Introduction to ZFS

What is ZFS?

- ZFS is the next generation enterprise file system and volume manger,
- Robust, scalable and simple to administer.
- Self healing, transactional
- Two components
  - Pool manager
  - File system manager

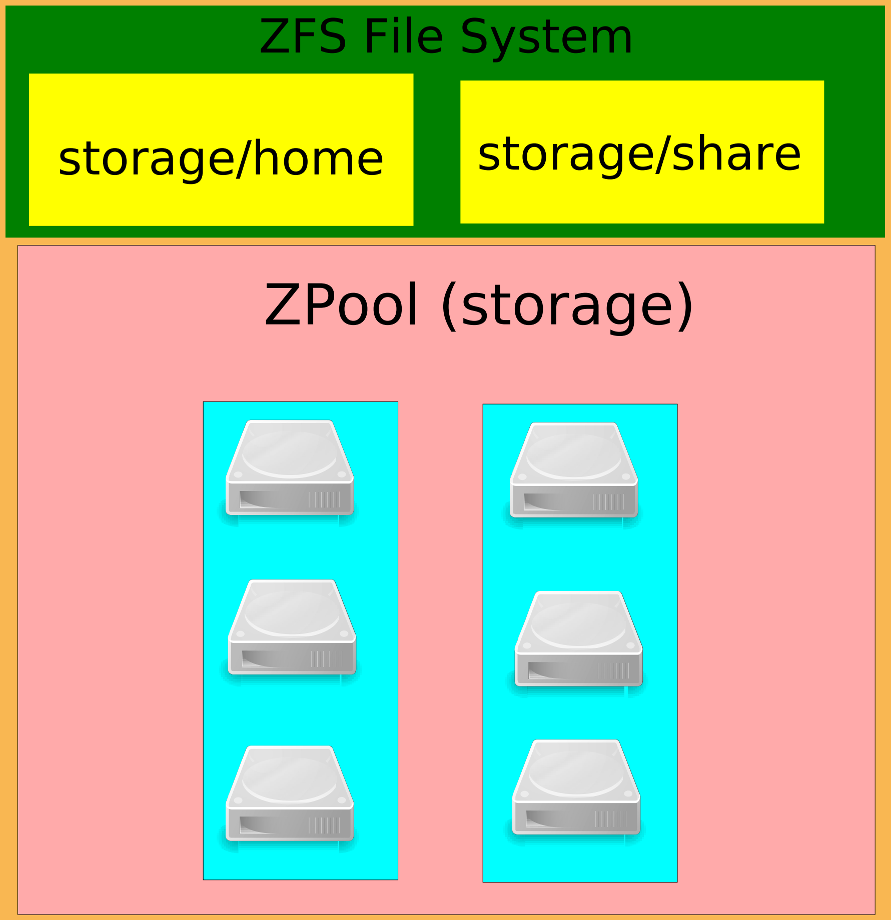
Who Am I?

- Mark Clarke
  - work at Jumping Bean, an solutions integration company,
  - Social Media
    - Twitter - @mxc4
    - G+ - MClarke4@gmail.com
    - LinkedIn
    - http://www.jumpingbean.co.za

Where did it come from?

- Released by Sun Microsystems now owned by Oracle
- Announced September 2004
- Work started in 2001
- Open source - Licensed under the CDDL (Common Development and Distributions License)
- Latest versions:
  - ZFS File System -5/6
  - ZFS Pool Version Number 28/34

ZFS Components



ZFS File System

storage/home

storage/share

ZPool (storage)

ZFS Features

- Combined file system and logical volume manager,
- Features:
  - Protection against data corruption,
  - Snapshots, clones
  - Automatic repair and
  - Continuous integrity checking

ZFS Features

- 128bit file system
  - Can address 1.84 × 10^19 more data than 64 bit system,
  - No practical limit to
    - File size,
    - Directory entries
    - Disk drives

Introduction to ZFS

- Only two commands
  - **zpool** – for creating/managing storage pools
  - **zfs** – for creating/managing file-systems

ZPOOL

ZPool

- Zpool handles the storage pool,
- Responsible for:
  - Data integrity
  - Self healing
  - Check summing
  - Performance
  - Vdev creation, management

ZPool

- Stripes data across vdevs
- ZPOOL components
  - ARC
  - L2ARC
  - ZIL (ZFS Intent Log)
  - COW (copy on writes)
  - Transaction groups

Why Do we Have Data Integrity Issues?

- Data faults/corruption occur because:
  - Bit rot,
  - Current spikes,
  - Firmware bugs
  - Phantom writes,
  - Misdirected read/writes
  - Raid “Write holes”

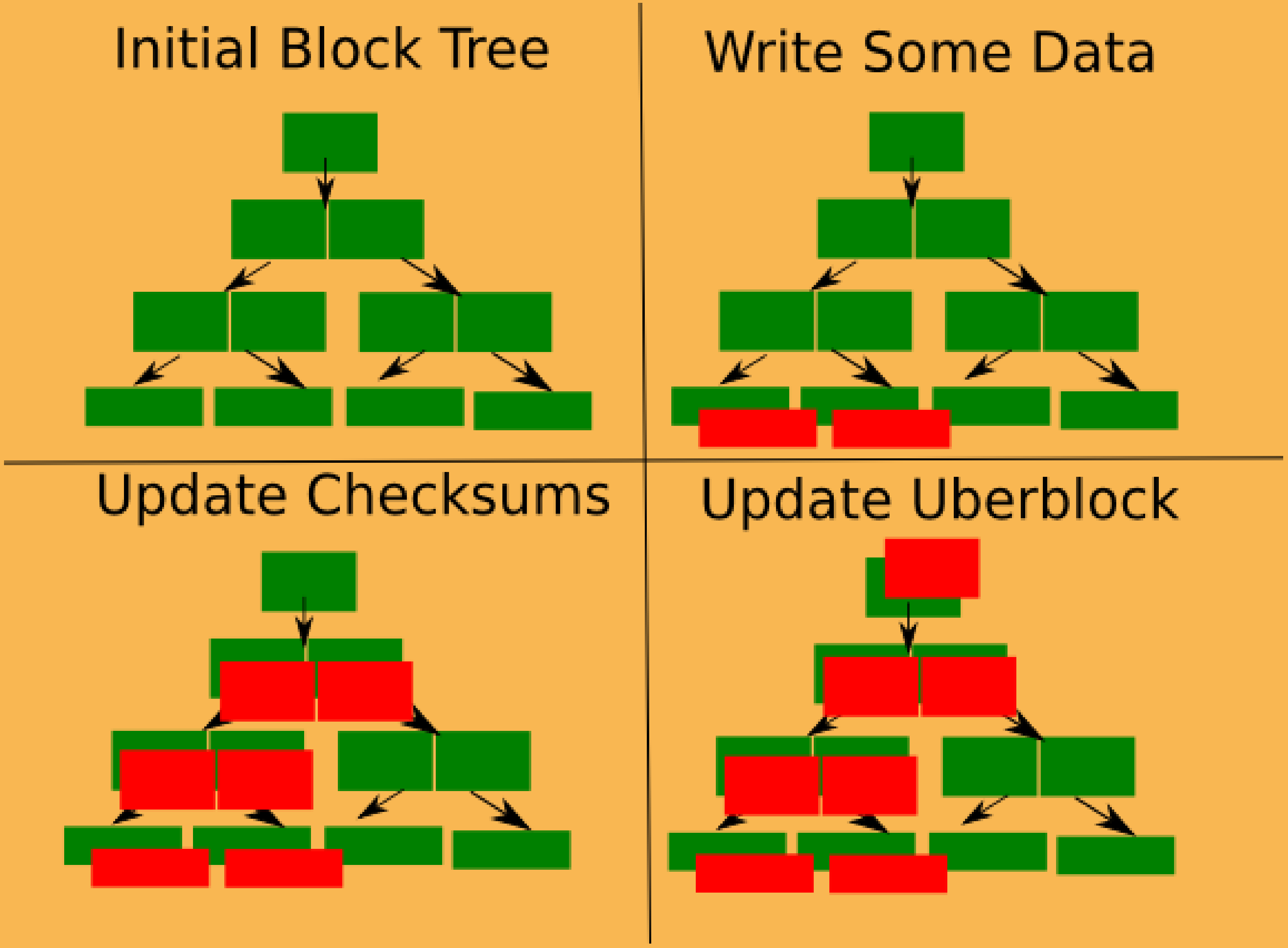
Data Integrity

- Silent data corruption
  - Errors undetected by firmware and/or operating system
  - Netapp study found 1 in 90 SATA drives had silent software corruption,
  - Faster disks/raid controllers + larger capacity = problem
  - Jeff Bonwick estimates silent corruption every 15 minutes at Greenplum

How ZFS handles Integrity

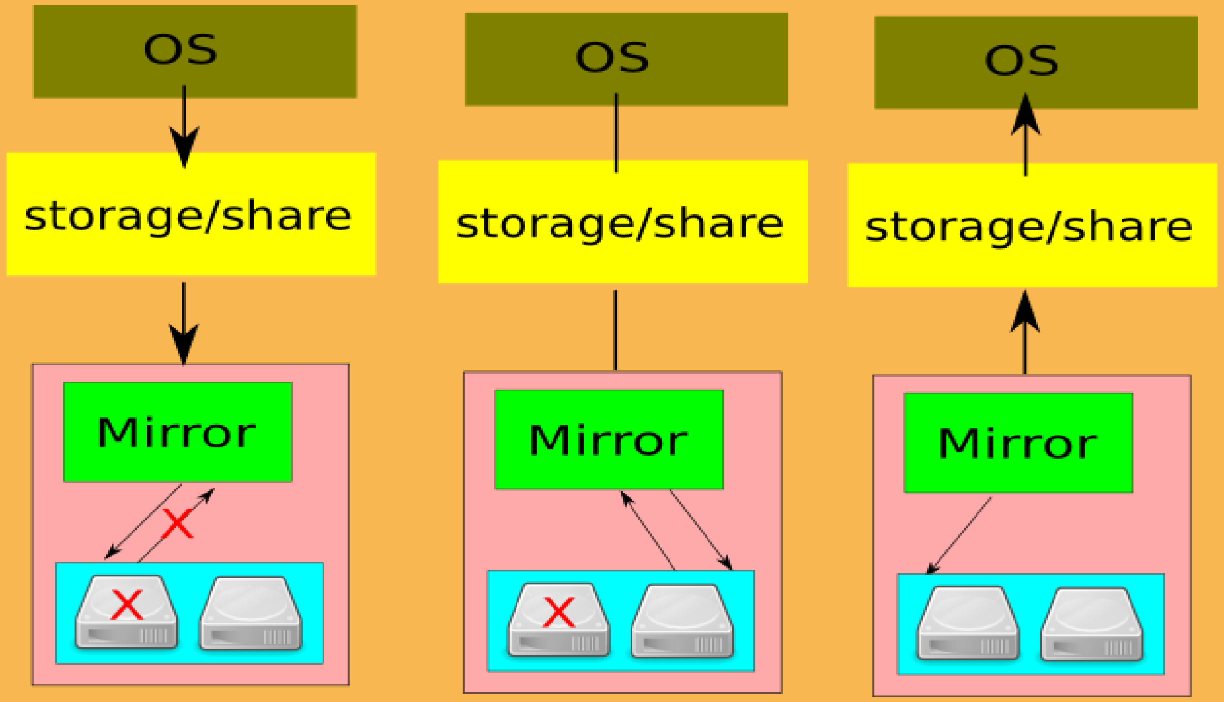
- ZFS Data Integrity handle by COW transactional data writes
  - Uses Hash Tree (Merkle Tree)
    - Each block checksummed, stored in pointer to block,
    - Each pointer checksums stored in pointer, etc – up to root block,
    - Uber block has check sum,
    - Checksum compared on block access,

ZFS Copy on Write



ZFS Auto Healing

- On check sum check error , zpool will rebuild data from redundant storage and repair bad blocks,



Zpool Vdevs

- Zpool works with virtual devices (vdevs)
- Vdevs can be:
  - Single disks
  - N-way mirrors
  - Raidz1,
  - Raidz2,
  - RaidDz3

ZPool Vdevs

- RAID-Z
  - RAID-Z1 - one disk failure ~ Raid 5
  - RAID-Z2 - up to two disk failures ~ Raid 6
  - Raid-Z3 – allows three disk failures
  - No write whole problem due to transaction group/cow

Zpool - Storage Design

- Zpool uses dynamic striping across vdevs,
- Can mix any combination of vdevs
- Design storage pool with redundancy, performance and maintainability in mind.

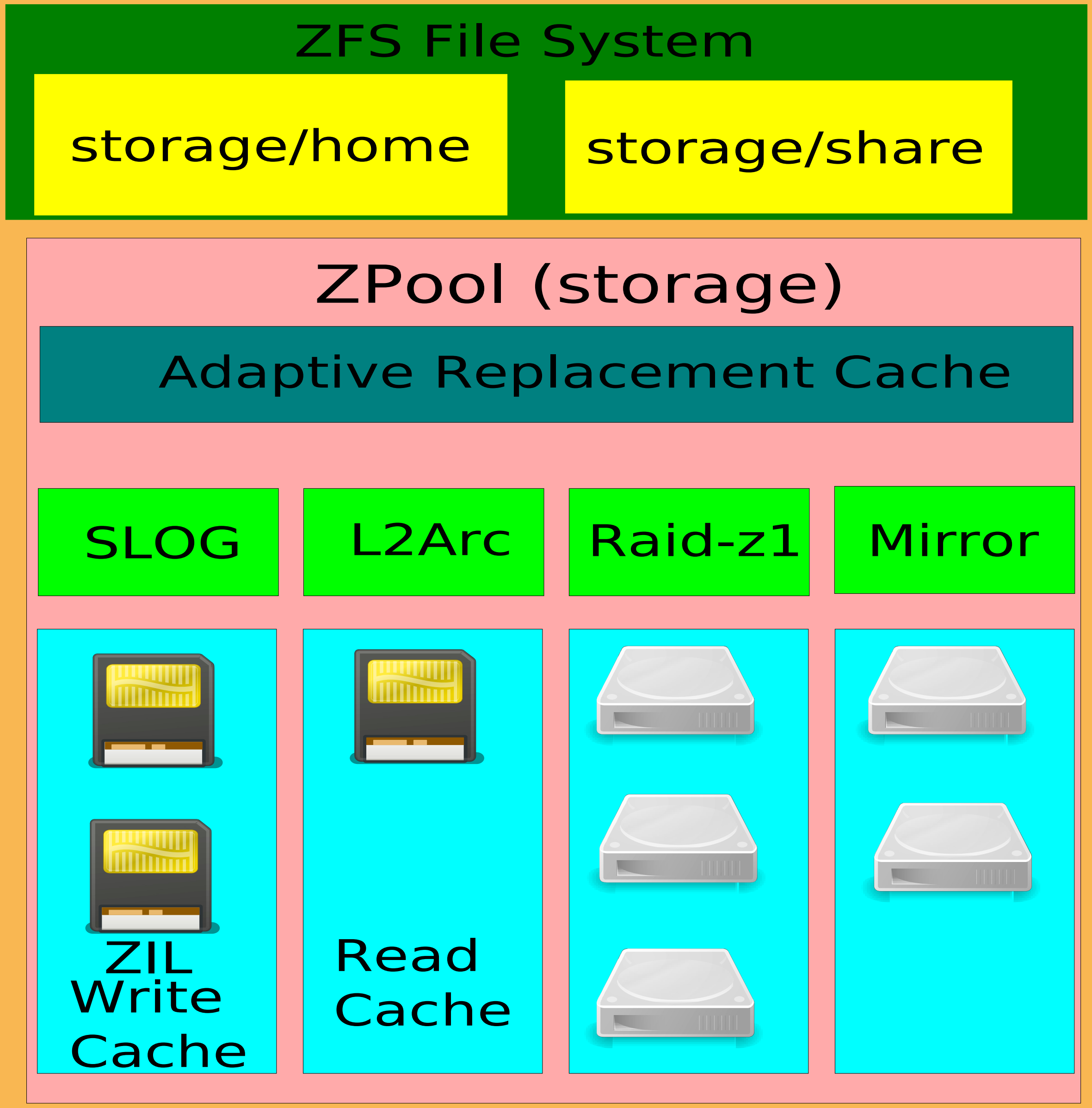
Zpool - ARC

- ARC is the adaptive replacement cache
- In memory cache,
- Zpool will use all memory -1G,
- Frees up memory when requested by other apps
- More ram = better performance
- ARC uses:
  - MRU,
  - MFU
  - MRU Ghosts – evicted pages
  - MFU Ghost – evicted pages

ZPOOL – Hybrid Pools

- Hybrid storage pools
  - Use SSD for
    - ZIL (write cache)
    - L2Arc cache (read cache)
  - Use disk for mass storage capacity
    - MLC – multi layer cell (L2Arc)
    - SLC – single layer cell (SLOG)

ZFS Hybrid Pools



ZFS File System

storage/home

storage/share

ZPool (storage)

Adaptive Replacement Cache

SLOG

L2Arc

Raid-z1

Mirror

ZIL
Write
Cache

Read
Cache

Zpool Demo

ZFS Datasets

What does ZFS File System Provide?

- ZFS file system provides
  - Compression
  - Encryption
  - Shares
  - De-duplication,
  - Quotas
  - Reservation,
  - Snapshot,
  - Clone,
  - Properties

ZFS - File Systems

- ZFS creates datasets,
- ZFS can create and mount file system with a single command,
- File systems mounted by default under pool name,
- Block devices can also be create on a Zpool and formatted with ext4 etc,

ZFS Properties

- Using ZFS properties one can:
  - Enable compression,
  - Enable CIFs/NFSv4 shares
  - Change mount point
  - Enable de-duplication – requires lots of memory

ZFS Snapshots & Clones

- Snapshot are efficient and cheap to create,
- Can rollback to snapshots easily
- Snapshots can be access via .zfs hidden directory
- Snapshots read only,
- Clones read & write

ZFS Send/Receive

- ZFS stream snapshots over stdin/stdout,
- ZFS send/receive can be done over WAN, no special hardware required,
- Can send/receive incremental snapshots

ZFS Dataset Demo

Q & A

The End