Oracle Solaris 11.4
ZFS Data Management

Cindy Swearingen, Principal Product Manager
ZFS Storage Product Management
February 2018

Oracle Confidential - Internal

Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Solaris 11.4 Data Management Feature Summary

Data Protocols and Services

- NFSv4.1 features
- SMB 2, 2.1, and 3.0 features
- Shadow migration improvements
- Deduplication v2.0
- Deduplicated send streams
- File level copy (cp –z)
- Compressed send streams
- Resumable replication
- Data privacy management

Performance

- Asynchronous destroy
- Meta device (DDT SSDs)
- More efficient RAIDZ space allocation
- Reduce resilver restarts
- Format fast startup

Management and Visibility

- Storage analytics
- Pool reGUID
- Device removal
- Read/write limits
- fsstat I/O latency
- dtrace file ops, iSCSI provider
- Mirrored device selection
- zpool get options
- zpool label subcommand
- ZFS storage management authorization

ORACLE

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal

Data Protocols and Services

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential - Internal

NFSv4.1 Server Side Improvements

Improves availability

- Exactly Once Semantics (EOS) – Provides reliable duplicate request cache to ensure client requests are executed once and only once.
- reclaim_complete – Allow NFS servers to resume service faster by ending grace period after all clients have recovered. Particularly important for high-availability environments.
- Planned GRACE-less Recovery (PGR) – Allows NFS 4/4.1 server to preserve state info across service restarts or graceful reboot, so the server doesn't enter GRACE period to recover state. NFS client can avoid a 90-sec downtime during service restarts and graceful system reboots.

SMB 2.0/2.1 Performance Improvements

- SMB 2.1 improvements for high-speed networks
  - SMB payload requests scale to 1MB instead of 64K
    - Allows better utilization of network bandwidth on high speed networks
    - Reduces CPU utilization on the server and client
  - Clients gain performance benefit of not losing local caching when the same file is opened multiple times

SMB 3.0 Enhancements

- AES-CMAC algorithm replaces SMB 2.0's HMAC-SHA256
- Persistent handles/transparent failover - Clients can recover from a server failure event by restoring saved file state from stable storage
- Multichannel support - Allows user session to associate with multiple TCP connections, taking advantage of all available network bandwidth and provides network fault tolerance when multiple paths are available between client and server.

Data Reduction Technology

Reduction Technology	S11.4
Compression (LZ4)	X
Compression (LZJB)	X
Compression (GZIP2)	X
Compression (GZIP)	X
Compression (GZIP9)	X
Deduplication 2.0	X

15X-20X
Data Reduction

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal

Deduplication DDT2 Improvements

- Deletion improvements
  Dynamic control of deduplicated frees to bite-sized chunks
- Reduce the number of metadata tables from 3 to 1
  Removes time/complexity to migrate entries back and forth
- Enable table to shrink as data is removed
- Reducing the average size of entries
- Improved hash table algorithm

SIZE IMPROVEMENT

	DDT	DDT2	Improvement
On Disk (Bytes / Entry)	620	127	4.9x
In Core (Bytes / Entry)	377	94	4.0x

SPEED IMPROVEMENT

	DDT	DDT2	Improvement
1st Copy After Reboot	501s	107s	4.7x
2nd Copy	300s	89s	3.4x

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal

Deduplication 2.0

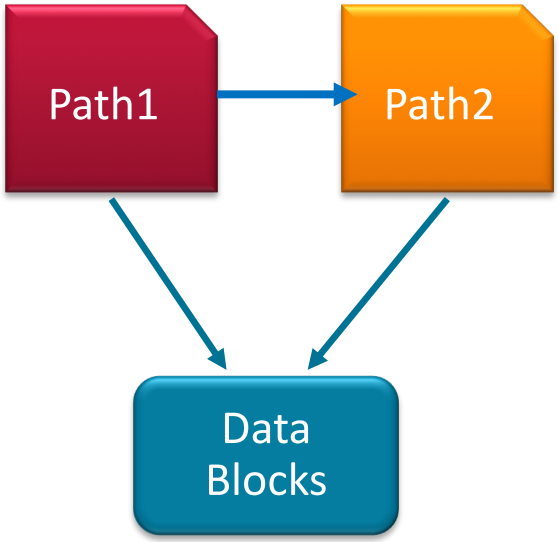
Best practices and lessons learned

- For backup use cases, not production workloads
- Leverage both dedup and LZ4 for best results
- DDT (SSDs) improve dedup performance
- Best ratios with large recordsize (128K to 1MB)
- S11 systems running dedupv1 automatically migrate to dedupv2 during S11.4 upgrade
- LUNs with 8K recordsize are not good candidates for dedupv2

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal

Space Efficient Data Copy (cp -z)

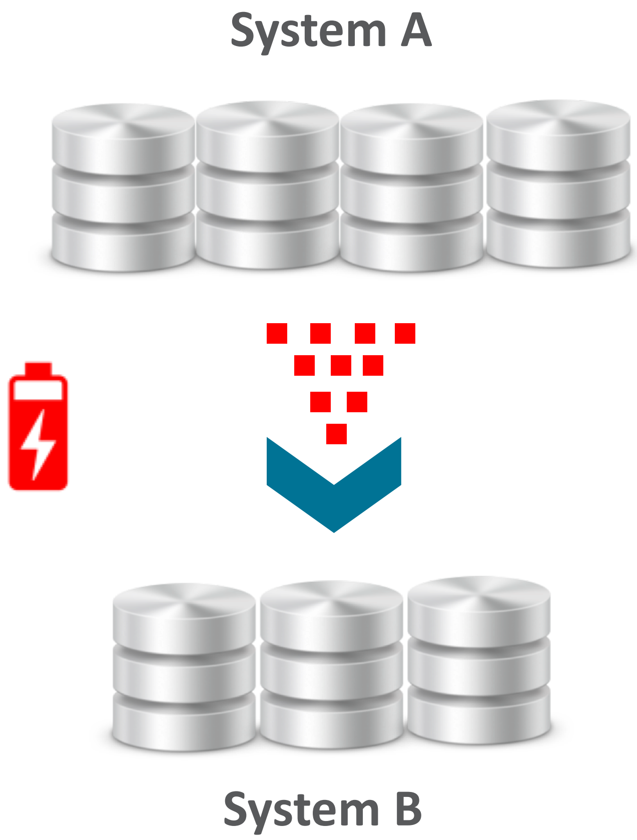
- Similar to file-level snapshots providing an efficient copy mechanism for VM environments
- Use less system resources and disk space
- Path1 overwrites Path2 with contents of Path1 but references same blocks rather than duplicating Path1 blocks (reflink)
  - Uses deduplication for initial copy of the file (xcopy)
  - Both files must be in the same ZFS pool



Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal

Raw Send Streams and Resumable Replication

- Raw send streams
  - Send streams are sent in raw format so that compressed data remains compressed
  - Uses less CPU and network bandwidth
- Resumable replication
  - Send operations can be time consuming
  - If a send stream is interrupted due to network transmission problems, it had to be restarted
  - If interrupted, specify that the send stream resumes where it left off



Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal

Raw Send Stream Syntax

- Specify `–w` option to indicate that a stream is sent in raw format
- Keep compressed data compressed when sending as a stream
- Encrypted data is not yet supported in the raw format

```
# zfs create -o compression=on
tank/compfs
# cp /usr/dict/words /tank/compfs
# zfs snapshot tank/compfs@snap
# zfs send tank/compfs@snap1 >
/tmp/snapstream
# zfs send -w compress
tank/compfs@snap >
/tmp/rsnapstream
# ls -lh /tmp/*stream
-rw-r--r-- 1 root root 155K Mar  9
14:14
/tmp/rsnapstream
-rw-r--r-- 1 root root 304K Mar  9
14:13
/tmp/snapstream
```

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Resumable Replication Syntax

- The `zfs send` and `zfs receive` commands modified to write and read checkpoints that identify snapshot data that is already received
- Indicate that a ZFS send stream should resume where it left off by specifying the `–C` option, if interrupted
- Uses less system resources to resume that to start over from scratch
- Display resumable file system information

```
# ssh sysA zfs receive –C pool | zfs
send
–RC tank/fs@snap1 | ssh sysA zfs
receive –F pool

# zfs list -HI resumable -o name
NAME      USED  AVAIL REFER TYPE
STATE
tank/fs   33.8M 120G  33.8M filesystem
resumable
```

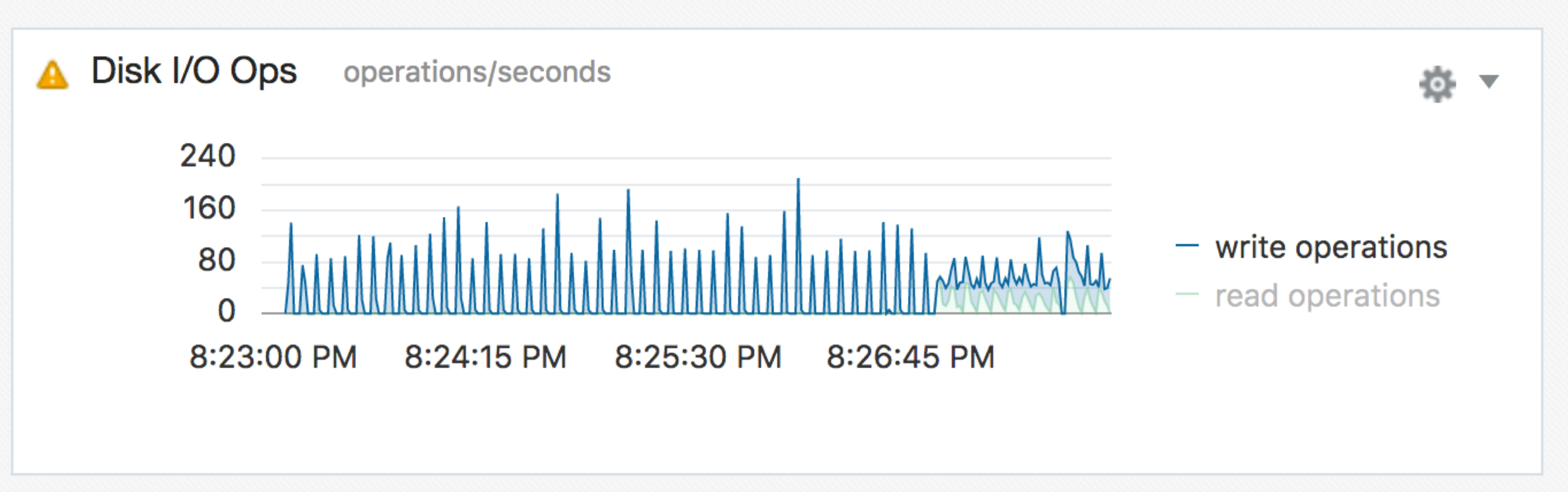
Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Pool Management and Visibility

Storage Analytics

Guided disk I/O diagnostics with Solaris Analytics

- Visualize disk storage I/O
- Quick drill down on areas of concern
- Analysis can be saved for further analysis



fsstat I/O Latency Visibility

- New option identifies read time and write time latency statistics
- Identifies latency info for read, write, and readdir
- Provides visibility into I/O latency to and from applications

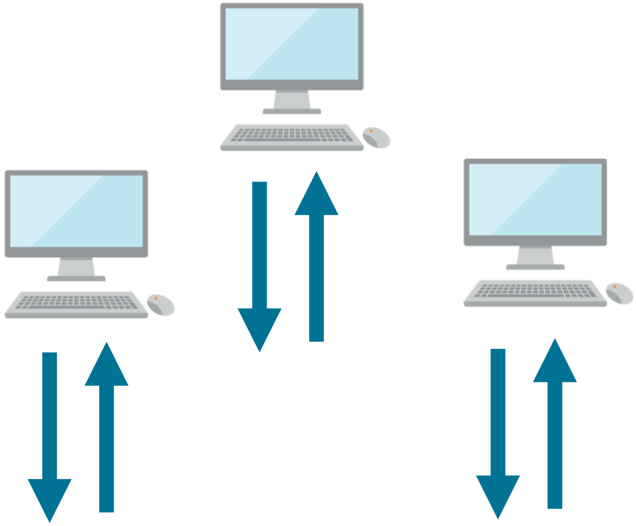
```
# fsstat -l zfs 1
 read read   read write write write rddir rddir rddir
  ops bytes  time   ops bytes  time   ops bytes  time
6.04M  178G 5.00n 1.15G  142T    4n 98.2K 70.2M 1.03u zfs
   27  252K 6.00n     0     0    0n     0     0    0n zfs
    0     0    0n   708 88.5M 21.0n     0     0    0n zfs
    0     0    0n 4.66K  596M    1n     0     0    0n zfs
    0     0    0n     0     0    0n     0     0    0n zfs
```

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal

Setting Read/Write Limits

I/O efficiency and management

- Balance I/O resources by setting read/write limits
  - Limit the rate in bytes/second at which a dataset and its descendents are read or written to.
  - Set a default limit for a dataset in bytes per second at which a dataset descendents are read or written to.
- Bandwidth is not guaranteed and the actual bandwidth may be limited by other factors including usage and limits set on other datasets



Copyright © 2017, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential - Internal

Pool Device Management

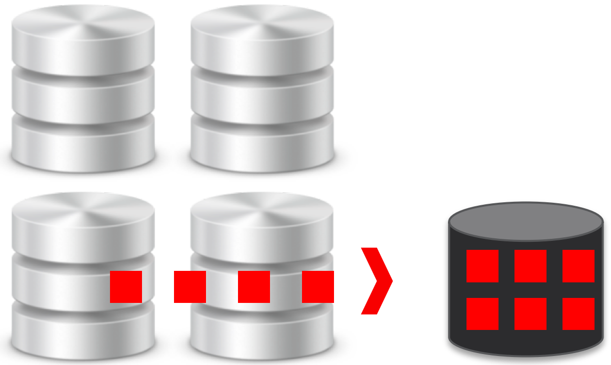
Increase mobility and agility

- Pool device removal
  - Reduce existing pool capacity
  - Remove an existing vdev from a mirrored pool
  - Resolve error when device is mistakenly added rather than attached

Pool Device Removal

Technical Details

- Removes mirror and raidz VDEVs and errant devices
- Physical top-level device is replaced with a pseudo device that lives within the pool
- Pseudo device handles only reads after replacement
- Data from removed physical device is redistributed among remaining devices at the removal time



Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal

Pool Device Flexibility

Pool device removal

- Monitor the progress of a remove operation until the resilvering completes to remaining devices
- Device is removed from the pool configuration and can be reused

```
# zpool remove tank mirror-1
# zpool status tank
.
.
.
NAME        STATE          READ  WRITE CKSUM
tank          DEGRADED        0      0      0
mirror-0       ONLINE         0      0      0
c1t1d0         ONLINE         0      0      0
c2t2d0         ONLINE         0      0      0
mirror-1       ONLINE         0      0      0
(removing)
c1t2d0         ONLINE         0      0      0
c2t2d0         ONLINE         0      0      0
```

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Scheduling Pool Scrubs

Improves routine pool maintenance

- More easily schedule routine pool scrubs
- Provide a scrubinterval time in days, weeks, or months
- Default value is 30 days
- If scrubinterval is set to manual, this feature is disabled
- Read-only lastscrub property identifies the last scrub

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

ZFS Toolbox Improvements

- More easily repurpose devices by wiping labels

```
# zpool label –C /dev/dsk/c3t2d0
```

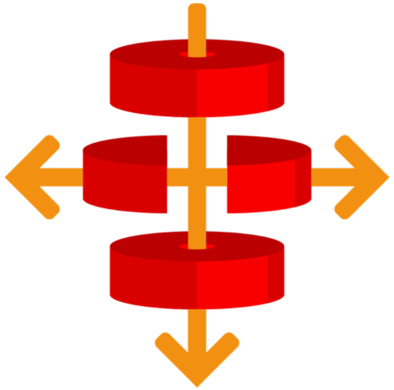
- Host hardware LUN snapshots on same system

```
# zpool reguid <pool>
```

Works only on imported, healthy pools with no faults

- Collect pool property values (scriptable):

```
# zpool get -H free,size datapool
datapool      free   452G   -
datapool      size   464G   -
```



Performance

Asynchronous Destroy

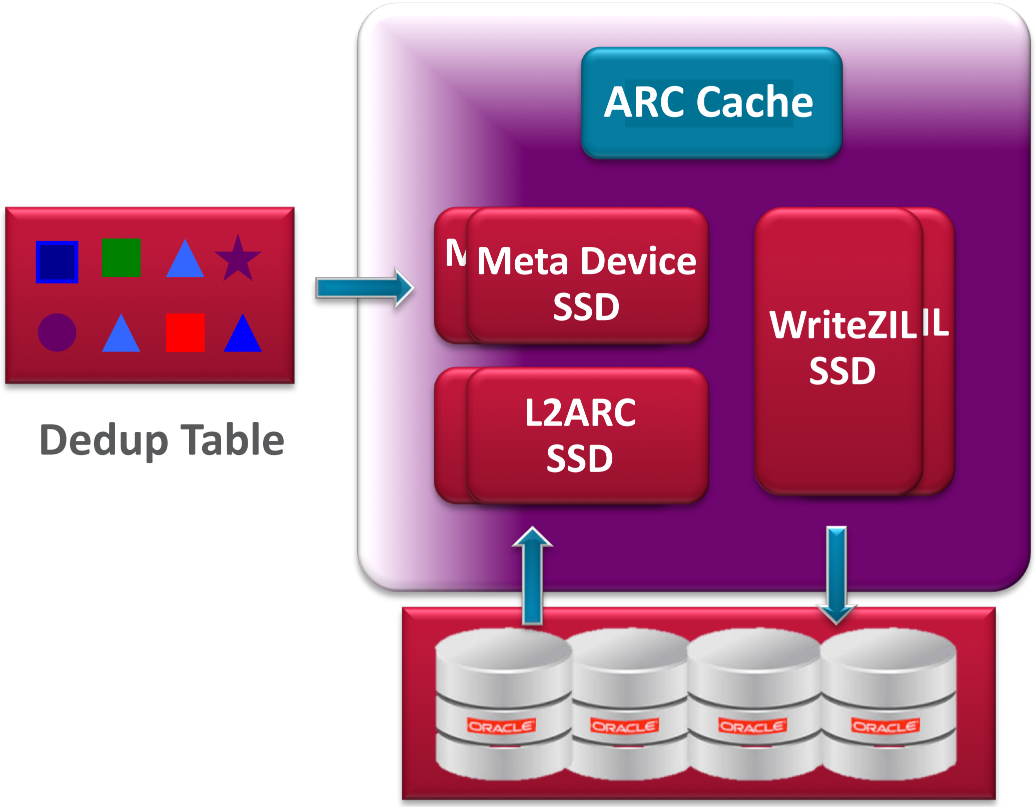
Improves dataset destroy operations

- Previous deletion operation was synchronous
  - Large dataset destroy could impact system operation
- Destroy operation now occurs asynchronously in the background
- Monitor destroy operations with zpool monitor command

Deduplication Meta Devices

Improves deduplication performance

- DDT entries are stored on SSDs to improve performance
- Read cache and DDT metadevices are separate SSDs
- Future work to share read/DDT workloads on same SSDs



ARC Cache

Meta Device SSD

L2ARC SSD

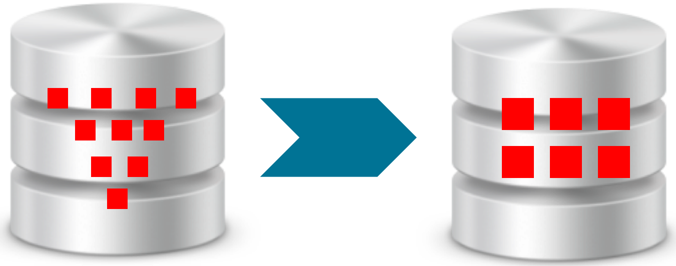
WriteZIL SSD

Dedup Table

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal

RAIDZ Space Efficiency

- Previous RAIDZ algorithm for rounding/unused blocks inefficient for 4K sector drives with small block size
- Reduced amount of padding for unused blocks
- Future work: Improve deduplication for small I/O workloads
- Future work: Improve small I/O workload performance for RAIDZ

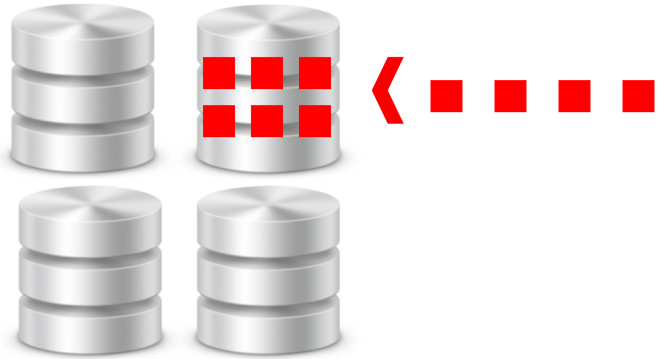


Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal

Improved Read Performance on Mirrored Pools

Smarter mirrored device selection

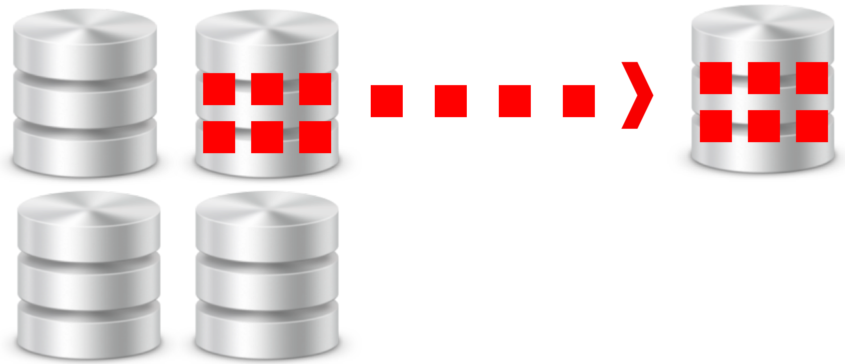
- Monitors mirrored pool devices to determine the device to favor for fastest read operation
- Tracks device queue depths and sends read requests to device with fewest pending I/Os
- Also monitors whether a device is slow and these devices are avoided for reads
- Slow device status is available in the zpool status



Resilvering Restarts Reduced

More efficient device failure recovery

- Resilvering improvement for large pools
- Separate populating and iterating phases during resilvering process
- Populating phase can take up to 50% or more of resilvering time
- Don’t restart populating phase if another device in same vdev fails



ORACLE

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal

Integrated Cloud

Applications & Platform Services

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. | Oracle Confidential - Internal

ORACLE®