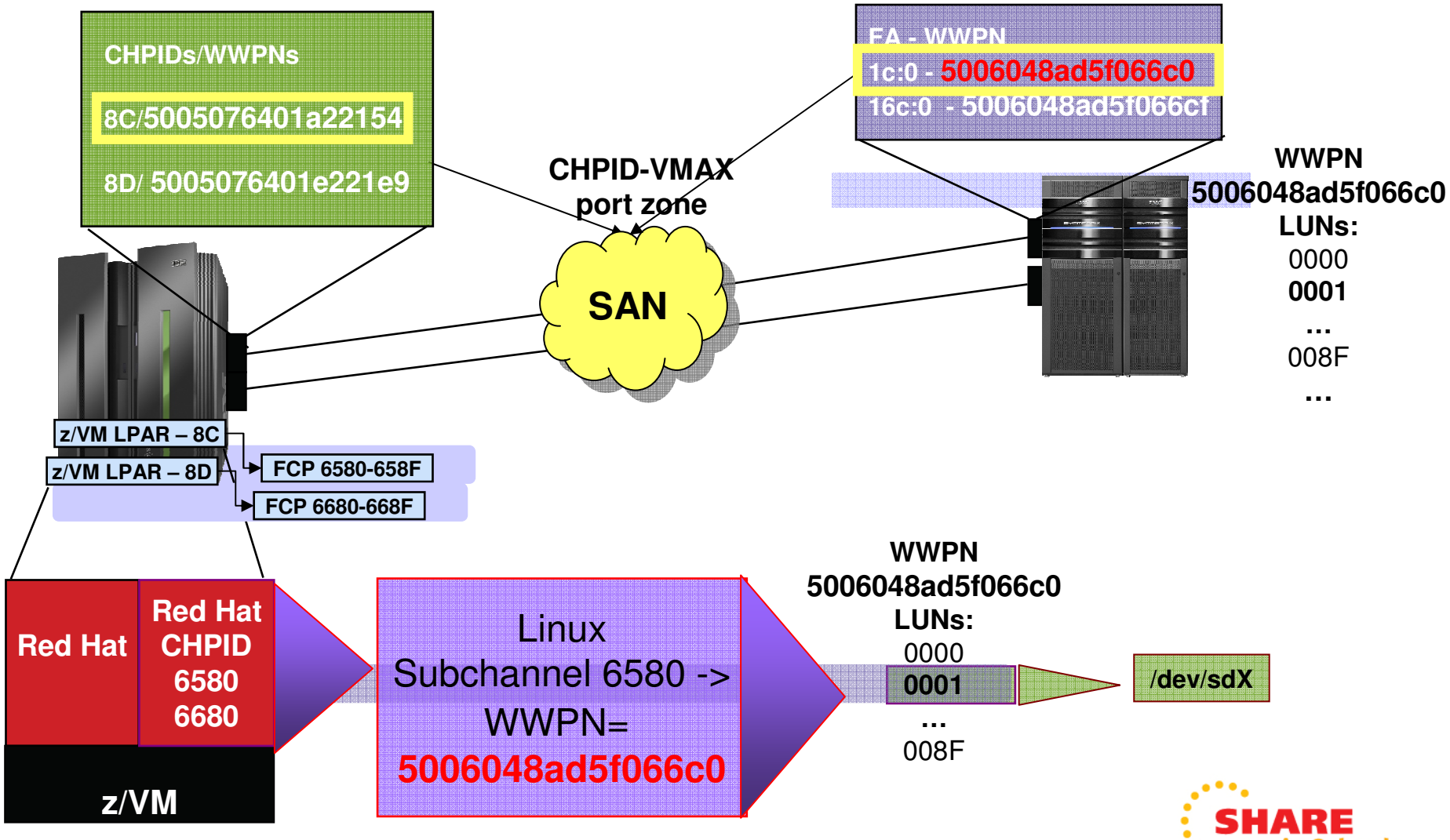
- Why is SCSI being implemented?
  - Performance – asynchronous I/O
  - Familiar to open systems users
  - Better use of physical devices
  - Ability to have larger devices
    - kernel dependent - currently 2TB max
  - Dynamic configuration – can add a new LUN without IOCDS change
- What are the challenges?
  - SAN - not familiar to everyone, zoning and masking required
  - To use NPIV or not
  - How to handle changing WWxN LUN information
  - Performance monitoring is at the Linux layer

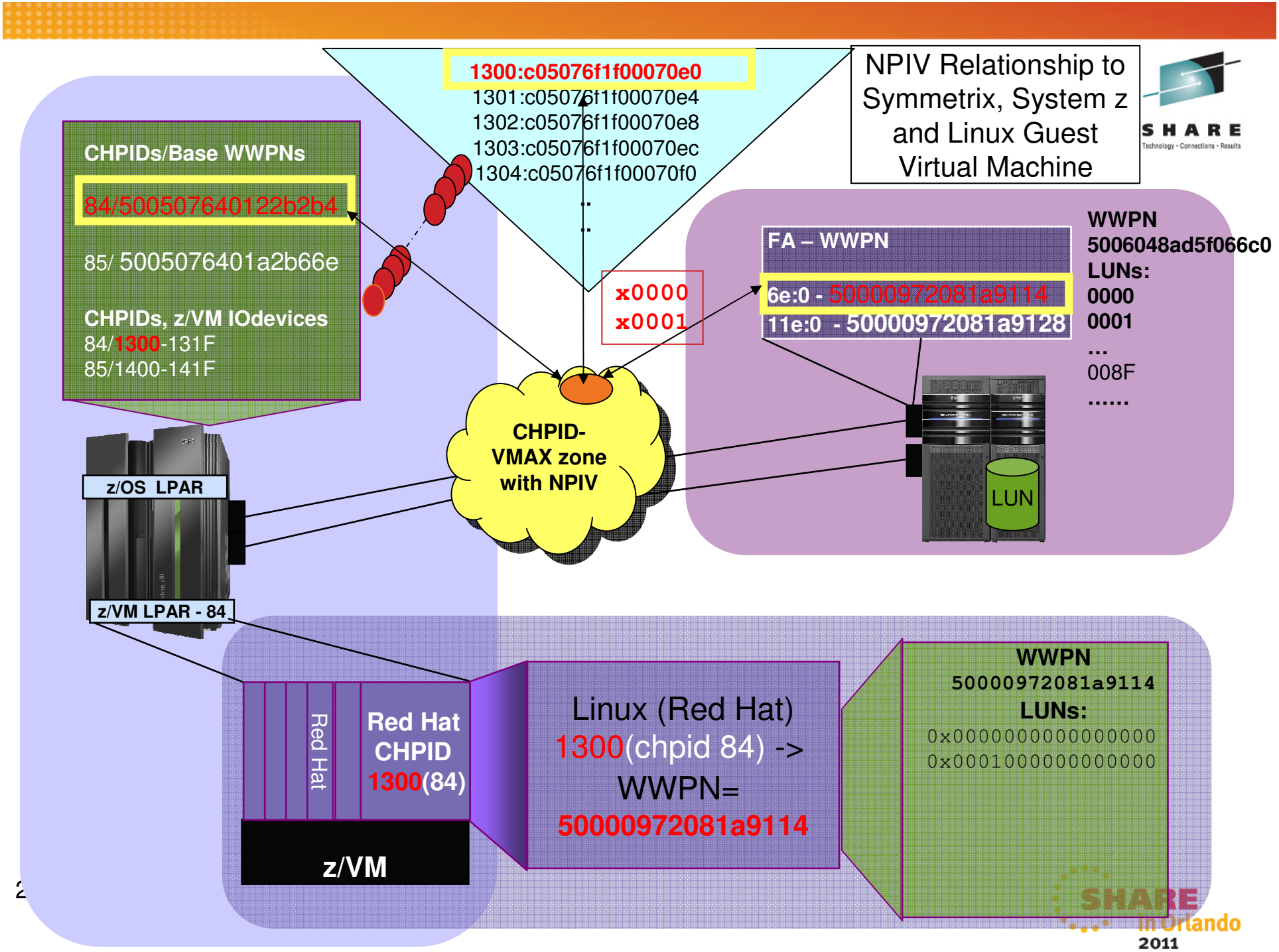
FCP Path Relationship without NPIV

(z/VM Channel/subchannel device) + (Symmetrix port WWPN + LUN (Symmetrix Logical Volume))

(6580) + (5006048ad5f066c0 + 0001) = /dev/sdX

6581-658F




NPIV Relationship to Symmetrix, System z and Linux Guest Virtual Machine
1300:c05076f1f00070e0
1301:c05076f1f00070e4
1302:c05076f1f00070e8
1303:c05076f1f00070ec
1304:c05076f1f00070f0
CHPIDs/Base WWPNs
84/500507640122b2b4
85/ 5005076401a2b66e
CHPIDs, z/VM IOdevices
84/1300-131F
85/1400-141F
x0000
x0001
FA – WWPN
6e:0 - 50000972081a9114
11e:0 - 50000972081a9128
WWPN
5006048ad5f066c0
LUNs:
0000
0001
...
008F
.....
LUN
CHPID-
VMAX zone
with NPIV
z/OS LPAR
z/VM LPAR - 84
Red Hat
Red Hat
CHPID
1300(84)
z/VM
Linux (Red Hat)
1300(chpid 84) ->
WWPN=
50000972081a9114
WWPN
50000972081a9114
LUNs:
0x0000000000000000
0x0001000000000000
SHARE
Technology · Connections · Results
SHARE
in Orlando
2011

SCSI Considerations with Replication

- WWxN will change
  - When using NPIV and a different FCP port (subchannel) than the source FCP port
  - Using the same FCP port/subchannel number on a different LPAR
  - Using a FCP port at a different site
  - No NPIV, different CHPID
- WWxN will not change with no NPIV and any port on same CHPID
  - This means all LUNs mapped and masked to CHPID WWxN may be seen through all FCP ports/subchannels on the CHPID

SCSI Considerations with Replication

- Use a different, unique WWxN (NPIV port) for your clone SCSI devices
  - For nonNPIV use a different CHPID
- How can I get Linux to recognize the new WWxN and find its data?
  - Update specific Linux files
  - Use scripting
  - Use Logical Volume Manager (LVM)

Minimize changes to Linux for failover

- Use Linux facilities already in place when using NPIV
  - /etc/zfcp.conf - List second site (DR) entries also along with Site 1
  - Correct paths will be found at each site
  - Updates are made in one location

```
# site 1 R1 path
0.0.1330 0x50000972081a9114 0x0000000000000000
0.0.1330 0x50000972081a9114 0x0001000000000000
.....................
#
# site 1 R1 path
0.0.1430 0x50000972081a9128 0x0000000000000000
0.0.1430 0x50000972081a9128 0x0001000000000000
.......
#
# site 2 R2 path
0.0.1010 0x50000972081acd59 0x0000000000000000
0.0.1010 0x50000972081acd59 0x0001000000000000
.....
# site 2 R2 path
0.0.1110 0x50000972081acd65 0x0000000000000000
0.0.1110 0x50000972081acd65 0x0001000000000000
.....................................
```

VM Directory – Production and Clone

- Production Site 1 and 2

```
USER PR192166

* FCP for R1 site
dedicate 1330 1330
dedicate 1430 1430
* FCP for R2 site
dedicate 1010 1010
dedicate 1011 1011
........
```

- Clone Site 1 and/or 2

```
USER CL192166

* FCP for R1 site – R1 CLONE
  dedicate 1331 1331
  dedicate 1431 1431
* FCP for Site 2 – R2 Clone
 dedicate 101a 101a
 dedicate 111a 111A
........
```

Red Hat Multipathing

- /etc/multipath.conf – basic configuration file
  - Created and maintained by the multipath program
  - /etc/multipath/bindings
  - /etc/multipath/wwids
- Both files contain wwid for each device with different entries for Site 1 and Site 2 ➔ different physical device
  - Site1

    360000970000192601700533030383737
  - Site2

    360000970000192601715533030333032

Use LVM with Replicated Copies

- LVM masks the changing SCSI multipath information
- Volume groups (VG) are made up of LVM physical volumes (PVs)
- LVM physical volumes are identified by PV UUID, not multipath device UUID/WWID
- Logical volumes(LVs) are associated to LVM volume groups
- Filesystems are associated to logical volumes in /etc/fstab
- All LVM entities are found, brought online and the filesystem mounted at Site 2, no different than Site 1

How can I test my replication environment?

- Clones/Snaps can be used at the Primary or DR site
  - Ensure consistency across all devices at time of clone creation if there are interdependencies
- System Considerations - Make sure you have a unique environment for your clone
  - Create a separate VM directory entry for clone use
  - CKD minidisks
    - make sure the VOLSER is unique if using fullpack minidisks
  - DEDICATE/ATTACH
    - make sure the same *virtual* address is used
  - Change the network – IP address, DNS as appropriate
  - Use different NPIV/WWxn ports than the production environment
  - Are there cron jobs you need to disable on the clone?

Application Considerations when Cloning

- Does it start up automatically?
- Does it connect to another application, IP address?
- Does it use a NFS mounted filesystem?
- Does it export information when it starts?
- Does it download or upload information when it starts or sometime during its instantiation?
- Does the application rely on a specific
  - Hostname
  - IP address
  - raw device
- Identify any application interdependencies

Linux Replication Considerations

- Both Local and Remote Replication have device considerations
  - CKD and/or FBA devices are supported
  - Use device-by-path, not device-id for device setup
  - Replicated devices have the same virtual addresses at both sites
  - SCSI LUN mapping is the same at both sites
  - Let LVM assist you in reducing changes for replicated copies

- Other considerations
  - Automate the process wherever possible
  - Standardize wherever possible, i.e., addressing scheme for system, application, other devices
  - Shared R/O Linux kernel –
    - May create unintended interdependencies between (application) environments
    - One environment can force another to upgrade
  - Don't forget about backups at the DR site

Discussion Topic Recap

- Replication methods
  - Sync vs. async
  - Manual vs. clone rpm
- Script customization for local and/or remote copies
- NPIV requirements
- Local vs. Remote replication considerations
- Use of LVM to handle replication failover
- Application considerations