SNIA

Education

ASPECTS OF DEDUPLICATION

Dominic Kay, Oracle
Mark Maybee, Oracle

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

© 2010 Storage Networking Industry Association. All Rights Reserved.

Abstract

- This tutorial will focus on block-level deduplication. While conceptually simple, an implementation can be quite complex as it must address multiple issues:
  - scalability - when the lookup table no longer fits in memory.
  - performance - impact of table lookups and writes dependent on reads.
  - space accounting - space now be shared between files and file systems.
  - administration - keeping the model simple.
- This tutorial will also
  - cover expanding the notion of deduplication beyond storage devices.
  - discuss in-memory and over-the-wire deduplication.

© 2010 Storage Networking Industry Association. All Rights Reserved.

Deduplication Defined

- Stores first unique domain, additional copies increase reference counts
- Improves storage efficiency
- Historically used for backups
  - Moving into archiving and primary storage
- Leads to reduced redundancy
  - A single corrupted block can have far greater impact
- In-line or post processing
- Can be done at the File, block, or byte level

© 2010 Storage Networking Industry Association. All Rights Reserved.

Other Tutorials

This tutorial covers research & work in progress. For a grounding in deduplication check out this SNIA Tutorial:

Understanding Data Deduplication (Thomas Rivera)

© 2010 Storage Networking Industry Association. All Rights Reserved.

ZFS Overview

- Pooled storage
  - Completely eliminates the antique notion of volumes
  - Does for storage what VM did for memory
- Transactional object system
  - Always consistent on disk – no fsck, ever
  - Universal – file, block, iSCSI, swap ...
- Provable end-to-end data integrity
  - Detects and corrects silent data corruption
  - Historically considered “too expensive” – no longer true
- Simple administration
  - Concisely express your intent

© 2010 Storage Networking Industry Association. All Rights Reserved.

Deduplication: The ZFS Approach

- Three elements:
  - On-disk dedup (delivered)
  - Over-the-wire dedup (delivered)
  - In-core dedup (concept stage)

© 2010 Storage Networking Industry Association. All Rights Reserved.

FS/Volume vs. Pooled Storage

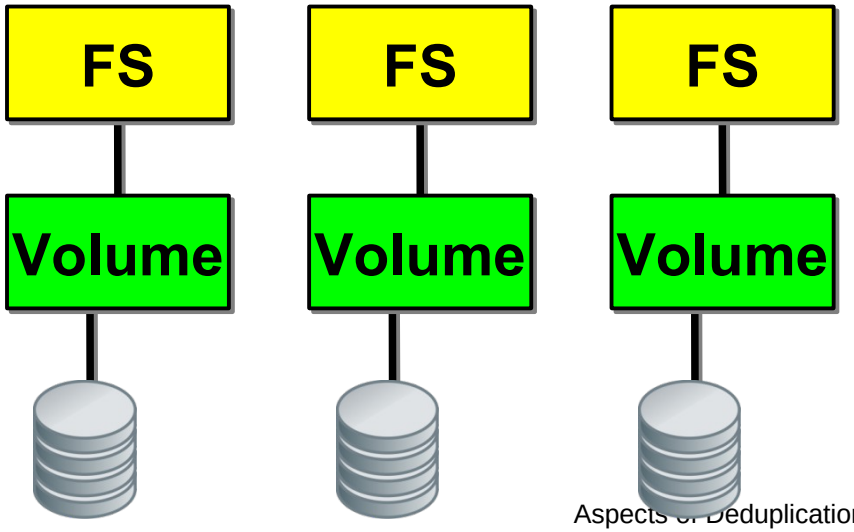
Abstraction: Virtual disk

Partition/volume for each FS

Grow/shrink by hand

Each FS has limited bandwidth

Storage is fragmented, stranded



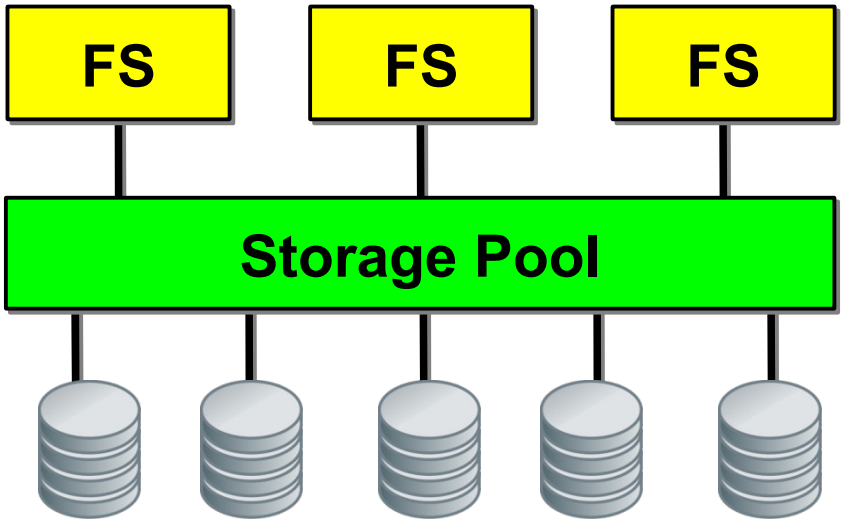
Abstraction: malloc/free

No partitions to manage

Grow/shrink automatically

All bandwidth available

All storage in the pool shared



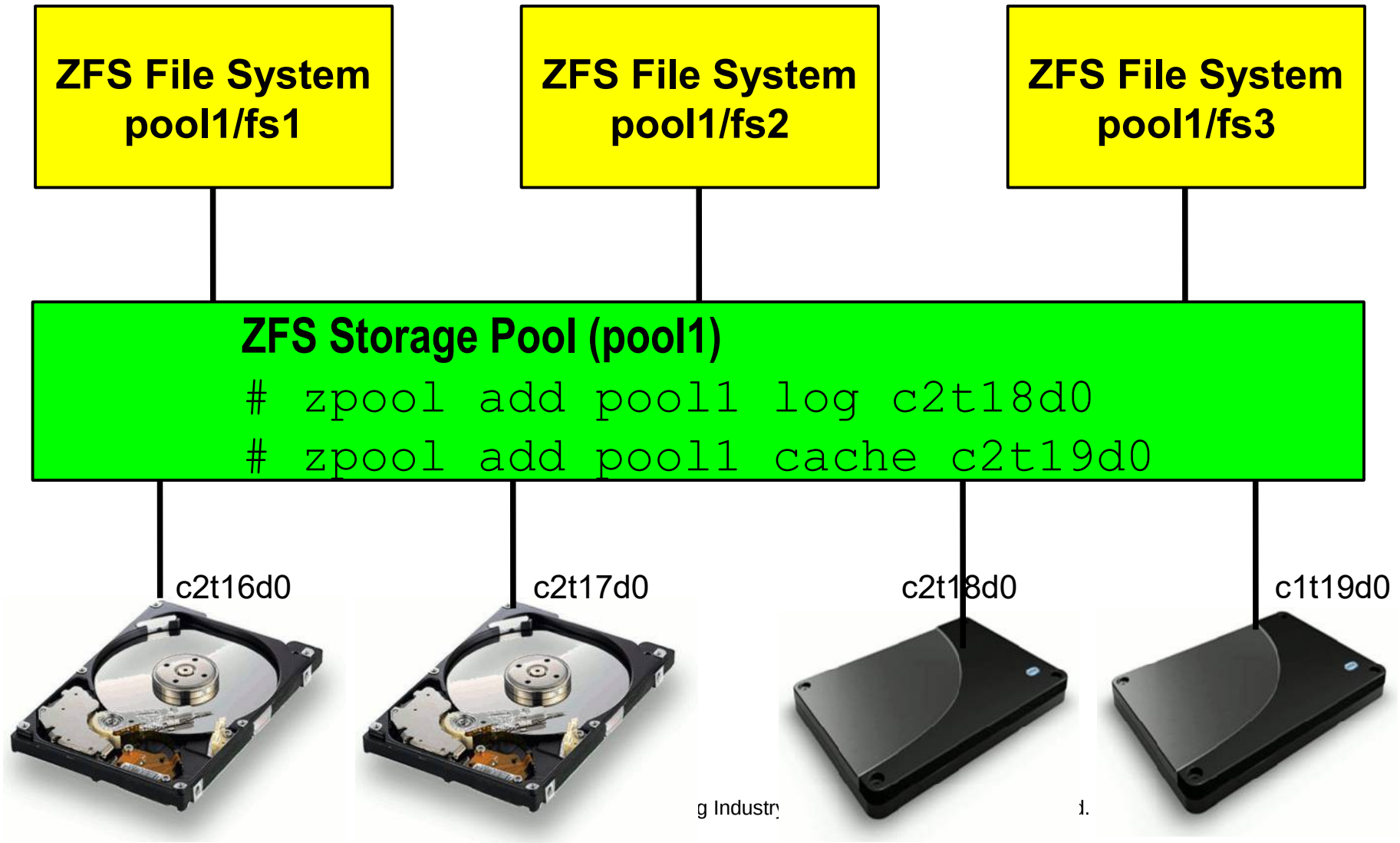
© 2010 Storage Networking Industry Association. All Rights Reserved.

Administrative Interfaces

- Design goals of ZFS dictate simple admin where possible.
- The pool/filesystem model dictates the administrative interface:
  - [zpool stuff here]
  - zfs set dedup=<on | off | checksum>[,verify]
  - zfs get dedup
- This model allows us to deal with mixed mode data stores [yadda more]

© 2010 Storage Networking Industry Association. All Rights Reserved.

ZFS & SSD: Hybrid Storage Pool



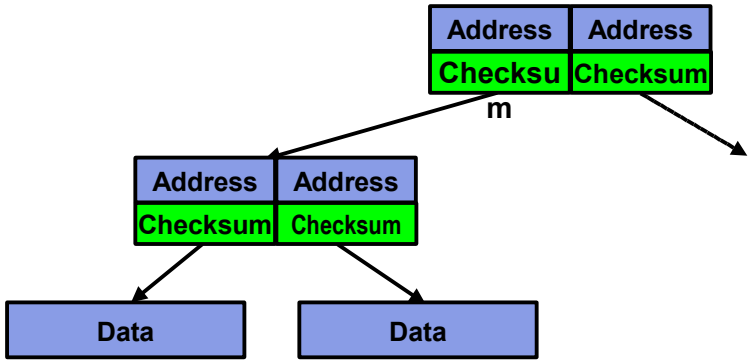
Dedup Table and its Placement

- In memory, on SSD, talk to Mark

© 2010 Storage Networking Industry Association. All Rights Reserved.

ZFS Data Authentication

- Checksum stored in parent block pointer
- Fault isolation between data and checksum
- Entire storage pool is a self-validating Merkle tree
- ZFS validates the entire I/O path
  - DMA parity errors
  - Driver bugs
  - Accidental overwrite
  - Misdirected reads and writes
  - Bit rot
  - Phantom writes



© 2010 Storage Networking Industry Association. All Rights Reserved.

Checksums

- The data validation checksums drive the deduplication table.

- **`zfs set dedup=<on|off|checksum>[,verify]`**
  - The acceptable values for the dedup property are as follows:
  - off (the default)
  - on (see below)
  - on,verify
  - verify
  - sha256
  - sha256,verify
  - fletcher4,verify
  - fletcher2,verify

© 2010 Storage Networking Industry Association. All Rights Reserved.

Ditto Blocks

- Data replication above and beyond RAID
  - Each logical block can have up to three physical blocks
    - Different devices whenever possible
    - Different places on the same device otherwise (e.g. laptop drive)
  - All ZFS metadata 2+ copies
    - Small cost in latency and bandwidth (metadata ≈ 1% of data)
    - Explicitly settable for precious user data
- Detects and corrects silent data corruption
  - In a multi-disk pool, ZFS survives any non-consecutive disk failures
  - In a single-disk pool, ZFS survives loss of up to 1/8 of the platter

© 2010 Storage Networking Industry Association. All Rights Reserved.

Ditto Blocks & Deduplication

- Automatic-ditto data protection
- Eliminates data redundancy concerns associated with deduplication
- Creates an extra copy of the block based on reference count threshold
- Setting the automatic-ditto threshold

```
# zpool set dedupditto=200 tank
```


© 2010 Storage Networking Industry Association. All Rights Reserved.

Variable Sized Block

- Yadda
  - Yadda

© 2010 Storage Networking Industry Association. All Rights Reserved.

ZFS Compression

- The tool for space optimization prior to deduplication.
- Several algorithms available [iterate]
- set and get (compression ratio) via filesystem properties.
- Apllies to data written after property is set and usual YMMV rules apply.

© 2010 Storage Networking Industry Association. All Rights Reserved.

Dedup Integrates with Compression

```
# zdb -DD tank
DDT-sha256-zap-duplicate: 110173 entries, size 295 on disk, 153 in core
DDT-sha256-zap-unique: 302 entries, size 42194 on disk, 52827 in core

DDT histogram (aggregated over all DDTs):

bucket              allocated                       referenced
______   ______________________________   ______________________________
refcnt   blocks   LSIZE   PSIZE   DSIZE   blocks   LSIZE   PSIZE   DSIZE
------   ------   -----   -----   -----   ------   -----   -----   -----
     1      302   7.26M   4.24M   4.24M      302   7.26M   4.24M   4.24M
     2     103K   1.12G    712M    712M     216K   2.64G   1.62G   1.62G
     4    3.11K   30.0M   17.1M   17.1M    14.5K    168M   95.2M   95.2M
     8      503   11.6M   6.16M   6.16M    4.83K    129M   68.9M   68.9M
    16      100   4.22M   1.92M   1.92M    2.14K    101M   45.8M   45.8M
    32      548   65.7M   34.0M   34.0M    22.4K   2.69G   1.40G   1.40G
    64      169   20.8M   11.2M   11.2M    13.8K   1.70G    940M    940M
 Total     108K   1.25G    787M    787M     274K   7.43G   4.15G   4.15G

dedup = 5.40, compress = 1.79, copies = 1.00, dedup * compress / copies = 9.67
```


© 2010 Storage Networking Industry Association. All Rights Reserved.

Sync & Async Deduplication

- Yadda
  - Yadda

© 2010 Storage Networking Industry Association. All Rights Reserved.

Dedup over the wire

- ZFS **`send`** and **`recieve`** example
  - Yadda

© 2010 Storage Networking Industry Association. All Rights Reserved.

In-memory Deduplication

- Yadda
  - Yadda

© 2010 Storage Networking Industry Association. All Rights Reserved.

Q&A / Feedback

- Please send any questions or comments on this presentation to SNIA: add your track reflector here

Many thanks to the following individuals for their contributions to this tutorial.

Name of contributor here	Name of contributor here
Name of contributor here	Name of contributor here
Name of contributor here	Name of contributor here
Name of contributor here	Name of contributor here
Name of contributor here	George Wilson
Name of contributor here	Cindy Swearingen

© 2010 Storage Networking Industry Association. All Rights Reserved.