Washington Systems Center - Storage

Accelerate with IBM Storage:

Data Reduction Pool (DRP) Overview/Best Practices

Byron Grossnickle
Spectrum Virtualize Specialist
Washington Systems Center

Accelerate with IBM Storage Webinars

The Free IBM Storage Technical Webinar Series Continues in 2019…

Washington Systems Center – Storage experts cover a variety of technical topics.

Audience: Clients who have or are considering acquiring IBM Storage solutions. Business Partners and IBMers are also welcome.

To automatically receive announcements of upcoming Accelerate with IBM Storage webinars, Clients, Business Partners and IBMers are welcome to send an email request to accelerate-join@hursley.ibm.com.

Located in the Accelerate with IBM Storage Blog:
https://www.ibm.com/developerworks/mydeveloperworks/blogs/accelerate/?lang=en

Also, check out the WSC YouTube Channel here:
https://www.youtube.com/channel/UCNuks0go01_ZrVVF1jgOD6Q

2019 Upcoming Webinars:

April 25 – Data Reduction Pools: Overview and Best Practices
Register Here: https://ibm.webex.com/ibm/onstage/g.php?MTID=e647561ca5cd145596bb1770fa1bd5fe7
May 2 - TS7760 Best Practices (A View from the Field)
Register Here: https://ibm.webex.com/ibm/onstage/g.php?MTID=e553c6faed4058a414f133151d5743f71
May 23 - How to use the IBM TS4500 and TS4300 Library Web GUI
Register Here: https://ibm.webex.com/ibm/onstage/g.php?MTID=edaae73a07257503198d0bdc04d18973e
May 30 - New Features with IBM Cloud Object Storage including a File Access Demo
Register Here: https://ibm.webex.com/ibm/onstage/g.php?MTID=eed0398a8690f50d436c35ba50c0135ac

Session Objectives

- DRP Overview
- DRP Planning
- DRP General Best Practices
- DRP On a Stand Alone Unit
- DRP on SVC

IBM Systems Flash Storage Offerings Portfolio

IBM Spectrum Virtualize | IBM Spectrum Accelerate | IBM Spectrum Scale | IBM Power Systems

NVMe end-to-end

Storwize V5010E / V5030E

Entry
SAS Hybrid & AFA

Storwize V5100/F

Entry
NVMe Accelerated
Hybrid & AFA Solutions

Storwize V7000

Enterprise for Everyone
NVMe Accelerated
Hybrid & AFA Solutions

FlashSystem 9110 / 9150

Enterprise Class
NVMe accelerated
Multicloud Enabled

FlashSystem A9000

Cloud Service Providers

FlashSystem A9000R

High End Enterprise

99.9999% Availability

Scale-out clustering
Simplified management
Flexible consumption model
Virtualized, flash-optimized, modular storage
Enterprise heterogeneous data services and selectable data reduction

Simplified management
Flexible consumption model
Large Grid scale
Full time data reduction

VersaStack™ Solution by Cisco and IBM

IBM FlashCore™ Technology Optimized

SVC

Enhanced data storage functions, economics and flexibility with sophisticated virtualization

NVMe FlashCore Module
Superior endurance & performance
- FIPS 140-2
- Hardware Compression

FlashSystem 900
Application acceleration

- Extreme performance
- Hardware Compression
- Targeting database acceleration

IBM Elastic Storage Server

Big Data

Consolidate file & object workloads
Faster data analysis
Global sharing

DS888xF

Business Critical

z/OS / AIX
Power HA
Power i HA

Business critical, deepest integration with z Systems
Superior performance and reliability
Three-site / Four-site replication
DS8888F, DS8886F, DS8884F, DS8882F

DRP Overview

New: Data Reduction Pools

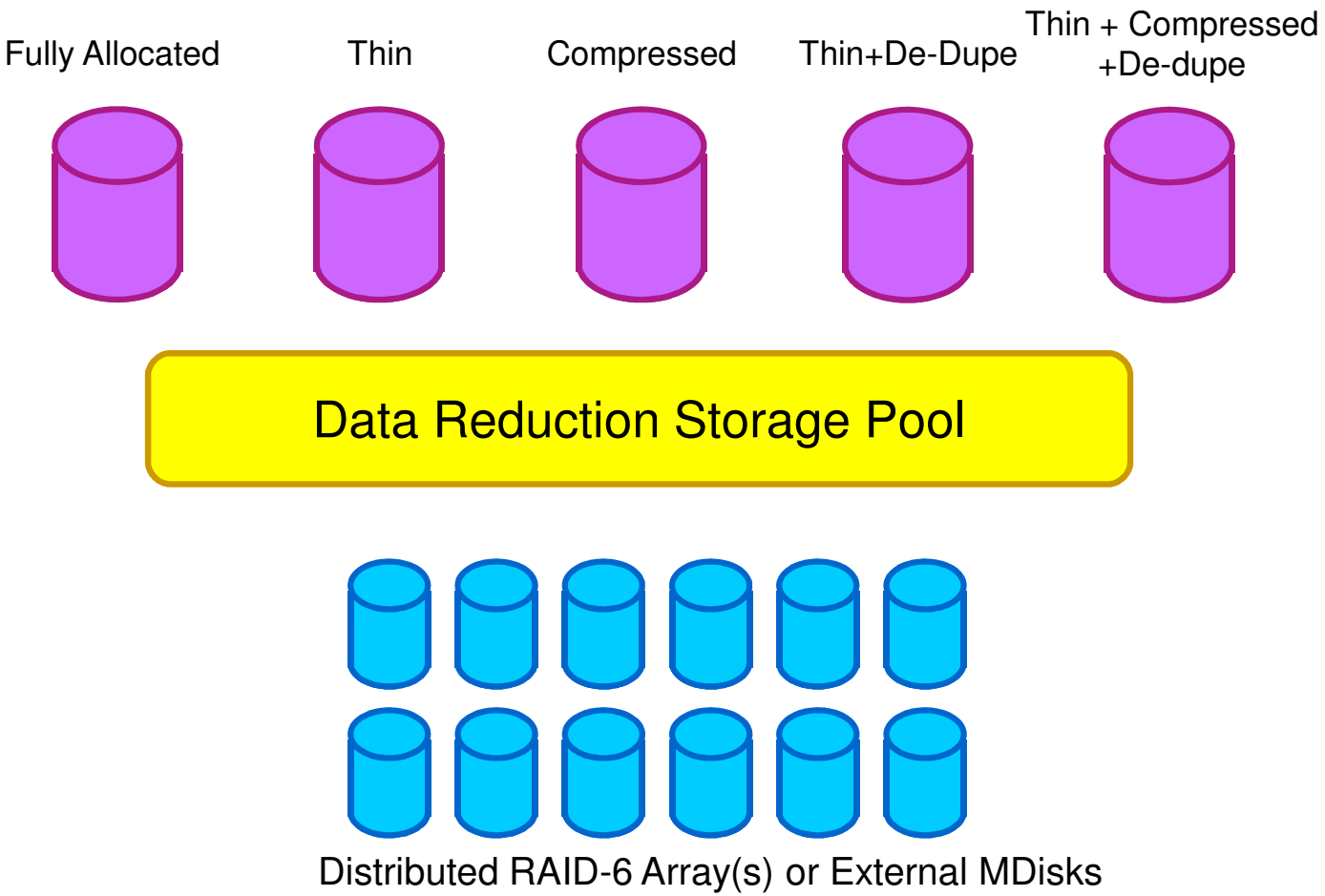
- Released in 8.1.2 – Recommended version 8.2.1.x
- New type of storage pool
- New implementation for thin provisioned volumes
- New implementation for compressed volumes
- Supports deduplicaton (8.1.3 and beyond)
- Supports up to 10K compressed volumes per system (the system volume limit)
- SCSI unmap garbage collection
- Data Reduction Pools supported on V5KG2/V5100, V7KG2/G2+/G3, SVC (DH8,SV1), V9000 (AC2, AC3), FS9100
  - No compression or deduplication support for 5010/5010E and 5020

- Note: Legacy storage pools and thin provisioned volumes are still supported
  - RtC is NOT supported on newer hardware: FS9100, V7KG3, V5100, V5030E

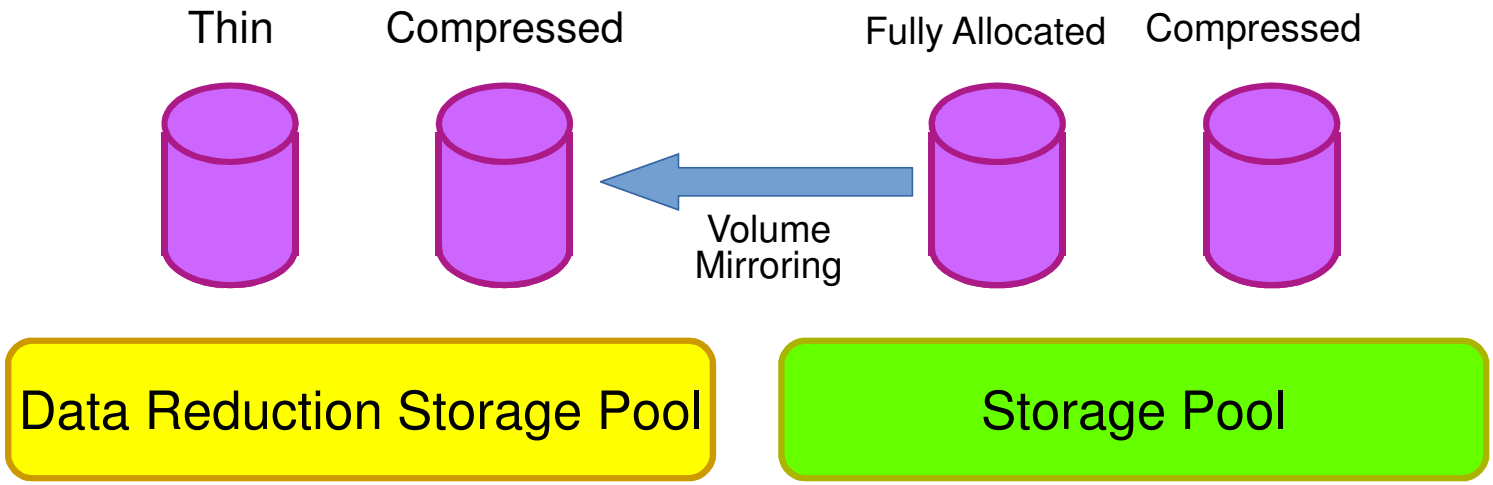
Using Data Reduction Pools

- New storage pool type designated at pool creation
- Create compressed + thin provisioned volumes as normal – pool type will determine if new or legacy implementation is used
  - Fully allocated volumes are also supported
    - Fully allocated volumes in a DRP use the same data path at in a legacy pool
      - This means that there is no performance advantage to using a legacy pool with fully allocated volumes
- Volume mirroring can be used to convert existing volumes to DRP
  - Note: Extent migration (migratevdisk) WILL NOT WORK
- Capacity and space saving reporting is added to the storage pool views
- Out of space warning thresholds are configured on the storage pool

New: Data Reduction Pools – Volume Types



Volume Migration



Note: If you are using RtC today, you must convert all RtC volumes to DRP compressed first, before you can enable deduplication

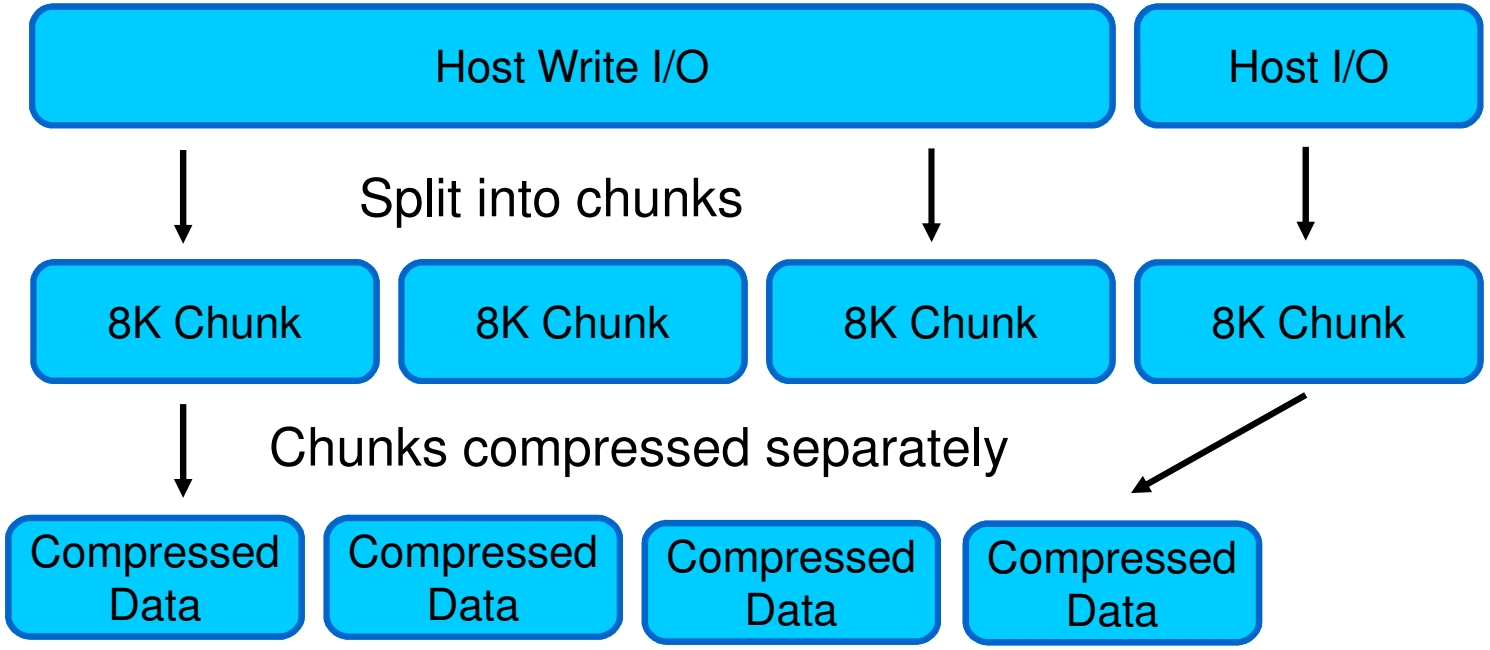
Benefits to Data Reduction Pools

- Designed to be scalable for next generation hardware, memory and CPU cores
  - Std SE and RtC do not scale with new generations of multi-core processors
- Compression integrated within I/O stack
  - Shared resource design
- When last RtC volume is converted, RtC cores that were dedicated to RtC will be used for all I/O processing
- Mirrored non-volatile metadata for compressed volumes means significantly improved failover/failback response times

Benefits to Data Reduction Pools - Continued

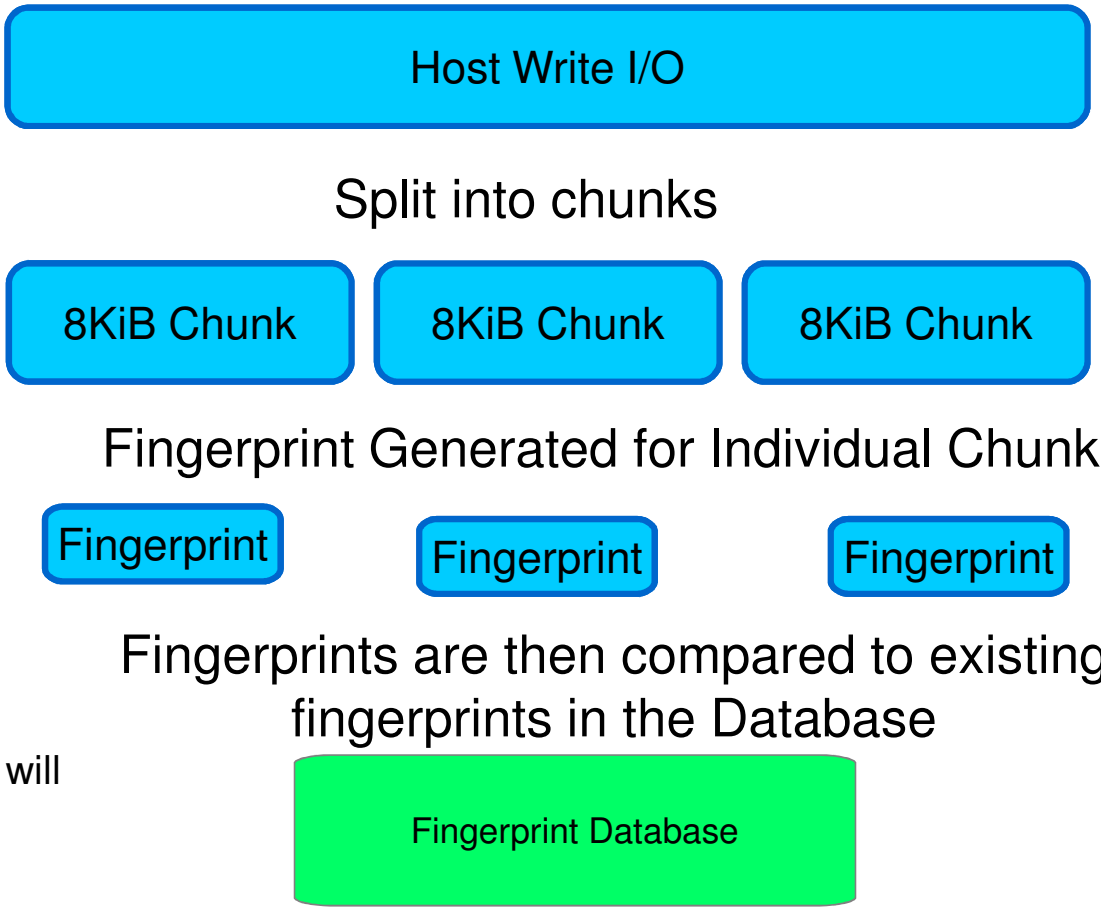
- No limit on the number of compressed volumes (system limit)
- Space reclamation – Unmap
- Smaller 8k chunk means less compression bandwidth for small I/Os
- Metadata and User data separated, better use of cache prefetch/destage.
  - Previously metadata was stored with user data
- More predictable latency compared to RACE
  - Less reliant on temporal locality
- Able to use maximum compression bandwidth from Intel offload cards
- Comprestimator Support

DRP Compression



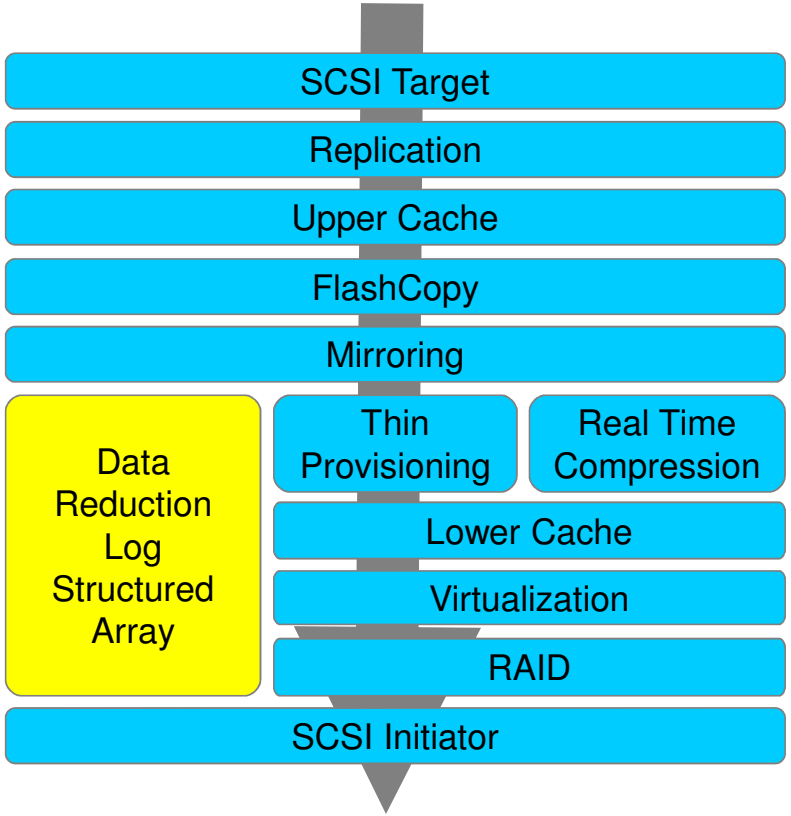
Compressed data is grouped into 256K blocks and written to storage
RtC was variable length input, fixed 32K output

Deduplication I/O

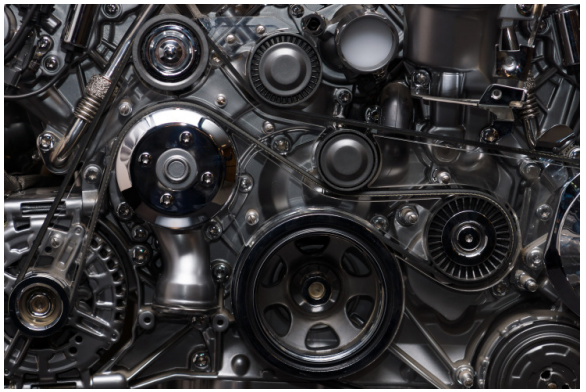


Note: A 4KiB block size will NOT deduplicate

SVC Internals – I/O Stack



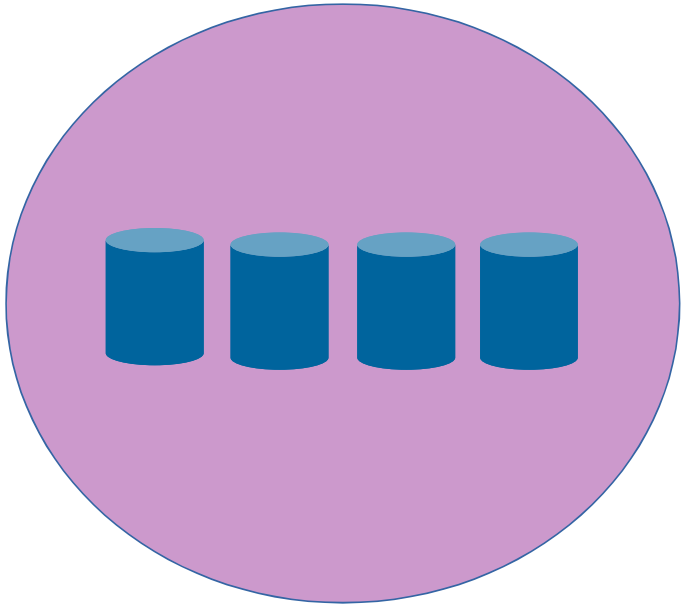
Internals



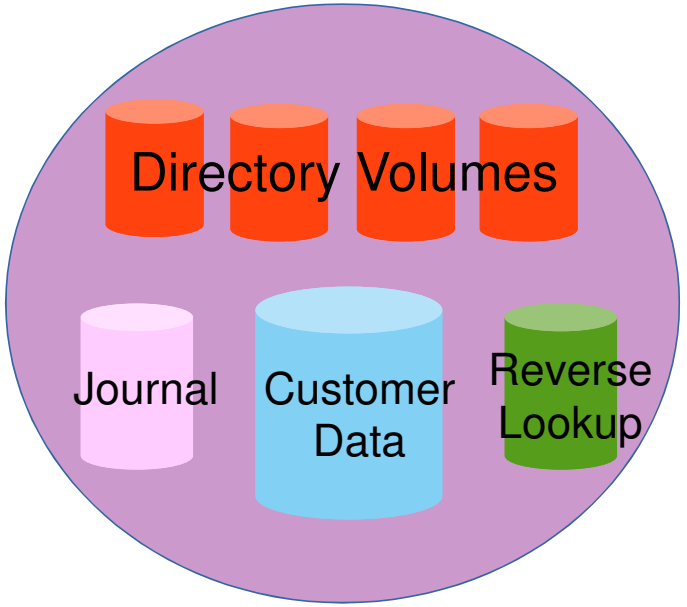
- CPUs
  - Data reduction uses same threads as main I/O process
    - No dedicated CPU cores for compression
    - No separate compression CPU utilization on Dashboard
- Memory
  - Data reduction shares memory with main I/O process
  - ~1GB memory taken from cache when data reduction is enabled
  - Fingerprint DB when deduplication enabled
    - Systems with 32GB per node = 12GB for fingerprint DB
    - Systems with 64GB per node = 16GB for fingerprint DB
    - Systems with 128GB+ per node = 32GB for fingerprint DB
- Compression Hardware
  - Shared with existing RrC compression and compression for IP replication
  - New DRP compression can drive Quick Assist hardware to its limits (RtC could not)
    - Coletto Creek hardware (assuming 2 per controller) and smaller Lewisburg chips – 4.8GB/s per controller (9.6GB/s I/O group)
    - Larger Lewisburg chip (FS9150) – 12.5GB/s per controller (25GB/s I/O group)

Inside a Data Reduction Pool

User view:
4 volumes within a data reduction pool



Internally:
4 Directory volumes – 1 per volume
Think of the host talking to the directory volume
1 Customer data volume (per IO group) – Counts against 10K
1 Journal volume (per IO group) – Counts against 10K
1 Reverse Lookup volume (per IO group) – Counts against 10K



Up to 48 volumes of the 10K used
3 volumes per pool X 4 I/O groups is 12 volumes
12 volumes X 4 DRP pools for maximum performance – 48 volumes of the 10K that can be used.

Note: The directory volumes are the only volumes visible and accessible to the end user

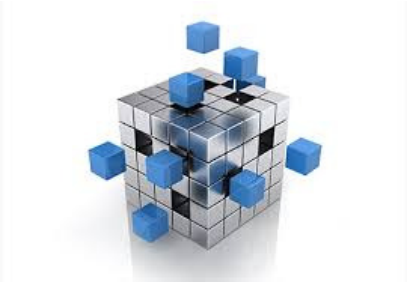
Why Lots of Internal Volumes?

Different volumes have different I/O patterns

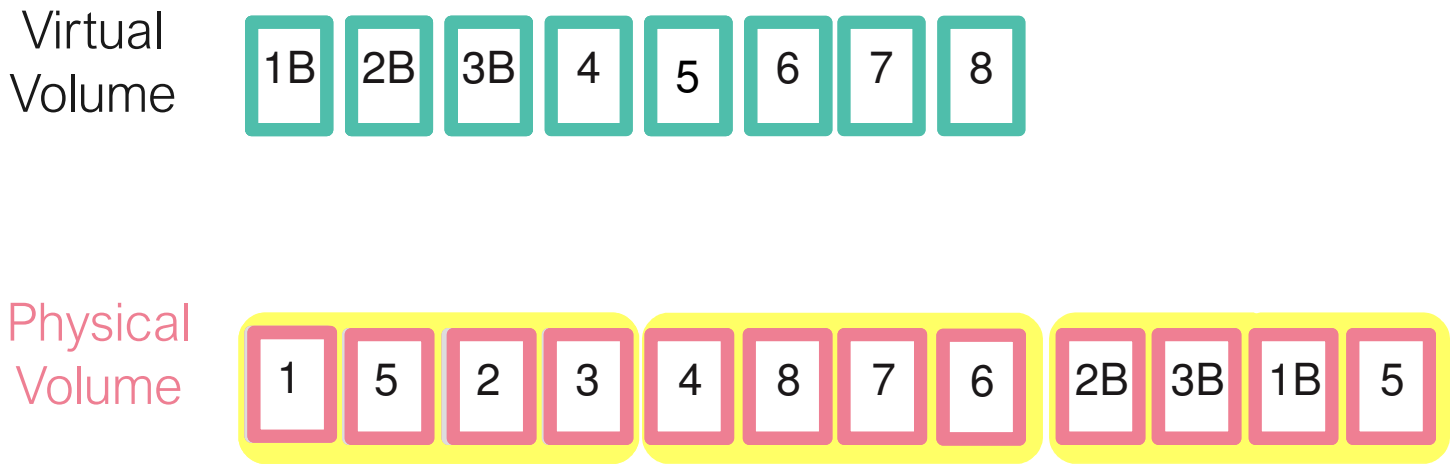
- Customer Data Volume – 98% of pool capacity
  - Where the customer data lives
  - Large sequential write I/Os, short random read I/Os
  - Writes are 256k into lower cache – coalesced into Full stride writes
- Directory Volumes – 1% of pool capacity
  - Directory of where everything in customer volumes lives
  - Short 4K random read/write I/Os
  - 1 per volume – Think of the host as talking to the directory volume
- Journal Volume - <2% of pool capacity
  - Read mainly for recovery scenarios (e.g. T3)
  - Large sequential write I/Os typically 256k into lower cache, only read for recovery scenarios (e.g. T3)
- Reverse Lookup Volume - <1% of pool capacity
  - Generally used for some metadata and garbage collection
  - Short semi-random read/write I/Os

Log Structured Array

- When data is overwritten in a compressed volume the new data usually compresses to a different size
- Solve this problem by writing new data somewhere new and deleting the old data leaving small holes
  - Unmap also creates small holes
  - De-duplication also creates small holes (when existing data is overwritten with new data that is a duplicate)
- Trying to fill in small holes is very inefficient – too many I/Os to keep reading and updating directory
- Solve this problem with garbage collection
  - Wait until an extent has many small holes
  - Move the remaining data in the extent (read and write somewhere new)
  - Once extent is empty either free back to VG or fill it with new data



Garbage Collection



Random write I/O fills extents sequentially
Overwritten data creates holes
Garbage Collection empties extents with lots of holes

Garbage Collection

- Uses free capacity in the pool to relocate valid user data that is left.
  - Tracked by reverse lookup volume
- Coalesces the user data left into 256k blocks. Lower Cache coalesces into Full-stride writes.
- Systems need to be sized to ensure no greater than 85% used DRP pool capacity.
- Garbage Collection Plans generated frequently based on (but not limited to):
  - Frequency of over-writes
  - Rate of New I/O
  - Rate of data being invalidated (eg. Unmap)
  - Active/Stale data
  - Amount of space required to move data.
  - Amount of space free.
  - Extents with largest number of holes
- Data grouped into frequently modified and not frequently modified.
- Extent freed or recycled
- Garbage collection rates based on pool fullness – IO amplification

New Statistics

- VDisk
  - used_capacity_before_reduction
- Mdiskgrp (Pool)
  - reclaimable_capacity – measures amount of GC work
  - used_capacity_before_reduction
  - used_capacity_after_reduction
  - deduplication_capacity_savings
  - compression_opportunity (GUI only)
  - overhead_capacity
- System
  - total_reclaimable_capacity
  - used_capacity_before_reduction
  - used_capacity_after_reduction
  - overhead_capacity
  - deduplication_capacity_savings
  - compression_opportunity (GUI only)

Licensing Implications of DRP

- Compression in DRPs is part of the base license in SVC.
  - When completely converted to DRPs, maintenance on RtC can be dropped
- FS9100 models have an all inclusive license
- Storwize products (except the 5030E) require a compression license
  - Included in the full bundle
  - DRP compression is included in the LMC license of the V5030E

Deleting a Deduplication Volume Copy

- IT WILL TAKE TIME!!!!!
  - Thin/Compressed volumes need to do the unmap commands in order to be deleted
  - De-duplicated volumes need unmap commands as well but also have to have their contents “re-homed”. Every volume in the system must be checked to make sure that it does not rely on any chunks of the volume being deleted and if they do that data must be re-homed.
    - Just remember – There is no free lunch!
    - This only applies to deduplicated volumes and not volumes that are merely compressed/thin

Note: Development is looking at ways to increase the speed of deleting deduplicated volumes

Note: Development is looking at ways to address customer questions around the process for deduplicating volumes

DRP Planning

Current Data Reduction Pool Limits

A maximum of 4 Data Reduction Pools per System

A maximum of 128K extents per ‘Customer Data Volume’ - per IO/Group. *

Thus the Pool extent size dictates the maximum physical capacity in a pool – after data reduction savings.

Currently 4GB is the recommended size

GUESS LARGE!!!!!!

*Since the Customer Data Volume only contains data for volumes owned by that I/O Group - each I/O Group has its own Customer Data Volume per DRP

Extent Size	Max Capacity in 1 DRP with 1 I/O Group	Max Capacity in 1 DRP with 4 I/O Group	Max Capacity per System with 4 DRP
1GiB	128TiB	512TiB	2PiB
2GiB	256TiB	1PiB	4PiB
4GiB	**512TiB**	**2PiB**	**8PiB**
8GiB	1PiB	4PiB	16PiB

Current Data Reduction Pool Minimum Capacity Limits

There are limits on the minimum size for a Data Reduction Pool to ensure Garbage Collection works acceptably.

Full reservation is taken from pool when first volume is created and written to.

Extent Size	Min Capacity in 1 DRP with 1 I/O Group	Min Capacity in 1 DRP with 4 I/O Group
1GiB	255GiB	1TiB
2GiB	0.5TiB	2TiB
4GiB	1TiB	4TiB
8GiB	2TiB	8TiB

Current Data Reduction Pool Restrictions

Maximum of 4 DRP

- It takes 4 to get maximum performance
- However depending on the system, 1 DRP pool will give you the best manageability and usable capacity

Child Pools are not supported in a Data Reduction Pool

- Hence VVOL is not supported in a Data Reduction Pool
- Also why Object Based Access Control ownership group is not supported

Volume cannot be shrunk in a Data Reduction Pool

No Volume move between IO Groups if Volume is in a Data Reduction Pool*

No split of a Volume Mirror to a copy in a different IO Group

Real/used/free/tier capacity not reported per volume - just per pool.

* FC or MM/GM can be used

Current Data Reduction Pool Restrictions

Cache mode is always read-write

Autoexpand always on

No ability to place specific volume capacity on specific MDisks

No extent level migration commands

No per-volume contingency buffer

Free capacity managed at pool level across all volumes

mkvdisk -rsize parameter is accepted but value is ignored

Other Migration Considerations

- For systems with Quick Assist hardware
  - No RtC and deduplication in the same I/O group at the same time
    - You will have to convert all RTC volumes to DRP compressed volumes BEFORE you convert to de-dupe
- For systems with software only based compression
  - No RtC and DRP compression in the same I/O group at the same time

DRP General Best Practices

DRP Pool Size/Extent Size

- Recommended to be a minimum of 20TB
- 100 – 200 TB = sweet spot
- Lower than 1PB per I/O Group
- 4 or 8GB extent size recommended
  - Newer systems (FS9100, V7KG2, V5100) default to 4GB
  - 4GB = 512TB physical capacity per I/O group
  - 8GB = 1PB physical capacity per I/O group
- 4 DRP pools maximum per system
  - Must balance # of pools with amount of storage/drives to be allocated
  - Must balance # of pools with performance
    - Maximum performance with 4 pools
      - On some stand alone systems with a limited # of drives this is impractical/impossible. On these systems 1 is recommended to maximize simplicity and usable capacity
  - In general, the more memory the better
  - In general, do not use DRP volumes without hardware acceleration
    - For example, although the 5030E supports it, there is no hardware acceleration. The V5100 would be a smarter choice for DRP volumes

DRP Pool Capacity

- If using DRP thin or compressed volumes (regardless of dedupe) your pool should be no more than 85% occupied
  - Why?
    - DRP thin or compressed volumes always do 256KB writes to new space, therefore you must have some space free to write to
    - We desire to hold off garbage collection as long as possible because it is possible that there will be less work to do by waiting
      - If a whole extent is garbage we have to move nothing and can just mark the space free
    - We slowly ramp up garbage collection as the pool fills up
      - Don’t expect garbage collection to do much if you have lots of space free
      - If necessary, you can “trick” the system by creating some fully allocated volumes and making the pool fuller than what it is to force garbage collection to do more work.
    - After 85% full garbage collection is running at full steam trying to free space and may impact performance
    - This guidance does not apply if you are using 100% fully allocated volumes
  - For systems with hardware compression offload (Quick Assist) it is recommended to use compressed volumes instead of thin. This becomes critical on systems with FCM’s for managing physical capacity on the back end.

DRP SCSI Unmap

- If you need to use SCSI unmap make sure that it is on
  - Depending on model and/or which software versions you upgraded from it could be off
  - lssystem |grep unmap
- 8.1 code has one level of unmap (unmap on or off)
- 8.2 code has two levels of unmap
  - Host unmap
  - Backend unmap
- To change the unmap use the chsystem command

DRP with Flash Core Modules

Performance Volumes

- As the name implies for those volumes that need maximum performance
- FCMs have built in, no penalty compression
  - 4.8TB module = 4:1
  - 9.6TB module = 2:1
  - 19.2TB module = 2:1
- FCM's advertise effective capacity to the DRP pool
- Create fully allocated (Capacity Savings:None) in the DRP pool.
  - This will bypass the overhead of the Log Structured Array and send data directly to the FCM's for no penalty compression

Note: Performance volumes and capacity volumes can be used in the same DRP to maximize performance and capacity. This sharing is recommended for stand alone units only

Capacity Volumes

- For volumes that do not need maximum performance
- These utilize the Log Structured Array
- Compression is more efficient than merely thin
- Compression/thin alone perform better than with deduplication
- The DRP pool space will see the benefit of SCSI unmap for these volumes

Performance Safe Transitions

- What if I am unsure whether I need a performance volume or a capacity volume?
  - Start with a performance volume
  - Look at the performance for those volumes you wish to potentially move
  - Instead of using the Modify Capacity Savings use the following manual process
    - Add a volume copy with the desired target configuration (thin/compressed +- dedupe)
    - As the copy starts to sync up, you should get an idea of how write performance is affected
    - When copies are in sync, change the primary copy to the target volume
      - At that point all reads are coming from the target volume and writes are going to both
      - If performance is not meeting requirements drop the target copy immediately – no harm/no foul
  - If and/or when you are satisfied with performance of the target copy, delete the original fully allocated copy
  - Once you get an idea of what kind of things perform well as capacity volumes you can use the Modify Capacity Savings

DRP General Performance Considerations

- Volumes still have an affinity to CPU cores
  - Make sure you have at least as many volumes on the system as you do CPU cores
- Make sure you have a good queue depth coming from your servers
- Zone so that you have no more than 4 paths per volume
  - 4 paths per volume is recommended
  - 8 path per volume is supported
- DRAID6 for your RAID arrays on internal drives
- If you are using FCMs the best performance will be experienced by using a DRP with fully allocated volumes
  - The fully allocated volumes have no performance overhead the Log Structured Array of DRP

DRP – Stand Alone Unit

Stand Alone Unit

- Assuming you are using a system with drives in the control unit only
  - Group all drives of the same type into a single DRAID6 array and put it in a single DRP pool
    - One DRAID6 array maximizes usable space and 1 DRP pool keeps things simple
- If you are creating a hybrid pool:
  - When creating your DRAID arrays, choose Internal Custom
    - Choose your drive class
    - Choose DRAID6
    - Choose the entire number of drives you want in the pool and allow the system to determine the best DRAID6 configuration
      - Remember that in DRAID your spare capacity is in the DRAID array itself so you should not have any drives marked as “spare” in your system
- Create at least as many volumes as what you have CPU cores to maximize system performance

DRP Capable Unit behind SVC

- Turn off NPIV (Target Port Mode)
  - On by default on: FS9100, V7KG3, V5100
  - Zone to physical WWPNs
- Group drives of same type into DRAID6 arrays in STANDARD pools
  - Use fully allocated volumes to give to the SVC
  - Do NOT over allocate the pool
- If you are using FCMs on an array behind SVC and you are using DRP with fully allocated volumes on SVC, start by provisioning the physical capacity to SVC and then cautiously add up to the effective capacity of the pool
  - Monitor the back end physical utilization to ensure you don't run out of space
- If you are using FCMs on an array behind SVC AND you plan to use DRP compressed volumes on SVC only allocate the amount of physical (not effective) capacity to the SVC

Note: In 8.3.0 there is a wizard that will do all of this

DRP On SVC

Using Data Reduction at 2 levels

If you create a solution where data reduction technologies are applied at both the storage and the virtualization appliance levels, then here are the rules you should follow

ALL DRP volumes should run with **compression** switched on

(Performance bottlenecks come with DRP metadata, not compression)
(The above statement does not apply to deduplication)

Physical storage behind SVC should be allocated **1:1**

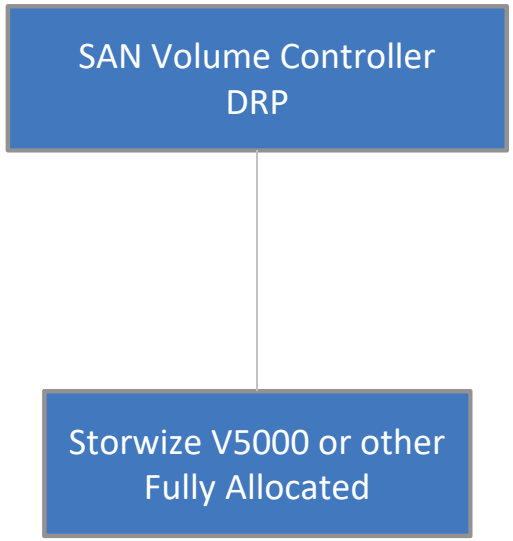
(e.g. on AE3 this means that physical capacity is mapped NOT effective capacity)

Fully allocated should be in their **own pool**

(~<20% of fully allocated can exist in the same pool and be lost in the error bars)

If you want to use DRP with an existing overallocated backend, you need to reclaim storage and configure it according to the best practices

DRP above simple RAID



Recommended!

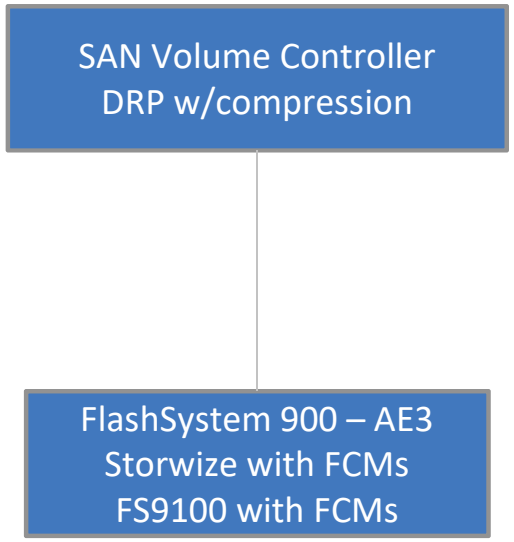
Use DRP at top level to plan for de-duplication and snapshot optimizations.

DRP at top level provides best application capacity reporting (volume written capacity).

Always use compression in DRP to get best performance.

Bottlenecks in compression performance come from meta-data overheads, not compression processing.

DRP above data reducing backend



Recommended!

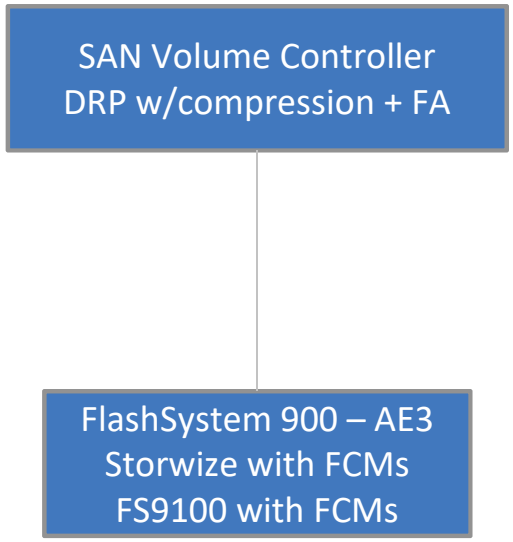
Assume **1:1** compression in backend storage – **do not overcommit!**

Small extra savings will be realised from compressing meta-data.

Using DRP with overallocated back end could lead to the DRP garbage causing out-of-space

Think : For existing systems, do you need to move to DRP to get the benefits of dedup, or is hardware compression good enough.

DRP + Fully-Allocated above data reducing backend



Not recommended!

Very difficult to understand physical capacity use of the Fully Allocated volumes.

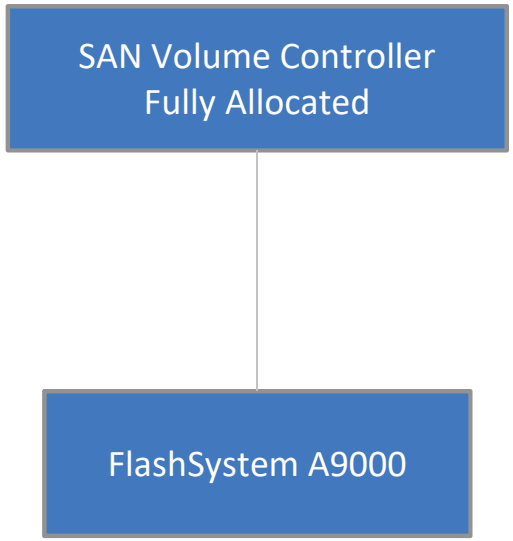
(Use with caution with a ~<20% mix of fully allocated)

DRP garbage collection will act as if Fully Allocated volumes are 100% used.

Temptation is to exploit capacity savings which might ***overcommit backend.***

The right answer is to use separate pools

Fully-allocated above single-tier data reducing backend



Use with appropriate precautions.

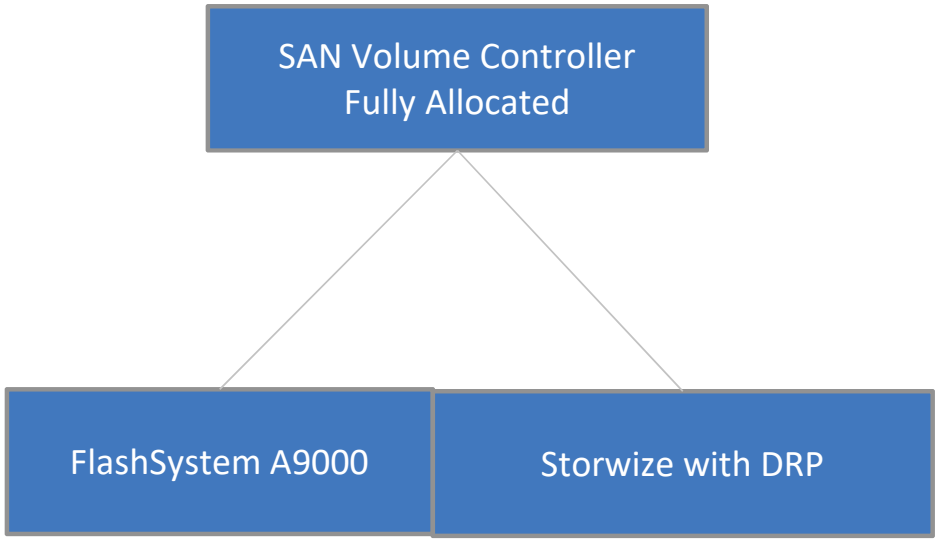
Need to track physical capacity use carefully to avoid out-of-space.

SVC can report physical use but does not manage to avoid out-of-space.

No visibility of each application's use at SVC layer.

If actual out-of-space happens there **is very limited ability to recover**. Consider creating sacrificial emergency space volume.

Fully-allocated above multi-tier data reducing backend



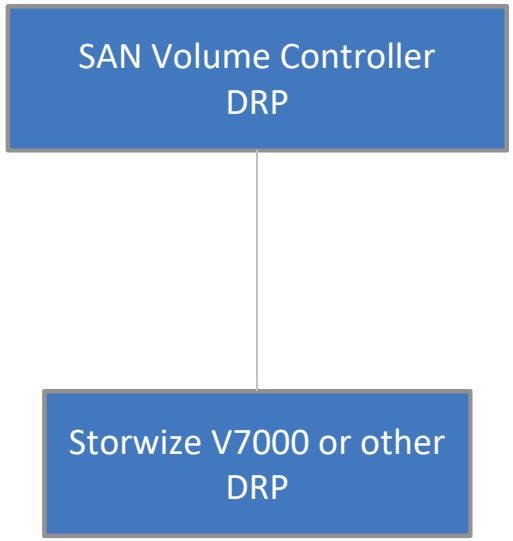
Use with great care!

Easy Tier is unaware of physical capacity in tiers of hybrid pool.

Easy Tier will tend to fill the top tier with hottest data.

Changes in compressibility of data in top tier can overcommit the storage leading to **out-of-space.**

DRP above DRP



Avoid!!

Creates two levels of IO amplification on meta-data.

Two levels of capacity overhead.

DRP at bottom layer provides no benefit.

Summary

Monitor Capacity

Configure alerts – do not get caught out

Running out of physical capacity will take volumes offline

Don’t over-complicate

If virtualizing a compressing storage system use a recommended design.

Other designs make it impossible to understand your capacity consumption.

You will run out of space!

THANK YOU!