Quo vadis ZFS

Martin Matuška
mm@FreeBSD.org

VX Solutions s. r. o.

EuroBSDcon 2010
10.10.2010

About this presentation

This presentation will give a very brief introduction into ZFS and try to answer the following questions:

- What are the newest features in ZFS?
- What operating systems do ship ZFS?
- Can we use, distribute and develop ZFS?
  Are there any legal issues?
- How does the future of the ZFS development look like?

Introduction

New features

Operating systems

Legal issues

Future of ZFS

Introduction

- What is ZFS
- ZFS history
- Main ZFS objects
- ZFS limits

What is ZFS?

ZFS is the ”Zettabyte filesystem”

Original ZFS features by design:

- pooled storage (integrated volume manager)
- transactional semantics (copy on write)
- checksums and self-healing (scrub, resilver)
- scalability
- instant snapshots and clones
- dataset compression (lzjb, gzip)
- simplified delegable administration

ZFS history

- 2005/10: OpenSolaris - ZFS introduced in revision 789
- 2005/12: Solaris Express - first release
- 2006/06: Solaris 10 update 6 - pool version 3
- 2007/04: FreeBSD (CURRENT) - pool version 6
- 2008/11: FreeBSD (CURRENT) - pool version 13
- 2009/10: Solaris 10 update 8 - pool version 15
- 2010/07: FreeBSD (CURRENT) - pool version 15
- 2010/08: OpenSolaris - closed, last revision 13149 (v28)
- 2010/09: Solaris 10 update 9 - pool version 22 (no dedup)

Main ZFS objects

The two main ZFS objects are:

- pool
- dataset

ZFS pool

A ZFS pool is a storage object consisting of virtual devices.
'vdevs' can be:

- disk (partition, GEOM object, ...)
- file (experimental purposes)
- mirror (groups two or more vdevs)
- raidz, raidz2, raidz3 (single to triple parity RAIDZ)
- spare (pseudo-vdev for hot spares)
- log (separate ZIL device, may not be raidz)
- cache (L2 cache, may not be mirror or raidz)

ZFS dataset

Each ZFS pool contains ZFS datasets.
ZFS dataset is a generic name for:

- file system (posix layer)
- volume (virtual block device)
- snapshot (read-only copy of filesystem or volume)
- clone (filesystem with initial contents of a snapshot)

ZFS limits

What are the limits of ZFS?

- ZFS is a 128-bit filesystem
- Maximum pool size: 256 quadrillion zettabytes ($= 256 * 10^{36}$ bytes)
- Maximum filesystem/file/attribute size: 16 exabytes
- Maximum pools/filesystems/snapshots: 2^{64}

New features

- ZFS pool and filesystem versioning
- New ZFS pool versions (v14-v28)
- Other new user-visible features

ZFS pool and filesystem versioning

- ZFS pools and filesystems have a version number
- incompatible structural changes lead to a version increase
- ZFS is backwards compatible
- forward compatibility is NOT provided
- version downgrade is NOT possible
- latest ZFS pool version: 28
- latest ZFS filesystem version: 5

New ZFS pool versions

Selected ZFS pool version upgrades between v14 and v28:

- version 15: user/group space accounting
- version 17: triple parity RAID-Z
- version 18: snapshot user holds
- version 19: log device removal
- version 21: deduplication
- version 22: zfs receive properties
- version 24: system attribute support
- version 25: improved scrub statistics

Other new user-visible features

Other important new features not touching pool versions:

- device autoexpansion (post-v16)
- ZFS pool recovery (post-v19)
- deduplication of zfs send streams (post-v21)
- splitting mirrors into separate pools (post-v22)
- ZIL synchronicity setting for datasets (post-v24)
- diff between snapshots (post-v28)

Operating systems

- Distributions based on Solaris/OpenSolaris
- Other operating systems / distributions

Systems based on Solaris/OpenSolaris

- OpenSolaris (discontinued)
- Oracle Solaris 10
- Nexenta Core
- OpenIndiana
- SchilliX
- Belenix

OpenSolaris

- The source of ZFS code for everyone else
- ZFS introduced on 31-Oct-2005 in revision 789
- Last release: OpenSolaris 0906 (Jun-2009)
- Last development release: build 134 (Mar-2010)
- Last public commit to ZFS on 18-Aug-2010 (rev 13147)
- wiki documentation not updated anymore
- Future: project discontinued
- Free successor: Illumos (releases: OpenIndiana)

Oracle Solaris

- Commercial OS - Licence Required
- ZFS introduced in Solaris 10 update 6 (Jun-2006)
- Latest release: update 9 (Sep-2010) with ZFS v22 (no dedup)
- Oracle® Solaris ZFS Administration Guide
- ”Oracle tells the future”

Nexenta Core

- OpenSolaris with debian package management
- Latest release: 3.0.1 (Sep-2010) with ZFS v26
- Compatible with OpenSolaris
- Quite stable, but weak documention
- Future: cooperation with Illumos

OpenIndiana, Belenix, SchilliX

- all OpenSolaris distributions
- OpenIndiana: ”continuation” of OpenSolaris (Illumos-based)
  Latest release: dev build 147 (Sep-2010)
- BeleniX: Indian LiveCD distribution
  Latest release: 0.8 beta 1
- SchilliX: German distribution (now Illumos-based)
  Maintained by Jörg Schilling and Fabian Otto
  (Fraunhofer-Institut für Offene Kommunikationssysteme)
  Latest release: 0.7.2 (Sep-2010)

Other Systems

ZFS originates from OpenSolaris - everybody elese has to port it.

- FreeBSD
- NetBSD
- MacOS X
- Linux (FUSE or standalone module)
- Debian (GNU/kFreeBSD) - "just" a distribution

FreeBSD

- ZFS introduced in Apr-2007 (pool version 6)
- Latest release: pool version 14 in 8.1-RELEASE
- Current state: pool version 15 in 9-CURRENT and 8-STABLE + some backported improvements (L2ARC, Metaslabs, ACL cache, ...)
- Testing: pjd's v28 patch in mailing lists
- Documentation: wiki, manual pages
- Support: mailing lists, forums
- Future: cooperation with Illumos?

NetBSD

- ZFS port in GSOC 2009 by Adam Hamšík (haad@netbsd.org)
- Integrated into NetBSD sources (HEAD branch)
- Works only on i386 and amd64
- Some functions not yet working (snapshots, permissions)
- Some bugs still need fixing (vnode reclaiming, ...)

MacOS X

- MacOS X ZFS project has been closed by Apple (Oct-2009)
- Dustin Sallings: mac-zfs on googlecode and github, installer available
- Beta quality

Linux

- ZFS-fuse project
  Version 0.6.9 - ZFS pool v23
- ZFS kernel modules by Brian Behlendorf
  Version 0.5.1 - pool v28, no ZFS Posix Layer (ZPL)
- ZFS Posix Layer (ZPL) from KQ Infotech
  Based on Brian Behlendorf's 0.4.7, ZFS pool v18, beta
- KQ Infotech (Anand Mitra) working on ZPL for 0.5.1

Legal issues

This section will cover the following topics:

- CDDL License
- Patent claims (Netapp lawsuit)

CDDL License

ZFS source code is licensed under the
Common Development and Distribution License (CDDL)

- based on Mozilla Public License (MPL)
- GPL incompatible
- if binaries are distributed, source code must be distributed
- but only from "Covered Software" = original + modifications
- if part of a "Larger Work", CDDL clauses must not be violated
- modifications must be CDDL, author ("Contributor") needs to be disclosed
- terminates if any patent infringements against author or contributors

Patent claims

There was a Lawsuit between Netapp and Sun Microsystems. Netapp claims included the following three important U.S. patents:

- 5,819,292 (copy on write) - almost completely nullified (final)
- 7,174,352 (snapshots) - almost completely nullified but non-final
- 6,857,001 (writable snapshots) - reexamination started

The lawsuit was settled in Sep-2010, both parties dropped their charges. Details are disclosed.

Future of ZFS

This section will cover the following topics:

- ZFS development at Oracle
- The Illumos Project
- FreeBSD ZFS developers
- Porting ZFS v15 to FreeBSD
- Other important backported features
- Ongoing ZFS work at FreeBSD

ZFS development at Oracle

A leaked internal memo from Oracle claims the following:

- Oracle will continue to develop ZFS but not in public
- ZFS code will remain CDDL licensed
- CDDL source code will be published with Solaris releases
- development sources will be available only to industry partners via OTN (Oracle Technology Network)

The Illumos Project

ILLUMOS

- project started by several former OpenSolaris developers
- sponsored and supported by Nexenta
- goal: provide a free ON source (and replace closed parts)
- distributions to build on Illumos: Nexenta, Belenix, Schillix

Where to get code for closed parts?
FreeBSD! (sed, tr, em, msk)

FreeBSD ZFS developers

- Pawel Jakub Dawidek (pjd@FreeBSD.org) (maintainer)
- Andriy Gapon (avg@FreeBSD.org)
- Xin Li (delphij@FreeBSD.org)
- Martin Matuska (mm@FreeBSD.org)
- External developers committing into p4

Porting ZFS v15 to FreeBSD

- Starting point: ZFS v14
- Evaluation of changes imported by Solaris 10 U8+
- Syncing with head work by pjd@
- Resolving tools and module compatibility problems
- Patch for public testing
- Import into -CURRENT (MFC to -STABLE)

Other important backported features

Features backported together with v15:

- Metaslab code rewrite (post v22)
- stat(), rrwlock and ACL caching speedup (post v16)

Ongoing ZFS work at FreeBSD

- Current state: pool version 15 in 8-STABLE
- Backported improvements from higher versions: L2ARC, Metaslabs, ACL caching, ...
- Testing: pool version 28 (patch by pjd@)
- Upgrade problems (tools and module incompatibility)
- Improving ARC and VM page daemon interaction (avg@)
- Cooperation with Illumos?

Thank you for your attention!

solutions

http://blog.vx.sk
http://www.vx.sk