Database Virtualization Technologies

DELPHIX®

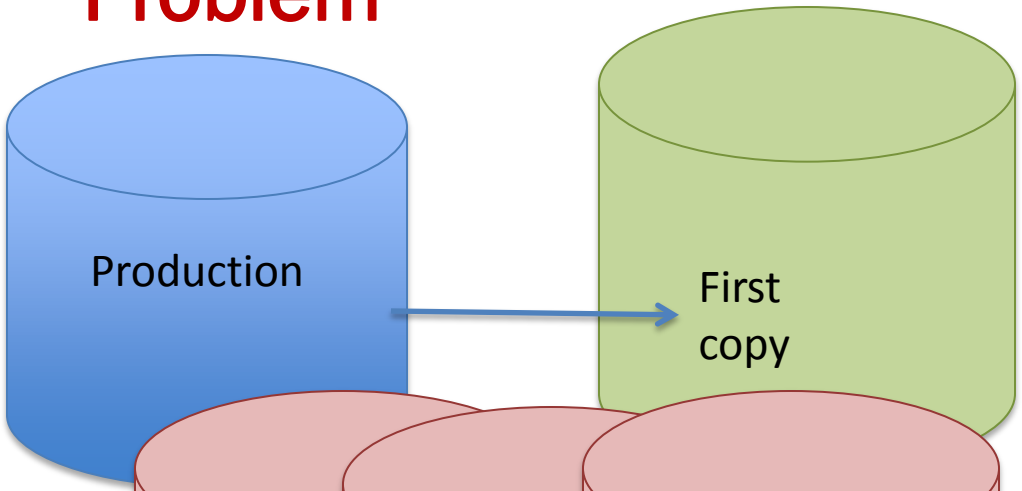
DELPHIX

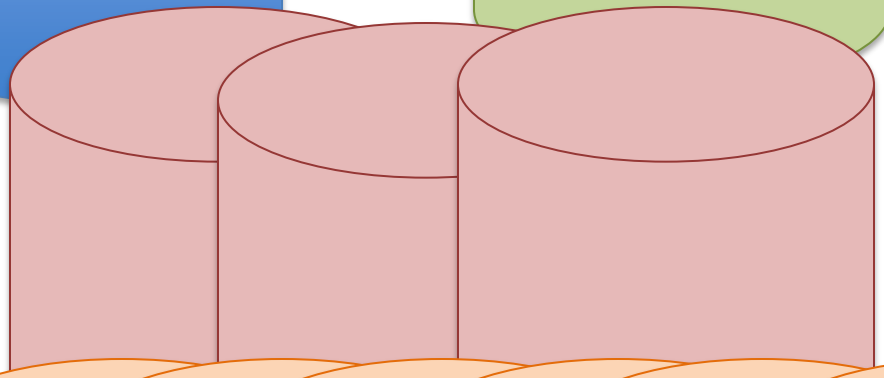
Database Virtualization

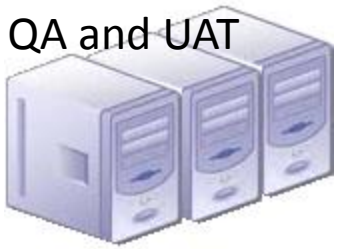
- Comes of age
  - CloneDB : 3 talks @ OOW
    - Clone Online in Seconds with CloneDB (EMC)
    - CloneDB with the Latest Generation of Database (Oracle)
    - Efficient Database Cloning with Clonedb (Cern)
  - Oracle 12c: new feature
  - Companies:
    - Delphix
    - EMC
    - NetApp
    - Vmware

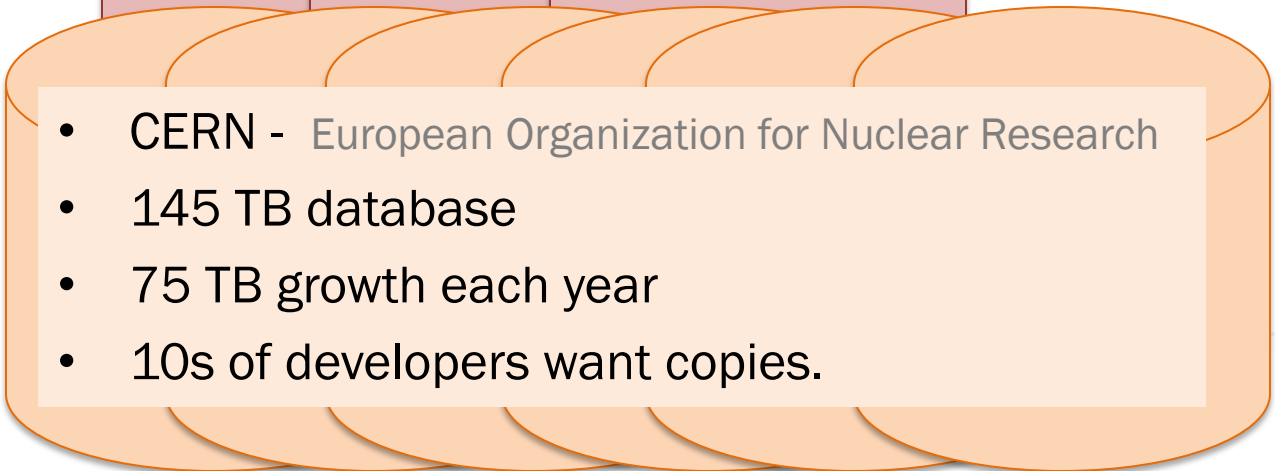
- What is it ?
  - database virtualization is for data tier
    as VMware is for compute tier

Problem







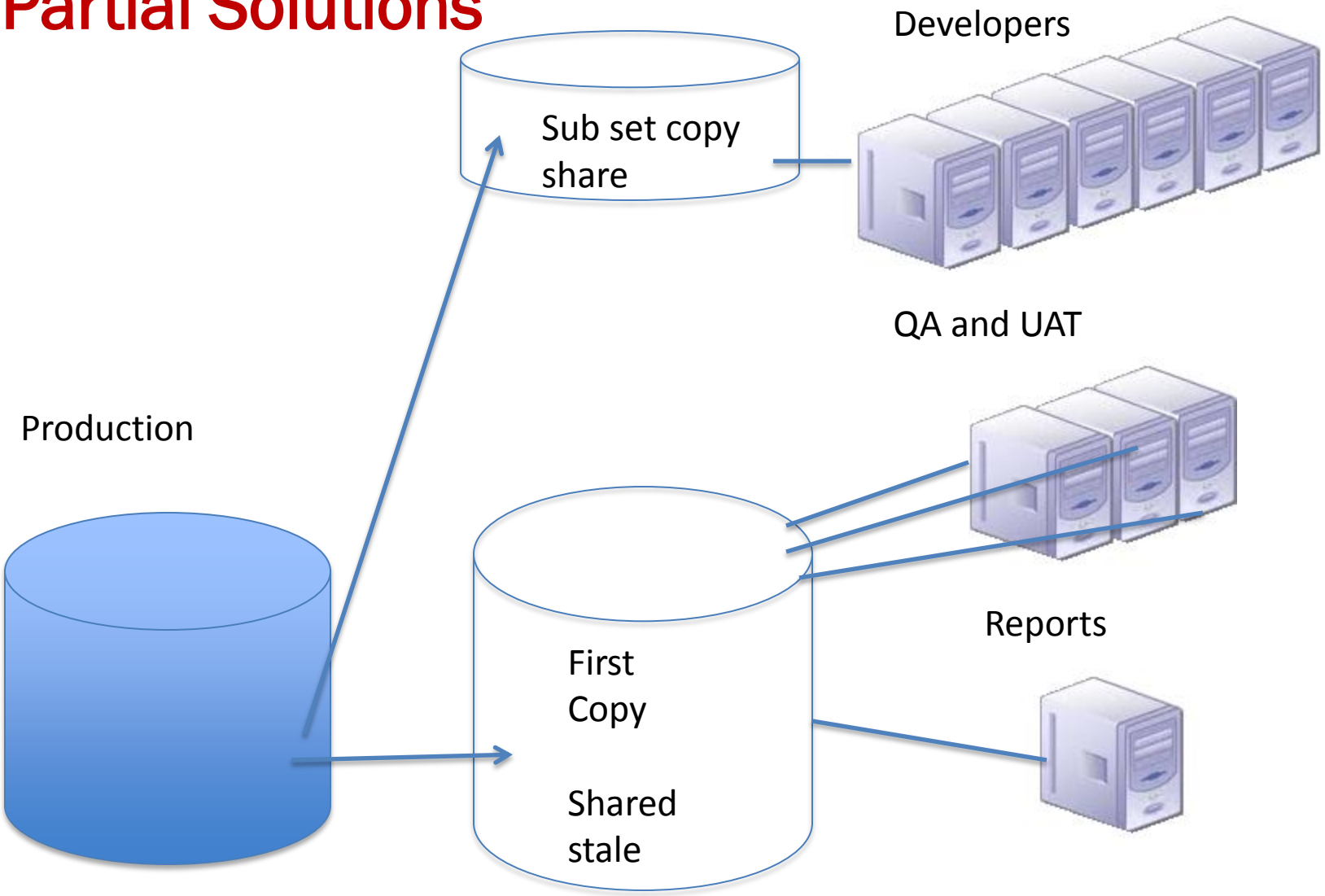


- CERN - European Organization for Nuclear Research
- 145 TB database
- 75 TB growth each year
- 10s of developers want copies.

Full copies problematic sometimes impossible

- Time consuming
  - Time to make copes, days to weeks
  - Meetings , days to weeks
    - System
    - Storage
    - Database
    - Network Admins
    - manager coordination
- Space consuming
  - 100 devs x 10TB production = 1 Petabyte
    - This is 100x actual unique data
    - Unique data is
      - 10 TB original
      - 2TB of changed data
      - ➢ = 12TB total unique data

Partial Solutions

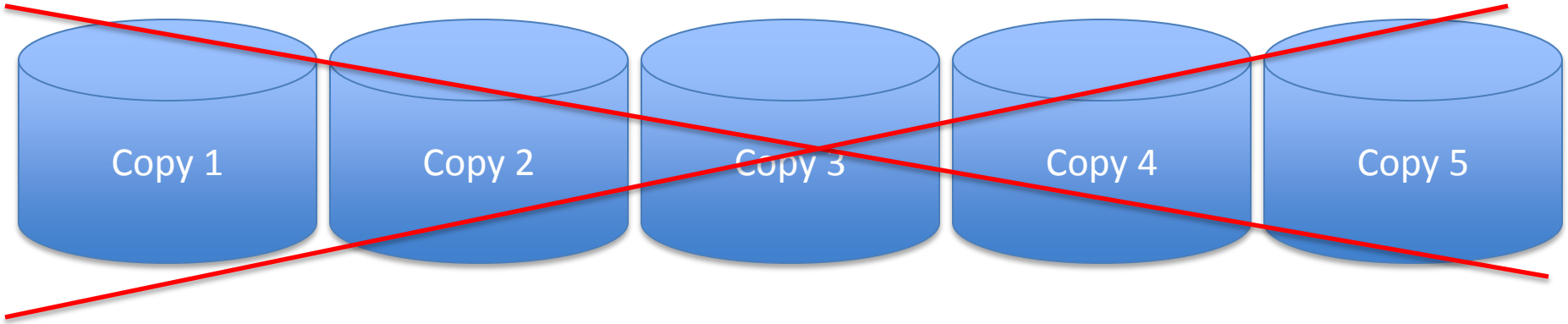


Partial solutions, create more problems

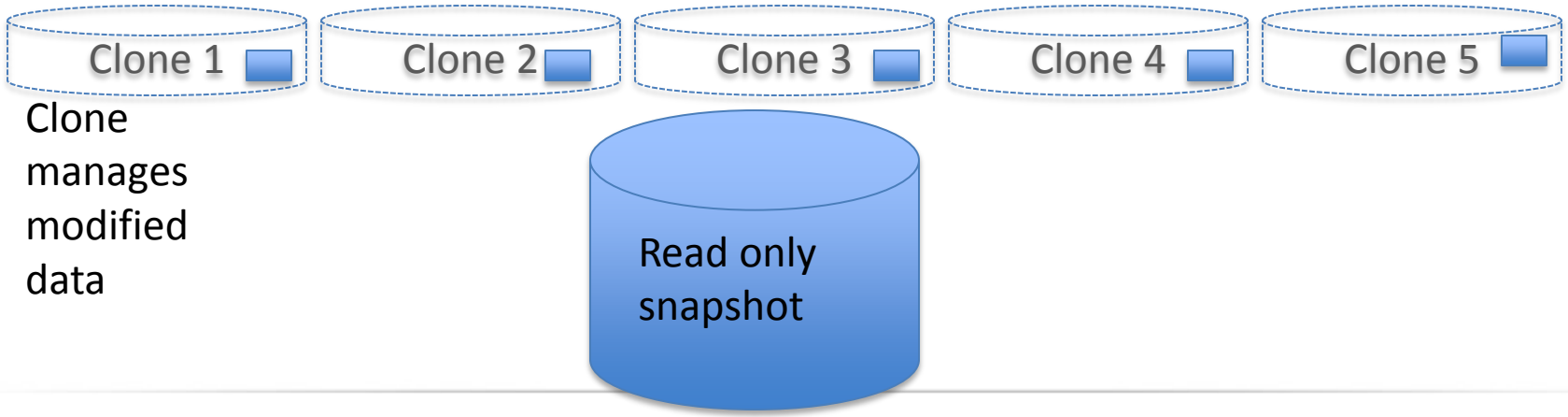
- Share copies -> slow projects down
  - long delays for new copies -> Stale copies
  - Stale copies give -> Incomplete results
  - Hard to get a new copy if everyone is sharing current copy
  - Shared copies slow down development
- Subset copies -> misleading and/or wrong
  - Incomplete results
  - Performance results may be wrong

Solution: Clone and Share

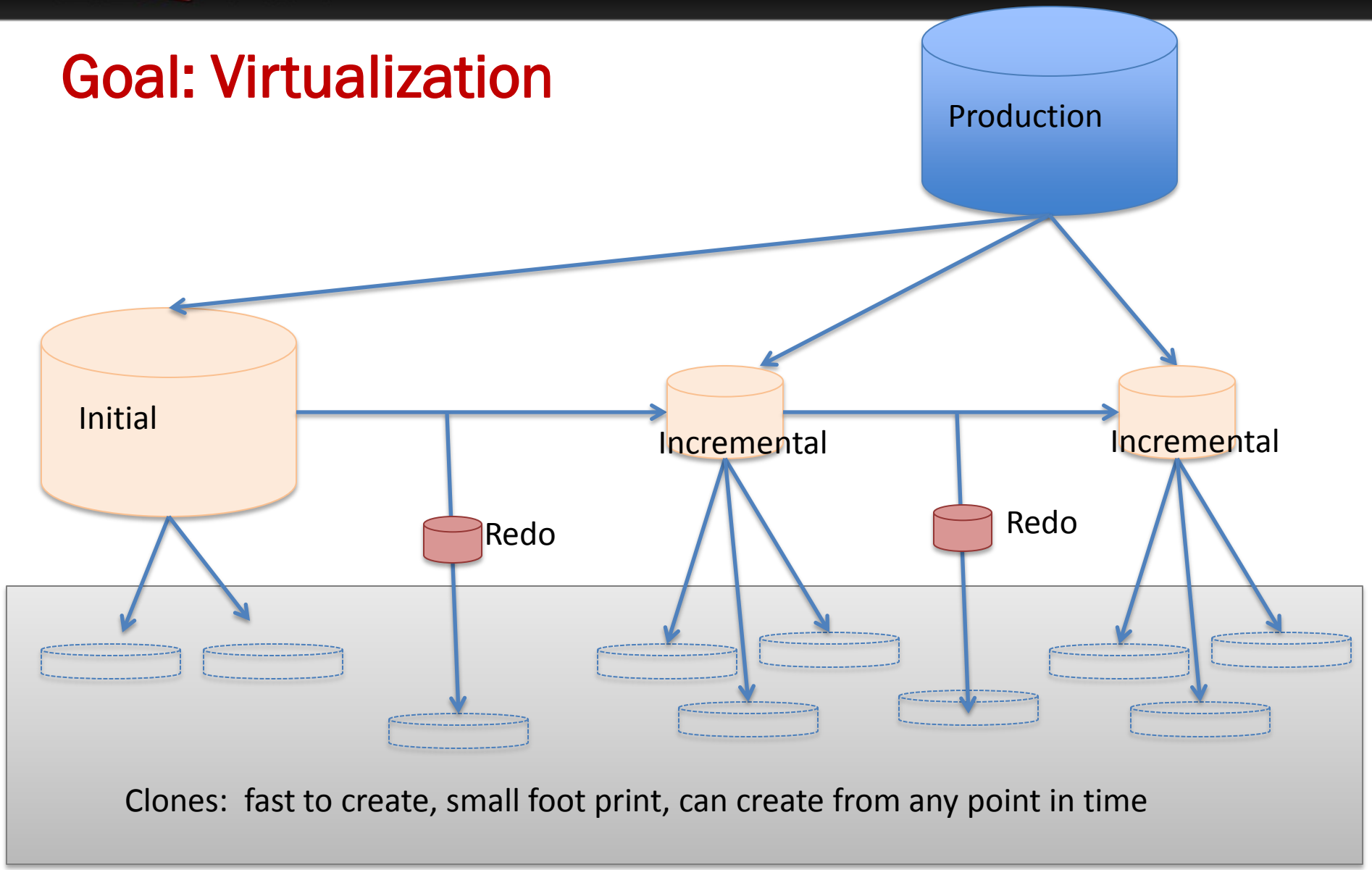
Instead of full copies of same data



One Read Only Copy plus thin layer of changes per clone



Goal: Virtualization



DELPHIX

Technologies:

1. CloneDB (Oracle)
2. ZFS Storage Appliance (Oracle)
3. Delphix
4. Data Director (Vmware)
5. EMC
6. NetApp
7. Oracle 12c Snap Manager Utility (SMU)

Virtualization : Advantages

- Space
  - Clones sharing a single snapshot
    - 100 copies of 10 TB goes from 1 Petabyte down to 3 TB with compression
- Speed
  - Eliminate Coordination
    - System, Storage, Database, Network Admins + manager coordination
  - Creation = time to start a database
- Agility

You Should be cloning now

If you have any of :

- Oracle 11.2.0.2+
- Oracle ZFS Storage Appliance
- NetApp

Gives you

- Storage savings
- More importantly time savings

Agility

How many copies are of database are made?
What size are these databases?
How often are the copies made?

DELPHIX

What do the technologies offer?

1. Snapshot
   - All (some more limited than others)
2. Roll Snapshot forward
   - NetApp, Delphix, ZFS
3. Clone
   - All (some more limited than others)
4. Provision
   - Oracle12c, Delphix
5. Automate
   - Delphix

Automation

- Source database changes
  - incremental backups
  - Redo collection
  - Retention windows
  - Expose file systems
- Create databases from clones
  - assigning SID
  - Parameters
  - file structure
  - recovery
  - Security
- Cloud ready
  - Hardware agnostic
  - Multi database support Oracle,SQL Server, Sybase, DB2, PostGres, MySQL
- Masking data
- Load Balancing
  - Provision databses on hardware with available resources

Types of solution – (part 1)

- Hardware Vendor verses Software
  - **Hardware lock in**: EMC, NetAPP, Oracle ZFS Storage Appliance
  - Software: CloneDB, Delphix, Data Director

- Database Specific versus General purpose Copies
  - Oracle Specific: CloneDB
  - General Purpose: EMC, NetApp, Oracle ZFS Appliance, Data Director
  - Multi Database Specific: Delphix*

*Oracle, SQL Server, User Data , other DBs coming

Types of solution – (part II)

- Golden Copy
  - Required: EMC, DataDirector, CloneDB
  - Not Required: Delphix, Oracle ZFS Appliance, NetApp (snaps of snaps)

- Performance Issues
  - Data Director
  - CloneDB

CloneDB

Tim Hall
www.oracle-base.com/articles/11g/clonedb-11gr2.php

1. RMAN backup (local or NFS)
2. Create an NFS mount
3. Setup dNFS and 11.2.0.2+

```
4.Clonedb.pl  initSOURCE.ora output.sql
5.sqlplus / as sysdba @output.sql
```

CloneDB

Tim Hall
www.oracle-base.com/articles/11g/clonedb-11gr2.php

- Setup dNFS and 11.2.0.2+
  - libnfsodm11.so
  - /etc/oranfstab
- `Clonedb.pl initSOURCE.ora output.sql`
  - export MASTER_COPY_DIR="/backuplocal" # backup location
  - export CLONE_FILE_CREATE_DEST="/clone" # requires NFS MOUNT
  - export CLONEDB_NAME="clone" # ORACLE_SID="clone“
- `sqlplus / as sysdba @output.sql`
  - startup **nomount** PFILE=/clone/initclone.ora
  - **Create control** file with backup location
  - dbms_dnfs.**clonedb_renamefile**('/**backup**/sysaux01.dbf' , '/**clone**/ora_data_clone0.dbf');
  - alter database **open resetlogs**;

CloneDB

Tim Hall
www.oracle-base.com/articles/11g/clonedb-11gr2.php

Source

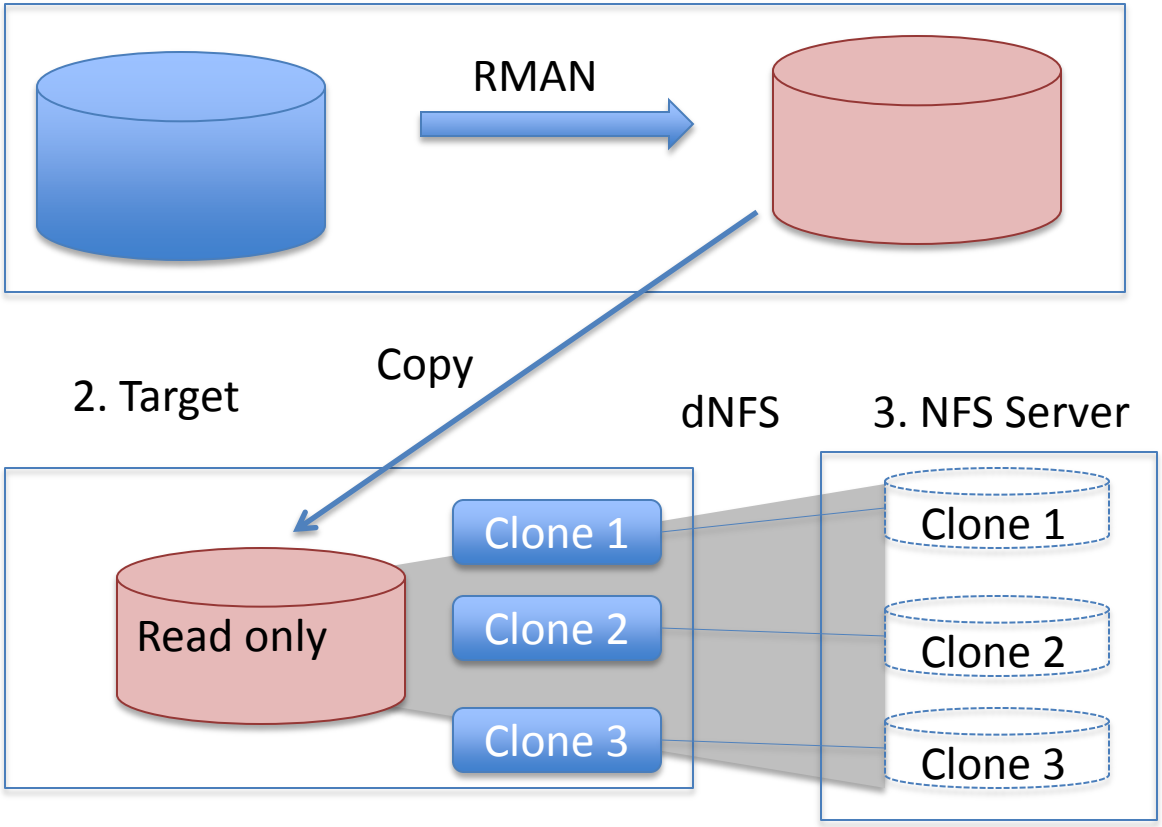
- RMAN backup as copy

Target

- Get a copy of RMAN backup (local or NFS)
- Create an NFS mount
- Setup dNFS and 11.2.0.2+
  - libnfsodm11.so
  - /etc/oranfstab
- `Clonedb.pl initSOURCE.ora output.sql`
  - `export MASTER_COPY_DIR="/backuplocal" # backup location NFS or not`
  - `export CLONE_FILE_CREATE_DEST="/clone" # requires NFS MOUNT`
  - `export CLONEDB_NAME="clone" # export ORACLE_SID="clone"`
- `sqlplus / as sysdba @output.sql`
  - startup **nomount** PFILE=/clone/initclone.ora
  - **Create control** file with backup location
  - dbms_dnfs.**clonedb_renamefile**('/**backup**/sysaux01.dbf' ,'/**clone**/ora_data_clone0.dbf');
  - alter database **open resetlogs**;

Clone DB : requires dNFS and 11.2.0.2+

1. physical



Three machines
1. Physical
2. NFS Server
3. Target

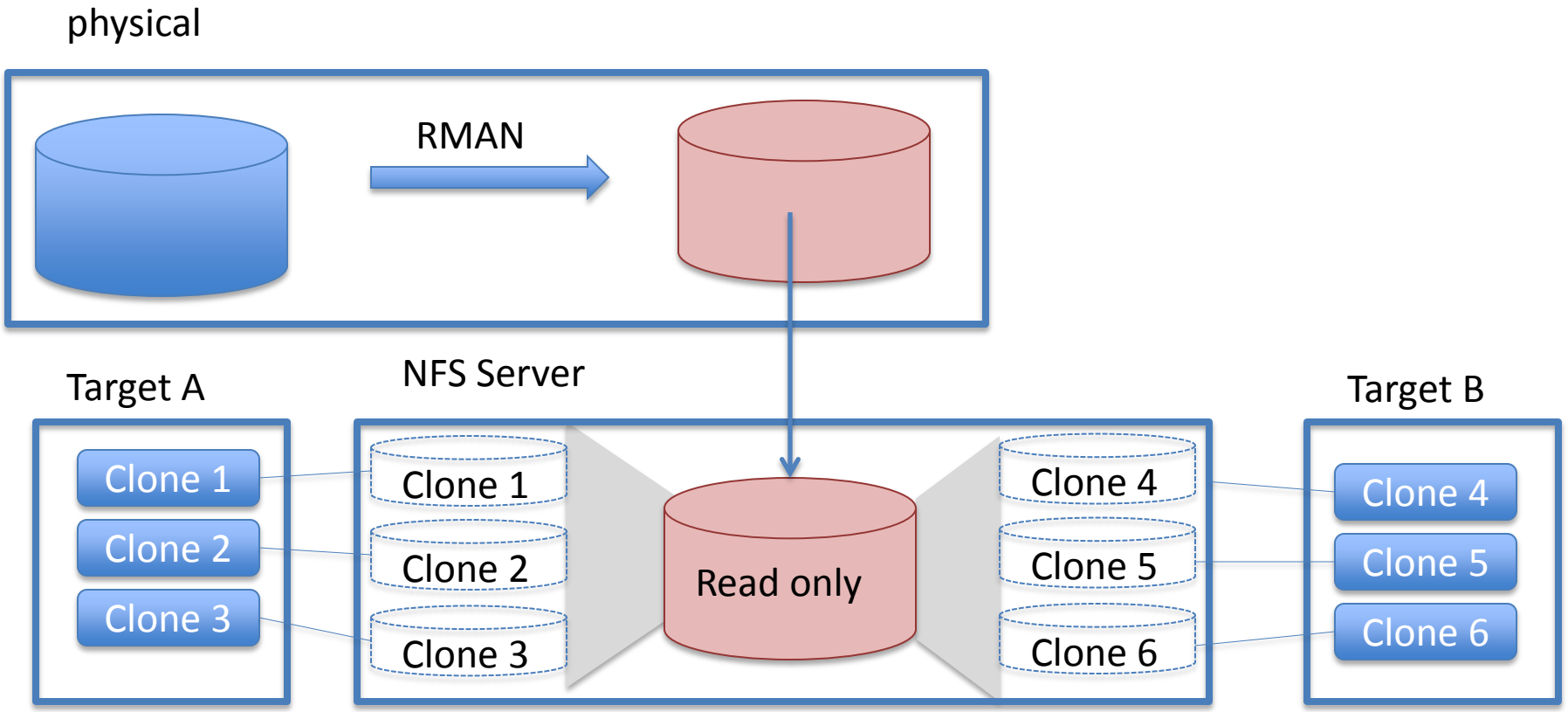
Problem:
No Versioning

830264 /backup/sysaux01.dbf
727764 /backup/system01.dbf
425388 /backup/undotbs01.dbf

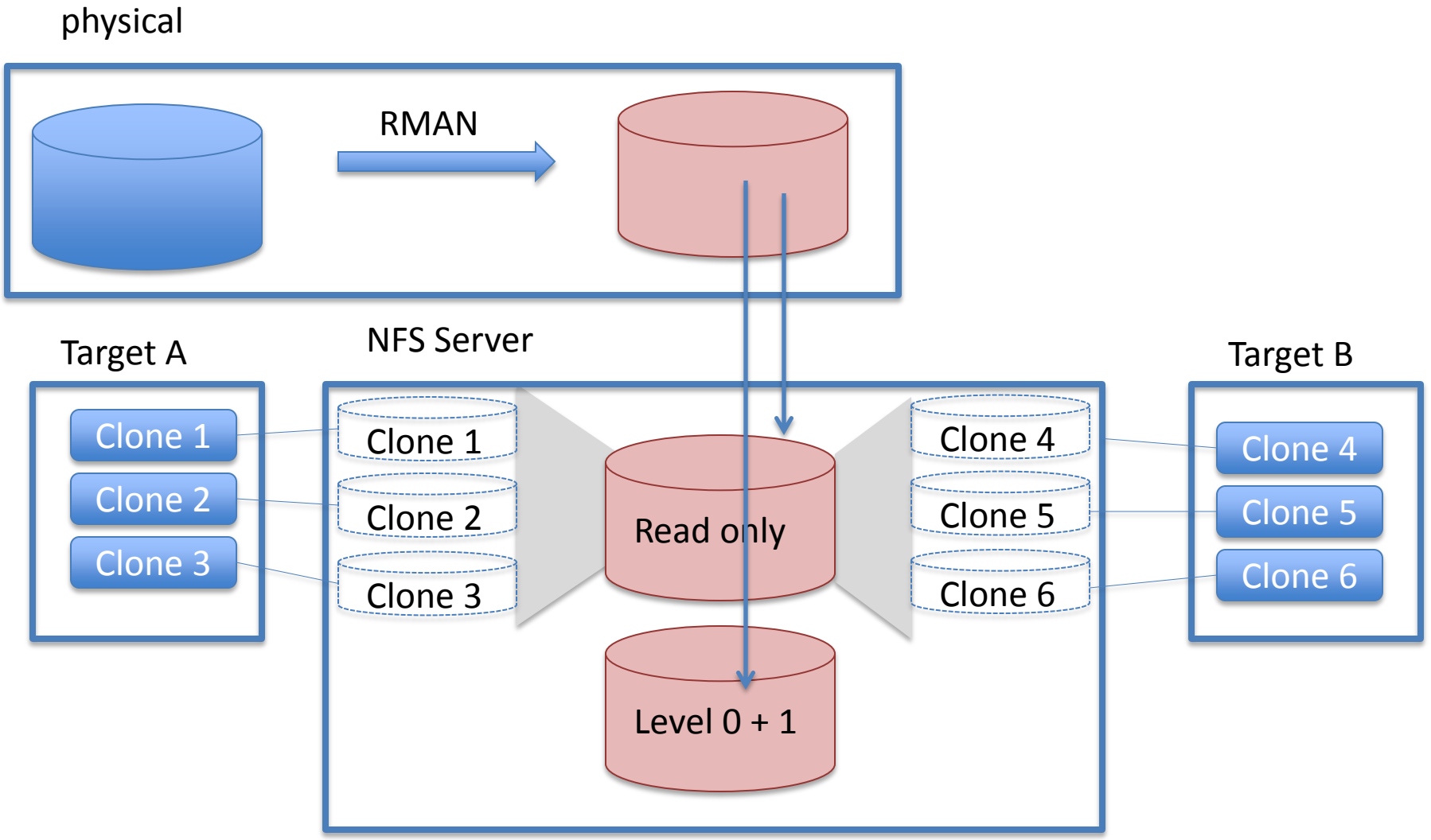
760 /clone/ora_data_clone0.dbf
188 /clone/ora_data_clone1.dbf
480 /clone/ora_data_clone2.dbf

du -sk

Clone DB : everything could be on NFS



Clone DB: refresh: either destroy or duplicate

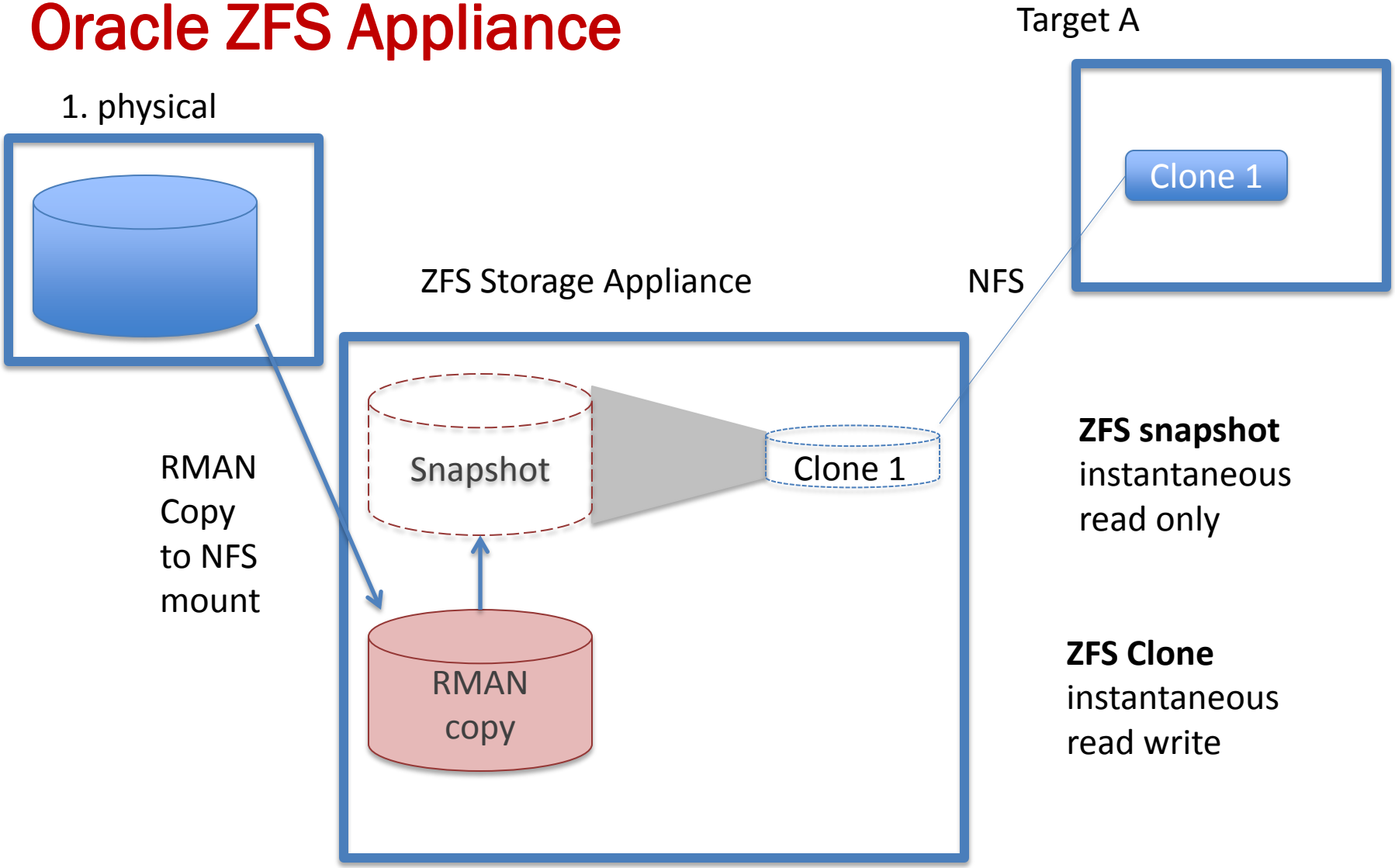


ZFS Appliance

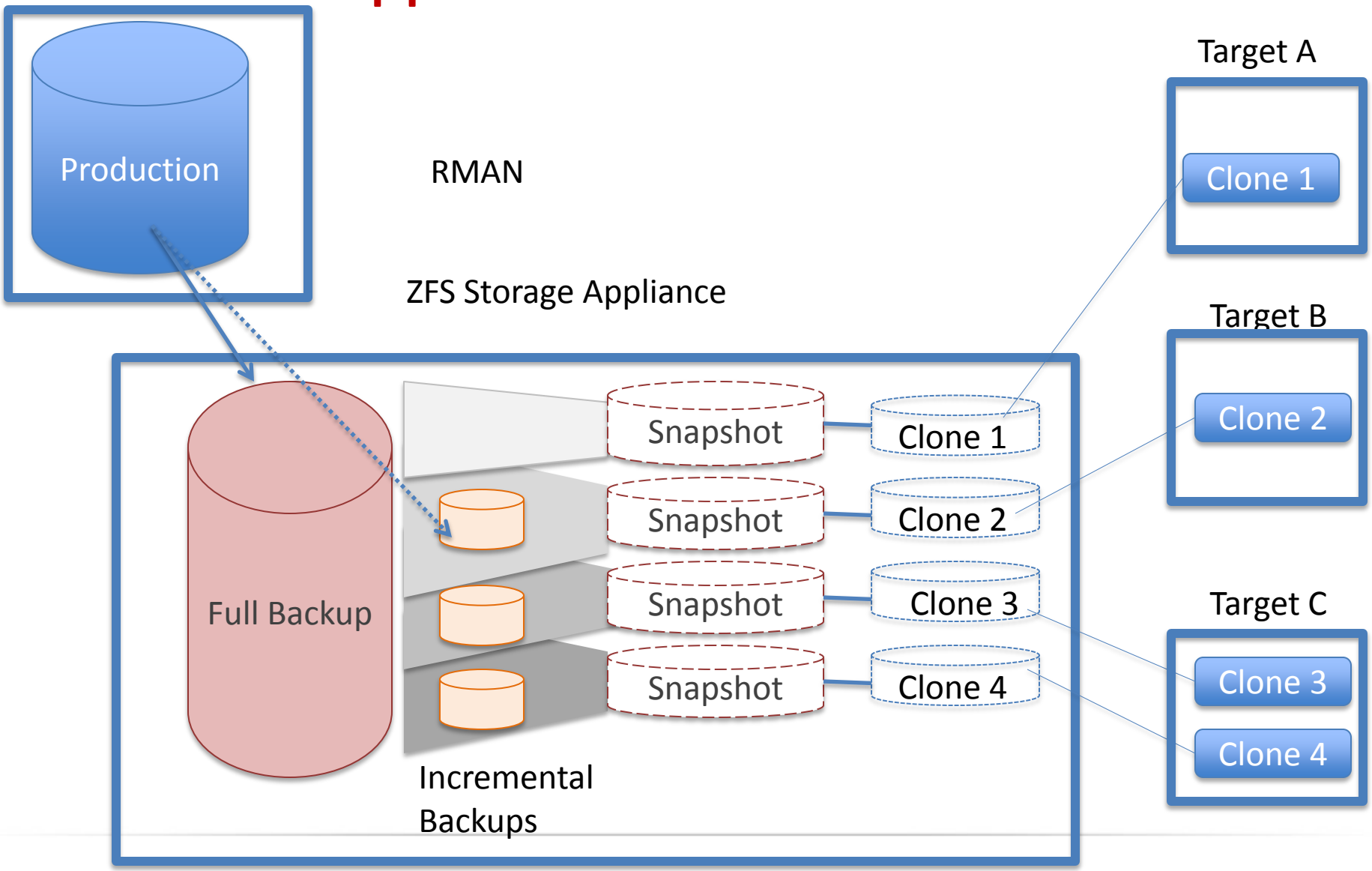
cloning-solution-353626.pdf
44 pages only partial solution

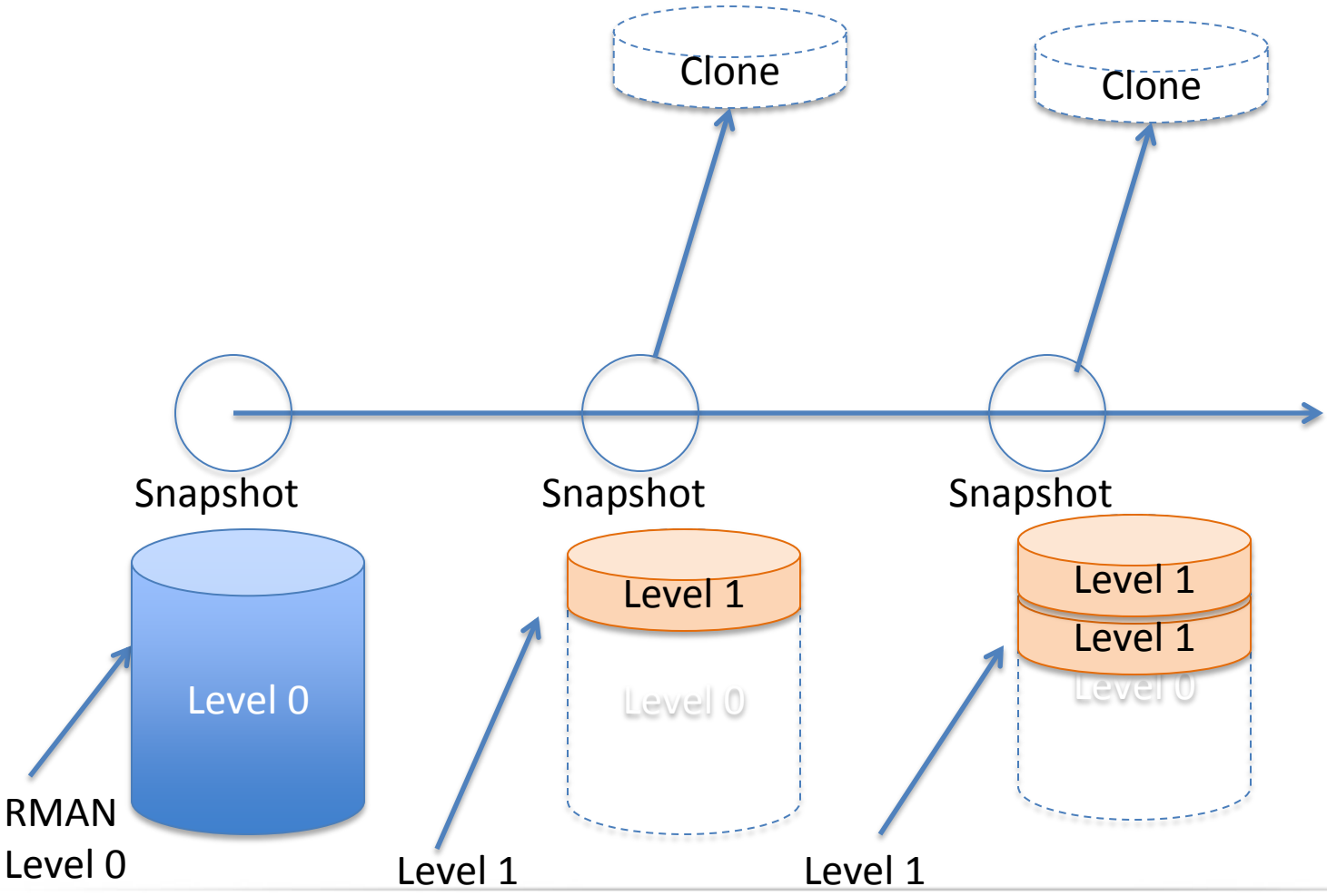
1. ZFS Appliance
   - Create backup project **db_master**
     - With 4 file systems: datafile, redo, archive,alerts
   - Create project for db_clone (with same 4 filesystems)
2. Source Database
   - NFS Mount Backup locations from ZFS Appliance
   - Backup with RMAN as copy, archive logs as well
3. ZFS Appliance
   - Login to Appliance shell, Snapshot backup location
     - Select **db_master**
     - Snapshots snapshot snap_0
     - Then each filesystem on db_master clone it onto db_clone
4. Target Host
   - Mount db_clone directories over NFS from ZFS Appliance
   - Startup and recover clone

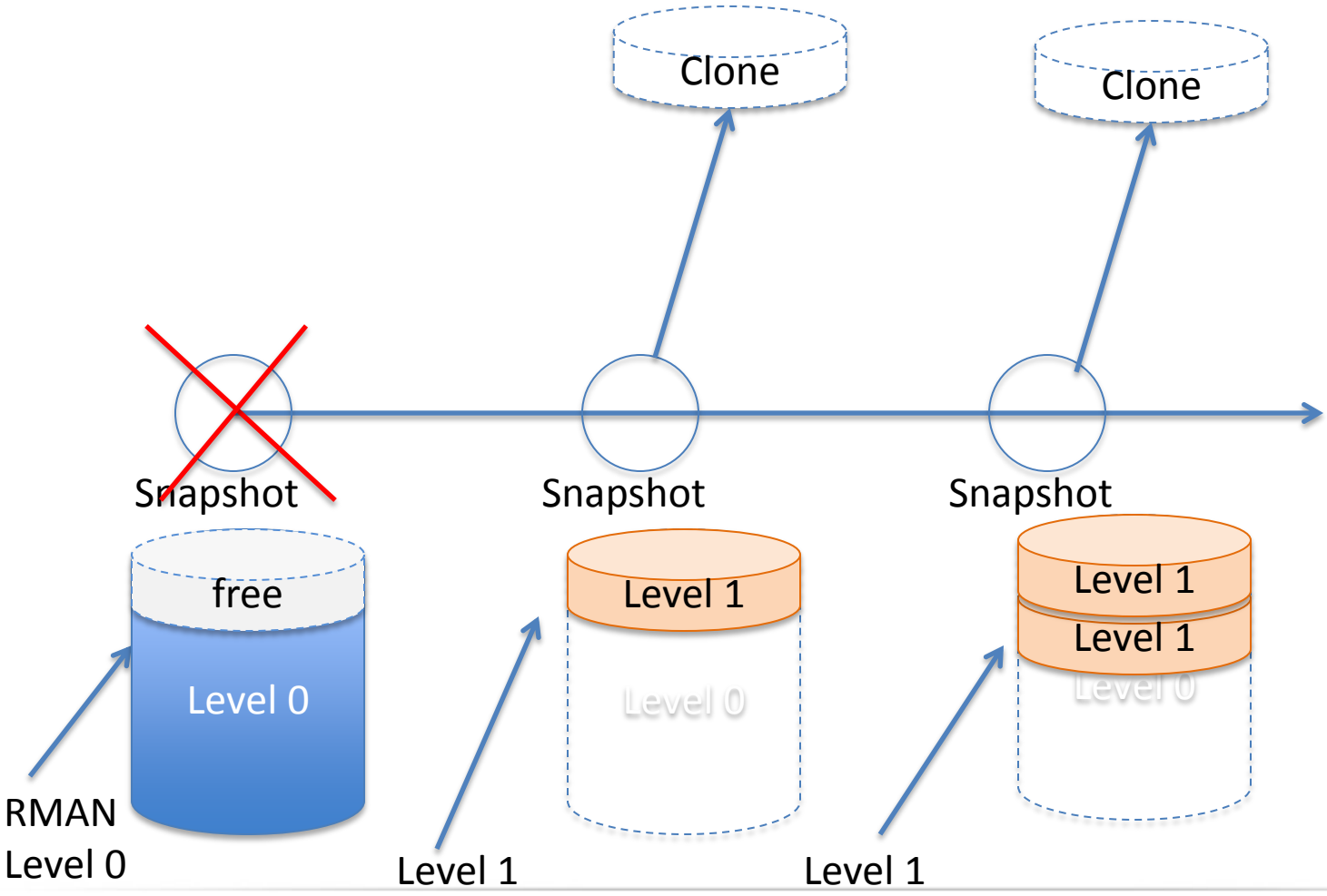
Oracle ZFS Appliance



Oracle ZFS Appliance: RMAN incremental




Clone
Clone
Snapshot
Snapshot
Snapshot
Level 0
Level 1
Level 0
Level 1
Level 1
RMAN
Level 0
Level 1
Level 1


Clone
Clone
Snapshot
Snapshot
Snapshot
free
Level 0
Level 1
Level 0
Level 1
Level 1
RMAN
Level 0
Level 1
Level 1

DELPHIX

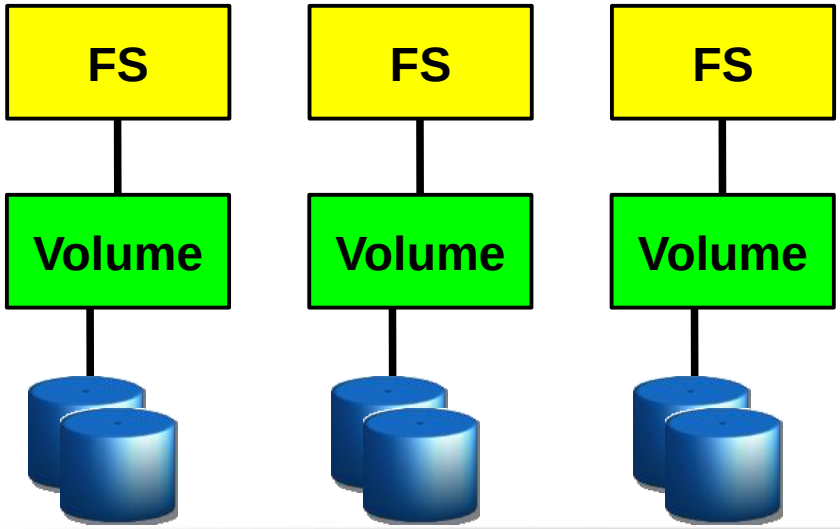
ZFS

- Prehistory: 1 disk = 1 filesystem
- ~1990: volume managers: N disks : 1 FS
- 2001-2005: ZFS development
- **2005: ZFS ships, code open-sourced**
- 2008: ZFS storage appliance ships
  - ZFS enables several ZFS-based startups including Delphix, Nexenta, Joyent,
- 2010: ZFS development moves to Illumos
  - headed by Delphix

FS/Volume Model vs. Pooled Storage

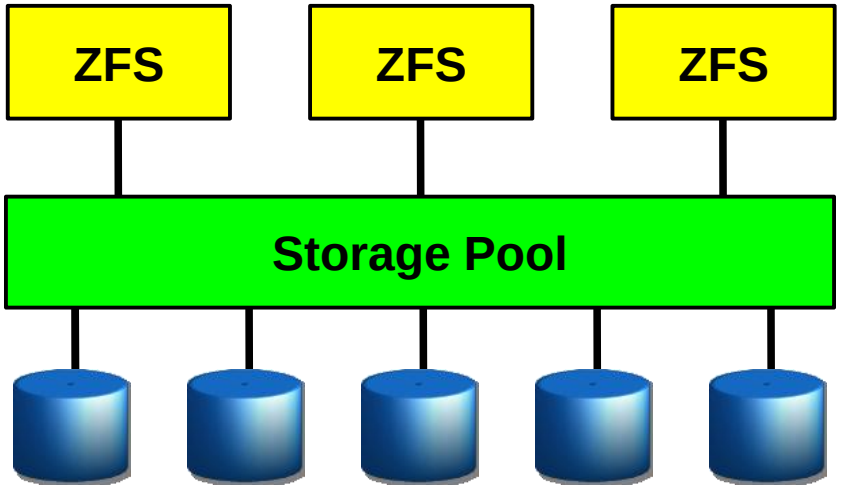
Traditional Volumes

- Abstraction: virtual disk
- Partition/volume for each FS
- Grow/shrink by hand
- Each FS has limited bandwidth
- Storage is fragmented, stranded



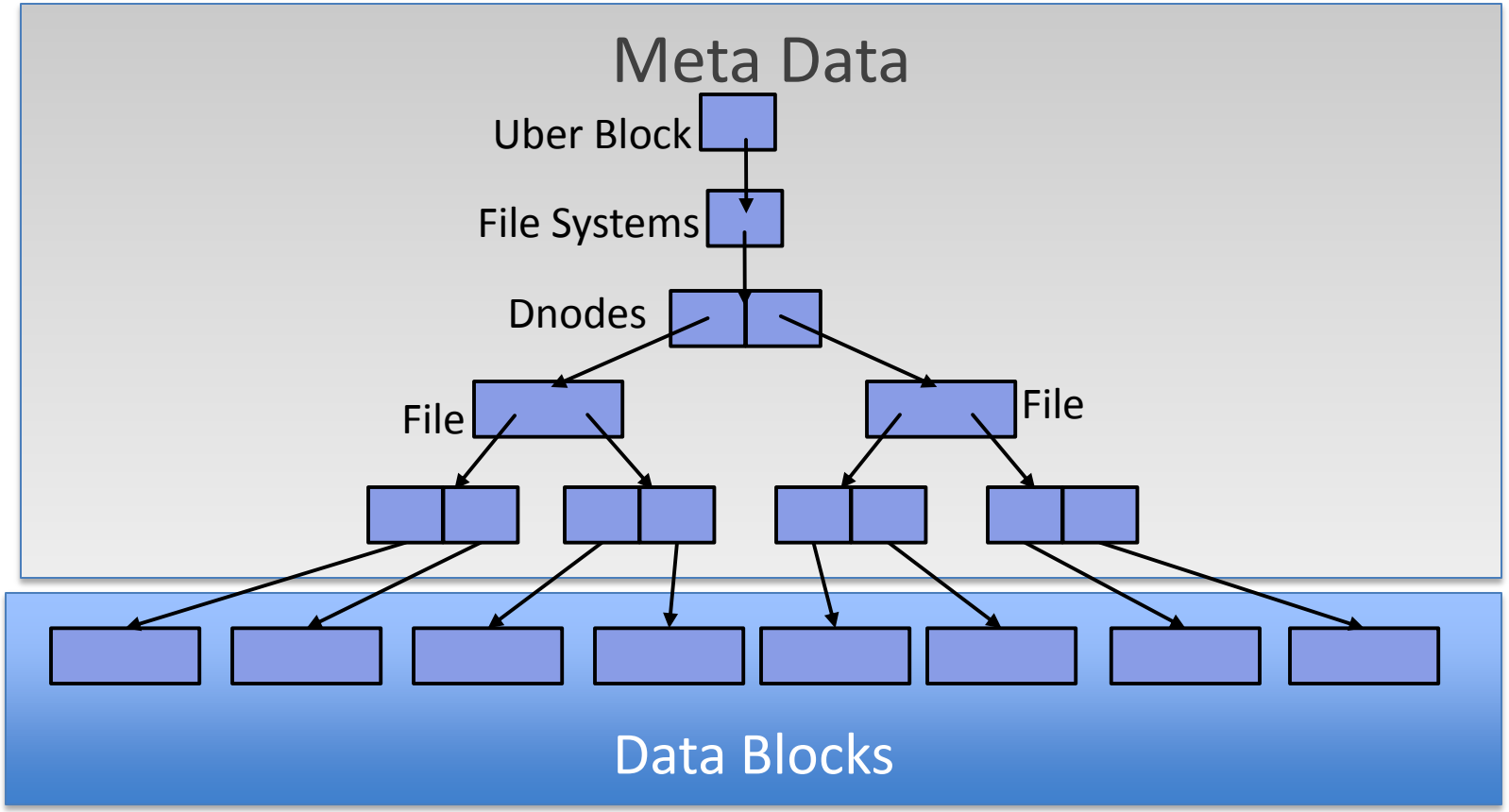
ZFS Pooled Storage

- Many filesystems in one pool
- No partitions to manage
- Grow automatically
- All bandwidth always available
- All storage in the pool is shared



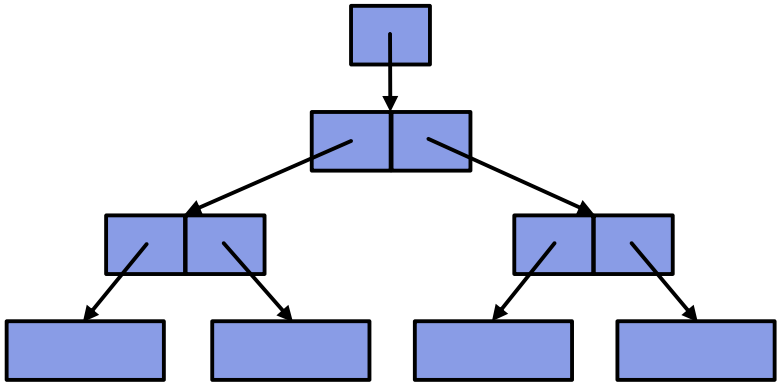
Always consistent on disk (COW)

1. Initial block tree

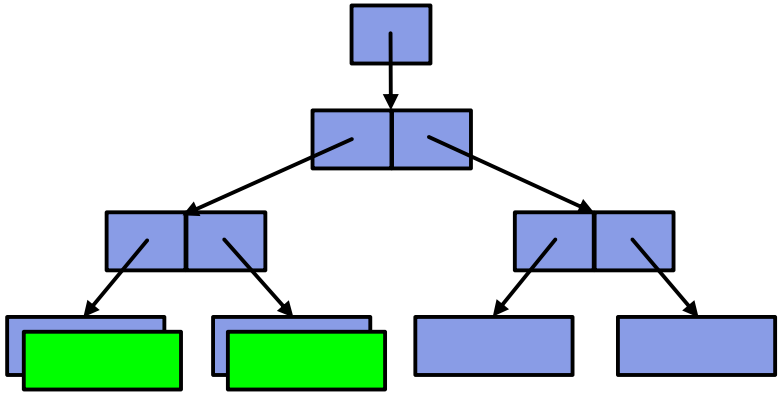


Always consistent on disk (COW)

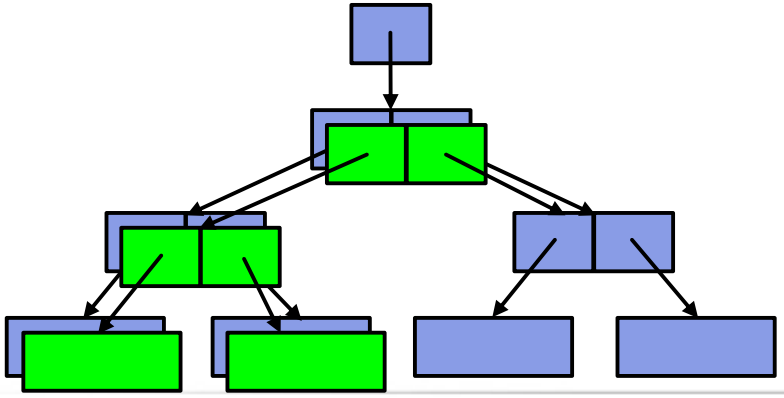
1. Initial block tree



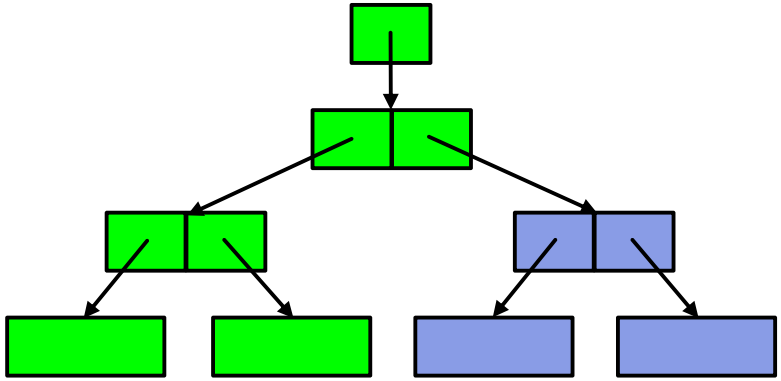
2. COW some blocks



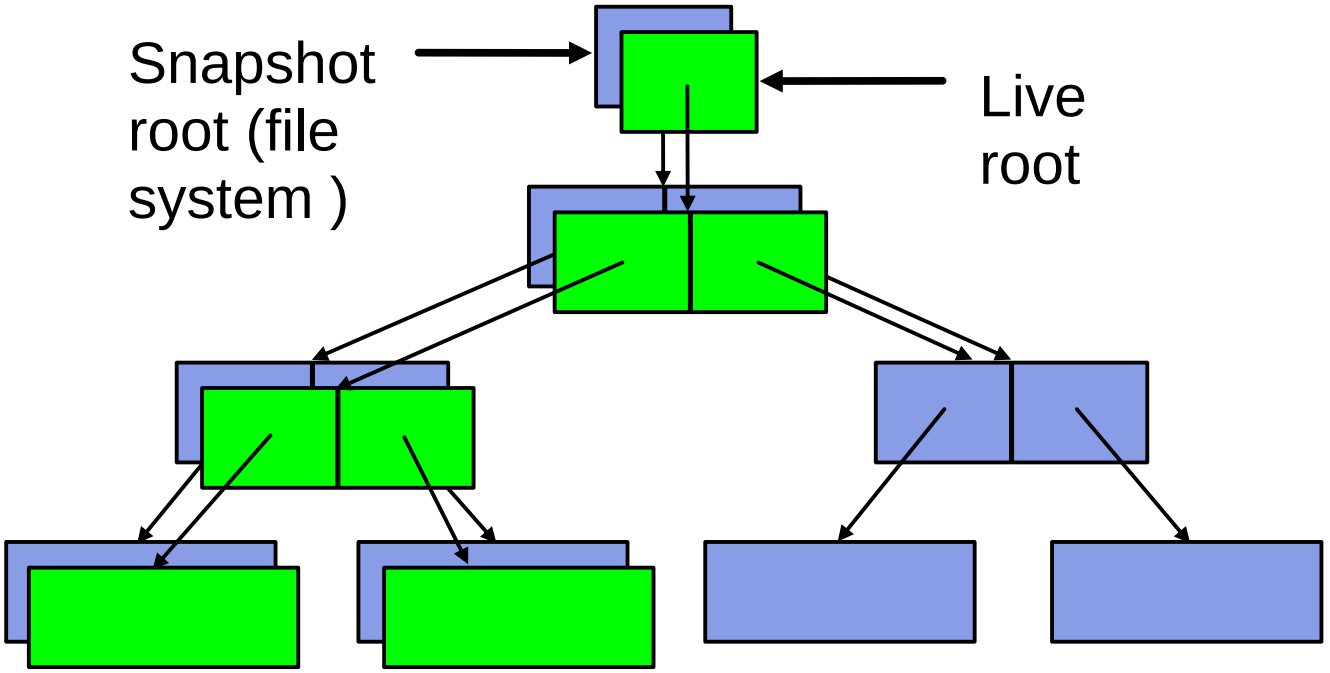
3. COW indirect blocks



4. Rewrite uberblock (atomic)

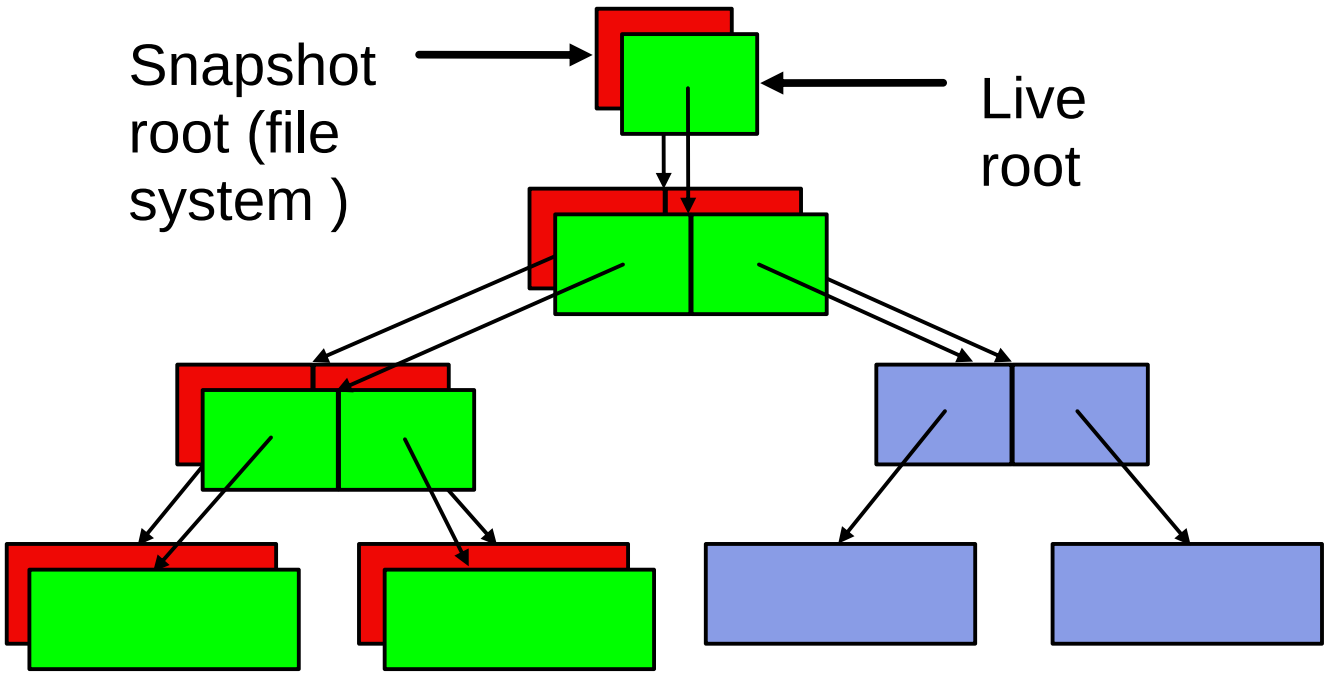


Bonus: Snapshots



Bonus: Constant-Time Snapshots

- Younger snapshots than blocks => keep
- No younger snapshots => free



Sync writes are written immediately out to
Intent log
Data and Metadata
Is batch written out later

ZFS Data Relationships

- Snapshot is a read-only point-in-time copy of a filesystem
  - Instantaneous
  - Unlimited
  - No additional space
- Clone is a writable copy of a snapshot
  - Instantaneous
  - unlimited
  - No additional space
- Send / receive : replication
  - Can send a full snapshot
  - Can send incremental changes between snapshots
  - Incremental send/receive quickly locates modified blocks

ZIL (ZFS Intent Log) Overview

- ZIL is per filesystem
- Logs filesystem modifications
- Log can used to replay filesystem changes
  - In the event of power failure / panic, the log records are replayed
- Log records are stored in memory until :
  - Sync write , ie fsync() or O_DSYNC
  - Transaction group commits

DELPHIX

ZFS at Delphix

- Compression
  - typically ~2-4x
- Block sharing
  - Via clones, Faster , cheaper than Deduplication which is too slow with overhead
- Link Source DB
  - create new filesystems for datafile, archive, etc.
  - set recordsize of datafile FS to match DB
- Snapshot Source
  - take ZFS snapshot of datafile fs
  - retain relevant log files in archive fs
- Clone Provision VDB
  - create clone of Source's datafile snapshot
  - share the dSource's blocks; no additional space used
  - new data takes space

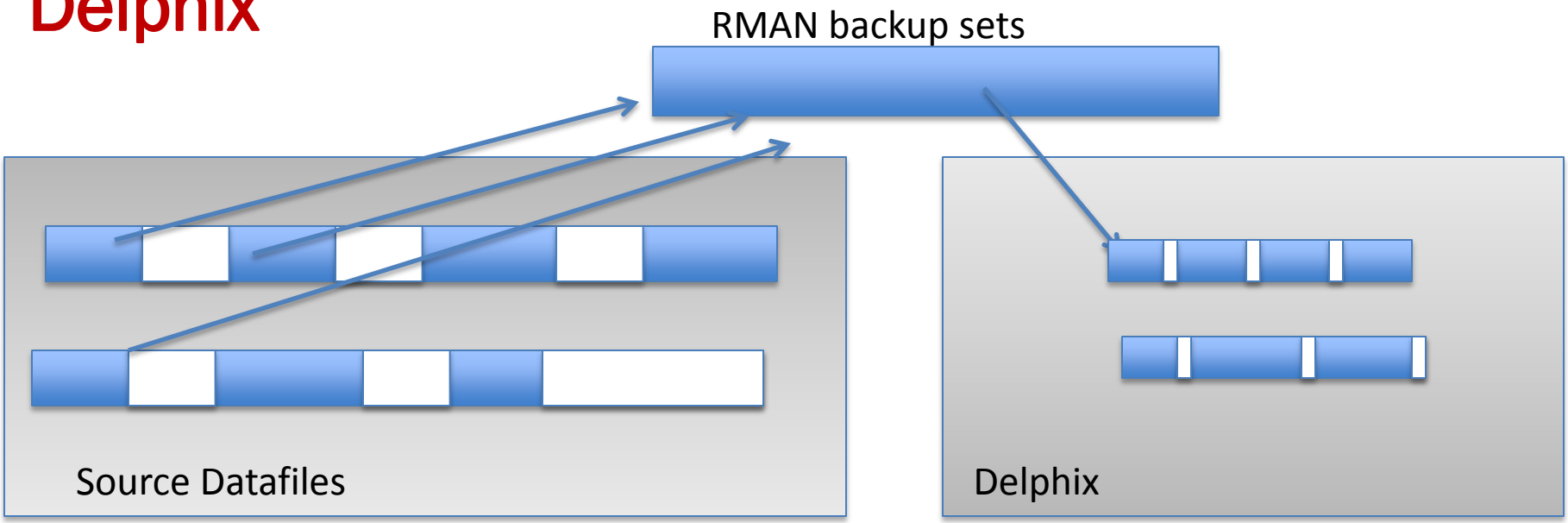
ZFS anti-patterns

- 128K for data blocks
- Full 80%
- Mixed size LUNs, with some full
  - Delphix has improved this with the Delphix appliance
- Scrubs run in middle of business day

ZFS improvements at Delphix

- Single copy ARC
- Multi-threaded space map compression
- NPM mode
- Fast Snap Shot delete 100x

Delphix

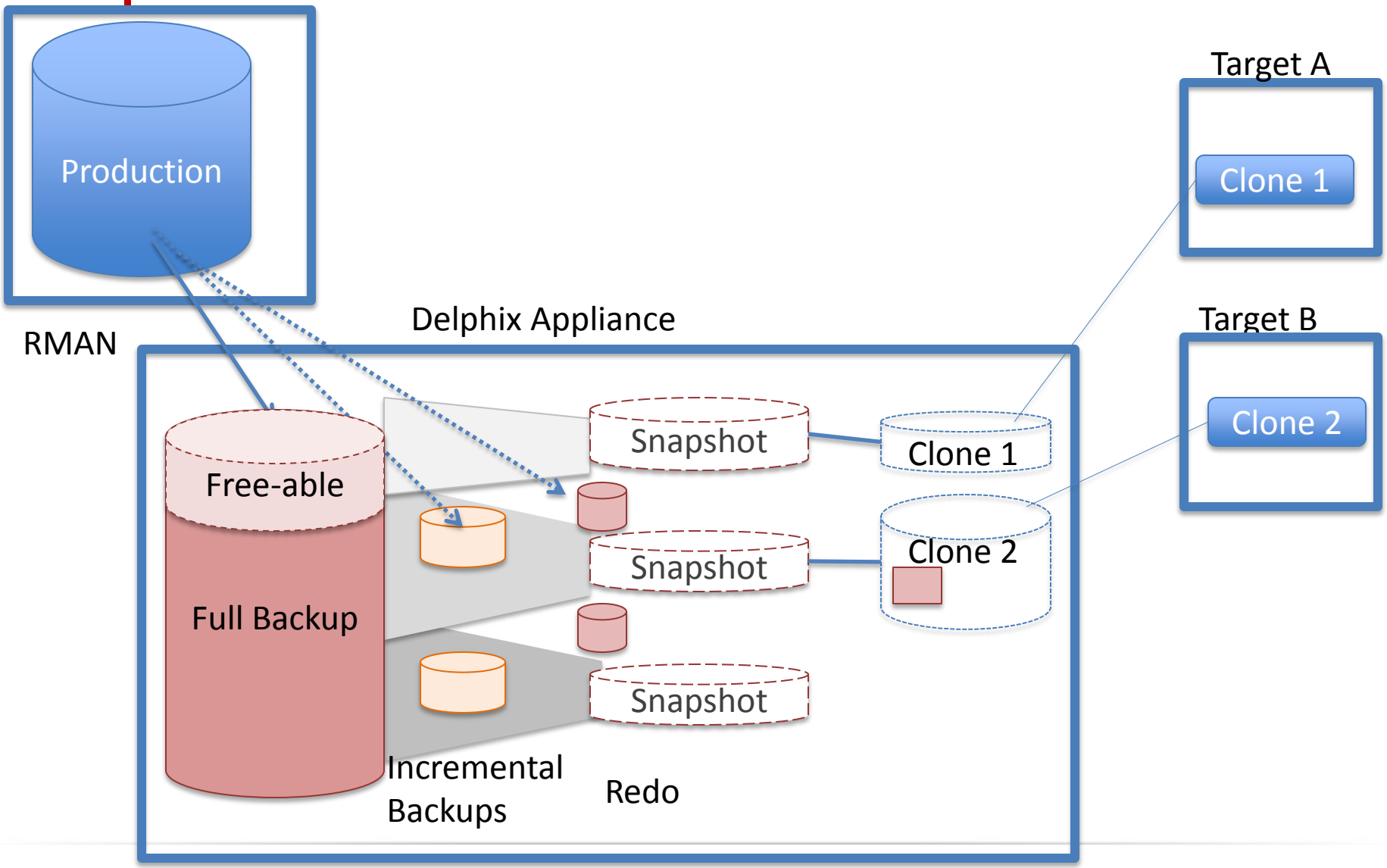


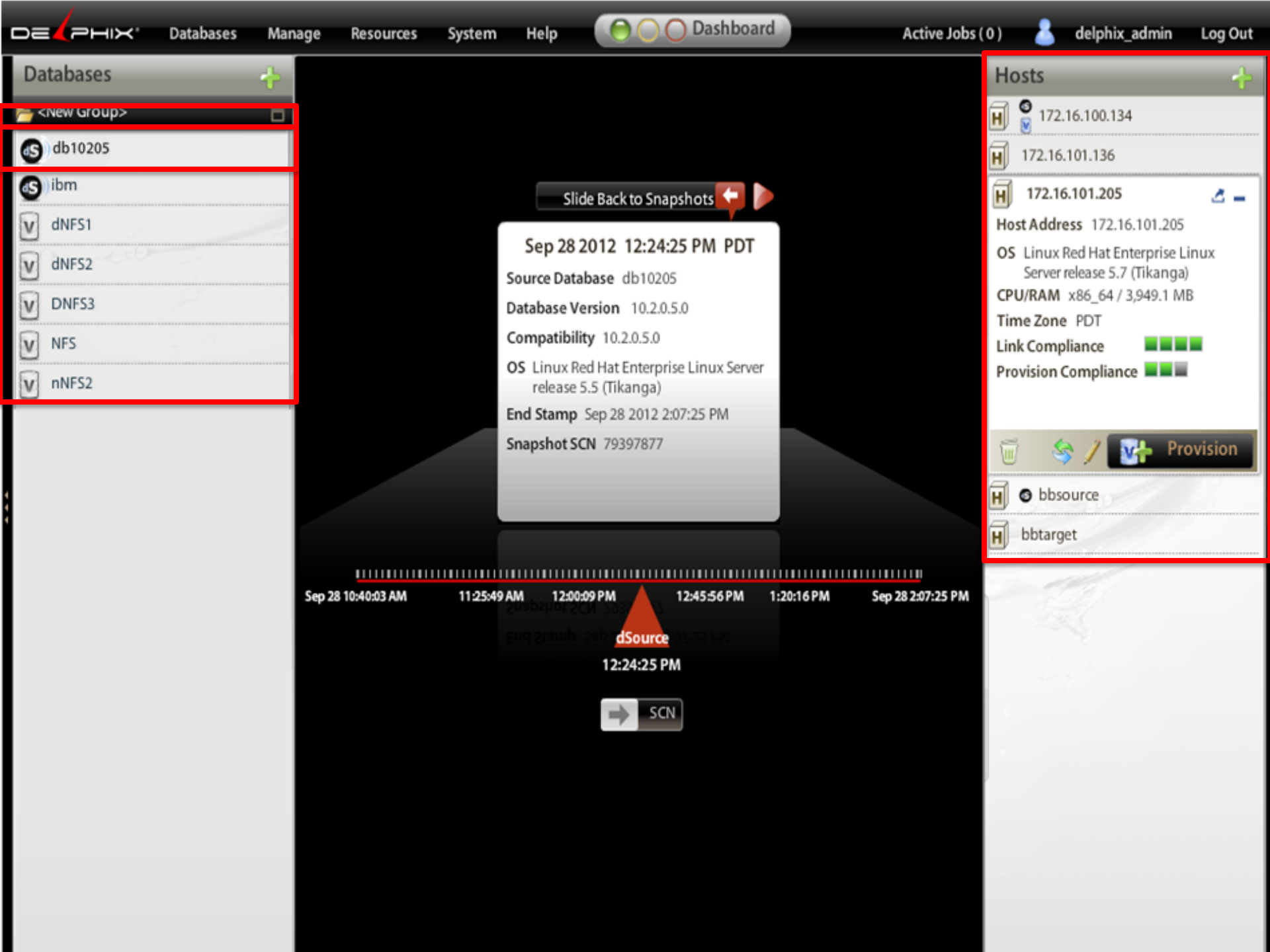
RMAN backup sets
- Allows control over send
- Unused blocks not sent

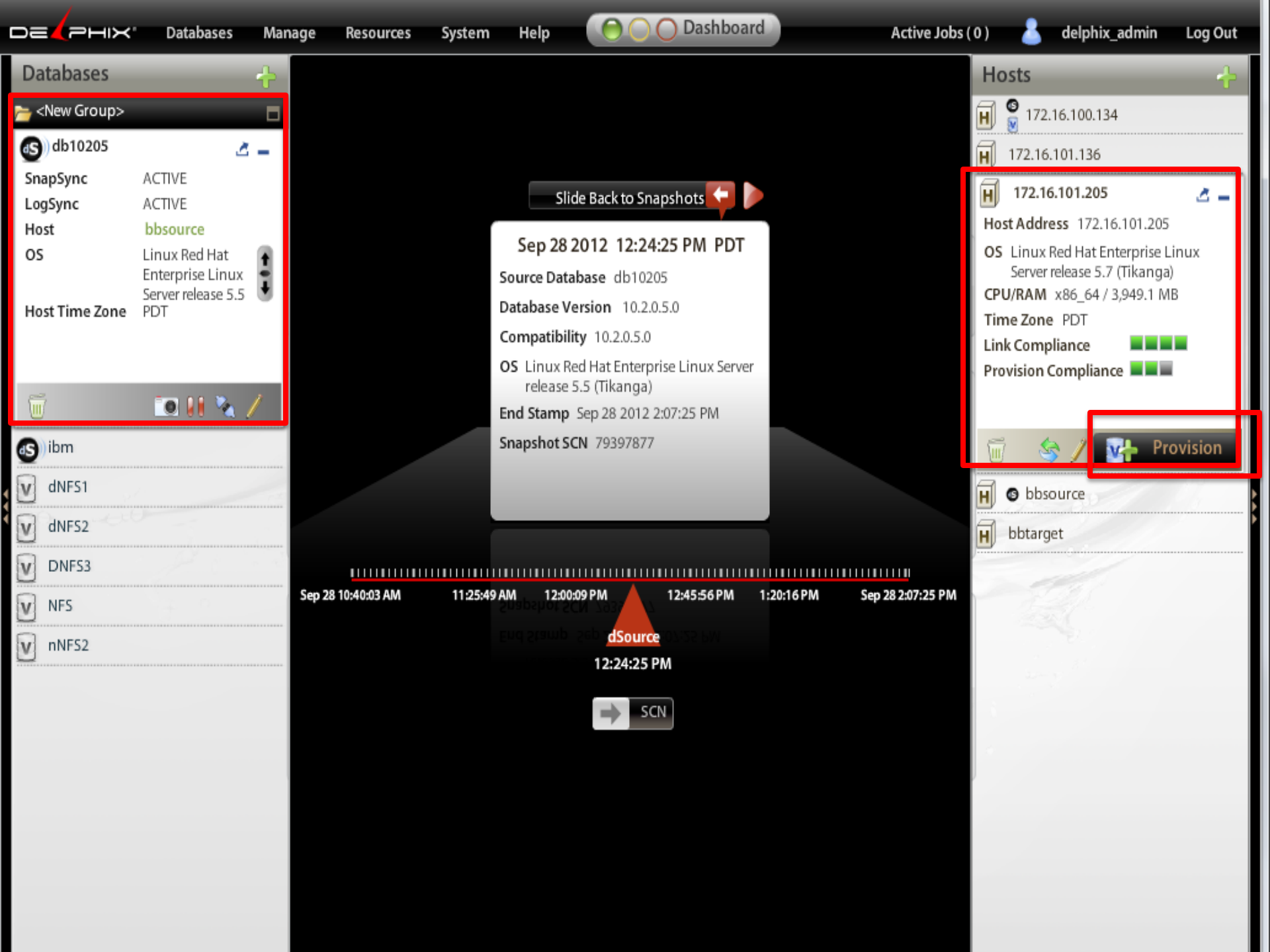
Delphix
- rebuilds the datafiles
- rebuilds unused blocks
- compresses datafiles
- highly compressed zero regions

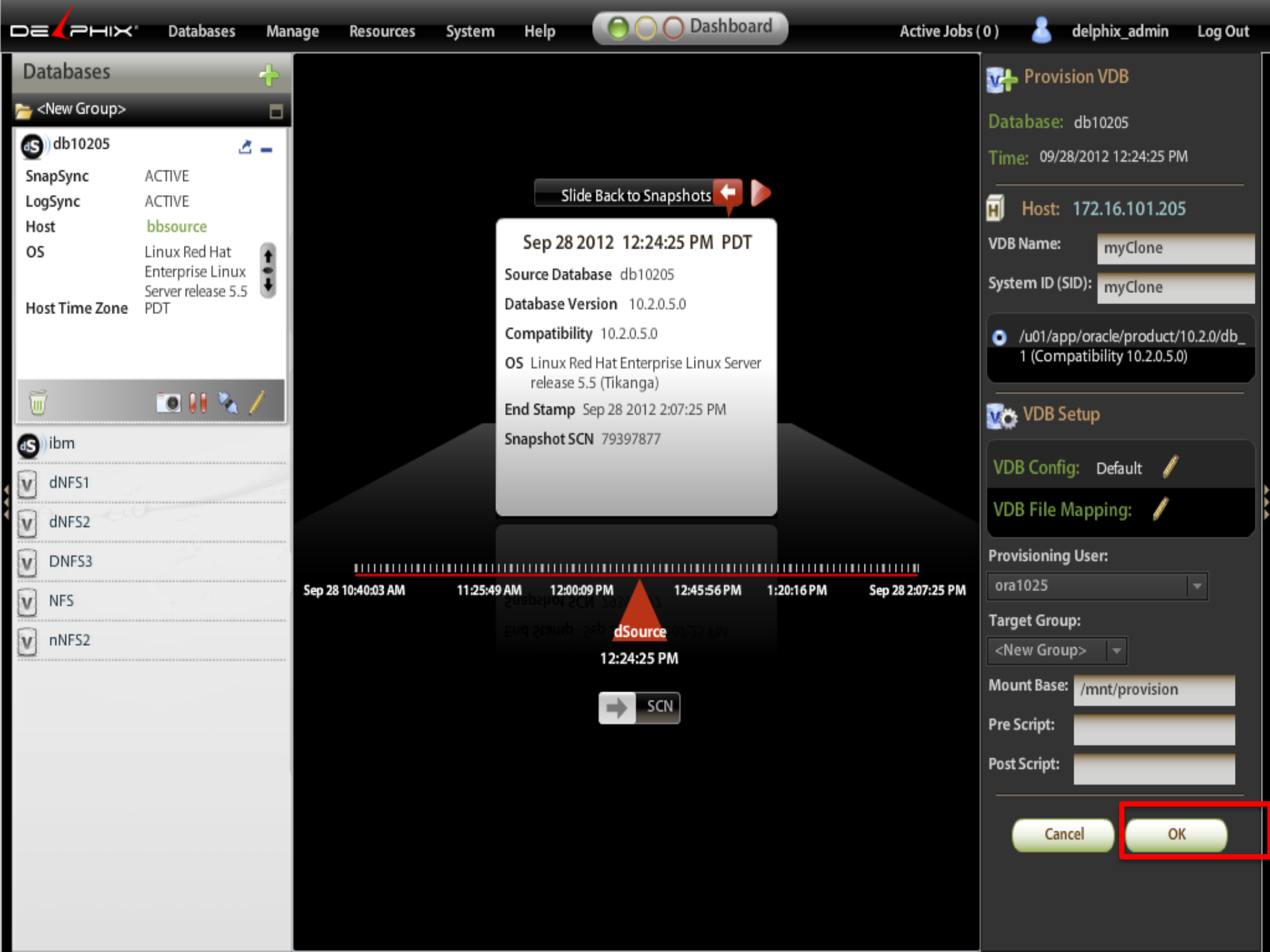
2-4x compression
This analysis shows lzjb compression comes at no performance cost:
http://dtrace.org/blogs/dap/2009/03/16/compression-on-the-sun-storage-7000/

Delphix


Production
RMAN
Delphix Appliance
Free-able
Full Backup
Incremental Backups
Redo
Snapshot
Snapshot
Snapshot
Clone 1
Clone 2
Target A
Clone 1
Target B
Clone 2


DELPHIX
Databases
Manage
Resources
System
Help
Dashboard
Active Jobs (0)
delphix_admin
Log Out
Databases
<New Group>
db10205
ibm
dNFS1
dNFS2
DNFS3
NFS
nNFS2
Slide Back to Snapshots
Sep 28 2012 12:24:25 PM PDT
Source Database db10205
Database Version 10.2.0.5.0
Compatibility 10.2.0.5.0
OS Linux Red Hat Enterprise Linux Server release 5.5 (Tikanga)
End Stamp Sep 28 2012 2:07:25 PM
Snapshot SCN 79397877
Sep 28 10:40:03 AM
11:25:49 AM
12:00:09 PM
12:45:56 PM
1:20:16 PM
Sep 28 2:07:25 PM
dSource
12:24:25 PM
SCN
Hosts
172.16.100.134
172.16.101.136
172.16.101.205
Host Address 172.16.101.205
OS Linux Red Hat Enterprise Linux Server release 5.7 (Tikanga)
CPU/RAM x86_64 / 3,949.1 MB
Time Zone PDT
Link Compliance
Provision Compliance
Provision
bbsource
bbtarget

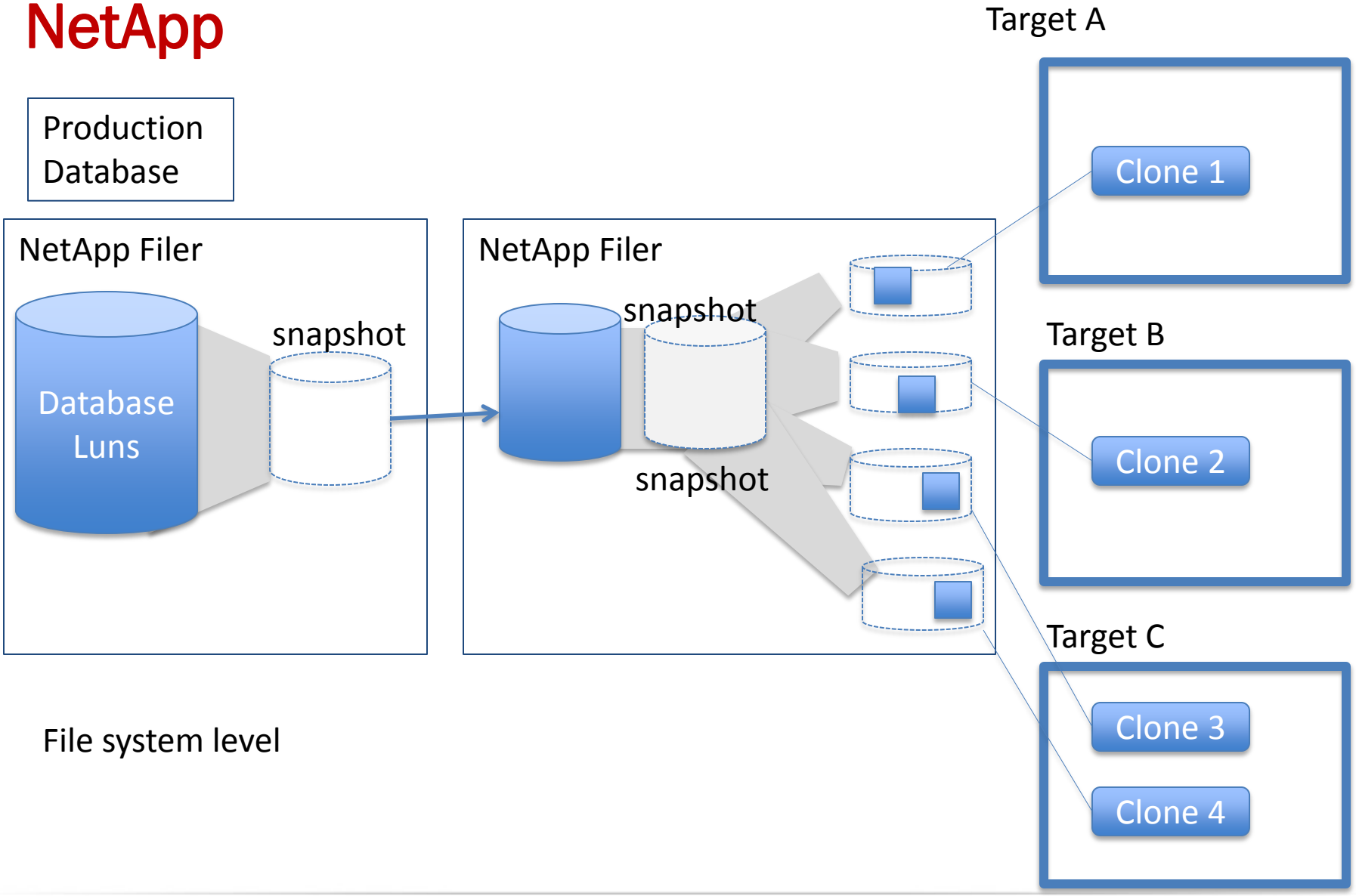

DELPHIX
Databases
Manage
Resources
System
Help
Dashboard
Active Jobs (0)
delphix_admin
Log Out
Databases
<New Group>
db10205
SnapSync ACTIVE
LogSync ACTIVE
Host bbsource
OS Linux Red Hat Enterprise Linux Server release 5.5
Host Time Zone PDT
ibm
dNFS1
dNFS2
DNFS3
NFS
nNFS2
Slide Back to Snapshots
Sep 28 2012 12:24:25 PM PDT
Source Database db10205
Database Version 10.2.0.5.0
Compatibility 10.2.0.5.0
OS Linux Red Hat Enterprise Linux Server release 5.5 (Tikanga)
End Stamp Sep 28 2012 2:07:25 PM
Snapshot SCN 79397877
Sep 28 10:40:03 AM
11:25:49 AM
12:00:09 PM
12:45:56 PM
1:20:16 PM
Sep 28 2:07:25 PM
dSource
12:24:25 PM
SCN
Hosts
172.16.100.134
172.16.101.136
172.16.101.205
Host Address 172.16.101.205
OS Linux Red Hat Enterprise Linux Server release 5.7 (Tikanga)
CPU/RAM x86_64 / 3,949.1 MB
Time Zone PDT
Link Compliance
Provision Compliance
Provision
bbsource
bbtarget


DELPHIX
Databases
Manage
Resources
System
Help
Dashboard
Active Jobs (0)
delphix_admin
Log Out
Databases
<New Group>
db10205
SnapSync ACTIVE
LogSync ACTIVE
Host bbsource
OS Linux Red Hat Enterprise Linux Server release 5.5
Host Time Zone PDT
ibm
dNFS1
dNFS2
DNFS3
NFS
nNFS2
Slide Back to Snapshots
Sep 28 2012 12:24:25 PM PDT
Source Database db10205
Database Version 10.2.0.5.0
Compatibility 10.2.0.5.0
OS Linux Red Hat Enterprise Linux Server release 5.5 (Tikanga)
End Stamp Sep 28 2012 2:07:25 PM
Snapshot SCN 79397877
Sep 28 10:40:03 AM
11:25:49 AM
12:00:09 PM
12:45:56 PM
1:20:16 PM
Sep 28 2:07:25 PM
dSource
12:24:25 PM
SCN
Provision VDB
Database: db10205
Time: 09/28/2012 12:24:25 PM
Host: 172.16.101.205
VDB Name: myClone
System ID (SID): myClone
/u01/app/oracle/product/10.2.0/db_1 (Compatibility 10.2.0.5.0)
VDB Setup
VDB Config: Default
VDB File Mapping:
Provisioning User:
ora1025
Target Group:
<New Group>
Mount Base: /mnt/provision
Pre Script:
Post Script:
Cancel
OK

Data Director : Linked Clones (Vmware)

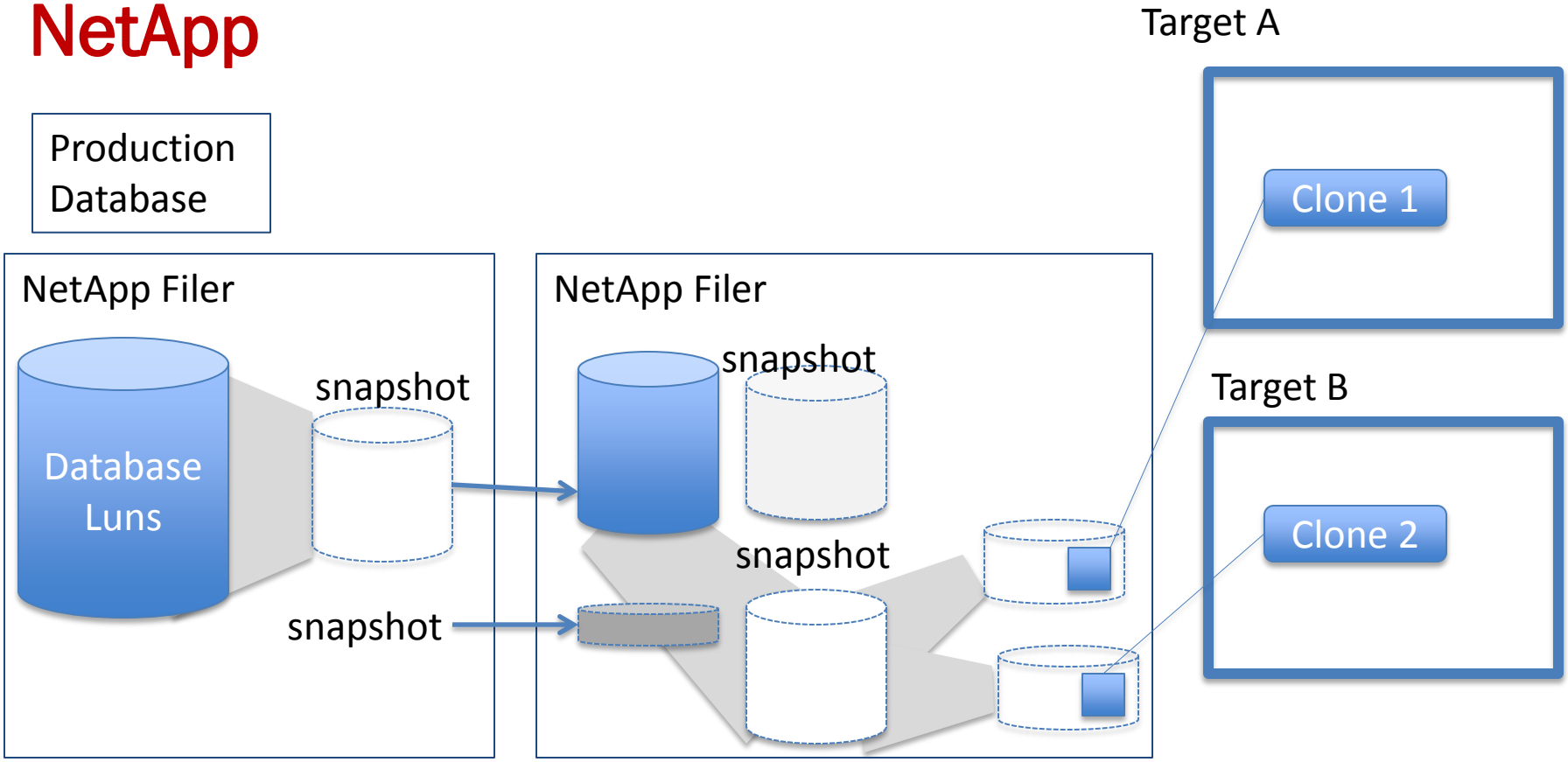
- Performance issues
  - "Having several linked clones can affect the performance of the source database and the performance of the linked clones." http://bit.ly/QOXbyE (on http://pubs.vmware.com)
  - "If you are focused on performance, you should prefer a full clone over a linked clone." http://www.vmware.com/support/ws5/doc/ws_clone_typeofclone.html
  - Performance worse with more snapshots
  - Can only take 16 snapshots
  - Performance worse with more concurrent users
- Golden Copy issue
  - original copy has to always exist
- x86 host databases only
  - Linux
  - OpenSolaris

NetApp

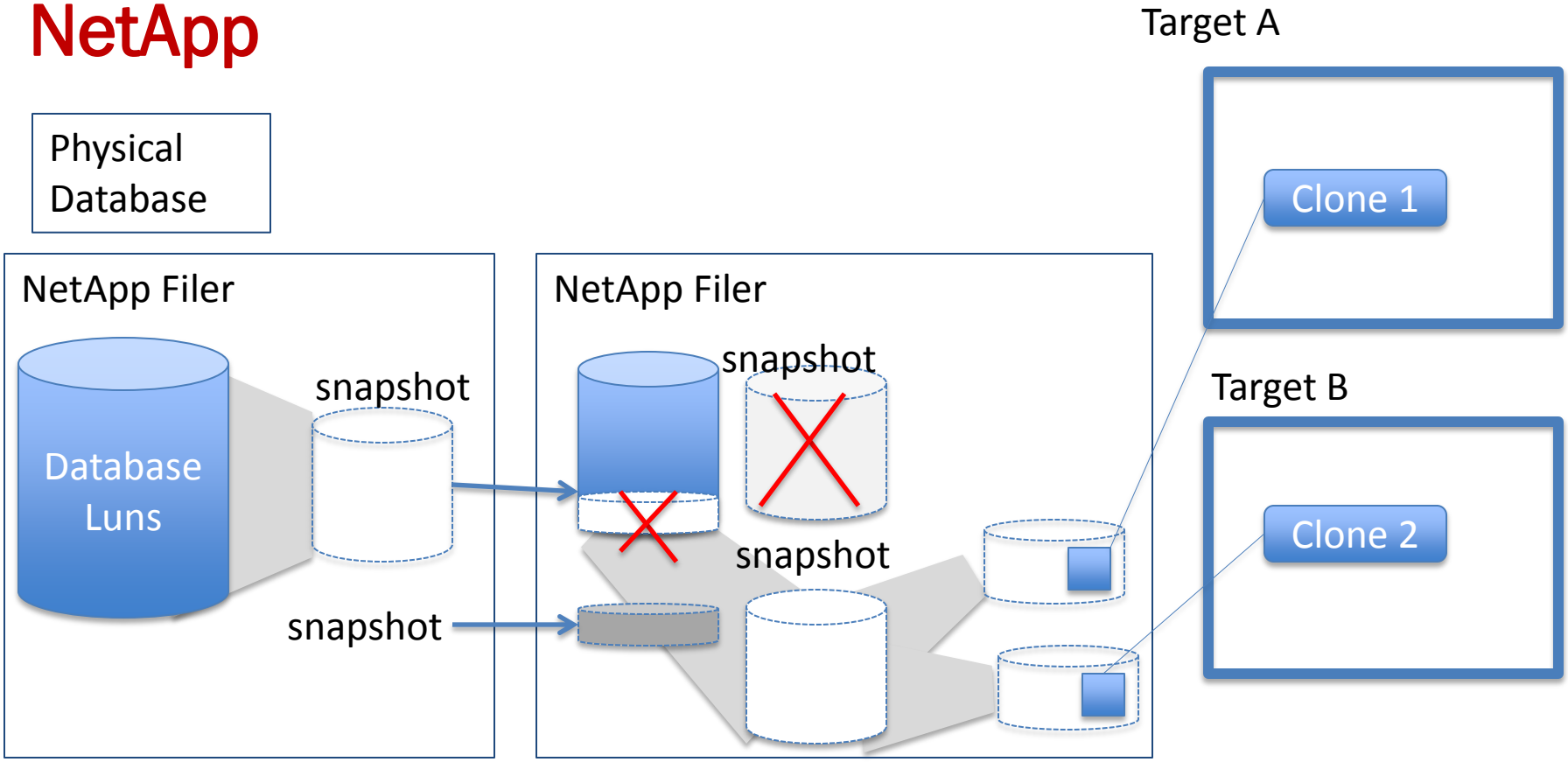


File system level

NetApp


Target A
Production Database
Clone 1
NetApp Filer
snapshot
Database Luns
snapshot
NetApp Filer
snapshot
snapshot
Target B
Clone 2

NetApp


Physical Database
NetApp Filer
Database Luns
snapshot
snapshot
NetApp Filer
snapshot
snapshot
Target A
Clone 1
Target B
Clone 2

NetApp Limits

Limit of 255 snapshots
snaps are limited to the same aggregate (storage pool)
Aggregates have size limits depending on controller

Controller	Size Limit
32 bit controllers	16TB
FAS3140/FAS3040/FAS3050	40TB
FAS3160/FAS3070	50TB
FAS6040/FAS3170	70TB
FAS6080	100TB

All sources have to be in the same aggregate to be snapshot together.

EMC

- Point of view: DR , backup and testing off of a full copy
  - Create BCV , a full copy (
  - Promote BCV to make accessible
  - Take snaps of BCV (limit 32?)
  - Zone and mask LUN to target host
  - Full copy of disk, now recover (may have to rename the LUNs)
- “Golden Copy”
  - Not a pointer based file system like NetApp and ZFS
  - EMC uses a save area, the amount of area for changes to the snapshot
  - No time flow
  - Ie initial snapshot has to stay
  - To get rid of “golden copy” have to recreated it with the new changes
- Snapshots
  - Can’t take a snap of a snap on low end
  - Can only take one level snap of a snap on high end

Oracle 12c

- Oracle Snap Manager Utility for ZFS Appliance
- Pay for option
- Requires source database hosted on ZFS appliance
- Principally a manual GUI
  - utility to snapshot source databases and provision virtual databases
- No concept of time flow
  - Virtual databases have to be provisioned of snapshots.

Conclusion

- EMC Timefinder, VMware Data Director
  - offer limited ability to benefit from cloning
- Clonedb
  - fast easy way to create many clones of the same copy
  - limited to 11.2.0.2 and systems with sparse file system capability
  - suffers the golden image problem
- NetApp Flexclone, Snap Manager for Oracle
  - offers a rolling solution
  - limited database awareness
  - file system clones
  - limited snapshots
  - Vendor lock-in
- Oracle ZFS Appliance
  - Vendor Lock-in
- Delphix
  - Agility : Automation, unlimited snapshots, clones, multi-database

DELPHIX

Matrix of features

	CloneDB	ZFS Appliance	Delphix	Data Director	NetApp	EMC
Time Flow	No	Yes	Yes	No	Yes	No
Hardware Agnostic	Yes	No	Yes	Yes	No	No
Snapshots	No	Unlimited	Unlimited	31	255	16 (96 read only)
Snapshots of snapshots	No	Unlimited	Unlimited	30	255	1
Automated Snapshots	No	No	Yes	No	Yes	No
Automated Provisioning	No	No	Yes	No	No	No
Any DB host O/S	Yes	Yes	Yes	No x86 only	Yes	Yes
Max DB size	None	None	None	~200G	16-100TB	?

Appendix

- CloneDB
  - http://www.oracle-base.com/articles/11g/clonedb-11gr2.php
- ZFS
  - http://hub.opensolaris.org/bin/download/Community+Group+zfs/docs/zfslast.pdf
- ZFS Appliance
  - http://www.oracle.com/technetwork/articles/systems-hardware-architecture/cloning-solution-353626.pdf
- Data Director
  - http://www.virtuallyghetto.com/2012/04/scripts-to-extract-vcloud-director.html
  - http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=1015180
- EMC
  - https://community.emc.com/servlet/JiveServlet/previewBody/11789-102-1-45992/h8728-snapsure-oracle-dnfs-wp.pdf
- NetApp
  - RAC provision example http://blog.flimatech.com/2008/02/07/how-to-create-a-netapp-flexclone-rac-database/
  - http://media.netapp.com/documents/snapmanager-oracle.pdf basic info

- END