Deduplication Best Practices
With Microsoft Windows Server 2012 and Veeam Backup & Replication 6.5

Joep Piscaer, VMware vExpert, VCDX #101
j.piscaer@virtuallifestyle.nl
@jpiscaer
Agenda

- Introduction
- Use case for deduplication
- Configure Data Deduplication in Windows Server 2012
- How to optimize backup repository for deduplication
- How to optimize backup jobs for deduplication
- How to leverage per-pool deduplication
- Demo
Agenda

- Introduction
- Use case for deduplication
- Configure Data Deduplication in Windows Server 2012
- How to optimize backup repository for deduplication
- How to optimize backup jobs for deduplication
- How to leverage per-pool deduplication
- Demo
Introduction

- Joep Piscaer
  - Consulting Architect at OGD ict-diensten
  - Know Veeam since 2007 and in love with them ever since (best. VMworld. parties. ever.)
Past Projects

- Past implementations of Veeam B&R
  - Commonly see a VMware virtualization layer with Windows VMs on top
  - Windows Server 2008 R2 as the base for Veeam implementation
  - Starting to work with Windows Server 2012

- Notable projects include
  - Bi-directional DR for 200-250 VMs with 2 infrastructures
  - 150+ VM backup and replication within a single large datacenter
  - Application consistent backups of Zarafa Collaboration Platform without bringing database down (or any other downtime)
  - Numerous smaller projects for DR or backup at customer sites
Agenda

- Introduction
- Use case for deduplication
- Configure Data Deduplication in Windows Server 2012
- How to optimize backup repository for deduplication
- How to optimize backup jobs for deduplication
- How to leverage per-pool deduplication
- Demo
What is deduplication?

LESS IS MORE
What is deduplication?

- Deduplication identifies identical data blocks in source VMs and stores each unique block only once.
Use case – why use deduplication?
Use case – why use deduplication?

- 41.9% of CIOs affected by cost of storage required for storing backups.
Use case – why use deduplication?

- Lack of disk space is a growing issue preventing adoption of replication.

- 22.6% of CIOs reported lack of disk space as an issue preventing adoption of replication in 2011; 35.5% report the same in 2013.
Use case – why use deduplication?

- Conclusions of Virtualization Data Protection Report 2013
  - Decrease cost of storage
  - Increase usage of existing storage
    Make backups as storage-friendly as possible

- To keep management cost down, don’t introduce new or complex solutions
  - Use ‘what we already know’ and ‘what we already have’ to solve these challenges
  - Prevent re-training of administrators
  - Increase ease of use of solution
Use case – why use deduplication?

▪ Why deduplication instead of feature x or technology y?
  ● Veeam captures entire VM including Guest OS, applications and data
  ● VMs usually share Guest OS type, middleware, etc.
  ● Hence: there tends to be a lot of identical data across VMs
  ● Deduplication slashes identical data

▪ But mix and match solutions if available
  ● Don’t forget compression
  ● Excluded unneeded data (VM and Guest OS swap files, etc)
Use case – why use deduplication?

- Backup files can be separated into two categories:
  - Recent backups for fast restore and recoverability
  - Older backups for archival purposes

- With Veeam, these two types are ‘connected’
  - The backup file chain contains ‘recent’ and ‘old’ backups
  - Nearly impossible to separate on the storage layer
Why use Windows Server 2012 deduplication?

- The killer use case for using Data Deduplication in Windows Server 2012 is longer-term (>60 day) retention or archival of VMs on the same on-site storage platform. Other use cases include off-site replication and increased performance of forward incremental jobs.

- For 30-60 day retention, consider reverse incremental backup mode which offers similar storage efficiency.
Agenda

- Introduction
- Use case for deduplication
- Configure Data Deduplication in Windows Server 2012
- How to optimize backup repository for deduplication
- How to optimize backup jobs for deduplication
- How to leverage per-pool deduplication
- Demo
Planning for disk-based deduplication

- Determine the amount of data being backed up
- Determine the data retention policy for each dataset
  - Determine if and when to use an archival system
- Determine the biggest full uncompressed backup file size
  - Windows Server 2012 Data Deduplication is post-process (and not in-line)
  - So you need at least this amount of free storage space to complete the job
- Determine the daily change rate
  - Determine if Data Deduplication can handle that amount of data
  - MSFT recommends to design for ~100GB/hr deduplication data processing
Planning for disk-based deduplication

- Estimate dedupe space savings on existing dataset
  - Use ‘Data Deduplication Savings Evaluation Tool’ (DDPEval.exe)
  - Installed in C:\Windows\System32 on 2012 host with Deduplication role
    But can be copied to any system running 2012, 2008 R2 and 7

```
Data Deduplication Savings Evaluation Tool
Copyright 2011-2012 Microsoft Corporation. All Rights Reserved.

Evaluated folder: E:\Test
Processed files: 34
Processed files size: 12.03MB
Optimized files size: 4.02MB
Space savings: 8.01MB
Space savings percent: 66
Optimized files size (no compression): 11.47MB
Space savings (no compression): 571.53KB
Space savings percent (no compression): 4
Files with duplication: 2
Files excluded by policy: 20
Files excluded by error: 0
```
Planning for disk-based deduplication

- Estimated sizing:
  - Applicable to environments with less than 1.5T of daily data ingestion daily (10.5T weekly)

- Limits
  - Large full backups (> 1.5T) require lengthy dedupe process
  - Consider running active fulls monthly to split ingestion of large amounts of data
Live Demo
Agenda

- Introduction
- Use case for deduplication
- Configure Data Deduplication in Windows Server 2012
- How to optimize backup repository for deduplication
- How to optimize backup jobs for deduplication
- How to leverage per-pool deduplication
- Demo
Tying it together with Veeam

- Veeam Backup & Replication 6.5 include full support for Microsoft Windows Server 2012 including ReFS volumes and global data deduplication.
- But support for Storage Spaces is experimental.
- To restore files from deduplicated volumes, backup server must be installed on Windows Server 2012 with Data Deduplication feature enabled in the OS settings.
Backup Repository settings

- Deduplicating storage compatibility
  - Align backup file data blocks
  - Decompress backup data before storing
Agenda

- Introduction
- Use case for deduplication
- Configure Data Deduplication in Windows Server 2012
- How to optimize backup repository for deduplication
- How to optimize backup jobs for deduplication
- How to leverage per-pool deduplication
- Demo
Backup Job settings

- Backup Mode
  - Incremental
  - Use ‘dedupe-friendly’ compression
  - Set storage optimization to ‘local target’

- In my experience, enabling Veeam dedupe and compression significantly speeds up job processing and do not interfere with Windows deduplication
Backup Job settings

- Why forward incremental?
  - Less ingestion of new data on deduplicating repository
  - Reversed incremental requires 3x I/Ops
    Deduplicating repository is usually I/O bound
Backup Job settings

- Synthetic or periodic active full?
  No difference after dedupe but synthetic prevents ‘full’ job to hit source. Synthetic fulls require data re-hydration while last full is read from disk and is building the new synthetic full. Try synthetic fulls first and monitor performance (time needed to build new synthetic full).

- Synthetic full only takes up space before deduplication process kicks in; after dedupe, synthetic full is deduped against previous synthetic full
  You want the previous synthetic full saved in deduped state the moment you create a new synthetic full, so make sure to keep enough restore points on disk.
Backup Job settings

Why enable inline data deduplication?

- Source-based deduplication
  - Takes place at the backup source instead of target
- Decreases the amount of data sent over the network
  - Amount of network traffic is a consideration, too
- Speeds up job processing significantly
- Uses a large block size so doesn’t interfere with Data Deduplication

- If we disable inline data deduplication:
  “The amount of data on disk remained the same, but the amount transferred across the wire is 40x more.”
Backup Job settings

- Why enable compression?
  - Source-based compression
  - Decreases the amount of data sent over the network
  - Speeds up job processing significantly
  - Dedupe-friendly compression uses a fixed dictionary and doesn’t interfere with Data Deduplication
  - Dedupe-friendly compression saves about 10-20% on the initial VBK/VIB file size. About the same (20-30%) is lost at Data Deduplication stage. Significantly faster restores and slightly faster instant recovery.
  - “Effectively you are trading some hard disk space overall (because of less dedupe) for some up front network and disk bandwidth savings.”
Agenda

- Introduction
- Use case for deduplication
- Configure Data Deduplication in Windows Server 2012
- How to optimize backup repository for deduplication
- How to optimize backup jobs for deduplication
- How to leverage per-pool deduplication
- Demo
Global Data Deduplication

- Deduplicate data across backup files produced by different backup jobs
  - Veeam uses per-job deduplication, but space savings isn’t the only consideration when choosing which VMs go into a job.
  - Microsoft deduplicates on a per-volume basis
    Multiple backup files / jobs stored on a single volume

- One less variable to consider when planning jobs
  - Makes ‘more small jobs’ a viable solution
  - Focus more on more functional separation of VMs in jobs
    - Separate per (multi-tier) application (vCenter Intenvory VM Folders)
    - Replicate off-site using Cloud Edition
Agenda

- Introduction
- Use case for deduplication
- Configure Data Deduplication in Windows Server 2012
- How to optimize backup repository for deduplication
- How to optimize backup jobs for deduplication
- How to leverage per-pool deduplication
- Demo
Live Demo
Demo – Dedupe: Jobs vs. Volumes

- 3 VM’s in 3 separate backup jobs
  - 24.2 GB before dedupe, 13.3 GB after dedupe
- Copy a large file (~8GB) to all three VMs
- Run backup jobs (regular incremental run)
  - 45.9 GB before dedupe, 19.5 GB after dedupe
- Run deduplication process on repository
  - Run dedup with ‘Start-DedupJob -type Optimization -volume E:’
  - Check status running dedupe job with ‘Get-DedupJob’
  - Check space savings after dedupe job with ‘Get-Dedupstatus’
- Source VM’s occupy 65 GB on source datastore
  - That’s a 3.33:1 deduplication rate!
Demo – Instant VM Restore

- Re-hydration of deduplicated backup files takes time and performance will suffer.
- Most recent backups typically not deduplicated yet because Data Deduplication is post-process, so instant recovery using vPower NFS won’t suffer in performance.
- Don’t set Data Deduplication schedule to dedupe (most) recent backup files as these are used for restores in 99% of the cases. Set data aging to > 1-2 days.
- Weigh the benefits of disk savings against increased RTO.
Introducing Veeam Backup Cloud Edition - #1 VM Backup is now cloud ready

Joep Piscaer
VMware vExpert, VCDX #101

j.piscaer@virtuallifestyle.nl
@jpiscaer